TOPIC: **BANK CUSTOMER CHURN PREDICTION**

**Introduction:**

Customer churn is a major concern for businesses in various industries, including banking, telecommunications, and e-commerce. It refers to the phenomenon of customers leaving a company or discontinuing their use of its products or services. Churn can have significant financial and reputational implications for businesses, as it can lead to lost revenue, decreased market share, and increased customer acquisition costs. Therefore, predicting and preventing customer churn is a critical task for businesses to maintain customer loyalty and profitability.

**Problem Statement:**

Several approaches have been proposed to address the problem of customer churn prediction. One common approach is to use machine learning algorithms to analyse customer data and identify patterns or predictors of churn.

Despite the availability of these solutions, predicting customer churn remains a challenging task for businesses. One of the main challenges is the complexity and heterogeneity of customer data, which can include structured and unstructured data from multiple sources. Another challenge is the dynamic nature of customer behaviour, as churn can be influenced by a wide range of factors such as changes in market conditions, competitive offerings, and customer preferences. Therefore, there is a need for more accurate and actionable churn prediction models that can help businesses understand the drivers of churn and implement effective retention strategies.

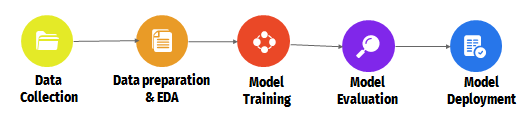
**Aims and Objectives:**

The goal of this project is to develop a machine learning model to

* Identify the key factors that influence customer churn in the banking industry, and provide recommendations to the bank on how they can improve customer retention
* Compare the performance of different machine learning algorithms for predicting customer churn, and recommend the best approach for the bank to use.
* Build and deploy a machine learning model to predict which bank customers are likely to churn, based on their demographic and behavioural characteristics, and evaluate the model's performance using appropriate metrics.

**Flow Process**

The steps taken for this project are illustrated in this flowchart below:



**Data Collection and Preprocessing**

* **Sources of data:** The bank customer churn dataset is a commonly used dataset for predicting customer churn in the banking industry. The dataset was obtained from Kaggle via the link below:

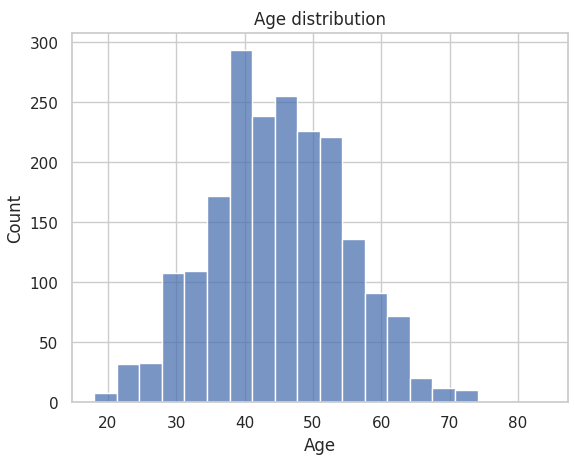
This dataset consists of 10002 rows and 14 columns and contains information about bank customers, including demographic information, account information, and transaction history.

* **Data cleaning and preprocessing techniques:** The data was cleaned and preprocessed by handling missing values, removing duplicates, encoding categorical variables, and scaling numerical variables.

**Exploratory Data Analysis and Visualization:**

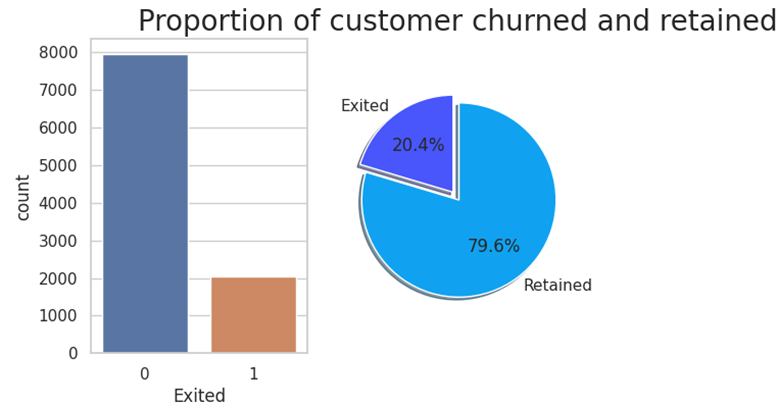
A summary of the numerical values was carried out. From the output, these are the main findings:

* The mean age of the customers is 38 years with the youngest and oldest customer being 18 and 92 years respectively.



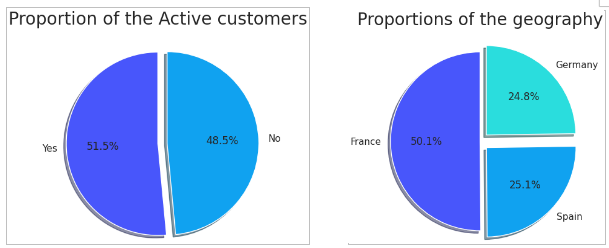
* Most of the customers have a tenure of 5 years in the bank.
* The mean number of products signed up by customers is 1, 1 being the minimum and 4 is the maximum.
* There are customers with zero balance in their bank accounts. However, the mean balance for the set is 76476.263.
* The mean estimated salary is 100,106. The minimum is 11.58 and maximum is 199,992. This means that the bank accommodates customers even with low paying jobs.

Subplots were used to visualise the distribution of the target variable, which is “Exited”. 70% are active members/customers while only 20% have churned the company. This in essence represents unbalanced data set for developing a machine learning model.

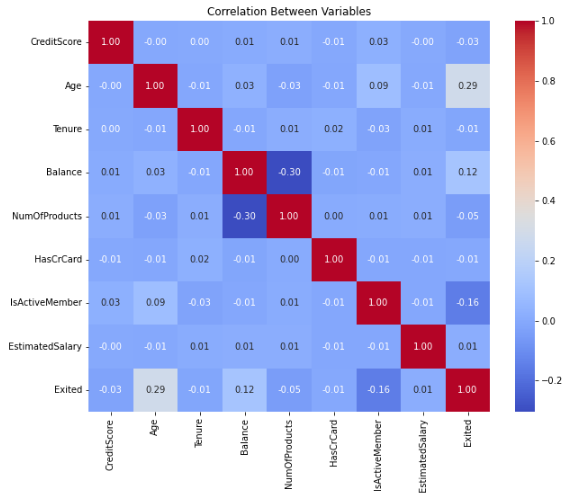


Secondly, histograms were used to visualise other numerical variables. The following are the observations made;

* The number of female customers who churned are more than those of male.
* While most of the customers are from France, the highest churned rate is from Germany. The company should look at the quality of services provided to these countries to see if it results in a high churn rate.
* A good number of customers do not have credit cards. It is also important to note that the highest churn rate is coming from customers with credit card. Again, the company should reassess the services provided to their customers to ensure they are of quality.
* Lastly, the number of inactive members is very high and this set also records a high churn rate. This is not surprising. The bank should into this and see why these customers are not active and implement strategies that will increase activity levels.



Lastly, a heatmap was fitted to explore the correlation between variables. From the map, most of the variables do not have a strong correlation with the target variable. However, age, and whether one is a member or not seems to have the highest correction with the target variable.



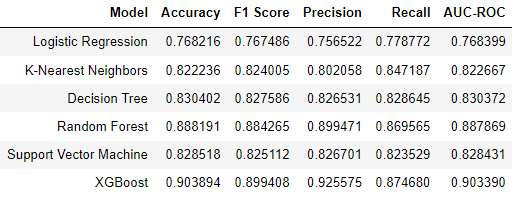
**Handling Imbalanced Data using SMOTE:**

SMOTE (Synthetic Minority Over-sampling Technique) is a popular technique used to address class imbalance in machine learning. In our dataset, the number of instances in the minority class (Churned) is much smaller than the number of instances in the majority class (Retained). This can lead to poor performance of machine learning models, as they tend to be biased towards the majority class. Therefore, SMOTE was used to address the imbalance to create new minority class samples by randomly selecting one or more of the k-nearest neighbours of each minority class sample, and using them to create new synthetic samples.

**Model Building, Selection and Evaluation:**

Model building involves choosing an appropriate machine learning algorithm and evaluating its performance on the dataset. We built different algorithms, such as, Logistic Regression, KNeighbors, Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machines, and compared their metrics.

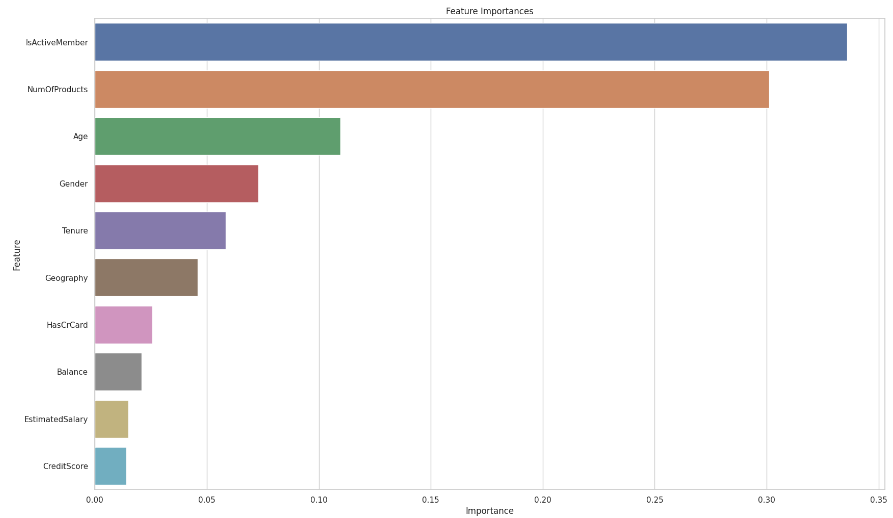
**Model Evaluation**: The metrics used were accuracy score, F1 score, precision score, recall score, and AUC-ROC to evaluate the performance of each model.

****

Based on the results, XGBoost model showed the best performance in all the metrics, with the highest Accuracy and F1 scores of 90.3% and 89.9% respectively.

**Feature Importance:**

The feature importance visual shows the relative importance of each feature in predicting customer churn. It shows that the most important factor in predicting customer churn is whether or not the customer is an active member. Customers who are not active members are more likely to churn. Also, customers with a high number of products and who are male are slightly more likely to churn. The fourth most important factor is the age of the customer, with older customers being more likely to churn.



**Results and Recommendations**

* **Summary of key findings:** Based on the analysis, we found that the most important factors that influence customer churn in the banking industry are customer satisfaction, product offerings, and customer demographics.
* **Recommendations**: Several recommendations for improving customer retention were provided, such as

1. Improving customer service,
2. Offering personalised products and services based on geography, gender and other features
3. Providing incentives for long-term customers,
4. Targeting high-risk customers, and
5. Improving product offerings.

**Conclusion:**

Predicting customer churn is important for banks because it allows them to identify customers who are at risk of leaving and take action to retain them. This can help banks improve customer retention; reduce costs associated with acquiring new customers, and increase revenue and profitability.

Future directions for research and improvement in customer retention should include using advanced machine learning techniques like deep learning and reinforcement learning.