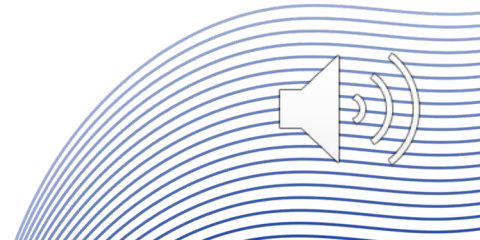
American Express - Default Prediction
Check Point 2

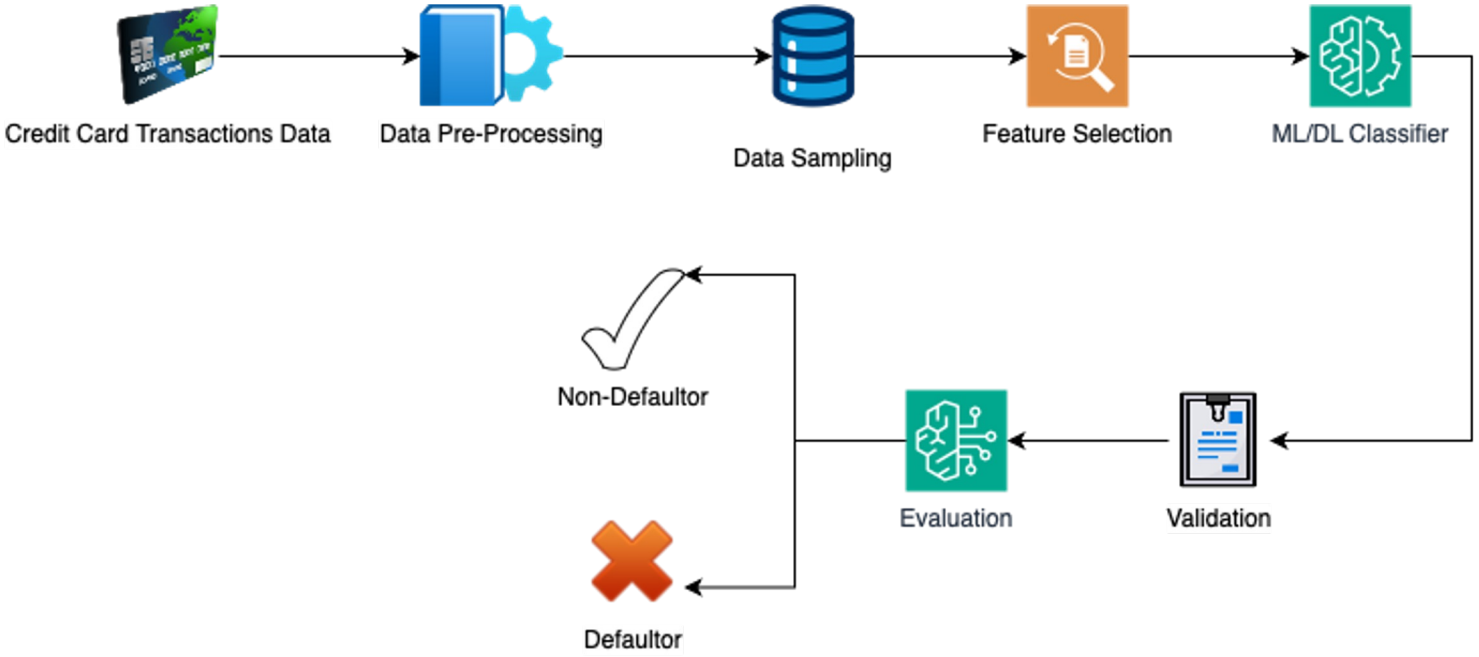
Group 24- Sarthak Garg, Yeshwanth Karra, Adithya Mahesh, Tanmay Parulekar

Problem Statement

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



Data Mining Pipeline



Classification Models

- Random Forest
- Logistic Regression
- LightGBM
- Neural Decision Tree
- XGBoost

Training models with minimal preprocessing

All 4 models go through very minimal preprocessing steps:

1. Considering one transaction per customer.
2. Removing columns with greater than 80% null values.
3. Imputing the rest null columns with median values.
4. Encoding the categorical variables.

Training models with extensive preprocessing

All 4 models go through extensive preprocessing steps:

1. Considering one transaction per customer.
2. Removing columns with greater than 80% null values.
3. Imputing the rest null columns with median values.
4. Encoding the categorical variables.
5. Performing correlation analysis.
6. Removing highly correlated features.
7. Identifying and removing irrelevant features through EDA

Models Performance Analysis

Random Forest

Metric	Minimal preprocessing	Extensive preprocessing
Accuracy	0.880741898	0.890037335
Precision	0.797430716	0.797741064
Recall	0.764882181	0.770730476
F1 Score	0.780799241	0.784003196

Models Performance Analysis

Logistic Regression

Metric	Minimal preprocessing	Extensive preprocessing
Accuracy	0.895659311	0.896240394
Precision	0.805275836	0.816460759
Recall	0.770141382	0.773058797
F1 Score	0.787279001	0.794167231

Models Performance Analysis

XGBoost

Metric	Minimal preprocessing	Extensive preprocessing
Accuracy	0.861642146	0.874369888
Precision	0.751106831	0.752991453
Recall	0.766090978	0.766129937
F1 Score	0.758543435	0.759503879

Models Performance Analysis

LightGBM

Metric	Minimal preprocessing	Extensive preprocessing
Accuracy	0.890029432	0.90208754
Precision	0.8004179	0.815065378
Recall	0.803523339	0.804364901
F1 Score	0.801967417	0.809679788

Models Performance Analysis

Neural Decision Tree

Metric	Minimal preprocessing	Extensive preprocessing
Accuracy	0.886751209	0.898666415
Precision	0.890884069	0.897372932
Recall	0.897751209	0.898666415
F1 Score	0.894306344	0.897849928

Choosing Baseline and Classification Models

From reviewing the performance metrics of all the models before and after the preprocessing, we decided to select the following models as the actual classification models:

1. Neural Decision Tree

2. LightGBM

And the following as baseline models:

1. Random Forest
2. Logistic Regression
3. XGBoost

K Fold Cross Validation

LightGBM

Average accuracy over 5 folds: 0.9016719972631104
Average precision over 5 folds: 0.813485928819214 1
Average recall over 5 folds: 0.8047751350682919
Average F1 score over 5 folds: 0.8091031513944082

K Fold Cross Validation

Neural Decision Tree

Average accuracy over 5 folds: 0.9011185217779729
Average precision over 5 folds: 0.9012873162730687
Average recall over 5 folds: 0.9011185217779729
Average F1 score over 5 folds: 0.9011588568015778

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

- Implement both classification and clustering algorithms sequentially using each others data.
- Training models with historical data.

Thank You