Customer Transaction Prediction

By: Rohitkumar Keswani

MSc Big Data Science

with Industrial Experience

200361138

Problem Statement

- Predicting which customers will make a future transactions with banks (financial institutions) irrespective of transaction in the past.

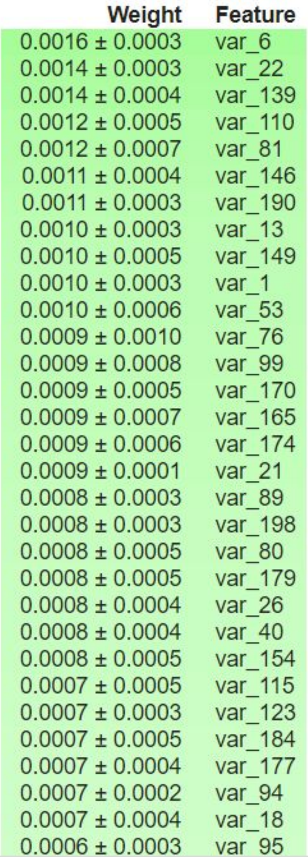
Dataset

- The dataset is of real customers of Santander bank, due to this the dataset is anonymized for the privacy reasons.

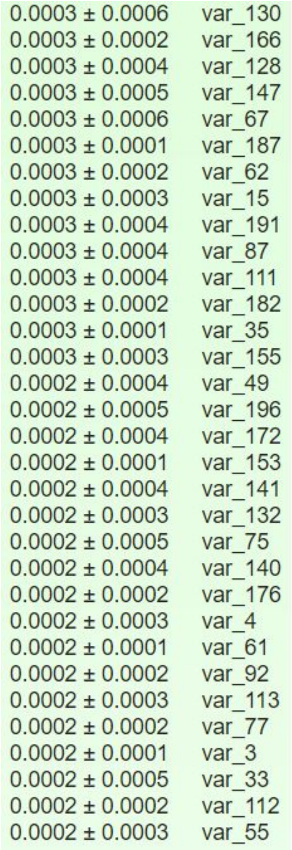
- The dataset contains 200 anonymized features and a binary target column with 200,000 rows of training data, the value of these features are also manipulated by some data pre-processing techniques which makes it harder to interpret.

Machine Learning Explainability

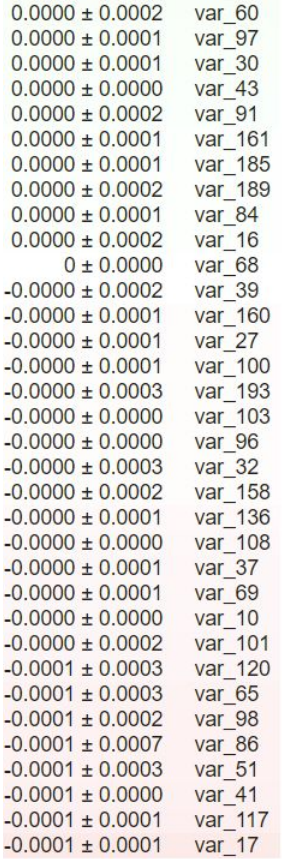
Permutation Importance



Weight	Feature
0.0016 ± 0.0003	var_6
0.0014 ± 0.0003	var_22
0.0014 ± 0.0004	var_139
0.0012 ± 0.0005	var_110
0.0012 ± 0.0007	var_81
0.0011 ± 0.0004	var_146
0.0011 ± 0.0003	var_190
0.0010 ± 0.0003	var_13
0.0010 ± 0.0005	var_149
0.0010 ± 0.0003	var_1
0.0010 ± 0.0006	var_53
0.0009 ± 0.0010	var_76
0.0009 ± 0.0008	var_99
0.0009 ± 0.0005	var_170
0.0009 ± 0.0007	var_165
0.0009 ± 0.0006	var_174
0.0009 ± 0.0001	var_21
0.0008 ± 0.0003	var_89
0.0008 ± 0.0003	var_198
0.0008 ± 0.0005	var_80
0.0008 ± 0.0005	var_179
0.0008 ± 0.0004	var_26
0.0008 ± 0.0004	var_40
0.0008 ± 0.0005	var_154
0.0007 ± 0.0005	var_115
0.0007 ± 0.0003	var_123
0.0007 ± 0.0005	var_184
0.0007 ± 0.0004	var_177
0.0007 ± 0.0002	var_94
0.0007 ± 0.0004	var_18
0.0006 ± 0.0003	var_95

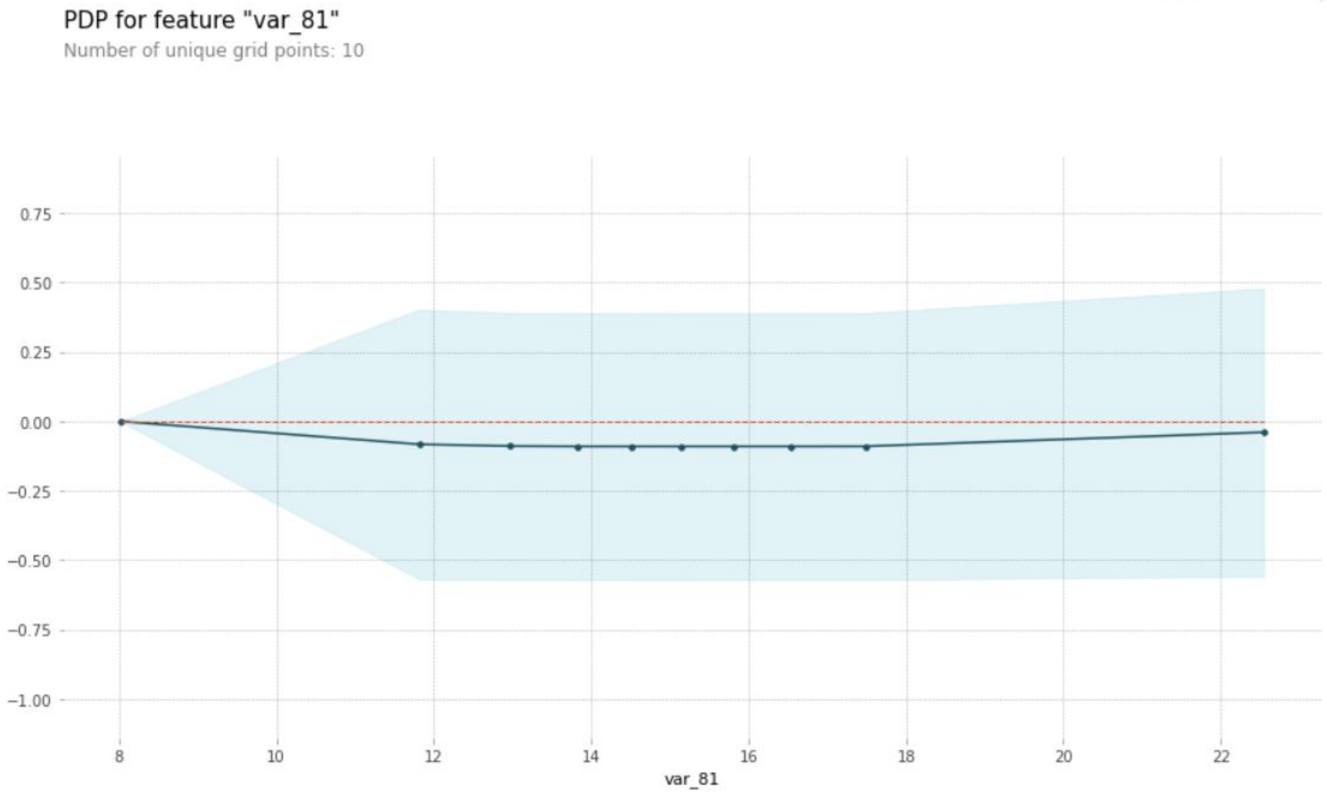


0.0003 ± 0.0006	var_130
0.0003 ± 0.0002	var_166
0.0003 ± 0.0004	var_128
0.0003 ± 0.0005	var_147
0.0003 ± 0.0006	var_67
0.0003 ± 0.0001	var_187
0.0003 ± 0.0002	var_62
0.0003 ± 0.0003	var_15
0.0003 ± 0.0004	var_191
0.0003 ± 0.0004	var_87
0.0003 ± 0.0004	var_111
0.0003 ± 0.0002	var_182
0.0003 ± 0.0001	var_35
0.0003 ± 0.0003	var_155
0.0002 ± 0.0004	var_49
0.0002 ± 0.0005	var_196
0.0002 ± 0.0004	var_172
0.0002 ± 0.0001	var_153
0.0002 ± 0.0004	var_141
0.0002 ± 0.0003	var_132
0.0002 ± 0.0005	var_75
0.0002 ± 0.0004	var_140
0.0002 ± 0.0002	var_176
0.0002 ± 0.0003	var_4
0.0002 ± 0.0001	var_61
0.0002 ± 0.0002	var_92
0.0002 ± 0.0003	var_113
0.0002 ± 0.0002	var_77
0.0002 ± 0.0001	var_3
0.0002 ± 0.0005	var_33
0.0002 ± 0.0002	var_112
0.0002 ± 0.0003	var_55



0.0000 ± 0.0002	var_60
0.0000 ± 0.0001	var_97
0.0000 ± 0.0001	var_30
0.0000 ± 0.0000	var_43
0.0000 ± 0.0002	var_91
0.0000 ± 0.0001	var_161
0.0000 ± 0.0001	var_185
0.0000 ± 0.0002	var_189
0.0000 ± 0.0001	var_84
0.0000 ± 0.0002	var_16
0 ± 0.0000	var_68
-0.0000 ± 0.0002	var_39
-0.0000 ± 0.0001	var_160
-0.0000 ± 0.0001	var_27
-0.0000 ± 0.0001	var_100
-0.0000 ± 0.0003	var_193
-0.0000 ± 0.0000	var_103
-0.0000 ± 0.0000	var_96
-0.0000 ± 0.0003	var_32
-0.0000 ± 0.0002	var_158
-0.0000 ± 0.0001	var_136
-0.0000 ± 0.0000	var_108
-0.0000 ± 0.0001	var_37
-0.0000 ± 0.0001	var_69
-0.0000 ± 0.0000	var_10
-0.0000 ± 0.0002	var_101
-0.0001 ± 0.0003	var_120
-0.0001 ± 0.0003	var_65
-0.0001 ± 0.0002	var_98
-0.0001 ± 0.0007	var_86
-0.0001 ± 0.0003	var_51
-0.0001 ± 0.0000	var_41
-0.0001 ± 0.0001	var_117
-0.0001 ± 0.0001	var_17

PDP Plots



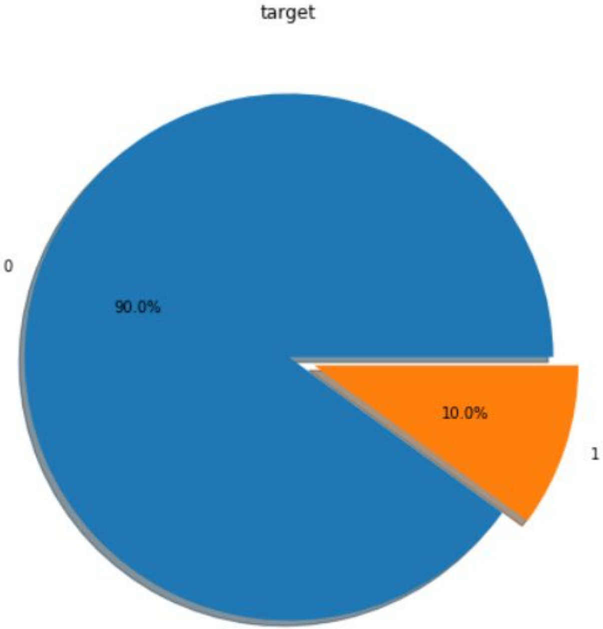
PDP for feature "var_81"

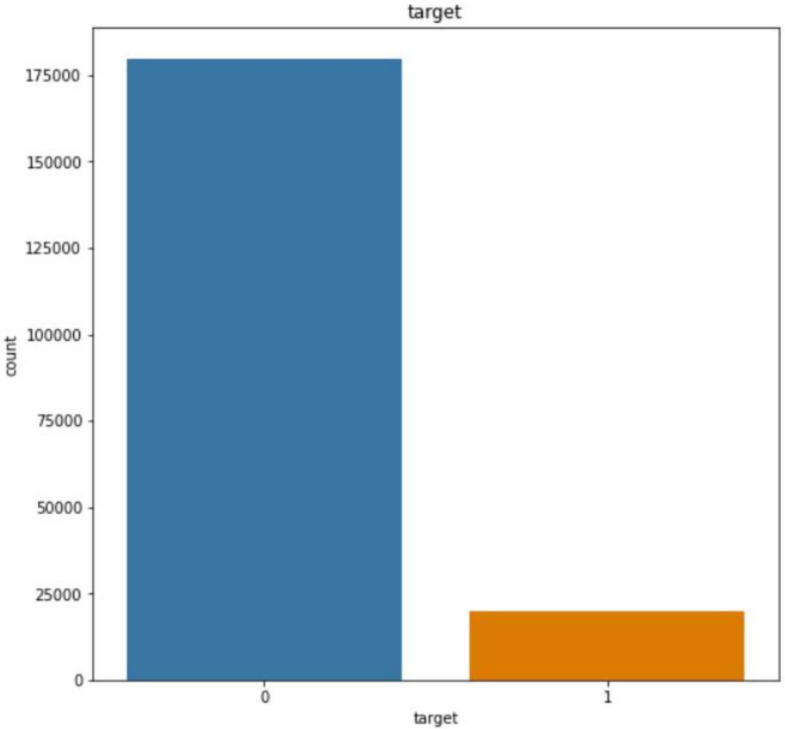
Number of unique grid points: 10

Data Pre-Processing

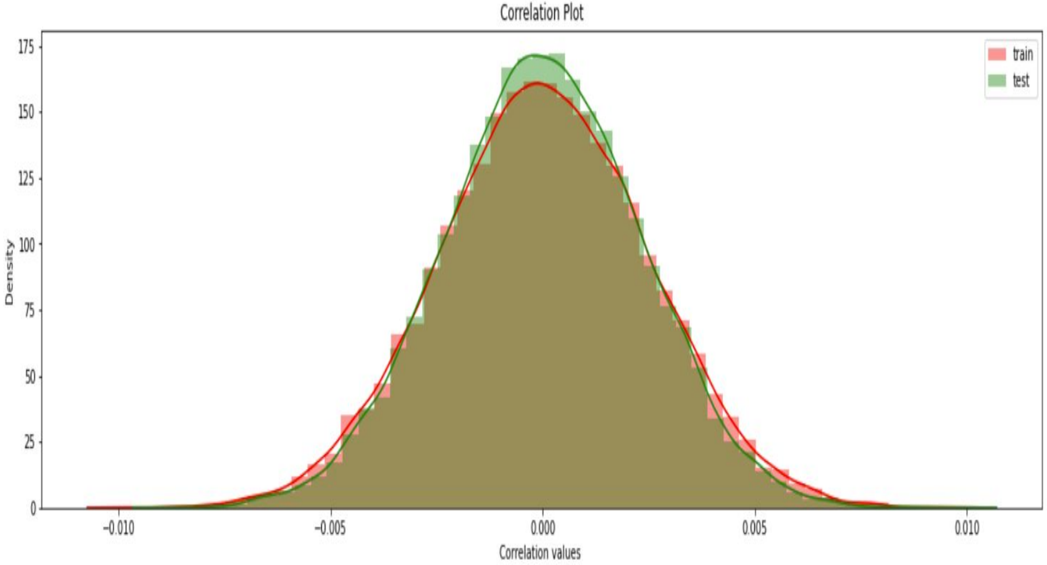
- There were no missing values in the dataset.

◎ The dataset is imbalanced. 90% of the data has 0 target value.



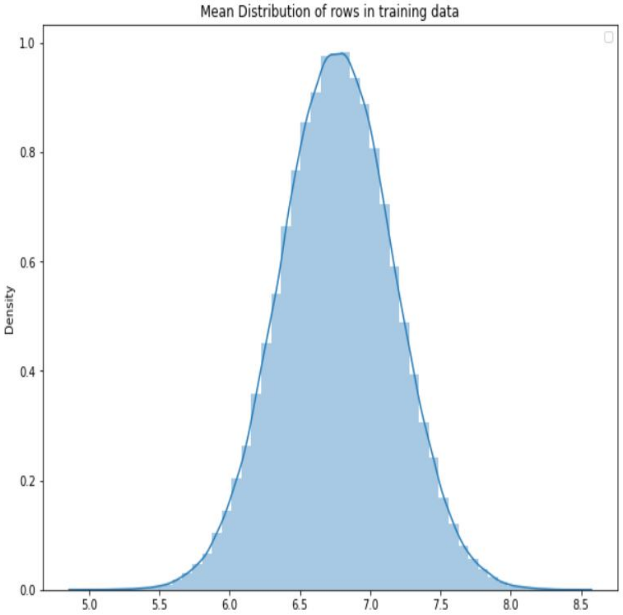


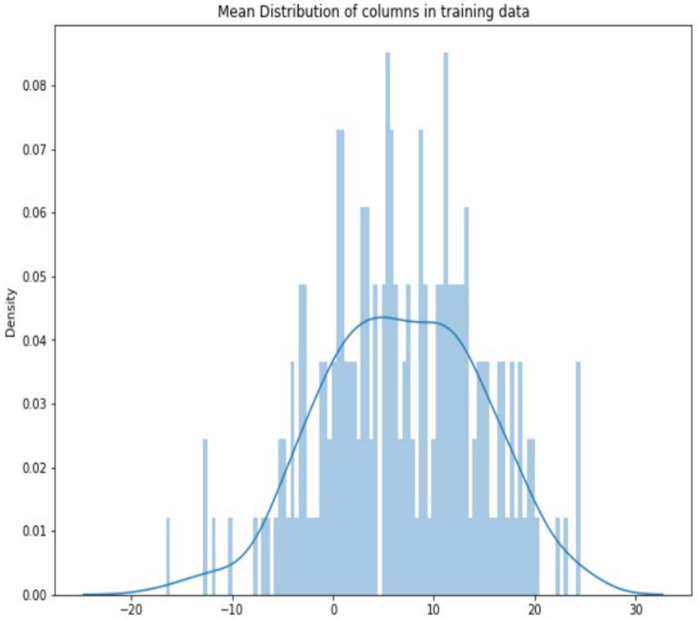
- There was no linear correlations either in training data or in the test data.



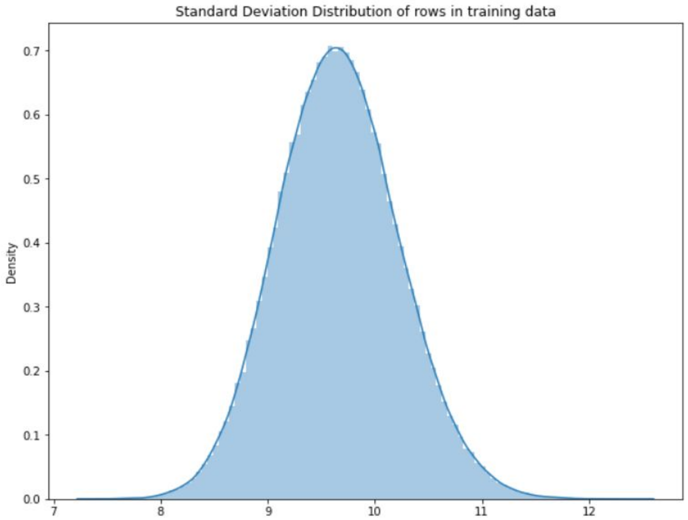
Distribution Plots

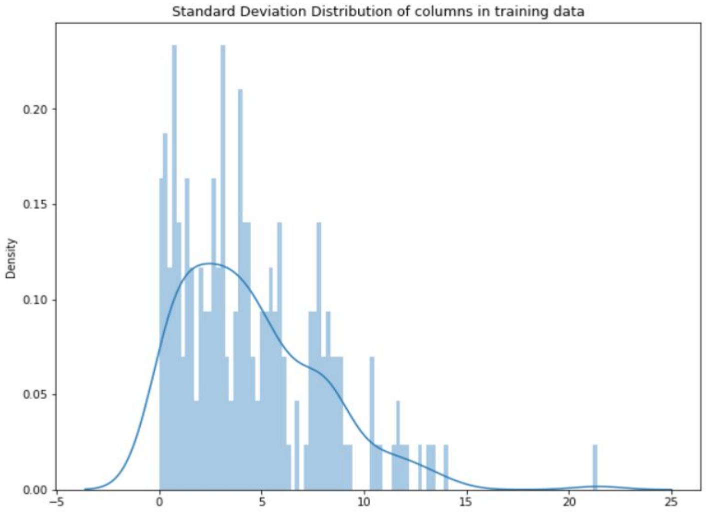
Mean Distribution Plots





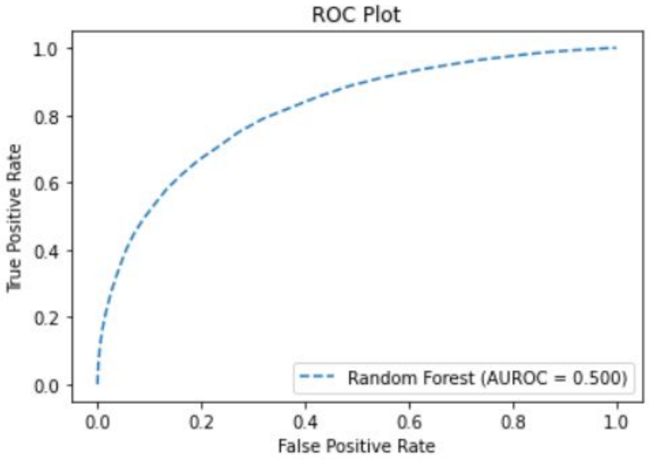
Standard Deviation Distribution Plots

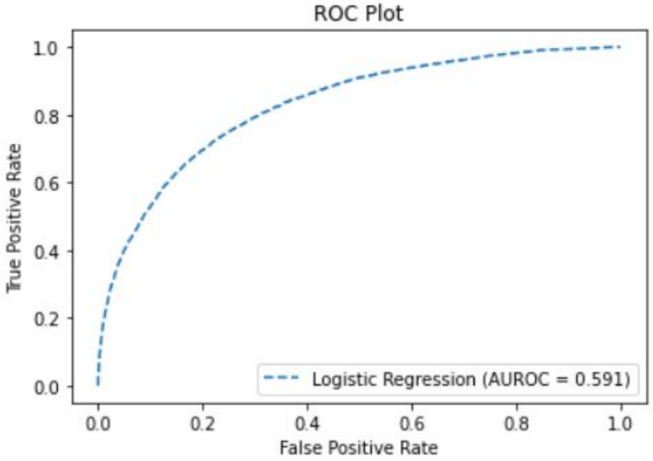




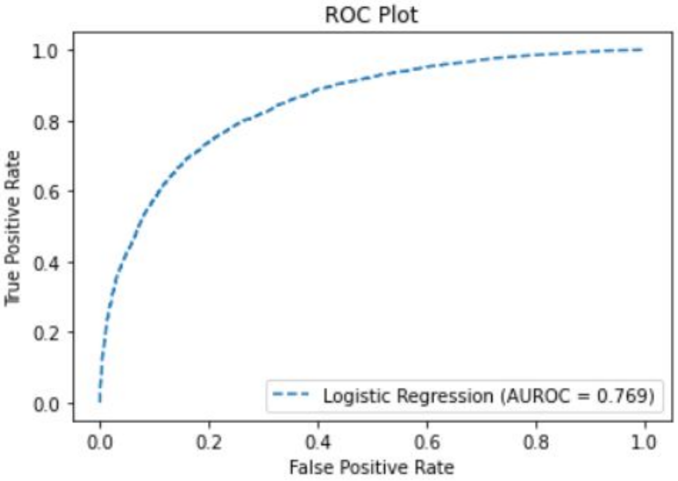
Machine Learning models

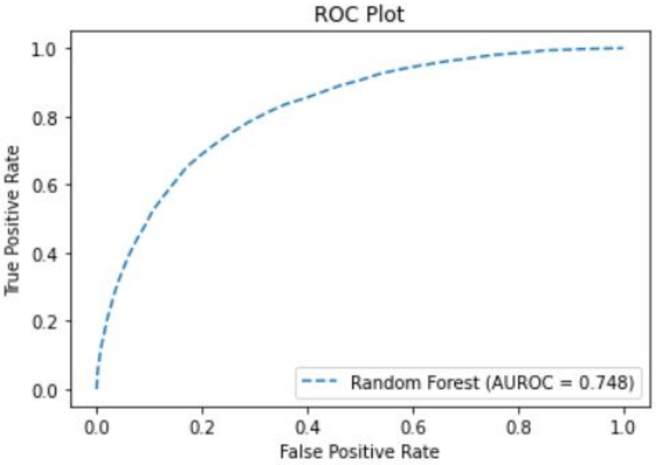
- Logistic Regression
- Random Forest
- Support Vector Machine
- Gaussian Naïve Bayes
- Artificial Neural Networks


ROC Plot
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUROC = 0.500)


ROC Plot
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUROC = 0.591)

Under Sampling


ROC Plot
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUROC = 0.769)

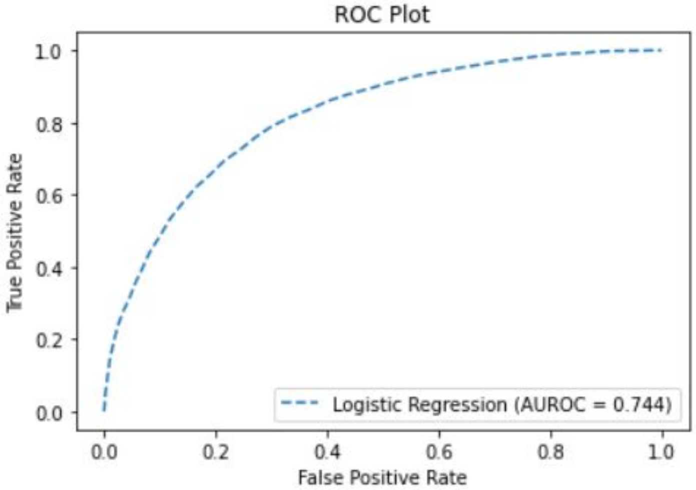

ROC Plot
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUROC = 0.748)

Limitations of Under Sampling

- In Under sampling the data samples of the majority class are deleted randomly, the data can be useful to create a robust decision boundary.

- There is no way that one can preserve information rich examples from the majority class

Over Sampling


ROC Plot
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Logistic Regression (AUROC = 0.744)

Limitations of Over Sampling

- The main drawback of over sampling is that it makes exact copies of the existing examples which makes overfitting more likely.

Data Augmentation

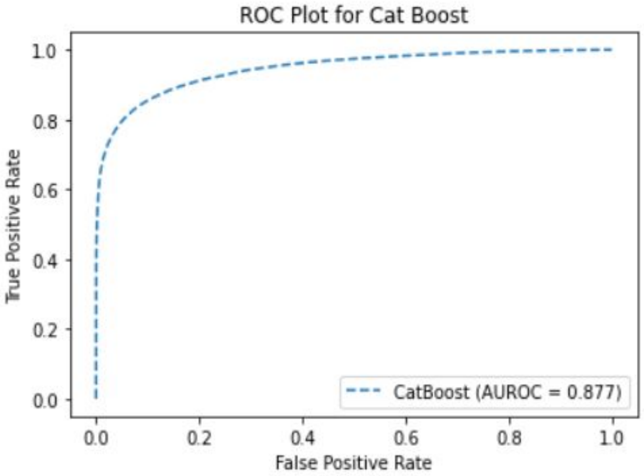
SMOTE

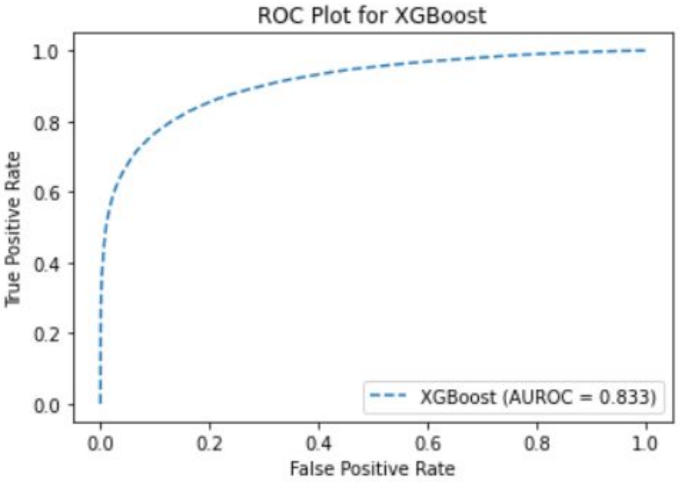
- Synthetic Minority Oversampling Technique is a Data augmentation technique, which synthesizes new examples for the minority class in the imbalanced dataset.

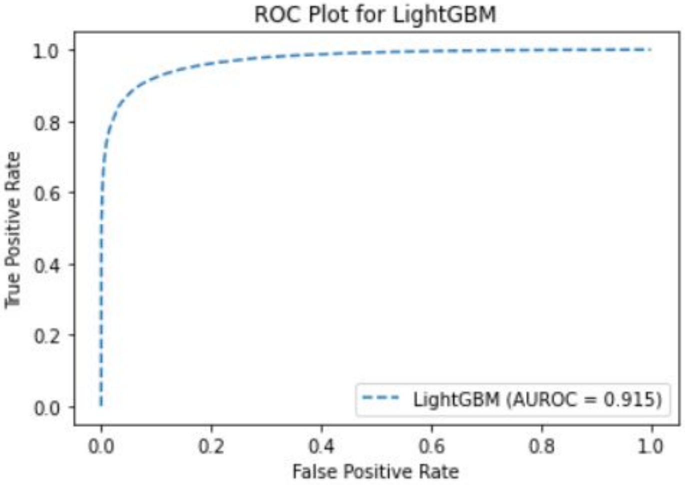
Ensemble methods

Boosting

- CatBoost
- XGBoost
- LightGBM


ROC Plot for Cat Boost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CatBoost (AUROC = 0.877)


ROC Plot for XGBoost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGBoost (AUROC = 0.833)


ROC Plot for LightGBM
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LightGBM (AUROC = 0.915)

Deployment

https://transaction-prediction.herokuapp.com/

Business Use-Case

- Customer Transaction Prediction model will help the Banks and financial institutions to build their business and marketing strategies.

- It will help them to give them personalized service which will enhance in user experience, which will contribute in the growth of the business.

Thank You