Data Analytics, Decision Support and Upholding Business Ethics in Chinese Automobile Market

Ruixin Ding, PhD, Management

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADAM SMITH BUSINESS SCHOOL

COLLEGE OF SOCIAL SCIENCE

10 2024

Abstract

After more than 70 years of development, Chinese automobile market has gradually become the world's largest automobile market. Especially after 2010, the rapid development of new energy vehicles has significantly changed the market competition pattern. Under the impact of new brands and new car models, all automakers have to rethink the applicability of the original business theory in the new environment in order to maintain the original market advantage or obtain new market competitiveness. At the same time, with the prevalence of social media, online reviews have gradually become an important source of information for consumers to make decisions and for enterprises to analyse consumer demands, but it has also gradually derived the unethical business behaviours of manipulating fake reviews. In addition, the rise and application of artificial intelligence technologies such as machine learning and deep learning have proved their important potential in business analytics. This research aims to address the above business challenges by applying suitable machine learning methods, and to provide theoretical support and methodological practice for business decision-making and ethical maintenance. The overall research consists of three subdivision studies. In the first study, it is mainly discussed what important business data are available in the automobile market and their potential applications, and finally a comprehensive automotive industry dataset is created. The second study investigates the impacts and changes of first-mover advantages and country-of-origin effects in the Chinese automotive market through a novel sales analysis framework. It demonstrates the positive impact of the first-mover advantage and country-of-origin effects on sales performance, as well as that the first-mover advantage can be broken by innovative latecomers. In the third study, a large number of fake reviews were identified by the

detection method based on BERT and PU-Learning we proposed, thus completing the study on the timing of manipulation of positive fake reviews. It shows the correlation between the timing of manipulation and car sales, brand sales, market size, and the duration of brands and models in the market. The whole research not only extends theories related to the first-mover advantage, country-of-origin effects, and the timing of manipulating fake reviews, but also demonstrates the superior performance of machine learning methods in business analytics. In addition to providing decision-making support for automakers' corporate strategies and marketing strategies, this research has important implications in formulating industrial policies, protecting consumer interests, enhancing the credibility of online review platforms, and maintaining a fair business environment.

Contents

List of Tables

List of Figures

Acknowledgements

First, I would like to thank my supervisors, Bowei Chen and James M. Wilson, for their guidance during my overall research. Secondly, I would like to thank Zhi Yan and Yufei Huang for their valuable suggestions in the first study and Sha Zhang for her suggestions in the second study during my PhD. Additionally, I would like to thank my friends for their assistance in the task of labeling fake reviews. Finally, I would also like to thank my parents for their financial support and my wife for her companionship during my PhD.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution

---

Ruixin Ding

Chapter 1

Introduction

1.1 Background

1.1.1 Automotive industry

As a key pillar of the global economy, the automotive industry has risen to become one of the most crucial sectors worldwide. Its significance is not only evident in its direct economic impact, but also in its broader contributions to various other industries and society at large (Rezaeinejad 2021). The production of automobiles requires the synergistic collaboration of multiple sectors, including steel, battery, internet-related technologies, rubber, and many others, meaning that the growth of the automotive industry drives advancement in these interrelated fields. Simultaneously, the entire automotive supply chain generates a vast number of job opportunities across the globe, which not only fuels global economic growth but also plays a vital role in maintaining social

stability. Additionally, as a major commodity in international trade, the automotive industry is indispensable to the export economies of numerous countries (Bloomfield 2017). Therefore, the automotive industry is one that demands comprehensive and long-term study and analysis.

Since the emergence of the first modern car in Europe in 1886, both the form of automobiles and the automotive industry itself have undergone tremendous changes. Driven by globalization, this industry, initially dominated by Europe and the United States, has gradually evolved into a global sector, with Asian countries such as Japan, South Korea, and China now holding significant positions in the global automotive market. Undoubtedly, competition in the global automotive industry is intensifying, and the pressures surrounding both research and development as well as car sales are increasing. Furthermore, the era of a singular technological focus in the automotive sector has come to an end, and technological diversity will be an important feature of the industry (Wells 2010). For example, the gradual maturation of technologies such as Advanced Driver Assistance Systems (ADAS) and vehicle-to-X communication systems has positioned smart vehicles as the emerging direction for the future development of the automotive industry (Cao et al. 2022). At the same time, given the contribution of new energy vehicles (NEVs) in mitigating environmental pressures, many countries are actively developing and promoting them. Moreover, amidst social development, industrial progress, and technological change, shifts in consumer demand and behavior warrant close attention. Millennials, in particular, have become the primary demographic for car purchases. Having grown up during the digital technology revolution and exposed to a wider range of information, this group has developed unique consumption attitudes and behaviors, characterized by distinct changes in their mindset and lifestyle (Lantos 2014). It is evident that traditional marketing strategies and approaches can no longer meet the demands of the industry.

While these factors pose significant challenges for automakers worldwide, they also present unprecedented growth opportunities for those able to adapt through timely changes in business models, operational management, and marketing strategies. The key to seizing these opportunities lies in accurate analysis of market shifts and consumer behavior within the automobile market.

1.1.2 Chinese automotive industry and automobile market

After more than 70 years of development, the Chinese automotive industry and automobile market occupies a pivotal role in the global automotive industry's evolution. China has emerged as the world's largest automobile market in terms of both production and sales (Wang et al. 2022), while also establishing itself as a key global automobile exporter. In 2024, car sales in the Chinese automobile market reached 25,580,000, while the number of exported vehicles hit a record high of 5,859,000. Clearly, with its profound background and abundant industry data, Chinese automobile market has become an indispensable and integral component of global automotive industry research.

The development of the Chinese automotive industry and automobile market is often defined as four phases: namely, start-up phase, growing phase, prosperity phase and stationary phase (Kenworthy et al. 2015). Firstly, the period from 1953 to 1978 marks the start-up phase, which primarily represents the origin of the Chinese automotive industry. In China's first Five-Year Plan, from 1953 to 1957, the government explicitly designated the automotive industry as one of the key areas for industrial revitalization. In 1953, with technical support from the Soviet Union, the establishment of China's first automobile manufacturing plant, the First Automobile Works (FAW), marked the beginning of the Chinese automotive industry (Yang et al. 2017). Subsequently, un-

der the influence of international circumstances and the loss of corresponding support, the Chinese automotive industry entered a phase of self-reliance. Through relentless efforts, several automakers were established, including the Second Automobile Works, China National Heavy Duty Truck Group, Guangzhou Automobile Company, Shanghai Automotive Industry Corporation, and Beijing Automotive Industry Corporation. However, all these automakers were directly controlled by the government. Initially, the vehicles produced by these companies were primarily trucks, with a limited number of passenger cars mainly supplied for government use. During this stage, production of most car models relied on imitating or replicating the mature car models of foreign brands, due to insufficient funding, a lack of automotive manufacturing experience, outdated industrial technology, and limited production capacity (Yuan and Broggi 2023). At the same time, political movements such as the Cultural Revolution that emerged during this phase severely disrupted the normal operation of the automotive industry, affected the development of automotive production technology, and hindered the training of relevant technical personnel (Field 1986). Additionally, starting in 1958, Chinese automakers gradually began producing passenger cars, marking the beginning of the development of passenger vehicles in the Chinese automobile market. However, the annual production of passenger cars was initially very low; it wasn't until 1973 that the annual output surpassed 1,000 vehicles. Even by 1978, the annual production had just exceeded 2,500 units. This undoubtedly severely restricted the development and proliferation of the automotive industry in China (Chu 2011). However, during the 25 years of overall development in the automotive industry, the accumulated experience, the gradually enriched range of automotive products, and the increasing infrastructure laid a solid foundation for the development of subsequent stages.

Secondly, the growing phase refers to the period from 1979 to 2000, with the establishment and success of joint ventures being the key industrial feature of this stage (Kenworthy et al. 2015). In 1978, the Chinese government issued a series of reform and opening up policies to promote economic development, which significantly reduced

the policy barriers for foreign investors to enter China, promoted the liberalization of foreign investment and accelerated the process of foreign investment in China's automobile industry (Luo and Zhi 2019). At this time, as a key industry supported by the Chinese government, the government encouraged domestic automotive manufacturers to establish joint ventures with advanced foreign automobile companies. This strategy of "exchanging market access for technology" accelerated the development of Chinese automotive industry (Li et al. 2016). After several years of negotiations and operations, many well-known foreign automotive manufacturers established strong cooperative relationships with Chinese companies. In 1983, the first joint venture, Beijing Jeep Corporation, was founded in China as a partnership between Beijing Automotive Industry Corporation and American Motors Corporation. This marked the beginning of joint-venture vehicle production and sales in China. The production process involved importing parts from American Motors and assembling the vehicles in China (Aiello 1991). Subsequently, Volkswagen AG established China's second joint-venture automobile company, Shanghai Volkswagen, in partnership with Shanghai Automotive Industry Corporation, beginning vehicle production in 1985. In 1991, Volkswagen further expanded its presence in China by forming a second joint venture with First Automobile Works. Through these partnerships, Volkswagen's joint ventures gained access to essential resources, such as advanced industrial equipment, technical support documents, and production blueprints, significantly boosting local vehicle production capabilities. The implementation of Volkswagen's dual-brand strategy in China gradually cemented its dominant position (Tang 2012; Zhao 2006). At the same time, Peugeot, Citroen, Honda, Buick and other auto brands have gradually entered the Chinese auto market in the form of joint ventures, which has enriched the choice of market products (Murray 1999; Nieuwenhuis and Lin 2015). In this process, the emergence of numerous automobile brands and models has lowered the threshold for automobile purchase, stimulated market consumer demand, enhanced the production capacity of Chinese automobile manufacturers, and opened a new chapter in automobile consumption.

Thirdly, from 2001 to 2010 was the prosperity phase, during which sales, production, and industry profit margins experienced continuous and significant growth. In particular, a more obvious change is that the number of domestic brands has rapidly increased from 14 to 180, showing a strong development trend (Li et al. 2016). Although the establishment of joint ventures in the previous stage solves technical problems for the automotive industry, it also reduces the market competitiveness of domestic brands, inhibits the technological innovation of domestic companies and hinders the development of domestic brands (Howell 2018). In 2001, China's official entry into the World Trade Organization (WTO) marked a new phase of market openness. This facilitated broader cooperation between Chinese automobile companies and foreign enterprises, but it also heightened the government's concerns about the competitiveness of domestic automotive products (Harwit 2001). Fortunately, during this phase, several positive changes within the automotive industry provided significant growth opportunities for domestic Chinese automakers. The reform of automotive industry policies lowered the entry barriers for domestic manufacturers. Increased market competition, falling car prices, rising household incomes, and government encouragement of car consumption stimulated consumer demand, rapidly expanding the passenger car market in China. Car sales also rapidly increased from around 2,300,000 in 2001 to 18,000,000 in 2010. Meanwhile, the influx of capital, the growth of supporting industries, the acceleration of infrastructure development, and the improvement of automotive services offered essential support for the establishment and growth of domestic automotive companies (Luo 2005). For Chinese domestic automakers, many experienced rapid growth by employing strategies such as reverse engineering, acquiring overseas automotive manufacturing assets, innovating product architectures, independently developing core components, implementing low-price strategies, and improving product services. These efforts gradually earned consumer recognition, and their market share increased from less than 10% to over 30% (Liu and Yeung 2008; Balcet et al. 2012; Chen and Kodono 2012).

Fourth, previous studies have defined the period after 2010 as the stationary phase. However, since this definition was based on a study from 2015, it is evidently insufficient to accurately describe the changes in Chinese automotive industry and market from 2011 to 2024. The most significant change during this phase has been the rapid development of NEVs, making transformation phase a more fitting characterization of this period. Facing the pressure of environmental pollution and the goal of carbon neutrality, NEVs have been paid attention to by many countries (Su et al. 2021). Although NEVs have been listed as a key research project by the Chinese government since the 1990s, only a small number of NEVs have gradually entered the market from 2006 to 2010. It was not until 2010 that the government listed the NEV industry as a strategic emerging industry, which accelerated the speed of NEVs entering the market (Gong et al. 2013). The popularity of NEVs in Chinese automobile market cannot be separated from the support of government policies. In China, the policies to promote the development of NEVs can be divided into guiding policies, supportive policies and normative policies (Dong and Liu 2020). For consumers, the government is committed to promoting low-carbon concepts and improving public awareness of environmental protection to guide consumers to buy new energy models. At the same time, the Chinese government has implemented supportive policies for consumers, such as exempting purchase tax and providing subsidies for NEVs. This has significantly enhanced consumers' purchase intention and expanded the market demand for NEVs (Wu et al. 2022; Ma et al. 2017). For NEV enterprises, the government has mainly implemented a series of supportive and normative policies, such as financial incentives, financial subsidies, technical support, and revision of industry standards (Qu and Li 2019; Yu et al. 2020). This motivates the R&D behavior of automobile manufacturers in the field of NEVs and accelerates the technological innovation of the NEV industry (Jiang and Xu 2023; Shao et al. 2021). With the rapid rise of the NEV industry, a large number of emerging brands such as XPeng and Li Auto have emerged in the Chinese market. Meanwhile, Tesla has entered China as a wholly foreign-owned brand, manufacturing and selling NEVs in China. This has transformed the operational and management methods of the traditional automotive industry (Xie et al. 2023). More importantly, in the field

Table 1.1: Sales statistics in Chinese automotive industry from 2011 to 2024.

Year	Car sales	Domestic sales	Export sales	NEV sales	Proportion of NEV sales
2011	18505100	17691100	814000	8200	0.04%
2012	19306400	18250300	1056100	12800	0.07%
2013	21984100	21066800	917300	17600	0.08%
2014	23491900	22581500	910400	74800	0.32%
2015	24597600	23869400	728200	331100	1.35%
2016	28028200	27319900	708300	507000	1.81%
2017	28878900	27988000	890900	777000	2.69%
2018	28080600	27039900	1040700	1256200	4.47%
2019	25769000	24744800	1024200	1206000	4.68%
2020	25311000	24316100	994900	1367300	5.40%
2021	26275000	24259800	2015200	3520500	13.40%
2022	26864000	23753400	3110600	6886600	25.64%
2023	30094000	25184000	4910000	9495200	31.55%
2024	31436000	25577000	5859000	12866000	40.93%

of NEVs, these emerging brands have broken consumers' preference for traditional car brands from developed countries, such as BMW and Mercedes-Benz (Hu and Yuan 2018). As a result, Chinese NEV industry has achieved a leading position globally in terms of technology, production and sales.

Based on historical sales data (see Table 1.1) released by the China Association of Automobile Manufacturers, the sales of NEVs have made a significant leap during this period. From a mere 0.04% market share in 2011, NEVs have experienced over a decade of continuous growth, reaching 40.93% in 2024. This underscores their success in Chinese automotive industry and market. Although the domestic market has yet to surpass its peak sales of 27,988,000 in 2017, it has still shown substantial growth compared to the highest sales figure of 17,511,300 recorded during the prosperity phase in 2010. Moreover, domestically produced vehicles have also achieved rapid sales growth in this period, now accounting for 65.2% of passenger car sales. Clearly, through this stage of development, Chinese automotive industry and market have made significant progress in four key areas: market scale, NEVs, domestic brand vehicles, and automobile exports. This progress has further solidified China's unparalleled significance in the global automotive landscape.

Obviously, the significance of the Chinese automobile market on a global scale is not solely limited to its large market size; it holds critical research and business value in multiple aspects. First, China's vast base of automobile consumers provides an abundance of sales data and consumer behaviour data. This benefits global automakers by allowing them to assess product competitiveness, evaluate brand performance in the market, gather valuable consumer feedback, and forecast future consumption trends in the automobile industry, thus enabling the formulation of more strategic product planning and business strategies. Additionally, this data serves as robust support for researchers conducting deeper studies in marketing and consumer behaviour. Second, due to the long-term support from the Chinese government for the automotive industry, a relatively comprehensive automotive production landscape has emerged in China. This includes a variety of brand types, such as domestic brands, joint venture brands, imported brands, and wholly foreign-owned brands (Yu et al. 2022; DING 2023). This diversity offers broader perspectives for brand value research, while the market performance of these brands also reflects their global influence. Third, the rapid development of Chinese automotive industry is closely tied to government policies (Petti et al. 2021). A comprehensive analysis of these policies, alongside Chinese automotive market development trajectory, provides other countries with valuable references for promoting their own automotive industries. In addition, more importantly, the market share of NEVs shows a strong growth trend and has exceeded 14% in the world(Khaleel et al. 2024). This shift has significantly impacted the competitive dynamics between China and the global automobile market, while also bringing about transformative changes in automotive sales channels and marketing models. Under this trend, China as the world's largest producer and seller of NEVs (He et al. 2022a), offers a wealth of comprehensive background information for studying future marketing priorities for NEVs, the development trajectory of NEVs, and the evolving competitive landscape between NEVs and traditional fuel vehicles.

In general, as a typical case of successful globalization, the Chinese automobile industry holds substantial guiding significance for the development of the global automotive sector (Ding and Akoorie 2013). Given the current influence of Chinese automobile market, conducting in-depth research on this market can offer a clearer reflection of the current state of the global automobile industry and future trends. This provides a strong basis for business strategy planning, marketing strategy formulation, and industry policy formulation.

1.1.3 Business data, business intelligence and analytics

Faced with constantly changing customer demands and intensifying market competition, driven by significant advancements in digital technology, digital transformation has become essential for ensuring the sustainable development of enterprises (Nadkarni and Prügl 2021). Digital transformation refers to the process by which an enterprise reshapes or redefines its existing business processes, operations, and management models using advanced digital technology. This enables innovation in business models and allows the company to meet evolving market demands (Cheng and Masron 2023). This shift has fundamentally altered consumer expectations and behavior, prompting managers to rethink traditional business theories, strategies, and operational approaches (Verhoef et al. 2021). It must be noted that in this data-driven transformation process, the identification of valuable data resources and how to analyze them using business intelligence techniques are important.

With the rise of social networks and the proliferation of information technologies, the rapid accumulation and presentation of data from various sources and formats in real life have increasingly captured people's attention (Chiang et al. 2018).Common business data often refers to the historical information and immediate information generated in

business activities, including market data, consumption data, marketing data, financial data, etc. Diverse data also means diversity of data forms, that is, it contains a variety of structured and unstructured data. As an important resource for knowledge discovery and knowledge creation, these data are important strategic assets with great economic and social value (Grover et al. 2018).

The value creation of business data exists in consumer analysis, product management, brand management, competitor analysis, marketing strategy analysis, and other aspects (Fan et al. 2015). Consumer-related data allows managers to gain deeper insights into consumer behavior and psychology, enabling them to generate accurate forecasts of future preferences. This, in turn, can be translated into competitive market advantages (Erevelles et al. 2016). Secondly, analyzing consumer feedback and sales data helps enhance customer responsiveness, establish accurate metrics for evaluating product performance in the market, and drive product innovation, ultimately leading to product success (Hajli et al. 2020). At the same time, the analysis of consumer data can reasonably control brand word-of-mouth, brand loyalty and brand risk, so as to realize the benign and sustainable development of the brand (Oh et al. 2019). In addition, analyzing business data helps identify potential competitors in the market, clarifying the relationships of substitution and complementarity with competing products, enabling more informed and strategic market decisions (Guo et al. 2017). Equally important, data-driven marketing has clear advantages in assessing marketing impact, attributing marketing effectiveness, and optimizing future marketing strategies (Kamath 2015). In general, integrating multiple types of business data and conducting comprehensive analysis will make the data create greater value.

Given the desire of business organizations to solve data-related problems, business intelligence and analytics have become an important area of concern for practitioners and researchers (Chen et al. 2012). The basic task of business intelligence and analytics is to obtain valuable business insights by analyzing key business data through data

analysis technology, business analysis system, business practice, application development, etc. The research related to business intelligence and analytics is often divided into three main aspects, namely big data analysis, text analysis, and network analysis (Lim et al. 2013). In practice, common traditional statistical methods such as general linear models, generalized linear models, and time series analysis models have often been applied in the past to analyze historical data to determine whether certain outcome patterns in the business domain were likely to have occurred by chance and to predict future trends (Miller and Haden 2006). It is evident that traditional statistical methods are often based on linear models to make reasonable inferences. When faced with large amounts of non-linear data, these methods struggle to capture the complex relationships between variables (Ij 2018). Moreover, because these methods focus on modeling and interpretation rather than on practical problems and data, they are prone to generating questionable conclusions or irrelevant theories (Delen and Zolbanin 2018). However, at present, machine learning methods, which are effective in pattern recognition and prediction, have demonstrated significant advantages in handling large-scale, high-dimensional, and nonl-inear data. Due to their outstanding adaptability, they have been applied to various fields of business analytics, such as using neural networks to predict consumer behavior (Chiang et al. 2006), using XGBoost to forecast product sales (Shilong et al. 2021), using generative adversarial networks to detect stock manipulation (Leangarun et al. 2018), using graph convolutional networks for text classification (Cui et al. 2022), and using reinforcement learning to optimize dynamic pricing (Wang et al. 2019), among others. Overall, these innovative data mining and artificial intelligence technologies have significantly accelerated the rapid development of the business intelligence and analytics field, while enhancing the efficiency and accuracy of business data analysis. At the same time, these methods are expected to create new value-creation mechanisms, helping businesses transform their models and optimize products and services. Clearly, the advantages of these technologies in

business analytics provide strong support for improving the quality of management decisions. But it also creates new challenges for managers and analysts. This requires them to master advanced analytical techniques, rich business theory, excellent ability to adapt to change, etc (Kowalczyk and Buxmann 2015).

Therefore, the effective application of diverse business intelligence and analytical technologies in studying a wide array of data from the automotive industry plays a crucial role in forecasting consumer demand, identifying market trends, and assessing competition in the future automotive landscape. This, in turn, offers valuable decision support for strategic planning and the formulation of marketing strategies. However, significant challenges remain. Establishing connections between various categories and types of data within the automobile market, utilizing business analysis tools with agility to adapt to the industry's evolving environment, and integrating business theory into the analytical process are pressing issues that require immediate attention.

1.2 Research gaps

In this study, in order to meet the needs of the digital transformation of the automotive industry, we mainly discuss the application of business intelligence and analysis technology represented by machine learning in the automotive industry and the extension of related business theories. Our main research objectives are to conducted in-depth analysis from three aspects: data management in the automotive industry, automobile market analysis and business fraud in online reviews by identifying the current research gaps in Chinese automobile market. Thus, it provides valuable data support for automotive industry analysis, effective decision support for automakers, and an ethical business environment for the entire automotive industry.

1.2.1 Chinese automotive industry data

For any industry, sales data is the most important business data. First of all, historical sales data is an intuitive reflection of past market demands, which can be used as the basis for predicting new product demand and analyzing market trends (Conrad 1976; Kalla et al. 2020). Second, sales data is instructive in optimizing products, pricing properly, and developing effective marketing campaigns. Because historical sales data directly reflects the market performance of the product. By identifying excellent products and product characteristics, it is beneficial for automobile companies to adjust product portfolio and improve enterprise benefits. As price is the product competition information and an important driver of sales, the analysis of sales data is also beneficial to enhance product pricing strategy (Huang et al. 2014). At the same time, the effect of product price discounts and other promotional activities can also be directly reflected in the sales data (Raju 1992). In addition, by combining other marketing variables, in-depth analysis of sales and the connection between them can help marketers discover more effective marketing mix and engage in precise marketing activities, thereby enhancing product and brand performance (Ataman et al. 2010). Overall, using sales data helps improve a company's operational efficiency across multiple areas, including production, sales, and marketing. However, there is currently no publicly available large-scale automotive sales dataset in the Chinese market, and the sales data used in past research contains only a limited number of variables, which cannot meet the growing demands of marketing.

Meanwhile, with the popularity of the Internet and the rise of social media, online consumer reviews have gradually become an important channel for consumers to convey their product experience. Online review often contains consumers' perceptual information, cognitive information, emotional information and social information. This provides potential consumers with real product information, reflects the decision-

making process of previous users, shows the emotions of users after experiencing the product, and promotes the high relevance of online reviews (Wang et al. 2023a). As an important form of online word-of-mouth communication, it has a profound impact on consumers' purchase decisions and product sales (Zhu and Zhang 2010; Gottschalk and Mafael 2017). Positive online consumer reviews can enhance the brand reputation of emerging brands and promote these brands to transform from weak to strong (Kostyra et al. 2016). Of course, the premise of this result is that the quality, design, price and so on of the product can be recognized by consumers. However, consumers now tend to pay more attention to negative online reviews when making consumption decisions. Negative online reviews tend to cut consumers' trust in products and brands, and significantly reduce their purchase intention (Cheung and Lee 2008). In addition to the aforementioned impacts, the consumer perception, preferences, and needs reflected in online reviews also serve as important support for companies to improve existing products. As a result, businesses are increasingly integrating online consumer reviews into their strategic and marketing decisions. Although many automotive-related social media platforms now encourage consumers to leave experiential reviews after purchasing a car, and have accumulated a large volume of online review data, many important product-centered, reviewer-centered, and review-centered variables are not included. These variables, such as car sales data, market sales data, the duration of consumer experience, and consumer review behavior, are often missing. This significantly limits the richness of insights derived from analyzing online reviews.

In addition, the effective application of news data mining in the past to predict the stock market, foreign exchange market, tourism market, etc., has significantly proved the business value of news data (Ashtiani and Raahemi 2023; Nassirtoussi et al. 2015; Park et al. 2021). News data often contains the latest industry policies, information on industry structure, competitive relationships between companies, technological innovation progress, current economic conditions, and related historical events (Han et al.

2018; Yamamoto et al. 2017). This helps companies to quickly understand the current and future industry development trends, and then make reasonable business decisions. However, due to the lack of news data in the automotive industry, previous studies often fail to combine relevant news data with business data for comprehensive analysis.

1.2.2 Competitive advantage of brands and products

As previously introduced in the background of the Chinese automobile market, both domestic and foreign automakers are facing increasing competitive pressure in China. First and foremost, the most evident competition is between NEVs and traditional fuel vehicles, as well as between emerging NEV brands and established traditional brands. Particularly, as consumers' environmental awareness gradually strengthens, government policies actively promote NEVs, government subsidies are continuously provided, and infrastructure for charging facilities is being constructed on a large scale, NEVs are gradually gaining recognition from many consumers (Wang et al. 2020; Fan et al. 2022). Under the integrated development of vehicle electrification, Internet and artificial intelligence technology, intelligent and connected vehicles have gradually become the development trend of the global automotive industry in the future (Zhang et al. 2022). Due to the high compatibility of electrification and intelligence, NEVs have obvious advantages in the field of intelligent connected vehicles. This undoubtedly significantly increases the competitive pressure and weakens the competitive advantage of traditional fuel vehicles. With the development of NEVs, some influential NEV startups have also emerged rapidly in the market, such as Tesla, NIO, XPeng and so on (Lib 2024). Some emerging NEV brands have adopted new technology-based marketing management methods that are different from traditional marketing, and have gained a

good response in the automobile market (Mangram 2012). This undoubtedly intensifies the impact on traditional brands. Therefore, it is unclear what market advantages each of the players—NEVs, emerging NEV brands, traditional fuel vehicles, and traditional brands—will have in the face of fierce competition.

Moreover, given the enormous consumer demand in the Chinese automobile market, numerous domestic and foreign brands have entered the market over the years. Similarly, due to changes in consumer attitudes and technological innovations in the automotive industry, consumer preferences for car brands may also be subtly shifting. For example, in the past, consumers often preferred brands from developed countries, but now, Chinese domestic automobile brands have made significant improvements in quality, design, technology, and reputation. Consequently, consumer preference for brands from developed countries may be declining, and biases against Chinese domestic automobile brands may be diminishing as well. At the same time, during the development of the automotive industry in China, political factors, unreasonable product strategies from automakers, and misguided marketing strategies may have gradually led consumers to develop biases against some brands. Therefore, it is currently unclear whether consumers' preferences for the brand country-of-origin(COO) in the Chinese automobile market have changed, and whether these preferences can be reflected in automobile sales data.

1.2.3 Business fraud in online reviews

The importance of online review data has been extensively discussed in many studies. However, the vast amount of online review data on social media platforms undoubtedly places a significant cognitive burden on consumers and creates considerable data analysis pressure for companies (Singh et al. 2017). While many studies aim to predict the usefulness of such data, these studies rely on the premise of obtaining reliable, authentic datasets. Unfortunately, the prevalence of fake reviews spans various consumer sectors, and the automotive market is no exception.

The most detrimental impact of fake reviews goes beyond this, as they seriously undermine business ethics. Online reviews serve as a crucial source of information for potential consumers to understand product features. Fake reviews can easily mislead consumers, lead to poor decision-making, and ultimately harm consumers' interests (Malbon 2013). At the same time, as a vital strategic asset for companies to analyze consumer behavior, predict market demand, and spread electronic word of mouth (eWOM), the presence of fake reviews significantly reduces operational efficiency and harms a fair market competition environment. In particular, online consumer reviews on social media are often overly positive and, in many cases, manipulated by companies or merchants (Wu et al. 2020). It has also significantly eroded consumer trust in social media platforms.

Although many studies have discussed the primary motivations for manipulating positive fake reviews and analyzed the textual characteristics of reviews, few have focused on the timing of manipulating positive fake reviews. Uncovering the timing of manipulation could significantly reduce consumers' cognitive burden regarding positive fake reviews and improve the efficiency of detecting fake reviews.

1.3 Research objectives

Building on the previous discussion of Chinese automotive industry data, the competitive advantage of brands and products, and business fraud in online reviews, as well as the identified research gaps, we have divided this study into three sub-studies. The primary research focuses on data management in the automotive industry, automotive market analysis, and business fraud in online reviews.

The first study aims to address the limitations in datasets related to the Chinese automobile market and is structured around the following three research objectives.

Objective 1.1: *To create a more comprehensive dataset by collecting sales data, online reviews, industry news and information data from the Chinese automobile market.*

Objective 1.2: *To enhance the dataset's practical value by incorporating additional relevant features.*

Objective 1.3: *To validate dataset's business value by conducting several preliminary empirical analysis.*

Given the increasingly intense competition among various automobile models and brands in the Chinese market, the second study focuses on sales prediction to gain deeper insights into market dynamics. The specific research objectives of this sub-study are as follows.

Objective 2.1: *To develop an innovative sales analysis framework based on machine learning, ensuring high predictive accuracy, strong interpretability, adaptability, and scalability.*

Objective 2.2: *To examine two common market advantages—first-mover advantage and brand country-of-origin effect—to identify significant changes in the current market landscape.*

Objective 2.3: *To evaluate the impact of these advantages on sales and determines*

their relative priority.

Objective 2.4: *To clarify the competitive advantages of different models and brands and assess whether innovative latecomers have the potential to surpass first-movers, by analysing the competitive landscape between NEVs and traditional fuel vehicles, as well as between emerging NEV brands and established brands.*

The third study focuses on identifying the timing of manipulated positive fake reviews to fill the research gap in this field. The specific research objectives are as follows.

Objective 3.1: *To identify potential factors associated with the timing of manipulated positive fake reviews by reviewing relevant literature.*

Objective 3.2: *To develop a high-performance framework for detecting fake reviews to obtain large-scale labelled online review data.*

Objective 3.3: *To establish effective fake review identification rules in the Chinese automobile market to generate labelled sample data that can enhance the accuracy of fake review detection.*

Objective 3.4: *To clarify the relationship between relevant factors and positive fake reviews to identify manipulation timing.*

In response to the rapid development of business intelligence and data analytics, the overall research, based on above three sub-studies, aims to achieve the following core objectives: optimizing existing dataset from the Chinese automotive industry to provide a solid data foundation for business analysis; improving the accuracy and applicability of market analysis and fraud detection by applying advanced machine learning methods; validating and expanding existing business theories; and providing effective support for the regulation of fake reviews. The alignment between the specific research objectives of the three sub-studies and the overall research objectives is illustrated in Table 1.2. Overall, this research explores data, methods, theory, and regulation across four dimensions to meet the practical demands of business intelligence development.

Table 1.2: The alignment between the specific research objectives of the three sub-studies and the overall research objectives.

Sub-study objective	Overall research objectives
Objective 1.1 Objective 1.2 Objective 1.3 Objective 3.2 Objective 3.3	To optimize existing dataset from the Chinese automotive industry to provide a solid data foundation for business analysis
Objective 2.1 Objective 3.2 Objective 3.3	To improve the accuracy and applicability of market analysis and fraud detection by applying advanced machine learning methods
Objective 2.2 Objective 2.3 Objective 2.4 Objective 3.1 Objective 3.4	To validate and expand existing business theories
Objective 3.1 Objective 3.2 Objective 3.3 Objective 3.4	To provide effective support for the regulation of fake reviews

1.4 Research overview

1.4.1 Study 1 - SRNI-CAR: A comprehensive dataset for analyzing the Chinese automotive market

In this part of the research, we first provide a detailed review of existing datasets and their limitations. Then, using data crawling techniques, we collected sales data, consumer online reviews data, and industry news and information data related to the Chinese automobile market from 2016 to 2022 across three well-known Chinese automotive media platforms. Additionally, we expanded the dataset by incorporating many valuable variables, such as product characteristics, brand attributes, and news topic features, which were not included in previous datasets. Based on the processed

complete dataset, we provided simple examples in the fields of sales forecasting and consumer behavior analysis to demonstrate the dataset's business and academic potential. Finally, we published the entire dataset at IEEE BigData 2023 (2023 IEEE International Conference on Big Data).

The simple application example of the created dataset in sales forecasting revealed the following key findings:

- In the overall Chinese automobile market, the official price and the actual transaction price of a vehicle have a greater impact on sales than discounts, with the official price being particularly influential.
- The launch date of the car model, the date the brand entered China, and the established date of the brand are key factors influencing car sales.
- The sentiment of review text has a greater impact on sales than the ratings in online reviews.
- In terms of car model size, Chinese consumers show a preference for compact car models.
- Regarding the country-of-origin of brands, Chinese consumers tend to favor car brands from Germany, Japan, and America.
- Among NEVs, pure electric vehicles and hybrid electric vehicles are more popular with Chinese consumers.

Additionally, based on the application example of online review data, we have obtained the following valuable findings:

- The level of user experience has a direct impact on the sentiment of online reviews.
- When consumers review a car's comfort, appearance, and performance, the influence of price factors is not significant.
- The sentiment of consumer reviews is largely influenced by the year the car model was purchased.

- The launch date of the car model, the date the brand entered China, and the established date of the brand are all directly related to the sentiment of online reviews.
- When reviewing car models, consumers are particularly focused on elements such as exterior design, the central console inside the car, acceleration capabilities, steering sensitivity, rear seat space, and seat comfort.

1.4.2 Study 2 - Do first-mover advantage and country-of-origin retain their impact in today's automotive market? Unveiling insights through machine learning

First-mover advantage and country-of-origin effect, as important marketing variables in the Chinese automobile market, prompted us to propose the following four research questions through a review of related literature and a re-examination of the current competitive landscape of the Chinese automotive market: (i) To what extent do first-mover advantages, related to both the model and the brand, affect sales performance in the automotive industry when many new participants enter the market? (ii) How significantly does the energy type of the vehicles affect their sales performance? (iii) In instances where first-mover advantages are present, can innovative later entrants in the market, such as NEVs and their emerging brands, mount a challenge and potentially offset these entrenched benefits? (iv) When many innovative car models and automakers enter the market, do consumers still show a preference for vehicles from developed countries? Are these preferences driven more by the country-of-origin effect, or do first-mover advantages play a larger role?

In this part of the research, we conducted an in-depth study based on sales data to examine the specific impact of car model first-mover advantage, brand first-mover advantage, the innovation advantage of late entrants, and the country-of-origin effect on sales performance in different segments of the Chinese automobile market. Additionally, we compared the differences in the importance of first-mover advantage and the country-of-origin effect in sales forecasting, as well as their varying impacts on high-end and low-end automobile market.

In previous studies related to sales analysis, many researchers employed traditional statistical methods or machine learning techniques. However, these methods often struggle to balance the accuracy and interpretability of sales predictions. To address this issue, we proposed a comprehensive two-stage sales analytics framework based on machine learning prediction models and the SHAP method in this research. We also improved upon the existing analyses by addressing the potential prediction biases caused by relying solely on a single sales performance indicator.

Based on the study of first-mover advantage and country-of-origin effect in predicting car sales performance across various segments of the Chinese automobile market, we have identified several important findings:

- In most segments, both brand first-mover advantage and model first-mover advantage positively impact car sales performance. However, in most segments, brand first-mover advantage plays a more critical role.
- Earlier brand establishment dates and market entry dates can both positively contribute to forming a brand's first-mover advantage, but their contributions are not always consistent. In most segments, the brand establishment date has a greater impact on sales performance.

- As late entrants in the Chinese automobile market, in some segments, NEVs and emerging NEV brands can disrupt the first-mover advantage held by traditional fuel vehicles and traditional brands.
- The country-of-origin effect has a direct impact on car sales performance.
- While most consumers still show a preference for car brands from developed countries, in a few segments, consumers are beginning to show a clear preference for Chinese domestic brands.
- In most segments, the impact of first-mover advantage on car sales performance is greater than that of the country-of-origin effect.
- There is no significant difference in the influence of first-mover advantage and country-of-origin effect on car sales performance between the high-end and low-end automobile markets.
- Either sales or sales ranking, when used alone as a performance indicator of car sales, will bring significant bias. The new performance indicators established by clustering based on sales and ranking can more accurately reflect the sales performance of car models in market segments.

1.4.3 Study 3 - When is the temptation too strong? Analyzing the timing of positive fake review manipulation

In the online review platform, positive reviews tend to occupy the vast majority. Therefore, to mitigate the bad impact of fake reviews in business ethics and make up for the lack of previous research on timing of manipulating fake reviews, we proposed the following research questions: When do companies or merchants manipulate positive fake reviews?

In this study, we first identified the ethical issues of fake reviews, and the harmful impact they have on consumers and the competitive environment of businesses. Since competition and financial incentives are the primary motivations behind the manipulation of positive fake reviews, we conducted an in-depth analysis of the timing for manipulating such reviews by discussing their relationship with car sales, market size, and the duration of both the model and brand in the market. Given that online review platforms typically do not label fake reviews, we reviewed past methods for detecting fake reviews and proposed a high-performance two-stage approach for detecting fake reviews based on BERT and PU learning. Finally, we applied the SHAP method to interpret the fake review predictions, uncovering important insights into the timing of manipulating positive fake reviews. In the process, we further expanded the product-centered features, review-centered features, and reviewer-centered features within the online review dataset. Additionally, we developed reliable rules for identifying fake reviews in the automobile market and created a relatively accurate labeled online review dataset.

By applying our proposed fake review detection method to predict online reviews in the Chinese automobile market and analyze the timing of manipulated positive fake reviews, we have uncovered the following key findings:

- In terms of car sales, firms are most likely to manipulate positive fake reviews when car sales reach a moderate level in the market.
- Regarding brand sales, firms are similarly more likely to manipulate positive fake reviews when the brand's sales reach a high level within the market.
- As for market size, while this factor is related to the timing of manipulating positive fake reviews, the manipulation timing varies significantly across different market segments.
- The mid-to-late stages of a car model's lifecycle are when manufacturers or dealers are most likely to manipulate positive fake reviews.

- In the first two years of brand establishment, firms are more likely to engage in the manipulation of positive fake reviews.
- In the early stage of a brand's entry into the Chinese market, firms are less likely to engage in the manipulation of positive fake reviews.
- Textual features are the most critical factor in identifying fake reviews.

1.5 Structure of the thesis

The remainder of this thesis is organized as follows. Chapter 2 presents my first study, which introduces the comprehensive dataset we created for the Chinese automobile market and its potential applications. Chapter 3 presents my second study, which discusses in detail the first-mover advantage and country-of-origin effect in the Chinese automobile market, and introduces a novel sales analysis framework based on machine learning. Chapter 4 presents my third study, which discusses in detail the timing of positive fake reviews and proposes a fake review detection method based on BERT and PU-learning. Chapter 5 summarizes the overall research, discussing the research contribution, implications, current limitations and directions for future research.

Chapter 2

SRNI-CAR: A comprehensive dataset for analyzing the Chinese automotive market

2.1 Abstract

The automotive industry plays a critical role in the global economy, and particularly important is the expanding Chinese automobile market due to its immense scale and influence. However, existing automotive sector datasets are limited in their coverage, failing to adequately consider the growing demand for more and diverse variables. This paper aims to bridge this data gap by introducing a comprehensive dataset spanning the years from 2016 to 2022, encompassing sales data, online reviews, and a wealth of information related to the Chinese automotive industry. This dataset serves as a valuable resource, significantly expanding the available data. Its impact extends to various dimensions, including improving forecasting accuracy, expanding the scope of business applications, informing policy development and regulation, and advancing academic research within the automotive sector. To illustrate the dataset's potential applica-

tions in both business and academic contexts, we present two application examples. Our developed dataset enhances our understanding of the Chinese automotive market and offers a valuable tool for researchers, policymakers, and industry stakeholders worldwide.

2.2 Introduction

The automotive industry has been one of the important economic sectors in many countries and regions. It also has an important role in global energy consumption, energy sustainability and environmental impact. At present, China has become the world's largest market for NEVs and one of the largest automobile markets. Studying the Chinese automobile market reveals its profound impact on the global automotive ecosystem in terms of market growth opportunities, innovation cooperation, market diversity, environmental sustainability, and more. For a long time, the automotive industry has been facing many challenges in the marketing process, such as fierce market competition, the diversity of consumer demand, and the pressure of environmental protection policies. With the growth of social media and digital advertising, consumer engagement on the Web has greatly increased, and rational and effective digital marketing has become another serious challenge in the automotive industry (Homburg and Wielgos 2022).

The imperative for precise and efficient forecasting, rooted in historical data, becomes evident to address the challenges mentioned above. A substantial cohort of automotive industry stakeholders including executives, marketers, and academics are actively in pursuit of more efficacious methods for market analysis. Traditional forecasting approaches face formidable hurdles when contending with vast datasets, often falling short in ensuring the fidelity of sales predictions (Cheriyan et al. 2018). While the advent of

cutting-edge data-driven methods such as data mining and machine learning has led to significant advancements in knowledge discovery and predictive capabilities, their efficacy hinges crucially on the availability of comprehensive and accurate datasets. However, currently available datasets are often fragmented, making it difficult for researchers to integrate industry information, automakers' behavior, consumer demand, market feedback, and sales forecasts. It is increasingly important to create a dataset that can meet the multiple business analytics needs of the automotive industry. Therefore, we collected car model sales data, consumer online review data, and automotive industry news and information data from different sources, and integrated them into a comprehensive dataset, called *SRNI-CAR*.

Our work makes two significant practical contributions. Firstly, we have curated a more comprehensive and scalable data resource. It not only consolidates industry news, development insights, automotive marketing data, consumer online reviews, and sales information but also introduces valuable variables previously absent, such as model launch dates and brand inception dates. Therefore, our dataset supports a broader spectrum of research possibilities compared to existing publicly available datasets in the automotive domain. Furthermore, it enhances analytical accuracy and interpretability. Secondly, our dataset possesses substantial business value in the automotive sector. Sales data aids automakers and marketers in discerning market trends, while review data facilitates the identification of consumer preferences, evaluation of marketing effectiveness, and product strengths and weaknesses analysis. Incorporating industry news, development insights, and automotive publicity empowers automakers to grasp industry trends and market competition. Integrated analysis across multiple data provides superior decision support for product planning, business expansion, and marketing strategy formulation. Our dataset can be used to improve product quality, and

align with consumer demands. Additionally, our dataset can be used by government policymakers, enabling the formulation of pertinent automotive industry policies and regulations. Thus, the development of this dataset holds profound significance for a myriad of stakeholders within the automotive industry.

Table 2.1: Summary of variables in the datasets for automotive market or consumer analytics.

Variable	**[1]**	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**	**Our dataset**
Vehicle attribute related variables (e.g., model type, model size, energy type, etc.)	✓		✓	✓	✓		✓
Brand related variables (e.g., brand energy type, brand country-of-origin, etc.)	✓	✓		✓			✓
Price variables (e.g., official price, transaction price, etc.)	✓		✓	✓		✓	✓
Date-related variables (e.g., year of purchase, year of review, model launch date, etc.)	✓	✓	✓	✓	✓	✓	✓
Consumer experience variables (e.g., experience duration, mileage, etc.)						✓	✓
User ratings (e.g., overall rating, exterior rating, interior rating, etc.)			✓			✓	✓
Online reviews (e.g., advantage, features comments, comfort comments, etc.)		✓	✓			✓	✓
Industry news and information (e.g., title, text, information label, pageview, etc.)							✓
Sales	✓	✓	✓	✓	✓		✓
Website search variables (e.g., search volume of brand, etc.)		✓					
Economic factor (e.g., consumer price index, gas prices, etc.)					✓		
Vehicle design (e.g., vehicle pictures, design fluency, etc.)				✓			

Note: [1] Xia et al. 2020, [2] Geva et al. 2017, [3] Chen et al. 2011, [4] Landwehr et al. 2011, [5] Sa-Ngasoongsong et al. 2012, [6] Wang et al. 2018

2.3 Review of existing automotive datasets

The automobile industry has garnered significant attention from researchers and analysts, particularly for sales forecasting. A review of prior studies reveals that researchers commonly incorporate a range of variables encompassing vehicle attributes, brand characteristics, pricing, temporal factors, user experiences, ratings, reviews, website search data, economic indicators, and vehicle design to prognosticate car sales. However, as shown in Table 2.1, none of the datasets employed in these studies encompass all these

pertinent variables. In the realm of nonlinear modeling, the reduction of variables often proves counterproductive to enhancing predictive model performance (Hülsmann et al. 2012). Consequently, the absence of comprehensive variables constitutes a notable deficiency within the existing datasets in the automotive industry.

The limitations of the current automotive datasets are noteworthy. Firstly, they often omit vital information about brand creation and model launch dates, which is crucial for understanding market dynamics (Markides and Sosa 2013). Secondly, they fail to differentiate between the growing number of new energy vehicle brands and do not clearly identify brand origins, despite their influence on consumer perceptions (Häubl 1996). Thirdly, these datasets often lack detailed comments and ratings for specific vehicle attributes, limiting their usefulness for sales forecasting and preference analysis (Geva et al. 2017). Fourthly, they usually provide only aggregate pricing data for vehicle series, omitting separate pricing for individual models and making it challenging to analyze the impact of discounts (Wang et al. 2018). Furthermore, important data related to the automotive industry, such as model sentiment and review articles, are often missing, and aligning them with sales data remains challenging (Urban et al. 1990). These limitations highlight the need for more comprehensive datasets encompassing critical temporal, attribute-specific, and contextual factors for robust automotive market analysis and forecasting.

2.4 Data collection, preparation and description

We collected sales and online review data from PCAuto and Dongchedi, and automotive industry news and information data from Autohome. They are three most authoritative and largest automobile media platforms in China. It is important to note that these platforms typically disclose sales data only for the past five years. Since our data collection took place in January 2021, the starting date of the dataset could only be set to January 2016 to ensure consistency in the time frame. The dataset was subsequently updated until December 2022 and used for the following research.

While the collected raw data encompass a broader range of variables than previous studies, they still lack essential elements necessary to address the evolving requirements of business analysis and academic research. To address these gaps, we introduced additional variables related to brand country-of-origin classification, vehicle entry order, and brand entry order. We also synchronized monthly sales data with online reviews to enhance our ability to derive valuable business insights. Furthermore, we manually incorporated actual transaction prices and official guide prices for specific car models within each series to investigate pricing and discount impacts. The actual transaction price can be directly extracted from the review data on automotive media platforms, as it is a required field when consumers post online reviews. To ensure data validity, we diligently identified and addressed missing and outlier values across three data sources. Sales data required minimal processing, containing no missing values, outliers, or duplicates. For online reviews, we selectively removed instances with multiple missing values and those with missing values in less than 1% of variables other than sales. Notably, data with only a missing value in sales were retained, comprising 9.68% of

the total dataset. It is important to note that imported car sales data are not included, resulting in the absence of corresponding sales data in the review dataset. Finally, we meticulously eliminated duplicates to produce the refined dataset intended for publication.

As shown in Table 2.2, SRNI-CAR consist of sales, online reviews, and automotive industry news and information, over the period from 2016 to 2022. The dataset can be downloaded from the following address:

https://srni-car.github.io

The sales data comprises 1,236 car series, including 518 sedans, 598 SUVs, and 121 MPVs, with 39,496 observations. Each observation represents the monthly sales data of a specific car series during the period from 2016 to 2022. It encompasses 155 car brands, with 107 traditional and 48 new energy brands, originating primarily from Germany, France, Korea, the Czech Republic, the USA, Japan, Sweden, Italy, the UK, and China. This data is valuable for research in sales forecasting, first-mover advantages, brand country-of-origin, and consumer preferences. It's stored as a 3.6 MB CSV file. The online review data is based on 217,292 comments from car owners in 358 cities, covering 13,039 specific models across 672 car series, ranging from 26,800 to 14.88 million yuan. It includes 10,977 traditional and 2,062 new energy models under 127 car brands, categorized into eight aspects by vehicle attributes. It can be effectively applied to research on sales forecasting, consumer behavior, product evaluation, and fake review detection, and is stored in a 480 MB CSV file. The automotive industry news and information data contains 83,590 items that covers a wide range of topics. Each data items is accurately labeled, aiding researchers in selecting relevant data. This data is stored as a 224.1 MB CSV file, beneficial for analyzing industry trends, marketing strategies, consumer perceptions, and automotive technology development.

Table 2.2: Description of variables in SRNI-CAR.

Data	Variables	Description
Sales	Car series	Name of the car series.
	Brand	Name of the brand.
	Year	Year in which the car series was sold.
	Month	Month in which the car series was sold.
	Car model type	Car model category: Sedan, SUV and MPV.
	Brand energy type	Brand category based on energy type of vehicle produced.
	Size	Vehicle size category: mini, minivan, minibus, small, compact, mid-size, larger than mid-size, full-size.
	Brand country of origin	Country in which the brand was created.
	Model launch date	Year when the car series was launched on the Chinese market.
	Brand establishment date	Year when the brand was created.
	Brand entered China date	Year when the brand officially entered the Chinese market.
	Sales	Total sales of the car series in the month.
Reviews	Car series	Name of the car series.
	Brand	Name of the brand.
	Size	Vehicle size category: mini, minivan, small, compact, mid-size, larger than mid-size, full-size.
	Car model type	Car model category: Sedan, SUV and MPV.
	User ID	Name that users use when making online reviews.
	Year of review	Year when the user reviews.
	Month of review	Month when the user reviews.
	Specific model purchased	Specific model of a car series purchased by a user.
	Official price	Official prices for specific models purchased.
	Car energy type	Vehicle energy type: gasoline vehicle, diesel vehicle, hydrogen vehicle, and so on.
	Brand energy type	Brand category based on energy type of vehicle produced.
	Brand country of origin	Country in which the brand was created.
	Brand establishment date	Year when the brand was created.
	Brand entered China date	Year when the brand officially entered the Chinese market.
	Model launch date	Year when the car series was officially launched on the Chinese market.
	Year of purchase	Year when the user purchased the model.
	Month of purchase	Month when the user purchased the model.
	Sales	Total sales of the car series in the month when the user purchased the model.
	Experience duration	Months between purchase date and review posting date.
	Province	Province in which the user purchased the model.
	City	City in which the user purchased the model.
	Transaction price	Real transaction price of the model.
	Average energy consumption	Gasoline, diesel, electricity, or hydrogen consumed for every 100 kilometers traveled.
	Mileage	Kilometers the user has driven the model at the date the review was posted.
	Overall rating	User's overall rating of the vehicle purchased.
	Exterior rating	User's rating of the exterior of the vehicle.
	Interior rating	User's rating of the interior of the vehicle.
	Space rating	User's rating of the space of the vehicle.
	Features rating	User's rating of the feature of the vehicle.
	Power rating	User's rating of the power of the vehicle.
	Energy consumption rating	User's rating of the energy consumption of the vehicle.
	Driving rating	User's rating of the driving of the vehicle.
	Comfort rating	User's rating of the comfort of the vehicle.
	Advantage	The advantages of the model as perceived by the user.
	Disadvantage	The disadvantage of the model as perceived by the user.
	Exterior comments	User's comments on the exterior of the vehicle.
	Interior comments	User's comments on the interior of the vehicle.
	Space comments	User's comments on the space of the vehicle.
	Features comments	User's comments on the feature of the vehicle.
	Power comments	User's comments on the power of the vehicle.
	Energy consumption comments	User's comments on the energy consumption of the vehicle.
	Driving comments	User's comments on the driving of the vehicle.
	Comfort comments	User's comments on the comfort of the vehicle.
News information	Title	Title of the information.
	Pageview	Number of times the information was viewed.
	Number of comments	Number of comments the information received.
	Text	Text content contained in the information.
	Release date	Date on which this information was published.
	Author	Person who posted the information.
	Source	Source of the information.
	Information type	Whether the information is original, compiled, a press release, or reprinted from another platform.
	Information label	Labels chosen by the author that summarizes the information, based on its content.

2.5 Automotive analytics examples

Two application examples are presented to showcase the potential of our dataset in automotive analytics.

2.5.1 Automobile sales forecasting

For automotive manufacturers, proficiency in sales forecasting holds critical significance in shaping both product development and marketing strategies. The intricate interplay among an array of factors influencing sales, including price dynamics (Assuncao and Meyer 1993; Busse et al. 2010), strategic discounting, consumer sentiment gleaned from online reviews, and numerical ratings (Hu et al. 2014), necessitates a comprehensive analytical framework. Moreover, analysts must diligently consider factors like first-mover advantages, brand country-of-origin, brand attributes, product attributes, and the increasingly salient energy characteristics of vehicles, especially within the context of China's fervent pursuit of NEVs (Zahoor et al. 2023). Unlike previous studies constrained by data limitations, our dataset offers an opportunity for an extensive and all-encompassing examination of these pivotal determinants of sales.

To assess the impact of these factors on sales, we harnessed online review data spanning 2019 to 2021 and opted for the XGBoost model, lauded for its efficiency and precision in sales forecasting (Shilong et al. 2021). We optimized model parameters using grid search cross-validation to bolster predictive performance. To add robustness, one can explore alternative machine learning methodologies but that is not the most significance of our example here. Categorical predictors were judiciously transformed into dummy variables. Furthermore, we employed SnowNLP for sentiment analysis of

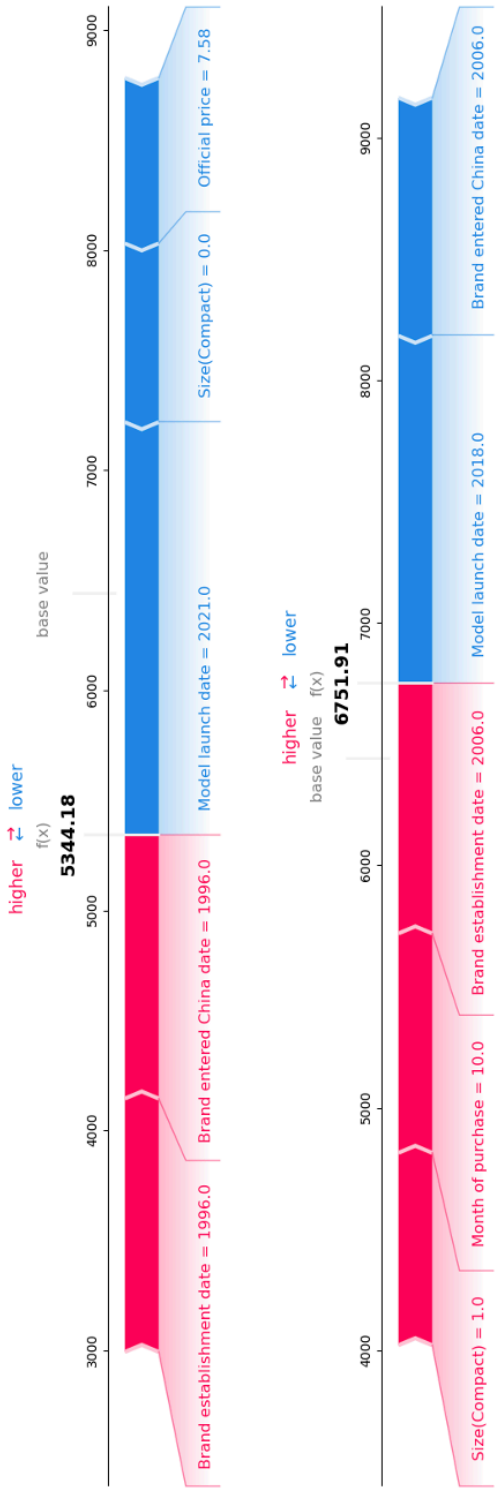

Figure 2.1: Top variables with the most significant contributions in two data instances.

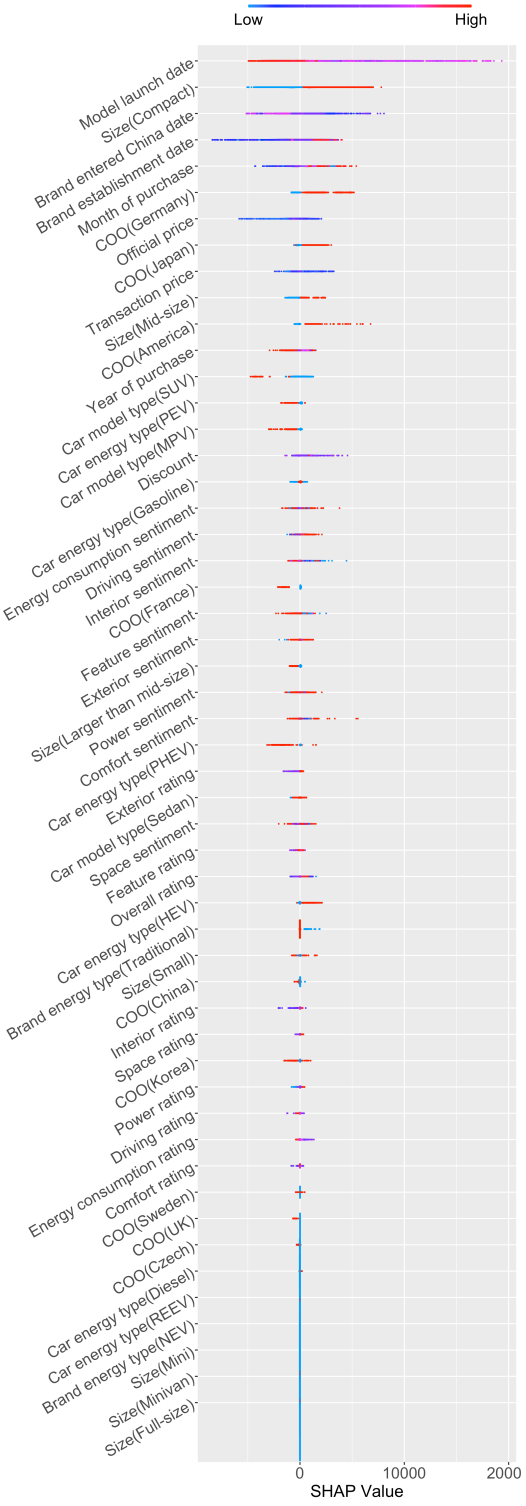

Figure 2.2: SHAP values corresponding to the variables influencing car sales.

textual comments, yielding sentiment scores on a scale from 0 to 1 as supplementary predictive variables. Finally, to interpret XGBoost model outputs, we applied the Shapley additive explanations (in short SHAP), mainly for its efficacy in elucidating feature contributions without the encumbrance of multi-collinearity concerns (Marcílio and Eler 2020; Baptista et al. 2022).

In our analysis, we were able to obtain localized insights, presented in the form of Figure 2.1, which depict the contribution of each variable for every data instance—commonly referred to as SHAP values. In a locally interpreted force plot, we can determine the contribution of features in this data instance by observing the length of the arrows. Additionally, red arrows represent features that positively impact the prediction, while blue arrows indicate a negative impact. For example, in the first data instance, we can observe that model launch date has the largest contribution to the prediction, but is negatively impacting it. On the other hand, brand entered China date, although its contribution is relatively lower, still positively impacts the prediction. Therefore, this data instance reveals that the launch date of the car model is a significant factor influencing its sales, albeit with a negative effect. Simultaneously, brand entered China date is also an important factor, despite its relatively lower contribution, as it positively affects the sales.

It is important to note that local interpretation only explains the contribution of these variables in a particular instance, while global interpretation based on variable importance reflects the impact of these variables on the overall data. Within the SHAP framework, the variable importance is gauged as the average of absolute SHAPHAP values across all data instances for each variable. Figure 2.2 comprehensively outlines the variables employed in our experiment, showcases the experimental outcomes, and ranks these variables based on their significance in predicting sales. Notably, it visually underscores how the contribution of each variable to sales evolves as the variable's value changes. In this plot, each point represents the SHAP value of the corresponding

variable in each instance. When the points are distributed at SHAP values greater than 0, it indicates a positive contribution of the corresponding variable in that instance; when the points are distributed at SHAP values less than 0, it indicates a negative contribution. The color of the points represents the relative magnitude of the variable's value across different instances, with the transition from blue to red corresponding to an increasing variable value. For categorical variables, they were divided into multiple dummy variables based on category labels in the experiment. These dummy variables take only two values: 1 represents that the model belongs to this category, while 0 indicates that it does not belong to this category.

We found several intriguing and valuable findings. Firstly, the sequence of model and brand entry wields a significantly higher influence on sales compared to other variables. Notably, alongside vehicle size, the launch date of a model, brand's entry date into China, and the brand's founding date emerge as the three most critical factors. At the same time, based on the distribution of the values of the model launch date and the brand's entry date into China as they vary with SHAP values, larger values tend to correspond to negative SHAP values, meaning that later years contribute more negatively to sales. This suggests that in the Chinese automotive market, there are both model first-mover advantages and brand first-mover advantages. However, it is important to note that, based on the the brand's founding date, late movers still have an advantage. Secondly, within the realm of price-related variables, although discounts exert a promotional effect on sales, official prices and transaction prices exert more substantial impacts, with official prices retaining the highest importance. However, it is worth noting that the relationship between price and sales is not linear. Thirdly, sentiment extracted from review text, categorized by vehicle attributes, exerts a notably higher impact on sales compared to corresponding rating data. This suggests that the review text serves as a more representative source of consumer sentiment regarding a model. Fourthly, in the Chinese automotive market, compact models and those from German, Japanese, and American brands tend to exhibit positive SHAP values,

meaning that models belonging to these categories have higher consumer popularity. Lastly, by observing the distribution of SHAP values for different car model energy types and brand energy types, consumers exhibit a preference for pure electric vehicles (PEVs) and hybrid electric vehicles (HEVs), alongside a high level of acceptance for NEV brands.

In summary, these findings underscore the imperative for automaker managers and marketers to meticulously analyze and weigh these factors when devising enterprise development and marketing strategies. Such insights are also instrumental for government departments in evaluating the effectiveness of new energy vehicle policies.

2.5.2 Consumer behavior analytics

Online reviews, generated by consumer post-purchase, significantly reflect and influence consumer decision-making (Chen and Xie 2008). Rating and review text are key components of online reviews, with ratings indirectly affecting sales through their interaction with the text (Li et al. 2019). Consumers tend to use ratings to shortlist products and rely on review text for final choices (Hu et al. 2014), highlighting the higher business value of review text in expressing customer satisfaction. Understanding how consumer behavior and preferences impact review sentiment is crucial for shaping marketing strategies.

Previous studies have shown that the level of user experience and price have an impact on automotive customer satisfaction. Mileage and experience duration are two important variables in the automotive industry to measure the level of experience (Wang et al. 2018). In addition, because consumers may have different expectations for different countries and different types of products, factors such as brand country-of-origin,

brand history, and car type may also have an impact on the sentiment of the review text. At the same time, previous studies did not take into account that the factors that affect the sentiment of reviews for different attributes of vehicles may be different. For example, perhaps the review for comfort is more likely to be influenced by the level of user experience than the review of exterior. SRNI-CAR makes it possible to forecast the sentiment of automobile consumer reviews and to study the above questions.

In this example, we conducted sentiment score forecasts for comments associated with distinct vehicle attributes, employing the SnowNLP, XGBoost, and SHAP. The comprehensive experiment was divided into eight segments, each aligned with specific vehicle attributes. These segments encompassed the forecasting of sentiment scores for comments pertaining to exterior, interior, space, features, power, driving, energy consumption, and comfort attributes.

The depicted analysis Figure 2.3 underscores the significance of various variables in predicting sentiment scores for comments linked to diverse vehicle attributes. The combination of these eight forecasts yields several pivotal insights. First, it is evident that the user's experience level, encompassing driving mileage and experience duration, influences consumer review sentiments, indirectly affirming its impact on overall customer satisfaction, with mileage proving more influential than experience duration among the eight vehicle attributes examined. Second, price-related factors indeed influence consumer reviews, with transaction and official prices holding greater sway than discounts, although their influence is less pronounced in reviews concerning comfort, exterior, and power attributes. Third, the year of vehicle purchase emerges as a crucial factor in forecasting review sentiment, an aspect hitherto unexplored, suggesting evolving consumer expectations. Fourth, vehicle-related attributes, such as size and energy type, demonstrate little relevance in exterior-related comments, indicating that Chinese consumers lack a distinct preference for specific model appearances. Fifth, model launch date, brand establishment date, and entry into the Chinese market date consistently wield

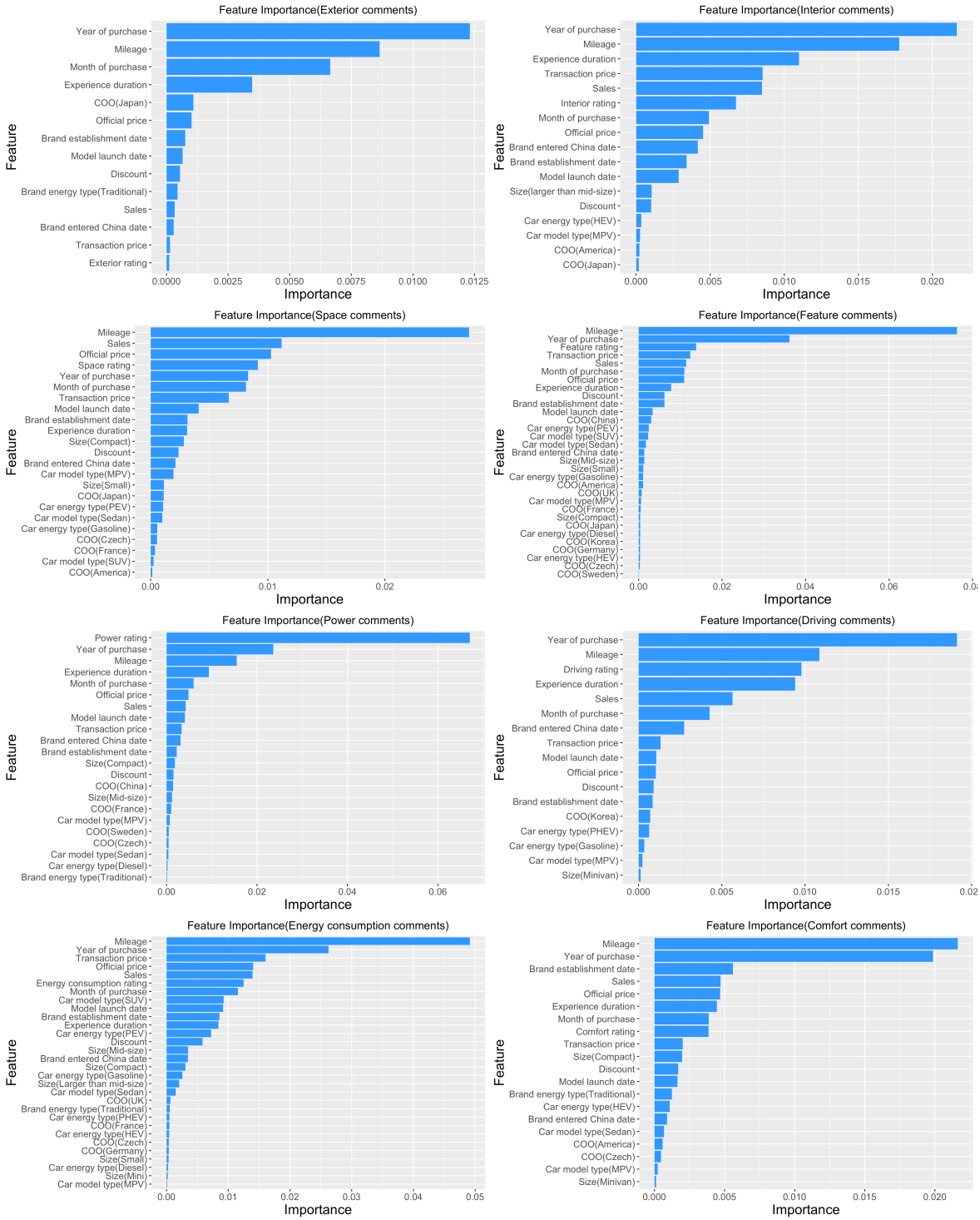

Figure 2.3: Importance of variables in predicting sentiment across eight review categories.

importance across all eight review categories, suggesting that consumer review sentiment may signify either a first-mover or late-mover advantage. Sixth, ratings assume the highest significance in reviews associated with power, with elevated importance in feature, driving, and space-related reviews, highlighting the divergence between ratings and textual comments across various review aspects. Last but not least, the country of origin, size, and energy type exhibit varying impacts on sentiment in comments across different aspects. Furthermore, as shown in Figure 2.4, we conducted a word frequency analysis of comments pertaining to the eight distinct attributes and subsequently generated word clouds. The findings elucidate consumers' distinct priorities within these attributes. Notably, consumers focus more on the exterior design when assessing a vehicle's appearance. Concerning interior design, particular attention is directed toward the vehicle's central console. In evaluating power performance, consumers prioritize acceleration capabilities, while precision in steering takes precedence in assessments of handling. Features such as the reverse camera and functionality garner significant attention. The spaciousness of rear-seat areas is a key consideration for consumers when assessing vehicle space. In the realm of comfort, the quality of seats emerges as a pivotal factor. Lastly, regarding energy consumption, consumers exhibit heightened scrutiny when driving in urban settings and on highways.

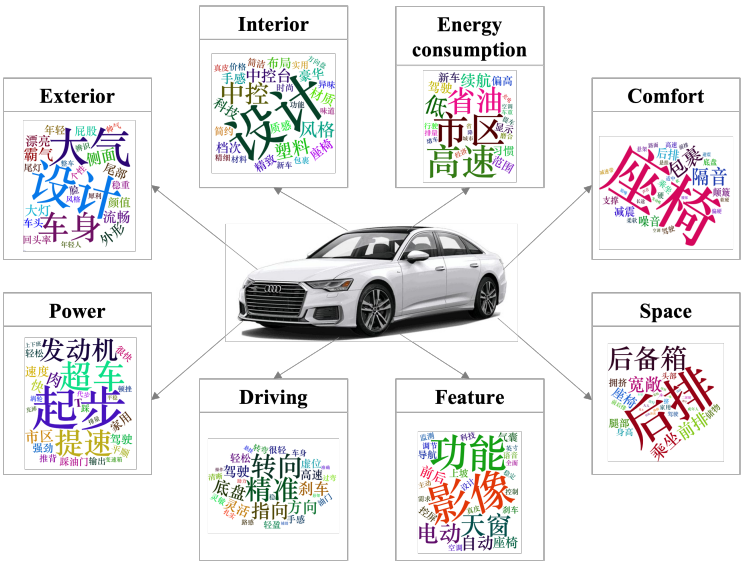

Figure 2.4: Word clouds created from review comments associated with the eight vehicle attributes.

In summary, when exploring consumer behavior within the automobile market, one must consider the multifaceted nature of review data. Our dataset and accompanying analysis provide invaluable insights for marketers, facilitating the evaluation of product competitiveness, the identification of suitable marketing strategies, and the implementation of precision marketing tactics to enhance overall marketing effectiveness.

2.6 Summary

We have diligently created a comprehensive dataset SRNI-CAR for Chinese automotive market. This paper details the data collection and processing procedures, expounds upon the intricacies of variables, and highlights the dataset's profound relevance in areas such as market demand, consumer behavior, preferences, and industry development. We believe that SRNI-CAR will benefit not only the academic research community but also automotive industry leaders, marketers, and government agencies. We remain committed to continuously improving this dataset, with plans for updates and expansions, all aimed at enhancing its practical applications. This paper serves as a means to make the dataset accessible to a wider audience, facilitating a deeper understanding of the complexities within the automotive sector.

Chapter 3

Do first-mover advantage and country-of-origin retain their impact in today's automotive market? Unveiling insights through machine learning

3.1 Abstract

While first-mover advantage and country-of-origin effects have long been recognized across industries, the automotive sector is undergoing significant transformation with the surge of new players, markets, and electric vehicles. Such radical changes challenge established beliefs, prompting a re-examination of these assumptions. This study examines how factors like market entry timing, technological advancements, and national origin influence consumer behavior in this dynamic landscape. We innovatively considered the impact of market entry timing from three aspects: brand establishment date, brand market entry date, and product launch date, and compared it with the country-of-origin effect. Leveraging advanced machine learning techniques, we propose a novel

two-stage framework to analyze automobile sales performance in the Chinese market, a major player in both traditional and new energy vehicles. Our findings reveal that first-mover advantages and country-of-origin effects persist, early market entry and brand heritage remain very important. However, there are also some noticeable shifts in the field. Innovative latecomers including new brands and car models, have the potential to disrupt established norms. In addition, we also found that first-mover advantage plays a more critical role in achieving strong sales performance than country-of-origin effects. At the same time, in most market segments, their impact is not significantly different for high-end and low-end car models. These insights provide substantial economic and strategic implications for the automotive sector and beyond, aiding in navigating rapidly changing markets.

3.2 Introduction

The automotive industry is a fundamental component of the global economy. For decades, the worldwide automobile market has been dominated by established brands from countries that industrialized early. These brands leveraged their first-mover advantage and the favorable impressions linked to their countries of origin to secure a leading position. Traditionally, being first and boasting a prestigious brand's country-of-origin have been acknowledged as important attractors for consumers across different sectors (Verlegh et al. 2005). However, the emergence of new challengers from countries like China, South Korea, and India, coupled with the growing popularity of NEVs (Shao et al. 2017; Li et al. 2023), is reshaping the competitive landscape (Fetscherin and Toncar 2010).

To address market challenges, most automakers continually update existing products and develop new ones. The development of these new products often requires substantial investment and faces immense competitive pressure and innovation risks. In the current environment where automotive products frequently and rapidly enter and exit the market, predicting product success has become increasingly important (Bharadwaj et al. 2017). Hence, the rapid transformation of the automotive industry, driven by the surge of new players, markets, and technology, prompts us to rethink previous assumptions in automotive sector.

This study delves into the dynamics of the automotive industry as numerous new products and brands continue to flood the market by investigating the impact of first-mover advantage and country-of-origin on consumer purchasing decisions within the Chinese automobile market. China is one of the largest and fastest-growing automobile markets in the world, with a significant push towards NEVs (Teece 2019). At the same time, in recent years, a large number of new Chinese car brands have also emerged in this market, including many NEV brands. It represents a marketplace where both global and domestic brands compete fiercely (Qian and Soopramanien 2014). Understanding consumer behavior in this market can provide insights into broader global trends and strategies for automotive companies.

Specifically, the following questions are aimed to answer in our study: (i) To what extent do first-mover advantages, related to both the model and the brand, affect sales performance in the automotive industry when many new participants enter the market? (ii) How significantly does the energy type of the vehicles affect their sales performance? (iii) In instances where first-mover advantages are present, can innovative later entrants in the market, such as NEVs and their emerging brands, mount a challenge and potentially offset these entrenched benefits? (iv) When many innovative car models and automakers enter the market, do consumers still show a preference for vehicles from developed countries? Are these preferences driven more by the country-of-origin effect,

or do first-mover advantages play a larger role? The intent behind these questions is to discover the complex interplay between market entry timing, technological innovation, and country-of-origin on consumer buying patterns in the dynamic environment of the automotive market.

By leveraging machine learning, we propose an innovative two-stage analytics framework, to enhance the robustness, accuracy, and interpretability of sales predictions and analytics. In the first stage, a meticulously crafted sales performance indicator offers a holistic view of the factors driving sales, mitigating biases inherent in relying solely on sales information. In the second stage, unlike traditional statistical approaches, we follow a machine learning pipeline to select the best classification method that can achieve superior accuracy by capturing nonlinear relationships (Bernstein et al. 2007) and eliminate confounding effects through Shapley values (Shapley 1953). In the process of innovative product management, this innovative, versatile and replicable framework demonstrates strong dynamic learning ability, adaptability and creativity (Verganti et al. 2020; Gama and Magistretti 2023). Therefore, this provides insights that extend beyond operational predictions to inform strategic decision-making with a deeper level of detail (Günther et al. 2022; Božič and Dimovski 2019; Athaide et al. 2024).

Our study not only contributes methodologically but also unveils fresh perspectives on the dynamics of first-mover advantage and country-of-origin effects when many new products and brands flood the market, thereby challenging preconceived notions regarding the interplay between a brand's establishment and its market entry. Firstly, we demonstrate that first-mover benefits, while not always simultaneous, can significantly impact sales performance. Our findings also confirm the persistence of first-mover advantages across both vehicle models and brands within the Chinese automobile market, addressing the limitation of previous studies that lacked a comprehensive analysis of this advantage from the three dimensions of model launch date, brand establishment date, and brand market entry date (Frynas et al. 2006). However, these entrenched ad-

vantages are surmountable, with innovative newcomers making discernible inroads. A shift in Chinese consumer preferences is evident, with reduced bias towards traditional fuel vehicles and an openness to new NEV brands. Secondly, the sustained preference for brands from developed countries, especially Germany and the United States, underscores the lasting influence of the country-of-origin effect. Additionally, we have compared the differences in contributions between brand first-mover advantage, model first-mover advantage, and brand country-of-origin, as well as their respective variances across different car classes and brand positions for the first time. Overall, the influence of the first-mover advantage appears to be more important than brand country-of-origin effects across most segmented markets. These insights are invaluable for automakers' innovative product management, strategic planning and marketing strategies, offering a compass for navigating the ever-changing market and consumer trends.

Next, Section 3.3 reviews the related literature and establishes hypotheses. Section 3.4 describes the data used in the study. Section 3.5 introduces the proposed analytics framework. Section 3.6 presents the analysis results. Section 3.7 discusses the results. Finally, Section 3.8 concludes the chapter.

3.3 Related work

This section first reviews studies focusing on the first-mover advantage and the influence of country-of-origin, where several hypotheses are developed that align with our research goals. It then continues with an examination of literature on automobile sales analytics and the metrics used to evaluate sales effectiveness.

3.3.1 First-mover advantage and country-of-origin

First-mover advantage – the edge gained by the initial significant occupant of a market segment – has been very well studied (Carpenter and Nakamoto 1989). This advantage often manifests as higher market share and profitability and is attributed to barriers such as brand recognition and customer loyalty that later entrants must overcome (Patterson 1993; Lambkin 1992; Coeurderoy and Durand 2004). Early entrants can cement technological leadership, corner resources, and secure customer attention more effectively than followers (Kardes et al. 1993). Nonetheless, the benefits of early entry may wear out as the market matures and competition intensifies (Huff and Robinson 1994). Meanwhile, late entrants could surpass first movers by adopting unique or enhanced product strategies (Besharat et al. 2016). In addition, in the automobile market, especially for foreign brands, there may be inconsistencies between the brand establishment date and the date the brand entry into other country markets (Zhao et al. 2005). A brand's international reputation is a critical factor influencing the timing of its entry into global markets (Omar et al. 2009). Surviving first-mover brands hold an advantage in establishing an exceptional brand reputation (Lieberman and Montgomery 1988). Therefore, in globalized industries, the brand establishment date may be more significant than its market entry date when entering new markets. Although previous studies have separately discussed the brand first-mover advantage based on these two key dates (Frynas et al. 2006; Patterson 1993), it is unclear whether their effects are consistent or which one provides a greater advantage. Meanwhile, previous research has shown that when the previous pioneer product continues to lead the next product, consumers show a strong preference for the previous pioneer product (Chang and Park 2013). This suggests that there also be product first-mover advantage in the market, such as shaping category standards, enhancing purchase intentions, and building brand loyalty (Bohlmann et al. 2002). Additionally, in markets characterized by incremental innovation, the brand first-mover advantage can significantly reduce the risk of complete failure when launching new products (Min et al. 2006). Therefore,

product first-mover advantages and brand first-mover advantages influence each other. However, we have reason to suspect that in the face of a large influx of innovative products into the market, brand first-mover advantages may play a more significant role. Nonetheless, few studies have compared the impacts of these two factors in the past. Given these dynamics and gaps, we hypothesize:

H1: *The first-mover advantages of a car brand and/or a car model have a positive impact on sales performance.*

H2: *The establishment date of a brand has a greater impact on sales performance in the automobile market than the date it enters a new market.*

H3: *The first-mover advantages of a brand contribute more to good sales performance than those of individual car models.*

When market and technological uncertainties are significantly high, the first movers will face great survival risks and new brands may gain greater brand preference (Wang and Xie 2014; Chang and Park 2013). Despite traditional barriers, companies like Tesla have demonstrated that new entrants, particularly in the NEV sector, are capable of disrupting well-established markets (Stringham et al. 2015). This shows that innovative late entrants can overcome the first-mover advantages of traditional vehicles, growing rapidly and reshaping market dynamics (Shamsie et al. 2004). Notably, supportive policies, increased investments from automakers, and a rising consumer embrace of NEVs have significantly propelled the swift expansion of the NEV sector (Zhang and Qin 2018). Concurrently, numerous NEV start-ups have surfaced, with some achieving

market valuations surpassing those of many established automotive brands (Jiang and Lu 2023). This trend suggests that NEVs and their emerging brands, as innovative late entrants, are perceived to possess more robust growth potential. Consequently, we put forward the following hypotheses:

H4a: *As innovative later entrants, NEVs can break the first-mover advantage of traditional fuel vehicles.*

H4b: *As innovative later entrants, emerging NEV brands can break the first-mover advantage of traditional automotive brands.*

A strong country-of-origin image can prevent potential consumers from turning to brands from other countries, while also moderating the effect of advertising on product evaluations, serving as both an informational cue and a source variable, highlighting its importance in marketing strategies (Verlegh et al. 2005). The positive country-of-origin image significantly influences consumers' intentions to purchase(Roth and Romeo 1992; Carneiro and Faria 2016). Historically, the automobile market has been dominated by brands from developed countries, enhancing the perceived quality of these brands among consumers. This perception has led consumers in emerging markets, such as China and India, to show a preference for automobiles from these developed countries (Sharma 2011; Hamzaoui Essoussi and Merunka 2007). However, the impact of country-of-origin can be diminished by exposure to other information cues (Agrawal and Kamakura 1999). Current research indicates that consumers in emerging markets, when using individual-oriented motivations such as personal satisfaction and enjoyment as consumption criteria, tend to prefer domestic brands (Zhang et al. 2024). Consumer

resistance to certain country-of-origin due to historical reasons and ethnocentrism will also affect automobile sales, marketing and advertising efforts. These occurrences create opportunities to change the competitive landscape (Sun et al. 2021; Russell and Russell 2006).

More importantly, consumer preferences for products with different levels of design novelty and technological innovation vary according to their country-of-origin. Design novelty and technological novelty are critical factors driving consumer purchase intentions (Arora and Arora 2017). Considering the extensive research on the country-of-origin effects and the notable shifts in consumer demand and product structures in the current automobile market, we propose the following hypothesis:

H5a: *Preference for automobiles from developed countries continues to be a significant trend among consumers.*

H5b: *Preference for automobiles from developed countries have undergone a significant shift.*

Car characteristics or brand image can be implicitly identified at different price levels according to the target automobile market segment (Saridakis and Baltas 2016). Consumer behavior varies between high-priced and low-priced products, with factors like preconceived beliefs and product attributes influencing consumers' price decisions. Consumers of high-priced products tend to prioritize perceived quality, with brand image having a significant impact on it (Lambert 1972; Jacoby et al. 1971). Because car models with larger-size or from luxury-brands often come with higher pricing (Carlson 1978), first-mover advantages and country-of-origin effects may have a greater impact on consumers of these high-end car models. Hence, the following hypotheses are then proposed:

H6a: *The first-mover advantages contribute more to good sales performance in high-end automobile markets.*

H6b: *Brand country-of-origin effects contribute more to good sales performance in high-end automobile markets.*

In addition, the product image established by market first-movers can influence COO effects. Consumers may develop a preference for the COO associated with first-movers in a specific product category (Gao and Knight 2007). Unfavorable COO effects may also weaken the positive impact of first-mover advantages. However, when faced with continuously growing competition, first-mover advantages may prove to be more enduring and significant than COO effects (Chen and Pereira 1999). Nonetheless, past studies have rarely conducted comparative and integrative analyses of first-mover advantages and COO effects based on market data. Therefore, we also proposed the following hypothesis:

H7: *The first-mover advantages contribute more to good sales performance in the automobile market than the brand country-of-origin effects.*

3.3.2 Automobile sales prediction and analytics

Accurate sales prediction and analytics provide valuable business insights that are crucial for a company's profitability and long-term survival, particularly during the growth phase of product innovation when major investment and marketing decisions are made (Martínez et al. 2020; Decker and Gnibba-Yukawa 2010). As summarized in Table 3.1, a range of methods have been employed in prior research on automobile sales

Table 3.1: Summary of sales predictive methods used in previous studies.

Author	**Date-related variable**	**Brand-related variable**	**Product-related variable**	**Other variables**	**Target variable**	**Method**
Du and Kamakura 2012	✓	✓		✓	Monthly sales	Structural dynamic factor-analytic model
Hülsmann et al. 2012	✓			✓	Yearly sales, quarterly sales, monthly sales	Linear regression, quantile regression, support vector machine, decision tree, k-nearest neighbor, random forest
Qian and Soopramanien 2014	✓				Quarterly sales	Linear regression, autoregressive integrated moving, exponential smoothing
Geva et al. 2017	✓	✓			Monthly sales	Least-squares linear regression, neural networks, support vector machines and random forest
Nunnari and Nunnari 2017	✓				Monthly sales	Neuro-fuzzy and feed-forward neural networks, non-linear autoregressive
Xia et al. 2020	✓	✓	✓	✓	Monthly sales	Linear regression, light gradient boosting, logistic regression, gradient boosting decision tree, decision tree, support vector machine, extreme gradient boosting
Afandizadeh et al. 2023	✓	✓	✓	✓	Monthly sales	Neural networks
Our study	✓	✓	✓		Monthly sales, performance indicator	Logistic regression, extreme gradient boosting, random forest, histogram-based gradient boosting, support vector machines and multilayer perceptron, SHAP

prediction and analytics. Time series analysis, including methods such as autoregressive integrated moving average (ARIMA) and exponential smoothing, is commonly used for short-term, seasonal forecasting of car sales. Additionally, linear regression has been used to examine causal factors that influence automobile (Du and Kamakura 2012; Qian and Soopramanien 2014). However, these traditional linear models may overlook complex interrelationships among variables, potentially compromising the precision of their predictions (Bernstein et al. 2007).

Compared with traditional statistical methods, predictive models based on machine learning have advantages in generating new theory, extending existing theories, comparing competing theories, developing measures, improving existing models, assessing relevance, and assessing predictability (George et al. 2016; Shmueli and Koppius 2011; Boudreau et al. 2004). And algorithmic decision-making based on machine learning are expected to make marketing smarter, more efficient, and more consumer-friendly (Herhausen et al. 2024). Machine learning has been recently applied to predicting automobile sales, using models such as neural network, support vector regression, random forest, and so on (Nunnari and Nunnari 2017; Afandizadeh et al. 2023; Geva et al. 2017). Since the linearity assumption may oversimplify real-world associational and structural relationships, many studies focusing on automobile sales forecasting have consistently demonstrated the superior predictive accuracy of machine learning in comparing machine learning models with traditional linear models (Xia et al. 2020; Hülsmann et al. 2012; Zhao and Hastie 2021). On the other hand, while high-performing machine learning models often rely on opaque, complex algorithms that are difficult to interpret, finding a middle ground between accuracy in interpretation and prediction continues to be a significant challenge (Rudin 2019; Messalas et al. 2020).

Most studies use sales as the sole indicator to gauge automobile sales performance. However, relying on a single indicator is not entirely appropriate, as it can be easily influenced by external factors (Behn 2003). For instance, an increase in sales could be influenced by a growing market size. The automobile market is also subject to notable seasonal fluctuations (Radas and Shugan 1998). To mitigate the bias resulting from relying on a single indicator, many studies have adopted a diverse array of performance indicators and multiple criteria decision-making approaches, which offer robust support for the development of improved decision-making processes (Ishizaka and Siraj 2018; Cooper and Kleinschmidt 1995; Moers 2005).

Hence, a combination of multiple indicators may produce a more accurate evaluation for automobile sales performance (Twyman et al. 2015). Sales ranking is also an important performance indicator as a statistical data in the retail industry that can indicate the relative position of a product's sales in the category (Garg and Telang 2013). There is an obvious correlation between sales ranking and product demand, but the relationship between sales ranking and sales is not linear (Carare 2012). Therefore, sales and sales ranking cannot represent the same meaning when they are used as single performance indicator. Aggregating these two indicators into a composite indicator to evaluate automobile sales performance may reduce the bias when sales is used as a single indicator.

3.4 Data

Data is collected from two automobile information platforms in China: PC Auto[1] and Dongchedi[2]. Our dataset includes monthly sales data for 1,009 car models in the Chinese automobile market from January 2016 to May 2021, with 29,233 observations in total. There are 103 domestic brands, 32 joint venture brands and 1 wholly foreign-owned brand in the dataset[3]. Among them, there are 3 domestic luxury brands, 11 joint venture luxury brands and 1 wholly foreign-owned luxury brand (see Appendix A). Due to the lack of a clear definition of luxury brands among domestic labels, the luxury brands in China were selected based on articles from Dongchedi[4]. In the Chinese automobile market, characterized by high consumption taxes and customs duties, there is no significant competition between imported cars and those from domestic brands, joint-venture brands or wholly foreign-owned brands. Given that the market share of imported cars is only around 5% (Shen et al. 2021), our study excludes imported cars.

Table 3.2 presents the variables and their definitions included in our dataset. Due to the heterogeneity of different market segments, we categorize the Chinese automobile market into 22 market segments based on brand luxury, car model size and car utility type. As car model size is derived from Dongchedi, we adopt the classification standard published therein (see Appendix B). In comparing the luxury and non-luxury

1. https://www.pcauto.com.cn
2. https://www.dongchedi.com
3. In China, joint venture brands in the automotive industry are partnerships between foreign and domestic companies to manufacture and sell vehicles within the country. These partnerships are largely driven by Chinese regulations that mandate foreign automakers to form joint ventures with local firms. This practice is prevalent among many of China's top-selling manufacturers, such as SAIC-GM-Wuling, Dongfeng Nissan, FAW-Volkswagen, and FAW Toyota, which are successful examples of joint ventures between Chinese and international auto manufacturers. Additionally, due to changes in Chinese regulations over the years, the Chinese government has allowed foreign brands to establish wholly-owned enterprises in China, enabling them to manufacture and sell vehicles locally. In other words, foreign automakers are now permitted to create wholly foreign-owned brands in China, with Tesla being the only brand in this category.
4. https://www.dongchedi.com/article/6930842800183476740
https://www.dongchedi.com/article/7054356619047502343?&aid=36&app_name=automobile&app=automobile&iid=3193462149352284&device_id=1283866180717960&zt=m_station_backflow

Table 3.2: Summary of variables in dataset.

Variables	Description
Car model	Name of the car model.
Brand	Brand name of the car model.
Minimum price	Minimum price for the car model.
Sales year	Year of sales of the car model.
Sales month	Month of sales of the car model.
Car utility type	Utility type of the car model: Sedan, SUV and MPV.
Car model size	Size of the car model: mini, small, compact, mid-size, large or full-size.
Brand luxury	Type of the car model brand: luxury and non-luxury.
Model launch year	Launch year of the first car model in the Chinese market.
Brand establishment year	Establishment year of the brand.
Brand enter China year	Year of brand officially entered the Chinese market.
Car model energy type	Energy type of the car model: fuel and NEV.
Brand energy type	Energy type of the brand: traditional brand and emerging NEV brand. The former means the car brand mainly produces traditional fuel vehicles at the beginning of its establishment while the latter means that this is an emerging car brand that only produces NEVs.
COO	Country of the automaker that established the brand.
Monthly sales	Monthly sales of the car model.
Sales ranking	Monthly sales rank of the car model in its segment market.

Table 3.3: Summary of the selected 12 market segments used for our data analysis.

Brand luxury	Car utility type	Car model size	Number of models	Number of observations
Luxury brand	Sedan	Compact	7	274
		Mid-size	15	586
		Large or full-size	12	528
	SUV	Compact	16	538
		Mid-size	23	628
		Large or full-size	7	115
Non-luxury brand	Sedan	Compact	207	6423
		Mid-size	67	2062
		Large or full-size	9	277
	SUV	Compact	180	5367
		Mid-size	93	2744
		Large or full-size	16	528

brands, we observe that there is no luxury branded mini sedan models, small sedan models, minivan models, and compact MPV models. Simultaneously, in the luxury brand market, the number of small SUV models and mid-size MPV models is limited. Additionally, it should be noted that internationally, car models larger than mid-size cars are defined as large or full-size models, while in China, they are often divided into two size-based subcategories. However, the subcategory with larger size contains very few models, usually imported ones, so we have also excluded it. In this study, large or full-size car models refer to the subcategory with relatively smaller size. Finally, as depicted in Table 3.3, to keep consistency between luxury and non-luxury markets, we ultimately retain only six classifications applicable to both luxury and non-luxury brands, resulting in a total of 12 market segments.

3.5 The proposed analytical framework

As depicted in Figure 3.1, we propose a comprehensive two-stage sales analytics framework, supported by a robust suite of machine learning algorithms. This framework aims to increase prediction accuracy, boost the interpretability of analytics, and provide critical business insights to automakers and marketers. The design of the framework is robust and flexible, allowing for the integration of diverse variables, clustering techniques, and predictive machine learning models, which makes it highly adaptable to various applications.

As mentioned in Section 3.3, some studies indicated that relying solely on a single performance indicator can introduce bias. To mitigate this, in the first stage of our study, we develop a reliable composite performance indicator using cluster analysis, incorporating both sales and ranking data. Popular clustering algorithms such as k-means, bisecting k-means, agglomerative clustering, and birch are compared. The op-

timal number of clusters is determined using the elbow method, which assesses curves of the sum of squares of error (SSE) across various market segments. To ensure uniform cluster counts and enhance clustering effectiveness, the piecewise regression method is applied to the SSE curve. Finally, the silhouette method is employed to choose the ideal number of clusters, based on the best clustering algorithm, the most effective sales performance indicators are obtained.

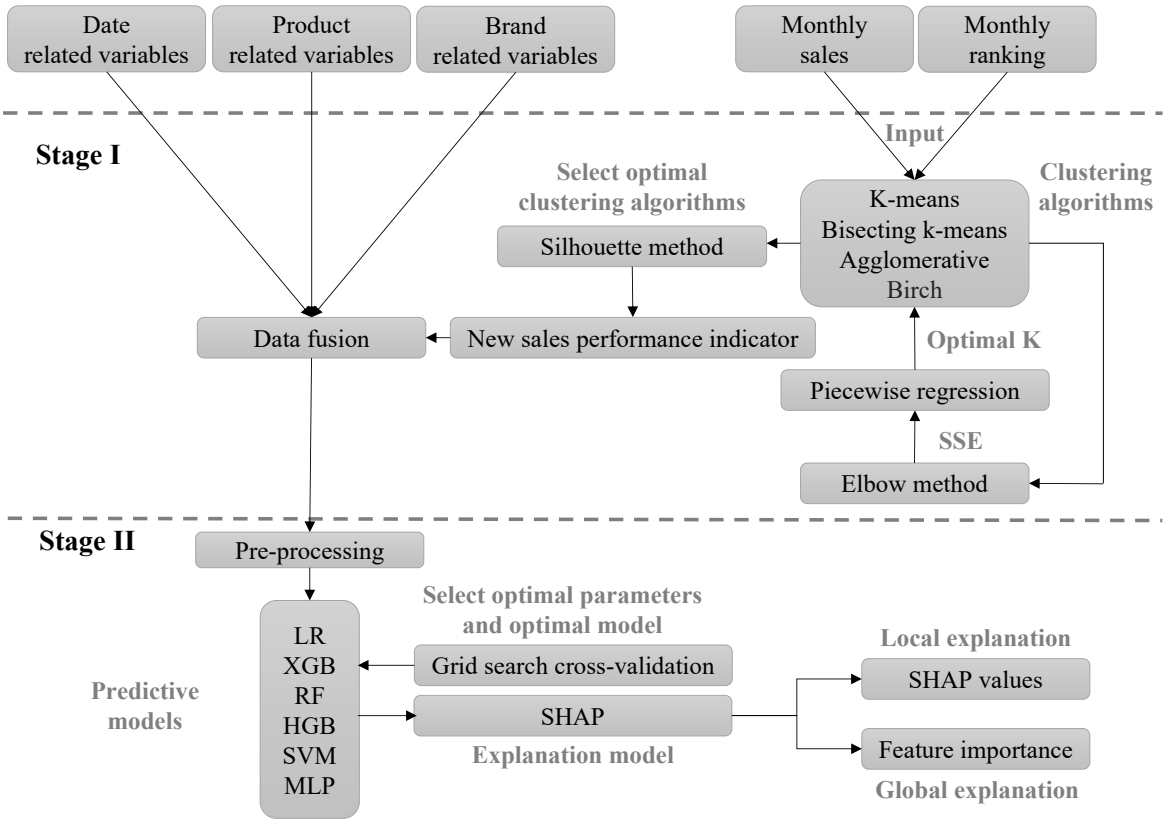

Figure 3.1: Schematic view of the proposed framework for automotive sales performance analytics.

In the second stage, predictive and explainable analytics are conducted to investigate the relationship between the composite performance metric and various influencing factors. Initially, data is pre-processed prior to training the machine learning models. Subsequently, machine learning classifier algorithms are utilized to predict automobile sales performance, including logistic regression (LR), XGBoost (XGB), random forest (RF), histogram-based gradient boosting (HGB), support vector machines (SVM), and multilayer Perceptron (MLP) for sales prediction. Since LR is present in both traditional statistical models and machine learning models, it serves as a benchmark for

comparing the performance of other machine learning models. In predictive analytics, grid search cross-validation is used to fine tune model hyperparameters. For interpreting the predictions made by the best model, SHAP (Lundberg and Lee 2017) is applied, which uses Shapley values from cooperative game theory to provide both local and global insights into how feature values contribute to predictions. Locally, SHAP values explain the individual impact of each feature on a prediction, where positive and negative values indicate positive and negative impacts, respectively. Globally, feature importance is determined by averaging the absolute SHAP values across all features, offering a comprehensive view of their effects.

3.6 Analysis of results

This section presents our analysis of the experimental results, confirming the hypotheses related to first-mover advantage, innovative later-mover strategies, brand country-of-origin effects, and biases resulting from reliance on a single performance indicator.

3.6.1 Sales performance indicator

Before generating new sales performance indicators, two car models randomly are selected from our dataset as examples, examined the fluctuations in sales, ranking, and market size to check whether employing either sales or ranking individually as performance indicators may result in biased outcomes. As illustrated in Figure 3.2, the sales of both models in February 2020 were notably lower than in other months, yet their rankings were higher. This discrepancy indicates that relying solely on sales as a measure of performance can lead to biased conclusions due to market size fluctuations.

Similarly, using rankings alone as the performance indicator does not accurately capture the actual sales dynamics. Instances may occur where a car model has high sales but a low ranking, or vice versa. Therefore, using either sales or rankings alone as performance indicators can introduce biases in evaluating sales performance.

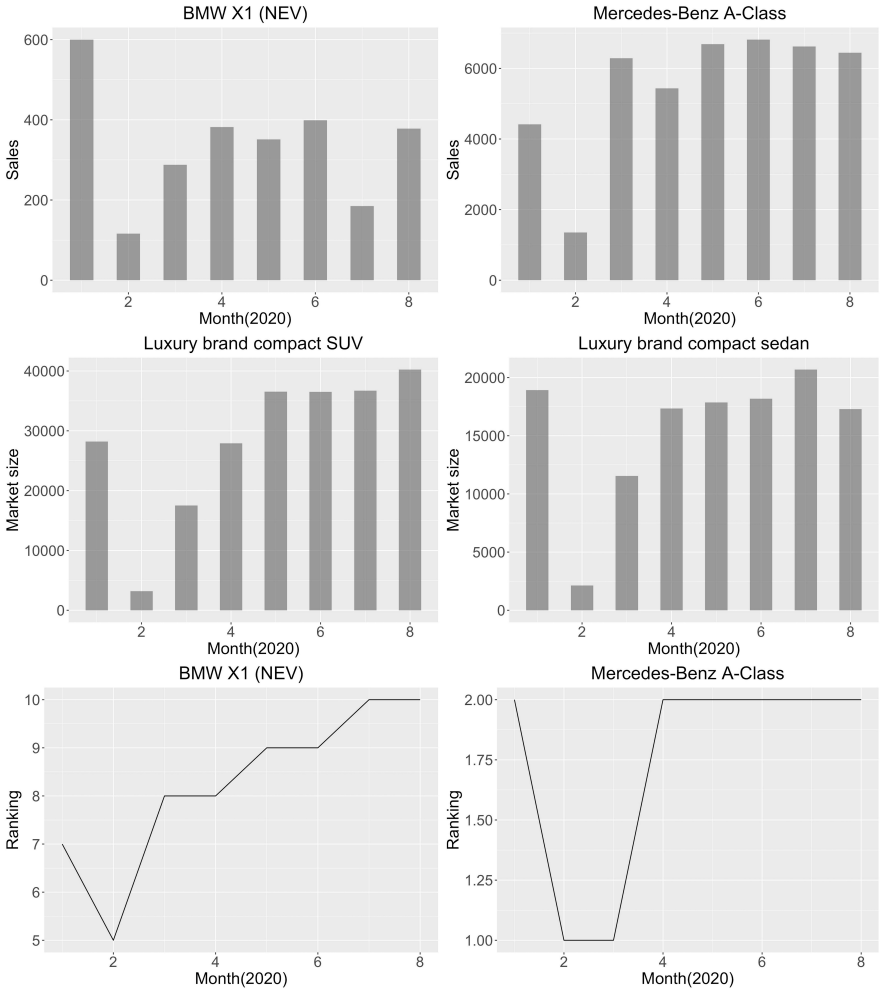

Figure 3.2: Illustrative examples of BMW X1 and Mercedes-Benz A-Class that shows bias of using solely sales or ranking as the sales performance indicator.

Figure 3.3 showcases the cluster analysis results for the sales performance indicator, revealing subtle differences in the optimal number of clusters across 12 market segments. The analysis identified three as the optimal number of clusters for each market segment following the fitting of the SSE curve. Among the algorithms tested, K-means clustering delivered the highest average silhouette score, validating its efficacy with the optimal clustering number and the silhouette method (detailed performances in Appendix C).

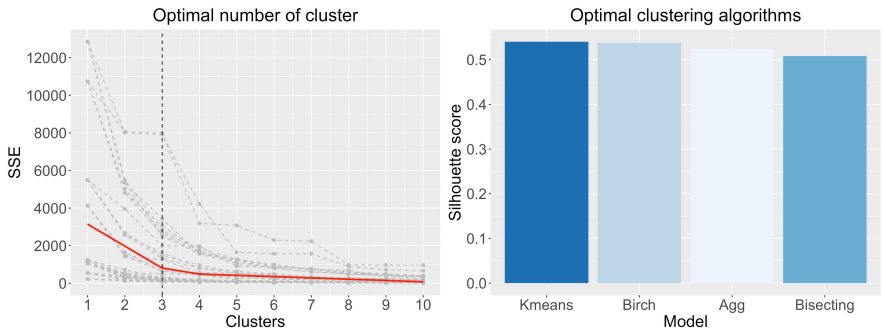

Figure 3.3: Cluster analysis of sales performance indicators.

Consequently, K-means clustering has been employed to generate the composite performance indicator, dividing sales performance into three categories: Good, Medium, and Bad. These classifications have been merged back into the original dataset to serve as the target variable for subsequent predictive analytics.

Table 3.4 shows the statistics for consistent and inconsistent sales performance evaluations within the clustering results when sales and ranking are taken as a single indicator of performance evaluation respectively. Meanwhile, it is compared with our clustering results using a composite indicator. It can be found that when the clustering results taking sales as a single indicator are consistent with those taking ranking as a single indicator, the clustering results using composite indicator are also highly consistent with their results. For example, in the data whose clustering results are all Good, our clustering results are all Good. Among the data whose clustering results are all Medium, 86.48% of the data based on the composite indicator is also clustered as Medium. Among the data whose clustering results are all Bad, 99.95% of the data based on the composite indicator are also clustered as Bad. Therefore, this proves that the use of composite indicator to evaluate performance has high accuracy.

Table 3.4: Compare the clustering results using sales and ranking as performance indicators respectively and the clustering results using new performance indicator.

Sales	**Ranking**	**Sales performance**	**Count**	**Percentage**	**Overall percentage**
Good	Good	Good	1514	100.00%	7.54%
		Medium	0	0.00%	0.00%
		Bad	0	0.00%	0.00%
		Total	**1514**	**100.00%**	**7.54%**
Good	Medium	Good	26	100.00%	0.13%
		Medium	0	0.00%	0.00%
		Bad	0	0.00%	0.00%
		Total	**26**	**100.00%**	**0.13%**
Medium	Good	Good	1449	43.07%	7.22%
		Medium	1915	56.93%	9.54%
		Bad	0	0.00%	0.00%
		Total	**3364**	**100.00%**	**16.76%**
Medium	Medium	Good	29	10.32%	0.14%
		Medium	243	86.48%	1.21%
		Bad	9	3.20%	0.04%
		Total	**281**	**100.00%**	**1.39%**
Medium	Bad	Good	0	0.00%	0.00%
		Medium	5	41.67%	0.02%
		Bad	7	58.33%	0.03%
		Total	**12**	**100.00%**	**0.05%**
Bad	Good	Good	0	0.00%	0.00%
		Medium	2328	100.00%	11.60%
		Bad	0	0.00%	0.00%
		Total	**2328**	**100.00%**	**11.60%**
Bad	Medium	Good	0	0.00%	0.00%
		Medium	4109	59.28%	20.47%
		Bad	2822	40.72%	14.06%
		Total	**6931**	**100.00%**	**34.53%**
Bad	Bad	Good	0	0.00%	0.00%
		Medium	3	0.05%	0.01%
		Bad	5611	99.95%	27.96%
		Total	**5614**	**100.00%**	**27.97%**
Total			**20070**		**100.00%**

It is noteworthy that in 11.60% of the automobile sales data, there is a significant disparity between the clustering results when sales are used as a single performance indicator versus when ranking is used. Specifically, when clustered based on sales, these data points are categorized as Bad, but when clustered by ranking, they are assessed as Good. This discrepancy clearly demonstrates the substantial bias that arises from using sales or ranking as standalone performance indicators. However, when employing a composite indicator, these data points are uniformly classified as Medium, which represents a more balanced and reasonable clustering outcome. Thus, we find that utilizing multiple performance indicators can significantly diminish the biases associated with single performance indicators.

3.6.2 Predictive and explainable analytics

After data preprocessing operations such as resampling and variable type conversion, the final target variable and predictor variables are shown in Table 3.5. For LR, RF, HGB, SVM and MLP, we select fundamental parameters and set the search range of those parameters based on the guideline provided by scikit-learn, and XGBoost is based on XGBoost documentation. Thus, the optimal parameters of these six models are obtained in this study (see Appendix D). By comparing the average accuracy of six models with optimal parameters in predicting automobile sales performance among all market segments in Figure 3.4, XGBoost is the optimal model in this study with the highest average accuracy reaching 0.889244.

By using the SHAP method to interpret the output of the XGBoost model in the experiment, SHAP values for each variable in each instance and the importance of each variable (see Appendix E) in predicting the sales performance of each classification in 12 market segments are obtained.

Table 3.5: Target variable and predictor variables.

Target variable	Predictor variables (Before processing)	Predictor variables (After processing)
Sales performance	Minimum price	Minimum price
	Sales year	Sales year
	Sales month	Sales month
	Model launch year	Model launch year
	Brand establishment year	Brand establishment year
	Brand enter China year	Brand enter China year
	Car model energy type	Model (Fuel) Model (NEV)
	Brand energy type	Brand (Traditional) Brand (NEV)
	COO	COO (America) COO (China) COO (Germany) COO (Japan) COO (Korea) COO (Italy) COO (France) COO (UK) COO (Sweden) COO (Czech Republic)

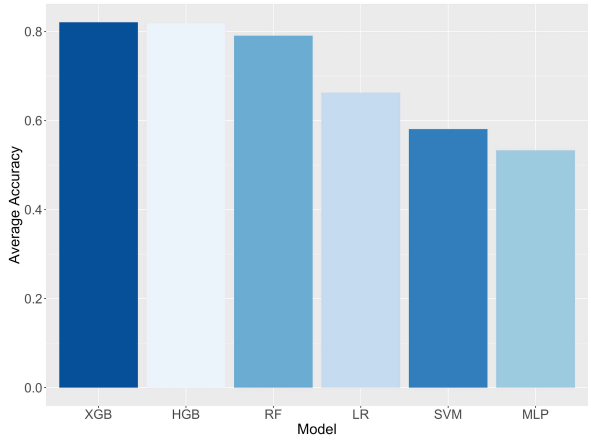

Figure 3.4: The average accuracy of six models with optimal parameters in predicting sales performance among 12 market segments.

3.6.2.1 The contributions of model launch year in different sales performance

Figure 3.5 illustrates the SHAP values of model launch year across three different sales performance categories in 12 market segments. These SHAP values represent the contribution of the model launch year to the prediction of corresponding sales performance category, showing their distribution across different launch years. Additionally, the figure highlights the trend of SHAP values as the model launch year becomes more recent. When the SHAP values exhibit an upward trend, it indicates that later model launch years are increasingly associated with the likelihood of the model's sales performance belonging to that category. Conversely, a downward trend suggests the opposite. If all SHAP values are zero or approach zero, it implies that the model launch year has no significant influence on predicting the sales performance category. It is important to note that although the model launch year may overall exhibit a negative contribution to certain sales performance categories when predicting sales performance, the determination of the first-mover advantage should be based on the trend of SHAP values as the model launch year progresses, rather than solely on actual SHAP values. For example, in a specific market segment, if all SHAP values of model launch year in the category of good sales performance are negative, it indicates that the launch year generally has a negative impact on predicting sales performance in this category. However, if SHAP values continues to decline as the model launch year progresses, it suggests that later-launched models experience an increasing negative effect, making it progressively harder for them to achieve good sales performance. This, in turn, implies that earlier-launched models are more likely to attain good sales performance, reflecting the presence of a first-mover advantage. Of course, the final conclusion should be drawn based on a comprehensive comparison of all sales performance categories.

Hence, we can identify the first-mover advantage through the following patterns. Firstly, models launched later show SHAP values trending downward in sales data for good-performing models and upward for bad-performing models, as seen in market segments such as the luxury brand compact sedan market, the luxury brand mid-size sedan market, the luxury brand large or full-size sedan market, the luxury brand compact SUV market,the luxury brand mid-size SUV market,the luxury brand large or full-size SUV market,the non-luxury brand compact sedan market, the non-luxury brand mid-size sedan market, the non-luxury brand mid-size SUV market and the non-luxury brand large or full-size SUV market. This pattern suggests that when many new products appear, earlier launches still tend to yield better sales outcomes and are less likely to perform poorly. Secondly, when models are launched later, SHAP values reveal no significant trend changes in the sales data of bad-performing models, yet indicate a decline for good-performing models, as seen in the non-luxury brand compact SUV market. This demonstrates that earlier launched models are more consistently successful, while launch year has a negligible impact on the performance of models with poor sales. In contrast, a late-mover advantage is apparent when models launched later improve their chances of achieving good sales performance and decrease their risk of bad performance. Obviously, except for the non-luxury large or full-size sedan market, the remaining 11 market segments clearly demonstrate first-mover advantages for car models.

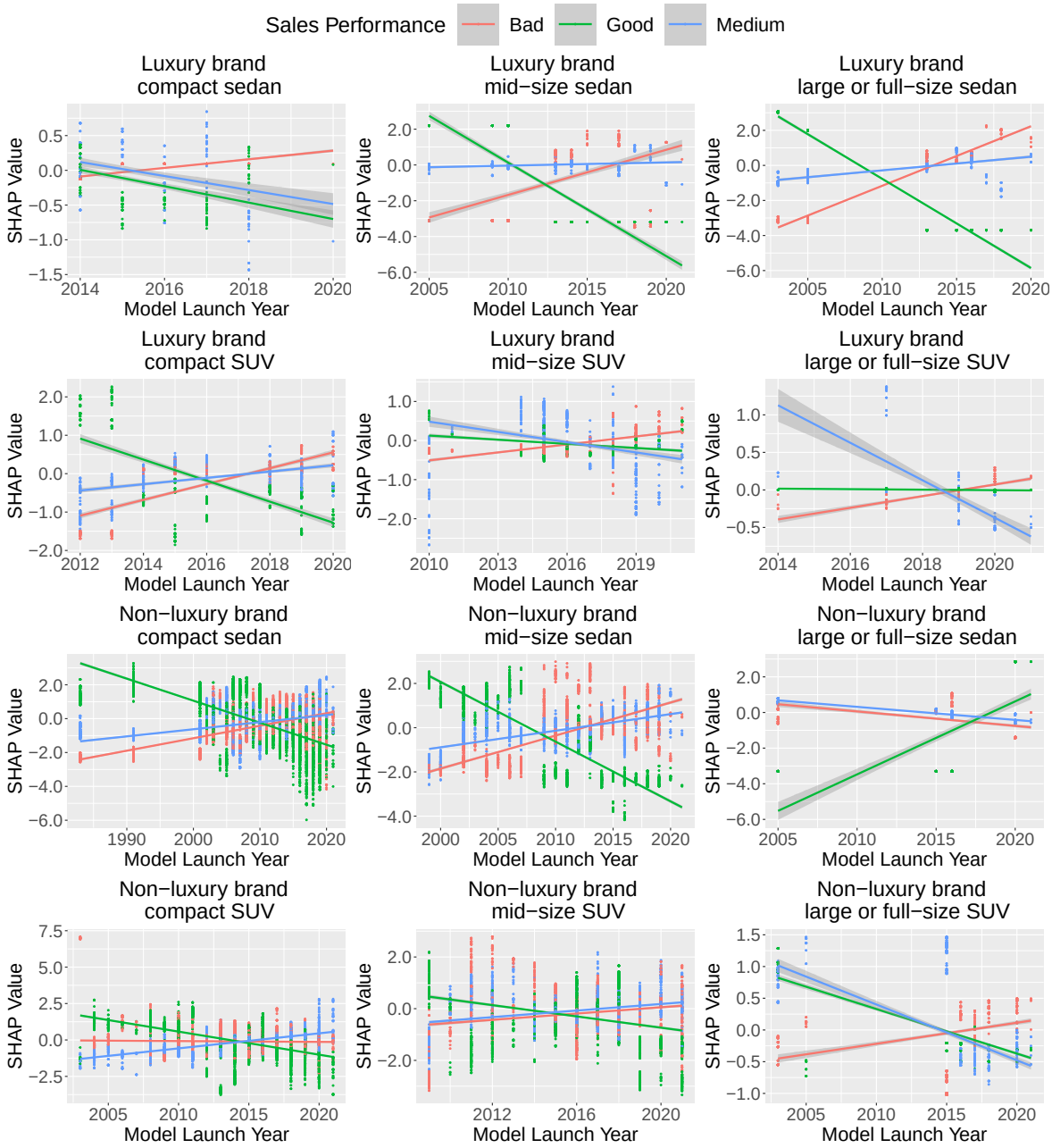

Figure 3.5: Distribution of SHAP values of model launch year with the change of model launch year in different sales performance classification in different market segments.

3.6.2.2 The contributions of brand establishment year and brand enter China year in different sales performance

In the Chinese automobile market, the establishment date of joint venture brands and wholly foreign-owned brands often differs from their market entry date in China. Therefore, when analyzing the first-mover advantages of brands, it is essential to separately examine the trends in SHAP values for both brand establishment year and brand enter China year. Similar to model first-mover advantages, certain criteria can be applied to determine their presence.

Firstly, models from brands established earlier or that entered the Chinese market earlier are generally more likely to exhibit good sales performance and less likely to show bad sales results. Secondly, if the relevant variables make no contribution in the good sales performance category, meaning the SHAP values show no trend and remain overall zero, then the brand first-mover advantage should be reflected in the ability to achieve medium sales performance. In this case, models from brands that were established or entered the Chinese market earlier are more likely to sustain medium sales performance and are less prone to bad sales results.

On the other hand, a late-mover advantage can be observed in scenarios such as: (i) when a brand is established later or enters the Chinese automobile market later, its models are more likely to exhibit good sales performance and are less likely to show bad sales performance; (ii) if a brand's good sales performance is unaffected by its establishment date or the date it entered the market, then models from brands that were established later or entered the market later are typically found to have medium sales performance and are unlikely to display bad sales results.

Furthermore, two important cases should be considered. First, the SHAP values of brand establishment year or brand enter China year may exhibit a consistent trend across both good and bad sales performance categories. Although the SHAP values are nonzero and show a clear trend, the fact that this trend is identical—meaning that the likelihood of achieving both good and bad sales performance increases or decreases simultaneously as brand establishment year or brand enter China year changes—indicates that there is no obvious first-mover or late-mover advantages in the market. Second, if the brand establishment year or brand enter China year has no contribution to any sales performance category—that is, if the SHAP values remain zero across all categories (good, medium, and poor)—it implies that these dates have no impact on sales performance.

Table 3.6 summarizes the presence of brand first-mover advantages in different market segments based on the distribution of SHAP values for brand establishment year as it changes and the distribution of SHAP values for brand enter China year as it changes in different sales performance classifications within various market segments (refer to Appendix F for details). Notably, the brand establishment date and date of brand entry into China did not exhibit a consistent trend in shaping the brand first-mover advantage in automobile sales performance across the 5 market segments. According to the brand establishment year, brand first-mover advantages exist in 9 market segments. Based on the brand enter China year, brand first-mover advantages are identified in 6 market segments. Therefore, we can determine the existence of brand first-mover advantages in sales performance even if the brand establishment date is often not consistent with the date of brand entry into a new market. Therefore, from both the perspective of car models and brands, H1 is supported.

Table 3.6: Summary of brand first-mover advantages based on brand establishment year and brand enter China year in 12 market segments.

Market segments	**Establishment year**				**Enter China year**			
	Good	*Medium*	*Bad*	***Summary***	*Good*	*Medium*	*Bad*	***Summary***
Luxury brand compact sedan market	↓	↑	↑	√	↔	↔	↔	—
Luxury brand mid-size sedan market	↑	↓	↑	O	↔	↓	↑	√
Luxury brand large or full-size sedan market	↓	↑	↑	√	↓	↑	↑	√
Luxury brand compact SUV market	↓	↓	↑	√	↓	↓	↑	√
Luxury brand mid-size SUV market	↓	↑	↑	√	↓	↑	↑	√
Luxury brand large or full-size SUV market	↑	↓	↓	×	↑	↑	↓	×
Non-luxury brand compact sedan market	↓	↓	↑	√	↓	↑	↑	√
Non-luxury brand mid-size sedan market	↓	↓	↑	√	↓	↑	↓	O
Non-luxury brand large or full-size sedan market	↔	↓	↑	√	↔	↑	↓	×
Non-luxury brand compact SUV market	↑	↓	↑	O	↑	↑	↑	O
Non-luxury brand mid-size SUV market	↓	↑	↑	√	↓	↑	↑	√
Non-luxury brand large or full-size SUV market	↓	↓	↑	√	↓	↑	↓	O

Note: First, ↑, ↓, and ↔ represent the increasing trend, decreasing trend, and overall zero values of the SHAP values for the corresponding variables as the brand establishment year or the brand enter China year becomes later, respectively. The presence of first-mover advantages can be inferred from the above trends. For example, the SHAP value of brand establishment year shows a decreasing trend over time in good sales performance category, while it shows an increasing trend for those with medium and bad sales performance. This suggests that the earlier a brand was established, the more likely its models are to achieve good sales performance, and the less likely they are to fall into the medium or bad performance categories — indicating the existence of first-mover advantages. Second, √ represents first-mover advantages, × represents late-mover advantages, O represents no obvious first-mover or later-mover advantages, — represents no impact.

3.6.2.3 The contributions of car model energy type and brand energy type in different sales performance

In all 12 market segments we selected, there are 10 market segments that have NEVs and there are 8 market segments that have NEV brands (see Appendix G). By comparing model launch year of traditional fuel vehicles and NEVs, except for the luxury brand large or full-size SUV market, the launch date of NEVs is generally later than that of traditional fuel vehicles in other 9 market segments. By comparing brand establishment year and brand enter China year of traditional brands and emerging NEV brands, emerging NEV brands are later than traditional brands in all 8 market segments. Consequently, NEVs and emerging NEV brands represent innovative later entrants in Chinese automobile market.

For a binary categorical variable, the two generated dummy variables are mutually exclusive. In SHAP calculations, the SHAP values of these two dummy variables are usually numerically equal but have opposite signs. Therefore, in the SHAP output, only one of the dummy variables is typically reported to represent the overall contribution of the binary variable. Clearly, the SHAP values of the dummy variables Model(NEV) and Brand(NEV) represent the contributions of car model energy type and brand energy type, respectively. Based on the distribution of SHAP values of car model energy type in Figure 3.6, it is easy for us to compare the differences in contributions between two different types of car model energy in achieving various sales performance categories. Clearly, in the luxury brand large or full-size sedan market and non-luxury brand large or full-size sedan market, car energy type has no impact on achieving any of the three sales performance categories. That is, compared to new energy vehicles, traditional fuel vehicles do not exhibit any advantages in sales performance. Therefore, in these two segments, new energy vehicles, as innovative later entrants, have broken the first-mover advantage of traditional fuel vehicles, supporting H4a.

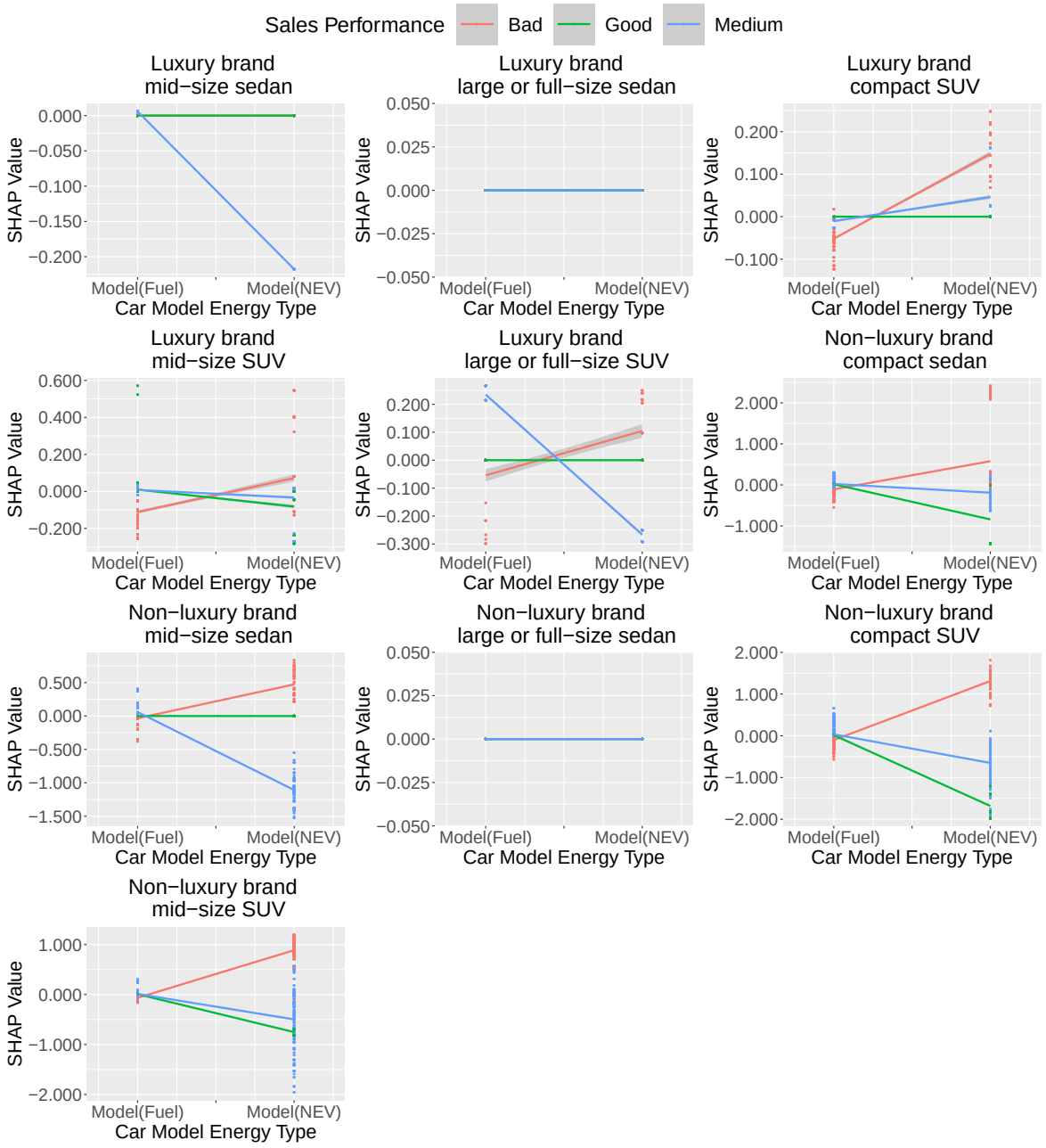

Figure 3.6: Distribution of SHAP values of car model energy type in market segments that have new energy vehicles.

Moreover, in the luxury brand mid-size sedan market, while car model energy type does not contribute to good or bad sales performance, traditional fuel vehicles still hold an advantage in the medium sales performance category. Additionally, in the remaining seven segments, new energy vehicles are more likely to fall into the bad sales performance category. Thus, in these eight segments, the first-mover advantage of traditional fuel vehicles in sales performance remains.

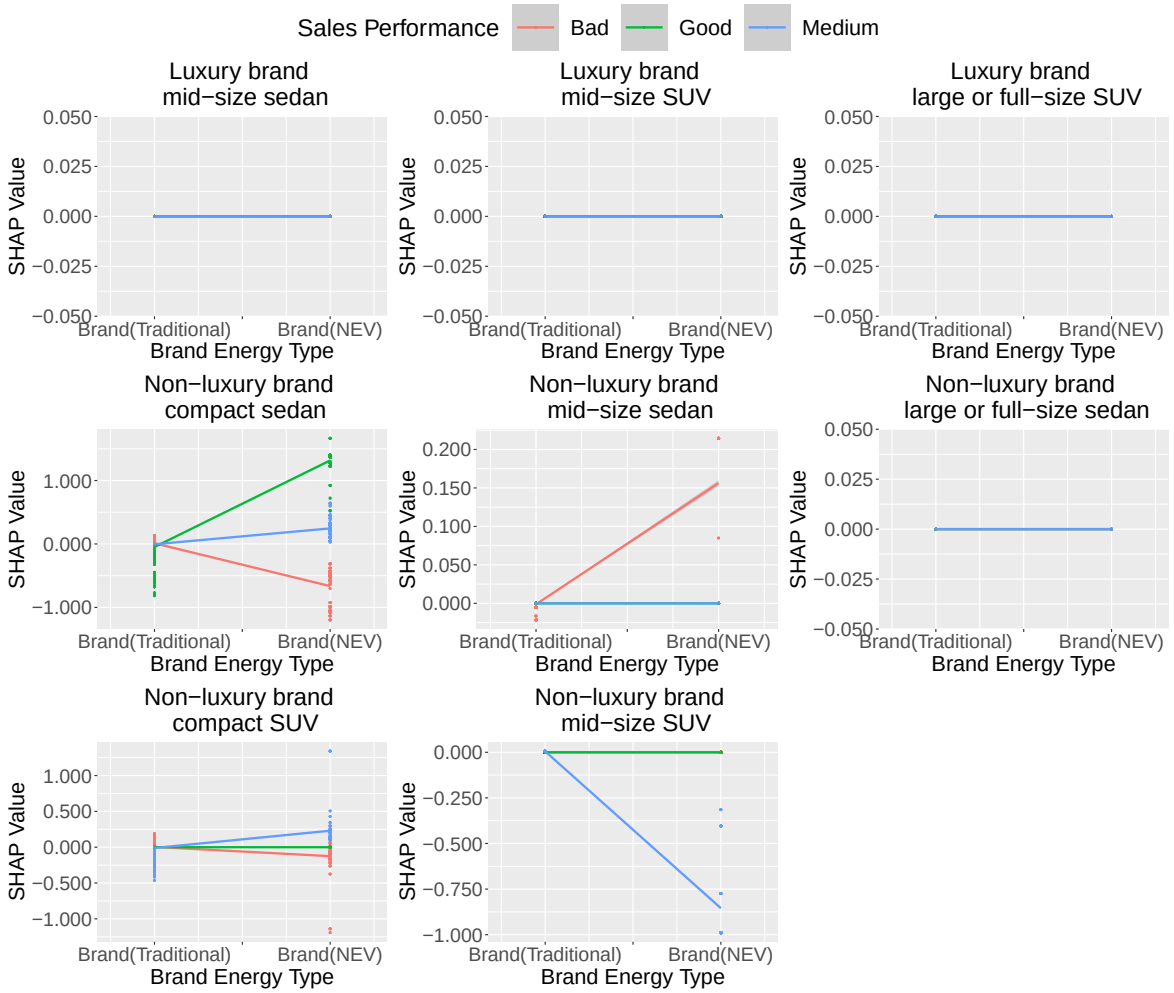

Figure 3.7: Distribution of SHAP values of brand energy type in market segments that have emerging new energy vehicle brands.

Similarly, Figure 3.7 shows that in the luxury brand mid-size sedan market, luxury brand mid-size SUV market, luxury brand large or full-size SUV market and non-luxury brand large or full-size sedan market, brand energy type does not have any influence on automobile sales performance. This indicates that in these 4 market segments, NEVs produced by emerging NEV brands have the same opportunities to achieve good, me-

dium, and bad sales performance as traditional fuel vehicles or NEVs produced by traditional brands. Secondly, in the non-luxury brand compact sedan market, compared to traditional brands, models from emerging new energy vehicle brands are more likely to achieve good and medium sales performance and less likely to fall into the bad sales performance category. In the non-luxury brand compact SUV market, while brand energy type has no impact on achieving good sales performance, models from emerging new energy vehicle brands are still more likely to achieve medium sales performance and less likely to experience bad sales performance. Therefore, this proves that as innovative later entrants, emerging NEV brands have broken the brand first-mover advantage of traditional brands in 6 market segments, supporting H4b. However, in the non-luxury brand mid-size sedan market and non-luxury brand mid-size SUV market, traditional brands still have more advantages.

3.6.2.4 The contributions of brand country-of-origin in different sales performance

Then, as shown in Table 3.7, we summarize the impact of each brand country-of-origin on automobile sales performance in all market segments, based on distribution of SHAP values of different brand country-of-origin. Obviously, the brand country-of-origin has no influence on the sales performance only in the luxury brand compact sedan market and the non-luxury brand large or full-size sedan market. This may be related to the fact that only brands from 2 to 3 countries sell these two types of car models in these market segments in China. In addition, only in the luxury brand mid-size SUV market, China, as the brand origin, has a positive impact on automobile sales performance, but in the remaining market segments, brands that have a positive impact on sales performance are all from developed countries. This indicates that the establishment of many new brands from China has not significantly changed consumer

preferences for brands from developed countries. Therefore, in most market segments, H5a is supported. H5b is supported only in the luxury brand mid-size SUV market. Notably, consumer preferences for brand country-of-origin are heterogeneous across different market segments.

Table 3.7: Countries of origin that have significant positive or negative effects on automobile sales performance in 12 market segments.

Market Segments	**Positive Effect**	**Negative Effect**
Luxury brand compact sedan		
Luxury brand mid-size sedan	America	France, Japan, Sweden, UK
Luxury brand large or full-size sedan	Germany, Sweden	America, UK
Luxury brand compact SUV	America, Sweden	Germany
Luxury brand mid-size SUV	Germany, America, China, Sweden	Japan, UK
Luxury brand large or full-size SUV	America	
Non-luxury brand compact sedan	America, Germany, Czech Republic	China, France, Italy
Non-luxury brand mid-size sedan	America, Germany, Japan	France
Non-luxury brand large or full-size sedan		
Non-luxury brand compact SUV	Germany, Czech Republic, Japan	America, France, China
Non-luxury brand mid-size SUV	America, Germany, Czech Republic, Japan	Korea, China
Non-luxury brand large or full-size SUV	Germany	Japan

3.6.2.5 Comparison of features' contributions

To further explore the differences in the effects of brand establishment date and the date of brand entry into new market, we compare their respective contributions in each instance in predicting good sales performance across different market segments. As shown in Table 3.8, the contribution of brand establishment year and brand enter China year significantly differs in 91.67% of cases. Only in the non-luxury brand large or full-size sedan market, both did not contribute to the good sales performance. Obviously, in 66.67% of market segments, brand establishment date is significantly more influential than brand entry date into a new market, while the reverse holds true in only 25% of market segments. Therefore, H2 is supported in most market segments.

In addition, variables related to brand first-mover advantages, model first-mover advantages, and brand country-of-origin are integrated individually, and in Figure 3.8, their combined feature importance across all segments is calculated to compare the difference between high-end and low-end markets. Considering that large or full-size car models are larger and more expensive than compact and mid-size car models within the same brand, and luxury brand models have higher brand premiums in the same car size category, luxury brand market, large or full-size sedan market, and large or full-size SUV market can be regarded as the high-end car market in this study. Clearly, for luxury brands, brand first-mover advantages have a greater impact in the compact sedan, large or full-size sedan, compact SUV, and large or full-size SUV markets compared to non-luxury brands. Model first-mover advantages also have a greater impact in the mid-size sedan and compact SUV markets compared to non-luxury brands, while the country-of-origin effect is only higher in the compact SUV and mid-size SUV markets compared to non-luxury brands.

Table 3.8: Comparing the contribution of brand establishment year and brand enter China year in in each instance in predicting good sales performance in 12 market segments based on t-test.(All values are rounded to four decimal places.)

Market Segments	**Establishment year**		**Enter China year**		**p-value**	
	Mean	**Variance**	**Mean**	**Variance**	**One-tail**	**Two-tail**
Luxury brand compact sedan	2.3899	0.1946	0.0000	0.0000	0.0000	0.0000
Luxury brand mid-size sedan	0.6955	0.6368	0.0000	0.0000	0.0000	0.0000
Luxury brand large or full-size sedan	0.1107	0.0077	0.4087	0.0150	0.0000	0.0000
Luxury brand compact SUV	2.4783	2.5320	0.1648	0.0742	0.0000	0.0000
Luxury brand mid-size SUV	0.9907	0.1667	0.1890	0.0169	0.0000	0.0000
Luxury brand large or full-size SUV	1.1668	0.3247	0.7254	0.3330	0.0000	0.0000
Non-luxury brand compact sedan	0.9730	0.6266	0.5578	0.2034	0.0000	0.0000
Non-luxury brand mid-size sedan	0.6388	0.2903	0.8575	0.3440	0.0000	0.0000
Non-luxury brand large or full-size sedan	0.0000	0.0000	0.0000	0.0000	NA	NA
Non-luxury brand compact SUV	0.9981	0.5802	0.3631	0.1787	0.0000	0.0000
Non-luxury brand mid-size SUV	1.0165	0.3917	0.3180	0.0573	0.0000	0.0000
Non-luxury brand large or full-size SUV	0.0943	0.0214	0.6735	0.3675	0.0000	0.0000
The proportion of market segments with significant differences (Significance level is 0.05.):					91.67%	91.67%

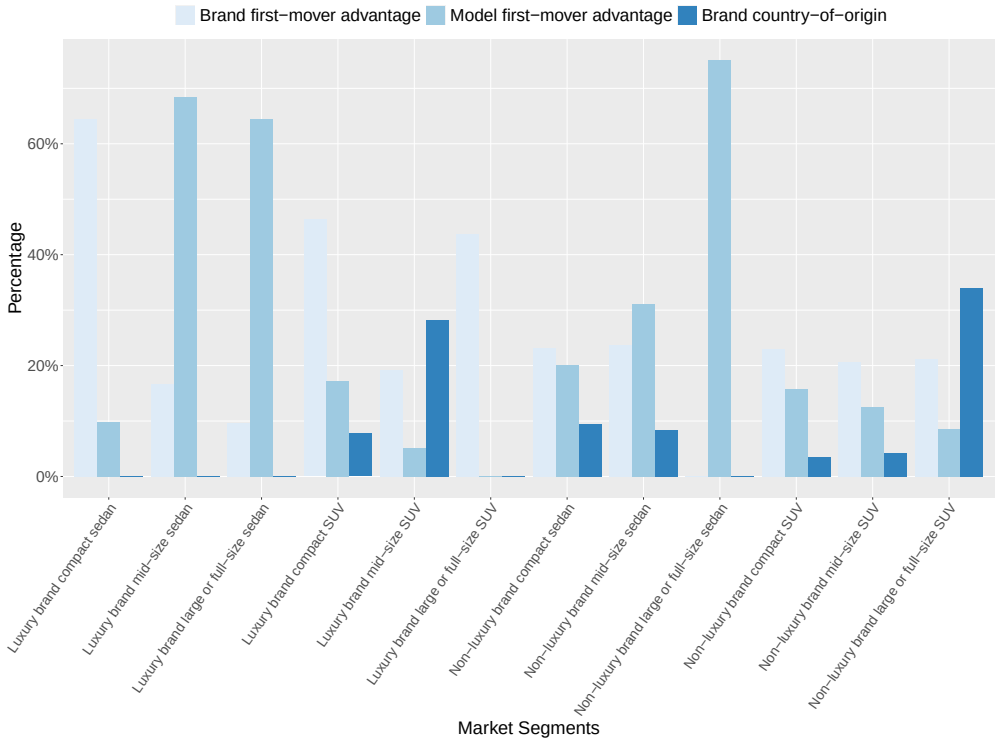

Figure 3.8: The feature importance proportions of brand first-mover advantage, car model first-mover advantage and brand country-of-origin in predicting good sales performance across different market segments.

In the large or full-size sedan market, the impact of brand first-mover advantage and country-of-origin effects is not greater than in the other two smaller-size car markets, model first-mover advantage is only slightly lower than mid-size sedan market in luxury brand sedan market. In the mid-to-large SUV market, the impact of brand first-mover advantage is lower than in the compact SUV market but higher than in the mid-size SUV market; model first-mover advantage is lower than in the other two smaller-size car markets; and the country-of-origin effect is only higher in the non-luxury brand market compared to the other two smaller- size car markets. Therefore, the impacts of first-mover advantage and brand country-of-origin effects in the high-end car market do not consistently show a significantly higher trend than in the low-end market, indicating that H6a and H6b are not supported.

Table 3.9: Comparing the contribution of the brand first-mover advantage, model first-mover advantage, and brand country-of-origin in each instance in predicting good sales performance in 12 market segments based on ANOVA and Tukey HSD test.(All values are rounded to four decimal places.)

Market segments	Mean(Standard Error)			p-value in Tukey HSD test(All ANOVA p-values $\ll 0.0001$.)		
	Brand first-mover advantage	Model first-mover advantage	Brand country-of-origin	Brand first-mover advantage VS Model first-mover advantage	Brand first-mover advantage VS Brand country-of-origin	Model first-mover advantage VS Brand country-of-origin
Luxury brand compact sedan	2.3899(0.0245)	0.3643(0.0106)	0.0000(0.0000)	0.0000	0.0000	0.0000
Luxury brand mid-size sedan	0.6955(0.0326)	2.8470(0.0192)	0.0000(0.0000)	0.0000	0.0000	0.0000
Luxury brand large or full-size sedan	0.4774(0.0058)	3.1964(0.0292)	0.0000(0000)	0.0000	0.0000	0.0000
Luxury brand compact SUV	2.5495(0.0647)	0.9462(0.0217)	0.4265(0.0105)	0.0000	0.0000	0.0000
Luxury brand mid-size SUV	1.1089(0.0153)	0.2960(0.0051)	1.6322(0.0131)	0.0000	0.0000	0.0000
Luxury brand large or full-size SUV	1.6860(0.0879)	0.0060(0.0008)	0.0000(0.0000)	0.0000	0.0000	0.9962
Non-luxury brand compact sedan	1.4204(0.0100)	1.2278(0.0093)	0.5738(0.0064)	0.0000	0.0000	0.0000
Non-luxury brand mid-size sedan	1.3928(0.0152)	1.8344(0.0149)	0.4967(0.0071)	0.0000	0.0000	0.0000
Non-luxury brand large or full-size sedan	0.0000(0.0000)	3.1429(0.0102)	0.0000(0.0000)	0.0000	1.0000	0.0000
Non-luxury brand compact SUV	1.2711(0.0102)	0.8776(0.0081)	0.1969(0.0036)	0.0000	0.0000	0.0000
Non-luxury brand mid-size SUV	1.1882(0.0117)	0.7262(0.0116)	0.2458(0.0049)	0.0000	0.0000	0.0000
Non-luxury brand large or full-size SUV	0.7591(0.0247)	0.3064(0.0110)	1.2132(0.0200)	0.0000	0.0000	0.0000

Finally, Table 3.9 presents a comparative analysis of the contribution of the brand first-mover advantage, the model first-mover advantage, and the brand country-of-origin in each instance in predicting good sales performance across 12 market segments. First, an ANOVA test was conducted to simultaneously examine the differences in contributions among these three key factors. The results showed that in all 12 market segments, the p-values were significantly lower than 0.0001. Therefore, the Tukey HSD test was applied as a post hoc analysis of the ANOVA results to examine pairwise differences in the contributions of these three factors. Obviously, regarding the comparison between brand-first-mover advantage and model-first-mover advantage, results of the Tukey HSD test indicates that there is a significant difference in the contribution of the two. Upon separate analysis of all market segments, the contribution of brand first-mover advantage was significantly higher than that of model first-mover advantage in 8 market segments, and only in 4 market segments was the opposite phenomenon. Consequently, H3 is supported in most market segments.

Additionally, in predicting good sales performance, the contribution of the brand first-mover advantage is significantly higher than that of the brand country-of-origin in 75% of market segments, significantly lower in 17% of market segments, and neither shows a significant contribution in 8% of the market segments. Similarly, the contribution of the model first-mover advantage to good sales performance is significantly higher than that of the brand country-of-origin in 75% of market segments, significantly lower in 17% of market segments. However, in 8% of market segments, although the model first-mover advantage contributes more to good sales performance than the brand country-of-origin, the difference is not significant. The results of the Tukey HSD test also confirm the significance of these differences, which shows that H7 is supported in most market segments.

3.7 Discussion

Our study primarily explores theories concerning first-mover advantages and the country-of-origin effects in markets characterized by the rapid entry of innovative products and brands, addressing existing gaps in these areas. We innovatively examined the relationship between first-mover advantages and automobile sales through three variables: the establishment date of the brand, its entry date into new markets, and the product launch date. Our findings indicate that first-mover advantages, both at the brand and car model levels, significantly influence sales performance. Additionally, our research on NEVs and emerging NEV brands demonstrates that innovative late entrants can disrupt the first-mover advantages held by established models and brands. Moreover, the presence of limited players in certain market segments introduces greater market uncertainty. This uncertainty about market and consumer demand provides opportunities for later entrants, potentially explaining the absence of first-mover advantage in these specific markets (Lieberman and Montgomery 1988).

Regarding the country-of-origin of brands, our study shows that it independently affects consumer purchase intentions, with varied impacts on sales performance. While Chinese consumers still generally favor brands from developed countries across most segments of the Chinese automobile market in the face of the establishment of many new brands, preferences for specific countries vary significantly. Additionally, previous research indicates that domestically manufactured vehicles by Chinese brands tend to generate greater brand excitement compared to those by foreign brands produced domestically (Fetscherin and Toncar 2010). This heightened enthusiasm may be shifting consumer preferences toward Chinese brands, particularly evident in the luxury mid-size SUV market.

It is worth emphasizing that our study addresses the limitations of previous research, which lacked comprehensive comparisons between the impacts of first-mover advantage and brand country-of-origin. We enhance these comparisons by examining their respective contributions across different car classes. First, our findings reveal that both the establishment date and market entry date of a brand significantly impact sales performance, with the establishment date having a more pronounced effect in a greater number of market segments. Second, our analysis shows that in most markets, the first-mover advantages of brands and car models play a more crucial role in achieving strong sales performance than the country-of-origin effect. Moreover, the first-mover advantage of brands is often more pivotal than that of car models in many markets. Lastly, while previous research indicates that consumers of high-priced products prioritize brand image, our analysis indicates that in most automobile market segments, the first-mover advantage and brand country-of-origin do not have significantly different impacts on high-end models compared to low-end models.

Finally, we introduce a novel two-stage framework that utilizes cluster analysis to refine performance metrics and integrates predictive models with SHAP for generating interpretable insights. This framework effectively uncovers complex underlying relationships between variables and provides clear, actionable explanations. As an adaptable and interpretable machine learning approach, it fulfils the needs of businesses eager to leverage machine learning techniques for analysing business issues. Armed with precise sales forecasts, valuable business insights, and a deep understanding of consumer demands, companies can swiftly adjust their product offerings and marketing strategies to secure a competitive edge.

3.8 Summary

Base on the proposed novel analytics framework using machine learning, our study in this paper provides significant insights into the role of first-mover advantages and brand country-of-origin effects, underscoring their importance as crucial intangible assets and marketing variables. We find that strong country-of-origin effects, positive first-mover advantages, and transformative innovation are key to automakers. A deep understanding of how these factors influence various market segments enables automakers to pinpoint competitive edges, develop more effective product portfolios, and craft efficient marketing strategies.

To enhance the applicability and robustness of the conclusions, future research could extend these findings by incorporating a broader range of markets. Studies could also explore additional variables that influence consumer purchasing decisions, such as environmental concerns or geopolitical factors, which were outside the scope of this analysis. Furthermore, integrating alternative analytical methods or mixed approaches could provide a more comprehensive view of the data and help reconcile the depth of machine learning insights with the clarity of traditional statistical analysis. Such studies would not only validate the current findings but also offer a richer, more nuanced understanding of the dynamics at play in global automobile markets.

Chapter 4

When is the temptation too strong? Analyzing the timing of positive fake review manipulation

4.1 Abstract

The unethical practice of posting fake online consumer reviews is increasingly prevalent, yet the role of time in this manipulation remains underexplored. This study examines when firms are most likely to generate positive fake reviews, considering key performance indicators such as sales, market size, brand market presence duration, and product lifecycle. We introduce a two-stage machine learning approach for fake review detection and apply it to the Chinese automobile market. Our findings reveal that firms are most likely to post positive fake reviews when a product or brand reaches relatively high sales levels, during the mid-to-late stages of a vehicle model's lifecycle, and in the early stages of brand creation. Conversely, firms are less inclined to manipulate reviews when a brand first enters a new market, and no clear link exists between market size and fake review timing. This study contributes to business ethics by emphasizing the temporal dimension in fake review detection. Methodologically, our machine learning

approach offers an automated, scalable solution for identifying fake reviews on social media. Theoretically, we highlight time as a crucial factor for regulators and platforms to monitor, enhancing fraud detection and intervention strategies. Our findings inform ethical business practices, regulatory policies, and the development of responsible digital marketplaces.

4.2 Introduction

With the rise of social media platforms, online consumer reviews have become a vital component of marketing communication, shaping consumer perceptions and purchase decisions (Chen and Xie 2008). While these reviews are intended to reflect genuine consumer experiences, they have also become a tool for deceptive marketing practices (Logsdon and Patterson 2009). The increasing prevalence of fake reviews, where firms manipulate ratings to influence potential buyers, raises significant concerns in business ethics. Misleading information in marketing has long been an ethical issue (Wilson et al. 2022), but the ease of digital manipulation amplifies its impact, necessitating stronger scrutiny.

While word-of-mouth marketing is ethically acceptable when based on voluntary and authentic feedback, firms now frequently incentivize consumers (Choi et al. 2017) or employ automated systems to generate positive fake reviews (Kumar et al. 2022). Such manipulation misleads consumers, distorts competition, and erodes trust in online reviews (Yoon and Lee 2014). When low-quality firms artificially boost their ratings, they create entry barriers for high-quality competitors, undermining fair market competition (He et al. 2022b). As a result, detecting and understanding fake reviews is critical for maintaining ethical business practices.

Despite extensive research on the motivations and consequences of fake reviews (Park et al. 2023; Sahut et al. 2024; Zaman et al. 2023; Aylsworth 2022), little attention has been paid to when firms manipulate reviews. Section 4.3 provides a systematic review and summary of existing studies in the relevant field, clearly identifying specific gaps in the current research. Moreover, while various detection methods exist, many social media platforms fail to implement them or even allow fake reviews to persist. Examining the timing of fake review manipulation offers valuable insights for consumers and platforms to better identify fraudulent activities. Given that most online reviews are predominantly positive (Chevalier and Mayzlin 2006), this study focuses on the strategic timing of positive fake reviews in relation to firms' key performance indicators, including sales, market size, the duration of brand market presence and product lifecycle.

To achieve this objective, we propose a two-stage machine learning analytics framework that combines Bidirectional Encoder Representations from Transformers (BERT) with Positive-Unlabeled (PU) learning. For the former, BERT (Kenton and Toutanova 2019) is a deep learning natural language processing algorithm based on the Transformer architecture that can learn contextual information bidirectionally, enabling it to understand the deep relationships within the text and obtain more accurate text representations. For the latter, PU learning (Bekker and Davis 2020) aims to infer the labels of unlabeled samples by combining the information from known positive examples and unlabeled samples. Our framework integrates text feature extraction, manual sample labeling, and automated fake review detection, offering a scalable solution for social media platforms. Additionally, by incorporating explainable analytics, our framework provides critical insights into the timing of positive fake review manipulation, helping regulators and platforms detect and mitigate unethical review practices more effectively.

We apply this method to the Chinese automobile market, one of the largest and most competitive automotive sectors globally, offering a rich source of online review data. Our study addresses a key gap in research by analyzing the timing of positive fake review manipulation and its connection to firms' key performance indicators, such as sales, market size, and the longevity of models and brands. Our findings indicate that firms are most likely to manipulate positive fake reviews when vehicle or brand sales reach relatively high levels. However, the timing of this manipulation varies across market segments as market sizes shift. Additionally, firms tend to engage in fake review manipulation during the mid-to-late stages of a model's lifecycle and the early phases of brand establishment. In contrast, brands entering a new market are less likely to engage in fake review manipulation initially, suggesting that firms may prioritize building trust before resorting to deceptive strategies.

Methodologically, our study advances fake review detection in three ways. First, by integrating manual labeling rules, BERT's ability to interpret text numerically, and PU Learning's efficiency in processing large-scale unlabeled data, the proposed framework achieves superior detection accuracy. It provides an automated, scalable solution for social media platforms, enhancing their ability to identify and mitigate unethical business practices while supporting regulatory enforcement. Second, the study highlights the importance of product-centered, review-centered, and reviewer-centered features, demonstrating their crucial role in improving fraud detection accuracy. Finally, the manual labeling rules and labeled review dataset generated in this study serve as a valuable resource for future research in fake review detection and analysis.

Theoretically, our study contributes to business ethics by identifying timing as a critical factor in predicting when firms engage in fake review manipulation. More broadly, it offers a new perspective on unethical business practices, suggesting that the primary driver of fraudulent behavior is not an immediate business necessity but the perceived

absence of risk. Our findings indicate that regardless of market position or size, firms are more likely to manipulate fake reviews when the risk of detection is low. This insight underscores the need for stronger regulatory measures and detection frameworks to deter unethical behavior in digital marketplaces.

The remainder of this chapter is structured as follows: Section 4.3 reviews the literature on fake reviews, their ethical implications, and detection methods. Section 4.4 details our proposed detection model. Section 4.5 presents the data, methodology, and results. Section 4.6 discusses key findings, while Section 4.7 outlines contributions to business ethics. Finally, Section 4.8 summarizes the study.

4.3 Related literature

Fake online reviews, characterized by false, misleading, or deceptive content, are often manipulated by businesses or merchants to gain a competitive advantage (Wu et al. 2020). While consumers may occasionally post fake reviews for personal reasons, the vast majority are strategically incentivized by firms, using monetary rewards such as discounts or coupons (Zaman et al. 2023). Consequently, most fake reviews are orchestrated or influenced by enterprises, particularly positive fake reviews, which are designed to artificially enhance brand perception and consumer trust. Online reviews, essentially serving as authentic product information written by users after using the product, play a significant reference role for potential consumers (Chen and Xie 2008). Positive online reviews help consumers form favorable perceptions of products, increasing purchase intentions (Vana and Lambrecht 2021). Meanwhile, the product and brand information conveyed in positive online reviews contributes to the formation of a strong brand image in consumers' minds, assisting them in differentiating companies based on quality, thereby having a significant impact on brand equity and enhancing

their purchase intentions (Ananthakrishnan et al. 2023; Chakraborty 2019). Previous studies have demonstrated the effectiveness of positive fake reviews in enhancing consumer willingness, with the condition that the fake reviews must be highly deceptive and difficult to detect (Song et al. 2023; Zhao et al. 2013).

However, the manipulation of positive fake reviews constitutes an unethical form of covert marketing (Nelson and Park 2015), misleading consumers into purchasing products they may not have otherwise chosen (Park et al. 2023). Such practices undermine consumer autonomy and violate business ethics (Aylsworth 2022). While past research has explored the motives, methods, and consequences of fake review manipulation, there has been limited discussion on the temporal aspects—when firms are most likely to engage in this unethical behavior. Given that timing is a critical component of marketing strategy (Chen and Xie 2008), it is important to investigate how different market conditions and firm performance indicators influence the likelihood of fake review manipulation. As illustrated in Figure 4.1, we explore the strategic timing of positive fake review manipulation across multiple dimensions, including market size, product sales, brand sales, product launch duration and brand market presence duration. Based on these factors, we develop a theoretical framework and formulate the following hypotheses.

Larger market sizes intensify competition, increasing pressure on firms to differentiate themselves and maintain profitability (Vives 2008). Moreover, heightened competition often leads to lower average mark-ups and greater productivity demands, further motivating firms to engage in deceptive practices (Badinger 2007). As firms in larger markets face stronger incentives to manipulate positive reviews to maintain or improve their standing, we hypothesize that firms operating in larger markets are more inclined to generate positive fake reviews.

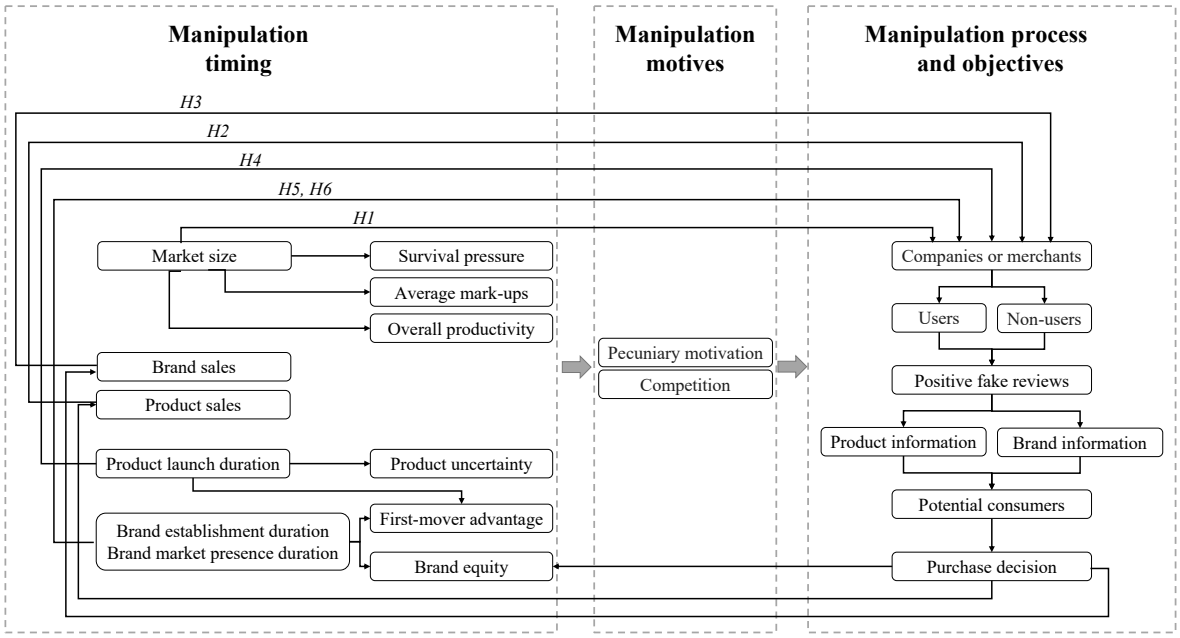

Figure 4.1: Theoretical framework for firms manipulating positive fake reviews.

H1: *Firms are more likely to manipulate positive fake reviews when market size is large.*

An increase in market share often helps companies achieve higher profit margins and reduce marketing costs (Buzzell et al. 1975). The growth of market share is closely linked to the increase in product sales and brand sales. In many cases, positive fake reviews are seen as substitutes for advertising and have a significant effect on boosting sales (He et al. 2022b). Some studies have shown that the impact of online reviews on profits is decreasing (Zhao et al. 2013). However, positive online reviews still play a key role in enhancing brand strength and brand equity, which subsequently boosts the sales of the brand and all its products. Meanwhile, low sales often reflect weak brand strength (Ho-Dac et al. 2013). This means that all firms still have a strong incentive to manipulate positive fake reviews in order to enhance brand strength and sales, particularly those behind products or brands with low sales.

More importantly, existing research has shown that perceived brand strength is negatively correlated with the likelihood of consumers doubting positive reviews, meaning that positive reviews associated with low-sales products or brands are more likely to be viewed with suspicion (Ko and Bowman 2023). At the same time, consumers' decision-making styles are often significantly influenced by product involvement (Bauer et al. 2006). When dealing with high-involvement products like automobiles—those with high prices, complex features, and long lifespans – consumers tend to approach related online review information with greater caution (Geva et al. 2017; Wang et al. 2023b). Consumers of such products have a higher ability to perceive risks (Hong 2015). However, when consumers perceive a higher intent to manipulate online reviews, it leads to more negative product attitudes, thereby reducing their purchase intention (Karabas et al. 2021). Therefore, when sales are low, brands and merchants face greater negative consequences from manipulating positive fake reviews, which leads to a weaker motivation to engage in such manipulation. In contrast, strong brands with high sales carry less risk in this regard, giving them a stronger motivation to manipulate reviews. In addition, it is well known that high sales of certain products do not necessarily indicate high brand sales. Brand sales can also be influenced by the number of product types offered. Therefore, when discussing the relationship between sales and the manipulation of positive fake reviews, it is essential to distinguish between product sales and brand sales. Based on the above perspectives, we propose the following hypotheses:

H2: *Firms are more likely to manipulate positive fake reviews when product sales are at relatively high levels.*

H3: *Firms are more likely to manipulate positive fake reviews when brand sales are at relatively high levels.*

Firms launching new products must establish credibility and consumer interest, making positive online reviews a critical tool for early adoption. Since new products face uncertainty, firms may use fake reviews to mitigate perceived risks and attract early adopters (Vana and Lambrecht 2021). This is particularly crucial in industries where first impressions significantly shape long-term consumer trust. However, early-stage fake reviews also face higher scrutiny, making timing and volume key strategic considerations. Hence, we posit:

H4: *Firms are more likely to manipulate positive fake reviews in the early stages of a product launch.*

Newly established brands require strong brand equity to compete effectively. Positive online reviews can enhance brand image and consumer trust, encouraging firms to engage in fake review manipulation during early brand development stages (Mitra and Jenamani 2020). As consumers rely heavily on early reviews to assess new brands, firms may strategically manipulate fake reviews to gain an initial advantage. However, if overused, fake reviews could backfire by raising consumer skepticism. Thus, we propose:

H5: *Firms are more likely to manipulate positive fake reviews in the early stages of brand creation.*

When a brand enters a foreign market, initial brand awareness and trust-building are critical (Swami and Dutta 2010). Firms may prioritize legitimate marketing efforts initially but later resort to fake reviews to sustain momentum. The unfamiliarity of foreign consumers with the brand makes them more dependent on online reviews, increasing the incentive for firms to manipulate fake reviews. However, brands that enter highly regulated or culturally skeptical markets may face higher risks in engaging in such practices. Therefore, we hypothesize:

H6: *Firms are more likely to manipulate positive fake reviews in the early stages of a brand entering the foreign market.*

These hypotheses will be empirically tested using our proposed machine learning analytical framework. Detecting fake reviews has been a persistent challenge, leading to the development of various machine learning approaches. Traditional supervised learning models, such as support vector machines (SVMs) and Naïve Bayes classifiers, require large amounts of labeled training data to achieve high accuracy (Khurshid et al. 2018). Unsupervised methods, including LSTM-autoencoders (Saumya and Singh 2022) and topic-sentiment models (Dong et al. 2018), can identify anomalies in textual patterns but often struggle when fake reviews closely resemble genuine ones. Semi-supervised techniques, such as co-training (Li et al. 2011) and PU learning (Bekker and Davis 2020), strike a balance by leveraging a small set of labeled data while inferring patterns in unlabeled data. Beyond classification methods, feature engineering has played a crucial role in improving fraud detection. Most studies rely on review-centric and reviewer-centric features, such as linguistic characteristics, sentiment polarity, and reviewer credibility (Kumar et al. 2022). However, product-centered features, despite their potential to enhance detection accuracy, remain underutilized in existing research. Additionally, deep learning models, particularly BERT-based text representation, have demonstrated superior performance in capturing contextual nuances within reviews (Cui et al. 2021; Kenton and Toutanova 2019; Subakti et al. 2022). Our study builds on prior work by incorporating a novel combination of product-centered, review-centered, and reviewer-centered features within a two-stage machine learning approach. By integrating BERT for text representation and PU learning for fraud detection, our method enhances both detection accuracy and scalability, addressing key gaps in existing research. Furthermore, by focusing on the timing of fake review manipulation, we extend the ethical discourse on when and why firms engage in deceptive practices.

4.4 Method

Our study proposes a two-stage machine learning approach that integrates BERT and PU learning to detect fake reviews. The framework systematically refines labeled samples and iteratively improves classification accuracy. Additionally, we employ SHapley Additive exPlanations (SHAP) to provide interpretability for the model's predictions. As illustrated in Figure 4.2, our approach consists of two key stages: sample initialization and fake review detection with iterative sample updating. The first stage establishes a small but reliable set of labeled samples, while the second stage expands the dataset using an iterative PU-learning mechanism, refining classification performance through multiple iterations.

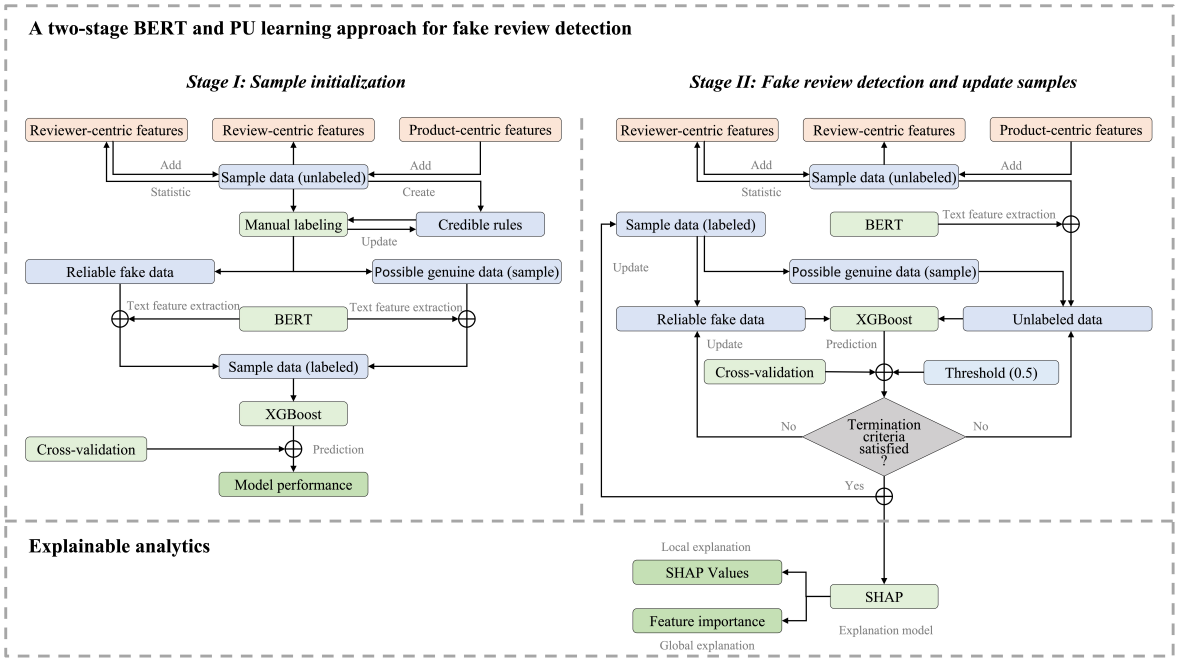

Figure 4.2: Fake review detection method and explainable analytics.

In the stage of sample initialization, the labeled sample data is primarily established through three main tasks. Firstly, richer variables are expanded from the traditional review-centric data by incorporating product-centric and reviewer-centric features. For the review text, 768-dimensional features, such as named entities, syntactic relations, word context information, and sentiment information, are extracted using BERT and

added to the labeled review data. Secondly, precise rules are employed to manually identify reliable fraudulent reviews and potential genuine reviews from the unlabeled sample data. The initial rules for identifying fraudulent reviews are formulated based on a deep understanding of the review data and fundamental automobile features. These rules are continuously updated during the manual labelling process to better capture the characteristics of fraudulent reviews. Thirdly, the effectiveness of the data is further validated through binary classifier applied to the labeled sample data for prediction. Considering XGBoost's high efficiency and excellent performance (Chen and Guestrin 2016), we adopted it as the classifier in this study.

In the stage of fake review detection and sample updating, the first task is to complete the same feature expansion work as in the initial stage. Secondly, the focus shifts to continuously identifying reliable positive examples from unlabeled data, which are initially considered negative instances, using PU learning with a small-scale reliable positive example dataset. This process automatically detects fake reviews and updates the labeled sample data for subsequent fraud detection tasks. Initially, reliable fake data within the sample dataset are treated as positive examples. Newly added data, augmented with corresponding features, along with potential genuine data from the sample dataset, are collectively considered negative examples. Iterative predictions are then made using a classifier based on the existing positive and negative examples. During each iteration, the data predicted to be positive examples is added to the initial positive examples for that round and removed from the negative examples. The next prediction cycle then commences, repeating this process iteratively until the termination condition is met, thereby completing the fake detection process. Ultimately, the continuously identified reliable positive examples during the iterative process represent fake reviews, while the remaining negative examples represent potentially genuine reviews.

Finally, in explainable analytics, based on the results of the last iteration, SHAP methods are used to obtain local and global explanations of the contributions of important variables in determining whether a review is fake, while avoiding the influence of multicollinearity (Lundberg and Lee 2017). Among this, the SHAP values for each variable output by the SHAP method represent the contribution of that variable to the prediction of the instance being classified as a positive example.

4.5 Experiments and analysis of results

To evaluate our proposed fake review detection framework, we conduct a series of experiments using an extensive dataset from the Chinese automobile market. This section outlines the data sources, preprocessing steps, and the machine learning methods employed to identify fake reviews. We first describe the dataset construction and initial labeling process, followed by the iterative fake review detection approach using PU learning. Finally, we present key findings through explainable analytics, highlighting the relationships between fake review manipulation and factors such as sales performance, market size, and brand maturity.

4.5.1 Data and sample initialization

Raw datasets in this study are sales data and unlabeled online review data over the period from 2016 to 2022 that we previously published in IEEE BigData 2023 (2023 IEEE International Conference on Big Data). To enhance the dataset, we incorporated reviewer-centric, review-centric, and product-centric variables, including review anonymity status, brand establishment duration, brand market entry duration, model

launch duration, discount information, sales at the time of review, brand sales at the time of review, market segment sales, overall review text, and overall sentiment score. The sales-related variables were derived through matching techniques and basic statistical analysis of previously disclosed sales datasets. Since the overall review text consolidates multiple textual fields into a single variable, we excluded the original text variables from further analysis in the fake review detection process. Sentiment scores for the overall review text were computed using SnowNLP, a widely used Python library for Chinese sentiment analysis based on Bayesian classification methods (Tang et al. 2020). Additionally, we employed BERT to extract 768-dimensional contextual features from the review text, enhancing the model's ability to capture linguistic nuances. The final processed dataset consists of 36 variables (see Appendix H), with key variables of interest summarized in Table 4.1.

Table 4.1: Important variables and the objectives for using them.

Variables	**Objective**
Sales at review	To study the relationship between the timing of automakers or dealers manipulating positive fake reviews and car model sales.
Brand sales at review	To study the relationship between the timing of automakers or dealers manipulating positive fake reviews and brand sales.
Market segment sales at review	To study the relationship between the timing of automakers or dealers manipulating positive fake reviews and market size.
Model launch duration	To study the relationship between the timing of automakers or dealers manipulating positive fake reviews and model launch duration.
Brand establishment duration	To study the relationship between the timing of automakers or dealers manipulating positive fake reviews and brand establishment duration.
Brand enter China duration	TTo study the relationship between the timing of automakers or dealers manipulating positive fake reviews and brand market presence duration.

Among them, by considering these variables comprehensively, we can derive initial rules for manually labeling data. First, the average fuel consumption of gasoline passenger cars cannot be 0 liters per 100 kilometers nor can it exceed 100 liters per 100 kilometers. Similarly, the average electricity consumption of pure electric vehicles cannot be 0 kWh

Table 4.2: Rules for detecting fake online reviews.

R1	Average energy consumption is 0 or significantly higher than normal levels.
R2	Review lag time is less than 0.
R3	Mismatch in energy consumption type.
R4	Purchase price mentioned in the review text does not match the recorded purchase price.
R5	Discounts mentioned in the review text do not match the recorded price.
R6	Mismatch in car model features.
R7	Mileage mentioned in the review text does not match the recorded mileage.
R8	Purchase date mentioned in the review text does not match the recorded purchase date.
R9	Average energy consumption mentioned in the review text does not match the recorded average energy consumption.
R10	Car model mentioned in the review text does not match the purchased car model.
R11	Clear semantic contradictions in the review text.

per 100 kilometers nor can it exceed 100 kWh per 100 kilometers. For plug-in hybrid and extended-range electric vehicles, although they can be driven by both electric power and fuel, it is also impossible for them to consume neither electricity nor fuel, or have an average energy consumption far beyond normal values. Second, since we collect reviews from users who purchased and used the car model, the purchase date of the car model cannot be later than the date the review was posted. Third, the model's characteristics should correspond to its energy type. For example, the average energy consumption of a fuel vehicle should be the fuel consumption per 100 kilometers, not the electricity consumption per 100 kilometers. Fourth, for consumers who post genuine reviews, the information about the car model mentioned in the review should be consistent with the various details that the online review platform requires to be separately filled out, such as the model name, model characteristics, mileage, purchase date, and average energy consumption. Similarly, genuine online reviews are often carefully written and published by consumers based on real experiences after using a specific car model, and are unlikely to contain contradictory statements. For example, the review might highlight the model's spaciousness when discussing its advantages, but then seriously criticize the car model's small space when talking about its space. Therefore, based on the understanding of car models in the Chinese automobile market, we established 11

rules (see Table 4.2) to identify fake reviews and manually labelled the sampled data after sampling. To better study the timing of manipulating fake reviews, we excluded reviews where car sales could not be matched due to inconsistent car model names. Similarly, since there is no available sales dataset for imported cars in the market, and imported cars account for less than 5% of sales in the Chinese automotive market (Shen et al. 2021), we also removed reviews related to imported cars. The final online review data we used contained 179,532 unlabelled data and 26,634 labelled data, with the labelled data consisting of 12,547 fake reviews and 14,087 possibly genuine reviews.

Based on the processed 36 variables, we use XGBoost to make classification predictions on the manually labelled sample data. For this basic classifier, the F1 score is about 0.72 and the accuracy is about 0.78, which proves that the sample data can be effectively applied to the fake review detection task.

4.5.2 Fake review detection and sample updating

Fraud detection tasks often face the challenge of imbalanced sample distributions, making it essential to select appropriate evaluation metrics. We used the F1 score as the primary metric to assess model performance throughout the iterative process. To optimize efficiency, we established a termination criterion where the F1 score threshold was set to 0.9, and the process would stop if no performance improvement was observed for 10 consecutive iterations. As shown in Figure 4.3, the model's performance improved over successive iterations. By the eighth iteration, the classifier achieved an F1 score exceeding 0.9 and an accuracy above 85%, indicating that the model effect-

ively distinguished fake reviews from genuine ones. In total, the experiment identified 90,326 fake reviews. When combined with the manually labeled fake reviews, the dataset contained a total of 102,873 fake reviews, of which 102,791 were classified as positive fake reviews.

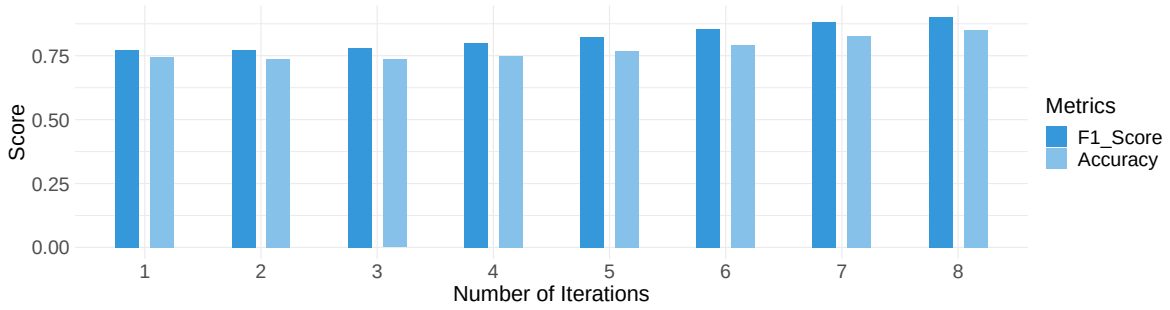

Figure 4.3: Model performance at each iteration.

4.5.3 Explainable analytics

By using the SHAP method to interpret the results of the last iteration of the classifier, the contribution of each variable in each instance to predicting fake reviews (SHAP values) can be obtained. According to the SHAP method's definition, feature importance is determined by averaging the absolute values of a variable's SHAP values across all instances. Hence, we can calculate the importance of each variable in predicting fake reviews and the importance of each variable in predicting fake reviews in positive reviews (see Appendix I). It is evident that most of the product-centered features, review-centered features, and reviewer-centered features we added have demonstrated significance, indicating their crucial role in improving the accuracy of fake review detection.

As target variables for studying the timing of manipulating positive fake reviews, we extracted the corresponding feature importance of Brand enter China duration, Sales at review, Brand sales at review, Model launch duration, Market segment sales at review review and Brand establishment duration. In addition, as shown in Table 4.3,

Table 4.3: Feature importance of the target variable, along with the proportion and rankings of feature importance.

Feature	Importance	Proportion	Rank
Brand enter China duration	0.62499891	10.27%	4
Sales at review	0.30668266	5.04%	6
Brand sales at review	0.1265666	2.08%	12
Model launch duration	0.04624445	0.76%	17
Market segment sales at review	0.04564869	0.75%	18
Brand establishment duration	0.03453301	0.57%	19

the proportion and ranking of the corresponding feature importance in all variables are calculated. It is obvious that these six variables show importance in predicting fake reviews, especially the importance of Brand enter China duration, Sales at review and Brand sales at review with high proportion and ranking. This suggests a direct correlation with the release timing of positive fake reviews.

4.5.3.1 The relationship between product and brand sales and positive fake reviews

Table 4.4: Data statistics in the car model sales clustering and brand sales clustering.

Sales	Clusters	Minimum sales	Maximum sales	Average sales	Number of data points
Car model	Low sales	0	6779	2088	196453
	Medium sales	6783	20585	11476	150424
	High sales	20654	62339	29804	40113
Brand	Low sales	0	73529	23836	198720
	Medium sales	73741	155748	123532	107437
	High sales	157430	346739	189405	80833

First, to better observe the relationship between product sales and positive fake reviews, as shown in Figure 4.4, we cluster the monthly sales of the corresponding car model at the time of review and show the distribution of corresponding SHAP values. Based on the elbow method, car sales are classified into three categories: low sales, medium sales, and high sales (see Table 4.4). Based on the ANOVA test and Tukey HSD (Honestly Significant Difference) test for SHAP values under different sales levels, the results

clearly indicate that there are significant differences between the three groups of SHAP values. In the low sales group, the vast majority of SHAP values are less than 0, indicating that the low sales most strongly contribute to the judgment of fake reviews as negative. Moreover, in the low sales group, most SHAP values are significantly smaller than those in the medium and high sales groups, with a highly significant difference. Therefore, in comparison, when a car model's sales are at a low level, the likelihood of automakers or dealers manipulating positive fake reviews is lower. However, when compared to low sales levels, car models with relatively high sales levels are indeed more likely to have positive fake reviews manipulated by automakers or dealers, thus supporting H2. It is important to note, though, that the likelihood of manipulating positive fake reviews is highest when the car model's sales are at a medium level.

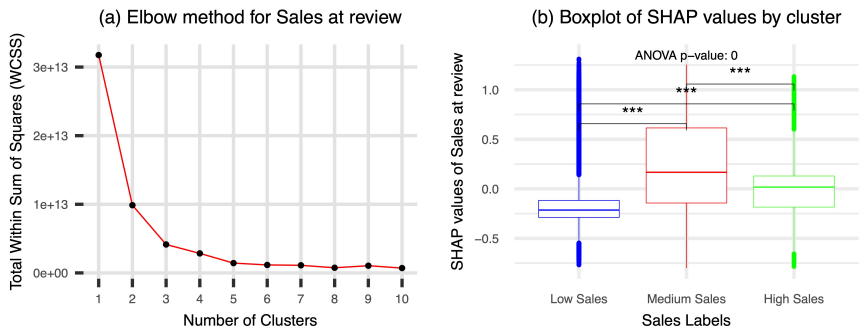

Figure 4.4: Clustering of car sales at the review and the distribution of corresponding SHAP values.(Note: The p-value displayed is rounded to four decimal places, with values rounded to 0.0000 shown as 0. The differences between data of different groups in the boxplot are represented by different symbols: *** represents a p-value ≤ 0.001, ** represents a p-value $0.001 < p \leq 0.01$, * represents a p-value $0.01 < p \leq 0.05$, ns represents a p-value > 0.05.)

Additionally, we categorized brand sales at the time of review into two groups (i.e., low sales and high sales, see Table 4.4) using the same method. Figure 4.5 illustrates the distribution of SHAP values for brand sales across different categories. Clearly, when a brand's sales are at a low level, the majority of corresponding SHAP values are less than 0, indicating a negative influence on the prediction of fake reviews. In contrast, when a brand's sales are at a high level, the opposite situation is observed. Given that

the brand sales levels were divided into only two categories, we conducted a t-test, and the results also confirmed that the differences are highly significant. Therefore, when a brand is at a high sales level, automakers or dealers are more likely to manipulate positive fake reviews, thus supporting H3.

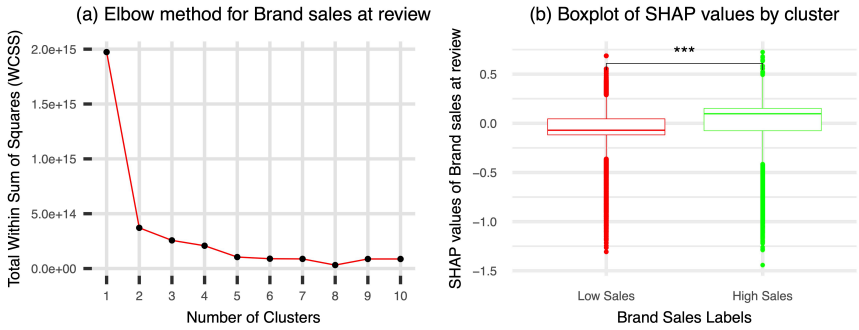

Figure 4.5: Clustering of brand sales at the review and the distribution of corresponding SHAP values.(Note: The differences between data of different groups in the boxplot are represented by different symbols: *** represents a p-value ≤ 0.001, ** represents a p-value $0.001 < p \leq 0.01$, * represents a p-value $0.01 < p \leq 0.05$, ns represents a p-value > 0.05.)

4.5.3.2 The relationship between market size and positive fake reviews

To study the relationship between the timing of manipulating positive reviews and market size, we divided the overall market into 14 segments based on Car Model Type and Size. It is important to note that internationally, car models larger than mid-size are typically defined as large-size or full-size models. However, the Chinese market further divides this category into two size classifications based on factors such as models' length and wheelbase. Therefore, in this study, the full-size SUV market(Small) refers to the smaller subcategory within this size range, while the full-size SUV market(Large) refers to the larger subcategory within the same range. The monthly sales of the market segment at the time of the review represent the size of the segment at that time. The

Sedan and MPV markets are similar to this as well. As shown in Table 4.5, by applying the elbow method to the sales data of each market segment (see Appendix J), we can determine the optimal number of clusters and the corresponding cluster labels for each segment.

Table 4.5: Market segments and number of clusters.

Market segments	**Number of clusters**
Mini sedan market	3-(low sales, medium sales, high sales)
Small sedan market	3-(low sales, medium sales, high sales)
Compact sedan market	3-(low sales, medium sales, high sales)
Mid-size sedan market	3-(low sales, medium sales, high sales)
Full-size sedan market (Small)	3-(low sales, medium sales, high sales)
Small SUV market	3-(low sales, medium sales, high sales)
Compact SUV market	3-(low sales, medium sales, high sales)
Mid-size SUV market	4-(low sales, lower-medium sales, upper-medium sales, high sales)
Full-size SUV market (Small)	4-(low sales, lower-medium sales, upper-medium sales, high sales)
Full-size SUV market (Large)	2-(low sales, high sales)
Minivan	3-(low sales, medium sales, high sales)
Compact MPV market	3-(low sales, medium sales, high sales)
Mid-size MPV market	3-(low sales, medium sales, high sales)
Full-size MPV market (Small)	4-(low sales, lower-medium sales, upper-medium sales, high sales)

Figure 4.6 shows the distribution of SHAP values across different market sizes in five sedan submarkets. Firstly, the results of the ANOVA test indicate that, only in the small sedan market, there is no significant difference in SHAP values across different market size categories. Secondly, in the other sedan market segments, post-hoc analysis using the Tukey HSD test on the ANOVA results revealed several significant differences. Only in the mini sedan market, when the sales in the segment are at a non-low level, are automakers or dealers more likely to manipulate positive fake reviews, thereby supporting H1. In contrast, in the remaining sedan segments, except for the mid-size sedan market where the likelihood of manipulating positive fake reviews is highest when the segment sales are at a medium level, in the other two sedan segments, automakers or dealers are more likely to manipulate positive fake reviews when the segment sales are at a low level. Therefore, in these sedan segments, H1 is not supported.

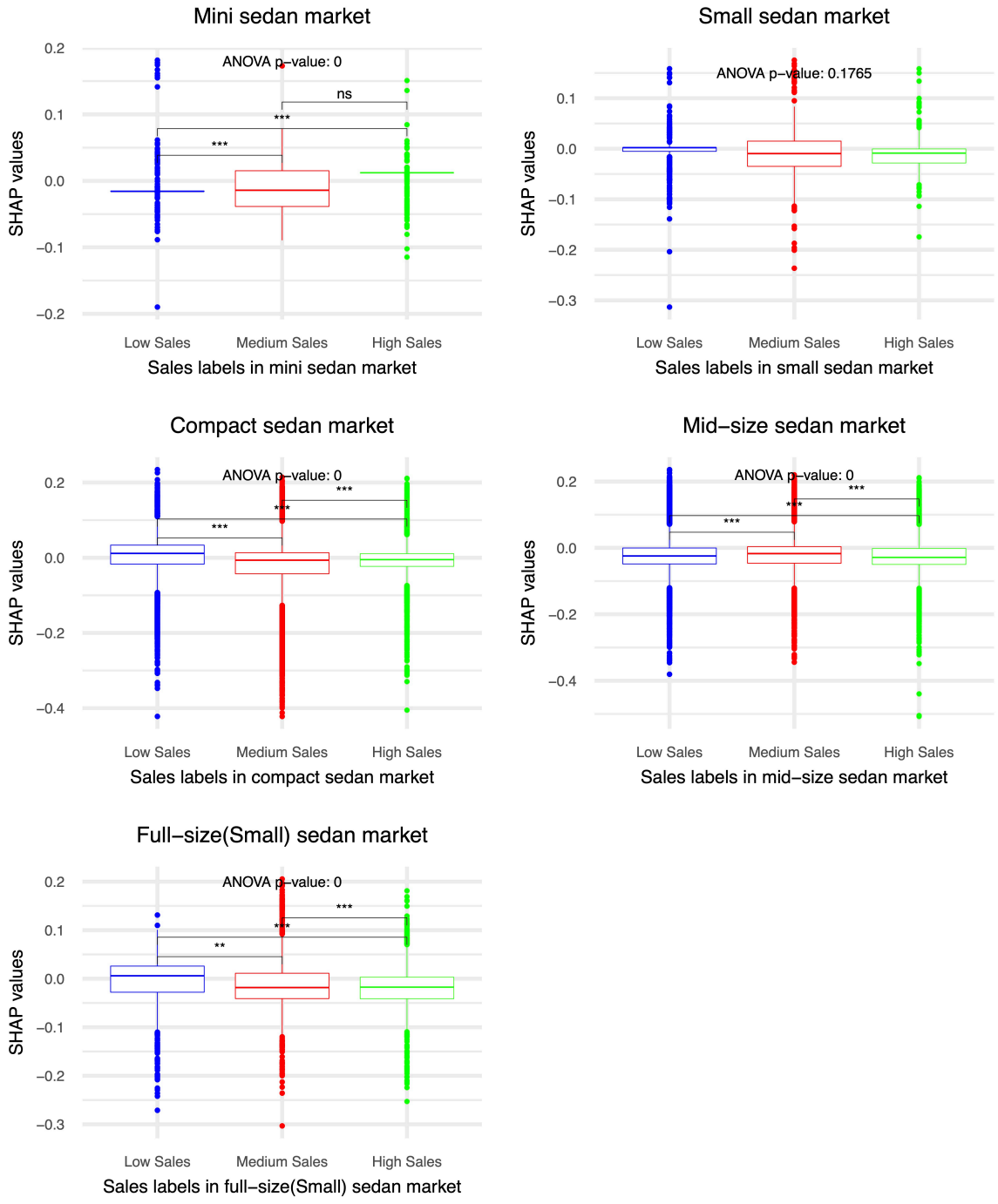

Figure 4.6: Clustering of different sedan markets sales at the review and the distribution of corresponding SHAP values.(Note: The p-value displayed is rounded to four decimal places, with values rounded to 0.0000 shown as 0. The differences between data of different groups in the boxplot are represented by different symbols: *** represents a p-value ≤ 0.001, ** represents a p-value $0.001 < p \leq 0.01$, * represents a p-value $0.01 < p \leq 0.05$, ns represents a p-value > 0.05.)

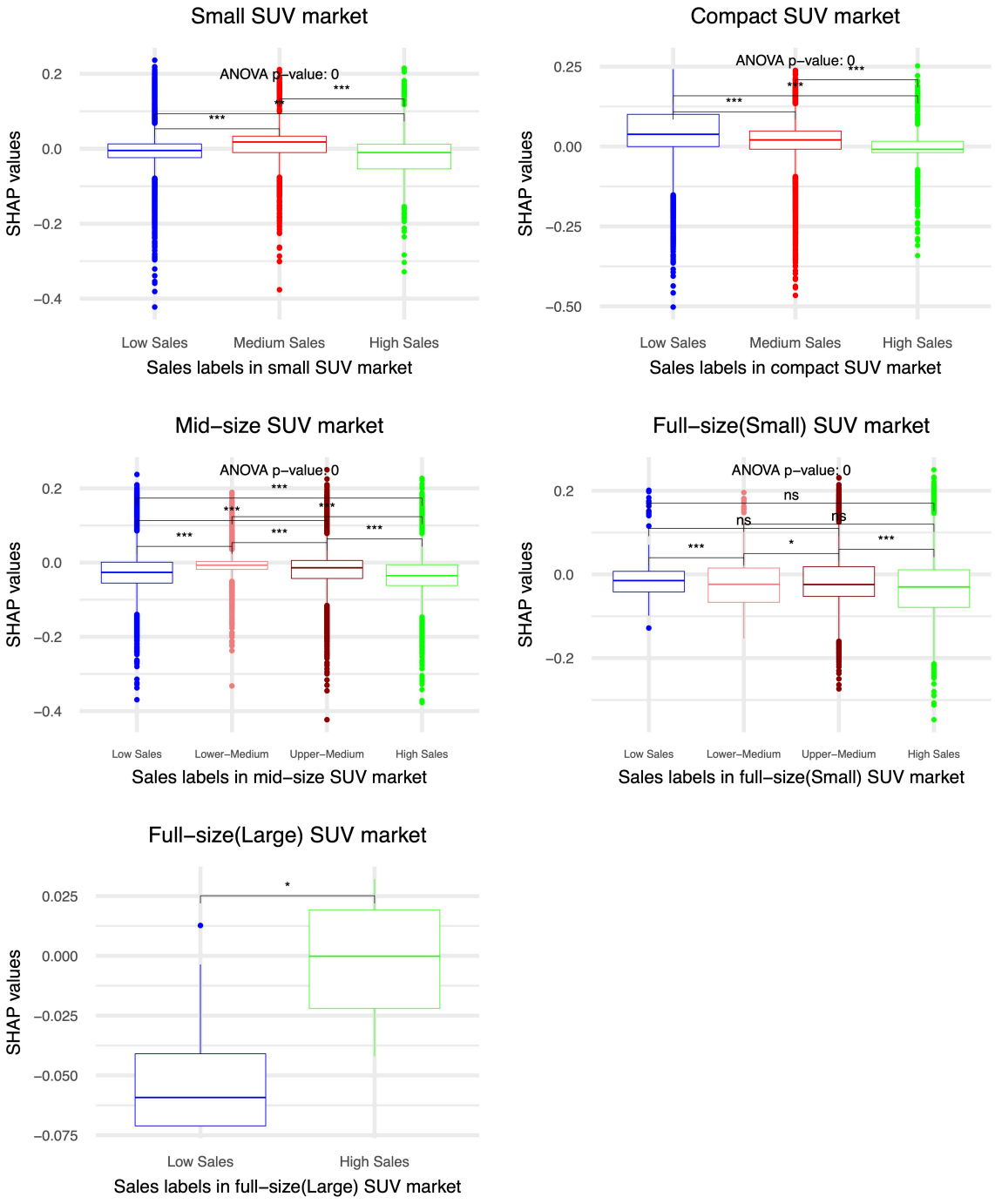

Figure 4.7: Clustering of different SUV markets sales at the review and the distribution of corresponding SHAP values.(Note: The p-value displayed is rounded to four decimal places, with values rounded to 0.0000 shown as 0. The differences between data of different groups in the boxplot are represented by different symbols: *** represents a p-value ≤ 0.001, ** represents a p-value $0.001 < p \leq 0.01$, * represents a p-value $0.01 < p \leq 0.05$, ns represents a p-value > 0.05.)

In Figure 4.7, the distribution differences of SHAP values across different market size under the SUV category are compared using the same method. It is important to note that in the full-size (Large) SUV market, the submarket sales were divided into only two categories, so a t-test was used to compare the differences in SHAP values between these categories. In the other submarkets, ANOVA and Tukey HSD tests were applied. Clearly, only in the full-size (Large) SUV market is H1 supported, meaning that when the sales in the segment are at a high level, automakers or dealers are more likely to manipulate positive fake reviews, with a significant difference compared to when sales are at a low level. In the small SUV market and mid-size SUV market, when the submarket's sales are at a medium level, i.e., under a medium market size, automakers or dealers are more likely to manipulate positive fake reviews, followed by when the submarket's sales reach lower levels, with these differences being highly significant. In the compact SUV market, the lower the sales in the segment, the more likely positive fake reviews are manipulated. In the full-size (Small) SUV market, although Tukey HSD test results indicate some noticeable differences, these differences do not show any clear pattern related to the manipulation of positive fake reviews. Therefore, the aforementioned SUV market segments do not support H1.

For the four MPV market segments in Figure 4.8, the results of the ANOVA test show that the distribution of SHAP values across different market scales does not exhibit significant differences in the minivan market and full-size (Small) MPV market. Therefore, in these two submarkets, the likelihood of automakers or dealers manipulating positive fake reviews does not differ significantly across market size. In the compact MPV market, when the submarket sales are at a medium level, the likelihood of automakers or dealers manipulating positive fake reviews is the lowest, and the distribution of SHAP values shows a clear difference compared to the other two categories. In the

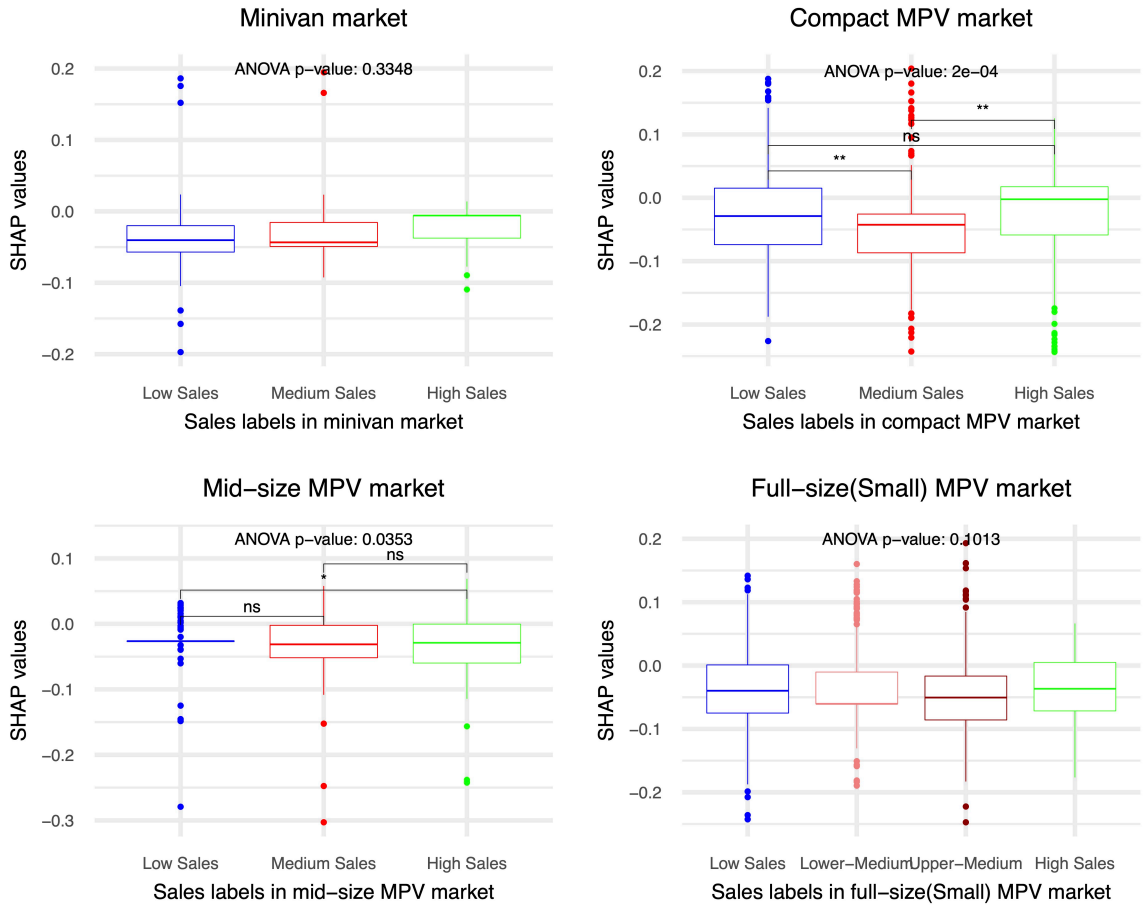

Figure 4.8: Clustering of different MPV markets sales at the review and the distribution of corresponding SHAP values.(Note: The p-value displayed is rounded to four decimal places, with values rounded to 0.0000 shown as 0. The differences between data of different groups in the boxplot are represented by different symbols: *** represents a p-value ≤ 0.001, ** represents a p-value $0.001 < p \leq 0.01$, * represents a p-value $0.01 < p \leq 0.05$, ns represents a p-value > 0.05.)

mid-size MPV market, we only observe a slightly higher likelihood of manipulating fake reviews in the small market size compared to the large market scale, but the difference is not significant when compared to the medium market size. Therefore, H1 is not supported in the MPV market.

4.5.3.3 The relationship between model launch duration and positive fake reviews

Through the SHAP method, we can obtain the distribution of SHAP values for Model Launch Duration, allowing us to study the timing of when automakers and dealers manipulate positive fake reviews in relation to the duration since a model's launch. To more clearly analyze the trend of SHAP values with Model Launch Duration, as shown in Figure 4.9a, we calculated the average SHAP values for each year. Overall, the average SHAP values of the model launch duration in the first three years are minimal and the contribution is negative. This indicates that during the early stages of a model's launch, the likelihood of automakers and dealers manipulating positive fake reviews is low, meaning H4 is not supported.

In addition, it should be noted that automakers often update car models that have been on the market for many years according to the lifecycle of the model. Although they tend to retain the same model's name during this process, the updated version may have entirely different characteristics. Previous studies have shown that the average lifecycle for model updates is 5.6 years, and this trend is decreasing (Volpato and Stocchetti 2008). Therefore, we analyzed the changes in the average SHAP values of Model Launch Duration over the past 15 years in 5-year intervals (see Figure 4.9b). It is evident that the changes in average SHAP values within each 5-year period over these

15 years exhibit some similarities. In the fourth year of each cycle, the average SHAP values for Model Launch Duration reach their highest point, consistently contributing positively. This indicates that during the mid-to-late stages of a model's lifecycle, the likelihood of automakers or dealers manipulating positive fake reviews is the greatest.

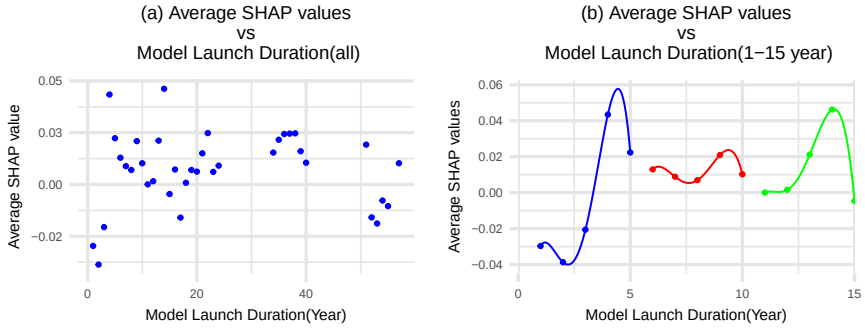

Figure 4.9: The average SHAP value changes with the model launch duration.

4.5.3.4 The roles of brand establishment duration and brand entering China duration

Next, as shown in Figure 4.10, by observing the changes in the average SHAP values for Brand Establishment Duration and Brand Entering China Duration, we study the timing of potential manipulation of positive fake reviews related to these factors. For Brand Establishment Duration, when it is less than 3 years, its average SHAP values are greater than 0, indicating that during the first two years after a brand is established, automakers are more likely to manipulate positive fake reviews. Therefore, H5 is supported. However, overall, although no clear trend of change is observed, the average SHAP value is higher when the brand has been established for approximately 30 and 120 years. By reviewing the experimental data, it is evident that this is mainly contributed by the four brands: Cadillac, Buick, Skoda, and Jetta, which show more positive SHAP values.

For Brand Entering China Duration, during the early stages of a brand's entry into the Chinese market, its average SHAP values are noticeably less than 0, indicating a negative contribution to identifying positive fake reviews. Thus, in the early stages of entering the Chinese market, automakers and dealers are less likely to manipulate positive fake reviews, meaning H6 is not supported. Overall, there is no significant trend in the likelihood of manipulating positive fake reviews based on the duration since the brand entered the Chinese market. However, when the brand has entered the Chinese market for approximately 60 years, its average SHAP value is significantly higher than during other periods. This is mainly attributed to the more positive SHAP values contributed by Toyota and Hongqi brands.

For the six brands mentioned in the analysis above, Cadillac, Buick, Skoda, and Toyota all have excellent brand value internationally. As one of the most important national brands in China, Hongqi enjoys an excellent brand reputation in China. As for the Jetta brand, which was spun off from the Volkswagen brand in the Chinese market, the strong reputation of the Volkswagen brand also provides it with good brand value. Therefore, these brands, relying on their strong brand value, are better equipped to resist the risks posed by the manipulation of fake reviews.

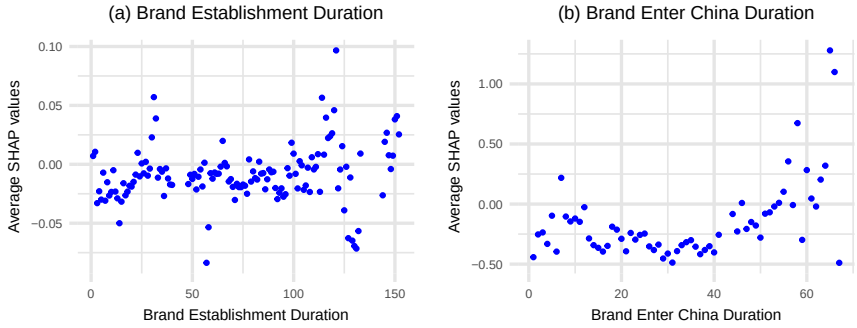

Figure 4.10: The average SHAP value changes with the brand establishment duration and the brand enter China duration.

4.6 Discussion

Previous studies have thoroughly discussed that financial motivation and competition are the main drivers for firms to manipulate positive fake reviews. However, our findings suggest that even when these motivations are strong, it is not necessarily the most likely timing for firms to manipulate positive fake reviews.

First, in terms of sales, when the sales of a product or brand are very low, companies face performance pressure, which should be a timing when financial motivation is particularly strong. However, the experimental results indicate that this is the least likely time for the manipulation of positive fake reviews. Surprisingly, when sales are at relatively high levels, firms are more likely to engage in the manipulation of positive fake reviews. This phenomenon may occur because, when sales are low, firms have a lower capacity to bear the risks associated with fake reviews. Meanwhile, the increase in fake reviews is likely to raise consumer suspicion, further weakening their purchase intention and causing severe negative publicity (Wu et al. 2020). It is worth noting that for some brands with low sales, low sales do not necessarily mean poor product performance. One possible reason is that these brands have fewer product categories in the market. Similarly, due to the limited range of products, firms behind these brands may have a lower capacity to bear risks. Overall, when sales are high, although financial motivation may not be as strong, consumers are less likely to question the online reviews of high-selling products and brands (Ko and Bowman 2023). Therefore, firms might manipulate some positive fake reviews to promote their products.

Secondly, from the perspective of market size, the timing of positive fake review manipulation by firms varies across submarkets. Overall, there is no clear pattern across all 14 submarkets. In the small sedan market, minivan market, and full-size (Small) MPV market, there is no significant difference in the likelihood of manipulating posit-

ive fake reviews at different market size levels. In the full-size (Small) SUV market and mid-size MPV market, although there are some significant differences in the likelihood of manipulating positive fake reviews at certain market size levels, no clear pattern emerges. In the compact MPV market, a medium-sized market size is the least likely time for manipulating positive fake reviews. Conversely, in the mid-size sedan market, small SUV market, and mid-size SUV market, the opposite situation is observed. Additionally, small market sizes are the most likely timing for firms to manipulate positive fake reviews only in the compact sedan market, full-size (Small) sedan market, and compact SUV market. In contrast, in the mini sedan market and full-size (Large) SUV market, firms are most likely to manipulate positive fake reviews at large market size levels. This may be because the market size in these two segments is smaller compared to others, indicating limited consumer demand for such products. Given the impact of brand credibility on consumer choices (Swait and Erdem 2007), firms are less likely to risk sacrificing brand credibility by engaging in fake review manipulation to achieve limited sales growth when the market size is at a lower level.

Third, from the perspective of a car model's lifecycle, the middle to later stages of a product's life is the most likely timing for firms to manipulate positive fake reviews, rather than at the very beginning of a model's release. This may be because, although positive fake reviews can help address the cold start problem of a new model, experienced consumers may become suspicious when faced with an abundance of positive reviews in the absence of clear market feedback, potentially turning the positive fake reviews into a negative influence (Zhuang et al. 2018). In the middle and later stages of a product's lifecycle, product performance often declines, and market competition intensifies. This heightened competitive pressure may be a key factor driving firms to post more positive fake reviews during this time.

Fourth, since the brand creation date and the timing of market entry may not align, the brand's presence in the market should be considered from two perspectives. From the perspective of brand creation, in the early stages of brand establishment, many companies may urgently need to manipulate positive fake reviews to solve the cold-start problem, thereby quickly shaping a positive brand image. However, from the perspective of the duration of the brand entering the market, brands that have just entered a new market are less likely to manipulate positive fake reviews. This may be because, prior to entering the new market, they already have a foundation of positive brand value but possess a lower risk tolerance in the new market, thus not wanting to bear the negative impact that such manipulative behavior might bring. It is important to note that for brands where the brand creation date and the market entry date align, balancing the positive and negative impacts of manipulating positive fake reviews is an important challenge they must face.

Fifth, in addition to the findings related to the timing of positive fake review manipulation, the results also show that among all the features used in the study, the features of the review text are the most important for predicting fake reviews, whether positive or all fake reviews. Secondly, the overall rating given by consumers is the second most important feature. This suggests that review text and ratings are the primary content manipulated by firms. However, consumers often use the overall rating of a product to filter potential options and rely on the review text to make their final purchasing decision from these options (Hu et al. 2014). This undoubtedly once again demonstrates the harmful impact of unethical manipulation of online reviews on consumers' purchasing decisions.

4.7 General contributions to business ethics

This study makes both methodological and theoretical contributions to business ethics literature. Methodologically, the proposed two-stage machine learning approach demonstrates the superior classification performance of the proposed fake review detection method. First, the analysis expands upon many product-centric, review-centric, and reviewer-centric features that had not been used in previous research. By integrating BERT for text feature extraction, the study shows the importance of feature engineering in improving the accuracy of machine learning classification and highlights their significant contribution to predicting fake reviews (Kumar et al. 2022). Second, the classification of manually labeled data validates the effectiveness of the created rules. These rules can be applied not only to the automotive market but also provide support for fake review labelling in other industries through the underlying rationale behind their creation. Additionally, the strong performance in F1 score and accuracy during the iterative process further underscores that PU-learning is a promising approach for detecting fake reviews. The semi-supervised learning approach, which integrates these methods, effectively balances the strengths and weaknesses of supervised and unsupervised learning in fake review detection.

Theoretically, this study is among the first to identify that the dimension of time is a critical factor in understanding when businesses are likely to engage in the unethical practice of creating and posting fake reviews. However, reflecting critically on the empirical results of this study across the different firm and market conditions, another surprising insight appears: the dimension of time is a critical factor that is strongly correlated with the motivation to post fake reviews, but it is not the main explanatory factor. It appears that the main reason why businesses decide to post fake reviews is the perceived absence of the risk and impact of being caught. Particularly, when the perceived negative impact of being caught is low – such as when sales volumes are high,

and the general brand position is strong – the motivation to act unethically is high. This suggests that unethical agencies in businesses are not actuated by an immediate need to respond to a business priority (e.g., the need to gain market entry, save a faltering brand, etc.) but by the perceived absence of risk of being caught. As such, some businesses have a default mode of operation were acting unethically is a central component that is primarily controlled by the active presence of external risk of being caught. The implication is that fake review detection methodologies such as the ones presented in this study are crucial to creating and upholding fair market practices, not just in terms of detecting fraud but as a preventive measure as it increases the risk of being caught which is the main ethical constraint. Put differently, to many businesses the temptation to act unethically is always too much unless there are robust measures in place to detect fraudulent behaviors.

4.8 Summary

Online consumer reviews, once a key source of information reflecting product quality, have gradually become a carrier for unethical marketing practices by firms. This study focuses on the unethical behaviors in online consumer reviews, particularly the timing of manipulating positive fake reviews, while also supplementing the existing literature on fake online reviews. To improve the accuracy of fake review detection and obtain valuable fake review data, the research introduces a fake review detection method based on BERT and PU-learning. The study holds significant practical implications in various aspects. Overall, it effectively protects consumer interests, enhances the efficiency of detecting fake reviews on social media platforms, maintains a fair competitive business environment, and mitigates the impact of unethical business practices.

The primary beneficiaries of this study are consumers. Since most social media platforms currently do not offer filtering services for fake reviews, misleading information is rampant, making it easy for consumers to suffer from poor purchasing decisions. In today's context, where fake reviews are more concealed, consumers can better identify fake reviews and reduce their cognitive load regarding online reviews by understanding the timing of manipulation in relation to the duration of product and brand in the market, sales, and market size. Secondly, the proposed method targets ethical social media platforms. These platforms can use this approach for dynamic and automated detection of fake reviews, which would help improve the credibility of online reviews, increase consumer trust, and promote the sustainable development of these platforms. Moreover, with the aid of appropriate penalties, the manipulation of fake reviews by firms on these platforms could be reduced. A lack of publicly available labeled sample data and reliable fake reviews has long been a challenge in academic research. Now, however, researchers can utilize the rules related to manual labelling fake reviews and sample review data identified in this study to further refine existing detection methods. Based on the identified fake reviews, they can further investigate the causes and consequences of fake review manipulation and expand more depth research in this field.

Inevitably, our study has some limitations. First, the timing of positive fake reviews manipulated by firms may also be related to other factors, such as emergencies in the auto market, political environment and so on. A richer study can be carried out in subsequent studies by expanding more relevant variables. Second, there is still the possibility of manipulating negative fake reviews in the market. However, since most of the reviews in social media are positive, there is no available large-scale dataset of negative reviews to conduct relevant research. In addition, in future research, we can find the connection between other characteristics and fake reviews based on reliable fake review detection methods and update the existing rules for identifying fake reviews to continuously improve the efficiency of detecting fake reviews.

Chapter 5

Conclusion

5.1 Research summary

This study systematically advances the research of business intelligence and analytics through three sub-studies, addressing the identified research gaps in Chinese automobile industry data, competitive advantage of brands and products, and business fraud in online reviews. The research spans four key dimensions: data collection, methodological framework construction, theoretical validation and expansion, and market behavior regulation.

The first study focuses on the collection, creation, and expansion of a comprehensive dataset for the Chinese automobile market, addressing the long-standing lack of publicly available datasets in this field. By exploring the potential applications of this dataset, the study identifies key factors influencing sales and consumer sentiment in reviews from a macro perspective, thereby demonstrating its significant business value.

The second study employs machine learning methods to develop an innovative sales analysis framework, applying it to sales forecasting and market dynamics analysis in the Chinese automotive industry. The findings reveal the persistence of first-mover advantages, the impact of innovative late entrants in challenging these advantages, the enduring influence of country-of-origin effects, subtle shifts in country-of-origin preferences, and the greater impact of first-mover advantages on sales compared to country-of-origin effects. These insights provide a clearer understanding of the evolving dynamics of these market advantages.

The third study successfully identifies four key factors closely related to the manipulation of positive fake reviews: sales, market size, brand market presence duration, and product lifecycle. By establishing 11 tailored fake review detection rules for the Chinese automobile market, the study generates a small but well-labeled dataset. When integrated with the proposed fake review detection method based on BERT and PU-learning, this study efficiently and accurately identified fake reviews within the dataset. Further analysis uncovers critical timings that firms are likely to manipulate positive fake reviews, thereby extending theoretical research in this domain.

Overall, these three types of research are interdependent and mutually complementary. First, from a market perspective, the first study provides reliable data support for research on the Chinese automotive industry, helping to enhance the breadth and accuracy of analysis. The second study focuses on sales analysis, predicting market trends and offering valuable business insights to guide production and marketing strategies. The third study examines the manipulation of fake positive reviews to regulate unethical marketing practices, thereby fostering a fair competitive environment. Together, data collection, market trend forecasting, and market behaviour regulation form a comprehensive analytical framework for the Chinese automobile market, with each of the three studies delving into these aspects from different perspectives. Second, from a business data analysis perspective, these studies align with the DIKW Pyramid (Data-

Information-Knowledge-Wisdom Pyramid), systematically transforming raw data into actionable wisdom. The first study processes and structures raw data, assigning it meaning and completing the transition from data to information. The second and third studies further analyse this information, extracting valuable business knowledge through sales forecasting and fake review detection while expanding relevant business theories. Finally, based on these research findings, an in-depth exploration of market dynamics and competitive environments provides wisdom-driven decision-making to navigate the complexities of the market.

5.2 Contribution

This study focuses on the construction of industry datasets, sales forecasting, fake review detection, and the exploration of relevant business theories. It advances business research in the automotive industry, promotes the application of business intelligence and data analytics, and expands existing business theories. The findings make significant contributions to methodological innovation, theoretical development, and practical applications.

5.2.1 Methodological contribution

The methodological contributions of this study are primarily reflected in the sales analysis framework proposed in the second study and the fake review detection framework introduced in the third study. By effectively integrating machine learning methods, both frameworks achieve efficient analysis while maintaining strong interpretability.

In the second study, a review of previous research reveals that sales prediction primarily relies on two types of methods: traditional statistical models, such as linear regression, logistic regression, and support vector machines, which offer strong interpretability but limited predictive accuracy (Hülsmann et al. 2012; Bernstein et al. 2007); and machine learning methods, such as random forests, gradient boosting decision trees (GBDT), and neural networks, which improve prediction accuracy but often lack transparency (Xia et al. 2020; Afandizadeh et al. 2023; Rudin 2019). Each of these approaches has its limitations, making it challenging for traditional analysis frameworks to balance both accuracy and interpretability. Additionally, existing studies often use sales as the sole target variable to evaluate product or brand performance, which can lead to prediction bias due to relying on a single indicator (Behn 2003). To address these issues, we propose an innovative two-stage machine learning-based analysis framework, significantly enhancing the robustness, accuracy, and interpretability of sales forecasting. In the first stage, we construct a sales performance indicator by comparing different clustering algorithms and incorporating both sales and ranking. Compared to using either sales or ranking alone, this approach effectively mitigates the impact of market fluctuations on predictions, reducing biases caused by relying on a single performance indicator. In the second stage, we compare various machine learning models during the sales prediction process, significantly improving analytical accuracy. Additionally, the incorporation of the SHAP method mitigates the impact of multicollinearity among variables while providing valuable global and local explanations, thereby enhancing model interpretability. This sales analysis framework demonstrates exceptional dynamic learning capabilities and innovation, while also offering strong scalability and adaptability. It can be extended beyond the automobile market to other industries and further optimized by integrating multiple models to achieve even better predictive performance.

Commonly used methods in research on fake review detection typically include supervised learning, unsupervised learning, and semi-supervised learning. Among them, supervised learning relies on large-scale labelled data to achieve accurate detection (Khurshid et al. 2018), while unsupervised learning does not require labelled datasets but often suffers from relatively low accuracy (Dong et al. 2018). In contrast, semi-supervised learning can perform well with a small amount of labelled data, yet the lack of such labelled review data remains a significant challenge in most markets. Additionally, previous research has primarily focused on review-centric and reviewer-centric features, often neglecting the role of product-centric features in identifying fake reviews (Kumar et al. 2022). In the third study, a two-stage fake review detection framework based on BERT and PU-learning was developed, demonstrating excellent classification performance as a semi-supervised learning approach. This framework expands on review-centric and reviewer-centric features, and introduces product-centric features that were not covered in previous research. Based on these features, 11 key rules for detecting fake reviews in the Chinese automobile market were identified, supporting the creation of a small-scale labelled review dataset, and classification predictions confirmed the effectiveness of these rules. The integration of BERT in text feature extraction underscores the importance of feature engineering in enhancing detection accuracy and highlights the significant contribution of product-centric features to identifying fake reviews. Furthermore, the application of PU-learning within this framework achieves high classification accuracy through continuous iteration, showcasing its potential in fake review detection. Overall, this analytical framework advances research in the field of fake review detection by focusing on four aspects: the construction of labeling rules, the creation of sample data, the optimization of feature engineering, and the application of classification models. It achieves a balance between prediction accuracy and reliance on labeled datasets, offering new research perspectives and practical solutions for fake review detection.

5.2.2 Theoretical contribution

Throughout the research, innovative analysis frameworks were employed to systematically examine key variables, thereby validating and expanding business theories concerning first-mover advantages, country-of-origin effects, and fake review manipulation. The core theoretical contributions of this research are primarily concentrated in the second and third studies.

In the second study, the dynamic changes in first-mover advantages and country-of-origin effects under the impact of newly introduced products and brands were analysed. First, this study innovatively provides a comprehensive examination of the impact of the first-mover advantage on sales performance from three dimensions: product launch date, brand establishment date, and brand market entry date. By distinguishing between product and brand first-mover advantages, the study expanded existing business theories. The findings confirmed the existence of first-mover advantages across these three dimensions, though their effects did not always manifest simultaneously. Additionally, in most market segments, brand establishment date contributed more significantly to brand first-mover advantage than brand market entry date. Furthermore, brand first-mover advantages exerted a greater influence on sales performance than produce first-mover advantages. Second, the study of NEVs and emerging NEV brands revealed a shift in consumer preferences, demonstrating a greater acceptance of innovative late entrants. This finding suggests that under the influence of technological changes and policy adjustments, the original first-mover advantage may be disrupted by innovative late entrants. Moreover, this study further validated the enduring influence of brand country-of-origin effects on sales performance, confirming their independent existence. Consumers continued to exhibit a preference for brands from developed countries, particularly German and American automobile brands. However, in a few market segments, consumer preferences regarding country-of-origin showed slight changes, with

an increasing acceptance of Chinese automobile brands. Finally, this study conducted a systematic comparison of the relative contributions of product first-mover advantages, brand first-mover advantages, and brand country-of-origin effects to sales performance, further extending related theoretical discussions. The findings revealed that, in most market segments, both brand and product first-mover advantages had a greater impact than country-of-origin effects. This conclusion deepens the understanding of the mechanisms by which first-mover advantages and country-of-origin effects shape market competition, providing new theoretical perspectives for future research.

The third study makes significant theoretical contributions to the fields of positive fake review manipulation and business ethics from the perspective of manipulation timing. Specifically, it innovatively explores the possible timing of firm engagement in positive fake review manipulation through four key dimensions: sales, market size, brand market presence duration, and product lifecycle. By uncovering the potential relationships between these factors and manipulation behaviour, this study enriches the theoretical framework surrounding fake review manipulation. The findings indicate that firms are most likely to engage in positive fake review manipulation when product or brand sales reach relatively high levels, during the mid-to-late stages of the product lifecycle, and in the early stage of brand establishment. Conversely, firms are least likely to manipulate reviews in the early stage of market entry. Additionally, the study reveals a direct relationship between market size and the likelihood of fake review manipulation; however, this relationship does not exhibit a consistent pattern across different market segments, suggesting that market conditions may influence firms' manipulation strategies. Furthermore, by reflecting on the empirical results regarding sales, market size, brand market presence duration, and product lifecycle, this study reveals the decision-making mechanisms behind unethical business practices from a novel perspective. The findings suggest that a firm's decision to manipulate fake reviews largely depends on its assessment of the risk of detection—when the perceived

risk is low or nearly non-existent, firms exhibit a significantly stronger motivation to engage in unethical behaviours. This discovery not only enhances the understanding of firm manipulation strategies but also offers a fresh theoretical perspective for research on business ethics.

5.2.3 Practical application

The practical contribution of this study is mainly reflected in two aspects: the creation of the dataset and the analytical methods employed, both of which are evident across the three sub-studies.

The primary contribution of the first study lies in data collection, creation, expansion, and validation. Although its theoretical and methodological innovations are relatively limited, it has made significant practical contributions. The most critical practical contribution is the integration of sales data, online reviews, industry news and information data, along with the supplementation of previously missing key variables, to construct a more comprehensive dataset for Chinese automotive industry. This dataset addresses the longstanding issues of data scarcity and fragmentation in the industry, where large-scale, publicly available datasets have been lacking. With its comprehensive content and diverse variables, this dataset not only enhances the accuracy and interpretability of academic research but also provides strong support for a broader range of research topics. Furthermore, it holds substantial value for business applications. Sales data aids in market demand forecasting, online reviews deepen the understanding of consumer preferences, and industry news and information help businesses better grasp industry trends. The integration of these three types of data enables enterprises to assess product advantages, refine brand positioning, evaluate marketing effectiveness, respond to industry changes, and enhance market competitiveness. Additionally, this

dataset serves as a valuable resource for policymakers, allowing them to gain a more comprehensive understanding of the industry landscape and assess the effectiveness of past policies. This, in turn, provides crucial support for formulating more precise and effective industry policies and regulations. Therefore, the development of this dataset not only has a profound impact on academic research but also benefits a wide range of stakeholders, including automotive manufacturers, marketers, and policymakers.

In the second study, its practical contribution lies primarily in the application value of the innovative sales analysis framework in business practice. First, the comprehensive indicator constructed based on sales and rankings provides enterprises with a more holistic performance evaluation method, enabling them to monitor product and brand market performance with greater precision. Second, this analysis framework demonstrates outstanding performance in accuracy, interpretability, and flexibility, showcasing significant practical value in market trend forecasting, competitive advantage and disadvantage analysis, and consumer preference insights. Moreover, the framework facilitates in-depth interpretation of prediction results, allowing business managers, marketers, and policymakers to gain unique business insights beyond existing commercial theories. This, in turn, enables them to make precise judgments and play a substantial role in strategic decision-making, marketing strategy optimization, and industry policy adjustments. Overall, whether in terms of the construction of comprehensive performance indicators, the sales analysis process, or the model interpretation approach, this study advances the application of machine learning in business intelligence and data analytics, providing a highly valuable reference framework for business analysis tasks across various industries.

In the third study, the rules for identifying fake reviews, the manually labelled sample data, and the fake review detection framework hold significant practical value. First, the developed identification rules are not only effective for labelling fake reviews in the Chinese automobile market but also offer a transferable approach based on product-

centric features, reviewer-centric features, and review-centric features, enhancing the efficiency and accuracy of manual annotation across different industries. Second, the manually labelled small-sample dataset provides crucial support for optimizing fake review detection methods, facilitating the assessment of unsupervised learning accuracy and laying the foundation for the improvement and application of semi-supervised learning techniques. More importantly, the proposed fake review detection framework enables automated detection of large-scale new online reviews while dynamically updating the sample dataset, offering online review platforms an efficient and sustainable solution that significantly enhances detection accuracy and efficiency. Overall, this study not only contributes valuable research resources to the academic community but also provides feasible technical support for online review platform managers and market regulators, advancing the practical application of fake review governance.

5.3 Managerial and policy implications

The theoretical, methodological, and practical contributions of this study have also had a profound impact on business data management, brand management, product management, online review management, market behavior regulation, and policy-making. Therefore, this study provides important managerial insights and policy support for various stakeholders, including business managers, marketing professionals, online review platform administrators, market regulators, and policymakers.

5.3.1 Managerial implications

As a crucial tool for data-driven decision-making and a key pathway for achieving sustainable business development, business intelligence and analytics play a significant role in guiding specific aspects of business data management. This study, based on the Chinese automobile market, provides valuable managerial insights into this field. First, the study highlights the contribution of variable expansion in improving data analysis accuracy and the role of diverse datasets in enriching research content. This underscores the importance of enhancing data diversity and constructing more comprehensive datasets as a top priority for business managers in managing commercial data. Second, the value creation process of both the sales analysis framework and the fake review detection framework in this study demonstrates that efficient analytical frameworks are fundamental to transforming data into actionable insights. Business managers should prioritize the development of flexible and scalable business data analysis frameworks. Such frameworks not only enhance adaptability and accuracy but also significantly improve the cost-effectiveness and efficiency of integrating new technologies. Furthermore, the key business insights derived from analysing sales data and fake reviews indicate that business managers must emphasize the interpretability of analytical results when selecting analytical frameworks. Interpretability is essential for enabling data-driven decision-making and generating business value, making it a core objective of business data management. Therefore, the business management community should optimize existing business data management strategies based on these three aspects to further enhance the business value of data, improve the efficiency of business intelligence and analytics applications, and elevate the quality of data-driven decision-making.

This study validates and expands on business theories related to first-mover advantages, country-of-origin effects, and manipulation of fake reviews, providing valuable reference for corporate managers and marketers engaged in targeted brand management and product management. Given the persistent influence of first-mover advantage and country-of-origin effect, as well as the potential for innovative latecomers to disrupt established market advantages, marketers in Chinese automobile market must carefully assess the impact of these factors on their brands and products to identify competitive strengths and weaknesses. When formulating marketing strategies, first-movers and brands benefiting from a positive country-of-origin effect should emphasize their longstanding heritage and brand origins, while strong latecomers should focus on highlighting product innovation to challenge first-mover advantages. At the same time, the heterogeneity of first-mover advantage and country-of-origin effect in different market segments highlights the importance of differentiated marketing, which is especially crucial for marketers to achieve precise marketing in various market segments. Similarly, for corporate managers, optimizing product development strategies based on entry timing and country-of-origin advantages can help reduce development costs and maximize returns. In particular, for Chinese domestic automobile brands, they should seize the opportunity in the luxury mid-size SUV market, where consumer preferences are shifting toward Chinese car brands, by increasing investment in the development and marketing of such models to further enhance brand value. Of course, for weaker brands, product innovation remains the top priority. Secondly, the prevalence of fake reviews must attract the attention of corporate managers and marketers, and they should take timely countermeasures based on competitors' manipulation strategies in areas such as sales, market size, brand market presence duration, and product lifecycle. Since online reviews are an important source for automotive companies to obtain market feedback, analyse their strengths and weaknesses, understand consumer preferences, and evaluate marketing effectiveness, the presence of fake reviews can easily mislead corporate decision-making, thereby causing severe negative impacts on the business. Improving the ability to identify fake reviews is key to helping corporate managers and marketers reduce the negative impact of fake reviews on business insights.

In addition, online review platforms are not only tools for businesses to obtain information and conduct marketing activities, but also an important channel for consumers to understand product characteristics and express their opinions on products. It has been proven that fake reviews are widespread and can not only diminish the trust between users and businesses, but also lead to a trust crisis between them and the online review platform. Therefore, to ensure the platform's sustainable development and commercial interests, platform managers must focus on identifying, labelling, and penalizing fake reviews to reduce the potential negative impact on both businesses and consumers, thereby gaining the long-term recognition of all relevant stakeholders. On the other hand, as an important channel for information dissemination, platform managers have the responsibility to regulate unethical business behaviours. Since perceived risks are a key driver for businesses engaging in unethical activities, strict governance of the platform will curb such behaviours at the source and effectively maintain a fair business environment.

5.3.2 Policy implications

This study also presents significant managerial implications from the perspectives of industry development and market regulation, providing valuable decision-making support for market regulators and policymakers. First, the analysis of sales can help policymakers assess the effectiveness of past policy implementations and provide data support for the formulation of new policies. The increasing competitive advantage of NEVs and NEV brands in certain markets proves the effectiveness of previous policies, such as purchase subsidies, exemption from vehicle purchase tax, and vehicle license plate support, in promoting NEV development. To further advance the growth of NEVs, the government should continue implementing and optimizing existing policies to facilitate a long-term shift in consumer preferences. Meanwhile, the diminishing bias against

Chinese domestic brands among consumers offers greater development opportunities for local automakers. Policymakers should introduce favourable support policies to further enhance the global influence of Chinese domestic brands, thereby fostering industry development and boosting the national economy.

Secondly, addressing the unethical practice of manipulating fake reviews requires not only effective monitoring by online review platforms but also joint efforts from market regulators and policymakers. Given the clear patterns in the timing of fake review manipulation, market regulators should comprehensively assess various factors to identify specific manipulation periods, thereby establishing a long-term, effective inspection mechanism for accurately detecting unethical behavior. Meanwhile, policymakers should formulate relevant policies from social and legal perspectives to strengthen penalties for unethical business practices. Ultimately, by exerting pressure through regulatory frameworks and government policies while increasing the risks associated with such behavior, the long-term protection of consumer rights, corporate interests, and a fair competitive environment can be ensured.

5.4 Limitation and future research

Although this study provides rich business insights and efficient data analysis methods, there are still some limitations that provide directions for future research. First of all, the studies on the first-mover advantage and country-of-origin effect only use structured data and sales performance as judgment criteria, but their influence is not limited to this. Future research can mine the specific impact of the first-mover advantage and country-of-origin effects on consumer purchase decisions by integrating unstructured data. The second limitation is that negative fake review are not included in the analysis of the manipulation timing of fake reviews. This is because the current online reviews

in Chinese automobile market are basically positive, so the small sample of negative fake review data is not enough to support the study. In the future, with the continuous accumulation of the number of negative reviews, a large number of fake negative reviews data can be obtained through the fake review detection method proposed by us. Then, the timing of manipulation can be discussed based on the motivation of publishing negative fake reviews. A third limitation is that the current data only cover China's car market between 2016 and 2022. Future research can investigate the dynamics of related theories by collecting data over a larger time span and on automobile markets in multiple countries around the world.

In addition, many important research directions can also be extended based on the analysis results and research methods in this study. For example, future research could apply the proposed sales analysis framework to analyze the impact of price on sales, facilitating reasonable product pricing. In the future, the impact of policies on the automobile market and consumer behavior can be studied by combining industry news data. Based on the relatively genuine reviews after fake review detection, future research can re-examine the business theories extended from online reviews in the past.

References

Afandizadeh, Shahriar et al. (2023). 'Using machine learning methods to predict electric vehicles penetration in the automotive market'. In: *Scientific Reports* 13.1, p. 8345.

Agrawal, Jagdish and Wagner A Kamakura (1999). 'Country of origin: A competitive advantage?' In: *International journal of research in Marketing* 16.4, pp. 255–267.

Aiello, Paul (1991). 'Building a joint venture in China: the case of Chrysler and the Beijing Jeep Corporation'. In: *Journal of General Management* 17.2, pp. 47–64.

Ananthakrishnan, Uttara, Davide Proserpio and Siddhartha Sharma (2023). 'I hear you: does quality improve with customer voice?' In: *Marketing Science* 42.6, pp. 1143–1161.

Arora, Anshu Saxena and Amit Arora (2017). 'WYSIWYG—seeing is believing: consumer responses to levels of design newness, product innovativeness, and the role of country-of-origin'. In: *Journal of International Consumer Marketing* 29.3, pp. 135–161.

Ashtiani, Matin N and Bijan Raahemi (2023). 'News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review'. In: *Expert Systems with Applications* 217, p. 119509.

Assuncao, Joao L and Robert J Meyer (1993). 'The rational effect of price promotions on sales and consumption'. In: *Management Science* 39.5, pp. 517–535.

Ataman, M Berk, Harald J Van Heerde and Carl F Mela (2010). 'The long-term effect of marketing strategy on brand sales'. In: *Journal of Marketing Research* 47.5, pp. 866–882.

Athaide, Gerard A et al. (2024). 'Marketing innovations and digital technologies: A systematic review, proposed framework, and future research agenda'. In: *Journal of Product Innovation Management*.

Aylsworth, Timothy (2022). 'Autonomy and manipulation: Refining the argument against persuasive advertising'. In: *Journal of Business Ethics* 175.4, pp. 689–699.

Badinger, Harald (2007). 'Market size, trade, competition and productivity: evidence from OECD manufacturing industries'. In: *Applied Economics* 39.17, pp. 2143–2157.

Balcet, Giovanni, Hua Wang and Xavier Richet (2012). 'Geely: a trajectory of catching up and asset-seeking multinational growth'. In: *International Journal of Automotive Technology and Management 19* 12.4, pp. 360–375.

Baptista, Marcia L, Kai Goebel and Elsa MP Henriques (2022). 'Relation between prognostics predictor evaluation metrics and local interpretability SHAP values'. In: *Artificial Intelligence* 306, p. 103667.

Bauer, Hans H, Nicola E Sauer and Christine Becker (2006). 'Investigating the relationship between product involvement and consumer decision-making styles'. In: *Journal of Consumer Behaviour* 5.4, pp. 342–354.

Behn, Robert D (2003). 'Why measure performance? Different purposes require different measures'. In: *Public administration review* 63.5, pp. 586–606.

Bekker, Jessa and Jesse Davis (2020). 'Learning from positive and unlabeled data: A survey'. In: *Machine Learning* 109.4, pp. 719–760.

Bernstein, Abraham, Jayalath Ekanayake and Martin Pinzger (2007). 'Improving defect prediction using temporal features and non linear models'. In: *Ninth international workshop on Principles of software evolution: in conjunction with the 6th ESEC/FSE joint meeting*, pp. 11–18.

Besharat, Ali, Ryan J Langan and Carlin A Nguyen (2016). 'Fashionably late: Strategies for competing against a pioneer advantage'. In: *Journal of Business Research* 69.2, pp. 718–725.

Bharadwaj, Neeraj et al. (2017). 'Predicting innovation success in the motion picture industry: The influence of multiple quality signals'. In: *Journal of Product Innovation Management* 34.5, pp. 659–680.

Bloomfield, Gerald T (2017). 'The world automotive industry in transition'. In: *Restructuring the global automobile industry*. Routledge, pp. 19–60.

Bohlmann, Jonathan D, Peter N Golder and Debanjan Mitra (2002). 'Deconstructing the pioneer's advantage: Examining vintage effects and consumer valuations of quality and variety'. In: *Management Science* 48.9, pp. 1175–1195.

Boudreau, Marie-Claude et al. (2004). 'Validating IS positivist instrumentation: 1997-2001'. In: *The handbook of information systems research*. IGI Global, pp. 15–26.

Božič, Katerina and Vlado Dimovski (2019). 'Business intelligence and analytics use, innovation ambidexterity, and firm performance: A dynamic capabilities perspective'. In: *The Journal of Strategic Information Systems* 28.4, p. 101578.

Busse, Meghan R, Duncan I Simester and Florian Zettelmeyer (2010). '"The best price you'll ever get": The 2005 employee discount pricing promotions in the US automobile industry'. In: *Marketing science* 29.2, pp. 268–290.

Buzzell, Robert D, Bradley T Gale and Ralph GM Sultan (1975). 'Market share-a key to profitability'. In: *Harvard business review* 53.1, pp. 97–106.

Cao, Dongpu et al. (2022). 'Future directions of intelligent vehicles: Potentials, possibilities, and perspectives'. In: *IEEE Transactions on Intelligent Vehicles* 7.1, pp. 7–10.

Carare, Octavian (2012). 'The impact of bestseller rank on demand: Evidence from the app market'. In: *International Economic Review* 53.3, pp. 717–742.

Carlson, Rodney L (1978). 'Seemingly unrelated regression and the demand for automobiles of different sizes, 1965-75: A disaggregate approach'. In: *Journal of Business*, pp. 243–262.

Carneiro, Jorge and Flávio Faria (2016). 'Quest for purposefully designed conceptualization of the country-of-origin image construct'. In: *Journal of Business Research* 69.10, pp. 4411–4420.

Carpenter, Gregory S and Kent Nakamoto (1989). 'Consumer preference formation and pioneering advantage'. In: *Journal of Marketing research* 26.3, pp. 285–298.

Chakraborty, Uttam (2019). 'The impact of source credible online reviews on purchase intention: The mediating roles of brand equity dimensions'. In: *Journal of Research in Interactive Marketing* 13.2, pp. 142–161.

Chang, Dae Ryun and Se-Bum Park (2013). 'The effects of brand strategy and technological uncertainty on pioneering advantage in the multigenerational product market'. In: *Journal of Product Innovation Management* 30.1, pp. 82–95.

Chen, Faen and Yukio Kodono (2012). 'SWOT analysis and five competitive forces of Chery automobile company'. In: *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*. IEEE, pp. 1959–1962.

Chen, Hsinchun, Roger HL Chiang and Veda C Storey (2012). 'Business intelligence and analytics: From big data to big impact'. In: *MIS quarterly*, pp. 1165–1188.

Chen, Hwei-Chung and Arun Pereira (1999). 'Product entry in international markets: the effect of country-of-origin on first-mover advantage'. In: *Journal of Product & Brand Management* 8.3, pp. 218–231.

Chen, Tianqi and Carlos Guestrin (2016). 'Xgboost: A scalable tree boosting system'. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chen, Yubo, Scott Fay and Qi Wang (2011). 'The role of marketing in social media: How online consumer reviews evolve'. In: *Journal of interactive marketing* 25.2, pp. 85–94.

Chen, Yubo and Jinhong Xie (2008). 'Online consumer review: Word-of-mouth as a new element of marketing communication mix'. In: *Management science* 54.3, pp. 477–491.

Cheng, Zhuo and Tajul Ariffin Masron (2023). 'Economic policy uncertainty and corporate digital transformation: evidence from China'. In: *Applied Economics* 55.40, pp. 4625–4641.

Cheriyan, Sunitha et al. (2018). 'Intelligent sales prediction using machine learning techniques'. In: *2018 International Conference on Computing, Electronics and Communications Engineering*. IEEE, pp. 53–58.

Cheung, Christy MK and Matthew KO Lee (2008). 'Online consumer reviews: does negative electronic word-of-mouth hurt more?' In: *AMCIS 2008 Proceedings*, p. 143.

Chevalier, Judith A and Dina Mayzlin (2006). 'The effect of word of mouth on sales: Online book reviews'. In: *Journal of marketing research* 43.3, pp. 345–354.

Chiang, Roger HL et al. (2018). *Strategic value of big data and business analytics*.

Chiang, Wei-yu Kevin, Dongsong Zhang and Lina Zhou (2006). 'Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression'. In: *Decision Support Systems* 41.2, pp. 514–531.

Choi, Sungwoo et al. (2017). 'The role of power and incentives in inducing fake reviews in the tourism industry'. In: *Journal of Travel Research* 56.8, pp. 975–987.

Chu, Wan-Wen (2011). 'How the Chinese government promoted a global automobile industry'. In: *Industrial and Corporate Change* 20.5, pp. 1235–1276.

Coeurderoy, Regis and Rodolphe Durand (2004). 'Leveraging the advantage of early entry: proprietary technologies versus cost leadership'. In: *Journal of Business Research* 57.6, pp. 583–590.

Conrad, SA (1976). 'Sales data and the estimation of demand'. In: *Journal of the Operational Research Society* 27.1, pp. 123–127.

Cooper, Robert G and Elko J Kleinschmidt (1995). 'Benchmarking the firm's critical success factors in new product development'. In: *Journal of Product Innovation Management: An International Publication of the Product Development & Management Association* 12.5, pp. 374–391.

Cui, Hongyan et al. (2022). 'Self-training method based on GCN for semi-supervised short text classification'. In: *Information Sciences* 611, pp. 18–29.

Cui, Yiming et al. (2021). 'Pre-training with whole word masking for chinese bert'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 3504–3514.

Decker, Reinhold and Kumiko Gnibba-Yukawa (2010). 'Sales forecasting in high-technology markets: A utility-based approach'. In: *Journal of product innovation management* 27.1, pp. 115–129.

Delen, Dursun and Hamed M Zolbanin (2018). 'The analytics paradigm in business research'. In: *Journal of Business Research* 90, pp. 186–195.

DING, Jiang-Min (2023). 'Analyzing Tesla's International Business Strategies: A Closer Look at the Korean and Chinese Markets'. In: *The Journal of Economics, Marketing and Management* 11.5, pp. 15–27.

Ding, Qiang and Michèle Akoorie (2013). 'The characteristics and historical development path of the globalizing Chinese automobile industry'. In: *Journal of Technology Management in China* 8.2, pp. 83–104.

Dong, Feng and Yajie Liu (2020). 'Policy evolution and effect evaluation of new-energy vehicle industry in China'. In: *Resources Policy* 67, p. 101655.

Dong, Lu-yu et al. (2018). 'An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews'. In: *Expert Systems with Applications* 114, pp. 210–223.

Du, Rex Yuxing and Wagner A Kamakura (2012). 'Quantitative trendspotting'. In: *Journal of Marketing Research* 49.4, pp. 514–536.

Erevelles, Sunil, Nobuyuki Fukawa and Linda Swayne (2016). 'Big Data consumer analytics and the transformation of marketing'. In: *Journal of business research* 69.2, pp. 897–904.

Fan, Ruguo et al. (2022). 'The effect of government policies and consumer green preferences on the R&D diffusion of new energy vehicles: a perspective of complex network games'. In: *Energy* 254, p. 124316.

Fan, Shaokun, Raymond YK Lau and J Leon Zhao (2015). 'Demystifying big data analytics for business intelligence through the lens of marketing mix'. In: *Big Data Research* 2.1, pp. 28–32.

Fetscherin, Marc and Mark Toncar (2010). 'The effects of the country of brand and the country of manufacturing of automobiles: An experimental study of consumers' brand personality perceptions'. In: *International Marketing Review* 27.2, pp. 164–178.

Field, Robert Michael (1986). 'The performance of industry during the Cultural Revolution: Second thoughts'. In: *The China Quarterly* 108, pp. 625–642.

Frynas, Jędrzej George, Kamel Mellahi and Geoffrey Allen Pigman (2006). 'First mover advantages in international business and firm-specific political resources'. In: *Strategic Management Journal* 27.4, pp. 321–345.

Gama, Fabio and Stefano Magistretti (2023). 'Artificial intelligence in innovation management: A review of innovation capabilities and a taxonomy of AI applications'. In: *Journal of Product Innovation Management*.

Gao, Hongzhi and John Knight (2007). 'Pioneering advantage and product-country image: evidence from an exploratory study in China'. In: *Journal of Marketing Management* 23.3-4, pp. 367–385.

Garg, Rajiv and Rahul Telang (2013). 'Inferring app demand from publicly available data'. In: *MIS quarterly*, pp. 1253–1264.

George, Gerard et al. (2016). *Big data and data science methods for management research*.

Geva, Tomer et al. (2017). 'Using forum and search data for sales prediction of high-involvement projects'. In: *Mis Quarterly* 41.1, pp. 65–82.

Gong, Huiming, Michael Q Wang and Hewu Wang (2013). 'New energy vehicles in China: policies, demonstration, and progress'. In: *Mitigation and Adaptation Strategies for Global Change* 18, pp. 207–228.

Gottschalk, Sabrina A and Alexander Mafael (2017). 'Cutting through the online review jungle—investigating selective eWOM processing'. In: *Journal of Interactive Marketing* 37.1, pp. 89–104.

Grover, Varun et al. (2018). 'Creating strategic business value from big data analytics: A research framework'. In: *Journal of management information systems* 35.2, pp. 388–423.

Günther, Wendy Arianne et al. (2022). 'Resourcing with data: Unpacking the process of creating data-driven value propositions'. In: *The Journal of Strategic Information Systems* 31.4, p. 101744.

Guo, Liang et al. (2017). 'Automated competitor analysis using big data analytics: Evidence from the fitness mobile app business'. In: *Business Process Management Journal* 23.3, pp. 735–762.

Hajli, Nick et al. (2020). 'Understanding market agility for new product success with big data analytics'. In: *Industrial Marketing Management* 86, pp. 135–143.

Hamzaoui Essoussi, Leila and Dwight Merunka (2007). 'Consumers' product evaluations in emerging markets: does country of design, country of manufacture, or brand image matter?' In: *International Marketing Review* 24.4, pp. 409–426.

Han, Songqiao, Xiaoling Hao and Hailiang Huang (2018). 'An event-extraction approach for business analysis from online Chinese news'. In: *Electronic Commerce Research and Applications* 28, pp. 244–260.

Harwit, Eric (2001). 'The impact of WTO membership on the automobile industry in China'. In: *The China Quarterly* 167, pp. 655–670.

Häubl, Gerald (1996). 'A cross-national investigation of the effects of country of origin and brand name on the evaluation of a new car'. In: *International Marketing Review* 13.5, pp. 76–97.

He, Hongwen et al. (2022a). 'China's battery electric vehicles lead the world: achievements in technology system architecture and technological breakthroughs'. In: *Green Energy and Intelligent Transportation* 1.1, p. 100020.

He, Sherry, Brett Hollenbeck and Davide Proserpio (2022b). 'The market for fake reviews'. In: *Marketing Science* 41.5, pp. 896–921.

Herhausen, Dennis et al. (2024). 'Machine learning in marketing: Recent progress and future research directions'. In: *Journal of Business Research* 170, p. 114254.

Ho-Dac, Nga N, Stephen J Carson and William L Moore (2013). 'The effects of positive and negative online customer reviews: do brand strength and category maturity matter?' In: *Journal of marketing* 77.6, pp. 37–53.

Homburg, Christian and Dominik M Wielgos (2022). 'The value relevance of digital marketing capabilities to firm performance'. In: *Journal of the Academy of Marketing Science* 50.4, pp. 666–688.

Hong, Ilyoo B (2015). 'Understanding the consumer's online merchant selection process: The roles of product involvement, perceived risk, and trust expectation'. In: *International journal of information management* 35.3, pp. 322–336.

Howell, Sabrina T (2018). 'Joint ventures and technology adoption: A Chinese industrial policy that backfired'. In: *Research Policy* 47.8, pp. 1448–1462.

Hu, Nan, Noi Sian Koh and Srinivas K Reddy (2014). 'Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales'. In: *Decision support systems* 57, pp. 42–53.

Hu, Zheng and Jiahai Yuan (2018). 'China's NEV market development and its capability of enabling premium NEV: Referencing from the NEV market performance of BMW and Mercedes in China'. In: *Transportation Research Part A: Policy and Practice* 118, pp. 545–555.

Huang, Tao, Robert Fildes and Didier Soopramanien (2014). 'The value of competitive information in forecasting FMCG retail product sales and the variable selection problem'. In: *European Journal of Operational Research* 237.2, pp. 738–748.

Huff, Lenard C and William T Robinson (1994). 'Note: the impact of leadtime and years of competitive rivalry on pioneer market share advantages'. In: *Management Science* 40.10, pp. 1370–1377.

Hülsmann, Marco et al. (2012). 'General sales forecast models for automobile markets and their analysis.' In: *Trans. Mach. Learn. Data Min.* 5.2, pp. 65–86.

Ij, H (2018). 'Statistics versus machine learning'. In: *Nat Methods* 15.4, p. 233.

Ishizaka, Alessio and Sajid Siraj (2018). 'Are multi-criteria decision-making tools useful? An experimental comparative study of three methods'. In: *European Journal of Operational Research* 264.2, pp. 462–471.

Jacoby, Jacob, Jerry C Olson and Rafael A Haddock (1971). 'Price, brand name, and product composition characteristics as determinants of perceived quality.' In: *Journal of applied psychology* 55.6, p. 570.

Jiang, Hong and Feng Lu (2023). 'New industry paradigms may overwhelm dynamic capabilities: Different competitive dynamics around Tesla and Chinese EV start-ups'. In: *Management and Organization Review* 19.1, pp. 157–169.

Jiang, Zhangsheng and Chenghao Xu (2023). 'Policy incentives, government subsidies, and technological innovation in new energy vehicle enterprises: Evidence from China'. In: *Energy Policy* 177, p. 113527.

Kalla, Rajasekhar, Saravana Murikinjeri and R Abbaiah (2020). 'An improved demand forecasting with limited historical sales data'. In: *2020 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, pp. 1–5.

Kamath, Anil (2015). 'Optimizing Marketing Impact through Data Driven Decisioning'. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1631–1631.

Karabas, Ismail et al. (2021). 'The impact of review valence and awareness of deceptive practices on consumers' responses to online product ratings and reviews'. In: *Journal of Marketing Communications* 27.7, pp. 685–715.

Kardes, Frank R et al. (1993). 'Brand retrieval, consideration set composition, consumer choice, and the pioneering advantage'. In: *Journal of Consumer Research* 20.1, pp. 62–75.

Kenton, Jacob Devlin Ming-Wei Chang and Lee Kristina Toutanova (2019). 'Bert: Pre-training of deep bidirectional transformers for language understanding'. In: *Proceedings of naacL-HLT*. Vol. 1, p. 2.

Kenworthy, Jeffrey, Peter Newman and Yuan Gao (2015). 'Growth of a giant: A historical and current perspective on the Chinese automobile industry'. In: *World Transport Policy and Practice* 21.2, pp. 40–56.

Khaleel, Mohamed et al. (2024). 'Electric vehicles in China, Europe, and the United States: Current trend and market comparison'. In: *Int. J. Electr. Eng. and Sustain.*, pp. 1–20.

Khurshid, Faisal et al. (2018). 'Enactment of ensemble learning for review spam detection on selected features'. In: *International Journal of Computational Intelligence Systems* 12.1, pp. 387–394.

Ko, Eunhee Emily and Douglas Bowman (2023). 'Suspicious online product reviews: An empirical analysis of brand and product characteristics using Amazon data'. In: *International Journal of Research in Marketing* 40.4, pp. 898–911.

Kostyra, Daniel S et al. (2016). 'Decomposing the effects of online customer reviews on brand, price, and product attributes'. In: *International Journal of Research in Marketing* 33.1, pp. 11–26.

Kowalczyk, Martin and Peter Buxmann (2015). 'An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study'. In: *Decision support systems* 80, pp. 1–13.

Kumar, Ajay et al. (2022). 'Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering'. In: *Decision Support Systems* 155, p. 113728.

Lambert, Zarrel V (1972). 'Price and choice behavior'. In: *Journal of marketing Research* 9.1, pp. 35–40.

Lambkin, Mary (1992). 'Pioneering new markets: A comparison of market share winners and losers'. In: *International Journal of Research in Marketing* 9.1, pp. 5–22.

Landwehr, Jan R, Aparna A Labroo and Andreas Herrmann (2011). 'Gut liking for the ordinary: Incorporating design fluency improves automobile sales forecasts'. In: *Marketing Science* 30.3, pp. 416–429.

Lantos, Geoffrey P (2014). 'Marketing to millennials: Reach the largest and most influential generation of consumers ever'. In: *Journal of Consumer Marketing* 31.5, pp. 401–403.

Leangarun, Teema, Poj Tangamchit and Suttipong Thajchayapong (2018). 'Stock price manipulation detection using generative adversarial networks'. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 2104–2111.

Li, Fangtao Huang et al. (2011). 'Learning to identify review spam'. In: *Twenty-second international joint conference on artificial intelligence*.

Li, Xiaolin, Chaojiang Wu and Feng Mai (2019). 'The effect of online reviews on product sales: A joint sentiment-topic analysis'. In: *Information and Management* 56.2, pp. 172–184.

Li, Yan Sheng, Xin Xin Kong and Miao Zhang (2016). 'Industrial upgrading in global production networks: The case of the Chinese automotive industry'. In: *Asia Pacific Business Review* 22.1, pp. 21–37.

Li, Yina et al. (2023). 'Designing government subsidy schemes to promote the electric vehicle industry: A system dynamics model perspective'. In: *Transportation Research Part A: Policy and Practice* 167, p. 103558.

Lib, Qian Liua Xiaoyan (2024). 'Analysis of Investment Value in China's New Energy Vehicle Industry-Taking BYD Company as an Example'. In: *Journal of Advanced Academic Research and Studies* 1.4, pp. 1–17.

Lieberman, Marvin B and David B Montgomery (1988). 'First-mover advantages'. In: *Strategic management journal* 9.S1, pp. 41–58.

Lim, Ee-Peng, Hsinchun Chen and Guoqing Chen (2013). 'Business intelligence and analytics: Research directions'. In: *ACM Transactions on Management Information Systems (TMIS)* 3.4, pp. 1–10.

Liu, Weidong and Henry Wai-chung Yeung (2008). 'China's dynamic industrial sector: the automobile industry'. In: *Eurasian Geography and Economics* 49.5, pp. 523–548.

Logsdon, Jeanne M and Karen DW Patterson (2009). 'Deception in business networks: Is it easier to lie online?' In: *Journal of business ethics* 90, pp. 537–549.

Lundberg, Scott M and Su-In Lee (2017). 'A unified approach to interpreting model predictions'. In: *Advances in neural information processing systems* 30, pp. 4765–4774.

Luo, Changyuan and Yan Zhi (2019). 'Reform and opening up in the new era: China trade policy review'. In: *The World Economy* 42.12, pp. 3464–3477.

Luo, Jianxi (2005). 'The growth of independent Chinese automotive companies'. In: *International Motor Vehicle Program, MIT* 6.

Ma, Shao-Chao, Ying Fan and Lianyong Feng (2017). 'An evaluation of government incentives for new energy vehicles in China focusing on vehicle purchasing restrictions'. In: *Energy Policy* 110, pp. 609–618.

Malbon, Justin (2013). 'Taking fake online consumer reviews seriously'. In: *Journal of Consumer Policy* 36, pp. 139–157.

Mangram, Myles Edwin (2012). 'The globalization of Tesla Motors: a strategic marketing plan analysis'. In: *Journal of Strategic Marketing* 20.4, pp. 289–312.

Marcílio, Wilson E and Danilo M Eler (2020). 'From explanations to feature selection: assessing SHAP values as feature selection mechanism'. In: *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images*. Ieee, pp. 340–347.

Markides, Constantinos and Lourdes Sosa (2013). 'Pioneering and first mover advantages: the importance of business models'. In: *Long range planning* 46.4-5, pp. 325–334.

Martínez, Andrés et al. (2020). 'A machine learning framework for customer purchase prediction in the non-contractual setting'. In: *European Journal of Operational Research* 281.3, pp. 588–596.

Messalas, Andreas, Christos Aridas and Yannis Kanellopoulos (2020). 'Evaluating MASHAP as a faster alternative to LIME for model-agnostic machine learning interpretability'. In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 5777–5779.

Miller, Jeff and Patricia Haden (2006). 'Statistical analysis with the general linear model'. In: *Creative Commons Attribution*.

Min, Sungwook, Manohar U Kalwani and William T Robinson (2006). 'Market pioneer and early follower survival risks: A contingency analysis of really new versus incrementally new product-markets'. In: *Journal of Marketing* 70.1, pp. 15–33.

Mitra, Satanik and Mamata Jenamani (2020). 'OBIM: A computational model to estimate brand image from online consumer review'. In: *Journal of Business Research* 114, pp. 213–226.

Moers, Frank (2005). 'Discretion and bias in performance evaluation: the impact of diversity and subjectivity'. In: *Accounting, Organizations and Society* 30.1, pp. 67–80.

Murray, Martin J (1999). 'Sources Of Capital For A Sino-Foreign Equity Joint Venture: A Case Study Of Shanghai General Motors Corporation'. In: *Journal for Economic Educators* 3.1, pp. 10–17.

Nadkarni, Swen and Reinhard Prügl (2021). 'Digital transformation: a review, synthesis and opportunities for future research'. In: *Management Review Quarterly* 71, pp. 233–341.

Nassirtoussi, Arman Khadjeh et al. (2015). 'Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment'. In: *Expert Systems with Applications* 42.1, pp. 306–324.

Nelson, Michelle R and Jiwoo Park (2015). 'Publicity as covert marketing? The role of persuasion knowledge and ethical perceptions on beliefs and credibility in a video news release story'. In: *Journal of Business Ethics* 130, pp. 327–341.

Nieuwenhuis, Paul and Xiao Lin (2015). 'China's car industry'. In: *The Global Automotive Industry*, pp. 109–126.

Nunnari, Giuseppe and Valeria Nunnari (2017). 'Forecasting monthly sales retail time series: a case study'. In: *2017 IEEE 19th conference on business informatics (CBI)*. Vol. 1. IEEE, pp. 1–6.

Oh, JooSeok, Timothy Paul Connerton and Hyun-Jung Kim (2019). 'The rediscovery of brand experience dimensions with big data analysis: Building for a sustainable brand'. In: *Sustainability* 11.19, p. 5438.

Omar, Maktoba, Robert L Williams Jr and David Lingelbach (2009). 'Global brand market-entry strategy to manage corporate reputation'. In: *Journal of Product & Brand Management* 18.3, pp. 177–187.

Park, Eunhye, Jinah Park and Mingming Hu (2021). 'Tourism demand forecasting with online news data mining'. In: *Annals of Tourism Research* 90, p. 103273.

Park, Sungsik, Woochoel Shin and Jinhong Xie (2023). 'Disclosure in Incentivized Reviews: Does It Protect Consumers?' In: *Management Science* 69.11, pp. 7009–7021.

Patterson, William C (1993). 'First-mover advantage: the opportunity curve'. In: *Journal of Management Studies* 30.5, pp. 759–777.

Petti, Claudio et al. (2021). 'Globalization and innovation with Chinese characteristics: the case of the automotive industry'. In: *International journal of emerging markets* 16.2, pp. 303–322.

Qian, Lixian and Didier Soopramanien (2014). 'Using diffusion models to forecast market size in emerging markets with applications to the Chinese car market'. In: *Journal of business research* 67.6, pp. 1226–1232.

Qu, Lu and Yanwei Li (2019). 'Research on industrial policy from the perspective of demand-side open innovation—A case study of Shenzhen new energy vehicle industry'. In: *Journal of Open Innovation: Technology, Market, and Complexity* 5.2, p. 31.

Radas, Sonja and Steven M Shugan (1998). 'Seasonal marketing and timing new product introductions'. In: *Journal of Marketing Research* 35.3, pp. 296–315.

Raju, Jagmohan S (1992). 'The effect of price promotions on variability in product category sales'. In: *Marketing Science* 11.3, pp. 207–220.

Rezaeinejad, Ismael (2021). 'Automotive industry and its place in the economy: case study Iran auto industry'. In: *Asian Journal of Economics, Finance and Management*, pp. 530–539.

Roth, Martin S and Jean B Romeo (1992). 'Matching product catgeory and country image perceptions: A framework for managing country-of-origin effects'. In: *Journal of international business studies* 23, pp. 477–497.

Rudin, Cynthia (2019). 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'. In: *Nature machine intelligence* 1.5, pp. 206–215.

Russell, Dale W and Cristel Antonia Russell (2006). 'Explicit and implicit catalysts of consumer resistance: The effects of animosity, cultural salience and country-of-origin on subsequent choice'. In: *International Journal of Research in Marketing* 23.3, pp. 321–331.

Sa-Ngasoongsong, Akkarapol et al. (2012). 'Multi-step sales forecasting in automotive industry based on structural relationship identification'. In: *International Journal of Production Economics* 140.2, pp. 875–887.

Sahut, Jean Michel, Michel Laroche and Eric Braune (2024). *Antecedents and consequences of fake reviews in a marketing approach: An overview and synthesis.*

Saridakis, Charalampos and George Baltas (2016). 'Modeling price-related consequences of the brand origin cue: An empirical examination of the automobile market'. In: *Marketing Letters* 27, pp. 77–87.

Saumya, Sunil and Jyoti Prakash Singh (2022). 'Spam review detection using LSTM autoencoder: an unsupervised approach'. In: *Electronic Commerce Research* 22.1, pp. 113–133.

Shamsie, Jamal, Corey Phelps and Jerome Kuperman (2004). 'Better late than never: A study of late entrants in household electrical equipment'. In: *Strategic Management Journal* 25.1, pp. 69–84.

Shao, Lulu, Jun Yang and Min Zhang (2017). 'Subsidy scheme or price discount scheme? Mass adoption of electric vehicles under different market structures'. In: *European Journal of Operational Research* 262.3, pp. 1181–1195.

Shao, Wei, Ke Yang and Xiao Bai (2021). 'Impact of financial subsidies on the R&D intensity of new energy vehicles: A case study of 88 listed enterprises in China'. In: *Energy Strategy Reviews* 33, p. 100580.

Shapley, Lloyd S (1953). 'A value for n-person games'. In: *Contribution to the Theory of Games* 2.

Sharma, Piyush (2011). 'Country of origin effects in developed and emerging markets: Exploring the contrasting roles of materialism and value consciousness'. In: *Journal of International Business Studies* 42, pp. 285–306.

Shen, Caixia et al. (2021). 'Comparison between uniform tariff and progressive consumption tax in the Chinese automobile industry'. In: *The Journal of Industrial Economics* 69.1, pp. 169–213.

Shilong, Zhang et al. (2021). 'Machine learning model for sales forecasting by using XGBoost'. In: *2021 IEEE International Conference on Consumer Electronics and Computer Engineering*. IEEE, pp. 480–483.

Shmueli, Galit and Otto R Koppius (2011). 'Predictive analytics in information systems research'. In: *MIS quarterly*, pp. 553–572.

Singh, Jyoti Prakash et al. (2017). 'Predicting the "helpfulness" of online consumer reviews'. In: *Journal of Business Research* 70, pp. 346–355.

Song, Yang et al. (2023). 'Do fake reviews promote consumers' purchase intention?' In: *Journal of Business Research* 164, p. 113971.

Stringham, Edward Peter, Jennifer Kelly Miller and Jonathan R Clark (2015). 'Overcoming barriers to entry in an established industry: Tesla Motors'. In: *California Management Review* 57.4, pp. 85–103.

Su, Chi-Wei et al. (2021). 'Can new energy vehicles help to achieve carbon neutrality targets?' In: *Journal of Environmental Management* 297, p. 113348.

Subakti, Alvin, Hendri Murfi and Nora Hariadi (2022). 'The performance of BERT as data representation of text clustering'. In: *Journal of big Data* 9.1, p. 15.

Sun, Qi et al. (2021). 'Consumer boycotts, country of origin, and product competition: Evidence from China's automobile market'. In: *Management Science* 67.9, pp. 5857–5877.

Swait, Joffre and Tülin Erdem (2007). 'Brand effects on choice and choice set formation under uncertainty'. In: *Marketing science* 26.5, pp. 679–697.

Swami, Sanjeev and Arindam Dutta (2010). 'Advertising strategies for new product diffusion in emerging markets: Propositions and analysis'. In: *European Journal of Operational Research* 204.3, pp. 648–661.

Tang, Rachel (2012). *China's auto sector development and policies: Issues and implications*. Congressional Research Service.

Tang, Tianwei, Liang Huang and Yan Chen (2020). 'Evaluation of Chinese sentiment analysis APIs based on online reviews'. In: *2020 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, pp. 923–927.

Teece, David J (2019). 'China and the reshaping of the auto industry: A dynamic capabilities perspective'. In: *Management and Organization Review* 15.1, pp. 177–199.

Twyman, Nathan W et al. (2015). 'Robustness of multiple indicators in automated screening systems for deception detection'. In: *Journal of Management Information Systems* 32.4, pp. 215–245.

Urban, Glen L, John R Hauser and John H Roberts (1990). 'Prelaunch forecasting of new automobiles'. In: *Management Science* 36.4, pp. 401–421.

Vana, Prasad and Anja Lambrecht (2021). 'The effect of individual online reviews on purchase likelihood'. In: *Marketing Science* 40.4, pp. 708–730.

Verganti, Roberto, Luca Vendraminelli and Marco Iansiti (2020). 'Innovation and design in the age of artificial intelligence'. In: *Journal of product innovation management* 37.3, pp. 212–227.

Verhoef, Peter C et al. (2021). 'Digital transformation: A multidisciplinary reflection and research agenda'. In: *Journal of business research* 122, pp. 889–901.

Verlegh, Peeter WJ, Jan-Benedict EM Steenkamp and Matthew TG Meulenberg (2005). 'Country-of-origin effects in consumer processing of advertising claims'. In: *international Journal of Research in Marketing* 22.2, pp. 127–139.

Vives, Xavier (2008). 'Innovation and competitive pressure'. In: *The Journal of Industrial Economics* 56.3, pp. 419–469.

Volpato, Giuseppe and Andrea Stocchetti (2008). 'Managing product life cycle in the auto industry: evaluating carmakers effectiveness'. In: *International Journal of Automotive Technology and Management* 8.1, pp. 22–41.

Wang, Fang, Zhao Du and Shan Wang (2023a). 'Information multidimensionality in online customer reviews'. In: *Journal of Business Research* 159, p. 113727.

Wang, Jying-Nan et al. (2018). 'Dynamic effects of customer experience levels on durable product satisfaction: Price and popularity moderation'. In: *Electronic Commerce Research and Applications* 28, pp. 16–29.

Wang, Kai-Hua et al. (2022). 'Is the oil price a barometer of China's automobile market? From a wavelet-based quantile-on-quantile regression perspective'. In: *Energy* 240, p. 122501.

Wang, Lei et al. (2020). 'What influences sales market of new energy vehicles in China? Empirical study based on survey of consumers' purchase reasons'. In: *Energy policy* 142, p. 111484.

Wang, Qi and Jinhong Xie (2014). 'Decomposing pioneer survival: implications for the order-of-entry effect'. In: *Journal of Product Innovation Management* 31.1, pp. 128–143.

Wang, Shuoyao, Suzhi Bi and Yingjun Angela Zhang (2019). 'Reinforcement learning for real-time pricing and scheduling control in EV charging stations'. In: *IEEE Transactions on Industrial Informatics* 17.2, pp. 849–859.

Wang, Shutian, Yan Lin and Guoqing Zhu (2023b). 'Online reviews and high-involvement product sales: Evidence from offline sales in the Chinese automobile industry'. In: *Electronic Commerce Research and Applications* 57, p. 101231.

Wells, Peter (2010). 'Sustainability and diversity in the global automotive industry'. In: *International Journal of Automotive Technology and Management* 10.2-3, pp. 305–320.

Wilson, Andrew E, Peter R Darke and Jaideep Sengupta (2022). 'Winning the battle but losing the war: Ironic effects of training consumers to detect deceptive advertising tactics'. In: *Journal of business ethics*, pp. 1–17.

Wu, Yan, Pim Martens and Thomas Krafft (2022). 'Public awareness, lifestyle and low-carbon city transformation in China: A systematic literature review'. In: *Sustainability* 14.16, p. 10121.

Wu, Yuanyuan et al. (2020). 'Fake online reviews: Literature review, synthesis, and directions for future research'. In: *Decision Support Systems* 132, p. 113280.

Xia, Zhenchang et al. (2020). 'ForeXGBoost: passenger car sales prediction based on XGBoost'. In: *Distributed and Parallel Databases* 38, pp. 713–738.

Xie, Yixiu et al. (2023). 'Success Factors for Emerging Brands in China's New Energy Vehicle Market: The Case of Li Auto'. In: *Industrial Engineering and Innovation Management* 6.11, pp. 7–11.

Yamamoto, Ayana et al. (2017). 'Company relation extraction from web news articles for analyzing industry structure'. In: *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, pp. 89–92.

Yang, Dongjin et al. (2017). 'Policy support for own-brand innovation in China's auto industry: panacea or placebo?' In: *Chinese Management Studies* 11.1, pp. 107–122.

Yoon, Jin Ma and Hyun-Hwa Lee (2014). 'Consumer responses toward online review manipulation'. In: *Journal of Research in Interactive Marketing* 8.3, pp. 224–244.

Yu, Feifei, Liting Wang and Xiaotong Li (2020). 'The effects of government subsidies on new energy vehicle enterprises: The moderating role of intelligent transformation'. In: *Energy Policy* 141, p. 111463.

Yu, Zhang, Syed Abdul Rehman Khan and Muhammad Umar (2022). 'Circular economy practices and industry 4.0 technologies: A strategic move of automobile industry'. In: *Business Strategy and the Environment* 31.3, pp. 796–809.

Yuan, Jia-Zheng and Carles Brasó Broggi (2023). 'The metamorphosis of China's automotive industry (1953–2001): Inward internationalisation, technological transfers and the making of a post-socialist market'. In: *Business History*, pp. 1–28.

Zahoor, Aqib et al. (2023). 'Can the new energy vehicles (NEVs) and power battery industry help China to meet the carbon neutrality goal before 2060?' In: *Journal of Environmental Management* 336, p. 117663.

Zaman, Mustafeed et al. (2023). 'Motives for posting fake reviews: Evidence from a cross-cultural comparison'. In: *Journal of Business Research* 154, p. 113359.

Zhang, Lei and Quande Qin (2018). 'China's new energy vehicle policies: Evolution, comparison and recommendation'. In: *Transportation Research Part A: Policy and Practice* 110, pp. 57–72.

Zhang, Rongjia et al. (2022). 'The innovation effect of intelligent connected vehicle policies in China'. In: *IEEE Access* 10, pp. 24738–24748.

Zhang, Zhe, Alex Yao and Zhiyong Yang (2024). 'Coach Versus Goldlion: The Effect of Socially Versus Personally Oriented Motives on Consumer Preference for Foreign and Domestic Masstige Brands in Emerging Markets'. In: *Journal of International Marketing* 32.3, pp. 101–115.

Zhao, Min (2006). 'Competitive strategy–differentiation of multinational enterprises in China: case study of American, European and Japanese automobile enterprises'. In: *Globalization, Competition and Growth in China*. Routledge, pp. 208–233.

Zhao, Qingyuan and Trevor Hastie (2021). 'Causal interpretations of black-box models'. In: *Journal of Business & Economic Statistics* 39.1, pp. 272–281.

Zhao, Yi et al. (2013). 'Modeling consumer learning from online product reviews'. In: *Marketing science* 32.1, pp. 153–169.

Zhao, Zheng, Jaideep Anand and Will Mitchell (2005). 'A dual networks perspective on inter-organizational transfer of R&D capabilities: international joint ventures in the Chinese automotive industry'. In: *Journal of management studies* 42.1, pp. 127–160.

Zhu, Feng and Xiaoquan Zhang (2010). 'Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics'. In: *Journal of marketing* 74.2, pp. 133–148.

Zhuang, Mengzhou, Geng Cui and Ling Peng (2018). 'Manufactured opinions: The effect of manipulating online product reviews'. In: *Journal of Business Research* 87, pp. 24–35.

Appendices

A Category of car brands in Chinese automobile market

This appendix provides details about the classification of automobile brands, which is mainly based on four categories: domestic luxury brands, joint venture luxury brands, domestic non-luxury brands and joint venture non-luxury brands. In addition, there is only one wholly foreign-owned brand in the Chinese automobile market. We summarized the list of luxury brands in China from some articles published in Dongchedi (https://www.dongchedi.com).

Luxury brands

- **Domestic brands:** Hongqi, NIO, Li Auto
- **Joint venture brands:** Audi, BMW, Mercedes-Benz, Jaguar, Cadillac, Land Rover, Volvo, Lincoln, Acura, Infinity, DS
- **Wholly foreign-owned brands:** Tesla

Non-luxury brands

- **Domestic brands:** Arcfox, MG, R, Seres, SWM, Wey, Aiways, Baojun, Borgward, Beijing, BAIC, BAIC Huansu, Wevan, BAIC BJEV, Beijing Automobile Works, Bestune, Bisu, BYD, BAIC Changhe, Dorcen, Dearcc, Dongfeng, DFMC, Dongfeng Fengguang, Forthing, Dong Feng Aeolus, Dongfeng Fukang, Venucia, Soueast, Ruilan, Foton, Qoros, Gacaion, Trumpchi, Gonow, GAC Group, Zedriv, Haval, Haima, Hanteng, Hycan, Horki, Huasong, Hawtai, Hawtai EV, Huanghai, Geely, Geometry, JAC, JMEV, Jetour, Jinbei, Traum, Karry, Cowin, Letin, Everus, Lifan, Leopaard, Lingbao, Leapmotor, LYNK&CO, LinkTour, Landwind, Neta, Luxgen, Ora, Chery, Chery New Energy, Gleagle, Roewe, Ruichi, Maxus, Ciimo, Tank, Denza, Weltmeister, SGMW, XPeng, NLM, Sitech, Exeed, Yema, FAW, Yingzhi, Jonway, Yusheng, Yogomo, YuluEV, Yudo, Carlarky, Kaicheng, Oushang, Changan, Great Wall, Zhidou, VGV, Zhonghua, ZX, Zotye
- **Joint venture brands:** Jeep, Mitsubishi, Toyota, Peugeot,Isuzu, Buick, Volkswagen, Fiat, Honda, Ford, Jetta, Renault, Suzuki, Mazda, Kia, Nissan, Sihao, Skoda, Hyundai, Chevrolet, Citroen

B Classification standards of car model size

In the automobile market, the car model size is often classified according to engine displacement, length and wheelbase. The table below shows the specific classification details of sedan models, SUV models and MPV models [5].

5. There are two key points that should be noted regarding the size classification standards. First, the classification of vehicle sizes is not entirely consistent across different automakers. Moreover, there is a growing trend for automakers to produce increasingly larger models within the same size category. As a result, there is currently no universally strict standard in the market for defining vehicle size categories, which leads to overlaps in engine displacement, vehicle length, and wheelbase across different segments. In our study, we adopted the classification standards published by the Dongchedi platform, which represent the typical range of engine displacement, length, and wheelbase associated with each size category, rather than mutually exclusive boundaries. Dongchedi (https://www.dongchedi.com) is one of the most authoritative automotive media platforms in the Chinese automobile market. Second, there is no uniform terminology internationally for the largest size category of vehicles; it is most commonly referred to as either a "large vehicle" or a "full-size vehicle." Therefore, in our research, we

Table B.1: Classification standards of car model size for sedan.

Car model size	Classification standard (Sedan)
Mini	Engine displacement <1.0 liter Length <3900 mm Wheelbase <2450 mm
Small	1.0 liter <Engine displacement <1.5 liter 3700 mm <Length <4450 mm 2350 mm <Wheelbase <2600 mm
Compact	1.0 liter <Engine displacement <2.0 liter 4300 mm <Length 4750 <mm 2500 mm <Wheelbase <2750 mm
Mid-size	1.4 liter <Engine displacement <3.0 liter 4600 mm <Length <5000 mm 2650 mm <Wheelbase <2950 mm
Large or full-size	Engine displacement >2.0 liter 4800 mm <Length <5250 mm 2800 mm <Wheelbase <3150 mm **or** Engine displacement >3.0 liter Length >5100 mm Wheelbase >2800 mm

Table B.2: Classification standards of car model size for SUV.

Car model size	Classification standard (SUV)
Small	Length <4450 mm Wheelbase <2650 mm
Compact	4200 mm <Length <4600 mm 2500 mm <Wheelbase 2700 mm
Mid-size	4500 mm <Length <4900 mm 2700 mm <Wheelbase <2950 mm
Large or full-size	4800 mm <Length <5200 mm 2800 mm <Wheelbase <3150 mm **or** Length >5100 mm Wheelbase >2900 mm

refer to this category as ”large or full-size vehicle.” However, in the Chinese sedan and SUV markets, automakers have further subdivided this category into two subgroups: a relatively smaller subgroup (full-size (Small)) and a relatively larger one (full-size (Larger)). As such, for the ”large or full-size” category, we show two separate classification standards.

Table B.3: Classification standards of car model size for MPV.

Car model size	**Classification standard (MPV)**
Minivan	Length <3500 mm
Compact	Length <4800 mm Wheelbase <2850 mm
Mid-size	4750 mm <Length <4950 mm 2850 mm <Wheelbase <3000 mm
Large or full-size	Length >5300 mm Wheelbase >3400 mm

C Results of clustering algorithms in different market segments

This appendix shows the clustering results of clustering algorithms in 12 market segments. Observing these clustering results also helps us preliminarily select the algorithm with better clustering effect in the experiment.

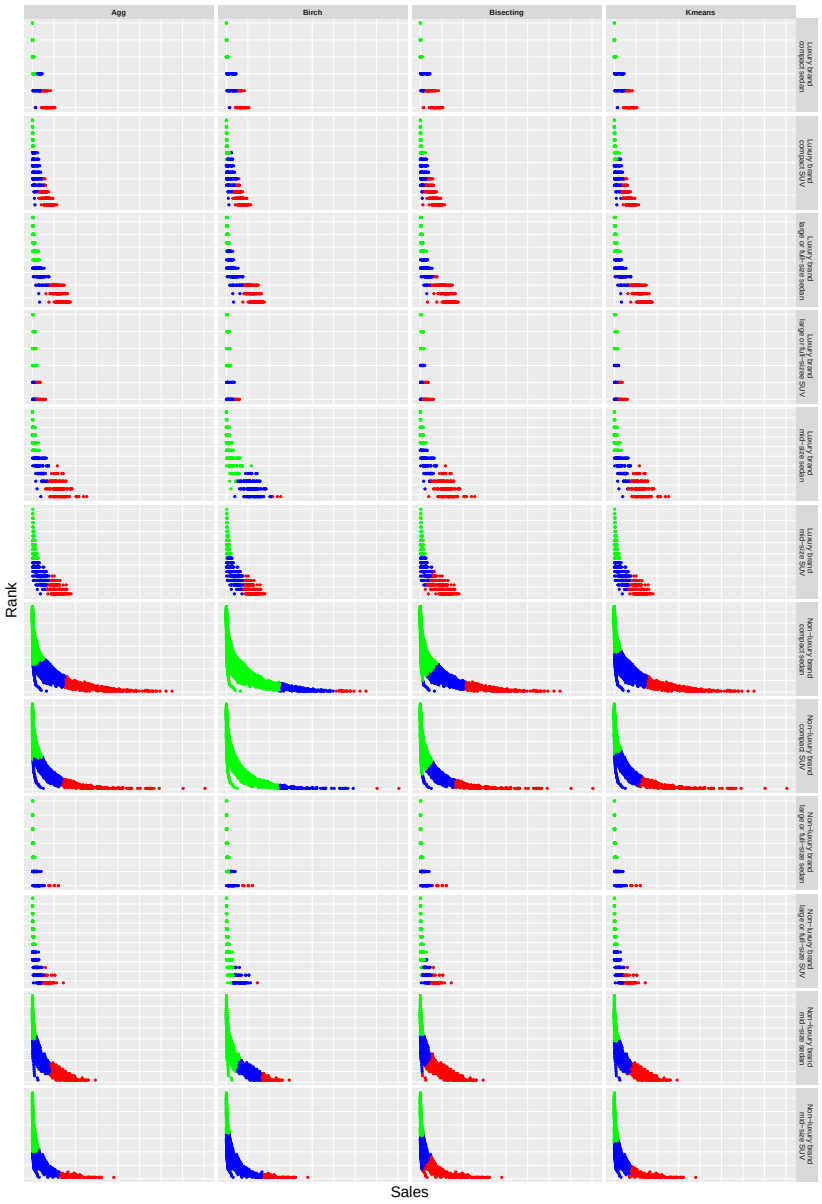

Figure C.1: Results of four clustering algorithms in 12 market segments.

D Parameter search range and optimal parameter setting

This appendix provides the search range of parameters adopted by machine learning models in our experiment and the optimal parameters for different models in different market segments. The parameter search range is set according to the default parameters in scikit learn (https://scikit-learn.org/stable/).

Table D.1: Parameter search range.

Model	Parameter	Range
LR	penalty	[”l1”, ”l2”]
	C	[0.01, 0.1, 1, 10, 100]
XGB	max_depth	[4, 5, 6, 7, 8]
	min_child_weight	[1, 2, 3, 4, 5]
	gamma	[0, 0.1, 0.2, 0.3, 0.4]
RF	max_depth	[1, 2, 3, 4, 5]
	min_samples_leaf	[1, 2, 3, 4, 5]
	n_estimators	[100, 150, 200, 250, 300]
HGB	learning_rate	[0.01, 0.05, 0.1, 0.15, 0.2]
	max_depth	[1, 2, 3, 4, 5]
	min_samples_leaf	[5, 10, 15, 20, 25, 30, 35]
SVM	C	[0.01, 0.1, 1, 10, 100]
MLP	solver	[”adam”, ”sgd”, ”lbfgs”]
	hidden_layer_sizes	[(10,), (50,), (100,), (100, 100)]
	alpha	[0.0001, 0.001, 0.01, 0.1, 1]

Table D.2: Optimal parameter setting in luxury brand compact sedan market.

Model	Optimal parameter setting
LR	{’C’: 1, ’penalty’: ’l2’}
XGB	{’gamma’: 0, ’max_depth’: 4, ’min_child_weight’: 3}
RF	{’max_depth’: 5, ’min_samples_leaf’: 1, ’n_estimators’: 100}
HGB	{’learnin_rate’: 0.01, ’max_depth’: 3, ’min_samples_leaf’: 20}
SVM	{’C’: 100}
MLP	{’alpha’: 1, ’hidden_layer_sizes’: (100,), ’solver’: ’adam’}

Table D.3: Optimal parameter setting in luxury brand mid-size sedan market.

Model	Optimal parameter setting
LR	{'C': 1, 'penalty': 'l2'}
XGB	{'gamma': 0.1, 'max_depth': 7, 'min_child_weight': 1}
RF	{'max_depth': 5, 'min_samples_leaf': 3, 'n_estimators': 100}
HGB	{'learning_rate': 0.01, 'max_depth': 5, 'min_samples_leaf': 15}
SVM	{'C': 100}
MLP	{'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'solver': 'adam'}

Table D.4: Optimal parameter setting in luxury brand large or full-size sedan market.

Model	Optimal parameter setting
LR	{'C': 10, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 4, 'min_child_weight': 3}
RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 250}
HGB	{'learning_rate': 0.15, 'max_depth': 5, 'min_samples_leaf': 10}
SVM	{'C': 100}
MLP	{'alpha': 0.0001, 'hidden_layer_sizes': (50,), 'solver': 'lbfgs'}

Table D.5: Optimal parameter setting in luxury brand compact SUV market.

Model	Optimal parameter setting
LR	{'C': 100, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 7, 'min_child_weight': 2}
RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 300}
HGB	{'learning_rate': 0.2, 'max_depth': 5, 'min_samples_leaf': 20}
SVM	{'C': 100}
MLP	{'alpha': 1, 'hidden_layer_sizes': (100, 100), 'solver': 'lbfgs'}

Table D.6: Optimal parameter setting in luxury brand mid-size SUV market.

Model	Optimal parameter setting
LR	{'C': 100, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 5, 'min_child_weight': 1}
RF	{'max_depth': 5, 'min_samples_leaf': 2, 'n_estimators': 100}
HGB	{'learning_rate': 0.15, 'max_depth': 5, 'min_samples_leaf': 10}
SVM	{'C': 100}
MLP	{'alpha': 0.0001, 'hidden_layer_sizes': (10,), 'solver': 'adam'}

Table D.7: Optimal parameter setting in luxury brand large or full-size SUV market.

Model	Optimal parameter setting
LR	{'C': 10, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 5, 'min_child_weight': 2}
RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 100}
HGB	{'learning_rate': 0.15, 'max_depth': 5, 'min_samples_leaf': 5}
SVM	{'C': 100}
MLP	{'alpha': 0.0001, 'hidden_layer_sizes': (50,), 'solver': ' lbfgs '}

Table D.8: Optimal parameter setting in non-luxury brand compact sedan market.

Model	Optimal parameter setting
LR	{'C': 10, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 8, 'min_child_weight': 1}
RF	{'max_depth': 5, 'min_samples_leaf': 2, 'n_estimators': 150}
HGB	{'learning_rate': 0.2, 'max_depth': 5, 'min_samples_leaf': 5}
SVM	{'C': 100}
MLP	{'alpha': 1, 'hidden_layer_sizes': (100, 100), 'solver': 'lbfgs'}

Table D.9: Optimal parameter setting in non-luxury brand mid-size sedan market.

Model	Optimal parameter setting
LR	{'C': 0.1, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 8, 'min_child_weight': 2}
RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 150}
HGB	{'learning_rate': 0.1, 'max_depth': 5, 'min_samples_leaf': 25}
SVM	{'C': 100}
MLP	{'alpha': 1, 'hidden_layer_sizes': (100,), 'solver': ' lbfgs'}

Table D.10: Optimal parameter setting in non-luxury brand large or full-size sedan market.

Model	Optimal parameter setting
LR	{'C': 0.01, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 8, 'min_child_weight': 1}
RF	{'max_depth': 5, 'min_samples_leaf': 2, 'n_estimators': 100}
HGB	{'learning_rate': 0.2, 'max_depth': 3, 'min_samples_leaf': 30}
SVM	{'C': 0.01}
MLP	{'alpha': 1, 'hidden_layer_sizes': (100,), 'solver': 'adam'}

Table D.11: Optimal parameter setting in non-luxury brand compact SUV market.

Model	Optimal parameter setting
LR	{'C': 100, 'penalty': 'l2'}
XGB	{'gamma': 0, 'max_depth': 8, 'min_child_weight': 1}
RF	{'max_depth': 5, 'min_samples_leaf': 3, 'n_estimators': 150}
HGB	{'learning_rate': 0.2, 'max_depth': 5, 'min_samples_leaf': 5}
SVM	{'C': 100}
MLP	{'alpha': 0.1, 'hidden_layer_sizes': (100,), 'solver': 'lbfgs'}

Table D.12: Optimal parameter setting in non-luxury brand mid-size SUV market.

Model	Optimal parameter setting
LR	{'C': 0.1, 'penalty': 'l2'}
XGB	{'gamma': 0.1, 'max_depth': 8, 'min_child_weight': 1}
RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 250}
HGB	{'learning_rate': 0.2, 'max_depth': 5, 'min_samples_leaf': 5}
SVM	{'C': 10}
MLP	{'alpha': 0.001, 'hidden_layer_sizes': (10,), 'solver': 'adam'}

Table D.13: Optimal parameter setting in non-luxury brand large or full-size SUV market.

Model	Optimal parameter setting
LR	{'C': 0.01, 'penalty': 'l2'}
XGB	{'gamma': 0.4, 'max_depth': 8, 'min_child_weight': 2}
RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 300}
HGB	{'learning_rate': 0.01, 'max_depth': 4, 'min_samples_leaf': 5}
SVM	{'C': 100}
MLP	{'alpha': 0.001, 'hidden_layer_sizes': (50,), 'solver': 'lbfgs'}

E Global explanation results in predicting sales performance

This appendix provides the global explanation results in our experiment, namely the importance of each feature in 12 market segments. These represent the impact of these features on car sales performance in 12 market segments.

Table E.1: Feature importance in luxury brand compact sedan market.

Feature importance	Bad	Good	Medium
Minimum price	2.9486	0.0953	1.1521
Sales year	0.9591	0.3942	1.0297
Sales month	0.4966	0.4643	0.7545
Model launch year	0.1088	0.3643	0.3510
Brand establishment year	0.0217	2.3899	0.4482
Brand enter China year	0.0000	0.0000	0.0000
Model(Fuel)	0.0000	0.0000	0.0000
Brand(Traditional)	0.0000	0.0000	0.0000
COO(France)	0.0000	0.0000	0.0000
COO(Germany)	0.0000	0.0000	0.0000

Table E.2: Feature importance in luxury brand mid-size sedan market.

Feature importance	Bad	Good	Medium
Minimum price	0.4357	0.1371	0.7898
Sales year	0.7602	0.2422	0.9209
Sales month	0.4096	0.2342	0.4859
Model launch year	1.8960	2.8470	0.2565
Brand establishment year	0.6872	0.6955	0.2251
Brand enter China year	0.0624	0.0000	0.2674
Model(NEV)	0.0000	0.0000	0.0190
Model(Fuel)	0.0000	0.0000	0.0000
Brand(NEV)	0.0000	0.0000	0.0000
Brand(Traditional)	0.0000	0.0000	0.0000
COO(America)	0.0000	0.0000	0.0538
COO(China)	0.0000	0.0000	0.0000
COO(France)	0.0222	0.0000	0.0096
COO(Germany)	0.0000	0.0000	0.0000
COO(Japan)	0.0063	0.0000	0.0052
COO(Sweden)	0.0096	0.0000	0.0997
COO(UK)	0.0000	0.0000	0.0369

Table E.3: Feature importance in luxury brand large or full-size sedan market.

Feature importance	Bad	Good	Medium
Minimum price	0.6682	0.2620	0.8883
Sales year	0.7088	0.6109	0.5964
Sales month	0.5344	0.4053	0.6673
Model launch year	1.6352	3.1964	0.6317
Brand establishment year	0.2483	0.1107	0.2127
Brand enter China year	0.1761	0.4087	0.3019
Model(NEV)	0.0000	0.0000	0.0000
Model(Fuel)	0.0000	0.0000	0.0000
Brand(Traditional)	0.0000	0.0000	0.0000
COO(America)	0.0787	0.0000	0.0398
COO(China)	0.0000	0.0000	0.0000
COO(Germany)	0.2326	0.0000	0.0045
COO(Sweden)	0.1042	0.0000	0.0128
COO(UK)	0.1678	0.0000	0.0549

Table E.4: Feature importance in luxury brand compact SUV market.

Feature importance	Bad	Good	Medium
Minimum price	0.7364	0.4285	0.8370
Sales year	1.3351	0.5820	0.9688
Sales month	0.6251	0.5676	0.8026
Model launch year	0.4684	0.9462	0.3389
Brand establishment year	1.1693	2.4783	0.2500
Brand enter China year	0.4007	0.1648	0.2819
Model(NEV)	0.0664	0.0000	0.0159
Model(Fuel)	0.0000	0.0000	0.0000
Brand(Traditional)	0.0000	0.0000	0.0000
COO(America)	0.0939	0.4265	0.0223
COO(China)	0.0000	0.0000	0.0000
COO(France)	0.0000	0.0000	0.0000
COO(Germany)	0.0000	0.0000	0.0340
COO(Japan)	0.0000	0.0000	0.0000
COO(Sweden)	0.1189	0.0000	0.0879
COO(UK)	0.0000	0.0000	0.0000

Table E.5: Feature importance in luxury brand mid-size SUV market.

Feature importance	Bad	Good	Medium
Minimum price	1.1334	1.3690	0.8877
Sales year	1.7564	0.7349	1.0429
Sales month	0.5353	0.6093	0.6953
Model launch year	0.2477	0.2960	0.6447
Brand establishment year	0.4293	0.9907	0.4620
Brand enter China year	1.4697	0.1890	0.1527
Model(NEV)	0.1305	0.0219	0.0260
Model(Fuel)	0.0000	0.0000	0.0000
Brand(NEV)	0.0000	0.0000	0.0000
Brand(Traditional)	0.0000	0.0000	0.0000
COO(America)	0.0935	0.3753	0.0483
COO(China)	0.0562	0.0000	0.0706
COO(France)	0.0000	0.0000	0.0000
COO(Germany)	0.0000	1.4122	0.2525
COO(Japan)	0.0637	0.0206	0.0748
COO(Sweden)	0.0000	0.0000	0.0114
COO(UK)	0.1159	0.0000	0.0870

Table E.6: Feature importance in luxury brand large or full-size SUV market.

Feature importance	Bad	Good	Medium
Minimum price	1.1251	0.7189	0.4333
Sales year	0.5937	0.6474	1.2623
Sales month	0.6250	0.7971	0.9209
Model launch year	0.1162	0.0060	0.3812
Brand establishment year	0.5339	1.1668	0.2209
Brand enter China year	1.0084	0.7254	0.0424
Model(NEV)	0.0770	0.0000	0.2492
Model(Fuel)	0.0000	0.0000	0.0000
Brand(NEV)	0.0000	0.0000	0.0000
Brand(Traditional)	0.0000	0.0000	0.0000
COO(America)	0.0038	0.0000	0.1841
COO(China)	0.0000	0.0000	0.0000
COO(Germany)	0.0000	0.0000	0.0000
COO(Sweden)	0.0000	0.0000	0.0000

Table E.7: Feature importance in non-luxury brand compact sedan market.

Feature importance	Bad	Good	Medium
Minimum price	1.0636	1.0846	0.4966
Sales year	0.6961	0.7318	0.4161
Sales month	0.4550	0.8902	0.4486
Model launch year	0.7885	1.2278	0.6073
Brand establishment year	0.5770	0.9730	0.2741
Brand enter China year	0.1680	0.5578	0.3064
Model(NEV)	0.1677	0.1127	0.0495
Model(Fuel)	0.0000	0.0000	0.0000
Brand(NEV)	0.0237	0.0719	0.0097
Brand(Traditional)	0.0000	0.0000	0.0000
COO(America)	0.1349	0.2426	0.0159
COO(China)	0.6442	0.1822	0.0042
COO(Czech Republic)	0.0034	0.0000	0.0077
COO(France)	0.0669	0.1450	0.0068
COO(Germany)	0.1274	0.0391	0.0813
COO(Italy)	0.0200	0.0000	0.0062
COO(Japan)	0.1077	0.0836	0.0781
COO(Korea)	0.0650	0.0548	0.0170

Table E.8: Feature importance in non-luxury brand mid-size sedan market.

Feature importance	Bad	Good	Medium
Minimum price	1.1175	0.7961	0.5371
Sales year	0.6392	0.6843	0.4627
Sales month	0.5699	0.6872	0.4887
Model launch year	1.0575	1.8344	0.5392
Brand establishment year	0.6266	0.6388	0.3280
Brand enter China year	0.4734	0.8575	0.4262
Model (NEV)	0.0624	0.0000	0.1288
Model (Fuel)	0.0000	0.0000	0.0000
Brand (NEV)	0.0039	0.0000	0.0000
Brand (Traditional)	0.0000	0.0000	0.0000
COO (America)	0.0425	0.2646	0.0322
COO (China)	0.0000	0.0000	0.0000
COO (Czech Republic)	0.0011	0.0000	0.0261
COO (France)	0.3233	0.0394	0.0090
COO (Germany)	0.0247	0.0608	0.0560
COO (Japan)	0.1969	0.0147	0.0777
COO (Korea)	0.0436	0.2382	0.0525

Table E.9: Feature importance in non-luxury brand large or full-size sedan market.

Feature importance	Bad	Good	Medium
Minimum price	1.2961	0.0537	0.9021
Sales year	0.7815	0.5338	0.8129
Sales month	0.5424	0.4528	0.9578
Model launch year	0.6508	3.1429	0.3125
Brand establishment year	1.3221	0.0000	1.3530
Brand enter China year	0.0247	0.0000	0.0117
Model (NEV)	0.0000	0.0000	0.0000
Model (Fuel)	0.0000	0.0000	0.0000
Brand (NEV)	0.0000	0.0000	0.0000
Brand (Traditional)	0.0000	0.0000	0.0000
COO (China)	0.0000	0.0000	0.0000
COO (Germany)	0.0000	0.0000	0.0000
COO (Japan)	0.0000	0.0000	0.0000

Table E.10: Feature importance in non-luxury brand compact SUV market.

Feature importance	Bad	Good	Medium
Minimum price	1.0524	1.6472	0.6646
Sales year	0.6330	0.6542	0.5042
Sales month	0.4458	0.7717	0.4589
Model launch year	0.4906	0.8776	0.4021
Brand establishment year	0.6029	0.9981	0.3835
Brand enter China year	0.3383	0.3631	0.1521
Model (NEV)	0.1783	0.1208	0.0760
Model (Fuel)	0.0000	0.0000	0.0000
Brand (NEV)	0.0079	0.0000	0.0150
Brand (Traditional)	0.0000	0.0000	0.0000
COO (America)	0.0105	0.0045	0.0263
COO (China)	0.0280	0.0212	0.0440
COO (Czech Republic)	0.0034	0.0000	0.0025
COO (France)	0.0249	0.0286	0.0270
COO (Germany)	0.0651	0.0877	0.0199
COO (Japan)	0.5126	0.0654	0.0477
COO (Korea)	0.0206	0.0270	0.0815

Table E.11: Feature importance in non-luxury brand mid-size SUV market.

Feature importance	Bad	Good	Medium
Minimum price	1.0861	1.7420	0.6392
Sales year	1.0942	1.0773	0.4888
Sales month	0.5914	0.7118	0.5522
Model launch year	0.5629	0.7262	0.3915
Brand establishment year	0.6140	1.0165	0.3206
Brand enter China year	0.2039	0.3180	0.1531
Model (NEV)	0.1319	0.0659	0.0692
Model (Fuel)	0.0000	0.0000	0.0000
Brand (NEV)	0.0000	0.0000	0.0287
Brand (Traditional)	0.0000	0.0000	0.0000
COO (America)	0.0934	0.0784	0.0753
COO (China)	0.0072	0.0000	0.0021
COO (Czech Republic)	0.0078	0.0000	0.0149
COO (France)	0.0000	0.0000	0.0000
COO (Germany)	0.0893	0.0865	0.0198
COO (Japan)	0.0780	0.0761	0.1449
COO (Korea)	0.0689	0.0468	0.0108

Table E.12: Feature importance in non-luxury brand large or full-size SUV market.

Feature importance	Bad	Good	Medium
Minimum price	1.3591	0.2906	0.3786
Sales year	0.5207	0.6161	0.5260
Sales month	0.3185	0.3859	0.4489
Model launch year	0.2755	0.3064	0.6141
Brand establishment year	0.1617	0.0943	0.3241
Brand enter China year	0.1679	0.6735	0.2071
Model (Fuel)	0.0000	0.0000	0.0000
Brand (Traditional)	0.0000	0.0000	0.0000
COO (America)	0.0000	0.0000	0.0000
COO (China)	0.0000	0.0000	0.0000
COO (Germany)	0.3107	1.2132	0.0769
COO (Japan)	0.0249	0.0000	0.0000

F Local explanation results for brand first-mover advantages

This appendix provides local explanation results for Brand Establishment Year and Brand Enter China Year. By observing the changes of SHAP values of Brand Establishment Year and Brand Enter China Year in different sales performance, we can clarify the impact of these two variables on car sales performance. This is an important basis for us to judge the existence of brand first-mover advantage.

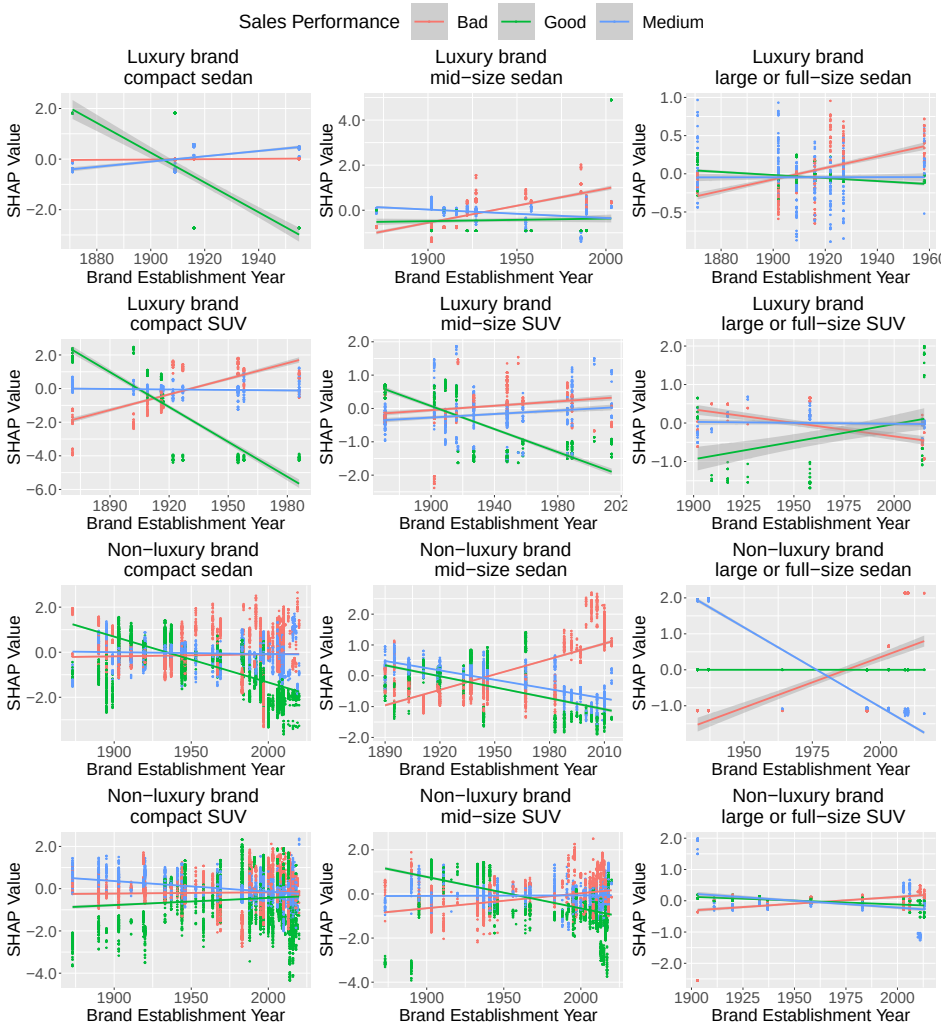

Figure F.1: Distribution of SHAP values of brand establishment year with the change of brand establishment year.

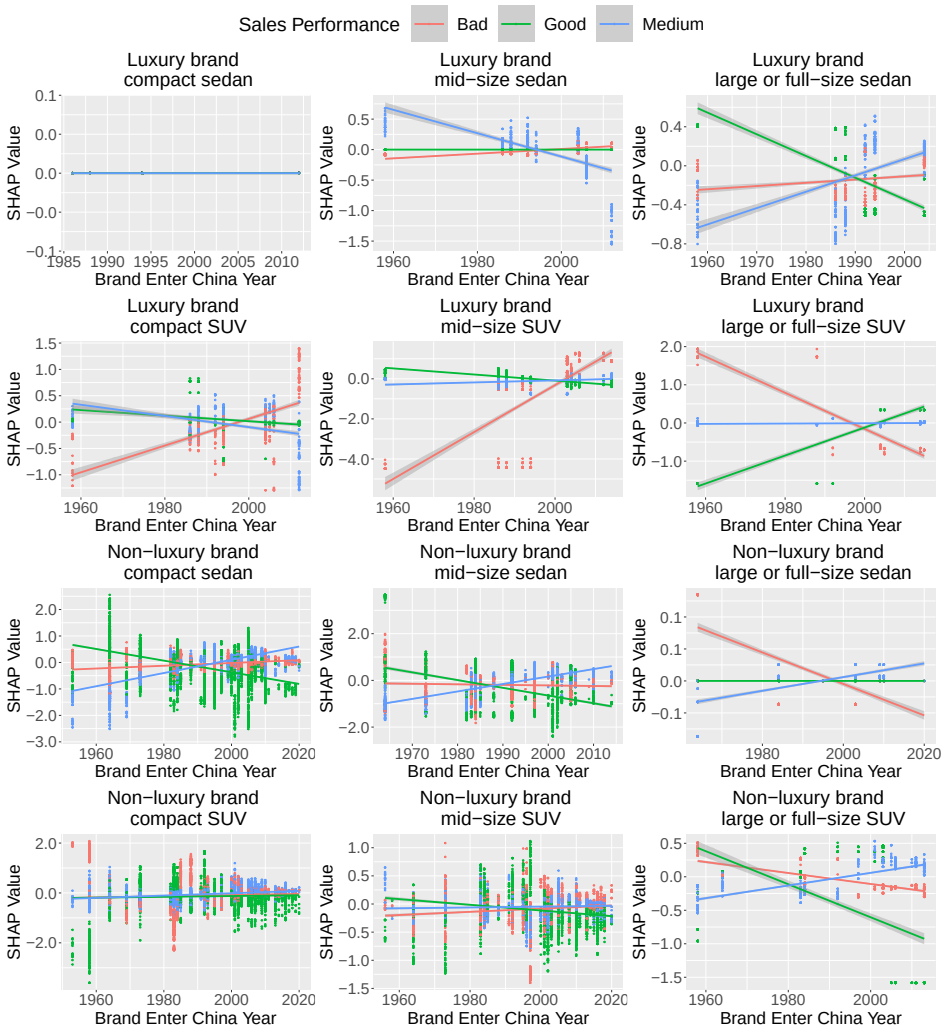

Figure F.2: Distribution of SHAP values of brand enter China year with the change of brand enter China year.

G Market segments that have NEVs and emerging NEV brands

This appendix provides statistics on the number of new energy vehicle models, the quantity of new energy vehicle sales data, the number of emerging new energy vehicle brands and the number of observations of emerging new energy vehicle brand sales data. Based on these statistical results, we screened out the market segments with new energy models and emerging new energy vehicle brands. These segments are used to study whether innovative late entrants can break the first-mover advantage.

Table G.1: Market segments that have new energy vehicles in selected 12 market segments.

Brand luxury	**Car utility type**	**Car model size**	**Number of models**	**Number of observations**
Luxury brand	Sedan	Compact	0	0
		Mid-size	3	42
		Large or full-size	3	69
	SUV	Compact	3	71
		Mid-size	7	86
		Large or full-size	3	53
Non-luxury brand	Sedan	Compact	51	1028
		Mid-size	15	176
		Large or full-size	2	14
	SUV	Compact	35	456
		Mid-size	18	264
		Large or full-size	0	0

Table G.2: Market segments that have emerging new energy vehicle brands in selected 12 market segments.

Brand luxury	Car utility type	Car model size	Number of models	Number of observations
Luxury brand	Sedan	Compact	0	0
		Mid-size	1	16
		Large or full-size	0	0
	SUV	Compact	0	0
		Mid-size	2	37
		Large or full-size	2	49
Non-luxury brand	Sedan	Compact	5	147
		Mid-size	3	48
		Large or full-size	1	3
	SUV	Compact	7	133
		Mid-size	8	91
		Large or full-size	0	0

H Variables description in analysis of fake reviews

This appendix mainly introduces all the original variables used in the detection of fake reviews and their definitions. This contributes to a deeper understanding of the results of the analysis.

Table H.1: Original variables in online review dataset.

Variables	Description
Car series	Name of the car series.
Brand	Name of the brand.
Size	Vehicle size category: mini, minivan, small, compact, mid-size, full-size(Small) , full-size(Large).
Car model type	Car model category: Sedan, SUV and MPV.
User ID	Name that users use when making online reviews.
Year of review	Year when the user reviews.
Month of review	Month when the user reviews.

Variables	Description
Specific model purchased	Specific model of a car series purchased by a user.
Official price	Official prices for specific models purchased.
Car energy type	Vehicle energy type: gasoline vehicle, diesel vehicle, hydrogen vehicle, and so on.
Brand energy type	Brand category based on energy type of vehicle produced.
Brand country of origin	Country in which the brand was created.
Brand establishment date	Year when the brand was created.
Brand entered China date	Year when the brand officially entered the Chinese market.
Model launch date	Year when the car series was officially launched on the Chinese market.
Year of purchase	Year when the user purchased the model.
Month of purchase	Month when the user purchased the model.
Review lag time	Months between purchase date and review posting date.
Province	Province in which the user purchased the model.
City	City in which the user purchased the model.
Transaction price	Real transaction price of the model.
Average energy consumption	Gasoline, diesel, electricity, or hydrogen consumed for every 100 kilometers traveled.
Mileage	Kilometers the user has driven the model at the date the review was posted.
Overall rating	User's overall rating of the vehicle purchased.
Exterior rating	User's rating of the exterior of the car model.
Interior rating	User's rating of the interior of the car model.
Space rating	User's rating of the space of the car model.
Features rating	User's rating of the feature of the car model.

Variables	Description
Power rating	User's rating of the power of the car model.
Energy consumption rating	User's rating of the energy consumption of the car model.
Driving rating	User's rating of the driving of the car model.
Comfort rating	User's rating of the comfort of the car model.
Advantage	The advantages of the model as perceived by the user.
Disadvantage	The disadvantage of the model as perceived by the user.
Exterior comments	User's comments on the exterior of the car model.
Interior comments	User's comments on the interior of the car model.
Space comments	User's comments on the space of the car model.
Features comments	User's comments on the feature of the car model.
Power comments	User's comments on the power of the car model.
Energy consumption comments	User's comments on the energy consumption of the car model.
Driving comments	User's comments on the driving of the car model.
Comfort comments	User's comments on the comfort of the car model.
Is anonymous review	Whether the user made the online review anonymously.
Brand establishment duration	The date difference between the year the user reviews and the year the brand is established.
Brand enter China duration	The date difference between the year the user reviews and the year the brand entered Chinese market.
Model launch duration	The date difference between the year the user reviews and the year the car model is launched.

Variables	Description
Discount	The car price discount enjoyed by the owner at the time of purchasing the model.
Overall review text	The text after merging all comments and the text variables used to assist in understanding the content of all comments.
Overall text sentiment	Sentiment score of the overall review text.
Sales at review	Sales of the model in the month when the user wrote the review.
Brand sales at review	Sales of the brand in the month when the user wrote the review.
Market segment sales at review	Sales of the market segment in the month when the user wrote the review.

Note: The table shows all variables and their definitions in the processed review dataset, in which variables without shading are used for subsequent fake review detection.

I Feature importance in analysis of fake reviews

In this appendix, the importance of all features in predicting fake reviews is mainly presented, as well as their importance in positive fake reviews. This is helpful to study the manipulation behavior, manipulation timing and manipulation motivation of fake reviews.

Table I.1: Feature importance for all and positive reviews

Feature	Importance	
	All reviews	***Positive reviews***
Overall review text	1.47785378	1.47753603
Overall rating	0.84486567	0.84475490
Average energy consumption	0.71696812	0.71701235
Brand enter China duration	0.62492985	0.62499891
Model launch date	0.45315606	0.45308815
Sales at review	0.30664197	0.30668266
Month of review	0.30606079	0.30605931
Month of purchase	0.19691887	0.19691912
Transaction price	0.17072946	0.17074219
Official price	0.13819222	0.13820115
Review lag time	0.13407788	0.13406184
Brand sales at review	0.12656699	0.12656660
Overall text sentiment	0.10543651	0.10543650
Year of review	0.09345236	0.09343297
Discount	0.08259058	0.08258335
Car energy type	0.07855183	0.07855234
Model launch duration	0.04623997	0.04624445
Market segment sales at review	0.04564521	0.04564869
Brand establishment duration	0.03453660	0.03453301
Brand entered China date	0.03127672	0.03127622
Brand country of origin	0.01708815	0.01708584
Driving rating	0.01372392	0.01372309
Mileage	0.01170892	0.01170965
Brand establishment date	0.01010938	0.01010900
Car model type	0.00811641	0.00811734
Size	0.00443876	0.00443879
Year of purchase	0.00265472	0.00265400
Exterior rating	0.00228713	0.00228611
Power rating	0.00196068	0.00196016
Comfort rating	0.00074332	0.00074342
Energy consumption rating	0.00003588	0.00003584
Interior rating	0.00000000	0.00000000
Space rating	0.00000000	0.00000000
Features rating	0.00000000	0.00000000
Is anonymous review	0.00000000	0.00000000
Brand energy type	0.00000000	0.00000000

J Results of the Elbow method used in analysis of fake reviews.

This appendix presents the exploration of the optimal number of clusters for sales across all automobile market segments using the elbow method in the study of the manipulation timing of fake reviews, along with the detailed analysis results.

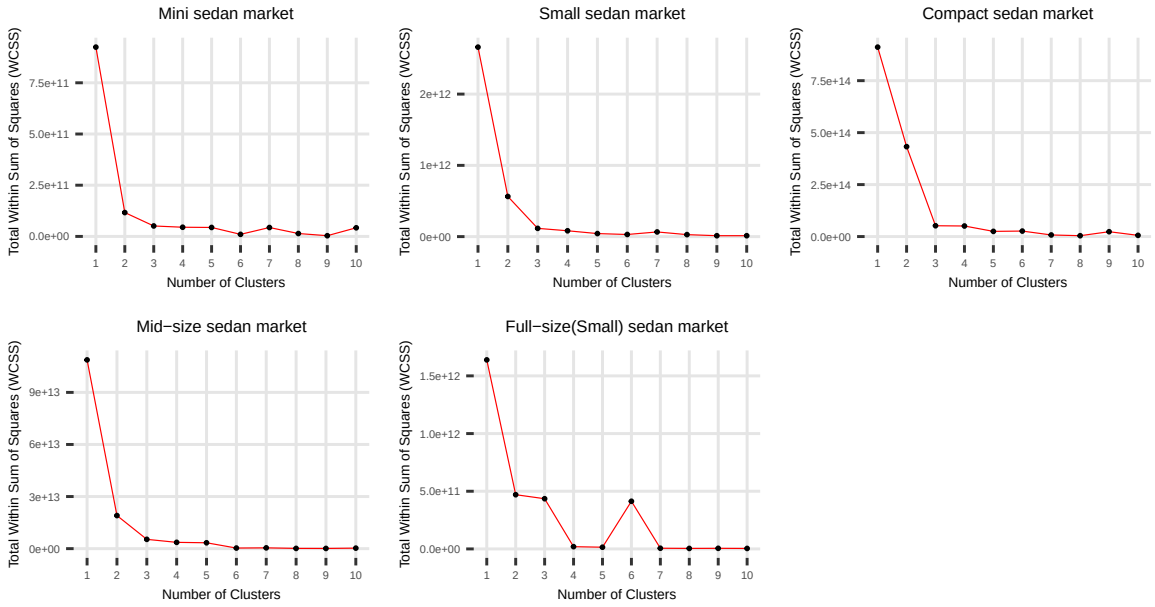

Figure J.1: Results of the Elbow method used in sales clustering for different sedan market segments.

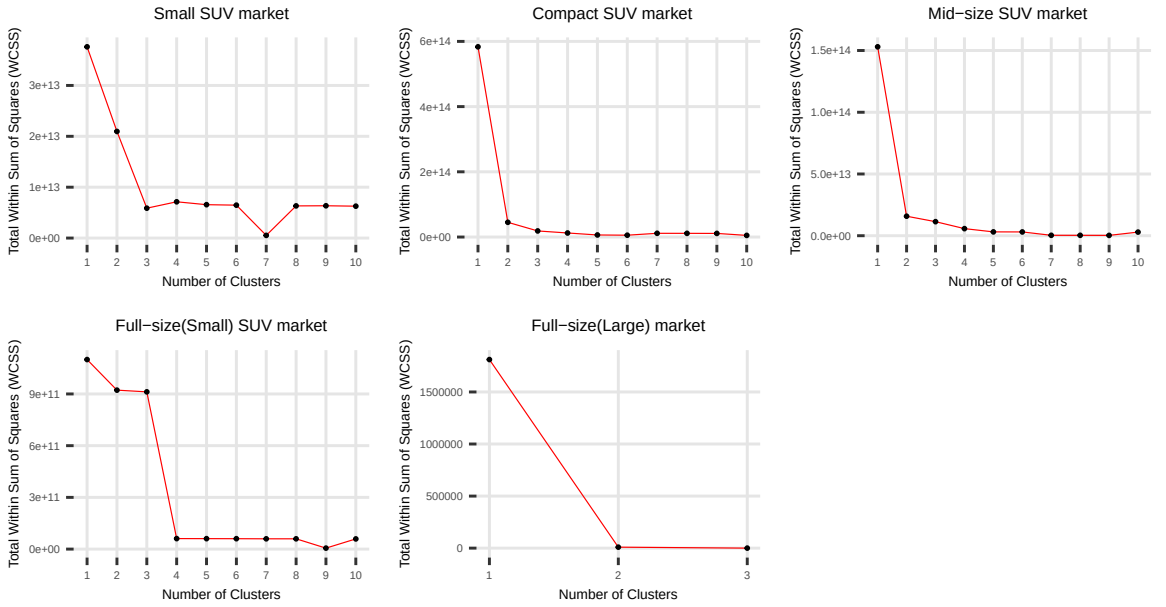

Figure J.2: Results of the Elbow method used in sales clustering for different SUV market segments.

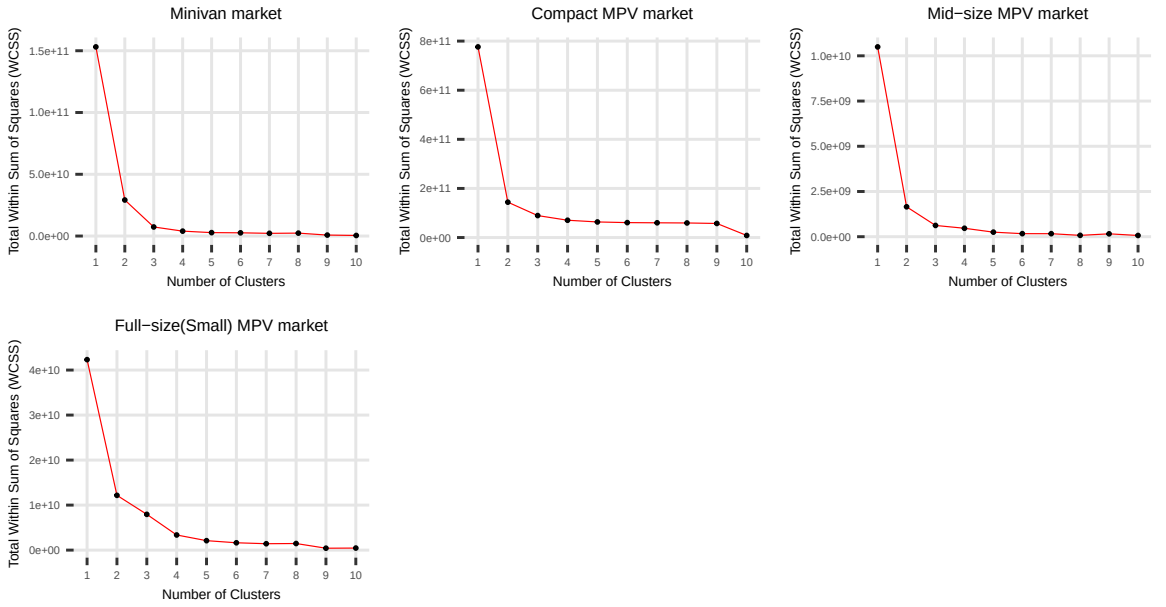

Figure J.3: Results of the Elbow method used in sales clustering for different MPV market segments.