American Express Project Report

Group 7

ME19B005	ME19B028	ME19B084
ME19B110	ME19B127	ME19B130

Problem Statement

We at Amex want you (the students) to develop a credit risk model that predicts the likelihood of a customer to default a payment. The Data present for a customer is at any given point of time. We need you to develop a model to predict the likelihood of the customer defaulting after 12 months.

Dataset Details

Training Data_2021.csv – Training data (83K Observations)
Test Data_2021.csv – Dataset that needs to be scored by the students.

The dataset has the customer application and bureau data with the default tagging i.e., if a customer has missed a cumulative of 3 payments across all open trades, his default indicator is 1 else 0. Data consists of independent variables at the time T0 and the actual performance of the individual (Default/Non-Default) after 12 months i.e., at time T12.

Terminologies

Classification:

In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes.

Ensemble:

Ensemble methods is a machine learning technique that combines several base models to produce one optimal predictive model.

Skewness:

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.

Outliers:

An outlier is a value or point that differs substantially from the rest of the data.

Our Approach:

Data Cleaning:

- Looking at the dataset we found there were three different types of missing values namely, “missing”, “na”, “NaN”. We replaced all three with “NaN” so that the column type can be float.
- There is one categorical column in the dataset (“mvar47”) and we changed it into binary.
- Some of the columns had a lot of missing values so we dropped columns having more than 56% of missing values.
- We also dropped highly correlated columns(more than 90% correlated columns).
- We used logarithmic transformation to deal with the skewness in the columns.

Machine learning:

- Looking at the dataset we found that it is imbalanced, i.e., it has 70% of zeros and 30% of ones, which may affect the model performance so we used the ensemble method to create batches of balanced datasets to train models and take majority vote for our final prediction. We found that 31 models gave good performance.
- We tried various machine learning algorithms and found xgboost gives the best performance. We did parameter tuning manually.
- We used “CalibratedClassifierCV()” to calibrate our model with 2 fold cross-validation.
- Since logarithmic transformation deals with skewness, we standardize the data.

Performance matrix:

As the training dataset is imbalanced, we used f1_score to measure the performance of our model in the validation dataset. But during the final submission, we used the entire data for training(without doing validation split), as increasing training data improved the performance of our model.

<u>Conclusion</u>

After all cleaning and ensembling, we got an overall performance of 60.66% in test data, which is considerably high when we compared to all other teams which had participated in the competition.