American Express - Default Prediction
Check Point 1

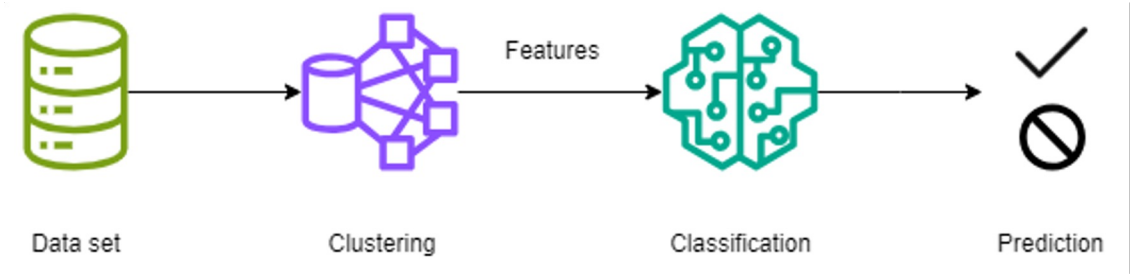
Group 24- Sarthak Garg, Yeshwanth Karra, Adithya Mahesh, Tanmay Parulekar

Problem Statement

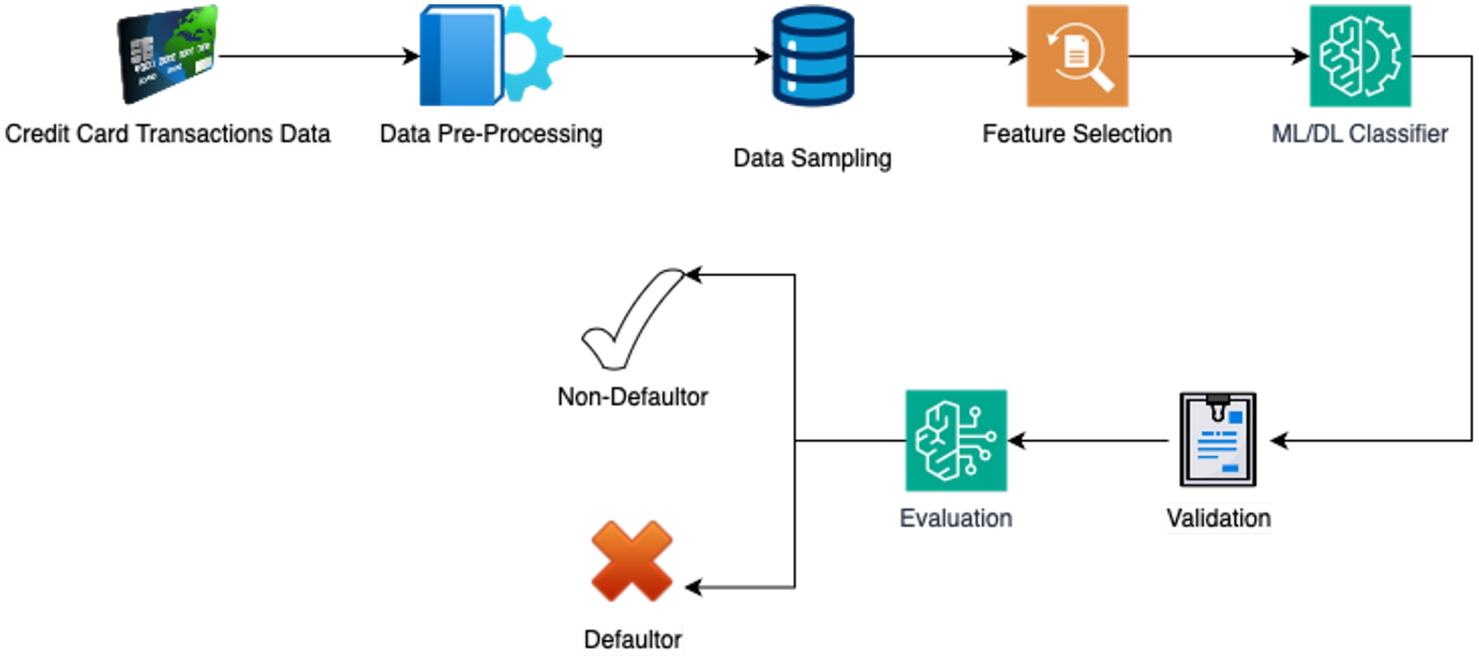
- Develop a machine learning model to predict credit default
- Improving lending decisions & enhancing the customer experience

Essential Task

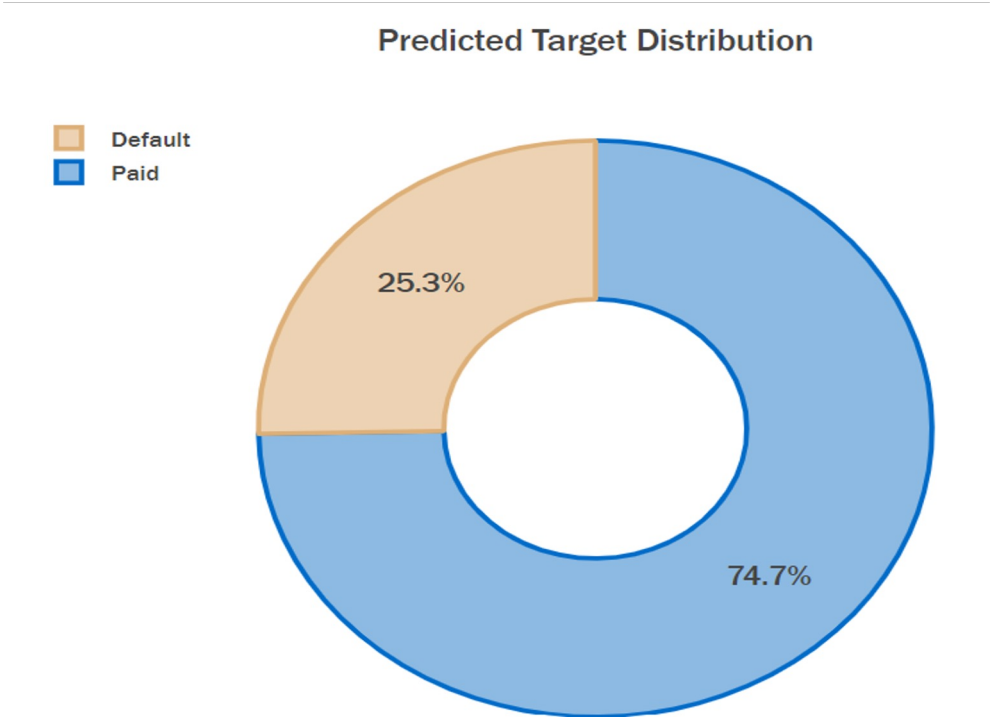
- Clustering
- Classification



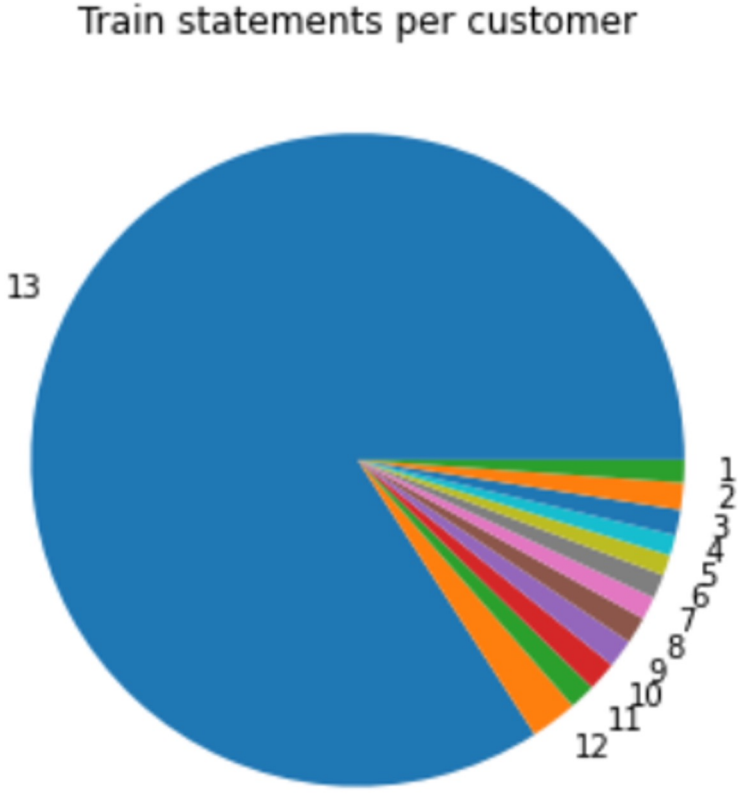
Data Mining Pipeline



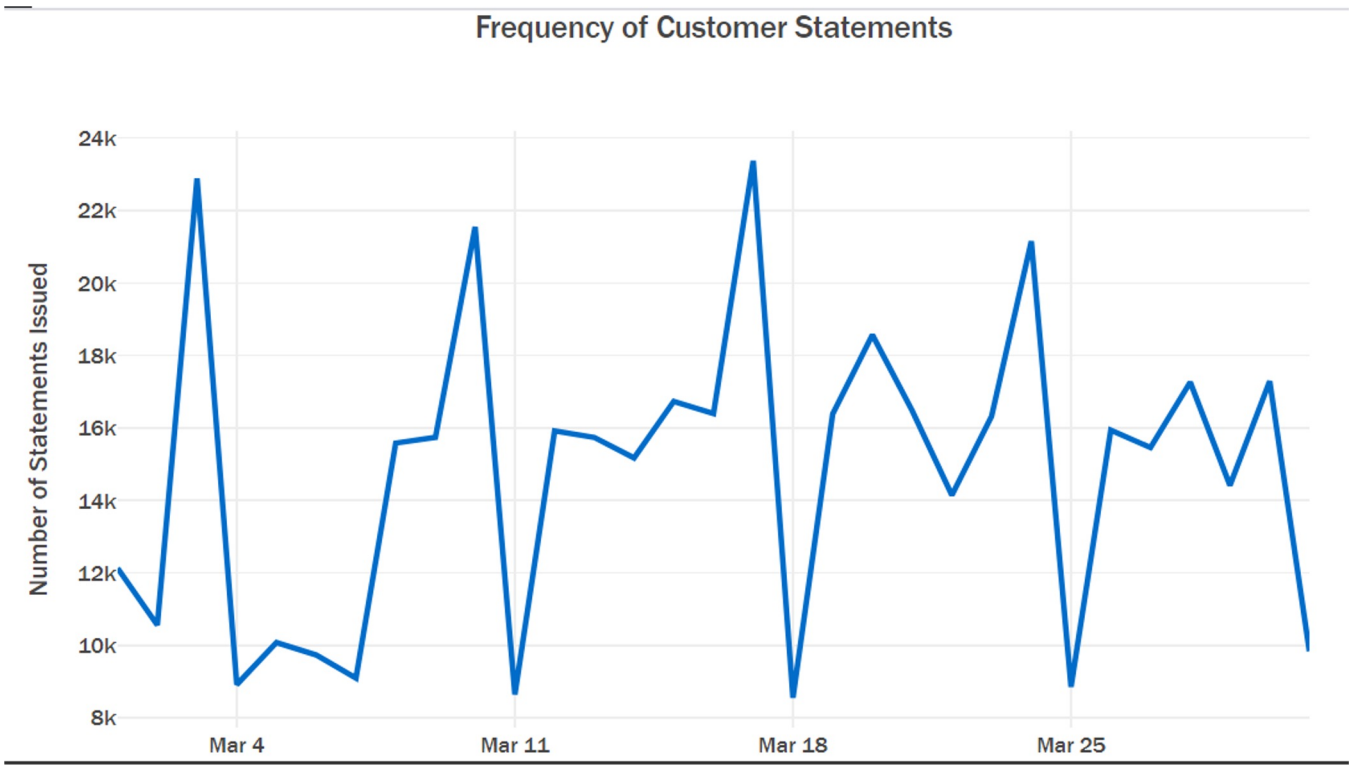
Exploratory Data Analysis



Exploratory Data Analysis

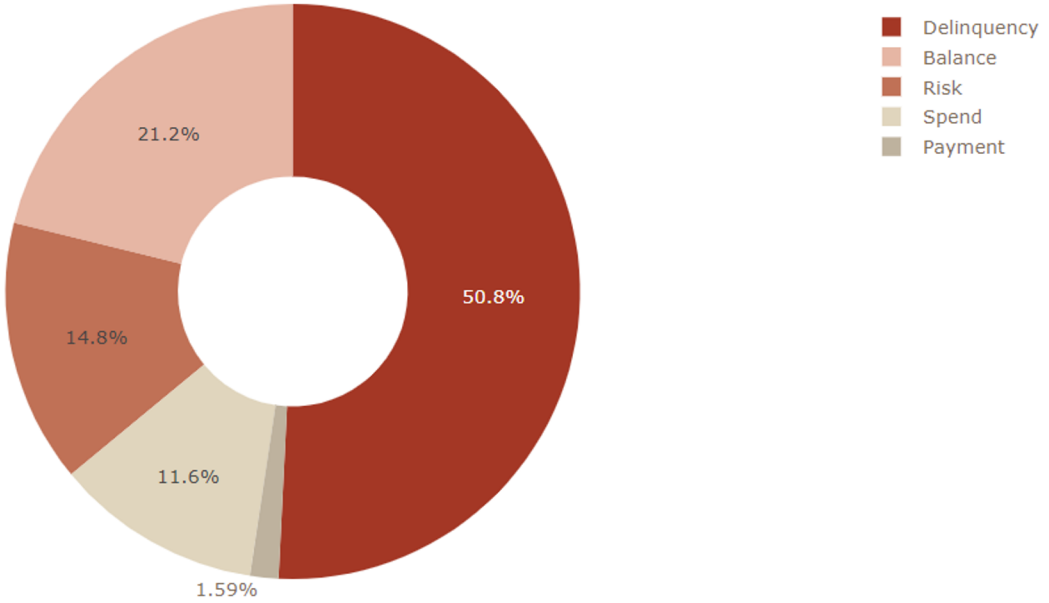


Exploratory Data Analysis

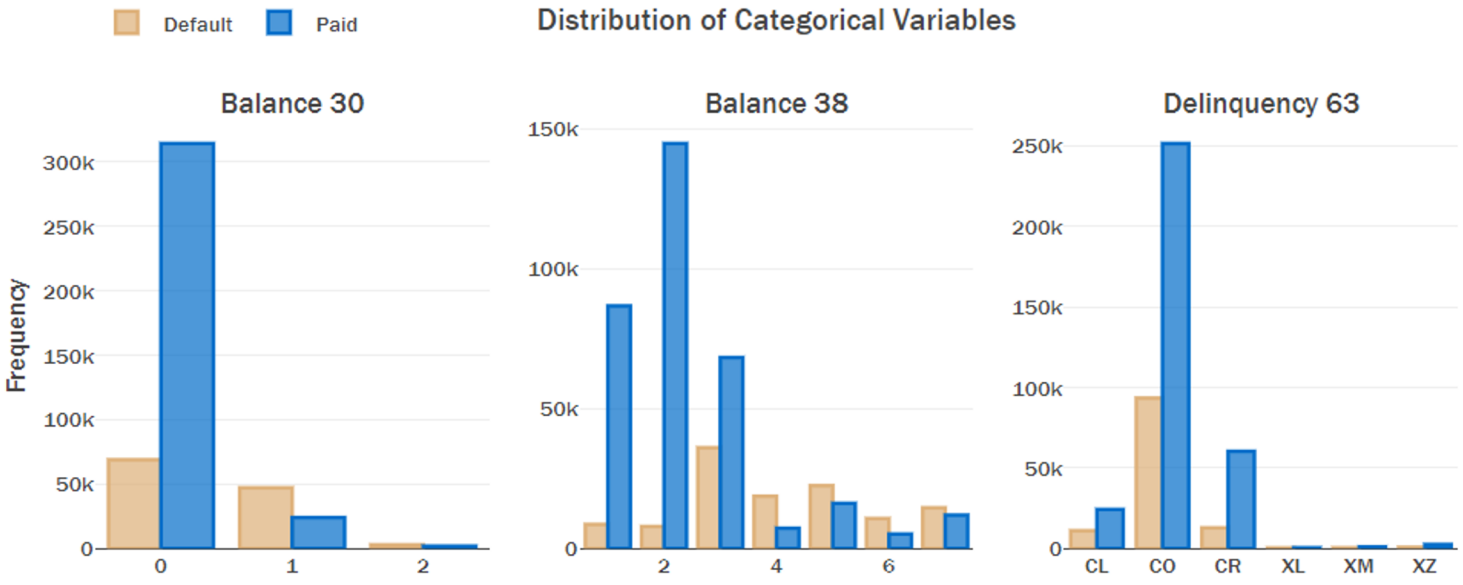


Exploratory Data Analysis

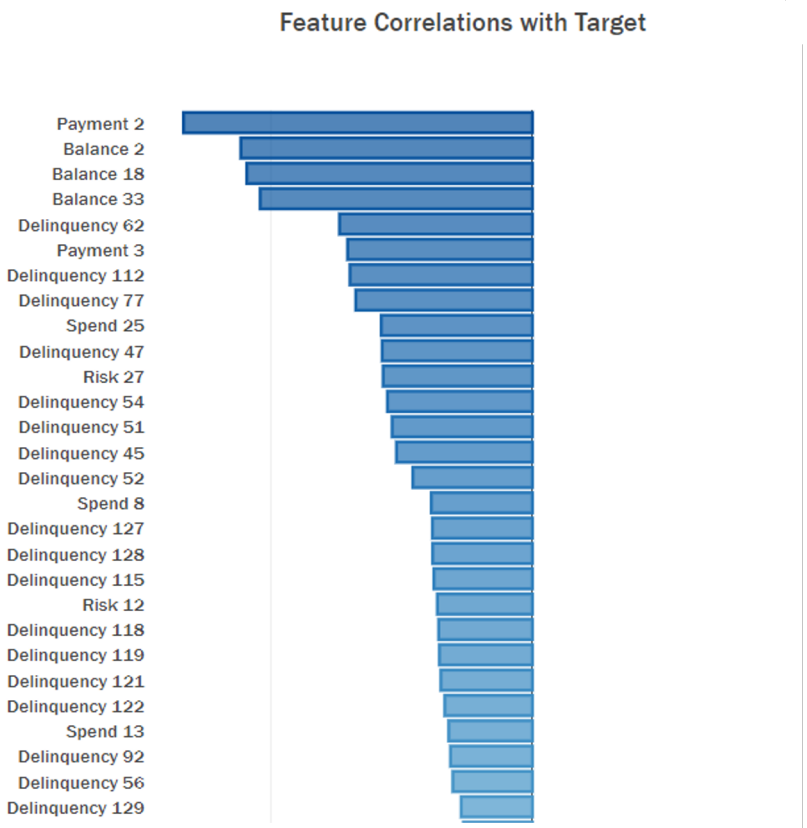
Feature Distribution

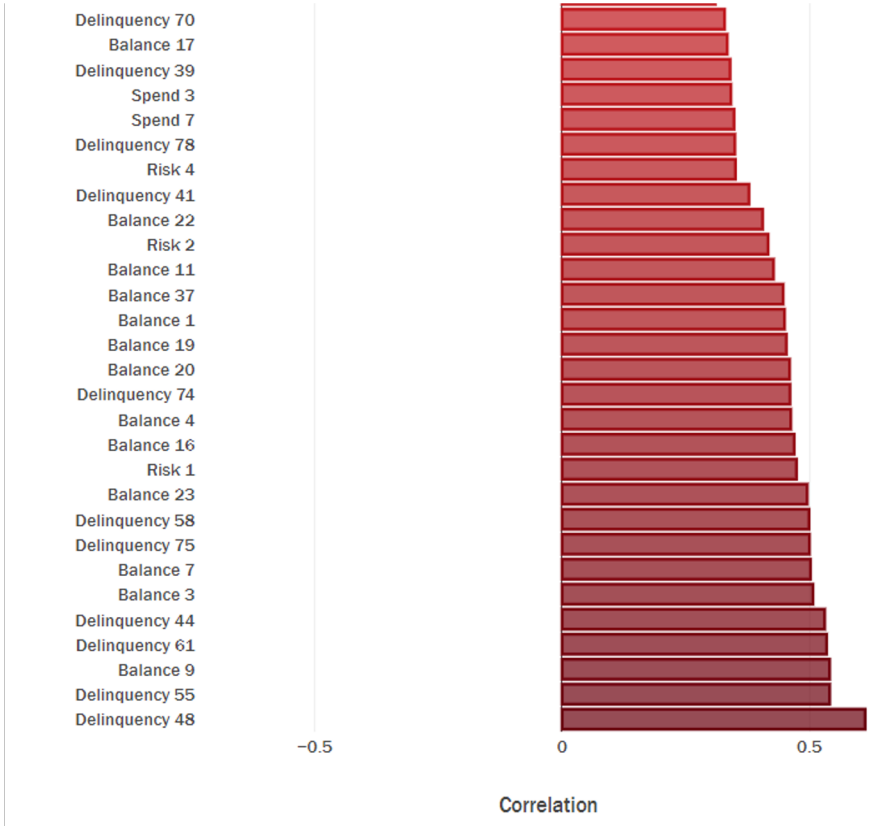


Exploratory Data Analysis



Exploratory Data Analysis





Pre-Processing

- Convert every columns to their right datatype.
- Keeping only features which give high correlation with our target
- Check for null values percentage per feature and remove ones which have more than 80%.
- Impute median values for other features with less than 80% null value.

Model Accuracy

- Random Forest
  Accuracy - 85.11

- Logistic Regression
  Accuracy - 86.90

Potential Issues

- As we are dealing with SOL due to our large dataset size, there are instances when the SOL crashes making it difficult to work with.

- As we are dealing with a completed competition, we will have to split the train dataset into train, and test. Although it is a very large dataset it still restricts the generalization of our models.

Future Works

- Try out more models
- Integrate the use of historical data for each person.
- Implement both classification and clustering algorithms sequentially using each others data.

Thank You