Credit Card Default Prediction with Data Modeling

Zhaohong Wang[1(✉)], Cheng Han Wen[2], Wenda Zhou[3], and Jun Zhang[4]

[1] Department of Statistics, 725 S Wright St, Champaign, IL 61820, USA
zw59@illinois.edu
[2] Department of Computer Science, 630 W. 168th St, New York, NY 00555-9642, USA
[3] Department of Applied Engineering Physics, 109 Clark Hall, Ithaca, NY 14853, USA
[4] School of Humanities, 25 Zhujiang Ave, Hexi District, Tianjin 300222, China

Abstract. This paper aims to build a predictive model to identify credit card default and minimize losses for financial institutions. The study uses data from the Credit Card Approval Prediction dataset on Kaggle, with 36,457 rows and 17 predictors. The credit card default is an unbalanced outcome variable, with most customers paying their credit card balance on time. The authors compare four models (logistic regression, KNN, random forest, and XGBoost) in terms of AUC, F1 score, accuracy, precision, and recall. The random forest and XGBoost models perform the best with AUC scores of 0.771 and 0.753, respectively. The findings suggest that the use of predictive models can help financial institutions identify good and bad customers and make better decisions regarding issuing credit cards.

Keywords: Credit card default · Predictive modeling · Credit risk assessment

1 Introduction

Credit card default is the failure to make at least minimum payment on the statement balance over an extended period of time. The credit card default is an indicator of whether a customer is a bad or good customer. Failing to predict the future default can lead to immense business loss for the credit card companies. Therefore, it is important for the credit card companies to predict the probability of future defaults and to decide whether to issue a credit card to an applicant. The aim of this project is to build the best prediction model that predicts credit card default with various customer information from the credit card application process. The dataset we used is Credit Card Approval Prediction data from Kaggle. The raw data consists of 439k rows and 19 columns. After data preprocessing, the final dataset used for analysis had 36,457 rows and 17 predictors such as gender, house/car ownership, income, education level, marital status and occupation. The credit card default is an unbalanced outcome variable where the signals are rare since most of the customers made their credit card payment before the due date. A number of prediction models were fit and compared in terms of AUC, F1 score, accuracy, precision and recall. Models we considered are logistic regression, KNN, random forest, and XGBoost. Our best models were the random forest and the XGBoost with AUC of 0.771 and 0.753 respectively.

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 1494–1503, 2023.
https://doi.org/10.2991/978-94-6463-198-2_155

2 Exploratory Data Analysis Data Set

The raw data set consisted of two separate data sets. The first is the application data set that contains information about the applicants such as gender, age, income, occupation type, and housing (Table 1). These are the variables we used to predict credit card defaults. The table shows the list of predictors included in the application data set. The second data set is the credit record data set that contains information about how long the customer's credit card remained active, and how many times the customer made the credit card payment past due. This credit record data set is a repeated measures data that contains credit records from multiple months for the same customer.

2.1 Descriptive Statistics

Before we move on to the model building step, we'd like to do more analysis to have a better understanding of all the predictors in Table 2 and how they are related to each other and the target variable (Table 3).

First, we looked at descriptive statistics of the continuous predictors. The scales of continuous variables are very different. This may affect the analysis when we calculate the euclidean distances. Therefore, we transformed the continuous variables to the same scale. The 'number of children' and 'income' variables are highly skewed. This indicates that there may exist outliers that can affect the model. The 'number of children' variable has a high kurtosis and median of 0 so that we can assume that most of the customers do not have children. The maximum value of 'days_employed' is 365243 which corresponds to 1000 years. This does not make sense so we transformed those values to 0.

We also compared the box plots of the continuous predictors between the good and bad customers. The Fig. 1 and Fig. 2 shows the box plot of income variables against the

Table 1. Variable Names

Type	Variable Names
Continuous	Income, age, days employed, family members
Binary	gender, car, realty, workphone, phone, email
Categorical	education, income type, marital status, housing, occupation

Table 2. Descriptive Statistics

	Mean	SD	Median	Min	Max	Skew	Kurtosis
children	0.42	0.77	0	0	19	3.61	47.84
income	181,252.92	99,375.49	157,500	27,000	1575,000	2.66	15.72
age	–15,991.09	4,246.02	– 15,607	– 25,152	– 7,489	–0.15	–1.05
days_employed	61,720.20	139,646.27	– 1,374	– 15,713	365,243	1.71	0.94
family_members	2.18	0.93	2	1	20	1.85	18.24

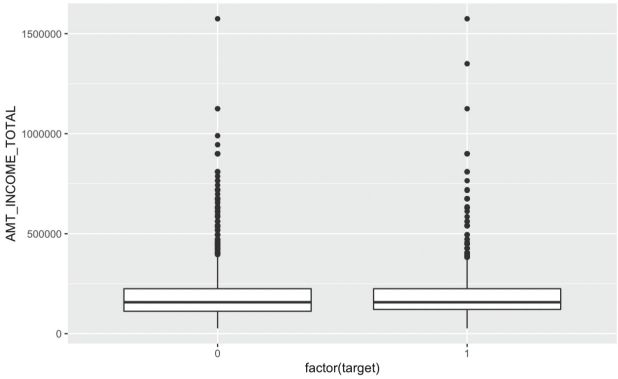

Fig. 1. Boxplots of Income

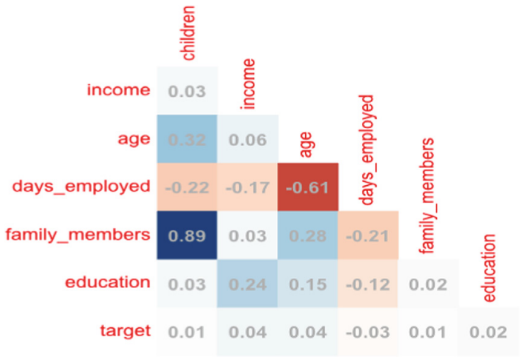

Fig. 2. Correlation Plot

two groups. Referring to the two figures, the distributions of income in two groups did not seem to be distant from each other.

2.2 Correlation

The 'number of children' and 'family members' have a strong linear relationship with correlation coefficient value of 0.9. And they contained redundant information. This can cause collinearity and rank deficiency problems. Therefore, we discarded the children variable from the data set. None of the predictors seems to have a strong linear relationship with the binary target variable. This may affect our linear models for predicting the target variable.

Table 3. Default Status Variable

Default	Meaning	Default	Meaning
0	1–29 days past due	4	120–149 days overdue
1	30–59 days past due	5	Overdue or bad debts, write-offs for more than 150 days
2	60–89 days overdue	C	paid off that month
3	90–119 days overdue	X	No loan for the month

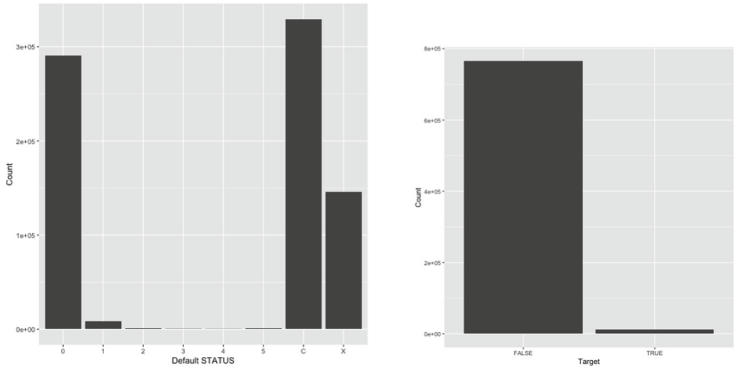

Fig. 3. Histograms of Target Variable

3 Data Cleaning and Data Preprocessing

3.1 Defining the Target

The target credit card default status variable from the credit record data set had eight categories. 0 indicates that the customer has paid off the bill before the due. C indicates that the customer has paid off the bill before the due. And X indicates that the customer has no loan. We combined these three as good customers. And we defined a bad customer as someone who did not make the payment for more than 30 days past due. Then we made the target variable as a binary outcome variable that takes on a value of either 0 or 1. 1 indicates that the customer had at least 30 days past due. After defining the target, the target variable became an unbalanced variable. Most of the customers made their credit card payment before the due date or at least within a month past due (Figs. 3 and 4).

3.2 Missing Values

There were 11k missing values in the occupation type variable. This was about one third of the data set. Given the larger proportion of the missing values and existing 17 types of occupations, we thought recoding the missing values as "others" is better than trying

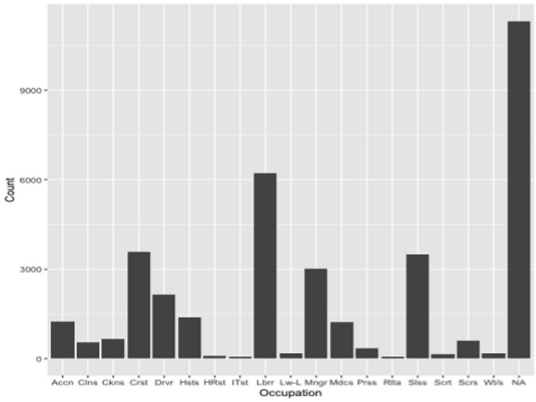

Fig. 4. Histogram of Occupation Type

to do imputation or discarding the missing rows. Therefore, we decided to recode the missing values in the occupation variable as "others".

3.3 Duplicates

There are 2 types of duplicates in the dataset. One type is simply duplicates, two or more rows have all duplicate values and we believe it's due to some data error in data collection. We simply just drop all the duplicate rows and make sure there is only 1 row for each unique ID. The second type of duplicates is more complicated. Some rows have different IDs while all other variables are exactly the same, including income, date of birth, employee days, etc. These values are specific for each person, so we believe these rows are from the same customer, and these customers have multiple applications or credit cards. We think that dropping the second type duplicate rows might oversample the signal of the target, since customers who defaulted in one credit card may not default in their other credit cards.

3.4 Regrouping

In the dataset, there are some categorical variables with very sparse data, some levels only contain small amounts of observations, so they may not be able to show a significant signal to the target variable in the model building process. By regrouping, we create a new level by combining some of the sparse levels together and make sure there are enough observations after combining (Fig. 5).

In order to make sure the groups after regrouping have the same patterns, we only combine levels with similar distributions. In the dataset, we did regrouping to 4 variables: occupation type, housing status, education level, marriage status. Here is one example of regrouping, the first bar plot shows the distribution of housing status before regrouping and the second plot shows the distribution after regrouping. There are originally 6 levels, and except House/apartment level, all other 5 levels have limited observation (some levels less than 200). And after regrouping, even though it's still unbalanced, all levels have more than 2000 observations (Fig. 6).

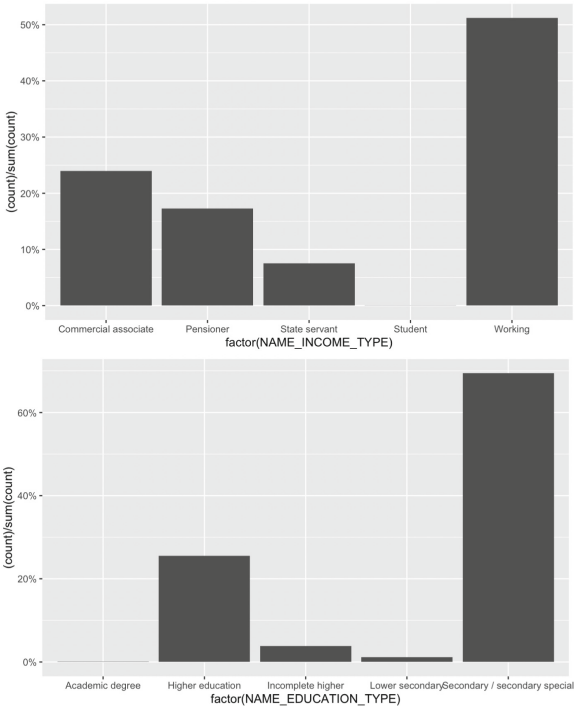

Fig. 5. Histograms of Income Type (left) and Education (right)

4 Model Building

We randomly sampled 70% of the individuals as the training data set and the rest as the testing data set. The training dataset has around 25,000 rows and the test set has around 10,000 rows. We used n-fold cross-validation for the models in which the tuning parameters are required. Accuracy, AUC, f1 score, recall and precision are used as model evaluation criteria. We have tried 3 types of models: Linear classification model, K-nearest neighbor model and tree based classification model.

4.1 Linear Classification Model

For linear classification models, we have tried Logistic regression and penalized logistic regression models [2]. We used the logistic regression model as the benchmark model, and used the penalized logistic regression as feature selection. The AUC of both linear classification models are around 0.55 and the accuracy is around 55%. Due to the low AUC and low accuracy, these 2 linear classification models are not predictive at all. The diagnostic plots for the linear models also suggest the assumptions of logistic regression models are violated. The normal QQ plot indicates the violation of linearity assumption (Figs. 7 and 8).

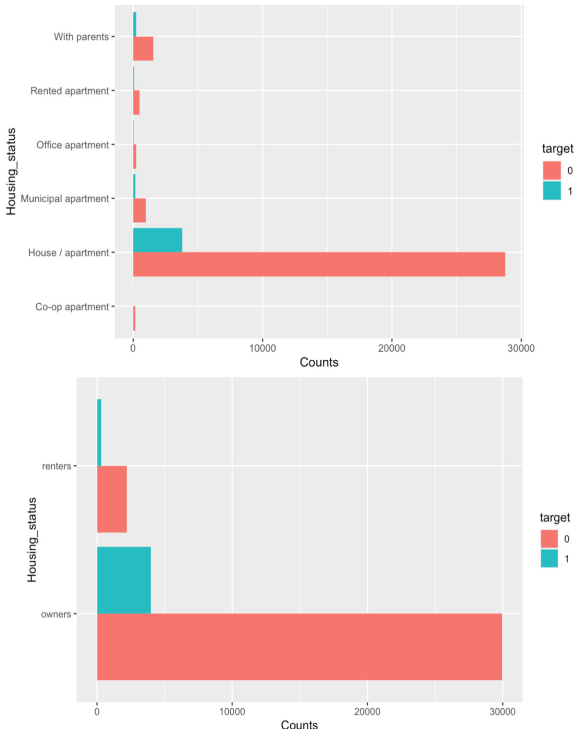

Fig. 6. Histograms of House Status Before (top) and After (bottom) Regroupping

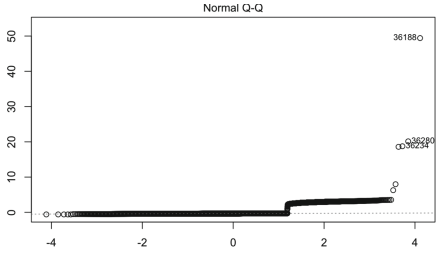

Fig. 7. Normal Q-Q plot of Logistic Regression Model

4.2 KNN

We used 5-fold cross-validation for tuning the k value. The best k value with the smallest classification error was 9. KNN performed way better than the logistic regression models. The AUC value increased to 0.712. This improvement in performance may be caused by the duplicates. When we randomly split the training and the testing data, there will exist duplicates in both data sets. The distance between those duplicates will be zero, and therefore the target will always be correct for the duplicates. This implies that this KNN model might work for the test data set with duplicates, but may not perform well for a new data set if it has no duplicate in the training data set.

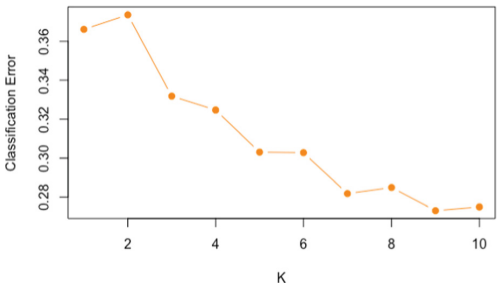

Fig. 8. Cross Validation Plot of KNN

Table 4. Model Performance with Cut-Off Threshold 0.5

Model	Accuracy	AUC	F1 score	precision	recall
LR	0.563	0.550	0.213	0.491	0.136
PLR	0.544	0.539	0.212	0.511	0.134
KNN	0.877	0.712	0.258	0.178	0.470
RF	0.887	0.771	0.266	0.171	0.597
XGBoost	0.886	0.753	0.379	0.291	0.544

* LR = Logistic Regression; PLR = Penalized Logistic Regression; RF = Random Forest

4.3 Tree Based Classification Model

For tree based classification models, we chose to try random forest and XGBoost. When training the model, we used cross-validation to do feature selection and parameter tuning to get the best model. Both models have pretty good performance which indicates there may be some features not linear related to the target. So the tree based models can better capture the patterns.

5 Model Performance and Cutoff Threshold Selection

Since the output of the classification models are the probabilities and eventually we need to define good and bad customers, approve and deny credit card applications, we need to choose a threshold to cut off. In this step, we also used cross validation to get higher F1 scores for the training set. The Table 4 shows the model performance with cut off threshold = 0.5.

And Table 5 shows the model performance with their best cut off threshold. Among all the 5 models we tried, Random Forest and XGBoost are the ones with best performance metrics, and they have similar accuracy, AUC and F1 scores. Overall, we believe these 2

Table 5. Model Performance with Best Cut-Off Threshold

Model	Cut-Off	Accuracy	AUC	F1 score	precision	recall
LR	0.11	0.442	0.550	0.216	0.641	0.130
PLR	0.38	0.346	0.539	0.219	0.762	0.128
KNN	0.2	0.840	0.712	0.348	0.385	0.348
RF	0.19	0.864	0.771	0.435	0.436	0.434
XGBoost	0.36	0.876	0.753	0.442	0.409	0.481

* LR = Logistic Regression; PLR = Penalized Logistic Regression; RF = Random Forest

Table 6. Confusion Matrices of Random Forest (left) and XGBoost (right)

	Target	Target
Prediction	0	1
0	8175	577
1	1450	736

	Target	Target
Prediction	0	1
0	9045	776
1	580	537

models are better than the other 3 models. And comparing those 2 tables with different cut-off thresholds, it's clear the threshold selection can improve F1 scores and precision and make them better models.

By looking at the confusion matrices in Table 6, we believe there is still a lot of space for each model that needs to be improved. Even though random forest and XGBoost have around 90% accuracy, the precision is not good enough. There are still enough false positives which may cause the loss of potential customers, and the false negatives can cause massive business losses [3].

6 Conclusion

In conclusion, both random forest and xgboost classification models have pretty good test set performance. They have relatively higher AUC, accuracy, F1 score, and precision than other 3 models. But on the other hand, there are still a lot of false predictions. So,

the models may not work for new data collected from a potential customer or work in real business productions. Since we only have limited information for each customer, the models are performing pretty well. In order to make an improvement and build better and successful models for business, we still need to collect more information and bigger sample size.

Data source. https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction..

References

1. Ying Chen, Ruirui Zhang, "Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network", Complexity, vol. 2021, Article ID 6618841, 13 pages, 2021. https://doi.org/10.1155/2021/6618841
2. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010 Nov;107(44):776–82. doi: https://doi.org/10.3238/arztebl.2010.0776. Epub 2010 Nov 5. PMID: 21116397; PMCID: PMC2992018.
3. Weaver CGW, Basmadjian RB, Williamson T, McBrien K, Sajobi T, Boyne D, Yusuf M, Ronksley PE. Reporting of Model Performance and Statistical Methods in Studies That Use Machine Learning to Develop Clinical Prediction Models: Protocol for a Systematic Review. JMIR Res Protoc. 2022;11(3):e30956. doi: https://doi.org/10.2196/30956. PMID: 35238322; PMCID: PMC8931652.
4. Chen, N., Ribeiro, B. & Chen, A. Financial credit risk assessment: a recent review. Artif Intell Rev 45, 1–23 (2016). https://doi.org/https://doi.org/10.1007/s10462-015-9434-x
5. Doko F, Kalajdziski S, Mishkovski I. Credit Risk Model Based on Central Bank Credit Registry Data. Journal of Risk and Financial Management. 2021; 14(3):138. https://doi.org/https://doi.org/10.3390/jrfm14030138
6. Carlos Andres Zapata Quimbayo, Carlos Armando Mejía Vega. (2023) Credit risk in infrastructure PPP projects under the real options approach. Construction Management and Economics 41:4, pages 293-306.