# Bank Churn Prediction

宋怡葶

# Table of contents

# Problem Introduction

- **Problem**: Increasing churning rate for their credit card services
- **Objective**:  Predict clients that are leaving

# Data Source

Kaggle Churn Modelling Dataset

| Attribute | Data Type | Description |
|---|---|---|
| Attrition_Flag | object | Internal event (customer activity) variable - if the account is closed then 1 else 0 |
| Customer_Age | int64 | Demographic variable - Customer's Age in Years |
| Card_Category | object | Product Variable - Type of Card (Blue, Silver, Gold, Platinum) |
| Months_on_book | int64 | Period of relationship with bank |
| Total_Trans_Amt | int64 | Total Transaction Amount (Last 12 months) |

# Features

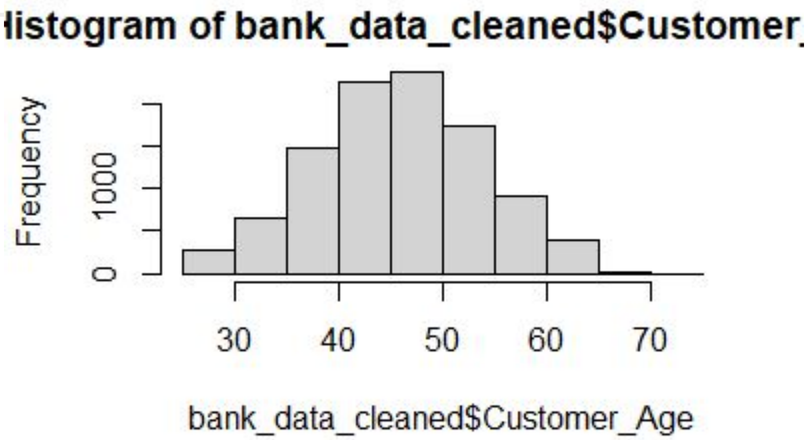
## Numerical Data(Discrete)

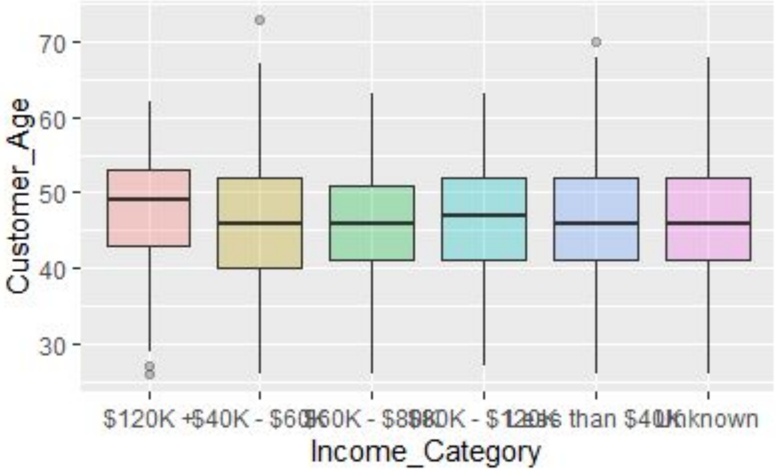
"Card_Category","Total_Relationship_Count","Customer_Age_cat","Months_on_book_cat"

## Numerical Data(Continuous)

"Months_Inactive_12_mon","Contacts_Count_12_mon","Credit_Limit","Total_Revolving_Bal","Avg_Open_To_Buy", "Total_Amt_Chng_Q4_Q1","Total_Trans_Amt","Total_Trans_Ct","Total_Ct_Chng_Q4_Q1"

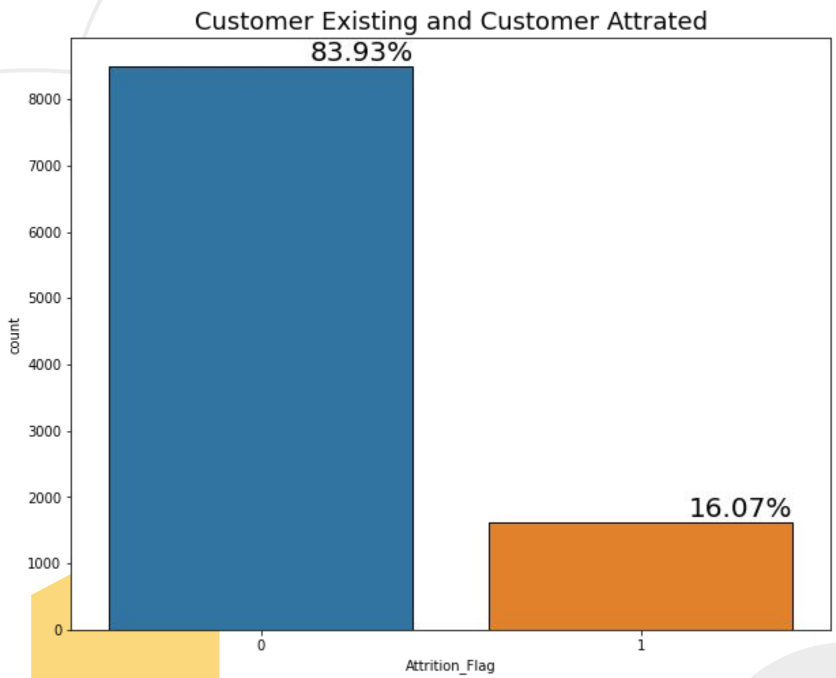
# Exploratory Data Analysis



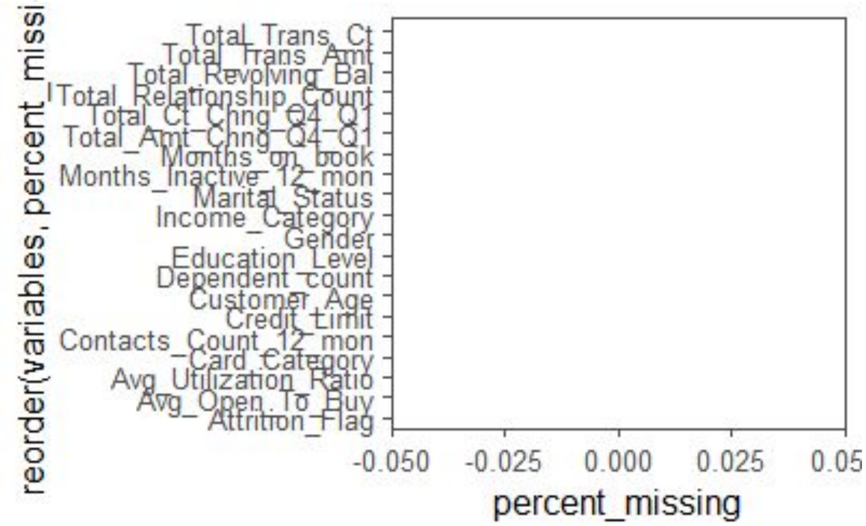

→ Numerical Variable - Age

# Exploratory Data Analysis



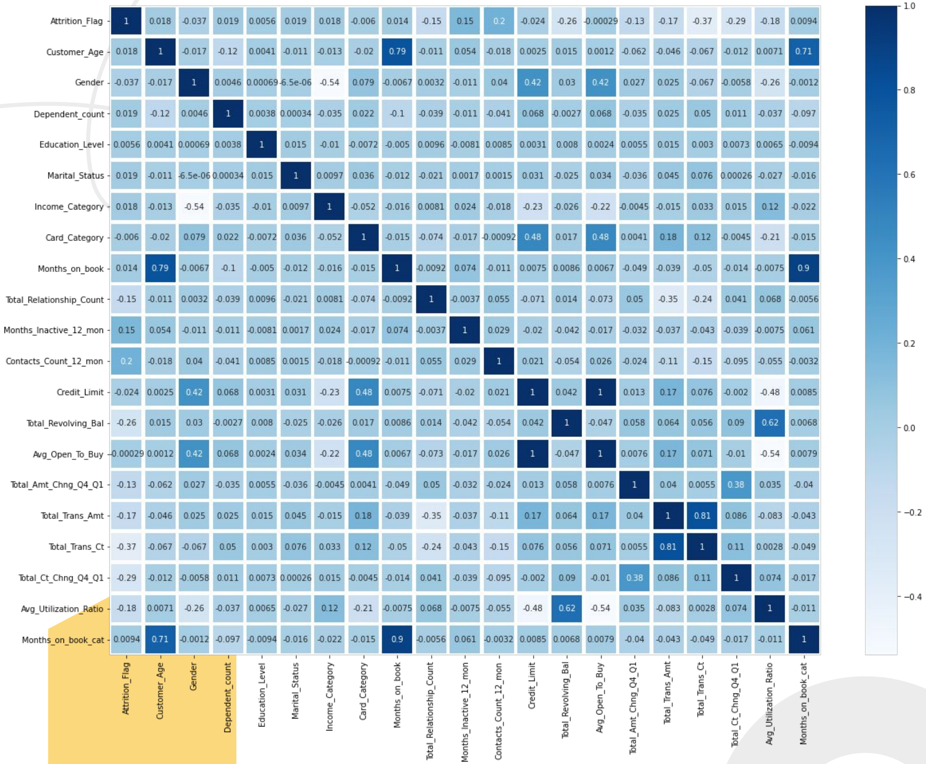

- ➔ Imbalance problem occurred in target variable

# Exploratory Data Analysis




→ The dataset has no missing value

# Exploratory Data Analysis




→ Variables are not highly correlated with other features.

# Exploratory Data Analysis - Outlier

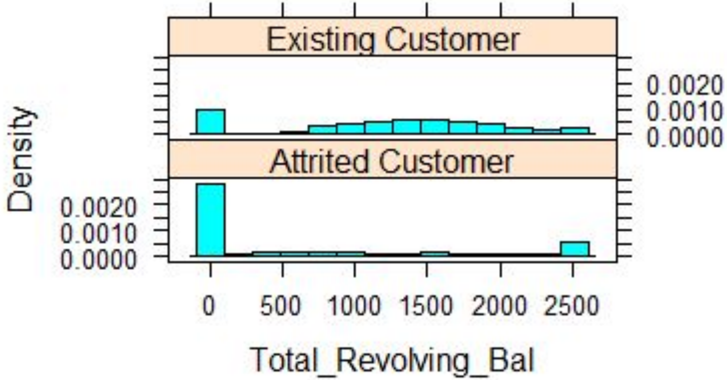
```
Outlier Customer_Age
IQR: 11.0
Q1: 41.0
Lower_Limit: 24.5
median: 46.0
Q3: 52.0
Upper_Limit: 68.5
```

| | Attrition_Flag | Customer_Age | Gender | Dependent_count | Education_Level | Marital_Status |
|---|---|---|---|---|---|---|
| **251** | 0 | 73 | 1 | 0 | 3 | 1 |
| **254** | 0 | 70 | 1 | 0 | 3 | 1 |
| **11** | 0 | 65 | 1 | 1 | 6 | 1 |
| **18** | 0 | 61 | 1 | 1 | 3 | 1 |
| **27** | 0 | 63 | 1 | 1 | 6 | 1 |

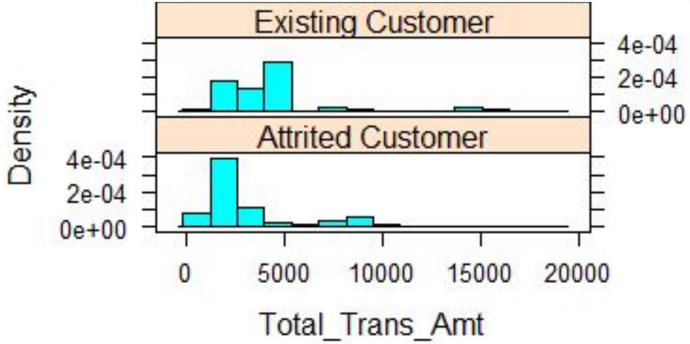
➔ Checking the presence of Outliers with IQR method

- The outliers will not be removed.

# Exploratory Data Analysis





# Data Preprocessing

# Methodology

methodology

- Label Encoding
- Numerical to Categorical
- PCA
- SMOTE

# Label Encoding

- **Encoding the following variables:**

  'Gender', 'Education_Level', 'Marital_Status', 'Income_Category', 'Card_Category'

- **For instance:**

| Education_Level | Marital_Status | Income_Category | Card_Category |
|---|---|---|---|
| 3 | 1 | 2 | 0 |
| 2 | 2 | 4 | 0 |
| 2 | 1 | 3 | 0 |
| 3 | 3 | 4 | 0 |
| 5 | 1 | 2 | 0 |

# Numerical to Categorical

- **Converting the following variables:**

  'Customer_Age' & 'Months_on_book''

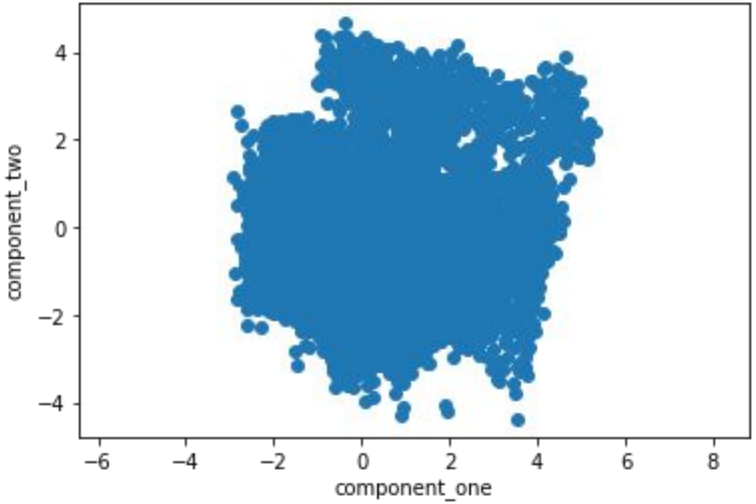
- For instance:

```
age_Conditions = [
        (data["Customer_Age"] < 25),
        (data["Customer_Age"] >= 25) & (data["Customer_Age"] < 35),
        (data["Customer_Age"] >= 35) & (data["Customer_Age"] <45),
        (data["Customer_Age"] < 55) & (data["Customer_Age"] >=45),
        (data["Customer_Age"] >= 55) & (data["Customer_Age"] < 65),
        (data["Customer_Age"] >=65)

]
age_Categories = [0,1,2,3,4,5]
data['Customer_Age_cat'] = np.select(age_Conditions, age_Categories)

print(data['Customer_Age_cat'])
```

# Scaling the dataset + PCA

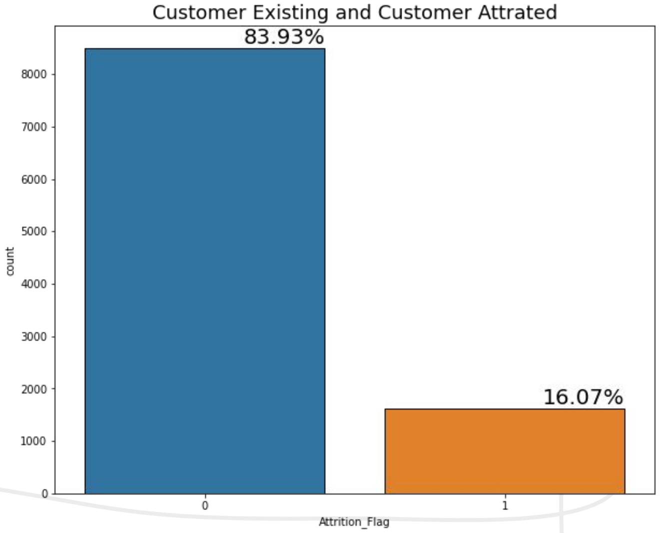



**Steps:**

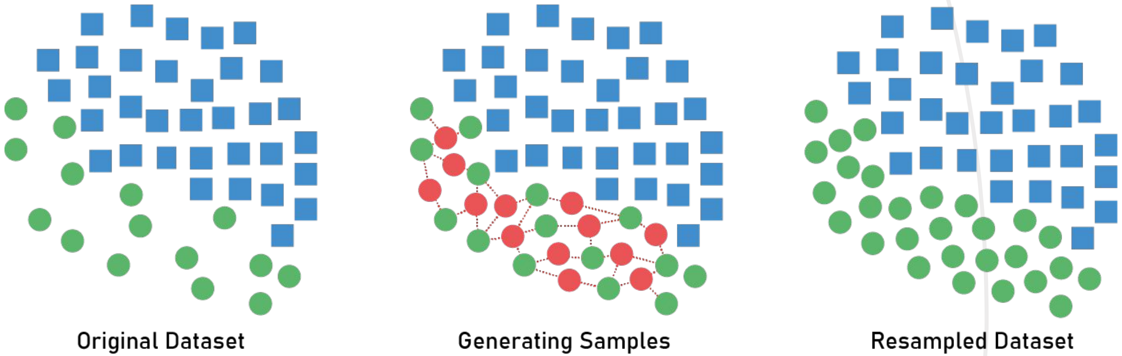
- Splitting into two datasets: continuous & discrete
- Scaling the dataset with continuous variables
- Reducing dimensions to 2D
- Merging the dataset

# Upsampling with SMOTE

**Problem: The target variable is imbalance.**

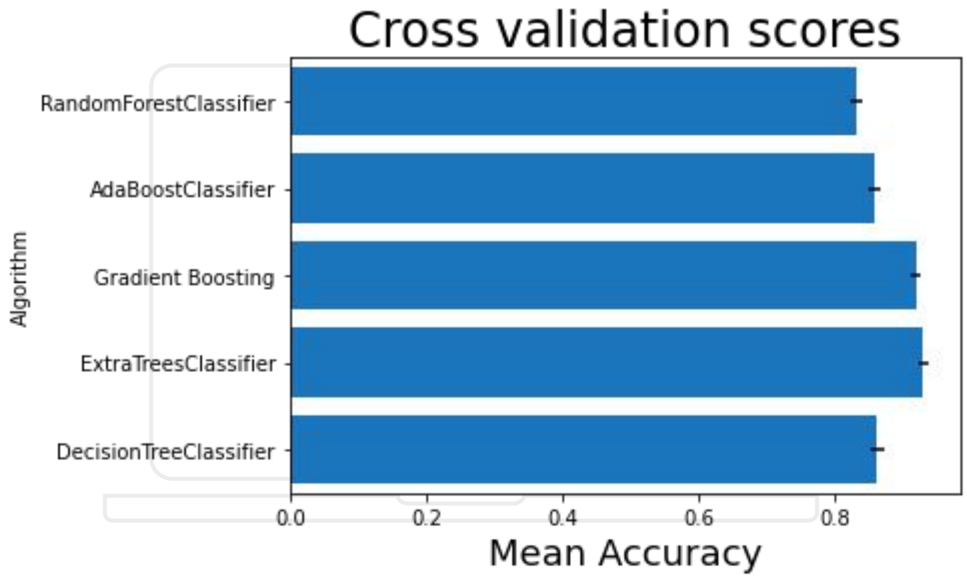



**Solution: Deploy SMOTE to “unsample” dataset**



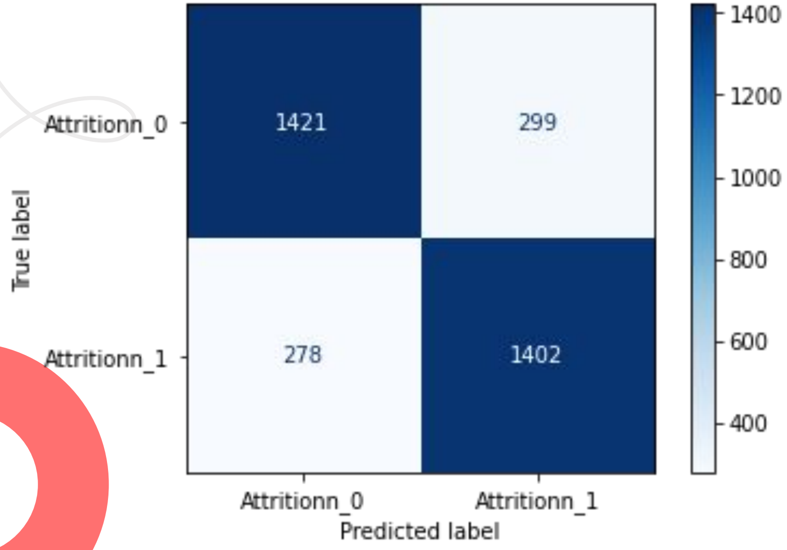
# Analysis

# Scores of different models



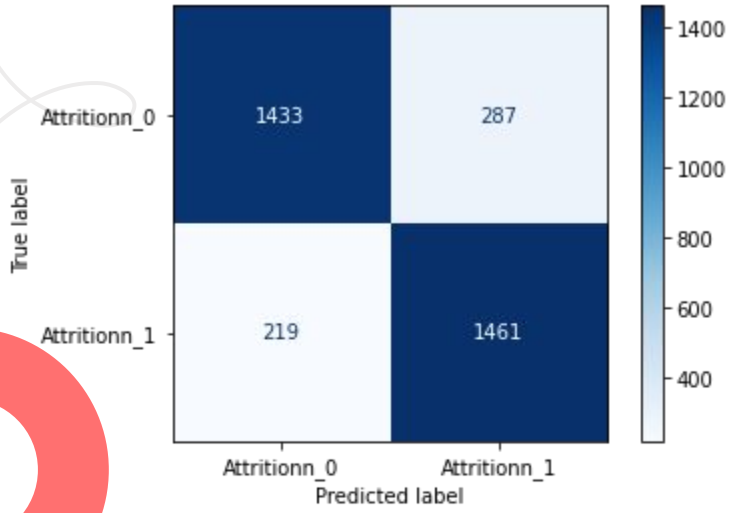

➔ Extra Tree Classifier has the highest score

# Confusion Matrix - AdaBoost Classifier



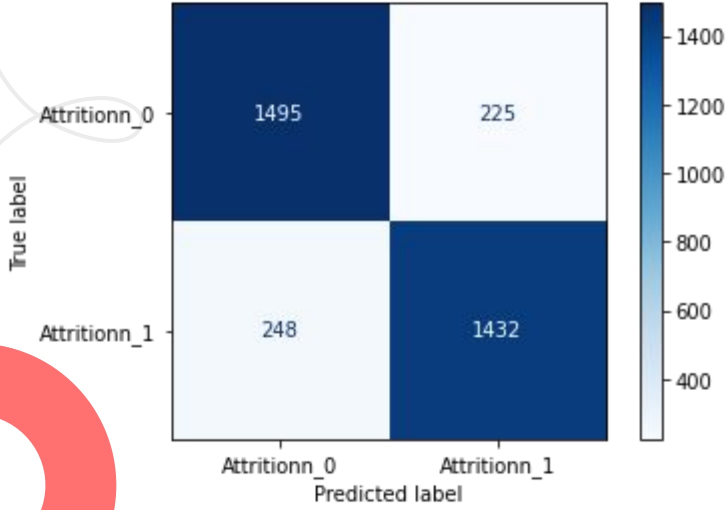

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.83 | 0.83 | 1720 |
| 1 | 0.82 | 0.83 | 0.83 | 1680 |
| accuracy | | | 0.83 | 3400 |

# Confusion Matrix - Decision Tree Classifier



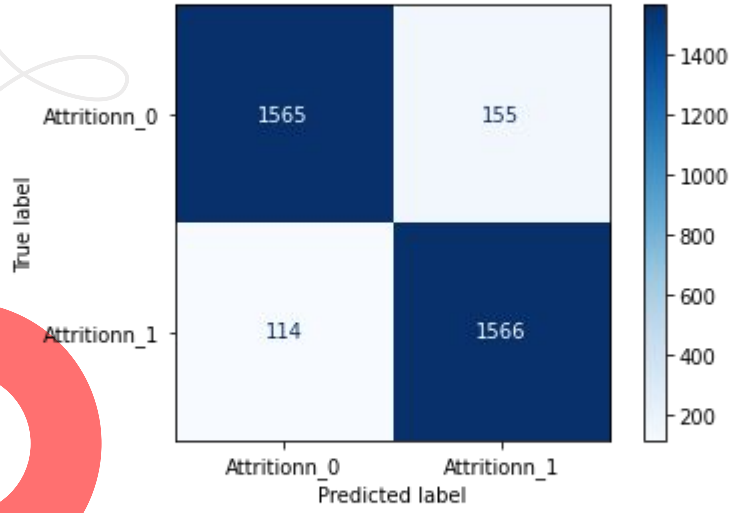

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.83 | 0.85 | 1720 |
| 1 | 0.84 | 0.87 | 0.85 | 1680 |
| accuracy | | | 0.85 | 3400 |

# Confusion Matrix - Gradient Boosting



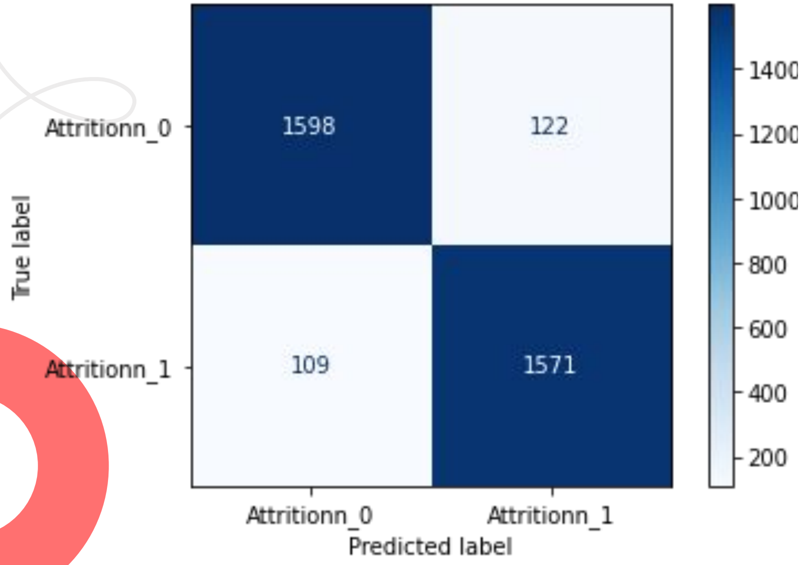

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.87 | 0.86 | 1720 |
| 1 | 0.86 | 0.85 | 0.86 | 1680 |
| accuracy | | | 0.86 | 3400 |

# Confusion Matrix - Random Forest Classifier



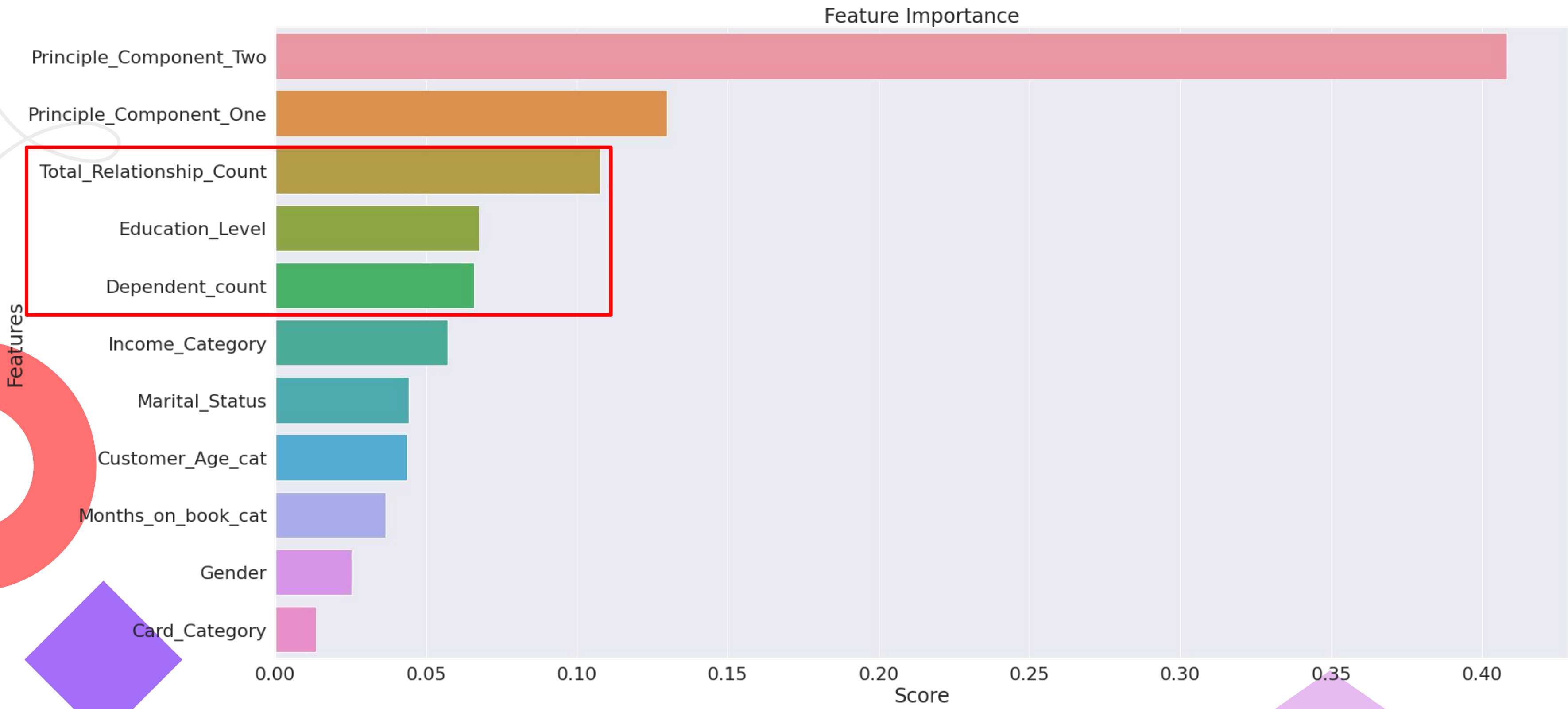

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.91 | 0.92 | 1720 |
| 1 | 0.91 | 0.93 | 0.92 | 1680 |
| accuracy | | | 0.92 | 3400 |

# Confusion Matrix - Extra Tree Classifier

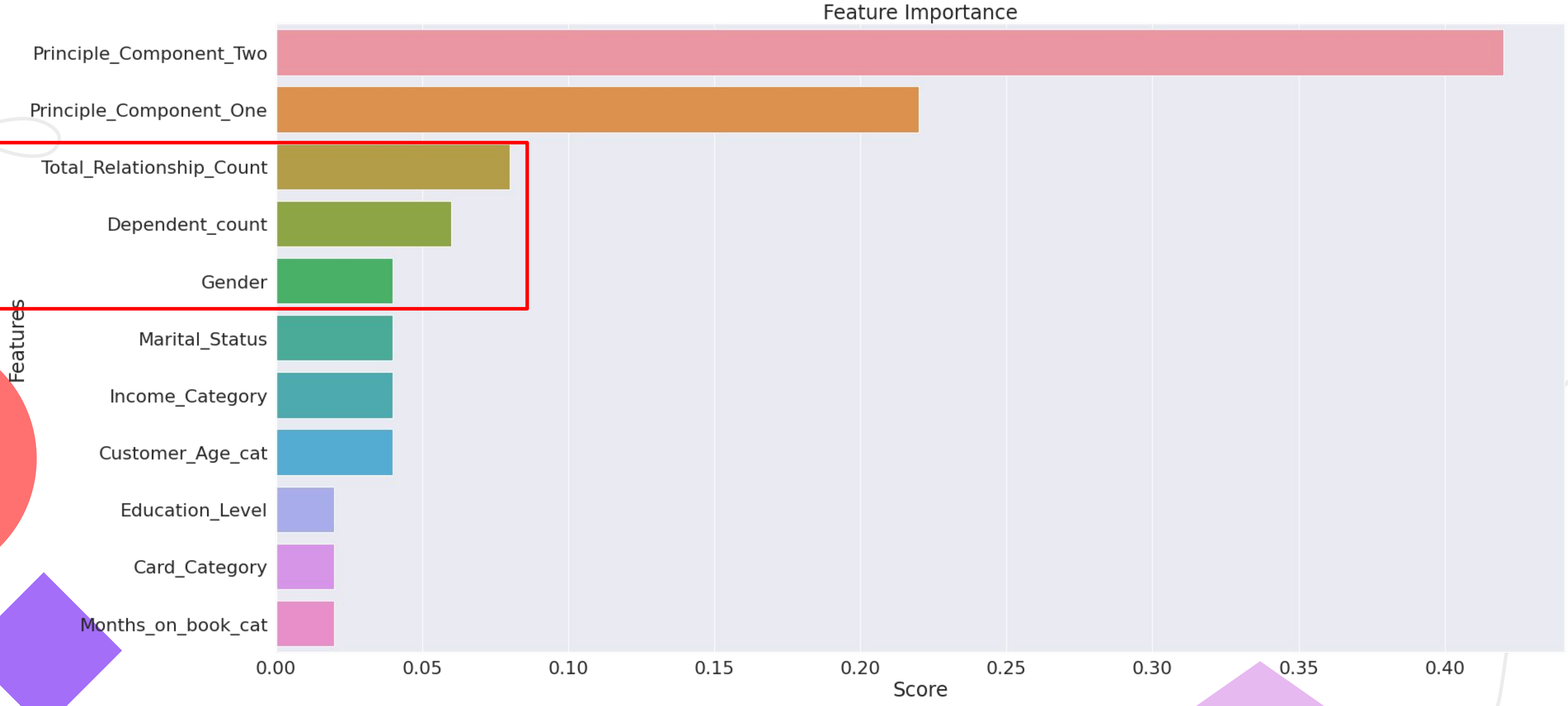



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.93 | 0.93 | 1720 |
| 1 | 0.93 | 0.94 | 0.93 | 1680 |
| accuracy | | | 0.93 | 3400 |

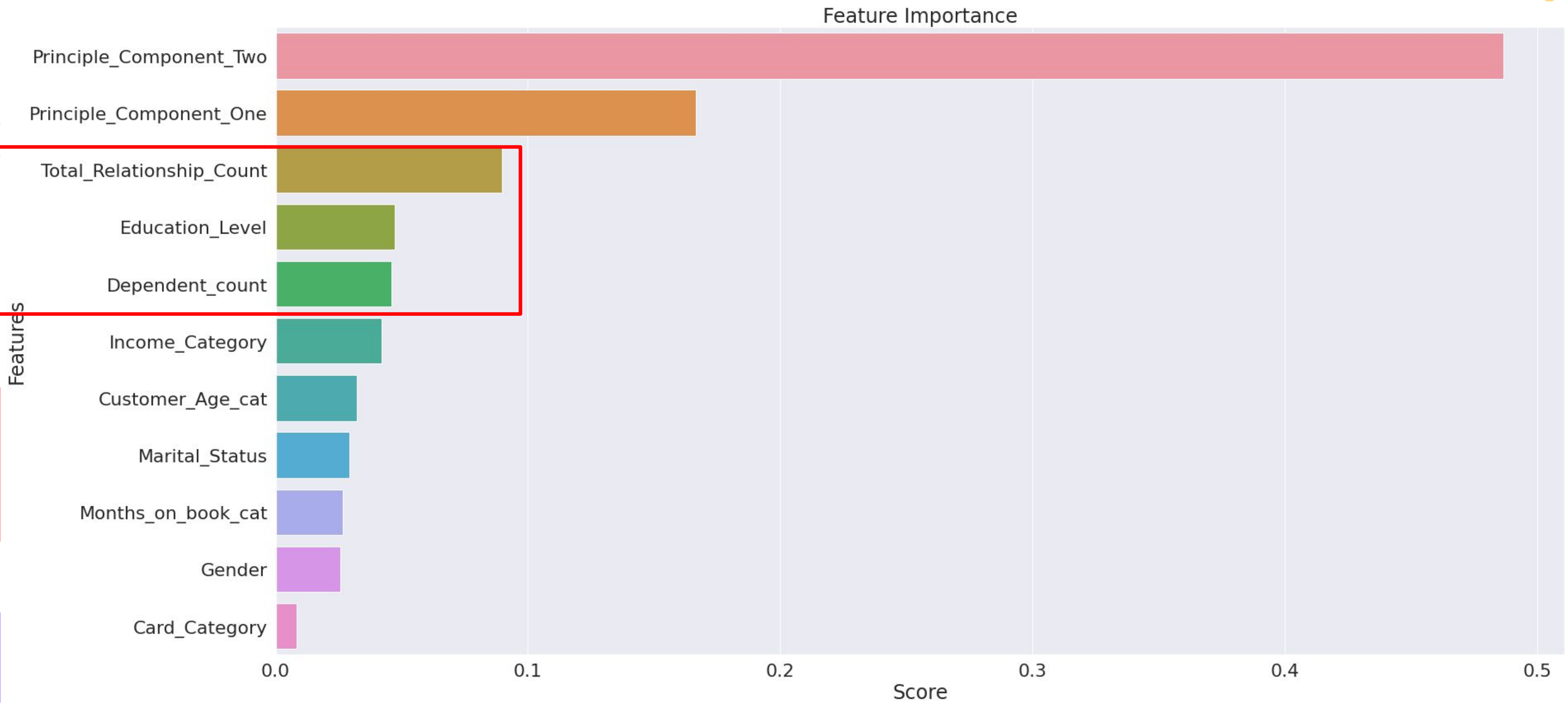
# Feature Importance - Extra tree classifier



# Feature Importance - AdaBoost classifier



# Feature Importance - Random Forest classifier



# Conclusion

# Conclusions

## Dataset Problem

Solving imbalance problem of the dataset can help increasing accuracy rate of models

## Model Selection

Extra tree classifier perform better than all the other algorithms in terms of accuracy rate and recall rate.

## Important Features

“Total relationship count”, “education level” and “dependent count” are important features.

## Marketing Implications

Knowing which customers will leave allows us to offer promotions to keep them from leaving or switching banks.

# Thank you for your attention