# Credit Scoring Approval Predict

Project Group W13

Kan Pei Ju Xu Shuya Zhang Yuqing

***Abstract*—The increasing credit card usage has pushed the banks to consider carefully before approving credit card applications. Banks usually use their judgment to determine whether to approve the credit card application of clients who satisfy certain criterias. Some machine learning algorithms have also been applied in credit scoring. The main objective of this paper is to build five models based on the Kaggle datasets, logistic regression (LR), decision trees(DT), support vector machine (SVM), artificial neural network(ANN) and extreme gradient boosting (XGBoost), which are commonly used to support the credit card approval decisions. Our results reveal that the overall performance of the Xgboost model is better than the other four models. Secondly, we introduce SHAP values based on XGBoost to interpret the determining factors from the model's result and hope it helps to provide understandable reasons to applicants.**

## I. Introduction

For banks or financial institutions, credit scoring models are used to help them to know how likely an individual is to repay his debts. And help them to do the decision-making of accepting and rejecting credit card applications. The wrong prediction of the credit score of the customers means economic losses. Therefore, the high accuracy rate should be one of the key criteria of the credit scoring models.

In recent years, due to the explosive growth of clients data, the model use of credit scoring models have increased significantly. More and more banks are considering adopting machine learning to compute credit scores. This approach should, in theory, deliver better performance than traditional methods--but it generates a model which is hard to explain: they are not interpretable given their complexity.

It is extremely important to interpret a prediction of a model's output to consumers and tell them what they should do to get approved by the bank. In some applications, simple models, like linear model, logistic model are often preferred because of their ease of interpretation, even if they may be less accurate than some machine learning models. However, the benefits of using machine learning have been increased by the expanding availability of big data, so a wide variety of approaches have been recently proposed to address this.

To address this problem, we use SHAP (SHapley Additive exPlanations) for interpreting the model's predictions. SHAP assigns each feature an importance value for each prediction. It includes two original characteristics. (1) SHAP can identify a new class of additive feature importance measures. (2) Theoretically, its results have been proved that there is a unique solution in this class with a set of desirable properties[1]. Based on SHAP, we provide some visualizations to explain the predictions of the model's output.

Therefore, this study has two aims. The first one is to construct a high accuracy model based on the dataset. Second, we make it easy to interpret to people who have no related background and experience such as the clients.

## II. Credit Scoring models

### A. Logistic Regression

The Logistic Regression model is one of the most useful statistical techniques for classification problems. Logistic Regression model can be used to predict the likelihood of a result that can just have two states (0 or 1) based on one or several numerical and categorical variables. It tries to find the best fitted parameters of the model to decide the probability of the binary response. If the probability is larger than the threshold value, then the observation is accepted as 1. If the probability is smaller than the threshold value, then the observation is rejected as 0.

$$p = \frac{exp(\beta_0+\beta_1 x_1+.....+\beta_n x_n)}{1+exp(\beta_0+\beta_1 x_1+.....+\beta_n x_n)}$$

where $p$ is the probability of the binary response being 1. n is the number of variables in the regression model, $x_i$ is the variable i and $\beta_i$ is the coefficient of variable i.

The advantages of the logistic model are that it is simple, fast and low-memory usage and has good interpretability. However, its disadvantages are that it requires a lot of model assumptions and the accuracy is usually lower than machine learning methods.

### B. Decision tree

The Decision Tree algorithm is one of the supervised learning algorithms. But unlike other supervised learning algorithms, the decision tree algorithm can not only solve prediction problems but solve classification problems.

Decision Tree is a tree-structured classifier, where the internal nodes represent the features of a dataset, the branches represent the decision rules and the leaf nodes represent the result. The decision tree model divides the data many times according to the specific cut-off value in the feature. Different subsets of the dataset are created by splitting, and each instance belongs to one subset. The final subsets are called terminal nodes or leaf nodes, and the middle subsets are called internal nodes or partition nodes. Therefore, it can be used for classification. The Decision Tree is usually used to create a model that could be used to predict the category of the target variable by learning decision rules from the training data.

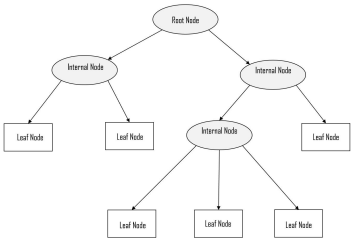


Fig. 1. Decision Tree.

The decision tree machine learning method is easy to understand because of its tree-structure and can be easily transformed to a set of corresponding rules. And they can classify both categorical data and numerical data and without priori assumptions of the data[2]. However, the decision tree is also easy to overfit the data and it's not a good approach for high-dimension data.

### *C. Support Vector Machine*

Support Vector Machine (SVM) learning method is based on the basic theory of Structural Risk Minimization (SRM). The purpose of SVM is to find the best hyperplane of two kinds of categories in the input space. The plane in the middle of two groups of data of two classes is the best hyperplane. By measuring the boundary of the hyperplane and finding its maximum point, the best hyperplane separator between the two categories can be found. The margin is the distance between the hyperplane and the closest pattern in each class [3].

SVM has the advantages that it has good accuracy and is suitable for small sample dataset, however SVM does not perform well with big dataset and is sensitive to missing values.

### *D. Artificial Neural Network*

Neural Network is a kind of soft computing that imitates the working mode of the human brain and can provide stimulation, processing and output.

Generally, Artificial Neural Network models consist of the activation function and the input layer, the hidden layer and the output layer. The input layer chooses the input signals and transfers these signals to the hidden layer and then the output layer can give the output of the final prediction . A neural network may have many layers and it can increase the number of layers to solve more complex problems.

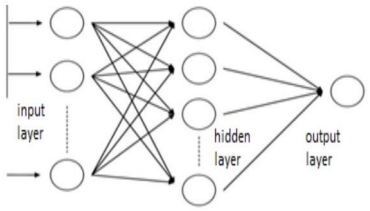


Fig. 2. Artificial Neural Network.

Artificial Neural networks are increasingly found to be useful in modeling non-stationary processes due to their associated memory characteristics and generalization capabilities[4]. And the disadvantages of ANN are that ANN requires a large number of initial parameters and takes a long time to do the modeling successfully.

### *E. XGBoost*

XGBoost is an optimized distributed gradient boosting library that is very portable, efficient and flexible. XGBoost achieves machine learning algorithms based on the gradient Boosting framework. Gradient boosting is a supervised learning algorithm, which tries to combine an ensemble of evaluations from a set of simpler and weaker models to predict the target variable.XGBoost provides a parallel tree boosting algorithm that can train models quickly and accurately. Also, XGBoost introduced linear regression and logistic regression with L1 and L2 regularization terms to control the complexity of the model. In many machine learning cases, XGBoost performs very well because XGBoost can handle various data types and relationships and many hyperparameters robustly.

But at the same time, because the XGBoost model has many parameters and the process of adjusting parameters is complex, people need to be very careful about the principle of Xgboost in order to use Xgboost well.

## III. DATA

The dataset is from Kaggle: Credit Card Approval Prediction. The description of the variables are as follows:

TABLE I. DESCRIPTION OF VARIABLES

| Table Head | Table Column Head | | |
|---|---|---|---|
| | Variable Name | Description | Type |
| 1 | ID | Client number | numerical |
| 2 | CODE_GENDER | Gender | binary |
| 3 | FLAG_OWN_CAR | Is there a car | categorical |
| 4 | FLAG_OWN_REALTY | Is there a property | binary |
| 5 | CNT_CHILDREN | Number of children | categorical |
| 6 | AMT_INCOME_TOTAL | Annual Income | numerical |
| 7 | NAME_EDUCATION_TYPE | Education level | categorical |
| 8 | NAME_FAMILY_STATUS | Marital status | categorical |
| 9 | NAME_HOUSING_TYPE | Way of living | categorical |
| 10 | DAYS_BIRTH | Age by days | numerical |
| 11 | DAYS_EMPLOYED | Days of work experience | numerical |
| 12 | FLAG_MOBIL | Is there a mobile phone | binary |
| 13 | FLAG_WORK_PHONE | Is there a work phone | binary |
| 14 | FLAG_PHONE | Is there a phone | binary |
| 15 | FLAG_EMAIL | Is there an email | binary |
| 16 | OCCUPATION_TYPE | Occupation type | categorical |
| 17 | CNT_FAM_MEMBERS | Family size | numerical |
| 18 | MONTHS_BALANCE | Record month. (0,-1,-2,…,-60) 0 is the current month, -1 is the | numerical |

| Table Head | Table Column Head | | |
|---|---|---|---|
| | Variable Name | Description | Type |
| | | previous month | |
| 19 | STATUS | Loan status.<br>0: 1-29 days due<br>1: 30-59 days due<br>2: 60-89 days due<br>3: 90-119 days due<br>4: 120-149 days due<br>5: Overdue<br>C: paid off<br>X: No loan | categorical |

The variable STATUS is the response variable that we would like to predict.The other variables are independent variables which are credit card applicants' background information.

## A. *Exploratory Data Analysis*

Exploratory Data Analysis is a good way to help us understand our dataset and think about how to deal with it for model building.

The distribution of the response variable STATUS is as Fig. 3.

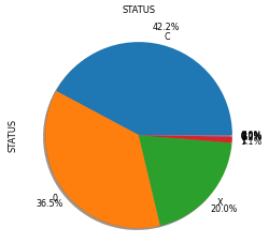


Fig. 3. Response distribution.

Then the boxplot of AMT_INCOME_TOTAL, DAYS_EMPLOYED and CNT_FAM_MEMBERS show that these three variables have some outliers as shown in Fig 4. These need to be addressed in the data next section

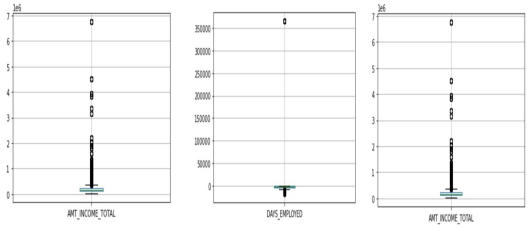

Fig. 4. Boxplot of AMT_INCOME_TOTAL , DAYS_EMPLOYED and CNT_FAM_MEMBERS

## B. *Data Processing*

### a) *Variables handling*

Categorical variables have a limited number of categories and no order, and do not have numerical data size implications. Therefore, in modeling, categorical variables cannot be dealt with directly in the same way as numerical variables. So we first need to convert categorical variables to numeric types.

Then transform the initial variable into a meaningful form. The variable DAYS_BIRTH is the number that counts backwards from the current day, 0 means current day and -1 means yesterday. So we converted it into Age = (DAYS_BIRTH/365)*(-1).

The variable DAYS_EMPLOYED is the number count backwards from the current day(0). If positive, it means the person is currently unemployed. So we converted it into the form : if DAYS_EMPLOYED >= 0 then Experience = 0 and if DAYS_EMPLOYED < 0 then Experience = (DAYS_EMPLOYED/365)*(-1).

### b) *Outliers and Missing values*

The variable OCCUPATION_TYPE has almost half of the missing values and is hard to convert to numeric, so we dropped it.

In the previous EDA process, we found that AMT_INCOME_TOTAL, DAYS_EMPLOYED and CNT_FAM_MEMBERS have outliers, so we remove these outliers by z-score.

## C. *Feature Engineering*

The length of time since the last approval of a credit card is also important information, so we set a new variable Begin_Month which is the smallest value of MONTHS _BALANCE.

And then we converted the original variable names to the more intuitive names that are easier to understand, such as converting CODE_GENDER to Gender.

## D. *Variable Selection*

The dataset has a lot of raw data variables and that may increase the complexity of the problem analysis and bring inconvenience to the model. Therefore, variable selection is an important process. Here we use the Random Forest method to select the important variables. Through the random forest process, we can calculate the importance of each feature, and then select the most important features. The importance of assessment of features is as Table2. The importance of Own_Work_Phone is 0, so this variable was dropped when building models.

TABLE II. IMPORTANCE OF FEATURES

| | Variable Name | Importance | | Variable Name | Importance |
|---|---|---|---|---|---|
| 1 | Gender | 0.019316 | 13 | Education_secondary | 0.001741 |
| 2 | Experience | 0.012053 | 14 | Family_Member_Count | 0.001146 |
| 3 | Is_Working | 0.011602 | 15 | Income_Type_Student | 0.001112 |
| 4 | Own_Mobile | 0.011058 | 16 | Age | 0.001042 |
| 5 | Income_Type_Pensioner | 0.008200 | 17 | Housing_Type_House / apartment | 0.001039 |
| 6 | begin_month | 0.005096 | 18 | Family_Status_Single | 0.001037 |
| 7 | Own_Car | 0.002243 | 19 | In_Relationship | 0.000970 |
| 8 | Own_Realty | 0.002205 | 20 | Education_ | 0.000966 |

| | Variable Name | Importance | | Variable Name | Importance |
|---|---|---|---|---|---|
| | | | | Academic degree | |
| 9 | Own_Email | 0.002203 | 21 | Housing_Type_With parents | 0.000555 |
| 10 | Income | 0.002162 | 22 | Income_Type_Working | 0.000042 |
| 11 | Own_Phone | 0.002005 | 23 | Education_Higher education | 0.000002 |
| 12 | Family_Status_Married | 0.001750 | 24 | Own_Work_Phone | 0.000000 |

### E. *Feature Scaling*

Many machine learning algorithms are very sensitive to "relative scale of features", because it is an essential step for these algorithms to calculate the distances between data. So , standardization of these features is a necessary process before building the models.

## IV. Model and Evaluation

This section explains the process of the training models and evaluates the results of the predictive models

### A. *Target Setting*

In the original data, the variable STATUS has several values and the distribution is as shown in Fig 1. Then we converted status= 0,C,X to target = 1 (good), and converted status=1,2,3,4,5 to target = 0 (bad), according to the bank's tolerance period for the credit loan repayment. And the pie chart of 'target' is as Fig. 5.

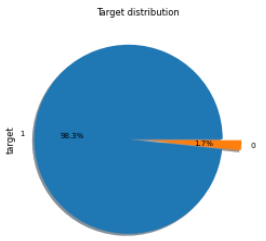


Fig. 5. Target distribution.

### B. *Data Balancing*

One problem with our data is that the ratio of credit card applications approval to failure is uneven as Fig 3. The problem with the imbalanced dataset is that there are too few examples of the few classes and too many examples of the majority class, so the model can not effectively learn the decision boundary.

SMOTE (synthetic minority oversampling technique) is a good method to balance the imbalanced dataset. SMOTE is a synthetic sampling technique. Starting from a few samples, find adjacent samples and synthesize new minority samples to make the number of minority samples consistent with that of most samples. It can balance the distribution of classes without adding redundant information[5].

Therefore, we applied the SMOTE method to balance the dataset by combining the oversampling minority class and undersampling majority class. And then we divided the balanced dataset into the training set and the testing set by 7:3 for the following modeling.

### C. *Data Modeling*

This section we use the training set to train the logistic regression model, decision tree model, Support Vector Machine model, Artificial Neural Network model and XGBoost model by Python separately.

### D. *Model Evaluation*

After the modeling process, we need to evaluate the classification results of different models and select the best classification model. For this binary classification problem, we first need to sum up the results of the model output on the testing set into a confusion matrix.

TABLE III. CONFUSION MATRIX

| | | actual | |
|---|---|---|---|
| | | **accept** | **reject** |
| **predicted** | **accept** | TP[a] | FP[b] |
| | **reject** | FN[c] | TN[d] |

a. TP is the number of results that actual classification is positive and the predicted classification is also positive

b. FP is the number of results that actual classification is negative and the predicted classification is positive

c. FN is the number of results that actual classification is positive and the predicted classification is negative

d. TN is the number of results that actual classification is negative and the predicted classification is negative

Then we have Accuracy = (TP+TN)/(TP+TN+FP+FN) and the larger the Accuracy, the more accurate the classification result of the model.

However, 'Accuracy' is not a good indicator to measure results of classification while the sample is unbalanced. So, we applied the ROC curve and AUC value to improve the evaluation of imbalanced sample classification and show more accurate evaluation. ROC curve is a graphical graph that explains the diagnostic ability of a binary classifier system by plotting the true positive rate (TPR) against the false positive rate (FPR) where TPR = TP/(TP + FN) and FPR = FP/(FP + TN).

AUC is the area under the ROC curve. Generally the value of AUC is between 0.5 and 1, and the classifier with the higher AUC is better. Generally, if the AUC value between 0.5 and 0.7 we think the classification isn't effective, if the AUC value between 0.7 and 0.85 we think it's fairly good and if the value of AUC is greater than 0.85, the classifier can be considered to be very good[6]. Therefore, we set 0.7 as the success measure for model validity.

The Accuracy value and ROC curve of each model are as Fig. 6 and Fig. 7.

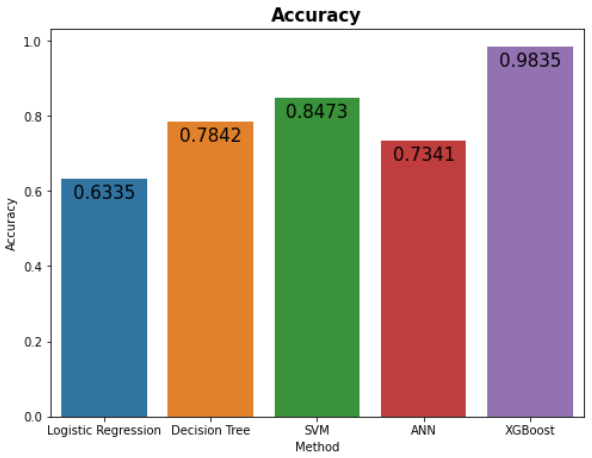


Fig. 6. Accuracy of models.

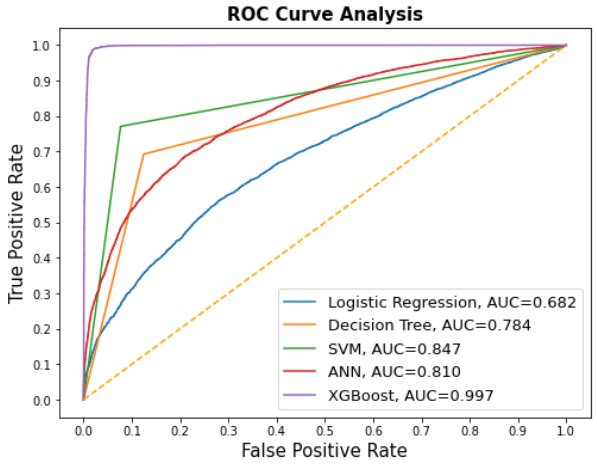


Fig. 7. ROC curve of models.

From Fig. 6 and Fig. 7, we found that XGBoost performs the best with the highest accuracy 0.98 and AUC 0.99 among all the models. So we choose XGBoost as our best classification model. The second-best model is SVM with the highest accuracy 0.85 and AUC 0.85. And the logistic model performs the worst with the accuracy 0.63 and AUC 0.68, which is lower than the other machine learning models.

## V. Interpretation

### A. Logistic Regression Model

Logistic Regression model is considered as an interpretable model. In the results of the logistic regression model, variables with p-values less than 0.05 are considered as significant variables and the odds can provide important information on how variables affect the output of the model.

TABLE IV. Categorical Variables

| Variable Name | P-value | Odds |
|---|---|---|
| Gender | 0.035089 | 0.836460 |
| Own_Car | < 0.001 | 1.122471 |
| Own_Realty | 0.100692 | 1.145649 |
| Own_Phone | < 0.001 | 1.057816 |
| Own_Email | < 0.001 | 1.058330 |
| Is_Working | < 0.001 | 1.011028 |
| In_Relationship | < 0.001 | 1.019305 |
| Edu_Academic degree | < 0.001 | 1.257538 |
| Education_Higher education | 0.129869 | 0.949938 |
| Education_secondary | 0.153695 | 1.037500 |
| Income_Type_Pensioer | < 0.001 | 0.982463 |
| Income_Type_Student | < 0.001 | 1.153316 |
| Income_Type_Working | < 0.001 | 1.011028 |
| Family_Status_Married | < 0.001 | 1.019305 |
| Family_Status_Single | < 0.001 | 0.981061 |
| Housing_Type_House/apartment | < 0.001 | 0.959606 |
| Housing_Type_With parents | < 0.001 | 1.042094 |

For categorical variables, Own_Car, Own_Realty, is_Working, Edu_Academic degree, Income_Type_Pensioer, Family_status_Married, Housing_Type_House/apartment, Housing_Type_With parents are significant.

TABLE V. Continuous Variables

| Variable Name | P-value | Odds |
|---|---|---|
| Age | 0.061350 | 1.082046 |
| Experience | < 0.001 | 1.217455 |
| Income | < 0.001 | 1.079242 |
| Family_Member_Count | < 0.001 | 1.155739 |
| begin_month | 0.833108 | 1.818198 |

For continuous variables, such as Experience, Income and Family_Member_Count variables are significant.

Even though the logistic regression model is easier to interpret, its accuracy and AUC is only 63% and 0.6689 separately, not high enough. Its outcomes are not trustworthy.

### B. XGBoost Interpretation

The model evaluation shows that the XGBoost model performs better than other models – but it generates a model which is usually hard to explain. Therefore, we generated

the SHAP value plot, which provided an understanding of the magnitude of variables' impact on the model from the length of the bar chart and the direction from the color.


Fig. 8. Intuitive illustration of SHAP(image source: SHAP Github)

The idea behind the SHAP value is to provide interpretations of machine learning models, which is to use fair allocation results from game theory and assign credit for a model's output among its input variables. The input features of a model with players in a game have to match the function of the model. In the game theory, a player can choose to join a game or not, and we have to define what it means for an input variable: to "join" a model means that variable has been "in the model" if we know the value of that feature, and it has not been in the model if we don't know the value of the variable. To evaluate a model when only a subset of features S are part of the model, we integrate the other features using a conditional expected value formulation. This formulation is as follows[7]:

$$E[f(X)|do(X_S = x_S)]$$

In the form we have already known the values of the variables in S because we set them. It is so complicated to compute the SHAP values which are NP-hard in general, so we use the SHAP package in Python to plot the feature's value.

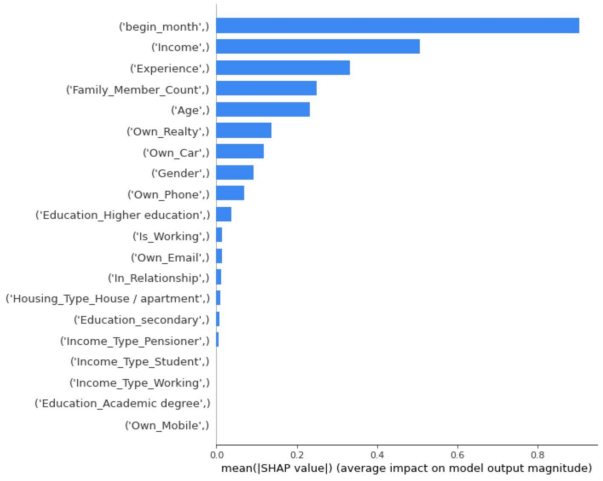


Fig. 9. Feature importance of model.

From the bar chart, we can obtain the variables with high feature importance, Begin Month, Income, Working experience, Family Size, Age, Own property, Own car.

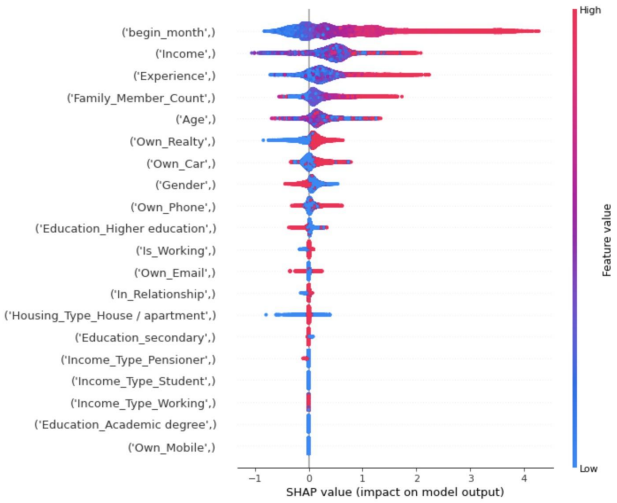


Fig. 10. Summary SHAP value plot

In this plot Fig.10, Feature importances are ranked in descending order. The horizontal axis displays whether the effect of that value is associated with a lower or higher prediction value ( y value). The color in vertical, indicates whether that variable is low (blue) or high ( red) for that observation.

To state it clearly, large begin_month have large y value predictions; higher Income leads to higher y value predictions; the ownership of property leads to higher y value where higher y value indicates the higher probability of approval (or closer to 1).

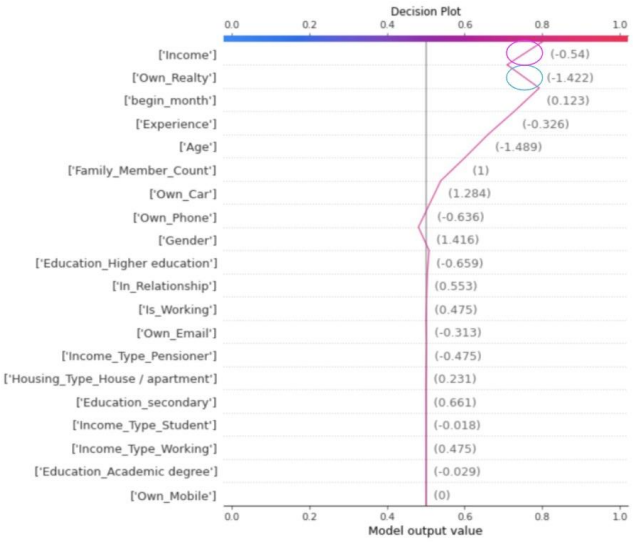


Fig. 11. Decision plot.

Fig.11 is the Decision Plot of one applicant in our dataset, which indicates that the features Income moving the decision to a positive value for this application, as the pink circle, while features Own_realty (observation: no realty) moving the decision to the negative value for this application, as the blue circle.

Another example, we randomly selected an applicant who has been rejected in our model.

TABLE VI. THE DATA OF THE APPLICANT

| Variable Name | Observation |
|---|---|
| Gender | male |
| Begin_month | 26 months from last loan |
| Income | 180,000.0 |
| Experience | 9 |
| Family_Memer_count | 2(married) |
| Age | 31 |

According to this table, we tried to provide answers for the applicant: What should the applicant do to get approved by the bank?

Then we created 2 control groups to test:

TABLE VII. DATA OF THE CONTROL GROUP

| Control Group | 1 | 2 |
|---|---|---|
| Gender | male | male |
| Begin_month | 26 months from last loan | 26 months from last loan |
| Income | 200,000.0 | 230,000.0 |
| Experience | 13 | 11 |
| Family_Memer_count | 2(married) | 2(married) |
| Age | 35 | 33 |

The model outcomes of the two control groups are approved. For the group1, if the applicant increases his income by 20,000 and work years by 4, then this applicant would get approved. For the group2, the applicant just needs to increase his income by 50,000 and work years by 2, then this applicant would get approved.

## VI. CONCLUSION

### A. High accuracy and AUC

The AUC scores in the above models except logistic regression are over 70%, which reached our success measure. Therefore, the decision tree, SVM, ANN and XGBoost models are good algorithms that provide credit card approval classification with high accuracy and AUC. And among them, we provided XGBoost and SVM as the best models to be implemented in credit card scoring models, which their AUC are over 85%.

### B. Interpretability

In the XGBoost SHAP value, we found that applicants who: have higher income, more working experience, nearer month of last approval of credit card have more possibilities to get approval. And according to the SHAP value, we are able to provide reasons and suggestions for the applicants instead of providing the untransparent result from the black-box machine learning models only.

## VII. LIMITATIONS AND FUTURE STUDIES

In the data processing step, we drop the variable: Occupation_Type, a categorical variable, because more than half of this item is missing value and the types are too many to deal with. Therefore, it is impossible for us to obtain information and the impact of Occupation may cause in the models. In the future study, it is encouraged to take the Occupation_Type variable into consideration, which we can try to group all the missing values and mark them as "Not Recorded", so that regression analysis can be performed after processing without losing sample size and then we can also evaluate the impact of missing values to decide whether to drop it.

Secondly, we didn't consider the intersection terms in the model, such as Gender*Income. We believe there might be some significant intersection terms that have important impacts on the classification model.

Finally, it would be better if we assign a credit score for each applicant, the score would be more straightforward for them to know the distance between approval and rejection. Moreover, the applicants with different score levels may apply for different loans. In the future study, in order to implement the research of credit scoring, we might need to quantify all variables and make sure the specific contributions of the application results even for every classification categorical variable.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

The first author conducted exploratory data analysis and built the ANN and SVM models. The second author conducted data description and visualization and constructed the LR model. The third author summarized the methods in literature and constructed Decision Tree and XGBoost models. The team finished the modeling evaluation and interpretation. All authors wrote the paper and approved the final version.

## REFERENCES


[1] Scott M.lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions" arXiv preprint arXiv:1705.07874.

[2] Yap Bee Wah and Irma Rohaiza Ibrahim (2010), "Using data mining predictive models to classify credit card applicants," 2010 6th International Conference on Advanced Information Management and Service (IMS), 394-398.

[3] Sugiyarto, I., Sudarsono, B., & Faddillah, U. (2019). Performance Comparison of Data Mining Algorithm to Predict Approval of Credit Card. Sinkron : Jurnal Dan Penelitian Teknik Informatika, 4(1), 149-157.

[4] Tian-ShyugLee, Chih-ChouChiu, Yu-ChaoChou and Chi-JieLu (2004), "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines" Computational Statistics & Data Analysis Volume 50, Issue 4, Pages 1113-1130.

[5] Alberto Fernández, Salvador García, Francisco Herrera and Nitesh V. Chawla. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Int. Res. 61, 1 (January 2018), 863–905.

[6] Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 3, pp. 299-310, March 2005, doi: 10.1109/TKDE.2005.50.

[7] Scott M.lundberg,"An introduction to explainable AI with Shapley values" Available at An introduction to explainable AI with Shapley values — SHAP latest documentation[Last accessed on 14 January 2021]