Banking Customer Loyalty Analysis & Modeling

Francisco Rodriguez

Background

Despite bank's continuous efforts to **attract and retain customers**, the banking industry faces a persistent challenge in the form of **customer turnover**, leading to financial losses and reduced customer satisfaction.

It is advantageous for banks to **identify the factors** influencing a customer's **decision to depart** from the institution.

Objectives

- Examine the **rate of customer turnover** within the bank, as it is essential to comprehend **the reasons behind** customer departures.
- Develop a **Machine Learning Model** capable of identifying the **key factors** that greatly **impact** customer turnover or attrition.
- Choose **the most reliable model** that will attach a probability to the churn to make it easier for customer service to **target right customer** in order to **minimize their efforts** to prevent customers churn.

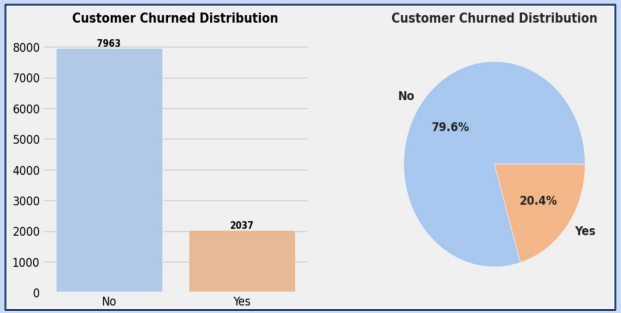
Note: The following analysis has been inspired by the following Kaggle project, from https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling

Data Exploration

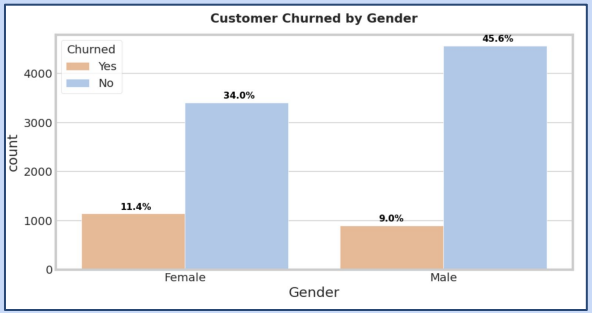
14 Columns - 10000 Rows

RowNumber	int64	It is likely a unique identifier for each record and does not contribute directly to the analysis.
CustomerId	int64	It can be used to track and differentiate individual customers within the dataset.
Surname	object	It provides information about the family name of each customer.
CreditScore	int64	It is a numerical value that assesses the creditworthiness of an individual based on their credit history and financial behavior.
Geography	object	It provides information about the customers' geographic distribution, allowing for analysis based on regional or national factors.
Gender	object	It categorizes customers as either male or female, enabling gender-based analysis if relevant to the churn prediction.
Age	int64	It represents the customer's age in years and can be used to analyze age-related patterns and behaviors.
Tenure	int64	It typically represents the number of years or months the customer has been associated with the bank.
Balance	float64	It reflects the amount of money in the customer's bank account at a specific point in time.
NumOfProducts	int64	It can include various offerings such as savings accounts, loans, credit cards, etc.
HasCrCard	int64	It is a binary variable with a value of 1 if the customer possesses a credit card and 0 otherwise.
IsActiveMember	int64	It is a binary variable indicating whether the customer is an active member (1) or not (0) within the bank.
EstimatedSalary	float64	It provides an approximation of the customer's income level, which can be relevant for analyzing churn behavior.
Exited	int64	It indicates whether a customer has churned (1) or not (0) from the bank. It is the variable we aim to predict using the other features.

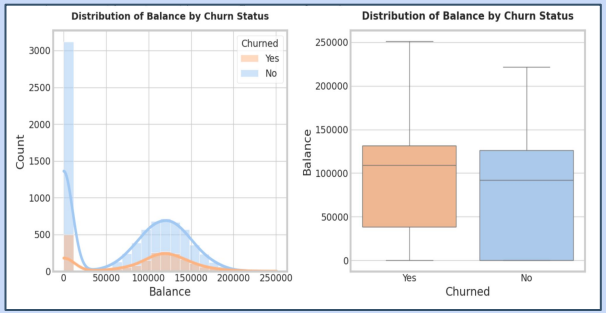
Data Exploration



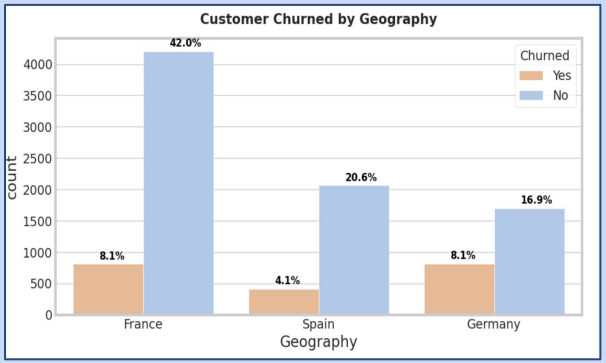
There is huge class-imbalance which can lead to bias in model performance.



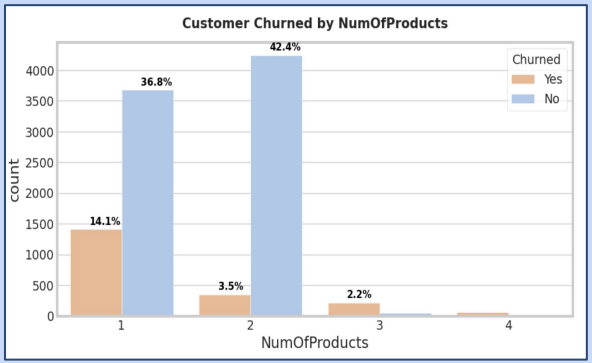
The churned probability is more for Female Customers compared to male customers.



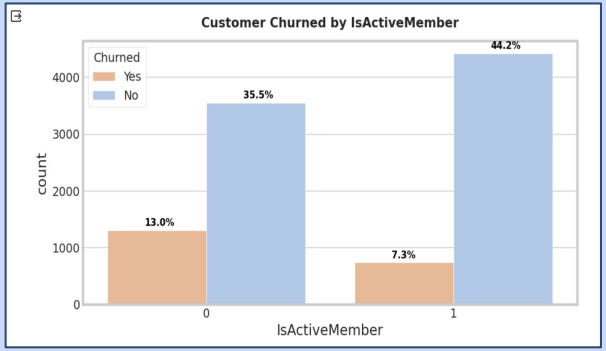
Customers with zero balance are more likely to deactivate their account.



There are almost equal customers from Spain & Germany, but the Churn rate is almost double in Germany when compared with spain.

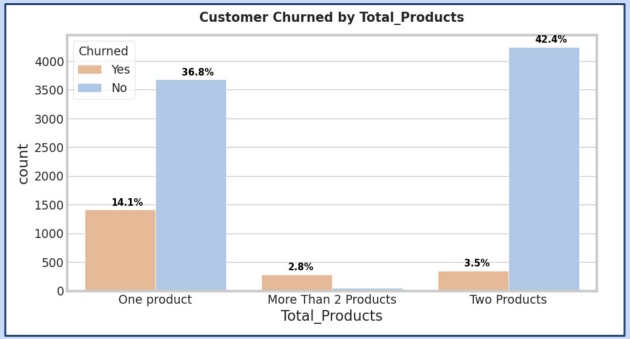


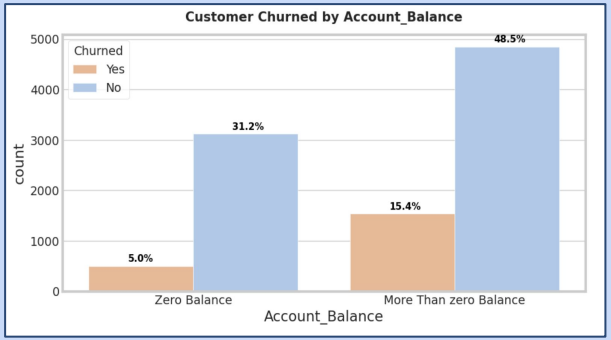
There is very high churn rate in customers having 1 product



Customers which are not active are more likely to deactivate their banking facilities.

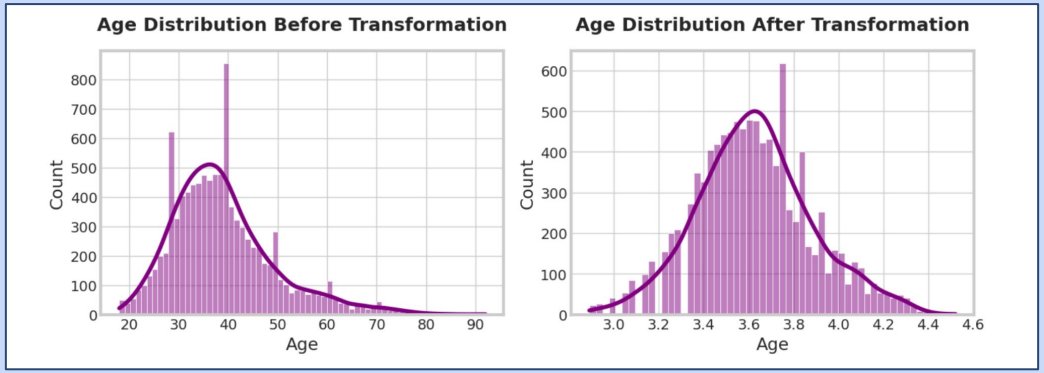
Before Modeling





Modification in Total_Products

Modification in Account_Balance



Skewness can negatively impact the performance of certain machine learning algorithms, like Decision Tree Models.

Before Modeling

Drop | Modify | No changes

RowNumber	int64	It is likely a unique identifier for each record and does not contribute directly to the analysis.
CustomerId	int64	It can be used to track and differentiate individual customers within the dataset.
Surname	object	It provides information about the family name of each customer.
CreditScore	int64	It is a numerical value that assesses the creditworthiness of an individual based on their credit history and financial behavior.
Geography	object	It provides information about the customers' geographic distribution, allowing for analysis based on regional or national factors.
Gender	object	It categorizes customers as either male or female, enabling gender-based analysis if relevant to the churn prediction.
Age	int64	It represents the customer's age in years and can be used to analyze age-related patterns and behaviors.
Tenure	int64	It typically represents the number of years or months the customer has been associated with the bank.
Balance	float64	It reflects the amount of money in the customer's bank account at a specific point in time.
NumOfProducts	int64	It can include various offerings such as savings accounts, loans, credit cards, etc.
HasCrCard	int64	It is a binary variable with a value of 1 if the customer possesses a credit card and 0 otherwise.
IsActiveMember	int64	It is a binary variable indicating whether the customer is an active member (1) or not (0) within the bank.
EstimatedSalary	float64	It provides an approximation of the customer's income level, which can be relevant for analyzing churn behavior.
Exited	int64	It indicates whether a customer has churned (1) or not (0) from the bank. It is the variable we aim to predict using the other features.

Before Modeling

17 Columns - 10000 Rows

Dependent Variable

Independent Variables

CreditScore	int64
Age	int64
Tenure	int64
HasCrCard	int64
IsActiveMember	int64
EstimatedSalary	float64
Churned	int64
Geography_France	int64
Geography_Germany	int64
Geography_Spain	int64
Gender_Female	int64
Gender_Male	int64
Total_Products_More Than 2 Products	int64
Total_Products_One product	int64

Total_Products_Two Products	int64
Account_Balance_More Than zero Balance	int64
Account_Balance_Zero Balance	int64

Before Modeling

Dependent and Independent Variables

```
X = df.drop(columns=["Churned"])
y = df["Churned"]
```

Splitting the Dataset

```
[ ] from sklearn.model_selection import train_test_split

    # Split the dataset into training and testing sets
    X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state=0,shuffle=True)
```

```
print("Shape of X_train is:",X_train.shape)
print("Shape of X_test is: ",X_test.shape)
print("Shape of y_train is:",y_train.shape)
print("Shape of y_test is: ",y_test.shape)
```

```
Shape of X_train is: (7000, 16)
Shape of X_test is:  (3000, 16)
Shape of y_train is: (7000,)
Shape of y_test is:  (3000,)
```

Target variable with equal number of records

```
[ ] from imblearn.over_sampling import SMOTE
    smt = SMOTE(random_state=42)
```

```
[ ] # Aplicando SMOTE para el ajuste del conjunto de datos
    X_train, y_train = smt.fit_resample(X_train, y_train)
```

```
[ ] print(X_train.shape, y_train.shape)

    (11168, 16) (11168,)
```

```
[ ] y_train.value_counts().to_frame()
```

	Churned
1	5584
0	5584

Modeling

Logistic Regression

Logistic regression is employed in banking churn analysis for its **capability in predicting customer churn** and identifying **influential factors.**

It helps develop **targeted retention strategies** and facilitates informed decision-making in customer retention efforts.

Decision Tree

Decision trees are employed in banking churn analysis for their simplicity and interpretability, allowing stakeholders to understand the decision-making process.

They efficiently partition data based on feature splits, helping **identify critical factors influencing customer churn.**

Random Forest

Random forests are utilized in banking churn analysis for their **robustness against overfitting and ability to handle large datasets with high dimensionality.**

By aggregating predictions from multiple decision trees, random forests provide **more accurate and stable predictions of customer churn probabilities.**

Logistic Regression

Modeling

Accuracy: 0.5165

```
from sklearn.linear_model import LogisticRegression

# Create a logistic regression classifier object
lm = LogisticRegression(random_state=0)

# Fit the model on the train set
lm.fit(X_train, y_train)

# Perform prediction on the test set
y_pred_lm = lm.predict(X_test)

# Calculate the accuracy of the model
score_test_lm= lm.score(X_test, y_test)
score_train_lm= lm.score(X_train, y_train)

print("Accuracy Train_lm:", score_train_lm)
print("Accuracy Test_lm:", score_test_lm)
print("Predictions_lm:", y_pred_lm)
```

```
Accuracy Train_lm: 0.5165651862464183
Accuracy Test_lm: 0.5343333333333333
```

Results

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
CreditScore	−0.0002	0.0003	−0.6263	0.5311	−0.0009	0.0005
Age	4.0008	0.1215	32.9319	0.0000	3.7627	4.2390
Tenure	−0.0288	0.0118	−2.4375	0.0148	−0.0520	−0.0057
HasCrCard	−0.1561	0.0725	−2.1536	0.0313	−0.2982	−0.0140
IsActiveMember	−1.2034	0.0710	−16.9565	0.0000	−1.3425	−1.0643
EstimatedSalary	0.0000	0.0000	1.5167	0.1294	−0.0000	0.0000
Geography_France	−4.6208	0.3148	−14.6781	0.0000	−5.2378	−4.0038
Geography_Germany	−3.5131	0.3159	−11.1213	0.0000	−4.1322	−2.8940
Geography_Spain	−4.5408	0.3193	−14.2198	0.0000	−5.1667	−3.9149
Gender_Female	−3.6715	0.3182	−11.5396	0.0000	−4.2951	−3.0479
Gender_Male	−4.2560	0.3188	−13.3479	0.0000	−4.8809	−3.6310
Total_Products_More Than 2 Products	−0.8381	0.3826	−2.1905	0.0285	−1.5880	−0.0882
Total_Products_One product	−3.4584	0.3221	−10.7374	0.0000	−4.0897	−2.8271
Total_Products_Two Products	−5.2605	0.3295	−15.9651	0.0000	−5.9063	−4.6147
Account_Balance_More Than zero Balance	−2.9758	0.3441	−8.6486	0.0000	−3.6502	−2.3014
Account_Balance_Zero Balance	−2.8272	0.3485	−8.1119	0.0000	−3.5103	−2.1441

Logistic Regression

Odds Ratio

	Variable	Odds Ratio
0	CreditScore	0.999823
1	Age	1.000001
2	Tenure	0.999995
3	HasCrCard	0.999998
4	IsActiveMember	0.999997
5	EstimatedSalary	1.000001
6	Geography_France	0.999997
7	Geography_Germany	1.000000
8	Geography_Spain	0.999998
9	Gender_Female	0.999999
10	Gender_Male	0.999997
11	Total_Products_More Than 2 Products	1.000000
12	Total_Products_One product	1.000001
13	Total_Products_Two Products	0.999996
14	Account_Balance_More Than zero Balance	1.000000
15	Account_Balance_Zero Balance	0.999997

Interpretations

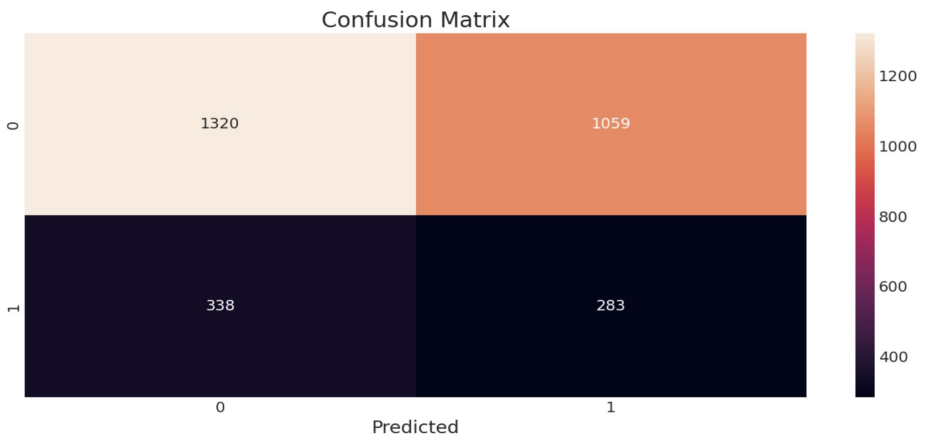
CreditScore: Despite the small negative coefficient, the odds ratio of 0.999823 suggests that an increase in credit score slightly decreases the odds of churn. However, t**his effect is very small and may not be significant in practice**

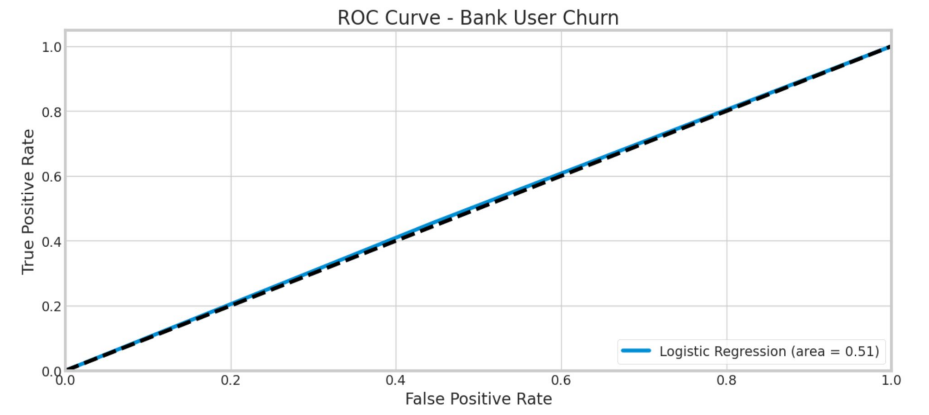
Age: The significant positive coefficient and odds ratio of 1.000001 indicate that aging of customers slightly increases the odds of churn. This suggests that **older customers may be more likely to leave the service.**

Total_Products: The number of products a customer has seems to have an impact on the odds of churn. **Customers with more than 2 products are less likely to churn compared to those with fewer products.**

IsActiveMember: Active members have a significant impact in reducing the odds of churn, as indicated by the odds ratio of 0.999997. **This suggests that strategies to encourage member activity could be effective in retaining customers.**

Logistic Regression





- The model correctly identified 283 churn cases and 1320 non-churn cases, but misclassified 1059 non-churn cases as churn and 338 churn cases as non-churn.
- Precision is around 50.35%, indicating how many predicted churn cases were correct, while Recall is approximately 50.53%, indicating how many actual churn cases were correctly identified.
- The model's overall accuracy is 53.43%.
- The ROC curve AUC value of 0.51 suggests poor performance in distinguishing between churn and non-churn cases.

Decision Tree

Modeling

Accuracy: 0.8883

```
best_params = grid_search.best_params_

# Create a DecisionTreeClassifier object with the best parameters
dtree = DecisionTreeClassifier(**best_params)

# Fit the model with the best parameters
dtree.fit(X_train, y_train)
```

```
DecisionTreeClassifier(max_depth=8, min_samples_leaf=4, random_state=42,
                       splitter='random')
```

```
# now make prediction on the y_test population, using X_test

y_test_pred = dtree.predict(X_test)

# now make prediction (i.e., get fitted values) on the y_train population, using X_train

y_fitted = dtree.predict(X_train)
```

```
# Measure Accuracy on the train population using accuracy_score() function :

accuracy_train = accuracy_score(y_train,y_fitted)

print("Accuracy_train:", accuracy_train)
```

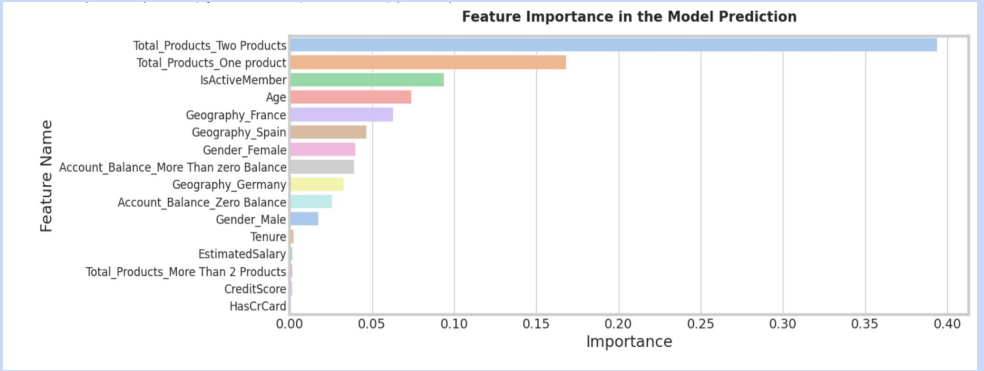
```
Accuracy_train: 0.8883416905444126
```

```
# Calculate the accuracy of the model using accuracy_score() function

accuracy_test = accuracy_score(y_test, y_test_pred)

print("Accuracy_Test:", accuracy_test)
```

```
Accuracy_Test: 0.8283333333333334
```



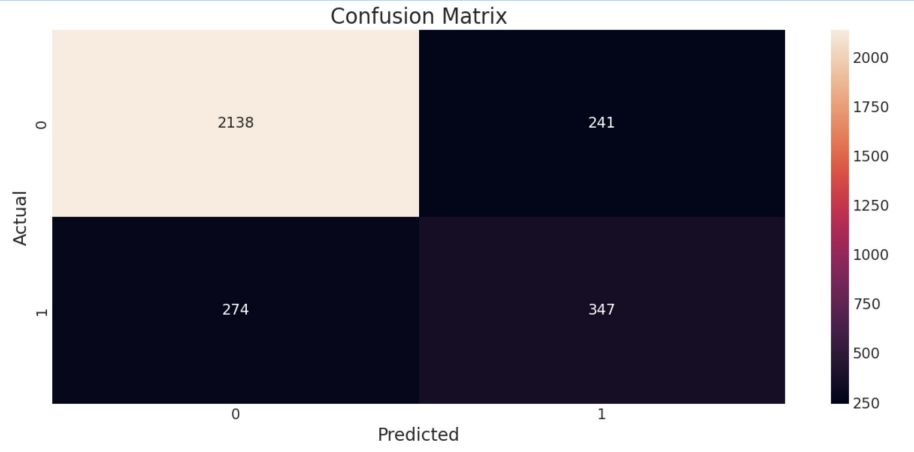
The key factors that significantly influence the deactivation of customers banking facilities are:

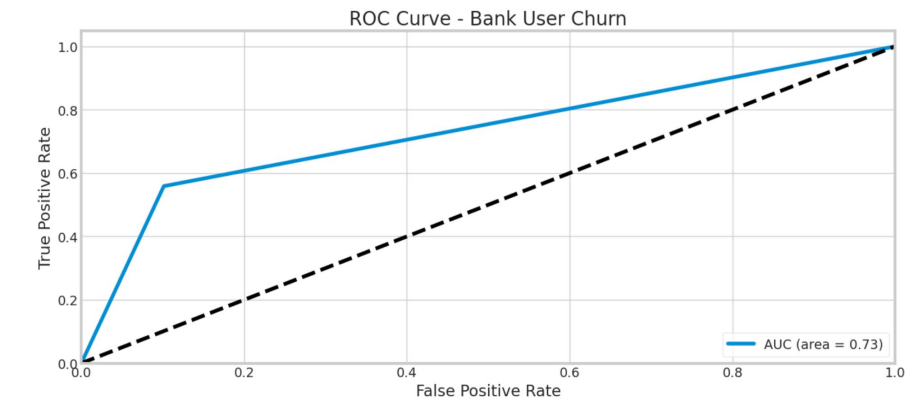
- Total_Products, Age, IsActiveMember, Geography, Balance and Gender.

The minimal impact of features on the deactivation of customers' banking facilities are:

- CreditScore, HasCrCard, Tenure and EstimatedSalary

Decision Tree





- The model **successfully identified positive cases with a high number of true positives,** demonstrating its accuracy in classification.
- However, **notable false negatives suggest the model may have missed some actual positive instances**, indicating room for improvement and further refinement.
- The ROC curve **AUC value of 0.73 indicates moderate performance in distinguishing between deactivated and non-deactivated cases.** While not excellent, it suggests that the model has some ability to discriminate between the two classes (Churn and Not Churn).

Random Forest

Modeling

Accuracy: 0.9058

```
[ ]  rfc = RandomForestClassifier(**best_parameters)
     rfc.fit(X_train, y_train)
```

```
RandomForestClassifier(criterion='entropy', max_depth=8, min_samples_leaf=3,
                       min_samples_split=5)
```

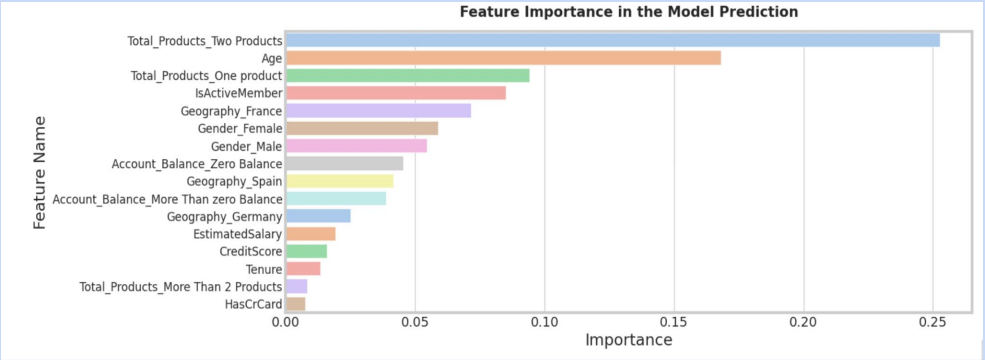
```
[ ]  # Evaluate the model
     print('Accuracy_rfc:', rfc.score(X_test, y_test))

     # Predict the labels of the test set
     y_pred_rfc = rfc.predict(X_test)

     # Calculate the accuracy of the model
     score_test_rfc = rfc.score(X_test, y_test)
     score_train_rfc = rfc.score(X_train, y_train)

     print("Accuracy Train_rfc:", score_train_rfc)
     print("Accuracy Test_rfc:", score_test_rfc)
     print("Predictions_rfc:", y_pred_rfc)
```

```
Accuracy_rfc: 0.8396666666666667
Accuracy Train_rfc: 0.9058022922636103
Accuracy Test_rfc: 0.8396666666666667
Predictions_rfc: [0 0 0 ... 0 0 1]
```



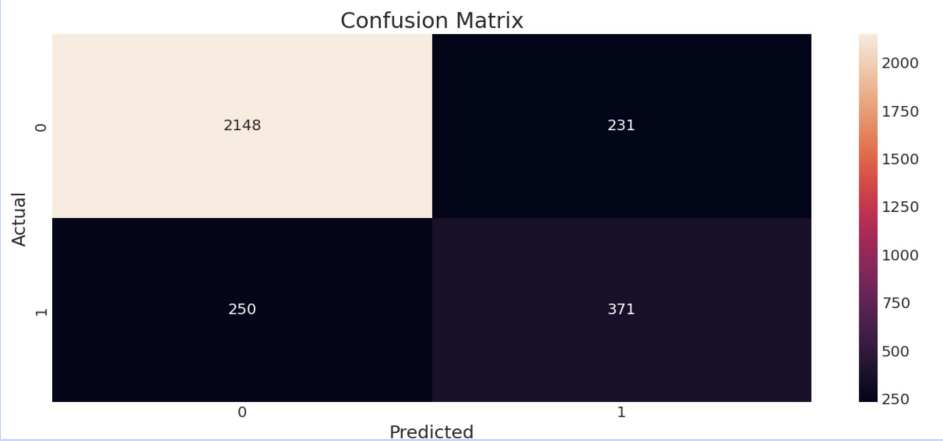
The key factors that significantly influence the deactivation of customers banking facilities are:

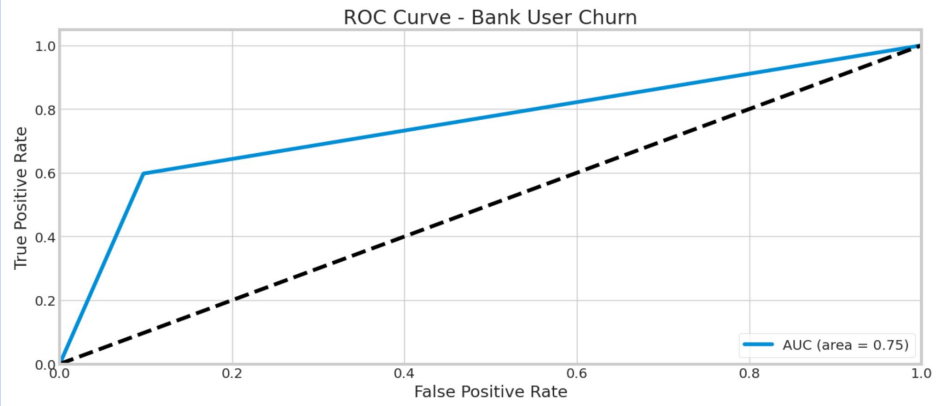
Total_Products, Age, IsActiveMember, Geography, Gender and Balance.

The minimal impact of features on the deactivation of customers' banking facilities are:

HasCrCard, Tenure, CreditScore and EstimatedSalary

Random Forest





- The model effectively identifies positive instances with **a significant number of true positive forecasts**, demonstrating its **ability to classify the desired outcome accurately.**
- However, the presence of a **relatively high number of false negatives suggests the model may have missed some genuine positive instances,** highlighting the need for further refinement to improve its accuracy in detecting all positive cases.
- The ROC Curve value of 0.75 indicates the model's **strong ability to distinguish** between positive and negative instances, showcasing its proficiency in making precise predictions regarding churn and non-churn scenarios.

Summary

```
*****************************************
********* Accuracy during Training ********
*****************************************

Accuracy Train_lm.    : 0.5165651862464183
Accuracy Train_dt.    : 0.8883416905444126
Accuracy Train_rfc.   : 0.9058022922636103

*****************************************
********* Accuracy during Testing ********
*****************************************

Accuracy Test_lm.     : 0.5343333333333333
Accuracy Test_dt.     : 0.8283333333333334
Accuracy Test_rfc.    : 0.8396666666666667
```

- **The logistic regression** model shows **the lowest performance**, with relatively low accuracies on both training and testing sets.
- T**he decision tree** and **random forest** models outperform the logistic regression model, with significantly **higher accuracies on both training and testing sets.**
- Among the tree-based models, **the random forest model achieves the highest accuracy** on both training and testing sets, indicating its **robustness and generalization ability.**
- All models perform better on the training set compared to the testing set, **suggesting a degree of overfitting**, especially notable in the decision tree model.

Conclusions and Recommendations

Conclusions

- **The key factors** that significantly influence the deactivation of customers banking facilities are **Total_Products, Age, IsActiveMember, Geography, Balance and Gender.**
- The **decision tree** and **random forest** models **achieved high accuracy scores**, with **training accuracies around 88% to 91%** and t**esting accuracies ranging from 83% to 84%**. This indicates a good fit to the training data and the **ability to generalize well to unseen instances**.
- The AUC values for the **models range from 0.71 to 0.75,** indicating a **moderate discriminatory power**. While not extremely high, these values suggest that the models are r**easonably effective in distinguishing between positive and negative instances.**
- The **decision tree and random forest** models demonstrate **strong performance across multiple evaluation metrics**, indicating their effectiveness in making accurate predictions and capturing the desired outcomes.

Recommendations

- **Incentivize** customers to **have multiple banking products**, as indicated by the **significance** of Total_Products in **influencing churn.**
- Implement **customer engagement strategies** such as rewards, incentives, and personalized communication to encourage active participation, especially among **older customers (Age)** and those with **higher balances (Balance**).
- Tailor **retention efforts** based on **geographical regions**, considering the influence of **Geography** on churn behavior.

Banking Customer Loyalty Analysis & Modeling

Francisco Rodriguez