Analysis on churning Data Set

Prepared by: Nithish Reddy Kurapati

Student Number: 23005994

GitHub Repository:
https://github.com/nk23aco/Applied_data_science_4

The dataset analysed in this research consists of 10,000 rows and 14 columns, enabling a comprehensive analysis of different characteristics and their effects on customer behaviour across numerous variables in the dataset of banking sector. This study has identified several connections between these variables and their causes.

Within the dataset, there are fourteen columns are there, three contains categorical data such as Surname, Geography, and Gender, and the rest comprised of numerical data that include attributes like credit score, age, tenure, number of products, estimated salary, has credit card, is active member and exited. The analysis has excluded columns with many missing data points.

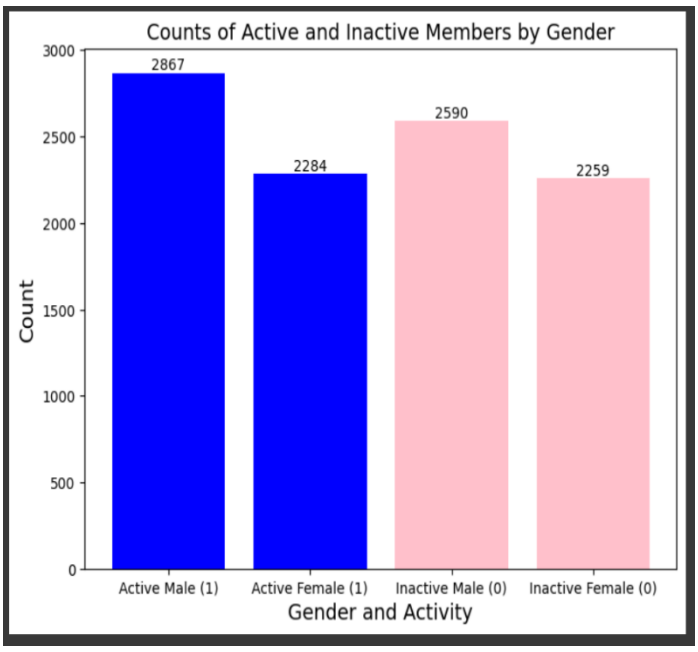

Figure 1. Bar Graph

It can be seen from Figure 1 that the bar graph displays the total amount of customers by activity and gender. The data indicates that 45.4% of consumers are female, while 54.6% of customers are male. Looking more closely reveals that there are 583 more active male clients than female consumers. There are 331 more inactive male clients than females. According to this graph, males account for most of both active and inactive consumers.

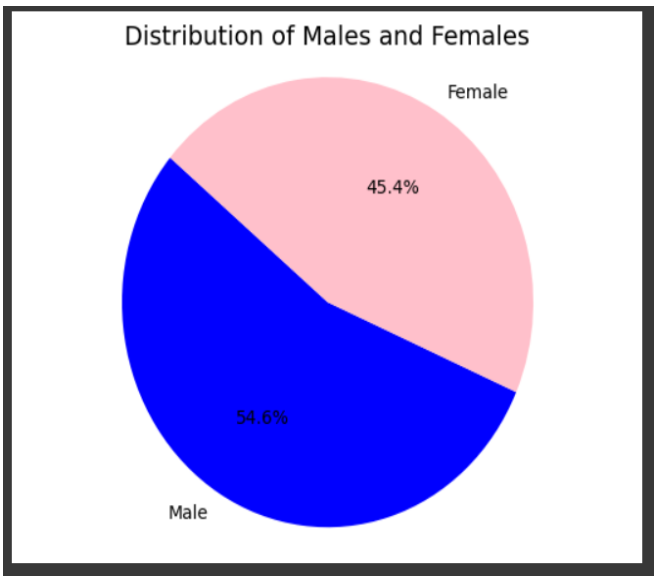

Figure 2. Pie Chart

Figure 2 shows that whereas 54.6% of the dataset's customers are men, 45.4% of them are female.

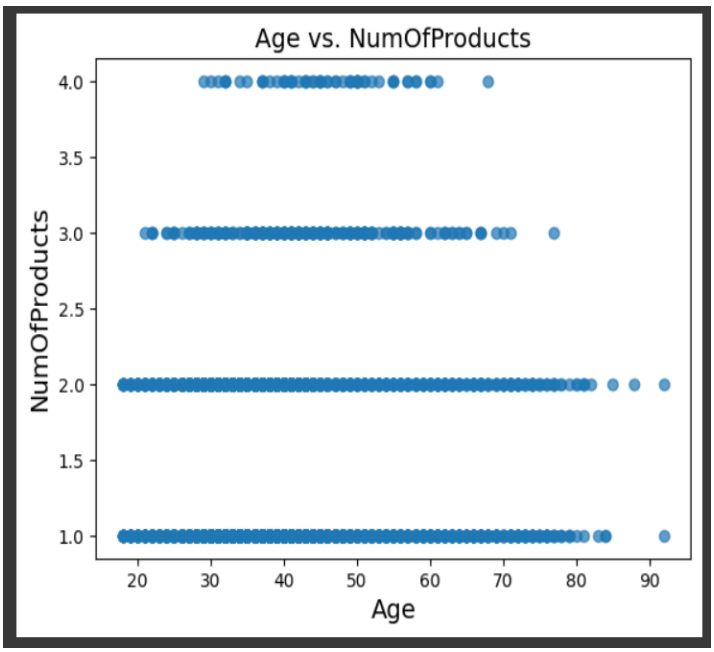

Figure 3. Scatter Plot

A scatter plot with the number of products on the y-axis and age on the x-axis is displayed in Figure 3. It is clear most customers across all age groups normally possess two products. The age range of customers who purchase three products, between 20 to 70, and those who acquire four items, ranging from thirty to seventy, has, however, clearly expanded. When examining the quantity of points and trends, most customers purchase two items, with very few selecting to purchase four.

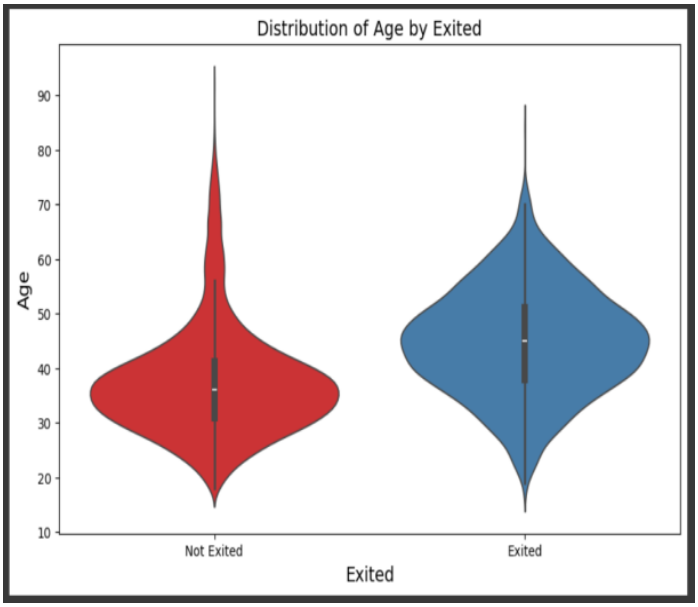

Figure 4. Violin plot

It is clear from the violin plot that more than half of leaving customers is between the ages of 40 and 50. In contrast, more clients stick with the bank longer when they are around their ages of 30 and 40. Customers who remain have an average age of 35, and those who have left have an average age of 45. Moreover, the whiskers of departing consumers are longer than those of remaining customers.

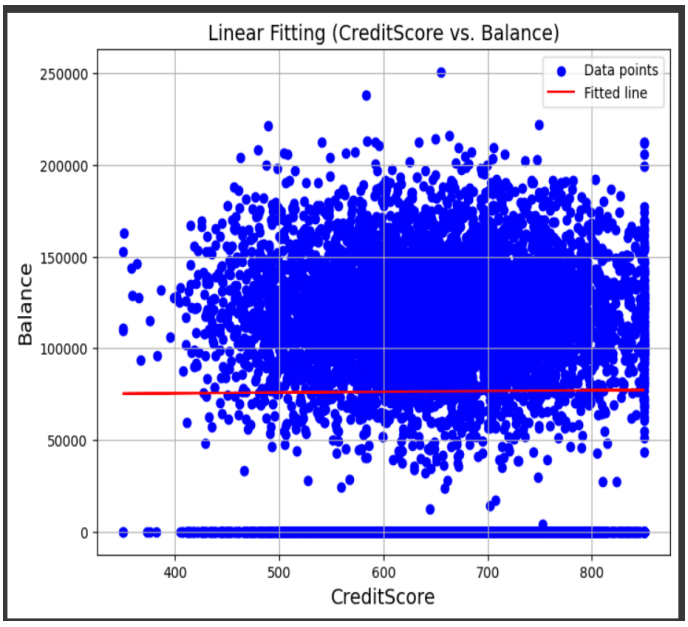

Figure 5: Line fitting

From the above figure, we can notice that we plotted balance on y-axis and credit score on x-axis. In the line fitting, that data points of credit score and balance are spread acoustically in the plot. Points are plotted in between credit 400 to 600 largely in range of balance 50,000 to 1.75 lakh. The best fit line in the scatter plot of the data points passes in at 75,000 rupees.

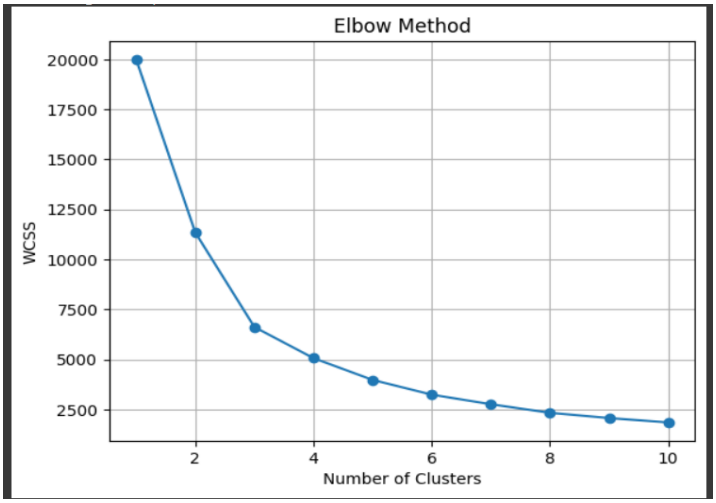

Figure 6. Elbow method

From above elbow method, we used to find the k-value as an optimal value to perform clustering. We are plotting, within-cluster sum of squares on y-axis and number of clusters on x-axis. I have chosen optimal value as k=4 based on the performance. if Number of clusters increases, then predominantly the WCSS decreases.

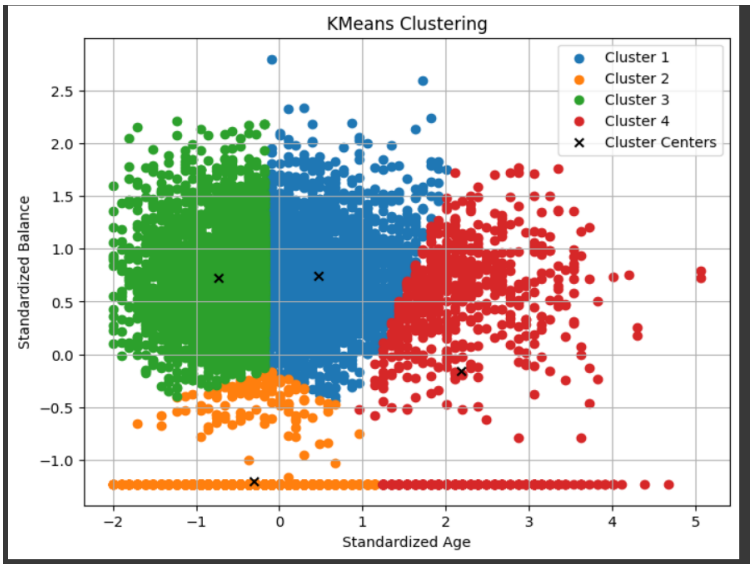

Figure 9: K-means clustering

The elbow approach determines that four clusters are the ideal number. As a result, there are four clusters in the dataset. Following standardisation, data points with comparable behaviours are combined into a single cluster. We can find the behaviour of the data points by using distance metric formulae. If the data points are close to each other, it means it is one cluster and given a single colour visualisation. The remaining data points, representing separate clusters, are given distinct colours to denote their respective groups.

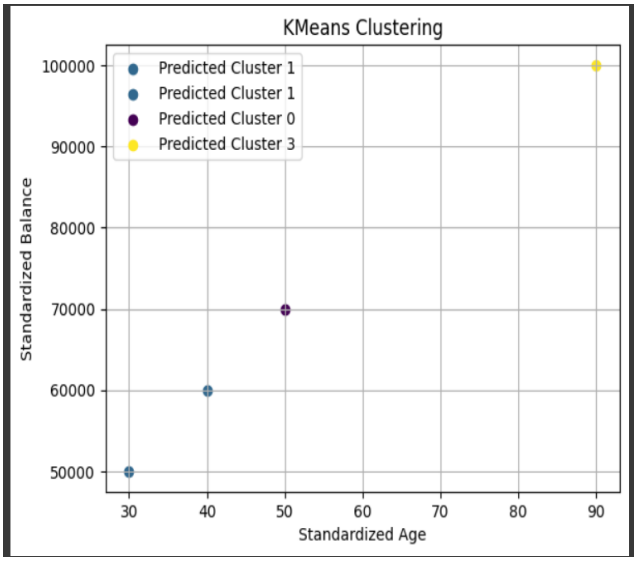

Figure 10: Predicted model for new data points.

I take just a handful of data points to predict the model's cluster of origins. [30, 50000], [40, 60000], [50, 70000], and [90, 100000] are the data points. Two of the data points are predicted by our model to belong to cluster 1 (blue in colour), one to cluster 0 (violet in colour), and one to cluster 3 (yellow in colour).

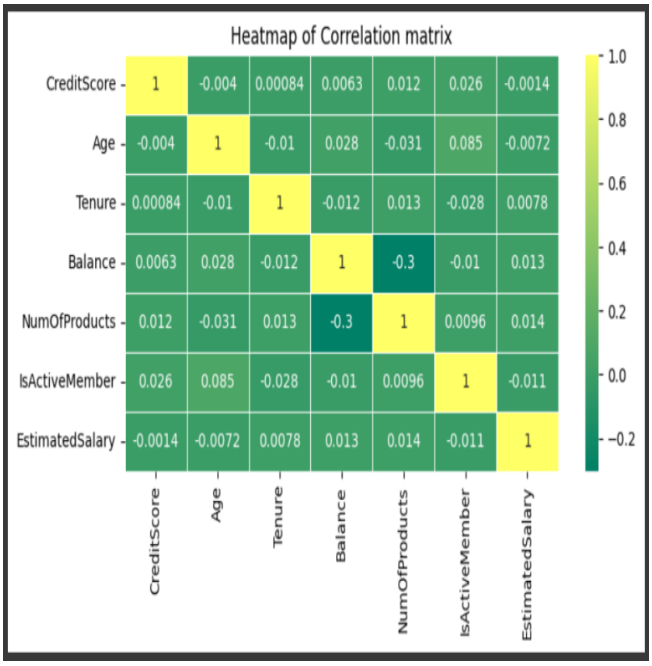

Figure 11: Heatmap

Statistical Methods:

Mean() :

Upon computation, we have found that the average credit score of bank customers is 650.52, their average age is 38.9, their average account life is 5, they maintain an average of 1.5 products, half of them is active, their average estimated salary is 100090.23, and they maintain an average balance of 76,485.88.

Median() :

Based on my analysis, we know that half of the customers are having age more than 37 years and rest of them are below 37 years, maximum life of bank account is 5, balance is 97198, the estimated salary is 100193, number of products is 1, credit score is 652.

Standard Deviation:

The average of credit score, age, tenure, balance, is active member is more than standard deviation, it makes sense us that the data points are close to the mean of the respective variables.

The standard deviation of balance, number of products, estimated salary is higher than the average it means the data points are not too close each other(dispersion of data points are largely dispersed).

Skewness:

Age, number of products, are positively distributed i.e. greater than 0, it shifts towards right skewness and credit score, balance, is active member is less than zero which means it shifts towards left skewness. Estimated salary and tenure is nearly symmetrical distributed, its value is close to (0.002 and 0.01 ≅ 0). towards to right skewness.

Kurtosis:

Features such as age and the number of products have data points with outliers, which indicates that the data points are not distributed normally. The remaining features have negative values, indicating that the peak is weaker than the normal distribution.

Describe ():

From the describe(), we can analyse the dispersion of the data points on the plot. We can understand the data by using describe method. The describe() method returns statistical summary values such as count, percentiles, mean, median, and range.

Corr ():

There is a positive correlation between credit score against tenure, balance, number of products, is active member.

There is a negative correlation between age against credit score, tenure, number of products, estimated salary.

There is a strongly correlation between balance and number of products