Data Source

[Bank Churn Dataset (kaggle.com)](https://www.kaggle.com/datasets/rangalamahesh/bank-churn)

Customer retention is a key focus for banks in ensuring the longevity of their business. ABC Multinational Bank, in particular, is keen on retaining its account holders. The objective here is to analyze the customer data of the bank's account holders with the ultimate goal of predicting and mitigating customer churn.

Below is the customer data of account holders at the fictional ABC Multinational Bank and the aim of the data will be predicting the Customer Churn.

About Dataset

This dataset contains just over 165,000 records and has the following columns:

1. customer\_id, unused variable.
2. credit\_score, used as input.
3. country, used as input.
4. gender, used as input.
5. age, used as input.
6. tenure, used as input.
7. balance, used as input.
8. products\_number, used as input.
9. credit\_card, used as input.
10. active\_member, used as input.
11. estimated\_salary, used as input.
12. churn, used as the target. 1 if the client has left the bank during some period or 0 if he/she has not.

Data Import and Cleaning with Python

Data was imported into a Jupyter notebook from csv file downloaded from Kaggle. The data types of each column were checked and found to be appropriate. No null entries were found in the dataset. The Customer ID, Numerical ID and Surname columns were dropped as they are not likely to be useful in our analysis. The data was then exported as a csv file for visualization in Tableau.

Machine Learning – Pre-processesing

Using Python, the balance of the target column was checked. Around 4 of every 5 records were for customers who did not leave the bank. This imbalance was accounted for while setting up our predictive model. Categorical data was converted into numerical data. The data was split into features and targets, and then into training and testing sets. The numerical data in the feature test and train sets were then scaled.

Machine Learning – Model Selection

To address the imbalance in the target column mentioned above, Random Over Sampling was used.

A number of different models were assessed as a binary classifier to predict if a customer left or remained with the bank. These included Logistic Regression, Random Forest, Support Vector Machine, K Nearest Neighbor, XG Boost and a TensorFlow Neural Network model, The confusion matrix based on the test set was used to select the best performing model.

Machine Learning – Model Results

The Random Forest model proved to be the most accurate model, followed by XG Boost.

The confusion matrix for the Random Forest model is shown below

precision recall f1-score support  
  
 Stay 0.98 0.90 0.94 32493  
 Leave 0.91 0.98 0.94 32564  
  
 Accuracy 0.94 65057

**The headline result is that the model correctly predicts the outcome correctly in more than 9 of every 10 cases.**

The report shows that the model has an accuracy of 0.94 which shows a very high ratio of correctly predicted observations to the total number of observations.

For Stay, the precision is 0.98 which means 98% predicted stays were actually stays. The recall is 0.9 which means 90% of the actual stays were predicted correctly.

For Leave, the precision is 0.91 which means 91% of predicted leaves were actually leaves. The recall is 0.98 which means 98% of the actual leaves were predicted correctly.

The graph below shows the calculated feature importance results showing which features has the most influence on the predicted outcome of the model.

