Default Car Loan Prediction with Classification Model

Pornpanit Rasivisuth

Introduction

Peer-to-Peer (P2P) lending is an alternative financing solution for individuals and organisations. The P2P lending is the marketplace platform that matches individual lenders with borrowers, and lenders will receive interest in return. The demand for P2P lending has increased than ever before, given the global transaction volume in the consumer segment of $92,695.5m in 2021 and is expected to grow by 7.75% in 2025 (Statista, 2021). Similar to traditional banking lending, lenders bear the default risk that the borrowers cannot pay principal and interest rate resulting in loss of revenue (Foottit, Doyle, & Tomlinson, 2016). The P2P lending uses a bank-like evaluation process using credit-scoring approaches and identifies potential borrowers with acceptable credit risk. The default risk identification can be automated by machine learning methods and the availability of a large dataset.

This report is inspired by Turiel and Aste (2020) work to access the historical dataset of loans published by the Lending Club, an American P2P lender, and then used logistics regression and support vector machine classify defaulting loans. The report extends the choice of algorithms to include tree-based methods such as decision tree and random forest. The objective is to identify the best classification model that detects defaulting loans for car financing and reduces false-positive cases that may lose potential customers. This work also examines the data sampling techniques and hyperparameters tuning to improve the prediction result given an imbalanced dataset. Further analysis on feature selection is also discussed to understand the relationship with the probability of default computational used by lenders (Vadgama, 2020).

Data and Methodology

The Lending Club dataset (George, 2019) was filtered to capture the car loan information resulting in 8,929 entries with 21 features, including target variable, i.e., charged_off. The charged_off attribute contains a binary value in which one indicates the defaulting loan and zero for a fully-paid loan of principal and interest rate. The training features (see appendix A) consist of two types of attributes related to borrowers (e.g., length of employment) and loan characteristics such as instalment and the total amount of loan. The dataset must be analysed and applied with feature engineering to obtain meaningful insights prior to the prediction. All processes were done in Python, and the source code is available in appendix B. The analysis identified few features that are insignificant for this problem. For example, the purpose column can be removed as all rows have the same value. Assuming that the date when using this model is unknown, training the data considering the loan issue date does not generate any value for this use case; therefore, the issue_d was also removed. Various feature engineering techniques were used, including filling in missing data and converting categorical data to numerical data by one-hot encoding. One-hot encoding creates a new binary attribute per categorical value when it is more suitable for non-ordinal features, e.g., application_type (Gron, 2017). The data was also normalised due to its

varying scale. As a result, the dataset size was updated to 8,898 with 24 attributes, including the target variable.

The next step is to split the dataset into training and testing sets with the proportion of 9:1 with shuffling to eliminate time-dependency of issue date. The dataset is imbalanced with 85.2% of fully-paid debt and 14.8% of defaulting loans. Hence, the prediction algorithms may not be able to distinguish between two classes due to the small minority class in the training set. Oversampling and synthetic minority oversampling (SMOTE) techniques were applied to solve this issue. Oversampling adds copies of the minority class, i.e., defaulting loan with replacement to have the same size as the fully-paid samples. Alternatively, SMOTE combines both oversampling and undersampling by generating synthetic samples from the minority class. Both techniques will increase the default loan label from 1,200 to 6,808 entries, equal to the fully-paid samples. This report compared random forest performance given imbalanced data and two balancing techniques and the best method was selected to examine other classification algorithms, including decision tree, logistics regression (LR), and support vector machines (SVMs).

The decision tree consists of the path containing classification rules and assigns the class label on each leaf node. This decision tree is a fundamental component of the random forest, an ensemble method that overcomes the decision tree's overfitting problem by averaging many trees. In contrast, the logistic regression identifies the best hyperplane that distinguishes binary regions using a logistic function and yields the probability, indicating if the data is classified into a particular class. The support vector machine also follows a similar approach to LR, but it aims to find the optimal hyperplane to maximise the distance of examples closed to the hyperplane.

There are various hyperparameters for each model, which can be automatically optimised. Grid search and randomised search are two standard techniques to find the combination of hyperparameters values that gives the best score. The difference is that randomised search randoms combinations of the model's hyperparameters to save computational time rather than evaluate all possible combinations (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The K-fold cross-validation optimises the find-tuning process by splitting the training set to k subsamples and iterate to train and validate the hyperparameters search for k times in the loop. LR and decision tree apply randomised search to obtain computational time benefit. The grid can be set to maximise a particular performance metric.

The classification model and hyperparameter tuning were evaluated by F1 scores, a mean of precision, and recall. It is a suitable metric for an imbalanced dataset to maximise true positive classification while taking both false positive and false negative into account. Accordingly, Recall and area under the Receiver Operating Characteristic curve (AUC-ROC) can support prediction performance discussion. The recall is the measurement of the model that correctly identifies true positive, i.e., defaulting loan. The AUC-ROC informs how well the classification performs by calculating the area under the ROC curve, constructed from the benchmark of true positive rate and false positive rate against different classification thresholds. The accuracy score is irrelevant for this study as the model may yield the best accuracy by labelling all datasets to fully-paid loans but not capturing any defaults. Lastly, feature importance highlights the attributes that have the most effect on prediction for different models. The feature importance takes the training dataset; thus, the feature importance may not affect the out-sample result. The analysis gives insights into how features picked by the artificial model may be similar to traditional approaches - the probability of default, used by lenders. The computation takes the borrowers' information, such as credit history, and assigns a credit score, i.e., FICO score.

Model	Hyperparameters	F1 Train (D:P)	F1 Test (D:P)	Recall Test (D:P)	AUC Test
SVMs	C: 50, gamma: scale, kernel: rbf	76.7%:74.9%	34.7%:80.1%	62.6%:70.6%	66.6%
LR	C: 2.494810169511268, penalty: l1, solver: liblinear	67.3%:66.0%	33.1%:75.9%	67.8%:64.1%	66.0%
Random Forest	max_depth: 8, max_features: 0.5, n_estimators: 1000	84.4%:82.8%	32.0%:81.0%	53.9%:72.8%	63.3%
Decision Tree	min_samples_split: 3, min_samples_leaf: 1, max_depth: 9, criterion: gini	78.2%:76.1%	29.3%:76.6%	56.5%:66.1%	61.3%

Table 1: Result of four classification models trained on a balanced dataset through oversampling and hyperparameter tuning. D stands for default loans and P for paid loans.

Results and Discussion

The simple random forest model with Gini impurity and 100 estimators trained on imbalanced data gave an AUC-ROC of 50.0% with 0% for both F1 score and recall. This result showed that the imbalanced data with 14.8% of minority class caused poorly defaults prediction given out-sample, despite having AUC-ROC training score at 99.9%. Importantly, the training score indicated an overfitting issue as it yielded a minimal in-sample error, but the data labels were badly classified, given another subset of data. After balancing the training data through oversampling, the model showed an improved F1 score of default loan at 17.4% and 54.5% for AUC-ROC. It implied that rebalancing the training dataset can improve the algorithm to capture the minority classes. However, the recall score of 10.4% detected that the model still misclassified default to fully-paid loans, along with its in-sample AUC-ROC of 1, which shows an overfitting problem. Meanwhile, the SMOTE technique provided a slightly lower F1 score for default loans at 14.8% and out-sample AUC-ROC at 53.7%. A combination of oversampling and undersampling does not necessarily improve the prediction result. SMOTE also had an overfitting problem with no training error. Due to these reasons, the SMOTE method was disregarded and use the oversampling technique with hyperparameter tuning for further analysis.

Maximising the F1 score in hyperparameter tuning for four classification algorithms yielded a better score than using default parameters provided by the scikit-learn library. None of the models was observed to have an overfitting issue. SVMs with Radial Basis Function gave the best F1 score among the four models. Testing F1 scores were 34.7% for defaults and 80.1% for fully-paid with AUC-ROC at 66.6% (figure 1). The recall score was also balanced for two classes. However, the AUC-ROC was not as high as expected due to the misclassification of 228 fully-paid loans to default, resulting in increasing false positives. The second-best model is LR with the testing F1 score of 33.1% for defaults and AUC-ROC at 66.0%. The out-sample recall score of defaults was surprisingly higher than SVMs by 5.2%, but its confusion matrix showed higher false positive cases of 278, thus, decreasing the F1 score. The threshold of LR can be adjusted from 0.50 to 0.45 to improve minority class prediction. But it resulted in additional false positives and produced the F1 score of default at 31.6%, which is lower than the original decision boundary threshold. Despite the nature of tree-based modeling that should capture the data

pattern better than the hyperplane approaches, both random forest and decision tree gave low F1 scores of defaults at 31.9% and 29.3%, respectively. Both algorithms detected fewer defaulting loans. Nevertheless, the random forest performed well in terms of precision at 70.3% due to its lowest misclassification false positives. Furthermore, it had the highest in-sample AUC-ROC score of 83.6% among other models. Again, the report aims to capture most defaulting loans. Thus, the random forest is not the best fit to meet the criteria.

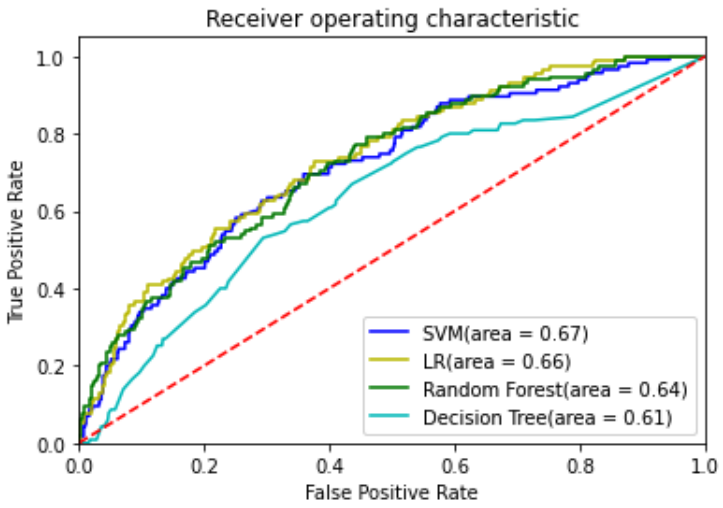

Figure 1: ROC curve of four classification models against true positive and false positive rates.

The result of feature importance by default random forest gave top eleven features with relative importance over 40% (see figure 2), despite weak correlation of features against target variable. Notably, the feature selection of SVMs is not available due to the limitation of the importance coefficient for the non-linear kernel. As expected, the FICO score matched the scoring factor used in the probability of default calculation used by lenders, along with relevant borrowers attributes such as debt to income ratio (dti), a log value of annual income, and employment length. The algorithm also assigned high importance coefficient for the loan's characteristics features, including instalment and term. Unexpectedly, the forest appointed the highest importance coefficient on the revolving line utilisation rate (revol_util) as part of FICO. However, categorical attributes, e.g., the vertification_status of income and home_ownership, were considered a low contribution by the random forest model. It could be said the model select additional attributes based on loan characteristics and did not weigh any categorical features, different from the traditional approach, that impact the classification rule assigned on each node on the tree. An additional finding was an improvement of random forest given top important features with F1 score for defaults of 33.5% and AUC-ROC at 64.6%. The overall score of the random forest is still lower than SVMs.

Conclusion

Overall, balancing the training data helps the algorithm to correctly predict default loan predictions given a low number of samples and improve its F1 score. The oversampling method

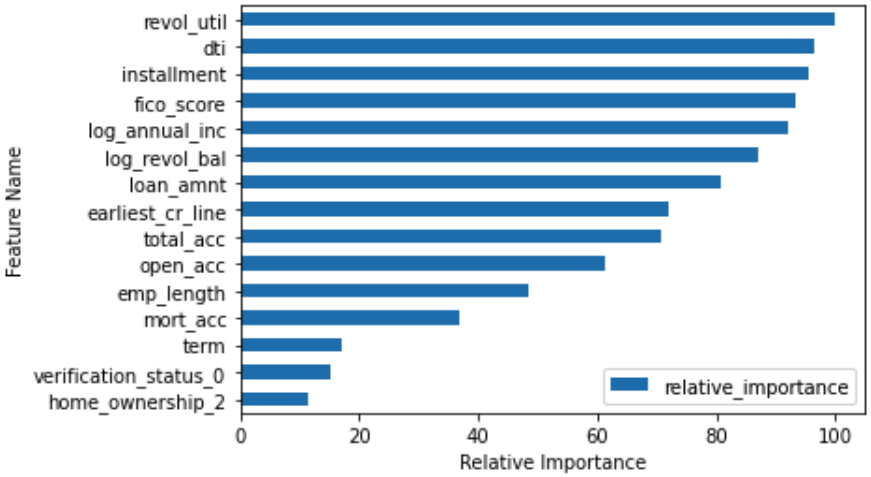

Figure 2: Relative Importance (%) generated by random forest model.

alone gave a better performance than a combination of oversampling and upsampling techniques used in SMOTE. Furthermore, hyperparameters tuning can automatically find combinations of parameters that solve overfitting issues during the training and improve the out-sample result. The search space can be extended to find the better hyperparameter set for each particular model with a computational time trade-off.

In terms of model prediction, the support vector machine is the best classification algorithm that identifies optimal hyperplane and yield the best F1 score of default at 33.1%. However, the out-sample recall score at 62.0% showed that some data was misclassified as default loans to reduce the credit risk. Notably, the F1 count was still lower than 50.0%, and the confusion matrix also gave the number of false positive cases with a precision score of 69.5%. This finding indicates that the SVMs model is still underperforming. One possible reason is the trade-off between recall and precision, which was discovered while adjusting the decision boundary threshold of logistics classification. In general, the model is still not preferable for this problem as P2P lenders will lose potential borrowers who are able to pay the debt, plus lenders still face default risk. The Youden index can find the optimal cutoff point on the ROC curve by considering true positive rate by sensitivity and specificity for true negative rate into account (Youden, 1950). The credit analyst can utilise it to balance credit risk and profitability. Examing other algorithms can be future work to use ensemble methods that combine multiple classification models or network-based approaches, i.e., a neural network for car loan defaults prediction.

The finding of feature importance showed that the random forest gave attributes weights different from the traditional approach by considering loan characteristics. There were none of the categorical features considered critical to distinguish between the two classes. Removing these attributes also improves the F1 score of the random forest model. Yet, the out-sample score was still lower than SVMs, which took all features into account. Notably, the dataset itself represents a critical role in classification and other machine learning problems. Considering external elements such as interest rate, the loan's grade assigned by the analyst, and spending habits that can predict borrowers' ability to pay the debt may improve the prediction result.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Foottit, I., Doyle, M., & Tomlinson, N. (2016). *A temporary phenomenon marketplace lending an analysis of the uk market.* Deloitte. Retrieved from `https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-fs-marketplace-lending.pdf`

George, N. (2019). *All lending club loan data.* Retrieved from `https://www.kaggle.com/wordsforthewise/lending-club`

Gron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build intelligent systems* (1st ed.). O'Reilly Media, Inc.

Sayah, F. (2021). *Lending club loan defaulters prediction.* Kaggle. Retrieved from `https://www.kaggle.com/faressayah/lending-club-loan-defaulters-prediction`

Statista. (2021). *Marketplace lending (consumer) - worldwide: Statista market forecast.* Retrieved from `https://www.statista.com/outlook/338/100/marketplace-lending--consumer-/worldwide`

Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, *7*(6), 191649.

Vadgama, N. (2020). *Comp0164 digital finance lecture notes in alternative finance.* FUniversity College London.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

Appendix

A Data Dictionary

The description of features was extracted from Sayah (2021)'s notebook and Turiel and Aste (2020)'s work.

Column Name	Description
loan_amnt	Total amount of car loan value
term	The number of months for loan payment. Possible values are 36 and 60 terms
installment	The monthly payment of loan.
emp_length	The length of borrower's employment history in years. If borrower employes less than a year, the value is zero. 10 means that they have been employed more than 10 years.
home_ownership	Borrower's home ownership status.
verification_status	Borrow's income source status whether it has been verified.
issue_d	The month and year when the loan was issued.
purpose	The purpose of loan. This dataset has single value of car.
dti	Ratio of debt and income.
earliest_cr_line	The earliest year when borrower's reported credit line was opened.
open_acc	The number of open credit lines.
pub_rec	The number of derogatory public records.
revol_util	Revolving line utilization rate.
total_acc	The total number of current credit lines.
application_type	Either individual application or a joint application.
mort_acc	The number of mortgage accounts.
pub_rec_bankruptcies	The number of public record bankruptcies.
log_annual_inc	The log value of borrower's annual income.
fico_score	Fair Isaac Corporation score which measures borrower's credit scores based on various factors such as payment history and utilisation of existing credit limits.
log_revol_bal	Total credit revolving balance.
charged_off	Default loan is 1 and 0 for full-paid loan for both principal and interest.

B Source Code

The source code is available on: https://github.com/P-Ras/ML-CW1. The repository will be set to private when the assignment feedback is published.