Bank Customer Churn Prediction

Milestone: Data Collection, Data Visualization, Data Exploration, and Data Processing

Group 28

Mukund Bankar

Ashok Thiruvengadam

857-654-4832 (Tel of Student 1)

213-294-6484 (Tel of Student 2)

bankar.m@northeastern.edu

thiruvengadam.a@northeastern.edu

Percentage of Effort contributed by Student 1: 50%

Percentage of Effort contributed by Student 2: 50%

Signature of Student 1: Mukund Bankar

Signature of Student 2: Ashok Thiruvengadam

Submission Date: 12-09-2022

Index

Bank Customer Churn Prediction

Problem Setting:

Customer churn, also known as customer attrition, is the phenomenon in which a customer leaves an organization. This customer attrition is likely to influence a bank's income for various reasons, including service, facilities, and assistance provided to consumers. When compared to client retention, the expense of marketing to gain new consumers is prohibitively expensive.

Problem Definition:

To grow a bank's business model, it is necessary to forecast the likelihood of client attrition. This project's primary goal is to create a supervised machine-learning model that will help in the classification of churn consumers.

Data Source:

The data was obtained from the well-known open-source repository Kaggle.

URL: https://www.kaggle.com/datasets/santoshd3/bank-customers

Data Description:

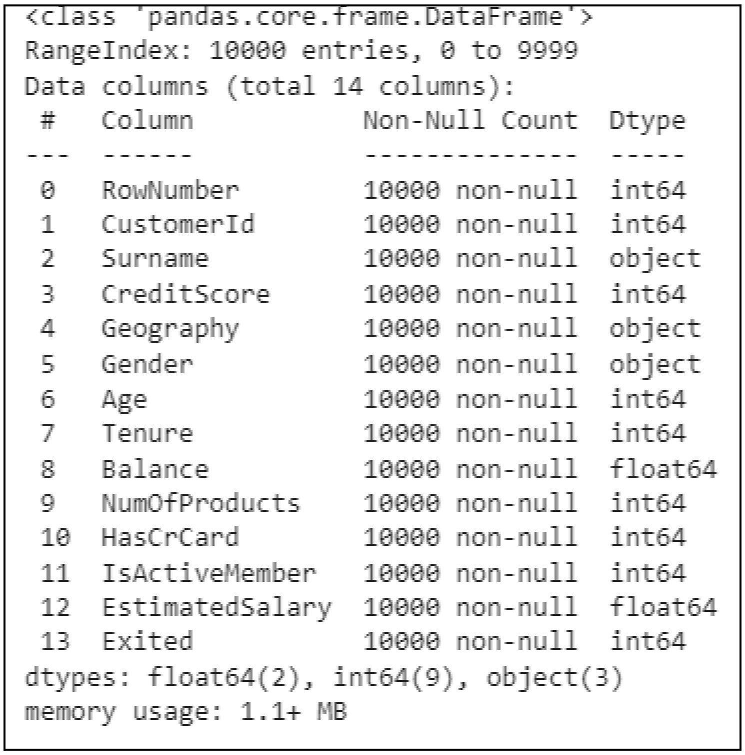
The data consists of 10000 Customers across the globe. This data has the following columns:

No	Feature Name
1	RowNumber
2	CustomerId
3	Surname
4	CreditScore
5	Geography
6	Gender
7	Age
8	Tenure
9	Balance

10	NumOfProducts
11	HasCrCard
12	IsActiveMember
13	EstimatedSalary
14	Exited

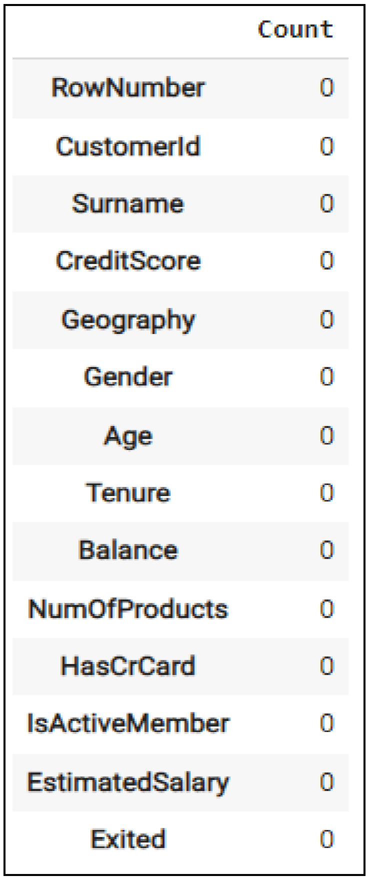
Data Exploration:

Data exploration is the initial phase in data analysis and is used to examine and display data in order to find insights from the beginning or to suggest regions or trends investigate further. Data is statistically and visually analyzed using Python packages. Finding the form of the table, correlation matrix, null values, and unique values are all part of the data exploration process.



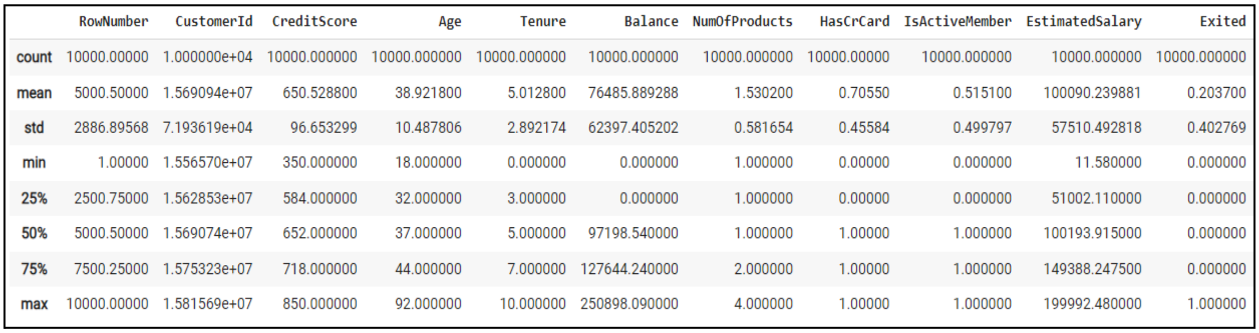
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   RowNumber        10000 non-null  int64  
 1   CustomerId       10000 non-null  int64  
 2   Surname          10000 non-null  object 
 3   CreditScore      10000 non-null  int64  
 4   Geography        10000 non-null  object 
 5   Gender           10000 non-null  object 
 6   Age              10000 non-null  int64  
 7   Tenure           10000 non-null  int64  
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64  
 10  HasCrCard        10000 non-null  int64  
 11  IsActiveMember   10000 non-null  int64  
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64  
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

There are 14 columns in all, 1 of which is the target column and 13 of which are predictors.



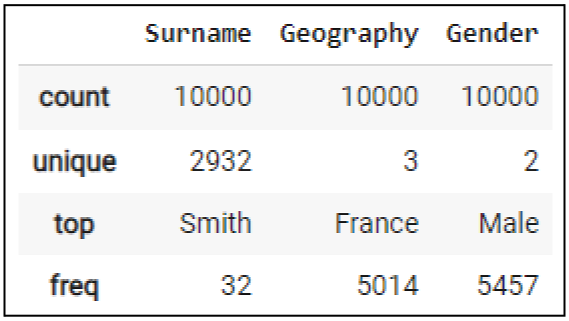
	Count
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

When the dataset's null values are checked, it is evident that the dataset is virtually clean. However, this may not be the case because the dataset may contain some outliers. It is simple to identify outliers using a boxplot.



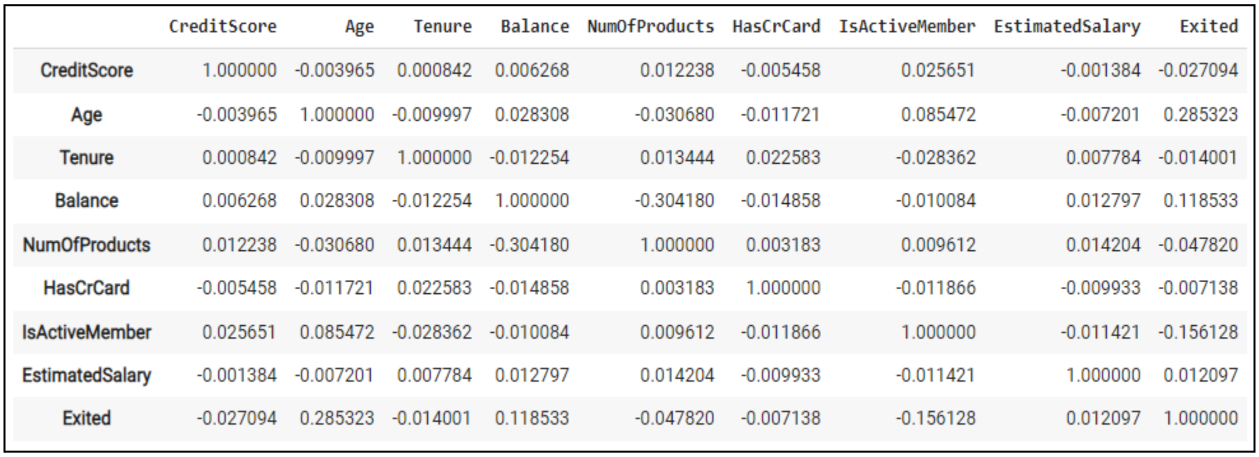
	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Some statistical inferences can be discovered using the Pandas library's describe() function. The average age of a bank account holder is roughly 39, with credit scores ranging between 350 and 850. Some consumers have been with the company for fifteen years.

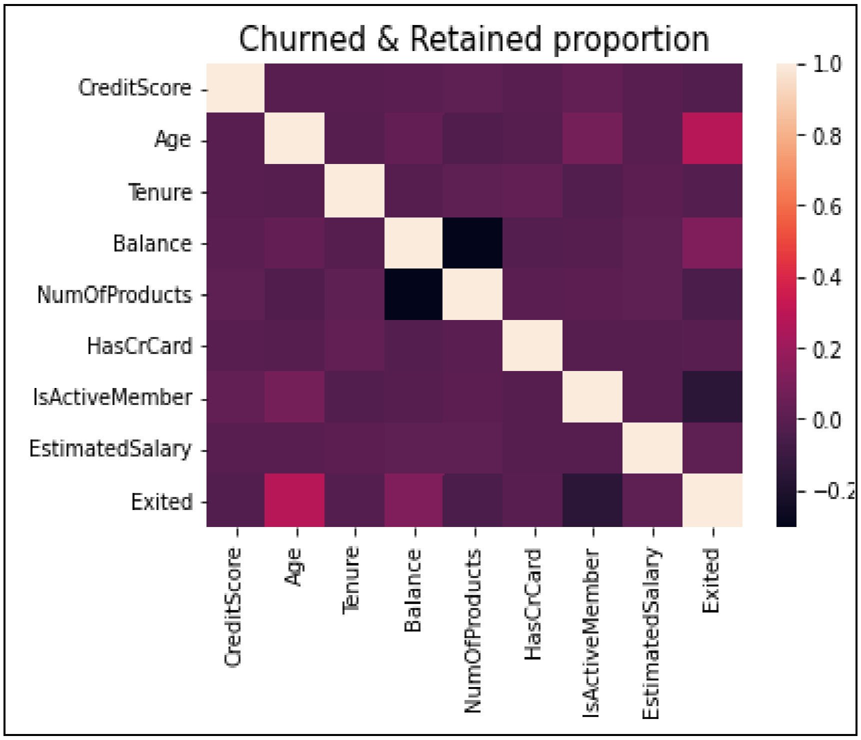


	Surname	Geography	Gender
count	10000	10000	10000
unique	2932	3	2
top	Smith	France	Male
freq	32	5014	5457

By describing the table's alphanumeric properties, three columns were found to have string values, of which the geography and gender columns are suitable for prediction. The geography column contains three distinct values: France, Spain, and Germany. Gender is divided into two categories: male and female.



	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
CreditScore	1.000000	-0.003965	0.000842	0.006268	0.012238	-0.005458	0.025651	-0.001384	-0.027094
Age	-0.003965	1.000000	-0.009997	0.028308	-0.030680	-0.011721	0.085472	-0.007201	0.285323
Tenure	0.000842	-0.009997	1.000000	-0.012254	0.013444	0.022583	-0.028362	0.007784	-0.014001
Balance	0.006268	0.028308	-0.012254	1.000000	-0.304180	-0.014858	-0.010084	0.012797	0.118533
NumOfProducts	0.012238	-0.030680	0.013444	-0.304180	1.000000	0.003183	0.009612	0.014204	-0.047820
HasCrCard	-0.005458	-0.011721	0.022583	-0.014858	0.003183	1.000000	-0.011866	-0.009933	-0.007138
IsActiveMember	0.025651	0.085472	-0.028362	-0.010084	0.009612	-0.011866	1.000000	-0.011421	-0.156128
EstimatedSalary	-0.001384	-0.007201	0.007784	0.012797	0.014204	-0.009933	-0.011421	1.000000	0.012097
Exited	-0.027094	0.285323	-0.014001	0.118533	-0.047820	-0.007138	-0.156128	0.012097	1.000000

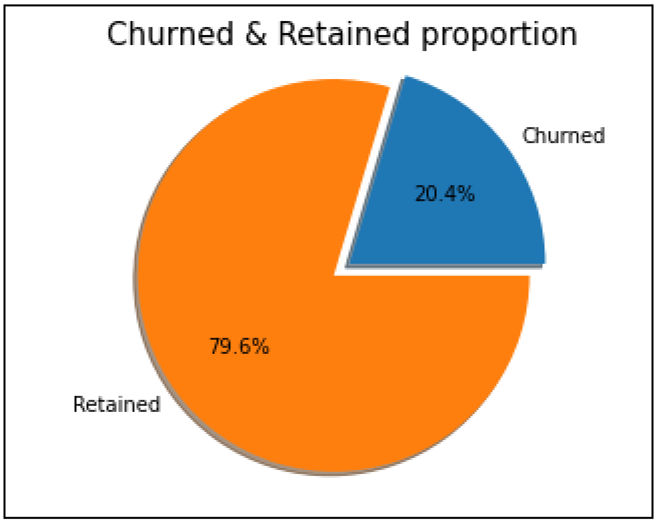


A correlation matrix was utilized to determine the relationship between the other characteristics in the column. Upon further investigation, there was little association apparent, however, Age has some correlation link with the churn prediction column.

Data Visualization:

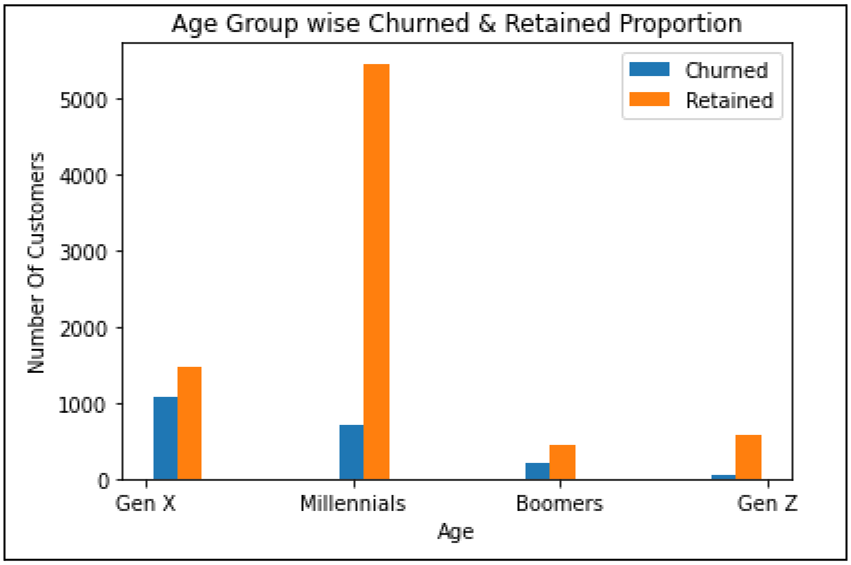
The presentation of data in a pictorial or graphical style is known as data visualization. It allows decision-makers to see analytics visually portrayed, allowing them to understand complex ideas or uncover new trends. You can take the notion a step further with interactive visualization by leveraging technology to drill down into charts and graphs for additional detail, dynamically modifying what data you see and how it's handled.

A. Churn and Retained Proportion



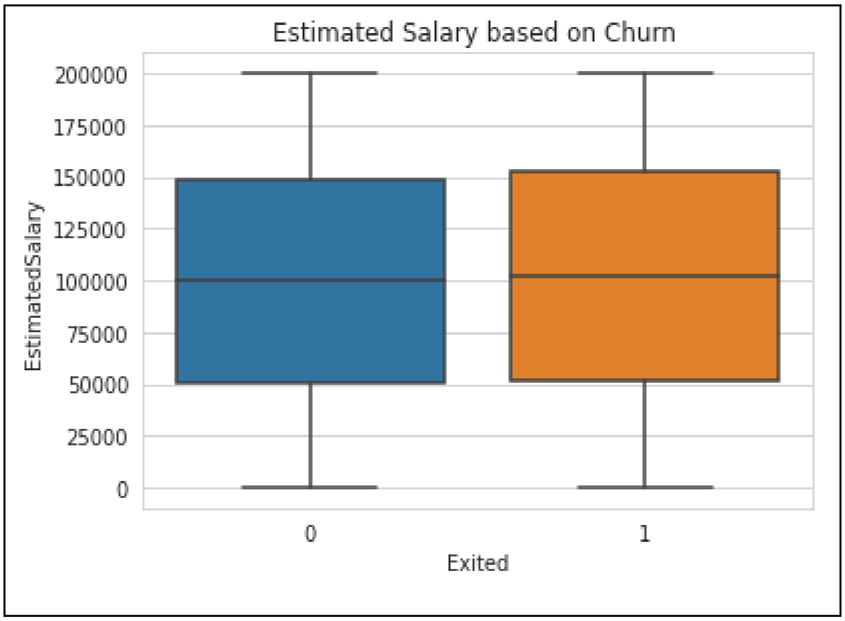
According to this graph, 79.6% of customers were kept while 20.4% churned.

B. Age Group Wise Churned & Retained Proportion



Millennial clients, those aged 26 to 40, account for the majority of customers. Furthermore, the proportion of consumers that churn is excessively high for the Gen X age group compared to all other age groups.

C. Estimated Salary Based on Churn:



Outliers in the dataset can be identified more easily using a box plot. When compared to the EstimatedSalary column, almost all of the attributes in the table had a normal range. There were no outliers in the dataset when EstimatedSalary was shown.

Data Mining Tasks:

Our dataset had no null values, however, there were a few columns that needed to be removed, including RowNumber, Surname, and CustomerID. These columns were not relevant for model development since there was no relationship between them and the goal column.

```
data = data.drop(columns=["RowNumber", "CustomerId", "Surname"])
```

Label Encoding:

With the help of SKlearn's preprocessing library, gender, geography, and age group columns were encoded with the numeric label as dimensionality reduction doesn't work on non-numeric columns.

1. **Label Encoding of Gender Column:**

```
Before Encoding:  ['Female' 'Male']
After Encoding:  [0 1]
```

2. **Label Encoding of Geography Column:**

```
Before Encoding:  ['France' 'Spain' 'Germany']
After Encoding:  [0 2 1]
```

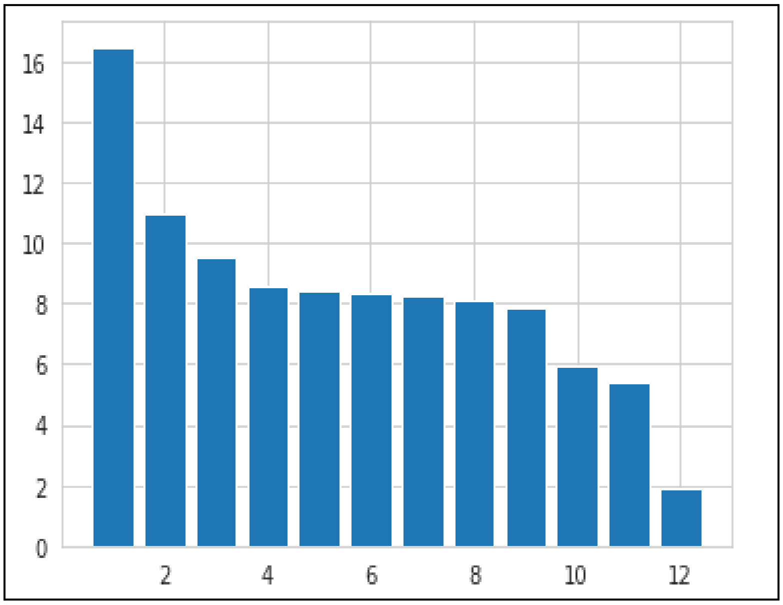
3. **Label Encoding of AgeGroup Column:**

```
Before Encoding:  ['Gen X' 'Millennials' 'Gen Z' 'Boomers']
After Encoding:  [1 3 2 0]
```

Dimensionality Reduction:

It is advised to scale the data in the usual format before lowering the dimension since this allows the model to only use columns that affect the target column. Scaled data also aids PCA in determining the optimum covariance matrix.

```
Explained Variance:
     [1.97776764 1.3203782  1.1462487  1.03301754 1.01520873 1.00093189
     0.99140536 0.97334139 0.94814303 0.71657325 0.65000503 0.22817937]
Proportion Variance:
     [0.16479749 0.11002051 0.09551117 0.08607619 0.08459227 0.08340265
     0.08260885 0.08110367 0.07900402 0.05970847 0.05416167 0.01901305]
Cumulative Variance:
     [0.16479749 0.274818   0.37032918 0.45640536 0.54099763 0.62440028
     0.70700913 0.7881128  0.86711682 0.92682529 0.98098695 1.        ]
```



It is evident that the majority of the columns have a greater correlation with the target column, thus we may remove the final column to increase performance.

Data Partitioning:

Data splitting is commonly used in machine learning to avoid overfitting. Typically, the original data in a machine learning model is divided into two or three parts, such as training, testing, and validation. For this dataset, the data is divided into two parts, 70% for training and 30% for testing, using the sklearn library's train_test_split() function. It is recommended to use randomized records from the data over the split using the random_state parameter.

Data Mining Models/Methods:

The primary goal of this project is to predict whether or not a customer will churn, which can be accomplished with Supervised Machine Learning Classification algorithms. The top three algorithms are chosen from a number of algorithms and studies based on their advantages and disadvantages.

1. **Logistic Regression Classifier:**

   Logistic Regression is a statistical technique for analyzing a dataset in which one or more independent variables influence the result. The target variable is modeled using a logistic function. There are several types of this model, including binomial, multinomial, and ordinal, with binomial being the best fit for the dataset because the target column has only two values: whether or not the customer will churn. This algorithm uses the sigmoid curve to predict the output of the target variable.
2. **KNN Classifier:**

   The KNN algorithm is a simple supervised machine learning algorithm that is used for classification and regression problems. The KNN algorithm uses distance measures to find common/similar things to the nearest K points for classification and regression. The reasons for selecting this algorithm are its O(n) time complexity and its ability to work efficiently on small datasets. By consuming less time and improving efficiency, this algorithm will be able to predict customer churn.
3. **Random Forest Classifier:**

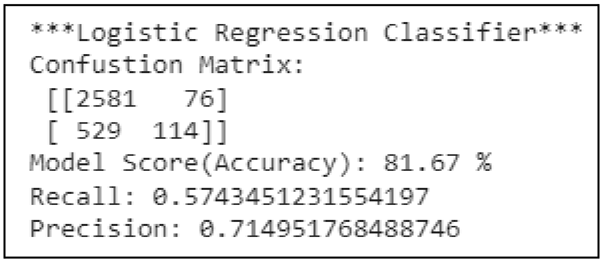
   Random forest is one of the most effective and robust supervised machine learning algorithms for classification and regression. As an ensemble, this algorithm utilizes a set of decision trees. When building each individual decision tree, this algorithm predicts classes using bagging and feature randomness. This algorithm randomly selects samples from the dataset and builds a decision tree for each sample. Data is predicted through each tree after it has been constructed. Finally, the final class of data is predicted using voting.

Performance Evaluation:

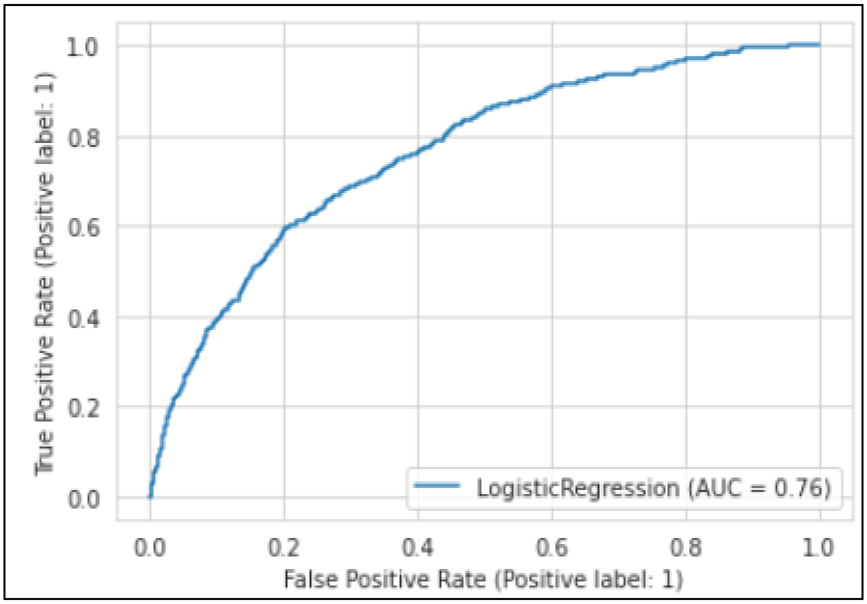
Building a classification algorithm is always a fun project to do when you are getting into Data Mining. But, evaluating a classification algorithm is confusing as there are a lot of parameters involved while creating a model which might impact the outcome of a model. Based on the certain set of parameters model is evaluated based on the following parameters:

1. Confusion Matrix
2. Model Score
3. Recall
4. Precision

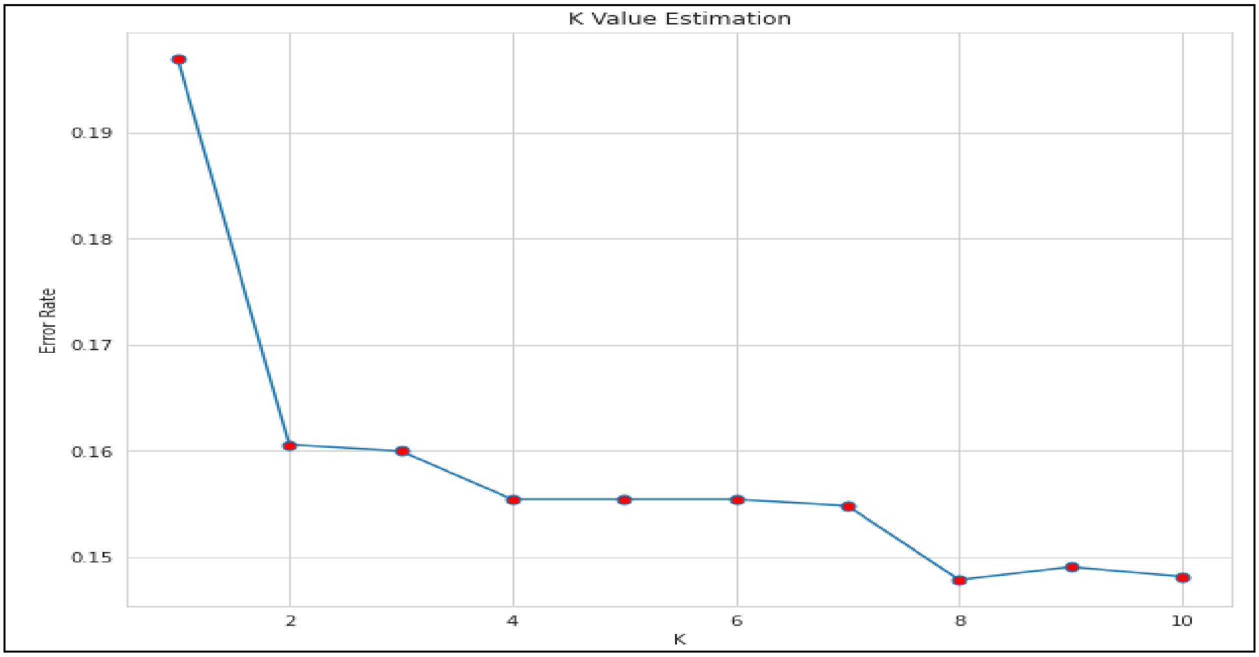
In the Logistic Regression Model, with the help of a sigmoid curve, the test data is getting classified. This model is one of the most robust algorithms because of which the accuracy is low compared to KNN and Random Forest Classifier. The ROC curve was so far from the top-left corner, with an average AUC value of 0.76 indicating normal classification between churned customers.



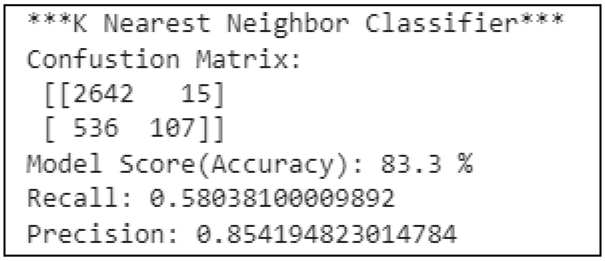
```
***Logistic Regression Classifier***
Confustion Matrix:
 [[2581   76]
 [ 529  114]]
Model Score(Accuracy): 81.67 %
Recall: 0.5743451231554197
Precision: 0.714951768488746
```



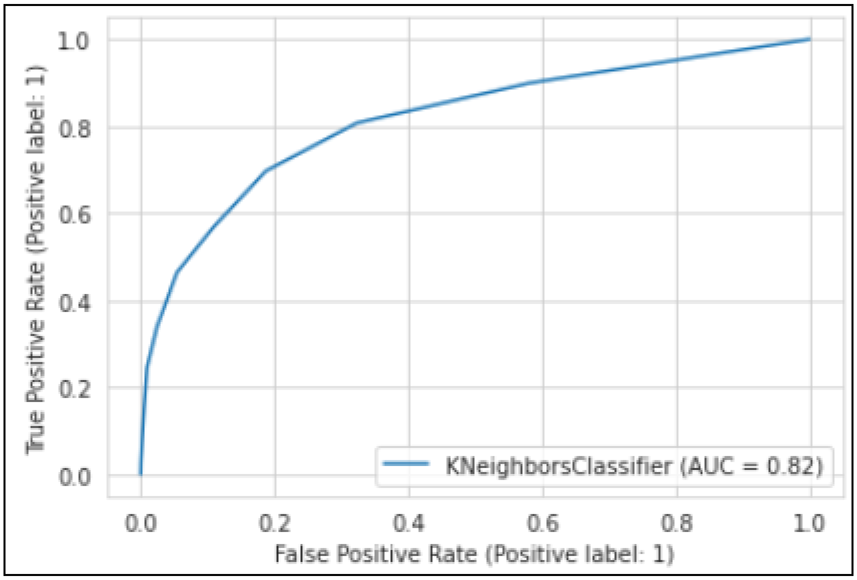
In the K-Nearest Neighbor algorithm, the main task while creating the model is to select the perfect value of K. By using the elbow method, the K value is getting selected within the range of 1 to10.



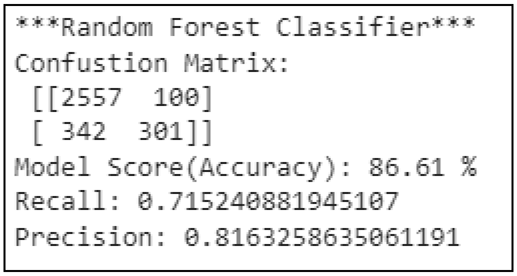
For K=4,5,6, the model was given almost the same and consistent, accuracy. To avoid the miscalculation for the even value of k, we have chosen k=5 for the perfect classification. Even after the evaluation, KNN was having decent accuracy of 83.3% only, while the value of precision was getting better compared to the Logistic Regression model. The ROC curve was so far from the top-left corner, with an improved AUC value of 0.82 indicating typical classification between churned customers. The ROC curve was so far from the top-left corner, with an average AUC value of 0.82 indicating normal classification between churned customers.



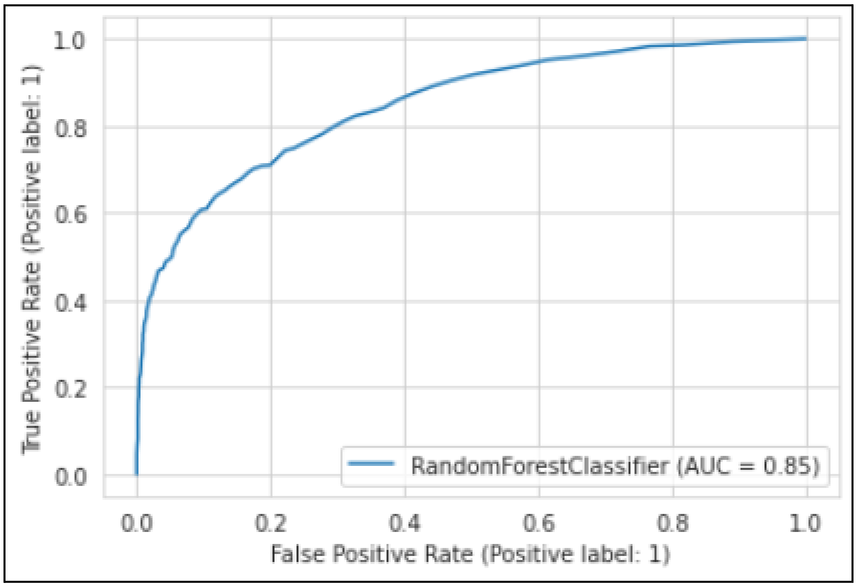
```
***K Nearest Neighbor Classifier***
Confustion Matrix:
 [[2642   15]
 [ 536  107]]
Model Score(Accuracy): 83.3 %
Recall: 0.58038100009892
Precision: 0.854194823014784
```



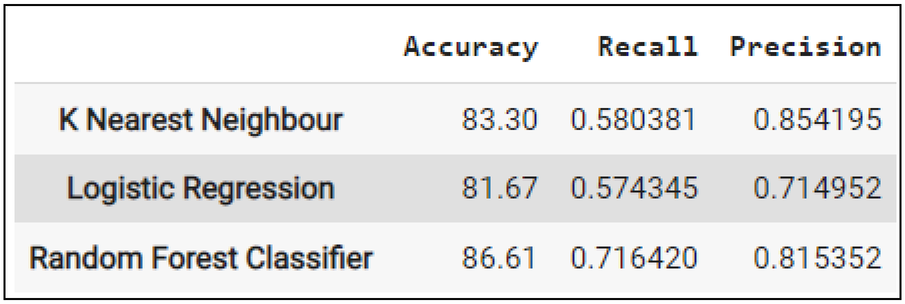
In the Random Forest Classifier, the model was giving the best performance among all the other 2 models. By choosing the n_estomators = 100, the accuracy of the model was increased to 86.61%. Along with the increase in accuracy, significant changes in the Recall and Precision values of the model have been observed. The ROC curve was so getting close to the top-left corner, with an improved AUC value of 0.85 indicating normal classification between churned customers.



```
***Random Forest Classifier***
Confustion Matrix:
 [[2557  100]
 [ 342  301]]
Model Score(Accuracy): 86.61 %
Recall: 0.715240881945107
Precision: 0.8163258635061191
```



Overall Model Result:



	Accuracy	Recall	Precision
K Nearest Neighbour	83.30	0.580381	0.854195
Logistic Regression	81.67	0.574345	0.714952
Random Forest Classifier	86.61	0.716420	0.815352

The primary goal of this project was to identify the churning of a customer. The recall and precision matrices are used to evaluate this accuracy. The Random Forest Classifier has the highest accuracy of all three models (86.61%), followed by K Nearest Neighbour and Logistic Regression. The K Nearest Neighbour algorithm has the highest precision for churned customers, but the Random Forest Classifier has the highest recall value. There is still scope for increasing the accuracy of the model by more than 90%.

Bank Customer Churn Prediction

Problem Setting:

Customer churn, also known as customer attrition, is the phenomenon in which a customer leaves an organization. This customer attrition is likely to influence a bank's income for various reasons, including service, facilities, and assistance provided to consumers. When compared to client retention, the expense of marketing to gain new consumers is prohibitively expensive.

Problem Definition:

To grow a bank's business model, it is necessary to forecast the likelihood of client attrition. This project's primary goal is to create a supervised machine-learning model that will help in the classification of churn consumers.

Data Source:

The data was obtained from the well-known open-source repository Kaggle.

URL: https://www.kaggle.com/datasets/santoshd3/bank-customers

Data Description:

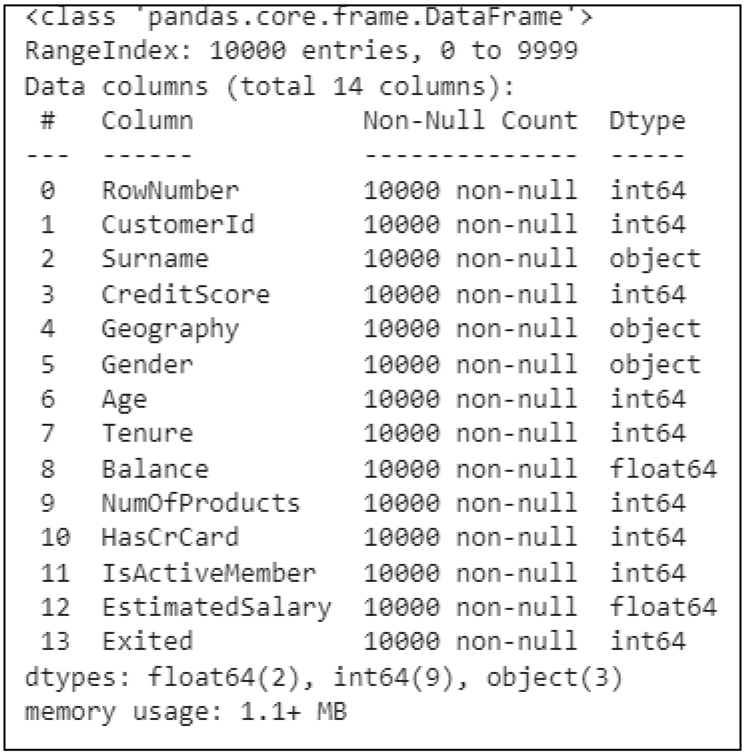
The data consists of 10000 Customers across the globe. This data has the following columns:

No	Feature Name
1	RowNumber
2	CustomerId
3	Surname
4	CreditScore
5	Geography
6	Gender
7	Age
8	Tenure
9	Balance

10	NumOfProducts
11	HasCrCard
12	IsActiveMember
13	EstimatedSalary
14	Exited

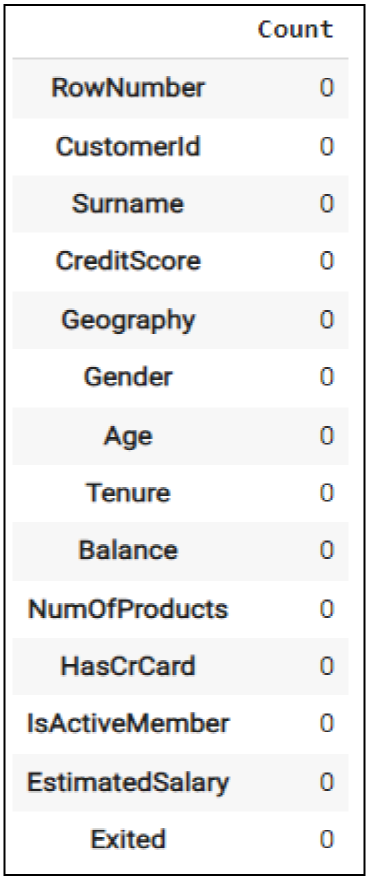
Data Exploration:

Data exploration is the initial phase in data analysis and is used to examine and display data in order to find insights from the beginning or to suggest regions or trends investigate further. Data is statistically and visually analyzed using Python packages. Finding the form of the table, correlation matrix, null values, and unique values are all part of the data exploration process.



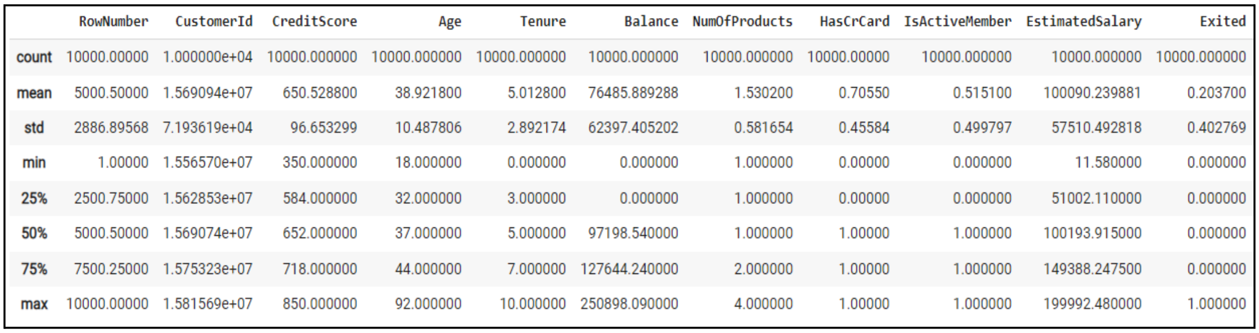
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   RowNumber        10000 non-null  int64  
 1   CustomerId       10000 non-null  int64  
 2   Surname          10000 non-null  object 
 3   CreditScore      10000 non-null  int64  
 4   Geography        10000 non-null  object 
 5   Gender           10000 non-null  object 
 6   Age              10000 non-null  int64  
 7   Tenure           10000 non-null  int64  
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64  
 10  HasCrCard        10000 non-null  int64  
 11  IsActiveMember   10000 non-null  int64  
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64  
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

There are 14 columns in all, 1 of which is the target column and 13 of which are predictors.



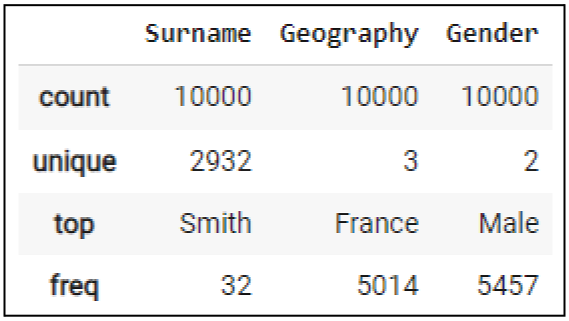
	Count
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

When the dataset's null values are checked, it is evident that the dataset is virtually clean. However, this may not be the case because the dataset may contain some outliers. It is simple to identify outliers using a boxplot.



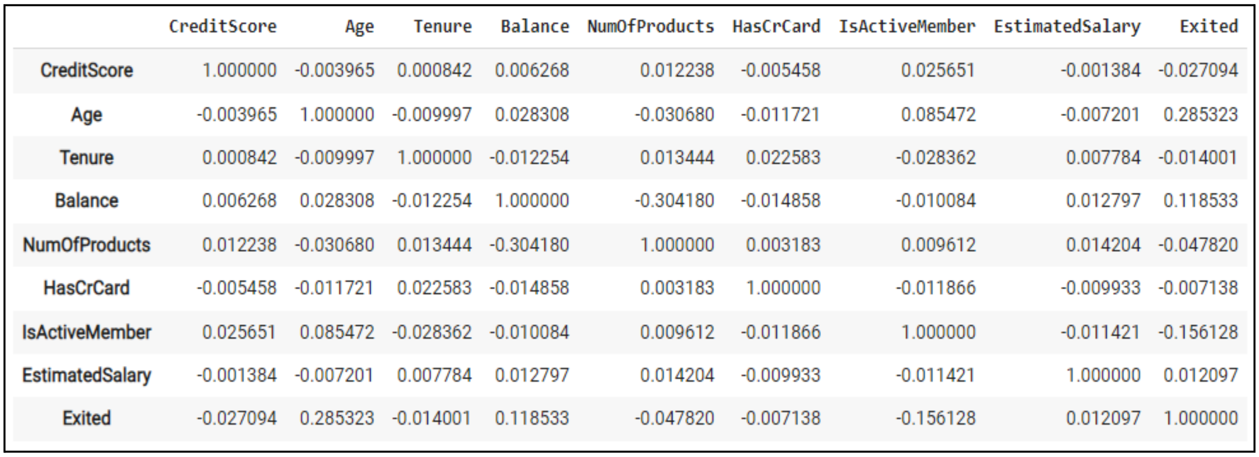
	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Some statistical inferences can be discovered using the Pandas library's describe() function. The average age of a bank account holder is roughly 39, with credit scores ranging between 350 and 850. Some consumers have been with the company for fifteen years.

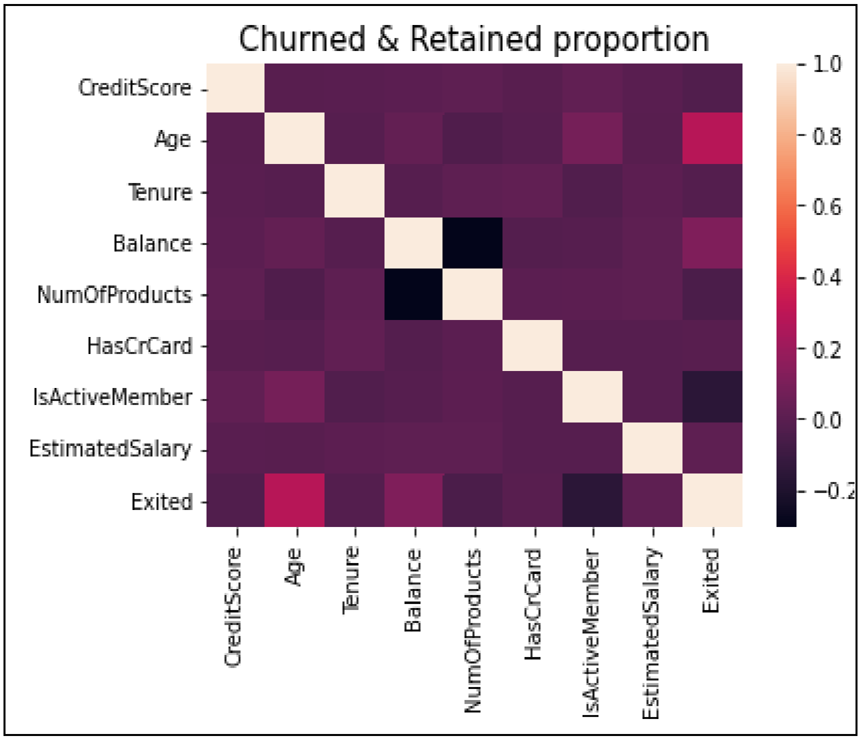


	Surname	Geography	Gender
count	10000	10000	10000
unique	2932	3	2
top	Smith	France	Male
freq	32	5014	5457

By describing the table's alphanumeric properties, three columns were found to have string values, of which the geography and gender columns are suitable for prediction. The geography column contains three distinct values: France, Spain, and Germany. Gender is divided into two categories: male and female.



	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
CreditScore	1.000000	-0.003965	0.000842	0.006268	0.012238	-0.005458	0.025651	-0.001384	-0.027094
Age	-0.003965	1.000000	-0.009997	0.028308	-0.030680	-0.011721	0.085472	-0.007201	0.285323
Tenure	0.000842	-0.009997	1.000000	-0.012254	0.013444	0.022583	-0.028362	0.007784	-0.014001
Balance	0.006268	0.028308	-0.012254	1.000000	-0.304180	-0.014858	-0.010084	0.012797	0.118533
NumOfProducts	0.012238	-0.030680	0.013444	-0.304180	1.000000	0.003183	0.009612	0.014204	-0.047820
HasCrCard	-0.005458	-0.011721	0.022583	-0.014858	0.003183	1.000000	-0.011866	-0.009933	-0.007138
IsActiveMember	0.025651	0.085472	-0.028362	-0.010084	0.009612	-0.011866	1.000000	-0.011421	-0.156128
EstimatedSalary	-0.001384	-0.007201	0.007784	0.012797	0.014204	-0.009933	-0.011421	1.000000	0.012097
Exited	-0.027094	0.285323	-0.014001	0.118533	-0.047820	-0.007138	-0.156128	0.012097	1.000000

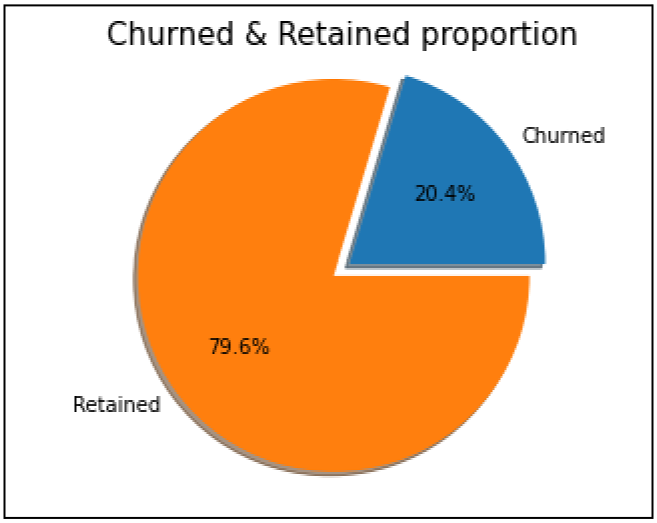


A correlation matrix was utilized to determine the relationship between the other characteristics in the column. Upon further investigation, there was little association apparent, however, Age has some correlation link with the churn prediction column.

Data Visualization:

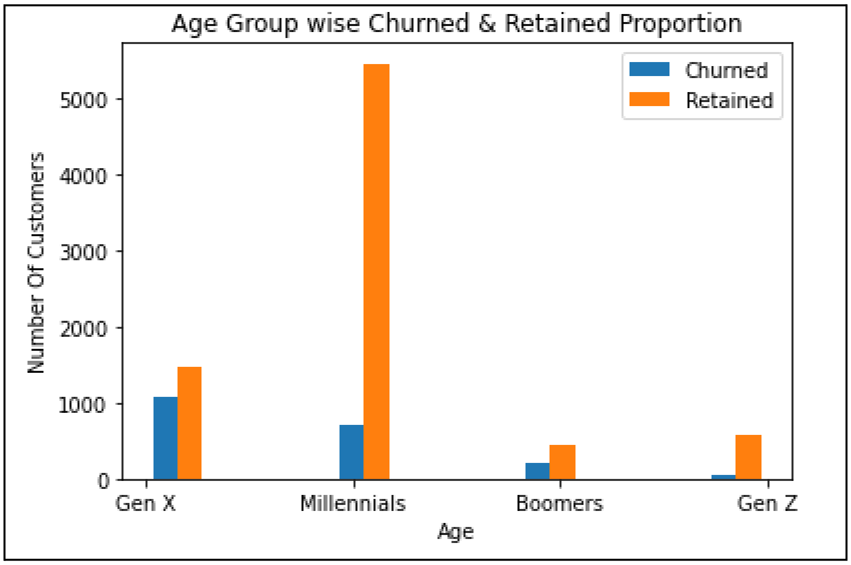
The presentation of data in a pictorial or graphical style is known as data visualization. It allows decision-makers to see analytics visually portrayed, allowing them to understand complex ideas or uncover new trends. You can take the notion a step further with interactive visualization by leveraging technology to drill down into charts and graphs for additional detail, dynamically modifying what data you see and how it's handled.

A. Churn and Retained Proportion



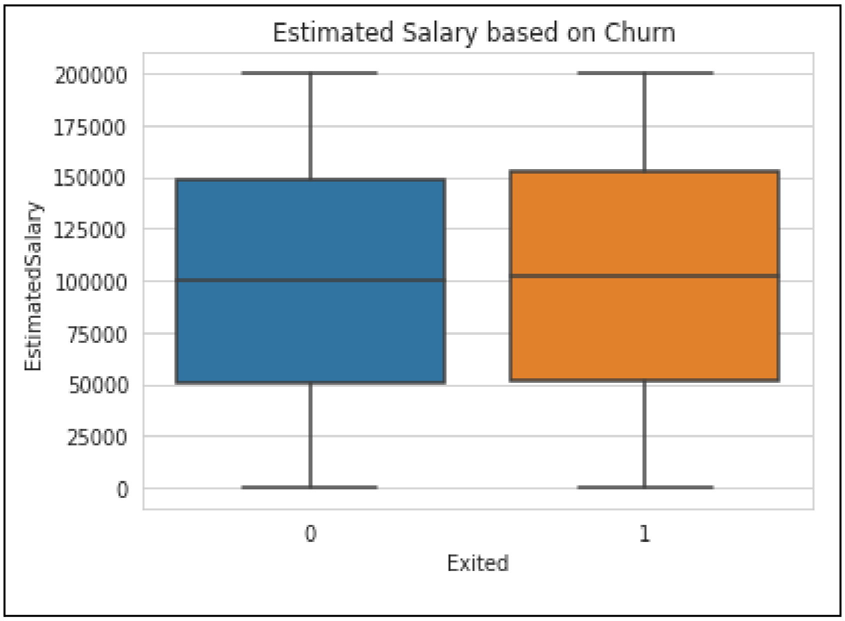
According to this graph, 79.6% of customers were kept while 20.4% churned.

B. Age Group Wise Churned & Retained Proportion



Millennial clients, those aged 26 to 40, account for the majority of customers. Furthermore, the proportion of consumers that churn is excessively high for the Gen X age group compared to all other age groups.

C. Estimated Salary Based on Churn:



Outliers in the dataset can be identified more easily using a box plot. When compared to the EstimatedSalary column, almost all of the attributes in the table had a normal range. There were no outliers in the dataset when EstimatedSalary was shown.

Data Mining Tasks:

Our dataset had no null values, however, there were a few columns that needed to be removed, including RowNumber, Surname, and CustomerID. These columns were not relevant for model development since there was no relationship between them and the goal column.

```
data = data.drop(columns=["RowNumber", "CustomerId", "Surname"])
```

Label Encoding:

With the help of SKlearn's preprocessing library, gender, geography, and age group columns were encoded with the numeric label as dimensionality reduction doesn't work on non-numeric columns.

1. **Label Encoding of Gender Column:**

```
Before Encoding:  ['Female' 'Male']
After Encoding:  [0 1]
```

2. **Label Encoding of Geography Column:**

```
Before Encoding:  ['France' 'Spain' 'Germany']
After Encoding:  [0 2 1]
```

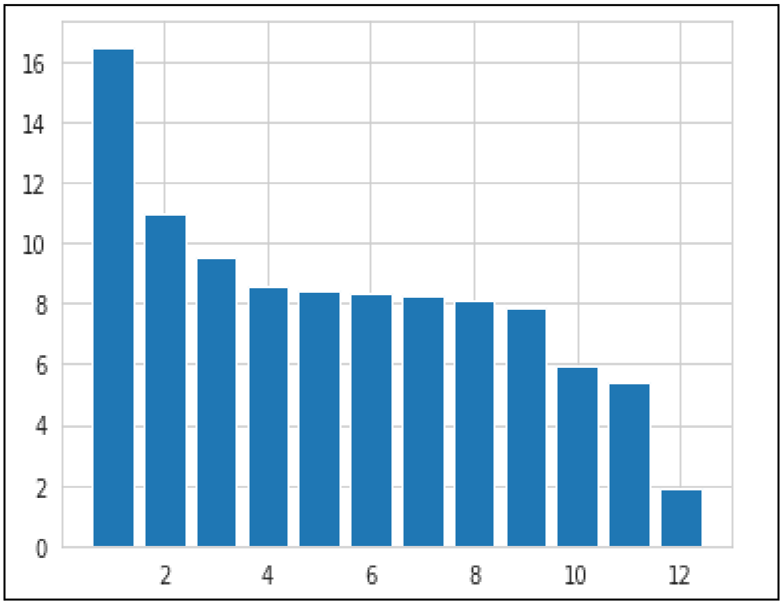
3. **Label Encoding of AgeGroup Column:**

```
Before Encoding:  ['Gen X' 'Millennials' 'Gen Z' 'Boomers']
After Encoding:  [1 3 2 0]
```

Dimensionality Reduction:

It is advised to scale the data in the usual format before lowering the dimension since this allows the model to only use columns that affect the target column. Scaled data also aids PCA in determining the optimum covariance matrix.

```
Explained Variance:
     [1.97776764 1.3203782  1.1462487  1.03301754 1.01520873 1.00093189
     0.99140536 0.97334139 0.94814303 0.71657325 0.65000503 0.22817937]
Proportion Variance:
     [0.16479749 0.11002051 0.09551117 0.08607619 0.08459227 0.08340265
     0.08260885 0.08110367 0.07900402 0.05970847 0.05416167 0.01901305]
Cumulative Variance:
     [0.16479749 0.274818   0.37032918 0.45640536 0.54099763 0.62440028
     0.70700913 0.7881128  0.86711682 0.92682529 0.98098695 1.        ]
```



It is evident that the majority of the columns have a greater correlation with the target column, thus we may remove the final column to increase performance.

Data Partitioning:

Data splitting is commonly used in machine learning to avoid overfitting. Typically, the original data in a machine learning model is divided into two or three parts, such as training, testing, and validation. For this dataset, the data is divided into two parts, 70% for training and 30% for testing, using the sklearn library's train_test_split() function. It is recommended to use randomized records from the data over the split using the random_state parameter.

Data Mining Models/Methods:

The primary goal of this project is to predict whether or not a customer will churn, which can be accomplished with Supervised Machine Learning Classification algorithms. The top three algorithms are chosen from a number of algorithms and studies based on their advantages and disadvantages.

1. **Logistic Regression Classifier:**

   Logistic Regression is a statistical technique for analyzing a dataset in which one or more independent variables influence the result. The target variable is modeled using a logistic function. There are several types of this model, including binomial, multinomial, and ordinal, with binomial being the best fit for the dataset because the target column has only two values: whether or not the customer will churn. This algorithm uses the sigmoid curve to predict the output of the target variable.
2. **KNN Classifier:**

   The KNN algorithm is a simple supervised machine learning algorithm that is used for classification and regression problems. The KNN algorithm uses distance measures to find common/similar things to the nearest K points for classification and regression. The reasons for selecting this algorithm are its O(n) time complexity and its ability to work efficiently on small datasets. By consuming less time and improving efficiency, this algorithm will be able to predict customer churn.
3. **Random Forest Classifier:**

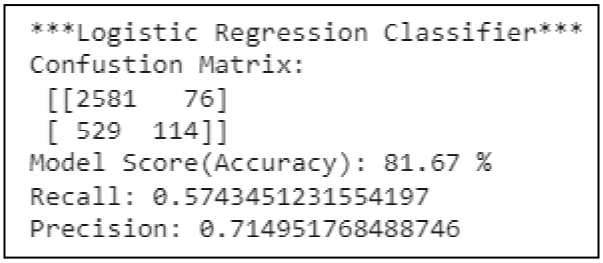
   Random forest is one of the most effective and robust supervised machine learning algorithms for classification and regression. As an ensemble, this algorithm utilizes a set of decision trees. When building each individual decision tree, this algorithm predicts classes using bagging and feature randomness. This algorithm randomly selects samples from the dataset and builds a decision tree for each sample. Data is predicted through each tree after it has been constructed. Finally, the final class of data is predicted using voting.

Performance Evaluation:

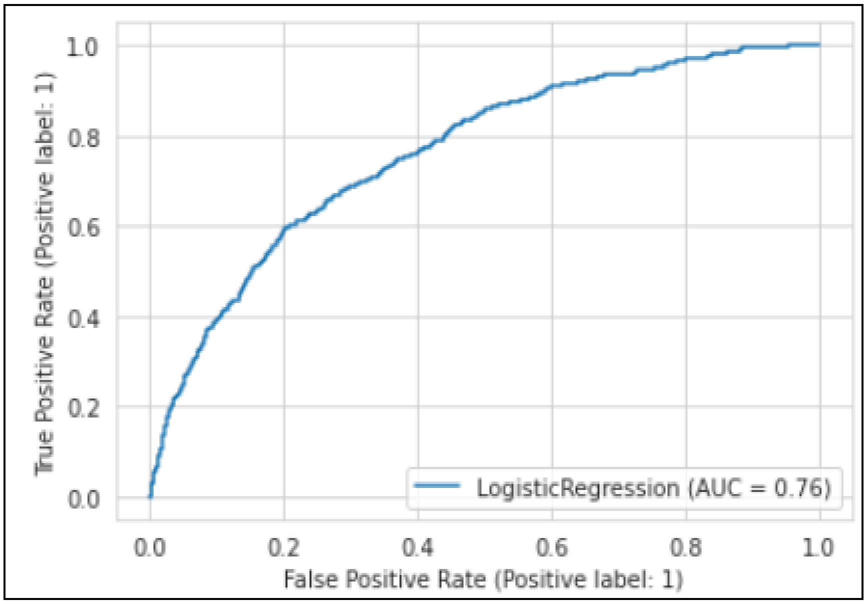
Building a classification algorithm is always a fun project to do when you are getting into Data Mining. But, evaluating a classification algorithm is confusing as there are a lot of parameters involved while creating a model which might impact the outcome of a model. Based on the certain set of parameters model is evaluated based on the following parameters:

1. Confusion Matrix
2. Model Score
3. Recall
4. Precision

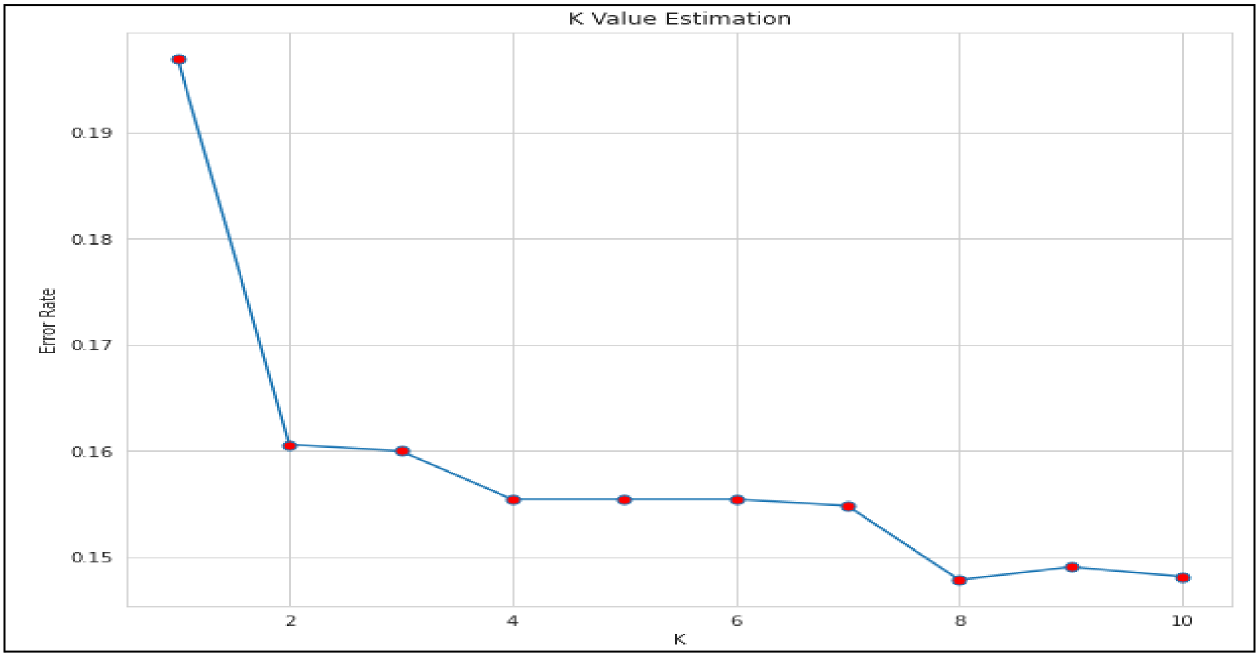
In the Logistic Regression Model, with the help of a sigmoid curve, the test data is getting classified. This model is one of the most robust algorithms because of which the accuracy is low compared to KNN and Random Forest Classifier. The ROC curve was so far from the top-left corner, with an average AUC value of 0.76 indicating normal classification between churned customers.



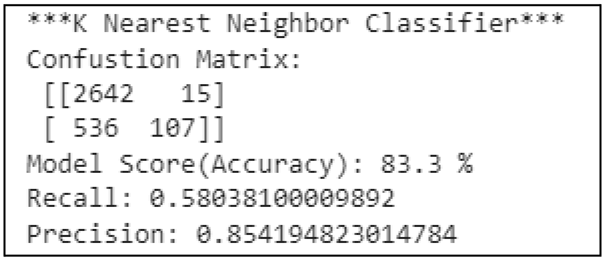
```
***Logistic Regression Classifier***
Confustion Matrix:
 [[2581   76]
 [ 529  114]]
Model Score(Accuracy): 81.67 %
Recall: 0.5743451231554197
Precision: 0.714951768488746
```



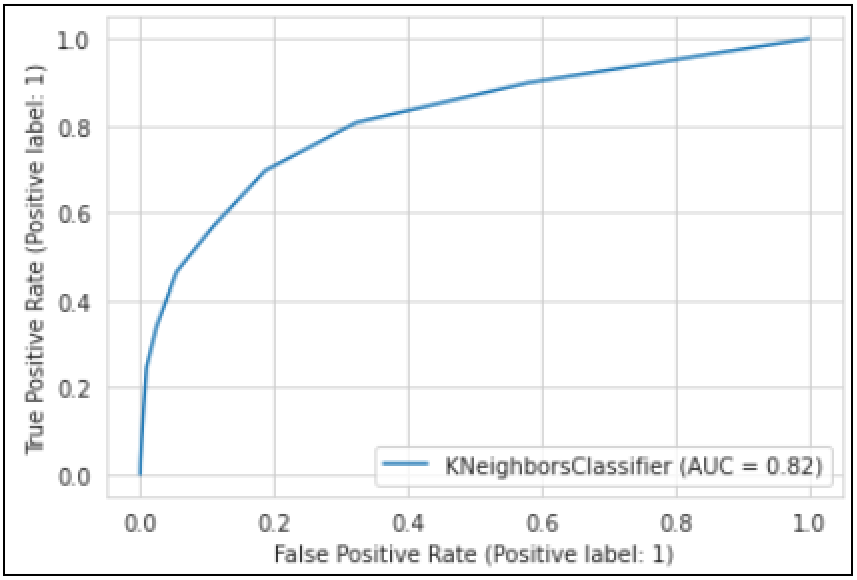
In the K-Nearest Neighbor algorithm, the main task while creating the model is to select the perfect value of K. By using the elbow method, the K value is getting selected within the range of 1 to10.



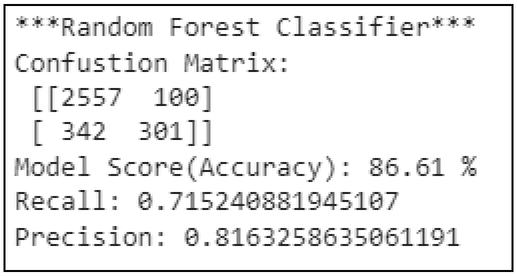
For K=4,5,6, the model was given almost the same and consistent, accuracy. To avoid the miscalculation for the even value of k, we have chosen k=5 for the perfect classification. Even after the evaluation, KNN was having decent accuracy of 83.3% only, while the value of precision was getting better compared to the Logistic Regression model. The ROC curve was so far from the top-left corner, with an improved AUC value of 0.82 indicating typical classification between churned customers. The ROC curve was so far from the top-left corner, with an average AUC value of 0.82 indicating normal classification between churned customers.



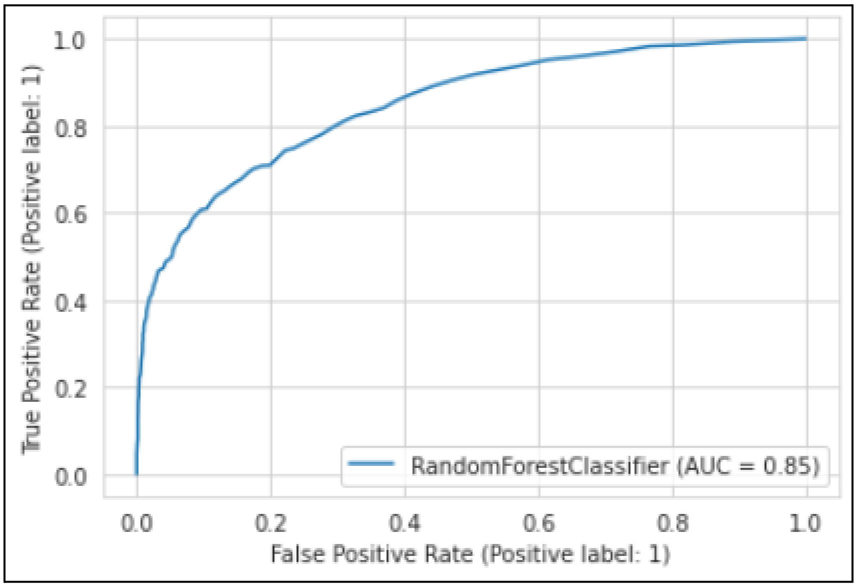
```
***K Nearest Neighbor Classifier***
Confustion Matrix:
 [[2642   15]
 [ 536  107]]
Model Score(Accuracy): 83.3 %
Recall: 0.58038100009892
Precision: 0.854194823014784
```



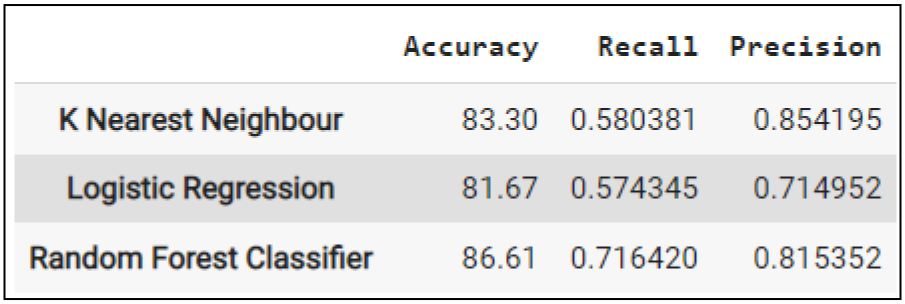
In the Random Forest Classifier, the model was giving the best performance among all the other 2 models. By choosing the n_estomators = 100, the accuracy of the model was increased to 86.61%. Along with the increase in accuracy, significant changes in the Recall and Precision values of the model have been observed. The ROC curve was so getting close to the top-left corner, with an improved AUC value of 0.85 indicating normal classification between churned customers.



```
***Random Forest Classifier***
Confustion Matrix:
 [[2557  100]
 [ 342  301]]
Model Score(Accuracy): 86.61 %
Recall: 0.715240881945107
Precision: 0.8163258635061191
```



Overall Model Result:



	Accuracy	Recall	Precision
K Nearest Neighbour	83.30	0.580381	0.854195
Logistic Regression	81.67	0.574345	0.714952
Random Forest Classifier	86.61	0.716420	0.815352

The primary goal of this project was to identify the churning of a customer. The recall and precision matrices are used to evaluate this accuracy. The Random Forest Classifier has the highest accuracy of all three models (86.61%), followed by K Nearest Neighbour and Logistic Regression. The K Nearest Neighbour algorithm has the highest precision for churned customers, but the Random Forest Classifier has the highest recall value. There is still scope for increasing the accuracy of the model by more than 90%.