Group 10 Machine learning project report

Topic: Credit Card Default Prediction

Team member: Zihao Li, Lunjing Yuan, Haorui Cheng, Mengyao Song

1. Introduction

Credit card business is one of the core businesses in the banking industry. Although banks gain very high profits from credit card business, they have been facing a sharp rise for credit card default rate. Hence, how to effectively predict the default of credit card holders has become one of the most challenging issues in the financial industry.

Our project constructs bank credit card default prediction models by using different machine learning algorithms. In particular, we build prediction models through, Decision Tree, Logistic Regression and Ridge and then compare the prediction effects of those three models through evaluation indexes including accuracy, recall and ROC-AUC.

2. Methodology

- Datasets

This data source contains 2 tables including application record and credit record, with 21 columns in total. The default of credit card clients' dataset can be found in kaggle at https://www.kaggle.com/rikdifos/credit-card-approval-prediction. In this dataset the definition of 'good' or 'bad' client is not given. So, we construct our label "decline" based on our analysis.

- Data preprocessing and feature selection

Before we combined 2 datasets, we first cleaned them separately. We dropped duplicated rows and built a linear model to fill nan value. Besides, we dealt with outliers in a few columns. Most importantly, we created labels, using 1 for the person who did not pay off their bill after 3 months of due date and using 0 for others.

In our analysis, feature selection will be undertaken to reduce the number of variables. So we could use the most important features to build models. For this project, the Information Value (IV), which is a common method in the financial world, will be utilized.

3. Data Visualisation

We can gain a deeper understanding of the data and visualize each data point. From the graphical illustrations showing in our Jupyter Notebook file, we summarized following results:

- Gender, realty, phone features have significant differences for credit approval, but car, work phone, email are not

- Few of the applicants have more than 2 children. The decline rate of applicants with 0, 1 or 2 children are not quite different
- Pensioner has the highest decline rate
- People with higher education have a lower probability of rejection
- Stable marriage group has lower decline rate
- Customer in municipal apartment has higher decline rate
- Low skill workers have the highest decline rate

4. Performance Evaluation

- SMOTE oversampling

After preprocessing, the data is still not ready to be used for training classifiers because of the imbalanced label; in other words, only few credit card users pay off their bills after the due date; the ratio of label 1 and 0 is 1:180; therefore, the classifiers trained by this imbalance data will have inevitable bias; luckily, the SMOTE NC algorithm, the method which can generate data by KNN algorithm to make balanced label(Wijaya and Cornellius Yudha etc , 2021), can solve this problem. After oversampling, the label 1 is equal to label 0;

- Accuracy score

Based on our results, in this credit card dataset, the Logistic regression classifier algorithm performs the best as it gets the highest accuracy score. Actually, every model implemented got a high score on training and testing data without any adjustment of hyperparameters. Finally, three classifiers are left for prediction: Ridge, Logistic regression and Decision tree. With GridSearching of hyperparameters, the Ridge got 0.94 on training data and 0.91 on test data respectively; the Logistic Regression got 0.96 on training and 0.93 on testing; and Decision tree got 0.96 on training and 0.92 on testing; the accuracy proves that the overfitting problem does not exist in the models

- ROC Curve and the Area Under Curve (AUC) score

However, since credit cards are high-risk products, some dedicated methods of evaluating standards will be more persuasive than accuracy for the situation of application of credit card. The ROC and recall, two scoring metrics that focus on true positive value, can evaluate the classifiers rigorously. However, three classifier performing excellently in accuracy does not work well in ROC and recall scoring;in roc_auc score, the Ridge got 0.99 in training data but only 0.48 on testing data; Logistic regression got 0.99 on training data but 0.48 on testing; and Decision tree got 0.98 on training data but 0.54 on testing; for recall part, the Ridge got 0.97 on training data and 0.19 on testing, the Logistic regression got 0.98 on training data and 0.06 on testing; and Decision tree got 1 and 0.44 on training and testing data respectively; obviously,

three classifiers stuck by underfitting; However, changing models cannot solve the underfitting problem here because three models' recall and roc auc scores are very low here. Hence, implementing larger data may work.

5. Conclusion and future works

The goal of our project is to use the machine learning algorithms to predict credit card default in the banking industry. According to the results, Logistic regression has the highest prediction accuracy score 93%. Decision Tree and Ridge also have good performance.

- Limitation

One of the limitations in our project is on the information value part. Information value was initially considered to evaluate the predicting power for each variable then we can filter out the variables which are more useful to our model. Hence, we construct an IV function for calculating IV value of each variable according to the formula below.

$$IV = \sum (\%\ \text{of non-events} - \%\ \text{of events}) * WOE$$

The result of information value is supposed to follow the rule in this graph(Bhalla, 2015):

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

However, when we manipulated it in practice, the chart we built indicated that the IV value of most variables is between 0.02 to 1 which is not significant. With investigating our process, we speculate this mismatch is due to different data preprocessing methods. During our data preprocessing, we use an inner join method to merge the two tables, but there is an ID mismatch issue between two tables and we must drop amounts of the IDs, which ends to only around 10 thousand entries.

- Future Work

The project is primarily focused on prediction of credit card applicants with machine learning. We have a 91% accuracy ridge classifier on the test data set. By the analysis, banks are able to evaluate the credit risk of their clients before issuing them credit cards. Furthermore, machine learning could give banks a visualization of their clients so that they can individualize unique products for different clients. By analyzing the level of customer risk and applying it to the model, banks are enabled to be more cautious in various decisions(Yashna et al., 2018). Most importantly, they can also improve their influence and monetization in the industry through the increasingly precise machine learning analysis.

Reference

1. Bhalla, Deepanshu. Weight of Evidence (WOE) and Information VALUE (IV) Explained. www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html.

2. Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802.

3. Wijaya, Cornellius Yudha. “5 SMOTE Techniques for Oversampling Your Imbalance Data.” *Medium*, Towards Data Science, 14 Feb. 2021, towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5.