Churn for Bank Customers

Data Science Capstone Project

Anmi Chou Shigeta

Overview

In this project, the focus is set on predicting whether a specific customer will continue to use the bank's services or not. This allows the bank to determine the factors that lead to customers leaving their services for other financial services, and an in-depth analysis can help the financial institutions in retaining the customers.

The dataset provided had 10,000 rows and 14 attributes such as CreditScore, Gender, Age, Tenure, EstimatedSalary, and more. A classifier is being built to determine which customers will Exit and which will not.

Problem Statement

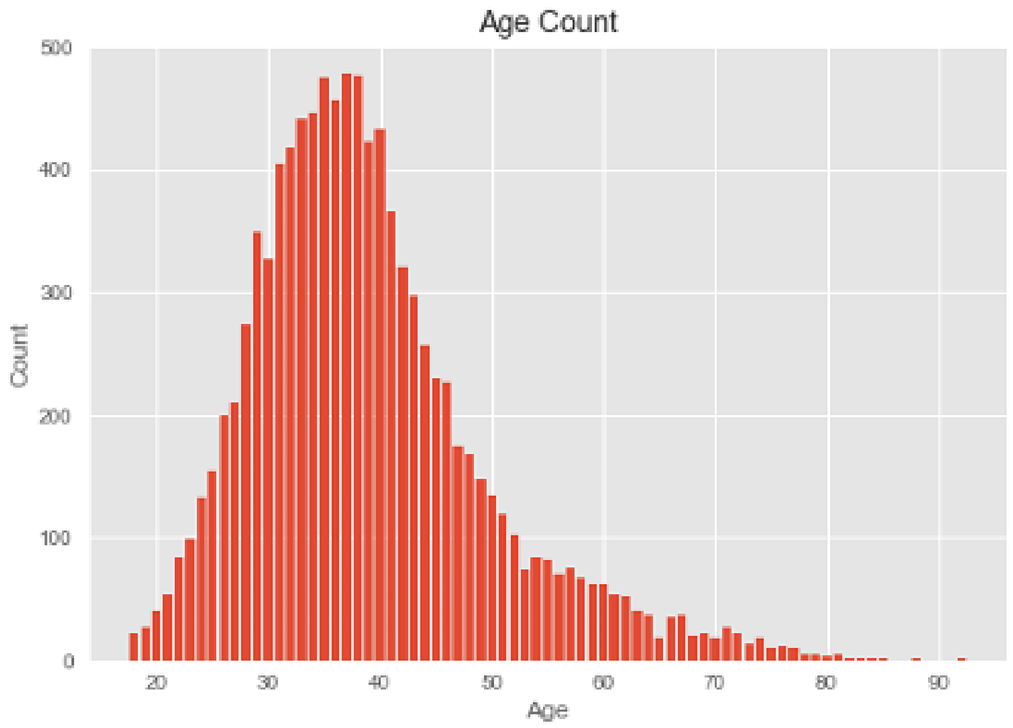
Provide a data-driven solution to predict whether a customer will churn.

Data Cleaning

The dataset consisted of a few unnecessary columns such as RowNumber, CustomerID, and Surname which added little value in classifying the customers. These were dropped, as these only added noise to the dataset. It was also checked whether the dataset had any missing values, which it did not have.

The next task was to categorise the numerical and categorical variables. The categorical variables were Geography and Gender, which were both label encoded. This data was then passed to the classifier, with the ground truth being the “Exited” attribute and all the remaining columns being the inputs.

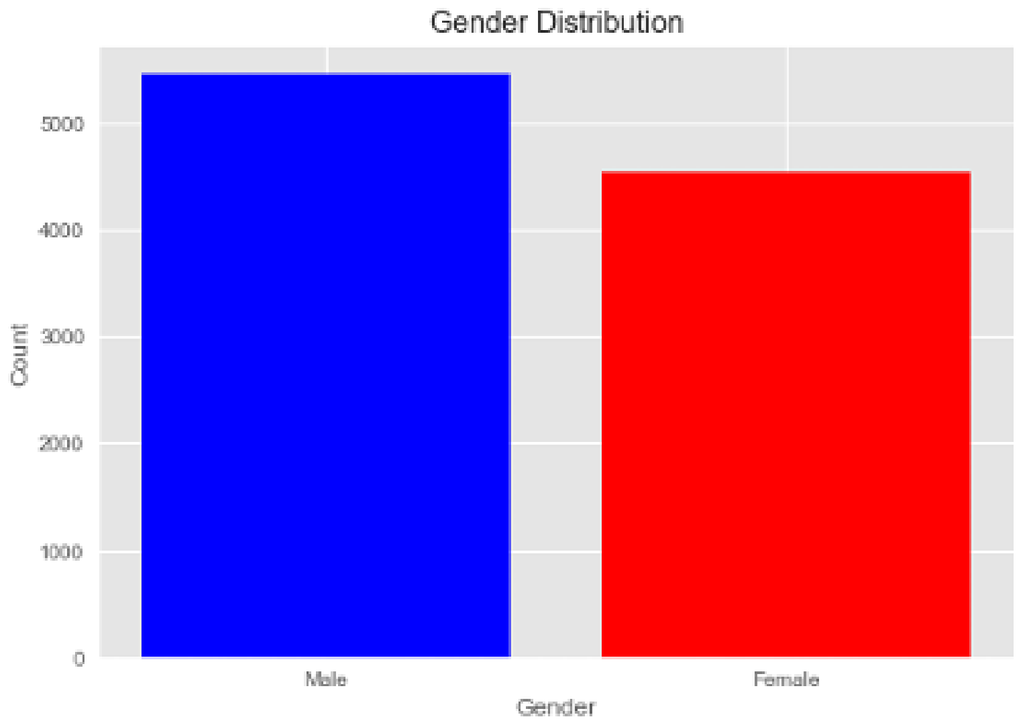
Findings from EDA



Age

- Majority of users of this financial service are between the ages 25 to 50. The highest number of customers are from age 36.

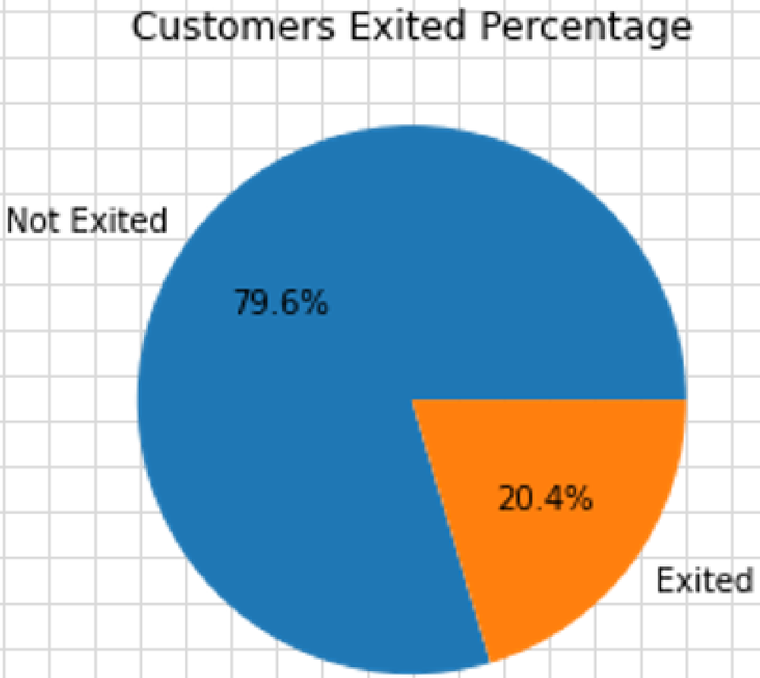
Findings from EDA

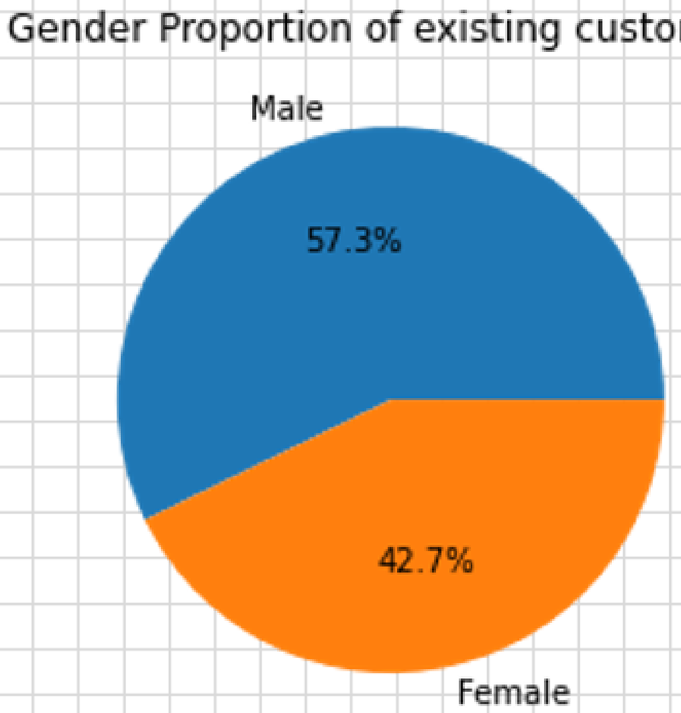


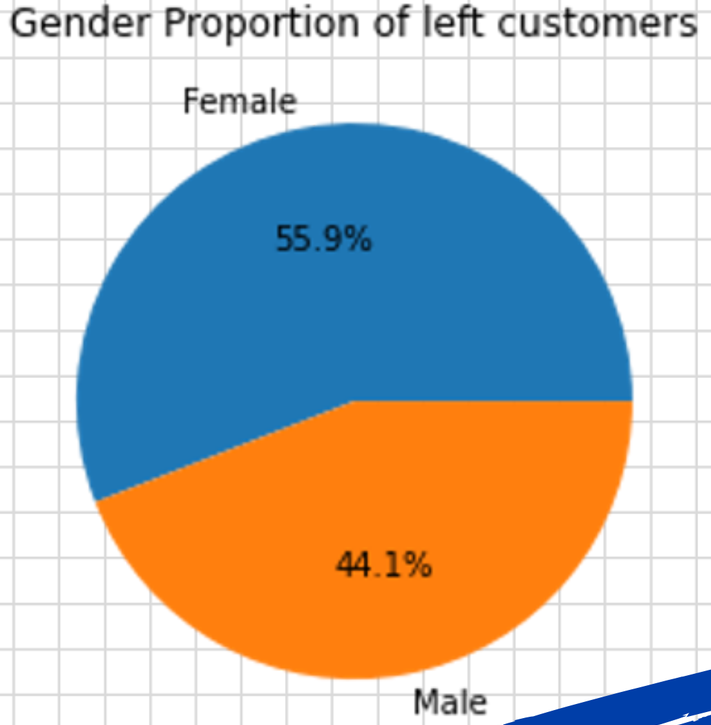
Gender

- The proportion of male to female customers is almost the same – 54.6% Males to 45.4% Females.
- In the provided dataset, most of the customers have been retained (79.6%).
- In the customers retained, the ratio between male to female is similar to the original data.
- However, in the customers left, an insight can be seen that most of them are female (55.9%), even though the total female population is lesser than the male.

Findings from EDA





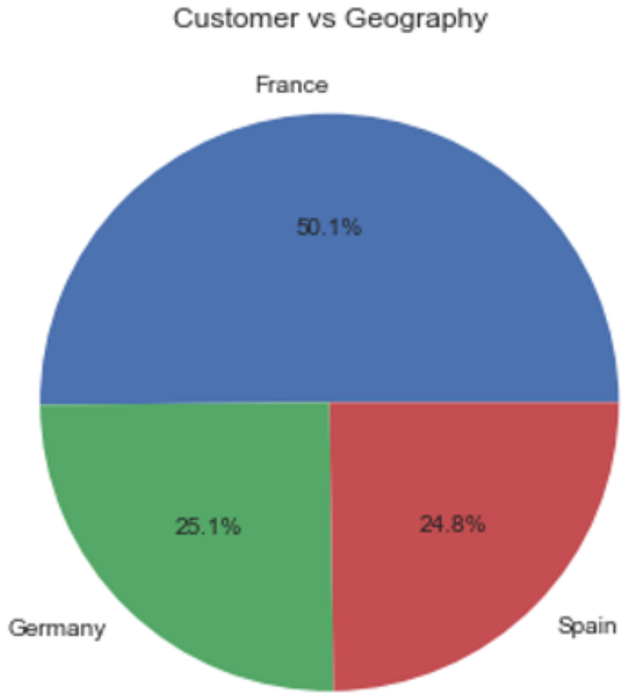


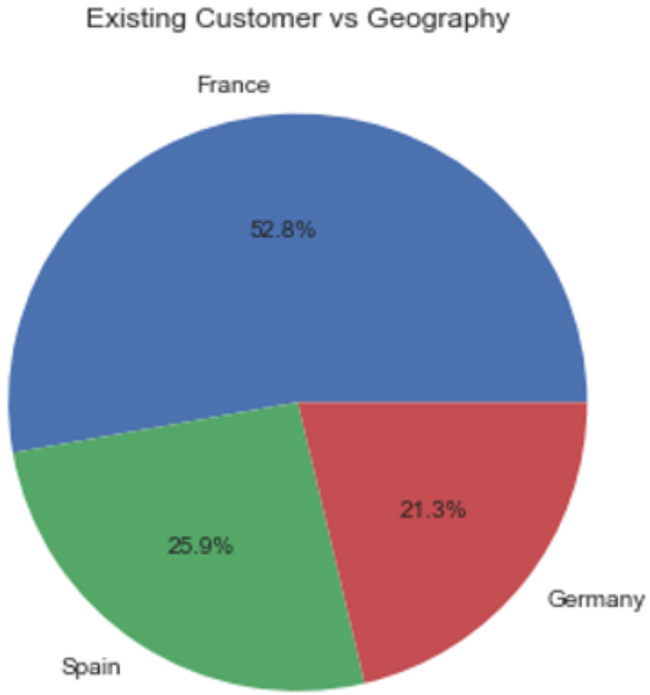
Findings from EDA

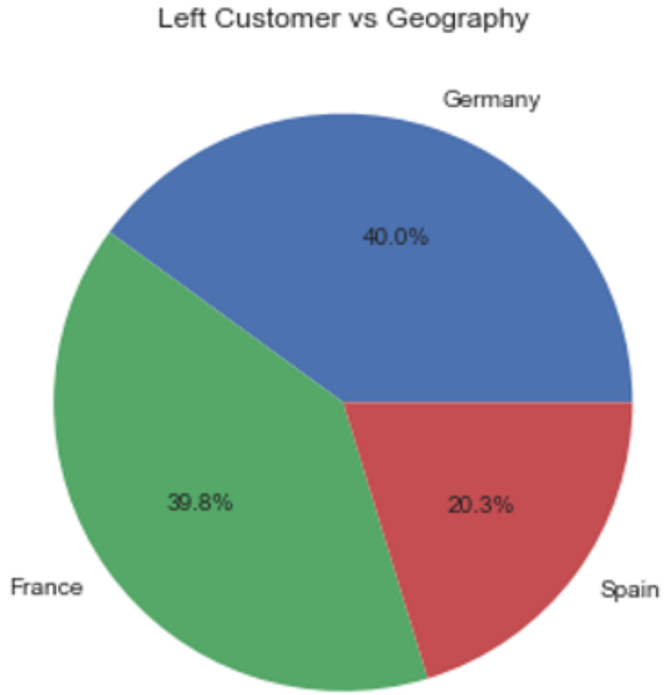
Geography

- The customers of the database are primarily from three geographical regions: France, Germany, and Spain.
- 50.1% of them are from France while the remaining are more or less evenly split between Germany and Spain.
- To develop further insights, the data was separated between existing customers and leaving customers and the results were plotted segregated on region.
- An interesting fact is that while only a quarter of the original customers are from Germany, the leaving customers are dominantly tied between Germany (40.0%) and France (39.8%).
- For the existing customers, it can be seen that Germany is lagging behind, which indicates that customers belonging from Germany tend not to prefer this bank.

Findings from EDA





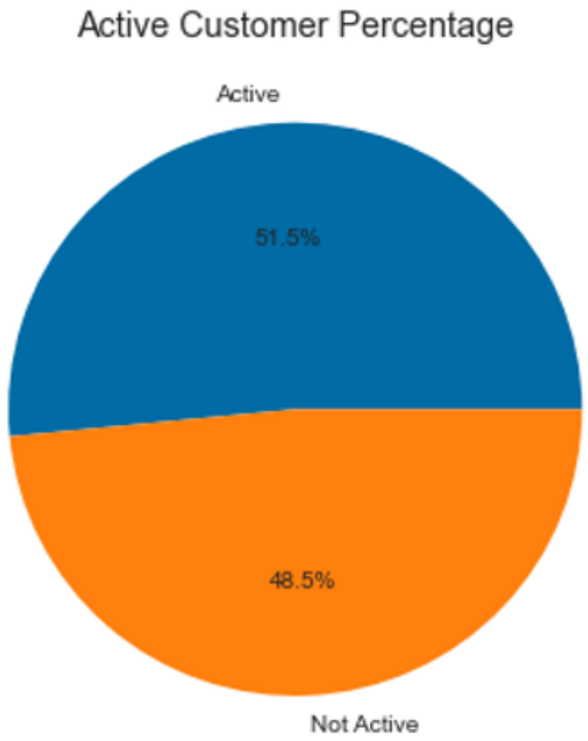


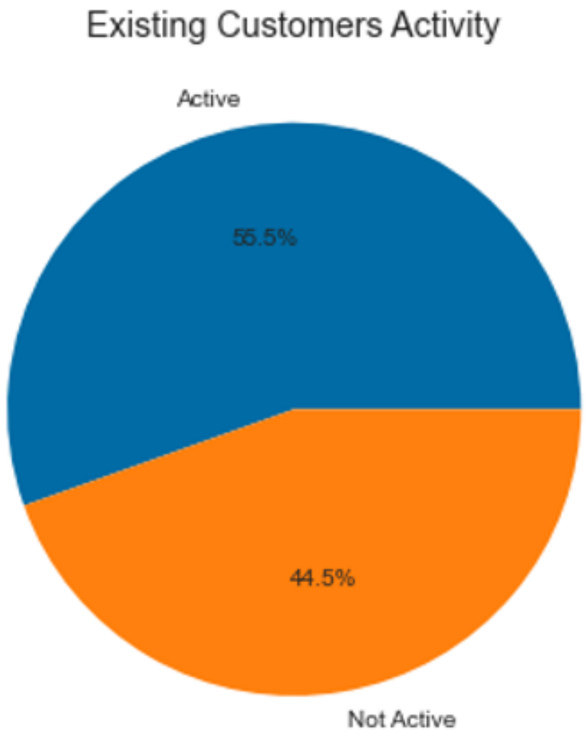
Findings from EDA

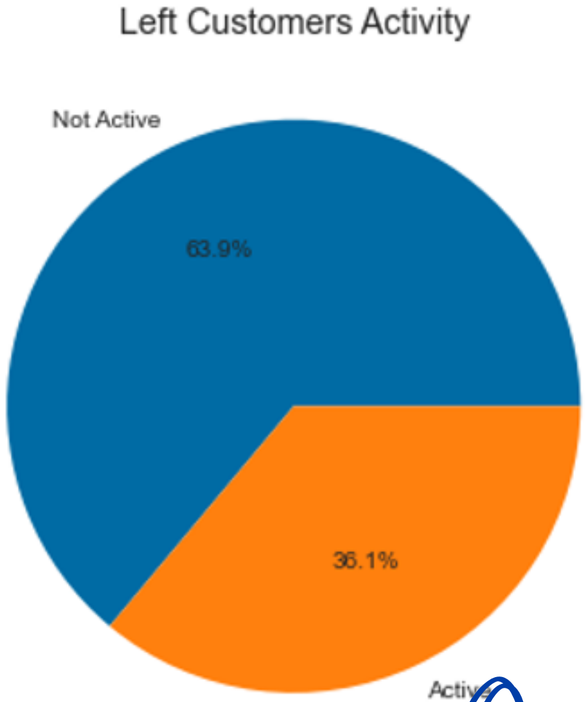
Activity

- This insight is derived from the user activity – whether they are active customers or not.
- The total active customers are similar to the inactive customers (51.5% to 48.5%).
- Intuitively, it can be inferred that active customers are more likely to stay loyal to the bank while inactive customers tend to leave.
- From the database, 55.5% of existing customers are active while 44.5% are inactive.
- In case of leaving customers, a more drastic change can be observed. 63.9% customers are inactive while 36.1% are active.

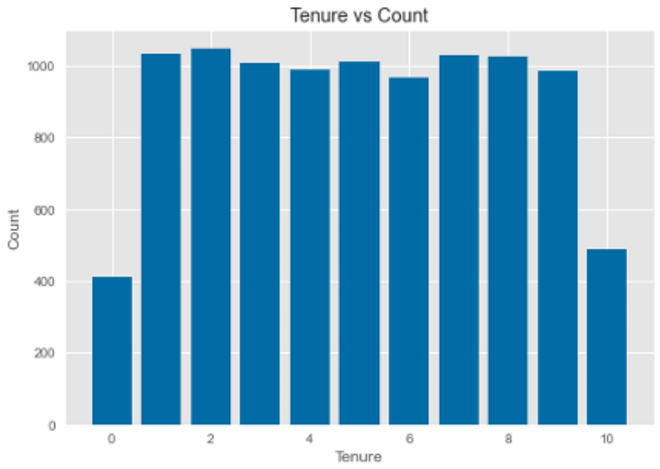
Findings from EDA

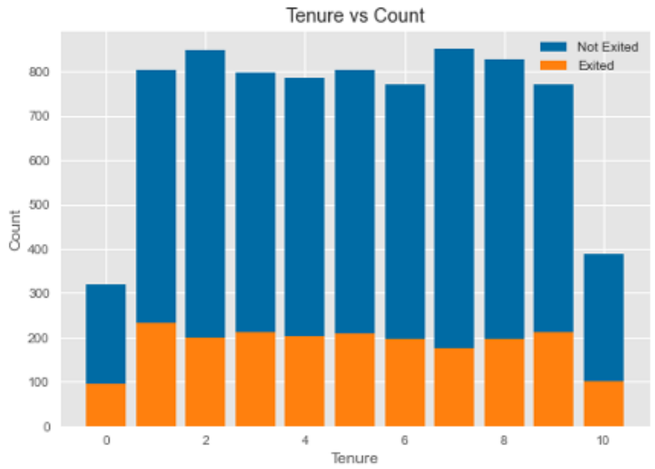






Findings from EDA

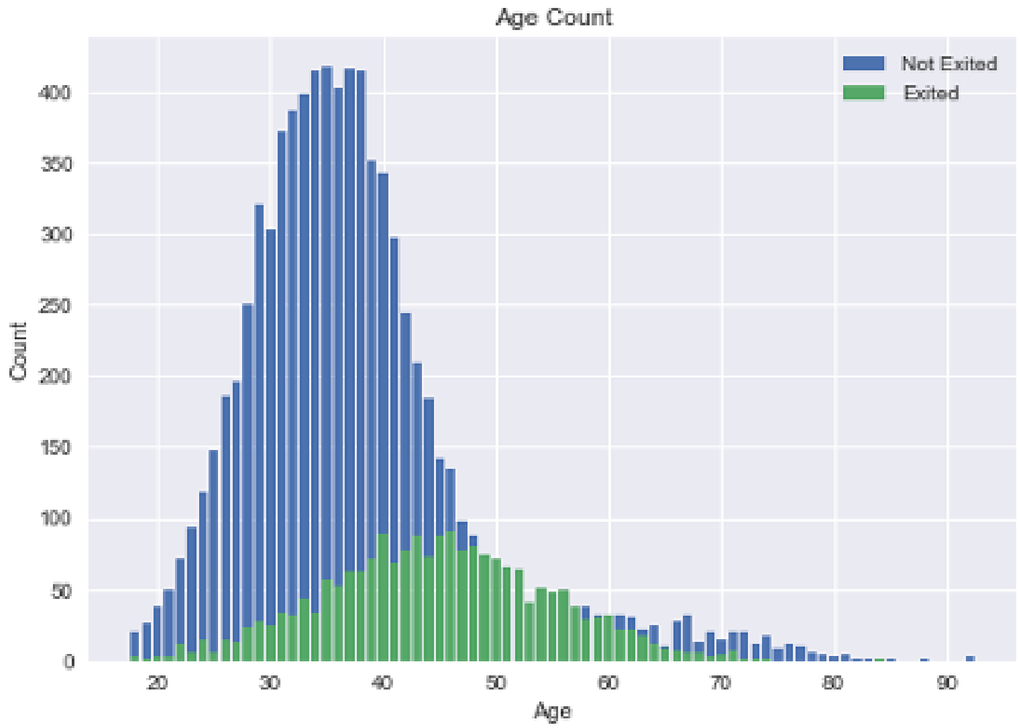




Tenure

- The data was grouped on the tenure i.e. the number of years customers have been using services from the bank.
- It shows that mostly the customers are spread somewhat evenly between years 1-9 with least customer at year 0 followed by year 10.
- The percentage of customers leaving compared to staying (ratio between heights of bar charts) remains the same with years.

Findings from EDA



Distribution

- The age distribution of both existing and attrited customers seem to follow a similar distribution spread.

Classification Model: KNN Neighbours Classifier vs RandomForest Classifier

- ·As the data was dependent on multiple factors, the conventional classifiers such as Logistic Regression and Linear SVM, may not perform well on it. To deal with this problem, the initial approach was using a KNN Neighbours classifier.

- ·After reviewing the results of the KNN Neighbours classifier, it was evident that the dataset could perform better on a classifier that deals with complex attributes and multiple dependencies. This is why RandomForest Classifier was chosen, and the results showed an improvement.

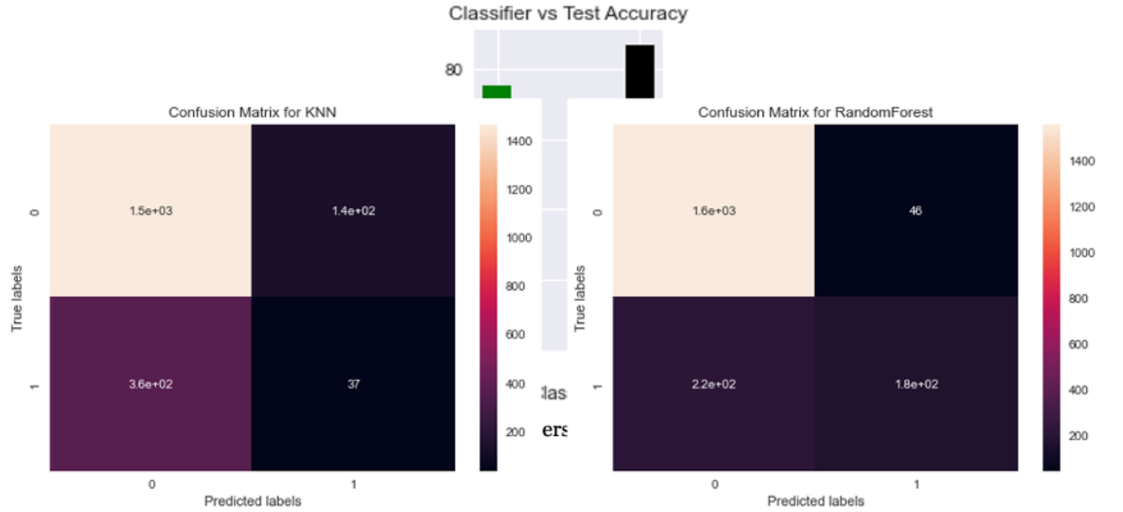
- The KNN Neighbours had a 75.2% test accuracy while a RandomForest Classifier had a test accuracy of 86.8%. Test accuracy here refers to the number of correctly predicted labels divided by total number of labels.

Classification Model: KNN Neighbours Classifier vs RandomForest Classifier

- ·Since a class imbalance existed, the F1-score was also calculated to see how well the classifier performed on individual classes. The KNN had a F1-score of 0.86 and 0.13 while the RandomForest Classifier had a F1-score of 0.92 and 0.57.

- Thus, it was proven that RandomForest Classifiers had an overall better performance.

- The accuracy for both classifiers can be compared as:



Classification Model: KNN Neighbours Classifier vs RandomForest Classifier

The classification report of both the classifier is as follows:

KNN Classifier

	precision	recall	f1-score	support
0	0.80	0.91	0.86	1607
1	0.21	0.09	0.13	393
accuracy			0.75	2000
macro avg	0.51	0.50	0.49	2000
weighted avg	0.69	0.75	0.71	2000

RandomForest Classifier

	precision	recall	f1-score	support
0	0.88	0.97	0.92	1607
1	0.79	0.45	0.57	393
accuracy			0.87	2000
macro avg	0.83	0.71	0.75	2000
weighted avg	0.86	0.87	0.85	2000

Results/Conclusion

- In conclusion, there are 20.4% of customers who have churned.

- The exploratory data analysis led to meaningful discoveries. The insights are discussed in depth above, but some notable ones are:
  - While the male to female percentage is more or less the same, more females (55.9%) tend to leave as customers than males (44.1%).
  - 63.9% customers who churned were inactive while 36.1% were active. The bank can monitor customers who tend to be inactive and provide them initiatives to continue using the services, as they are more vulnerable to leaving.
  - Most of the customers are in the ages 25 to 45. The bank needs to retain more users of the older age, perhaps by providing them with benefits such as retirement savings plans.
  - The data analysis showed that customers from Germany were most likely to leave, even though they constituted only a quarter of the total customer base. The bank can look into the matters and determine the cause of their dissatisfaction.

Results/Conclusion

- Developed a classifier that allowed me to predict whether a customer will leave or not with an accuracy of 86.8%. This can help in identifying potentially vulnerable customers who are planning to leave, and we can act in ways to retain them.