Project Proposal: American Express - Default Prediction

Author: Kausik Chattapadhyay

Date: 05/08/2024

DSC-680 Applied Data Science

Topic: Developing a Machine Learning Model for Credit Default Prediction Using American Express Data.

Business Problem: The project aims to leverage machine learning techniques to build a predictive model that assesses the probability of a customer defaulting on their credit card balance within American Express. By accurately predicting default events, American Express can optimize lending decisions, enhance risk management strategies, and improve the overall customer experience.

Modern life relies heavily on the convenience of credit cards for daily transactions, offering benefits like cashless payments and deferred purchases. However, the challenge for card issuers lies in predicting repayment, a complex problem with room for improvement, as seen in this competition focusing on credit default prediction. This predictive capability is crucial for managing risk in lending businesses, leading to better customer experiences and improved financial outcomes. American Express, as a leading payments company, seeks to enhance its credit default prediction model through this competition, offering participants the opportunity to contribute to a more efficient lending process and potentially gain recognition and rewards.

Datasets:

https://www.kaggle.com/competitions/amex-default-prediction/data

The objective is to predict the probability that a customer does not pay back their credit card balance amount in the future based on their monthly customer profile. The target binary variable is calculated by observing 18 months (about 1 and a half years) performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days (about 4 months) after their latest statement date it is considered a default event.

The dataset contains aggregated profile features for each customer at each statement date. Features are anonymized and normalized, and fall into the following general categories:

D_* = Delinquency variables

S_* = Spend variables

P_* = Payment variables

B_* = Balance variables

R_* = Risk variables

with the following features being categorical:

['B_30', 'B_38', 'D_114', 'D_116', 'D_117', 'D_120', 'D_126', 'D_63', 'D_64', 'D_66', 'D_68']

Task is to predict, for each customer_ID, the probability of a future payment default (target = 1).

Methods: The project will follow a structured approach:

1. **Data Exploration and Preprocessing:** This phase involves loading the datasets, handling missing values, outliers, and performing exploratory data analysis (EDA) to understand the distributions and relationships between features and the target variable.
2. **Feature Engineering:** Feature engineering will be crucial for creating meaningful predictors. Techniques such as one-hot encoding for categorical variables, creating new features based on ratios, trends, or rolling averages, and normalization or scaling of numerical features will be applied.
3. **Model Selection and Training:** We will experiment with various machine learning algorithms suitable for binary classification, including Random Forest, Gradient Boosting Machines (GBM), XGBoost, and possibly deep learning models like neural networks. Using techniques like grid search or random search will optimize model performance.
4. **Model Evaluation:** The evaluation metric, a combination of Normalized Gini Coefficient and default rate captured at 4%, will be used to assess model performance. Cross-validation and validation sets will help in evaluating the model's generalization capability.

5. **Prediction and Submission:** Once the model is trained and evaluated, we will use it to predict the probability of default for customers in the test dataset and prepare submissions in the required format.

Ethical Considerations: Ethical considerations in this project include ensuring data privacy and confidentiality by anonymizing customer information, addressing potential biases in the data or model predictions to avoid discriminatory outcomes, and transparently communicating the purpose and implications of the credit default prediction model to stakeholders, including customers.

Challenges/Issues: Challenges may arise from handling imbalanced data due to the subsampling of negative labels, feature selection from many variables to avoid overfitting, addressing potential model interpretability issues, and ensuring model fairness and transparency.

References: The project will draw insights from academic papers on credit risk modeling, industry best practices in financial risk management, regulatory guidelines such as Fair Credit Reporting Act (FCRA) and Consumer Financial Protection Bureau (CFPB) guidelines, and ethical frameworks for AI and machine learning developed by organizations like IEEE and ACM.

https://www.kaggle.com/competitions/amex-default-prediction/overview