# Predicting Vehicle Insurance Purchasing by Machine Learning Algorithms

Yangyang Wan *

School of Statistics and Data Science, Nankai University, Tianjin, China

* Corresponding author: 2012446@mail.nankai.edu.cn

**Abstract.** This report intends to address the problem of predicting customers' willingness to be insured. To solve the problem, this report uses SMOTE over-sampling and Near Miss under-sampling to solve the data imbalance, establishes eight basic or ensemble models, including Logistic Regression, Decision Tree, etc., and compares the model strengths and weaknesses by using f1_score as a measure. The results of the models represent that the effects of over-sampling are better than under-sampling, and the results of the ensemble models are overall better than the basic models. The best method is over-sampling combined with Adaptive Boosting or Extreme Gradient Boosting. The highest f1_score among all the results is only 0.4, which means that all the methods mentioned in this report are limited in their ability to solve this problem. The methods for solving data imbalance, the prediction models, and the ensemble algorithms mentioned in the report are of high application value. This report expects the existence of models and methods that can significantly improve the prediction effects of this dataset.

**Keywords:** Vehicle Insurance, Machine Learning, SMOTE.

## 1. Introduction

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

As socio-economic development continues to grow, different types of property insurance are emerging. As automobiles are more and more popular, and road traffic accidents are transformed into a common social threat and safe hazard, the vehicle insurance industry is developing rapidly. Vehicle insurance is a kind of commercial insurance that is liable for compensation for personal injury or property damage caused by natural disasters or accidents to vehicles. The operation of vehicle insurance will have an important impact on the actual profit of insurance companies, and its demand will also play a decisive role in the future development trend of insurance companies [1].

In this instance, an insurance company that provides health insurance to its customers wants to know if customers from the past year would also be interested in the vehicle insurance offered by the company so that they can formulate sales strategies and communication tactics to maximize the interests of insurance companies. The problem needs to be solved by building machine learning classification prediction models. And this report follows the idea that modeling is result-oriented, and pays full attention to the ability to apply and generalize models.

Four basic models are applied, including Logistic Regression, K-nearest Neighbors, Support Vector Machine, and Decision Tree. In addition to these, in order to improve the prediction effects, three ensemble models are established based on Decision Tree, including Random Forest based on the Bagging algorithm, and Adaptive Boosting and Extreme Gradient Boosting based on the Boosting algorithm.

The results after over-sampling are better than after under-sampling, and the best prediction after using over-sampling is using the Adaptive Boosting with an f1_score of 0.4 and a precision of 0.27. In addition to this, Support Vector Machine and Random Forest have an f1 value of 0.39 and Extreme Gradient Boosting has a precision of 0.31.

## 2. Data Processing

The dataset is obtained from the Kaggle website code called Vehicle Insurance EDA and boosting models. The dataset containing the vehicle insurance information is open source and publicly available [2]. In this report, 1000 of the original data are randomly selected to use. No null values in the dataset. There are a total of 12 variables in the dataset, containing 9 categorical and 3 numerical variables. The categorical variables include a binary target variable (1/0) which is interested or not interested, and a unique indicator variable. The information on each variable is shown in Table 1.

**Table 1.** Description of Variables

| Names | Descriptions |
|---|---|
| id | Unique Identification of Samples, Categorical |
| Gender | Categorical, including Male/Female |
| Age | Numerical, Values from 20-80 |
| Driving_License | Categorical, including 1/0 for having or not |
| Region_Code | Categorical, including 50 values |
| Previously_Insured | Categorical, including 1/0 for insured in the last year or not |
| Vehicle_Age | Categorical, including 1-2 Year, < 1 Year and > 2 Years |
| Vehicle_Damage | Categorical, including Yes/No |
| Annual_Premium | Numerical, Values from 2630-97986 |
| Policy_Sales_Channel | Categorical, including 48 values |
| Vintage | Numerical, the Number of Days that the Customer has been associated with the Company, Values from 10-299 |
| Response | Response Variable, Categorical, including 1/0 |

### 2.1. Outliers Handling

The boxplots for numerical variables are shown in Fig. 1, which demonstrates the presence of outliers in the Annual_Premium variable. The outliers are dealt with using the triple standard deviation method.

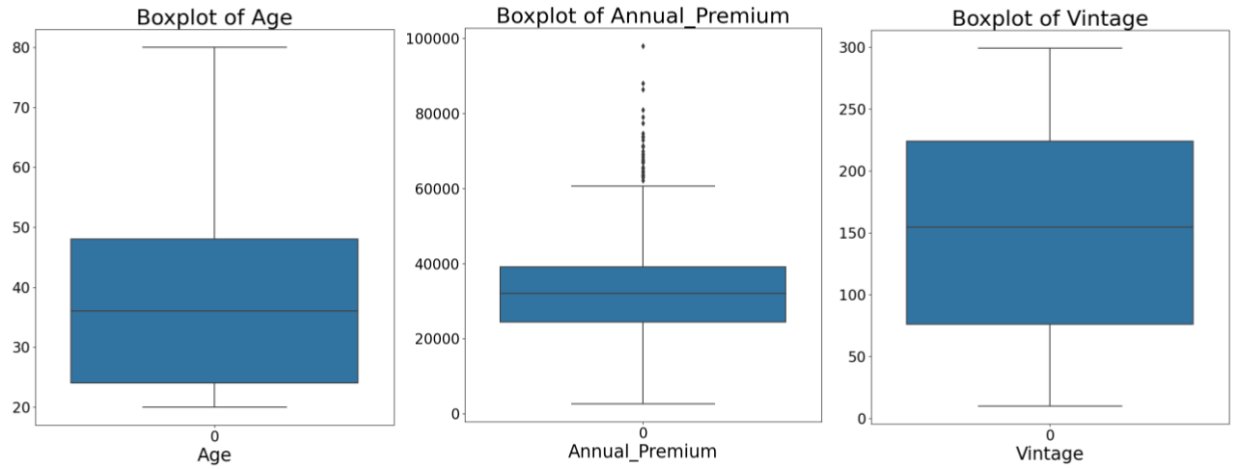


**Figure 1.** Boxplots of Numerical Variables

### 2.2. String Encoding

Recoding of categorical variables, which means that reassigning the values of all categorical variables sequentially starting from 0 [3].

### 2.3. Feature Selection

A tree model-based feature importance assessment method is used to select variables with importance greater than 0.05. The deleted variable is Gender, Driving_License, and Vehicle_Age. Besides, the id variable is meaningless and should also be excluded.

### 2.4. Feature Merging

There are too many possible values in Region_Code and Policy_Sales_Channel, and the sample size for taking some of them is too small to be meaningful, so they are merged. Combine the values in these two variables for which the number of sample values is less than 10 [4]. After performing feature merging, these two variables need to recode again.

### 2.5. Standardization

The range of values in the numeric variables is too large, so they are converted to around 0 using the mean-standard deviation method [5].

## 3. Exploratory Data Analysis

After data processing, the original dataset is reduced to a dataset with 971 samples and 8 variables. The information on each variable is shown in Table 2.

**Table 2.** Description of Variables

| Name | Type |
|---|---|
| Age | Numerical, values from -1.19 to 2.71 |
| Region_Code | Categorical, including 276 10, 151 22, and so on |
| Previously_Insured | Categorical, including 503 0 and 468 1 |
| Vehicle_Damage | Categorical, including 494 0 and 477 1 |
| Annual_Premium | Numerical, values from -1.82 to 2.14 |
| Policy_Sales_Channel | Categorical, including 358 4, 213 0, 169 2, and so on |
| Vintage | Numerical, values from -1.72 to 1.77 |
| Response | Response Variable, categorical, including 860 0 and 111 1 |

### 3.1. Response Variable

As shown in Fig. 2, the response variable has about 80 percent of the total number of values that take 0, which means that there is a high imbalance in the dataset.

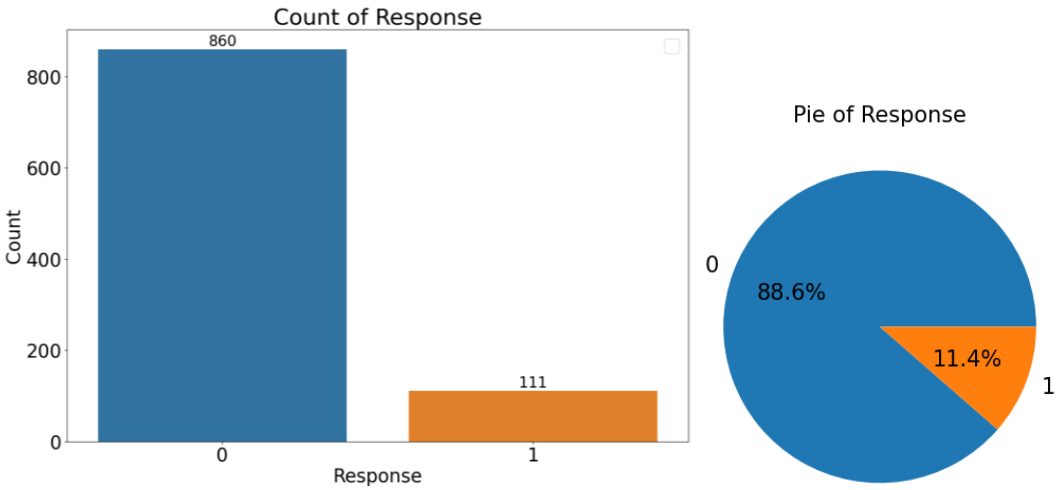


**Figure 2.** Distribution of Response

### 3.2. Categorical Variables and Response Variable

The chi-square tests are performed separately for each categorical variable and the response variable, and the chi-square statistics and p-values obtained are shown in Table 3. Setting 0.05 as the threshold, only Region_Code has a small association with the response variable, and the p-values of the remaining three variables are much less than 0.5, implying that there are extremely strong association between them and the response variable.

**Table 3.** Results of chi-square tests with response variable

| | Region_Code | Previously_Insured | Vehicle_Damage | Policy_Sales_Channel |
|---|---|---|---|---|
| Chi-square | 23.05 | 114.44 | 114.20 | 49.65 |
| P value | 0.40 | $1.05\times10^{-26}$ | $1.18\times10^{-26}$ | $1.26\times10^{-7}$ |

In order to visualize the association more closely, the distributions of these three variables grouped by response variable are shown in Fig. 3. Based on the distribution, it is surmised that if there is a sample whose Previously_Insured takes 1 or Vehicle_Damage takes 0 or Policy_Sales_Channel takes 1,3,5,6,7,8, it is highly likely to have a response variable that takes 0.

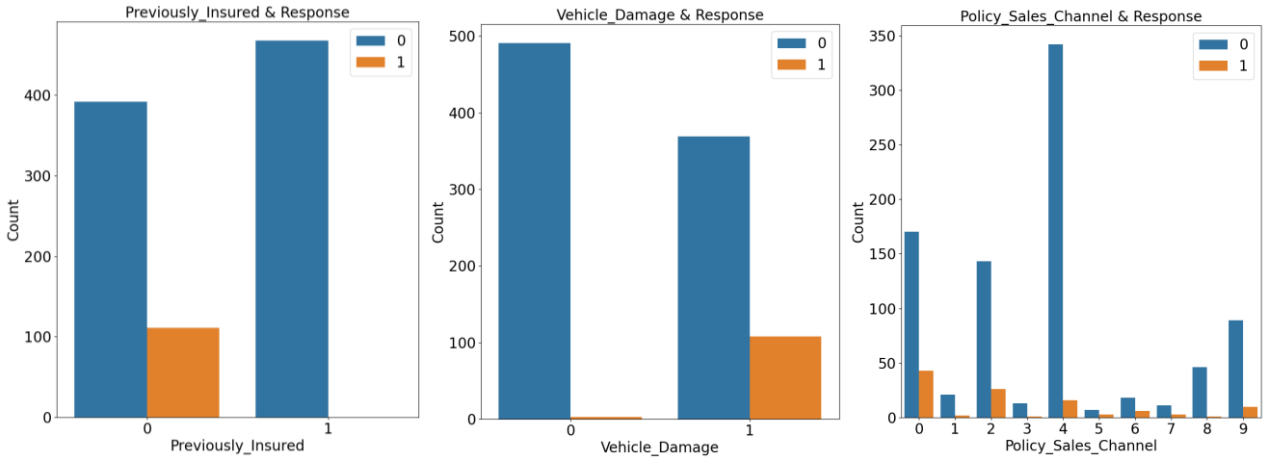


**Figure 3.** Distribution of Categorical Variables

### 3.3. Numerical Variables and Response Variable

As shown in Fig. 4, the distribution trends for each numerical variable are essentially the same when grouped by the response variable.

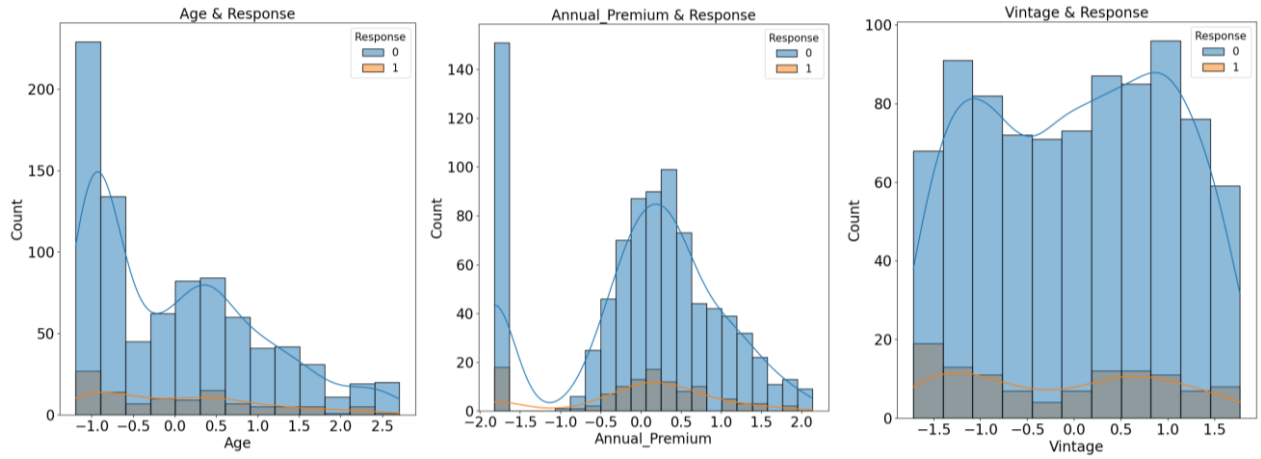


**Figure 4.** Distribution of Numerical Variables

### 3.4. Correlation Heatmap

As shown in Fig. 5, the importance of the existing features is assessed based on the tree model, and these Vintage and Annual_Premium have the highest importance of 0.25 and 0.22 respectively.

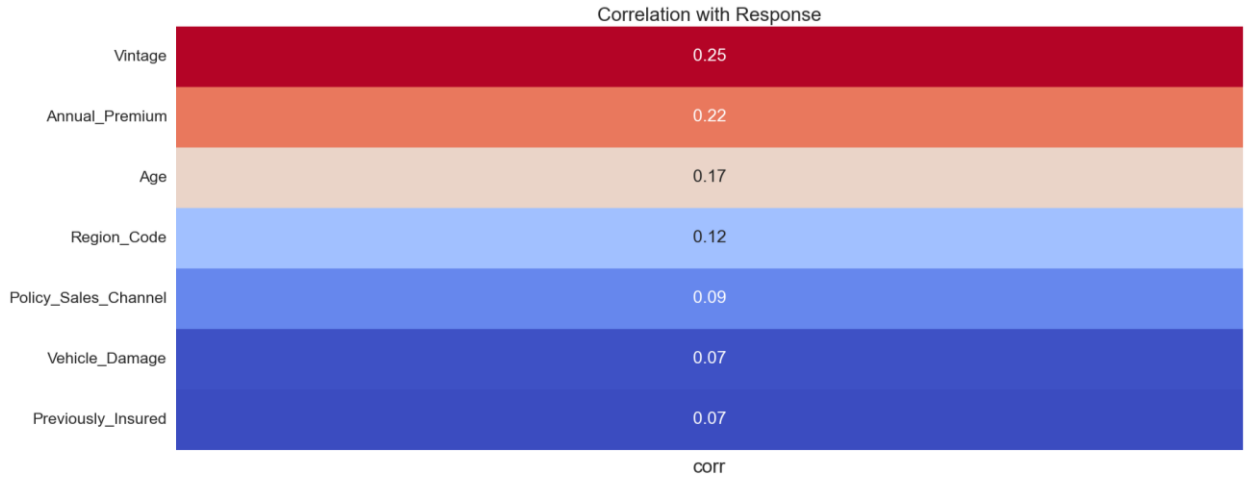


**Figure 5.** Correlation with Response

## 4. Models Building

Firstly, the basic models are built, including Binary Logistic Regression (LR), K-nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT), and then the ensemble models are built based on Decision Tree combined with two ensemble algorithms, Bagging and Boosting, including Random Forest (RF) (based on Bagging algorithm), and Adaptive Boosting (ADA) and Extreme Gradient Boosting (XGB) (based on Boosting algorithm) [6-7]. For each model, the same ideas of building and optimizing are used.

### 4.1. Data Splitting

All the data are divided into training and testing sets in the ratio of 6:4, and instead of dividing the validation set separately, cross-validation is used in each model building process. After data splitting, there are 582 samples in the training set and 389 samples in the testing set [8].

In order to avoid too large a difference in the distribution of categories in the dependent variable between the training set and the testing set, stratified sampling is used, and the ratio of 0 and 1 in both datasets after sampling is about 8.1:1.

### 4.2. Grid Search

To find the optimal parameters, the grid search algorithm is applied for each model [9]. In order to measure the model effect more accurately, 5-fold cross-validation (CV) is applied to the process, which means that the dataset is divided into five equal-sized subsets, four of which are used as the training set, and the remaining one is used as the validation set.

For each hyper-parameter combination, the performance of the models will be calculated using the specified evaluation metrics. Since the data is highly unbalanced, it is incorrect to use only accuracy as a measure of model effectiveness.

In this report, accuracy, precision, recall, f1_score, roc_auc, and balanced_accuracy is used as the metrics for evaluating model performance. Balanced_accuracy is a more robust evaluation metric for category-imbalanced datasets, which will take into account the number of samples in each category and give some trade-offs for unbalanced data.

Besides, f1_score is used as the best model performance metric, which means that at the end of the grid search, the model with the best f1_score would be selected as the final model.

### 4.3. Over-sampling and Under-sampling

To solve the problem of data imbalance, over-sampling and under-sampling should be used. It is important to note that oversampling and under-sampling can only be applied to the training sets, and not to the cross-validation and testing sets [10]. Therefore, this report uses a pipeline to integrate samplings and model instances to fit the training sets.

SMOTE over-sampling and NearMiss under-sampling are used. SMOTE generates new samples based on interpolation of similarities between minority class samples, and NearMiss removes the closest samples in the majority class samples to the minority class samples. Both methods essentially achieve data balance.

### 4.4. Model Parameters

Focusing on some of the model parameters is beneficial in addressing data imbalance. For example, class_weight can artificially assign different weights to different classes, thus placing more emphasis on samples with larger weights. Parameters such as scale_pos_weight can serve the same purpose.

## 5. Analysis of Results

The above methods are applied to each model to train the optimal model and the optimal models are applied to the testing set for prediction. The confusion matrixes and ROC curves are not represented in this report. For further insights, please reach out directly to the author for additional information.

The accuracy, f1_score, precision, and recall for all models after over-sampling or under-sampling are shown in Table 4 and Table 5 respectively.

**Table 4.** Predicted Results after Over-sampling

| | Accuracy | | | F1_score | | | Precision | | | Recall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | CV | Test | Train | CV | Test | Train | CV | Test | Train | CV | Test |
| LR | 0.64 | 0.59 | 0.66 | 0.39 | 0.37 | 0.39 | 0.24 | 0.23 | 0.25 | 0.99 | 1.00 | 0.95 |
| KNN | 0.95 | 0.72 | 0.75 | 0.78 | 0.29 | 0.27 | 0.72 | 0.21 | 0.20 | 0.87 | 0.51 | 0.41 |
| SVM | 0.64 | 0.64 | 0.66 | 0.39 | 0.39 | 0.39 | 0.24 | 0.24 | 0.25 | 0.99 | 0.98 | 0.95 |
| DT | 0.71 | 0.66 | 0.73 | 0.42 | 0.34 | 0.38 | 0.28 | 0.22 | 0.26 | 0.91 | 0.76 | 0.75 |
| RF | 0.62 | 0.63 | 0.64 | 0.38 | 0.38 | 0.39 | 0.23 | 0.23 | 0.24 | 1.00 | 0.98 | 1.00 |
| ADA | 0.72 | 0.71 | 0.74 | 0.42 | 0.34 | 0.40 | 0.28 | 0.23 | 0.27 | 0.87 | 0.68 | 0.77 |
| XGB | 0.95 | 0.78 | 0.82 | 0.82 | 0.33 | 0.38 | 0.75 | 0.26 | 0.31 | 0.90 | 0.46 | 0.50 |

**Table 5.** Predicted Results after Under-sampling

| | Accuracy | | | F1_score | | | Precision | | | Recall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | CV | Test | Train | CV | Test | Train | CV | Test | Train | CV | Test |
| LR | 0.57 | 0.47 | 0.57 | 0.35 | 0.31 | 0.34 | 0.21 | 0.19 | 0.20 | 1.00 | 0.98 | 0.98 |
| KNN | 0.51 | 0.45 | 0.52 | 0.21 | 0.20 | 0.18 | 0.13 | 0.12 | 0.11 | 0.57 | 0.59 | 0.45 |
| SVM | 0.59 | 0.59 | 0.60 | 0.36 | 0.36 | 0.36 | 0.22 | 0.22 | 0.22 | 1.00 | 0.98 | 1.00 |
| DT | 0.70 | 0.47 | 0.66 | 0.42 | 0.29 | 0.34 | 0.27 | 0.18 | 0.22 | 0.96 | 0.85 | 0.77 |
| RF | 0.25 | 0.14 | 0.27 | 0.23 | 0.21 | 0.24 | 0.13 | 0.12 | 0.13 | 1.00 | 1.00 | 1.00 |
| ADA | 0.69 | 0.65 | 0.67 | 0.34 | 0.28 | 0.29 | 0.22 | 0.19 | 0.19 | 0.69 | 0.59 | 0.61 |
| XGB | 0.62 | 0.50 | 0.62 | 0.37 | 0.32 | 0.37 | 0.23 | 0.19 | 0.23 | 1.00 | 0.94 | 1.00 |

The results of the above projections are analyzed as follows:

If the accuracies and f1_scores on the testing set are used as measures, it is clear that the results of under-sampling are much worse than those of over-sampling. This is because a large amount of data information is lost during the under-sampling process, thus weakening the model effect.

Considering the f1_scores on the testing set in over-sampling, it is clear that KNN has the worst results. Combining the model effects on the training set and the cross-validation set, the KNN model shows significant overfitting. The KNN model in under-sampling also has the lowest f1_score.

Therefore, it can be assumed that the KNN model is limited in its ability to handle unbalanced data, and the disadvantage may be more obvious in practical applications.

Except for the KNN model, the other models show similar results in oversampling, with generally high recall values and low precision values, which means that many actual values of 1 are predicted to be 0, again stemming from a high imbalance in the data.

Overall, the ensemble models are better than the basic models. This is because the ensemble models combine multiple weak learners so that they can reduce the bias and variance of the models, thus improving the overall prediction effects while reducing the risk of overfitting.

Using the f1_scores on the testing set as a measure, all models produce results of 0.4 and below, which is not considered a good result. It is therefore inferred that there are problems with the dataset itself and that the use of machine learning models and sampling algorithms does not lead to a significant improvement in prediction.

## 6. Conclusions and Suggestions

In this report, 7 machine learning models are used on the dataset, and in general, the results are all not very good, with low f1_socres and precisions. It is therefore inferred that the dataset suffers from an inability to solve or optimize, which affects the fitting and application of the models.

However, the methodology applied in this report is reliable and valid. For highly unbalanced data, over-sampling and under-sampling are the most effective and commonly used processing methods.

Using the f1_score as a measure, over-sampling combined with Logistic Regression, Support Vector Machine, Decision Trees, Random Forest, Adaptive Boosting and Extreme Gradient Boosting have relatively good results and minimal differences, with the optimal being Adaptive Boosting.

Although the ensemble algorithms optimize the prediction results, the optimal f1_socre is only 0.4 and the optimal accuracy is only 0.31, which cannot be considered a good result.

It is therefore expected that more efficient models will be available to solve the problem. The new model should have stronger generalization capabilities and the ability to solve data imbalance.

## References

[1] Zhu, Y. Research on the Influencing Factors of the Demand for Automobile Commercial Insurance Market in China. Zhejiang University, Thesis for master's degree, 2022.

[2] Vehicle Insurance EDA and Boosting Models. www.kaggle.com/code/yashvi/vehicle-insurance-eda-and-boosting-models. Accessed on 2023/8/4.

[3] Wang, Y. H. Research on machine learning-based network device identification methods. Guangzhou University, Thesis for master's degree, 2023.

[4] Liu, Y., Suo, L. Research on prediction of claims payment in commercial medical insurance based on machine learning methods: a new perspective introducing health behavior preferences. Journal of Central China Normal University (Humanities and Social Sciences Edition), 2023, 62 (04): 81 - 93.

[5] Li, K. X. Research on machine learning-based cost estimation for office buildings. Shandong Jianzhu University, Thesis for master's degree, 2023.

[6] Cheng, W., Yuan, D., Xiong, P., et al. Construction and evaluation of water quality index prediction model based on multiple machine learning algorithms. Journal of Environmental Sciences, 2023: 1 - 9.

[7] Wang, Z., He, L. Application of ensemble machine learning in predictive maintenance dataset. China Science and Technology Information, 2023, (14): 112 - 114.

[8] Fan, J. Q. Research on financial fraud detection model based on privacy-preserving machine learning. Chinese Academy of Fiscal Sciences, Thesis for master's degree, 2023.

[9] Chen, Y. Research on car credit default prediction based on grid search and random forest. Science and Technology and Industry, 2023, 23 (09): 116 - 121.

[10] Credit Fraud, Dealing with Imbalanced Datasets. www.kaggle.com/code/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets. Accessed on 2023/8/4.