MLPH final report – Predicting default risk

Kexin Li (kl4447)

1.Introduction

In Taiwan, in 2005, about half a million people did not repay their credit card debts and it finally led to a serious credit card crisis.[1] People called themselves "credit card slaves", which refers to those who cannot repay the minimum amount of credit card debt. This crisis caused many people to lose their homes and finally end their lives because of the pressure. The origins of the credit risk explosion even started with the Asian Financial Crisis in 1999.[1] Back in 1999, banks significantly lower and lowered their card issuance standards to solve the pressure of the Asian Financial Crisis. The issuance of credit card and loan balances has rapidly grown after that year due to excessive reduced risk controls and credit limits for clients. The CAGR on the number of issued credit cards and revolving credit balances of credit cards in Taiwan exceeded 20% from 1998 to 2005.[1] This dramatically spread economic phenomenon was finally out of control at the end of 2005 and lasted four years to end. This credit crisis caused the regulatory authorities in Taiwan to pay serious attention to the Banking Association.

2.Related Work

Most of the literature research on credit card default prediction are related to neural network or SVM to construct a set of scoring indicators of applicant credit, but those researches are all related to the early prevention or complicated methods.[2,3] This project will use real bank account data and apply machine learning methods that are easy to understand to predict the risk of credit card default. The unique part is that both under sampling and oversampling methods will be applied to deal with the imbalanced data in this research to find out which method is compatible with banking data. The motivation of the project is to offer a warning system to banks by telling bankers whom clients are facing the risk of being unable to repay in the next month to avoid another crisis happen in Taiwan.

3.Methods

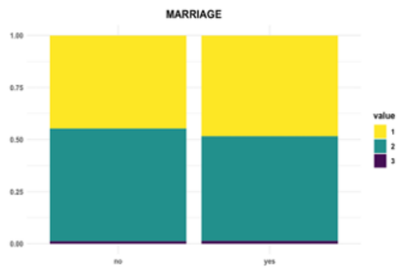
This project will focus on the real banking data of credit card clients collected by UCI in Taiwan from April 2005 to September 2005. The project goal is to predict if clients will default on their card payments in the next month. The correlation between each variable and the response variable will be detected and the relationship could be seen from graphs. The variables of repayment status from April to September and the amount of given credit will influence the result of prediction a lot during the process. The prediction would apply supervised learning methods including logistic regression, k nearest neighbor, and random forest to make the classification prediction. In this report, a default probability of more than 50% is marked as a default client, and a default probability of less than 50% is a normal client. After applying the models to train the data, the random forest gives the highest accuracy by cross validation by comparing the accuracy of cross validation from logistic regression and k nearest neighbor. Then, we test the prediction performance by applying the best accuracy values that we got from cross validation such as the best k number of 9 at k nearest neighbor and the best feature of 4 at each node. According to AUC and F1 score, random forest is the best model for us to make predictions of default risk. The sample imbalance might have a great effect on the model performance. The prediction model performance is significantly improved on the balanced data set constructed by oversampling, so we also applied under-sampling method on balancing the data and compare the model performance.

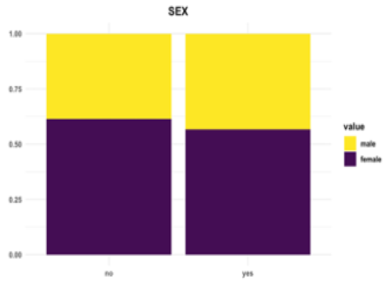
4.Data and Experiment setup

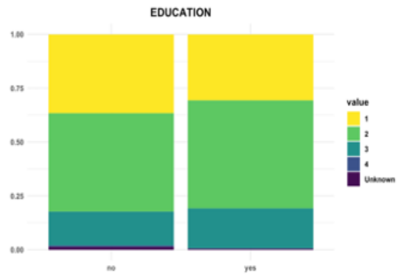
4.1.explortary data analysis

The amount of given credit in NT dollars, gender, education, marital status, age, repayment status from April to September, amount of bill statement from April to September, and amount of previous payment from April to September are the predictors. The response variable is whether the clients will default next month. There is no missing value in the dataset. The distribution of the response variable default is class-imbalanced. The number of clients who cannot repay on time is 6636 which takes about (22%), while 23364(78%) clients repaid on time. The proportion of the default clients and clients in a good stand is about 1:4 which is very imbalanced. This project will balance the data in the training set.

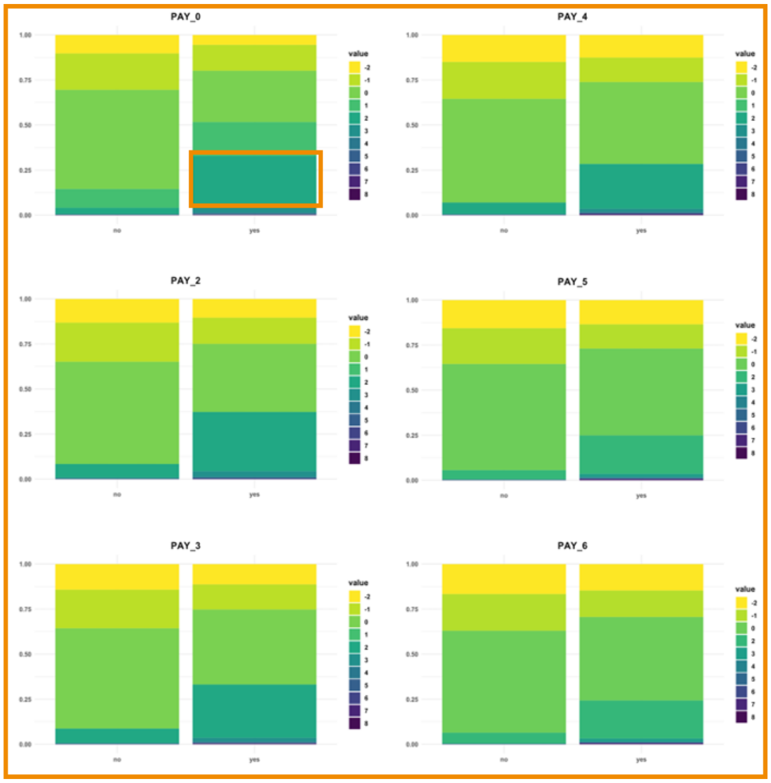
This project factored in the categorical variables and use bar plots to check the correlation between each categorical feature and the response variable. The following three plots show the distribution of all levels of features among the two groups is quite similar. These three features may not have effects on predicting default risk.



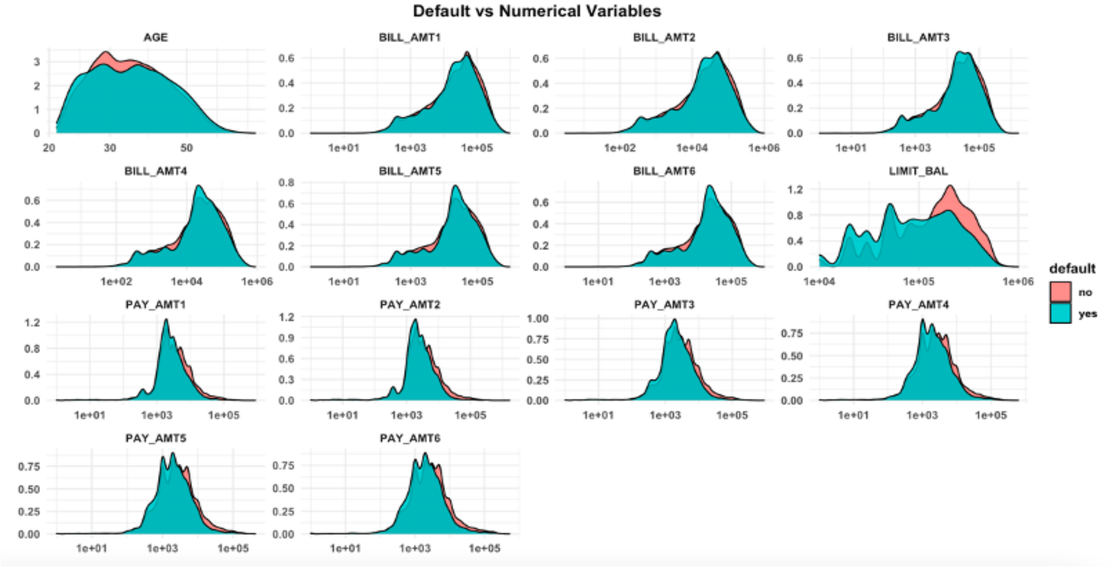




The distribution of each level of repayment status from April to September is quite varied. For example, clients who have payments delayed for two months are more likely to default. These six features may have effects on predicting default risk.



Density plots are used to describe the distribution of the frequency of numerical variables. The following 14 plots shows the correlation between each numerical variable and the response variable. The distribution of the frequency of two classes in response variables is quite similar except for the variable limit_bal which is the amount of given credit in NT dollars. This means that the limit_bal may have effect on predicting default risk.



4.2.data processing

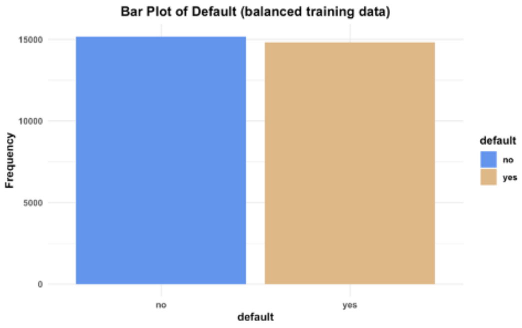
The dataset was split into training data and test data with 70% as the training data and 30% as the test data. After splitting the data, standardization is applied to the training data to prevent features with a large scale of value to become the most determinant of the final classification. For example, in the models of logistic regression, the k nearest neighbor is sensitive to the magnitude of features, so we need to avoid the loss of information to standardize the numerical variables. The test set will also be standardized by using the same standardization rule. The standardized rule of numerical features followed the following formula:

$$Z = \frac{x - \mu}{\sigma}$$

There is one more important step in the data processing which this dataset needs to solve the important problem of class imbalance. This problem was observed during the dataset introduction, and it continually exists after splitting because the split follows randomness. The proportion of no and yes classes in the current training set is still 1:4. We would apply oversampling method to train the model at first and then use the undersampling method to train the model again to compare the results.

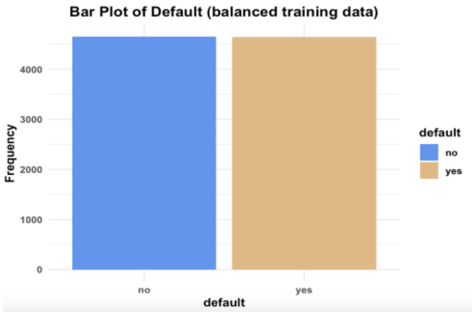
4.2.1.oversampling

The oversampling is applied to balance the training set samples and avoid the classification prediction will be affected by excessive “no” values. The balanced classes were created by artificially synthesizing samples with about 15,000 samples in each class.



4.2.2.undersampling

The undersampling is applied to balance the training set samples by using the amount of “yes” value as a benchmark to cut down the “no” values. The balanced classes were created with about 4650 samples in each class. Although the training data is reduced a lot by undersampling method, the k-fold cross validation could help to increase the accuracy of prediction.



4.3.cross validation

This project would train models based on the concept of cross-validation. The performance of the model is quite affected by the randomness of splitting when we used training data to train our selected algorithms, so the variance will be large. Cross-validation can solve the problem of the randomness of splitting. 5-fold cross-validation would apply in the training process, which randomly divided the training set into 5 parts, 4 of them are used for training the model each time, and the remaining one is used for prediction. All the training data are used in the training process by cross-validation, so the results will be more accurate.

5.Results

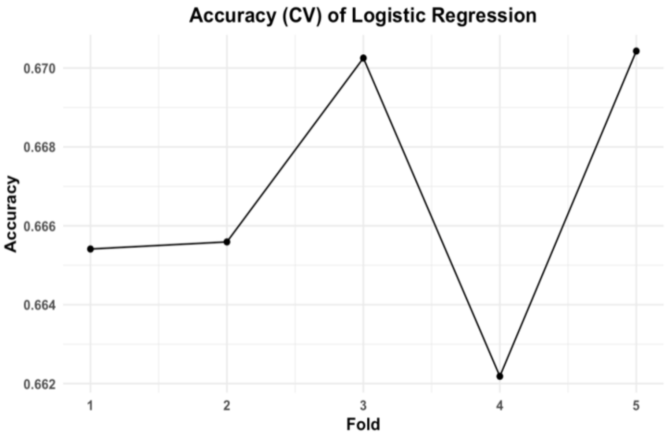
5.1.oversampling

The advantage of logistic regression is that the original dataset does not need to follow any specific distribution, which can avoid the uncertainty caused by assumptions of distribution. Also, the training calculation speed of logistic regression on large datasets is very fast. Using logistic regression as the first model can provide a benchmark of performance to help evaluate other classifiers. The next algorithm is k nearest neighbor which is easy to understand and not sensitive to outliers. Cross-validation is to determine the optimal parameter K in the training data, which is 9. Random forest is robust in the imbalanced data, so it is useful to solve real cases. It usually runs efficiently on large datasets and gives higher accuracy than other algorithms. The number of features tried at each node and the number of trees to grow is adjusted and the number of trees starts at 100 and gradually increases to 500 and the number of features is randomly sampled from 1 to 10. After the training process, the random forest gives the highest accuracy by cross-validation comparing the accuracy from logistic regression and knn.

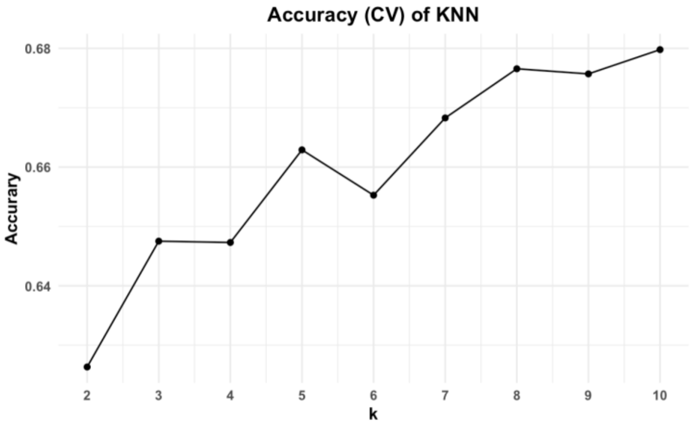
Algorithm	**Accuracy**
Logistic Regression	65.68%
K-nearest Neighbor	73.82% (best k = 9)
Random Forest	82.37% (mtry = 4, ntree = 300)

5.2.undersampling

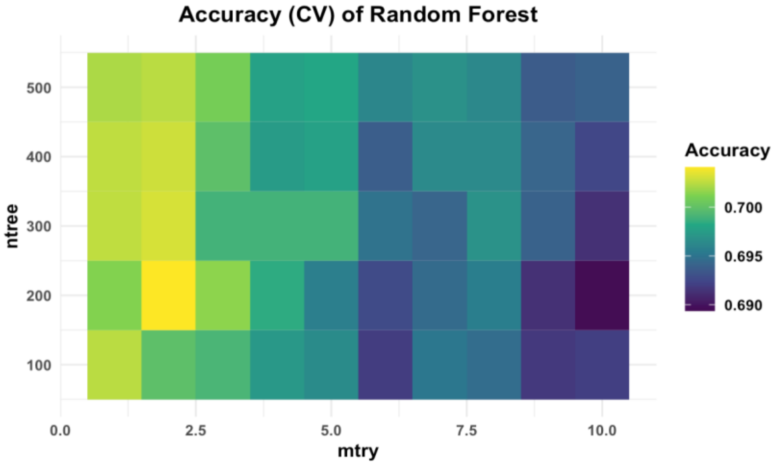
The average accuracy of the model is about 66.68% by applying cross-validation.



The graph shows the highest cv accuracy (67.98%) is obtained when k is equal to 10.

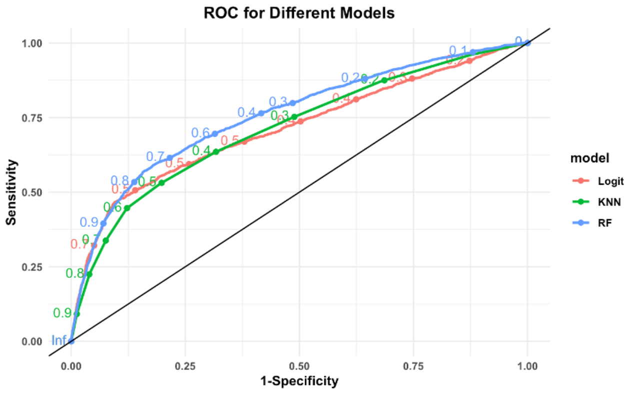


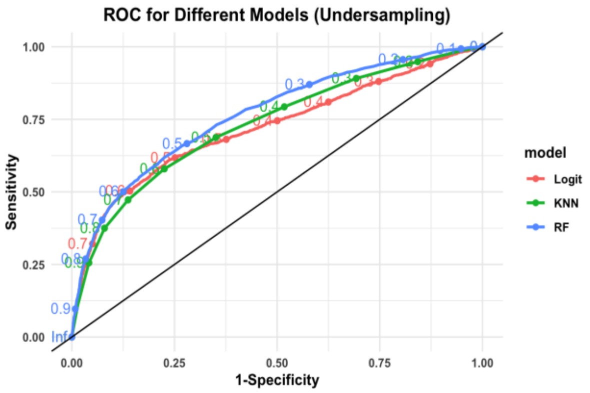
The best feature is 2 at each node with 200 trees to grow with the best accuracy of 70.41%. The results show that the random forest still gives the highest accuracy in undersampling method by cross-validation comparing the accuracy from logistic regression and KNN.



5.3.model evaluation

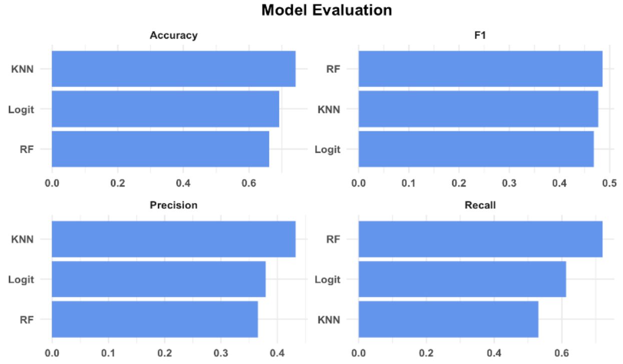
After training the models, we would like to make our prediction on the test set to have model evaluation. ROC gives random forest is the best model which has the largest area under curve. Random Forest is also the best model by comparing the AUC on both algorisms trained by oversampling method and undersampling method. The algorisms trained by undersampling method have higher accuracies than algorisms trained by oversampling.

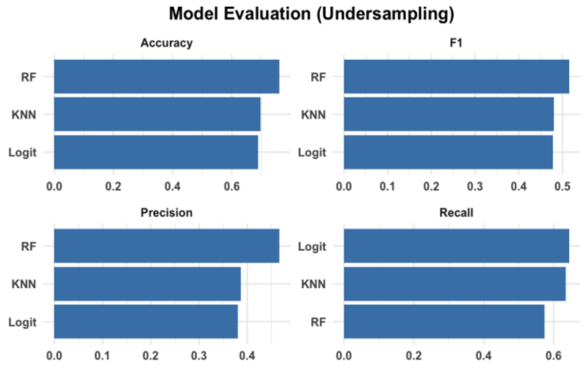




Model	**Oversampling AUC**	**Undersampling AUC**
Random Forest	0.7581	0.7694
K Nearest Neighbor	0.7154	0.7303
Logistic Regression	0.7137	0.7196

The purpose of this report is selecting clients who are likely to default, and banks act on that. Therefore, recall is a more important index to consider. However, a low precision will make some clients who are in good standard to falsely fall to the default list, which is also not good for banks. Therefore, F1 score is a better evaluation standard, which is the weighted average of Precision and Recall. Random forest is the best model by comparing the F1 Score which giving comprehensively consideration on Precision and Recall are also better than accuracy in the imbalanced class distribution. The results from model evaluation by undersampling is also significantly higher than oversampling.





	Oversampling				**Undersampling**			
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
KNN	0.742	0.433	0.531	0.477	0.697	0.387	0.635	0.481
Logistic	0.693	0.379	0.613	0.468	0.689	0.381	0.644	0.478
RF	0.662	0.366	0.719	0.485	0.761	0.467	0.574	0.515

6.Discussion

This analysis set up models by applying the logistic regression, k nearest neighbor, and random forest on the banking data of credit card clients. After comparing the performance of the above models which were trained by undersampling and oversampling datasets, the random forest model is the best classification algorism and has the best model performance. Constructing balanced training data can significantly improve the prediction accuracy of the model. Although the balanced training set has only 4650 samples in the undersampling method, cross-validation is applied to improve the prediction ability of the samples. The application of standardization and cross-validation also increased accuracy and effectively prevent models from having problems of overfitting.

We could use feature selection in the future studies to obtain the most important features that affect the judgment of model and removing irrelevant features and rerun the model to improve the robustness and increase the accuracy of the model. Also, the data collection by requesting more data from other financial organizations could solve the problem of imbalanced data and obtain more real data resources to build models.

Reference

1. Lawi, A. and Aziz, F., 2018. *Classification of Credit Card Default Clients Using LS-SVM Ensemble*. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/document/8780427> [Accessed 10 April 2022].

2. Hassan, M., 2020. *Credit Card Default Prediction Using Artificial Neural Networks*. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/343126888_Credit_Card_Default_Prediction_Using_Artificial_Neural_Networks> [Accessed 10 April 2022].

3. Kao, M., 2011. *Impact of the financial crisis and risk management on performance of financial holding companies in Taiwan*. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/290024392_Impact_of_the_financial_crisis_and_risk_management_on_performance_of_financial_holding_companies_in_Taiwan> [Accessed 15 April 2022].

Contribution：Kexin Li for this project.