Predicting Credit Card Defaults using Machine Learning

Project Proposal

Team Members:

- Chia Tien TANG
- Nicolas REBOULLET
- Ariel TORJMANE
- Nhu PHAM

1. Introduction/Motivation

The project aims to address the critical issue of predicting credit card defaults, which is central to managing risk in the consumer lending business. Defaulted credit cards can lead to significant financial losses for banks and negatively impact the overall financial ecosystem. Early prediction of credit card defaults can help banks take preventive measures, offer financial counseling, or adjust credit limits, ultimately leading to better customer experience and sound business economics.

The dataset used for this project is obtained from American Express and contains time-series behavioral data and anonymized customer profile information. The primary objective is to leverage machine learning algorithms to predict the likelihood of a cardholder defaulting based on historical data and customer attributes.

2. Problem Definition

The problem revolves around binary classification, where the goal is to predict whether a credit card holder is likely to default or not. The target variable is binary, with "default" and "non-default" as the two classes. The dataset provides features related to customer behavior, transactions, and profiles, which can be used to build predictive models.

3. Importance

Credit card defaults have far-reaching implications for financial institutions and the broader economy. By accurately predicting defaults, banks can:

- Optimize credit card approvals.
- Make dynamic credit limit adjustments based on predicted default risks.
- Provide targeted financial counseling for high-risk individuals.

4. Potential Applications

The project's outcomes can find applications in various sectors, including:

- Banking and financial services.
- Credit risk assessment.
- Debt management.
- Regulatory compliance.

5. Related Work

Previous research in credit scoring has utilized traditional methods like logistic regression. Recent advancements include ensemble methods and deep learning techniques to improve accuracy. We will build upon this body of work and leverage modern machine learning techniques to enhance the predictive power of our models.

6. Dataset

We used the Amex credit card transaction dataset provided by Kaggle, which originally contained a large amount of data. However, due to computational capacity limitations, we focused our analysis on a subset of the data, specifically the first 5000 rows.

7. Methodology

Approach

Section 1 - Feature Engineering

Data Overview

- Features for training: 26 features in total, including float64(18), int64(2), object(6).

Feature Selection & Engineering (Step by Step):

- Unbalance of Data: The imbalance in the 'Credit_Score' distribution is acknowledged.
- Prevent Overfitting and Privacy Conflict: Dropped features 'Customer_ID', 'Name', and 'SSN' due to privacy concerns and potential overfitting. The 'Month' column was also removed as it was deemed not relevant for credit scoring.
- Additional Feature Engineering:
  - Features such as 'delinquency_count', 'delinquency_severity', 'total_spend_recent', and 'average_monthly_spend' were created, providing more insights into customer behavior.
  - Advanced feature engineering techniques like dimensionality reduction with PCA were considered, indicating the potential usefulness of PCA in this scenario.

Section 2 - Model Tuning and Comparison

Model Name | Hyperparameter | Accuracy Rate (CV=5)

1. Logistic Regression
   - Hyperparameters: max_iter=1000, penalty="l2", solver="lbfgs", class_weight = "balanced"
   - Accuracy Rate (CV=5): 0.9998 (approx.)

Additional Insights from Data Analysis:

- Outlier Analysis: Histograms and box plots for selected columns ('D_39', 'B_1', 'D_41', 'D_145') were created to understand the data distribution and identify outliers.
- Dimensionality Reduction: PCA analysis suggested that dimensionality reduction might benefit the dataset, considering a significant number of components (109) required to explain 95% of the variance.

2. Decision Tree

- Hyperparameters: (Specify the hyperparameters used)
- Accuracy Rate: 1.0

Additional Insights from Data Analysis:

- **Perfect Accuracy**: The Decision Tree model achieves a perfect accuracy of 1.0, indicating that it correctly classifies all 10,000 samples in the dataset. This remarkable performance suggests that the model has learned the training data very well.
- **Precision and Recall**: Both precision and recall for both classes (0.0 and 100.0) are also perfect at 1.00. This means that the model not only correctly identifies all positive cases (high recall) but also avoids making false positive predictions (high precision).
- **F1-Score**: The F1-scores for both classes are also perfect at 1.00, demonstrating an excellent balance between precision and recall for both classes.

Macro and Weighted Averages: Both macro average and weighted average F1-scores are also perfect at 1.00. These values reflect the overall model performance, taking into account the class distribution.

2. SVM:

- Hyperparameters: C=1.0, kernel='rbf', gamma='auto'.
- Accuracy Rate (CV=5): 0.998 (approx.)

Additional Insights from Data Analysis:
Outlier Analysis: Histograms and box plots for selected columns ('D_39', 'B_1', 'D_41', 'D_145') were created to understand the data distribution and identify outliers.
Dimensionality Reduction: PCA analysis suggested that dimensionality reduction might benefit the dataset, considering a significant number of components (109) required to explain 95% of the variance.

3. Gradient Boosting Machine:

- Accuracy Rate: 0.86
- Additional Insights from Data Analysis:
- **Class Imbalance**: The classification report reveals a significant class imbalance, with a higher number of samples in class 0 compared to class 1. Addressing this imbalance may be crucial for model performance.
- **Precision-Recall Trade-off**: The model exhibits a trade-off between precision and recall. Class 0 has a higher precision (0.87) and recall (0.94), indicating that it is better at correctly classifying samples of class 0. In contrast, class 1 has a lower precision (0.79) and recall (0.61), suggesting that the model is less accurate in predicting class 1 instances.
- **F1-Score**: The F1-score, which balances precision and recall, provides a comprehensive measure of model performance. The F1-score for class 0 (0.91) is higher than that for class 1 (0.69), reflecting the trade-off mentioned earlier.
- **Accuracy**: The overall accuracy of the GBC model is 0.86, indicating that it correctly classifies approximately 86% of the samples. However, given the class imbalance, accuracy may not be the sole metric to assess the model's quality.
- **Macro and Weighted Averages**: The macro average F1-score (0.80) and weighted average F1-score (0.85) provide a holistic view of the model's performance across both classes. These values consider the class distribution and should be taken into account when evaluating the model.

8. Conclusion

- The Logistic Regression model showed remarkably high performance, which might indicate overfitting. It's recommended to review the model to ensure it generalizes well to unseen data.
- The Decision Tree model, while achieving perfect accuracy, likely overfits the training data. This model's complexity and the decision criteria should be reviewed for better generalization.
- Random Forest, through cross-validation, demonstrated consistent performance, making it a promising model for this dataset. However, the specific parameters and final evaluation on the test set are not provided.
- The report acknowledges the unbalance in the target variable and the importance of feature engineering. However, specific strategies to address class imbalance in the model training process are not detailed in the code.
- Advanced feature engineering techniques like PCA and the creation of interaction features were explored to enhance model performance.

9. Algorithms/Models/Tools

We will compare the performance of various models, including logistic regression, gradient boosting, and support vector machines (SVM).

11. Relation to Prior Work

Our approach builds on traditional credit scoring models but aims to incorporate modern machine learning techniques for improved accuracy and interpretability. We will explore and evaluate a range of models to find the one that best suits our problem.

112. Evaluation

Evaluation Metrics

We will use a range of evaluation metrics to assess the performance of our models. These metrics include:

- Accuracy: To measure the overall correctness of predictions.
- Precision: To measure the proportion of true positive predictions among all positive predictions.
- Recall: To measure the proportion of true positive predictions among all actual positives.
- F1-score: A balance between precision and recall.
- Area under the ROC curve (AUC-ROC): To assess the model's ability to distinguish between default and non-default cases.

13. References

- Jing Zhou, Wei Li, Jiaxin Wang, Shuai Ding, & Chengyi Xia (2019, August). Default prediction in P2P lending from high-dimensional data based on machine learning. ElSEVIER journal.
- Anas Arram, Masri Ayob, Musatafa Abbas Abbood Albadr, Alaa Sulaiman, Dheeb Albashish (2023, October).