Bank Customer Churn Prediction

Link to github Repository : https://shorturl.at/ALNT7

1. Data Description :

This data is taken from kaggle, here is the link to the original data : https://shorturl.at/jtFUX

Our data contains 14 columns including one target variable, which defines whether the customer has exited or not.
The 13 remaining columns are the following :

- RowNumber : indicates the row ID (int)
- CustomerId : indicates the customer ID (int)
- Surname : indicates the customer name (string)
- CreditScore : a score that shows how likely you are to pay a loan back on time (int)
- Geography : where is our customer from (string : country)
- Gender : Indicates the gender of our customer (categorical : male-female)
- Age : indicates the age of our customer (int)
- Tenure : indicates for how long the customer has been our bank's client (int)
- Balance : indicates the balance of our customer (float)
- NumOfProducts : indicates the number of products/services the customer has (int)
- HasCrCard : indicates whether our customer has a credit card or not (categorical : 0-1)
- IsActiveMember : indicates whether our customer is active or not (categorical : 0-1)
- EstimatedSalary : indicates the estimated salary of our customer (float)

2. Main objective :

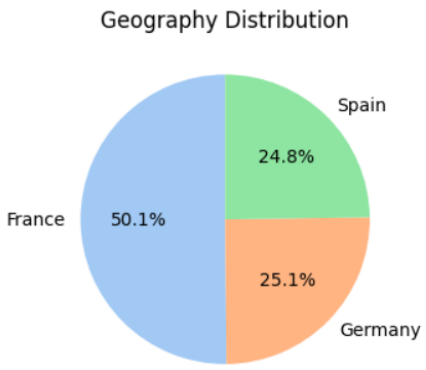
In this project, our models will focus on prediction of our customers' churn. It will help our stakeholders anticipate the customer behavior and try to find solutions to maintain the customer based on our findings. Other parts of the project will definitely reveal the relationship between the features and the target variable which would explain why such a behavior would occur.
Actions will then be taken upon our understanding !

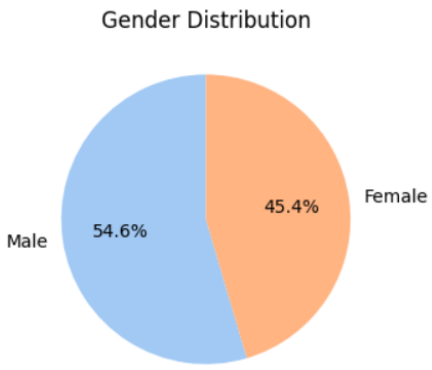
3. Data Exploration :

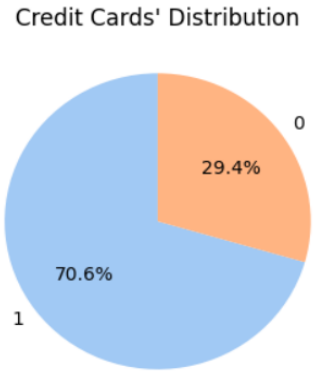
In this section, we will incorporate visual aids from my notebook to facilitate our data exploration and elucidate the relationships among the features themselves or between the features and the target variable. Prior to delving into this visual analysis, we will commence by acquainting ourselves with the structure of our dataset, scrutinizing the data types, and determining which columns merit removal due to their lack of relevance in constructing our forthcoming predictive machine learning model.

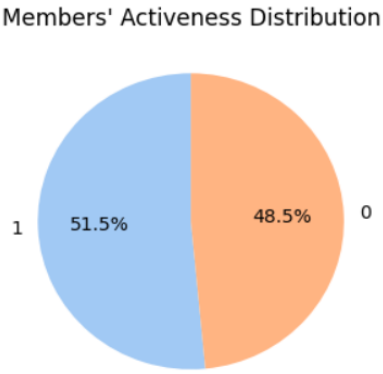
The complete notebook will be attached with this document.

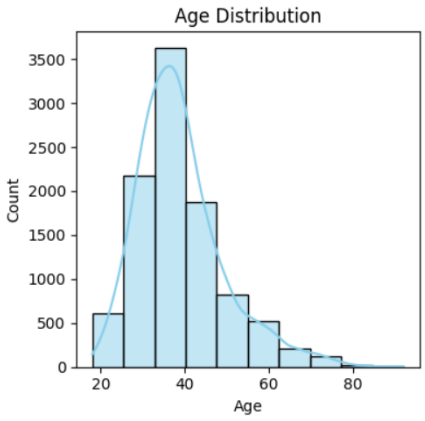
3.1 Basic Visuals :

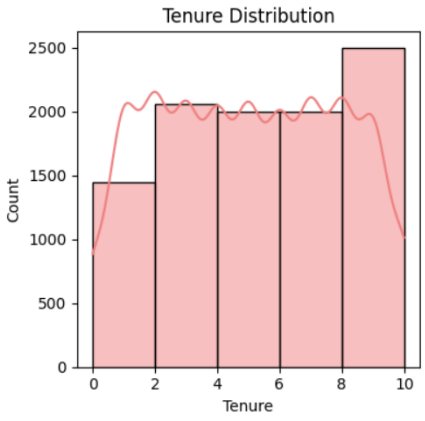


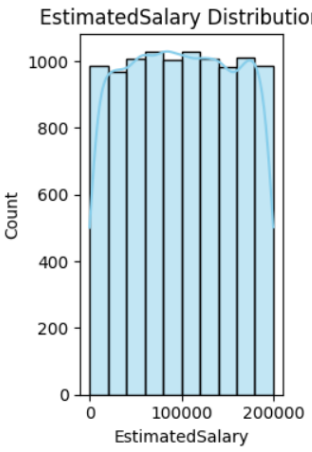


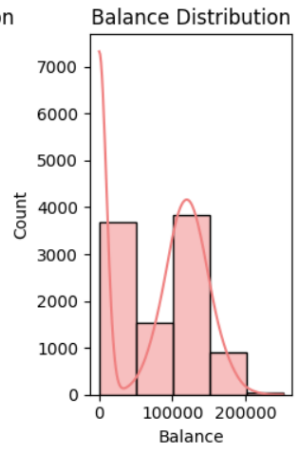


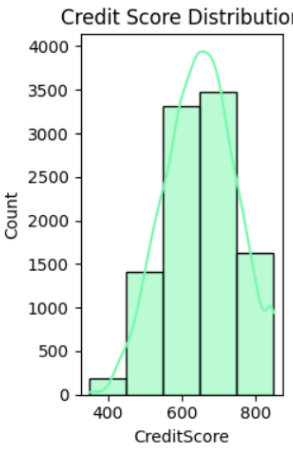




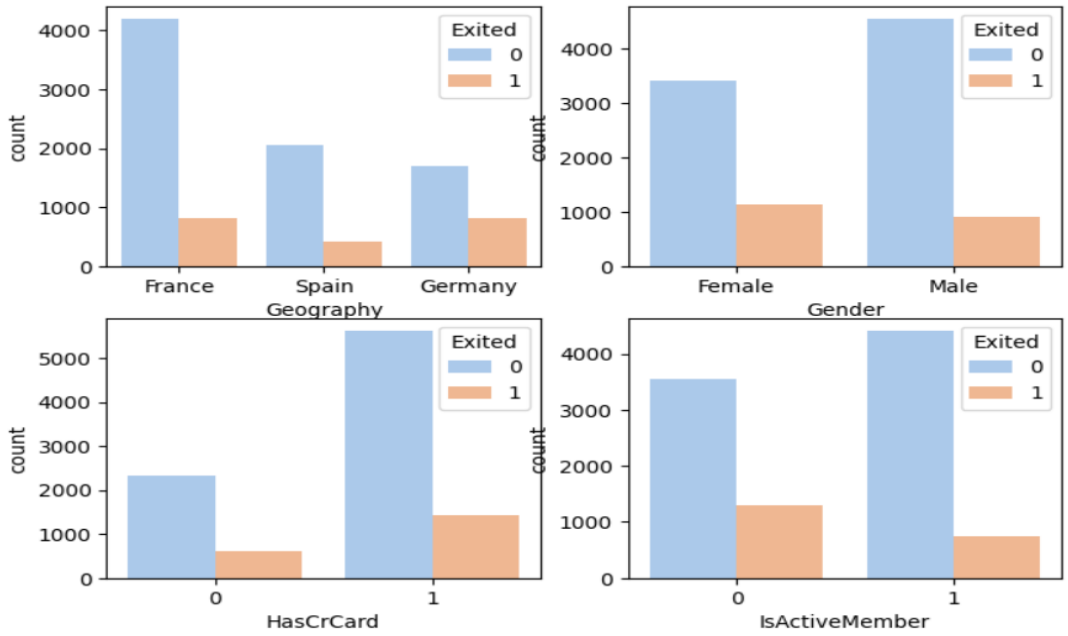








3.2 Correlations, Feature Engineering and Further Findings :

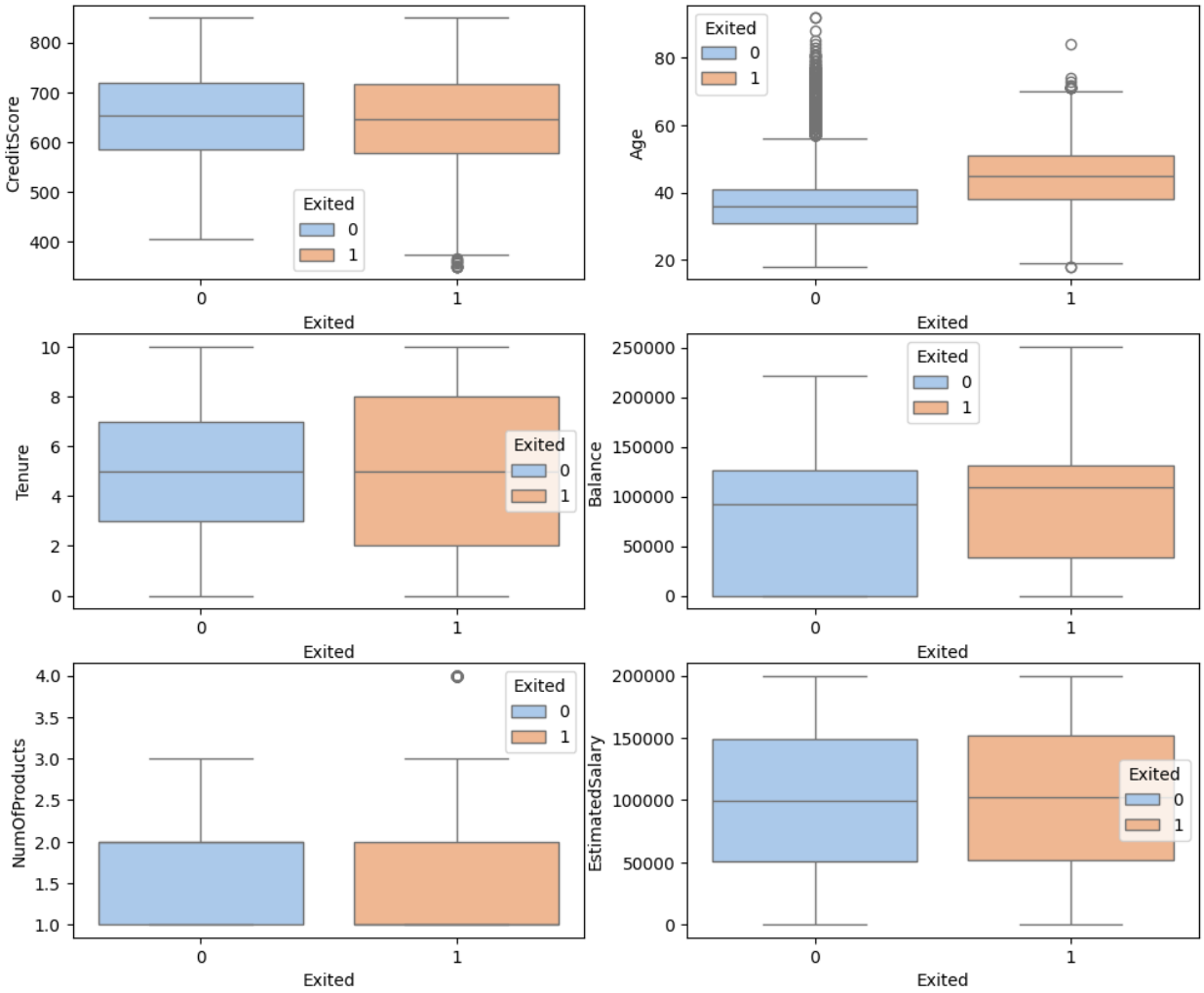


- *Geographic Churn Discrepancy*: The bank's customer base is predominantly from France. Paradoxically, the regions with fewer clients appear to experience a higher proportion of customer churn, suggesting a potential challenge, possibly related to resource allocation in those areas.

- *Gender Disparity in Churn*: Notably, the churn rate is notably higher among female customers compared to their male counterparts.

- *Credit Card Influence on Churn*: While the majority of customers hold credit cards, a considerable number of churned customers also possess this financial tool. The correlation between credit card ownership and churn warrants further investigation.

- *Inactive Members and Churn*: Alarming is the fact that inactive members demonstrate a considerably higher churn rate. The prevalence of inactive members overall highlights an area for improvement, and strategies to engage these members could be pivotal in reducing customer churn.

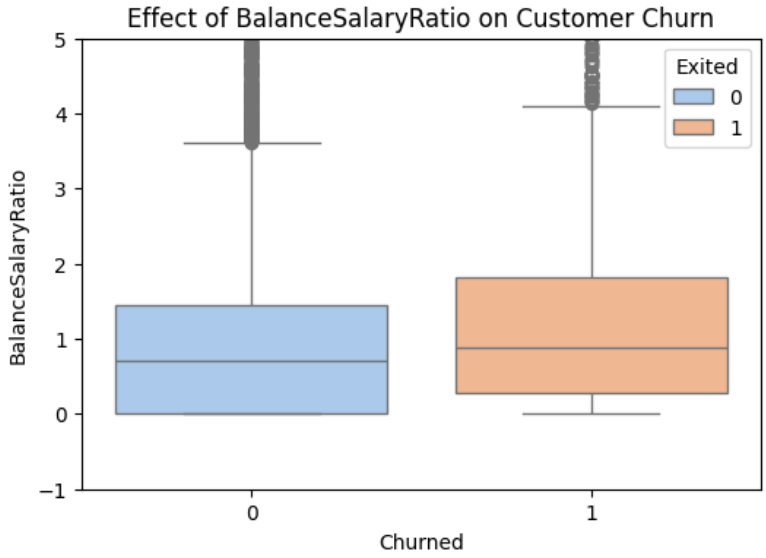


- The older customers are churning at more than the younger ones alluding to a difference in service preference in the age categories. The bank may need to review their target market or review the strategy for retention between the different age groups.

- With regard to the tenure, the clients on either extreme end (spent little time with the bank or a lot of time with the bank) are more likely to churn compared to those that are of average tenure.

- Worryingly, the bank is losing customers with significant bank balances which is likely to hit their available capital for lending.

- The "**BalanceSalaryRatio**" feature, which is calculated by dividing a customer's bank balance by their estimated annual salary, is important because it provides insights into the financial stability and behavior of customers.



- The observation made is that, contrary to the initial assumption that salary has little effect on the likelihood of a customer churning, the boxplot shows that customers with a higher "BalanceSalaryRatio " are more likely to churn. This observation is concerning to the bank because it implies that customers with a larger balance relative to their estimated annual salary are more likely to leave the bank, which could impact the bank's source of loan capital. In other words, it suggests that customers who may be more financially stable (as indicated by a higher "BalanceSalaryRatio") are still choosing to churn, which could have financial implications for the bank's lending activities.

4. Data Preparation:

- First we had to change the categorical variable of gender into a binary class with 0 assigned to male and 1 to female
- For the geography feature, we applied the one-hot encoding, resulting in additional 3 columns, one for each country.

5. Building the models:

We have utilized three different classification models which are the Logistic Regression, Support Vector Machine and Random Forest.

5.1 Logistic Regression :

We used two different models here, one primal logistic regression model and one with 2-degree polynomial kernel. Each was chosen beforehand via a parameter-grid.

	precision	recall	f1-score	support
0	0.81	0.98	0.89	1607
1	0.45	0.08	0.14	393
accuracy			0.80	2000
macro avg	0.63	0.53	0.51	2000
weighted avg	0.74	0.80	0.74	2000

	precision	recall	f1-score	support
0	0.83	0.97	0.89	1607
1	0.60	0.21	0.31	393
accuracy			0.82	2000
macro avg	0.72	0.59	0.60	2000
weighted avg	0.79	0.82	0.78	2000

5.2 Support Vector Machine :

	precision	recall	f1-score	support
0	0.81	0.98	0.89	1607
1	0.45	0.08	0.14	393
accuracy			0.80	2000
macro avg	0.63	0.53	0.51	2000
weighted avg	0.74	0.80	0.74	2000

5.3 Random Forest :

	precision	recall	f1-score	support
0	0.88	0.96	0.92	1607
1	0.76	0.46	0.58	393
accuracy			0.87	2000
macro avg	0.82	0.71	0.75	2000
weighted avg	0.86	0.87	0.85	2000

Random Forest Classifier yielded the best results.

6. Enhancements:

The imbalance in our dataset is a critical issue that significantly impacts the performance of our classification models. When we talk about data imbalance, we mean that one class of our target variable is disproportionately represented compared to the other class. In our case, it's evident that the number of customers who haven't churned far outweighs those who have.

This data imbalance can pose substantial challenges for machine learning algorithms. Classification models trained on imbalanced data tend to have a bias towards the majority class, making it difficult for them to accurately predict the minority class. As a result, our models may struggle to identify customers who are at risk of churning, which is a crucial task for any business.

To mitigate this issue and improve our models' performance, we can employ various data balancing techniques. **Oversampling** involves generating more instances of the minority class, effectively leveling the playing field and allowing the model to learn from a more balanced dataset. **SMOTE** (Synthetic Minority Over-sampling Technique) is a method that creates synthetic examples of the minority class by interpolating between existing instances.

Another approach is undersampling, where we reduce the number of instances in the majority class to match the minority class. **Hybrid methods**, which combine oversampling and undersampling, can also be effective. Additionally, techniques like cost-sensitive learning assign different misclassification costs to the classes to help the model prioritize the minority class.

By applying these data balancing strategies, we can enhance the performance of our classification models and make them more adept at correctly identifying customers at risk of churning, ultimately leading to more effective customer retention strategies and better business outcomes.