Forecasting China's Car Sales: Integrating Multivariate Analysis with Weibo and Autohome Data

Shilan Chen (695495)

Erasmus University Rotterdam

Course Number: Data Science and Marketing Analytics

Dr. O. Karabag

July 28, 2024

Abstract

Accurate sales forecasting can help car manufacturers develop more appropriate market strategies and implement better supply chain management. In the big data era, social media opinions, word-of-mouth data, and other information are increasingly useful in sales forecasting. This paper proposes a multivariate model including social media sentiment and engagement data, user reviews, search trends, and macroeconomic factors to forecast car sales for 14 car models in the Chinese market. In the research, three machine learning models were applied, and their performance was compared on different datasets. The study found that incorporating social media and user review data improved the performance of China's car sales forecasting models. Notably, user review data from Autohome significantly enhanced the overall prediction accuracy of the models, with an average improvement of 54.42%. Although the improvement was modest, social media data from Weibo still managed to enhance the model prediction performance by 10.58% on average. Additionally, the study found that XGBoost outperformed Random Forest and SVR in terms of prediction accuracy within the multivariate model using user-generated content data.

Table of Content

Lists of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation

The automotive industry is one of the largest and most influential industries in the world, playing a crucial role in the global economy. The importance of the Chinese automotive industry in the global automotive market has also been increasing. According to Wikipedia, the automotive industry in mainland China has been the largest in the world both in terms of sales and ownership since 2008.

The development of the automotive industry affects economic growth, employment, and technological innovation. The sales and ownership of automobiles also influence various aspects of people's lives, such as the trend for travel and tourism, roads, and patterns of housing (Abu-Eisheh & Mannering, 2002). Conversely, the decision to purchase an automobile is important for consumers and is influenced by various economic and social factors, and automobile ownership affects both developing and developed countries (Abu-Eisheh & Mannering, 2002).

Forecasting automobile sales and demand is crucial for automobile manufacturers. Sales forecasting is the foundation for all business activity planning. From a market perspective, sales forecasting can help marketing personnel better assess the competitive situation and future market share, and then formulate corresponding marketing strategies. From the perspective of supply chain and inventory management, sales forecasting can help companies plan for the procurement of components and develop inventory plans in advance, thereby improving cost efficiency.

In the field of automobile sales and demand forecasting, the primary methods employed are statistical methods such as ARIMA (Autoregressive Integrated Moving Average) and machine learning techniques. These methods utilize various data types and have distinct advantages in terms of prediction accuracy and robustness. Time series analysis, like ARIMA, and regression models are widely applied in automobile sales forecasting, either on historical series of sales or on economic and demographic variables. For instance, Shahabuddin (2009) employed multiple regression models to forecast US automobile sales, incorporating durable industrial demand, personal consumption, population, discount rate, and GDP. Gao et al. (2018) utilized a Vector Error Correction Model (VECM) to analyze the connections between Chinese automobile sales and economic variables such as the Consumer Confidence Index (CCI), steel production, Consumer Price Index (CPI), and gasoline prices.

With the advent of big data, pre- and post-purchase consumer data is available in huge amounts. One of the typical big data is real-time search trends. The search engine calculates absolute search volume for a keyword (e.g., Baidu index) and relative search volume (e.g., Google Trends), which are commonly used in the latest forecasting research (Tang et al.,2022).

Integrating search trends provides real-time insights into consumer behavior and preferences, enhancing the model's ability to predict sales trends accurately. Fantazzini and Toktamysova (2015) used Google search data combined with economic indicators in models and showed that the predictive power increased when real-time search trends for automobile sales in Germany were included. Similarly, Wachter et al. (2019) showed that pre-purchase online search data is extremely useful in understanding consumer behavior for the generation of accurate sales forecast generation.

Another type of big data is user-generated content (UGC). UGC involves any kind of digital content created by consumers and then shared either online or offline. It can manifest in the forms of text, audio, video, or images and appear on various platforms, such as social media, blogs, and review sites (Kim & Johnson, 2016; Timoshenko & Hauser, 2019). UGC usually provides explicit and elaborate accounts of consumer experiences, opinions, and assessments of products or services. It gives ideas about what customers need, their preferences, and their levels of satisfaction through reviews, ratings, comments, and other multimedia elements (Pan & Zhang, 2011; Sun, 2012).

Studies have shown UGC data has a significant impact on sales (Bahtar & Muda, 2016, Kim & Johnson, 2016; Timoshenko & Hauser, 2019). People are increasingly influenced by social media recommendations and opinions on social media. It is the human quality of imitation that makes people susceptible to social groups; thus, information from actual customers is more important than the supplier's source. (Huang & Chen, 2006). For example, Jabr and Zheng found that product ratings are a strong sales driver in competitive markets (Jabr et al.,2014). Moe and Trusov noted that ratings not only directly influence sales but also affect future ratings, which in turn impacts future sales (Moe et al., 2011).

Researchers and marketers are increasingly using big data to predict sales trends. Sentiment analysis, and engagement metrics (likes, comments, shares) are used to gauge consumer interest and predict sales trends. Kolchyna (2017) specifically pointed out that sources of data used were from Twitter, Facebook, and Google Trends, indicating that the

combination of data from these sources yielded a high degree of accuracy in next-day sales forecasts for many brands by virtue of giving timely information on changes in consumer demand.

Despite these developments, most existing research is focused solely on economic indicators, search trends, online reviews, or social media data and talks generally about a few popular automobile brands. This limitation underlines the demand for comprehensive models incorporating diverse sources of data that could improve accuracy and comprehensiveness. Therefore, the motivation behind this thesis is to leverage different sources of UGC big data to understand customer behavior and intentions, with a focus on evaluating the predictive power of UGC data in the automotive sector. Specifically, among all kinds of online big data that have been used to predict sales, this study considers customer online review rating data and social media textual data, to which individuals may have been exposed along their life journey.

1.2 Research Questions

The objective of this research is to empirically investigate the value of UGC big data in the construction of sales prediction models for the automotive industry. Specifically, it seeks to determine how well UGC big data can help increase the accuracy of prediction models.

In light of this, this thesis will be guided by the following primary research question: *To what extent does UGC data increase the accuracy of sales forecasting models within the context of the Chinese automotive industry?* To answer this question, the thesis will restrict its attention to the Chinese auto market and develop a benchmark model. Drawing on previous studies, the benchmark predictors will incorporate macroeconomic indicators such as the Consumer Confidence Index (CCI), gasoline prices, and search trends. Search trends will be analyzed using car model names as keywords. The comparison model will incrementally include UGC data as features to construct a prediction model, followed by a comparative analysis of the results. According to Statista's latest data, Weibo is one of the social media platforms in China with the highest user activity; it provides user’s comments, and engagements on a specific topic. Autohome is a primary social networking site for car reviews, offering rating data based on overall model scores and specific feature ratings; it also provides information such as the purchase price of the car, the purchase location of the car from users, etc. Therefore, the comparison model will systematically incorporate

Autohome customer data, including ratings and purchase price, as well as Weibo data, including sentiment and engagement metrics, into the benchmark model.

A relevant sub-question for this thesis regarding the model comparison is: *For the Chinese automotive industry, which has a greater impact on car sales: car review data or social media data?* To address this, three models will be developed and compared with the baseline model. Each of these models will integrate data from Autohome, Weibo, or both, thus making it possible to analyze their relative impacts on car sales. This approach will help distinguish which platform's data—Autohome's car reviews or Weibo's data—has a greater influence on predicting car sales.

The contributions of this study will involve:

1. **Academic Contributions:** This research will establish whether there is an improvement in prediction accuracy by applying a multivariate car sales prediction model that will incorporate data from UGC sources. It shall, therefore, add to the existing body of knowledge on using UGC for predictive analytics. The influence of the different data sources on car sales will be checked, hence giving insight into how various types of UGC data can be integrated into sales forecasting models.
2. **Practical Contributions:** Targeting the Chinese market, this study will determine if user data influences car sales by 14 brands across major marketing channels applied by automobile manufacturers like Autohome and Weibo. This could be very useful information for marketers within the Chinese automobile market.
3. **Social Contributions:** Results will greatly benefit the marketing experts in making proper and workable marketing strategies and budget allocations. Knowing the influence of user-generated content on car sales will make it easier for more focused and productive marketing efforts.

The findings will significantly benefit marketing professionals in formulating effective marketing strategies and budget allocation plans.

Chapter 2: Literature Review

The literature review identifies how sales forecasting has developed over the years—from the more traditional, statistically-based techniques of moving averages and ARIMA models to more advanced machine learning methods in recent times, including Random Forests and SVM. These modern techniques address the limitations of earlier methods in handling non-linear patterns and incorporating complex data types, including customer behavior and economic data. The review then focuses on the enhancing role of UGC from social media and review sites in contributing to an increase in the accuracy of forecasting with the help of sentiment analysis. The specificity of the Chinese automotive market is examined, considering unique factors like government policies, environmental concerns, and social attributes. Finally, the review identifies challenges and research gaps in integrating UGC data, proposing advanced techniques to address these issues and improve forecasting accuracy.

2.1 The Evolution of Sales Forecasting

The foundation of inventory and sales forecasting was built on early statistical methods, which primarily used historical sales data. Simple moving averages and exponential smoothing were among the first techniques employed to analyze and predict sales trends. ARIMA models, introduced by Box and Jenkins, became widely adopted due to their robust statistical foundations and ability to model linear trends and seasonality (Box & Jenkins, 1970). However, early statistical methods often assume linear relationships and struggle to capture non-linear patterns, which are common in dynamic markets like retail and automotive sales (Alon et al., 2001). Also, the methods become sensitive to volatility and seasonality and require a long history of data to generate relatively accurate forecasts. (Box & Jenkins, 1970; Makridakis et al., 1998). Besides, external factors relating to economic conditions and consumer behavior are hard to be included (Alon et al., 2001; Tang et al., 2022).

Realizing the inadequacy of these traditional statistical methods, especially on non-linear and multisource data, machine learning techniques were introduced, such as Random Forests and support vector machines (SVMs), which could use even more complex data types, including customer behavior and economic data. Liu et al. (2013) observe that AI methods, especially neural networks, are much more accurate because they learn from data and model complex relationships within them.

Hybrid models were developed to combine traditional statistical techniques with modern AI methods to utilize their strengths. For example, time series analysis fused with such machine learning techniques as SVMs and neural networks enhanced the predictive performance very significantly (Alon et al., 2001).

Data types used in these models expanded into customer behavior data, demographic information, and real-time data from search engines and social media (Liu et al., 2013; Tang et al., 2022). Social media platforms generate a large volume of data that mirrors consumer sentiment, preferences, and upsurge trends, thus turning out to be very useful resources for the improvement of forecast accuracy (Liu et al., 2013; Zhu et al., 2019). However, few studies have been conducted on integration across different UGC platforms, especially in the Chinese automobile market.

2.2 Use of UGC Data in Forecasting Models

2.2.1 The impact of social media data on forecasting

Social media-related UGC has a significant impact on consumer behavior and sales performance because such sources provide credible, detailed information at the core of decision-making. For example, positive brand-related UGC posted on Facebook enhances consumer engagement, world-of-mouth behavior, and potential brand sales by provoking brand-related emotional and cognitive responses (Kim & Johnson, 2016). In comparison with marketer-generated content, the influence of UGC is usually found to be stronger (Goh et al., 2013).

Social media data is widely used in sales forecasting, with researchers incorporating sentiment (e.g., sentiment score) and engagement metrics (e.g., likes, shares, mentions, comments) into forecasting models. Sentiment analysis, typically conducted using Natural Language Processing (NLP) methods, is employed to analyze text data to capture consumer sentiments and needs, as noted by Malakooti (2013). Sentiment analysis in forecasting models has proved promising. Mohan et al. (2022) and Liapis et al. (2021) extended the application of sentiment analysis to other time series prediction areas, highlighting its role in improving the accuracy of forecasts in various contexts. Amin et al. (2024) demonstrated the importance of sentiment data from Twitter in short-term stock market prediction.

Studies have shown that incorporating both sentiment and engagement metrics significantly enhances forecasting accuracy. Kolchyna (2017) did a large analysis of 75

brands from diverse sectors and measured the impact of social media (Twitter, Facebook) on sales. The study found that Facebook engagement metrics improved sales forecast for 53% of the brand, and Twitter features, including engagement metrics and sentiment metrics, had similar effects (Kolchyna,2017, p. 192). Cui et al. (2017) study the enhancement of operational decisions in both supply chain and operations management using social media data. It includes interaction data from sites like Facebook and is analyzed through advanced natural language processing techniques to obtain sentiments and engagement. Findings reveal that including social media data in sales forecasts vastly enhances their accuracy of predictions from about 13% to over 23% compared to conventional ways. In another example, Verdonck (2019) analyzed Instagram data, including the number of likes, comments, posts, and caption sentiment, to predict car sales. The results showed that Instagram engagement metrics were significantly correlated with car sales, demonstrating that social media UGC is a strong predictor of consumer behavior and sales performance in the automobile industry.

2.2.2 The impact of product review data on forecasting

Another critical type of UGC in relation to consumer behavior and sales is product reviews and ratings. According to Pan and Zhang, detailed reviews and high ratings help to give consumers more knowledge about a product's quality and performance, which may decrease uncertainty and perceived risk of purchasing decisions (Pan & Zhang, 2011). Product reviews also influence purchase decisions with an enhanced level of trust and perceived usefulness, which increases purchase intent. Liu, Lee, and Srinivasan developed models in their 2019 research to test how the content of consumer reviews and ratings impacted sales conversion with deep learning techniques. In their work, dimensions regarding the quality of goods and value for price are withdrawn from over 500,000 reviews, which were rated against more than 600 product categories. Their results indicated that reviews have a great influence when the average rating of a product is high and the variance of ratings is low (Liu et al., 2019).

For the Chinese automobile industry, Wang et al. (2022) investigated how online reviews influence the offline sales of high-involvement products. Specifically, they utilized monthly sales data for 34 car models from January 2016 to January 2020, along with online review data from Autohome. The finding was that the valence of online reviews has an inverted U-shaped relationship with car sales, where the initial positive reviews increased the

sales, while further increases in the positive reviews reduced the sale of cars. Moreover, it was found that the volume of reviews had a positive significant influence on the sales of a car.

2.3 The Specificity of the Chinese Auto Market

The Chinese automobile market has some peculiarities, driven by rapid economic growth, urbanization, and government policies. The industry has dramatically shifted towards New Energy Vehicles (NEVs) in the past few years, due to environmentally induced concerns and policy incentives. From 2014 to 2020, NEVs gained significant market share, which turned China into the biggest consumer of NEVs in the world. Notably, the Chinese automobile market is also characterized by significant local protectionism, where joint ventures (JVs) and state-owned enterprises (SOEs) enjoy higher market shares in their headquarters provinces compared to the national level (Barwick et al., 2021).

Several studies have shown that a number of factors have significant effects on Chinese consumers' purchase intentions for automobiles. Government policies, including subsidies, tax incentives, and other preferential regulations, become major factors contributing to the formation of consumer demand in the context (Chen et al., 2020; Wang et al., 2021; Barwick et al., 2021). Economic factors such as vehicle price (Wang et al., 2020) and oil price (Wang et al., 2022) are major considerations, with many consumers being price sensitive. Improvement in technological innovation and infrastructure development has also tackled some of the concerns of consumers, such as the improvement in battery range and expansion of charging networks (Li, 2020). Chinese consumers show distinct preferences based on brands and their country of origin (Wu et al., 2019); their purchase intention is also influenced by social norms (Wang et al., 2021) and vehicle features (Wu et al., 2023). The factors influencing the purchase intentions of Chinese consumers regarding automobiles can be summarized as a combination of the following elements.

Government Policies: Financial incentives, such as subsidies and tax exemptions, are crucial in encouraging consumers to purchase cars. In addition, non-monetary policy incentives, which include unrestricted driving, and preferential treatment for local brands, have been established as impactful drivers of consumer decisions. (Wang et al., 2021; Barwick et al., 2021).

Environmental Concerns: Environmental concerns and awareness by consumers of pollution lead to the uptake of NEVs. The desire to save the environment favorably influences the intention to purchase (Wu et al., 2023).

Social Attributes: Social norms and face consciousness - the desire to gain social approval - are of importance. In other words, consumers would be motivated by family opinions, friends, and peers that may either support or tell them against the adoption of a car model. (Wang et al., 2021).

Vehicle Attributes: Performance, safety, and technological features—such as driving range and space—are key considerations in this area (Wu et al., 2023).

Economic Factors: The cost of the car, which includes the purchase price and operating costs, is a major consideration. Most consumers consider the best balance between cost and performance, with many of them being price-sensitive concerning the operating costs (Wang et al., 2020).

Recent research on automobile sales forecasting in China has adopted diversified data resources and methods of forecasting. Li et al. (2019) applied text-mining technology to analyze online search data from the Baidu Index and Sohu's automobile sales data, which disclosed a long-term equilibrium between online search data and automobile sales, where a regression model explained 76% of the variance. The work from Dai et al. 2023 combined univariate and multivariate models by integrating the Prophet model with a BP neural network for the prediction of sales with respect to fuel vehicles, electric vehicles, and plug-in hybrid vehicles using economic, social, and technical factors. Ding et al. (2023) developed a multivariate forecasting model, which combines the backpropagation neural network, recurrent neural network, and long short-term memory models by means of online reviews, sentiment analysis, and historical sales data.

2.4 The limitations of incorporating UGC data and the research gap

In the proposal framework, the collected UGC data includes user-generated textual content and its engagement metrics (number of likes, shares, comments) from the social media platform Weibo, as well as average ratings, feature ratings, and engagement metrics (views, volume of ratings) from the car forum Autohome.

Several potential biases and limitations arise when using textual data from social media and product reviews to forecast Chinese car sales.

Firstly, integrating UGC insights into real-time decision-making processes requires sophisticated text mining and sentiment analysis techniques, which may not always capture the full nuance of consumer opinions (Li et al., 2022). This study aims to address sentiment analysis challenges by applying advanced techniques like ERINE 4.0, released in December 2023. ERNIE, a large language model developed by Baidu, has shown promising results. Experimental outcomes indicate that ERNIE outperforms baseline methods like BERT in various Chinese textual datasets, achieving higher accuracy (Sun et al., 2019; Sun et al., 2021).

Secondly, in terms of review data, the popularity effect can skew the perceived helpfulness of reviews, with more popular reviewers providing more objective but potentially more negative ratings (Goes et al., 2014). To address this, this study will incorporate the engagement metrics of ratings (e.g., the likes and views of the ratings) into the predictors to balance ratings with their popularity.

Thirdly, many studies focus on a limited combination of data sources, such as historical sales (Ding, 2023), economic indicators (Gao et al., 2018; Dai et al., 2023), Baidu search trends (Li et al., 2019; Zhang et al., 2017), social media data from Weibo (Yang et al., 2020; Zhang et al., 2022), or online reviews (Wang et al., 2023). Often, these studies do not explore the potential of integrating diverse data sources, which may limit the scope and accuracy of the predictions. This study will incorporate economic indicators, search trends, social media metrics, and vehicle attribute ratings into a predictive model to capture the impact of the economic situation, user behavior, and different characteristics of car sales comprehensively.

Additionally, this study will use three machine learning models—random forest, XGBoost, and SVR—to evaluate the performance of different models on this type of data. By integrating UGC from these platforms, the study seeks to provide a more comprehensive and culturally relevant understanding of consumer behavior in the Chinese automotive market.

Chapter 3: Data

3.1Framework

A comprehensive framework was designed to extract maximum benefits from user-generated content (UGC) and other data sources to enhance China's car sales forecasting. The framework includes several key components: data collection, data processing, sentiment extraction, and the building of forecasting models, as illustrated in Figure 1.

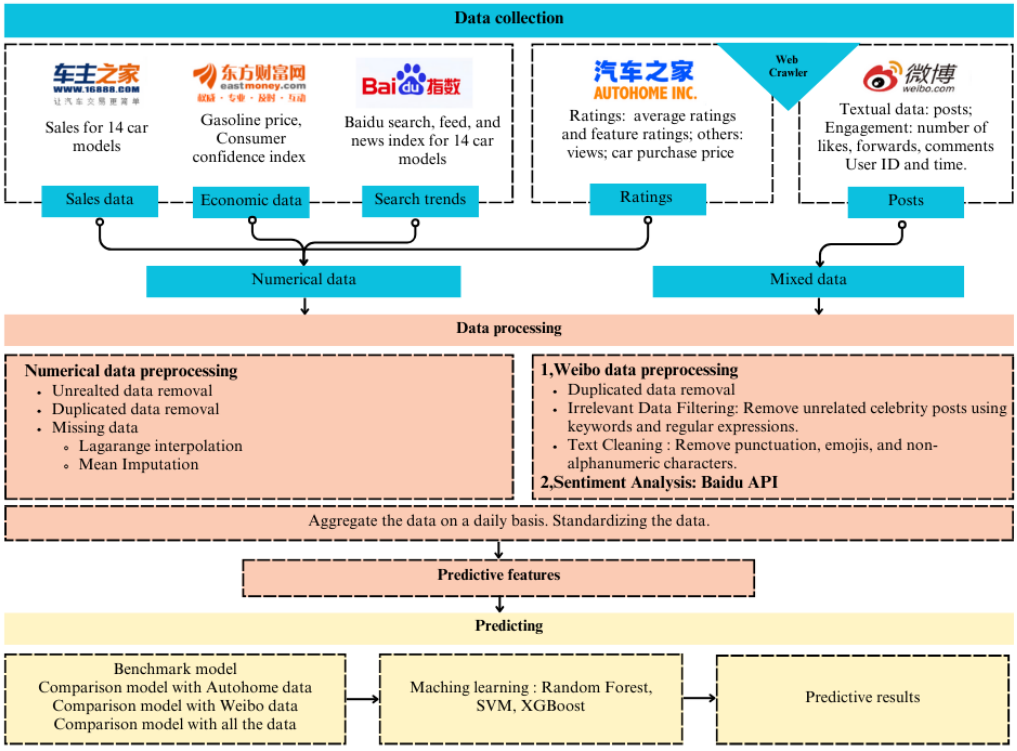


Figure 1. Framework for incorporating user-generated content (UGC) data.

To select car models that best represent the Chinese automotive market and to reflect the dynamic changes in sales of electric and fuel vehicles, 14 best-selling models from various brands were chosen based on the 2023 sales rankings by the China Association of Automobile Manufacturers. These brands include foreign, joint-venture, and domestic brands, covering both fuel and electric vehicles. The research period was from 1st January 2021 to 31st December 2023 since some electric vehicles were launched in 2020 and 2021.

In the data collection phase, sales data, gasoline price, and Consumer Confidence Index (CCI) were all downloaded from various websites. In addition, Baidu Index data were collected through searching by keywords, whereas UGC data in both Autohome and Weibo were obtained using web scraping tools. To maintain consistency in data collection, all keywords used were the names of car models.

In data processing, various methods were employed to convert all data into daily data to maintain uniformity in the scale. In addition, during this phase, sentiment analysis results for Weibo text data were obtained using the Baidu ERINA API.

During the model-building phase, predictors like sales data, gas prices, and CCI were used for the baseline dataset. Three comparison datasets were then constructed and trained on random forest, XGBoost, and SVR algorithms. The performance of these models was evaluated using metrics such as RMSE, R^2, and MAE.

3.2 Data Collection

To accurately represent the specificity of the Chinese automotive market, I selected a diverse range of best-selling models of 2023 from mainland China. This selection includes four joint-venture brand models: Dongfeng Nissan, FAW Toyota, Guangqi Honda, and SAIC Volkswagen. Popular electric vehicles such as the Model Y, NIO ES6, Li ONE, and XiaoPeng P7 were also chosen. Additionally, Hybrid models that have shown strong performance in recent years, such as the BYD Song Plus and AION Y, were also included. Conventional fuel vehicles with significant market presence in China, such as the Corolla, CR-V, Langyi (Lavida), and Nissan Xuanyi (Sylphy), were also considered. For a comprehensive analysis across different price ranges, high-end models like the BMW 5 Series and Mercedes-Benz GLC were incorporated into the study. The car models included in this study are listed in Table 1.

Table 1. Car models and brands list

Model Name	Car brand	Model Name	Car Brand
Model Y	Tesla	BMW 5 Series	BMW
AION Y	GAC	BYD Song Plus	BYD
NIO ES6	NIO	Corolla	FAW Toyota
Xiaopeng P7	XPeng	CR-V	Guangqi Honda
Audi A6	Audi	Hafu H6(Haval H6)	Haval
Benz GLC	Mercedes-Benz	Langyi	SAIC Volkswagen
Xuanyi	Dongfeng Nissan		

To maintain consistency in data collection, all keywords used on Baidu, Weibo, and Autohome when querying were the names of car models listed in Table 1. For example, By querying "model Y" in the Baidu index platform and selecting the time range, the Baidu index of Model Y in this period was obtained.

Due to the late launch of some models—for instance, AION Y, which went on sale formally in March 2021—and in order to acquire data for all models as much as possible, all data collected for this study was from January 1st, 2021, to December 31st, 2023.

3.2.1 Historical Sales

The data was collected from the car sale platform (https://www.16888.com/) and compared with the data of the China Automobile Dealers Association to ensure the accuracy of the sales data. The sales data were recorded by month and in units of vehicles, as shown in Table 2.

Table 2. Summary statistics of monthly sales

	Monthly Sales (units)			
Car Model	**Mean***	**Std**	**Min**	**Max**
AIONY	12110	7834	2203	27132
Audi A6	13127	5224	2159	29492
BMW 5 series	13023	3434	8331	20579
BYD Song	17014	6986	5253	33712
Benz GLC	11072	4066	3400	18211
CR-V	17014	6986	5253	33712
Corolla	21247	8234	9397	38370
Hafu H6	24226	7848	14267	46368
Langyi	31160	9410	8442	55268
Li ONE	6241	4438	462	14087
Model Y	31714	18105	960	69098
NIO ES6	3897	2303	174	11118
XiaoPeng P7	4622	1920	1022	9183
XuanYi	36445	9392	22855	61170

*Note: The mean is the average value of the specific data, std stands for standard deviation, and it measures the dispersion of data from the mean. Min is the smallest value, and max is the largest. 25% (25th percentile) indicates that 25% of the data points are below this value, and 75% (75th percentile) indicates that 75% of the data points are below this value. All abbreviations used in the tables follow these explanations.

3.2.2 Macroeconomic Indicators

The Consumer Confidence Index (CCI), also referred to as the consumer sentiment index, serves as an indicator of consumer confidence strength. It is based on survey responses regarding households' expected financial situation, their sentiment about the general economic situation, unemployment, and their capability to save (OECD, 2023). The CCI

reflects consumers' expectations about the future economic situation. When confidence is high, people are more likely to make large purchases like cars, as they feel more secure about their financial future (Pavithra & Velmurugan, 2023)

For this study, the monthly consumer confidence index data was sourced from the Oriental Fortune Network data center at http://data.eastmoney.com/cjsj/xfzxx.html. The index value ranges between 0 and 200, with 100 being the critical point indicating the strength of consumer confidence. An index value above 100 indicates that consumer confidence is in the strong confidence zone. Conversely, an index value below 100 indicates that consumer confidence is in the weak confidence zone (National Bureau of Statistics of China, 2023). The summary statistics of CCI in the study time frame are shown below.

Table 3. Summary statistics of CCI (2021-2023)

Mean	Std	Min	25%	75%	Max
101.65	16.48	85.50	87.08	120.28	127

Gasoline prices significantly influence the overall cost for fuel car owners and impact consumers' decisions when purchasing cars. An increase in the price of gasoline increases the cost of running fuel vehicles; as indicated by Klier & Linn, 2010, consumers will delay purchasing new fuel vehicles or buy more fuel-efficient vehicles. High gasoline prices improve the relative advantage of hybrid and electric vehicles—cutting the cost per mile of driving such cars to far below that of a gasoline-powered car. Data for the average monthly prices of gasoline at yuan per ton were retrieved from the Oriental Wealth Network at http://data.eastmoney.com/cjsj/oil_default.html , reflecting price adjustments over time.

Table 4. Summary statistics of Gasoline price (2021-2023), yuan per ton.

Mean	Std	Min	25%	75%	Max
8818.06	937.64	6645	8120	9575	10850

In this study, gasoline prices were used as a surrogate measure of the effect of energy costs on car demand, while the consumer confidence index indicated the general economic climate and consumers' willingness to buy cars. The summary statistics are shown in Table 4.

3.2.3 Search Trends

As outlined in CTR's "2023 China Search Engine Industry Research Report", as of 2021, 829 million individuals in China utilized search engines, representing 80.3% of Internet

users. Baidu holds the largest market share among traditional search engines in China. This paper utilizes the Baidu index for each car model as the search data, extracting the average monthly PC and mobile comprehensive index data by querying car model names as keywords.

The Baidu index data includes daily metrics such as user search frequency (search), browser SEO advertising exposure (feed), and official channel news exposure (news) for each car model. Baidu is China's predominant search platform, and these metrics collectively represent the online popularity and marketing intensity of each car model on the internet. Table 5 displays the average values and standard deviation of search, feed, and news metrics for each car model from January 1, 2021, to December 31, 2023.

Table 5. Summary statistics of the Baidu index

	Feed		News		Search	
Car Model	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
AION Y	1954.11	3648.65	0.00	0.00	858.04	2650.59
Model Y	95584.84	175612.53	9.82	36.07	3846.80	3468.15
Corolla	195994.93	145162.63	7.03	18.12	8040.69	5106.84
Hafu H6	208099.21	122741.05	9.32	23.90	25495.83	26934.96
Benz GLC	159149.38	123216.80	0.16	0.83	2633.24	633.00
Audi A6	322347.07	344833.13	0.29	1.47	7315.64	2526.49
BYD Song	52590.70	91588.49	1.10	6.87	1050.55	201.49
BMW 5 series	263387.13	158060.76	4.36	10.31	5721.14	2754.12
XiaoPeng P7	34094.22	165442.98	0.15	1.26	4087.20	4037.23
Langyi	192250.24	93837.96	0.00	0.00	3824.62	969.50
CR-V	46104.47	70363.78	0.00	0.00	3794.74	1561.24
Li ONE	83455.92	239847.92	0.01	0.14	14923.89	10772.95
NIO ES6	15049.40	28868.21	0.58	3.55	1697.03	4461.47
XuanYi	253308.73	235801.96	0.32	1.54	4539.14	1780.34

3.2.4 Car Model Ratings

In this research, review data from Autohome was collected by web crawler tools. Autohome is one of the largest auto forums in China. Users can post a self-written comment and rating on the forum about the car purchased. Since different users purchase from different car dealers located at different places, the purchase price may be different even for the same car model. Therefore, users will fill in their car purchase price, location, and other information. A typical review from Autohome is shown in Figure 2:

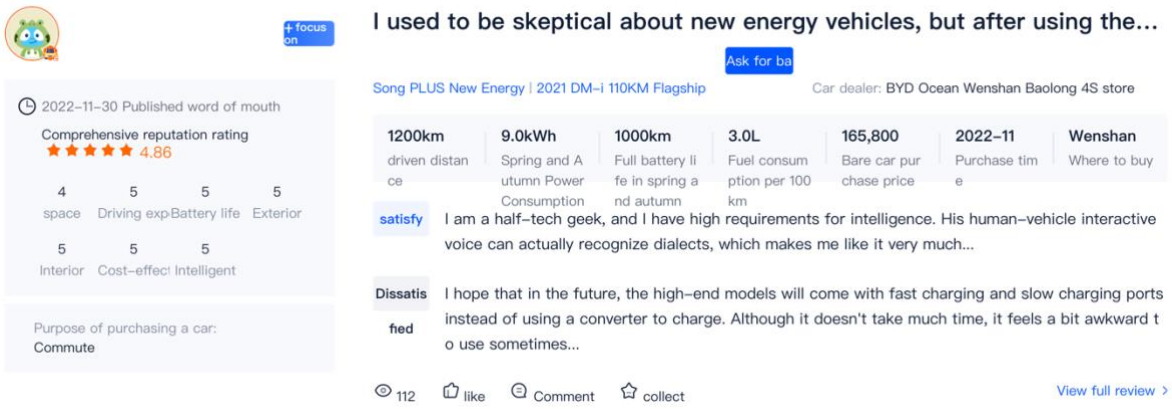


Figure 2. Screenshot of a typical review from Autohome on BYD Song Plus, translated by Google.

The ratings encompass the average score of the vehicle model and specific characteristics such as space, endurance mileage, appearance, interior, and cost performance, as shown in Table 6.

Table 6. A review of Model Y on December 21st, 2023, translated by DeepL.

User ID	Car owner from Chongqing XXXX
Release Date	2023-12-21
Average Score	4.71
Views	223785
Comments	32
Likes	29
Driving Experience	5
Space	5
Endurance	5
Appearance	5
Interior	4
Cost performance	5
Integellence	4
Milledge	2625km
Fuel Consumption per 100 km	N/A
Purchase Price (RMB)	266,400
Purchase Date	2023-11
Purchase Location	Chongqing

Satisfaction	At the end of November, I happily got the new Model Y. The best part is, they added more features without increasing the price. Let me tell you what they added...
Dissatisfaction	The only thing I'm not satisfied with is that it doesn't have Gaode (AutoNavi) navigation...
Purchase Purpose	For commuting, shopping, and road trip.

Note: All the ratings are on a scale of 1 to 5.

3.2.5 Social Media Data:

Data from QuestMobile shows that by September 2023, the number of active Internet users in China has reached 1.224 billion. Among them, the top three applications in terms of monthly active users of social media content platforms are WeChat, Douyin, and Weibo, with monthly active users reaching 1.045 billion, 743 million, and 485 million, respectively. Among the three platforms, Weibo is widely used for news dissemination, public opinion sharing, and social networking, making it a valuable platform for gathering user-generated content and engagement metrics. This study only collected user opinions and their respective engagement metrics on Weibo for consideration of data collection feasibility.

Through the data crawler tool, all the Weibo data of 14 models by querying car model names was collected. The data includes user ID, publication time, Weibo content, and engagement metrics such as the number of likes, comments, and shares. The original Weibo data contains 795,164 rows. Table 7 presents a sample of Weibo data, as indicated below.

Table 7. A data sample mentioned the Audi A6 from Weibo, translated by DeepL.

User ID	Name	Release Date	Shares	Comments	Likes	Content
XXX	XXX	2023-06-06	10	125	161	Many netizens say BMW, Audi and Mercedes don't understand the Chinese market, that's a bit of a stretch .....

3.3 Data Processing

Since the various individual data sets have been collected according to different time scales, while the Baidu index data is recorded daily, all the other data sets should be aggregated daily so that their scales are compatible.

3.3.1 Car Sales Data Preprocessing

The original data set contains 490 rows, and each is related to the total sales for one month of a specific car model, measured in units. I predict daily sales for two main reasons: First, sales data that happen on a monthly basis are too sparse for experimental purposes; hence, using daily data will allow for a richer dataset for research. Second, social media metrics and search trends data are recorded on a daily or hourly basis; translating monthly sales into daily sales can provide more granular insights for a better understanding of the relations between sales and UGC data.

A common method to convert monthly sales data into daily sales is to divide the total monthly sales by the number of days in the respective month. While the simple average daily sales provide a basic estimation, polynomial interpolation is used to capture more complex patterns and trends within the data, especially for highly volatile sales data. Polynomial interpolation involves fitting a polynomial function to the data points, which can more accurately reflect non-linear trends and variations in time series data. This will guarantee that the interpolated values just pass through the given points to give a smooth, continuous expression for the trend of sales (Lepot M et al.,2017).

Basically, the polynomial interpolation takes 4 steps to transform a monthly sales dataset into a daily sales dataset.

1. **Daily Averages:** Calculated for the average daily sales of each car model by dividing its monthly sales figure by the days in each month, as shown in Table 8.
2. **Map to Dates:** Created a new dataset where these daily averages were assigned to specific dates within each month, typically choosing the 15th day to evenly distribute data points.

Table 8. Summary statistics of daily average sales

	Daily Average sales(units)			
Car Model	**Mean**	**Std**	**Min**	**Max**
AIONY	397	257	71	899
Audi A6	431	170	72	951
BMW 5 series	428	109	278	664
BYD Song	560	229	175	1087
Benz GLC	364	132	111	605
CR-V	560	229	175	1087

Corolla	699	270	313	1279
Hafu H6	796	254	474	1496
Langyi	1022	302	281	1783
Li ONE	205	145	15	454
Model Y	1042	594	32	2303
NIO ES6	128	75	6	359
XiaoPeng P7	152	62	33	296
XuanYi	1197	299	737	1973

3. **Fit Sales Patterns**: Used the sales data on the 15th day of each month as training data. Fitted a nonlinear function to model sales patterns over time for each car model. The polynomial function used in this study is defined by the following formula (Fox, 2015):

$$P(x_i) = a_n x_i^n + a_{n-1} x_i^{n-1} + \cdots + a_1 x_i + a_0$$

Where a_n, $a_{n-1} \ldots a_0$ are the coefficients of the polynomial function. And x_i represent the number of days since the start date of a specific month, n is the degree of the polynomial.

The objective function to be minimized is:

$$\text{minimize} \sum_{i=1}^{m} (P(x_i) - y_i)^2$$

Where $P(x_i)$ is the predicted sales on day *i*, *m is the total number of days in a specific month*, y_i is the average sales of that month.

4. **Optimize Model**: Experimented with different polynomial degrees, specifically ranging from 3 to 5, to find the best-performing model. Evaluated the models using mean absolute error (MAE) to select the one that most accurately predicts daily sales amounts.

The polynomial function was implemented in Python. Specifically, I employed the 'np.polyval' function to evaluate the polynomial and the 'curve_fit' function from the Scipy library to optimize the polynomial coefficients. From the range of polynomial degrees between 3 and 6, the 4-degree polynomial curve demonstrated the best performance. Figure 3 depicts the 4-degree polynomial fit applied to average daily sales data for Benz GLC, Hafu H6, and CR-V models. Additional polynomial fits for other vehicle models are presented in Appendix 1.

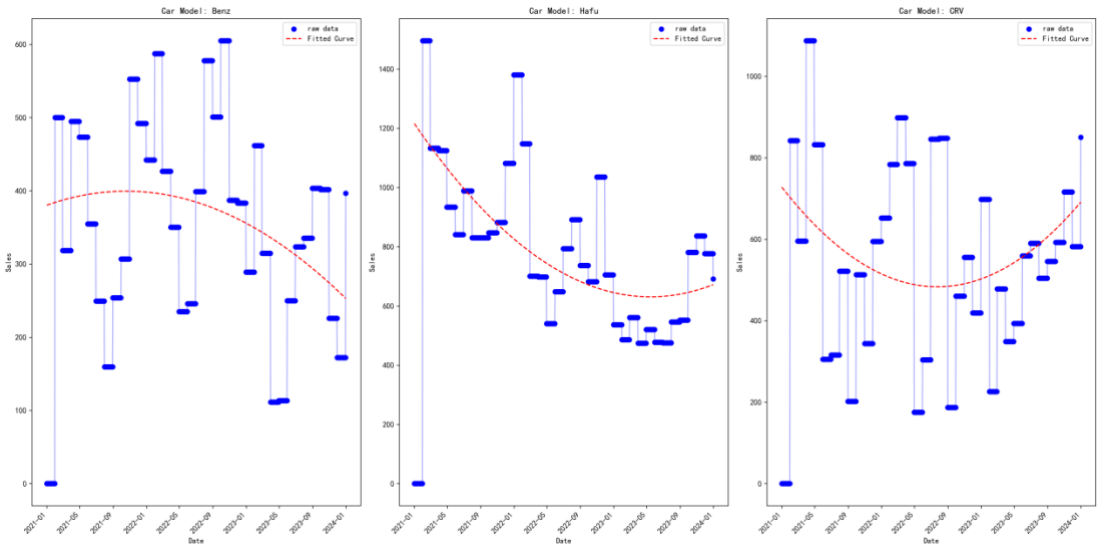


Figure 3. The 4-degree polynomial curve on average daily sales (from left to right: Benz GLC, Hafu H6, CR-V). Blue dots represent the original monthly sales, and the red curve represents the fitted polynomial curve.

3.3.2 Macroeconomic Data Preprocessing

The consumer confidence index (CCI) is only available monthly. To ensure that each day in the dataset reflects the consumer confidence level of its respective month, the monthly CCI values were simply copied to each day of that month. This means that every day within a particular month has the same CCI value.

The prices of gasoline were obtained monthly, interpolation was employed to complete the missing value. Linear interpolation is employed due to its straightforwardness and the stability of gasoline prices. From the trend of gasoline prices shown in Figure 4, it can be seen that this method ensured that the data of gasoline prices were complete with no breaks, which is important for the analysis of the effect of energy costs on car demand.

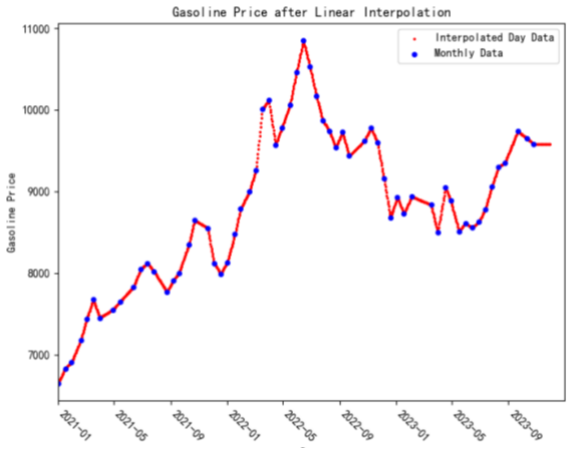


Figure 4. Interpolated daily gasoline price

3.3.3 Weibo Data Preprocessing

The data from Weibo has gone through several cleaning steps, adding to its relevance and usability:

Removal of Duplicates: The original Weibo data contains 795,164 entries. I first removed rows with complete duplicates of the fields ‘User ID’ and ‘content’, remaining only the first occurrence of the rows. Subsequently, I eliminated rows containing any NA values. These preprocessing steps resulted in a final dataset of 633,450 entries.

Text Cleaning: Punctuations, Emojis, and non-alphanumeric characters were removed. It was an essential step to make the data noise-free and to convert it into a form so that textual analysis can be done on this data (Singh, 2018, p. 2). Initially, a regular expression pattern was created to match special characters in Weibo data, and then special characters from the content of Weibo posts were identified and removed. This cleaning process enabled a more accurate and effective analysis of the textual data.

Irrelevant Data Filtering: Posts that are not relevant to car models were filtered out by the use of keywords and regular expressions. In order to get rid of spam data, first, I identified and extracted rows where different User IDs have the same Weibo content and formed a data set of 153,749 rows. Looking into these repetitive Weibo contents, it turned out that many of them were related not to cars but heavily to celebrities. For example, Users posted a large number of TV dramas of a certain celebrity repeatedly on Weibo but actually tagged it as #Saic Volkswagen Langyi brand ambassador XXX. By manually browsing through the high-frequency repeated Weibo content, 35 keywords that contained celebrity names and other irrelevant terms were filtered out. These keywords were then used in regular expressions to filter out high-frequency repetitive content that is unrelated to cars. This resulted in a final Weibo dataset of 550,058 entries.

3.3.4 Sentiment Analysis Using Baidu API

Under the broad field of Natural Language Processing, Bidirectional Encoder Representations from Transformers (BERT) have been much utilized within the last five years. This breakthrough in NLP was introduced by Jacob Devlin and his co-authors in 2018 and was developed by Google (Devlin et al.,2019). It is one of the major improvements made in NLP. The architecture of BERT is basically founded on the Transformer model that

treats words as relative to all other words in a sentence and does not process them sequentially.

The ERINE model, developed by Baidu in 2019, is an enhanced version of the BERT model. Baidu's ERNIE improves language representation through creative masking strategies for integrating external knowledge. For Chinese language datasets, Sun et al. (2019) have illustrated appreciable improvements in tasks such as sentiment analysis and public opinion monitoring over Bert. Considering the huge amount of Weibo data and the data itself with a lot of noise, this study chooses to apply Baidu ERINE 4.0 API for sentiment analysis to ensure the speed and accuracy of data processing.

After conducting sentiment analysis through the Baidu API, the sentiment tendency of Weibo content is categorized into three groups: 0 for negative, 1 for neutral, and 2 for positive sentiments. Additionally, the API provides:

Confidence: It indicates the reliability of the classification, ranging from 0 to 1.

Positive Probability: It denotes the likelihood that the text is positive, ranging from 0 to 1.

Negative Probability: It denotes the likelihood that the text is negative, ranging from 0 to 1.

The specific criteria for labeling sentiments of the API are as follows:

0 (Negative): When the Positive Probability is less than 0.45 and the Negative Probability is greater than 0.55, the text is labeled as negative.

1 (Neutral): When the Positive Probability is greater than or equal to 0.45 and the Negative Probability is less than or equal to 0.55, the text is labeled as neutral.

2 (Positive): When the Positive Probability is greater than 0.55 and the Negative Probability is less than 0.45, the text is labeled as positive.

Since Weibo data is collected in seconds, it was necessary to aggregate all data to a daily level. Subsequently, the cleaned data was aggregated on a daily basis. A new variable named ‘mentions’ was created to represent the number of mentions per day for each car model. Additionally, the average sentiment score of all posts in a single day was calculated and denoted as the sentiment mean. The variables included in the dataset are shown in Table 9.

Table 9. Integrated Weibo data of 14 car models on June 17, 2022

Car Model	Like Sum	Comment Sum	Share Sum	Negative Prob Mean	Positive Prob Mean	Confide-ce Mean	Sentiment Mean	Positive Counts	Neutral Counts	Negative Counts	Mentions
AIONY	65	82	3020	0,41	0,59	0,83	1,33	2	0	1	3
Audi A6	165	28	127	0,18	0,82	0,82	1,75	10	1	1	12
BMW 5 series	3205	227	75	0,33	0,67	0,83	1,32	63	2	32	97
BYD Song	23	27	8	0,19	0,81	0,81	1,67	7	1	1	9
Benz GLC	193	99	30	0,19	0,81	0,83	1,6	12	0	3	15
CR-V	182	29	6	0,42	0,58	0,78	1,13	9	0	7	16
Corolla	74	54	29	0,16	0,84	0,87	1,72	46	1	7	54
Hafu H6	88	30	40	0,19	0,81	0,81	1,74	16	1	2	19
Langyi	428	55	109	0,18	0,82	0,8	1,7	60	1	10	71
Li ONE	17339	5113	3491	0,15	0,85	0,84	1,74	54	0	8	62
Model Y	12267	4722	2497	0,67	0,33	0,61	0,51	86	45	292	423
NIO ES6	6254	1632	8264	0,86	0,14	0,84	0,19	16	6	178	200
XiaoPeng P7	10950	2361	1517	0,14	0,86	0,88	1,78	24	0	3	27
XuanYi	7219	3690	3655	0,36	0,64	0,83	1,33	10	0	5	15

3.3.5 Autohome Data Preprocessing

For the Autohome data, variables with more than 70% missing values, such as intelligence, Mileage, and Fuel Consumption per 100 km, were initially removed. Variables irrelevant to this study, including purchase location, purchase date, and purchase purpose, were also excluded. Since the Autohome dataset already included ratings for various car attributes, textual comments such as ‘Satisfaction’ and ‘Dissatisfaction’ were excluded as well.

During the study period, not every car model had review data for each day. Autohome displays reviews in chronological order, with the most recent comments listed first, and it’s not possible to change the review order. The recency of reviews is important when deciding which product to buy; a survey conducted by Sammy Paget in 2024 found that 47% of consumers rated the 'sort by newest' function from Google review as ‘highly useful’. Given Autohome’s display logic, it is logical to fill in missing rating scores using the most recent available data, as users tend to focus on the latest reviews rather than an average or median score from all reviews. Therefore, forward and backward filling methods were applied to the remaining data, using the most recent available data to fill in missing values for the intervening dates. The final Autohome dataset includes the following variables shown in Table 10.

Table 10. An Autohome data sample of CR-V and Hafu H6

Release Date	Car Model	Average Score	Views	Number of Reviews	Likes	Drive Experience
2021-01-05	CR-V	3,88	209218	40	141	3
2021-01-06	Hafu H6	4.625	17598	9	26	4.5
Release Date	**Car Model**	**Space**	**Appearance**	**Interior**	**Cost Performance**	**Purchase Price (yuan)**
2021-01-05	CR-V	5	4	4	4	192800
2021-01-06	Hafu H6	5	5	4	5	124900

3.3.6 Final datasets

After preprocessing, the various datasets were combined into a single dataset containing 15,330 entries and 27 columns. To examine the impact of different sources of user-generated content (UGC) data on predicting sales, four distinct datasets were prepared. The predictive features included in each dataset are as in Table 11 . These four datasets were

then used in predictive models to determine whether the inclusion of UGC data enhances sales predictive performance.

Table 11. Predictive features for different datasets

Dataset	Predictive features
Benchmark	Baidu index (Baidu Feed, Baidu Search, Baidu news), Gasoline price, CCI
Benchmark + Autohome	Benchmark features, Autohome features (Views, Number of Reviews, Likes, Average score, Drive Experience, Space, Appearance, Interior, Cost Performance, Purchase Price)
Benchmark + Weibo	Benchmark features, Weibo features (Mentions, Negative Counts, Neutral Counts, Positive Counts, Sentiment Mean, Confidence Mean, Positive Probability Mean, Negative Probability Mean, Shares, Likes, Comments)
Combined	Benchmark features, Autohome features, Weibo features

3.4 Preliminary Data Analysis

3.4.1 Baidu Index Trend

According to the dynamic data from Baidu search counts, there were apparent trends in customer attention and interest in different car models. Figure 5 depicts Baidu Search trends for three models. Additional trends for other vehicle models are presented in Appendix 2. Car models like CR-V, BMW 5 series, Langyi, XuanYi, and Li ONE have shown a declination trend in consumer searches. On the other hand, BYD Song is showing a growing pattern of interest among customers.

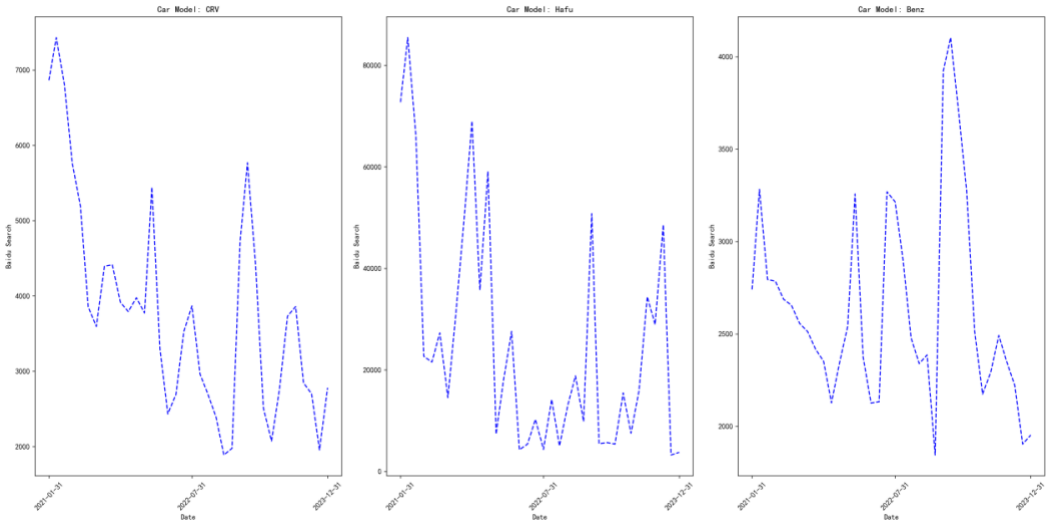


Figure 5. Baidu Search trends for CR-V, Hafu H6, and Benz GLC models (left to right) from Jan 2021 to December 2023.

For Benz GLC, wavering trends show up occasionally with spikes, probably due to the release of new models or promotional activities. The same goes for AION Y and NIO ES6, but less often. Archiving this information is quite important in understanding the popularity and market performance of these car models.

3.4.2 Gasoline Prices and Consumer Confidence Index

As shown in Figure 6, the line chart of the trend of gasoline prices and the Consumer Confidence Index (CCI) indicates the two variables moving in opposite directions. According to Su et al. (2023), consumer confidence can both positively and negatively influence oil prices. The positive effect suggests that higher consumer confidence can boost oil demand. However, the negative impact, influenced by economic policies such as economic downturns and dollar appreciation, can weaken this relationship. This is shown in Figure 6 where a low gasoline price is associated with increased consumer confidence, while a high gasoline price translates to decreased consumer confidence. Additionally, the low consumer confidence in 2023 tells more of the uncertain outlook for the economy and the effects of energy prices on consumer attitudes.

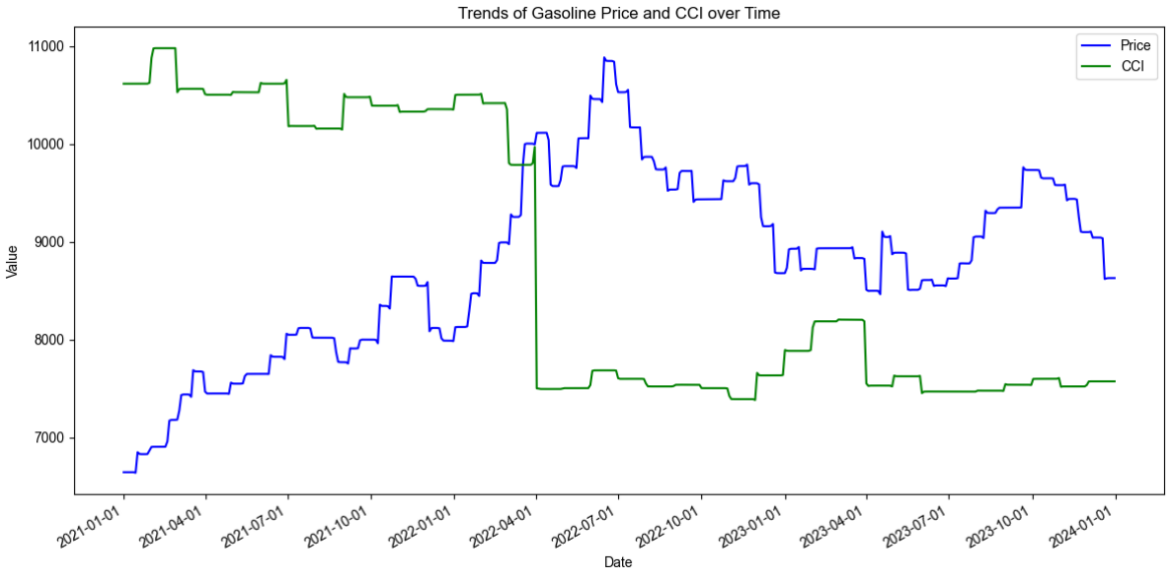


Figure 6. Line chart of gasoline prices and Consumer Confidence Index

3.4.3 Sentiment and Mentions for Various Car Brands on Weibo

The analysis of Weibo posts reveals significant trends in customer sentiments towards various car models. Overall, the sentiment is predominantly positive, with a substantial portion of posts exhibiting negative sentiments. As shown in Figure 7, for most car models, the values represented by the red bars (positive count) are significantly higher than those of

the negative count for the majority of the period from 2021 to 2023. This indicates a generally positive perception of the car models among the public. It is worth noting that the Weibo post sentiment of the BMW 5 series, Li One, and Model Y shows a trend of two-level differentiation; the number of negative posts is more prominent than that of other brands.

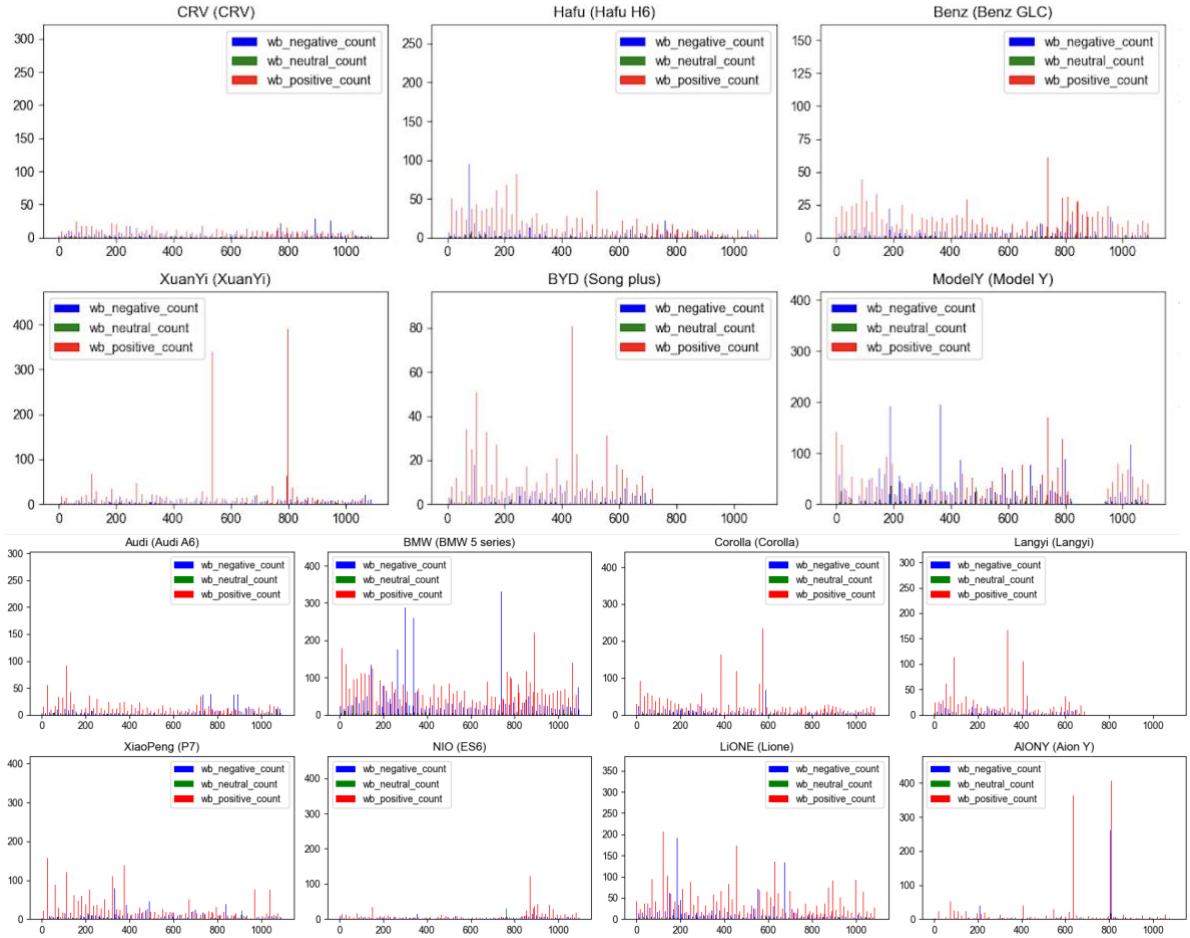


Figure 7. Sentiment distribution of different car Models on Weibo

The data also shows fluctuating levels of attention paid to different car brands. Models like the BMW 5 series and Model Y frequently appear on Weibo, indicating that they are highly considered by individuals. In contrast, models such as CR-V, Hafu H6, Benz GLC, Langyi, BYD Song Plus, and AION Y are mentioned less often. This trend highlights increased interest in well-known fuel car brands like BMW and new electric car models like Li ONE and Model Y. The volume of mentions for various car models can be found in Appendix 3.

Chapter 4: Methodology

4.1Model Selection

The most common automobile forecasting methodologies in use today are statistical models, AI models, and hybrid models. According to Tang's survey (2022), out of 5,463 articles that used social media data to build forecasting models, 60.94% utilized AI models. Major ones include neural networks, support vector machines, and Random Forest, while major causes can be attributed to their powerful ability to deal with nonstationary, nonlinearity, and complexity (Tang et al., 2022). Neural networks, particularly deep learning models, are not attractive in analyzing small and tabular data (Dong et al., 2022). while Boosted tress is highly effective for small to medium-sized datasets and can handle imbalanced data well (McElfresh et al., 2024)

The final dataset comprises 15,330 observations. Additionally, the data is highly imbalanced, with some car models like Model Y having high popularity on platforms such as Weibo and Baidu, while small brands are mentioned or searched far less frequently. Considering the data size and structure, this study chose Random Forest, Support Vector Regression (SVR), and XGBoost as forecasting models based on their suitability for small and imbalanced datasets.

4.1.1 Random Forest

Random Forest, developed by Breiman (2001), is an ensemble learning method conducting an ensemble of multiple decision trees toward the construction of a more robust and accurate model.

The decision tree is a supervised learning algorithm applicable to both classification and regression. It works by modeling decisions on given input data. It has a structure that resembles a tree form. Each node essentially symbolizes an internal decision based on one attribute. Each branch relates to an attribute value; a leaf node is the final decision or prediction. For the following processes, a decision tree will be constructed:

1. Initially, using metrics such as Mean Squared Error (MSE) as the metric for choosing the best attribute to split data into regression tasks. The formula of MSE can be expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i represents the actual values and $\hat{y}_i$ represents the predicted values.

2. The optimal attribute is selected based on the MSE; the dataset is then divided into subsets.
3. Apply the above steps recursively to each subset until some stopping criterion is met, such as when a maximum depth or minimum number of samples at a leaf is reached.

Random Forest is an ensemble learning method that enhances the performance of decision trees by combining multiple trees. It is based on the concept of Bootstrap Aggregating (Bagging), which aims to reduce variance and improve generalization. The basic steps of Bagging involve (James et al., 2021, p. 380):

1. Given a training set $D = (x_1, y_1), \ldots, (x_n, y_n)$, where *x* represents the predictive features, and y represents the daily sales.
2. Sample T sets of n elements from D (with replacement) as D_1, D_2,...,D_T. Each D_i will have the same size as the original dataset but may contain duplicates.
3. For each bootstrap sample D_i($i = 1, \ldots, T$): Train a regression model (e.g., decision tree) $f_i(x)$ on D_i, this model learns to predict daily sales based on the input features.
4. Get prediction from all T models: $f_1(x)$, $f_2(x)$, …, $f_t(x)$.

The final prediction for regression problems can simply be taken as the average of all the predictions obtained from all the models. As can be mathematically instantiated, this is expressed by (Hastie et al., 2009) :

$$Final\ Prediction = \frac{1}{T}\sum_{\mathrm{i=0}}^{\mathrm{n}} f_i(x)$$

The Random Forest algorithm is appropriate for this research's dataset since it is robust and handles high dimensionality. This dataset is small, noisy, and imbalanced; therefore, the ensemble approach can help reduce the effect of noise and variability in the data (Biau, G., & Scornet, 2016).

I implemented all the machine learning algorithms in Python using the Scikit-learn library; the ‘RandomForestRegressor’ is utilized to perform random forest regression. The key parameters that were optimized include:

- **n_estimators:** The number of trees in the forest. Having more trees generally improves performance but increases computation time.
- **max_depth:** The maximum depth of each tree. Deeper trees can fit a more complex pattern but may overfit.
- **min_samples_split :** the minimum number of samples required to split an internal node. Higher values prevent overfitting.
- **min_samples_leaf:** This is the minimum number of samples required at a leaf node. Increasing the values avoids creating too many small leaves (Koehrsen,2018).

4.1.2 XGBoost

XGBoost (Extreme Gradient Boosting) is the efficient and scalable implementation of the gradient boosting framework. As per Tianqi Chen and Carlos Guestrin (2016), its major advantages are overfitting avoidance by penalizing complicated models and automatic exploration of different possibilities with built-in cross-validation.

Gradient boosting is an ensemble machine learning algorithm that combines a number of weak models to improve general predictive accuracy. This contributes to the solving of regression and classification problems in multiple domains. ‘Gradient’ in gradient boosting refers to the gradient of the loss function used to minimize errors during training, while "boosting" refers to how the algorithm combines weak predictive models into a single, strong learner (Mason et al., 1999). The steps of the Gradient Boosting Algorithm are:

1. **Initialization:** It is initiated with some initial prediction, usually the mean of the target values in case of regression tasks.
2. **Iterative Boosting:**

- Compute the gradient (first derivative) and the hessian (second derivative) of the loss function with respect to the predictions.
- Fit a decision tree to the gradients and hessians.

- Update the predictions by adding the weighted output of the new tree.

3. **Regularization:** Apply L1 and L2 regularization to control the complexity of the model and prevent overfitting.

The major innovation of XGBoost is related to the use of a more regularized model formalization that produces better performance by controlling overfitting. The XGBoost objective function can be split into two parts: a loss function, including the difference between the predicted and true values, and a regularization term, which will penalize the model for its complexity, expressed below (Chen & Guestrin, 2016):

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where:

- $\text{Obj}(\Theta)$is the regularized objective to minimize.
- The loss function, denoted as $\mathcal{L}(y_i, \hat{y}_i)$, measures the discrepancy between the predicted $\hat{y}_i$ and the actual y_i. For the regression task, XGBoost typically uses the squared error loss function, also known as L2 loss:$\mathcal{L}(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$
- $\Omega(f_k)$ is the regularization term, given by $\Omega(f) = \gamma T + \alpha \sum|w_j| + \frac{1}{2}\lambda \sum w_j^2$, where T is the number of leaves in the tree, γ controls the complexity for each tree, $\alpha \sum|w_j|$ is the L1 regularization term and $\frac{1}{2}\lambda \sum w_j^2$ is the L2 regularization term. This formulation allows independent tuning of α and λ. In this study, the α was set to 0 and λ was set to 1 to improve the computational efficiency.

XGBoost incorporates L1 and L2 regularization to prevent overfitting, enhancing its generalization capability (Youraijournal, 2023), and ensuring robust predictions, which is crucial given the potential noise and variability in the UGC data.

The key hyperparameters that were optimized in this study include:

- **Learning Rate (eta):** It controls the step size at each iteration while moving toward a minimum of the loss function. A small value makes the model robust but at the cost of requiring more trees.

- **Number of Trees (n-estimators):** The number of trees in the model. Adding more trees can make the model better, but it also increases the chance of overfitting.
- **Max_depth:** The maximum depth of each tree. The deeper the trees are, the more complex patterns they can capture. However, with a higher propensity to overfit.
- **Subsample:** Fraction of samples used for training each tree. Lower values prevent overfitting.
- **Colsample_bytree:** Fraction of features used when constructing each tree. Lower values prevent overfitting.

4.1.3 Support Vector Regression

Support Vector Regression follows the same basics as Support Vector Machines. These techniques work on creating a best-possible hyperplane in a high-dimensional space for any given data, making it ideal for cases with non-linear relationships and complex datasets (Smola & Schölkopf, 2004). The SVR has been successfully adopted to solve many forecasting problems, having been applied to product demand forecasting (Guajardo et al., 2006) with remarkable results and also to forecast tourist arrivals (Pai et al., 2006).

Support Vector Regression (SVR) can be defined as the following equation (Vapnik,1999):

$$y = w \cdot \phi(x) + b$$

It utilizes a weight vector w , model inputs x, bias b, and a kernel function $\phi(x)$, which transforms non-linear inputs into a linear mode within a high-dimensional feature space.

The objective of SVR is to minimize the following regularized risk function:

$$\frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{n} max(0, |y_i - (w \cdot \phi(x_i) + b)| - \epsilon)$$

where:

- $\|w\|^2$ is the regularization term that penalizes large weights.

- C is a regularization parameter that controls the trade-off between the model complexity and the amount up to which deviations larger than ϵ are tolerated.
- ϵ is the epsilon-insensitive loss function that defines a margin of tolerance where no penalty is given to errors.

SVR can use various kernel functions to handle non-linear relationships in the data. The choice of kernel depends on the data's characteristics and the task's complexity. Commonly used kernels include:

Linear Kernel:

$$K(x_i, x_j) = x_i \cdot x_j$$

Polynomial Kernel:

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d$$

Radial Basis Function (RBF) Kernel:

$$K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2)$$

Sigmoid Kernel:

$$K(x_i, x_j) = \tanh(\gamma x_i \cdot x_j + r)$$

In this study, the kernel function was set as RBF because of its ability to create non-linear decision boundaries, making it more effective for capturing complex patterns than other kernels (Dibike et al., 2001; Lin et al.,2006).

The hyperparameters for SVR discussed in this study are as follows:

- **C (Regularization Parameter):** Controls the trade-off between achieving a low error on the training data and minimizing the model complexity. A larger value of C means the model will try to fit the training data more closely, potentially leading to overfitting.
- **Epsilon (ϵ):** Defines a margin of tolerance where no penalty is given to errors. It affects the number of support vectors used by the model.
- **Gamma (γ):** Controls how far the influence of a single training example reaches. the influence of a single training example. Low values mean ‘far’ and high values mean ‘close.’ It is used in non-linear kernels like RBF and polynomials. In the Scikit-learn library, when gamma is set to auto, the value

of gamma is calculated as: $\gamma = \frac{1}{n_{features}}$. When gamma is set to scale, the value of gamma is calculated as:$\gamma = \frac{1}{n_{features} \times VAR(x)}$, Here, $VAR(x)$ is the variance of the features in the training data.

4.2 Evaluation Metrics

Evaluation of machine learning model performance is very critical to be sure of the accuracy and reliability of the model outputs. Commonly used metrics include Root Mean Squared Error, Mean Absolute Error, and the Coefficient of Determination. All of these metrics convey specific knowledge about model performance.

Root Mean Squared Error (RMSE): RMSE is the square root of MSE, providing error metrics in the same units as the target variable. It can take on values between 0 and 1, and the lower the RMSE, the better the model performance.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i represents the actual values and $\hat{y}_i$ represents the predicted values, *n* is the number of observations.

Mean Absolute Error (MAE): MAE measures the average absolute difference between actual and predicted values, less sensitive to outliers than MSE. Similar to RMSE, a lower MAE value indicates better model performance.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Coefficient of Determination (R^2): R^2 indicates the proportion of variance in the dependent variable explained by the independent variables. Values range from 0 to 1, with higher values indicating better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

These metrics provide a comprehensive evaluation of model performance, ensuring robust and reliable predictions (Hassan et al., 2023).

4.3 Model Implementation

In this section, I cover the implementation of the Random Forest, XGBoost, and SVR algorithm using the Scikit-learn library in Python, focusing on normalization, hyperparameter tuning, cross-validation, and feature importance. These steps are crucial for optimizing model performance and getting robust results.

4.3.1 Normalization

Before model implementation, all numerical variables in four datasets are normalized so each feature has a mean of 0 and a standard deviation of 1. Normalization scales variables to a unit scale, reducing differences in data values and improving the model's convergence speed (Singh & Birmohan, 2020).

4.3.2 Hyperparameters Tuning

The hyperparameters should be adjusted to find the optimal performance of all three models. In this research, a random search was used for hyperparameter optimization for the three models. Random search examines random combinations of hyperparameters from specified distributions. This approach is particularly effective for high-dimensional parameter spaces and can often find a solution comparable to grid search in a fraction of the time (Bergstra and Bengio, 2012).

The validation of the algorithms was implemented with a cross-validation strategy with three splits. This allows the model to perform very robustly with regard to changes in subsets of the data. To balance the exploration of the parameter space with computational efficiency, the tuning of hyperparameters only involved fitting 90 times each model. The best value of hyperparameters had been chosen by minimizing R^2 on the validation set. Following best practices in the development of machine-learning models (Hastie et al., 2009).

The tuned hyperparameters, along with their search space, can be found in Table 12.

Table 12. Results of Hyperparameter Tuning

Model	Hyperparameter Search Space	Optimal Hyperparameter Value
Random Forest	• n_estimators: 50 to 200 • max_depth: 3 to 15 • min_samples_split: 2 to 10 • min_samples_leaf: 1 to 5	• n_estimators: 88 • max_depth: 9 • min_samples_split: 2 • min_samples_leaf: 1
XGBoost	• n_estimators: 80 to 200 • learning_rate: 0.01 to 0.2	• n_estimators: 102 • learning_rate: 0.0762

	• subsample: 0.6 to 1.0 • colsample_bytree: 0.6 to 1.0 • max_depth: 3 to 10	• subsample: 0.8675 • colsample_bytree: 0.8493 • max_depth: 9
SVR	• C: Regularization parameter (0.1 to 200.0) • gamma: scale, auto	• C: 100 • gamma: auto

4.4.3 Cross Validation

After the best hyperparameters for Random Forest, XGBoost, and SVR were identified, all the datasets were first split into training and test sets with a ratio of 80:20, the models were trained on four training sets employing a 5-fold Cross-Validation strategy. The description of how this technique is applied in general goes as follows: training data is partitioned into five subsets; a model is trained on four subsets and validated on the remaining subset. The process is repeated five times so that each subset is used exactly once as validation data (Hastie, Tibshirani, & Friedman, 2009). This will not only reduce the risk of overfitting by applying cross-validation but also obtain more realistic performance measures.

Afterward, the average R^2 on the validation set was calculated to evaluate the model's performance; The best model was chosen by the best R^2 score. This best model was then retrained on the entire training set to ensure it learned from the maximum amount of data available. Once retrained, the model's performance was evaluated on a separate test set, which was not used during the training or cross-validation process. The performance metrics, including R^2, MAE, and RMSE, were calculated on the test set to assess how well the model generalizes to unseen data. Feature importances were extracted for models that support it: Random Forest and XGBoost.

4.4.4 Feature Importance

Feature importance is also one of the vital components while using a machine learning model because it tells which input variables are most important for making predictions. Both XGBoost and Random Forest have methods for calculating feature importance.

The most popular and primary method for estimating feature importance in Random Forest regression is through the use of Mean Decrease in Impurity, commonly referred to as Gini importance or as Mean Decrease in Mean Squared Error for regression tasks. This

approach averages the total decrease in MSE across all trees within the ensemble (Breiman, 2001).

XGBoost offers multiple methods for calculating feature importance in regression tasks, including gain, weight, and cover, with the gain being the default (Chen & Guestrin, 2016). At each split node, this technique calculates the gain (reduction in MSE) achieved by the split, then accumulates this gain for each feature used in splitting and sums the accumulated gains across all trees for each feature. The last step is to normalize the importance scores, so they sum to 1.

The Mean Decrease in Impurity method in Random Forests can be biased towards variables with many categories or continuous features, as they have a higher chance of reducing impurity (Strobl et al., 2007). XGBoost Gain metric is less prone to such biases because it considers the improvement in the objective function (e.g., accuracy) directly.

In this study, feature importance plots were employed to provide insights into which features have the most significant influence on car sales.

Chapter 5: Results

5.1 Model performance

Each of the four datasets was split 80:20 into training and test sets. I trained Random Forest, XGBoost, and SVR models on the training sets using five-fold cross-validation to ensure reliable performance estimates. The average results on the test sets for each model are summarized in Table 13.

Table 13.Training results of RF, XGBoost, and SVR

Model	Dataset	MAE	RMSE	R^2
Random Forest	1:Benchmark	142.31	220.93	0.64
	2:Benchmark+Autohome	35.36	80.68	0.95
	3:Benchmark+ Weibo	122.88	195.02	0.72
	4:Combined	45.02	94.02	0.94
XGBoost	1:Benchmark	151.35	223.19	0.63
	2:Benchmark+Autohome	37.86	70.38	0.96
	3:Benchmark+ Weibo	125.55	192.10	0.73
	4:Combined	44.34	83.50	0.95
SVR	1:Benchmark	239.98	313.88	0.28
	2:Benchmark+Autohome	119.73	214.93	0.66
	3:Benchmark+ Weibo	220.78	294.63	0.36
	4: Combined	138.26	233.56	0.60

The above table indicates that RF and XGBoost are more suitable for this kind of data. Both algorithms exhibit good stability across various datasets, with XGBoost demonstrating exceptional learning and generalization capabilities. In sharp contrast, SVR performs less stable and gives relatively small R^2 values for both the benchmark model and the one including Weibo data. This suggests that the SVR training processes do not effectively capture the data trends, leading to comparatively poorer performance in practice.

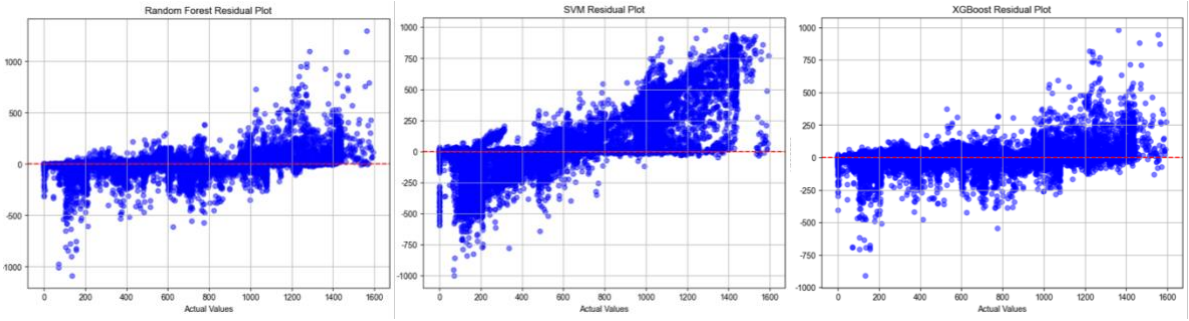


Figure 8.The residual plots for the Random Forest, SVR and XGBoost.

The residual plots (the differences between the interpolated sales and the predicted sales), as illustrated in Figure 9, indicate that both models from XGBoost and Random Forest

outperform in stability and accuracy compared to the model from SVR. In particular, XGBoost has the most stable residuals with fewer outliers, which will make the model most suitable for the data. The variance for the model of SVR is higher, and a more expressed pattern in residuals is noticed, hence posing a challenge to catch up with data trends properly.

Various reasons contribute to the underperformance of SVR:

Sensitivity to Noisy Data and Outliers: Similar to the classification counterpart SVM, SVR does not have support inbuilt mechanisms to deal with noisy data or outliers effectively. The classifier generated by SVR relies on a limited subset of the data, known as support vectors, making it particularly susceptible to noise and outliers within the training set. This sensitivity to data irregularities has been highlighted by Sabzekar and Naghibzadeh in their 2013 study.

Hyperparameters sensitivity: The performance of SVR is highly dependent on the selection of penalty parameters and kernel functions. Different choices can significantly affect the model's prediction performance, and there is no universally accepted method for selecting the optimal SVR parameters (Sun et al., 2021). In other words, RF and XGBoost all had stable performance on different datasets, while SVR had poorer performance, which may be due to sensitivity to noisy and outlying data with the hyperparameter selection problem in handling UGC data from Weibo.

5.2 Dataset Comparison

All models show a rather consistent trend of R^2 and RMSE for various datasets. Most notably, the second dataset, including data from Autohome, significantly boosts model performance. Compared to the baseline dataset, the addition of Autohome data led to an average RMSE reduction of 63.49% for Random Forest, 68.46% for XGBoost, and 31.52% for SVR. This indicates that Autohome data significantly improved the overall prediction accuracy of the models, with an average improvement of 54.42%.

However, combining with Weibo data produces some modest improvement compared to the baseline model but incomparably smaller than the boost brought by Autohome data, with an average improvement of 10.58%. The reason behind this may be due to the different nature of the two platforms. Autohome offers a vast amount of automotive information, including news, reviews, and car owners' feedback, while Weibo mainly offers public opinions, life sharing, and news. Compared with Weibo, potential car buyers are more likely

to browse through Autohome. This indicates data from Autohome may play a greater role in influencing people's car-buying decisions.

It is important to note that all models using the combined dataset show a slight decrease in performance compared to the second model, as illustrated by the slight decrease in R^2. Snijders and Bosker, 1999 suggest two reasons for such a decrease in a larger model: i) chance fluctuation or sampling variance, which would be more marked with small sample sizes, and ii) model misspecification in which the new predictor introduced is redundant relative to one or more predictors already in the model. This redundancy might mean the Weibo data does not provide additional useful information, potentially leading to overfitting or multicollinearity.

The findings reveal that the Autohome data, being rich and comprehensive, provides valuable information that helps the models better capture the underlying trends in the automobile market. However, Weibo, being a social media platform, may contain more noise and less structured information compared to Autohome. This can contribute to the relatively smaller impact on model performance. This suggests that Weibo may not be as rich or relevant for predicting daily sales in the Chinese automobile market.

5.3 Feature Importance

The feature importance analysis for the XGBoost and Random Forest models disclosed several key insights into the factors influencing car sales predictions, as illustrated in Figure 10.

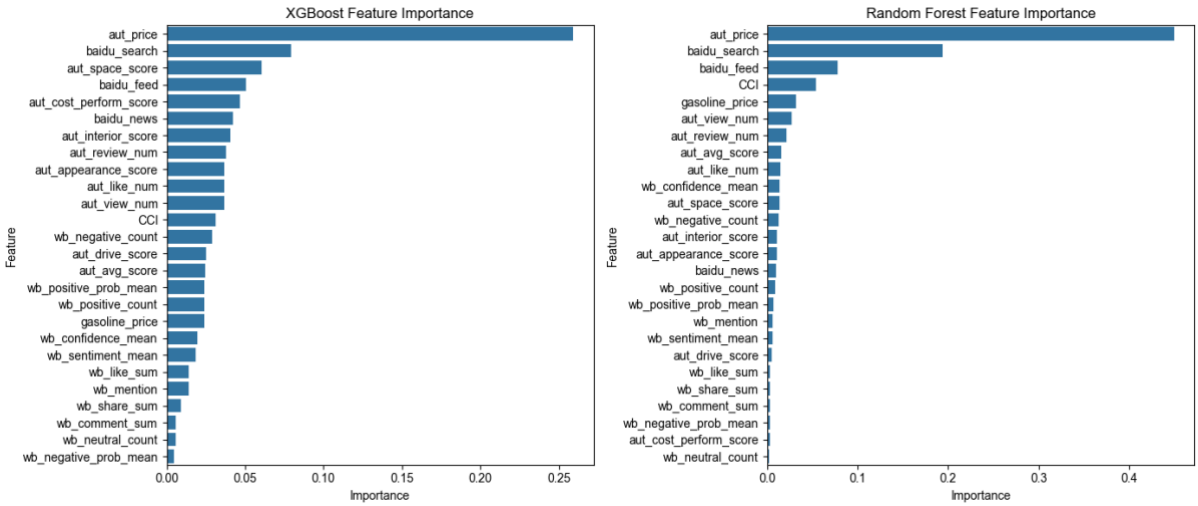


Figure 9. Feature importance of XGBoost VS. Random Forest

Macroeconomic indicators: CCI and gasoline prices play a pretty significant role with regard to predictions of car sales within Random Forest but with a much lesser

importance within XGBoost. Those are some of the important indicators explaining broader economic conditions related to consumer purchasing power and behavior and, therefore, impact car sales.

Baidu Index: The Baidu search index and feed index are very influential in both models, though Baidu news plays a more influential role in XGBoost than in Random Forest. This means the user's search behavior, engaging with personalized content recommendations (e.g. SEO), and being interested in news articles related to a car model name are all important to accurate forecasting for providing real-time understanding of consumer behavior and market trends.

Autohome Data: The key factor that both models highlight is the importance of the Autohome data, which contributes to the greatest extent in predicting automobile sales. The purchase price is the most influential feature, highlighting customers' price sensitivity. Additionally, it shows that review engagement metrics, such as the total number of reviews, review impressions, and the number of likes, are highly significant. Users often interpret high engagement metrics as a sign of quality and credibility, which may influence their perceptions and behaviors (Cheung et al.,2012). The average rating score affects both models. In the case of the XGBoost model, other specific ratings, such as the space rating and cost performance rating for car models, also highly affect the predicted sales. This result shows the significance of detailed product reviews in influencing consumer purchasing decisions.

Weibo Data: The influence of Weibo data in car sales predictions is more subtle. In a Random Forest model, the importance of Weibo engagements, like the number of likes, comments, forwards, and volume of posts, is almost negligible, with the exception of the number of negative Weibo posts representing sentiment and its confidence interval. This suggests that while social media sentiment contributes to the model, it is less critical than direct product rating attributes and search data. Compared with positive and neutral Weibo posts, the XGBoost model assigns greater importance to the number of negative comments. Such negative sentiment expressed on social media might have a much greater effect on the prediction of sales, indicating probably the role of public opinion and social media activities in purchasing behavior.

This analysis emphasizes how proper feature selection and maximal data sources may lead to improvement in the accuracy and reliability of China’s car sales forecasting.

Chapter 6: Conclusion and Discussion

6.1 Conclusion and Recommendations

Based on the main results, I can now address the primary research question: *To what extent does UGC data increase the accuracy of sales forecasting models within the context of the Chinese automotive industry?* This study demonstrates that Autohome data significantly improved the overall prediction accuracy of the models, with an average improvement of 51.59%. In contrast, Weibo data contributes to an average improvement of 9.17%. In addition, the study demonstrates that XGBoost and Random Forest are highly effective for car sales prediction. XGBoost shows the most stable and accurate prediction, followed by Random Forest, whose result stands quite close. SVR shows high variance and relatively poor trend capture. These findings further underline the role of UGC data from specialized platforms, such as Autohome, in enhancing the predictive accuracy of sales forecasting models within the context of the Chinese automotive sector.

The study also answered the second research question: *For* the *Chinese automotive industry, which has a greater impact on car sales: car review data or social media data?* Results show that though Weibo data and Autohome data have both improved the performance of the benchmark model, the impact of Autohome data is more significant. Specifically, the number of reviews, average ratings, and the user purchase prices published on Autohome all have a larger impact on the performance of models. The sentiment metrics of Weibo posts are more influential than engagement metrics and mentions on China's car sales forecasting.

Furthermore, the results show that Baidu Search is highly influential in combining datasets containing Autohome and Weibo data. CCI and gasoline prices also play a pretty significant role in predicting China's car sales. The findings have increased the need to include macroeconomic indicators and search trends data to increase the accuracy and reliability of the models in China's car sales forecasting.

The recommendations for predicting car sales in the Chinese market are as follows:

Firstly, more elaborated models could be explored using other data present within Autohome, including the ranking of car sales, automotive news, and car video or text reviews. Seasonality and holiday effects could be taken into consideration, categorical variables for holidays might also enhance the performance of the model.

Secondly, when analyzing the impact of user-generated content on sales, the focus should be on content from automotive-specific forums or automotive influencers on social media platforms. This approach will ensure that what is obtained is more specialized and relevant information that may have great influences on car sales predictions.

Thirdly, For Chinese automobile manufacturers, it is crucial to pay more attention to negative news on Weibo, as it may have a significantly greater impact on sales than neutral or positive content. Additionally, conducting feature importance analysis on Autohome data and then focusing on car attributes that have a larger influence on sales can be beneficial.

6.2 Limitations and Future Research

The relatively low R^2 values for the SVR results reflect poor model performance. This may be due to several reasons. First, the Weibo dataset is probably full of noise, which may prevent the model from learning effectively. Noisy data usually comes from the web crawling process, which may capture a great proportion of spam accounts that provide irrelevant or misleading information. The next issue is the small size of sales data. This easily leads to model overfitting, capturing the noise but missing the trend. Meanwhile, according to a study, there are still some misinterpretations existing for the current mainstream Chinese sentiment analysis APIs (Tang et al., 2020). Some Weibo data may have been misclassified using the Baidu API for sentiment analysis; there was a huge volume of raw Weibo textual data hence making it not feasible to carry out manual sentiment verification.

To address these limitations, future work should focus on improving data collection methods and enhancing noise reduction techniques. For instance, more complex algorithms in data cleaning could potentially help weed out more noise from the social media data, and a better filtering of spam accounts. Also, the adoption of more precise sentiment analysis or the application of at least two types of sentiment analysis tools could provide more reliable sentiment data. An increase in the amount of data and model upgrades using new data will enhance their accuracy and reliability even further. With more data, the problems of small samples will be reduced, and the generalization capacity of the models will increase. These are the limitations whose addressing will let models provide more accurate and reliable predictions, beneficial for their practical application in different areas of business activities.

Reference

Abu-Eisheh, S. A., & Mannering, F. L. (2002). Forecasting automobile demand for economics in transition: A dynamic simultaneous equation system approach. *Transportation Planning and Technology, 25*(4), 311-331.

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services, 8*(3), 147-156. https://doi.org/10.1016/S0969-6989(00)00011-4

Amin, M. S., Ayon, H., Ghosh, B. P., Chowdhury, M. S., Bhuiyan, M. S., Jewel, R. M., & Linkon, A. A. (2024). Harmonizing macro-financial factors and Twitter sentiment analysis in forecasting stock market trends. *Journal of Computational Science and Technology Studies, 6*(1), 7. https://dx.doi.org/10.32996/jcsts.2024.6.1.7

Barwick, P. J., Li, S., & Wallace, B. (2021). Local protectionism, market structure, and social welfare: China's automobile market. *American Economic Journal: Economic Policy, 13*(1), 1-35. https://doi.org/10.1257/pol.20180416

Bahtar, A. Z., & Muda, M. (2016). The Impact of User-Generated Content (UGC) on Product Reviews towards Online Purchasing – A Conceptual Framework. *Procedia Economics and Finance, 37*, 337–342. https://doi.org/10.1016/S2212-5671(16)30134-4

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281-305. Retrieved from https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*, 197-227.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Chen, T., & Guestrin, C. (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). ACM.

Chen, Y., Lin Lawell, C.-Y. C., & Wang, Y. (2020). The Chinese automobile industry and government policy. *Research in Transportation Economics*, *84*, 100849. https://doi.org/10.1016/j.retrec.2020.100849.

Cheung, C. M., & Thadani, D. R. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support*

Systems, *54*(1), 461-470.

China Daily. (2022). https://caijing.chinadaily.com.cn/a/202201/11/WS61dd1990a3107be497a01a76.html

CTR. (2023). 2023 China Search Engine Industry Research Report. Retrieved from https://36kr.com/p/2298513609122179

Cui, L., Zhang, H., & Huang, L. (2017). Enhancing operational decisions in supply chain and operations management through social media data. *Journal of Operations Management, 49-51*, 21-33.

Dai, D., Fang, Y., Wang, S., & Zhao, M. (2023). Prediction of China automobile market evolution based on univariate and multivariate perspectives. *Systems, 11*(8), 431. https://doi.org/10.3390/systems11080431

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. http://arxiv.org/abs/1810.04805

Dibike, Y. B., Velickov, S., Solomatine, D., & Abbott, M. B. (2001). Model induction with support vector machines: introduction and applications. Journal of Computing in Civil Engineering, 15(3), 208-216.

Ding, Y., Wu, P., Zhao, J., & Zhou, L. (2023). Forecasting product sales using text mining: A case study in new energy vehicle. *Electronic Commerce Research*. https://doi.org/10.1007/s10660-023-09701-9

Ding, Z. (2023). Sales Forecast of New Energy Vehicles in China Based on LSTM Model. *Frontiers in Business, Economics and Management*, *10*(3), 47–49. https://doi.org/10.54097/fbem.v10i3.11209

Dong, Y., Zhou, S., Xing, L., Chen, Y., Ren, Z., Dong, Y., & Zhang, X. (2022). Deep learning methods may not outperform other machine learning methods on analyzing genomic studies. Frontiers in Genetics, 13, 992070. https://doi.org/10.3389/fgene.2022.992070

Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics, 170*, 97-135. https://doi.org/10.1016/j.ijpe.2015.09.014

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage publications.

Gao, J., Xie, Y., Cui, X., Yu, H., & Gu, F. (2018). Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data: A method based on econometric model. *Advances in Mechanical Engineering, 10*(2), 1-11.

https://doi.org/10.1177/1687814017749325

Goh, K.-Y., Heng, C.-S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *Information Systems Research, 24*(1), 88-107. https://doi.org/10.1287/isre.1120.0469

Goes, P. B., Lin, M., & Au Yeung, C. M. (2014). Popularity effect in user-generated content: Evidence from online product reviews. *Information Systems Research, 25*(2), 222-238. https://doi.org/10.1287/isre.2014.0514

Guajardo, J., Weber, R., & Miranda, J. (2006). A forecasting methodology using support vector regression and dynamic feature selection. *Journal of Information & Knowledge Management, 5*(4), 329-335.

Hassan, M. M., Yasmin, F., Khan, M. A. R., Zaman, S. G., & Bairagi, A. K. (2023). A comparative assessment of machine learning algorithms with the least absolute shrinkage and selection operator for breast cancer diagnosis. *Journal of Applied Mathematics and Decision Sciences*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer. Retrieved from https://hastie.su.domains/Papers/ESLII.pdf

Huang, J. H., & Chen, Y. F. (2006). Herding in online product choice. *Psychology & Marketing, 23*(5), 413-428. https://doi.org/10.1002/mar.20119

Jabr, N., & Zheng, Z. (2014). The impact of product ratings on sales in competitive markets. *Journal of Marketing Research, 51*(3), 1-12. https://doi.org/10.1509/jmr.13.0219

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

Kim, A. J., & Johnson, K. K. P. (2016). Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook. *Computers in Human Behavior, 58*, 98-108. https://doi.org/10.1016/j.chb.2015.12.047

Klier, T., & Linn, J. (2010). The price of gasoline and new vehicle fuel economy: Evidence from monthly sales data. *American Economic Journal: Economic Policy, 2*(3), 134-153.

Kolchyna, O. (2017). Measuring the impact of social media on sales: A study of 75 brands. *Journal of Marketing Analytics, 5*(3), 192-206. https://doi.org/10.1057/s41270-017-0021-4

Koehrsen, W. (2018). A guide to time series forecasting with ARIMA in Python 3. Towards Data Science. Retrieved from https://towardsdatascience.com/a-guide-to-time-series-

forecasting-with-arima-in-python-3-324d10c05a

Lepot, M., Aubin, J.-B., & Clemens, F. H. L. R. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, *9*(10), 796. https://doi.org/10.3390/w9100796

Li, J. (2020). Charging Chinese future: The roadmap of China's policy for new energy automotive industry. *International Journal of Hydrogen Energy*, *45*(20), 11409–11423. https://doi.org/10.1016/j.ijhydene.2020.02.075

Li, Y., Yu, L., & Wen, R. (2019). Predicting sales by online searching data keywords based on text mining: Evidence from the Chinese automobile market. *Journal of Physics: Conference Series*, *1325*(1), 012071. https://doi.org/10.1088/1742-6596/1325/1/012071

Li, X., Wu, L., & Wang, Y. (2022). Information overload and consumer decision-making in the era of big data. *Journal of Business Research*, 139, 1-12. https://doi.org/10.1016/j.jbusres.2021.09.001

Liapis, C. M., Karanikola, A., & Kotsiantis, S. (2021). A Multi-Method Survey on the Use of Sentiment Analysis in Multivariate Financial Time Series Forecasting. *Entropy*, 23(12), 1603. https://doi.org/10.3390/e23121603

Lin, J. Y., Cheng, C. T., & Chau, K. W. (2006). Using support vector machines for long-term discharge prediction. *Hydrological sciences journal*, *51*(4), 599-612.

Liu, N., Ren, S., Choi, T. M., Hui, C. L., & Ng, S. F. (2013). Sales forecasting for fashion retailing service industry: A review. *Mathematical Problems in Engineering*, 2013, 1-9.

Liu, X., Lee, D., & Srinivasan, K. (2019). Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research*, *56*(6), 918-943. https://doi.org/10.1177/0022243719866690

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. *Advances in neural information processing systems*, *12*.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting methods and applications. John wiley & sons.

Malakooti, B. (2013). Operations and production systems with multiple objectives.

McElfresh, D., Khandagale, S., Valverde, J., C, V. P., Feuer, B., Hegde, C., Ramakrishnan, G., Goldblum, M., & White, C. (2024). *When Do Neural Nets Outperform Boosted Trees on Tabular Data?* (arXiv:2305.02997). arXiv. http://arxiv.org/abs/2305.02997

Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings

forums. *Journal of Marketing Research*, 48(3), 444-456. https://doi.org/10.1509/jmkr.48.3.444

Mohan, S., Solanki, A., Taluja, H., Anuradha, & Singh, A. (2022). Predicting the Impact of the Third Wave of COVID-19 in India Using Hybrid Statistical Machine Learning Models: A Time Series Forecasting and Sentiment Analysis Approach. *Computers in Biology and Medicine*, 105354. https://doi.org/10.1016/j.compbiomed.2022.105354

National Bureau of Statistics of China . (2023), Consumer confidence index (CCI). Retrieved from https://www.stats.gov.cn/zs/tjws/tjbk/202301/t20230101_1912948.html

OECD. (2023). Consumer confidence index (CCI). Retrieved from https://www.oecd.org/en/data/indicators/consumer-confidence-index-cci.html

Pai, P. F., Wei-Chiang, H., Ping-Teng, C., & Chen-Tung, C. (2006). The application of support vector machines to forecast tourist arrivals in Barbados: An empirical study. *International Journal of Management*, *23*(2), 375.

Pan, Y., & Zhang, J. Q. (2011). Born unequal: A study of the helpfulness of user-generated product reviews. *Journal of Retailing*, 87(4), 598-612. https://doi.org/10.1016/j.jretai.2011.05.002

Pavithra, M., & Velmurugan, R. (2023). Factors associated with consumer confidence. In *E3S Web of Conferences* (Vol. 449, p. 04013). EDP Sciences.

QuestMobile. (2023). 2023 China Internet Core Trends Annual Report. Retrieved from https://www.questmobile.com.cn/research/report/1737028262113153026

Sabzekar, M., & Naghibzadeh, M. (2013). Fuzzy c-means improvement using relaxed constraints support vector machines. *Applied Soft Computing*, *13*(2), 881–890. https://doi.org/10.1016/j.asoc.2012.09.018

Sammy Paget. (2024), https://brightlocal.com/research/local-consumer-review-survey

Singh, D., & Birmohan, S. (2020). Investigating the impact of data normalization on classification performance. *Journal of Machine Learning Research*, *21*(1), 1-20.

Singh, S. (2018). Natural language processing for information extraction (arXiv:1807.02383). *arXiv*. http://arxiv.org/abs/1807.02383

Shahabuddin, S. (2009). Forecasting automobile sales. *Management Research News*, *32*(7), 670–682. https://doi.org/10.1108/01409170910965260

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199-222.

Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological methods & research*, *22*(3), 342-363.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*, 1-21.

Su, C. W., Wang, D., Mirza, N., Zhong, Y., & Umar, M. (2023). The impact of consumer confidence on oil prices. *Energy Economics*, *124*, 106820.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, 58(4), 696-707. https://doi.org/10.1287/mnsc.1110.1438

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., & Wu, H. (2019). *ERNIE: Enhanced Representation through Knowledge Integration* (arXiv:1904.09223). arXiv. http://arxiv.org/abs/1904.09223

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., Liu, W., Wu, Z., Gong, W., Liang, J., Shang, Z., Sun, P., Liu, W., Ouyang, X., Yu, D., ... Wang, H. (2021). ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation (arXiv:2107.02137). *arXiv*. http://arxiv.org/abs/2107.02137

Sun, Y., Ding, S., Zhang, Z., & Jia, W. (2021). An improved grid search algorithm to optimize SVR for prediction. *Soft Computing*, *25*(7), 5633–5644. https://doi.org/10.1007/s00500-020-05560-w

Tang, T., Huang, L., & Chen, Y. (2020). Evaluation of Chinese sentiment analysis APIs based on online reviews. *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 923-927. https://doi.org/10.1109/IEEM45057.2020.9309968

Tang, L., Li, J., Du, H., Li, L., Wu, J., & Wang, S. (2022). Big Data in Forecasting Research: A Literature Review. *Big Data Research*, *27*, 100289. https://doi.org/10.1016/j.bdr.2021.100289

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1), 1-20. https://doi.org/10.1287/mksc.2018.1120

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988-999.

Verdonck, M. (2019). Predicting car sales using Instagram data: An analysis of engagement metrics. Journal of Business Research, 101, 1-10. https://doi.org/10.1016/j.jbusres.2019.03.012

Wang, K.-H., Su, C.-W., Xiao, Y., & Liu, L. (2022). Is the oil price a barometer of China's automobile market? From a wavelet-based quantile-on-quantile regression

perspective. *Energy*, *240*, 122501. https://doi.org/10.1016/j.energy.2021.122501

Wang, Y., Li, J., & Zhang, H. (2020). The impact of government policies on the adoption of new energy vehicles: The case of China. *Energy Policy*, 139, 1-12. https://doi.org/10.1016/j.enpol.2020.111326

Wang, Y., Li, J., & Zhang, H. (2021). The role of social influence in the adoption of new energy vehicles: Evidence from China. *Energy Policy*, 149, 1-12. https://doi.org/10.1016/j.enpol.2020.112023

Wang, X., Lin, Z., & Zhu, T. (2022). The impact of online reviews on offline sales of high-involvement products: Evidence from the Chinese automobile industry. Journal of Business Research, 139, 1-12. https://doi.org/10.1016/j.jbusres.2021.09.001

Wachter, S., Mittelstadt, B., & Floridi, L. (2019). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76-99. https://doi.org/10.1093/idpl/ipx005

Wu, F., Sun, Q., Grewal, R., & Li, S. (2019). Brand Name Types and Consumer Demand: Evidence from China's Automobile Market. *Journal of Marketing Research*, *56*(1), 158–175. https://doi.org/10.1177/0022243718820571

Wu, Z., He, Q., Li, J., Bi, G., & Antwi-Afari, M. F. (2023). Public attitudes and sentiments towards new energy vehicles in China: A text mining approach. *Renewable and Sustainable Energy Reviews*, *178*, 113242. https://doi.org/10.1016/j.rser.2023.113242

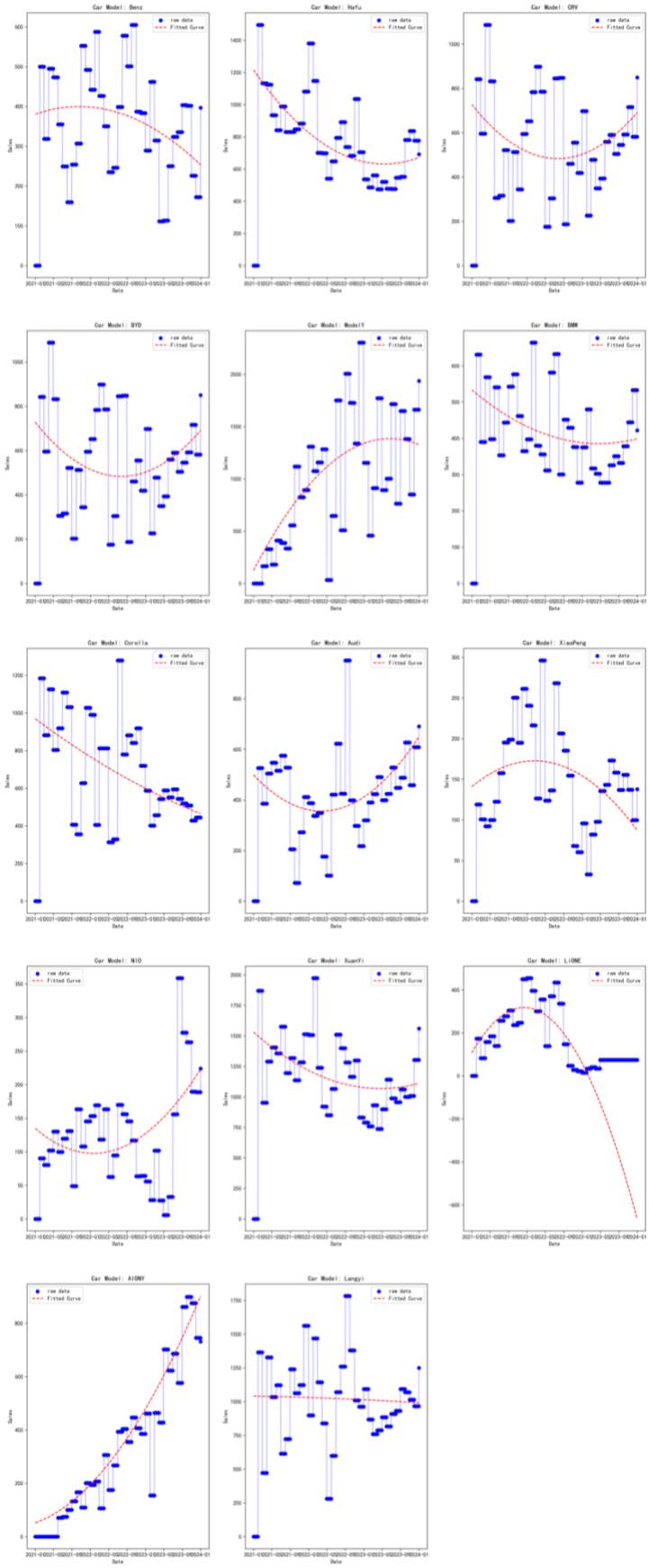
Yang, Z., & Zhang, C. (2020, July). Automobile sales forecast based on web search and social network data. In *Proceedings of the 2020 11th International Conference on E-business, Management and Economics* (pp. 37-41).

Youraijournal. (2023). XGBoost in machine learning: The ultimate guide.

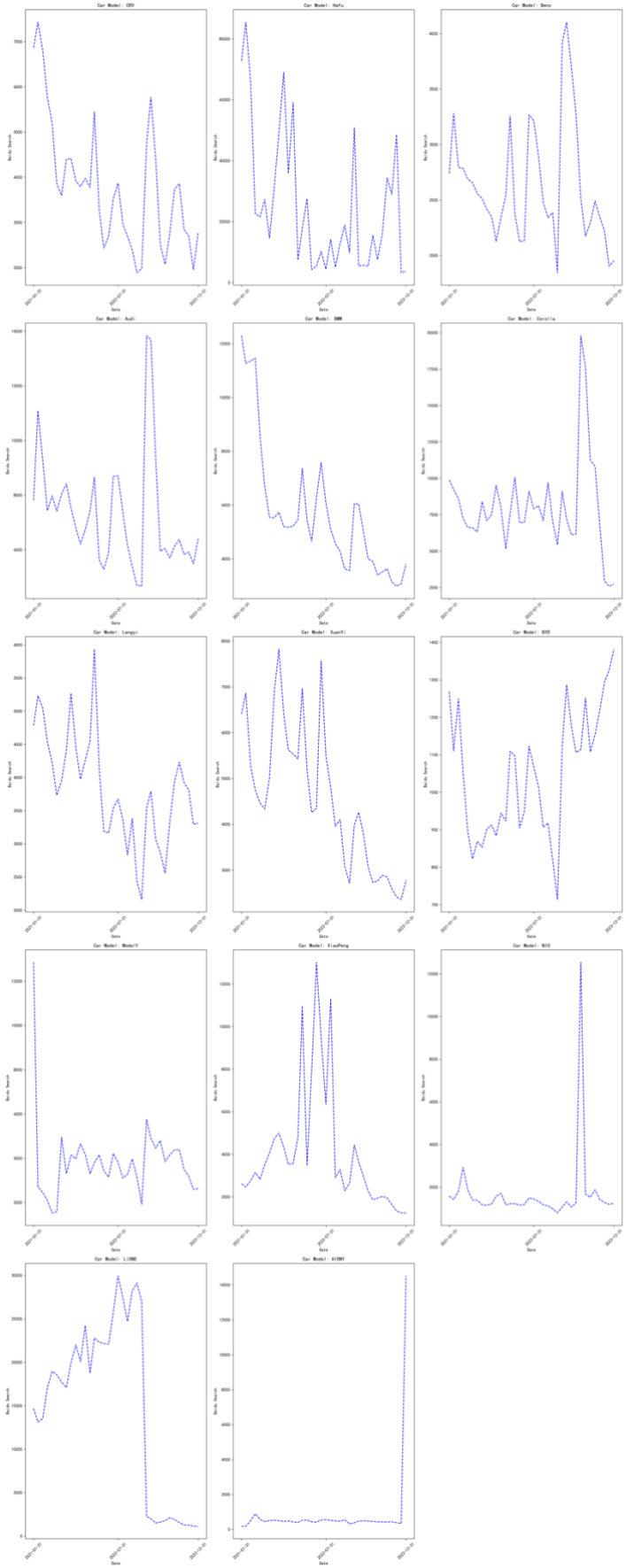
Zhang, Y., Zhong, M., Geng, N., & Jiang, Y. (2017). Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China. *PLOS ONE*, *12*(5), e0176729. https://doi.org/10.1371/journal.pone.0176729

Zhu, X., Zhang, G., & Sun, B. (2019). A comprehensive literature review of the demand forecasting methods of emergency resources from the perspective of artificial intelligence. *Natural Hazards*, *97*, 65-82.

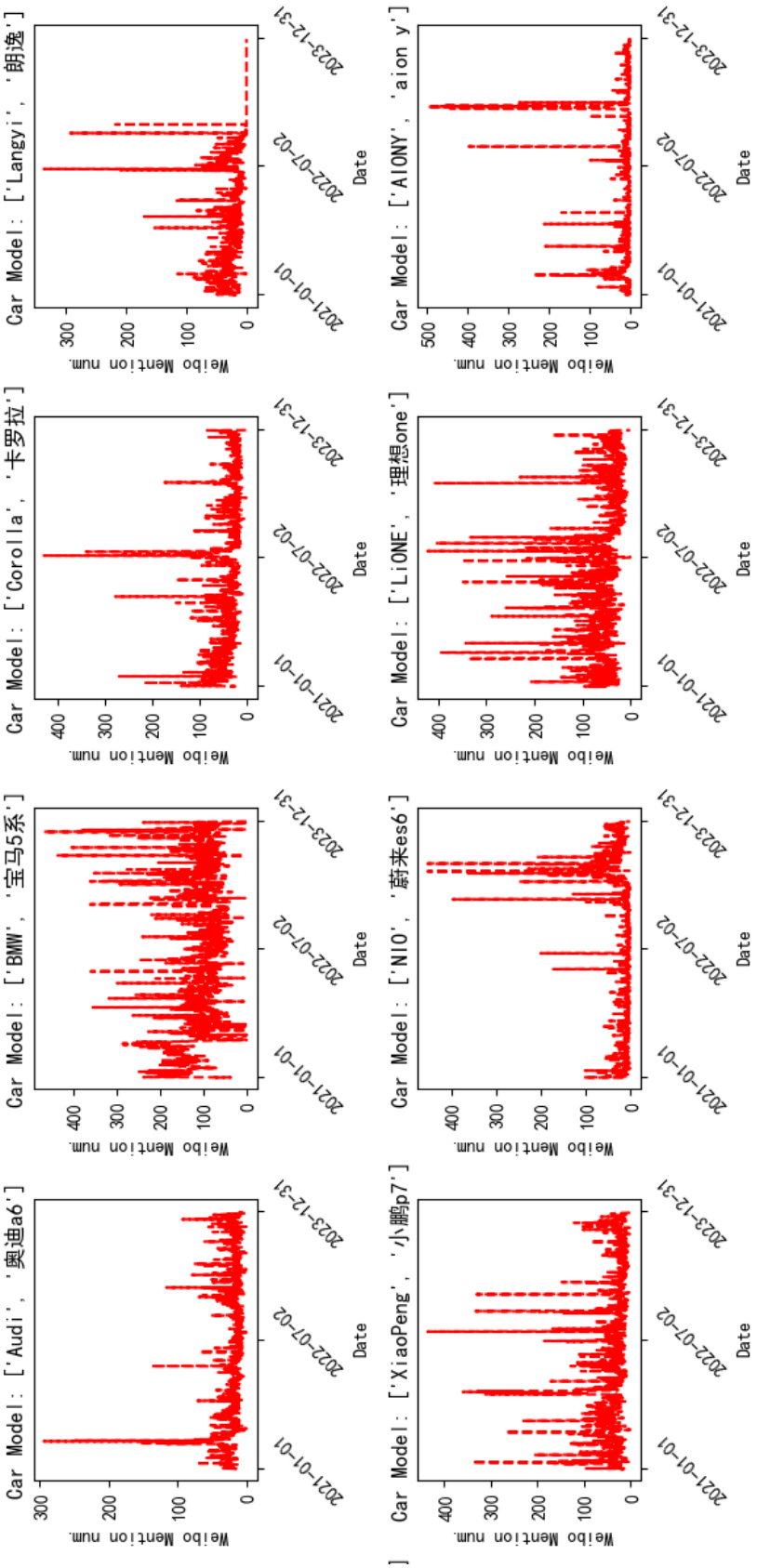
Appendix 1: The 4-degree Polynomial Curve on Average Daily Sales

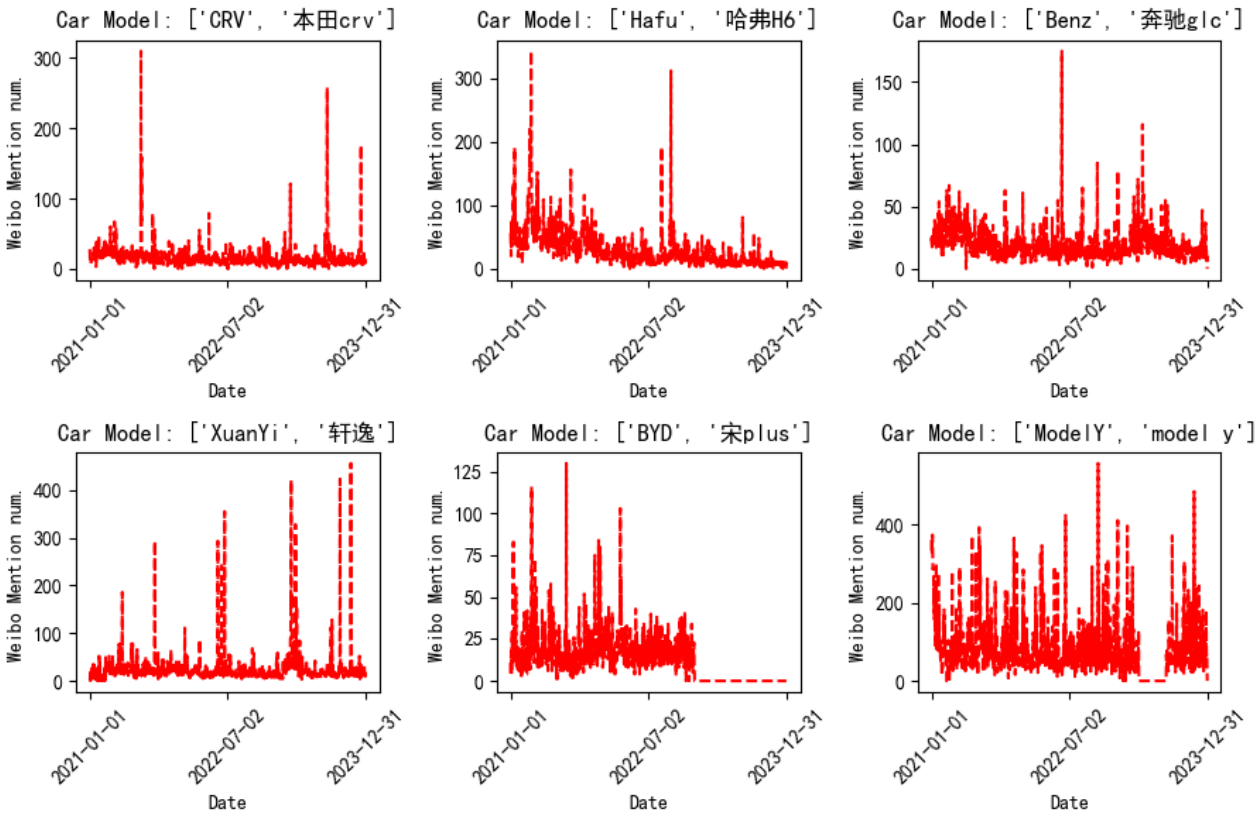


Appendix 2: Baidu Search Trends for Various Car Models



Appendix 3: The Number of mentions for various car brands on Weibo




Car Model: ['CRV', '本田crv']
Car Model: ['Hafu', '哈弗H6']
Car Model: ['Benz', '奔驰glc']
Car Model: ['XuanYi', '轩逸']
Car Model: ['BYD', '宋plus']
Car Model: ['ModelY', 'model y']
Weibo Mention num.
Date
2021-01-01
2022-07-02
2023-12-31