Bank Churn Predictor:

The purpose of the project is to develop a binary classifier using Logistic regression, Random Forest or deep learning model to create a predictive tool that estimates the likelihood of bank customers attrition trends.

Machine learning (ML) Project

The Team

Bethelhem Arefayne

Maksym Andreiko

Umeadi Dungor

Problem Statment

A bank managers are worried why the customers are leaving their bank and they would like to identify the patterns.

Business goals are:

- Increase customer satisfaction
- Decrease churn rate

Project Overview

- Develop model that would help a bank predict the customers who are going to get churned.
- Based on the data the problem is identified as classification problem.

Data Overview

Dataset includes 10127 rows and 23 attributes with 7 categorical and 16 numerical variables. The data has no missing values and duplicates.

Dataset title: Credit Card Customers
Data source: Kaggle
Link:
https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

Steps

Tools used: Python is the main technology that was used for this project. Some of the libraries that were imported were Pandas - used for data processing and plotly and matplotlib for data visualizations. Tableau was also used for additonal data visualizations.

Step 1

Data Preparation

- Uploaded csv file
- Imported libraries
- Cleaned data

Step 2

EDA

- Customer demographics
- Data distribution

Step 3

Data Process

- Classify problem
- select feature
- Scale data
- Train data
- Balance data

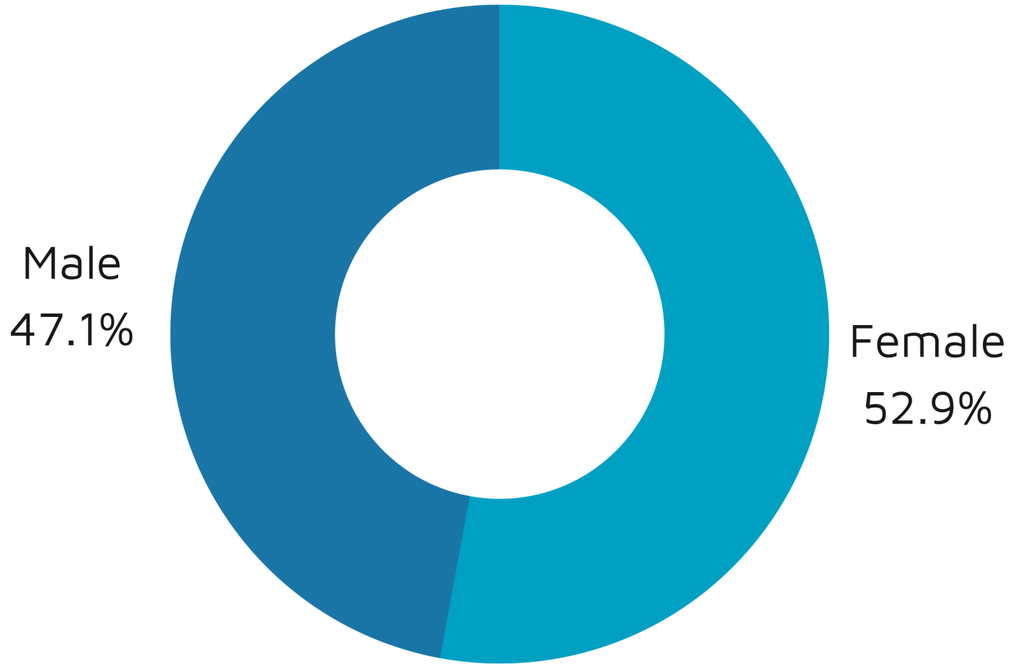
Step 4

ML Model

- Select model/s
- Make predictions
- Evaluate
  - Accuracy score
  - Confusion matrix

Gender Rate

Female customers are slightly higher than the male customers.

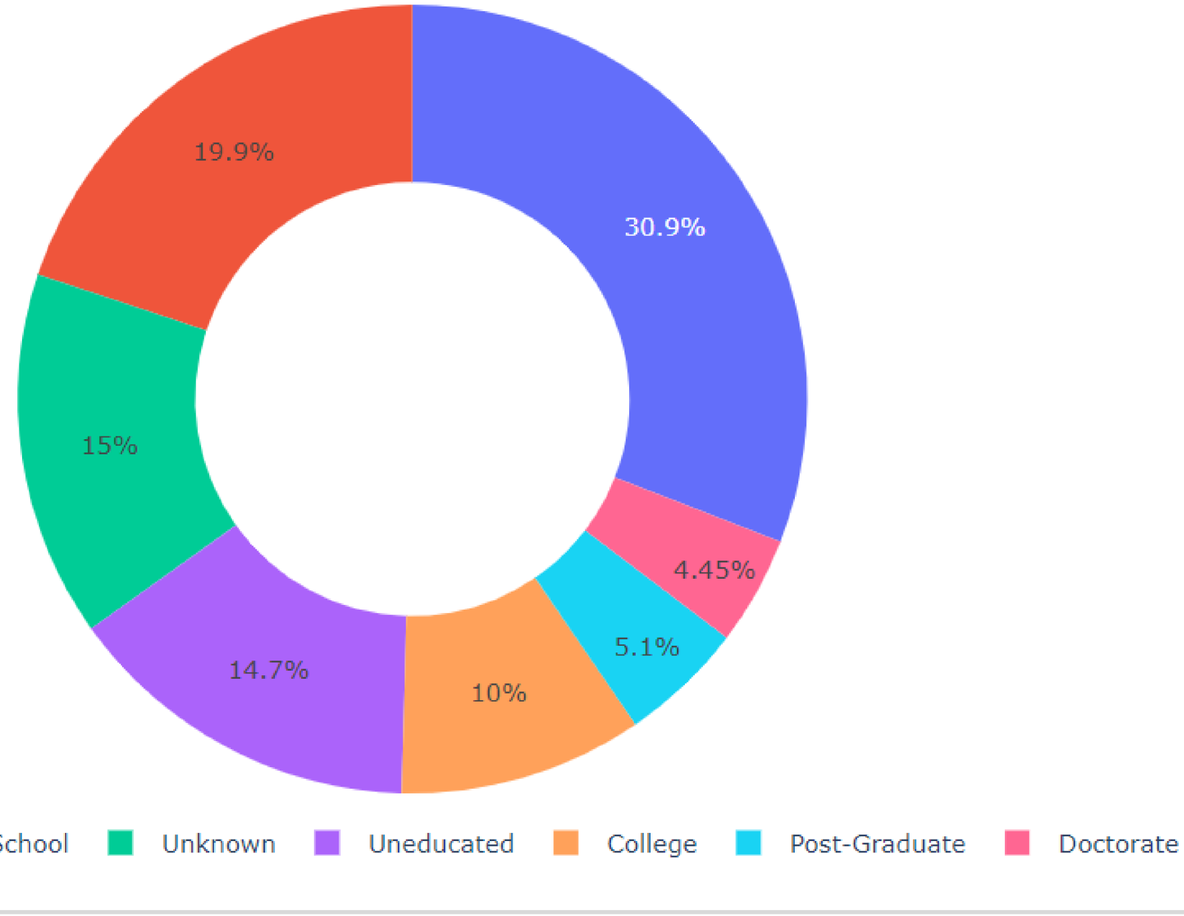


Education Level

70% of customers are educated

15% of customers are uneducated

15% of customers education level is unknown

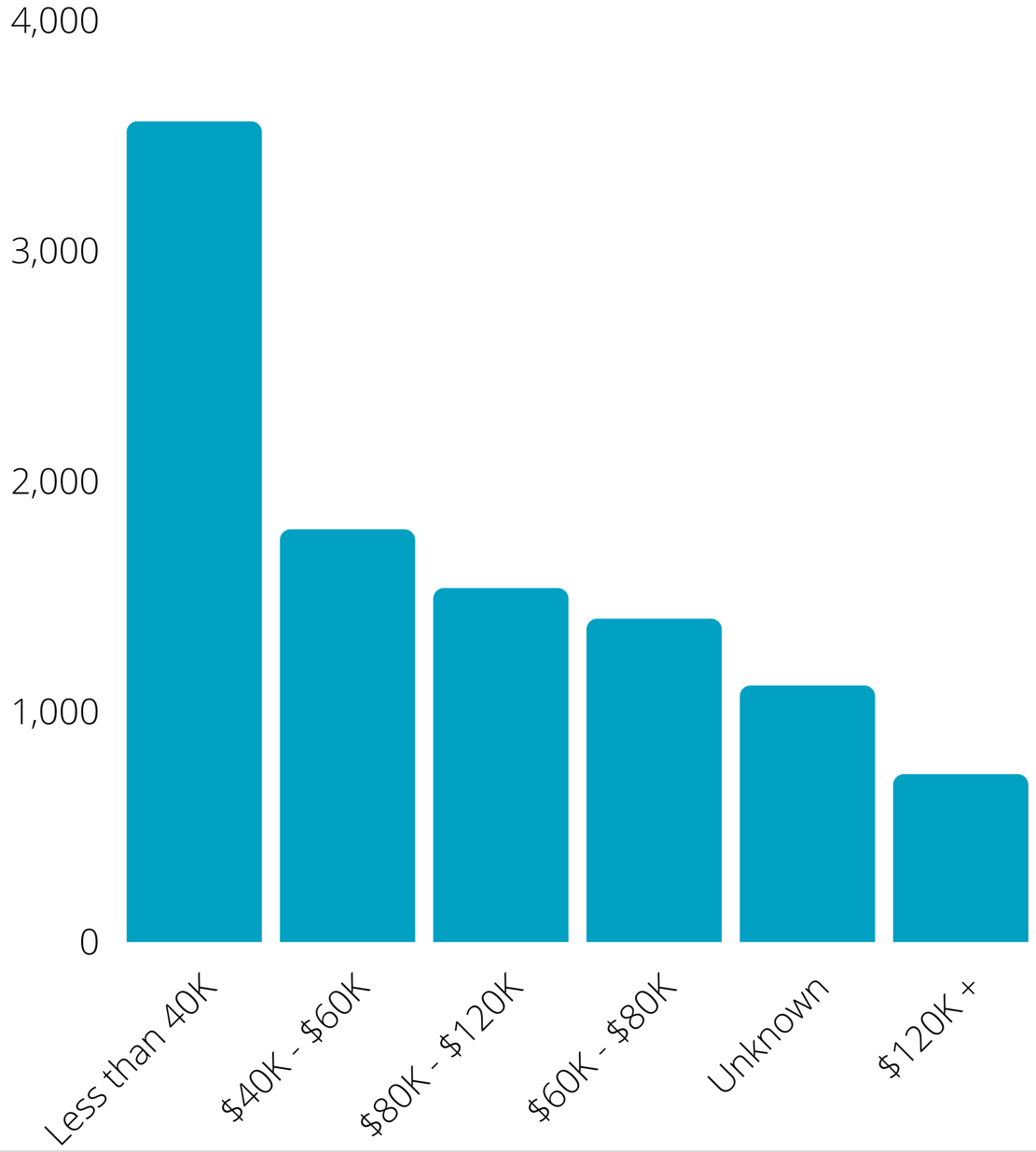


Income Category

35% of customers have less than 40k income

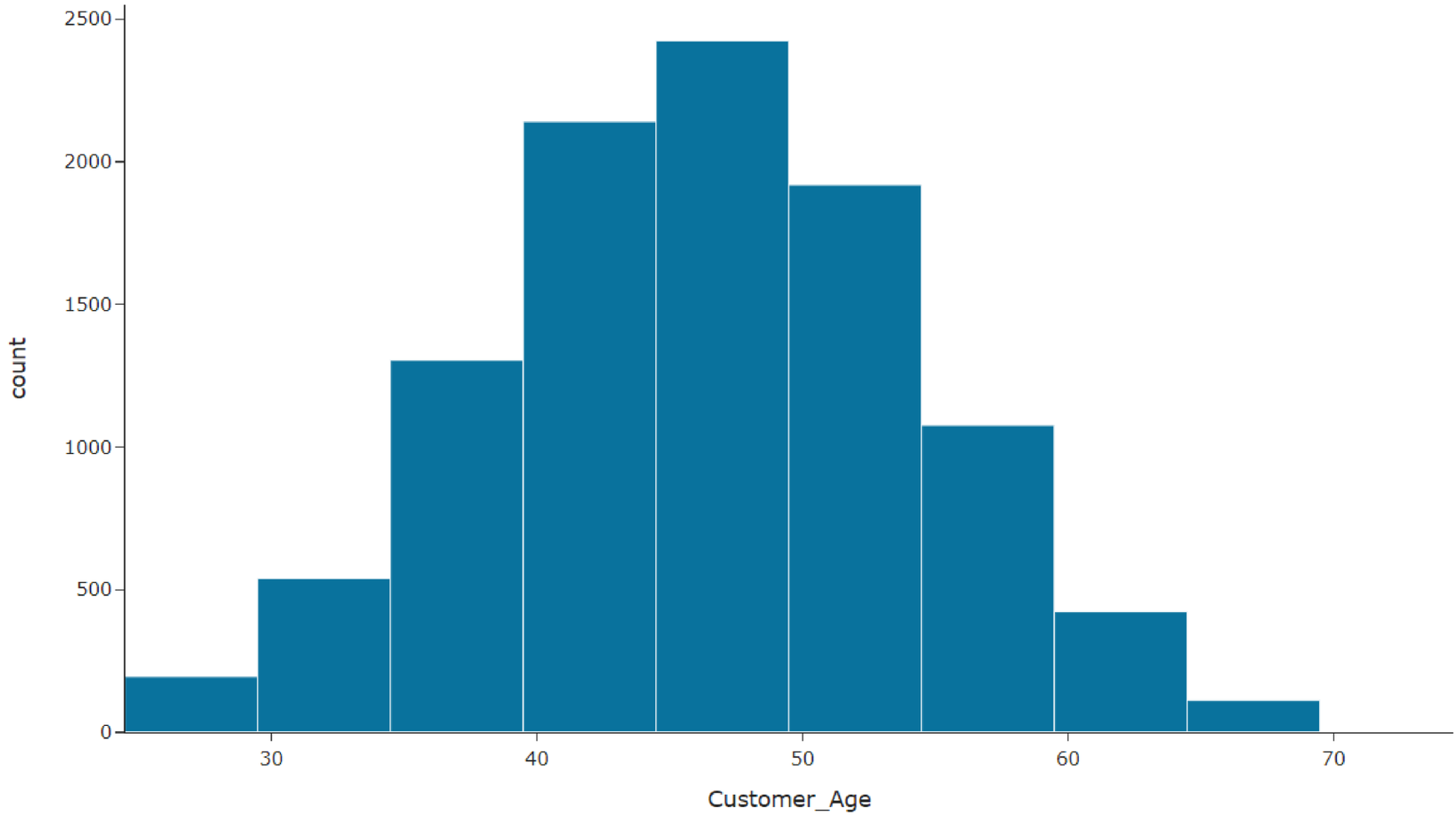
11% of customers unknown income

7% of customers have more than 120k income



Age Distibution

Average age of the bank customers is 46.

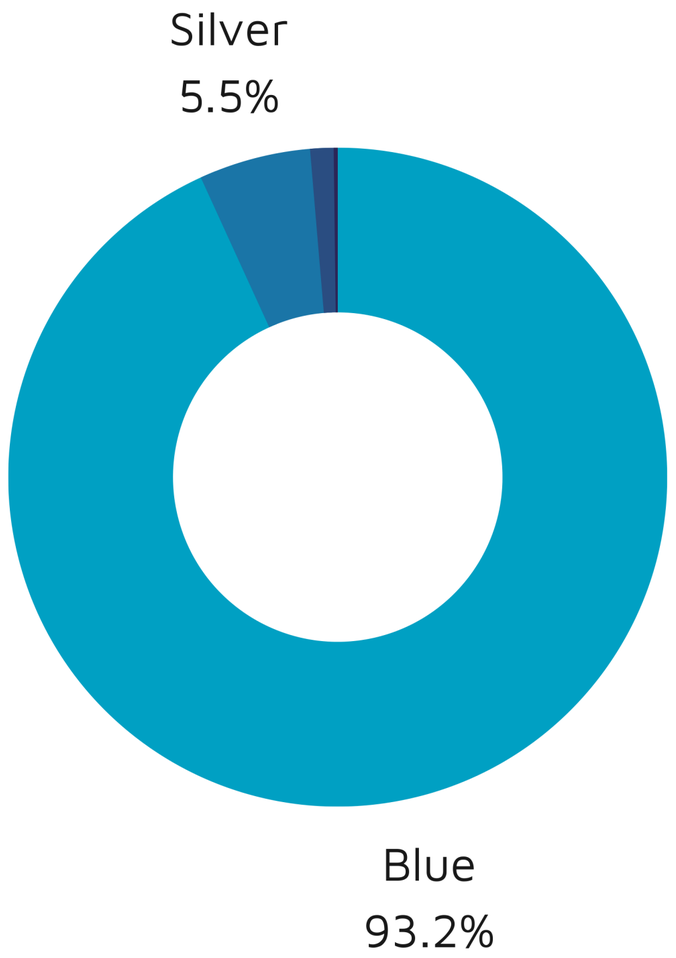


Card Category

93% of customers have blue cards

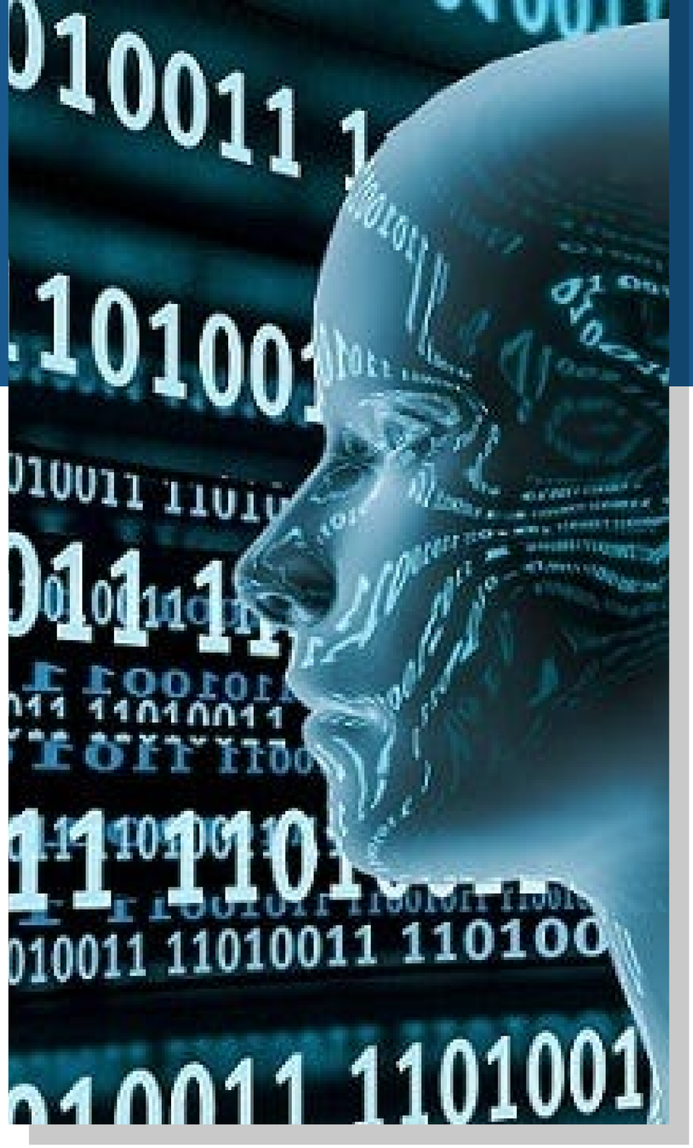
5% of customers have silver cards

2% of customers have gold and platinum cards



Model

After data exploration, it's concluded that the churn prediction is a classification problem.



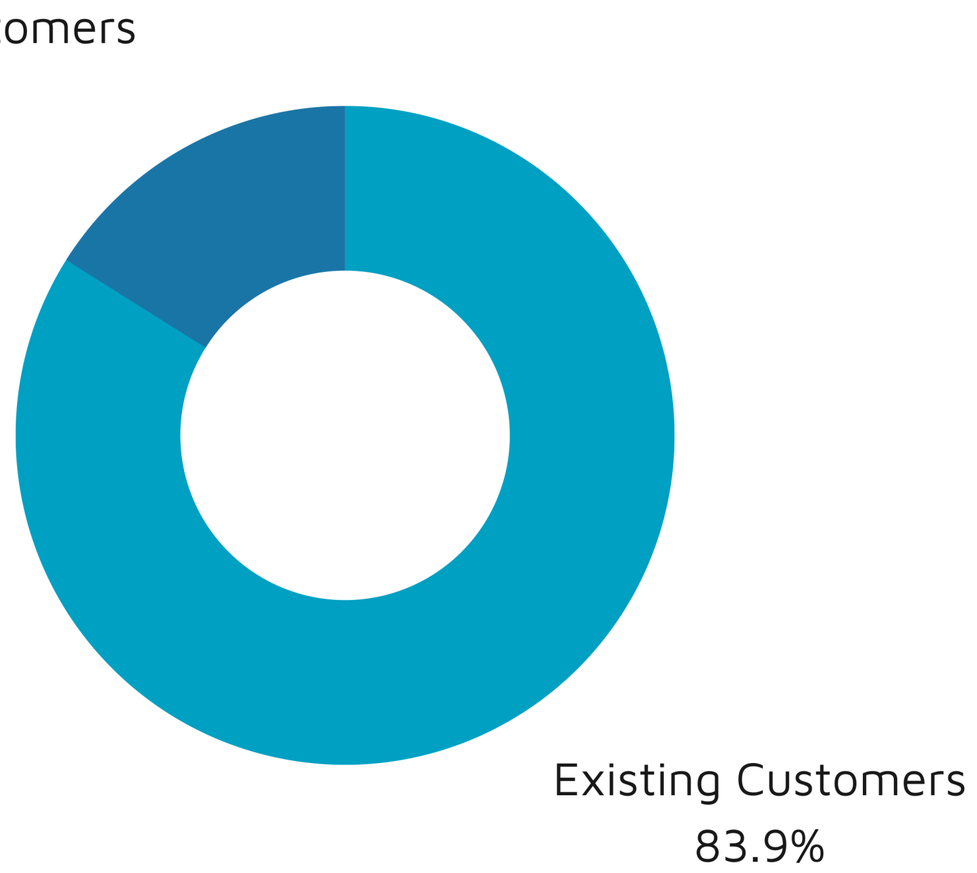
Selected Models

- Logistic Regression
- Random Forest
- Neural Network

Target Variable

Attrition Flag

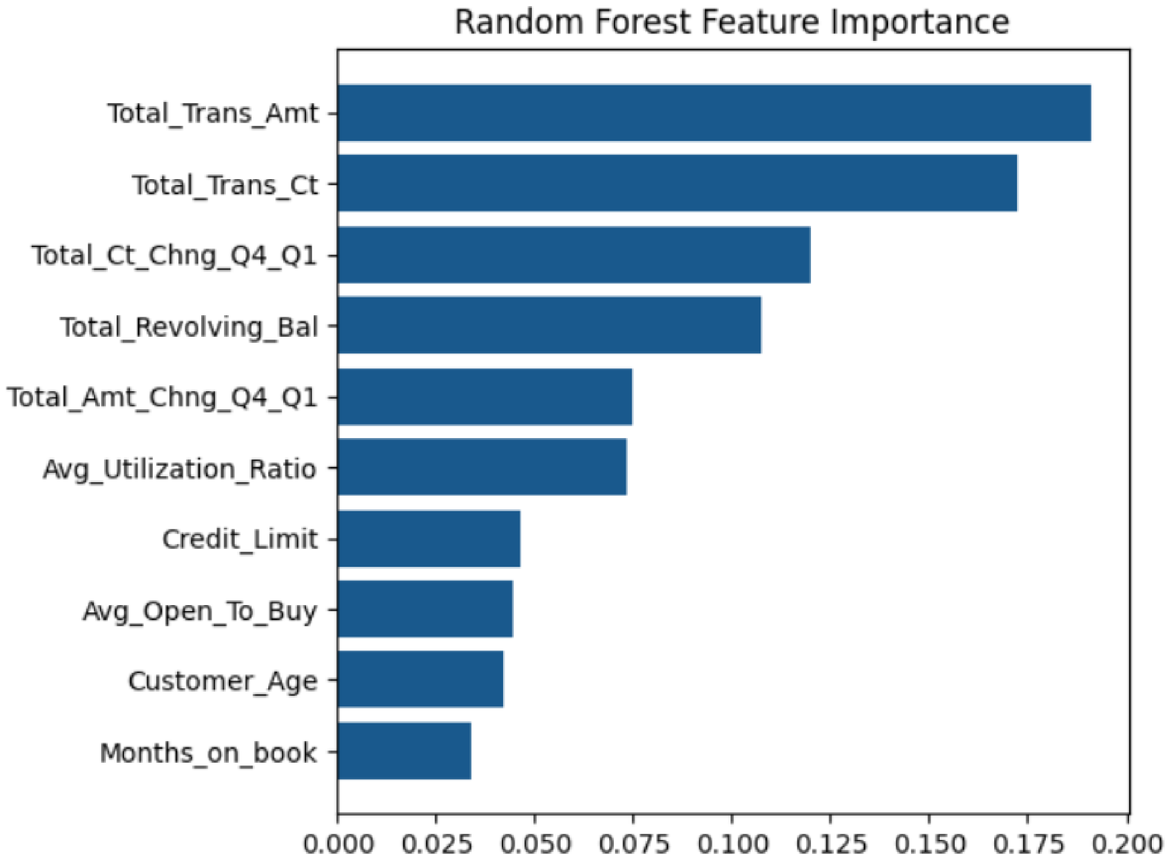
- Out of the 10,127 rows, 1627 customers have left the bank.
- This causes an imbalanced data.
- **Issue**: Data proportion makes prediction difficult.
- **Solution**: Balance data using RandomOverSampler.



Random Forest

Feature Importance

19%	Total Transaction Amount
17%	Total Transaction Count
12%	Total Count Change from Q4-Q1



Code adapted from: https://www.kaggle.com/code/andreshg/churn-prediction-0-99-auc-h2o-sklearn-smote#4.-Feature-Selection

Model Comparison

Classification report Before balancing data

	Accuracy	Precision	Recall	f1-score
Logistic Regression Model	88%	70%	49%	57%
Random Forest Model	94%	88%	71%	79%

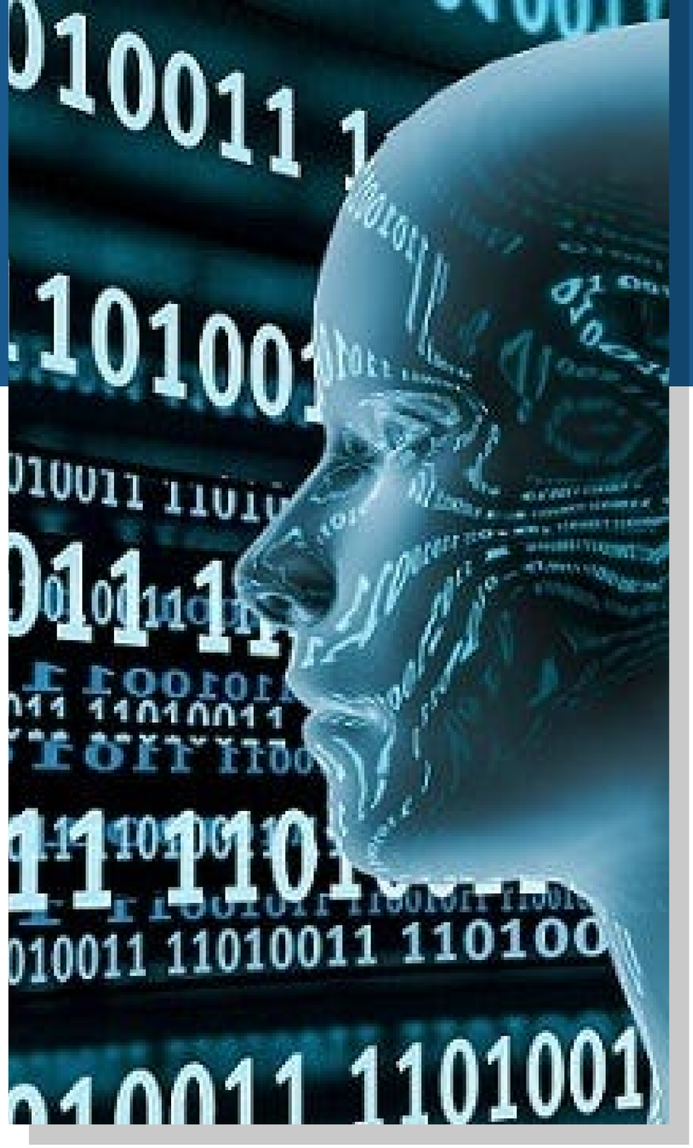
Model Comparison

Classification report After balancing data

	Accuracy	Precision	Recall	f1-score
Logistic Regression Model	83%	82%	84%	83%
Random Forest Model	99%	97%	100%	99%

Results

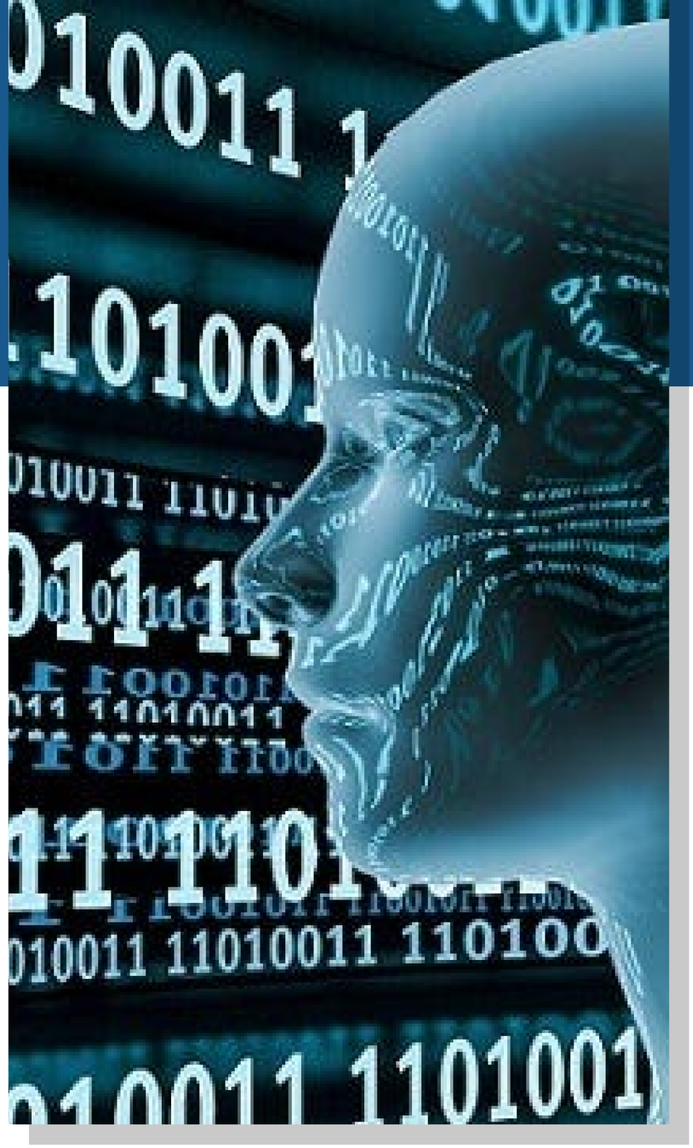
Classification report **After** balancing data



	Accuracy	Loss
Neural Network Model	93%	17%

Conclusion

Based on the results, it is recommended to use Random Forest Model with Resampled Training Data.



This model achieved a higher accuracy score and demonstrated consistent precision, recall, and f1-scores for both classes.
Its ability to accurately identify both existing and attrited customers is crucial for minimizing false positives and effectively mitigating potential churns. The higher recall score for Class 1 also indicates its ability to capture a larger proportion of actual churns, which is vital for identifying the customers who are likely to leave the bank.