Credit Card Customer Churn

Raj Shah (A20524266)

Akshay Singh (A20498211)

Ravi Teja Batchala (A20512513)

Contents

Introduction

It costs on average around 200 USD to acquire a credit card customer (up to and beyond 1,000 USD if they are affluent cards like “MasterCard World Elite” and “Visa Infinite”). The Apple Card doesn’t need any affiliates or marketing, yet analysts say that this new card has a customer acquisition cost of 350 USD and will take several years for Goldman Sachs to turn a profit on it.

That being said, it actually takes banks a few years to recover the acquisition investment of that customer and that's the main reason why they're motivated to predict which users are the most likely to churn, in order to proactively and not reactively retain the client.

Problem Statement

- Customer retention is critical for a good marketing and a customer relationship management strategy. The prevention of customer churn through customer retention is a core issue of Customer Relationship Management.
- Here, an analysis is done on purchasing behavior of bank customers.
- A detailed scheme is worked out to convert raw customer data into meaningful and useful data that suits the buying behavior, and in turn, converts this meaningful data into knowledge for which predictive data mining techniques are adopted.

Data Description

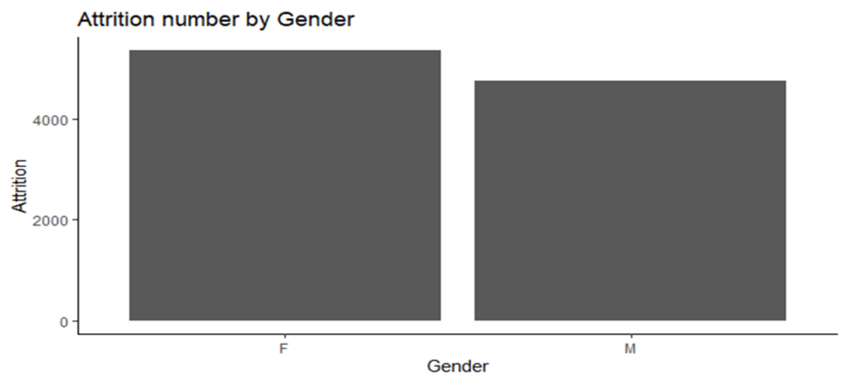
The dataset consists of records of 10,127 bank customers and 20 columns describing various features-

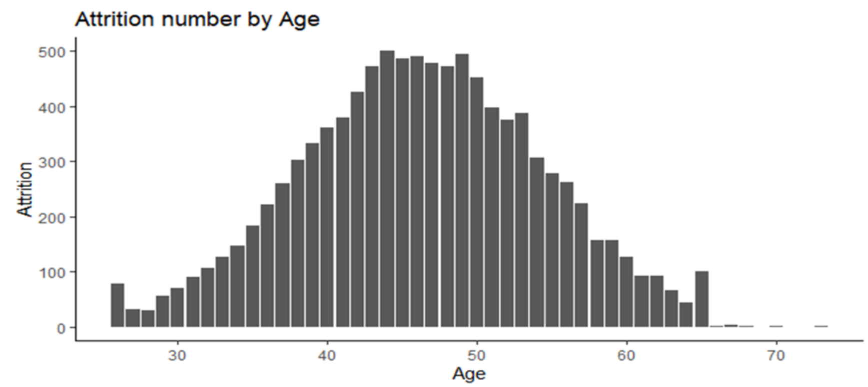
- **Clientnum** - Client number. Unique identifier for the customer holding the account
- **Attrition_Flag** - Internal event (customer activity) variable
- **Customer_Age** - Customer's Age in Years
- **Gender** - M=Male, F=Female
- **Dependent_count** - Number of people dependents
- **Education_Level** - Educational Qualification of the account holder (example: high school, college graduate, etc.)
- **Marital_Status** - Married, Single, Unknown
- **Income_Category** - Annual Income Category of the account holder (< 40K, 40K - 60K, 60K - 80K, 80K-120K, > 120K, Unknown)
- **Card_Category** - Type of Card (Blue, Silver, Gold, Platinum)
- **Months_on_book** Months on book (Time of Relationship)

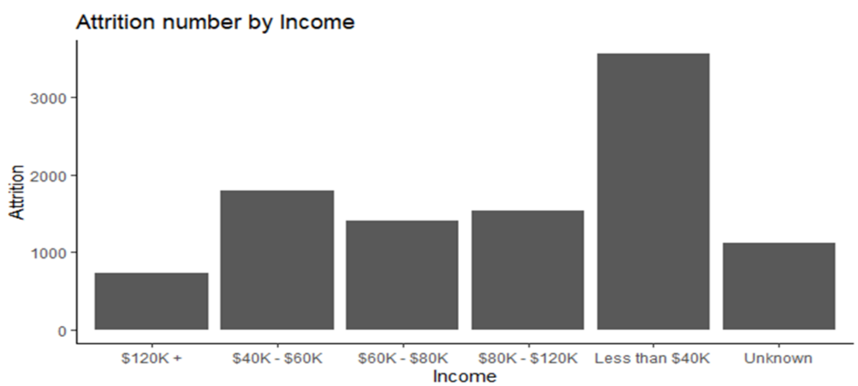
Data Description

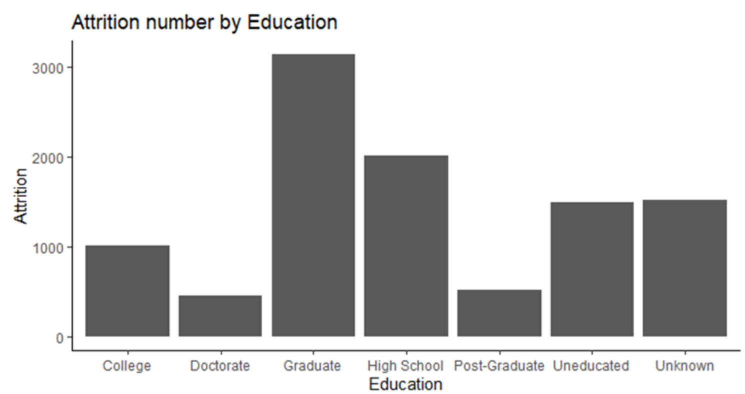
- **Total_Relationship_Count** - Total no. of products held by the customer
- **Months_Inactive_12_mon** - No. of months inactive in the last 12 months
- **Contacts_Count_12_mon** - No. of Contacts in the last 12 months
- **Credit_Limit** - Credit Limit on the Credit Card
- **Total_Revolving_Bal** - Total Revolving Balance on the Credit Card
- **Avg_Open_To_Buy** - Open to Buy Credit Line (Average of last 12 months)
- **Total_Amt_Chng_Q4_Q1** - Change in Transaction Amount (Q4 over Q1)
- **Total_Trans_Amt Num** - Total Transaction Amount (Last 12 months)
- **Total_Trans_Ct Num** - Total Transaction Count (Last 12 months)
- **Total_Ct_Chng_Q4_Q1 -** Change in Transaction Count (Q4 over Q1)
- **Avg_Utilization_Ratio** - Average Card Utilization Ratio

Exploratory Data Analysis

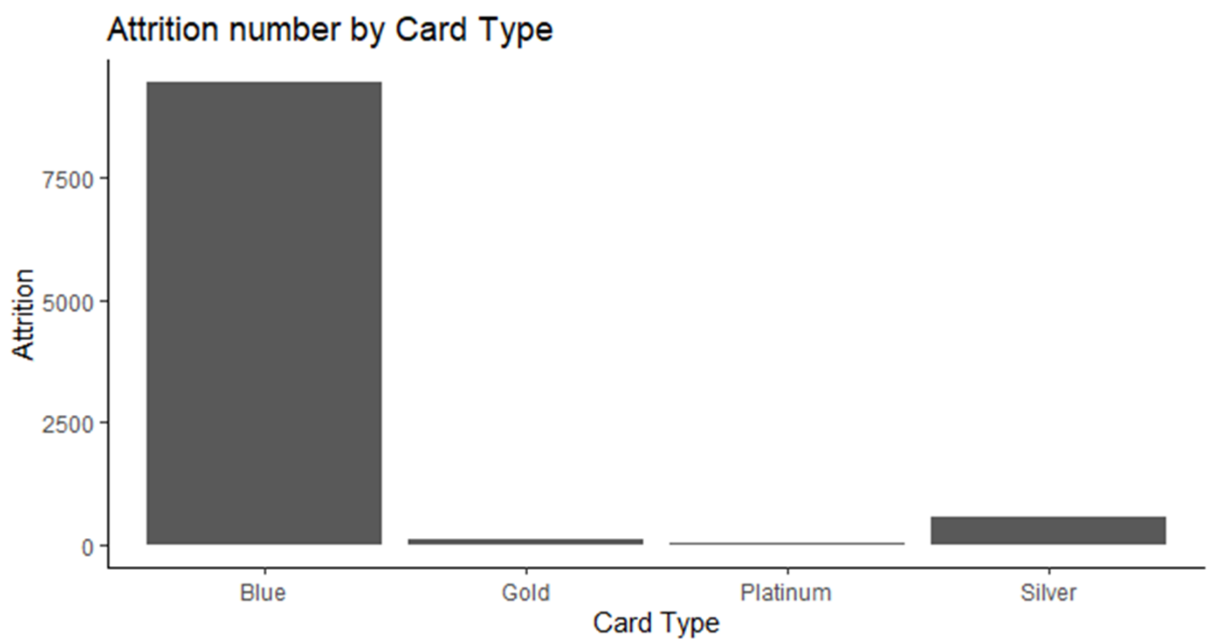




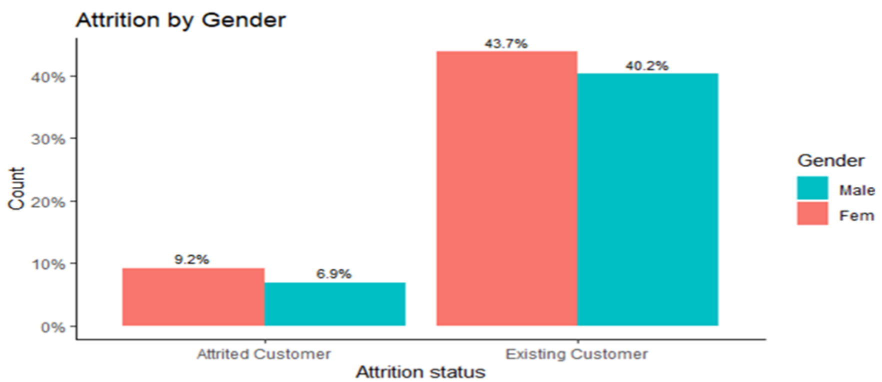


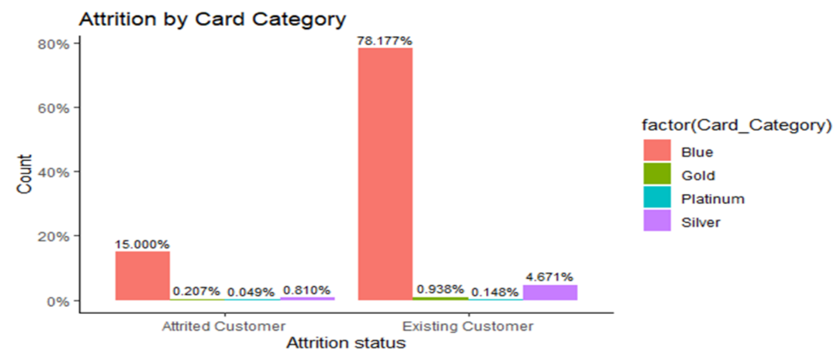


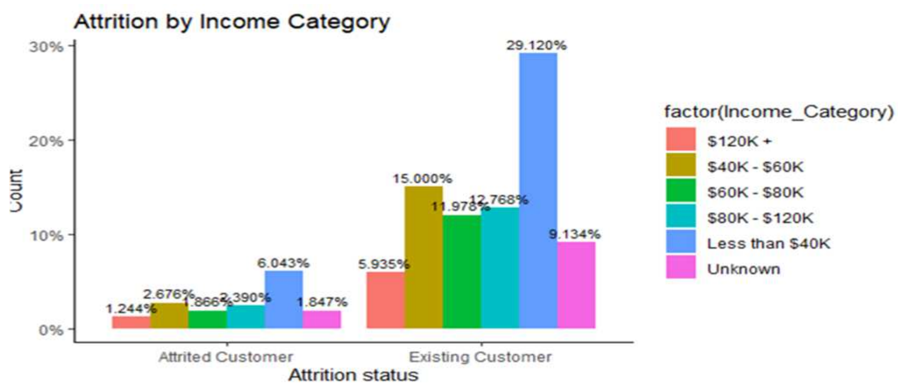
Exploratory Data Analysis

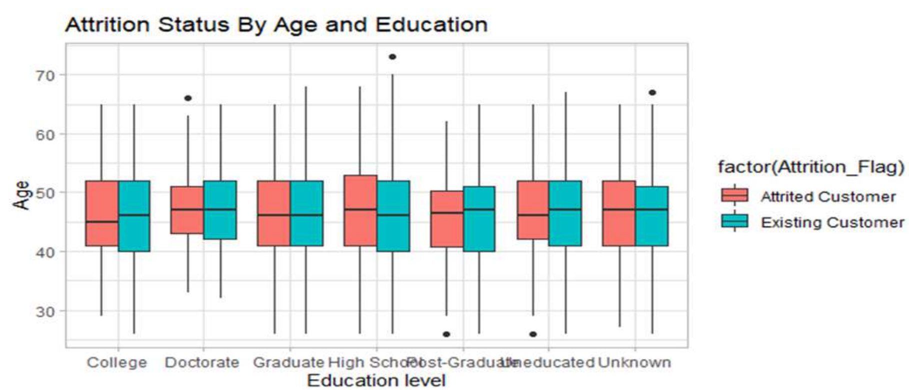


Exploratory Data Analysis

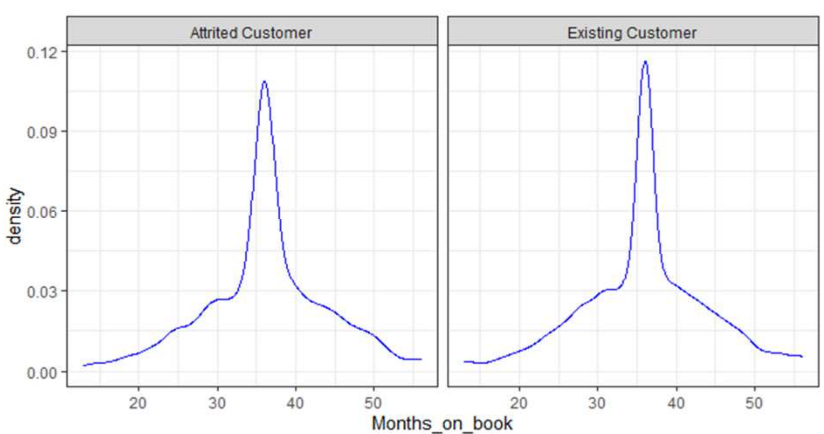


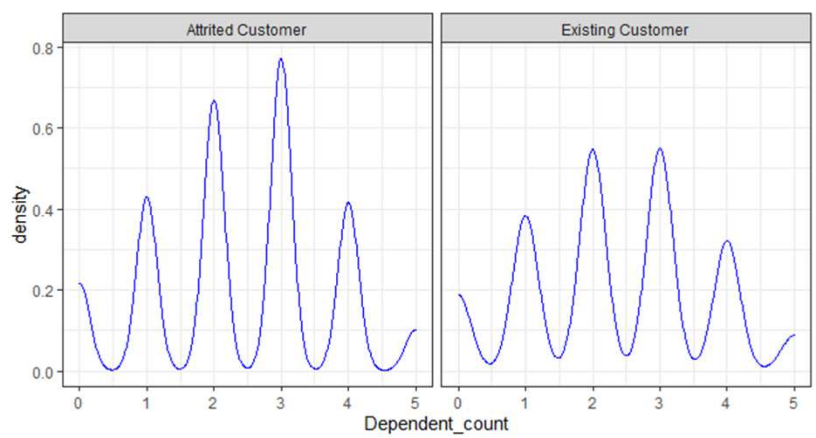






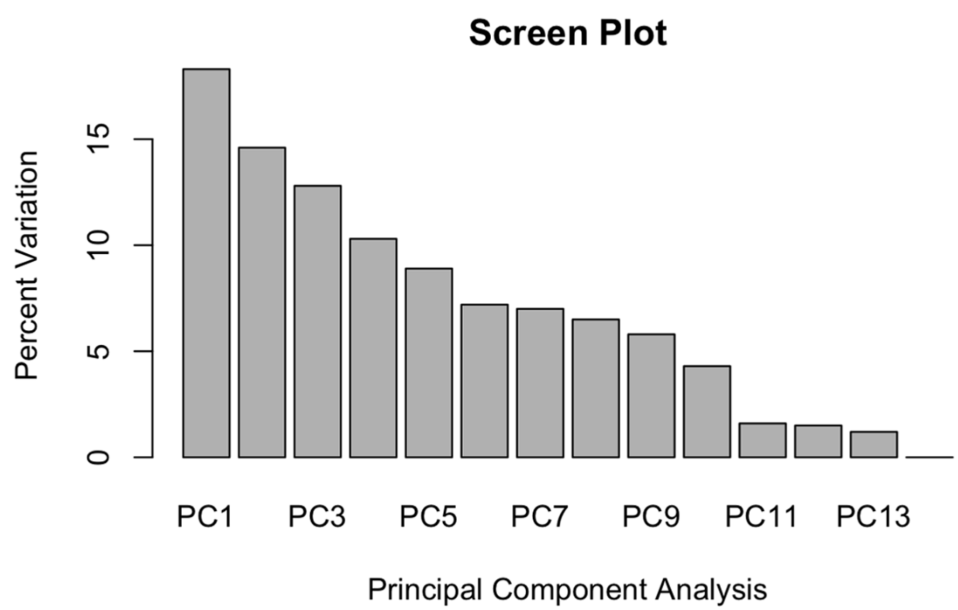
Exploratory Data Analysis





Principal Component Analysis

To extract the principal components (that can be used to train the models), we perform Principal Component Analysis on the dataset containing 14 feature data in which 7 features came out to account for most variation in data.



ML Models

Random Forest

```
Confusion matrix:
                  Attrited Customer Existing Customer class.error
Attrited Customer              1086               216  0.16589862
Existing Customer                82              6718  0.01205882
Confusion Matrix and Statistics

                   Reference
Prediction          Attrited Customer Existing Customer
  Attrited Customer               265                20
  Existing Customer                60              1680

               Accuracy : 0.9605
                 95% CI : (0.9511, 0.9686)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8457

 Mcnemar's Test P-Value : 1.299e-05

            Sensitivity : 0.8154
            Specificity : 0.9882
         Pos Pred Value : 0.9298
         Neg Pred Value : 0.9655
             Prevalence : 0.1605
         Detection Rate : 0.1309
   Detection Prevalence : 0.1407
      Balanced Accuracy : 0.9018

       'Positive' Class : Attrited Customer
```

Model with Non-PCA data

```
Confusion matrix:
                  Attrited Customer Existing Customer class.error
Attrited Customer               675               627  0.48156682
Existing Customer               109              6691  0.01602941
Confusion Matrix and Statistics

                   Reference
Prediction          Attrited Customer Existing Customer
  Attrited Customer               166                18
  Existing Customer               159              1682

               Accuracy : 0.9126
                 95% CI : (0.8994, 0.9245)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6066

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.51077
            Specificity : 0.98941
         Pos Pred Value : 0.90217
         Neg Pred Value : 0.91363
             Prevalence : 0.16049
         Detection Rate : 0.08198
   Detection Prevalence : 0.09086
      Balanced Accuracy : 0.75009

       'Positive' Class : Attrited Customer
```

Model with PCA data

ML Models

(SVM) Support Vector Machine

```
Confusion Matrix and Statistics

                    Reference
Prediction           Attrited Customer Existing Customer
  Attrited Customer                 62                 5
  Existing Customer                263              1695

               Accuracy : 0.8677
                 95% CI : (0.8521, 0.8821)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : 0.0002322

                  Kappa : 0.2766

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.19077
            Specificity : 0.99706
         Pos Pred Value : 0.92537
         Neg Pred Value : 0.86568
             Prevalence : 0.16049
         Detection Rate : 0.03062
   Detection Prevalence : 0.03309
      Balanced Accuracy : 0.59391

       'Positive' Class : Attrited Customer
```

Model with Non-PCA data

```
Confusion Matrix and Statistics

                    Reference
Prediction           Attrited Customer Existing Customer
  Attrited Customer                 94                12
  Existing Customer                231              1688

               Accuracy : 0.88
                 95% CI : (0.865, 0.8938)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : 1.563e-07

                  Kappa : 0.3879

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.28923
            Specificity : 0.99294
         Pos Pred Value : 0.88679
         Neg Pred Value : 0.87962
             Prevalence : 0.16049
         Detection Rate : 0.04642
   Detection Prevalence : 0.05235
      Balanced Accuracy : 0.64109

       'Positive' Class : Attrited Customer
```

Model with PCA data

ML Models

Naive Bayes

```
Confusion Matrix and Statistics

                    Reference
Prediction           Attrited Customer Existing Customer
  Attrited Customer                203               127
  Existing Customer                122              1573

               Accuracy : 0.877
                 95% CI : (0.8619, 0.891)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : 1.156e-06

                  Kappa : 0.5465

 Mcnemar's Test P-Value : 0.7999

            Sensitivity : 0.6246
            Specificity : 0.9253
         Pos Pred Value : 0.6152
         Neg Pred Value : 0.9280
             Prevalence : 0.1605
         Detection Rate : 0.1002
   Detection Prevalence : 0.1630
      Balanced Accuracy : 0.7750

       'Positive' Class : Attrited Customer
```

Model with Non-PCA data

```
Confusion Matrix and Statistics

                   Reference
Prediction          Attrited Customer Existing Customer
  Attrited Customer               146                28
  Existing Customer               179              1672

               Accuracy : 0.8978
                 95% CI : (0.8838, 0.9106)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : 2.646e-14

                  Kappa : 0.5329

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.44923
            Specificity : 0.98353
         Pos Pred Value : 0.83908
         Neg Pred Value : 0.90330
             Prevalence : 0.16049
         Detection Rate : 0.07210
   Detection Prevalence : 0.08593
      Balanced Accuracy : 0.71638

       'Positive' Class : Attrited Customer
```

Model with PCA data

ML Models

Decision Tree

```
Confusion Matrix and Statistics

                   Reference
Prediction          Attrited Customer Existing Customer
  Attrited Customer               246                59
  Existing Customer                79              1641

               Accuracy : 0.9319
                 95% CI : (0.92, 0.9424)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7406

 Mcnemar's Test P-Value : 0.1058

            Sensitivity : 0.7569
            Specificity : 0.9653
         Pos Pred Value : 0.8066
         Neg Pred Value : 0.9541
             Prevalence : 0.1605
         Detection Rate : 0.1215
   Detection Prevalence : 0.1506
      Balanced Accuracy : 0.8611

       'Positive' Class : Attrited Customer
```

Model with Non-PCA data

```
Confusion Matrix and Statistics

                   Reference
Prediction          Attrited Customer Existing Customer
  Attrited Customer               171                74
  Existing Customer               154              1626

               Accuracy : 0.8874
                 95% CI : (0.8728, 0.9009)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : 5.090e-10

                  Kappa : 0.536

 Mcnemar's Test P-Value : 1.678e-07

            Sensitivity : 0.52615
            Specificity : 0.95647
         Pos Pred Value : 0.69796
         Neg Pred Value : 0.91348
             Prevalence : 0.16049
         Detection Rate : 0.08444
   Detection Prevalence : 0.12099
      Balanced Accuracy : 0.74131

       'Positive' Class : Attrited Customer
```

Model with PCA data

ML Models

Logistic Regression

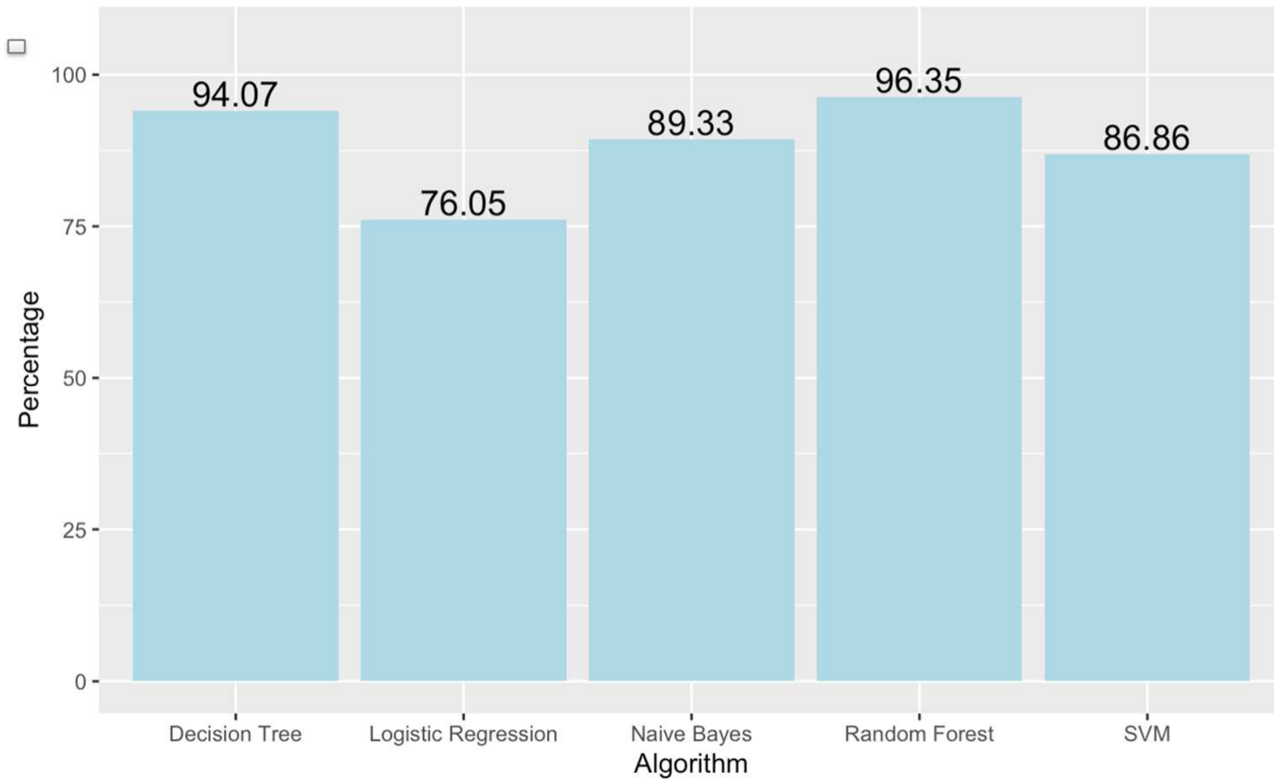
```
          target
y_pred    1    2
     1   44  204
     2  281 1496

               Accuracy : 0.7605
                 95% CI : (0.7413, 0.7789)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : 1.0000000

                  Kappa : 0.017

 Mcnemar's Test P-Value : 0.0005586

            Sensitivity : 0.13538
            Specificity : 0.88000
         Pos Pred Value : 0.17742
         Neg Pred Value : 0.84187
             Prevalence : 0.16049
         Detection Rate : 0.02173
   Detection Prevalence : 0.12247
      Balanced Accuracy : 0.50769

       'Positive' Class : 1
```

Model with Non-PCA data

```
Confusion Matrix and Statistics

          target
y_pred    1    2
     1  165   41
     2  160 1659

               Accuracy : 0.9007
                 95% CI : (0.8869, 0.9134)
    No Information Rate : 0.8395
    P-Value [Acc > NIR] : 1.038e-15

                  Kappa : 0.5676

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.50769
            Specificity : 0.97588
         Pos Pred Value : 0.80097
         Neg Pred Value : 0.91204
             Prevalence : 0.16049
         Detection Rate : 0.08148
   Detection Prevalence : 0.10173
      Balanced Accuracy : 0.74179

       'Positive' Class : 1
```

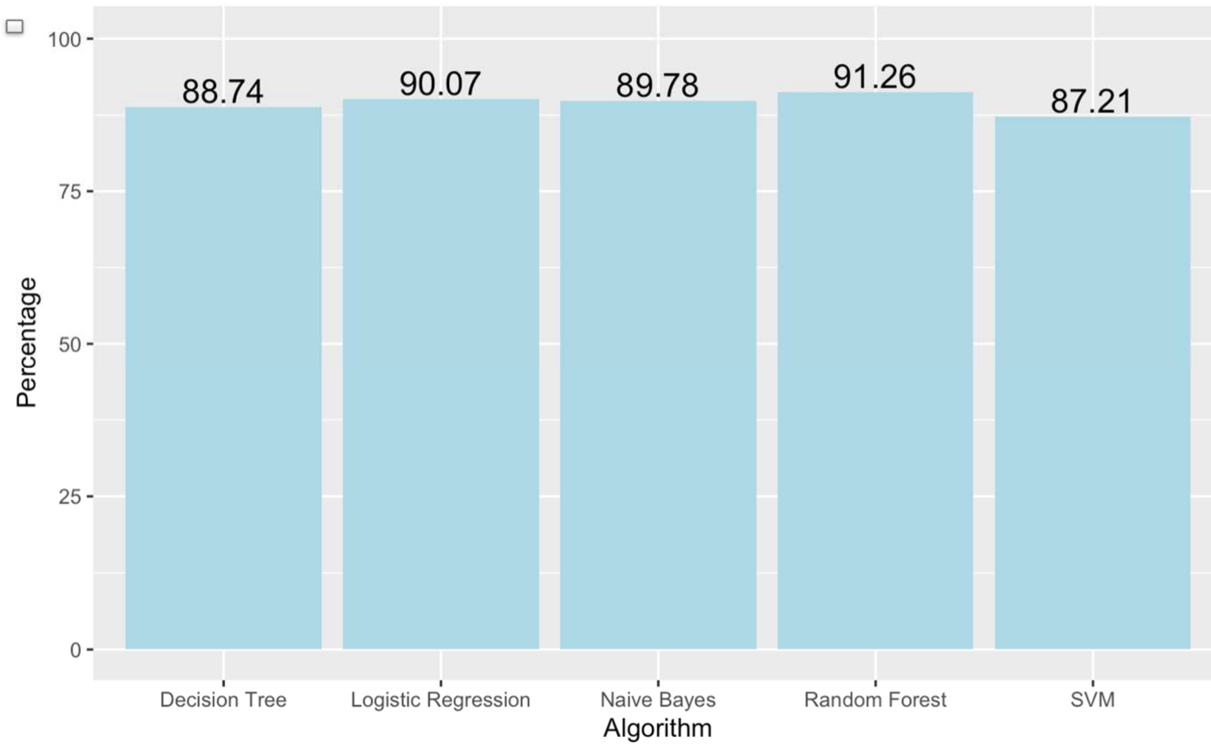
Model with PCA data

Comparison of ML Models



Models using Non-PCA data

Comparison of ML Models



Models using PCA data

Conclusion

- From the previous graphs, we can clearly conclude that the Random Forest Model have outperformed the remaining models, let it be in the regular data or the transformed data with reduced dimensions and principal components.
- Also, as we observe the graphs, we can see few models performed a bit better with the transformed PCA data instead of the regular ones and vice versa. Dataset is of 20 dimensions (features), which might not be considered large enough to perform dimentionality reduction or even PCA, but performing PCA altered the model accuracy to an extend. Accuracy of all the models before PCA differed a lot from each other, but after PCA all the models gave the same accuracy in the range of 90%. This shows that all the models were able to train and fit the data correctly and capture the variablitiy of the features, which was not the case when we considered all the features from the dataset.

Future Scope

- The data set is minimal, hence more data can be included.
- The machine learning model that are built, can be used in real life web applications.
- More complex algorithms like the deep learning models can be trained and used for the prediction.
- In few cases, the models can be fine-tuned according to the bank's requirement.

References

[1] Dataset -

https://www.kaggle.com/code/varunbarath/credit-card-customers-bank-churners/data

[2] Logistic Regression -

https://www.javatpoint.com/logistic-regression-in-machine-learning

[3] SVM -

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[4] Decision tree -

https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

THANK YOU