# CREDIT CARD CUSTOMER ATTRITION (CHURN) PREDICTION

CAPSTONE PROJECT TWO

DATA SCIENCE CAREER TRACK

SPRINGBOARD

SHAHJAHAN AHMED

OCTOBER 27, 2021

# Problem Statement

Predict the credit card customer attrition rate and find the potential churning customers by analyzing data for a specific period (last twelve month) of time to minimize churn rate by providing better service for company’s growth and financial stability.

# Introduction

❑ Churn Prediction is one of the most popular Big Data use cases in banking sector. It consists of detecting account holders who are likely to close their account or stop using the services.

❑ The ability to predict that a particular credit card holder is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for a Bank.

❑ It helps the stakeholders to make proactive changes to the retention efforts that drive down churn rates.

# Stakeholders

❑ Board of Directors

❑ Chief Executive Officer

❑ Chief Financial Officer

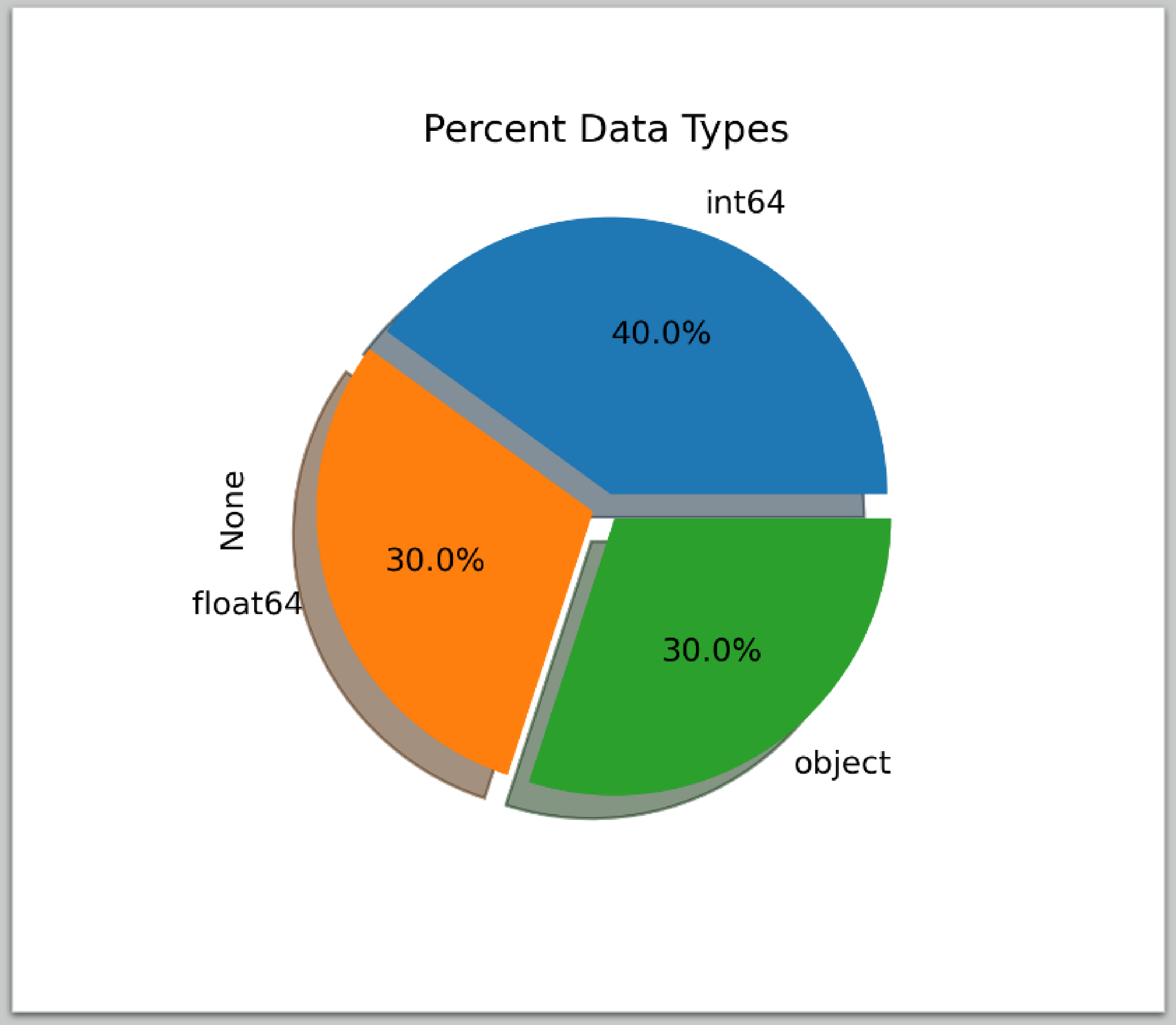
❑ Territorial Heads

❑ Branch Managers

❑ Financial Analyst

# Dataset Summary

- ❑ The pie chart showing the distribution of attributes Based on Data Type.
- ❑ Approximately 70% of the attributes contained numerical data.
- ❑ 30% of the attributes contained categorical data.
- ❑ Among numerical attributes, few columns are discrete.



# Exploratory Data Analysis Steps

- ❑ Distribution of Numerical Attributes
- ❑ Finding Normality
- ❑ Distribution of Categorical Attributes
- ❑ Visualizing Outliers
- ❑ Visualizing Correlations

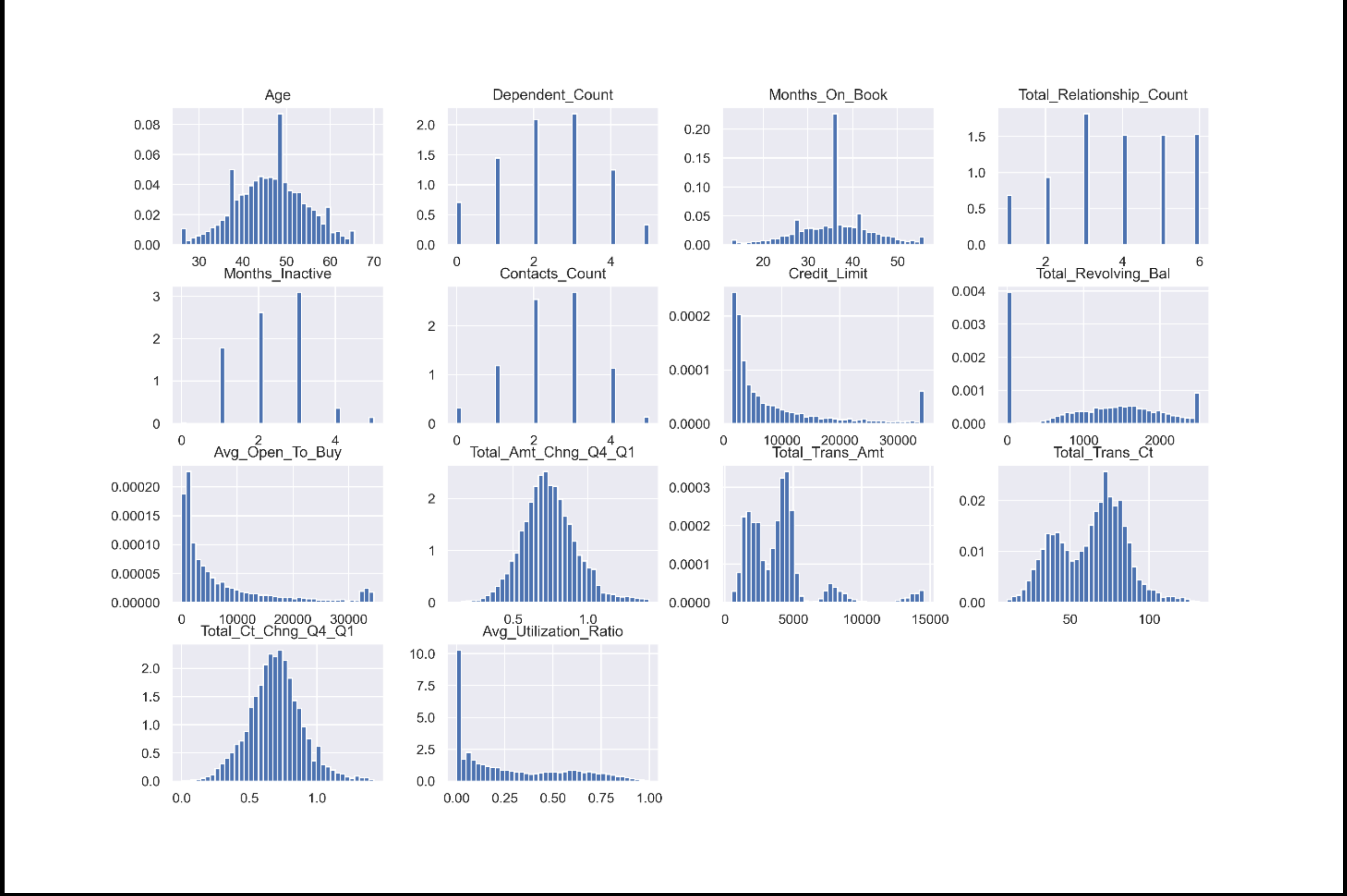
# Numerical Attributes

❑ The histogram is showing the distribution of numerical attributes.

❑ The distribution of few attributes are approximately normal.

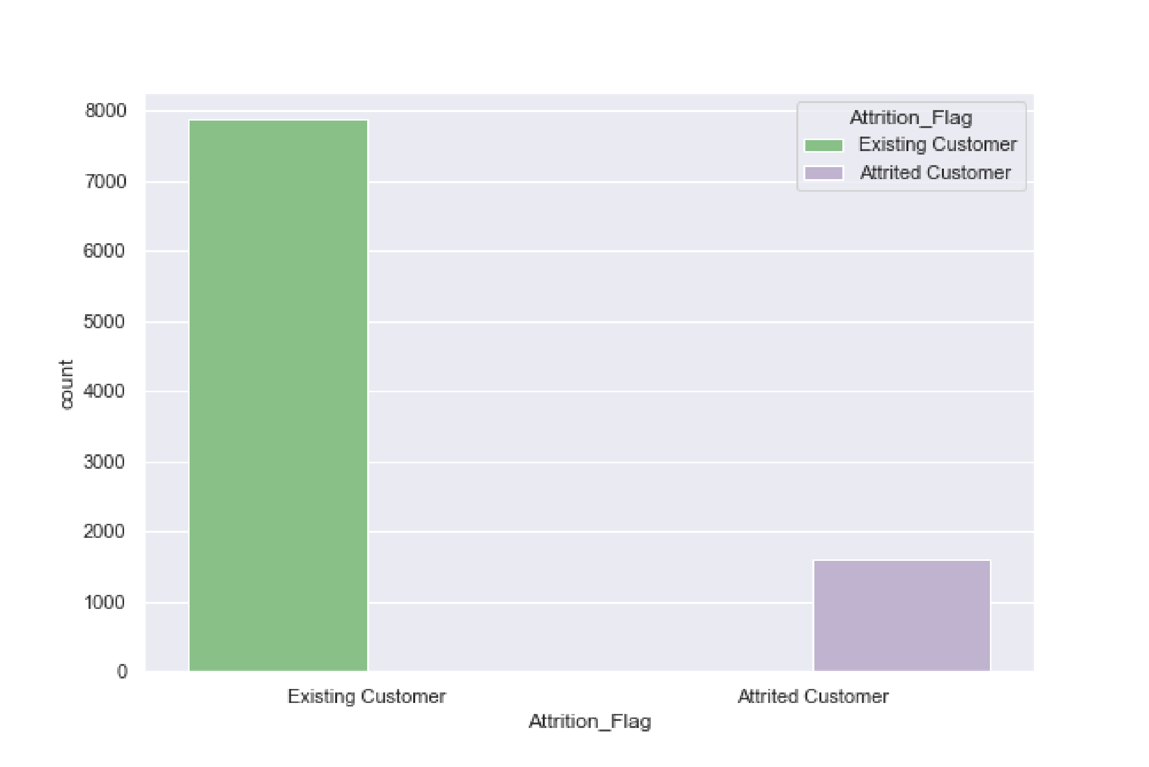
❑ Few attributes are skewed to the right.

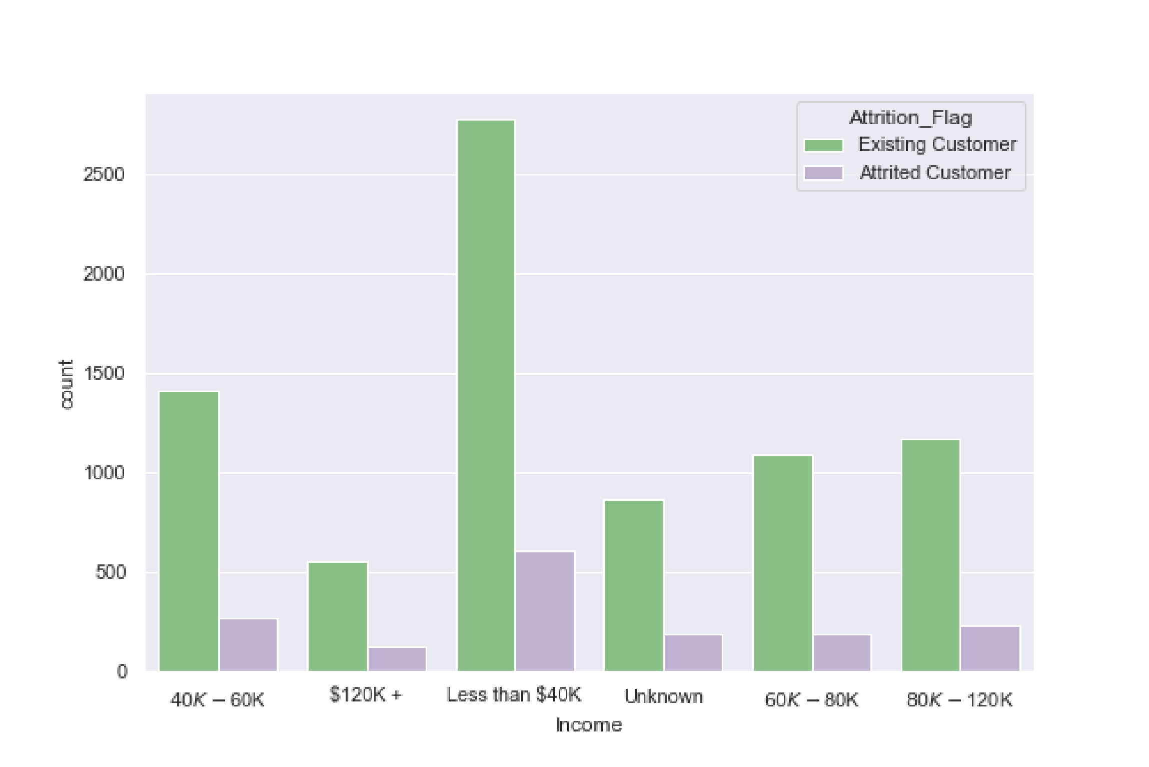
❑ Few attributes are bimodal, and the rest does not have any relationship at all.



# Distribution of Categorical Attributes

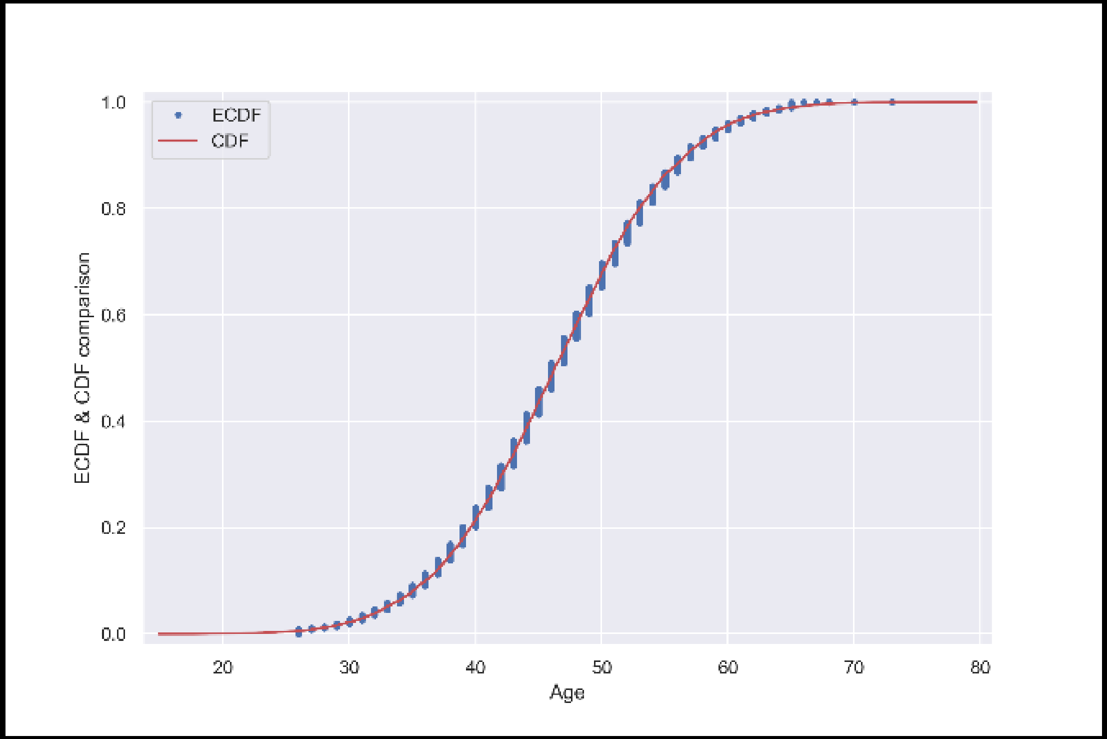
- ❑ The distribution of categorical attributes
- ❑ The target attributes contained class imbalance data
- ❑ A significant number of customers Income is below $40K.





# Visualizing Normality using ECDF

- □ The distribution of Age is normal.
- □ Few other columns distribution are also approximately normal.
- □ The rest of the attributes have no normality.



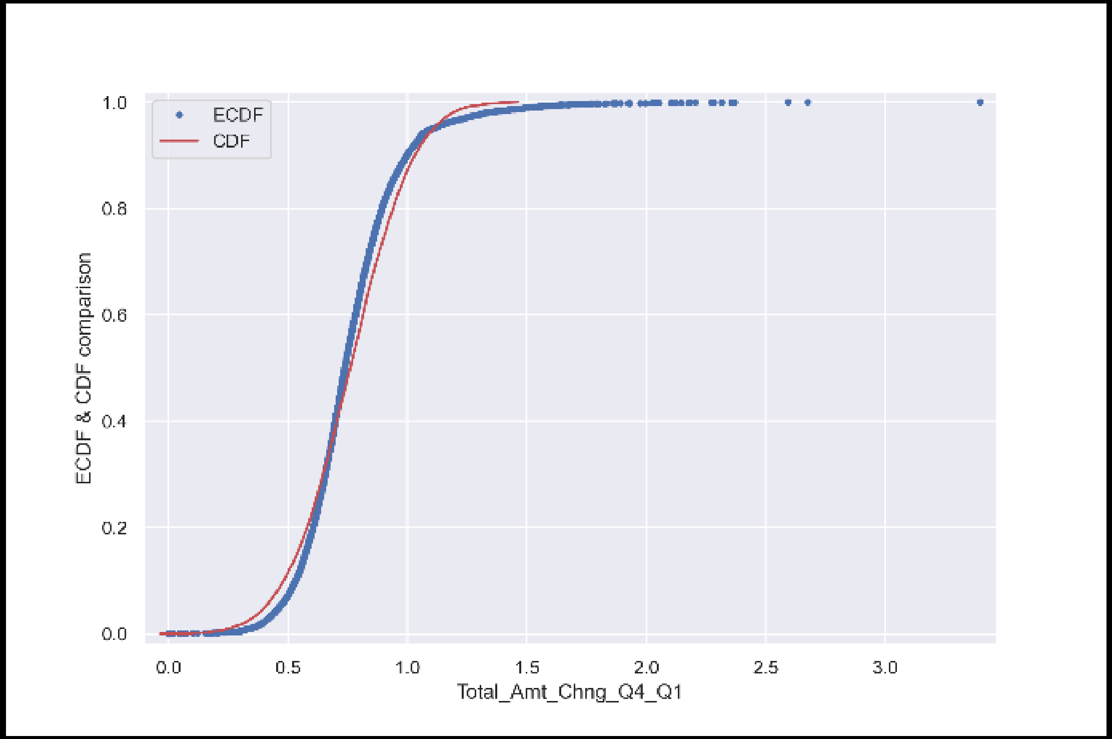
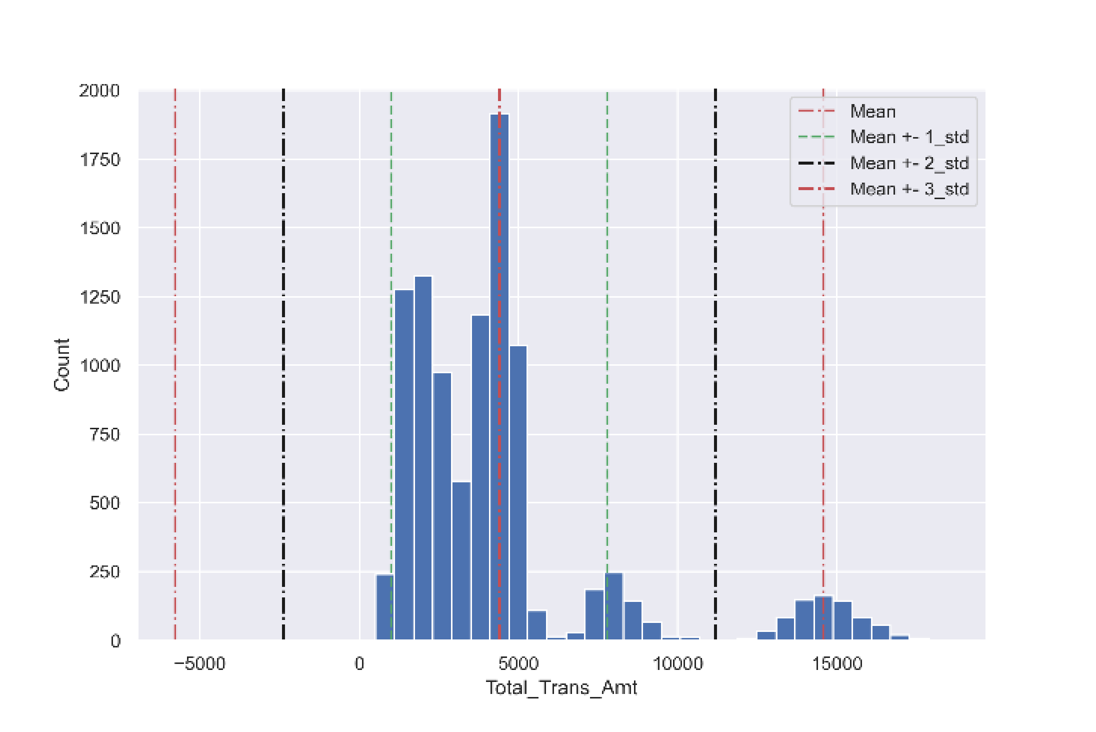


Fig. Theoretical cdf and ecdf comparison of numerical columns

# Visualizing Outliers

- Standard deviation ($Mean \pm 3 * Standard\_Deviation$) based outlier detection.
- By observing the histogram and vertical lines (Mean, $Mean \pm 3 * Standard\_Deviation$) it is clear that some numerical attribute contains outliers.



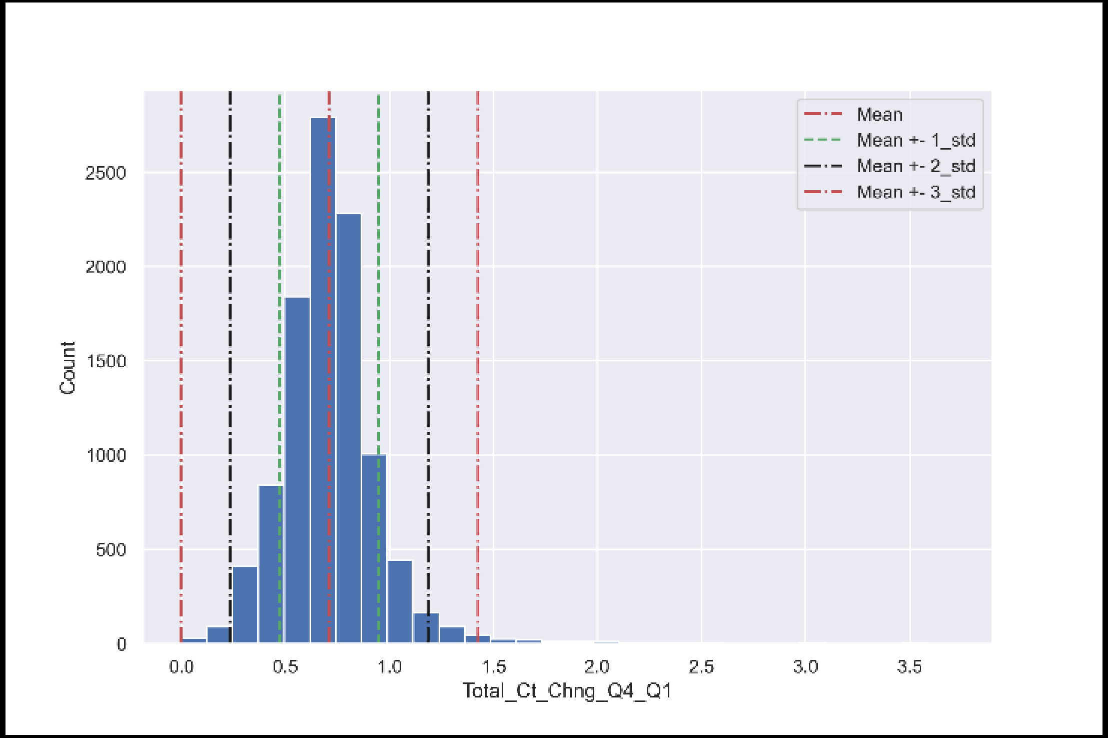


Fig. $Mean \pm 3 * Standard\_Deviation$ based outlier detection.

# Visualizing Correlations

- ❑ Heatmaps are essential for visualizing the correlation between columns.
- ❑ Few columns have significantly high positive correlations
- ❑ Few others have correlations, that are not too high.

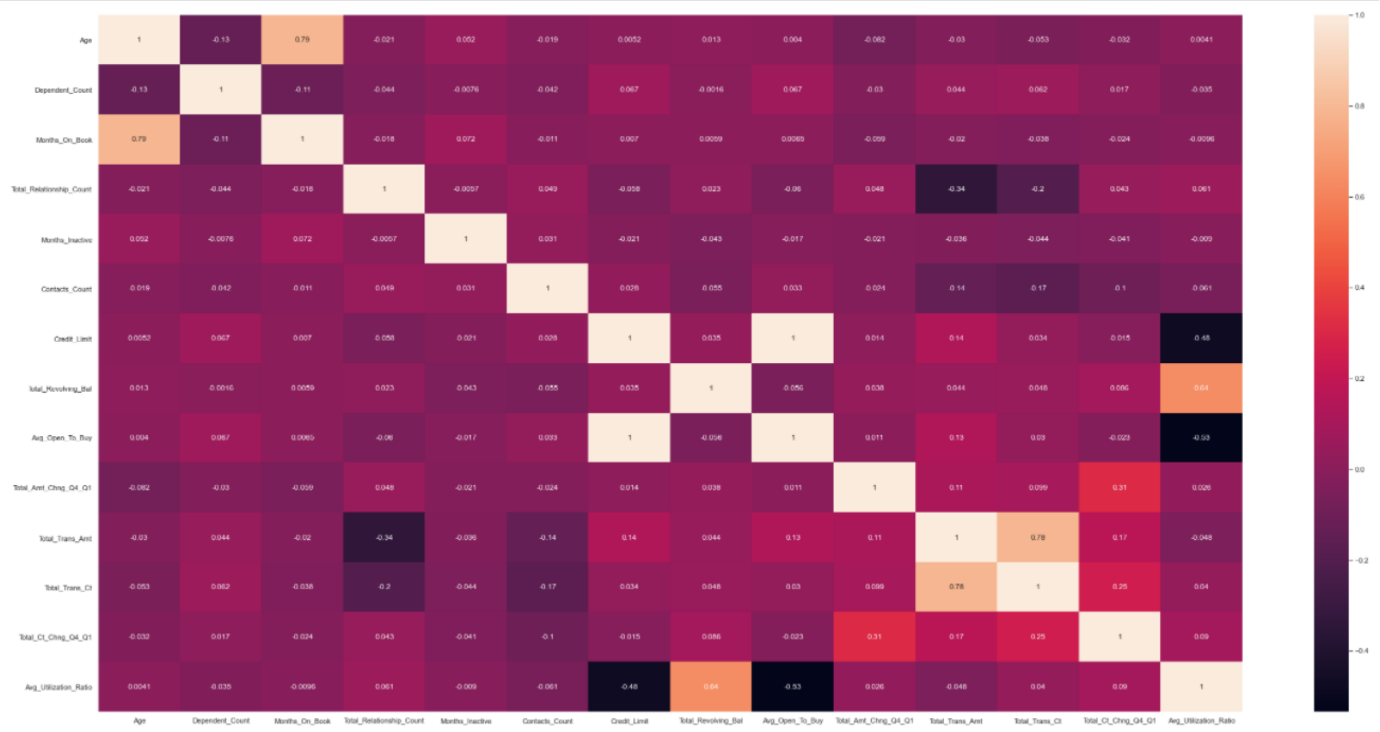


Fig. Correlation between numerical attributes

# Modeling Steps

❑ Preprocess the data

❑ Tuning hyperparameters for each model

❑ Train and Test each model using best hyperparameters

❑ Evaluate using evaluation metrics

❑ Find the best model

# Preprocessing

- ❑ Encoded the target attribute ‘Attrition_Flag’ to binary 1 and 0 (‘1’ for 'Attrited Customer' and ‘0’ for 'Existing Customer’).

- ❑ Separated the dependent variable y, and the independent variables X from the dataset.

- ❑ Split the data (X and y) into train and test for performing training and evaluation for each classifier.

- ❑ Used StandardScaler() function for scaling the training and test set separately.

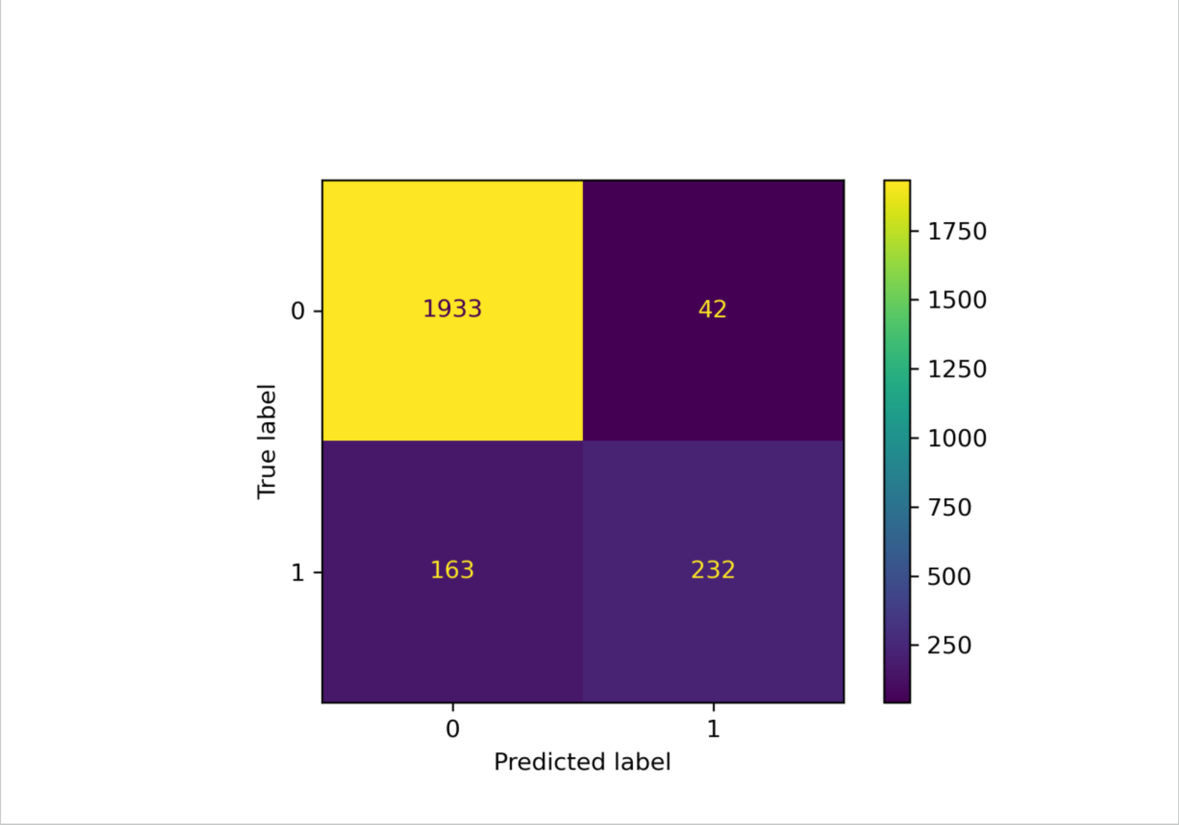
- ❑ Used the pandas get_dummies() function for one hot encoding of the categorical attributes.

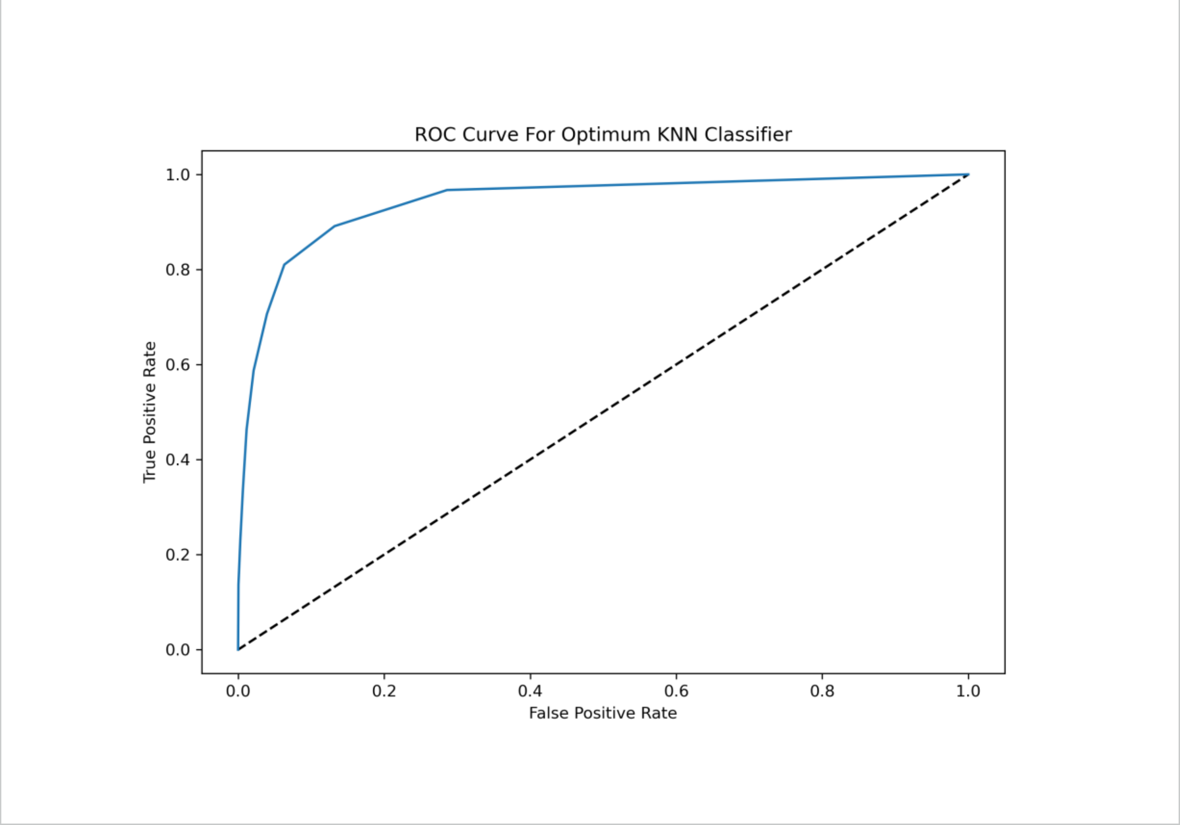
# ML Models used for Customer Churn prediction

- ❑ KNN Classifier
- ❑ Logistic Regression Classifier
- ❑ Decision Tree Classifier
- ❑ SVM Classifier
- ❑ Random Forest Classifier
- ❑ Gradient Boosting Classifier
- ❑ XGBoost Classifier

# Customer Churn Prediction using KNN Classifier

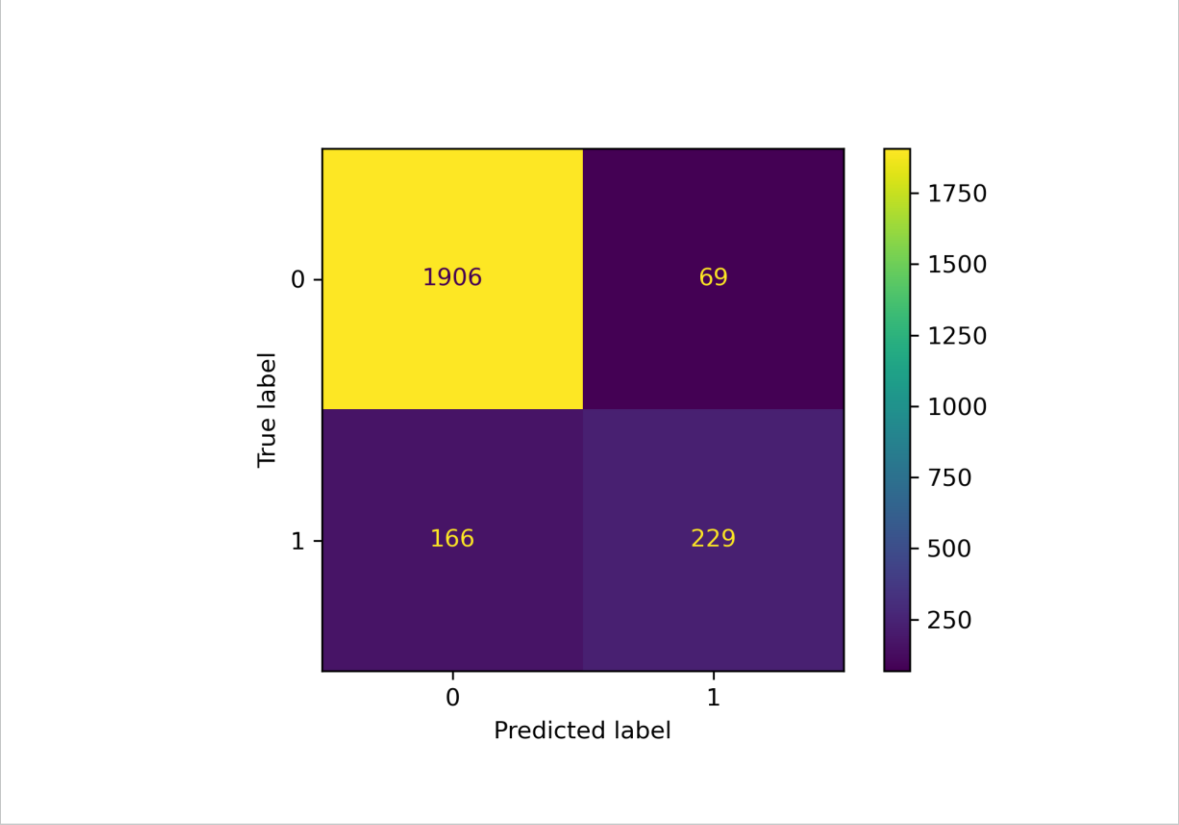
- Training Accuracy: 0.951
- Test Accuracy: 0.904
- Precision Score: 0.791
- Recall Score: 0.603
- F1 Score: 0.902
- Area Under ROC: 0.895

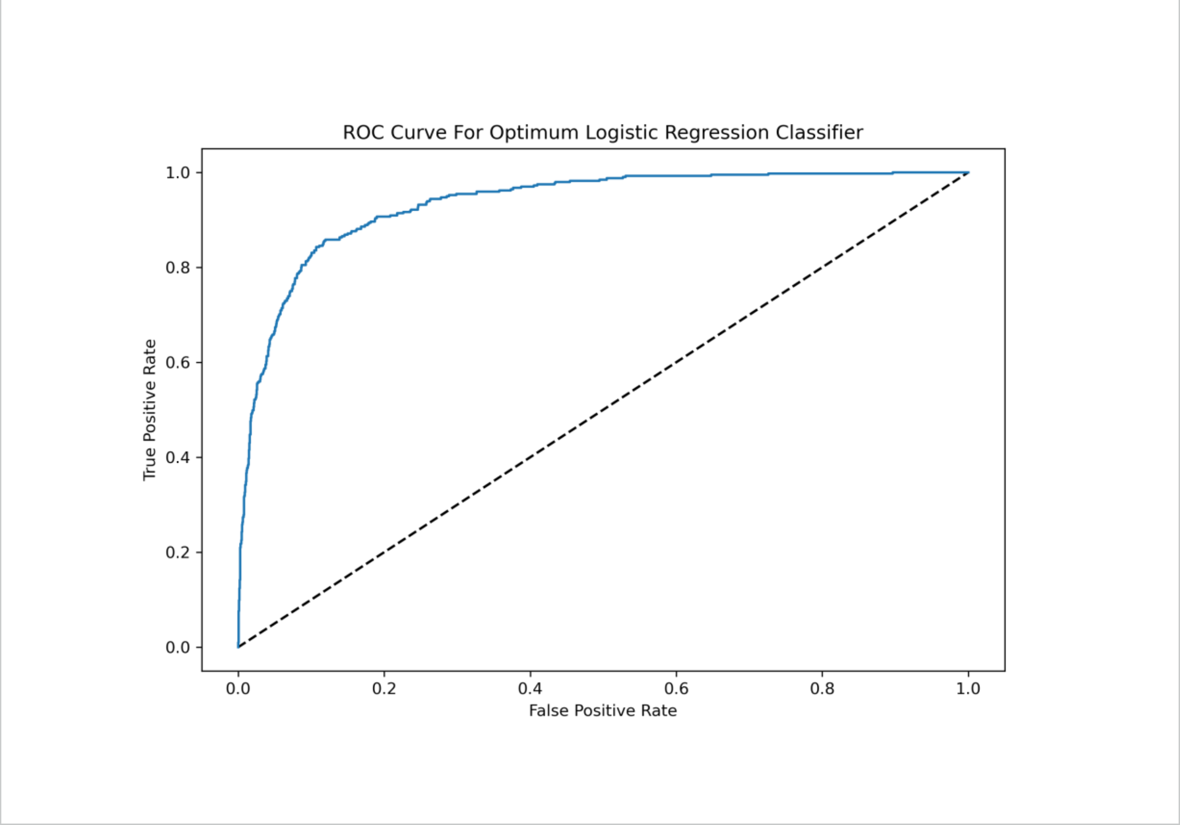




# Customer Churn Prediction using Logistic Regression Classifier

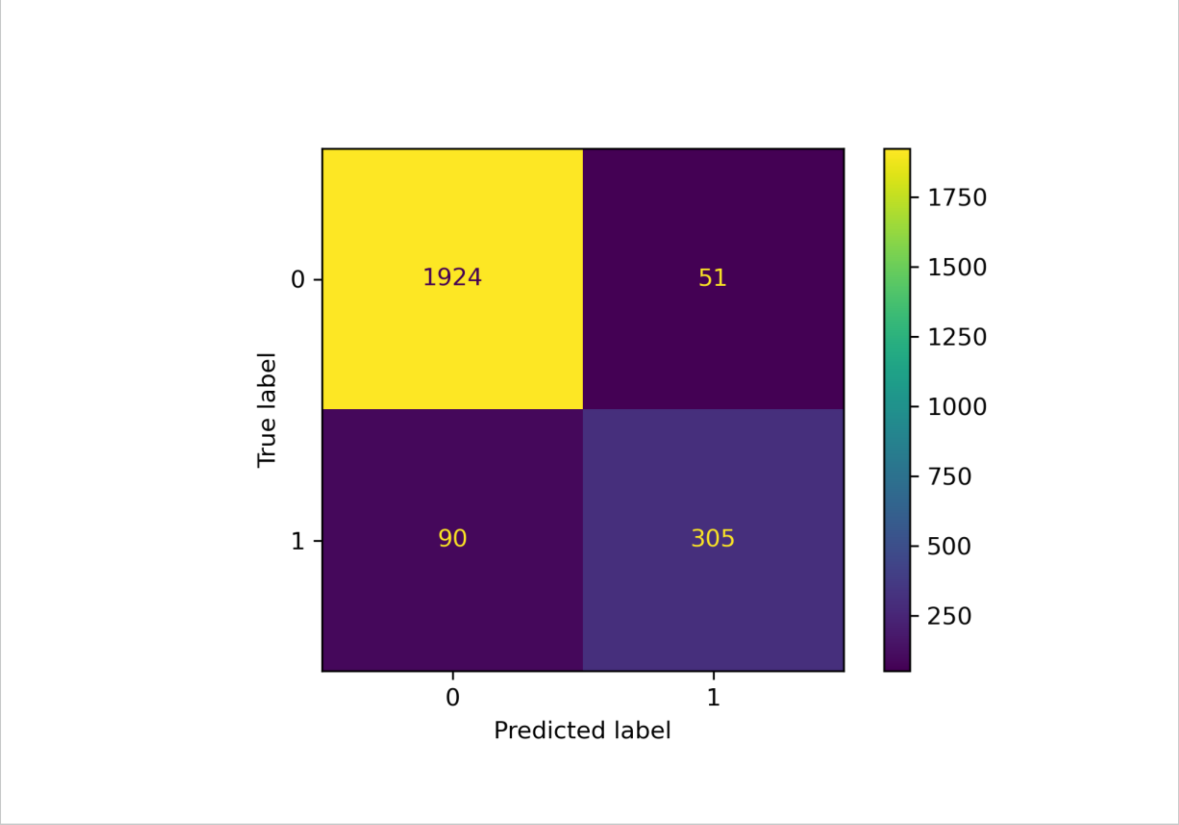
- ○ Training Accuracy: 0.900
- ○ Test Accuracy: 0.901
- ○ Precision Score: 0.768
- ○ Recall Score: 0.580
- ○ F1 Score: 0.895
- ○ Area Under ROC: 0.934

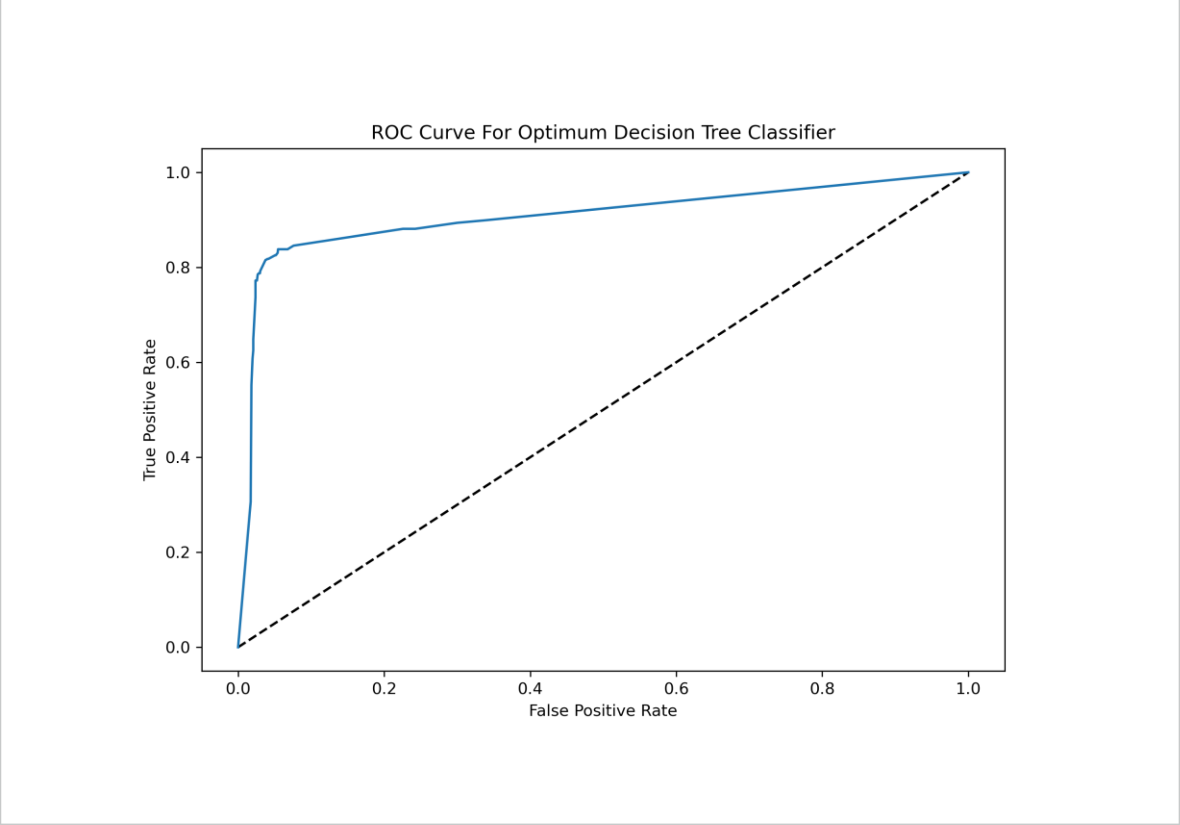




# Customer Churn Prediction using Decision Tree Classifier

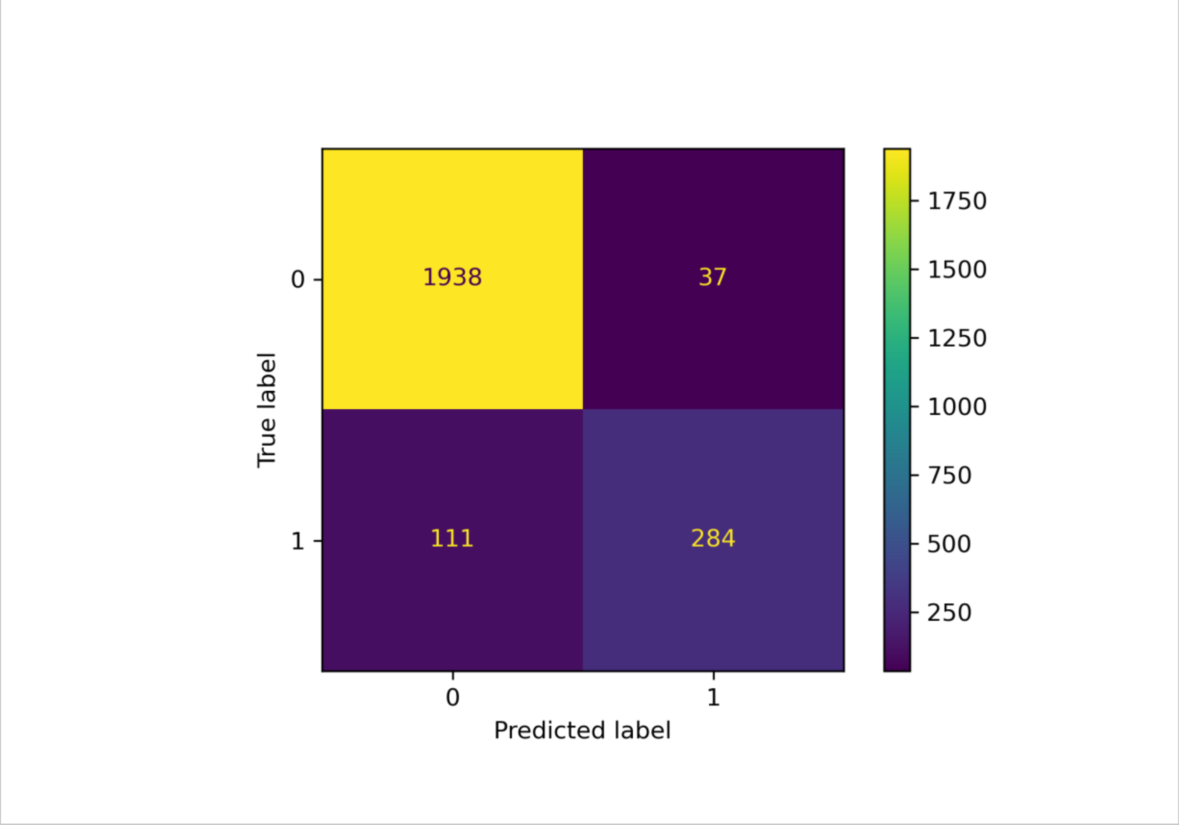
- ○ Training Accuracy: 0.982
- ○ Test Accuracy: 0.938
- ○ Precision Score: 0.853
- ○ Recall Score: 0.762
- ○ F1 Score: 0.937
- ○ Area Under ROC: 0.900

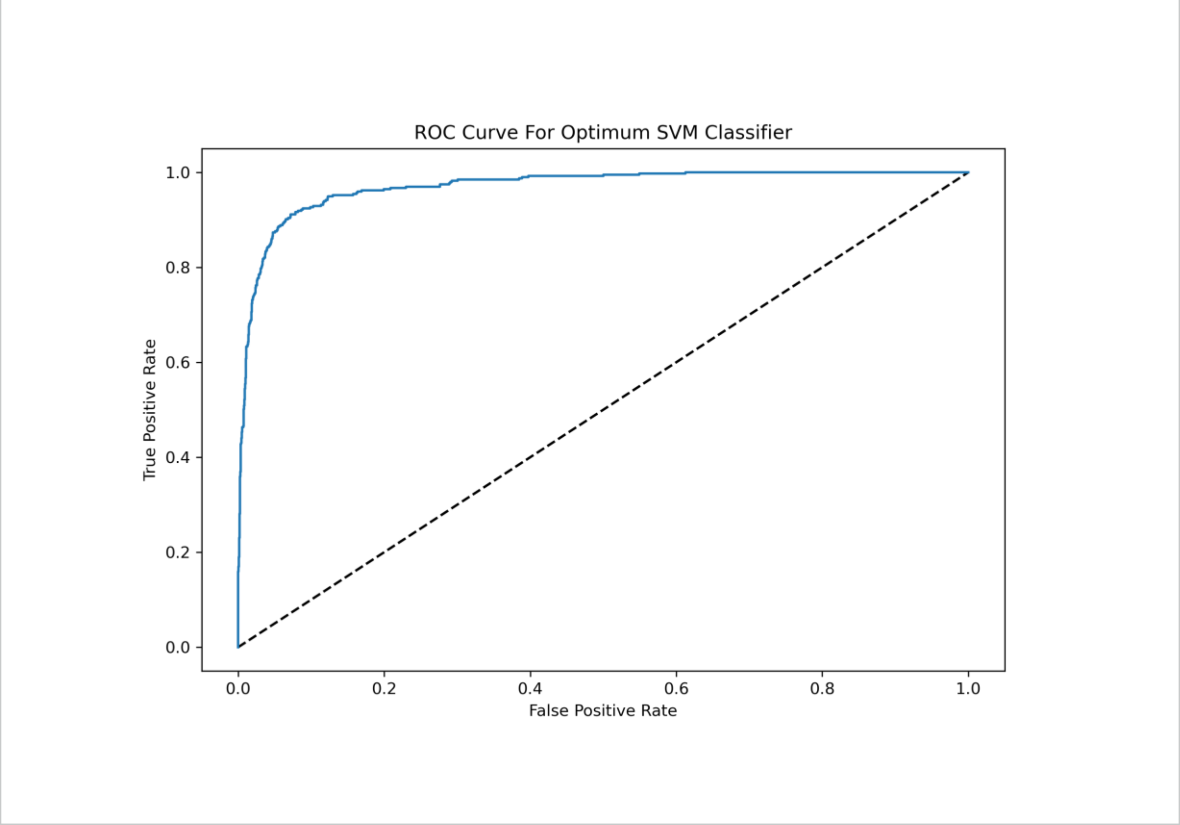




# Customer Churn Prediction using SVM Classifier

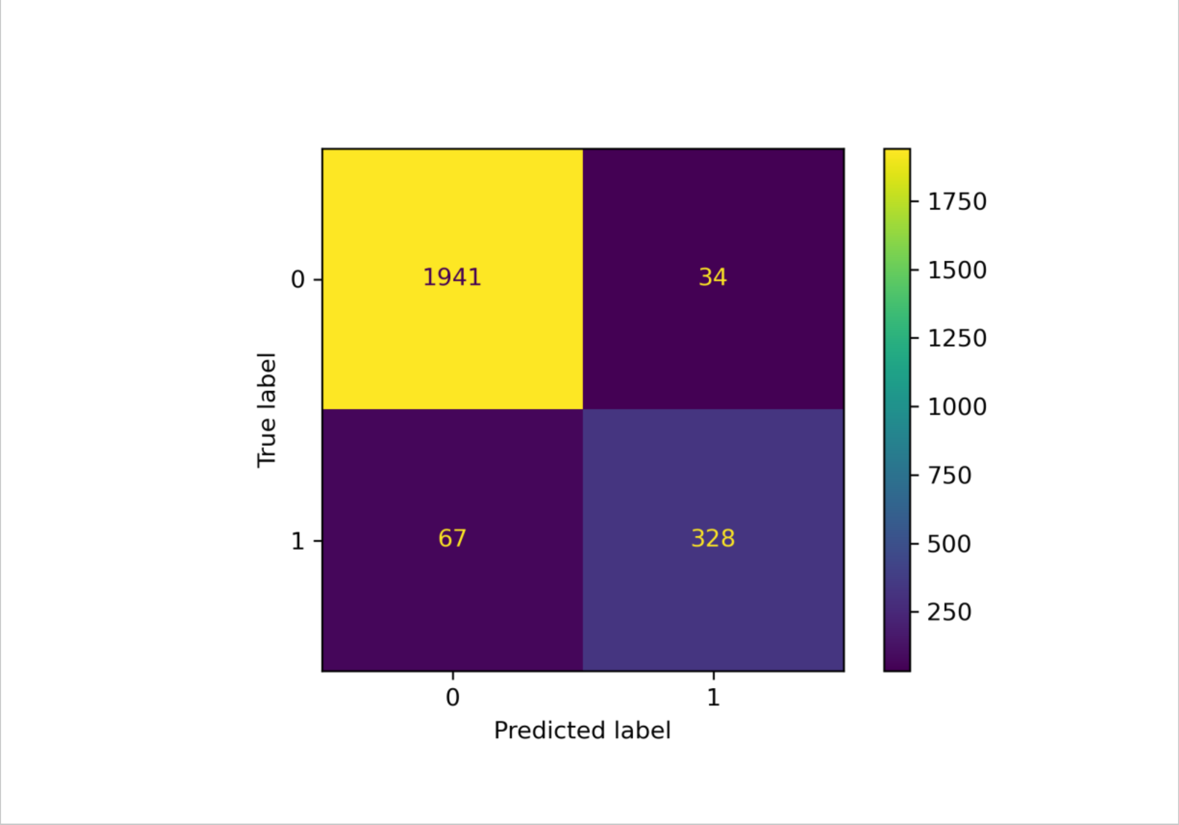
- ○ Training Accuracy: 0.966
- ○ Test Accuracy: 0.938
- ○ Precision Score: 0.885
- ○ Recall Score: 0.719
- ○ F1 Score: 0.935
- ○ Area Under ROC: 0.970

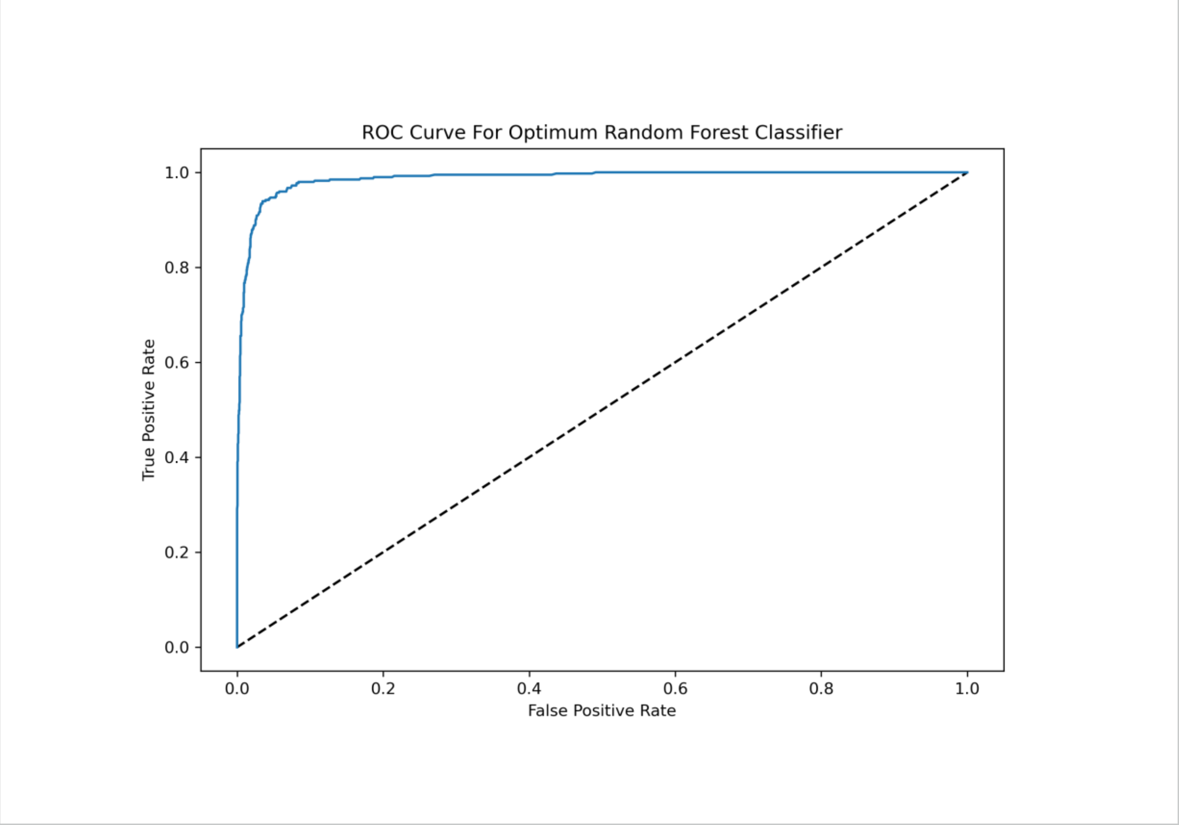




# Customer Churn Prediction using Random Forest Classifier

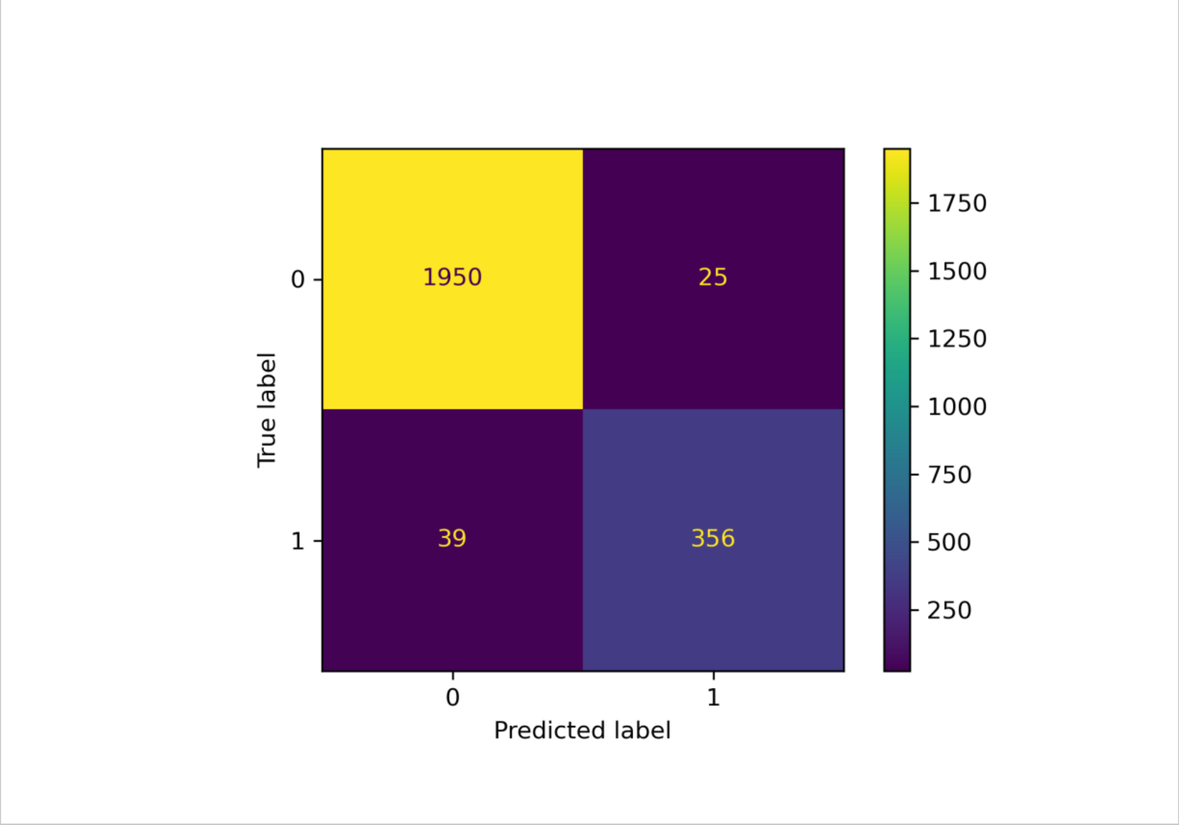
- Training Accuracy: 1.000
- Test Accuracy: 0.958
- Precision Score: 0.909
- Recall Score: 0.830
- F1 Score: 0.957
- Area Under ROC: 0.987

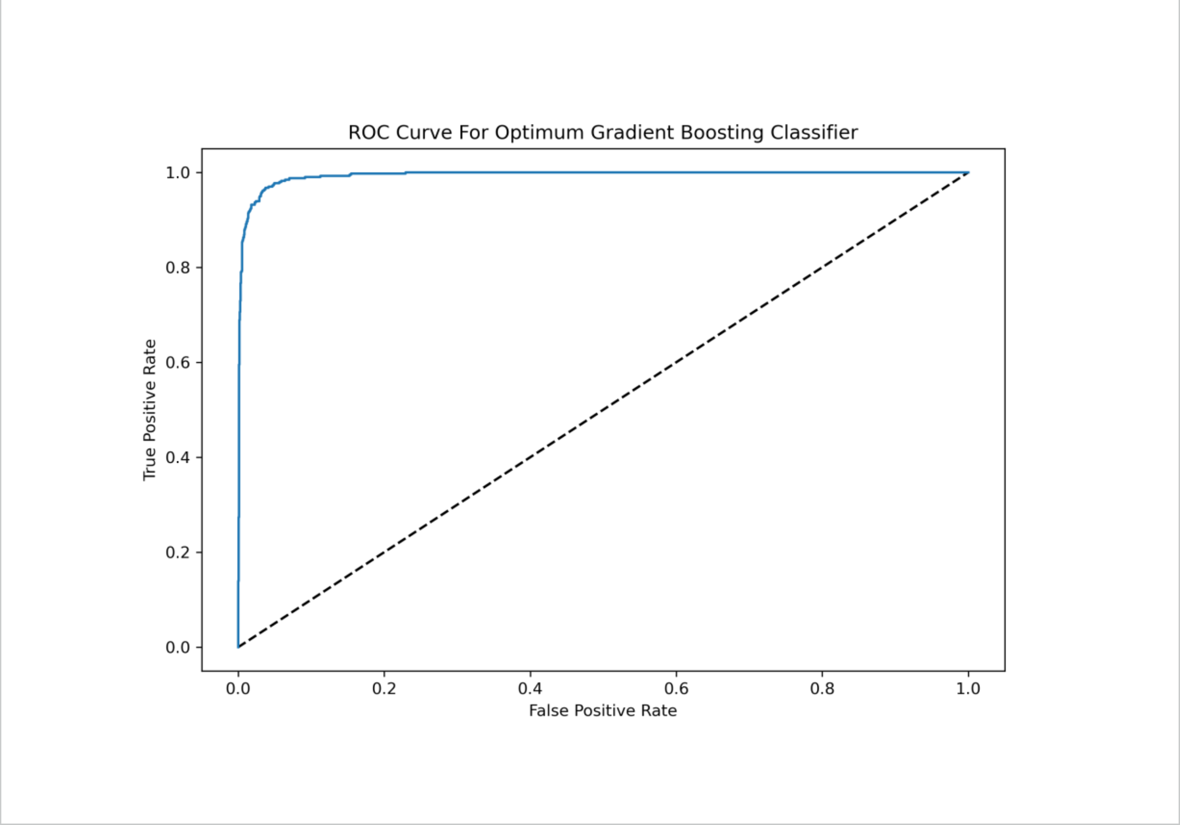




# Customer Churn Prediction using Gradient Boosting Classifier

- Training Accuracy: 0.997
- Test Accuracy: 0.973
- Precision Score: 0.934
- Recall Score: 0.901
- F1 Score: 0.973
- Area Under ROC: 0.994

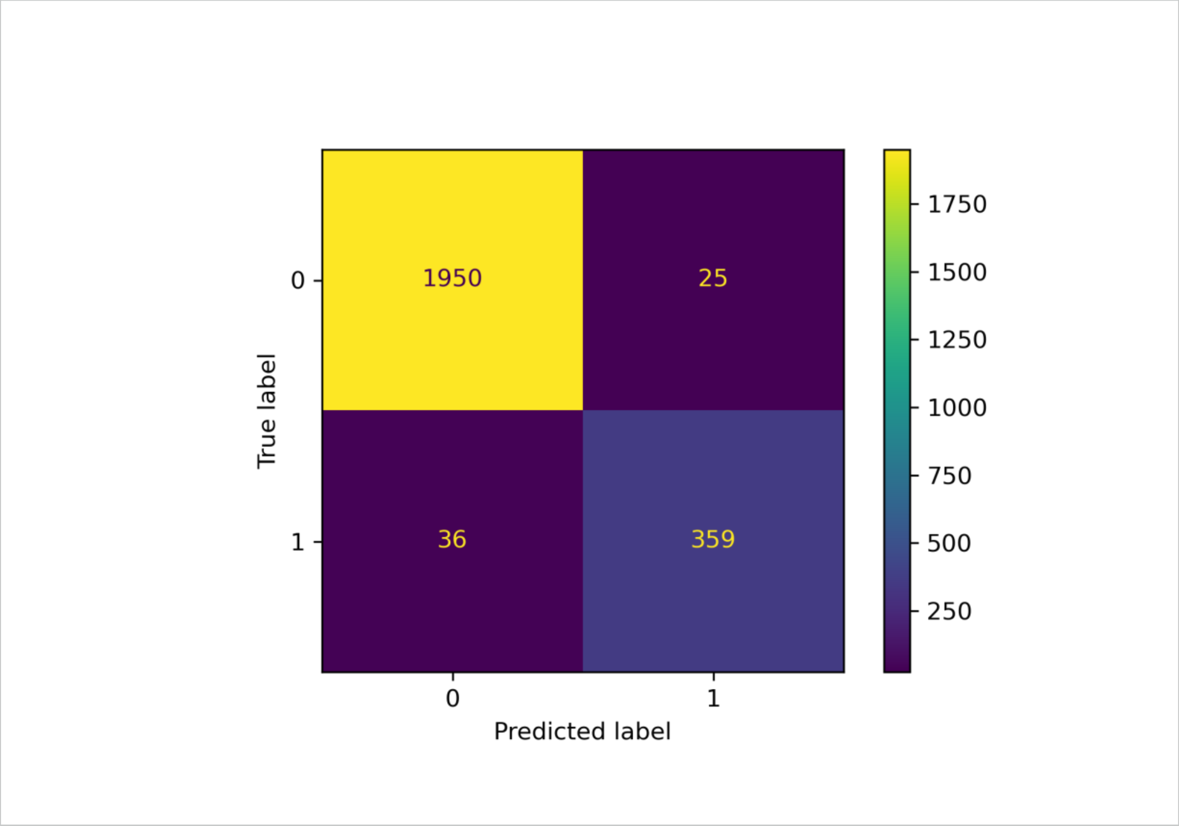


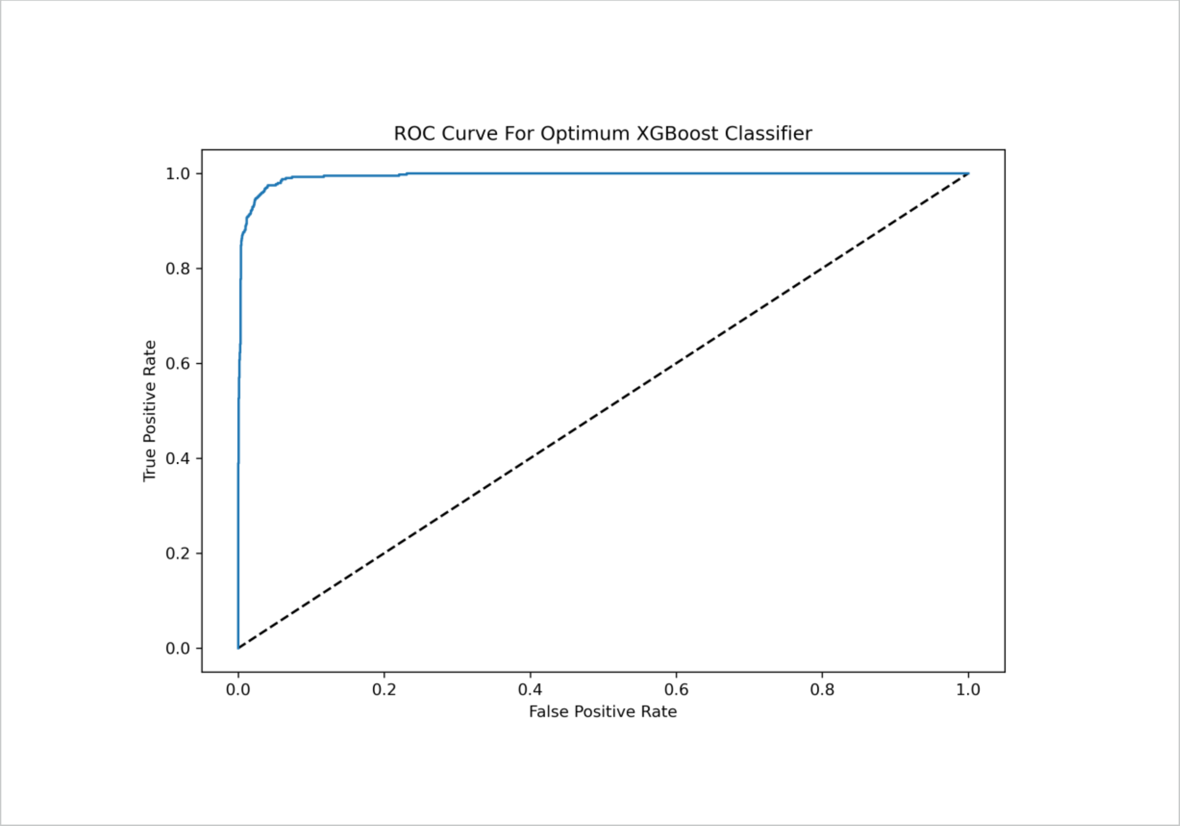


# Customer Churn Prediction using XGBoost Classifier

- Training Accuracy: 0.999
- Test Accuracy: 0.974
- Precision Score: 0.935

- Recall Score: 0.909
- F1 Score: 0.974
- Area Under ROC: 0.994





# Best Model for Predicting Customer Churn and Findings

❑ The XGBoost Classifier outperformed well based on all model evaluation metrics I have used.

❑ I have used the metric ‘Recall’ as my main evaluation metric.

❑ To predict all possible churning customer the false negatives should be low, and recall should be high.

❑ Despite the class imbalance data, the ‘Recall’ value of 0.909 or 91% indicates that my model worked well on predicting possible churning customers.

❑ Also, the ‘F1 Score’ of 0.974 and the Area under roc curve of 0.994 indicates that my model performed well.

# Table showing the comparison among all Models Outcome

| Evaluation Metrics<br>ML Models | Training Accuracy | Test Accuracy | Precision Score | Recall Score | F1 Score | Area Under ROC |
|---|---|---|---|---|---|---|
| KNN | 0.951 | 0.907 | 0.791 | 0.603 | 0.902 | 0.895 |
| Logistic Regression | 0.900 | 0.901 | 0.768 | 0.580 | 0.895 | 0.934 |
| Decision Tree | 0.982 | 0.938 | 0.853 | 0.762 | 0.937 | 0.900 |
| SVM | 0.966 | 0.938 | 0.885 | 0.719 | 0.935 | 0.970 |
| Random Forest | 1.000 | 0.958 | 0.909 | 0.830 | 0.957 | 0.987 |
| Gradient Boosting | 0.997 | 0.973 | 0.934 | 0.901 | 0.973 | 0.994 |
| XGBoost | 0.999 | 0.974 | 0.935 | 0.909 | 0.974 | 0.994 |

# Conclusion

❑ The target attributes have class imbalance data, and approximately 16% of the customers are labeled with ‘Attrited Customer’ and the remaining 84% customers are labeled with ‘Current Customer’.

❑ Despite of the class imbalance, the ‘Recall’ value of 0.909 or 91%, ‘F1 Score’ of 0.974, and AUROC of 0.994 indicates that the XGBoost classifier accurately predicted the possible churning customers.

# Future Work

- ❑ There are also few correlations, and outliers in the dataset
- ❑ Due to those irregularities my model did not get 100% accuracy on all the metrices.
- ❑ Spending more time on preprocessing to tackle those issues will be helpful to get better accuracy on each of the evaluation metrics I have used.

Thank You