

# Predicting Bank Customers Attrition

Data Science Bootcamp Final Project  
25 February 2021

**Nada Alzahrani**



# Table Of Content



**Project  
Motivation**



**Exploratory  
Data Analysis**



**Preprocessing**



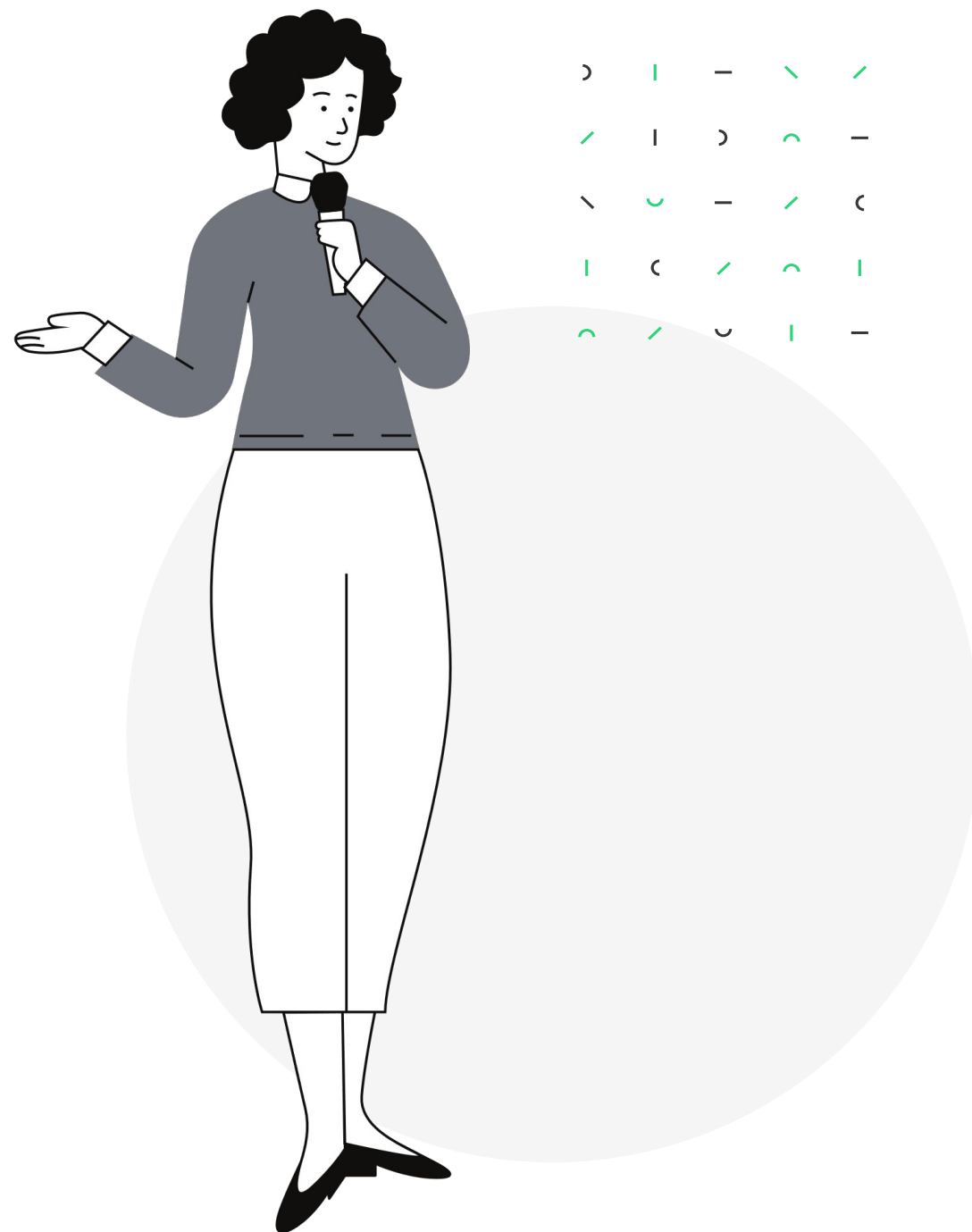
**Modeling**



**Results**

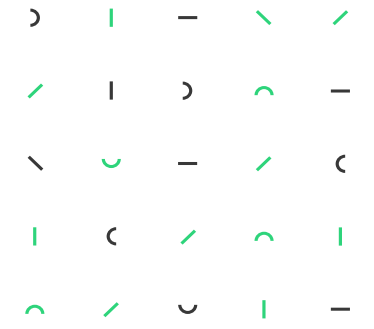
# Project Motivation

# 01



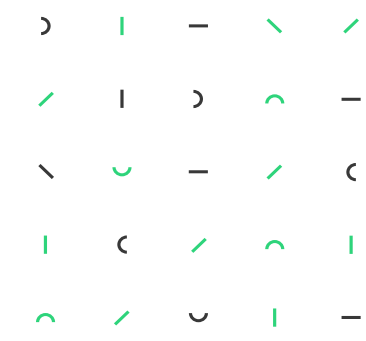
Customer attrition which is also known as customer churn, customer turnover is defined as the loss of customers in a business, it is one of the biggest concerns especially in banking since customers are considered as the most valuable part of it.

We are using the bank customer's data to predict possible attrited customers using Machine Learning to help prevent any possible attrition that may happen in the future.



# Exploratory Data Analysis

02



## Data Information

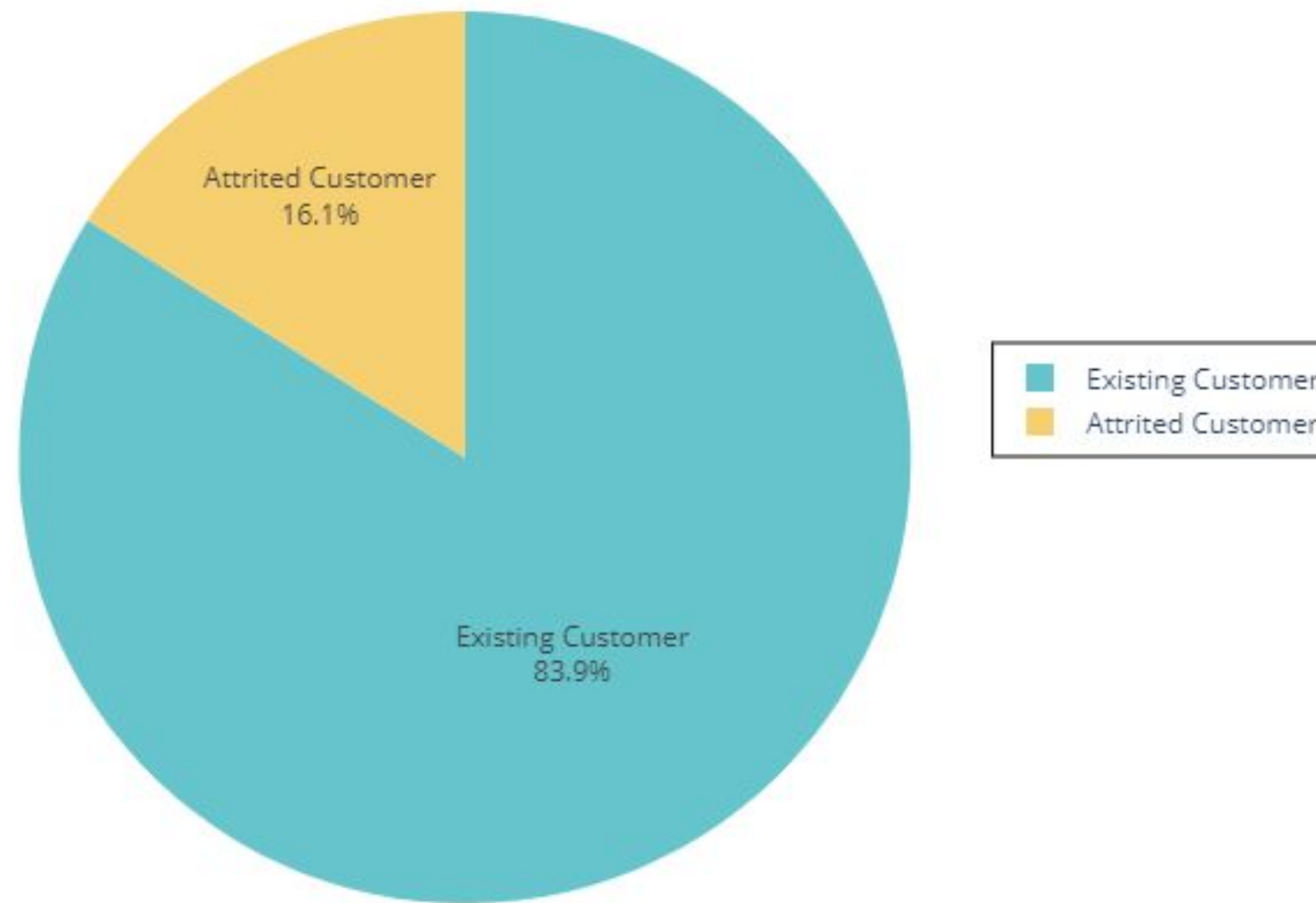
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Attrition_Flag                        10127 non-null  object
1   Customer_Age                         10127 non-null  int64
2   Gender                               10127 non-null  object
3   Dependent_count                      10127 non-null  int64
4   Education_Level                      10127 non-null  object
5   Marital_Status                      10127 non-null  object
6   Income_Category                     10127 non-null  object
7   Card_Category                       10127 non-null  object
8   Months_on_book                      10127 non-null  int64
9   Total_Relationship_Count            10127 non-null  int64
10  Months_Inactive_12_mon              10127 non-null  int64
11  Contacts_Count_12_mon              10127 non-null  int64
12  Credit_Limit                       10127 non-null  float64
13  Total_Revolving_Bal                10127 non-null  int64
14  Avg_Open_To_Buy                    10127 non-null  float64
15  Total_Amt_Chng_Q4_Q1               10127 non-null  float64
16  Total_Trans_Amt                    10127 non-null  int64
17  Total_Trans_Ct                     10127 non-null  int64
18  Total_Ct_Chng_Q4_Q1                10127 non-null  float64
19  Avg_Utilization_Ratio               10127 non-null  float64
dtypes: float64(5), int64(9), object(6)
memory usage: 1.5+ MB
```





## Class Distribution

Type Of Customers Based On Their Account Status



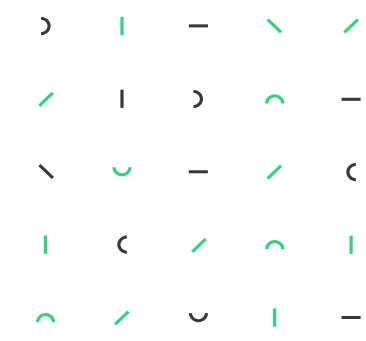
04



## Age Range

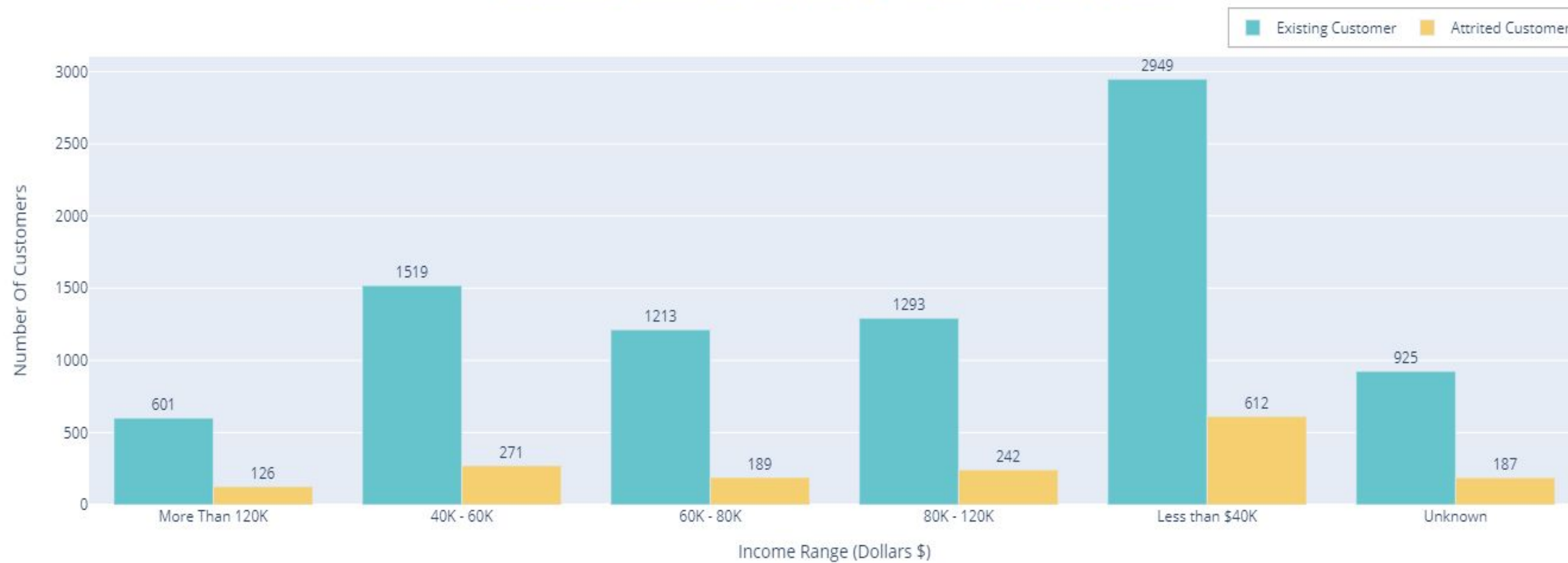


05



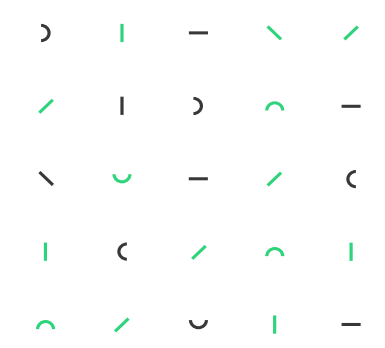
## Annual Income

Customer's Annual Income By Their Account Status



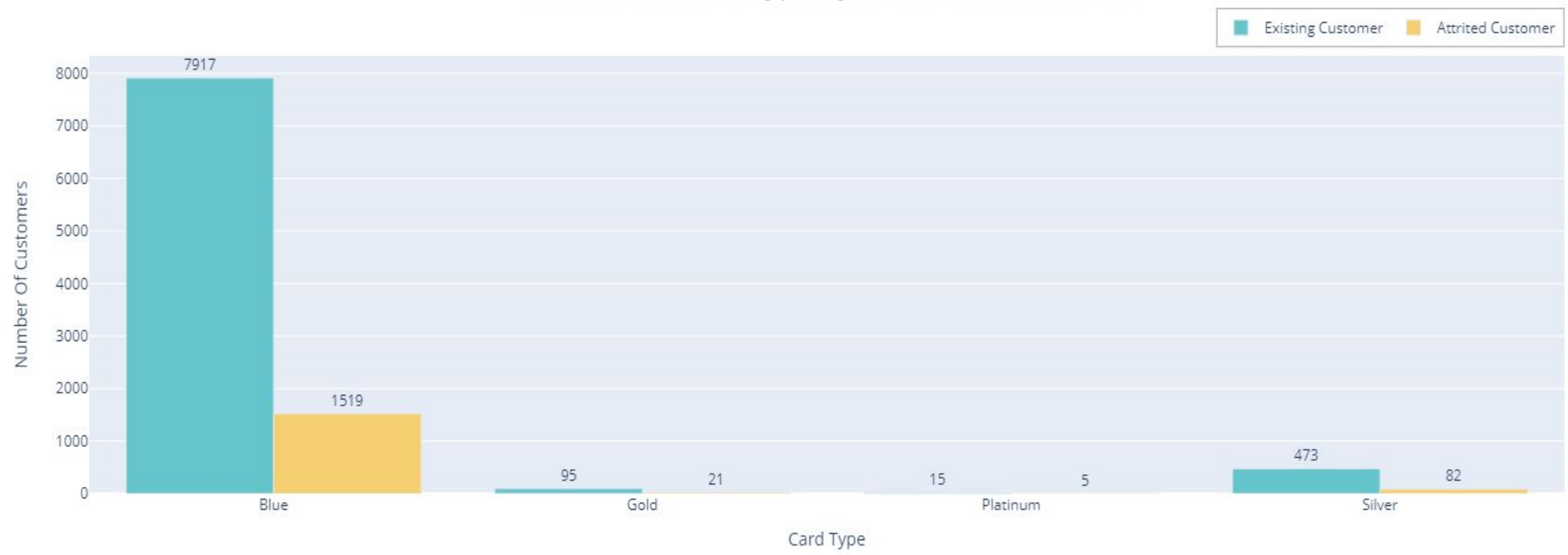
06

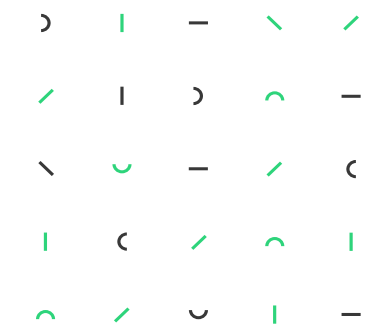




# Credit Card Type

Customer's Card Type By Their Account Status





# Period Of Relationship With The Bank

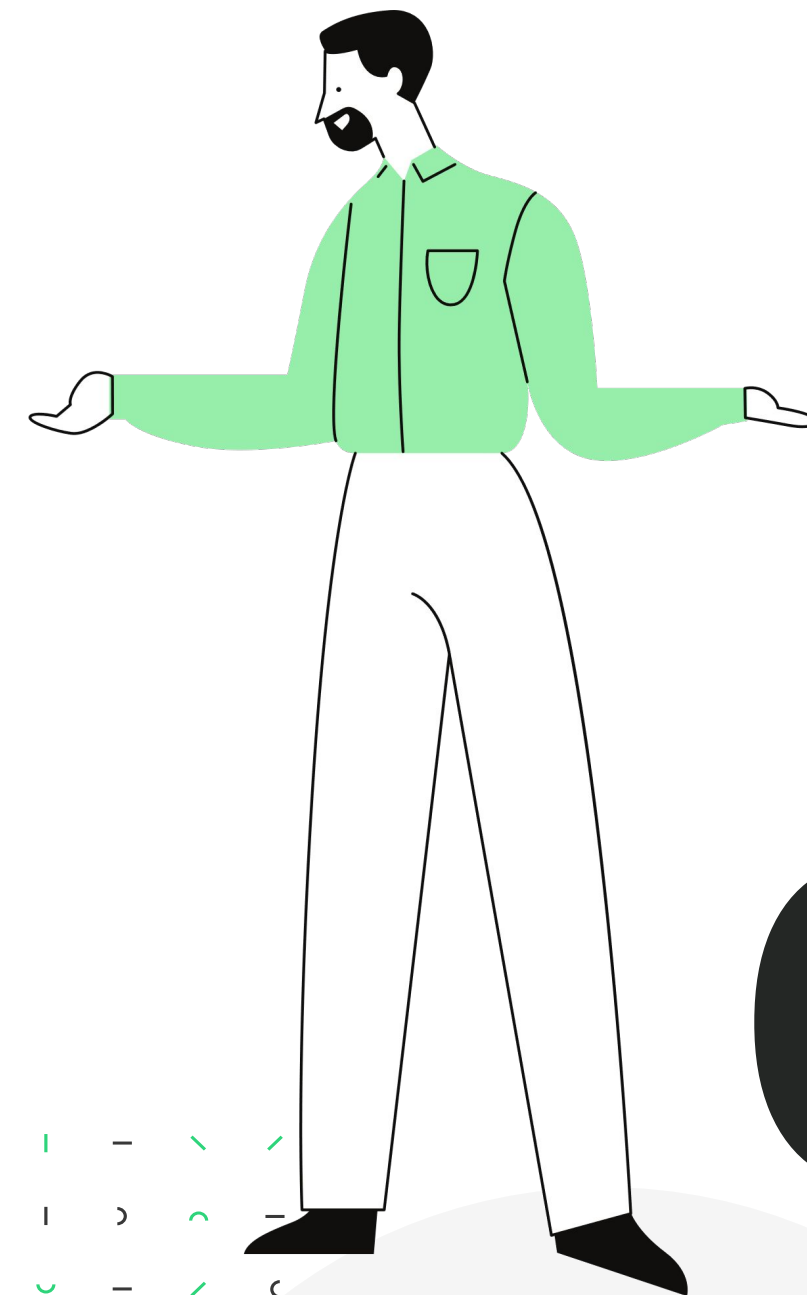
Period Of Relationship With The Bank



08

# Preprocessing

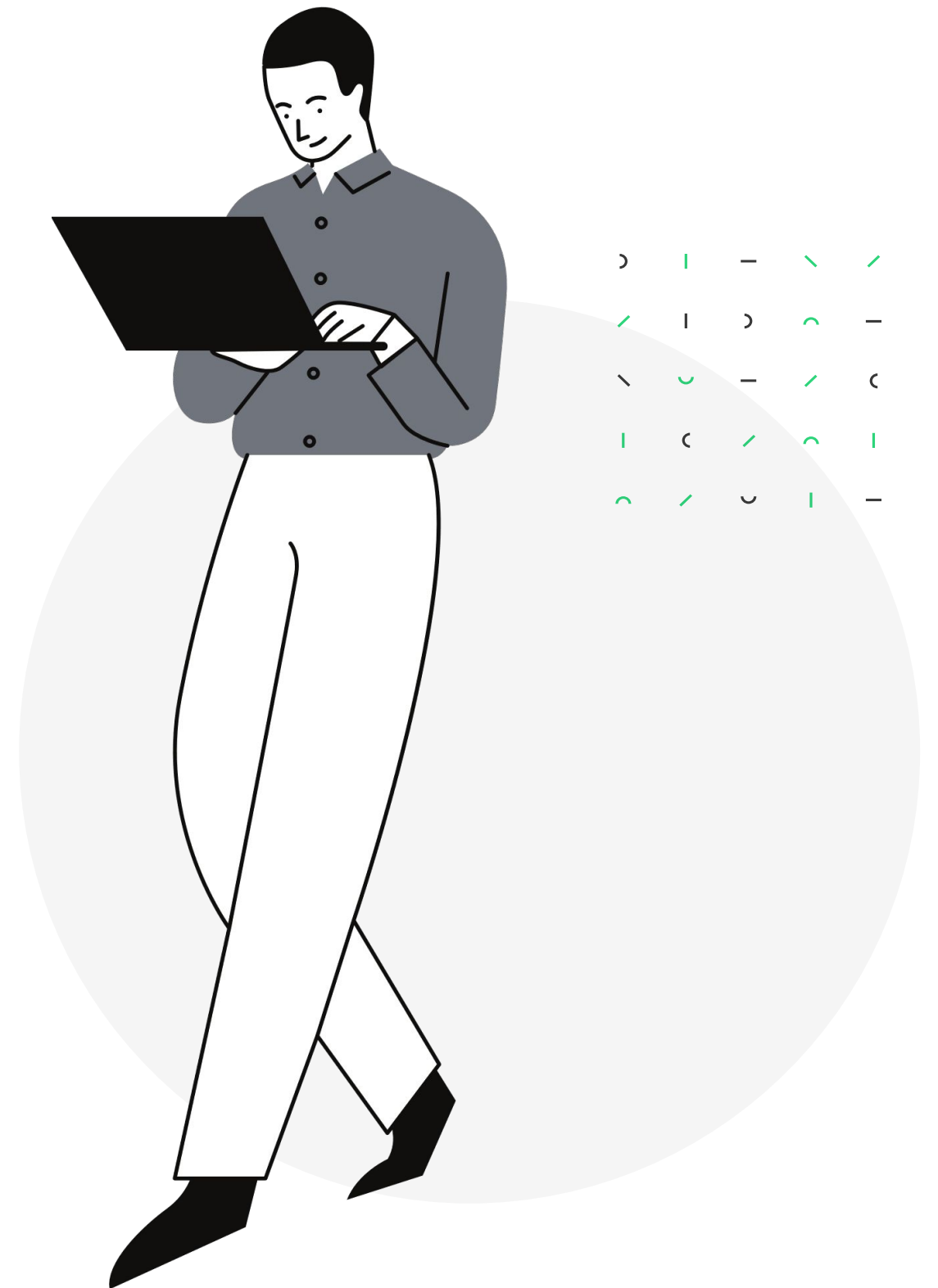
- **Encoding Categorical Features**  
Using Label Encoder.
- **Split Data 80% Train, 20% Test**



09

# Modeling

- **Baseline Model.**
- **Ensemble learning Method.**
- **Tuning Highest Scoring Model With GridSearchCV.**



# Ensemble Results

Model: Logistic Regression, Score: 0.8810108832096327

Model: knn, Score: 0.8914288569583075

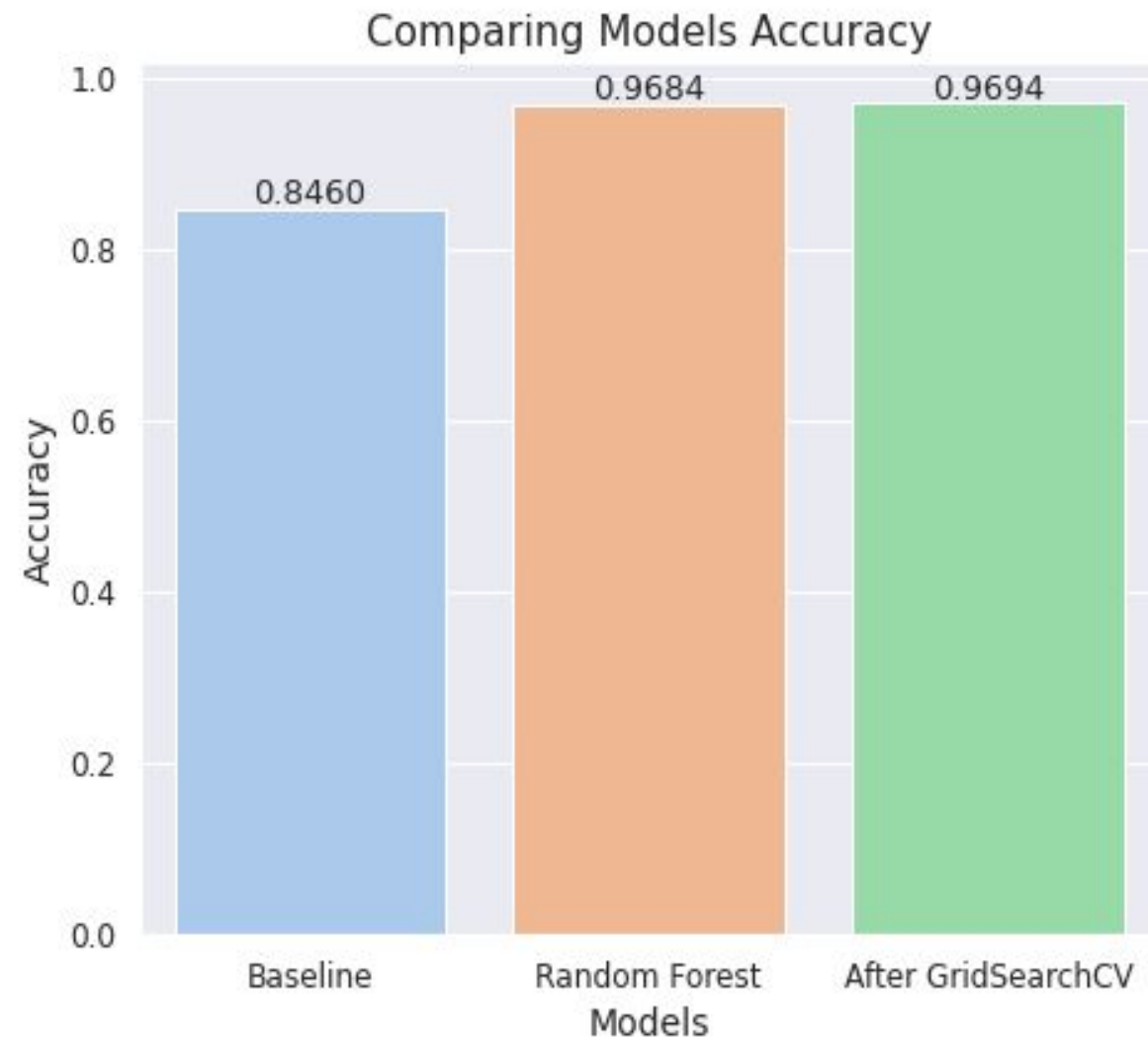
Model: Decision Tree, Score: 0.9388765797716111

Model: Random Forest, Score: 0.963020023643255

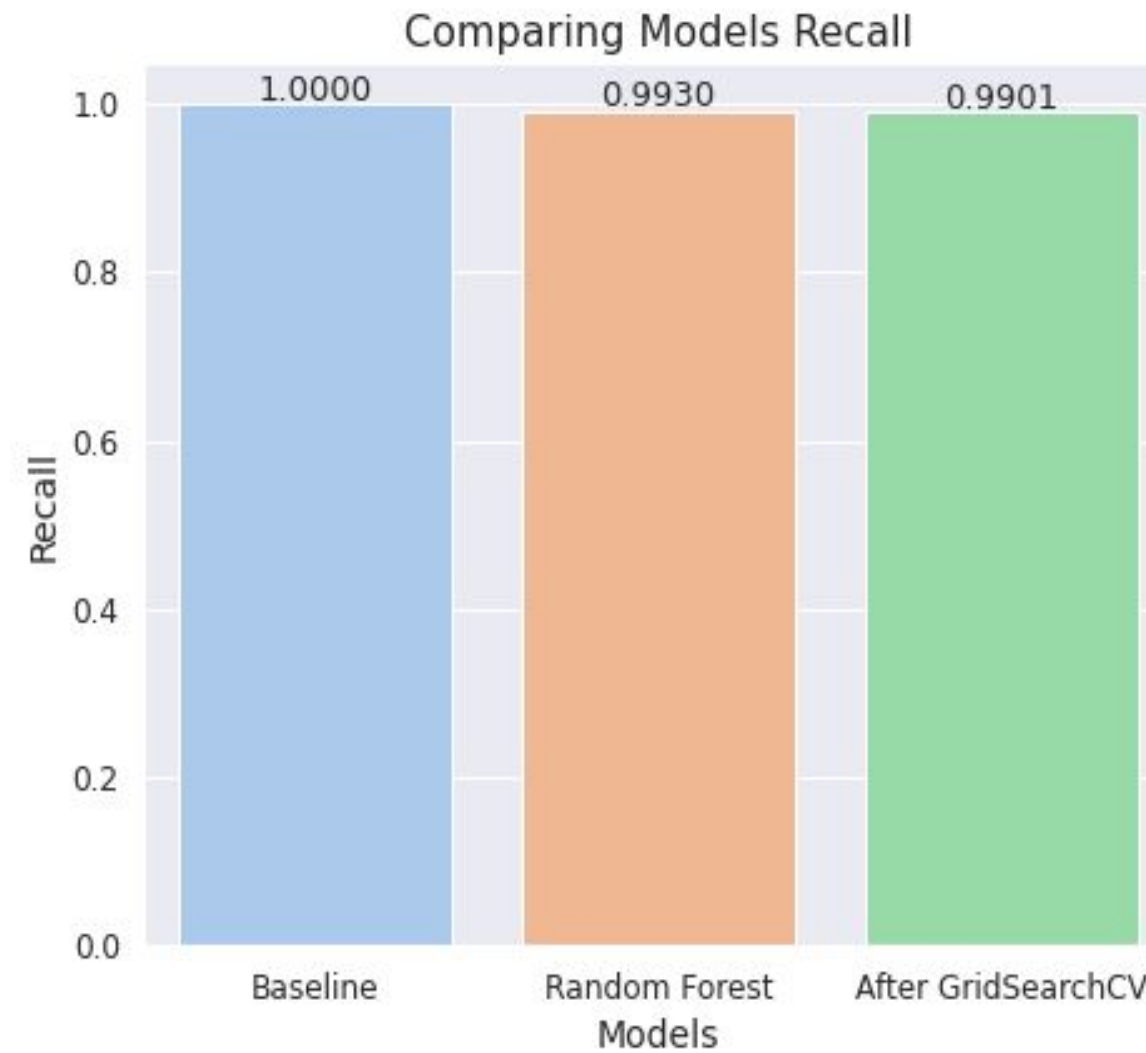
- Since **Random Forest Classifier** Scored the highest among other model i will use it and tune it to achive the best results

# Results

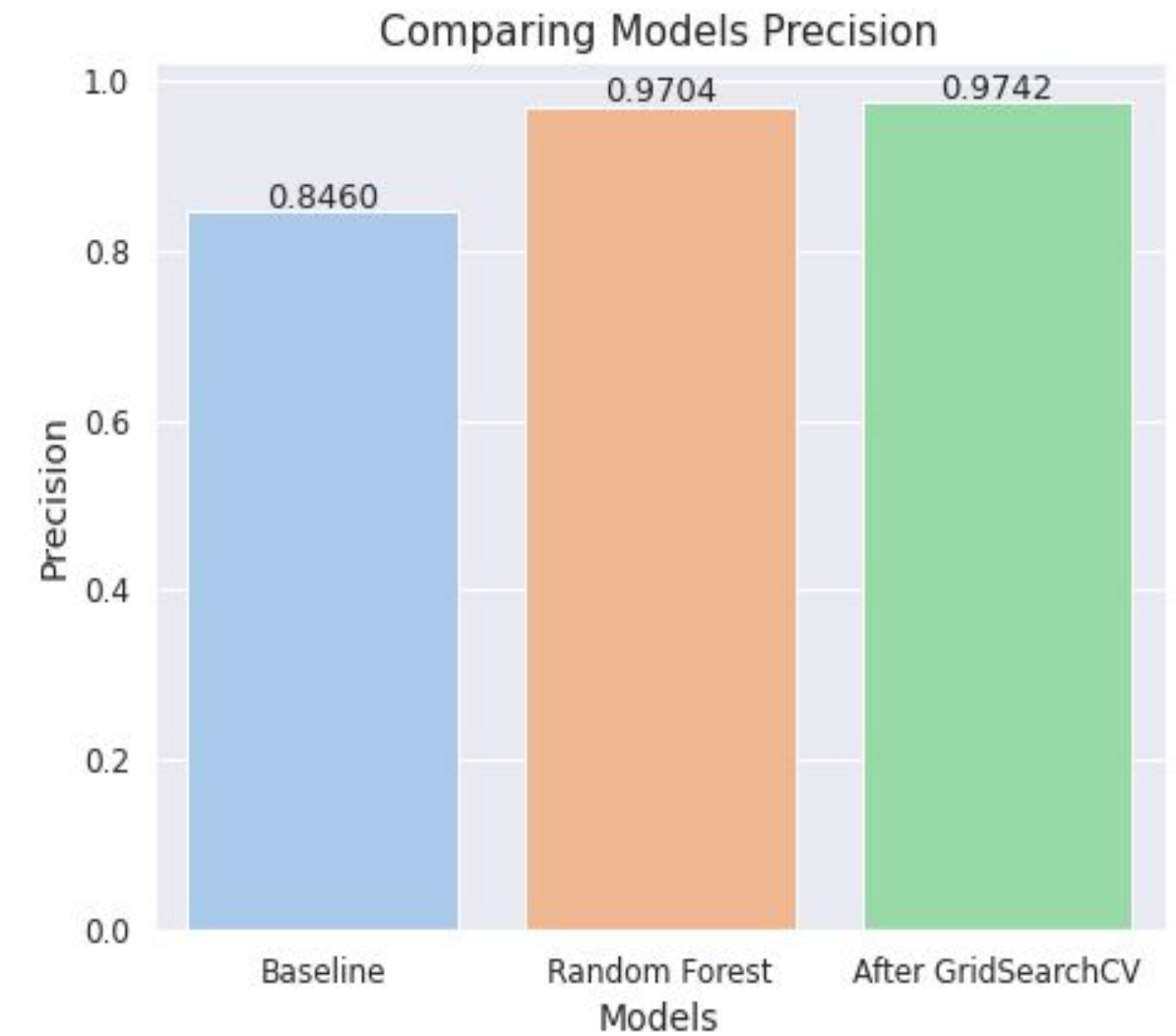
## Accuracy



## Recall

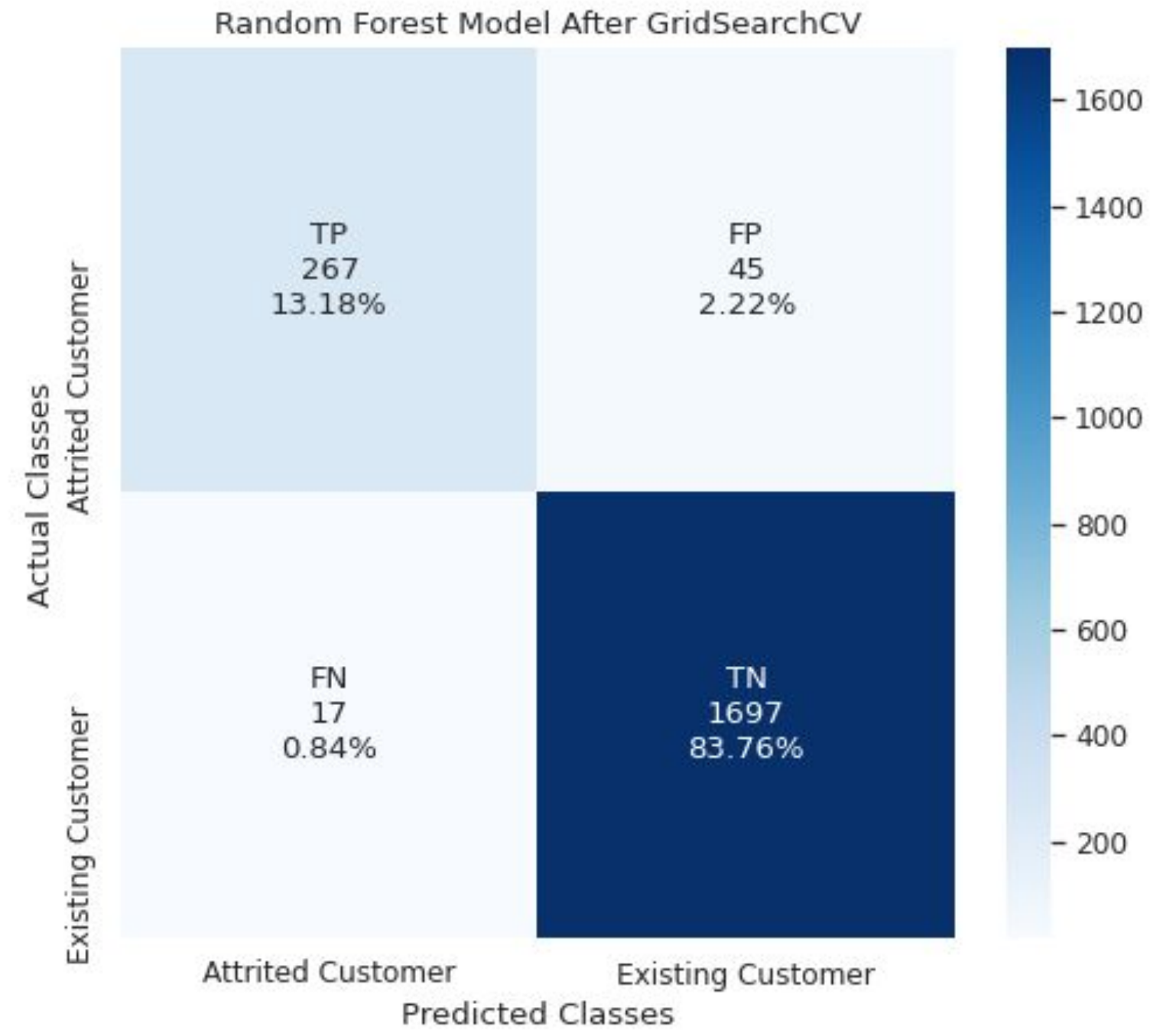
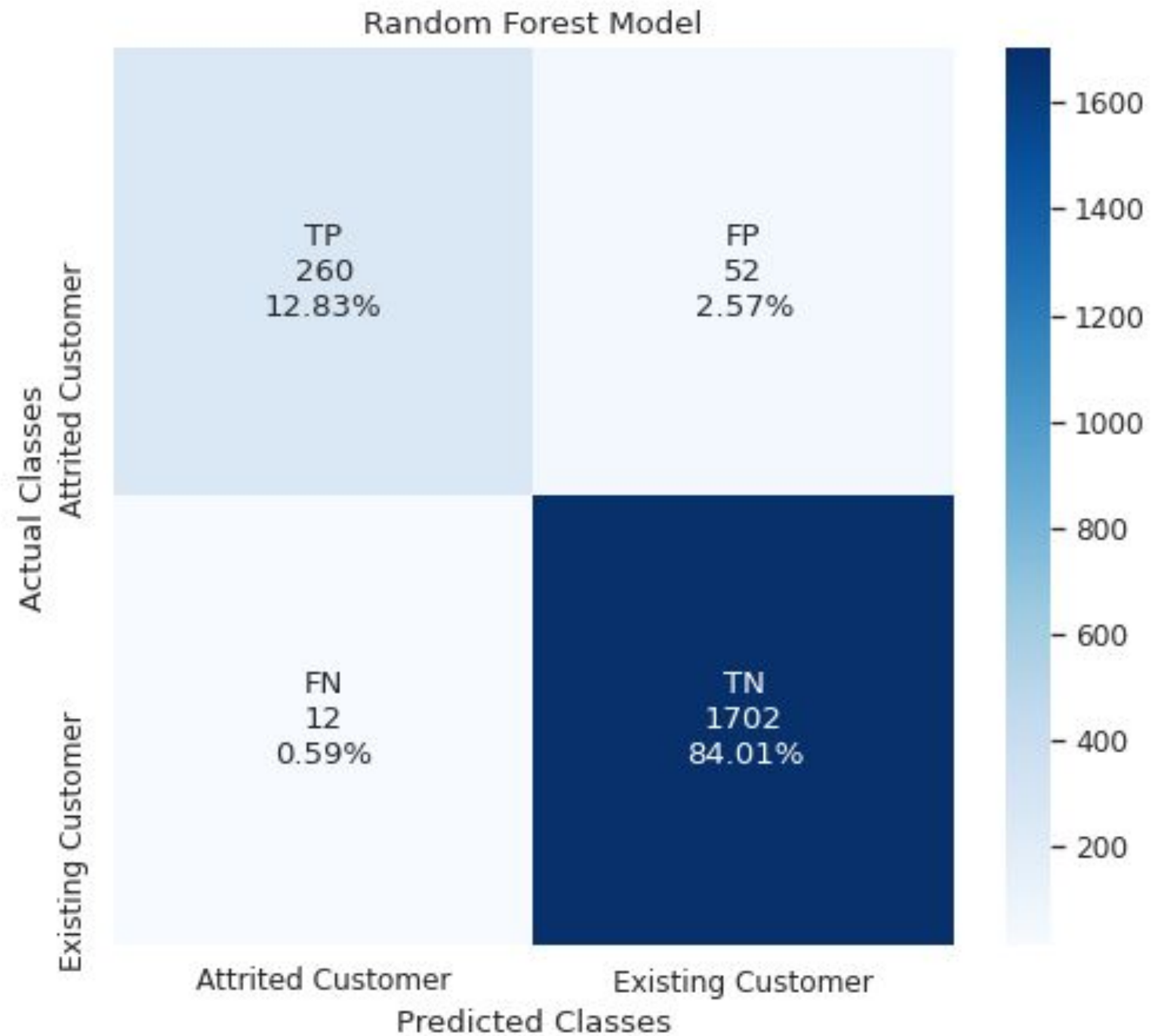


## Precision

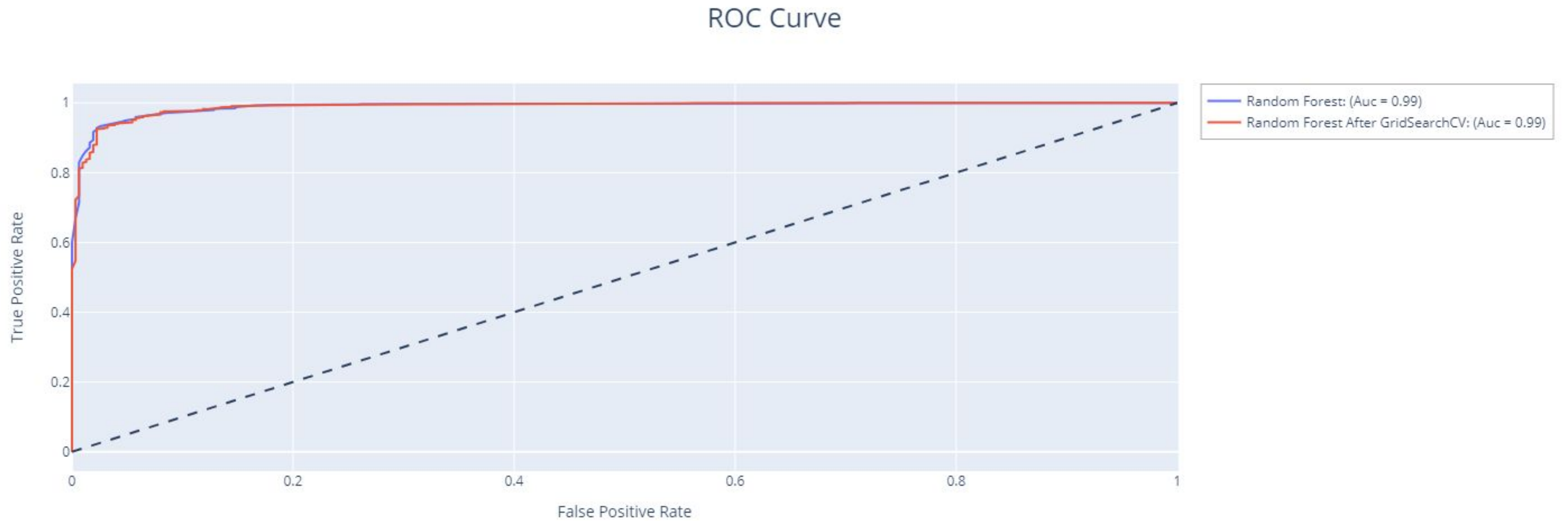




# Confusion Matrix



# Roc Curve



**Thank you**  
**Any Questions?**

