Credit Card Behaviour Score Development

CONVOLVE 3.0
Team Name: Ping Pong Pajeets
Team Members:
Divyam Kulshrestha
Ankit Raj
Kishlay Raj

Abstract

This document outlines our approach to developing a predictive model for determining credit card behavior scores. The objective was to identify the likelihood of customer defaults using a robust machine learning framework. The report details data preprocessing, model selection, key insights, and evaluation metrics to provide a comprehensive solution aligned with the problem statement's requirements.

Introduction

Bank A seeks to enhance its risk management framework by introducing a behavior score model for credit card customers. This score predicts the probability of defaults, leveraging customer attributes and transactional data. Our objective was to create a model that effectively identifies potential risks while maintaining high accuracy and robustness.

Data Preprocessing

To prepare the data for modeling, the following preprocessing steps were performed:

1. **Handling Missing Values:** Columns with more than 70% missing values were dropped. For the remaining missing values, the median of the respective columns was used for imputation. This approach ensured a robust dataset by preserving central tendencies while minimizing distortions.

2. **Zero Dominance:** Approximately 400 columns with over 98% zero values were identified and removed. This highlighted the sparse distribution of data and ensured the model focused on meaningful patterns.

3. **Feature Correlation:** Highly correlated features (correlation coefficient > 0.9) were dropped to reduce multicollinearity, retaining only one representative feature per group. This step improved model interpretability and computational efficiency.

4. **Low Variance Features:** Columns with negligible variance were removed as they contributed little to predictive power, reducing noise in the data.

5. **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to transform the dataset. The analysis revealed that 256 components explained 95% of the variance, significantly reducing dimensions while retaining critical information.

6. **Class Imbalance Handling:** The Synthetic Minority Oversampling Technique (SMOTE) was employed to address class imbalance. This method synthetically generated samples for the minority class, enhancing the model's ability to predict rare default events.

Following these steps, the final dataset consisted of 256 features, ready for modeling.

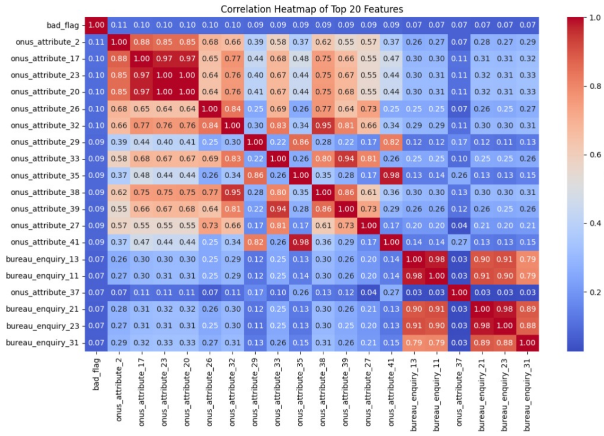

Figure 1: Correlation Matrix

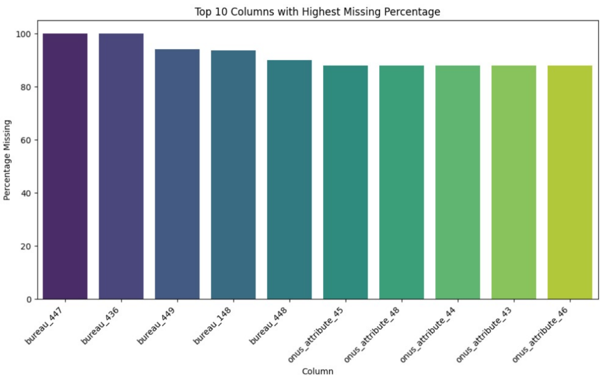

Figure 2: Top 10 Sparse Columns

Model Development and Algorithms Used

To predict default probabilities, we employed the following machine learning algorithms, each tailored to leverage the processed data:

1. **Logistic Regression:** A foundational algorithm for binary classification, logistic regression models the relationship between input features and the probability of default using a logistic function. It provides straightforward interpretability and serves as a baseline.

2. **Random Forest:** This ensemble learning method constructs multiple decision trees during training, aggregating their outputs for a final prediction. It effectively handles overfitting and identifies feature importance, aiding in deeper insights.

3. **Support Vector Classifier (SVC):** By finding the optimal hyperplane, SVC maximizes the margin between classes. With the kernel trick, it captures complex, non-linear relationships within the data, enhancing prediction accuracy.

4. **XGBoost:** An advanced gradient boosting framework designed for speed and performance. XGBoost iteratively minimizes a loss function and incorporates regularization to prevent overfitting, making it highly effective for imbalanced datasets.

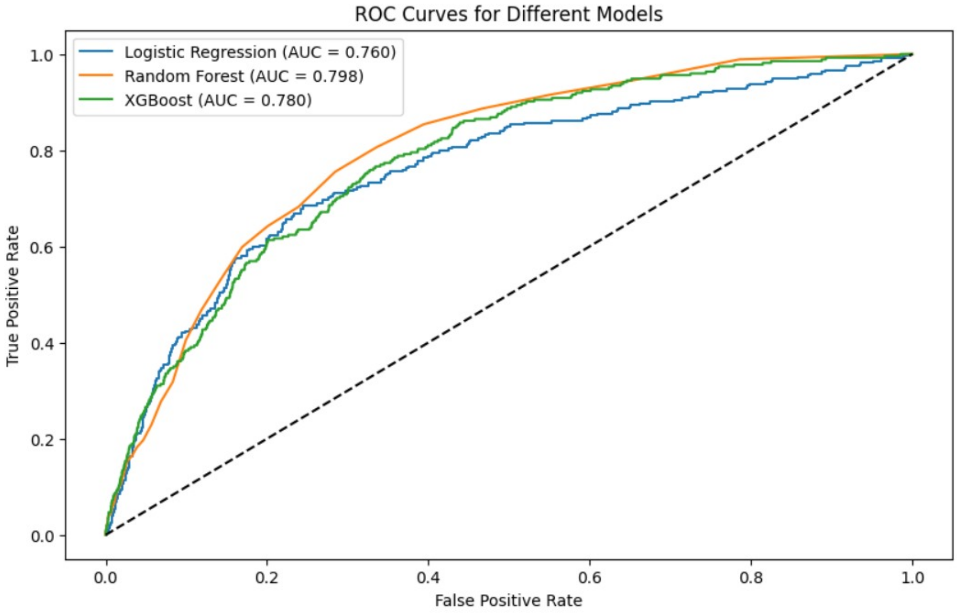

Figure 3: Comparison between different models

Model Evaluation

Each model was rigorously evaluated using the following metrics:

- **ROC-AUC Score:** Measures the model's ability to distinguish between default and non-default classes, with higher scores indicating better performance.
- **Log Loss:** Quantifies the uncertainty of probabilistic predictions. Lower values reflect more confident and accurate models.
- **Precision, Recall, and F1 Score:** Precision assesses the proportion of true positive predictions among all positive predictions, while recall evaluates the model's ability to identify defaults. The F1 score balances precision and recall, offering a comprehensive performance metric.

Hyperparameter tuning via grid search and cross-validation ensured optimal configurations for each algorithm.

Insights and Observations

- Customers with high bureau inquiry frequencies exhibited a higher likelihood of default.
- PCA effectively reduced dimensions while retaining essential variance, significantly improving model training efficiency.
- Approximately 400 columns contained over 98% zero values, highlighting the sparse nature of the dataset and the need for preprocessing.

Conclusion

Our approach successfully developed a robust predictive model for credit card behavior scores. By leveraging comprehensive preprocessing, effective dimensionality reduction, and advanced algorithms, the solution meets Bank A's requirements. This model provides actionable insights for risk management, ensuring its applicability to real-world scenarios.

References

1. Scikit-learn Documentation: `https://scikit-learn.org/`
2. XGBoost Documentation: `https://xgboost.readthedocs.io/`
3. SMOTE Technique Paper: `https://arxiv.org/pdf/1106.1813.pdf`