Customer Churn Prediction

-To Predict the Customer Churn in banking

- **Harish Jamallamudi**
- **Balamurali B**
- **Yenkatarajalaxmimanohar Meda**
- **Amruth Reddy Nagireddy Palli**
- **Sai Sreeja Yalamanchi**
- **Anirudh Boddu**

Content

- **Title and Author**
- **Content**
- **Data Description**
- **Data Attributes**

Data Description

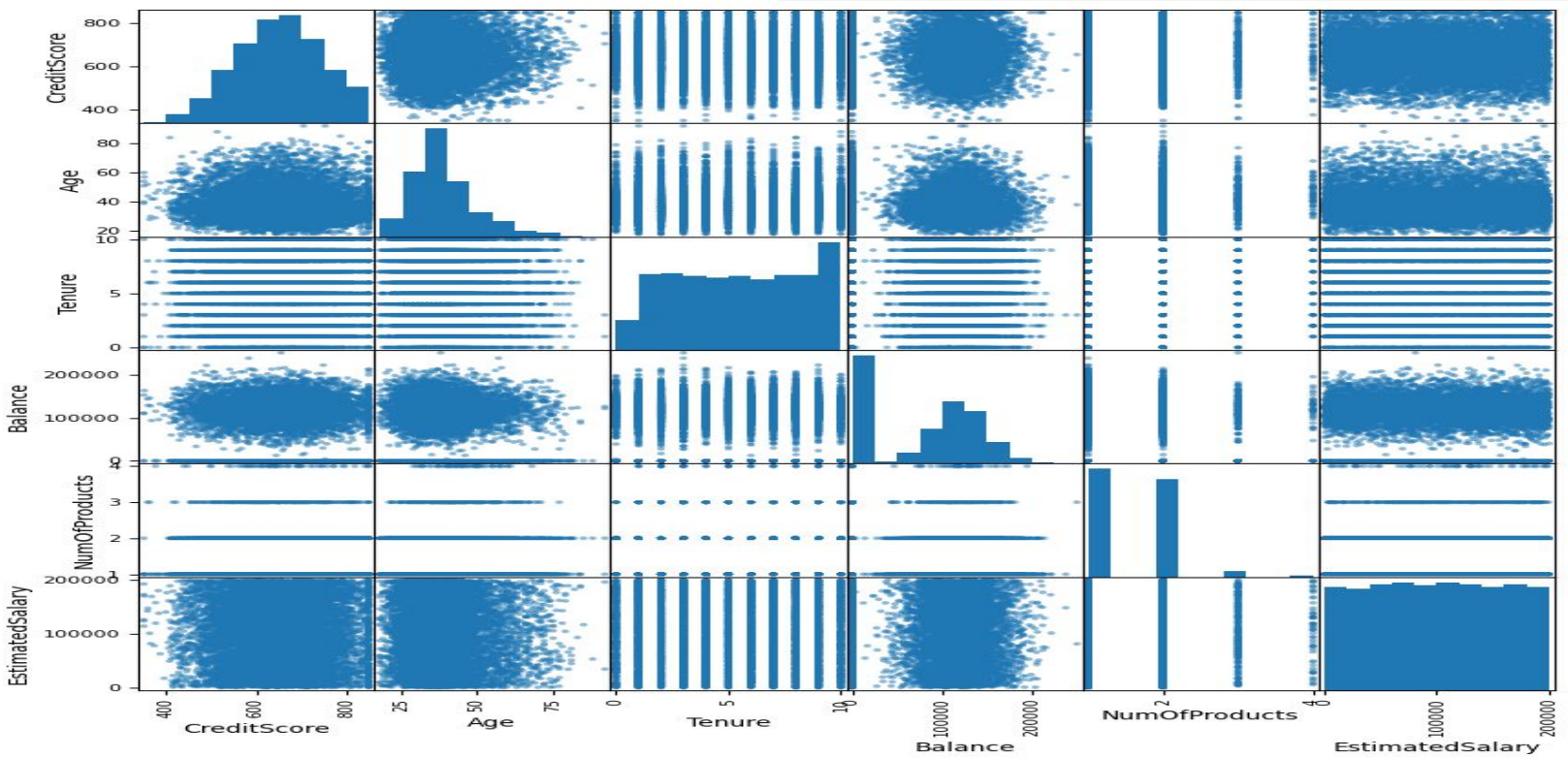
This dataset includes details regarding the banking services and products the customer holds with the bank. This may encompass information about the kind of accounts they have like savings or checking accounts as their credit cards, loan accounts, and any other financial products they utilize. Moreover, it could also consist of data related to their transaction history, such as how they make transactions the amount of money deposited or withdrawn, and the specific types of transactions they engage in. Collectively this data can provide insights into how involved the customers are with the bank.

The dataset is available at : https://www.kaggle.com/datasets/shubh0799/churn-modelling

Data Attributes

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

SCATTER PLOT



Interpretation

Scatter plot is used to find the relationship between features like linear relationship or non linear relationship and clusters etc. From above scatter plot on the quantitative data from dataset “churn modelling”, we decided to remove age because, it has linear relationship between credit score and age, and creditscore has gaussian curve. Remaining attributes present in dataset are

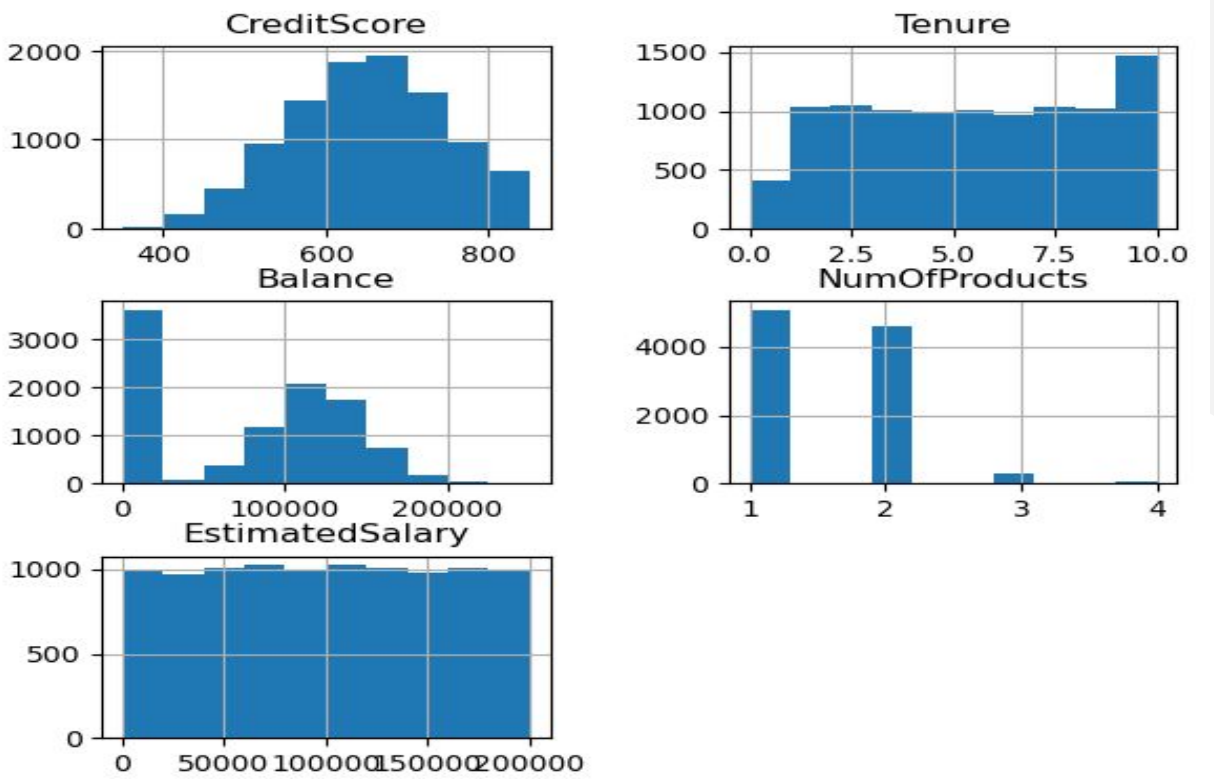
Qualitative

- Geography
- Gender
- HasCrCard
- IsActiveMember
- Exited

Quantitative

- CreditScore
- Tenure
- Balance
- NumOfProducts
- EstimatedSalary

HISTOGRAMS


CreditScore
2000
1000
0
400
600
800
Tenure
1500
1000
500
0
0.0
2.5
5.0
7.5
10.0
Balance
3000
2000
1000
0
0
100000
200000
NumOfProducts
4000
2000
0
1
2
3
4
EstimatedSalary
1000
500
0
0
50000
100000
150000
200000

Interpretation

A histogram is an effective way to display numerical data grouped into intervals. It provides a visual representation of data distribution, using bars to depict the frequency of values within each interval. The X-axis shows the intervals (bins), while the Y-axis represents the frequency of data falling into those bins.

Histogram of Credit Score is symmetric unimodal, as it has normal distribution

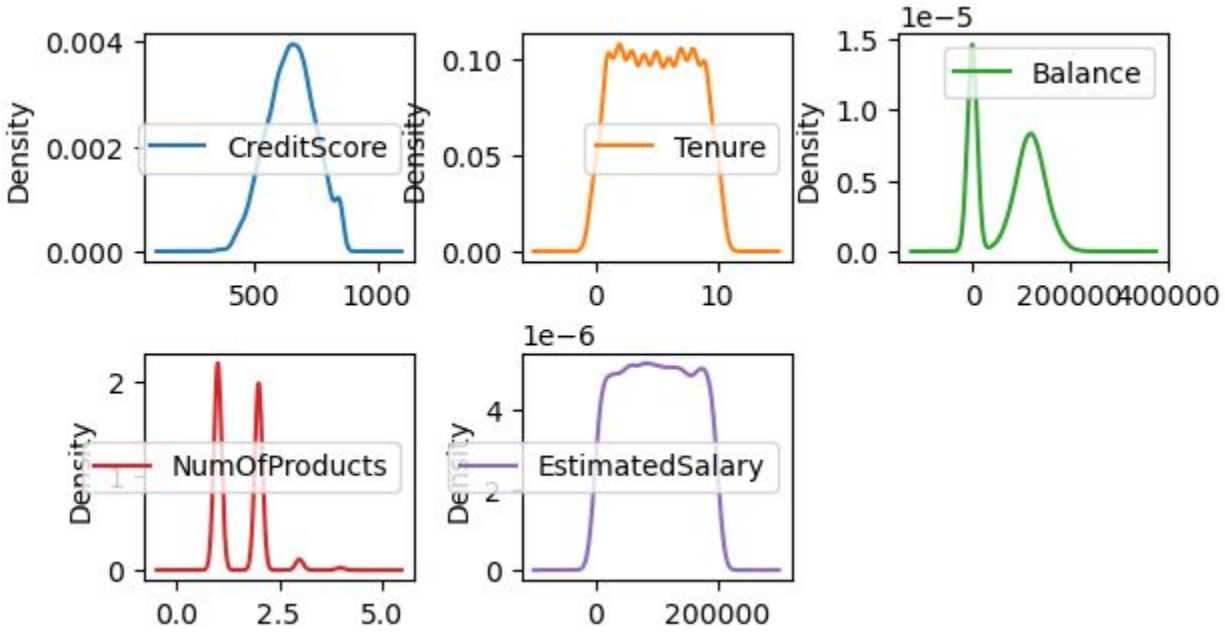
Histogram of Tenure is Uniform because every value in a specified range has equal probability of occurring

Histogram of Balance is symmetric with left skewed as majority of data is at centre but one outlier at left

Histogram of Estimated Salary is Uniform as it has no peaks or tail as it is same for entire time

Histogram of NumOfProducts is Discrete Distribution at most three products.

Density Plot



Interpretation

A density plot, a form of data visualization, employs 'kernel smoothing' to depict a smooth, continuous version of a histogram derived from the dataset.

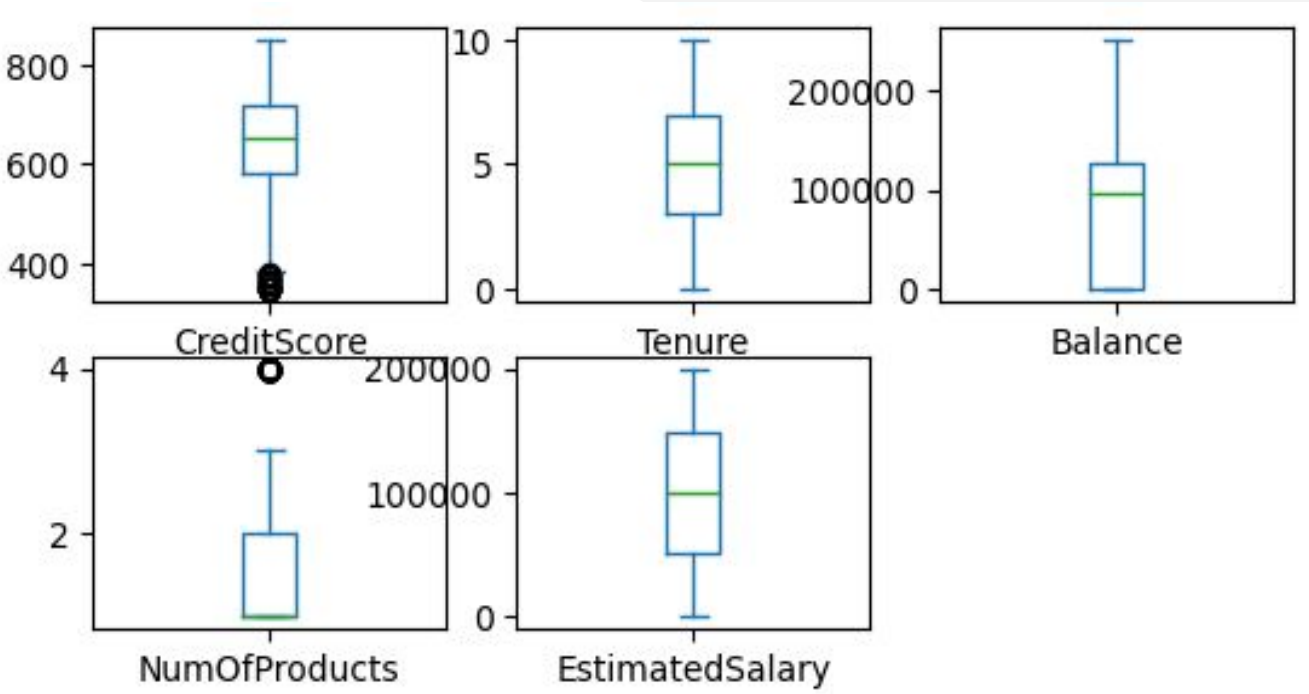
The x-axis represents the values of the specified columns:CreditScore, Tenure,Balance, NumOfProducts, and EstimatedSalary.

The y-axis represents the density of these values, indicating how frequent these values occurs in the dataset.

Each curve represents the density plot for a specific column, showing the distribution of values for the column. The areas under the curves represent the estimated probability density of the respective variables

The density plot provides insights into distribution and concentration of values for each specified variable,helping to understand their probability density and patterns within the dataset.

Box plot


800
600
400
CreditScore
10
5
0
Tenure
200000
100000
0
Balance
4
2
NumOfProducts
200000
100000
0
EstimatedSalary

INTERPRETATION

A Box Plot, often referred to as a Whisker plot, summarizes a dataset by displaying key properties: minimum, first quartile, median, third quartile, and maximum values. It uses a box to represent the interquartile range (from the first to the third quartile) and includes a vertical line at the median. The x-axis represents the data, while the y-axis shows the frequency distribution.

In credit score there are some outliers in the side of lower quartile and median exist far from at center of quartile and third quartile.

Four quartiles are equally distributed in Tenure, it has range of 10 starting from 0.

Balance and Number of products do not have lower values and no of products has median as lowest data point and starts from 0.

Alike Tenure, Estimatedsalary also has equally distributed values and range starts from 0 to 2 lakhs.

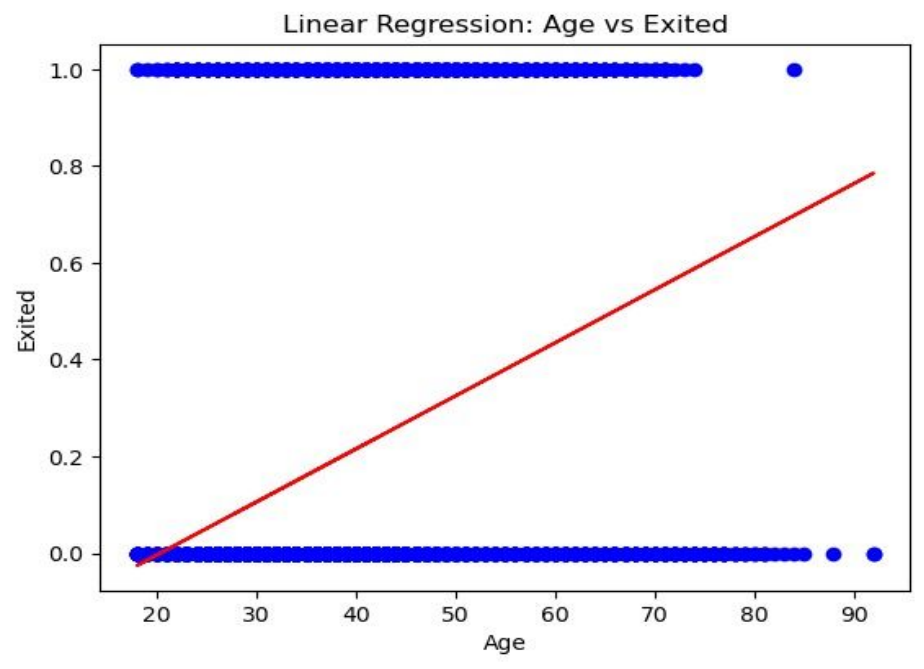
Code:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
from sklearn.linear_model import LogisticRegression
```

```
df = pd.read_csv('Churn_Modelling.csv')
X = df['Age'].values.reshape(-1, 1)
Y = df['Exited'].values.reshape(-1, 1)

regression_model = linear_model.LinearRegression()
regression_model.fit(X, Y)

y_pred =  regression_model.predict(X)

plt.figure
plt.scatter(X, Y, color='blue')
plt.plot(X, y_pred, color='red')
plt.xlabel('Age')
plt.ylabel('Exited')
plt.title('Linear Regression: Age vs Exited')
plt.show()

X1 = df[['Age']]
Y1 = df['Exited']
model = LogisticRegression(solver='liblinear')
model.fit(X1, Y1)
y_pred = model.predict(X1)
conf_matrix = confusion_matrix(Y1, y_pred)

print('confusion matrix \n',conf_matrix)
print()
print(regression_model.coef_)
print(regression_model.intercept_)
```

Confusion Matrix

7677	286
1964	73



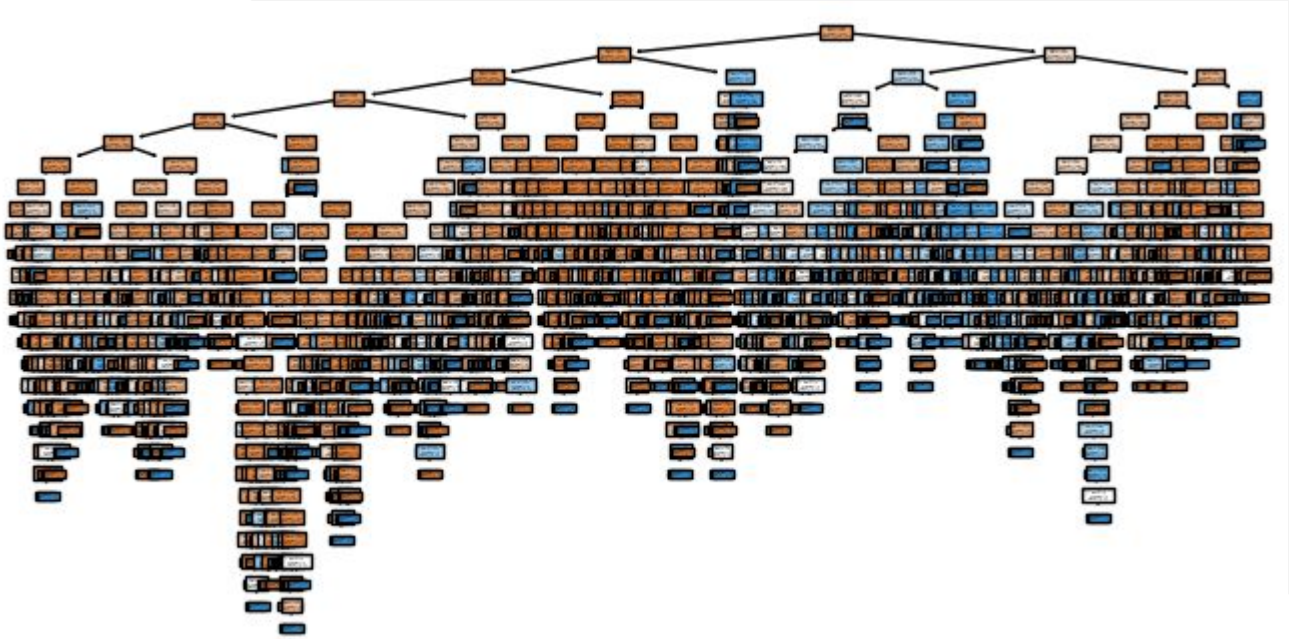
7677 are correctly predicted as 0 using x, 286 predictions are wrongly corrected as 1 when it is 0, 1964 are wrongly predicted as 0 when they are 1, and finally 73 correctly predicted as 1.

The linear regression plot determines that all the data is just distributed either at 0 or 1 means. People leave irrespective of age. And this output cannot be determined using linear regression. Linear equation for above line is y=0.01095741x-0.22278197

Code:

```
import pandas as pd
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
from sklearn.tree import plot_tree
```

```
churn_data = pd.read_csv('Churn_modelling.csv')
encoder = preprocessing.LabelEncoder()
churn_data['Geography'] = encoder.fit_transform(churn_data['Geography'])
churn_data['Gender'] = encoder.fit_transform(churn_data['Gender'])
churn_data
X = churn_data.iloc[:, 3:13]
y = churn_data.iloc[:, 13]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
Model = DecisionTreeClassifier().fit(X_train, y_train)
plt.figure(figsize=(8, 4))
plot_tree(Model, filled=True)
plt.show()
```



Here, the decision tree is very vast and got accuracy for testing data as 77%, So, Decision tree can satisfy the prediction of person whether he is gonna exit from bank or not with 0.77 probability. We can go with other models to improve accuracy furthermore. But Decision tree started its root nood gini value as 0.3263 which is relatively very less impure dataset, so, decision tree can respond very good for required prediction.