Project Report

Bank Customer Churn

By

Wiam Boumaazi and Dipendra Bharati

Table of Contents:

1- Introduction:

Background - We wanted to know what could be the most significant factors that affected whether the customer continued their subscription to their credit card provider

Problem Description -

Objectives -The objective was to find the major factors/features that made the customers continue using their credit card.

2- Dataset

2.1- Exploratory Data Analysis

We obtained the dataset for this project from Kaggle.

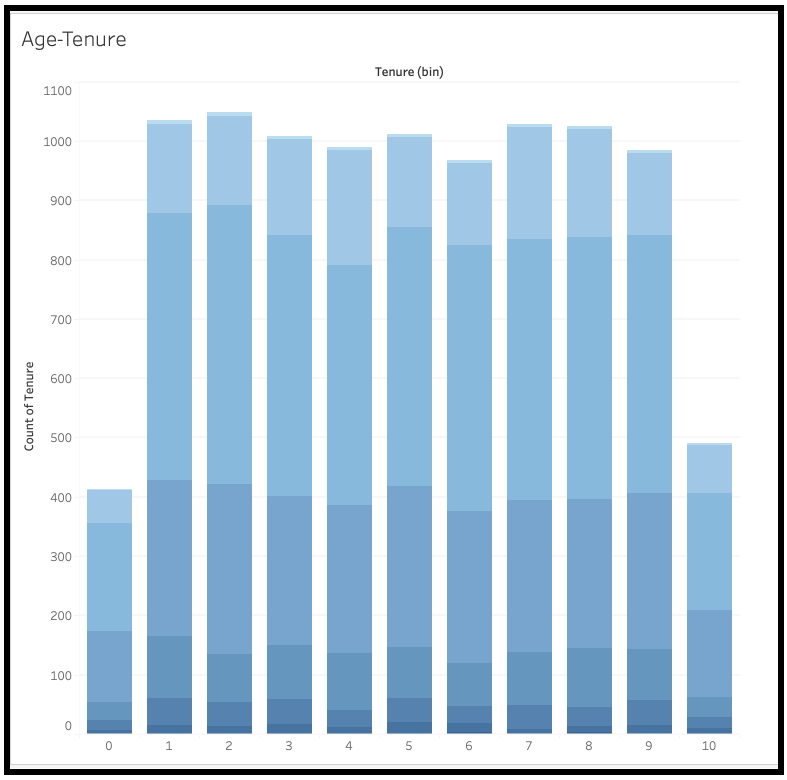
The dataset consisted of 10,000 rows and twelve feature columns. In terms of labels, there were 7963 positive class (IsActiveMember set to value 1) samples and 2037 negative class (isActiveMember set to value 0) samples.

The dataset had thirteen feature columns. These features were Credit Score, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, EstimatedSalary, Exited, CustomerId and RowNumber. Out of them, two features: Geography and gender were categorical features.

2.2- Dataset Visualization

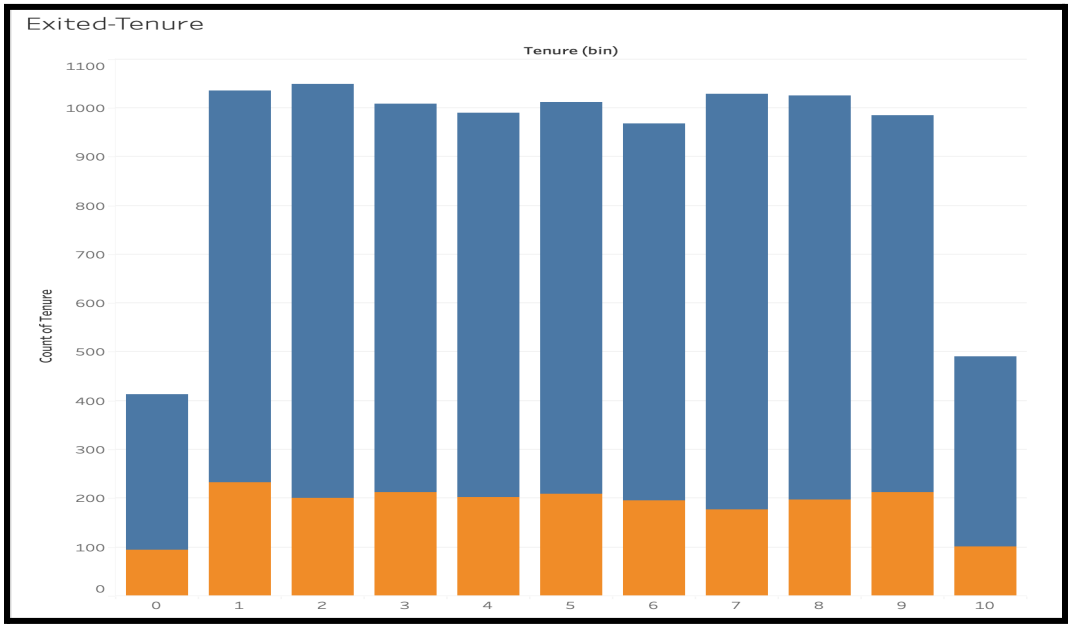
- **Age vs Tenure**

People with age ranging between 30 and 40 have the highest count of tenure. However, the distribution of the amount of customers by age is similar for every number of tenure years.



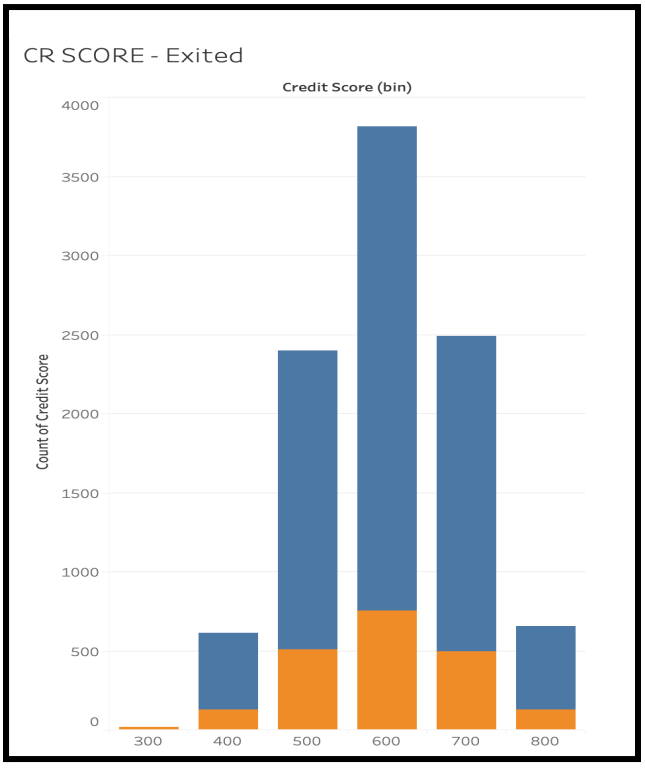
- **Exited vs Tenure:**

The distribution of the amount of customers by label (Exited / not Exited) is similar for every number of tenure years.



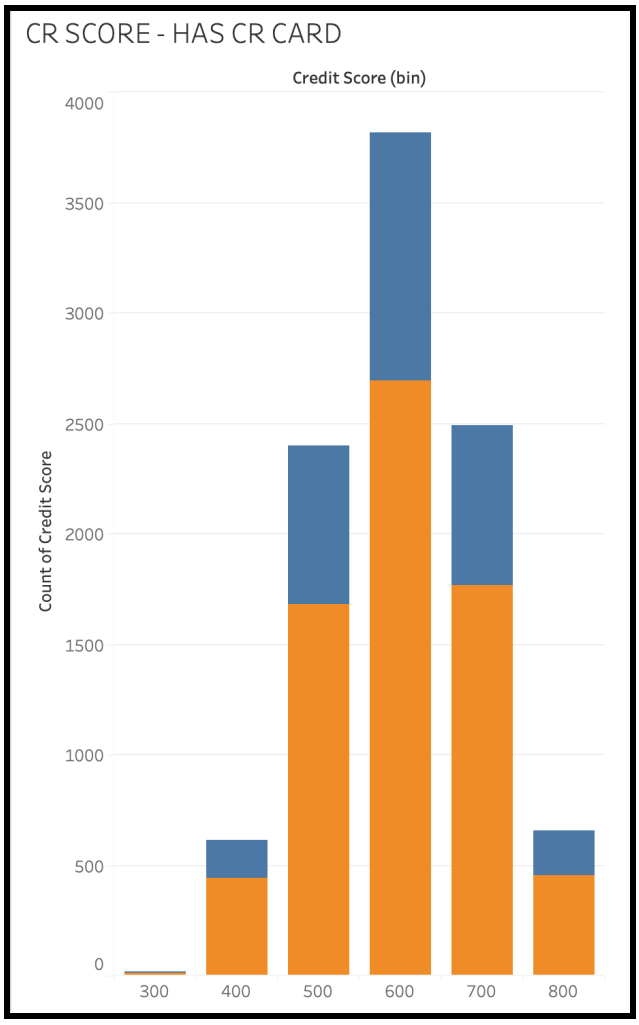
- **Credit Score vs Exited:**

The Number of customers with a credit score of 600 is higher than the number of customers with a credit score of 500 and 700. On the other hand, Customers with a credit score of 400 and 800 are widely lower than the other categories.



- **Credit score vs Credit card**

The Number of customers who have a credit card is slightly lower than those who don’t have a credit card among each category of credit score.



2.3- Data Pre-processing

In this step, we encoded the categorical features, Geography and Gender using the LabelEncoder module from Scikit Learn.

We also performed feature reduction by dropping columns which were not contributing to the prediction. The columns were Exited, CustomerId and RowNumber. The final dimension dataset was 10000 rows and ten columns.

We were planning on upsampling the negative class using SMOTE but due to time constraint proceeded to go with the existing dataset.

3- Methods:

3.1- ML Classifiers:

In this project we applied three ML classifiers: K-Nearest Neighbor, Random Forest, and XGboost.

Random Forest:
Random forest is a supervised classification and regression algorithm. The algorithm creates a bunch of trees randomly to constitute a forest. If the number of trees in the forest is high, it will result in a higher accuracy. In other words, Random forest is a group of decision trees that try to make a prediction. Each prediction resulting from a tree is considered a vote. The final prediction is basically the class label that has received the greater number of votes. Each decision tree is meant to result in a class decision as a vote. It might happen that all the decision trees within a random forest come with all similar class labels and that may affect the accuracy of the algorithm. In order to avoid that, there is a method called Bootstrapping/ bagging. Bootstrapping is an estimation method used to make predictions on a data set creating smaller datasets out of the training dataset.

K Nearest Neighbor:
K Nearest Neighbor is a ML algorithm that could be used for both binary and multi-class classification. This algorithm classifies the data point on how its neighbor is classified. Each new data point would be classified with KNN classifier based on the similarity measure of the earlier stored data points. K in KNN identifies the number of the nearest neighbors that is used to classify new data points. We can choose the number K with parameter tuning until finding the best results. However, there is also an initial try for the K number that can be calculated with the following formula: K= sqrt (total number of data points).

Xgboost:
Xgboost, also known as Extreme Gradient Boosting, is used for supervised learning for regression and classification tasks. It uses the `Boosting` ensemble method for classification. It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. It adds another tree to the existing model only if it reduces the overall loss. It is based on the gradient descent procedure.
It has following different kinds of parameters:

1. General Parameters - The guide the overall functioning of the algorithm.

2. Booster Parameters - These parameters guide the individual booster(tree) at each step. It includes max_depth, min_child_weight, gamma, reg_alpha etc
3. Learning Task Parameters - These parameters guide the optimization of the algorithm. It includes learning rate, batch size, epoch, early stopping etc

3.2 Evaluation metrics:

- Recall = The number of rows correctly assigned to class I compared with the total number of rows belonging to class I:
    - TP / TP + FP
- Precision = The number of rows correctly assigned to class I compared with the total number of rows predicted as belonging to class I:
    - TP / TP + FP
- F1 score = 2 * (precision * recall) / (precision + recall)
- F1 score for multi-class= F1 score is calculated for each single class
- Sensitivity = The proportion of correctly identified positive predictions:
    - TP / TP + FN
- Specificity =The proportion of correctly identified negative predictions:
    - TN / TN + FP
- G_means = a metric for imbalanced classification that seeks balance between sensitivity and specificity:
    - Sqrt (Sensitivity * Specificity)
- AUC = the area under the ROC curve that has sensitivity in the y-axis and 1-specificity on the x-axis
- 5-fold cross validation

The following table presents a synopsis of our experiment:

Synopsis of Experiment Setup

Programming Language	Python 3.7.9
Libraries	Sklearn, pandas, numpy, xgboost, knn, shap, matplotlib
Tools	Jupyter Notebook, Google collab, Tableau, Excel

ML Classifier	XGBoost, KNN
Feature Ranking	Gini Impurity, SHAP

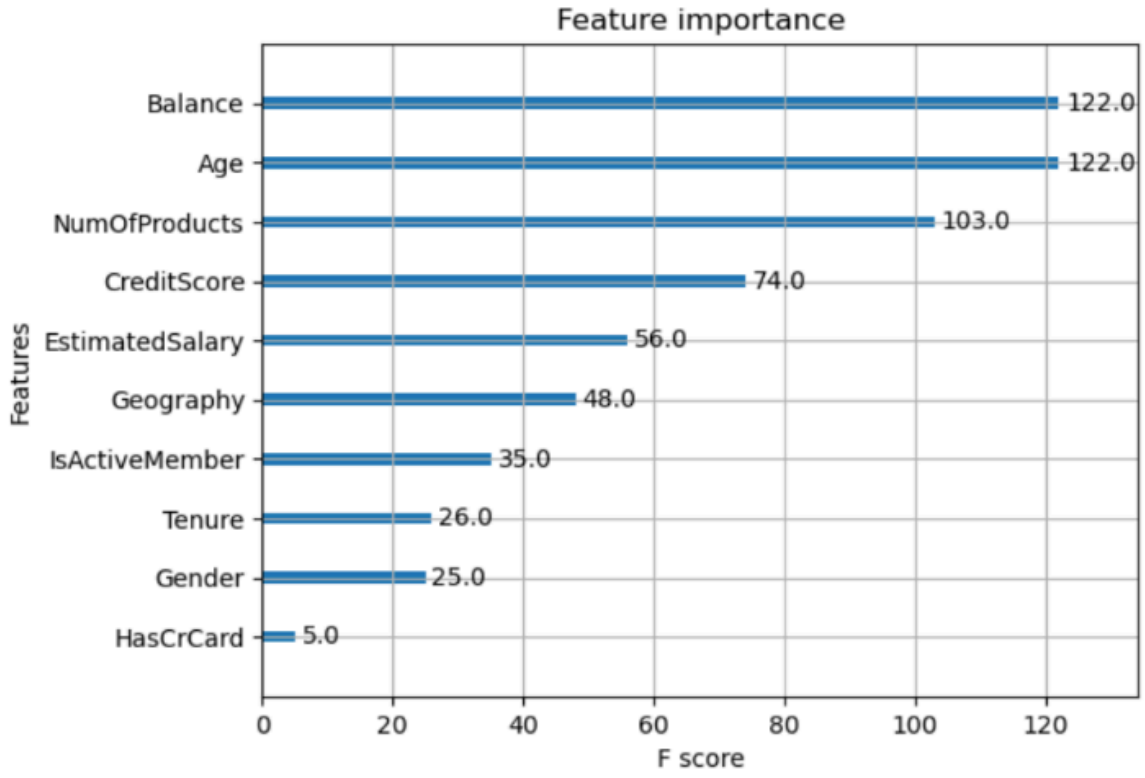
4- Results Analysis

Cross validation (CV) is one of the techniques used to test the effectiveness, and avoid sample errors of a machine learning model. We applied a 5-fold CV to performance metrics. The entire dataset was split into two separate dataset, 80% for model training and 20% for testing.

Classes	**IsActiveMember** (1), Not **IsActiveMember** (0)
Total Data points	7963(**IsActiveMember**), 2037(Not **IsActiveMember**)
Training Dataset	80%
Testing Dataset	20%
Cross validation	5 fold Cross Validation on training set

IsActiveMember/not IsActiveMember											
Model	**Thres-hold**	**Sensitivity (%)**		**Specificity(%)**		**G-mean(%)**		**AUC(%)**		**F1**	
		Test	**C.V**	**Test**	**C.V**	**Test**	**C.V**	**Test**	**C.V**	**Test**	**C.V**
Random Forest	**N**	0.74	0.77	0.79	0.75	0.77	0.76	0.85	0.85	0.59	0.56
	Y	0.75	0.73	0.77	0.81	0.76	0.77	0.85	0.85	0.59	0.57
KNN	**N**	0.39	0.38	0.63	0.65	0.50	0.50	0.51	0.52	0.10	0.08
	Y	0.39	0.38	0.63	0.65	0.50	0.50	0.51	0.52	0.10	0.08
Xgboost	**N**	0.8	0.8	0.75	0.75	0.77	0.77	0.86	0.86	0.60	0.60
	Y	0.8	0.8	0.75	0.75	0.77	0.77	0.86	0.86	0.60	0.60

The feature importance graph which shows the contribution of each feature towards the prediction is shown in the graph below:



5- Conclusion and Future Work

From this research project, based on our dataset, we concluded that Age of the customers and the balance they had in their bank account were the primary factors that affected whether or not they kept on using their Credit Card.