UNIVERSITY OF EDINBURGH
Business School

2022-23

CMSE11428 PREDICTIVE ANALYTICS AND MODELLING OF DATA

Implementation and Comparison of Credit Scoring Models for the Prediction of Loan Defaulters

GROUP 1

Word count: 3900 excluding headings, figures, tables, and references

Table of Contents

Executive Summary

Credit scoring functions as a critical determinant of prosperity and solvency of the clients in the financial and banking world. As the financial institution aims at maximising its value, the default risk among financial debt is closely monitored to ensure that it meets optimal tolerable level and condition of the offered loan. Following traditional credit scoring models, this paper employs and assesses the performance among models, specifically Logistic Regression, Random Forest, and Gradient Boosting, based on specific one-year credit card users' data. This study undertakes pre-processing of data, which includes eliminating irrelevant data and missing values, variables, excluding outliers and noise, applying PCA, transforming numerical values, and normalising the variables. The cleansed data is then inputted to examine in specified credit scoring model. Regarding the result, despite its highest AUC score of 0.6216, the Logistic Regression model tends to produce low performance of stable and reliable prediction, evaluated by precision-recall line and prediction distribution. The Gradient Boosting technique with grid search parameter tuning method results in a better performance, while it has its limitation in the case of threshold increase. Lastly, the Random Forest model is considered a great baseline. As the precision-recall curve being well above threshold, it indicates good performance of defaulting classification among borrowers. Comparatively, the feature importance in particular models is shown. Logistic Regression model does not demonstrate significant importance, while the Random Forest and Gradient Boosting models generate similar importance to age, personal income, payment day, and number of dependants, which are quite in line. Therefore, it helps confirm the prediction capability according to the importance of these variables. For future work recommendations, the quality of data and variant advanced modelling techniques are suggested to facilitate further effective mechanisms in developing the credit scoring process.

1 Introduction

Despite the technological evolution observed in the banking sector, credit risk is at paramount among other major risks associated with this industry. Consequently, a corresponding necessity emerges regarding the need for a systematic and analytical tool to evaluate potential borrowers' applications. In this context, credit scoring models which utilize a set of specific characteristics to alert banks when a potential risk is identified tend to come into use. As defined by Sadatrasoul, et al. (2013), credit scoring models are used to assess the creditworthiness of applicants by determining a credit score based on the observed characteristics of the borrower. Such practices have been established as the loaning process does not necessarily generate maximum profits and because a vast number of stakeholders are also involved in the process (Dinha & Kleimeier, 2007).

Given the importance of such models, this paper focuses on the implementation of several approaches to predict potential defaulters utilizing data of more than 40,000 loan applicants. To this end, the results of three credit scoring models, logistic regression, random forest, and gradient boosting were compared. The remainder of this report is organized as follows: Section 2 discusses the problem and the underlying complications, whereas Section 3 defines the techniques used along with their corresponding literature background. The results and recommendations are presented and discussed in Section 4 and 5. Lastly, Section 6 provides a brief conclusion of the work performed.

2 Problem Statement

Over the years, banking customer characteristics have become increasingly complicated and dynamic, challenging how financial institutions (FIs) gauge customers' credit risk. Given that high credit risk directly signals probable default of financial products, especially loans, well-designed credit risk modelling is an imperative factor determining banks' stability and profitability. Following process improvement of credit scoring, this project aims at offering accurate and solid business recommendations to help our bank client – credit analysis department – lessen its default rate. As a result, the enhanced algorithm of credit assessment will help the client suitably identify credit risk regarding its relevant customer information and detect undesirable class of customers. The bank customers with potential default risk will be recognized concerning level of tolerable criteria. This will, in turn, contribute to the higher accuracy of loan decisions.

The various predictive modelling techniques are employed to better verify the most reliable credit scoring framework, given the provided one-year credit card users' data. The challenge in the project is missing and unknown descriptive information in the dataset. Therefore, the pre-processing of data is necessary in order to extract only fitting and indicative data that have an influence on the customers' credit. Thereafter, the key puzzle in identifying sensible predictive credit scoring model will be tackled, setting apart expected uncollectable loans from the good debt. The project's outcome offers bank client suggestions of significant indicative variables and specific guidelines to develop the credit scoring model.

3 Techniques and Literature Review

3.1 Pre-processing

Data pre-processing is deemed to be an integral aspect in the process of implementing credit scoring models (Nalić & Švraka, 2018). There are different kinds of variables to be found within any bank's data set. However, research shows that factors such as age, income as along with other financial and employment information have been included in the credit scoring process (Chen & Huang, 2003). In previous credit scoring studies, key variables were found to be in forms such as numerical continuous, numerical discrete and categorical variables. However, Amoo & Raghupathi (2015) suggested that the standardization of values between -1 and 1 for numerical continuous data, and between 0 and 1 for numerical discrete and categorical data enables the model to work efficiently with different kinds of variables (Amoo & Raghupathi, 2015).

In circumstances with data imbalances, where the model's prediction quality can be highly affected. The machine learning, synthetic minority over-sampling technique for nominal and continuous variables (SMOTENC) can be applied to help solve the problem.

The SMOTENC works by oversampling the minority groups to impartially depict all classes in dataset (Ampountolas, et al., 2021). Thereafter, the feature selection can be employed to filter irrelevant variables from the model. For instance, in the study conducted by Ampountolas, et al. (2021) several variables were eliminated, in case a high correlation is found, one of the variables is omitted. Thus, improving the quality of the model by removing confounding effects and eventually avoid overfitting or under fitting situations. Sayjadah, et al. (2018) also used feature selection approaches to undertake dimension reduction and identify several significant factors driving credit card default.

Regarding microfinance's credit scoring research, statistical techniques are used to evaluate the correlation between categorical variables. Thereby, this can help in the final determination of the predictor variables to be used in the algorithm. Following this step, the categorical variables are transformed into binary or dummy variables through the gain ranker approach such that useful variables are selected for the classification process while eliminating unsuitable ones . As a result, this project considers data transformation, oversampling, and feature selection to clean the dataset before modelling for pre-processing the data.

3.2 Modelling

- ***Logistic Regression***

The logistic regression is considered to be one of the most acceptable and traditional techniques when dealing with credit scoring assessment due to its ability in providing the needed comprehensive capacity about creditworthiness to regulators. Amoo & Raghupathi (2015) conducted a study on credit scoring using logistic regression. Their research applies various combinations of variables to the logistic regression model. This combination of variables includes both variables that had been approved by credit reporting agencies and variables suggested by Amoo & Raghupathi (2015). These consist of income, debt-to-income ratio, number of credits, years at residence amongst others. The paper concluded that logistic regression is accurate, dynamic and easily interpretable for credit scoring. However, it has been found to be sensitive to alteration in factors and weights (Amoo & Raghupathi, 2015).

The hybridisation of logistic regression and principal component analysis (PCA) technique was developed to assess the credit default risk for Savings and Credit Co-Operative Society (SACCOs) (Walusala, et al., 2017). The researching process consisted of two main stages. First, PCA was employed for transforming the original variables to new uncorrelated principal components. Second, with the principal component values, the logistic regression technique was used to predict credit rating scores. This model outperforms other comparable models in the study, such as Multiple Regression with PCA and Factor Analysis-Multinomial logistic Regression, with an overall accuracy of 85%. Similarly, Muhammed & Adinoyi (2019) paper studied binary logistic regression improvement for credit scoring with PCA, in the case of commercial bank in Nigeria. The paper showed that PCA results in better classification accuracy (Muhammed & Adinoyi, 2019). This research compared the dimension reduction in predictors between PCA and factor analysis and it showed evidence that PCA better reduces collinearity.

The significant downside of using PCA as variables in a regression model is loss of interpretability. Principal components are constructed based on linear combinations of original features in the dataset. This means one cannot meaningfully determine what factors may be leading to lower credit scores based on a regression model using principal components as features.

- ***Random Forest***

Random Forest algorithm has become prevalently utilised in credit scoring models due to its accurate performance and overfitting prevention. It is also known for its functionality in solving problems related to class imbalance (Brown, et al., 2012). Ampountolas, et al. (2018) showed that multi-class classifiers perform well in a micro-credit scoring case, compared to other machine learning algorithms. The article studied credit scoring methodologies for default risk estimation applying machine learning and deep learning models, tested on extracted data from Innovative Microfinance Limited in Ghana from 2012 to 2018 (Ampountolas, et al., 2021). The key predictors focused on customer information including demographic, loan conditions and behaviour. The conclusion suggests that machine learning ensemble classifiers, namely Random Forest, XGBoost, and Adaboost, produce the best performance and highest accuracy when compared to other models. This was further confirmed by the confusion matrix, classification reports through Precision, Recall, and F1-Score as well as sensitivity analysis (ROC and AUC) which will be discussed in detail as key performance evaluation metrics.

- ***Gradient Boosting Decision Trees***

Gradient Boosting Decision Trees could bring promising performance improvement in credit scoring, especially by sequentially step-wise loss optimization while keeping the training objectives unchanged, thereby becoming a prominent technique deployed by banks to conduct credit scoring, manage credit risk, and boost profitability. Wu & Hsu (2012) suggested a decision-support model based on the Decision Tree (DT) to help with the decision-making process in regard to the assessment of credit risks. They praised the model's flexible integration of relevance vector machine and DT for its promising prediction performance and interpretability. However, this method might result in other possible issues, such as high computational cost and the vector machine's complexity.

To train the model, Xia, et al. (2017) adjusted the XGBoost hyper-parameters adaptively using Bayesian hyper-parameter optimization. While logistic regression is an acceptable and competitive classifier for classifying credit data in general, Gradient Boosting decision tree (GBDT) and random forest (RF) are typically more effective and powerful classifiers than other ensemble and single models (Xia, et al., 2017).

As proven by the research articles, several machine learning techniques will also be adopted for the credit scoring analysis in this report. Initially, the hybrid application of logistic regression along with PCA technique will be applied. Subsequently, Random Forest as well as Gradient Boosting Decision Trees will be utilized all for higher precision outcomes.

3.3 Model Performance Evaluation

Different methods are used for the purpose of evaluating the prediction performance of credit scoring models. Starting with the confusion matrix, which is a method that compares the predicted values with their actual values within a specified cut-off value as defined by (Zeng, 2020). In general, a confusion matrix provides a tabular summary where terms known as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are presented. Additionally, other terminologies such as accuracy, precision, sensitivity, and F-measure also known as F1-Score can be derived from the confusion matrix to quantify the model's performance (Zeng, 2020).

Accuracy gives a measure of the degree to which the predicted results are close to the actual value (Rahmad, et al., 2020). Zhang, et al. (2007) conducted a study that compared multiple credit scoring models given that the performance evaluation criteria relied mainly on the model's accuracy. Thus, the model with the highest accuracy value eventually outperforms other models. Precision is defined as the ratio of the true positive values to the overall values in the positive class (Louzada, et al., 2016). Recall or also known as sensitivity, is the fraction result of dividing the true positive values over those values belonging to the positive class, in other words, over the true positives and false negatives (Chopra & Bhilare, 2018). Specificity, on the other hand, is the ratio of the true negative values to all values belonging to the negative class (Louzada, et al., 2016).

As a result, the ROC curve (Receiving Operator Characteristic) can be formed by plotting the sensitivity as a function of specificity at different cut-off points. The corresponding area under the ROC curve (AUC) is a good performance evaluation measure used by many scholars within scoring models (Abellán & Castellano, 2017). Taking both precision and recall into account, the F1-score can be described as the harmonic mean of both precision and recall (Singh, 2017). It is worth noting that measures like GINI coefficient, MSE, and RMSE are also used as other evaluation metrics (Abdou & Pointon, 2011). Hence, this report will utilize the precision and recall generated from the confusion matrix, as well as the AUC score to compare and evaluate each model's performance.

4 Results and Analysis

4.1 Data Pre-processing

Throughout this report, a dataset containing one year of information on credit card customers is to be used. Yet, for the convenience of the modeling, several data preparation and cleansing steps were performed. The first began with the elimination of characteristics that had only a single value. Then, those variables deemed to be correlated had their values combined. Another important aspect to note is that the dataset contained a noticeable number of missing values as indicated in Figure 1. Therefore, variables having more than fifty percent missing values were removed, whereas other variables, such as type of employment and months of residence, were imputed in accordance with their holistic characteristics. Noise and outliers were also screened for and removed. Additionally, a selection was made on variables that seemed to provide the same kind of insight such as professional state and residential state to remove duplication of information from the model for the sole purpose of incorporating the most relevant features within the study. Finally, a few data transformation methods were executed, which included the conversion of nominal values to numerical values along with the use of MinMax scaling technique to normalize each variable.

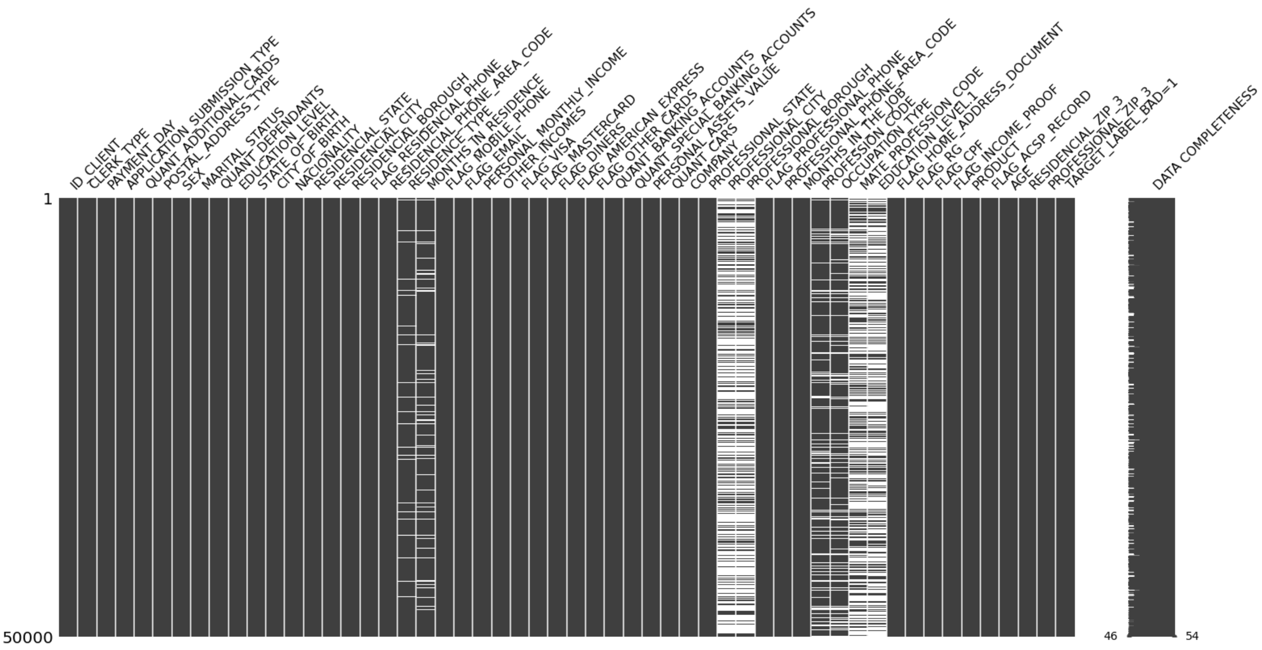

Figure 1: Visualization of missing values across features

To better understand the present dataset, count plots were used to present how many respondents held different target values for binary, categorical, and nominal variables. As shown in Figure 2, an obvious and distinct pattern was observed, demonstrating that about 73 percent of the respondents' target value to be 0. That implies that there is no clear relationship between the target type and an individual being part of a certain group. A class imbalance can be also seen across the dataset despite the feature engineering process.

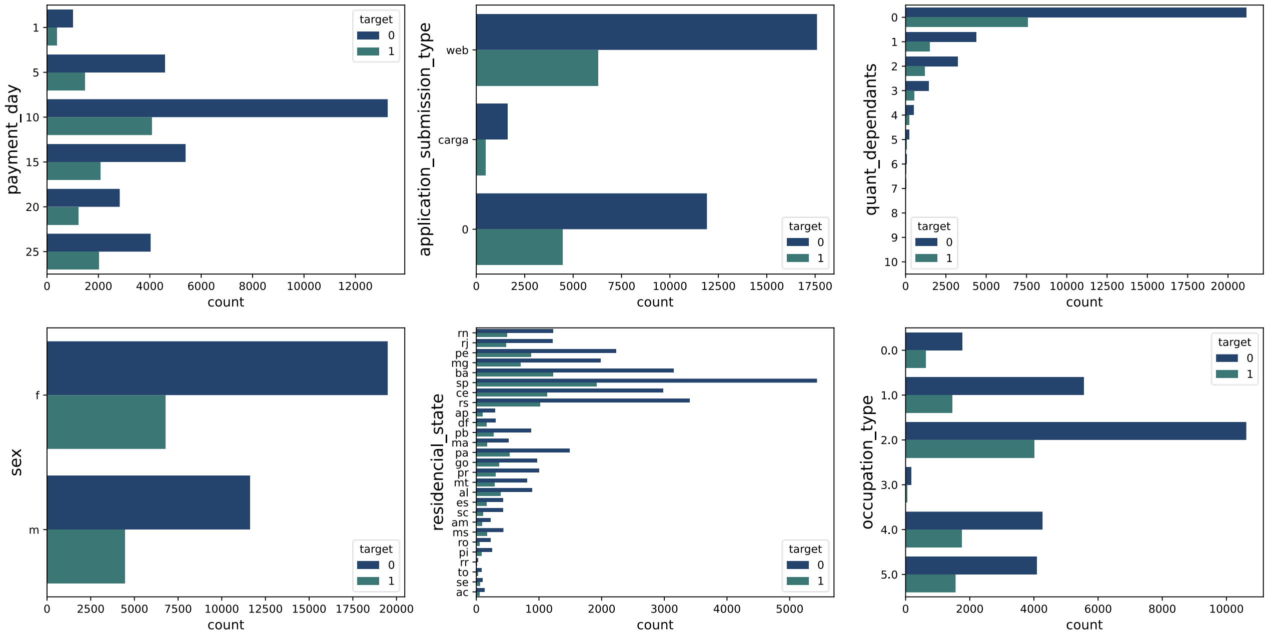

Figure 2: Count Plots for a sample of variables (Full Plots can be found in Appendix A)

Continuous variables such as age, monthly income, and months in residence for different classes were then compared. Figure 3 shows several histograms for both classes with class zero (highest frequency) and class one (lowest frequency). It shows that the shape of the distribution is not significantly different across both target classes. Additionally, most respondents earn between R$300 and R$810 each month, and most of them are, on average, between the age of 35 and 59.

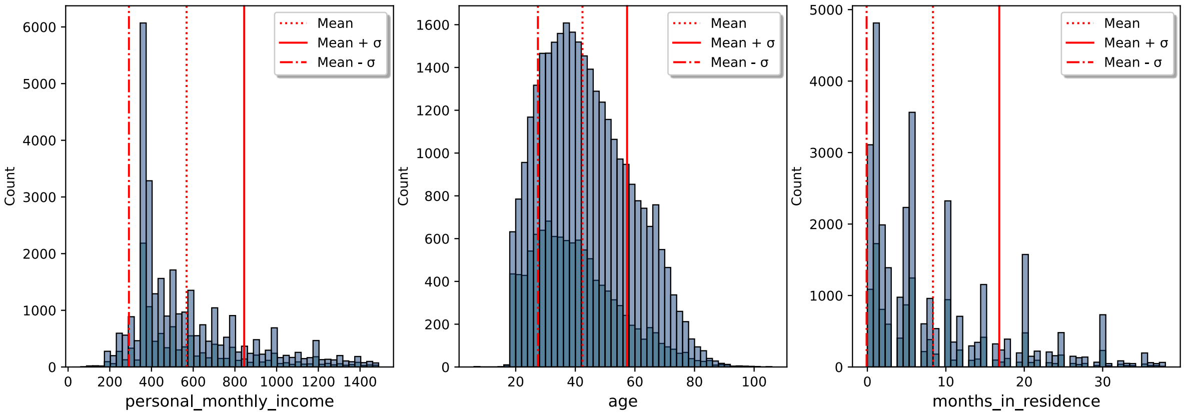

Figure 3: Histograms comparing classes of continuous variables

Figure 4 shows the interquartile range and distribution of the personal monthly income, age, and months in residence. A similar distribution of class values can be seen as in the histogram, with class 0 having a slightly larger Interquartile range. However, some outliers can still be observed after data cleaning.

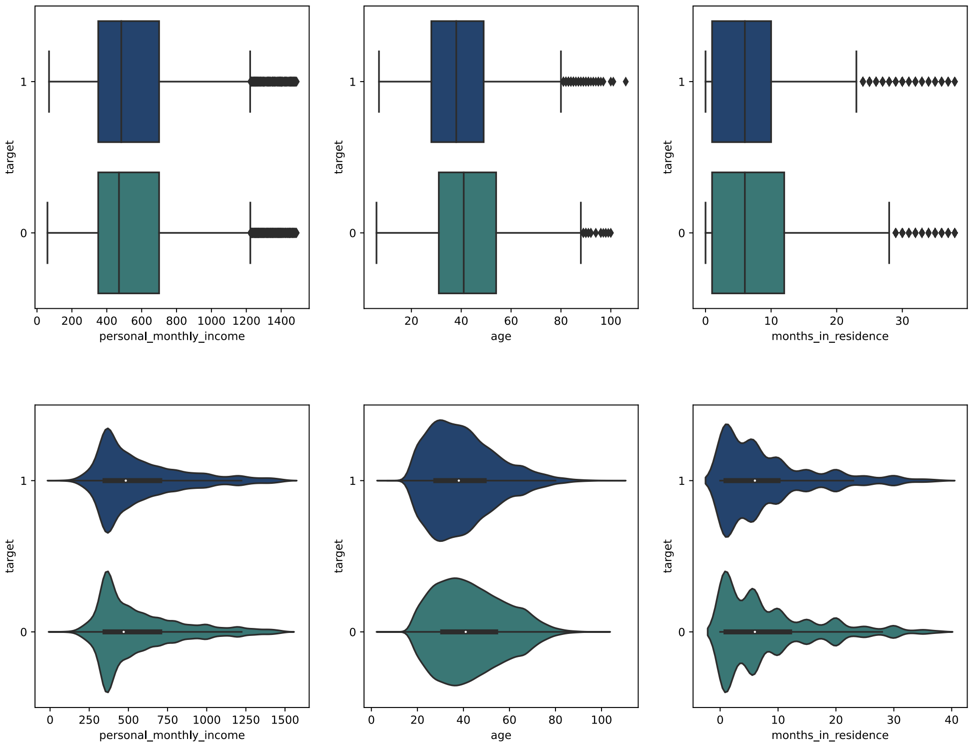

Figure 4: Box plots and Violin plots for personal monthly income, age, and months in residence

4.2 Modelling and Analysis

Before training and evaluating each of the models mentioned below, the data was split into training and test sets using sklearn's 'train_test_split' (Scikit-learn.org, 2018). In the case of Gradient Boosting cross validation was performed to split the data into both training and test sets.

- ***Logistic Regression Results***

Logistic regression has long been the standard practice in credit scoring. It is well recognized for being a good starting point in predicting potential defaulters out of a set of data. Nevertheless, it is rarely used as the primary approach and is usually used in combination with other techniques to improve performance. As a result of the large number of variables in our dataset, Lasso regularization was applied to obtain higher predictive power. Lasso regularization both shrinks values toward zero at the same time as it picks out features, making it a useful tool that is both convenient and interpretable (Wang, et al., 2015). The results obtained from the Logistic Regression are shown in Figure 5.

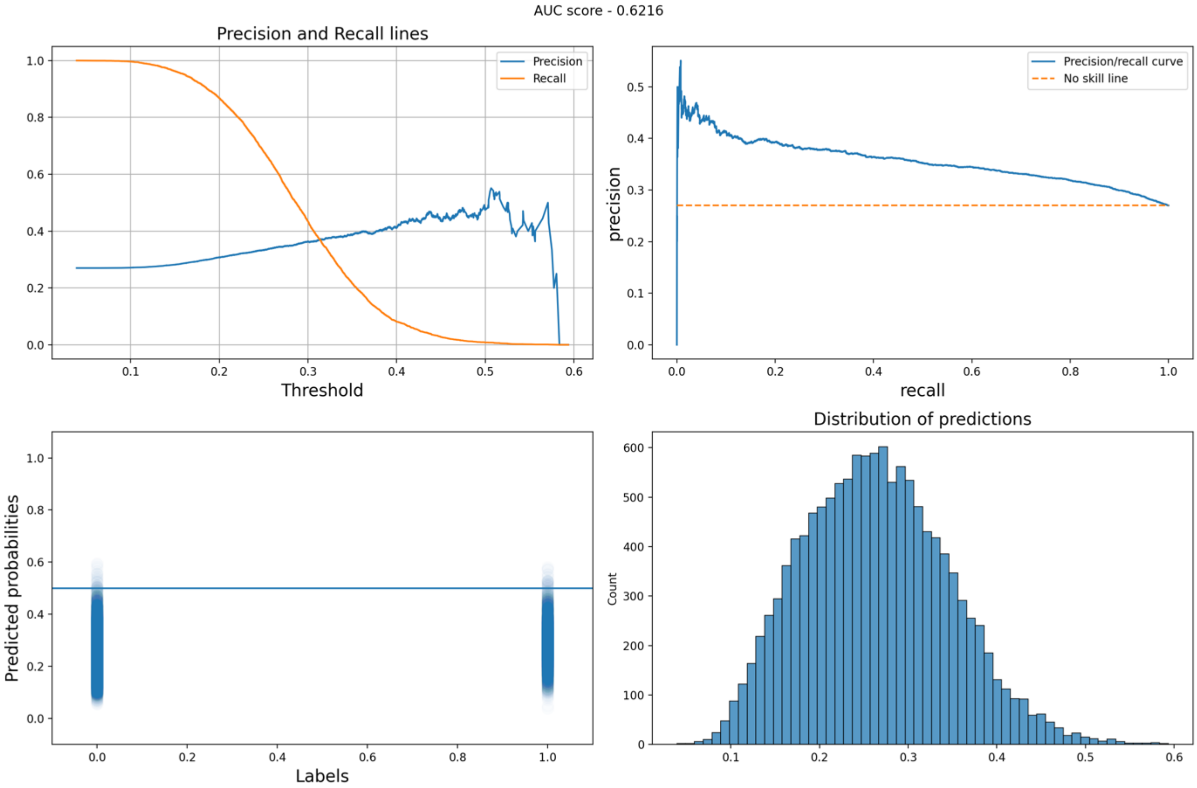

Figure 5: Logistic Regression Results

The precision-recall curve shows the tradeoff between precision and recall for different threshold levels (scikit-learn, n.d.). The precision and recall lines plot indicate that Logistic Regression produces results with low accuracy and that majority of the results are target = 0. As the threshold increases precision increases slightly and drops off to 0 at the 0.57 threshold. The distribution of predictions plot shows that most of the model predictions are probabilities of being target = 1 less than 0.5. This means that the model largely fails to predict many target = 1. The AUC of Logistic regression is 0.62 suggesting that the model can distinguish between target classes, however given the distribution of predictions this finding is likely to be unstable.

- ***Random Forest Results***

Following Logistic Regression, a Random Forest model was constructed. The model uses a minimum leaf split of 3, minimum samples split of 5, and unrestricted number of trees. The outcomes are presented in Figure 6.

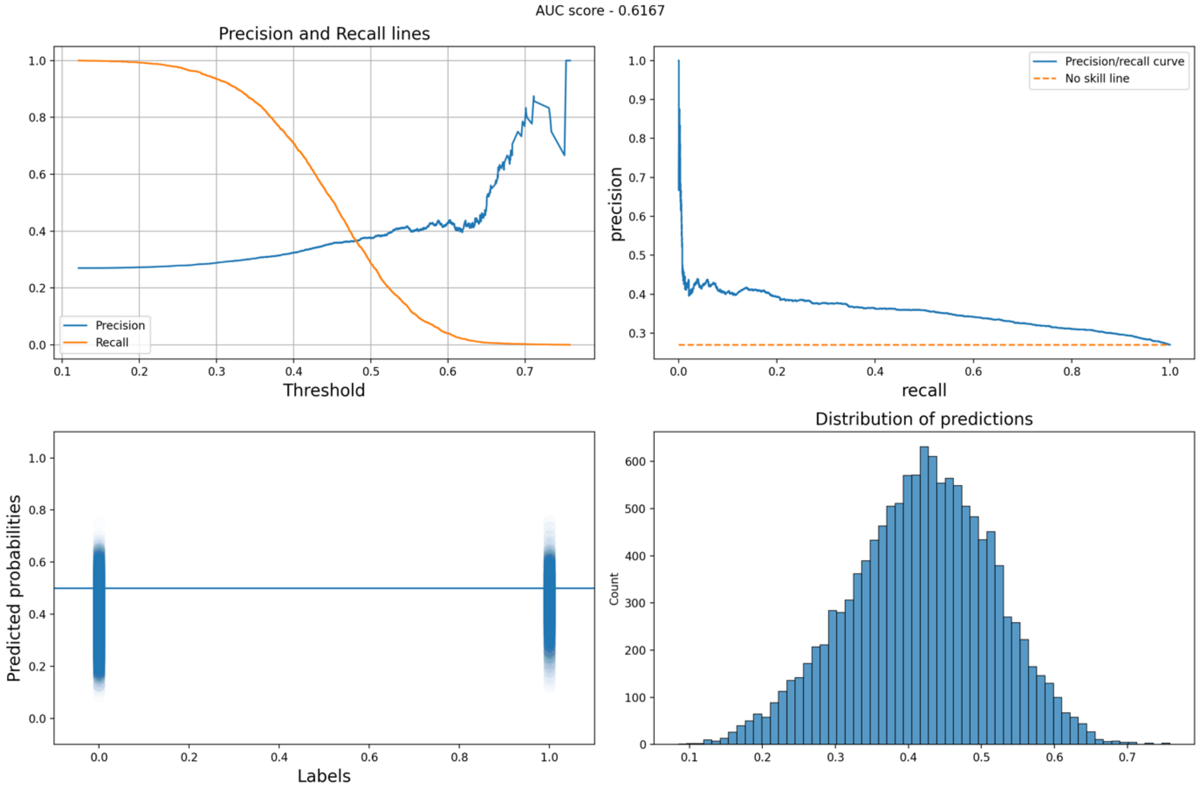

Figure 6: Random Forest Results

This precision-recall curve is like the Logistic Regression's, also indicating low accuracy and majority of results being target = 0. This model, compared to Logistic Regression, shows an improvement in precision (evident in the precision and recall lines plot) as precision increases gradually and at threshold = 0.62, starts to increase more rapidly. Distribution of predictions probabilities shows an improvement at predicting target = 1 as opposed to Logistic Regression. However, majority of probability target = 1 predictions are less than 0.5. This makes sense as majority of the data are target = 0. The AUC is like Logistic Regression but given the distributions of predictions this might be more reliable.

- ***Gradient Boosting Results***

Following Random Forests, a Gradient Boosting model was produced. The main challenge of working with this model is tuning its hyperparameters[1]. Finding optimal combinations of hyperparameters to a particular problem can be difficult and time consuming if the problem has many parameters. In creating this model, a Random Grid Search method, was used to define these parameters. The Random Grid Search method was run on five variables: feature_fraction, bagging_fraction, max_depth, boost (type of boosting), and learning rate. The best parameters found were: 0.1 for feature fraction, 0.3 for bagging fraction, 10 for maximum depth, "gbdt" boosting type, and learning rate of 0.05.

[1] Hyperparameters are values that are set by the practitioner that control the learning process of the model and the parameters that the model ultimately learns from (Nyuytiymbiy, 2021).

Results of the Gradient Boosting model with optimal parameters are represented in Figure 7.

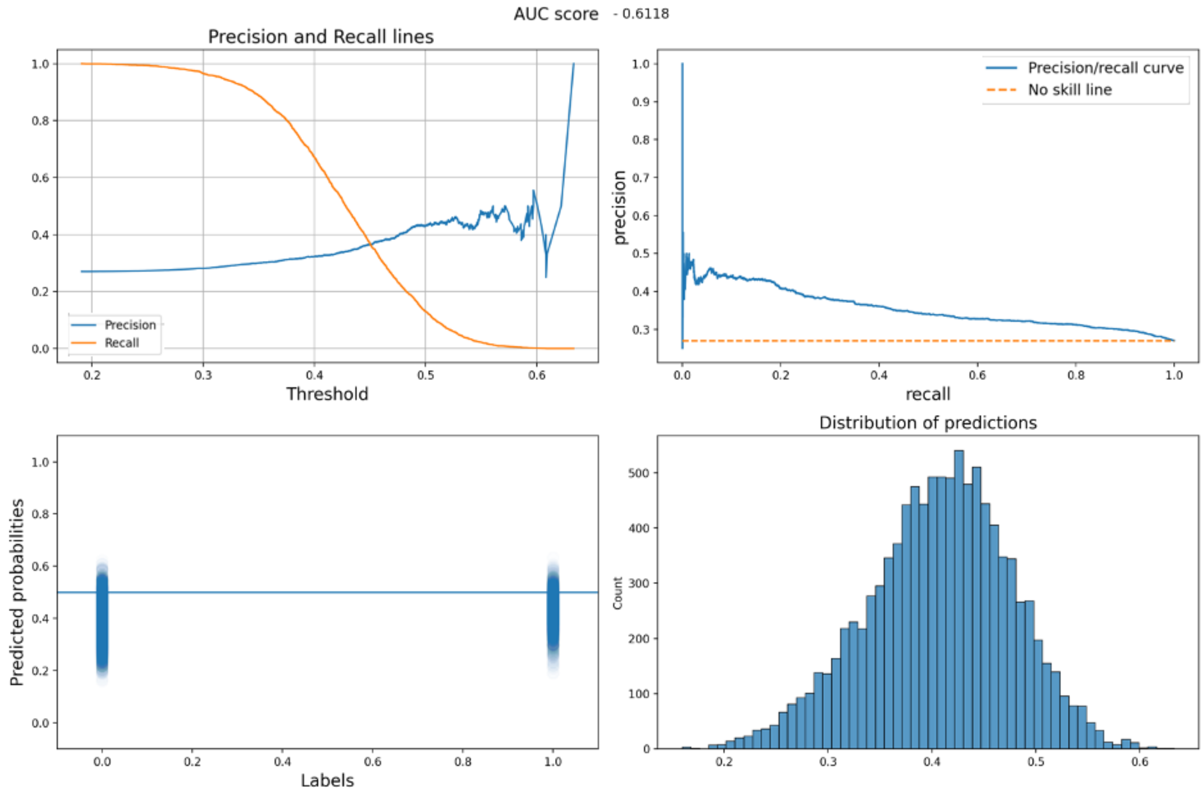

Figure 7: Gradient Boosting Results

The Gradient Boosting model's precision-recall is like Random Forests. The small area under the curve indicates that both precision and recall is relatively low. Like Gradient Boosting, the precision and recall lines indicate that precision increases gradually as threshold increases and then drastically when threshold reaches 0.6. A major disadvantage of this model is that it will fail to make predictions if the threshold exceeds 0.6. The distribution of predictions is very similar to Random Forests. The AUC score is almost equal to Random Forrest and Logistic Regression. This indicates that Gradient Boosting, Random Forests, and Logistic Regression perform equally at distinguishing between the target classes. However, Gradient Boosting and Random Forest's predictions are more reliable.

- ***Principal Component Analysis (PCA)***

A hybridised Logistic Regression with PCA model as well as Random Forest with PCA were also constructed. The variance explained by each principal component is show in Figure 8.

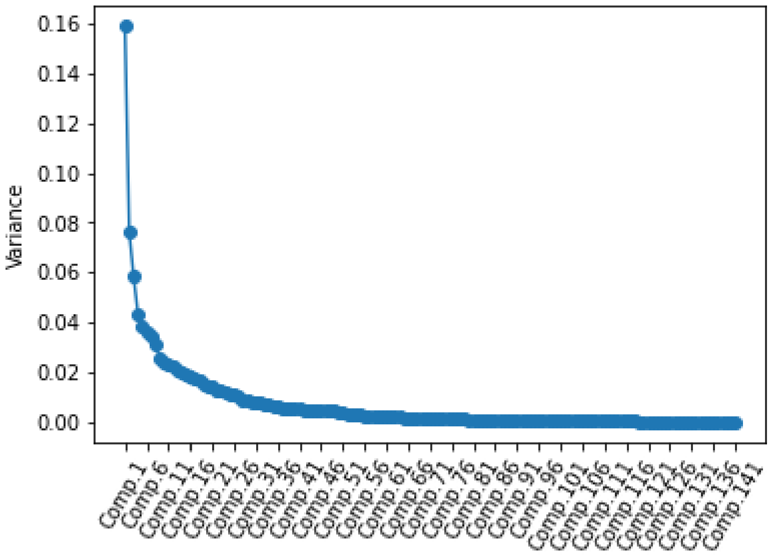

Figure 8: PCA Variance

Although PCA lowers computation time, it failed to significantly increase AUC scores in both cases. Furthermore, predictive power is lost when PCA is applied. Given the aim of this project interpretability was favoured over reducing computation time. As such the results of these models have been excluded from the main report (can be found in Appendix B).

- ***Feature Importance***

After constructing the above models, we now analyse feature importance. Table 1 shows the features that were the most influential across the different models.

Table 1: Feature importance across models

Logistic Regression		**Random Forest**		**Gradient Boosting**	
Feature	**Imp**	**Feature**	**Imp**	**Feature**	**Imp**
payment_day	0.006	age	0.1348	age	0.1174
marital_status_1	0.0003	personal_monthly_income	0.1187	payment_day	0.0658
occupation_type	0.0002	months_in_residence	0.0871	quant_dependants	0.0356
residence_type	0.0002	payment_day	0.0529	personal_monthly_income	0.0348
sex_m	0.0001	quant_dependants	0.0373	sex_f	0.0308

Logistic regression fails to give significant importance to any variable and can therefore be discarded from consideration. Random Forest and Gradient Boosting give quite similar importance to age, personal income, payment day, and number of dependants. Thus, as both of main models agree on selection of most significant variables, one can conclude that these have strong predictive power.

5 Recommendations

For further studies, there are two major elements that could be improved for finer quality of credit scoring models. These include improved data collection and using more advanced predictive algorithms. First and foremost, the structure of data collection should be revised, incorporating data relevance and data description. The viable method of collecting complete data also needs to be redesigned to prevent the generous number of missing values. This advancement will contribute to a robust base that will provide significant improvements in the performance and accuracy of the model. On top of that, the alternative algorithms, such as support vector machine (SVM) and novel credit scoring model (NCSM), could be experimented (Zhang, et al., 2018). They are extensively implemented in credit scoring models, yielded lower overdue rate and higher approval rate in the case of microfinance's credit scoring in emerging markets . Accordingly, the quality of data and variant modelling techniques will help facilitate further effective mechanisms in developing the credit scoring process in the future works.

6 Conclusion

The credit scoring process has proven to be an extremely valuable tool, which has been steadily improving over the years, for organizations such as banks and financial institutions. Hence, from a risk management perspective, credit scoring practices are beneficial at identifying potential risks from an early stage, thereby preventing damages and massive losses. The scope of this paper targeted a dataset that exhibited apparent gaps and imbalances between classes. Nevertheless, some extensive data pre-processing stages were carried out, covering missing values handling to variables standardization, solely for the purpose of enhancing the prediction performance. Data was then carefully divided into training and testing sets. In this study, three main algorithms were used: Logistic Regression, Random Forest, and Gradient Boosting Decision Trees. Additionally, Lasso regularization was used as a variable shrinkage technique under logistic regression to support the feature selection criteria. Results showed that both Random Forest and Gradient Boosting can be powerful algorithms capable of handling both a large number of variables and imperfect data structure. This was in contrast to logistic regression, which was unable to match the importance of features and demonstrated low accuracy. PCA was also incorporated into the logistic regression framework. In our case, however, this could lead to a loss of predictive power and was therefore discarded considering that accuracy and interpretability are of importance. Random forest and gradient boosting provided similar results in terms of AUC scores, thus contributing to significantly higher predictive power. Further future work, however, may be directed to other types of algorithms that have been widely used in credit scoring, such as Support Vector Machines and Neural Networks.

7 References

Abdou, H. A. & Pointon, J., 2011. Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. *Intelligent Systems in Accounting Finance & Management*, 18(2-3), pp. 59-88.

Abellán, J. & Castellano, J. G., 2017. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications,* Volume 73, pp. 1-10.

Amoo, T. & Raghupathi, V., 2015. Using logit model to predict credit score. *Proceedings for the Northeast Region Decision Sciences Institute (NEDSI),* pp. 1-7.

Ampountolas, A., Nde, T. N., Date, P. & Constantinescu, C., 2021. A Machine Learning Approach for Micro-Credit Scoring. *Risks,* 9(50), pp. 1-20.

Brown, I., Mues, C. & Thomas, L., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications,* 39(3), pp. 3446-3453.

Chen, M.-C. & Huang, S.-H., 2003. redit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, 24(4), pp. 433-441.

Chopra, A. & Bhilare, P., 2018. Application of ensemble models in credit scoring models. *Business Perspectives and Research*, 6(4), pp. 129-141.

Dinha, T. H. T. & Kleimeier, S., 2007. A credit scoring model for Vietnam's retail banking market. *International Review of Financial Analysis,* 16(5), pp. 471-495.

Louzada, F., & Fernandes, G. B., 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2), pp. 117-134.

Muhammed, M. M. & Adinoyi, I. A., 2019. Credit Scoring Prediction with Logistic Classifier using Latent Components from Principal Components and Factor Analyses. *Professional Statisticians Society of Nigeria,* Volume 3, pp. 546-553.

Nalić, J. & Švraka, A., 2018. *Importance of data pre-processing in credit scoring models based on data mining approaches.* Opatija, Croatia, IEEE; 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).

Rahmad, F., Suryanto, Y. & Ramli, K., 2020. Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification. *Materials Science and Engineering,* 879(1), pp. 1-11.

Ruiz, S., Gomes, P., Rodrigues, L. & Gama, J., 2019. Credit scoring for microfinance using behavioral data in emerging markets. *Intelligent Data Analysis,* Volume 23, p. 1355–1378.

Sadatrasoul, S. M., Gholamian, M. R., Siami, M. & Hajimohammadi, Z., 2013. Credit scoring in banks and financial institutions via data mining techniques: A literature review. *Journal of Artificial Intelligence and Data Mining,* 1(2), pp. 119-129.

Sayjadah, Y., Hashem, I. A. T., Alotaibi, F. & Kasmiran, K. A., 2018. Credit Card Default Prediction using Machine Learning Techniques. *Conference: 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA),* pp. 1-5.

Scikit-learn.org, 2018. *sklearn.model_selection.train_test_split — scikit-learn 0.20.3 documentation.* [Online]
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
[Accessed 4 December 2022].

Singh, P., 2017. *Comparative study of individual and ensemble methods of classification for credit scoring.* Coimbatore, India, IEEE, pp. 968-972.

Walusala, S., Rimiru, R. & Otieno, C., 2017. A Hybrid Machine Learning Approach for Credit Scoring Using PCA and Logistic Regression. *International Journal of Computer (IJC),* 27(1), pp. 84-102.

Wang, H., Xu, Q. & Zhou, L., 2015. Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. *PloS one,* 10(2), pp. 1-20.

Wu, T.-C. & Hsu, M.-F., 2012. Credit risk assessment and decision making by a fusion approach. *Knowledge-Based Systems,* Volume 35, pp. 102-110.

Xia, Y., Liu, C., Li, Y. & Liu, N., 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications,* Volume 78, p. 225–241.

Zeng, G., 2020. On the Confusion Matrix in Credit Scoring and Its Analytical properties. *Communications in Statistics - Theory and Methods,* Volume 49, pp. 2080-2093.

Zhang, D., Huang, H., Chen, Q. & Jiang, Y., 2007. A comparison study of credit scoring models. *Third International Conference on Natural Computation (ICNC 2007),* Volume 1, pp. 15-18.

Zhang, X., Yang, Y. & Zhou, Z., 2018. A novel credit scoring model based on optimized random forest. *In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC),* p. 60–65.

8 Appendix A (Count Plots)

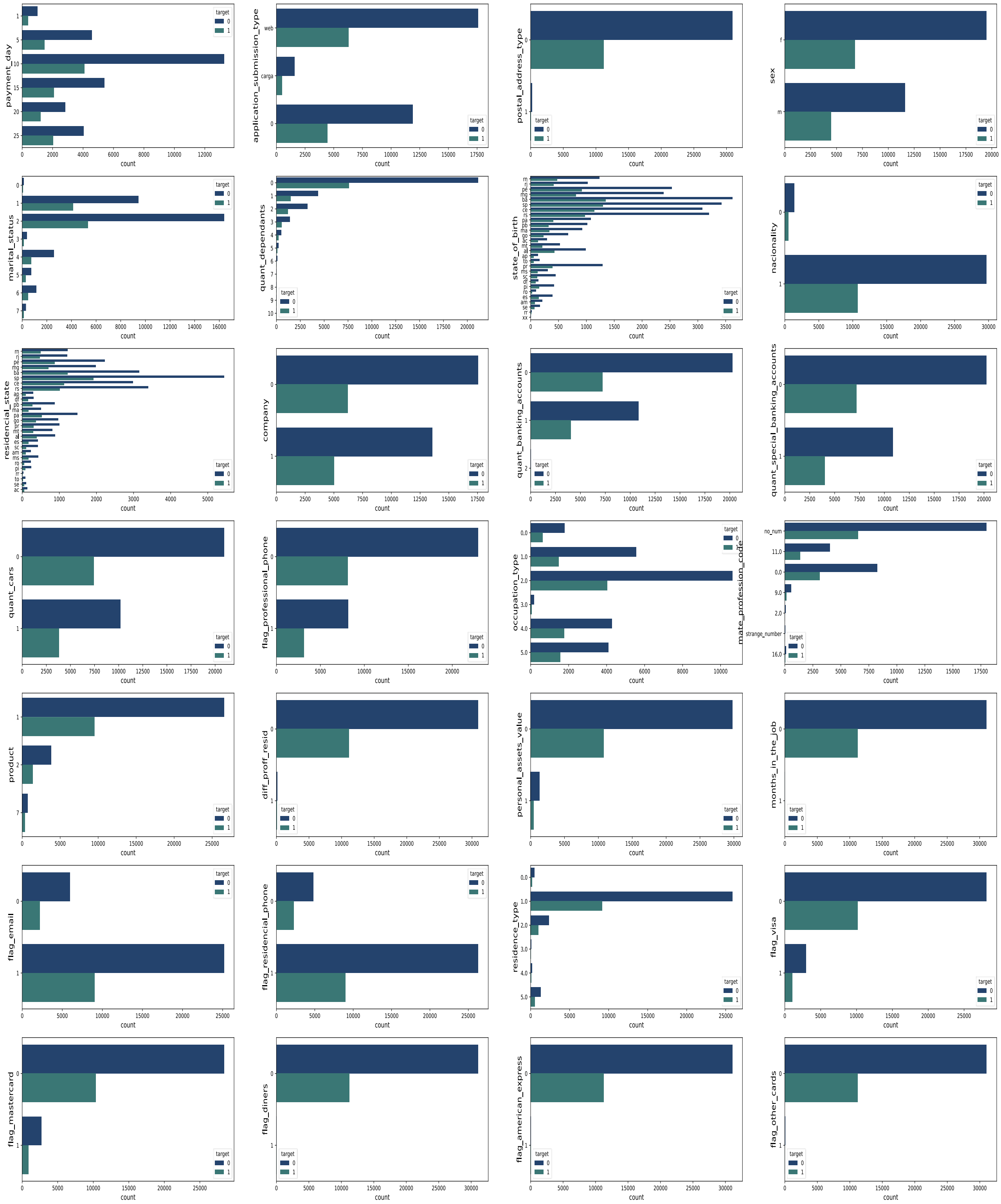

Figure 9: Count plots for all categorical and binary variables

9 Appendix B (Logistic Regression and Random Forest with PCA Results)

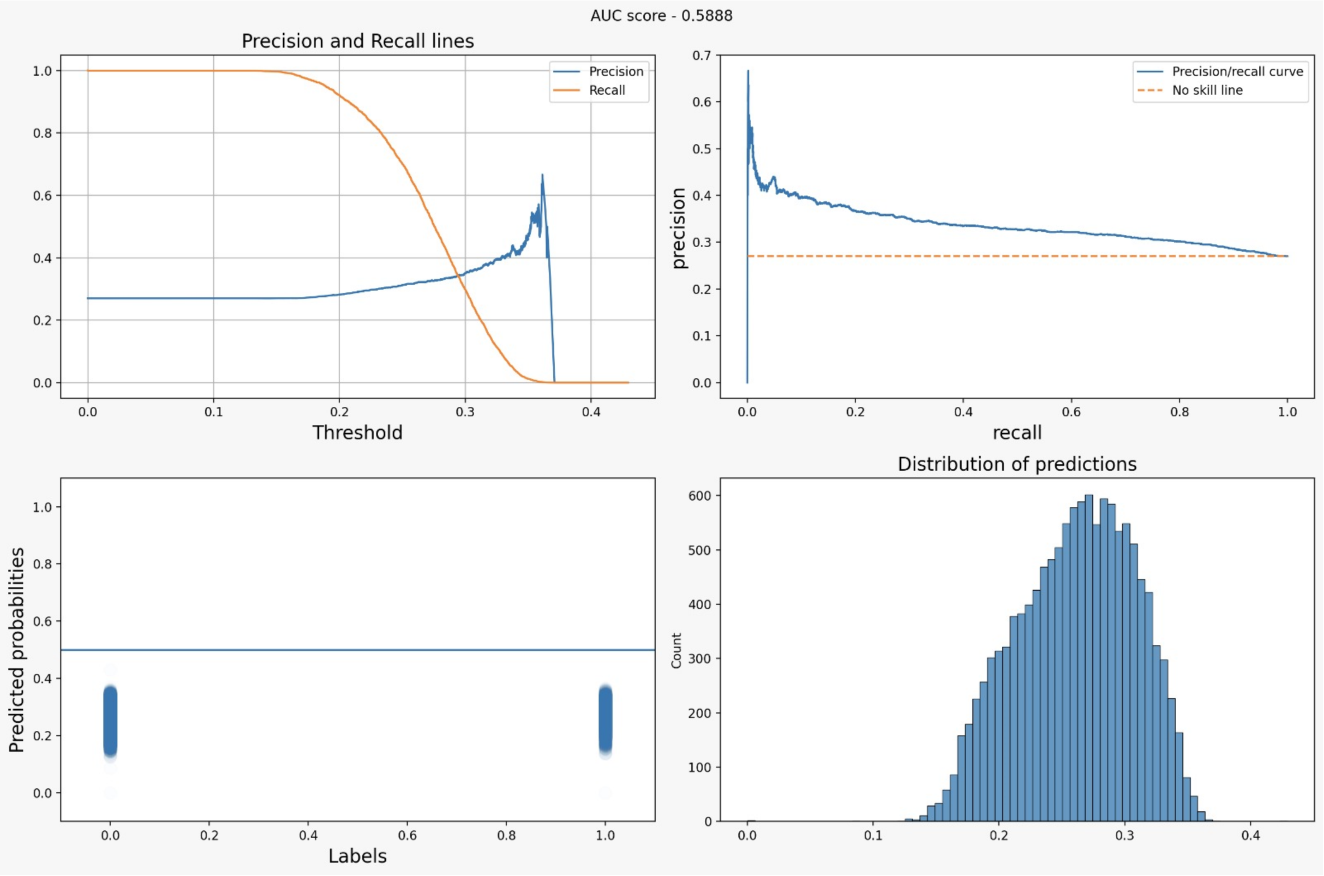

Figure 10: Logistic Regression with PCA Results

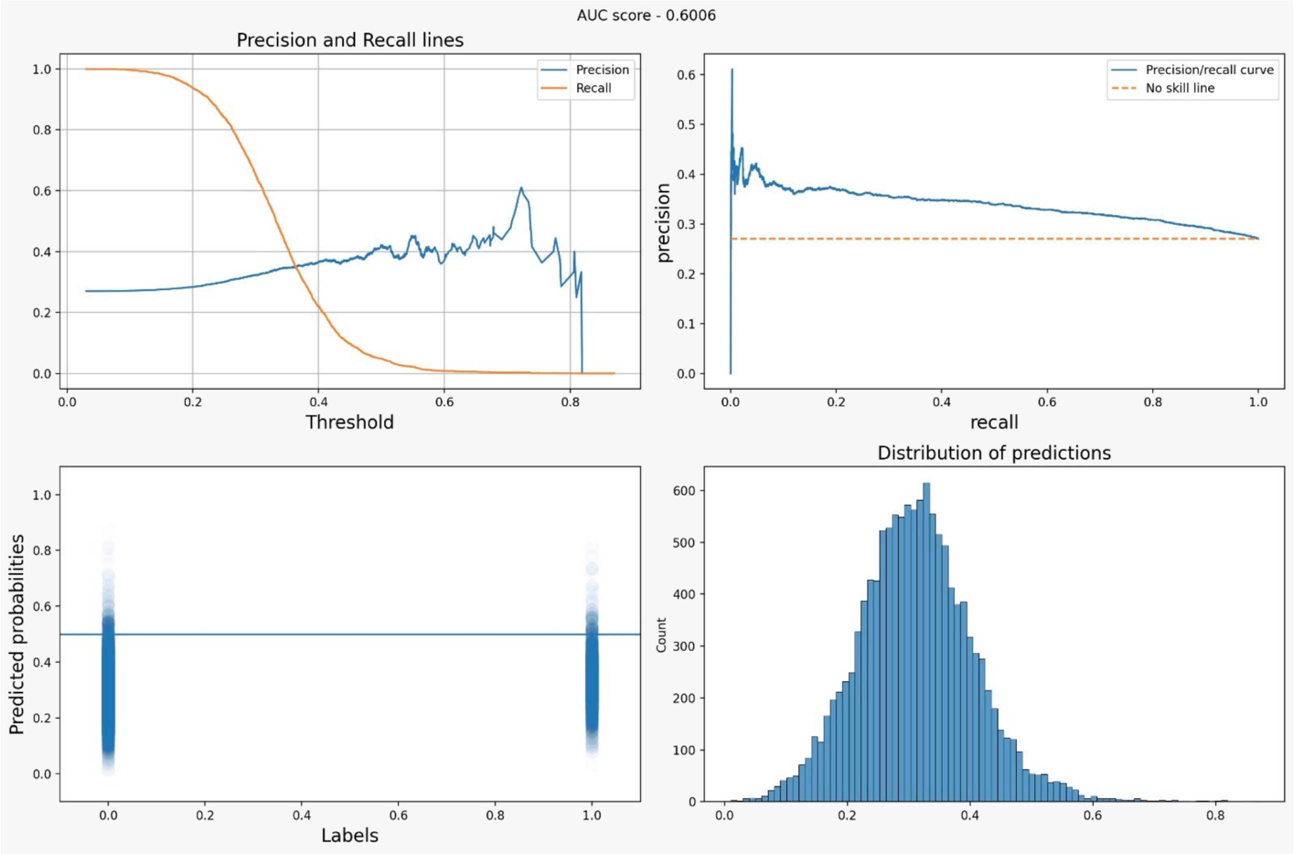

Figure 11 - Random Forrest with PCA Results