Using Data Mining Techniques to Assess the Impact of COVID-19 on the Auto Insurance Industry in China

JIANGSHAN WANG

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

December 2021

Abstract

Since coronavirus disease 2019 (COVID-19) was discovered at the end of 2019, the whole world has been severely affected. The insurance industry, regarded as an important factor in recovery, has also been affected by COVID-19. However, effective data mining techniques have rarely been utilized in the insurance industry in China, especially under the circumstances of COVID-19. Although some traditional statistical analysis methods have been applied to this area, the limitation of the lack of data distribution still cannot be efficiently overcome. With the machine learning technique proposed in this thesis, this limitation can be solved by using a stacking model with great generalization ability. In this research, the ElasticNet, LightGBM, and Random Forest approaches were employed as base learners; ridge and LASSO regression were used as meta-models to increase the prediction accuracy; and the SHAP value was utilized to explain the impact of COVID-19 on the insurance industry in China. The stacking meta-model in this thesis has a mean absolute percentage error (MAPE) of 12.57134, whereas the average value in the past week is 21.50972, and the MAPE of ElasticNet is 22.57935. In conclusion, COVID-19 affects the auto insurance industry in China.

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor Dr. Huaiping Zhu, for always unconditionally supporting and guiding my thesis work at any time when needed throughout the course of my master degree. Dr. Zhu has helped me by providing not only his invaluable professional knowledge, but also his patience. His vision and motivation have deeply and positively influenced my attitude toward completing research. He has taught me the methodology of conducting thesis research as well as the proper thesis structure, and it was a great privilege and honor to work with and learn from him.

Moreover, I want to thank Professor Zijiang Yang who have offered instructions and technical support of my research whenever I needed.

Last but not least, this thesis could not have been completed without the support of my parents and my best friend Joylie – I would like to thank them for always providing efficient methods of de-stressing. I greatly appreciate the spiritual support I have received from them throughout my life.

Table of Contents

List of Tables

List of Figures

1. Introduction

According to Worldometers data (n.d.), as of May 27, 2021, the global number of people infected with coronavirus disease 2019 (COVID-19) reached 160 million, and the death toll reached 3 million. It is still deemed a pandemic worldwide. Individuals, businesses, societies, and the global economy are severely affected. The main stock markets in the world have collapsed, such as the Financial Times Stock Exchange, Dow Jones Industrial Average, and Nikkei Index. Affected by the epidemic, the economies of most countries are in recession. The insurance industry plays a very important role in the national economy by ensuring the welfare of individuals, organizations, and enterprises. Further, a healthy and well-developed insurance industry improves the stability of the financial markets. Growing risk awareness and social protection promote the growth of the global insurance industry, thereby increasing premiums. According to Allianz Research, the global premium income reached 3.906 trillion Euros in 2019. Life insurance gradually became a stronger line of business against the backdrop of the aging population and was the major driver of global insurance growth in 2019. Through the joint efforts of various countries and the promotion of the COVID-19 vaccine, the

impact of COVID-19 on the world economy has lessened, and the international economy has gradually recovered. Data show that in the first quarter of 2021, international logistics and trade will improve further, and a clear recovery trend can be expected from the overall picture. The International Monetary Fund stated in the 2021 World Economic Report that the global economy was expected to decrease by 3.3% in 2020 and grow by 6% in 2021.

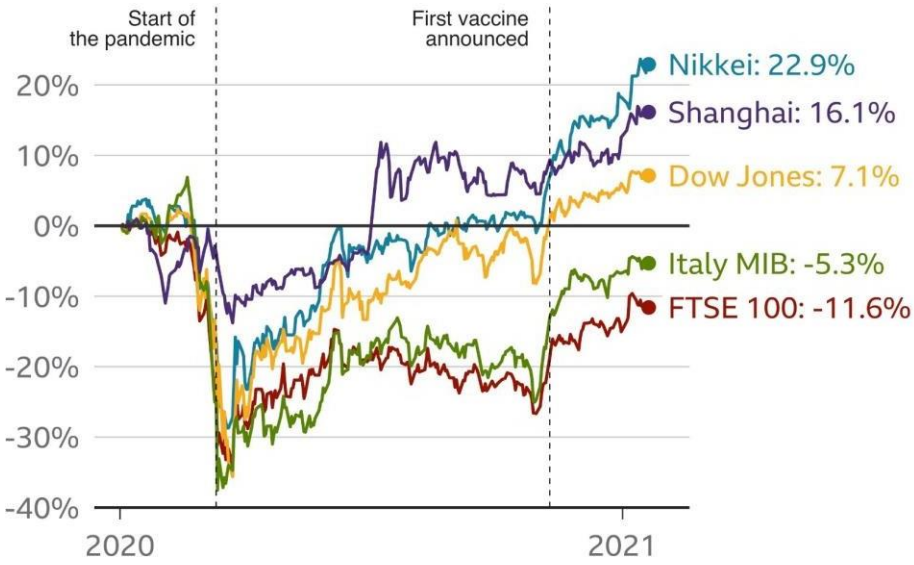

Figure 1. Influence of COVID-19 on the world economy

The pandemic decreased the capital of the major reinsurance companies in China by 8%, as well as the ratings of many insurance companies. The Chinese auto insurance industry underwent comprehensive reforms in 2020, a key stage of transformation and upgrading.

There is increasing interest in applying machine learning to assess the impact of the pandemic on the insurance industry, although the existing analysis of and research on auto insurance claims is very limited. Therefore, the objectives of this study were to analyze the impact of the pandemic on the auto insurance industry through the use of

machine learning technology and to predict the trend of auto insurance claims as well as the development of the industry. Among the related difficulties are the issues of how to extract valid information and historical data features effectively and how to use them to predict insurance claims. Mection, including and excluding pandemic factors, is also trending. Analysis of these two areas will be beneficial to upgrading and digital innovation in the industry, such as enabling reasonable determination of claim amount rates based on historical data.

1.1 Problem Definition

The insurance industry plays a very important role in economic recovery; hence, studying the impacts of pandemics on this industry and forecasting its development trend are of great importance.

Two issues were addressed in this study. The first was how to extract feature data and underlying information efficiently, instead of using only surface statistical information. Given the possibly poor quality of real-world data, such as missing values, outliers, and noise, the model could extract and employ undesirable data, resulting in poor performance. To prevent such problems, the data were firstly preprocessed, and outliners were removed. Feature selection was then performed on the clean data. To address the massive amount of data, representative features were reasonably chosen to build a feature set with a realistic number of features.

The second question was how to make more accurate predictions based on historical data. Specifically, the goal was to predict future data based on extracted data features and information combined with historical data. Given that different predictive models

have different assumptions and characteristics and that real data are often more complicated, it is necessary to formulate more reasonable assumptions when designing predictive models. Furthermore, the product must be adjusted based on the data used. The complexities of the setting and data require analysis of the relevance of different models.

Conducting experimental analysis to illustrate the degree of impact of the pandemic using the predictive model provides further insight into the effects of the pandemic on the auto insurance industry and their extent. Simultaneously, the model must remain interpretable to elucidate the importance of the impacts of the pandemic and other related factors.

1.2 Significance

As an important financial industry in China, the insurance industry not only is a professional risk manager, but also helps promote national economic development, maintain social stability, increase employment opportunities, and ensure the stability of the lives of individuals. The impact of COVID-19 on the insurance industry, as well as overall economic and social development, were of interest in this research. Understanding these effects will not only help the insurance industry respond actively, but also facilitate assessment of the effects of COVID-19 on the entire society and the national economy from the perspective of a sub-industry.

The auto insurance data of the property insurance company assessed in this research were time series data, and the application of regression algorithms to the prediction of auto insurance claims was studied. The ample amount of research on the application of feature engineering methods and prediction models to time series data is favorable for

this research. Most insurance companies still prefer traditional statistical models that have limited abilities to optimize available data and hence prevent accurate determination of auto insurance rates. Meanwhile, data analysis shows that artificial features cannot effectively extract in-depth information from data. Compared to traditional statistical models, machine learning algorithms enable more effective data feature extraction and provide more accurate predictions.

After incorporating pandemic-related features into the algorithm model, feature engineering and time series predictive methods were employed to extract meaningful features efficiently from a massive amount of data. To prove the advantage of this methodology, several comparative experiments were performed on the prediction results of the moving average model and other traditional algorithms. The application of machine learning algorithms to the prediction of auto insurance claims was analyzed in terms of algorithm reasoning, model construction, and result prediction. The final chapter accordingly provides suggestions on the design of new auto insurance products and the formulation of insurance technology policies.

1.3 List of Contributions

In this study, the following aspects were analyzed and data mining technology was utilized to assess the impact of COVID-19 on the auto insurance industry in China.

Research on auto insurance claims: The main focus was the impact of the pandemic on the auto insurance industry. Using machine learning algorithms on historical data to predict claim amounts enabled reasonable determination of the insurance premium rate.

Data preprocessing technology: Real-world data were firstly cleaned and preprocessed for optimization.

Feature selection: To extract feature elements effectively, feature selection was conducted on auto insurance claim data.

Combining pandemic features: Pandemic-related features were added to the algorithm model to analyze the impacts of the pandemic on the insurance claims industry.

Interpretability: Regarding the SHAP value, the importance and effects of pandemic features were examined.

Optimal predictive model: Based on related research, this thesis proposes an optimal predictive model for auto insurance industry data.

1.4 Thesis Outline

This thesis is organized as follows. Chapter 1 provides an overview of the research question addressed in this thesis. Chapter 2 presents a literature review of relevant data mining techniques and previous research regarding the time series data prediction algorithm utilized in this study. Chapter 3 describes the feature extraction techniques employed, as well as some relevant machine learning algorithms. Chapter 4 formally presents the proposed method. Chapter 5 introduces the dataset and evaluation metrics used in this research, discusses the feature extraction results, and describes the experiments conducted on the prediction models. Finally, Chapter 6 concludes this thesis and discusses future work and potential improvements of the proposed method.

2. Literature Review

2.1 Literature Review of Traditional Metering

Willmot (1986) found that, with many distribution models, the skewness of the number of claims makes it necessary to consider the model with a certain skewness when selecting the distribution model. Willmot adopted an inverse Gaussian distribution and a noncentral chi-squared distribution as the structure density function of the Poisson distribution and demonstrated that the established mixed Poisson distribution could well satisfy the skewness of the number of claims. According to Haberman (1990), a generalized linear model is an effective means of addressing issues such as the number and intensity of claims. As a well-known statistical tool, the generalized linear model (GLM) is typically used for interpretive analysis of automobile insurance claims, rate determination, and loss prediction in claim prediction. A loss of data is required to follow a certain distribution assumption; that is, the number of claims follows a Poisson distribution or a negative binomial distribution. In addition, the claim intensity follows a gamma distribution or an inverse Gaussian distribution, and the cumulative compensation follows a Tweedie distribution. Gerber (1992) proposed using

a generalized negative binomial distribution to study the number of claims, as the unconditional generalized gamma distribution family is a general negative binomial distribution. The generalized negative binomial distribution derived from the mixed generalized gamma distribution can be applied to the number of claims models. Walhin and Paris (1999) also established a new mixed Poisson distribution model by combining the advantages of the Poisson inverse Gaussian and negative binomial distributions and empirically proved that the mixed Poisson distribution formed through such a combination can fit the claim experience. The data had a very good effect. Because the number of claims has the characteristic of zero accumulation, many theoretical studies have been conducted on addressing the zero-inflation issue to solve the problem of the number of claims better and more reasonably. For example, Lambert (1992) used a zero-inflated Poisson distribution model for the first time to conduct a fitting study on empirical data and found the fitting effect to be significant. In addition, Gupta (1996) introduced a zero-inflated generalized Poisson distribution into empirical research to establish a reasonable model for the number of claims data with zero inflation and achieved good results. Yip and Yau (2005) used different zero-inflation models to fit a set of auto insurance claim data. After discussing the application of the ZINB, ZIP, ZIGP, and ZIDP models in non-life insurance, and through comparative analysis, they found that the ZID model fit auto insurance data more closely than the other models. Ismail and Jemain (2007) systematically introduced an over-dispersed negative binomial distribution regression model and a generalized Poisson regression model, compared the differences and connections between them, and delineated their respective conditions of use.

In summary, statistics and measurement models have been well developed in the

analysis and prediction of auto insurance claims and have been continuously optimized in a mathematical or statistical sense to achieve close fitting. Relatively speaking, statistical and measurement models are advantageous in explaining the effects of an epidemic on auto insurance claims; for instance, they are easy to apply and can intuitively describe the parameter estimation results. However, their shortcomings are obvious. It is necessary to determine the functional relationship between the dependent and explanatory variables in advance, and the form of the function is relatively limited. In addition, statistical and quantitative models cannot automatically identify the interaction between the explanatory variables, which also renders the establishment of the modeling process time-consuming, and statistical and econometric models rely on assumptions about the distribution. If the assumptions regarding the distribution are wrong, the sum of the squared errors of the fitted values may not be the smallest, which means that the model does not fit the actual effect well.

2.2 Literature Review of Predictive Models

Machine learning, the core of artificial intelligence, is an interdisciplinary procedure in many fields involving probability theory, statistics, approximation theory, convex analysis, computational complexity theory, and other disciplines. It focuses on studying ways to improve the performance of a system through computational methods and experience. A machine learning algorithm is applied for this purpose, which enables prediction and decision-making without explicit programming by constructing mathematical models from target datasets. Machine learning algorithms, as completely new prediction models, have already been applied successfully and produced valuable results in many different fields. With the general trend of applying data analysis results for car insurance risk management and pricing and assisting insurance companies in

transforming from price competition to improved management and innovative capabilities, research into the application of machine learning to the business data used by car insurance companies is of great significance in the face of voluminous and ever-improving data and increasingly perfect machine learning algorithms.

Leo (2012) proposed the use of gradient boosting (GB) for more effective adaptation to different distribution models (e.g., Gaussian, Poisson, and Bernoulli distributions). For simple problems, GB usually provides better data retention and maintains dimensionality. GB is often employed when dealing with cost-loss models. The author conducted empirical research on auto insurance loss data provided by large Canadian insurance companies and compared the analysis results with those obtained using a traditional GLM.

Guelman et al. (2012) introduced the GB algorithm and used the GB tree (GBT) method to model the claim frequency and intensity. The methods of undersampling and cross-validation were employed to solve the problem of an imbalance in the original data, and compared with the traditional generalized linear model, the GBT model has certain substitutions in the prediction of the two aforementioned problems. The author believes that the GB algorithm is advantageous in dealing with the characteristics of auto insurance data, such as multi-categorical variables, an independent variable correlation, and nonlinear characteristics, and provides a superior anti-interference effect against missing and unclean data. The fitting and prediction of certain auto insurance data show that the GB algorithm is better than the generalized linear model. Simultaneously, the relative importance and partial correlation are used to rank the importance of the variables.

Liu (2014) used a multi-class regression tree and AdaBoost algorithm to predict the frequency of auto insurance claims and compared the results to those of generalized linear models, neural networks, and support vector machines, proving the robustness of the AdaBoost algorithm in auto insurance business prediction. Paefgen et al. (2014) applied generalized linear models, decision tree models, and neural network models under multivariate combinations to classify whether drivers are at risk of traffic accidents. The characteristic variables mainly include the driving time, road type, average speed, and mileage; among the different models, the neural network model achieved the most accurate classification. Lee et al. (2015) proposed the use of the delta boosting method to improve the modeling of insurance loss data through a boosting tree and conducted actual data analysis to show that delta boosting achieves higher prediction accuracy than the GBT method and generalized linear regression approach.

Lee and Antonio (2015) utilized generalized linear models, generalized additive models, neural networks, decision trees, and other algorithmic models to predict the frequency of auto insurance claims. They found that although the prediction accuracy of neural networks was higher than those of the other approaches, there was a problem of tail overfitting. Mzhavia (2016) also applied a neural network algorithm to the driver risk classification of car insurance data and proposed more appropriate numbers of neurons in the input, output, and hidden layers for achieving the best classification performance. Gao (2018) utilized principal component analysis and bottleneck neural network to process UBI auto insurance acceleration and, based on actual loss data, established a Poisson generalized additive model to predict the frequency of the claims. Liu et al. (2014) applied the AdaBoost algorithm to predict auto insurance claims, used actual auto insurance data for an empirical analysis, compared this approach to neural

networks and generalized linear models, and proved that AdaBoost provides superior prediction and interpretability. It was concluded that the AdaBoost algorithm can be used to predict car insurance claims. In addition, Sakthivel and Rajitha (2017) explored the application of artificial neural networks to auto insurance. Through data verification, they concluded that the neural network model was advantageous compared to the zero-inflated Poisson model and zero-inflated hurdle model, and that the neural network model was preferable to the Bayesian reliability model.

An advantage of the machine learning algorithm is that it does not rely on distribution assumptions and can improve the accuracy of insurance loss prediction to a certain extent. However, it is time-consuming, and more human intervention is required in the modeling process. It requires more information from the user, and the output results are not as interpretable as those of the generalized linear model. The objective of this research was to consider whether an epidemic affects auto insurance claims, which is not only a prediction problem. Therefore, to overcome the poor interpretability issue of machine learning models, the SHAP value method needs to be introduced. This article will describe the working principle in detail later. In addition, most existing studies are based on individual machine learning algorithms. A single algorithm may only be suitable for dealing with certain types of problems owing to its own characteristics. To integrate the characteristics of various models, specific integrated learning ideas were employed in this study to construct a new model with optimal functions.

2.3 Literature Review of Ensemble Learning

Ensemble learning aims to complete a learning task by constructing and combining multiple learners, that is, by using a combination strategy to incorporate the prediction

results of a series of individual learners and predict new examples. The main idea of ensemble learning is group decision-making, and the concept of using multiple models has been present in human society for a long time. Schapire (1990) theoretically proved that integration can promote weak learners to strong learners. He also noted that under the premise of knowing the lower limit of the correct rate of a weak learner, a proper integration method can be used to promote a weak learner to a strong learner, because the ensemble is typically more accurate than a single constituent classifier. Since the 1990s, ensemble learning methods constructed through supervised learning have been popular. Researchers in various fields have explored ensemble methods from different aspects of integration.

The base classifier in a good ensemble classification system should possess the qualities of accuracy and diversity. (Chandra & Yao, 2006) This view has also been recognized by most people engaged in ensemble learning research. From this perspective, researchers of ensemble learning have proposed many methods of generating diverse individual classifiers for ensembles with their own features. The following are three types of strategies that are currently used frequently.

- Bagging strategy

Breiman (1996) proposed a well-known bagging method (bootstrap aggregation and bagging). This method is mainly based on repeatable sampling (bootstrap sampling), with each sampling round based on a definite probability. The samples are reselected through replacement, and thus, many different sample subsets can be generated. These different sample subsets are then used to train multiple base classifiers to obtain a certain diversity of ensemble classifiers. The bagging diversity strategy is simple and

effective, and derivative methods based on this strategy have achieved good classification results, with the most representative being the Random Forest method.

- Boosting strategy

In contrast to the parallel approach of bagging, the boosting method uses a serial approach to generate each base classifier; that is, the data subset for training the nth base classifier is determined by the classification performance of the first n–1 base classifiers. The premise of the boosting algorithm is to understand the lower bound of the correct rate of the weak classifier algorithm, which is difficult to achieve in practical applications. Considering this problem, Freund and Schapire (1997) proposed the classic AdaBoost algorithm. This algorithm is simple to use and highly effective; thus, it has been widely employed in the field of machine learning. The principle of the AdaBoost algorithm involves firstly assigning an initial weight (usually averaged) to each sample in the training dataset, then using the learning algorithm to train and obtain the first base classifier, and finally correcting it according to the misclassification of the training dataset. The weight of each sample is adjusted specifically to increase the weight ratio of the misclassified samples. The purpose is to classify the previously misclassified samples as correctly as possible in the next round of training. The above steps are iterated until all the samples are correctly divided or a threshold is reached.

- Stacking strategy

Stacking is a type of hierarchical integration framework that is often used with another learner as an integration method to relearn the outputs of individual learners. (Wang et al., 2019) In the stacking algorithm, individual learners are also called base learners. The learners utilized for integration were called meta-learners. Many variants

of stacking have been developed. Using the output class probability of base learners as the input for the meta-learners, empirical evidence has shown that multi-response linear regression as a meta-learner can learn the class probabilities extremely well. SCANN, a variant of the stacking method, employs correspondence analysis to find the correlation between base classifiers and eliminates these correlations by transforming the original meta-layer feature space (class prediction) to generate a new feature space. (Cui et al., 2021) A new meta-layer learning method can be used to combine classifiers. Meta decision trees utilize the characteristics of the probability distribution (such as entropy and maximum probability) predicted by the base classifier as the attributes of the meta-layer, rather than the probability distribution itself. (Hamori et al., 2018) As the stacking algorithm can construct a multi-layered set of individual classifiers, it has strong moldability and can build corresponding stacking algorithms according to specific classification problems; therefore, it is widely used in various practical tasks, such as identifying named entities and building sentiment classification models. In addition, the XGBTree model is employed along with the concept of stacking multi-model fusion to construct a two-level stacking algorithm framework for user portraits and to classify the CIFAR-10 image data to obtain better classification results. Evolving non-linear stacking ensembles are also utilized in GOPlayer data, and the literature has established an automatic stacking noise reduction coding classifier for image classification. Based on network intrusion data, the stacking algorithm has been improved in terms of the generation, selection, and combination of individual classifiers and effectively employs user-generated content to identify potential users.

3. Data Analysis Techniques

The model built in this research, like the predictive model employed in the auto insurance industry, consists of three modules: a data preprocessing module, feature selection module, and module for predicting unseen data based on historical data. It runs actual data from a Chinese auto insurance company dated January 1, 2016, to December 31, 2020. SHAP is used to perform feature impact analysis according to the size and

plus or minus characteristics and to conduct model fusion via ensemble learning to achieve better performance.

3.1 Data Preprocessing

The auto insurance data adopted in this study included 3,009,723 samples and 12 variables, as listed in Table 1.

Insurance data
Residential area
Policy number
Insurance type
Effective date
Expiration date
Premium (tax included)
Premium (tax excluded)
Number of claims made in the past
Claim date
Claim amount
Car brand
Car model

Table 1. Insurance variables

Among the variables, the residential area and policy number are customer information and are not relevant to prediction; thus, these were removed from the predictive model. The variable called “insurance type” comprises five different insurance types. The sample sizes of the DZA, DAA, DAT, DAA, and DGC insurance types were 1,517,908, 1,427,980, 63,820, 10 and 3, respectively. The effective and expiration dates refer to the start and end dates of the insurance. The premium comprises tax-included and tax-excluded situations. Figure 2 illustrates the premium distribution (tax included). It can be observed in Figure 2 that the premium (tax included) has a long-tailed distribution with the premium mostly in the lower range, corresponding to actual circumstances in the industry. From the number of claims made in the past, it can be deduced from Figure 3 that the DAT insurance type is the most likely to have claims. The claim date is the date on which the insurance is registered. Despite the small number of unusually large values, taking the logarithm of the claim amount produces a nearly normal distribution, as shown in Figure 4. Extremely large claim amounts affect the overall distribution and machine learning models that depend on the data distribution, such as the linear regression model.

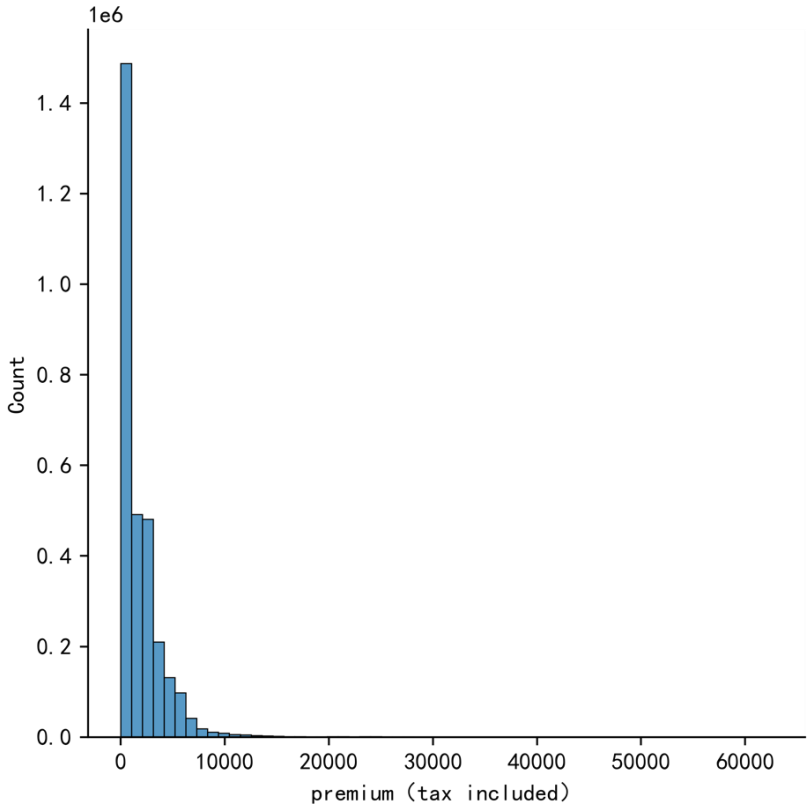

Figure 2. Premium distribution (tax included)

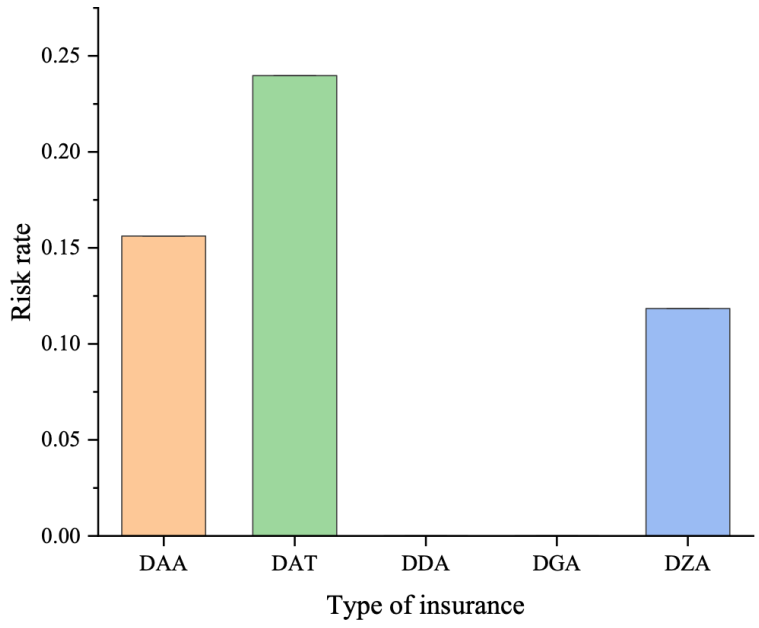

Figure 3. Claim probabilities for different insurance types

As shown in Figures 5 and 6, Shanghai Volkswagen and SAIC-GM-Buick are the two most insured car makers, whereas Excelle and Lavida are the two most insured car models.

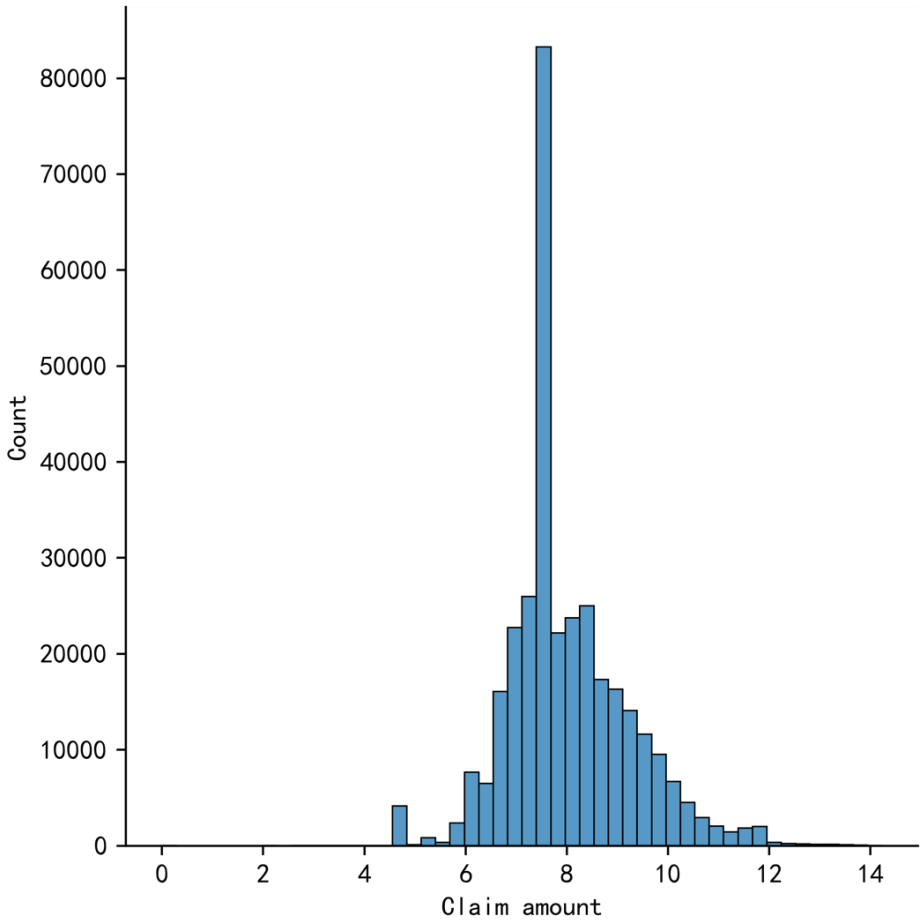

Figure 4. Claim amount distribution

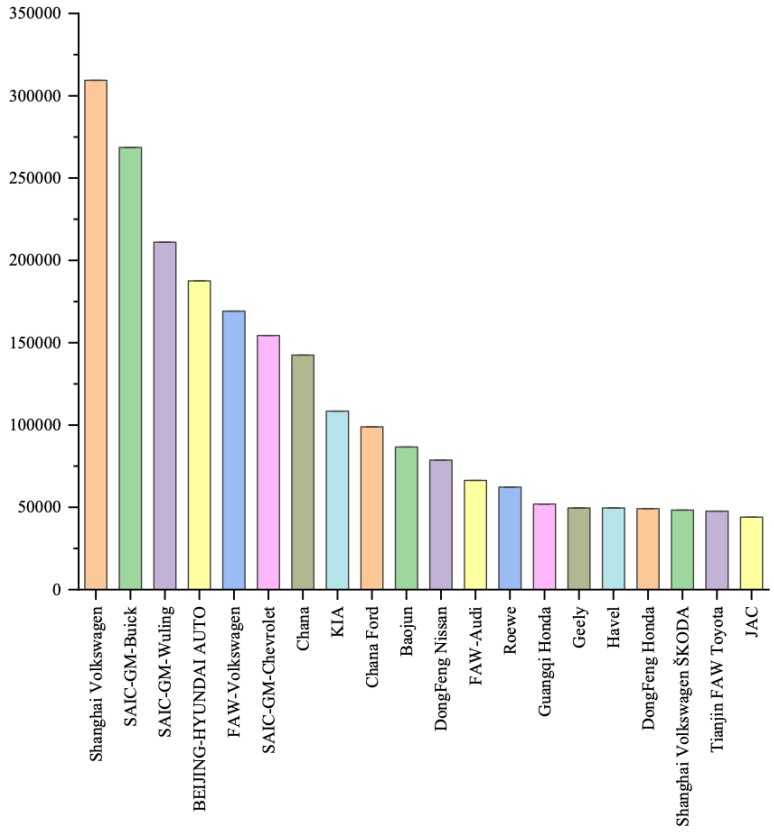

Figure 5. Insurance distribution for different car brands

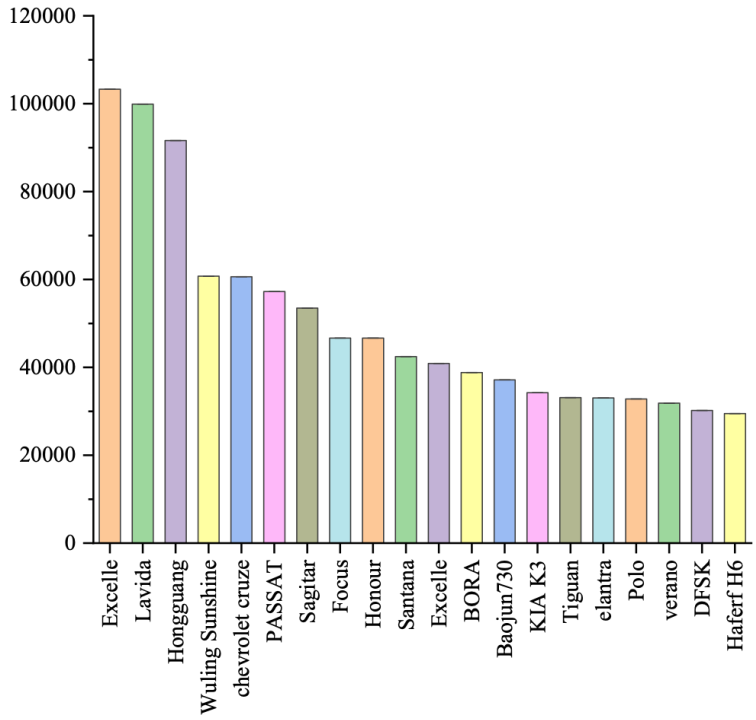

Figure 6. Insurance distribution for different car models

Figure 7 details the monthly changes in the number of insurance purchases and reveals that insurance purchases are in fact periodic: the most sales occur at the beginning and end of the year. A steady increase in insurance purchases can be observed from 2016 to 2020; however, the increases in 2019 and 2020 are smaller than those in 2016, 2017, and 2018. The data corresponding to 2019 and 2020 are also provided in a separate graph for deeper understanding of the changes during the pandemic period. Figure 2 shows that there were fewer insurance purchases in 2020 than in 2019 and that the pandemic did not result in fewer insurance purchases. It cannot be determined from the available information whether the pandemic affected insurance purchases.

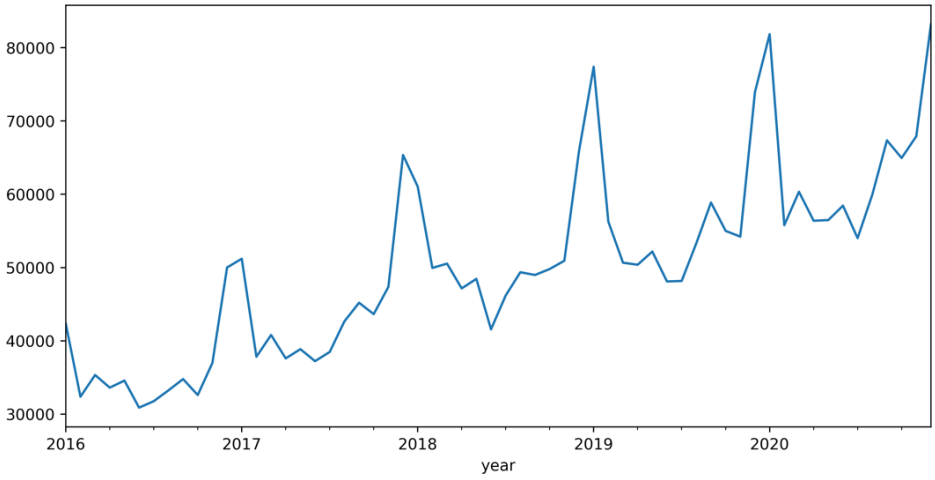

Figure 7. Changes in number of insured over time

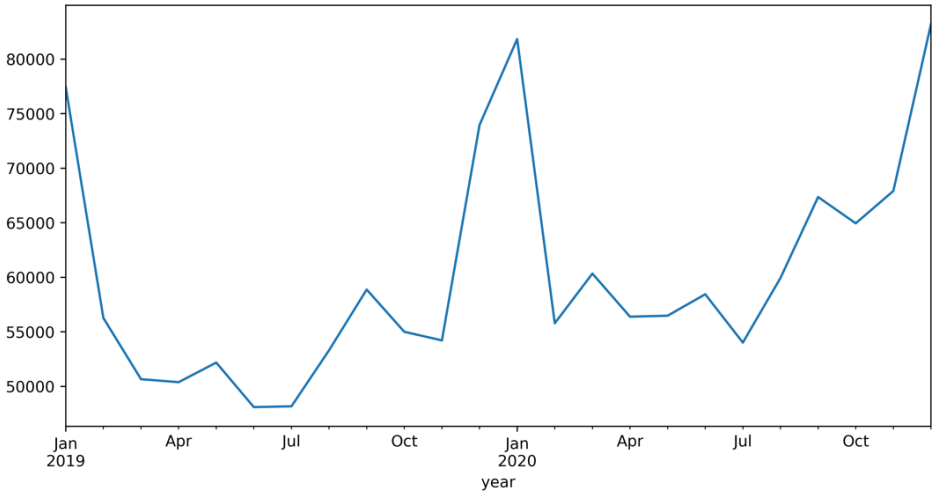

Figure 8. Changes in number of insured over time in 2019 and 2020

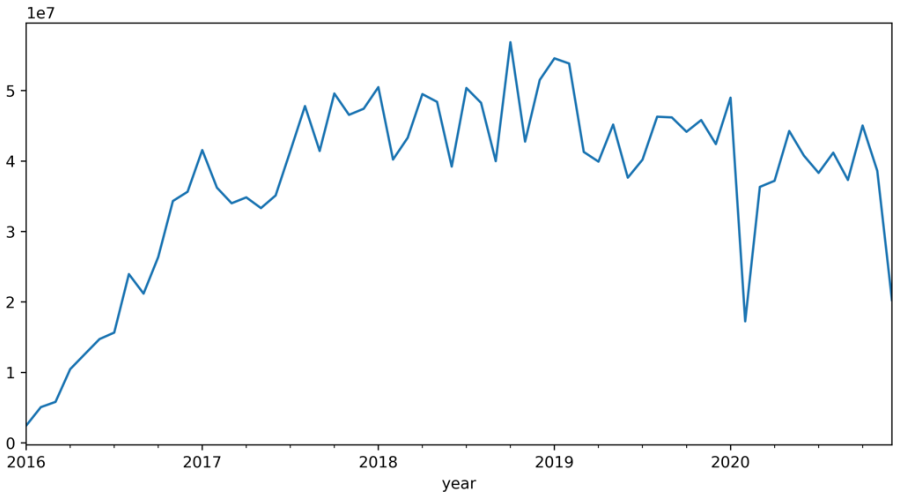

Figure 9. Changes in amount of compensation over time

Figure 5 illustrates the changes in claim amounts from 2016 to 2020. The claim amounts are relatively stable from 2017 to 2019 but decline in 2020. To investigate the changes in the 2019 and 2020 claim amounts further, the data are visualized in a separate graph. The following figure reveals a drastic decrease in 2020 claim amounts in comparison with 2019 and the previous years. This finding supports the hypothesis that the pandemic impacted the claim amounts.

Analysis of the three insurance variants suggests that it is constructive to examine the effects of the pandemic on the claim amounts and quantify such results with machine learning.

The pandemic-related Baidu index, i.e., the search rate on the most popular Chinese search engine Baidu, was also collected in this study. This quantity represents the interests of mainstream Internet users, as the search engine accounts for 81.26% of the PC market share and 80.62% of the mobile market share. Search rates for pandemic-related

keywords not only reveal changes in user needs and media trends, but also enable market analysis from an industrial perspective.

Table 2 shows the six attributes of the Baidu index. The pandemic search rate in China is the Baidu index for pandemic-related keywords in China, whereas the pandemic search rate in Jiangsu is the Baidu index for pandemic-related keywords in the city. The Baidu search index is based on the search rate of Internet users on Baidu, with keywords as the statistical objects, and scientifically analyzes and calculates the weighted sum of the search frequency of each keyword in the Baidu web search. According to different search sources, the search index is divided into PC and mobile search indices. In this research, the total PC and mobile search indices were used as variables. Pandemic news in China refers to all Baidu media search indices in China, whereas pandemic news in Jiangsu corresponds to all Baidu media search indices in the city.

The media index is based on the number of pandemic-related news reports by major Internet media collected by the Baidu news channel. The data source and calculation method are not directly related to the search index. Pandemic vaccine news in China represents all the Baidu information search indices in China, whereas pandemic vaccine news in Jiangsu represents all the Baidu information search indices in Jiangsu, China. Finally, the Baidu news search index is based on the Baidu intelligent distribution and recommended content data, whereas the information index is obtained by the weighted summation of the number of the reading, commenting, forwarding, liking, disliking, and other behaviors of netizens.

We preprocessed the data based on the observation and analysis of the dataset described above. The main purpose of data preprocessing was to organize and complete

the existing data. Specifically, the data were inspected and validated to remove duplicates, existing errors were fixed, and the reliability was checked. The data preprocessing included the following major components.

Firstly, duplicate data were eliminated and missing data were filled in to ensure that the data were complete. Thus, it was necessary to eliminate outliers. Based on specific scenarios, criteria were established to detect outliers. Inspecting the dataset, eliminating outliers, and improving the quality of the dataset ensured that the data were within a reasonable range.

In addition, the pandemic Baidu index data adopted in the current study was the search rate of the term "COVID-19" on Baidu, and they emanate from the Baidu index. Baidu is a major search engine and represents the degree of attention of the Chinese people. The collected Baidu index comprises six attributes, as presented in Table 2.

Pandemic data
Pandemic search rate in Jiangsu
Pandemic search rate in China
Pandemic news in Jiangsu
Pandemic news in China
Pandemic vaccine news in Jiangsu
Pandemic vaccine news in China

Table 2. Pandemic attributes

Considering the pandemic search rate in Jiangsu variable as an example, a variation graph of the daily amount of claims with predicted tags from January 1, 2020, to December 31, 2020 was constructed. The data were normalized to make them comparable. In a multi-indicator system, based on the nature of indicators, there are usually different units and data levels. When there are significant differences between the data levels of various indicators, adopting the original indicator values for analysis would strengthen the importance of indicators with higher numerical values and weaken the importance of those with lower numerical values. Therefore, it was necessary to normalize the original indicator data to ensure the reliability of the results. Normalization can eliminate the unit limitation of data and convert the data into unitless values, to enable the comparison and weighing between indicators with different units and data levels. This is consistent with the previous inference that the Baidu search index of the epidemic keyword is highly correlated with the amount of claims. The same phenomenon can be observed from Figures 8 and 9, which further demonstrates the proposed perspective.

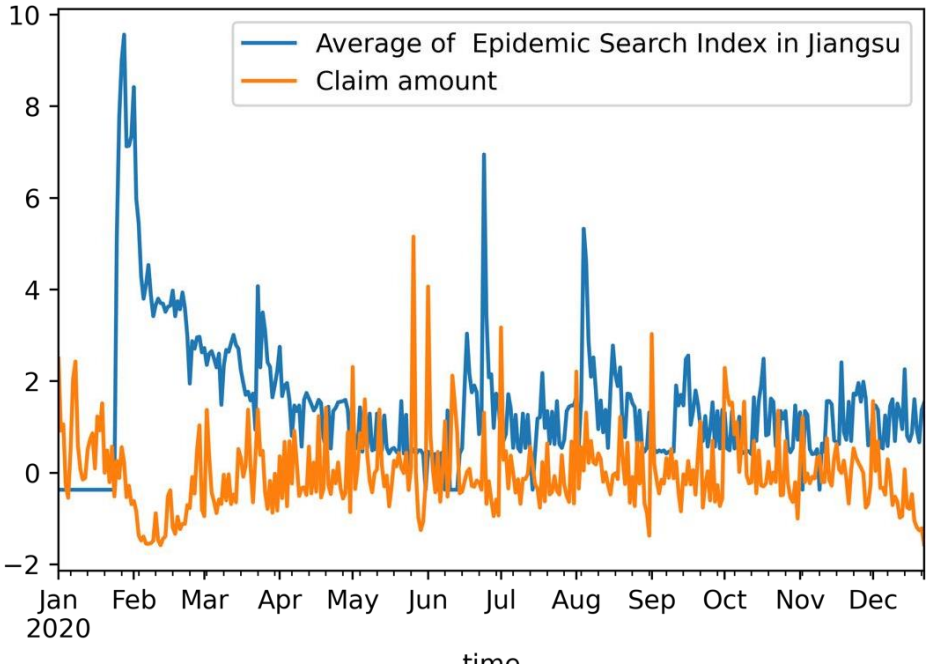

Figure 10. Jiangsu pandemic index and number of claims

As illustrated in Figure 7, from late January to late February 2020, the pandemic search index in Jiangsu is very high, and the total number of claims is comparatively low in the same year. More careful investigation revealed that the Jiangsu pandemic search index and the total amount of claims are negatively correlated; when the Jiangsu pandemic search index is high, the total number of claims is comparatively low.

3.2 Feature Selection

Feature selection is the process of selecting relevant feature subsets from a given feature set and comprises a feature subset search and feature subset evaluation. M sub-features were selected from N features (M < N) to achieve the optimal criterion function. The goals of feature selection are to select as few sub-features as possible to sustain the model performance and to ensure that the categorydistribution of the results is

as close to the actual distribution as possible.

Besides auto insurance data, the pandemic Baidu index data adopted in the current study included the search rate of the term “COVID-19” on Baidu and originated from the Baidu index. The Baidu index is a major search engine, and its index represents the degree of attention of the Chinese people. The collected Baidu index comprises six attributes, as illustrated in Table 3 below.

Considering the pandemic search rate in Jiangsu as an example, a variation graph of the daily amount of claims with predicted tags from January 1, 2020, to December 30, 2020 was created. The data were normalized to make them comparable. It is evident that from late January to late February 2020, the pandemic search index in Jiangsu is very high, and the total number of claims is comparatively low in the same year. More careful investigation revealed that the Jiangsu pandemic search index and total amount of claims are negatively correlated; when the Jiangsu pandemic search index is high, the total amount of claims is comparatively low.

Pandemic data
Pandemic search rate in Jiangsu
Pandemic search rate in China
Pandemic news in Jiangsu
Pandemic news in China
Pandemic vaccine news in Jiangsu

Table 3. Pandemic attributes

The purpose of using the feature selection technique is that in reality, the number of task data is large and the amount and dimension of features are both very high; hence, dimensionality problem often occurs, and choosing important features from among the features with several attributes will allow later learning tasks to be conducted solely on a few representative features. Meanwhile, some features irrelevant to the task can be removed by feature selection, shifting the focus of the learning task to the relevant features only. This approach reduce the complexity of the learning task significantly.

3.2.1 Feature Selection Process

The first part of feature selection is the feature subset search, meaning that candidate feature subsets are selected from a given feature set. By analyzing the dataset, the features listed in Table 4 were generated and selected.

Features
Number of claims = 7
Number of claims = 6
Number of claims = 5
Number of claims = 4
Number of claims = 3
Number of claims = 2

Number of claims = 1

Amount of claims = 7

Amount of claims = 6

Amount of claims = 5

Amount of claims = 4

Amount of claims = 3

Amount of claims = 2

Amount of claims = 1

Month

Is a holiday

Year 2016

Year 2017

Year 2018

Year 2019

Year 2020

Weekly mean number of claims

Weekly mean amount of claims

Mean number of claims in the past month

Mean amount of claims in the past month

Total amount of insurance

Number of registered insurances

Whether there is a pandemic

Jiangsu pandemic search index

Pandemic search index

Jiangsu pandemic news

Pandemic news

Pandemic information

Jiangsu pandemic search index in the past week

Mean pandemic search index in the past week

Mean Jiangsu pandemic news in the past week

Mean pandemic news in the past week
Mean pandemic information in the past week

Table 4. Model feature generation and selection results

Candidate features were selected from the entire candidate set gradually until the required number of features is reached. Therefore, two search methods were adopted to complete the search for feature subsets: forward and backward. A forward search start from an empty set and gradually add relevant features until the number of features meets the requirement, whereas a backward search starts from a complete set of features and eliminates irrelevant features gradually until the number of features meets the requirement.

The second part of feature selection is the subset evaluation process, meaning that the effectiveness of feature selection is evaluated based on the dataset and feature subsets. In actual tasks, the effectiveness of attribute set A is usually evaluated by the information gain of the attribute subsets:

$$Gain(A) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v).$$

In this formula, $\{D^1, \ldots, D^V\}$ are V subsets of the dataset divided based on the attribute subsets. The information entropy is defined as follows:

$$Ent(D^v) = -\sum_k p_k \log_2 p_k.$$

Therefore, the information gain can be utilized as the evaluation criterion.

Feature selection reduces the number of input variables and the dimensionality of the model and can improve the model performance. There are three types of feature selection mechanisms: filter, wrapper, and embedded mechanisms. The following section focuses on the filter method.

3.2.2 Filter Feature Selection

The filter method first passes through feature selection in the dataset, filters the initial features with feature selection techniques, and then completes the learning task using filtered features.

In the current study, the variances and correlation coefficients were adopted to filter the generated features. Using a threshold value of 0.1 for the variances and 0.005 for the correlation coefficients, 38 features were selected, as presented in Table 3.

A one-way analysis of variance tests whether there is any significant difference between the group means among multiple groups that are influenced by a single factor. In the current study, the Pearson correlation coefficient was used to measure the linear relations between variables. The Pearson correlation coefficient is commonly adopted in feature selection because it is fast and easy to compute. The range of the coefficient was [–1, 1]. In the current study, different feature selection techniques were adopted. In this study, two different feature selection techniques were used for feature selection.

3.3 Regression Predictive Models

To investigate the economic impact of the pandemic on the insurance industry, a comparative analysis was conducted using regression predictive models. Specifically, a similar predictive model was trained with data from 2016 to 2017, 2017 to 2018, and

2018 to 2019 and tested the predicted outcomes using the data from 2018, 2019, and 2020. Table 5 lists the results. From 2016 to 2019, the variation in the economic trend is insignificant; hence, the error caused by using data from the first two years to forecast for the following year is negligible. However, the error is significant when using data from 2018 and 2019 to forecast the situation in 2020. The influence of the pandemic renders the forecast based on the pre-pandemic data inaccurate. Based on the analysis above, multiple regression predictive models were investigated and tested and the proposed method was developed considering the influence of the pandemic.

		MAPE	Mean squared log error
2016–2017	predicting 2018	11.741	0.019
2017–2018	predicting 2019	13.076	0.021
2018–2019	predicting 2020	19.286	0.050

Table 5. Predictive model test

3.3.1 Ridge

Owing to linear dependence, linear regression was unfit and required regularization. Hence, linear regression via ridge regularization was adopted.

First, the ordinary least square regression formula is as follows:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}.$$

In this formula,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

The ridge regression adds a penalty term to the minimized loss function. L2 regularization takes the root sum square of all the elements in the weight vector. The loss function after regularization can be expressed as follows:

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} (\theta^T x_i - y_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2.$$

L2 regularization assumes the prior distributions of the parameters to be Gaussian distributions, which ensures model stability, meaning that the values of the parameters do not become too large or too small. The L2 norm is the root sum square of all parameters. To minimize the L2 norm, every element of W can be made very small, approximately 0. Unlike the L1 norm, the L2 norm only makes elements approximately 0, but not equal to 0. Smaller parameters indicate that the model is simple and not likely to overfit. Therefore, ridge regression is least-squares regression with an L2 norm penalty. The estimated target of ridge regression is called a shrinkage estimator.

3.3.2 LASSO

The ridge regression model is introduced above. However, if the input features have large dimensions and a sparse linear relationship, ridge regression is inappropriate, and LASSO regression must be considered.

The L1 norm assumes a Laplace distribution for the parameter prior distributions and ensures the sparsity of the model. Based on the least squares method, L1 regularization was adopted. The difference between L1 and L2 regularization is that they utilize different penalty terms. L1 regularization sums the absolute values of all elements in the weight vector W. The loss function after regularization can be expressed as follows:

$$\min_{\theta} \frac{1}{2m} \sum_{i=1} (\theta^T x_i - y_i)^2 + \lambda \sum_{j=1}^{n} \| \theta \|.$$

L1 regularization is featured by adding absolute values, which make the coefficients of a few features smaller, or even 0, for some coefficients with smaller absolute values. This strengthens the generalizability of the model. For high-dimensional feature data, especially those with sparse linear relationships, LASSO regression is adopted. LASSO regression is also very efficient for finding major features among several features.

3.3.3 ElasticNet

LASSO may obtain an overly sparse model and make the model difficult to fit because it screens out excessive variables. To solve the overly sparse issue with LASSO, Zou et al. proposed the ElasticNet method combining LASSO and ridge regression, while introducing ℓ_1 and ℓ_2 penalty functions. The equation is as follows:

$$\min \frac{1}{2m} a [\sum_{0} \sum_{j=1} (\theta^T x_i - y_i)^2 + \sum_{j=1}^{n} \| \theta \| + \lambda \sum_{2} \sum_{j=1}^{n} \| \theta \|^2].$$

Making $\alpha = \frac{\lambda_1+\lambda_2}{\lambda_1}$ 又 $\lambda = \lambda_1 + \lambda_2$, the above equation can be re-written as

$$min \frac{1}{2m} [\sum_{j=1}^{n} (\theta^T x^{(i)} - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \| \theta \| + (1 - \alpha \sum_{j=1}^{n} \| \theta \|_2^2)].$$

Evidently, ElasticNet has a penalty function of $\alpha \sum_{j=1}^{n} \| \theta \| + (1 - \alpha \sum_{j=1}^{n} \| \theta \|_2^2)$. It is easy to tell that in this case, the penalty function of ElasticNet solely retains the penalty function of ridge regression. Therefore, ElasticNet turns into ridge regression solely with compressed coefficients. When $\alpha = 1$, the penalty function of ElasticNet retains the penalty function of LASSO and becomes LASSO regression. When $0 < \alpha < 1$, ElasticNet retains both penalty functions from the ridge and LASSO regressions and has the characteristics of both. It compresses a few coefficients at a certain ratio and compresses the other coefficients to 0. Therefore, ElasticNet combines the characteristics of ridge and LASSO regression, addressing the multicollinearity issue and filter factors, while avoiding the over-sparsity of variables owing to over-compression in LASSO.

3.3.4 LightGBM

In comparison with the GB decision tree (GBDT) algorithm, LightGBM makes significant improvements in several aspects, including using the complexity of the decision tree as the regularizer when optimizing the objective function of the algorithm and using second-order Taylor expansion in the optimization of the objective function.

Assuming there is a supervised dataset $X = \{(x_i, y_i)\}_{i=1}^{n}$, the purposes of the LightGBM algorithm are to determine an approximate value $\hat{f}(x)$ for a certain function $f(x)$ and to minimize its specific loss function $L(y, f(x))$. The loss function is adopted to evaluate the effectiveness of the fit of the model, which can be written as

$$\hat{f} = argmin_f E_{y,x} L(y, f(x)).$$

The LightGBM model integrates k regression trees to fit the final model, as follows:

$$f_k(x) = \sum_{i=1}^{n} f_t(X).$$

A regression tree can be represented as w_{qx}, $q \in \{1,2, \ldots, J\}$, where w refers to the sample weight vector of the leaf nodes, q is the structure of the regression tree, and J is the number of leaves on the tree. Meanwhile, all the information from the previous (t–1) trees will be utilized at the t^{th} tree. Therefore, the objective function at the t^{th} iteration can be written as (Minastireanu & Mesnita, 2019):

$$\Gamma_t = \sum_{i=1}^{n} [L(y_i, F_{t-1}(x_i)) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_t(x)).$$

In this function, g_i and h_i are the first- and second-order gradient statistics of the loss function, respectively. As the regression tree is defined above, the complexity of a tree is

$$\Omega(f_t(x)) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2,$$

where J is the number of leaf nodes, γ is the coefficient of leaf nodes, and λ is the coefficient of L2 regularization. Therefore, the complexity of a decision tree can be considered depending on the number of leaf nodes and the following L2 norms. Assuming that $I_j = \{i|q(x_i) = j\}$ is the sample set assigned to the leaf nodes, the objective function can be rewritten as

$$\Gamma_t = \sum_{j=1}^{J} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma J.$$

For a specific tree structure $q(x)$, the optimized weighting score for each leaf node is

$$\Gamma_t = -\frac{1}{2} \sum_{j=1}^{J} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma J.$$

When the method of calculating the objective function is fixed, the optimization minimizes the objective functions for each tree. Therefore, it is necessary to calculate the gain of the leaf node splitting in trees, maximize the gain of node splitting, and choose the feature with the highest gain as the splitting feature. This process must be repeated until all conditions are satisfied. Assuming that LR I II $I = I_L \cup I_R$ is the sample set of parents, I_L and I_R are sample sets of left and right branches. Then, the gain of each node split is

$$G = -\frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_j \right)^2}{\sum_{i \in I_L} h_j + \lambda} + \frac{\left(\sum_{i \in I_R} g_j \right)^2}{\sum_{i \in I_R} h_j + \lambda} + \frac{\left(\sum_{i \in I_j} g_j \right)^2}{\sum_{i \in I} h_j + \lambda} \right].$$

Based on this, LightGBM will continuously perform deep-level optimizations, targeted at the increasingly large training data and higher data feature dimensions in this big data

era by using methods such as histogram algorithm, leaf-wise tree growth strategy, and

histogram subtraction for further acceleration. Using these techniques can significantly reduce the complexity of the algorithm and training time, thus improving the training efficiency and accuracy.

3.3.5 SHAP Value

Machine learning is developing at a phenomenal rate in the industry and permeating daily life. Complex models such as integrated models and deep neural networks have a wide range of practical applications, from recommending Douyin videos and the Google neural network-based translation system to the Xiaomi voice assistant. Despite the success of these applications, they have limitations and shortcomings. The lack of transparency prevents users from understanding the reasoning behind certain decisions. For example, a vehicle equipped with an advanced autopilot system brakes when passing by an ambulance parked on the side of a road with no siren. This unexpected behavior confuses the user, who will be curious about the reason for this behavior. Therefore, increasingly many industries and academics are becoming interested in explaining and conducting research on machine learning models.

During the training process for the auto insurance predictive model, in addition to optimizing the model performance, it is also extremely important to know how models work and what features play crucial roles in the actual process. Meanwhile, explainable models will enable the users of these models to trust and familiarize themselves with the decision process, as well as to be informed about the impacts of the pandemic on the auto insurance and the intensity of these impacts.

Compared to the logistic regression that assigns a weight to each modeled variable,

machine learning usually produces better outcomes than conventional scorecard models. Methodologically, training and deployment in machine learning can be recognized as being encapsulated in a black box: variables are input, and predicted probability values are obtained. Therefore, a balance point should be determined to maintain the optimal performance of machine learning and better explain the models. While completing the training with the machine learning models adopted in the current study, it was attempted to analyze the variables with important roles further and to determine whether they influenced the prediction outcomes. It was attempted to elucidate these factors through variable importance combined with the SHAP value.

Explanatory machine learning technology can be divided into two categories: model-specific and model-agnostic technology. Model-specific interpretability requires the construction of a self-explanatory model to explain its structure. These types of models include rule-based decision trees, linear models, logistic regression, and naive Bayes models. In contrast, model-agnostic models are applicable to any machine learning model after training and are mainly explained by input and output analysis. A significant difference between the two technologies lies in the trade-off between the model accuracy and the fidelity of interpretation. Model-specific models produce accurate explanations at the expense of performance, whereas model-agnostic models have limited interpretability but high accuracy.

Local interpretable model-agnostic explanations (LIMEs) replace objective models with interpretable and simple models such as decision trees and linear regression in local areas. LIMEs only add a slight disturbance to the input data to examine how the output of the model changes and train a simple interpretable model on this basis without delving into the interior of the model. The corresponding mathematical expression is as

follows:

$$explanation(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g). \quad (2.9)$$

Here, x represents a sample, g is an interpretable model based on x, f refers to the original model, $\Omega(g)$ is the complexity of interpretable model g, G indicates the ensemble of all interpretable models, and π_x stands for the range of x. LIMEs minimize model g and the loss function of original model f in π_x. One drawback of LIMEs is that they need to be confirmed and are sensitive to range; different ranges produce different interpretable models.

SHAP calculates the Shapley value through cooperative game theory to determine the contribution of features to predictions. The model generated a predictive SHAP value for each sample. Let the average prediction of the entire data set be the base of the whole model. The SHAP value for the jth feature in the ith sample is represented by formula 2.10:

$$y_i = y_{\text{base}} + f(x_{i,1}) + f(x_{i,2}) + \ldots + f(x_{i,k}). \quad (2.10)$$

The SHAP value is an additive explanatory model. If it is greater than 0, the feature increases the prediction and results in a positive impact. If it is smaller than 0, the feature decreases the prediction and results in a negative impact. The greatest advantage of the SHAP value is that it reflects the influence of the features of each sample on the prediction and averages the influence of each sample to obtain the overall feature importance.

3.4 Ensemble Learning

Figure 2-1 shows a structural diagram of the ensemble learning. The base models were products of training from machine learning algorithms, such as the logistic regression algorithm, decision tree algorithm, and LASSO regression algorithm, in the original dataset. They were trained and combined using ensemble methods to obtain the final output. If the ensemble algorithm contains only base models of the same type, for example, if the logistic regression ensemble consists only of logistic regression models or the decision tree ensemble contains only decision trees, the ensemble is described as homogeneous. Ensembles can also be composed of different types of base models simultaneously, such as support vector machines and naive Bayes classifiers. Such an ensemble is heterogeneous.

Ensemble learning outperforms the single-base model in generalization by combining multiple base models, particularly weak learners, which generalize only slightly better than random guessing. Many researchers choose ensemble learning algorithms for stronger learners to achieve better results; likewise, a strong learner was selected as the base model in this study.

Combination base models depend on specific problems; common combination strategies include averaging, voting, and learning methods. The averaging methods are is divided into simple and weighted average methods, and the voting methods are divided into weighted, relative majority, and absolute majority voting methods. For continuous label R in the regression problem, averaging is preferred; for discrete label Z in the classification problem, voting is preferred.

3.5 Stacking

Compared to using only one model for learning and prediction, stacking and blending involves combining multiple models in a certain manner to build a multi-model system and achieve better learning outcomes. Currently, applications based on weight combination and bagging methods are more common than applications based on stacking. Stacking involves a leveled model integration framework. Its basic idea is to adopt an individual prediction model as the meta-model and to connect it to a predictive model. It utilizes the outputs of the meta-model as inputs of the next model and the outputs of this model as the results. To ensure the efficiency of model integration, meta-models need to complement each other; hence, their structures and parameters cannot be completely similar.

The stacking framework firstly divides the original dataset into multiple subsets and inputs the subsets into base learners at the first level of the predictive model. The base learners output their prediction results. Then, the outputs from level 1 are adopted as inputs into the second level to train meta-learners at level 2. The models in level 2 output the final prediction outcomes. The stacking framework achieves improved accuracy by generalizing the output results from multiple models, as illustrated in Figure 11.

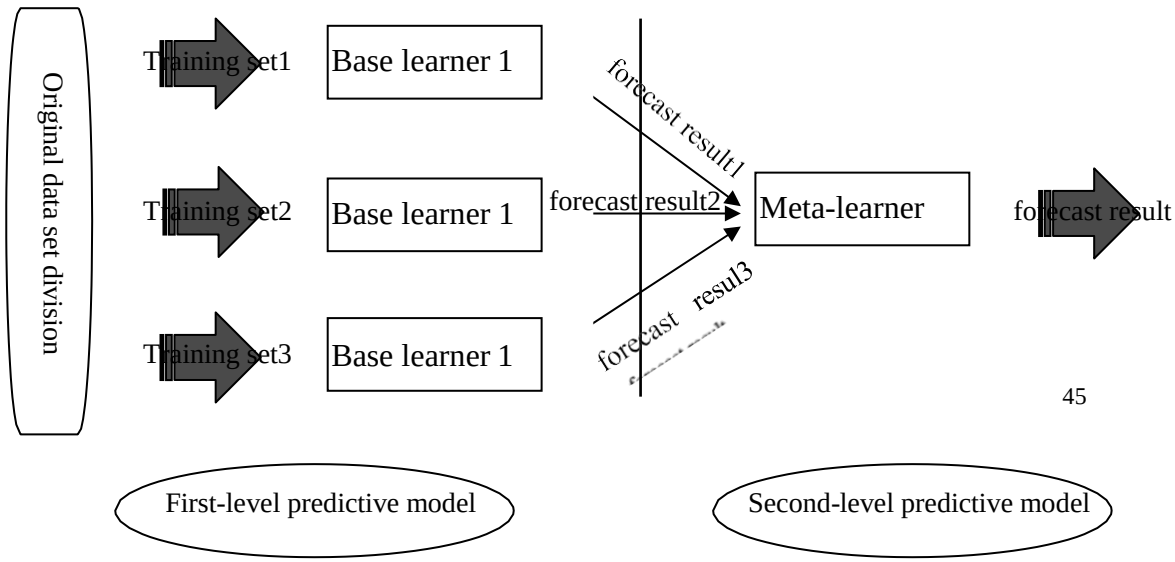

Figure 11. Stacking framework

4. Proposed Method

The claim amount prediction model includes three sections. Firstly, data pre-analysis and pre-processing, including outlier processing and standardization, are performed. Next, feature engineering technology, specifically, feature generation and feature selection were applied and features are extracted from the original data in full measure for model use. Finally, the processed data are passed to the insurance claims prediction model for training, and the trained model completes the prediction tasks. The overall model framework is shown in Fig. 9. This chapter introduces various components of the insurance claim prediction model.

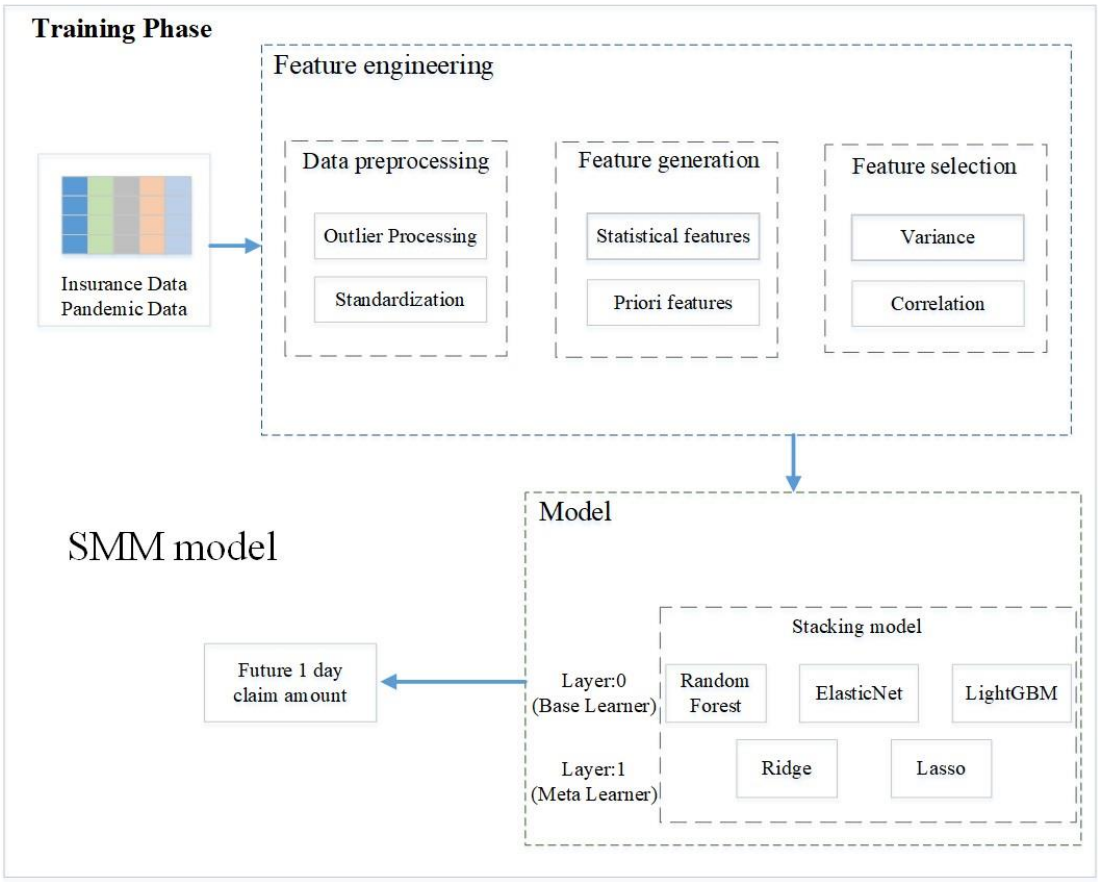

Figure 12. Insurance prediction model

4.1 Data Preprocessing

4.1.1 Treatment of Outliers

There are numerous days in the data in which the claim amounts sharply increase to the point that they much higher than those in the past few days. This type of data is difficult for prediction models to manage because of the steepness of its changes. The threshold σ is set to ensure the stability and continuity of the time series. If the absolute value of the difference between the claim amount of the day and the average claim amount of the past week is greater than σ, it is considered an outlier. In the entire dataset, data with outliers were filtered for data modeling.

The graph in Fig. 13 indicates that the difference between the average lines of the predicted and past predicted values at t = 8 is minor, whereas the difference between the average lines of the predicted and past predicted values at t = 9. Therefore, t = 8 has a stronger predictability.

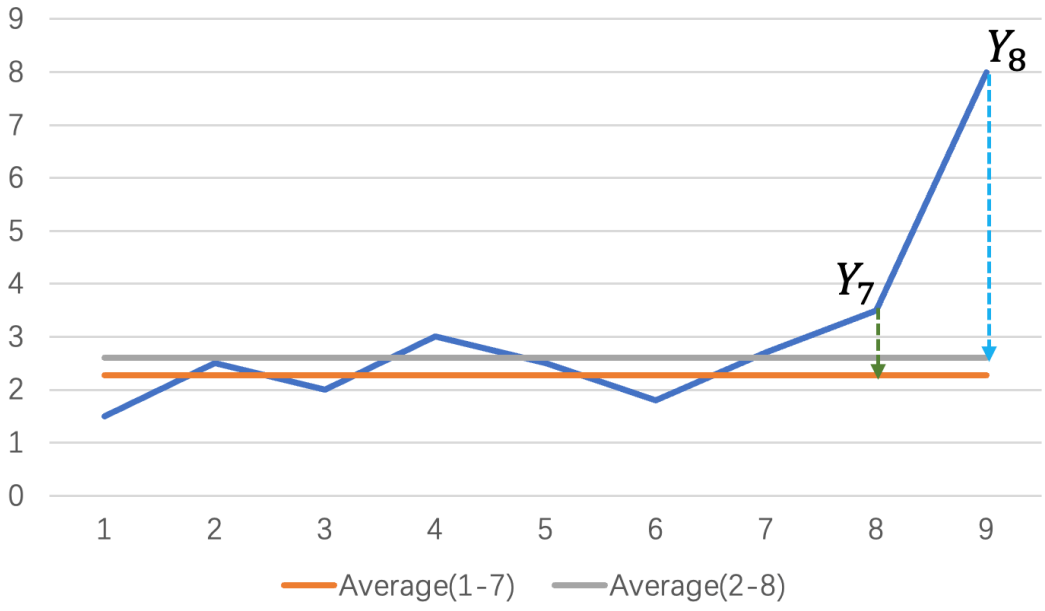

Figure 13. Differences between average lines of predicted and past predicted values

4.1.2 Data Standardization

Distance-based machine learning algorithms, such as KNN and SVM, are highly sensitive to feature values. If the value of a feature is several orders of magnitude higher than those of other features, it will completely dominate the machine learning algorithm, and the other features will be ignored. This characteristic significantly affects the predictive performance of the model. Therefore, standardization is an important process in feature engineering. Assuming that $\bar{X}$ is the mean of X, σ is the standard deviation of X, and its calculation is conducted on each column, that is, each feature is calculated independently using the following formula:

$$X' = \frac{X - \bar{X}}{\sigma}. \qquad (4.1)$$

As shown in Fig. 14, the average can be subtracted from the original data to obtain zero-centered data, and formula 4.1 can be applied using its standard deviation σ. Normalized data can be derived by dividing the zero-centered data by σ. The x- and y-axis scales are the same for all dimensions. Each dimension is then scaled to the same order of magnitude to prevent possible offsets in the machine learning model training.

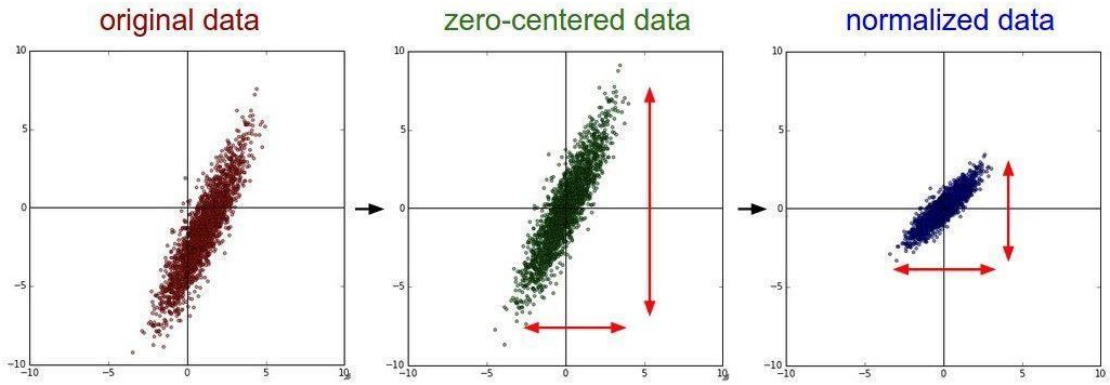

Figure 14 Data standardization

Each dimension is then scaled to the same order of magnitude to avoid the offset that occurs in the machine learning model training.

4.2 Feature Engineering

4.2.1 Feature Generation

In the insurance data set collected in this study, each piece of data represents the insuring behavior of a customer. If there is a claim, the claim date and amount are included. To build a machine learning model, a feature variable X and label variable y are required. In this study, historical insurance data and Baidu index data were used to predict the future claim amount y. The original dataset is summed up according to the number of days required to obtain the daily total claim amount label.

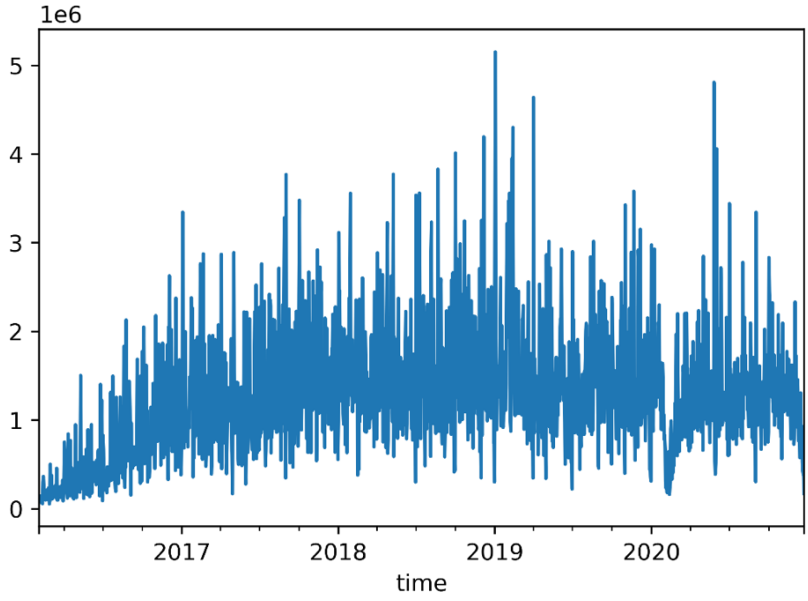

Figure 15. Daily total claim amount

Figure 10 shows the changes in the daily total claim amount over time. From 2016 to 2020, there is a steady increase in the daily total claim amount. The sudden drop in the daily total claim amount from January to February 2020 is likely due to the lock-down of Suqian and the subsequent decrease in traffic flow. This finding confirms the impact of the pandemic on the claim amount.

Feature X is required to establish the claim amount prediction model. Features from insurance data and Baidu index data are generated to predict the next-day claim amount, which can be considered to be related to the number of accidents in the past because data of this type have a time-series correlation. The historical K-day claim amount and number of risks are added to the feature set. According to the existing insurance pre-mium rate determination and literature on insurance pricing, the number of insured and total insured amount are directly related to the final claim amount. The greater the num-ber of insured, the greater the increase in the probability of future claims. Historical

claim amounts reflect, to a certain extent, future claim amounts. Generally, a high traffic volume significantly affects the probability of road accidents. After the pandemic outbreak, governmental control and public fear caused profound changes in these features. To measure the impact of the pandemic on insurance buying behavior and traveling, Baidu search rate data, which quantify the public perception of the pandemic and reflect attitudes toward insurance buying and traveling to a certain extent, were employed in this study.

In addition to historical K-day insurance and the Baidu index, statistical features were also included. Statistical features are the statistical results of historical features over a period of time, such as the average value of claims over the past seven days. Compared with the one-time result, the statistical features are more stable and less prone to interference from outliers. Figure 11 outlines the features extracted from the insurance and Baidu index datasets to obtain new features as input into the feature selection step.

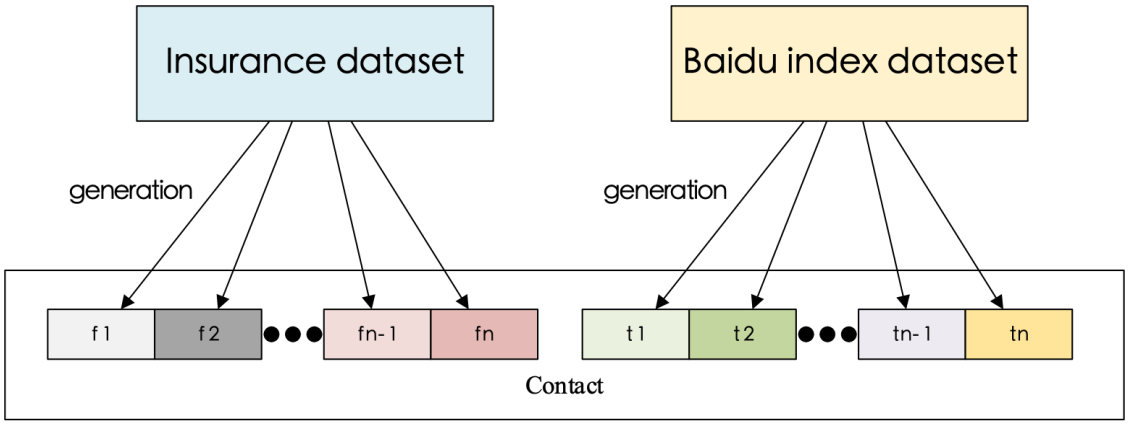

Figure 16. Feature Generation

4.2.2 Feature Selection

New features generated by feature engineering must be screened for effectiveness. Ineffective features that impair the operation of the algorithm are removed. In this study, two feature selection techniques were used in sequence: method filtering and correlation coefficient filtering. The scatter plot highlights the relationship between feature variance and its changes. When the variance of feature x1 is small, the change in y is large. For feature x2, the greater the variance, the greater the change in y. Feature x1 with smaller variance has almost no correlation with y and cannot be used to predict y. If the values of a feature are the same or nearly the same, the feature has no effect on the prediction; the variance of this feature is very small, and it needs to be filtered by variance. The Pearson coefficient is a measure of the correlation of two variables and can remove the dimensional influence of the two variables. As shown in Figure X, the Pearson coefficient between features x1 and y is very small, close to 0, whereas that between features x2 and y is large. The Pearson coefficient between the feature and the label was then calculated, and the feature whose Pearson coefficient was close to 0 was filtered. A coefficient greater than 0 suggests a positive correlation, and a coefficient less than 0 indicates a negative correlation. Using Pearson coefficients, features with little correlation can be eliminated, and the running time of the algorithm can be reduced.

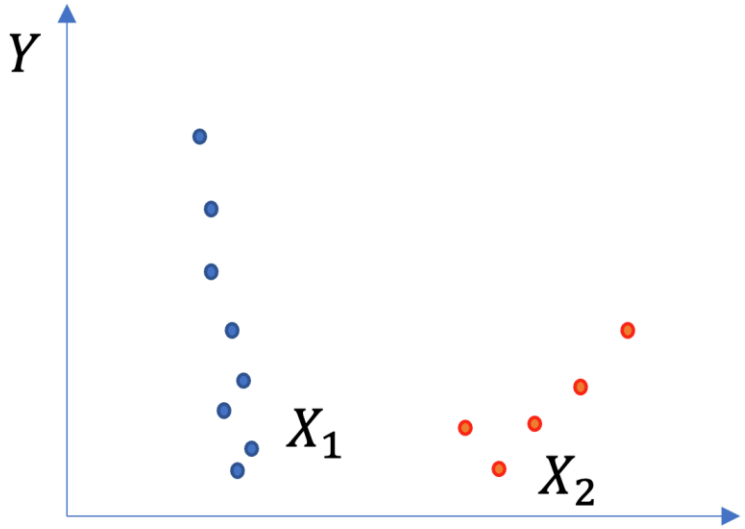

Figure 17. Variance visualization

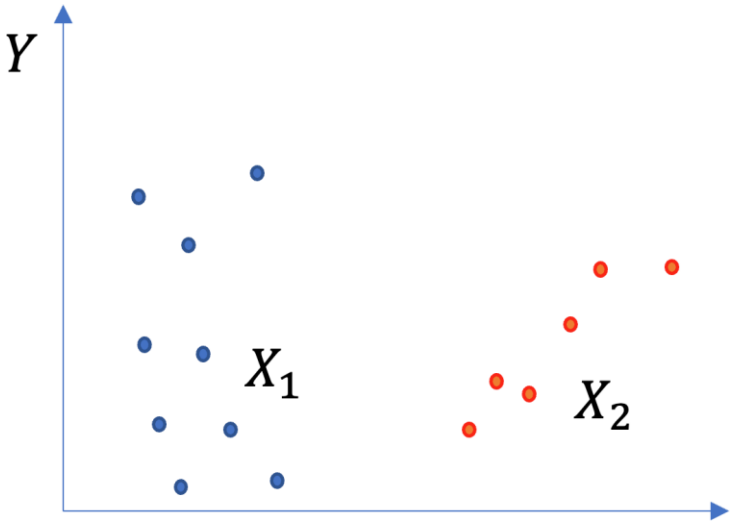

Figure 18. Pearson correlation coefficient visualization

4.2.3 Stacking Meta-Model (SMM)

The performance of insurance claims prediction models is greatly limited by traditional regression algorithms that have different assumptions and features that make it difficult to choose the best algorithm when there are numerous variables and complex conditions. An SMM based on stacking ensemble learning was introduced to eliminate such restrictions.

Figure 3 presents an algorithm that utilizes LightGBM, Random Forest, and ElasticNet models as the base model and Ridge as the meta-model and takes the input of the meta-model as the output of the base model. Compared with traditional regression algorithms, the SMM combines the advantages of multiple models to handle more complex situations and improves the predictive ability of the model. The stacking algorithm employs the prediction results of the base model as the input of the meta-model, and the output of the meta-model is the final prediction result. Although the predictive capability of the classic stacking algorithm may be reduced by the presence of the meta-model, the SMM can discard the meta-model as it has more than one such model. The SHAP tool is used to explain the proposed SMM algorithm and prediction results.

Algorithm 1 Stacking algorithm based on multiple base models

Input: training set D, test set T and base model.$D =$

$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}; T = \{x_{m+1}, y_{m+1}), (x_{m+2}, y_{m+2}), \ldots, (x_n, y_n)\}$

Process:

1. Train each base model on training set D.

2. Predict each base model on training set D and test set T to obtain new features D_{pred} and T_{pred}.

3. Train each meta-model on training set D and make predictions on test set T_{pred}.

4. Combine the predictions of each model with those of the average method.

Several improvements were made in this study based on the traditional stacking algorithm with a single meta-model. As shown in Algorithm 1, the outputs of each base model are used as the inputs of multiple meta-models, whose outputs are combined as the final prediction results. The generalization capability is improved at the cost of increased training effort. L1 and L2 standardized combined the linear ElasticNet, GBDT-based LightGBM, and bagging-based Random Forest models as base models and L1 standardized Lasso and L2 standardized ridge as meta-models. The stacking model combines the linear model, decision tree model, GBDT framework, and bagging algorithm. Each model has its own advantages and is complementary to the others. The three models were trained on the training set, and their predictions were used as input to train the meta-models, whose performances were tested on the test set.

As illustrated in Algorithm 2, some data transformation operations are performed in the SMM. Feature generation, feature selection, outlier processing, and data standardization operations are conducted sequentially to obtain a new dataset. The new dataset contains features that are highly correlated with the predicted label as input to the machine learning algorithm. Finally, a stacking model incorporating two meta-models is employed for claim amount prediction.

Algorithm 2 Claim amount calculation SMM using multiple base models

Input: insurance dataset D_1, Baidu search rate dataset D_2

Process:

1. Generate features from D_1 and D_2, use variance and correlation coefficients to perform feature selection, and obtain datasets $X \in R^{n \times m}$ and label $Y_{real} R^{n \times 1}$.

2. Set parameters for abnormal values of σ, filter undesirable data, and obtain dataset X_{filter} and label Y_{filter}.

3. Standardize the datasets as dataset X_{std} and label Y_{std}.

4. Split the dataset at time point t into dataset X_{train} with label Y_{train} (before t) and dataset X_{test} with label Y_{test} (after t).

5. Train three base models, the LightGBM, Random Forest, and ElasticNet models, based on training sets X_{train} and Y_{train}.

6. Predict the performance of each base model on training set X_{train} and test set X_{test}; obtain new features T^{pred}_{train} and T^{pred}_{test}.

7. Train two meta-models, the ridge and LASSO models, on the training sets T^{pred}_{train} and Y_{train} and predict test set T^{pred}_{test}.

8. Combine the predictions of the two meta-models using the average method.

5. Results and Discussion

This chapter presents the validation of the proposed method through in-depth experiments and analyses. Firstly, the time series data were divided into training and test sets by time-point division. The time point selected for the experiment was July 1, 2020. The data before this time point were used as training data, whereas the subsequent data were used as test data. A comparative analysis of the proposed method and other methods, namely, the LightGBM, Random Forest, ElasticNet, SVM, KNN, and Catboost

approaches, was conducted. In addition, a moving average model was included for comparison to verify the proposed method further.

A pre-analysis of the dataset was performed as described in this chapter. Further discussion and analysis of the experimental results of the proposed method and other algorithms were performed based on the dataset. Of particular interest are the advantages of the stacking model over the other models and the use of SHAP to interpret the results.

5.1 Dataset

In this research, Chinese auto insurance company data from January 1, 2016, to December 31, 2020 were employed, as well as the Baidu index of the epidemic from January 1, 2020, to December 31, 2020. In the model construction phase, the data from January 1, 2016, to June 30, 2020, were used for the training set, whereas the data from July 1, 2020, to December 31, 2020, were used as the test set.

In the experiment described in this chapter, Python3.6 was employed as the programming language and Anaconda was utilized as the programming environment. The Random Forest, SVM, KNN, Lasso, Ridge, and ElasticNet models were built based on the scikit-learn library. The LightGBM and Catboost models were constructed based on the LightGBM and Catboost libraries, respectively. The stacking model was built with the Mlxtend library as the base, and the residual distribution of the prediction model was analyzed based on Yellowbrick. Finally, the SHAP library was used as a base for the analysis and prediction of the SHAP value.

5.2 Evaluation Criteria

The regression task predicts the continuous real value; in other words, the output value is a continuous real value. The main evaluation indicators of the regression model are as follows.

The root mean square error (RMSE), also known as the root mean square deviation (RMSD), is more sensitive to outliers. If the regressor is irrational to the return value of a certain point, it returns a large error, which will have a great impact on the RMSE. In this case, the average value is not robust and is defined as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{n}}.$$

Different predicted values and outliers can be found in regression tasks. With its sensitivity to outliers in mind, the RMSE is not ideal as an evaluation criterion as it fails to measure the model performance objectively. Therefore, it is only used as a reference to assess the regression algorithm performance. Other evaluation criteria include the mean absolute percentage error (MAE) and mean absolute percentage error (MAPE).

- **MAE**

  This indicator is the expected value of the absolute error loss:

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

- **MAPE**

  This indicator is the expected value of the relative error loss. The relative error

is the percentage of the absolute error and true values:

$$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{|y_i|}.$$

The model is applied through a cyclic and iterative process, and only through continuous adjustment and improvement can it be adapted to online data and business goals. At the beginning of the selection model, the data distribution was assumed to be certain, but in practice, it changes with time, a phenomenon known as distribution drift. The verification index can track the performance of the model on a constantly growing dataset. If there is a decrease in performance, the model can no longer adapt to the current data and needs to be retrained. Generalization ability, the ability of the model to adapt to new data, was improved through established evaluation criteria in this research.

5.3 Feature Generation and Selection Experiment Results

As future and historical claims amounts are related in time, the claim amount and number of accidents in the past week were added to the feature set. Features such as the month, year, and holidays were also considered as they affect traffic flow. Using statistical knowledge, the average claim value sand numbers of accidents in the previous week and month were extracted. Finally, the total number of effectively insured and number of insured were added to the feature set. Statistical knowledge was applied to calculate the average Baidu index in the previous week, which was also added to the feature set.

The correlation coefficient between the extracted features was calculated, as shown

in Figure 12. The index characteristics related to the pandemic and the characteristics related to the number of claims are negatively correlated. The Baidu search index is utilized to measure pandemic-related features to quantify behavioral changes caused by changes in public understanding of the pandemic, including objective factors in the external environment and subjective factors from the public itself. It is probable that growing concern about the pandemic, as reflected by the increase in the Baidu search index, led to a decrease in travel frequency and a subsequent reduction in the claim amount.

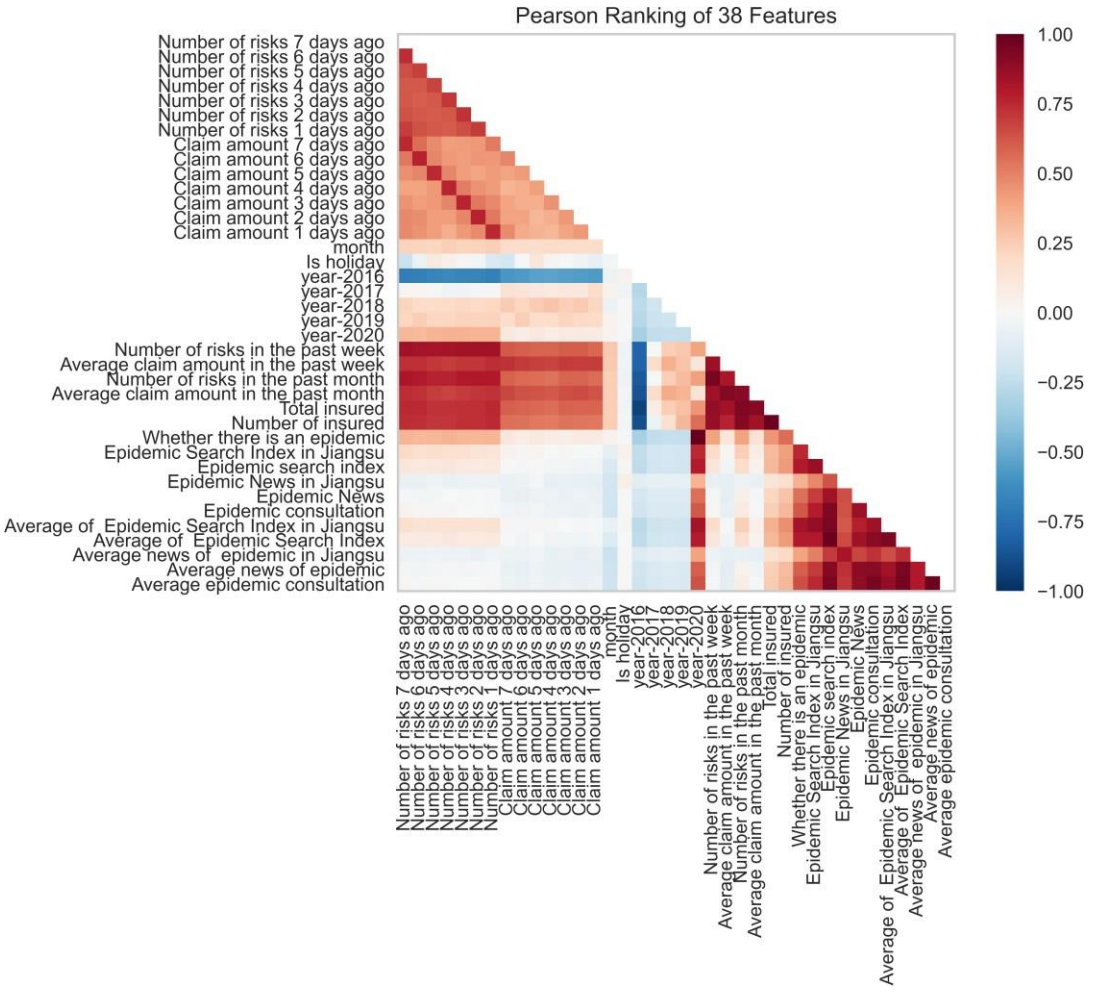

Figure 19. Feature related coefficients

Because high-dimensional data are difficult to visualize, a dimensionality reduction algorithm is used to transform the high-dimensional data into two-dimensional space. Figure 12 shows the transformation of the 38-dimensional data into two-dimensional space using the Isomap algorithm. In the graph, the upper right corner contains deep red dots, the lower right corner includes light red dots, and the lower left corner presents blue to dark blue dots. The smooth changes in the color dots illustrate the smooth and continuous changes in the daily total claim amount. Strong connections between the feature selection and daily total claim amount can be observed.

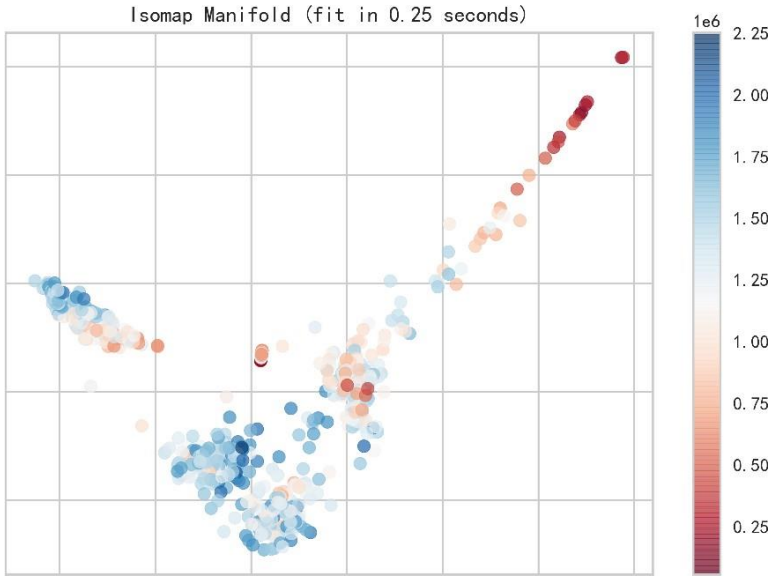

Figure 20 Isomap visualization

In the dataset, there were certain days on which the total daily claim amount is far higher than that in the previous week. If the total claim amount of the day exceeded the average total claim amount of the past seven days by 340,000, it was considered an outliner and was filtered. After filtering, there were 781 data in the training set and 103 data in the test set.

5.4 Claim Amount Prediction Experiment

The claim amount was predicted based on the extracted features. Three types of models were compared in terms of claim amount prediction: the traditional moving average model, traditional machine learning models, and stacking model. For the traditional moving average model, the average historical claim for K days was considered to be the predicted value of one day in the future. When K increases, the prediction will be smoother, and the predicted value will be insensitive to the actual changes in the data. The predicted value may fluctuate as K decreases. The experiment was performed using different K values. The traditional machine learning models considered included the decision tree-based LightGBM, Random Forest, and Catboost models as well as the linear ElasticNet, SVM, and KNN models. Finally, the stacking model was built with the same base model but different meta-models. The LightGBM, Random Forest, and Catboost models were the base models, whereas the ridge and LASSO models were the meta-models. The SMM model employs two meta-models, and the predicted mean value is regarded as the final predicted value.

Name	MAE	MAPE	Median absolute error	RMSE
Mean value of the previous week	156,025.3	14.79683	150,869.3	179,456.8
Claim amount of	323,256	27.1154	256,385.1	455,81

the previous day	.4	5		6.6
Mean value of the previous three days	211,192.8	18.88215	163,448.3	27,667 4.5
Mean value of the previous five days	193,015.9	17.69597	180,538.3	229,68 8.6
Mean value of the previous month	214,874.8	21.50972	178,966.6	261,22 3
LightGBM	213,147.2	18.78956	164,629.3	281,53 9.8
Random Forest	158,356.1	14.75239	152,307.7	182,50 3.1
ElasticNet	217,946.3	22.57935	20,8547	264,01 6.2
SVM	203,538	19.2450	141,188.9	272,08

	.5	7		9.7
KNN	205,287 .4	20.4869 7	172,285.8	253,15 9
Catboost	175,478 .1	17.4507 1	153,819.1	214,78 0.5
Stacking	146,607 .5	12.5713 4	130,924.6	179,36 0.7

Table 6. Prediction experiment results

The proposed prediction model was compared with the other models, and Table 6 presents the experimental results. To compare the moving average model, the average value of the previous K days was selected as the claim amount for the next day. The values of K were 1, 3, 5, 7, and 30 in the experiment. When K = 7, the various indicators are optimal with a MAPE of 14.79, indicating that the absolute value of the prediction error accounts for 14.79 of the true value. The MAE of 156,025.3 is equal to the average absolute error of 156,025.3. Among the aforementioned machine learning models, the stacking model is the best because its MAPE is only 12.57134; hence, it possesses apparent advantages compared to other models.

Features	MAE	MAPE	Median absolute error	RMSE

No epidemic features	153,193	13.14776	136,381.6	181,158.9
Epidemic features	146,607.5	12.57134	130,924.6	179,360.7

Table 7. Whether to use epidemic features

To illustrate the impact of epidemic characteristics on claim prediction further, the pandemic features were removed, and the experiment was performed again using the proposed model. Table 7 shows that the prediction performance decreases in the absence of pandemic features. The MAPE, MAE, median absolute error, and RMSE are increased by 0.5863, 6992, 9496.6, and 4206.6, respectively. In short, pandemic features improve the model performance and affect the claim amount. These findings highlight the effects of the pandemic on the claim amount. COVID-19 has had tremendous impacts on the entire national economy and public travel habits, ultimately affecting the amount of auto insurance claims. Changes in claim amount will not only affect the determination of future insurance rates to a large extent, but also promote changes in the overall business models of auto insurance companies.

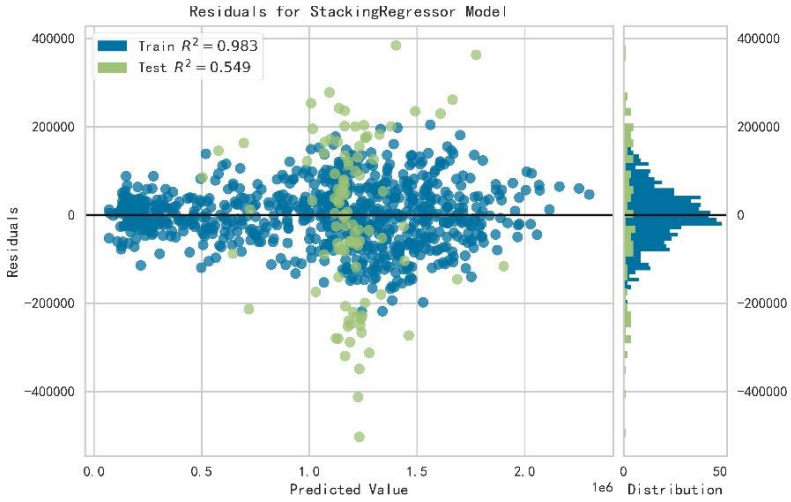

Figure 21. Residual visualization

The coefficients of determination in the training set was calculated, and prediction in the test set is visualized in Figure 14. The training set outperforms the test set; the residual distribution of the former more closely resembles a normal distribution than that of the latter.

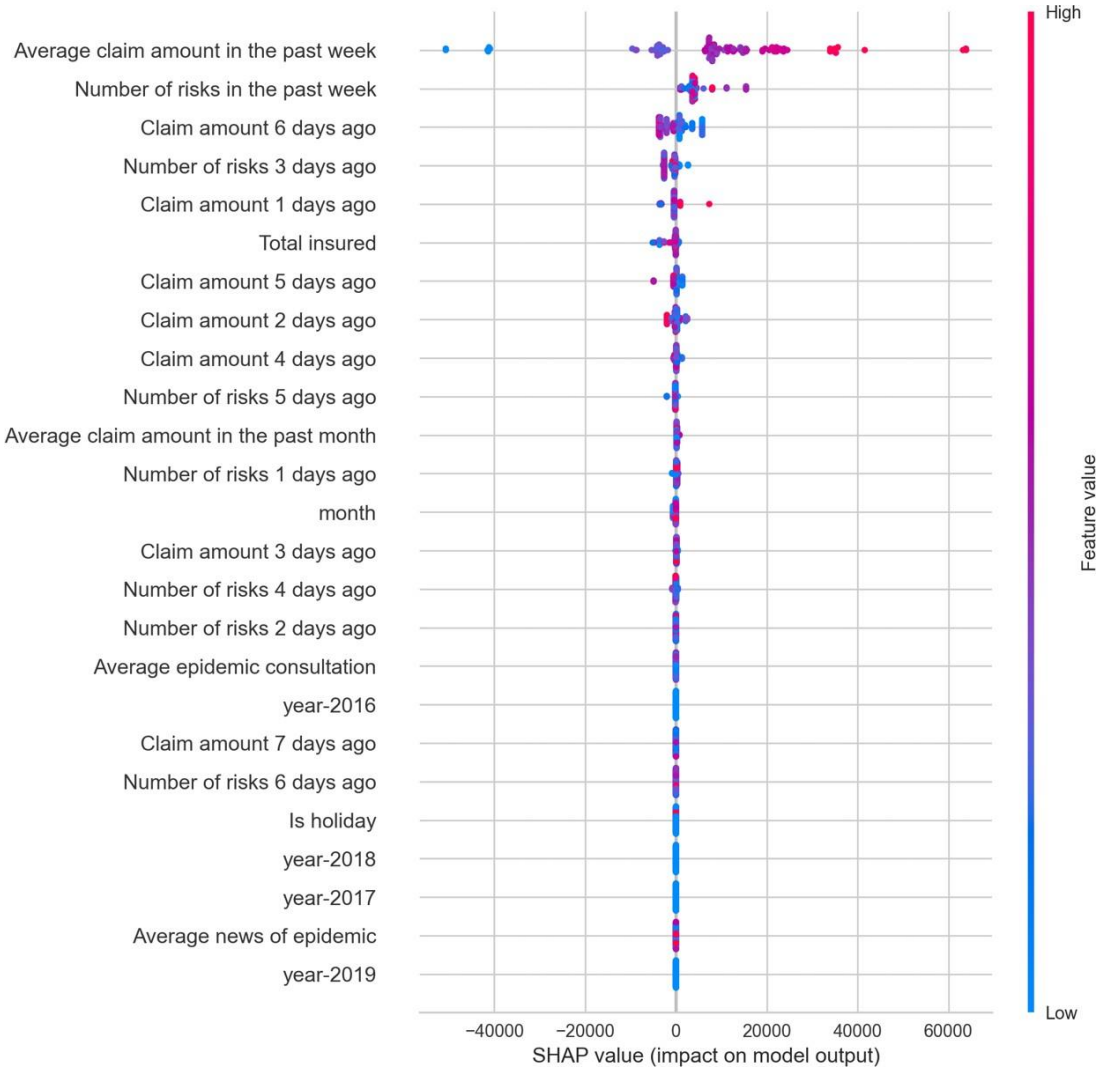

Figure 22. LightGBM model SHAP visualization

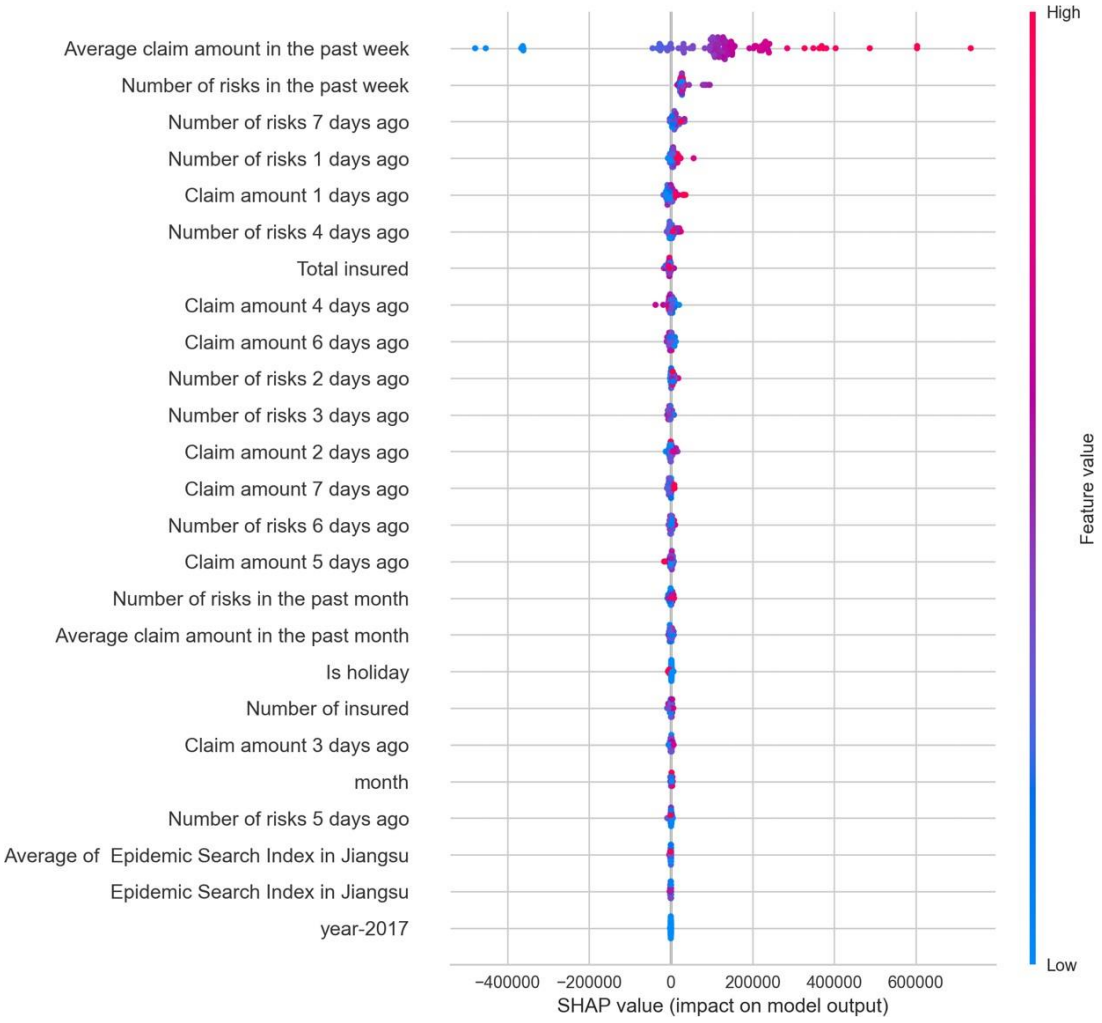

Figure 23. Random forest SHAP visualization

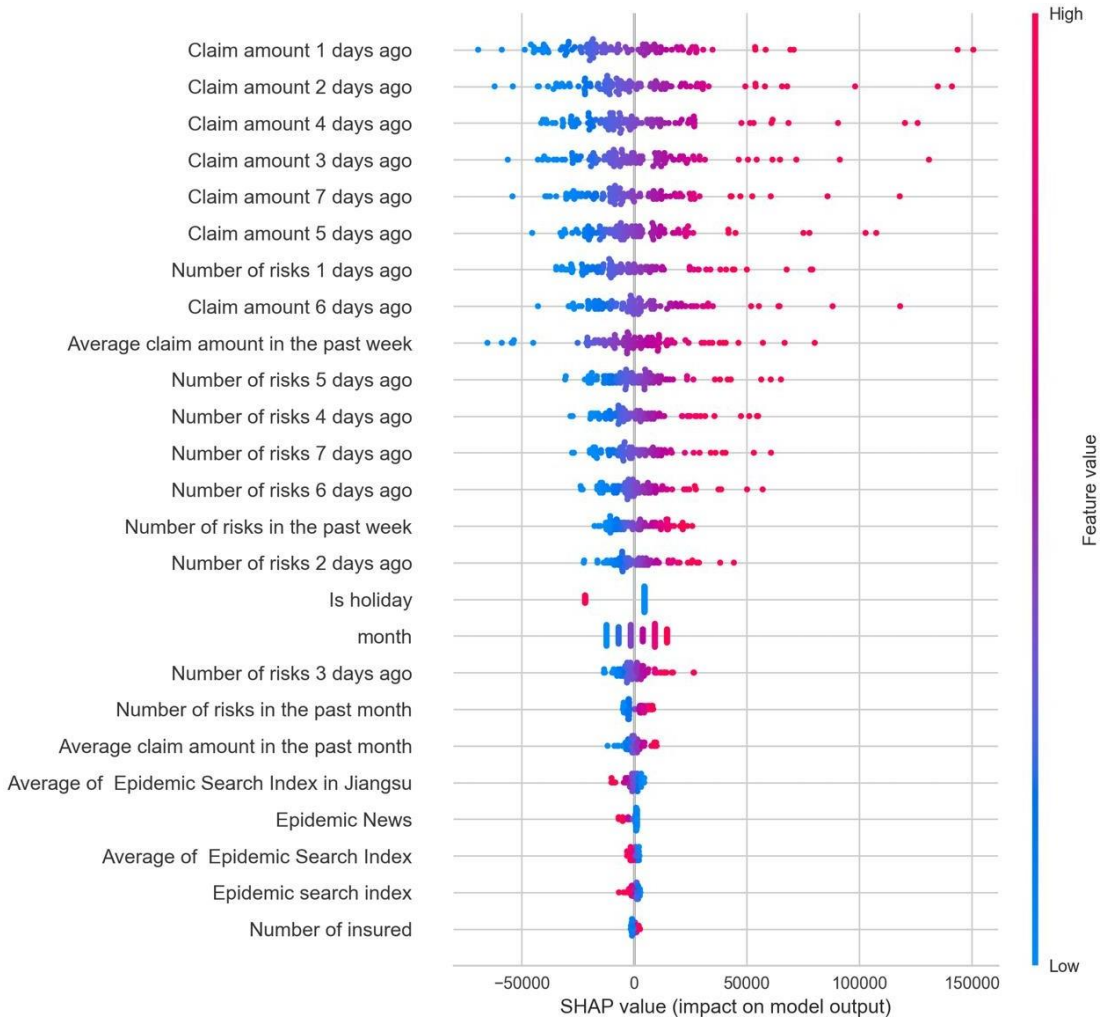

Figure 24. ElasticNet model SHAP visualization

Figures 15–17 present the SHAP values of the three base models of stacking, which are arranged in descending order. The greater the SHAP value, the more distributed the dots on the right side of the graph. The SHAP values are represented by the dots on the x-axis, where blue and red dots indicate smaller and larger values, respectively.

According to Figure 17, the six features with the greatest SHAP values are all features of the claim amount in the past few days. Given the linear nature of ElasticNet and its linear combination of features, it is probable that ElasticNet places greater emphasis on these six features, in particular, pandemic-related features such as pandemic

news and the average pandemic search rate. Figure 17 shows that when the mean pandemic search rate in Jiangsu Province is high, as indicated by the red dots, the SHAP value is less than 0. Meanwhile, when the mean pandemic search rate in Jiangsu is low, as indicated by the blue dots, the SHAP value is greater than 0. Similar results were found for the other pandemic features. The two models utilize weekly claim amounts, which display higher consistency than the claim amounts of the past few days. On the other hand, ElasticNet prefers a linear combination of the claim amount over the past few days over the mean value of the weekly claim amount. As a linear model, ElasticNet is relatively simple, and Figure 17 reveals that it performs more poorly than LightGBM and the random forest model. Other than the predictive performance, the features selected by the three models are quite different, which satisfies the stacking requirement of a great difference from the base model, enabling itself to benefit from the various strengths of each model.

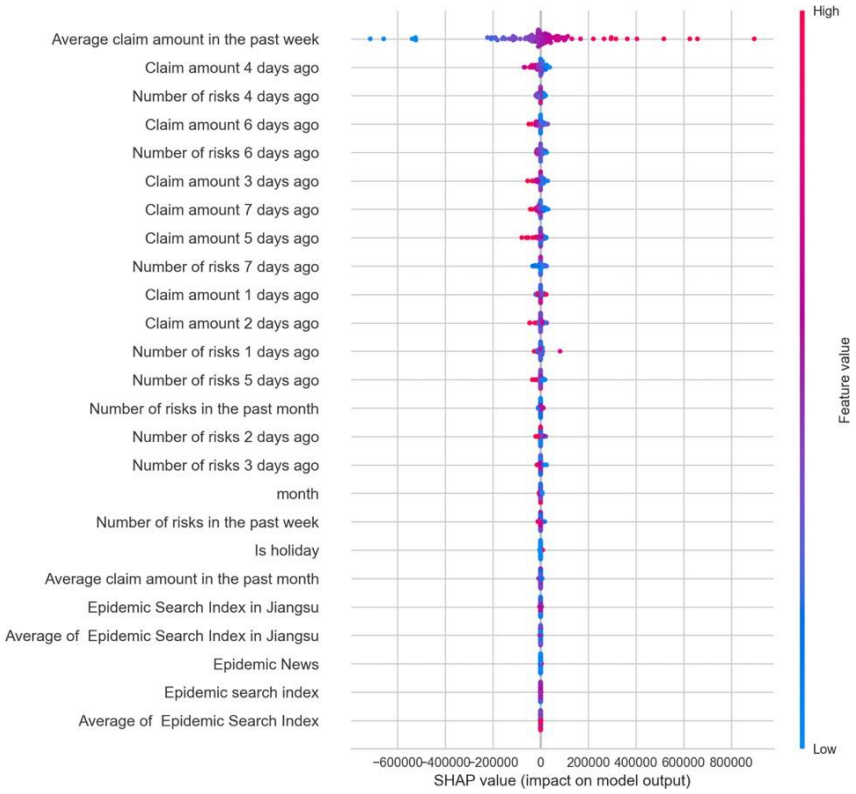

Figure 25. SHAP values of SMM models

In Figure 18, the feature with the largest SHAP value of the SMM is the average claim settlement value in the past week, and the following two features are the claim settlement amounts four and six days ago, in sequence. The distribution of the SMM SHAP value balances the SHAP values of the three base models. Further analysis demonstrates that the SMM combines the strengths of multiple base models.

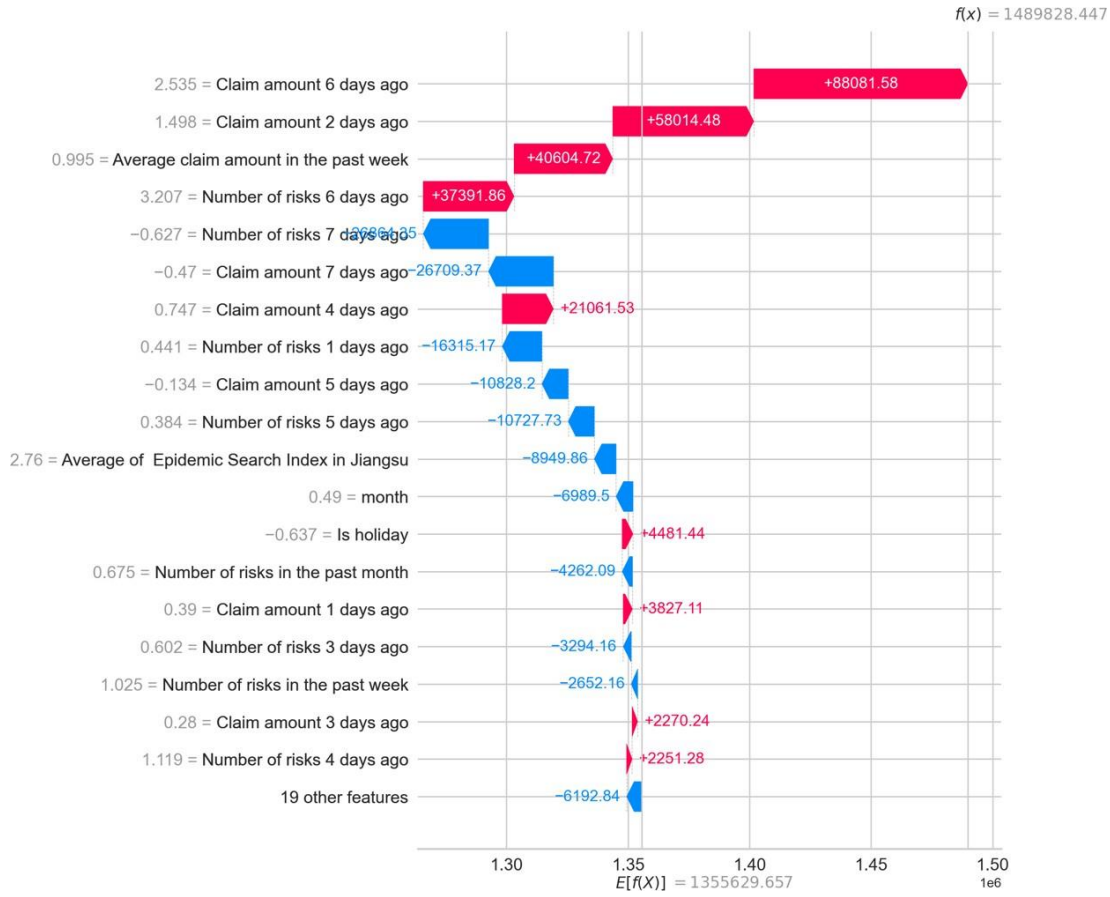

Figure 26. Predictive analysis

Predictive analysis on selective test bank data from August 7, 2020 was performed with SHAP. Figure 19 presents the average predictions of the model. The average prediction for August 7, 2020 is higher than the prediction of 134,198.790 obtained according to the SHAP formula for the 38 features. Figure 19 also reveals that the claim amounts two and six days previously, weekly average claim amount, and claim amount six days prior to an accident are higher, hence their higher SHAP values. The Jiangsu pandemic search rate average is the average value of the search index for the epidemic in Jiangsu in the past week. Although the value of 2.76 is high, the SHAP value is negative. Thus, the Jiangsu pandemic search rate average decreases the predicted value.

It can be concluded that increasing concern regarding the pandemic decreased the number of travelers and traffic flow. Thus, the Jiangsu pandemic search rate average negatively affects the prediction.

6. Conclusion

6.1 Summary

Despite its catastrophic scale, there has been no academic study of the impact of COVID-19 on auto insurance companies thus far. Under the new policy of commercial automobile insurance in China, insurance companies can use the Internet and big data to obtain massive amounts of data and achieve better risk control and rate determination. However, the traditional statistical measurement models used in actuarial science require the functional relationship between the dependent and explanatory variables to be determined in advance, and the functional form is limited and depends on the distribution assumption. An incorrect distribution assumption prevents the sum of the squared errors of the fitting values from reaching the standard performance level, resulting in a poor fitting effect. In this research, it was attempted to solve these problems by using auto insurance data from an insurance company in China and integrating existing machine learning technologies, specifically, the ElasticNet, LightGBM, and Random Forest models.

Such integration coupled with stacking greatly enhances the generalization ability of the model and the prediction precision. The SHAP value was introduced to explain the effects of pandemic factors on automobile insurance predictions. It is found that the auto insurance industry has been greatly affected by pandemic-related factors, which also have negative impacts on the economy.

The 38 features examined in this research were generated by feature engineering and filtered using variance and correlation coefficients. Visual pairwise feature analysis showed negative correlations between the pandemic-related index features and claim amount. A plausible explanation for this finding is that during the pandemic period, marked by increases in the pandemic-related indices, the willingness of people to travel decreased, closely followed by decreases in traffic flow and claim amount. The pandemic-related indices and claim amount were then transformed into two-dimensional space using a dimension reduction algorithm. Isomap formatting also suggested strong relationships between the pandemic features and claim amount.

On this basis, the SMM model constructed in this study was compared with other models. Comparison of the MAE and MAPE proved that the SMM model is superior to the moving average model, traditional machine learning models, and two other stacking models.

To explain the impacts of the pandemic features on the claim amount further, the experiment was repeated after removing the pandemic features. It was found that the MAE, MAPE, and RMSE increased significantly in the absence of the pandemic features, so it can be concluded that the epidemic characteristics have very important impacts on auto insurance claim amounts. Moreover, the analysis of the experimental results showed that the epidemic-related characteristics negatively affect the number of

claims.

Finally, this paper provides a visual explanation of the impacts of SHAP epidemic characteristics on the claim amount and further clarifies the role of the epidemic in auto insurance claim amount prediction. After considering the effects of epidemic characteristics on the auto insurance claim amount, the accuracy of auto insurance claims is improved. Hence, this study makes the following contributions:

(1) It enables risk heterogeneity reduction at different rates The fair premium burden causes the insurance premium paid by the insured to reflect the true risk level. The rate factor can only describe part of the potential loss, and the risk difference that the rate factor cannot express can be reflected through the claim experience.

(2) It will facilitate claim cost reduction and prevent frequent small claims. In an accident, if the damage is minor and the compensation received by filing a claim is lower than the next available discounted premium, the insured tends to pay this small amount in exchange for the renewal of premium discounts. This behavior of the insured is considered in the SMM when predicting the amount of auto insurance claims, which will not only bring reasonable and favorable renewal premiums to the insured, but also reduce the expense incurred by insurance companies in accepting small claims.

(3) It will help control and optimize risks. As there is no indemnity preferential treatment, insured drivers tend to pay more attention to safe driving and take the initiative to control risks. Meanwhile, the screening of the non-compensation preferential treatment system with no indemnity benefits will keep the good risks of insurance companies through premium discounts, so that insurance companies can better understand risk distributions and optimize covered risks.

6.2 Future Work

This paper proposed a fusion model based on stacking to predict the amount of auto insurance claims. Notwithstanding its excellent prediction performance, there are limitations to its application. Although the SHAP value can visualize the pandemic features, additional tools are needed to quantify the effects of these features in each model accurately. Furthermore, as the impact characteristics of the forecast of auto insurance claim amounts are too large, it may be necessary to determine more epidemic characteristics as well as other factors that can affect the vehicle insurance claim amount based on theory and practice in the future. Finally, from the perspective of application, only data from a single city in Jiangsu were utilized in this study, which yielded good application performance. The accuracy of the model could be tested in more cities in China in the future.

With the rapid development of the Internet and artificial intelligence technology, insurance data are growing exponentially, and their quality has improved tremendously. Various effective tools are needed to obtain valuable risk management information from a vast amount of data. As new prediction models, machine learning algorithms have remarkable application prospects in actuarial science and play a pivotal role in improving risk management in the insurance industry.

Bibliography

[1] Worldometer. (n.d.). Real time world statistics. https://www.worldometers.info/coronavirus/

[2] Willmot, G. (1986). Mixed compound Poisson distributions. ASTIN Bulletin: The Journal of the IAA, 16(1), 59-79.

[3] Haberman, S., & Renshaw, A. E. (1990). Generalized Linear Models in Actuarial Work. Journal of the Staple Inn Actuarial Society, 32, 171-172.

[4] Gerber, H. U. (1992). From the generalized gamma to the generalized negative binomial distribution. Insurance: Mathematics and Economics, 10(4), 303-309.

[5] Walhin, J. F., & Paris, J. (2000). The true claim amount and frequency distributions within a bonus-malus system. ASTIN Bulletin: The Journal of the IAA, 30(2), 391-403.

[6] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.

[7] Gupta, P. L., Gupta, R. C., & Tripathi, R. C. (2005). Score test for zero inflated generalized Poisson regression model. Communications in Statistics-Theory and Methods, 33(1), 47-64.

[8] Yip, K. C., & Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics, 36(2), 153-163.

[9] Ismail, N., & Jemain, A. A. (2007). Handling overdispersion with negative binomial

and generalized Poisson regression models. In Casualty actuarial society forum (Vol. 2007, pp. 103-58). Citeseer.

[10] Leo, A., Walker, A. M., Lebo, M. S., Hendrickson, B., Scholl, T., & Akmaev, V. R. (2012). A GC-wave correction algorithm that improves the analytical performance of aCGH. The Journal of Molecular Diagnostics, 14(6), 550-559.

[11] Guelman, L., Guillén , M., & Pérez -Marín , A. M. (2012). Random Forests for Uplift Modeling: An Insurance Customer Retention Case. In: Engemann, K. J., Gil-Lafuente, A. M., & Merigó, J. M. (eds) Modeling and Simulation in Engineering, Economics and Management. MS 2012.

[12] Liu, Y. (2014). Random forest algorithm in big data environment. Computer Modelling & New Technologies, 18(12A), 147-151.

[13] Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. Transportation Research Part A: Policy and Practice, 61, 27-40.

[14] Lee, S., Lin, S., & Antonio, K. (2015a). Delta boosting machine and its application in actuarial modeling.

[15] Lee, S., & Antonio, K. (2015b). Why high dimensional modeling in actuarial science?

[16] Mzhavia, T. (2016). Vehicle insurance claim data study and forecasting model using artificial neural networks: Tallinn University of Technology, Tallinn, Estonia.

[17] Gao, G., Wang, H., & Wüthrich, M. V. (2021). Boosting Poisson regression models with telematics car driving data. Machine Learning, 1-30.

[18] Liu, Y., Wang, B. J., & Lv, S. G. (2014). Using multi-class AdaBoost tree for prediction frequency of auto insurance. Journal of Applied Finance and Banking, 4(5), 45.

[19] Sakthivel, K. M., & Rajitha, C. S. (2017). A comparative study of zero-inflated, hurdle models with artificial neural network in claim count modeling. International Journal of Statistics and Systems, 12(2), 265-276.

[20] Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5(2), 197-227.

[21] Chandra, A., & Yao, X. (2006). Ensemble learning using multi-objective evolutionary algorithms. Journal of Mathematical Modelling and Algorithms, 5(4), 417-445.

[22] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

[23] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

[24] Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., & Jin, Y. (2019). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. Applied Soft Computing, 77, 188-204.

[25] Cui, S., Yin, Y., Wang, D., Li, Z., & Wang, Y. (2021). A stacking-based ensemble learning method for earthquake casualty prediction. Applied Soft Computing, 101, 107038.

[26] Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble learning or deep learning? Application to default risk analysis. Journal of Risk and Financial Management, 11(1), 12.

[27] Minastireanu, E., & Mesnita, G. (2019). Light GBM machine learning algorithm to online click fraud detection. Journal of Information Assurance and Cybersecurity, 263928.

[28] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.

[29] Velavan, T. P., & Meyer, C. G. (2020). The COVID – 19 epidemic. Tropical Medicine & International Health, 25(3), 278.

[30] Khan, N., & Faisal, S. (2020). Epidemiology of Corona virus in the world and its effects on the China economy. Available at SSRN 3548292.

[31] World Bank Group. (2021). Timor-Leste Economic Report, May 2021: Charting a New Path.

[32] Wang, Y., Zhang, D., Wang, X., & Fu, Q. (2020). How does COVID-19 affect China's insurance market?. Emerging Markets Finance and Trade, 56(10), 2350-2362.

[33] Volosovych, S., Zelenitsa, I., Kondratenko, D., Szymla, W., & Mamchur, R. (2021). Transformation of insurance technologies in the context of a pandemic. Insurance Markets and Companies, 12(1), 1-13.

[34] Babuna, P., Yang, X., Gyilbag, A., Awudi, D. A., Ngmenbelle, D., & Bian, D. (2020). The impact of Covid-19 on the insurance industry. International journal of environmental research and public health, 17(16), 5766.

[35] Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals, 139, 110059.

[36] Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant applications of machine learning for COVID-19 pandemic. Journal of Industrial Integration and Management, 5(4).

[37] Dreyer, A., Kritzinger, G., & Decker, J. D. (2007, June). Assessing the Impact of a Pandemic on the Life Insurance Industry in South Africa. In 1st IAA Life Colloquium, Stockholm.

[38] Fan, V. Y., Jamison, D. T., & Summers, L. H. (2018). Pandemic risk: how large are the expected losses?. Bulletin of the World Health Organization, 96(2), 129.

[39] O'Connor, C. M., Anoushiravani, A. A., DiCaprio, M. R., Healy, W. L., & Iorio, R. (2020). Economic recovery after the COVID-19 pandemic: resuming elective orthopedic surgery and total joint arthroplasty. Journal of Arthroplasty, 35(7), S32-S36.

[40] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Machine learning proceedings 1992 (pp. 249-256). Morgan Kaufmann.

[41] Ellis, M., Durand, H., & Christofides, P. D. (2014). A tutorial review of economic model predictive control methods. Journal of Process Control, 24(8), 1156-1178.

[42] Folks, J. L., & Chhikara, R. S. (1978). The inverse Gaussian distribution and its statistical application—a review. Journal of the Royal Statistical Society: Series B (Methodological), 40(3), 263-275.

[43] Sheu, M. L., Hu, T. W., Keeler, T. E., Ong, M., & Sung, H. Y. (2004). The effect of a major cigarette price change on smoking behavior in California: A zero-inflated negative binomial model. Health Economics, 13(8), 781-791.

[44] He, D., Habetler, T., Mousavi, M. J., & Kang, N. (2013, July). A ZIP model-based feeder load modeling and forecasting method. In 2013 IEEE Power & Energy Society General Meeting (pp. 1-5). IEEE.

[45] Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions. Communications in Statistics-Theory and Methods, 32(2), 281-289.

[46] Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378.

[47] Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, 308-324.

[48] Li, T. R., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R., & Fu, F. (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. Frontiers in Physics, 7, 98.

[49] Dietterich, T. G. (2002). Ensemble learning. The Handbook of Brain Theory and Neural Networks, 2(1), 110-125.

[50] Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2019, September). Explanation of machine learning models using improved Shapley Additive Explanation. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 546-546).

[51] Gong, X., Han, Y., Hou, M., & Guo, R. (2020). Online public attention during the early days of the COVID-19 pandemic: Infoveillance study based on Baidu index. JMIR Public Health and Surveillance, 6(4), e23098.

[52] Fang, J., Zhang, X., Tong, Y., Xia, Y., Liu, H., & Wu, K. (2021). Baidu Index and COVID-19 epidemic forecast: Evidence from China. Frontiers in Public Health, 9, 488.

[53] Azhagusundari, B., & Thanamani, A. S. (2013). Feature selection based on information gain. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2(2), 18-21.

[54] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

[55] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.

[56] Sánch ez-Maroño, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007, December). Filter methods for feature selection–a comparative study. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 178-187). Springer, Berlin, Heidelberg.

[57] O'brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. Journal of the American Statistical Association, 74(368), 877-880.

[58] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise Reduction in Speech Processing (pp. 1-4). Springer, Berlin, Heidelberg.

[59] Hastie, T. (2020). Ridge regularization: An essential concept in data science. Technometrics, 62(4), 426-433.

[60] Park, T., & Casella, G. (2008). The Bayesian lasso. Journal of the American Statistical Association, 103(482), 681-686.

[61] Eltoft, T., Kim, T., & Lee, T. W. (2006). On the multivariate Laplace distribution. IEEE Signal Processing Letters, 13(5), 300-303.

[62] Li, Q., & Lin, N. (2010). The Bayesian elastic net. Bayesian analysis, 5(1), 151-170.

[63] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 3146-3154.

[64] Seeniselvi, T., & Nirmala, M. (2019). A survey on data preparation and feature engineering in machine learning. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 8(5).

[65] Noble, W. S. (2006). What is a support vector machine?. Nature Biotechnology, 24(12), 1565-1567.

[66] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.

[67] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

Appendices

Appendix A. Coding Reference

1. Evaluation

```
def my_score(label,pre):
    rmse=np.sqrt(mean_squared_error(label,pre))
    mae=mean_absolute_error(label,pre)
    mape_score=mape(label,pre)
    r2_scores=r2_score(label,pre)
    msle=mean_squared_log_error(label,pre)
    median_ae=median_absolute_error(label,pre)
    print('rmse:{}'.format(rmse))
    print('mae:{}'.format(mae))
    print('mape:{}'.format(mape_score))
    print('r2:{}'.format(r2_scores))
    print('mean_squared_log_error:{}'.format(msle))
    print('median_absolute_error:{}'.format(median_ae))
    return
```

{'rmse':rmse,'mae':mae,'mape':mape_score,'r2':r2_scores,'mean_squared_log_error':msle,'median_absolute_error':median_ae}

2. Prepeocessing

```
df_dataset=pd.read_csv('dataset.csv',index_col=0)

label=np.load('labels.npy')

df_dataset.index=pd.to_datetime(df_dataset.index)

df_dataset['label']=label

good_index=[]

for i in range(df_dataset.shape[0]):

    if(np.abs(label[i]-df_dataset.iloc[i]['claim_7':'claim_1'].mean())<340000):

        good_index.append(i)

good_test_index=[]

for i in range(df_dataset.shape[0]):

    if(np.abs(label[i]-df_dataset.iloc[i]['claim_7':'claim_1'].mean())<340000):

        good_test_index.append(i)



df_good=df_dataset.iloc[good_index]

df_good_test=df_dataset.iloc[good_test_index]
```

X_train=df_good.loc[df_good['flag']=='train','number_of_insurance_7':'epi-
demic_news']

X_test=df_good_test.loc[df_good_test['flag']=='test','number_of_insurance_7':'epi-
demic_news']

y_train=df_good.loc[df_good['flag']=='train','label']

y_test=df_good_test.loc[df_good_test['flag']=='test','label']

good_index.extend(good_test_index)

df_data=df_dataset.iloc[good_index]

3. Prediction

ut=my_score(y_test,X_test['week_ago'].values)

out['name']='week_average'

out['random_state']=0

my_log.insert_dict(out)

out=my_score(y_test,X_test['claim_1'].values)

out['name']='day_ago'

out['random_state']=0

my_log.insert_dict(out)

4. Feature Selection

from sklearn.feature_selection import VarianceThreshold

```
df_corr=df_data.corr()['label']

indexs=df_corr[np.abs(df_corr)>0.005].index.to_list()

indexs.remove('label')
```

5. Stacking Regressor

```
best_seed=11

seed=32

clf1=lgb.LGBMRegressor(max_depth=8,learning_rate=0.001,random_state=seed)

clf2=RandomForestRegressor(max_depth=8,random_state=seed)

clf3=ElasticNet(random_state=seed)

clf4=Lasso(random_state=seed,alpha=0.01)

mclf=StackingRegressor(

[clf1,clf2,clf3],

clf4,

verbose=0,

use_features_in_secondary=False,

store_train_meta_features=False,

refit=False,

)
```

```
mclf.fit(X_train,y_train)

pre1=mclf.predict(X_test)

out=my_score(y_test,pre1)

pre_new=(pre+pre1)/2

out=my_score(y_test,pre_new)
```

6. SVM

```
from sklearn.svm import SVR

clf=SVR(kernel='rbf')

clf.fit(X_train,y_train)

pre=clf.predict(X_test)

out=my_score(y_test,np.abs(pre))

out['name']='SVM'

out['random_state']=7.

my_log.insert_dict(out)
```

7. KNN

```
from sklearn.neighbors import KNeighborsRegressor

clf=KNeighborsRegressor(n_neighbors=7)

clf.fit(X_train,y_train)
```

```
pre=clf.predict(X_test)

out=my_score(y_test,np.abs(pre))

out['name']='KNN'

out['random_state']=7.

my_log.insert_dict(out)
```

8. SHAP value

```
seed=7

clf1=lgb.LGBMRegressor(max_depth=8,learning_rate=0.001,random_state=seed)

clf2=RandomForestRegressor(max_depth=8,random_state=seed)

clf3=ElasticNet(random_state=seed)

clf4=Ridge(random_state=seed)

mclf=StackingRegressor([clf1,clf2,clf3],clf4,verbose=0,use_features_in_second-
ary=False,store_train_meta_features=False,refit=False,)

mclf.fit(X_train,y_train)

explainer=shap.TreeExplainer(mclf.regressors[0])

shap_values=explainer(X_test)

shap.summary_plot(shap_values,max_display=25,show=False)

plt.savefig('plot/Lightgbm_im.png',bbox_inches='tight',dpi=200)
```

```
explainer=shap.LinearExplainer(mclf.regressors[2],X_test)

shap_values=explainer(X_test)
```