**Project Report**

**BDM500 - Tamanna Eini Keleshteri**

**Deep Thumar - 154113211**

**Title: Bank Customer Churn Prediction with ML**

**Description:** The objective of the project is to analyse dataset. Prepare dataset adequately for machine learning tasks and to find the best possible model and to train that model to get best possible accuracy with AUC, Recall and F1 score. So it can predict the churn of other dataset.

**Index**

1.Introduction

1.1 Data Exploration

2.KPIs (Key Performance Indicators)

2.1 Credit Score

2.2 Demographics

2.3 Customer Relationship

2.4 Financial

3.Data Correlation

3.1 Correlation

3.2 Skewness

4.Feature Engineering

4.1 One Hot Encoding

4.2 Normalisation

5. Balancing the dataset

6. Model Building

6.1 Split

6.2 Models

6.3 Prediction

6.4 Prediction after Feature Engineering

6.5 Prediction after Balancing

7. Prediction from chosen Models

8. Conclusion

**1. Introduction**

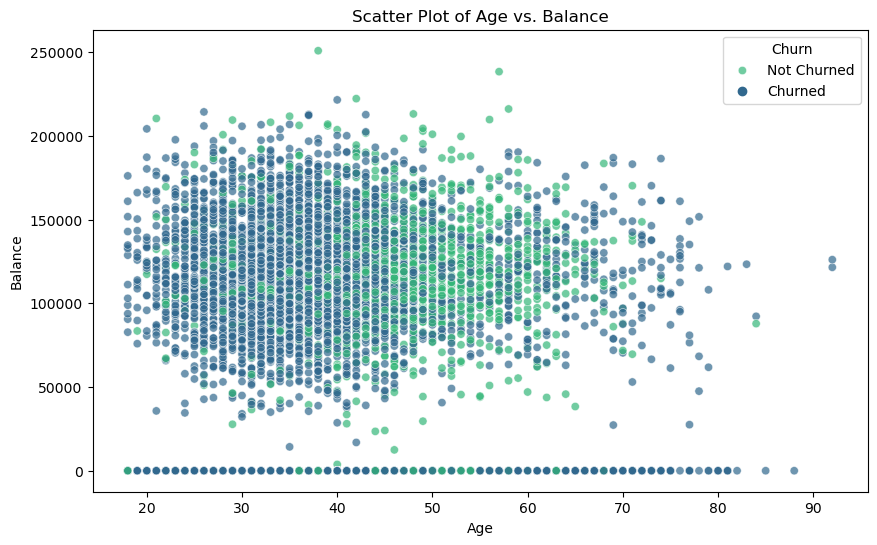
**1.1. Data set Exploration**

**Data:** Bank Customer churn Dataset

* Bank customer churn dataset comprises multiple features about the customers of the bank. Feature contains a variety of data, taking accounts of all kinds of customers.

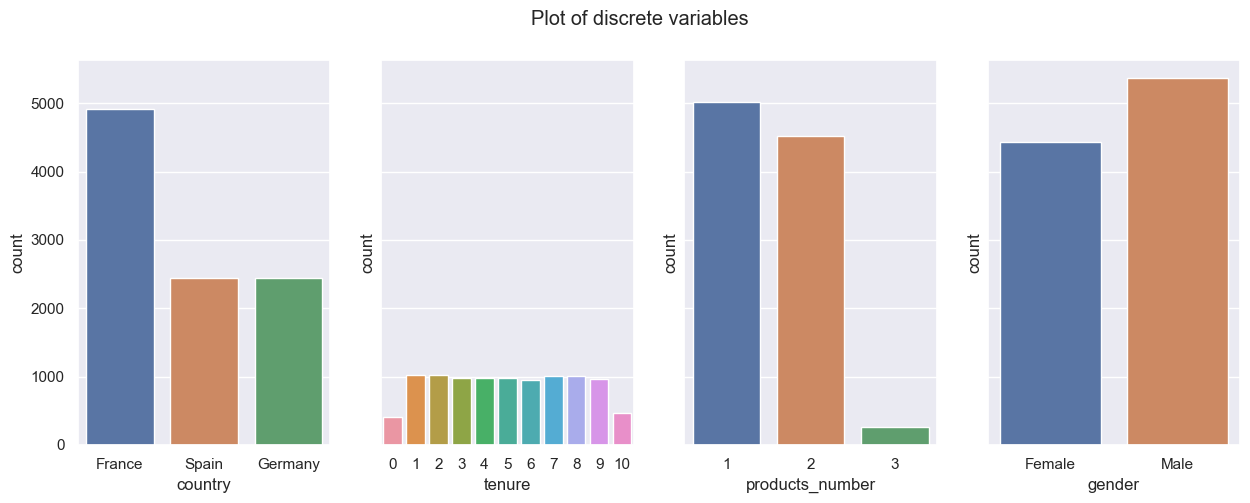
**Columns:** 'customer\_id', 'credit\_score', 'country', 'gender', 'age', 'tenure', 'balance', 'products\_number', 'credit\_card', 'active\_member', 'estimated\_salary', 'churn'.

* To take an example of how balance and age is showing churn rate:



**Figure 1.1 Scatter plot of Age vs. Balance**

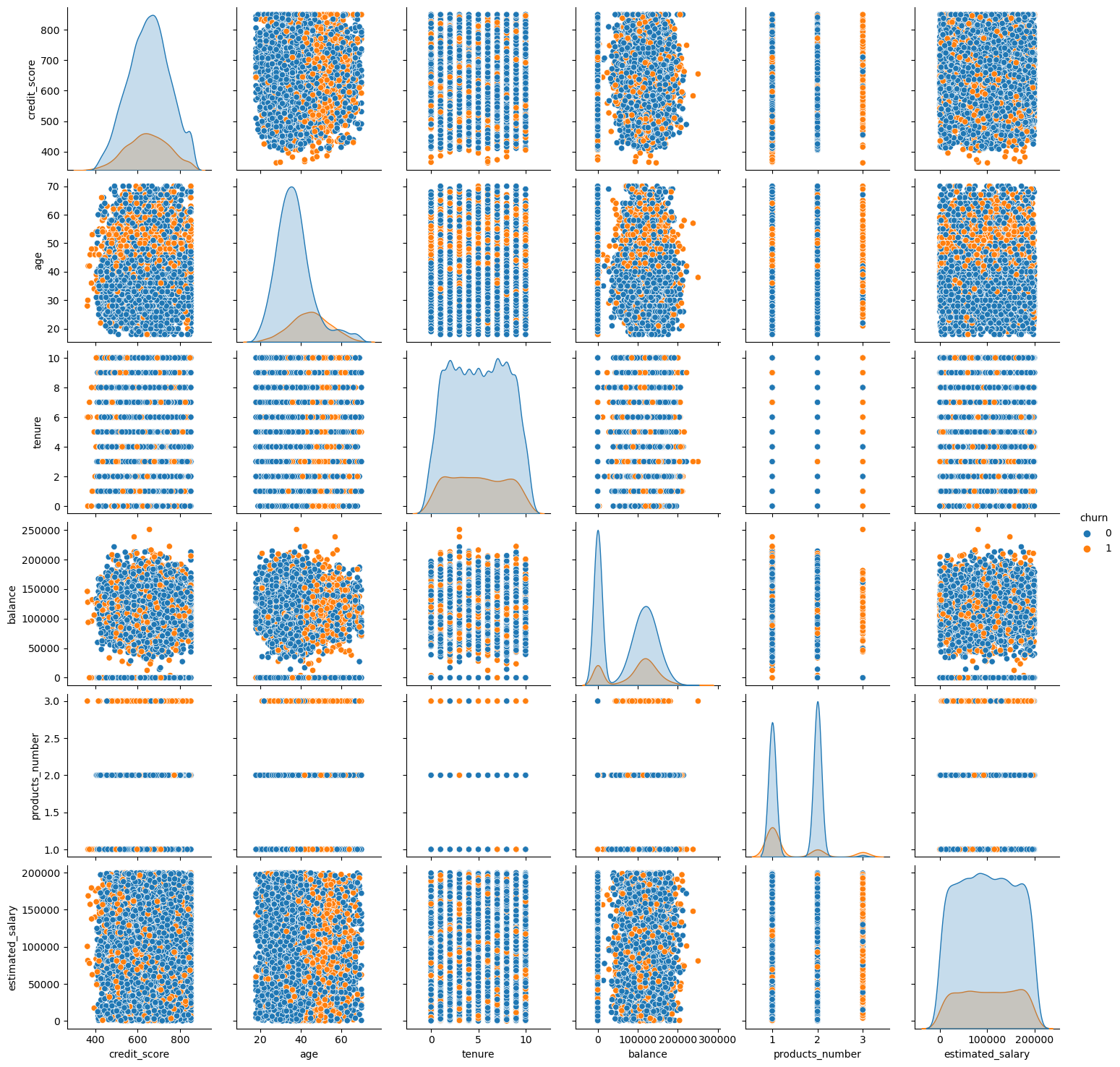
* This plot shows that the majority of people age 40 to 65 are not churned regardless of their balance while people under 40 are churned more so our focus can be on people under the age of 40

****

**Figure 1.1 Visualisation of discrete variable**

* The country column shows there are more customers from the france the spain and germany which contains almost same portion of customer
* Gender distribution is almost balanced
* Tenure is fairly distributed across 1 to 9 years.
* Also most of the customers have product number 1 and 2

**Shape:** There is total of 12 columns and 10000 rows



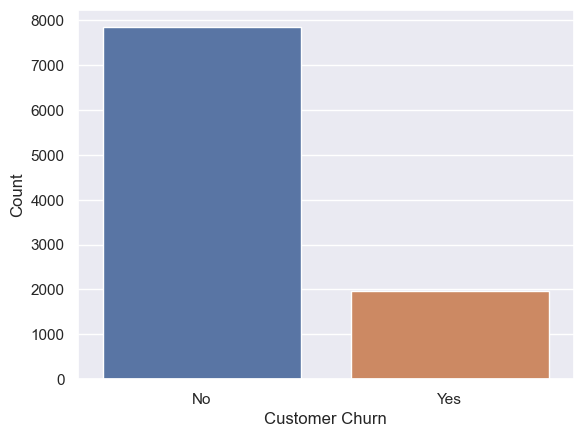
**Figure 1.1: Pair plot of features**

**Explanation:** By putting the churn in colours we can see that the dataset is imbalanced.

* We can also see that Churn is set for certain age range
* Also Churn is also correlated to the number of products. The more the products the higher the chances of customers leaving.

**Missing Values:** The is no missing values found in the dataset

**Outliers:** Outlier have been found with Z-score and have been removed for the dataset had very less outlier and it was irrelevant to the outcome



**Figure 1.2: Target variable**

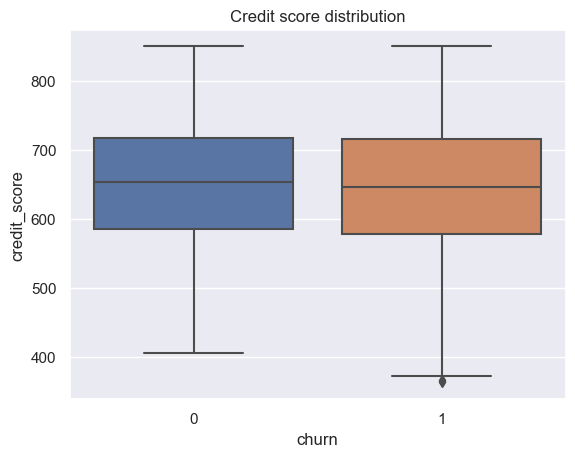
* **Target variable churn shows imbalance in its values which we will deal with as we go.**

**2. KPIs (Key Performance indicators)**

**2.1 Credit Score**

**Metric:** Credit score

* Average credit score for churning and non-churning customers



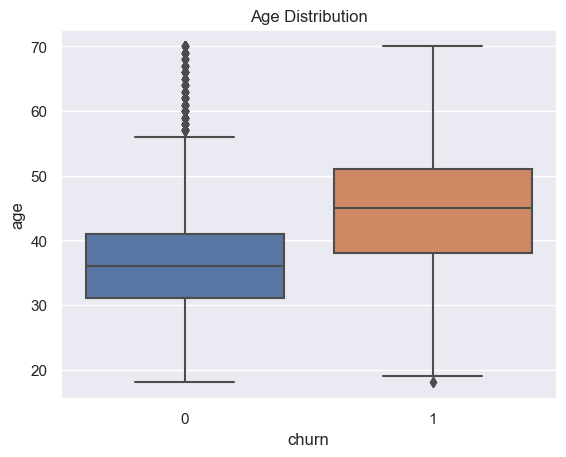
**Figure 2.1 Credit score**

* **In regards to credit score both churned and not churned customers are equally distributed between scores of 600 to 700.**

**2.2 Demographics**

**Metric:** Gender, Country and Age

* Gender, country, and Age distribution for churning and non-churning customers

****

**Figure 2.2 Age Distribution**

* This Box plot shows Customers who are churn are in between age 40 to 50 and not churned customers are in 40 to 30 years of age.

**2.3 Customer Relationship**

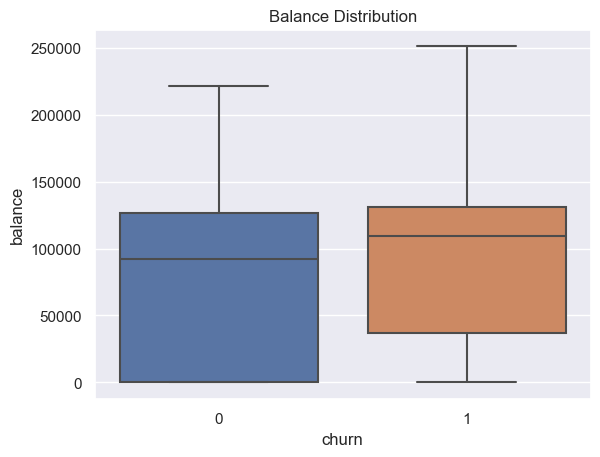
**Metric:** Tenure, Product Number, Active Credit Card, Active Member

* Average Tenure, Percentage of customer with Credit card, Percentage of Active member for churn and non-churn

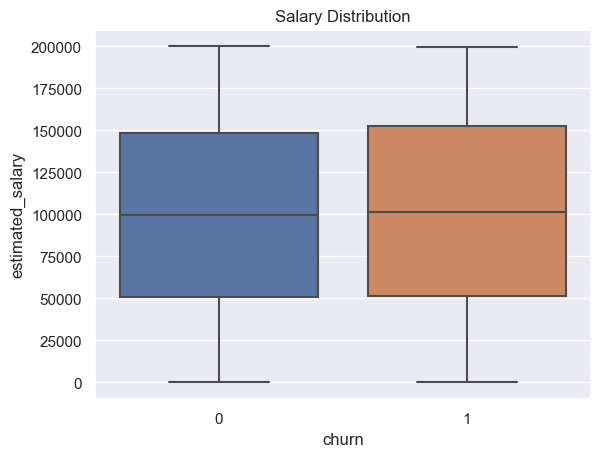
**2.4 Financial**

**Metric:** Balance, Salary

* Balance and salary correlation for churn and non-churn customers

****

**Figure 2.3 Balance Distribution**

****

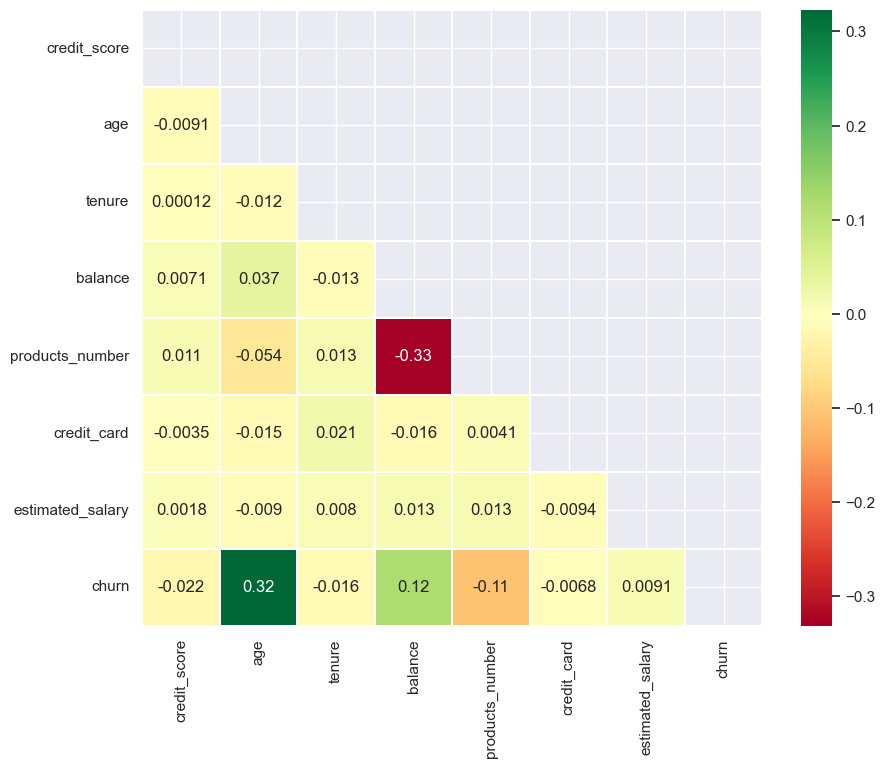
**Figure 2.4 Salary Distribution**

* Balance and salary distribution shows that these features are not affecting overall churn rate, in fact the customers who are churning has some balance in their account before churning.

**3. Data correlation**

**3.1 Correlation**

* The correlation between features is very less in this dataset but for making the model and prediction it would not be a factor.

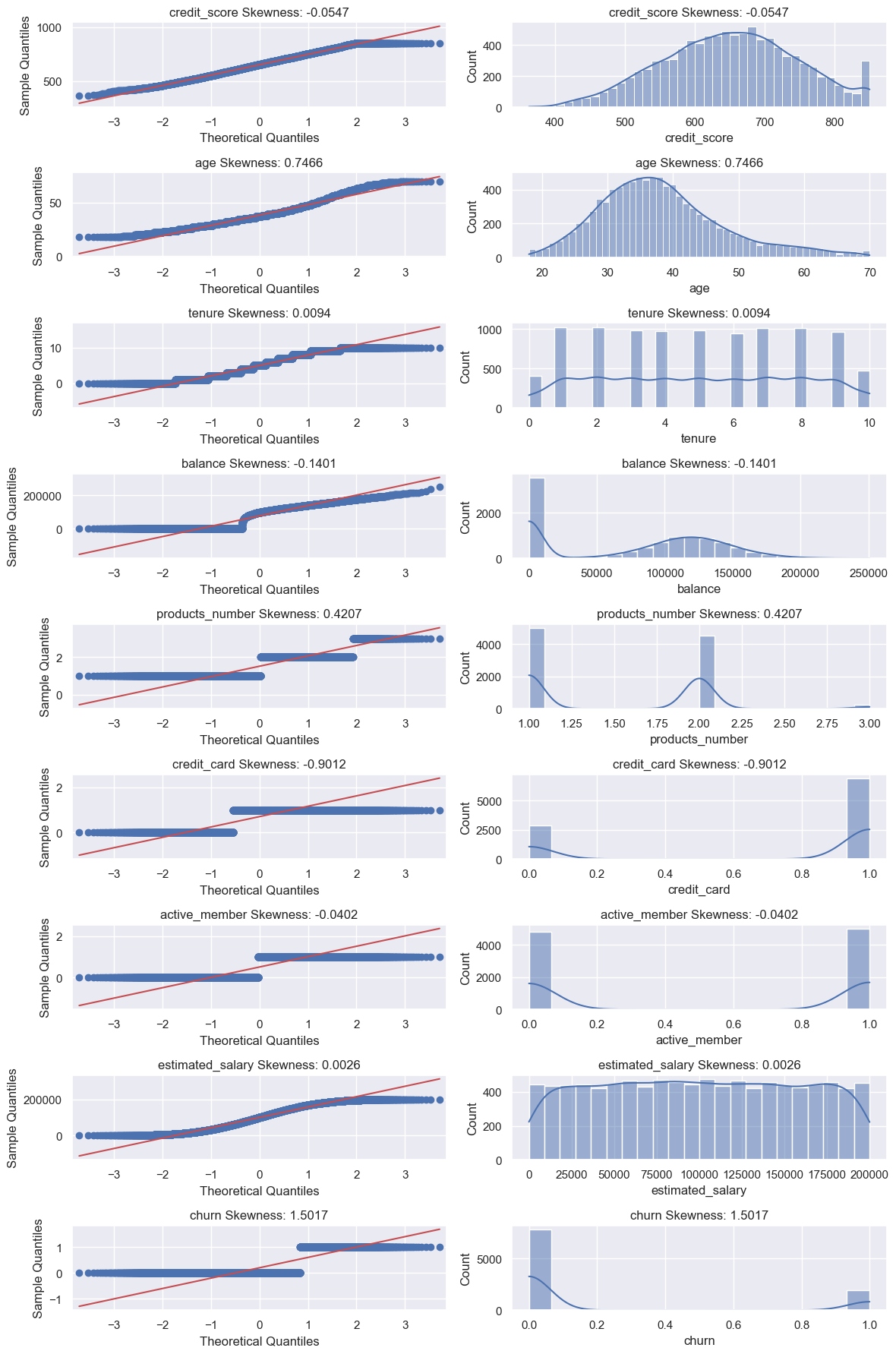


**Figure 3.1 Correlation Heatmap**

* Majority of the columns do not show major correlation.
* Whereas Age to churn shows positive correlation and Product number to balance shows fairly negative correlation which i’ve mentioned above.

**3.2 Skewness**

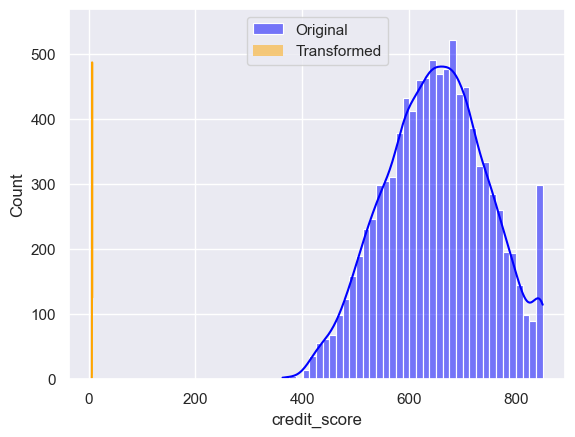
* Skewness of the dataset shows its asymmetry of distribution.



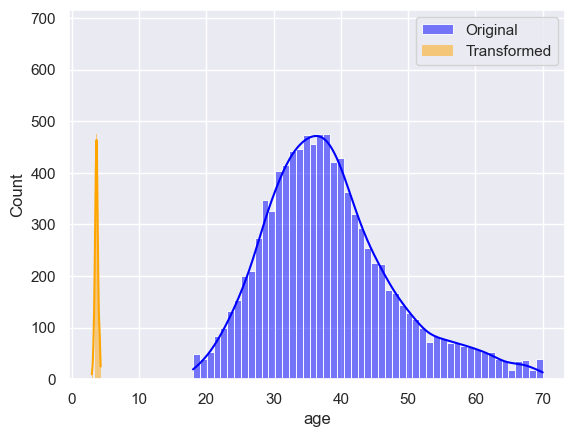
**Figure 3.2 Skewness Visualization**

* This chart shows all the features and its skewness which is daily distributed with Age and Credit score leaning left and right consecutively.

**Transformation:**

****

**Figure 3.3 After skewing credit score feature**

****

**Figure 3.4 After skewing Age feature**

* This chart shows credit score and age column after its transformation from the Log1p function of Numpy.

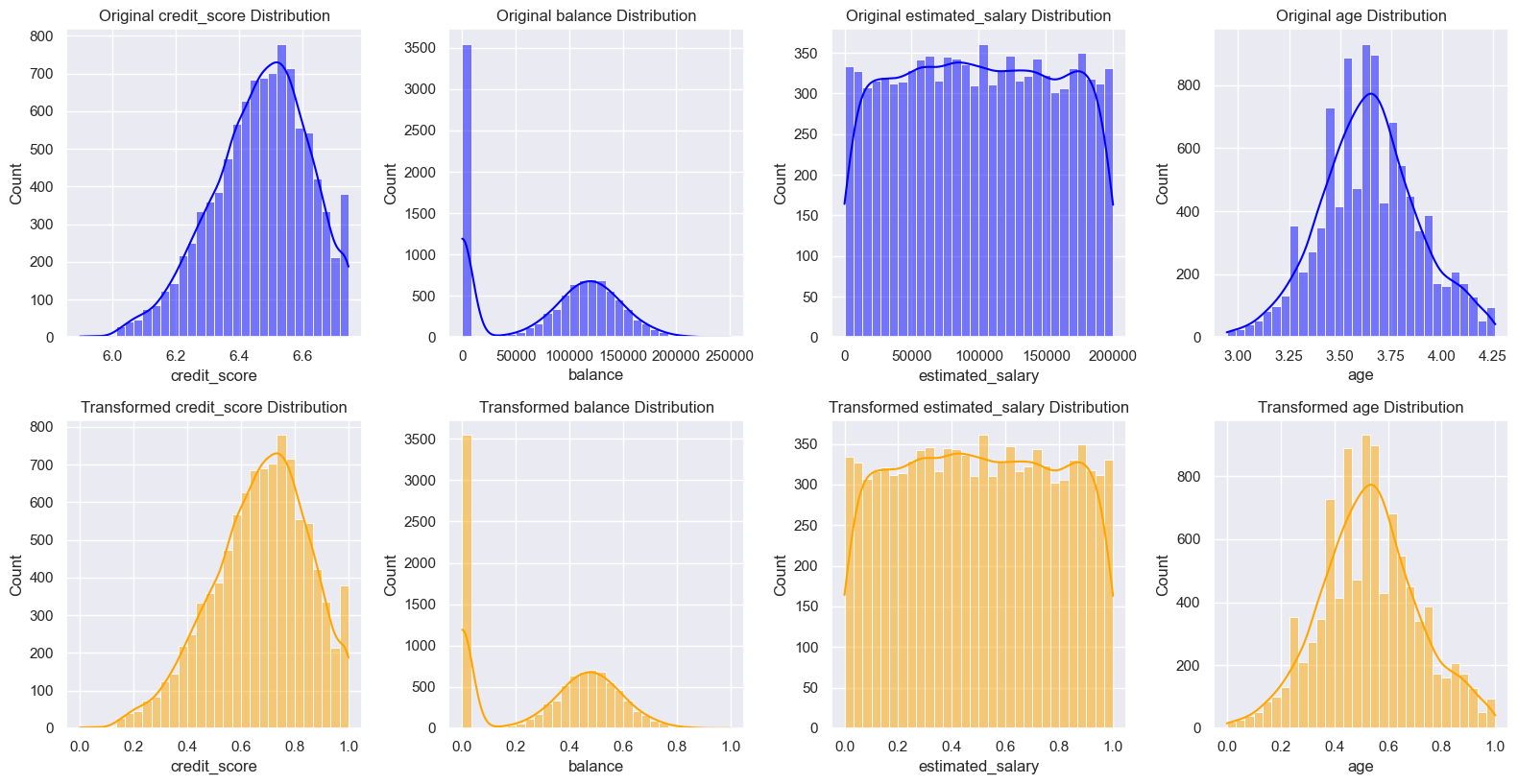
**4. Feature Engineering**

* Feature engineering is the process of normalising the features by using scale function on numerical features and one hot encoding all the categorical features.

**4.1 One Hot Encoding**

* One Hot Encoding is been done on gender and Country columns

**4.2 Normalisation**

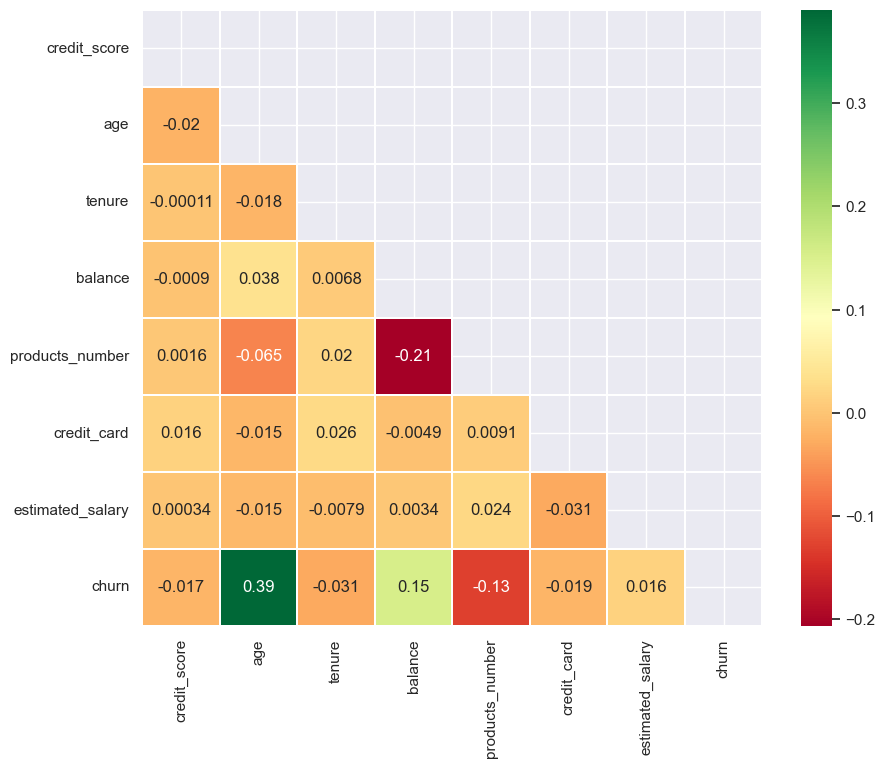
****

**Figure 4.1 Normalisation**

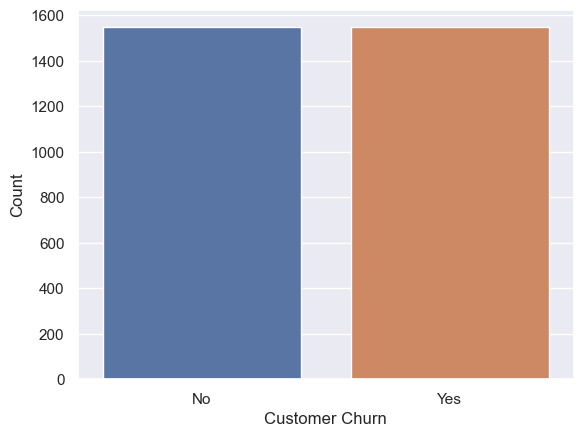
* This plot shows scaled numerical features

**5. Balancing the dataset**

* As shown before The data is imbalance and need to be balance for further predictions



**Figure 5.1 Heatmap after Balance**

****

**Figure 5.2 churn**

* As we can see The target variable is fairly balanced by now and we are ready for further modelling

**6. Model Building**

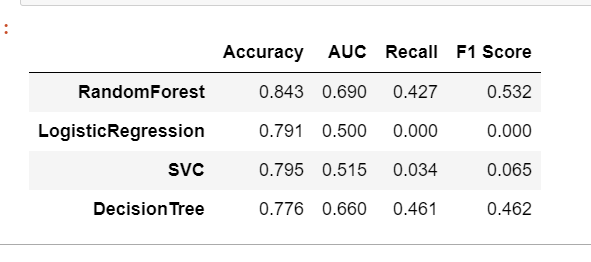
**6.1 split**

* First i ran train and test split on the data

**6.2 Models**

* I took 4 classification models for my dataset for the target variable is distributed in 0 and 1
* The models are Random forest, Logistic Regression, Support vector Machine and Decision tree classification.

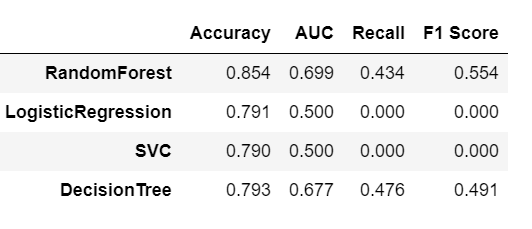
**6.3 Prediction**

****

**Figure 6.1 Predictions of Models on raw dataset**

* The result shows that Logistic regression and SVC is performing poorly for this dataset

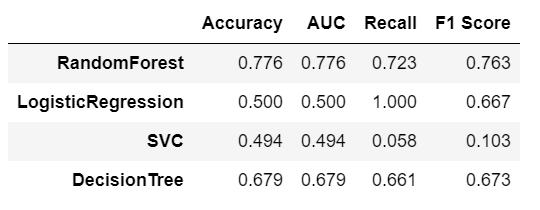
**6.4 Prediction after future engineering**

****

**Figure 6.2 Result after Feature engineering**

* After feature engineering the overall result hasn’t improved much and Logistic and SVC’s Recall and F1 score is showing 0 indicating it’s not able to predict positive values at all because of imbalance in the dataset

**6.5 Prediction after Balancing**

****

**Figure 6.3 Predictions after balance data**

* After balancing the event the accuracy is dropped but the recall and f1 score is improved on random forest and decision tree which shows proper result and other models we will disregard.

**7. Prediction from chosen models**

'credit\_score': [750, 680, 800],

'age': [35, 45, 28],

'tenure': [5, 10, 2],

'balance': [5000, 2000, 8000],

'products\_number': [2, 3, 1],

'credit\_card': [1, 0, 1],

'active\_member': [1, 1, 0]

For this values i made an predictions and it came as this,

Predictions for RandomForest: [0 0 1]

Predictions for DecisionTree: [0 0 1]

Both models are showing similar results with an accuracy of 77% for Random forest and 68% for Decision tree classifier.

**8. Conclusion**

In conclusion, this project involves in depth analysis on the bank customer churn dataset. The EDA (Exploratory data analysis) has been done on the dataset and key performance indicators such as demographics, credit score and customer relationship metrics we carefully analysed. Normality of the data and skewness is shown with transformation.

Also the dataset is gone through feature engineering and balancing which was necessary for the dataset on hand. The models were created after these processes and evaluation of the model is done through techniques such as accuracy, AUC, Fi-score and recall. Four models on classification were used, including Random forest, Logistic, Support vector machine and decision tree. The models can run on different dataset and will be able to predict with given accuracy.

* The goal of this project was to accurately predict customer churn on given features which I found is done perfectly.
* The data set was imbalanced so i found it makes the model not fitting for outside data due to it’s f1-score being so low so i had to balance the data first
* For model selection, i just took 4 highly used classification models and used them
* For the result I used accuracy, AUC, F1 score and recall. These are the best evaluation techniques I found because it tells you the condition of the dataset as well.
* For churn value prediction there are no time constraints. Any dataset with the same features will get predicted by these trained models.

**References:**

**For dataset:**

[**https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction**](https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction)

**For Models:**

[**https://www.geeksforgeeks.org/getting-started-with-classification/**](https://www.geeksforgeeks.org/getting-started-with-classification/)

**For Evaluation methods:**

[**https://www.geeksforgeeks.org/machine-learning-model-evaluation/**](https://www.geeksforgeeks.org/machine-learning-model-evaluation/)