T5 Data Science Bootcamp

Bank Unity

Predicting Churn Credit Card Customers

By Nouf Alotaibi

About the Company

Bank Unity is a financial institution licensed to receive deposits and make loans. It also provides financial services such as wealth management, currency exchange, and safe deposit boxes.

Problem

A manager at *Bank Unity* is disturbed by more and more customers leaving their credit card services. They would appreciate it if one could predict who is going to get churn, so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction.

Dataset

Source

Credit Card Customers Dataset

It was uploaded to Kaggle.com.

Records

10,127 Customer Records

Features

21 Features

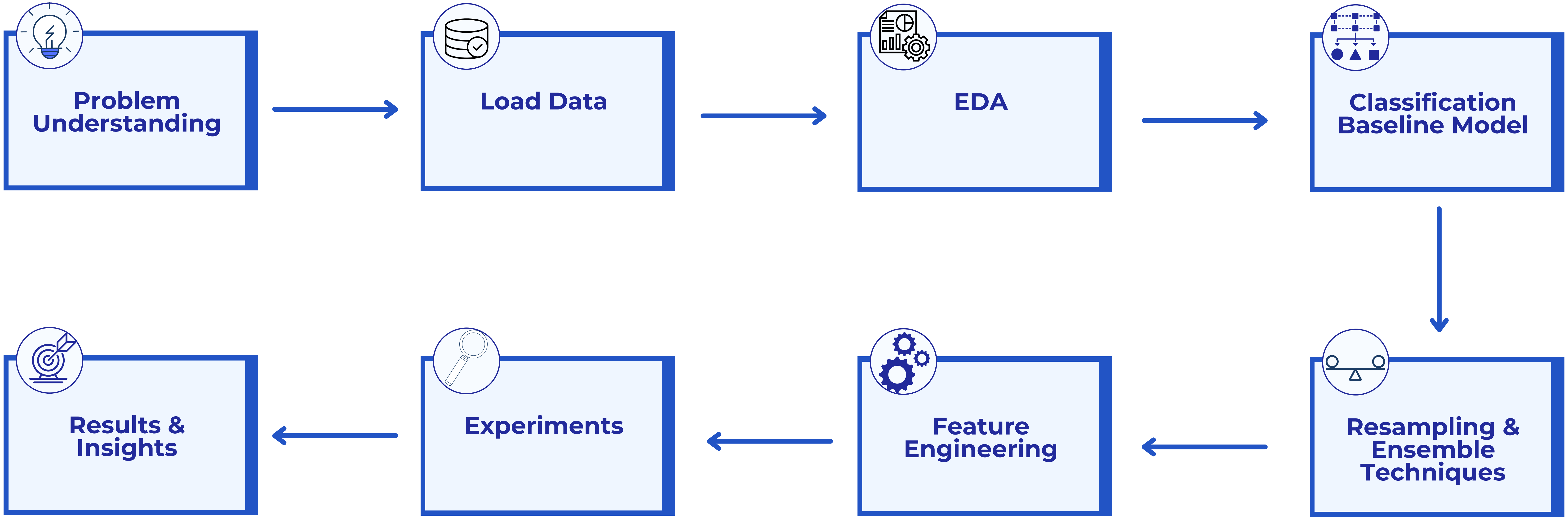
Such as customers' age, salary, marital status, credit card limit, credit card category, etc.

Target

Attrition_Flag

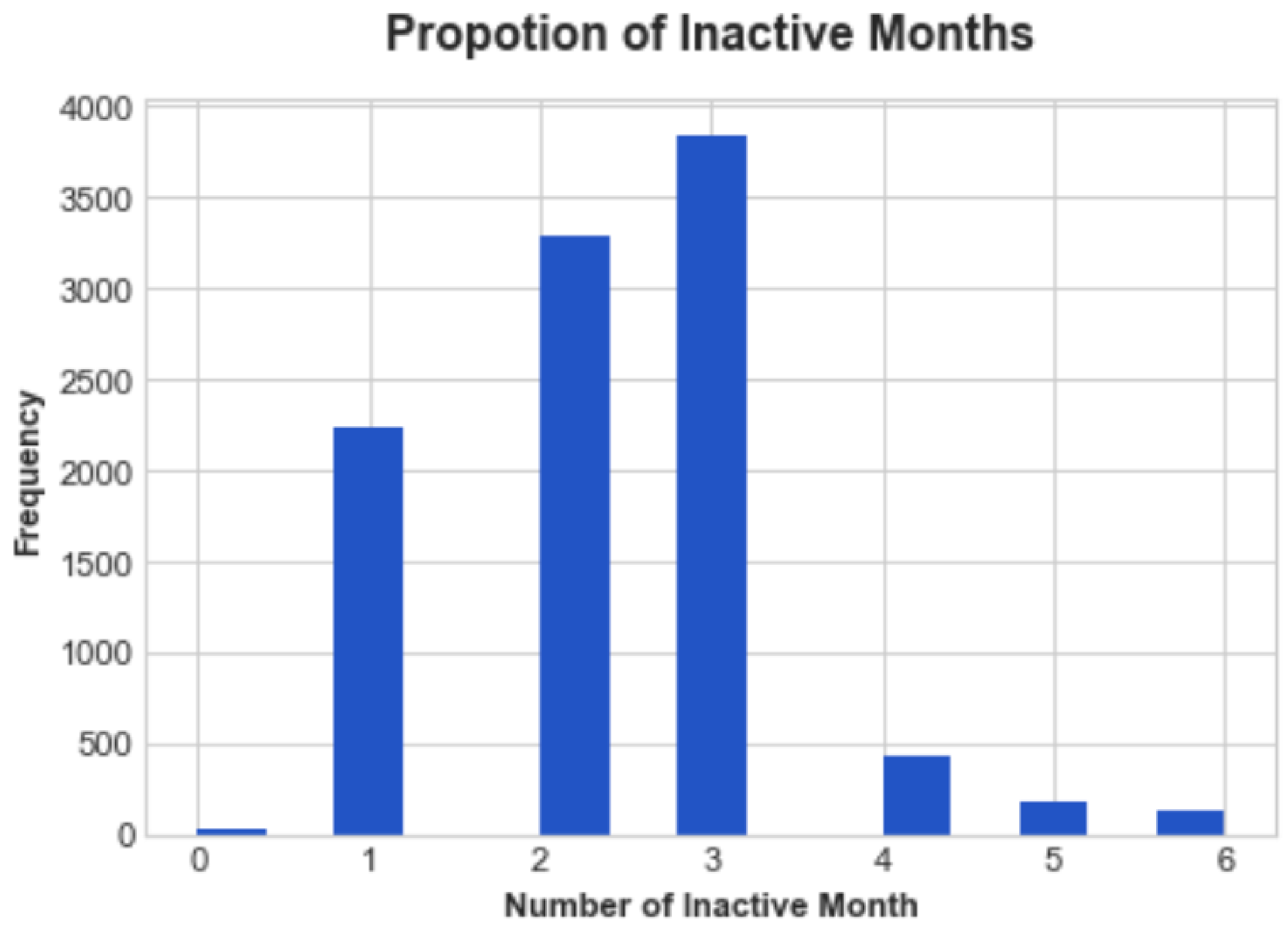
Positive if a customer is attrited, **Negative** if not.

Methodology



Exploratory Data Analysis

The highest number of months on book is about 36, we cannot assume normality of the feature.



The highest number of inactive months is 3, followed by 2 and 1.

Baseline Model

Actual Values \ Predicted Values	Existing	Attrited
Existing	1222	45
Attrited	135	116

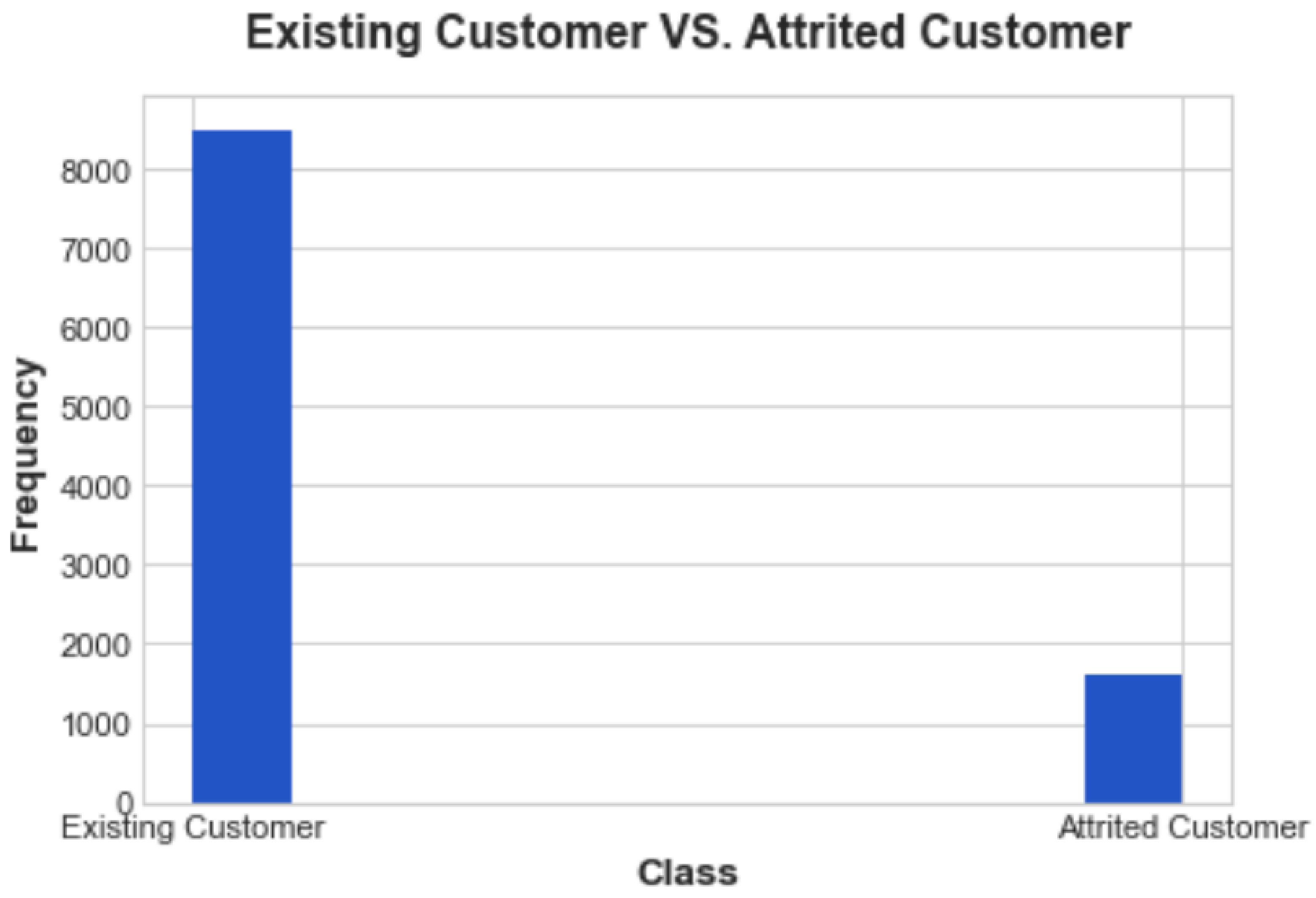
Training Scores

- Accuracy = 0.889
- Precision = 0.728
- Recall = 0.494
- F-1 = 0.588
- F-beta of 2 = 0.528

Validation Scores

- Accuracy = 0.881
- Precision = 0.72
- Recall = 0.462
- F-1 = 0.563
- F-beta of 2 = 0.498

Data Imbalance



Number of Observations

10127

Number of Attrited Customer

1627

Number of Existing Customer

8500

Event Rate

16.1%

Data Imbalance

Approaches for Imbalanced Data

Resampling Techniques

- Random Under-Sampling
- Random Over-Sampling
- SMOTE

Bagging Based techniques

- Random Forest

Boosting-Based techniques

- Ada Boost

Feature Engineering

Convert Categorical Columns to Dummy Variables

Resulted in decreasing the models' scores.

Factorize Categorical Columns

Resulted in increasing the models' scores.

Scale Numerical Columns

Using StandardScaler() on numerical columns.

Experiments

Classifier		Recall	Precision	F-1	F-2
Logistic Regression	train	0.813	0.82	0.817	0.815
	val	0.776	0.450	0.57	0.678
Decision Tree	train	1.0	1.0	1.0	1.0
	val	0.84	0.70	0.768	0.810
Random Forest	train	1.0	1.0	1.0	1.0
	val	0.90	0.83	0.865	0.886
Ada	train	0.967	0.959	0.963	0.965
	val	0.912	0.760	0.829	0.877

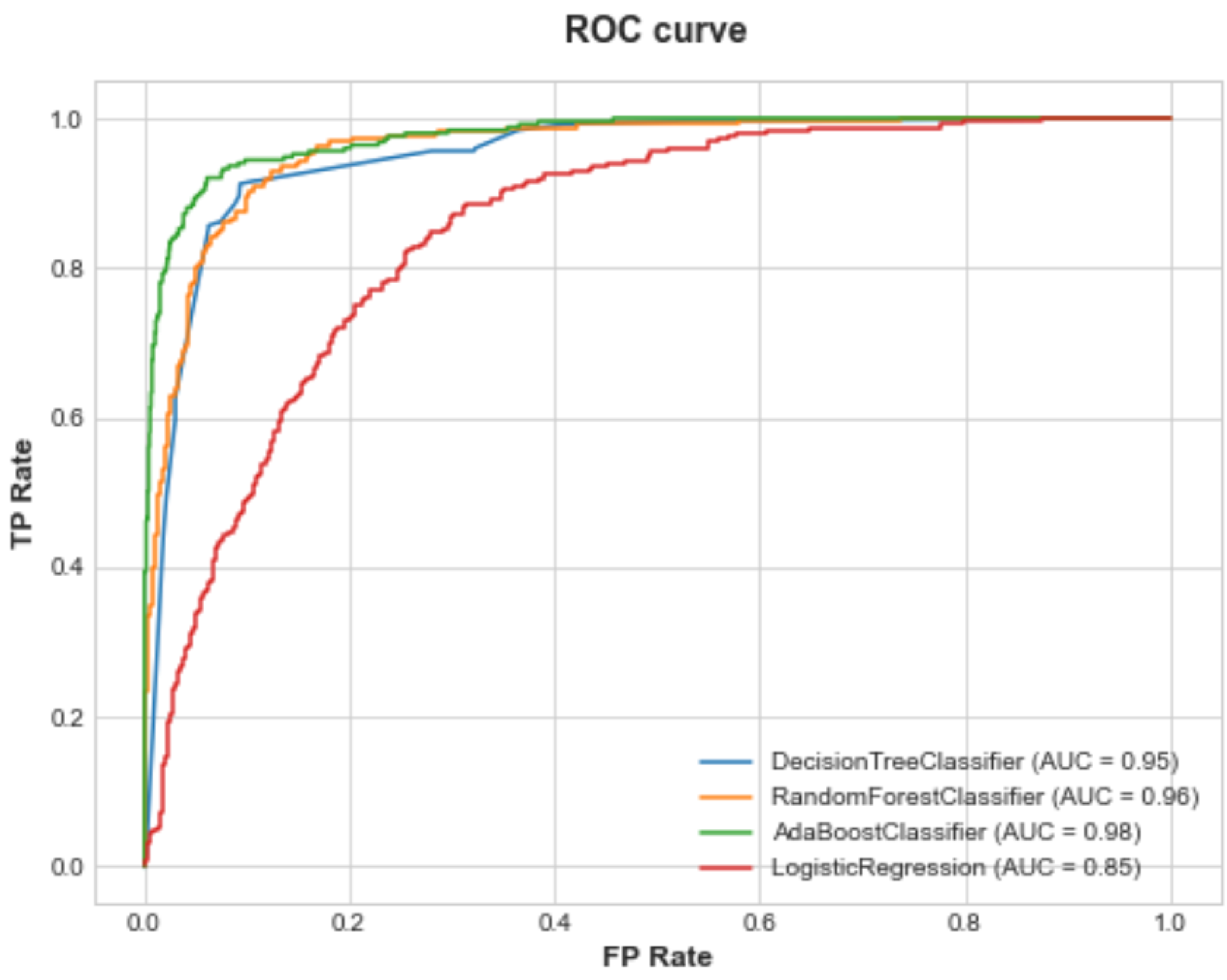
Split Data

%75 train,
%15 validation
%10 test

Resampling Technique

SMOTE

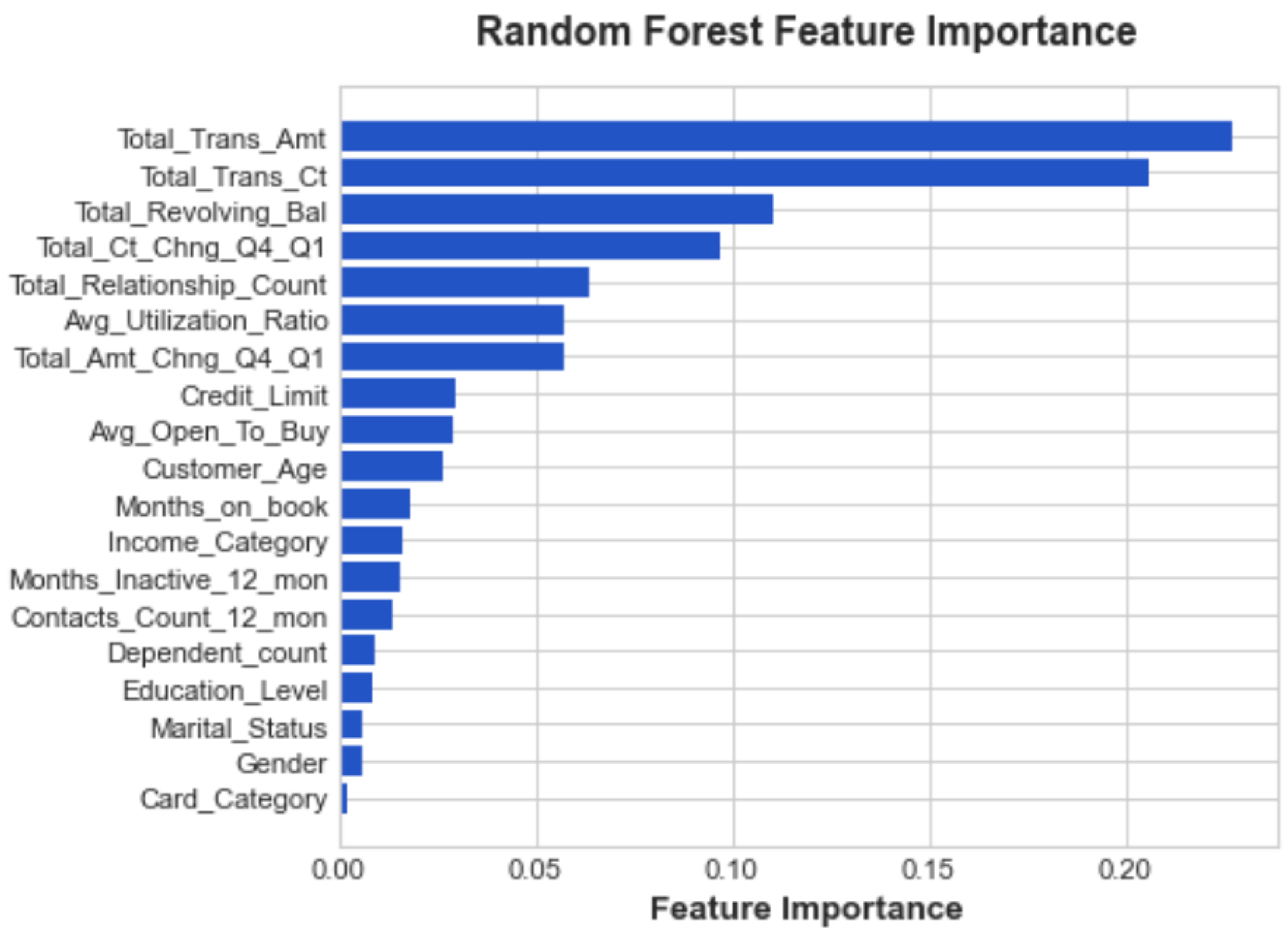
Experiments



Ada Boost Classifier has best results with **98%** AUC, followed by Random Forest and Decision Tree Classifiers with **96%**, **95%** AUC respectively.

Feature Importance

For Random Forest



Low Feature Importance

Gender

Card_Category

Marital_Status

Were removed.

Validation Scores

```
Accuracy = 0.956
Precision =  0.841
Recall = 0.904
F-1 = 0.871
F-beta of 2 = 0.891
```

Conclusion

1. **Classifier Performance Metrics of interest**
   Recall, Precision, F-1, and F-beta
2. **SMOTE for Handling Imbalance Data**
3. **AdaBoost is the Best Classifier**
   With **precision = 0.88, recall = 0.82, F-1 = 0.85, F-2 = 0.85** and it is consistent with the **AUC** metric **0.98** in the ROC.

Future Work

1. Try solving the imbalance data with **Anomaly Detection Algorithm**
2. Work on tuning the classifiers more, and try other classifiers.
3. Investigate more on why are customers churning

Thank You

Any Questions?