Credit Card Approval Prediction Final Report

Adel Abdalla, Hong Gi Baek, Fan Bu, Xinyi Chen, Peinan Zhou

Introduction

With the increasingly inclined of digital transactions, credit cards have become not only a popular payment method, but also a form of financial products that represents the creditworthiness of users. Utilizing credit cards is a deferred payment structure where consumers take on debt that accrues with time. It is important to repay the debt timely to avoid interest charges. While banks have measures to protect against late payments, they still face significant risks if many customers fail to pay their debts. To manage the risk and mitigate potential losses of unpaid debts, banks have to assess applicants' financial stability to determine their lending risk.

Although noteworthy progress has been made recently with the financial analysis techniques, the issue of credit card approvals remains complex and controversial. In this project, we outline a comprehensive approach to refining the credit card approval process by leveraging advanced machine learning algorithms that prioritize not only predictive accuracy but also model transparency.

Dataset

The dataset used in this project is the Credit Card Approval Prediction dataset from Kaggle. It provides client records for a financial service, with a variety of features such as gender, car and real estate ownership status, income, education level, marital status, housing type, employment history, contact information, occupation type, and family size. Additionally, there's another dataset containing monthly financial account records linked to clients, indicating the month of the record, the client's ID, and the status of their loan or account, including information on overdue payments and payment statuses.

EDA

During the EDA, we focused on understanding correlations, notably identifying a relevant correlation between numeric features (CNT_FAM_MEMBERS and CNT_CHILDREN), which, although intuitive, did not necessitate adjustments in our model due to its moderate strength. Our analysis of variable distributions highlighted that the majority of the applicants possessed secondary or higher education levels and were primarily aged between 27 and 43 years.

Data cleaning involved classifying applicants into 'good' and 'bad' credit categories, with those over 30 days overdue categorized as risky. Significant data gaps, especially the 30% missing in 'Occupation Type', were addressed through imputation techniques to maintain data integrity. Additionally, we managed outliers in income and employment durations to mitigate potential analysis distortions.

These EDA efforts revealed crucial insights, such as the lack of a strong correlation between age and income, challenging conventional assumptions used in credit scoring. This informed our decision to not weigh age heavily in our predictive model. The EDA also led to a better understanding of the demographic and financial profiles of potential cardholders, guiding our feature selection and data transformation strategies to enhance model accuracy and reliability.

Data Preparation

Based on our EDA findings, we created strategies on data preparation, including data cleaning and feature engineering. The steps are as follows: Feature Labeling, Merging Data, handling outlier, categorical variable encoding, imputing missing values, and sampling using SMOTE. Firstly, we created labels to distinguish "good" users and "bad" users by labeling every user that was ever overdue by 30 days (STATUS 1-5) as risky or default users (1) and the rest as non-risky (0). Then We merged the applicants' information dataframe with the overdue behavior dataset, and grouped by each applicant. Now, each row represents a credit card user. After merging the two datasets, we removed the outliers in

AMT_INCOME_TOTAL and DAYS_EMPLOYED (Figure 1). Then we encoded all the 8 categorical variables, with special treatment to OCCUPATION_TYPE because it has too many levels. We categorized the occupations and decreased the number of levels from 19 to 8 to prevent possible sparsity issues. About 30% of OCCUPATION_TYPE information is missing, so we used KNN to impute the missing values. Finally, because of the extreme imbalance in the classes (Figure 2), To get more information from the dataset, we performed SMOTE, an oversampling method, on the training set (Figure 3). After all treatment, we have 49,351 data in the training set and 6,986 in the testing set for modeling.

Modeling Methodologies

To predict the approval status for credit card applications, we implement a wide range of machine learning models to evaluate their effectiveness in the decision making process. We delve into both traditional algorithms and more complex models to seek a tradeoff between predictive accuracy and interpretability. The models enlisted for initial evaluation as our baseline are:

- **Logistic regression** - Use Binary Classification to estimate the probability of whether a credit card will be approved or not.
- **Decision trees** - A transparent model that recursively splits the data based on the features to create rules/thresholds for predicting approval or rejection, making decisions understandable.
- **Random Forest** - An ensemble approach that averages multiple decision trees to reduce overfitting and enhance performance
- **Gradient Boosting Machines (GBM)** - Sequentially builds decision trees, with each tree refining the predictions of the previous ones.
- **Support Vector Machines (SVM)** - Find the hyperplane that best separates approved and rejected credit card applications based on input features.

After establishing baselines with these models, we will delve into more sophisticated modeling techniques.

- **Neural Networks** - A vanilla feed-forward neural network with two hidden layers, which allows us to harness more complex non-linear relationships in the data without sacrificing too much interpretability.

Results and Findings

In our project, we aimed to fine-tune our predictive models to achieve high recall while maintaining a satisfactory level of precision. This strategy is particularly pertinent in credit risk assessment, where the consequences of false negatives—failing to identify potential risk—can be substantially costly. Among various classifiers evaluated, Neural Network and Random Forest emerged as superior in recall, precision, and F1 scores, outperforming Logistic Regression, GBM, SVM, and Decision Trees. Our baseline models performed as follows:

1. Decision tree: 85% Accuracy, .375 Precision, .37 Recall, .644 Roc-Auc, 389 F1 Score.
2. Random Forest: 86% Accuracy, .414 Precision, .368 Recall, .649 Roc-Auc, 389 F1 Score.
3. Gradient BOOSTING: 88% Accuracy, .38 Precision, .009 Recall, .508 Roc-Auc, .019 F1 Score.
4. Neural Network: 83.35% Accuracy, .28 Precision, .28 Recall, .685 Roc-Auc, .28 F1 Score.

We decided to try to improve our Random Forest as it performed best overall. We also decided to try to improve our Neural network, as they are renowned for their extremely large capacity.

Neural Network: A class weight was introduced to the loss function to prioritize the accurate classification of the minority class (risky customers), significantly shifted the model's focus towards the minority class, resulting in a more balanced detection capability as reflected by the achieved F1 score of

0.3782, and increasing the recall to .537. Note other attempts at model improvements included: batch normalization, skip connections, and larger capacity models. None, however, none of these helped improve performance with the exception of changing the class weights.

Random Forest: Adjusting the decision threshold to 0.25 amplified the model's ability to flag risky customers, thereby enhancing recall. Despite a ROC-AUC score similar to the Neural Network's, the Random Forest achieved a marginally higher F1 score, suggesting a slight edge in maintaining a balance between precision and recall.

Through strategic modifications to the model parameters aimed at countering the imbalanced class distribution, both models showed promising results in identifying credit risk. While the ROC-AUC scores imply a comparable level of class differentiation by both models, the Random Forest's higher F1 score particularly emphasizes its capacity to accurately identify risky customers, minimizing the incidence of false positives—an important advantage in the predictive modeling of credit risk.

Discussion

Challenges. A principal challenge in this project centered on the precision-recall trade-off encountered with our imbalanced dataset. Our initial approach using undersampling significantly improved the model's recall to 0.6. However, this method also led to an unacceptable decrease in precision, which dipped below 0.2, thus yielding a suboptimal F1 score. The underperformance of the undersampling approach prompted us to implement SMOTE. We adjusted the weight values in our Neural Network's loss function, and modified the decision threshold for our Random Forest model. These modifications were carefully calibrated to enhance the model's ability to accurately identify the positive class without skewing predictions too far in either direction.In credit risk prediction, where both over-predicting and under-predicting risk carry consequences, achieving equilibrium between these metrics is crucial. The models' performance illuminated the inherent conflict in optimizing these competing objectives within the constraints of our imbalanced dataset.

Model Fairness. We are aware of the growing concern over biased models in credit card approvals. Using features like gender, property ownership, income, education, and occupation in a credit card approval model could inadvertently introduce biases, potentially impacting certain demographic groups. To investigate the potential fairness problem in our model (mainly on gender), we checked the confusion matrix for each group as well as evaluated four metrics: Statistical Parity, Demographic Parity, Separation, and Sufficiency. Statistical Parity measures whether outcomes are equally distributed across different groups. Probability for credit Card Approval for Female is 0.83 and for Male is 0.81. Demographic Parity ensures that the predicted outcome is independent of the protected attribute. Among those predicted to be approved by our model, 69.0% are Female, 31.0% are Male. Separation assesses the extent to which predictions are independent of the protected attribute given the true outcome. Of those actually approved, the probability of approving using our model for F is 0.87, for Male is 0.86; Sufficiency evaluates whether the predicted outcome contains all the information relevant to the protected attribute. Of those predicted to be approved by our model, 87.0% were actually approved for Female and 86.0% were actually approved for Male. In summary, our model is not biased towards minority groups; in fact, it slightly favors female applicants in credit card approval.

There are potential next steps we could take to further our project. We could enhance the existing models by exploring advanced algorithms and hyperparameter tuning for improved accuracy. Additionally, incorporating alternative data sources, such as social media, could provide richer insights into consumer behavior. Developing a system for real-time predictions could significantly speed up the decision-making process, improving customer satisfaction. It’s also crucial to continuously monitor and

adjust the model to ensure it remains fair across all demographic groups and complies with relevant financial regulations and privacy laws.

Appendix

Figure 1. Handling Outlier

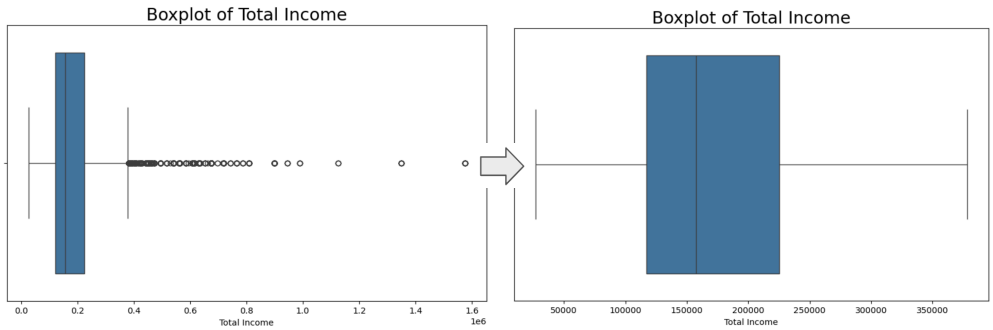

Figure 2. Dataset Imbalance (Original Merged Dataset)

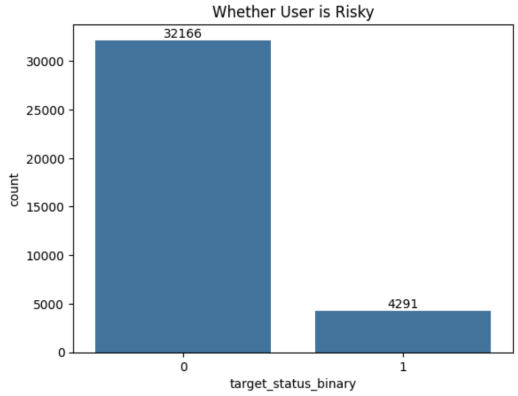

Figure 3. SMOTE on Training set