Soheil: sf3116 Allan C: ac5439 Fernando: fc2672 Jahnavi M: jm5667 Kevin: zy2582

Credit Card Default Prediction - Project Proposal

Background and problem statement

With the increasing rate of credit card losses each year, the growing concern of credit card defaults are at an all-time high. According to Goldman Sachs, credit card losses are expected to rise another 1.3 percentage points to 4.93% with Americans owing more than $1 trillion on credit cards. Banks have experienced more risk with lending decisions, so the economy, not just banks, depends on further optimized decision-making models.

Credit default predictions have been utilized to minimize risk in consumer lending; however, better models can always be created to outperform current models. From American Express's Default Prediction dataset, we will create an optimized model to predict credit card default. Our solution will also provide an improved customer experience by making it easier for customers with less predicted risk to be approved for credit.

Dataset

The data for this project is sourced from a 2022 Kaggle competition hosted by American Express. The target binary variable is if a customer defaults on their credit card. This variable is derived by collecting the 18 months performance window after the latest credit card statement. Default event is when a customer does not pay the amount due in 120 days after the statement date.

Due to the sensitivity of this dataset, all features are anonymized. The dataset includes five aggregated profile features:

D_* = Delinquency variables
S_* = Spend variables
P_* = Payment variables
B_* = Balance variables
R_* = Risk variables

Development data contains 190 features and 5.53m rows for 459k customers.
Link to the data: https://www.kaggle.com/competitions/amex-default-prediction/overview

Proposed ML techniques

Credit default prediction is a crucial task in the finance industry, and machine learning models are commonly used for this purpose due to their ability to handle complex patterns and large datasets. These are the four models we plan to train on the dataset, and compare their performance.

Logistic regression:

Logistic regression is a simple and interpretable algorithm. It provides a basic framework for understanding the relationship between input features and binary outcomes (default or non-default). As a baseline, it helps establish a minimal level of predictive performance.

Decision Tree:

Decision trees can capture nonlinear relationships between input features and the likelihood of credit default. They are easy to interpret graphically, allowing us to understand the decision-making process. They can also highlight key features that influence credit default predictions.

Random Forest:

Random forests consist of multiple decision trees, which are combined to improve accuracy and generalization. They mitigate overfitting by averaging predictions from multiple trees, leading to better performance on unseen data. They can also handle missing values in features, a common issue in real-world datasets.

XGBoost:

XGBoost is an optimized gradient boosting algorithm that provides high performance and efficiency. It includes regularization techniques, such as L1 and L2 regularization, to prevent overfitting. XGBoost offers built-in feature importance scores, helping us identify key features affecting credit default.