Prediction of Accumulated Auto Insurance Claims Based on Improved XGBOOST Modeling

Yang Li* and Zhao Jinyan

Hebei University of Technology, Tianjin, China
Email: 3179076588@qq.com (Y.L.); zhaojinyan@hebut.edu.cn (Z.J.Y.)
*Corresponding author

Manuscript received August 27, 2024; revised September 25, 2024; accepted October 27, 2024; published December 20, 2024.

***Abstract*—Generalized linear model is a commonly used method in the traditional cumulative claim prediction. With the advent of the era of big data, machine learning algorithm has achieved good results in this field, but it is still not very good in prediction accuracy, fitting effect and practical interpretation. In the era of big data, with the substantial increase in the number and dimension of data, the key to how to predict the pure premium amount more accurately lies in finding a model that is more suitable for the characteristics of data, and improving the prediction effect and accuracy of the model is a crucial issue. In order to solve the problem of low prediction accuracy of accumulated insurance claims, a kind of data equalization method is adopted to predict the balanced insurance data, and the XGBoost model with improved loss function is used to predict the insurance data after SOMTE equalization. The improved model is more suitable for the processed data, and finally it has better effect and more accurate prediction results than the original XGBoost model.**

***Keywords*—XGBoost, Synthetic Minority Over-sampling Technique (SMOTE), Huber loss, cumulative claim amount**

I. Introduction

With the development of science and technology and the change of realistic policies, the insurance industry began to actively explore the use of big data and advanced algorithms to optimize insurance pricing. For car owners and consumers, reasonable car insurance pricing can ensure that they get fair and transparent insurance products. Through the scientific pricing model, individual car owners can choose the appropriate insurance scheme according to their own risk level and demand, and realize the balance between risk protection and economic benefits. In a more perfect model, it has a positive incentive effect on correcting bad driving habits and enhancing risk awareness. This is very important for road safety and people's life and property safety. Meaning. For insurance companies, accurate automobile insurance pricing is the key to ensure profitability. Through in-depth research and accurate calculation of auto insurance premiums, insurance companies can manage risks more effectively, improve profitability and ensure sufficient funds to cope with claims risks.

To sum up, using big data and machine learning technology to study automobile insurance pricing can not only improve the efficiency and profitability of the insurance industry, but also provide customers with more reasonable and personalized insurance products and services, and promote the development of the insurance industry in an intelligent and refined direction, which is of great significance to China's insurance industry.

II. Literature Review

In the 1990s, actuarial research in the UK first introduced generalized linear models into non life insurance pricing. Ohlsson and Johansson introduced the application of GLM in pricing, and based on this, began to discuss the application of GLM's extended model in non life insurance pricing [1]. De Jong and Heller provided a detailed example on how to use generalized linear models in insurance [2]. These examples are mainly implemented through programming software SAS and R language. In previous studies, more consideration was given to the distribution types that fit the model and the classical mixed Poisson correlation distribution. In addition, Jorgensen and de Souza, Smyth and Jorgensen innovatively constructed the Tweedie GLM model to model claim frequency and claim intensity respectively [3]. For the first time, it was proposed that the individual claim amount follows a gamma distribution and the claim frequency follows a Poisson distribution. As this model considers both zero and continuous non-zero claims, it has been widely applied in actuarial science. Although Tweedie GLM has a wide range of applications, its logarithmic mean structure is limited to a linear form and is too demanding in practical use. Klein *et al.* extended GLM and applied the Generalized Linear Additive Model (GAMLSS) to analyze the claim data of a car insurance company in Belgium. They also established a zero adjustment model for the distribution of claim amounts and concluded that the results were superior to the GLM model [4]. In China, the practical application research of generalized linear models began in the 1970s, with logistic regression being a typical representative. Chen [5] systematically elaborated on the model and analyzed it with examples. Meng *et al.* introduced the generalized linear model into the determination of non life insurance classification rates and analyzed it in combination with specific applications. They proposed various methods for estimating model parameters, such as least squares, maximum two multiplication, direct method, marginal sum method, etc. They systematically compared and analyzed the model and parameter estimation methods [6]. Wang [7] used factors such as vehicle origin, vehicle type, sales source, vehicle age, number of seats in the vehicle, and driving area as rate factors, and established a generalized linear model for research based on specific examples. When exploring the factors affecting claim frequency, Zhang *et al.* [8] introduced the semi parametric smoothing method to predict the probability of claim occurrence, and compared it with the traditional GLM logistic model. The results showed that

the newly established GAM logistic model set had better performance. Due to the characteristics of over dispersion, zero inflation, and heterogeneity of insurance data in our country, Meng *et al.* [9] constructed a zero inflation mixed Poisson distribution model and constructed a mixture with a layered structure. When using generalized linear models, it is often based on the assumption that the frequency and intensity of claims are independent of each other. Meng and Li [10] established a dependency adjustment model to solve the problem of this hypothesis. Based on the independence of claim frequency and claim intensity, they obtained results that are more consistent with the actual situation through dependency adjustment.

III. Theoretical Basis

A. XGBoost

Algorithm Extreme gradient Boosting (XG Boost) near model and boosting model, and its basic model is the integration of Categorical Regression Tree (CART).Usually, the prediction accuracy of using a single decision tree for classification regression often fails to reach the expected goal, and the prediction results are too absolute fitting. XGBoost algorithm determines the prediction results by integrating the prediction results of multiple CART tree models, which greatly improves the prediction accuracy of the models. The basic idea of this algorithm is to establish a base classifier first, and then add new classifiers step by step. After each classifier is added, the value of its objective function is calculated, so as to ensure that the value of the objective function decreases gradually during the iteration process, so as to continuously improve the expression effect of the model.

B. Loss Function

All algorithms in machine learning depend on optimization. We call the function that needs to be optimized "objective function", and the minimized set of functions is called "loss function". For different types of data, we need to select the appropriate loss function to achieve better optimization results. Loss function can be roughly divided into two categories: classified loss function and regression loss function. Among them, our two common loss functions are: mean square error loss function and average absolute error loss function.

C. Data Imbalance Processing Methods

The unbalanced distribution of sample categories can be divided into two types: unbalanced distribution of big data and unbalanced distribution of small data. Uneven sample distribution will often lead to too few features in the classification with small sample size, and it is difficult to extract rules from it; Even if the final model is obtained, it will easily lead to the problem of over-fitting due to excessive dependence on limited data samples. When the model is applied to new data samples, the accuracy and stability of the model will be poor. I often use three methods to deal with unbalanced data: undersampling, oversampling and model algorithm.

IV. Empirical Application

A. Data Sources

The data used in this paper comes from the real data of data car data set in the insurance data package in R language. The data collected is the insurance loss data published by an insurance company, with a total of 67,856 pieces of data, including 4,621 pieces of claim data and 63,235 pieces of unclaimed data.

Table 1. Variable summary table

Variable name	Type of variable	Variable value
Motor vehicle value	continuation	[0, 35)
Exposure risk value	continuation	(0, 1)
Number of claims	classify	0, 1, 2, 3, 4
Accumulated claim amount	continuation	[0,55923)
Driver gender	classify	F, M
Driving area	classify	A, B, C, D, E, F
Whether to claim	classify	0, 1
Driver age	order	1, 2, 3, 4, 5, 6
Service life of motor vehicles	order	1, 2, 3, 4
Motor vehicle use	classify	Of the 13 level variables

1) Cumulative claim amount

Cumulative claim amount is the response variable selected in this paper, which has obvious characteristics of zero inflation and right deviation distribution. After excluding the policy data with zero cumulative claim amount, it can be obtained through calculation that the average cumulative claim amount of the remaining policies is 2014.4, the minimum claim amount is 200 USD, and the maximum claim amount is 55,922.1 USD, with a large relative difference, and most of the cumulative claim amounts are distributed within 3,000 USD, as shown in Fig. 1.

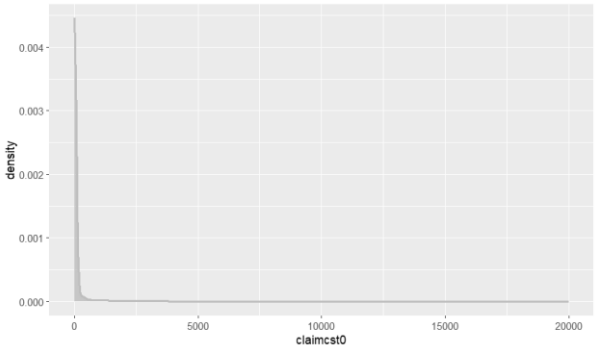

Fig. 1. Statistical chart of cumulative claim amount.

2) Value of motor vehicles

The vehicle value of motor vehicles is a continuous variable, which refers to the purchase value of insured vehicles, ranging from 0 to 350,000 US dollars. As can be seen from Fig. 2, the value of most vehicles in the data set ranges from 0 to 50,000 US dollars, and the value of a few vehicles exceeds 50,000 US dollars, as shown in Fig. 2.

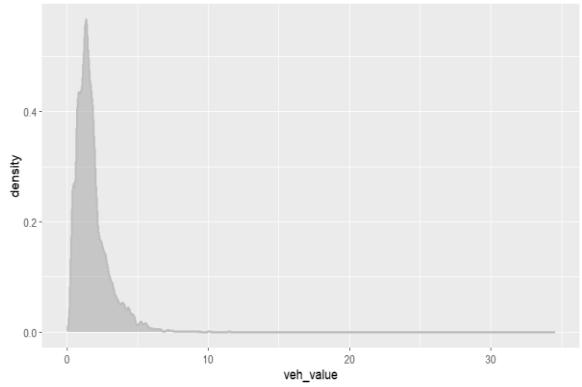

Fig. 2. Vehicle value distribution map of motor vehicles.

3) Exposure risk value

The exposure risk value refers to the risk coefficient of possible claims of the insured automobile, which is a continuous variable with a value between 0 and 1. As shown in Fig. 3, the exposure risk value of each insurance data in the data set is evenly distributed between 0 and 1, as shown in Fig. 3.

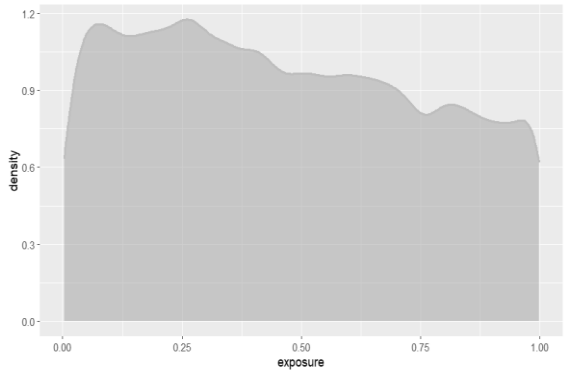

Fig. 3. Inquiry diagram of exposure risk value.

4) Whether to claim and the number of claims

In the variable of whether a claim has occurred, 0 means that no claim has occurred and 1 means that a claim has occurred. Among the 67,856 policies in the data sample, 63,592 policies have no claims, while only 4,264 policies have claims more than once. It can also be seen from the statistical chart of the number of claims that 67,565 of the 67,856 policies have no claims, while only 291 policies have more than one claim, accounting for less than 0.1%. Therefore, the cumulative claim amount of the response variable selected in this paper has obvious characteristics of zero expansion and right deviation distribution, as shown in Figs. 4 and 5.

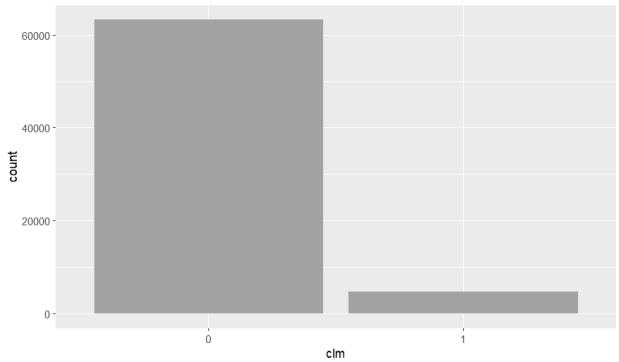

Fig. 4. Statistics of whether a claim has occurred.

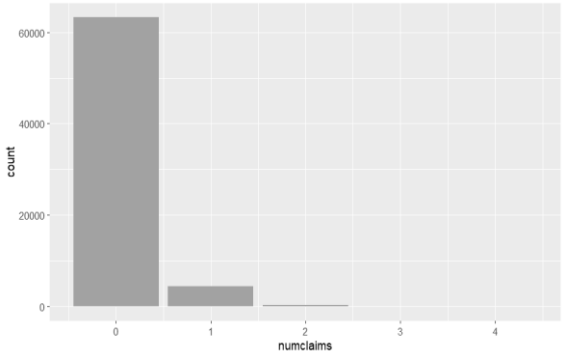

Fig. 5. Statistical chart of claim times.

5) Driver's gender, exercise area, driver's age and service life of motor vehicle

The gender of drivers mainly includes two categories, namely, F category: female and M category: male. In the sample data, the ratio of the two types is about 1: 1; The driving area is mainly divided into six areas: A, B, C, D, E and F, and the number of insurances in each area is relatively average. Among them, area C contains the most samples, and area F contains the least samples. The data sample divides the driver's age into four stages, and each stage contains a relatively balanced number of people, among which the number of people in the third group is the largest and the number of people in the first group is the least. The service life of motor vehicles ranges from 1 to 6 years, and the data distribution is mainly normal, which has no obvious influence on data research, as shown in Figs. 6–9.

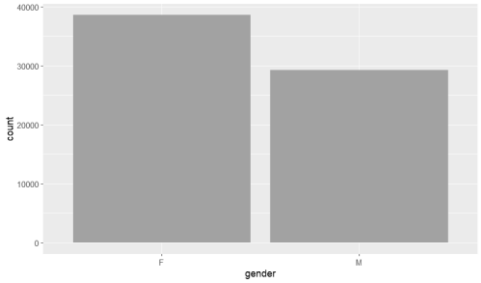

Fig. 6. Gender statistics of drivers.

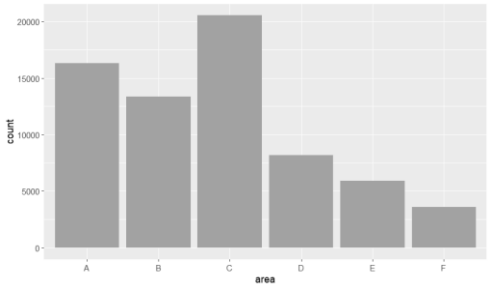

Fig. 7. Statistical map of driving area.

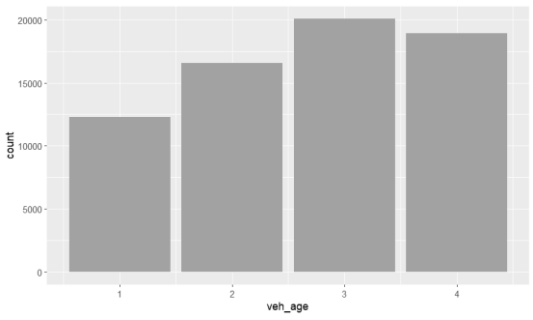

Fig. 8. Statistical chart of driver's age.

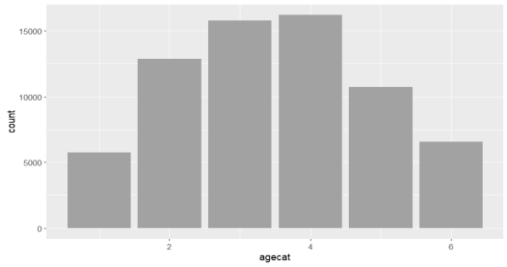

Fig. 9. Statistical chart of service life of motor vehicles.

6) Motor vehicle use purposes

There are 13 types of motor vehicle uses with uneven distribution. For example, SEDAN has 22,233 policies, accounting for 32.765% of the total policies. Similarly, the number of policies of HBACK and STNWG is more than 20%, while the number of policies of BUS, CONVT, MCARA and RDSTR is less than 1%, as shown in Fig. 10.

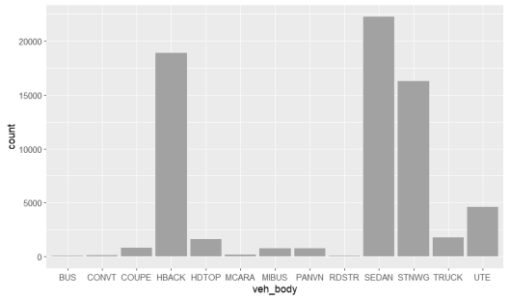

Fig. 10. Statistical chart of motor vehicle use and purpose.

B. Original XGBOOST Model

1) Data preparation processing

Because the insurance claim data has a large amount of unclaimed data, we choose tweedie distribution to construct the loss function, which is more in line with the characteristics of the data set. However, the value range of the dependent variable corresponding to tweedie is $[0,+\infty)$,so there is no need to standardize and normalize the original data set. We select 85% of the data as training data and the remaining 15% as test data.

2) Model building

We use the "xgb.train" function in the "xgboost" package to build the model. After many parameter adjustments, the final parameters were as follows: the learning rate was 0.4, the gamma value was 5, the maximum depth of the tree was set to 10, the control parameter of Tweedie distribution variance was 1.03, and the sub-sampling rate of each tree time series was 0.62. The XGBoost model is predicted on the test set, and the absolute average error of the results is calculated as the model measure, that is, MAE(XGBoost)=192.16.

C. Improved XGBOOST Model

1) Data preprocessing

After processing, the accumulated claim amount is non-zero, so we can't choose XGBoost model of tweedie loss function and use our improved XGBOOST model.

In the original data, there are 63,232 claims and 4,264 unclaimed samples. After SOMTE equilibrium, there are 33,827 claim samples and 34,029 unclaimed data. The ratio of claim sample data to unclaimed sample data is close to 1:1, and the sample data is balanced, which is more convenient for us to analyze. The process changes are shown in Figs. 11 and 12.

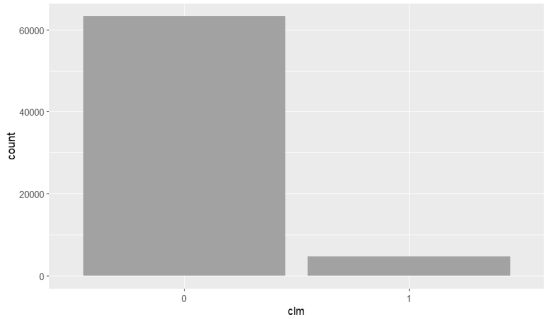

Fig. 11. Original data distribution map.

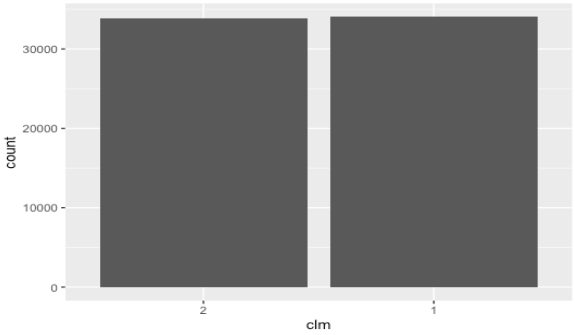

Fig. 12. Balanced data distribution map.

2) Model construction

We use the improved XGBoost model. Compared with the original model, we choose huber loss function, which has better applicability and stability.

We use the "xgb.train" function in the "xgboost" package to build the model. After many times of parameter adjustment, the parameters finally selected were as follows: learning rate was 0.4, gamma value was 5, the maximum depth of the tree was set to 10, the sub-sampling rate of each tree was 0.62, and the number of lifting wheels was 500, which was randomly planted. The XGBoost model is predicted on the test set, and the absolute average error value of the results is calculated as the model measure, that is, MAE(XGBoost)=1.135.

D. Model Comparison and Analysis

1) Error comparison

The root mean square error of XGBoost model with Huber loss as objective function is better than that of XGBoost model with mean square error loss as objective function and XGBoost model with Tweedie distribution as objective function. XGBoost model with Huber loss as the objective function has the smallest error, the highest prediction accuracy and the best prediction effect.

Table 2. Error comparison table

Object function	RMSE	MAE
Huber	2.6527	1.135
Tweedie	1092.84	255.76

2) Comparison of the importance of variables

And that importance or d of the output variables of the two model is roughly the same, as shown in the following Figs. 13 and 14.

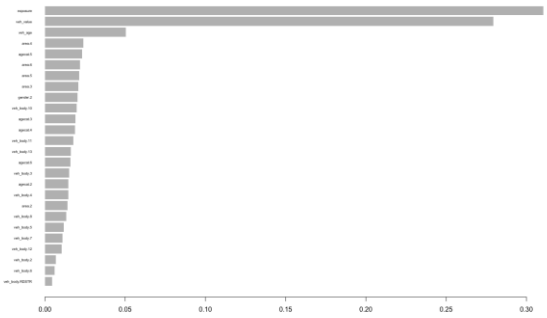

Fig. 13. Importance distribution diagram of original data variables.

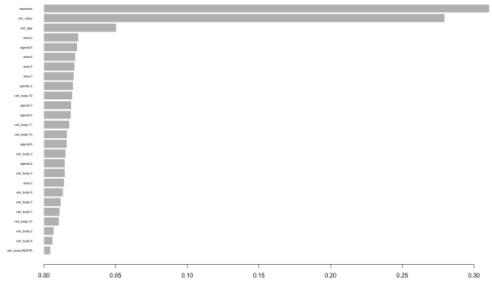

Fig. 14. Improved data variable importance distribution chart.

V. Conclusion

This paper takes "XGBOOST method to establish auto insurance pricing model" as the core, and uses different natural loss functions to model and predict the accumulated claim amount. Although this paper uses a variety of integrated learning methods to improve the performance of the model, but this paper needs to be further improved and improved in the following aspects:

1. Model building: Limited by the limited number of eigenvalues of data sets, more dimensions should be considered when constructing the model of claim probability and cumulative claim intensity.

2. Model optimization: There are many integrated learning methods, so we can consider more integrated learning methods and find a better and faster way to determine parameters.

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

Yang Li conceived the study, participated in its design, coordinated the research, performed the data collection and carried out the initial data analysis; Zhao Jinyan provided guidance and revision for the paper; both authors had approved the final version.

Funding

This paper is the research result of Hebei University of Technology's 2024 provincial postgraduate demonstration course project (project number KCJSX2024011_, funding number: 94/220079).

References

[1] M. Denuit and S. Lang, "Non-life rate-making with Bayesian GAMs," Insurance, vol. 35, no. 3, pp. 627–647, 2004.

[2] J. Pinquet, "Designing optimal bonus-malus systems from different types of claims," *Astin Bulletin*, vol. 28, no. 2, pp. 205–220,1998. DOI:10.2143/AST.28.2.519066.

[3] W. W. Sun, "Application of generalized additive model based on tweedie class distribution in determining car insurance rates," *Journal of Tianjin University of Commerce*, vol. 34, no. 1, pp. 60–67. DOI:CNKI:SUN:TSXY.0.2014-01-010.

[4] W. W. Sun and W. K. Chen, "Application of limited mixed distribution in determining car insurance rates," *Systems Engineering*, vol. 34, no. 5, pp. 144–153, 2016. DOI:CNKI:SUN:GCXT.0.2016-05-022.

[5] X. R. Chen and L. Yue, "Strong consistency and convergence speed of quasi maximum likelihood estimation in generalized linear models," *Science in China Series A Mathematics (in Chinese)*, vol. 34, no. 2, pp. 203–214, 2004. DOI: 10.1360/za2004-34-2-203

[6] Y. F. Huang and S. W. Meng, "A non life insurance reserve evaluation model based on thick tailed distribution," *Systems Engineering Theory and Practice*, vol. 40, no. 1, p. 13, 2020, DOI:CNKI:SUN:XTLL.0.2020-01-004.

[7] Z. H. Wei, T. B. Xia, and H. P. Wang, "Related party transactions, analyst behavior, and stock price synchronicity: an empirical study based on Chinese listed companies," *Accounting and Economic Research*, vol. 34, no. 05, pp. 3–27, 2020. DOI: 10.16314/j.cnki. 31-2074/f.2020.001

[8] L. Z. Zhang and W. W. Sun, "Logistic model analysis of factors influencing the probability of car insurance claims," *Insurance Research*, vol. 07, pp. 16–25, 2012. DOI:10.13497/j.cnki.is.2012.07.001.

[9] X. H. Wang, S. W. Meng, Y. S. Wang, "Pricing model and application of automobile insurance based on thick tail loss distribution," *Insurance Research*, vol. 04, pp. 67–78, 2017. DOI:10.13497/j.cnki.is.2017.04.005.

[10] Z. Li, "Cramer Lundberg approximation of Poisson risk model for dependent claims," *Journal of Mathematics*, vol. 30, no. 03, pp. 480–484, 2010. DOI: 10.13548/j.sxzz.200.03.017