A COMPARATIVE STUDY FOR CREDIT CARD APPROVAL PREDICTION

Anh Trinh Phuong

Abstract

Credit risk management is crucial for financial institutions. In the era of big data, machine learning has demonstrated significant improvements in the accuracy of credit risk modeling. The objective of this study is to predict default probability and classify credit applicants to aid banks in decision-making while mitigating bad debts. Using a bank's credit card approval dataset, the study developed five models including logistic regression, decision tree, random forest, lightGBM, XGBoost to find the most accurate model. The result records that XGBoost, with an accuracy of 96.2% and ROC AUC score of 0.96, gave the best predictive performance. It also implied that machine-learning-based models outperformed the traditional statistical model, logistic regression.

Keywords: Credit risk analysis, machine learning, logistic regression, decision tree, random forest, lightGBM, XGBoost

1. Introduction

Having existed for a long time ago, it was not until the 2007-2009 financial crisis that financial institutions, especially banks, became more aware of the importance of risk management. The cause of the risk is related to the high-risk mortgage loans that US financial institutions offer for low income real estate buyers without regard to the pay back ability, and the collapse of the housing market bubble. Since then, the development of credit risk management has become the priority of financial institutions in making credit decisions. In which, credit scorecards are a widely used method. Based on personal information and data provided by the credit applicants, the bank will predict the probability of default, and decide whether to issue a credit card or not. Today, in the Digital Age, credit risk is not executed only by statistical methods such as linear regression, logistic regression. With the help of big data and machine learning algorithms, credit risk tasks are performed more and more accurately, and faster (Lessmann et al., 2015; Doko et al., 2021). Also, due to the simplicity and convenience of use, machine-learning-based models have been extensively adopted in the financial industry.

The main purpose of this study is to give predictions of classifying credit applicants by default probability, helping banks in making credit decisions, reducing and avoiding bad debts. This work evaluates different models within statistics-based and machine-learning-based to find out the most accurate prediction model, given the credit card dataset used.

2. Theoretical background

2.1. Credit Risk and Credit Risk analysis

According to Brown & Moles (2014), credit risk is defined as “the potential that a contractual party will fail to meet its obligations in accordance with the agreed terms”.

One of its concerns is related to the risk of issuing credit cards. Nowadays, credit cards are a very popular payment tool. The card is issued by financial institutions (usually commercial banks), allowing users to make a payment first without having money, then pay back to issuers later. Due to the base on the credibility of users, credit risk management must be strictly implemented right from the card issuance stage. The credit scorecard assessment will help banks measure the applicants’ likelihood of default, thereby deciding whether to issue a card or not. According to Basel II, a default occurs when: (i) an obligor is unlikely to repay in full its credit obligations to the banking group, without recourse by the bank to actions such as realising security, (ii) the obligor is past due for more than 90 days on any material credit. The credit applicants will be predicted as 1 if defaulted, and 0 otherwise.

Today, banks are increasingly pushing for digital transformation, leveraging the contribution of data and its predicted power. To perform credit risk analysis, researchers need data about demographics, internal-financial data, collateral data, bureau data, macroeconomic variables and credit agreement. In analysis process, banks applied predictive analytics, which is the technology set that combines data science, machine learning and predictive and statistical modeling to generate predictions for different expert systems, such as predicting risk, liquidity, customer churn, fraud detection and revenue and for making informed decisions (Lackovic et al. 2016; Provost and Fawcett, 2013). Study of credit risk prediction using Naive Bayes, neural network and decision trees, indicated that decision trees gave the highest accuracy (Hamid & Ahmed, 2016). Another paper (Kovvuri & Cheripelli, 2019) stated that logistic regression and random forests equal in accuracy values, after using logistic regression, decision trees and random forests, Meanwhile, Gahlaut et al. (2017) showed that random forest is the best classified algorithm.

2.2. Methodology

Logistic Regression

Binary logistic regression, without any assumptions of normal distribution, is a popular statistical method based on the maximum likelihood to estimate the happening probability of an event. This method is widely applied in a variety of fields including biomedical, social science, marketing, and in terms of financial, logistic regression often adopted for credit risk assessment.

In 1995, Henly used a logistic model for credit scoring applications. Binary logistic regression also applied to predict the probability of default of SMEs (Behr et

al., 2004; Altman & Sabato, 2007; Yazdanfar & Nilsson, 2008). Although there is evidence supporting that artificial intelligence techniques have enhanced the outputs of credit risk analysis compared with classical statistics one (Abellan & Castellano, 2017), logistic regression with easy implementation and balanced error distribution remain favored.

Decision Trees

Decision Trees are one of the commonly used techniques for credit risk analysis. They are a type of supervised learning algorithm that is mostly used for classification and regression analysis. The aim of using a Decision Tree is to build a training model that can forecast the class or value of a target variable by learning basic decision rules from training data. In Decision Trees, we begin at the base of the tree to forecast a class label for a record. We compare the numbers of the root property and the attribute of the record. Based on the comparison, we follow the branch matching to that number and proceed to the next node (Chauhan, n.d.)

Decision Trees have gained popularity in the credit risk analysis field due to their ability to handle both categorical and numerical data, as well as their interpretability and ability to manipulate nonlinear relationships between variables. Additionally, decision trees can handle missing data and outliers, making them a robust tool for credit risk analysis.

Simha and Satchidananda (2006) used farm loan data from two Indian banks to forecast failure risk using both a decision tree learning strategy and the logistic regression technique. They discovered that decision trees produce excellent results (the overall default precision using decision trees was 90%, while logistic regression produced 83%), and that they are simple to comprehend and apply. Another credit scoring models research from Tap & Ong (2011) using the logistic regression model, decision tree, no model outperforms the other.

Random Forests

Random forests is a well-known ensemble learning method that combines various decision trees to improve model accuracy and stability. Individual trees in the forest have been trained on random subsets of data and features, and each tree has an equal share in the ultimate forecast determined from majority voting. Random forests have been widely used in a variety of areas, including finance, healthcare, and marketing, due to their ability to deal with high-dimensional data with complex connections between characteristics.

Kruppa et al. (2013) suggested that random forests models outperformed traditional approaches like logistics, or even when compared with closest k neighbors. Similar conclusion is stated by Gahlaut & Singh (2017) that after comparing algorithms such as decision tree, support vector machine, adaptive boosting model,

linear regression, random forest, and neural network for building predictive models, the best algorithm for risky credit classification is a random forests algorithm.

LightGBM

LightGBM is a gradient boosting framework that uses tree-based learning algorithms and is designed to handle large-scale datasets with high-dimensional features. It achieves this by using a novel technique called Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which help to reduce the training time and improve the accuracy of the model.

GOSS excludes a large proportion of data instances with small gradients and uses the remainder to estimate the information gain. It is shown that, because data instances with larger gradients are more significant in the computation of information gain, GOSS can estimate information gain with a much smaller data set. To reduce the amount of features, algorithms bundle mutually exclusive features with EFB. We show that finding the optimal bundling of exclusive features is NP-hard, but a greedy algorithm can obtain a reasonable approximation ratio and thus effectively reduce the number of features without affecting split point determination accuracy significantly (Ke et al., 2017).

The LightGBM algorithm is infrequently used in the financial industry. Since its introduction in 2017, LightGBM has attracted wide attention from the field. In most situations, its prediction accuracy outperforms that of other machine learning algorithms. The LightGBM algorithm outperforms the other three algorithms (KNN, Decision Trees, Random Forests) in the prediction results for business financing risk (Wang & Zhao, 2022). Gao & Balyan (2022) also showed that LightGBM has obtained the best outcome in evaluating the default risk of users on P2P platforms.

XGBoost

XGBoost, stands for Extreme Gradient Boosting, is another popular gradient boosting framework that has gained popularity due to its scalability and performance on large datasets with complex and heterogeneous features. It is the top machine learning library for regression, classification, and ranking problems, and it supports parallel tree boosting. XGBoost training, for the production of each tree method, through calculating node split and classification first whether to produce “gain” to determine whether the node is divided, and through parameter controls the depth of the tree, when a tree is generated after pruning were needed to prevent a fitting, the first m round of the generated tree will learn the real value and m-1 wheel model forecasts “residual”, so that the model prediction results gradually close to the real value. Moreover, individual trees are created using numerous cores in XGBoost, and data is organized to minimize lookup times. This reduced model training time, which improved model performance.

The personal credit risk assessment model based on XGBoost exceeds the other two categorization algorithms, decision tree and K-nearest neighbor, in terms of AUC and accuracy (Wang et al., 2022). In another research, Huang and Yen (2019) chose 16 financial variables from Taiwanese listed firms' financial records as input for six models, including the XGBoost model. According to their empirical findings, XGBoost offers the most accurate prediction of financial hardship.

Information Value and Weight of Evidence

The weight of evidence (WOE) and information value (IV) provide a great framework for exploratory analysis and variable screening for binary classifiers. WOE and IV have been used extensively in the credit risk world for several decades, and the underlying theory dates back to the 1950s.

The weight of evidence (WOE) indicates an independent variable's predictive ability in connection to the dependent variable. It is usually characterized as a measure of the separation of good and bad customers because it developed from the credit rating world. "Bad Customers" are consumers who have failed on a debt. and "Good Customers" refers to consumers who returned the debt.

$$\text{WOE} = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right)$$

Information value (IV) is a helpful method for selecting significant variables in a predictive model. It aids in ranking factors according to their significance. The IV is determined using the following formula:

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

Table 1 provides conventional values of IV statistics (Siddiqi, 2006)

Table 1. Conventional Interpretation of IV

Information Value (IV)	Predictive Power
< 0.02	Useless
0.02 - 0.1	Weak predictors
0.1 - 0.3	Medium predictors
0.3 - 0.5	Strong predictors
> 0.5	Suspicious

Model Evaluation Metrics

A model's performance can be evaluated using a variety of metrics, including the F1 value, recall rate, and precision rate. As assessment metrics, we primarily use the AUC value, F1 value, precision, accuracy, and recall rate.

Firstly, we define: (i) TP as the number of positive samples correctly predicted by the classification model, (ii) FN as the number of positive samples incorrectly predicted as a negative class by the classification model (Type II error), (iii) FP as the number of negative samples incorrectly predicted as a positive class by the classification model (Type I error), and (iv) TN as the number of negative samples correctly predicted by the model. These values are shown in the Confusion Matrix:

Table 2. Confusion Matrix

		Predict	
		0	1
True	0	TN	FP
	1	FN	TP

Based on the Confusion Matrix, many indicators can be calculated to evaluate the predictive ability of the classifier. Among these, accuracy is a more common assessment measure, representing the proportion of correctly classified samples to the total number of samples. Accuracy is computed by the following formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

For the default risk analysis, banks place a greater emphasis on identifying default users. At this point, the two evaluation indicators of recall and precision play a critical role. Precision is referred to as the ratio of true positive samples predicted by the classifier. And recall is defined as the ratio of positive samples correctly predicted by the classifier:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Accuracy and recall are indexes used to judge classifier predictive power on good data. Nevertheless, there are conflicts between the two, making it challenging to enhance both at the same time. Therefore, we use recall and precision to compute the F1-score, which takes both the recall and accuracy of the classifier forecast into consideration. The F1-score computation method is shown in the following equation:

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Another important index to describe the overall performance of a model is the AUC value, which stands for "area under the ROC curve". A receiver operating characteristic curve (ROC curve) is a graph that depicts the performance of a classification algorithm across all classification levels.

3. Data

3.1. Data Description

The dataset sourced from a competition named "Credit Card Approval Prediction" on Kaggle, an online platform for the Machine Learning and Data Science community. (Credit Card Approval Prediction Dataset)

Credit card dataset includes two tables: application_record and credit_record. Application_record data has 18 columns with 438 557 rows. And credit_record contains 3 columns with 1 048 574 rows. These two tables will be merged by "ID" column.

Table 3. Variables Description

Variable	Meaning
ID	Client ID
CODE_GENDER	Gender of the applicant (M for male, F for female)
NAME_EDUCATION_TYPE	Education level
NAME_FAMILY_STATUS	Marital status
NAME_HOUSING_TYPE	Housing status
NAME_INCOME_TYPE	Source of income
OCCUPATION_TYPE	Client job
AMT_INCOME_TOTAL	Total amount of income
CNT_FAM_MEMBERS	Number of family members
CNT_CHILDREN	Number of children

DAYS_BIRTH	Count backwardly number of days from current day to birth (negative values)
DAYS_EMPLOYED	Count backwardly number of days from current day to the first day employed (negative values; if positive, currently unemployed)
FLAG_OWN_CAR	Is there a car (Y for yes, N for no)
FLAG_OWN_REALTY	Is there a realty (Y for yes, N for no)
FLAG_MOBIL	Is there a mobil (Y for yes, N for no)
FLAG_EMAIL	Is there an email (Y for yes, N for no)
FLAG_PHONE	Is there a phone (Y for yes, N for no)
FLAG_WORK_PHONE	Is there a work phone (Y for yes, N for no)
MONTHS_BALANCED	Record month. The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on
STATUS	Historical loan status (0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month)

The following is some quantity statistics chart generated by Tableau, which gave better understand of the distribution of different subjects participating in the dataset.

Figure 1. Proportion of Gender

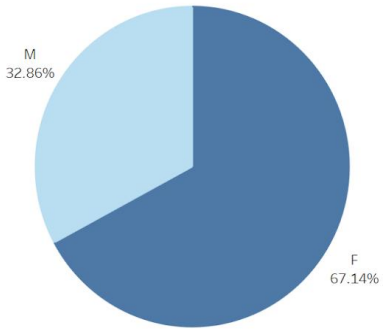
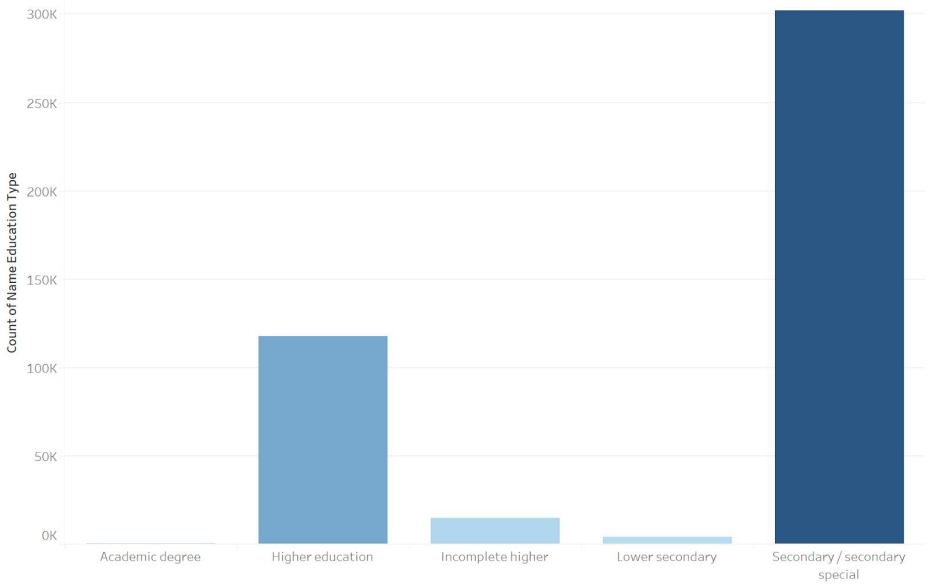


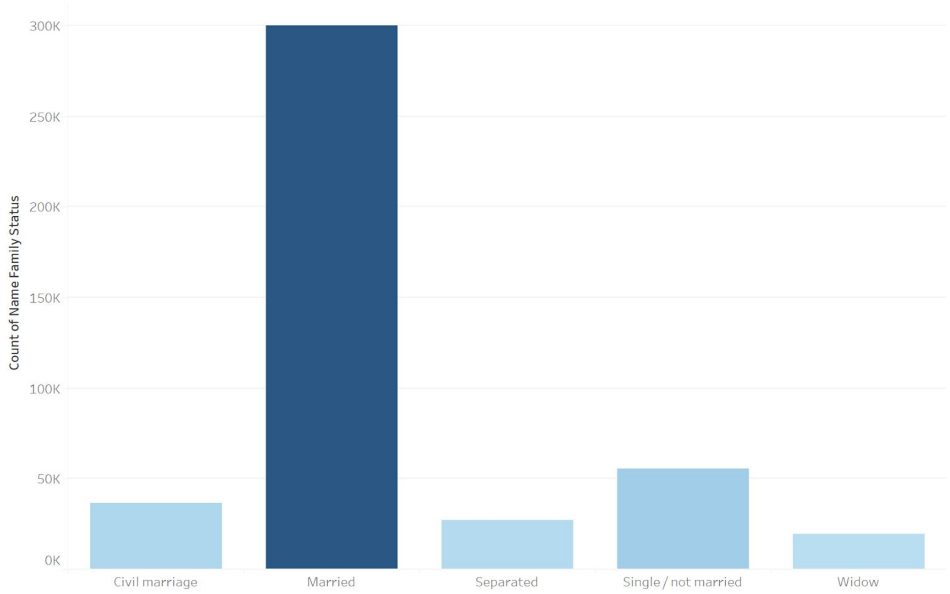
Figure 1 shows that the number of Female applicants (67.14%) exceeds the number of Male applicants (32.86%). This result implies that women are considered to have a higher demand for credit card services than men do.

Figure 2. Distribution of Education Types



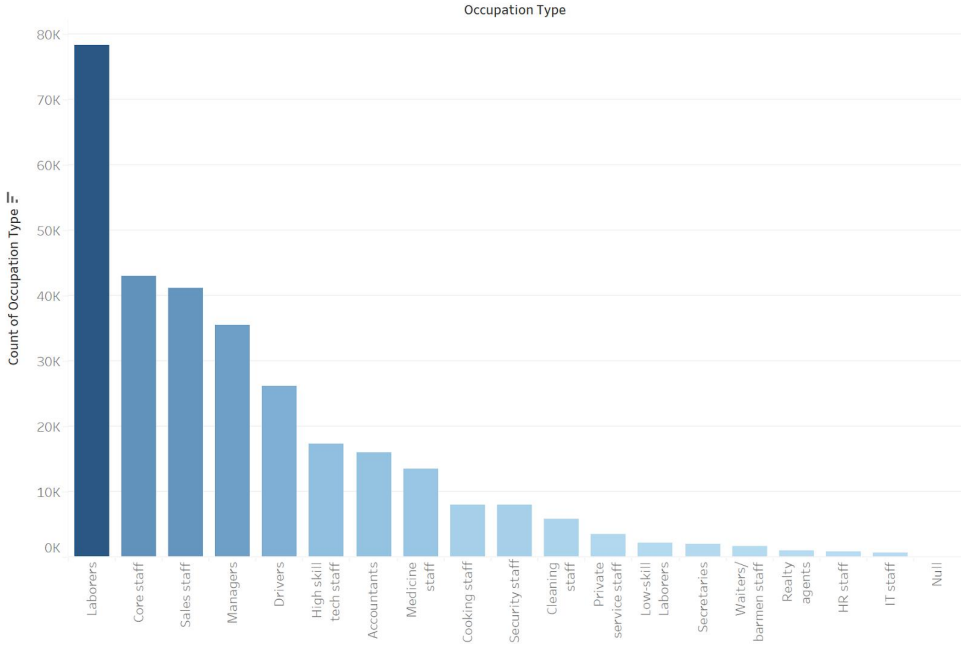
Of the five education levels, the majority of applicants are at the secondary or lower secondary education level, followed by the higher education, incomplete higher, lower secondary and finally the academic degrees. Thus, the set of customers in this dataset has a quite low educational level.

Figure 3. Distribution of Family Status



Most of the clients are married. A smaller portion is still single/ not married, civil marriage, seperated, or widow. Because "married" status accounts for the majority, family status information, including the number of family members, will be considered more carefully during the examination.

Figure 4. Distribution of Occupation types



There are 18 occupations recorded in the dataset. In which, there are nearly 80k applicants who are laborers, accounting for the largest proportion. Next are core staff, sales staff, manager, drivers, ranging from 27k to 43k. The remaining occupations account for a smaller amount. However, occupation information records many missing values. This problem will be solved later in the preprocessing step.

3.2. Data Preprocessing

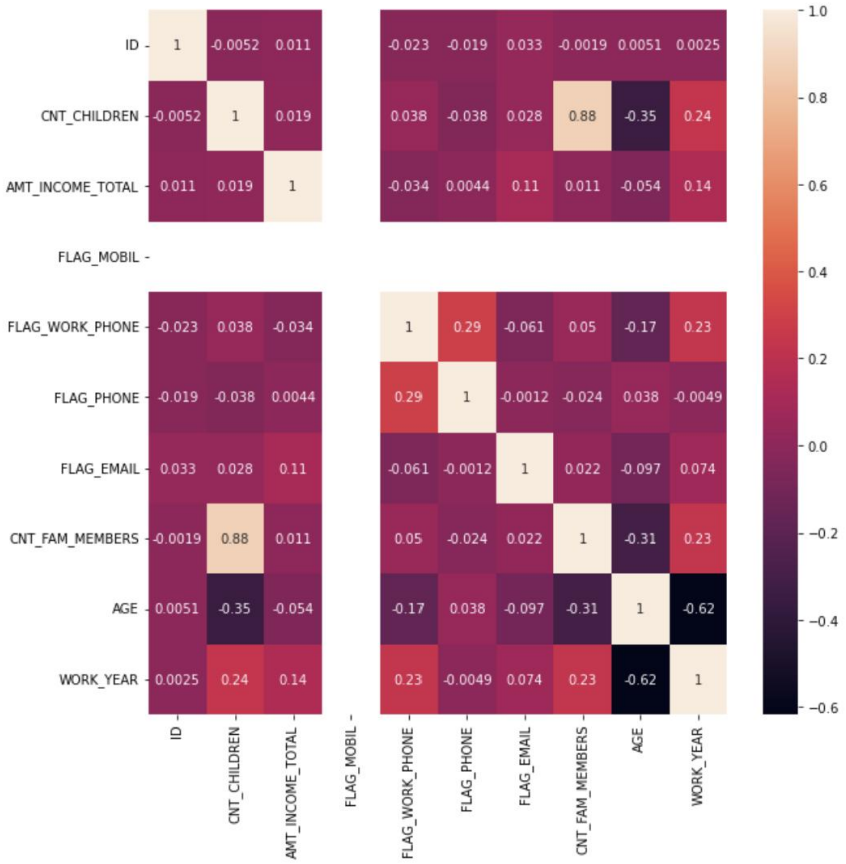
Missing Values

Missing values only exist in application_record dataset. OCCUPATION_TYPE contains 134 203 values "NaN", accounting for 30.6%. These

missing values would be replaced with appropriate values followed by a customized method.

Correlation Matrix

Figure 5. Correlation Matrix of application_record dataset



By calculating correlation matrix, the author decided to drop CNT_CHILDREN columns due to high correlation between this column versus CNT_FAM_MEMBERS (0.88) (Cohen, 1988).

Outliers Detecting

This study combined the observation of boxplot graphs and the Interquartile Rule with numeric variables. In terms of categorical variables, outliers are found by distribution graph.

Tukey's (1977) boxplot is a useful technique because it makes no distributional assumptions and does not rely on a mean or standard deviation. The lower quartile (Q1) is the 25th percentile of the statistics, and the higher quartile (Q3) is the 75th percentile. The interquartile range (IQR) is the spread between Q1 and Q3, which means IQR = Q3 - Q1. He also defined Q1 - (1.5 * IQR) and Q3 + (1.5 * IQR) as "inner fences", and observations between inner fences as "outside". Therefore, remaining data points are called "outliers".

Data Merging

Having been processed, two tables would be merged together by inner method on the "ID" column. The completed data includes 656 424 duplicated rows that need to be dropped, and no missing data. The final dataset has 21 columns with 32 834 rows.

Features Selection

The study utilized Information Value estimation, a commonly employed method in credit scoring issues, to determine the correlation between each feature and the dependent variable. (Zdravevski et al., 2014). After binning for continuous variables, IV calculations are shown in the below table:

Table 4. Calculation of IV values

Variable	IV
NAME_FAMILY_STATUS	0.102943
NAME_HOUSING_TYPE	0.060346
NAME_EDUCATION_TYPE	0.047155
NAME_INCOME_TYPE	0.042232
CNT_FAM_MEMBERS	0.038941
AGE	0.035022
WORK_YEAR	0.019894
FLAG_OWN_REALTY	0.01797
OCCUPATION_TYPE	0.00629
FLAG_OWN_CAR	0.00581
FLAG_PHONE	0.004742
CODE_GENDER	0.004263
AMT_INCOME_TOTAL	0.000723
FLAG_WORK_PHONE	0.000254
FLAG_EMAIL	0.00001

With IV threshold equals 0.02, 9 variables would be dropped from the dataset. However, the author finds that AMT_INCOME_TOTAL has significant financial meaning, this income variable will continue to be included in the model. Thus, there are 8 variables that will be removed after feature selection stage.

Variables encoding and transformation

Encoding variables ensures that all variables are properly represented in the model. One common encoding technique is to use one-hot encoding, where each category is represented as a binary variable. This technique involves creating a new binary variable for each category in the original categorical variable. Two widely used methods are Scikit-learn's OneHotEncoder and Pandas' get_dummies method.

Moreover, label encoding is a technique used to convert categorical variables into numerical ones by assigning each unique category a different integer value. However, it is important to note that label encoding may not be suitable for all types of categorical variables and may result in misleading interpretations.

For transformation, continuous variable would be transformed by the MinMaxScaler method. MinMaxScaler is a normalization technique that scales numerical data to a specific range, typically between 0 and 1. This technique is commonly used in credit risk analysis to ensure that numerical variables are on the same scale and have equal importance in the machine learning model (Maruma et al., 2022).

Data Balancing

Checking for distribution of “1” and “0” target values give a severe imbalanced result in Table 5.

Table 5. Distribution of TARGET variable

	Frequency	Percentage
1	32 571	99.2%
0	263	0.8%

Data with severe class imbalance will lead to imbalance classification, therefore, give a poorly performed predictive model. Oversampling the minority class is one method for dealing with unbalanced information. The most basic method includes duplicating cases from the minority class, even though these examples contribute no new information to the model. Instead, new instances can be created by combining old ones. The Synthetic Minority Oversampling Technique, or SMOTE for short, is a form of data enhancement for the minority class.

4. Findings and Discussion

In the results part, accuracy and ROC are the main measurements for classification performance. Furthermore, for the fact that it's impossible to have both high precision and recall, F1-score, as weighted average of precision and recall, is also considered in judging model performance.

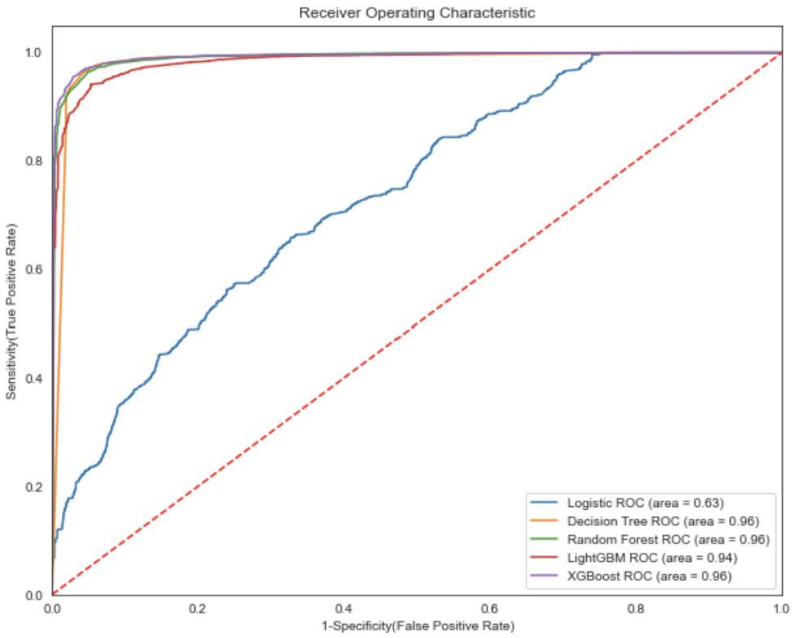
Table 6. Performance evaluation metrics of models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.6345	0.64	0.63	0.63
Decision Trees	0.95999	0.96	0.96	0.96
Random Forests	0.95753	0.96	0.96	0.96
LightGBM	0.94054	0.94	0.94	0.94
XGBoost	0.96213	0.96	0.96	0.96

Table 6 shows the performance evaluation indicators of five models: Logistic Regression, Decision Trees, Random Forests, LightGBM, and XGBoost. The worst result came from classical statistical model, linear regression, with accuracy of only 63.45%. The remaining models, which based on machine learning algorithms, gave excellent accuracy, about 94% - 96% (Allwright, 2022). Other metrics like precision, recall and F1-score are also in this range. Among these, XGBoost algorithm has the best performance, with accuracy of 96.213%, precision, recall, and F1-score of 0.96. Moreover, despite having high accuracy (94%), LightGBM performed weakest in terms of machine-learning-based models.

Above results are further supported by ROC curve and AUC values in Figure 5. AUC ranges from 0 to 1, and AUC closer to 1 means that its predictions are closer to 100% correct.

Figure 6. ROC curve for five models



5. Conclusion

In conclusion, this study employ five models, both statistics and machine learning, for Credit Card Approval dataset to classify applicants by default probability. The result indicates that machine-learning-based models (Decision Tree, Random Forest, LightGBM, XGBoost) outperformed classical statistical one (Logistic Regression). And the best performed model is XGBoost, which gave highest evaluation metrics.

Regarding limitations, the dataset used in this study has a mysterious background. Therefore, the author has not been able to clarify the research context, an important factor in financial analysis. In addition, because the accuracy achieved is quite high, the study has not optimized the parameters. Future studies can rely on these limitations and improve by using different datasets to make research results more diverse and inclusive, as well as perform hyperparameters tuning to boost the model predictive results.

Appendix

Viewers can visit my code here: Credit Card Approval Prediction Python Code

References

1. Abellán, J., & Castellano, F.J. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Syst. Appl., 73, 1-10.
2. Allwright, S. (2022, August 10). What is a good accuracy score? Simply explained. Stephen Allwright. https://stephenallwright.com/good-accuracy-score/
3. Altman, E.I., and Sabato, G. (2007). Modelling Credit Risk for SMEs: Evidence from The US Market. ABACUS, 43 (3), 332-357.

4. Behr, P., Güttler, A., and Plattner, D. (2004). Credit Scoring and Relationship Lending: The Case of German SME. Working Paper, University of Frankfurt, Version 15.

5. Brown, K., & Moles, P. (2014). Credit risk management. K. Brown & P. Moles, Credit Risk Management, 16.

6. Chauhan, N. S. (n.d.). Decision Tree Algorithm, Explained - KDnuggets. KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

7. Cohen, J. (1988). Statistical power analysis (2nd ed.). Hillsdale NJ: Erlbaum.

8. Doko, F., Kalajdziski, S., & Mishkovski, I. (2021). Credit Risk Model Based on Central Bank Credit Registry Data. Journal of Risk and Financial Management, 14(3), 138. https://doi.org/10.3390/jrfm14030138

9. Gahlaut, A., Tushar, & Singh, P.K. (2017). Prediction analysis of risky credit using Data mining classification models. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-7.

10. Gao, B., & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. Journal of Intelligent Systems, 31(1), 767–779. https://doi.org/10.1515/jisys-2022-0036

11. Hamid, A.J., & Ahmed, T.M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining.

12. Huang, Y. P., & Yen, M. F. (2019). A new perspective of performance comparison among machine learning algorithms for financial distress prediction. Applied Soft Computing, 83, 105663.

13. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS.

14. Kovvuri, R.S., & Cheripelli, R. (2019). Credit Risk Valuation Using an Efficient Machine Learning Algorithm. Learning and Analytics in Intelligent Systems.

15. Lacković, I.D., Kovšca, V., & Vincek, Z.L. (2016). Framework for big data usage in risk management process in banking institutions.

16. Lessmann, Stefan & Baesens, Bart & Seow, Hsin-Vonn & Thomas, Lyn. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research. (doi:10.1016/j.ejor.2015.05.030). 10.1016/j.ejor.2015.05.030.

17. Maruma, C., Tu, C., Nawej, C. (2022). Banking Credit Risk Analysis using Artificial Neural Network. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Seventh International Congress on Information and

Communication Technology. Lecture Notes in Networks and Systems, vol 447. Springer, Singapore. https://doi.org/10.1007/978-981-19-1607-6_76

18. Provost, Foster & Fawcett, Tom. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making. Big Data. 1. 10.1089/big.2013.1508.
19. Satchidananda, S.S., & Simha, J.B. (2006). Comparing decision trees with logistic regression for credit risk analysis.
20. Siddiqi, N. (2006). Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ: John Wiley & Sons, Inc.
21. Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160).
22. Wang, D., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. Information Sciences, 602, 259–268. https://doi.org/10.1016/j.ins.2022.04.058
23. Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. Procedia Computer Science, 199, 1128–1135. https://doi.org/10.1016/j.procs.2022.01.143
24. Yap, B.W., Ong, S.H., & Husain, N.H. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst. Appl., 38, 13274-13283.
25. Yazdanfar, D., and Nilsson, M. (2008). The Bankruptcy Determinants of Swedish SMEs. Institute for Small Business and Entrepreneurship, Belfast, Ireland.
26. Zdravevski, E., Lameski, P., Kulakov, A., & Gjorgjevikj, D. (2014). Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets. 2014 Federated Conference on Computer Science and Information Systems, 387-394.