***Abstract*- This project aims to do Exploratory Data Analysis by using the "Predicting Churn for Bank Customers" dataset from Kaggle, which includes information on 10,000 bank customers, to identify customers likely to churn and determine the key factors that influence their decision to leave the bank.**

Keywords—Churn Prediction, Correlation Matrix, Gradient Boosting, Confusion Matrix

I. Introduction

Customer churn is a significant problem for businesses, including the banking industry. The "Predicting Churn for Bank Customers" dataset from Kaggle provides valuable data to analyze the factors contributing to customer churn in the banking industry. This project uses the dataset to develop a predictive model to identify customers likely to churn and uncover the key factors driving their decision to leave. The insights generated from this study can help banks to better understand their customers and develop effective retention strategies.

II. Data Description

The dataset I used for this project is a Customer Churn dataset collected from Kaggle. This dataset consists of both numerical and categorical data with 10,000 rows and 14 attributes, as given below:

a) RowNumber: Index for the records
b) CustomerId: Unique identifier for each customer
c) Surname: Customer's last name
d) CreditScore: Customer's credit score
e) Geography: Customer's location (France, Germany, or Spain)
f) Gender: Customer's gender (Male or Female)
g) Age: Customer's age
h) Tenure: Number of years the customer has been with the bank
i) Balance: Customer's account balance
j) NumOfProducts: Number of bank products the customer is using.
k) HasCrCard: Whether the customer has a credit card (1) or not (0)
l) IsActiveMember: Whether the customer is an active member (1) or not (0)
m) EstimatedSalary: Estimated salary of the customer
n) Exited: Whether the customer has left the bank (1) or not (0)

III. Data Cleaning and Preprocessing

Before performing Exploratory Data Analysis, I first looked into the statistical values of the data in order to get insights like mean, major quantiles data, min-max values along with the standard deviation and variance for each numerical column in the dataset with the help of describe() function. This dataset does not have any missing values.

IV. Exploratory Data Analysis

I performed exploratory data analysis to identify patterns and trends in the customer churn data. Here, among 10000 customers, 5457 are male and 4543 are female.

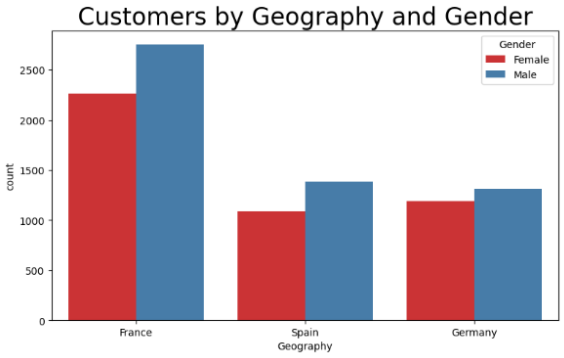

Fig. 1. Customers by Geography and Gender

From the above graph, we can see that most of the customers are from France. Among the 3 countries, all of them have a higher proportion of male customers in comparison.

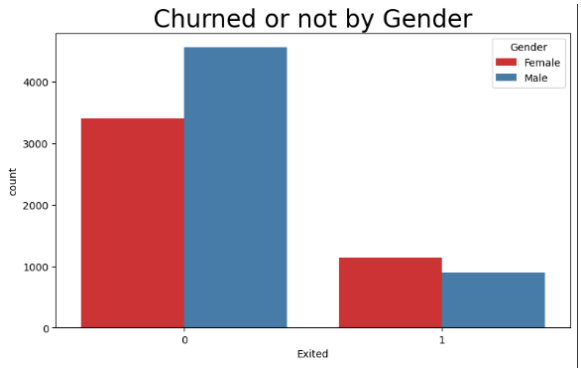

Fig. 2. Churned or not based on Gender

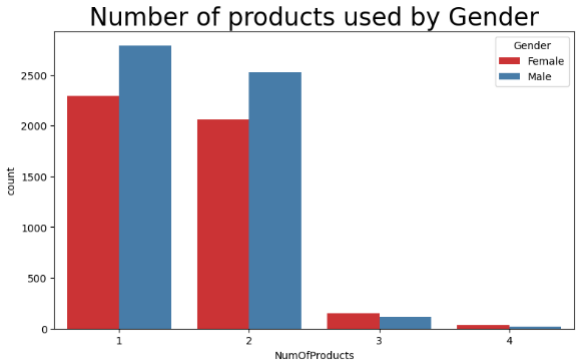

Fig. 3. Number of products based on Gender

From Fig 2, among the two genders, we can observe that female customer churning is more significant in comparison. Similarly, in Fig 3, most of the customers are using 1 and 2 number of products.

After that, I examined the distribution of the variables in the dataset using histograms and box plots. I found that the age and estimated salary variables are normally distributed, while the remaining variables have skewed distributions.

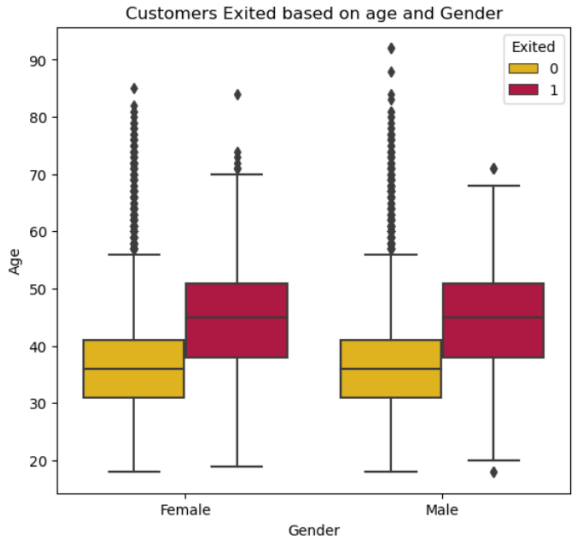

Fig. 4. Churned or not based on Gender

After further analysis, from the above box plot in Fig 4, we can say that the customers churned for each gender belong to the almost same age group between 38-52 years.

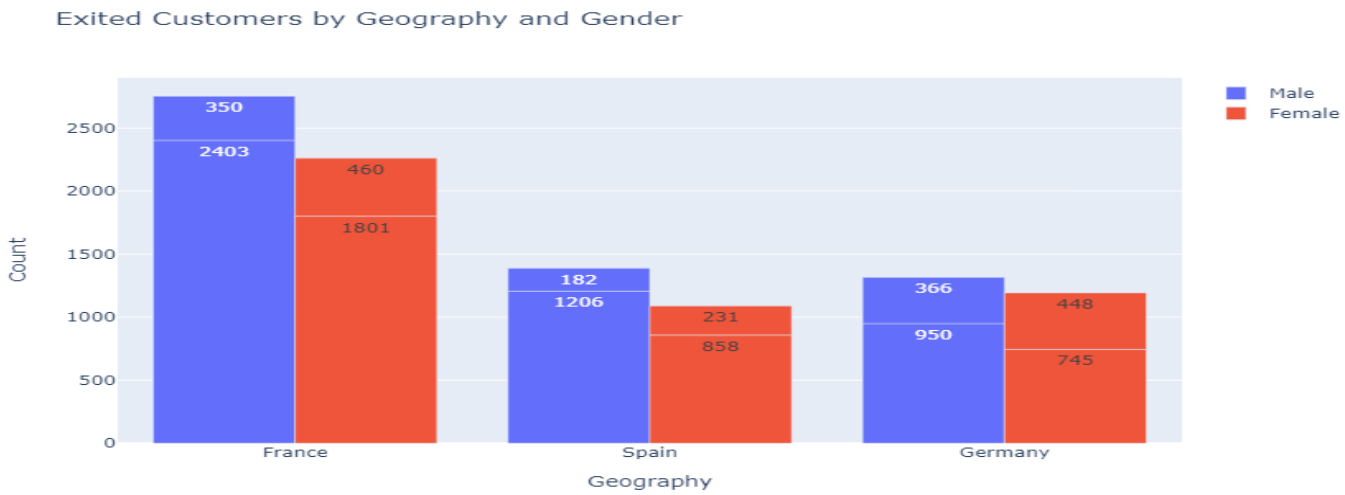

Fig. 5. Exited customers based on Gender

From the above Fig 5, we can clearly see that, although most of the customers are from France, the number of exited customers is higher in Germany, which is 366 for males and 448 for females. Overall, females tend to leave the bank in comparison to males.

53.3% of non-churn customers, and those with one product are 46.2%. Generally, the customers using four products are found to be churned more.

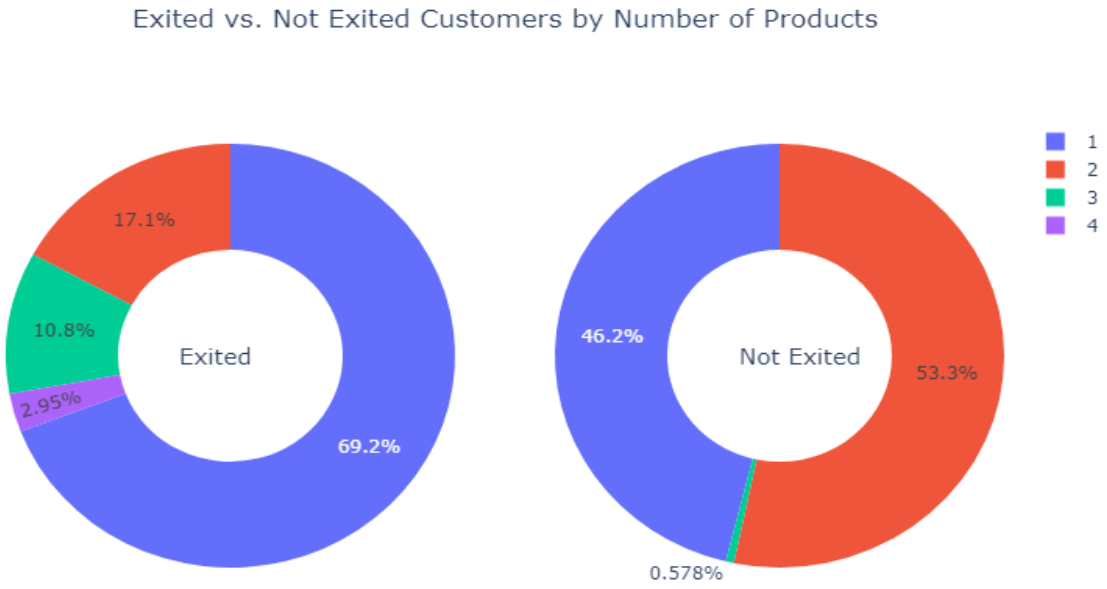

Fig. 6. Exited Vs Non-Exited Customers by Number of Products

Here, the above Fig 6 pie plot depicts that among customers who churn, the percentage of those who use only one product is relatively high at 69.2%, followed by those who use two at 17.1%, three at 10.8%, and four at 2.95%. Customers with two items comprise

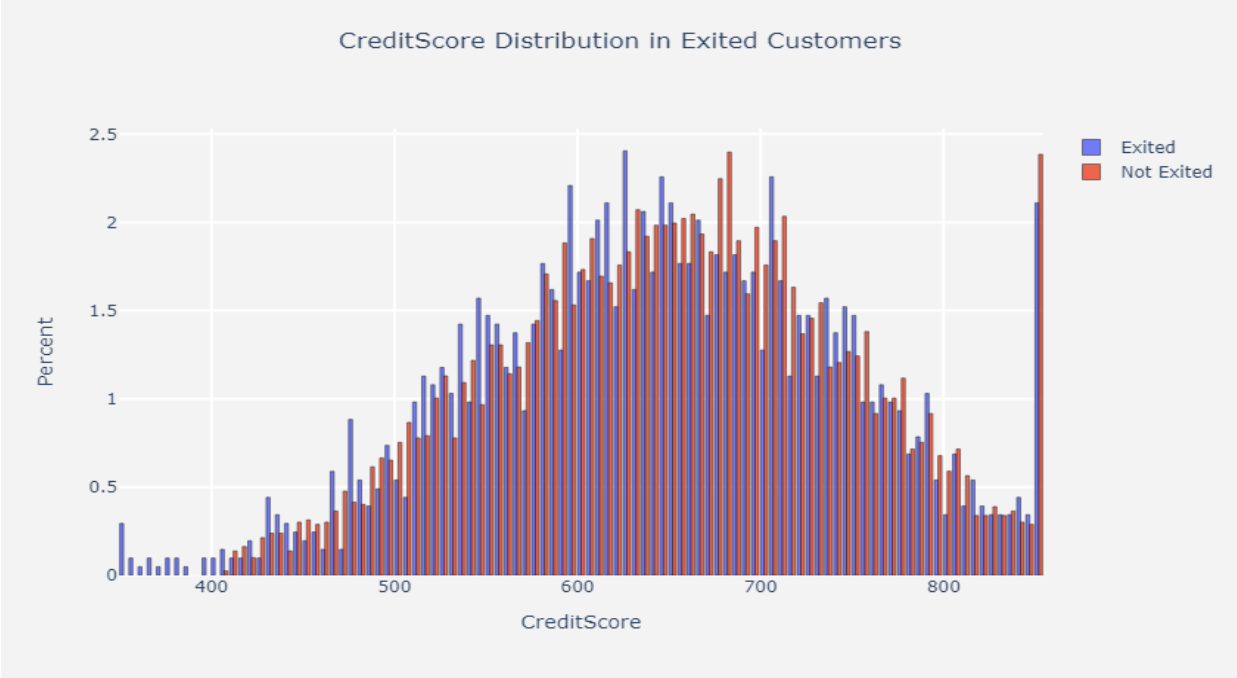

Fig. 7. Exited Vs Non-Exited Customers by Number of Products

Here, the maximum percentage (2.4 %) of churned customers has a credit score between 625 and 629. Looking thoroughly through this graph, we see that most of the churned customer's credit score lies 590-710.

From the above Age Distribution, we can see that the maximum age of the churned customer is 46. Customers leaving the company have an age group of 38 to 52 years.

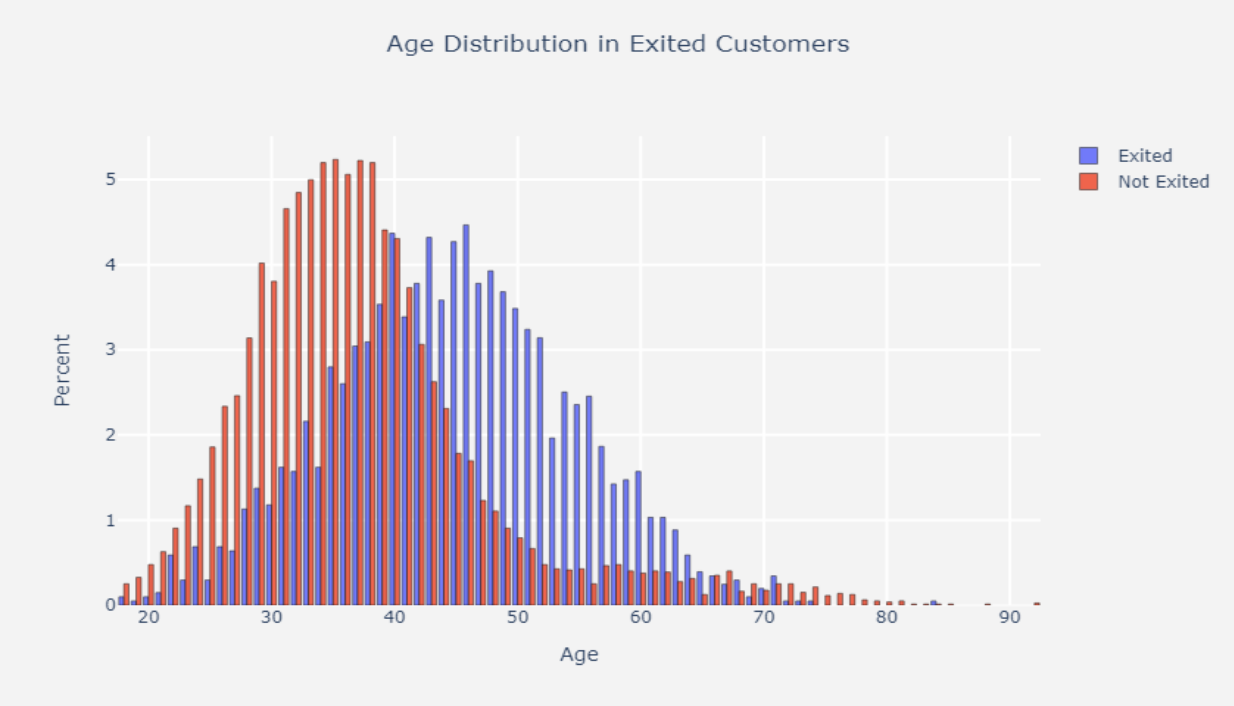

Fig. 8. Age Distribution in Exited Customers

V. Outlier Handling

Another step of our project is handling outliers that are present in the dataset.

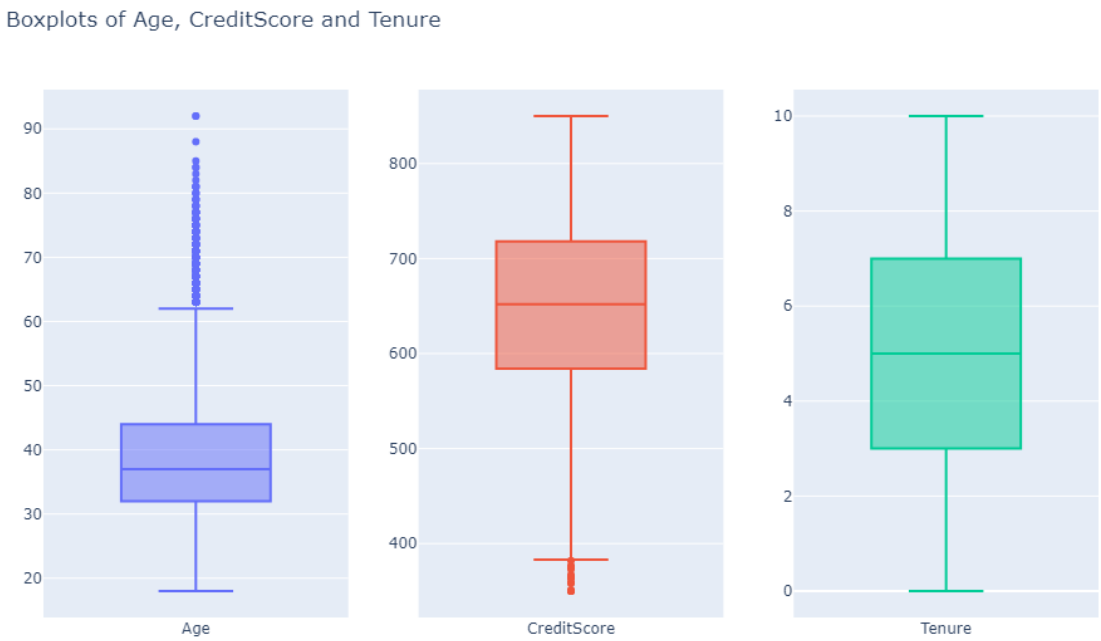

Fig. 9. Visualization of outliers in different columns

For all the numerical columns, Outliers can only be seen in Age and CreditScore columns. Although there are a few outliers in the age column, they might not be considered outliers as there are customers aged 80,85 and 92.

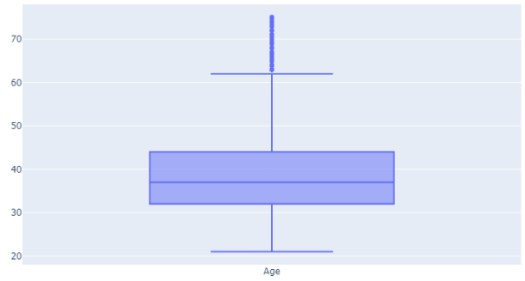

Fig. 10. Age column after removing outliers

So, we performed the flooring and capping method for handling the outliers, as shown in Fig 10, we used quantile by replacing the age lower than 21 with 21 and greater than 75 with 75.

VI. Feature Selection

After removing outliers, I removed some of the columns and only selected relevant features for the model. The surname column has 2932 unique values, meaning the customers belong to the same family group. Although it seems unique, we don't need this column. Removing 3 columns RowNumber, CustomerId and Surname, as they are irrelevant to our analysis. For selecting the crucial features, we used a random forest classifier and got the result below:

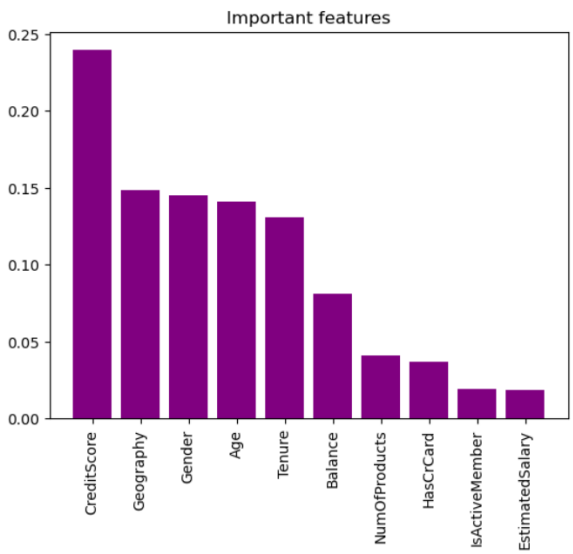

Fig. 11. Important Features

VII. Model Training

Since this dataset also consists of categorical columns, we did one hot encoding by using a label encoder for those columns. Some features are not of the expected data type. Specifically, NumOfProducts, HasCrCard, IsActiveMember, and Exited are all represented as object data types, which can cause issues when using specific machine learning algorithms. So, we changed the data types of those columns to numerical.

Before training the model, we split the dataset into 80 % for training and 20% for testing. The machine learning algorithm we used is Gradient Boosting Algorithm.

VIII. Performance Evaluation

For evaluating the performance of our model, we have used Confusion Matrix as given below:

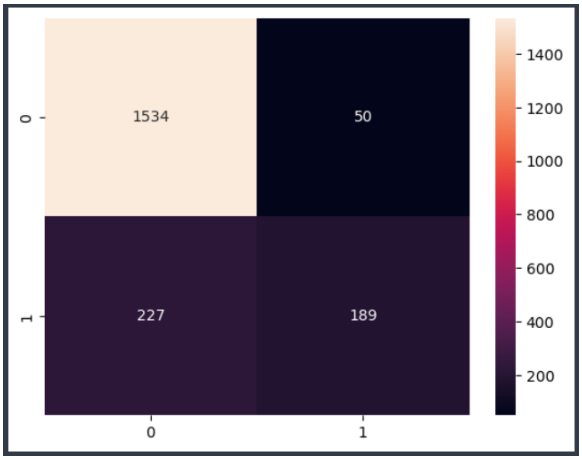

Fig. 12. Visual Representation of Confusion Matrix

Hence, From the confusion matrix of the Gradient Boost model, it can be seen that the True positive value is 1534, the True negative value is 189, the False positive value is 50, and the false negative value is 227. [2] The model's performance is measured by determining the accuracy, precision, recall and F1 score, which is given below:

	precision	recall	f1-score	support
0	0.88	0.96	0.92	1618
1	0.72	0.46	0.56	382
accuracy			0.86	2000
macro avg	0.80	0.71	0.74	2000
weighted avg	0.85	0.86	0.85	2000

Fig. 12. Classification Report

The accuracy of the Gradient Boosting model is 86.15 %.

IX. Conclusion

In this report, I presented the findings of our EDA on customer churn data. I identified several factors contributing to customer churn, including the customer's tenure, whether the customer has a credit card, and whether the customer is an active member. In this project, I focused on Exploratory Data Analysis and did data visualization to know the behaviour of customers and what is the primary factor in determining churned customers. I analyzed and extracted different patterns from the dataset.

a) Gender is not playing any role in customers leaving the company
b) It is the age group of customers between 40-50 years of age who are churned
c) Although there are a higher number of customers from France, the proportion of customers leaving the company is from Germany.

The insights gained from this analysis can help businesses take appropriate measures to reduce customer churn and improve customer retention.

REFERENCES

[1] Adam Mause. (2019). Predicting Churn for Bank Customers [Data set]. Retrieved April 21, 2023, from https://www.kaggle.com/datasets/adammaus/predicting-churn-for-bank-customers

[2] S. Narkhede, Understanding Confusion Matrix, May 2018 DOI: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd6