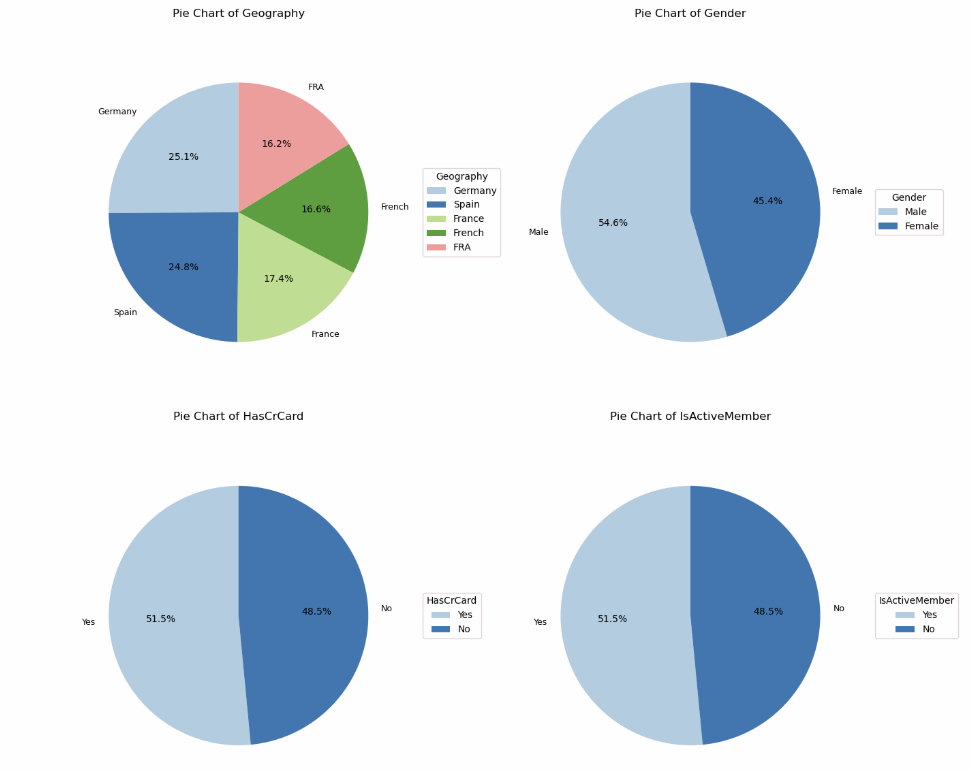


**Checking Data Imbalance**

**Data is quiet Imbalanced**

**Categorical Column Analysis**

**Distribution**[**¶**](http://localhost:8888/notebooks/Desktop/Learnabay%20Training/My%20Portfolio%20projects%20for%20resume/Placement%20Project%20By%20learnbay/Customer%20Churn%20Analysis%20and%20prediction%20Model/Bank%20Customer%20Churn%20Modelling.ipynb#Distribution)

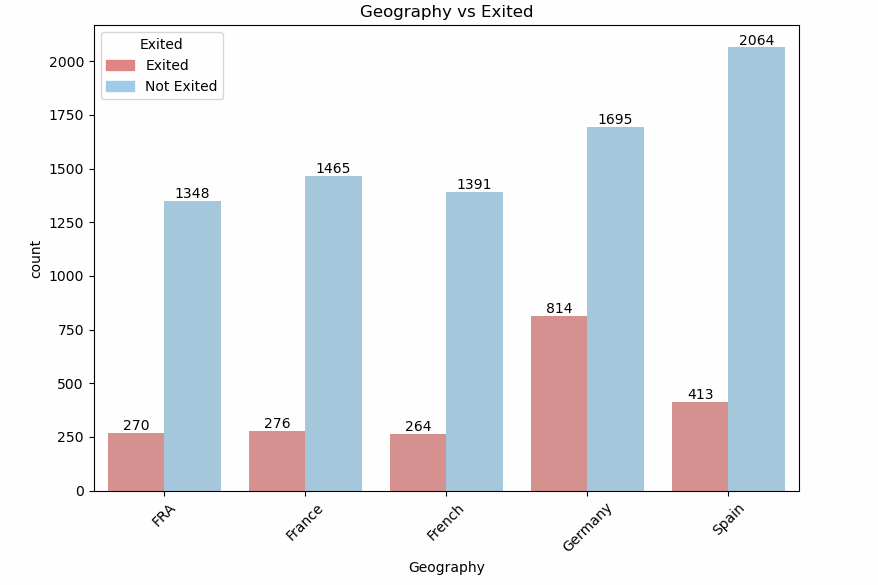
****

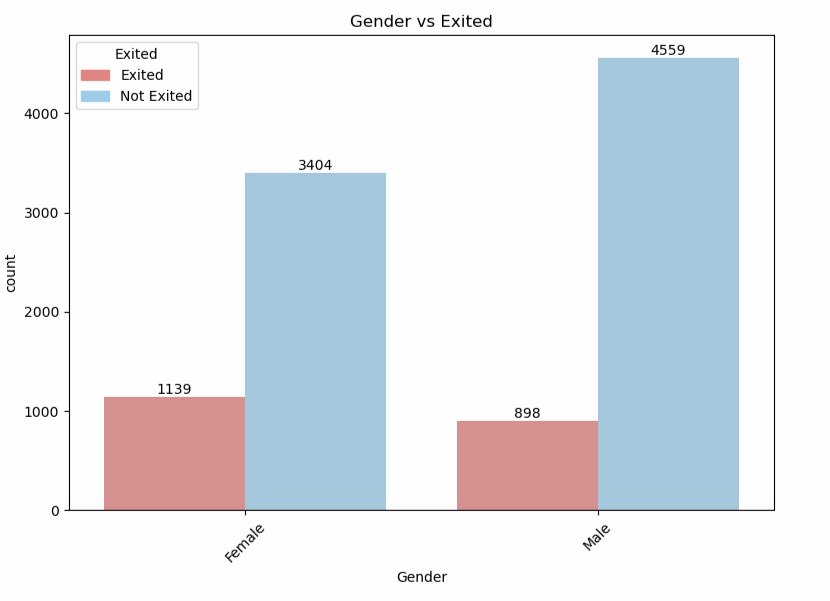
# Observation:

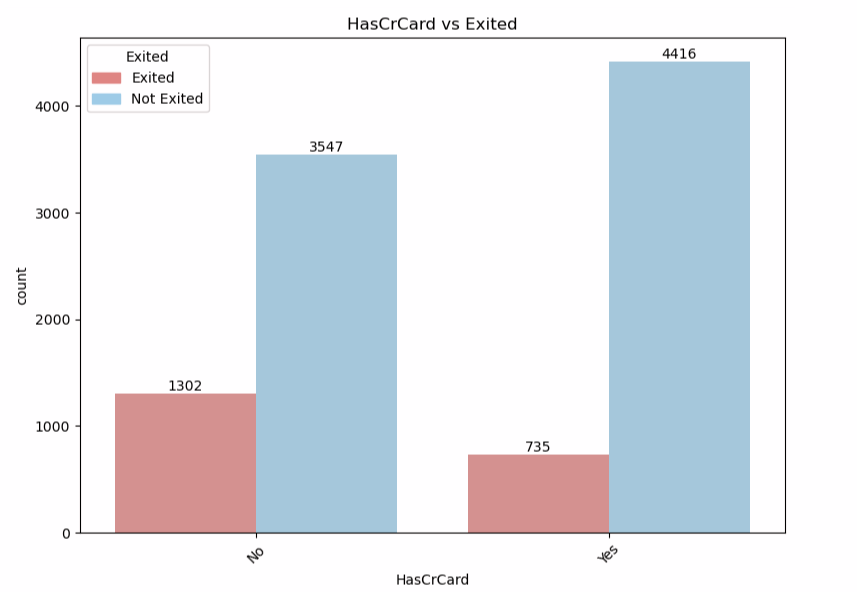
# -- there are more male customers as compared to female customers

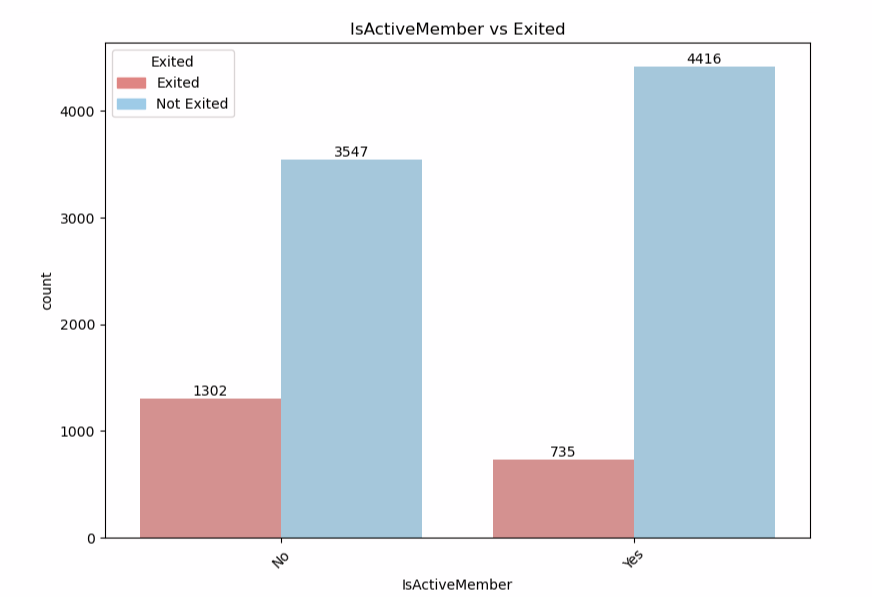
# -- distribution of other categorical variables are quiet uniform

Category, Subcategory and its relation with the target variable[¶](http://localhost:8888/notebooks/Desktop/Learnabay%20Training/My%20Portfolio%20projects%20for%20resume/Placement%20Project%20By%20learnbay/Customer%20Churn%20Analysis%20and%20prediction%20Model/Bank%20Customer%20Churn%20Modelling.ipynb#Category,-Subcategory-and-its-relation-with-the-target-variable)









# Observations :

# -- Germany has the most number of churners followed by spain

# -- Female Churners are more as compared to male

# -- Those not having credit cards tend to churn more

# -- those who are not active members tend to churn more

Statistical Association Test using Chi-Square.[¶](http://localhost:8888/notebooks/Desktop/Learnabay%20Training/My%20Portfolio%20projects%20for%20resume/Placement%20Project%20By%20learnbay/Customer%20Churn%20Analysis%20and%20prediction%20Model/Bank%20Customer%20Churn%20Modelling.ipynb#Statistical-Association-Test-using-Chi-Square.)

Chi2 p-value

Geography 301.678041 4.707920e-64

Gender 112.918571 2.248210e-26

HasCrCard 242.985342 8.785858e-55

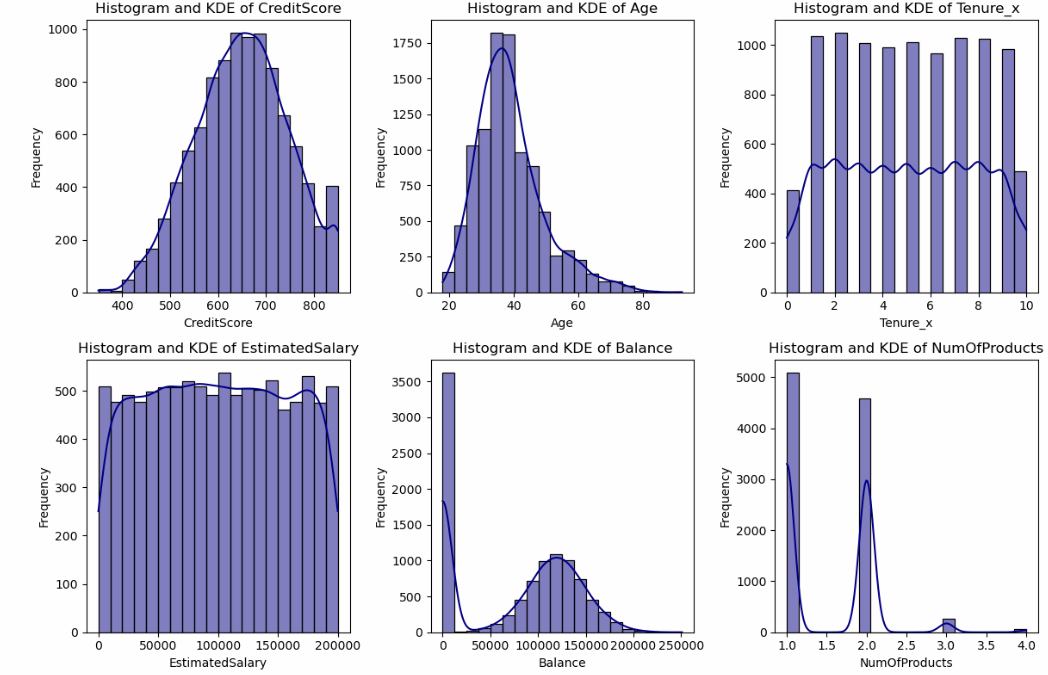
IsActiveMember 242.985342 8.785858e-55

'''observation - Chi-Square test results suggest strong associations between categorical features and the target variable (Exited), given the

very low p-values (all < 0.05). So we will include all the categorical columns in our data.'''

Numerical column Analysis

Column Distribution[¶](http://localhost:8888/notebooks/Desktop/Learnabay%20Training/My%20Portfolio%20projects%20for%20resume/Placement%20Project%20By%20learnbay/Customer%20Churn%20Analysis%20and%20prediction%20Model/Bank%20Customer%20Churn%20Modelling.ipynb#Column-Distribution)



# Observations:

# -- credit score data is slightly left skewed

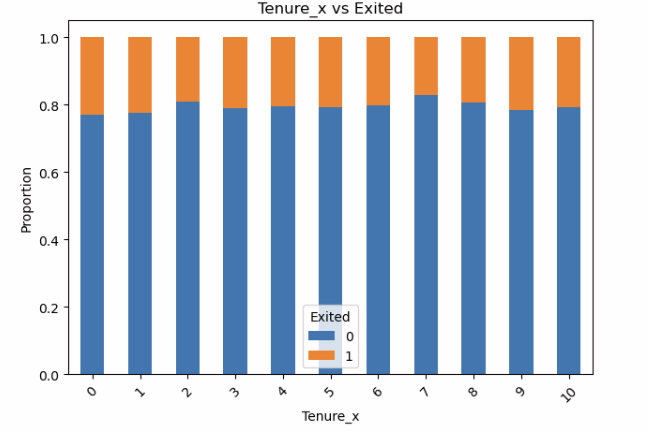
# -- Age is right skewed

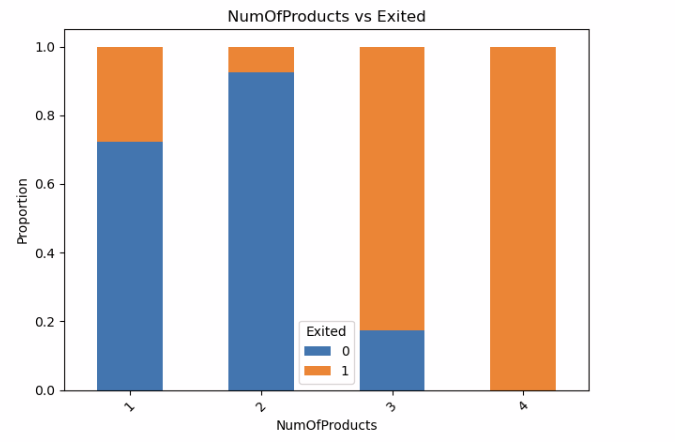
# -- Tenure and No. of products are discrete numerical columns

# --Estimated Salary has a flat curve

# --Balance is mostly noraml except skewed for 1 data point.

Checking relationship between discrete variables and target variable





# # Observations :

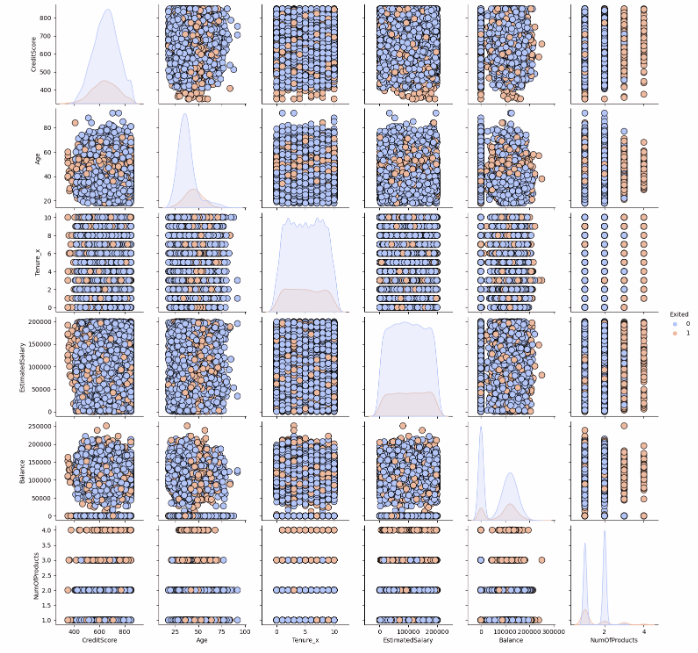
# -- there is no indicative pattern between tenure and Exited

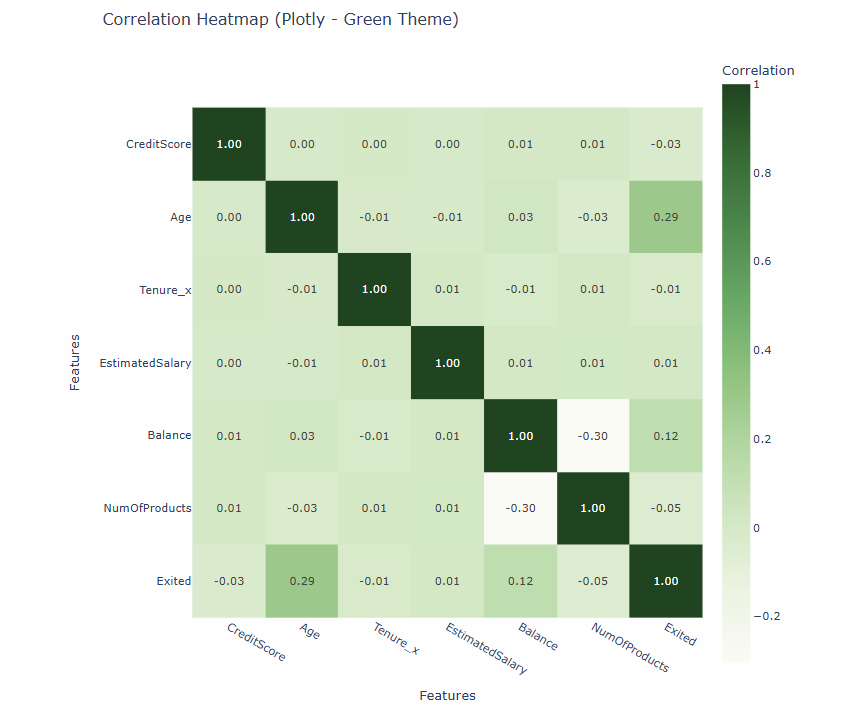
# -- Customer having 1 or 2 products tend to churn the least

# -- Customer with 3 or 4 have churned more

# -- customer with 4 products have all churned

Checking relationship of Numerical Varibles with each other and the target variable





# # Observation :

# # -- As shown by the scatter plots and the heatmap there is no linear relationship between the numerical variables as points are randomly scattered and

# the pearson r coefficient values is low to moderate

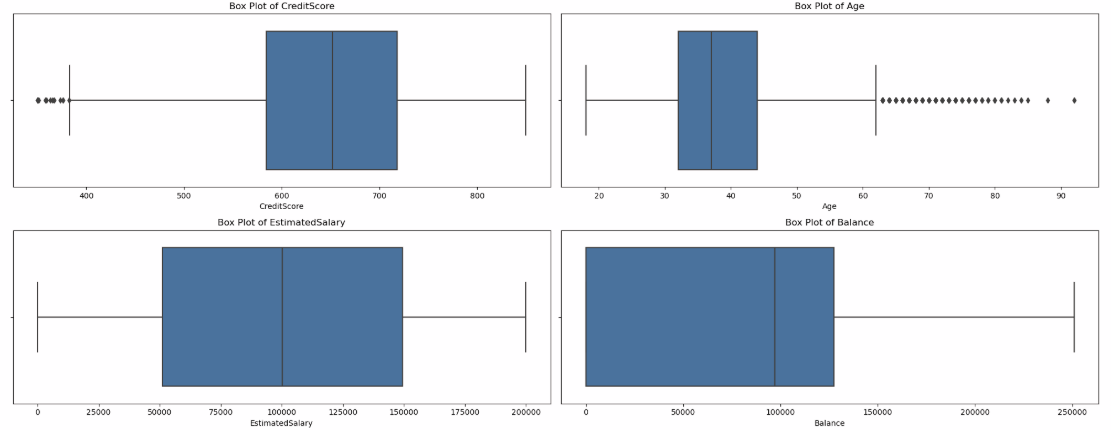
Statistical Significance Using T-test

# # Observation:

# --- As shown by the p values of t-test we can say that Credit Score,Age,Balance,NumOfProducts have significant relationship with the target variable

# -- where as Estimated Salary and Tenure\_x does not have much significance.

Boxplots to check for distribution and outliers



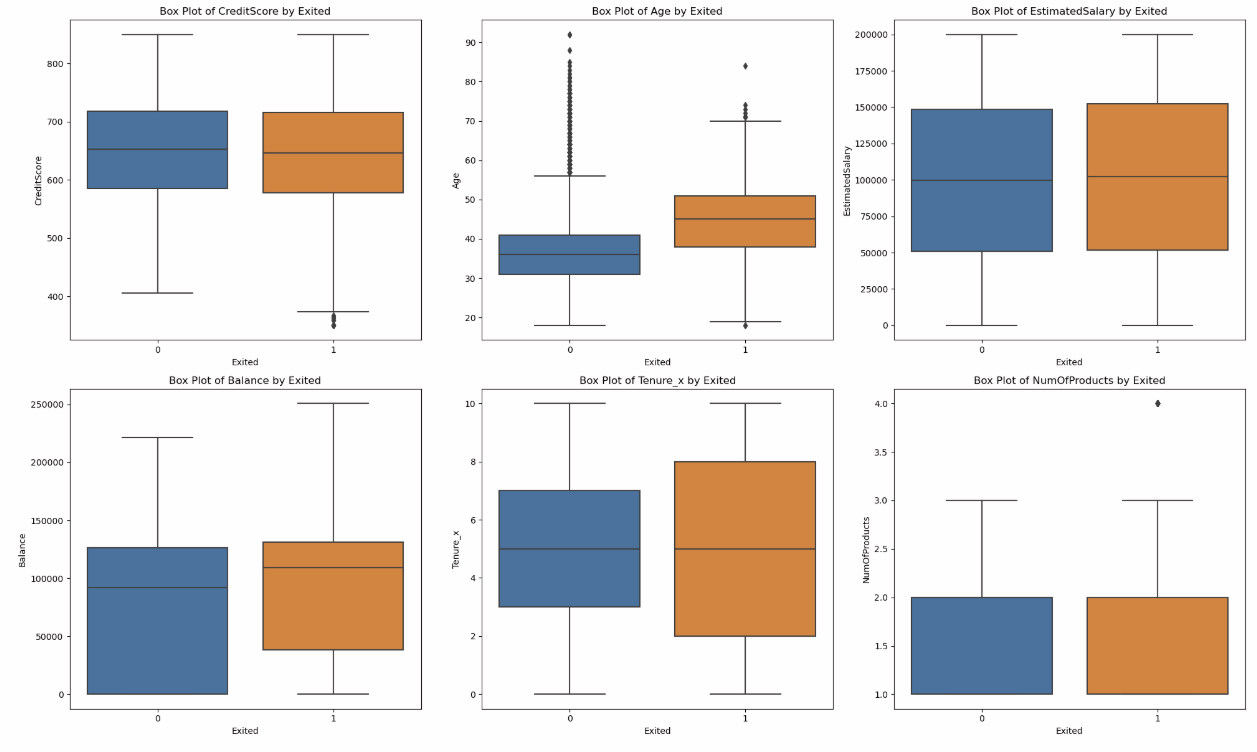
# # Observations:-

# -- We can see some outliers in the credit score and Age columns

# --Estimatedsal column look normal

# --Balance column is hightly right skewed

Multivariable box plots



Feature Engineering and Preprocessing strategy

# -- giving and indicator column to show whether balance is 0 or not.

# Imputation Strategy -

# -Impute using simple Imputer with median for age,estimate,balance,credit\_score.

# -Impute tenure by mean

# - for any other columns in the data make the input compulsory.

# Encoding Strategy -

# -- Gender, hascrcard, Isactive use onehot encoding ,( you can drop column to reduce multicollinearity)

# -- Label encoding for Geography

# Outlier strategy:

# -- compare between log transformation and IQR capping and check the results

# Scaling:

# -- simply do standard scaling

# Base Model:-

# -- try classifier models like Logistic regression,svm,random forest,xgboost all with the weight parameter as balanced

After this I made 2 different pipelines tried all the base models, selected the best model to tune which was svm and xgboost, finally went ahead with tuned and regularised xgboost for the final pipeline

Final train and test scores were like

📘 Train Classification Report:

precision recall f1-score support

0 0.94 0.83 0.88 6370

1 0.55 0.81 0.65 1630

accuracy 0.83 8000

macro avg 0.75 0.82 0.77 8000

weighted avg 0.86 0.83 0.84 8000

Train ROC AUC: 0.9042672226984233

📗 Test Classification Report:

precision recall f1-score support

0 0.93 0.82 0.87 1593

1 0.51 0.75 0.61 407

accuracy 0.80 2000

macro avg 0.72 0.78 0.74 2000

weighted avg 0.84 0.80 0.82 2000

Test ROC AUC: 0.8590454861641303

🔍 Comparison Summary:

Metric Train Test

Accuracy 0.8255 0.8030

Precision 0.5488 0.5108

Recall 0.8074 0.7543

F1 Score 0.6534 0.6091

ROC AUC 0.9043 0.8590

