Bank Customer Churn Prediction

MIS S381N - Data Science Programming Project

Neel Sheth - NDS967
Samarth Mishra - SM79247
Srividya Rayaprolu – LR34488
Tanushree Devi Balaji – TB33857

Agenda

Bank Customer Churn Classification

01 Introduction

02 Exploratory Data Analysis
Initial Analysis, Outliers, Feature Creation

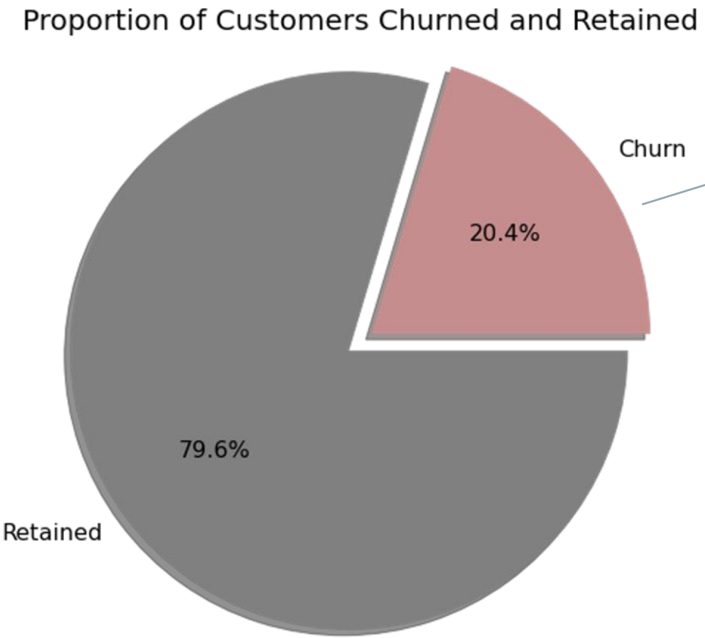
03 Modelling
KNN, Naïve Bayes, Logistic Regression, Trees

04 Model Selection
Accuracy Rate Comparison

05 Recommendations and Conclusion

INTRODUCTION

Overview of Problem Statement



A 20% churn rate translates to losing roughly 1.5M euros in bank balance per 100 customers

What is Churn Rate?
Measure of waning customer engagement

Objective:
Identify potential churners early on and formulate a retention strategy

Approach:
Build a model to predict churn propensity at a customer granularity

Data & Attributes

Location

Account Balance

Credit Card?

Credit Score

Age of the Customer

Estimated Salary

Activity

Number of Bank Products

Ref:https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/data

EXPLORATORY DATA ANALYSIS

Sanity Checks

```
df.nunique()
```

```
RowNumber          10000
CustomerId         10000
Surname             2932
CreditScore          460
Geography              3
Gender                 2
Age                   70
Tenure                11
Balance             6382
NumOfProducts          4
HasCrCard              2
IsActiveMember         2
EstimatedSalary     9999
Exited                 2
dtype: int64
```

Could be bucketed for enhanced readability

```
df.isnull().sum()
```

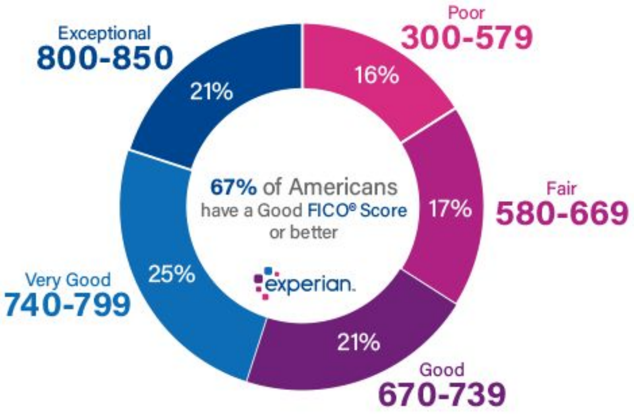
```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

Clean Dataset

Feature Creation

These features would be more meaningful for models like Naïve Bayes that solely rely on categorical inputs

Credit Score Buckets



Ref: https://www.experian.com

Age Buckets

18 to 25 years	*Young Adult*
26 to 35 years	*Adult*
36 to 68 years	*Middle Age*
69 to 80 years	*Early Retirement*
Over 81 years	*Old*

% of Credit Card Ownership

$$\frac{\#\ Credit\ Cards}{\#\ Total\ Products}$$

Balance Bucket

Avg ~ 76k euros

Zero

Below Average

Above Average

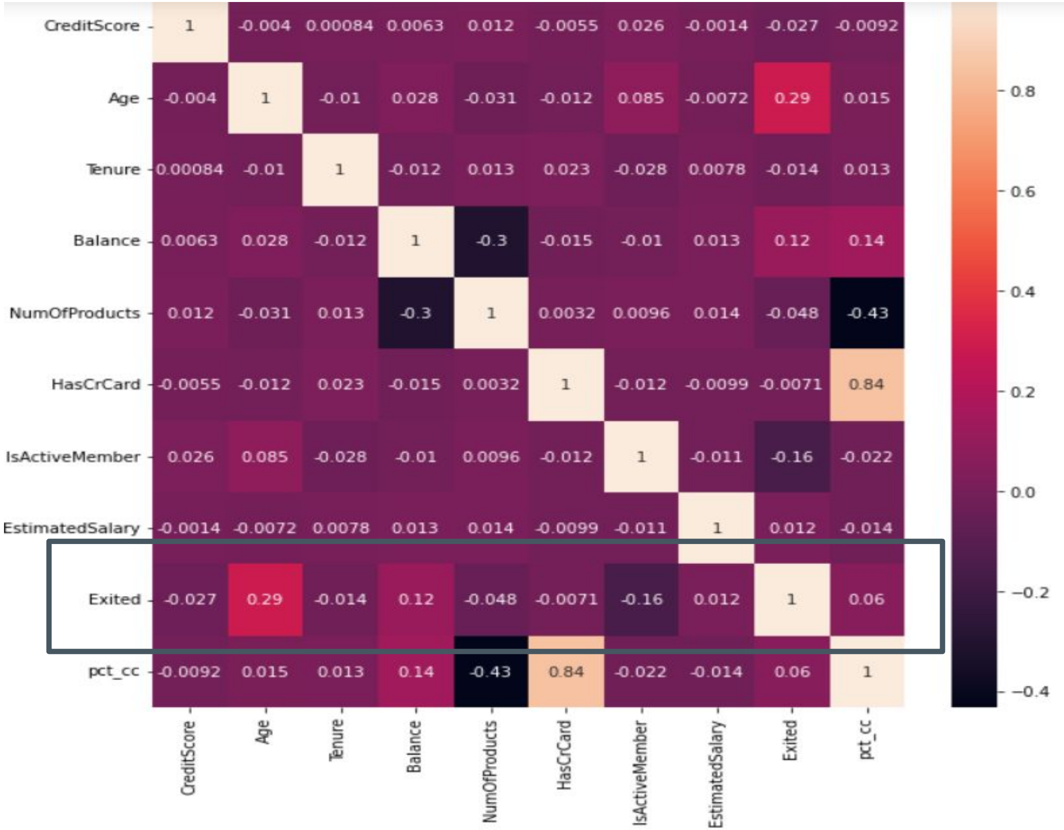
Salary Bucket

Avg ~ 100k euros

Below Average

Above Average

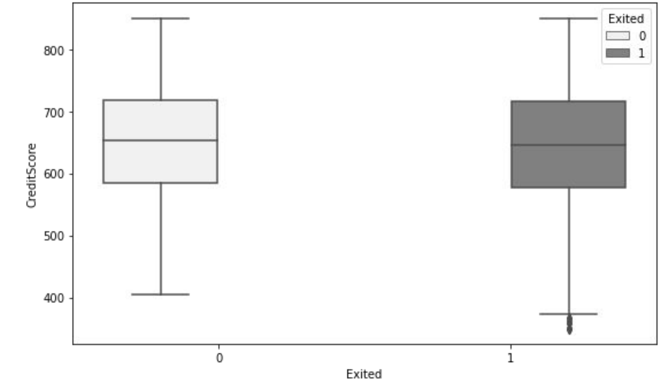
Correlation Matrix

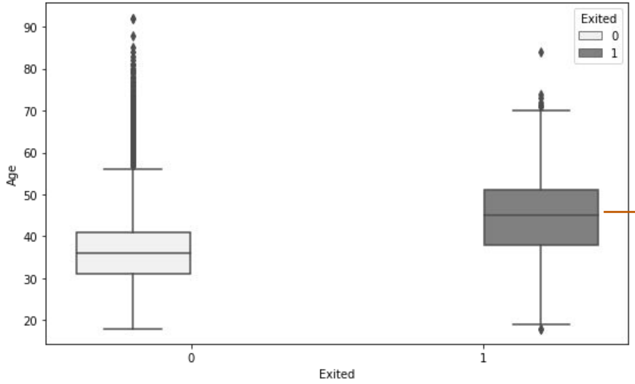


From the correlation matrix, we can see that none of the variables have strong linear relations with the 'exited' variable except age

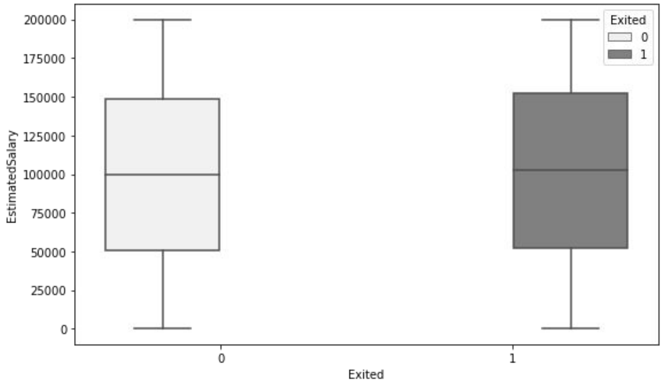
Numerical Variable Summary

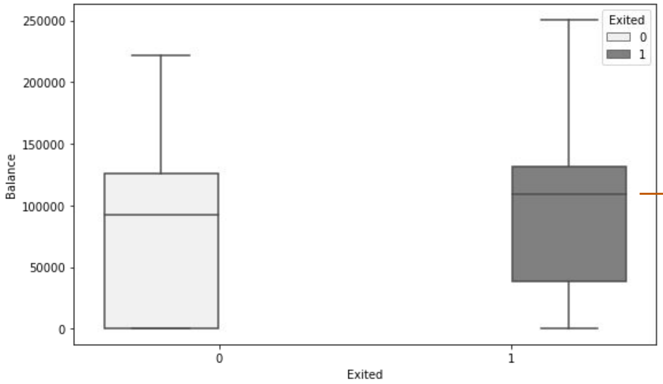
Predictors like Credit Score & Age have outliers which would be handled inherently by the models





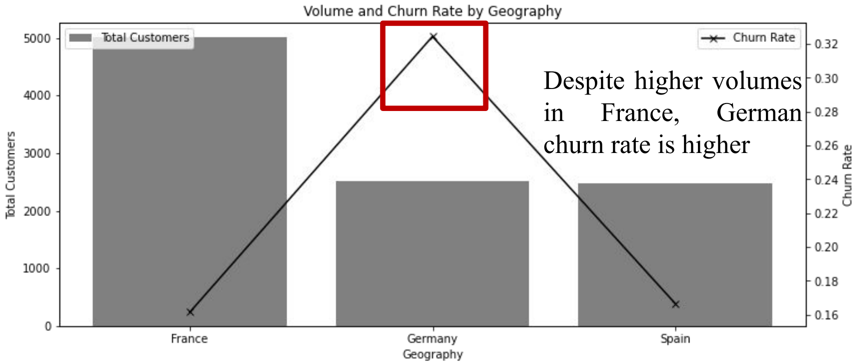
Churners seem 40-50 years old on average

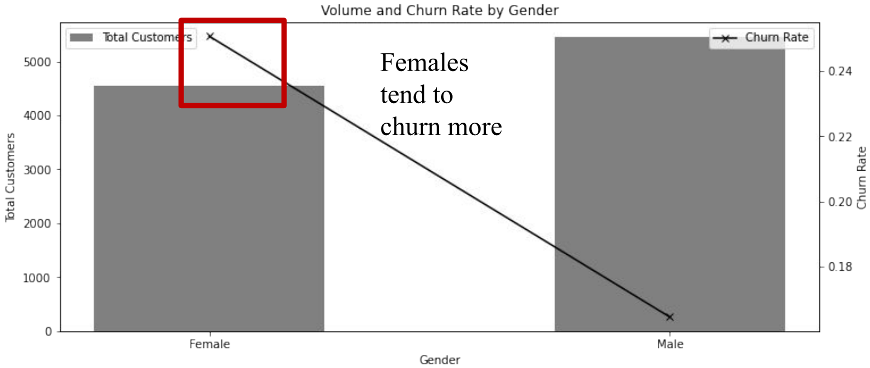


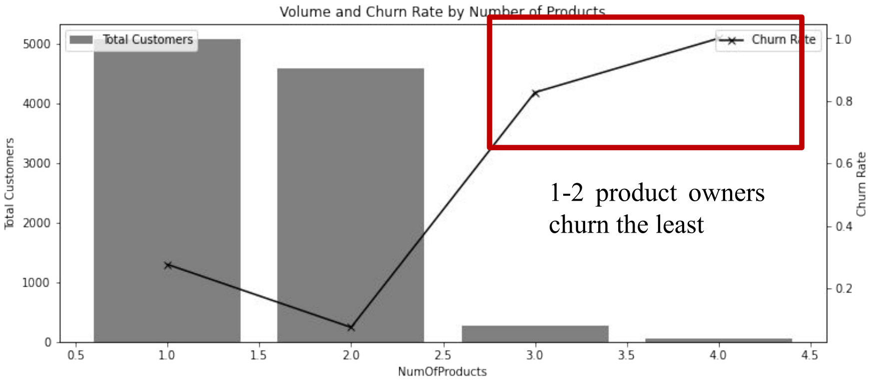


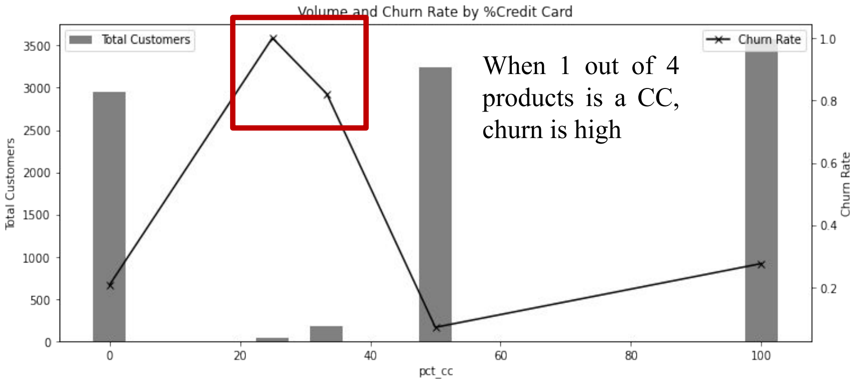
They also tend to have higher bank balances

Categorical Variable Summary

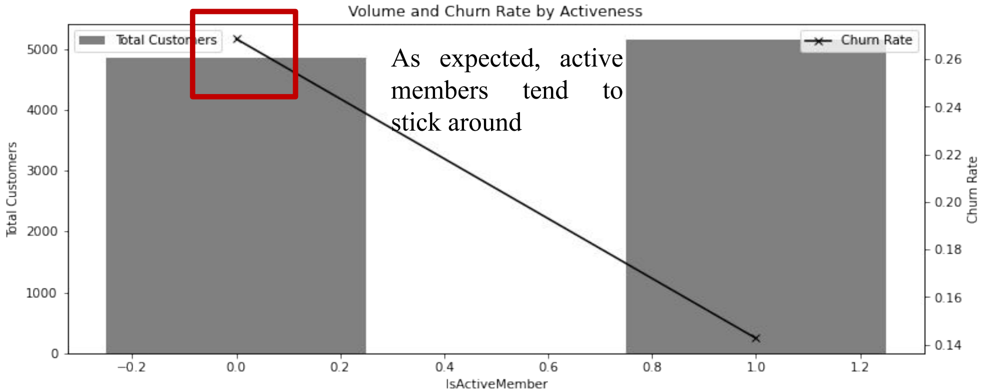




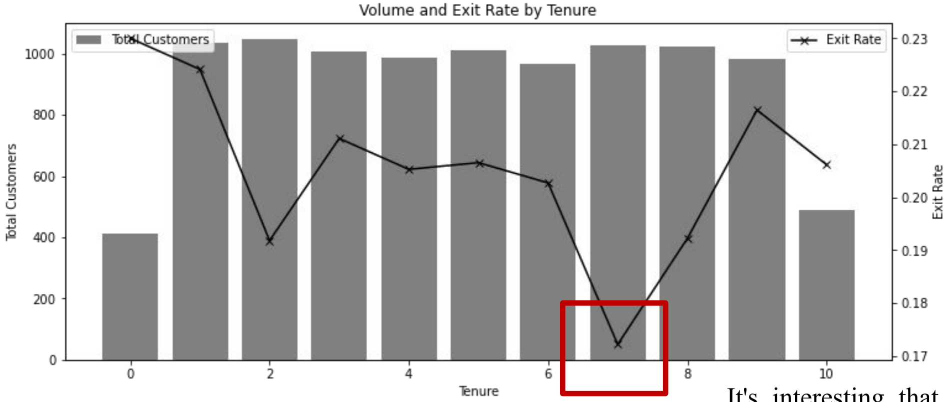




Categorical Variables Contd.



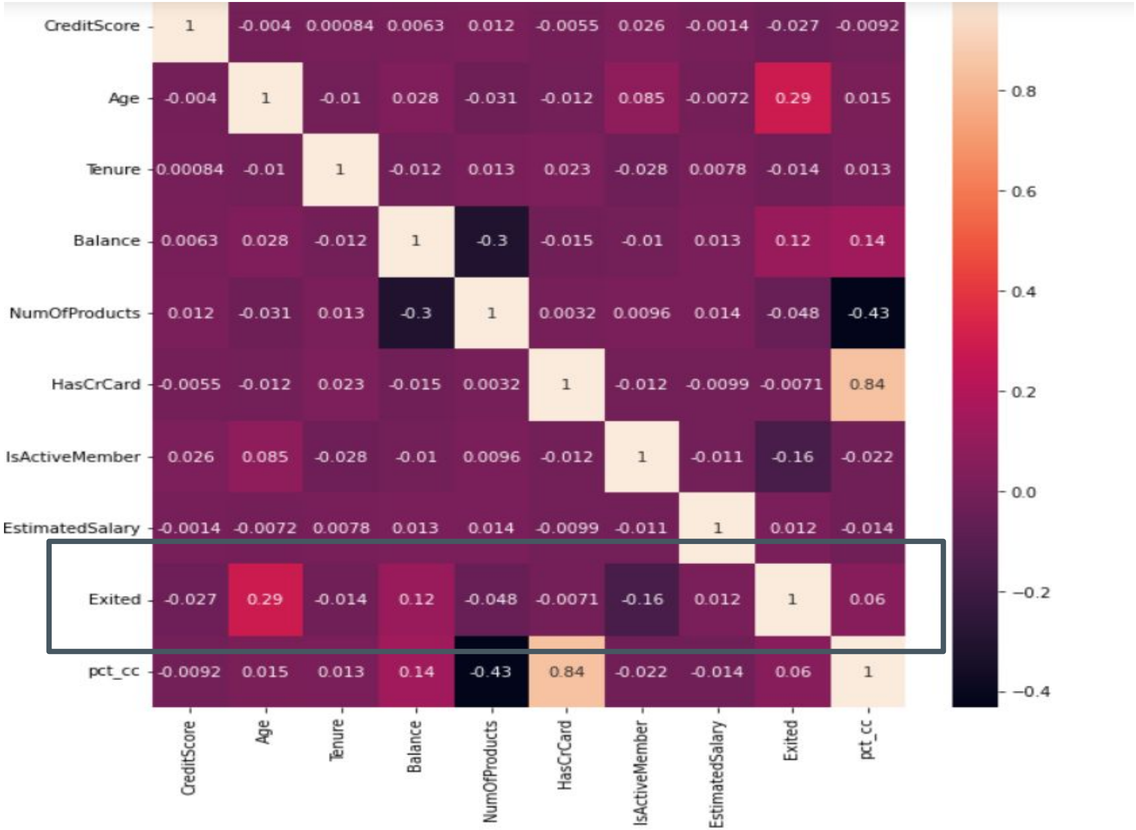
As expected, active members tend to stick around



It's interesting that the lowest churn is when the customer relationship is 7 years old

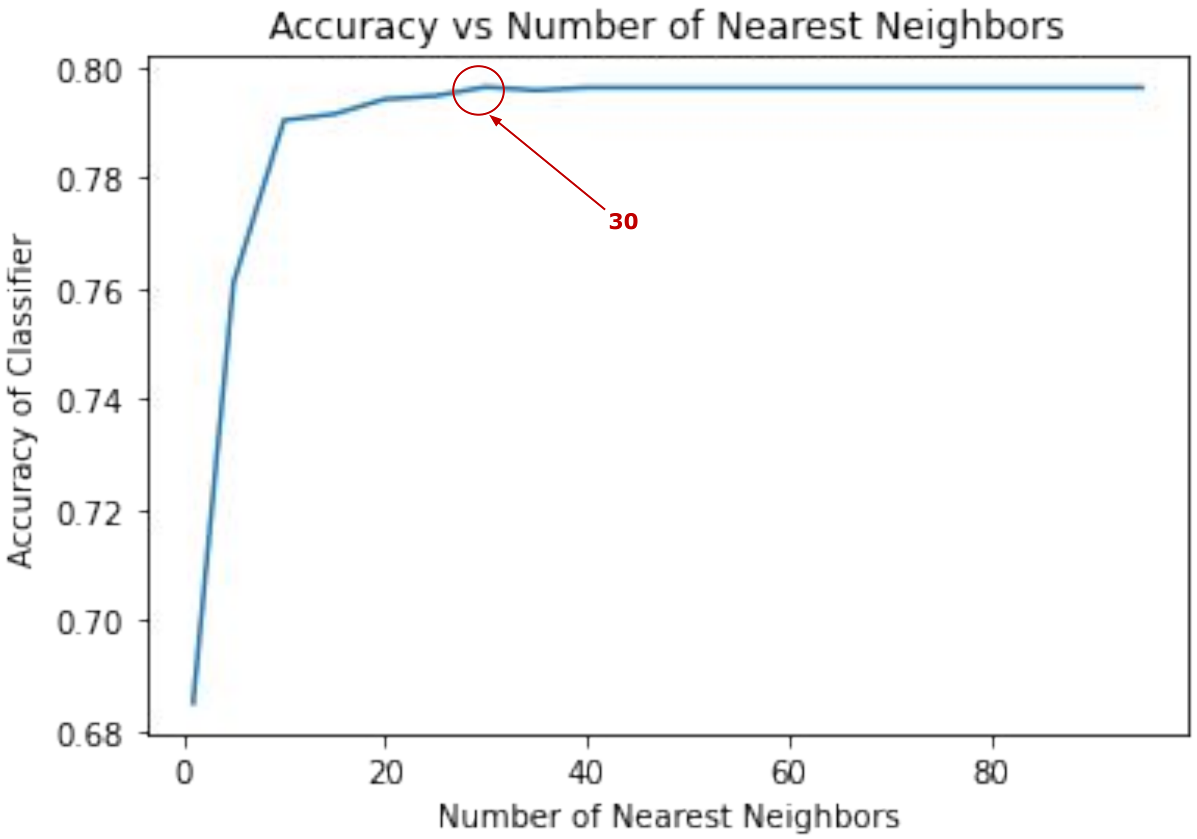
K-NEAREST NEIGHBORS

Picking the Best Features for KNN

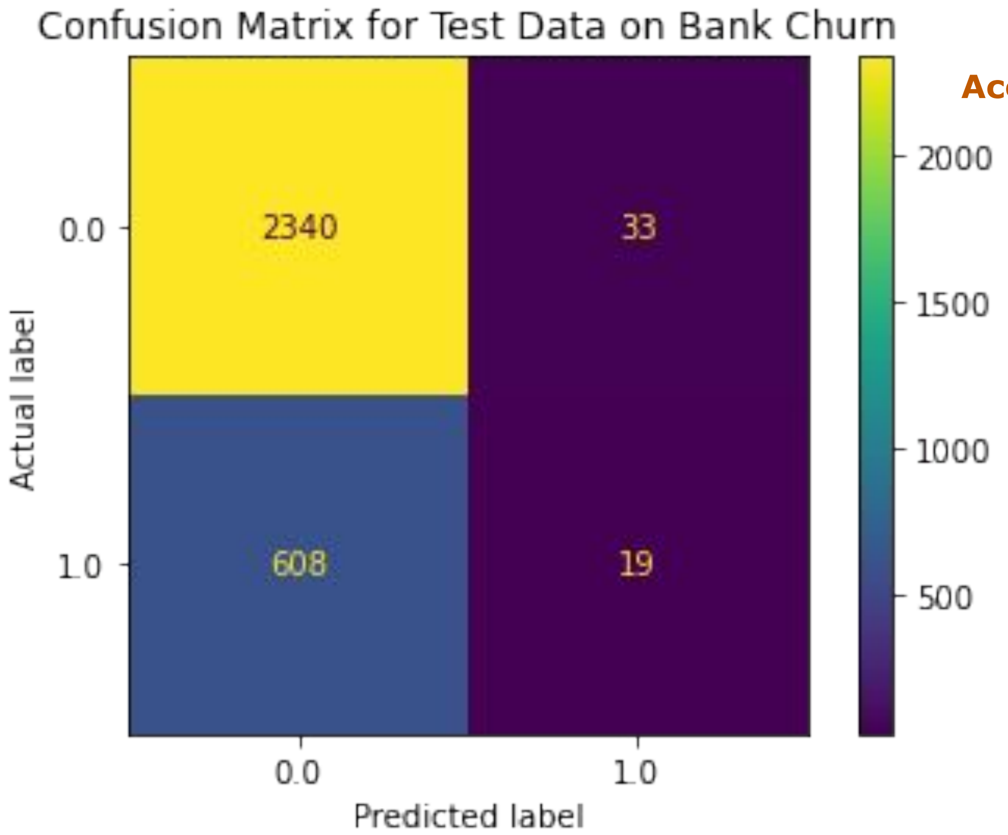


As shown earlier, the lowest correlations with Exited, the y variable, were 'HasCrCard', 'EstimatedSalary', and 'Tenure'. Dropping these improved the performance of the model.

Picking optimal K for KNN



KNN Confusion Matrix



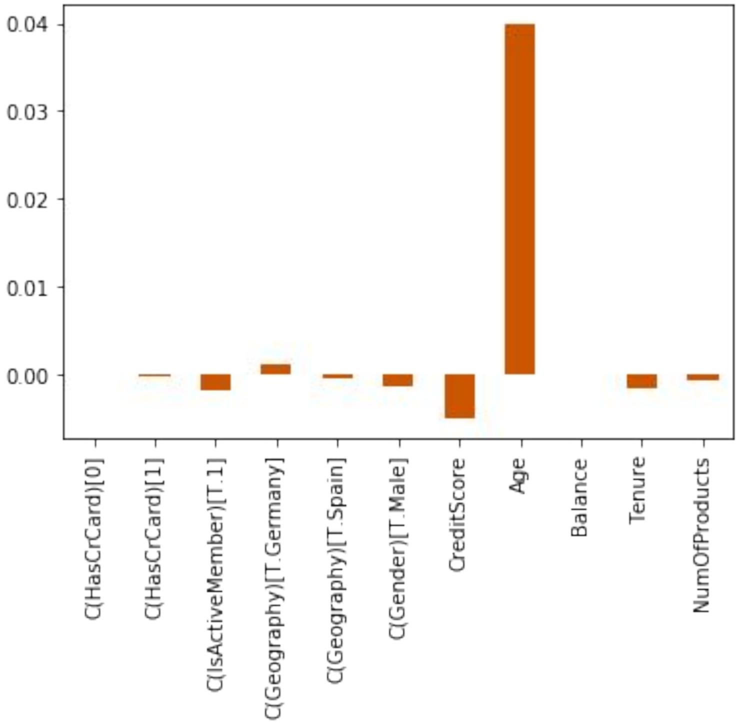
Accuracy: 0.804

LOGISTIC REGRESSION

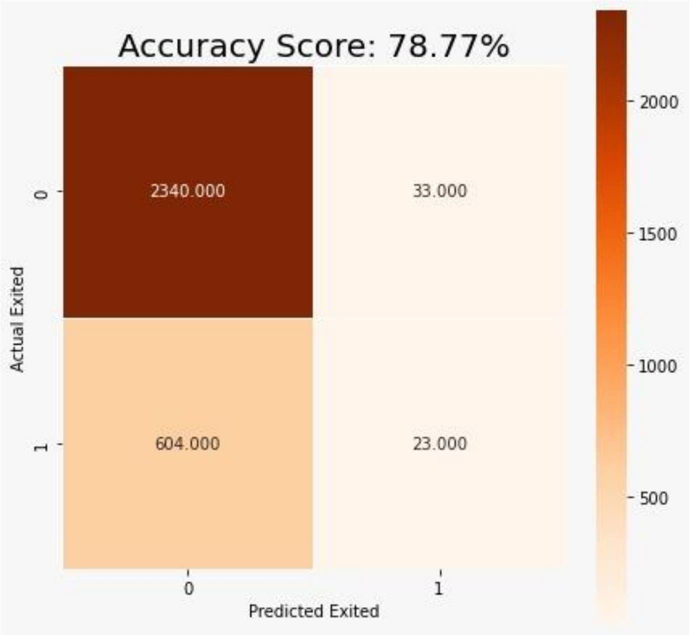
Logistic Regression

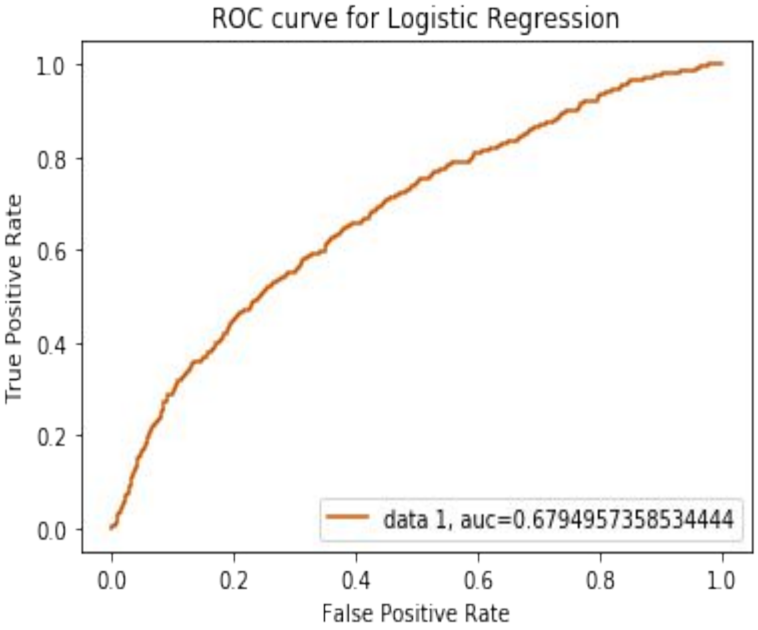
Simple Linear Logistic Model shows the variable importance as follows:

- Age is a very important Feature - Age[Middle Aged] seemed to affect the attrition the most (i.e) middle aged customers tend to exit more.
- Credit Score has a negative weight - Lesser the credit score, higher the probability of customer exit.
- Germany is the location with most exits
- Active members don't exit as much as inactive members.



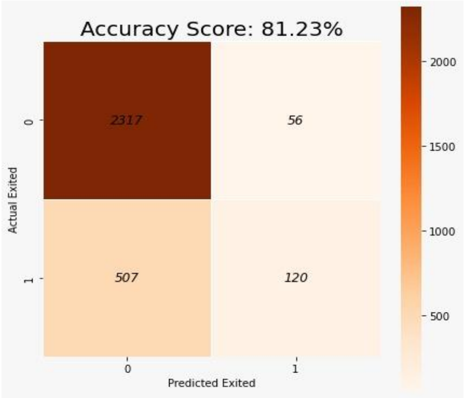
Logistic Regression

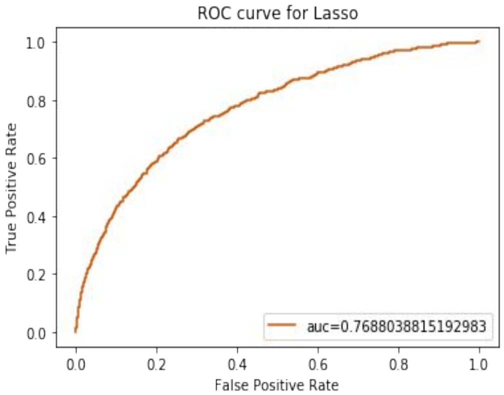




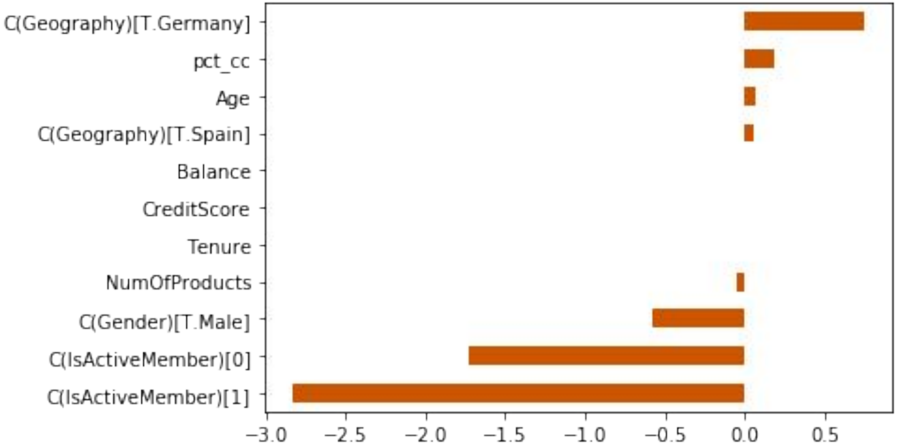
True Positive Rate/Sensitivity = 3.67%
False Positive Rate = 20.51%
Specificity = 98.6%

Logistic Regression - Lasso





- True Positive Rate/Sensitivity = 19.14%
- False Positive Rate = 2.35%
- Specificity = 97.64%
- Accuracy = 81.23%



- Proportion of Credit Cards/Total Number of Products Availed has an impact on the Churn
- Apart from Age and Location, Gender and IsActive Features seem to impact more in this model.

NAIVE BAYES

Naive Bayes Overview

Churn Data

Train	Test
70%	*30%*

37 features

	Train	Test
Accuracy	*82.4%*	*82.0%*

Key Predictors

	Positive class	Negative class	Positive_Negative_Ratio	Importance
Q("pct_cc_0.25")	0.002357	0.000020	118.677615	4.776411
Q("pct_cc_0.3333333333333333")	0.008800	0.000556	15.823682	2.761508
Q("pct_cc_0.5")	0.013593	0.042366	0.320851	1.136780
Q("Age_Bucket_Young Adult")	0.002671	0.007587	0.352098	1.043847
Q("Age_Bucket_Adult")	0.016657	0.044809	0.371744	0.989549
Q("Age_Bucket_Early Retirement")	0.000943	0.002185	0.431555	0.840360
Q("Geography_Germany")	0.044630	0.024073	1.853930	0.617308
Q("Age_Bucket_Old")	0.000079	0.000139	0.565131	0.570697
Q("Age_Bucket_Middle Age")	0.090831	0.056409	1.610227	0.476375
Q("IsActiveMember_1")	0.038894	0.061036	0.637221	0.450638

Top Predictors for Churn=1

	Positive class	Negative class	Positive_Negative_Ratio	Importance
Q("pct_cc_0.25")	0.002357	0.000020	118.677615	4.776411
Q("pct_cc_0.3333333333333333")	0.008800	0.000556	15.823682	2.761508
Q("Geography_Germany")	0.044630	0.024073	1.853930	0.617308
Q("Age_Bucket_Middle Age")	0.090831	0.056409	1.610227	0.476375
Q("pct_cc_1.0")	0.053115	0.035990	1.475829	0.389220
Q("IsActiveMember_0")	0.072052	0.050033	1.440087	0.364704
Q("Gender_Female")	0.063330	0.047808	1.324666	0.281161
Q("Balance_Bucket_Above Average")	0.078337	0.062745	1.248513	0.221953
Q("Tenure_10")	0.006364	0.005224	1.218363	0.197508

Owning 3-4 products but with only one credit card is a red flag

Top Predictors for Churn=0

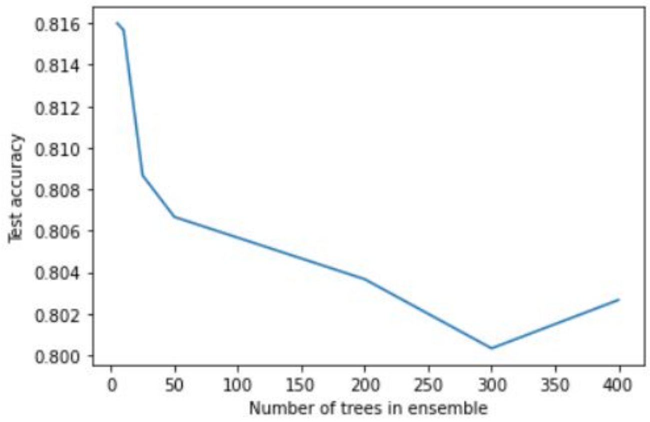
	Positive class	Negative class	Positive_Negative_Ratio	Importance
Q("pct_cc_0.5")	0.013593	0.042366	0.320851	1.136780
Q("Age_Bucket_Young Adult")	0.002671	0.007587	0.352098	1.043847
Q("Age_Bucket_Adult")	0.016657	0.044809	0.371744	0.989549
Q("Age_Bucket_Early Retirement")	0.000943	0.002185	0.431555	0.840360
Q("Age_Bucket_Old")	0.000079	0.000139	0.565131	0.570697
Q("IsActiveMember_1")	0.038894	0.061036	0.637221	0.450638
Q("Balance_Bucket_Below Average")	0.032608	0.048325	0.674767	0.393389

The more serious users (possibly students or young working professionals) are the ones with 2 products of which 1 is a credit card.

DECISION TREES AND ENSEMBLE METHODS

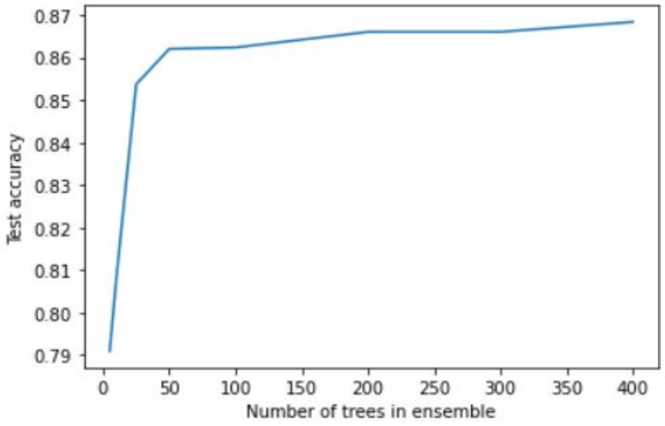
Variation of Testing Accuracy with Variation in Number of Trees (from 5 to 400)

Random Forest Classifier



10 trees seem to be enough for the Random Forest Classifier

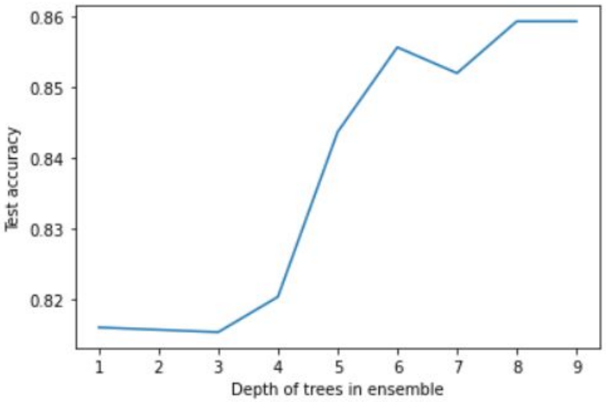
Gradient Boosting Classifier



50 trees seem to be enough for the Gradient Boosting Classifier

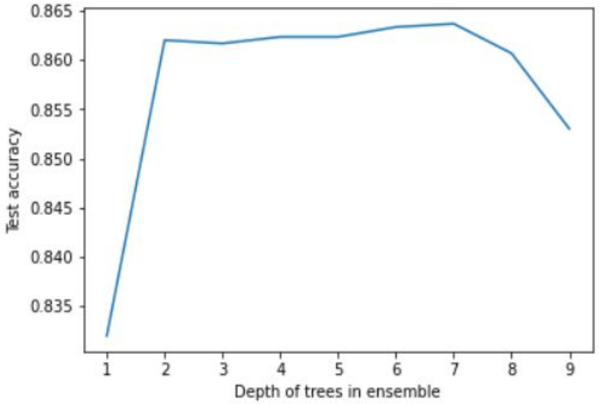
Variation of Testing Accuracy with Variation in Depth of Trees (from 1 to 10)

Random Forest Classifier



Trees of depth 8 seem to be enough for the Random Forest classifier

Gradient Boosting Classifier



Trees of depth 2 seem to be enough for the Gradient Boosting classifier

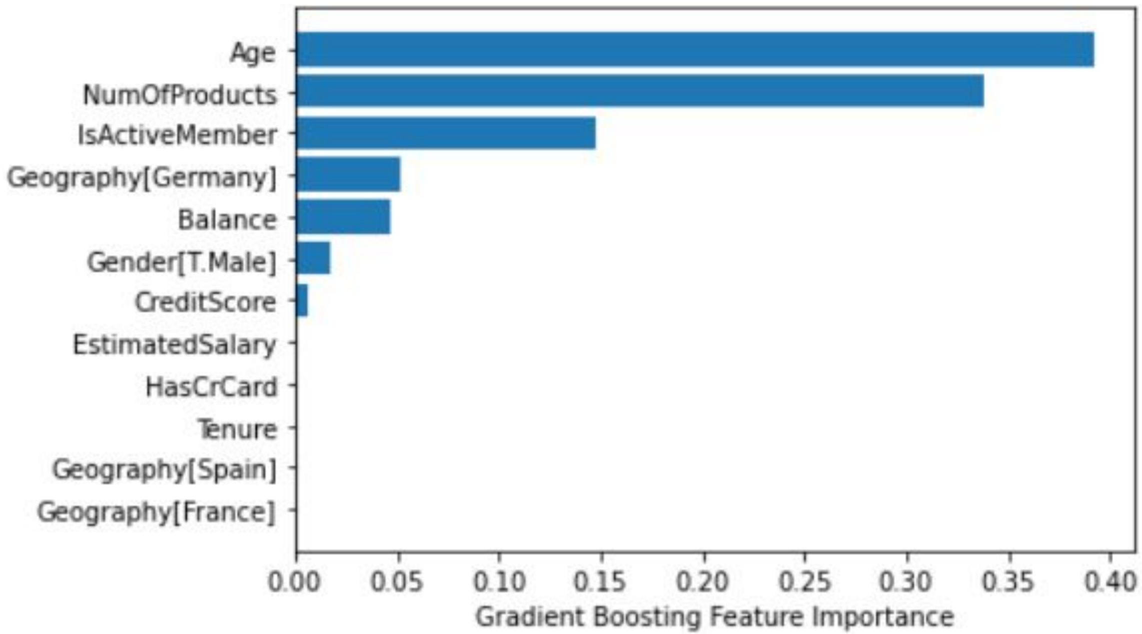
Optimal Parameters for Number and Depth of Trees and Training/Test Accuracy

Baseline Accuracy = 79.6%

Model	Number of Trees	Depth of Tree	Training Accuracy	Test Accuracy
Decision Tree	-	7	87.11%	85.90%
Bagging	10	8	88.47%	85.93%
Random Forest	10	2	87.61%	85.93%
Gradient Boosting	50	2	86.23%	86.20%

Gradient Boosting gives the best test accuracy across all models (**86.2%**)

Variable Importance (Gradient Boosting)



Age and ***NumOfProducts*** are observed to be the most important variables when it comes to feature importance. ***EstimatedSalary***, ***HasCrCard*** and ***Tenure*** do not play a major role in determining churned customers

Model Selection

Baseline Accuracy = 79.6%

Gradient Boosting fetches the best results for test accuracy

Model	Test Accuracy
KNN	80.40%
Logistic Regression	78.67%
Naive Bayes	82.00%
Decision Tree	85.90%
Bagging	85.93%
Random Forest	85.93%
Gradient Boosting	86.20%

INSIGHTS & RECOMMENDATIONS

Recommendations

The variables that are most meaningful across models are Number of Products and Age.

Consuming Model Results

Loyalty programs for middle aged users with over 3 products

Marketing Campaigns for moderate risk customers

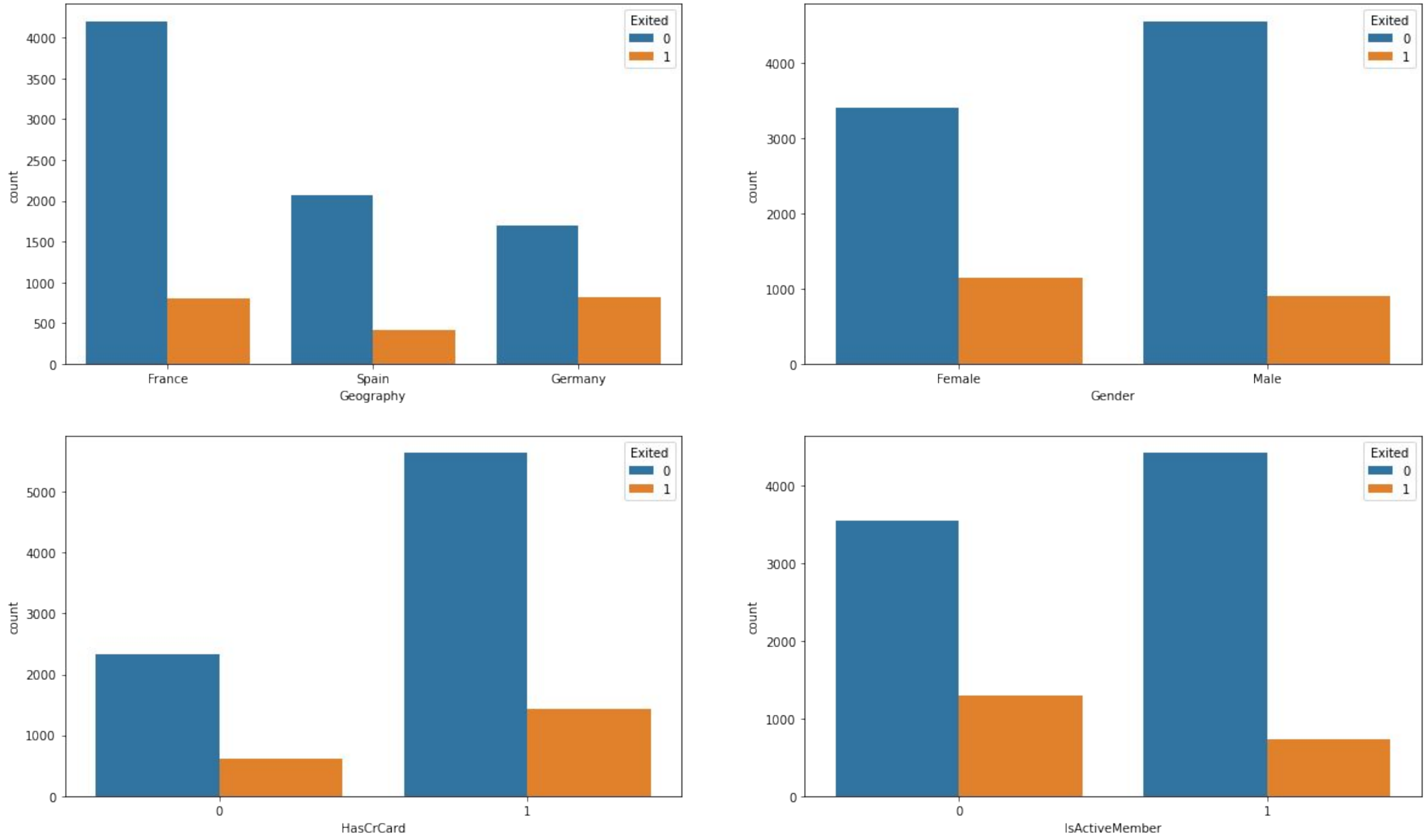
Caveats

- While models are good indicators of relative trends, hard to define causal relationships – A/B tests required
- The data size is specific to 1 bank and 3 regions – higher granularity (product level) + big data = greater generalisation

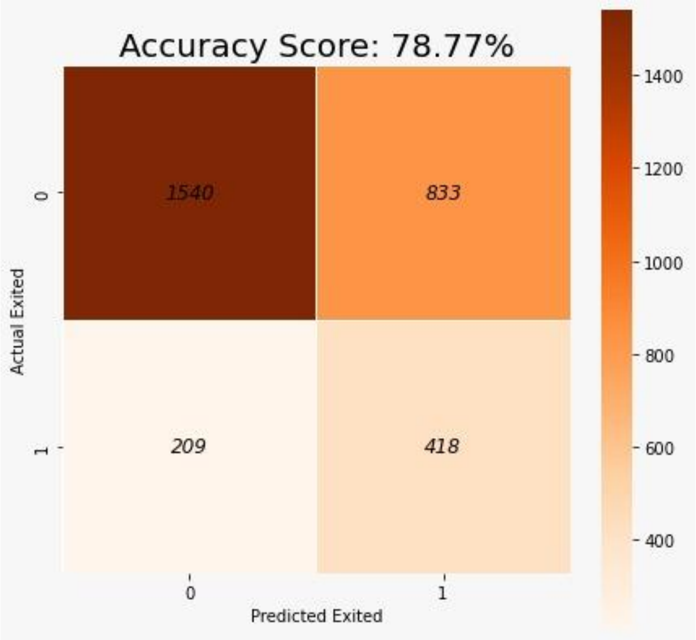
THANK YOU

APPENDIX

Categorical Variables



Logistic Regression with Threshold Optimization



- The optimal threshold achieved is 0.20. (Maximizing the difference between True Positive and False Positive Rate)
- If the probability given by the model > 0.2055 then we classify it as Exit. By reducing the threshold so low, we increase our TPR but our specificity decreases. We may lose precision ultimately.

True Positive Rate/Sensitivity = 66.67%
False Positive Rate = 35.10%
Specificity = 64.9%