Final Project: Supervised Learning

To find and predict the reasons behind customer churn for a multinational bank aiming to increase its market share in Europe

Introduction

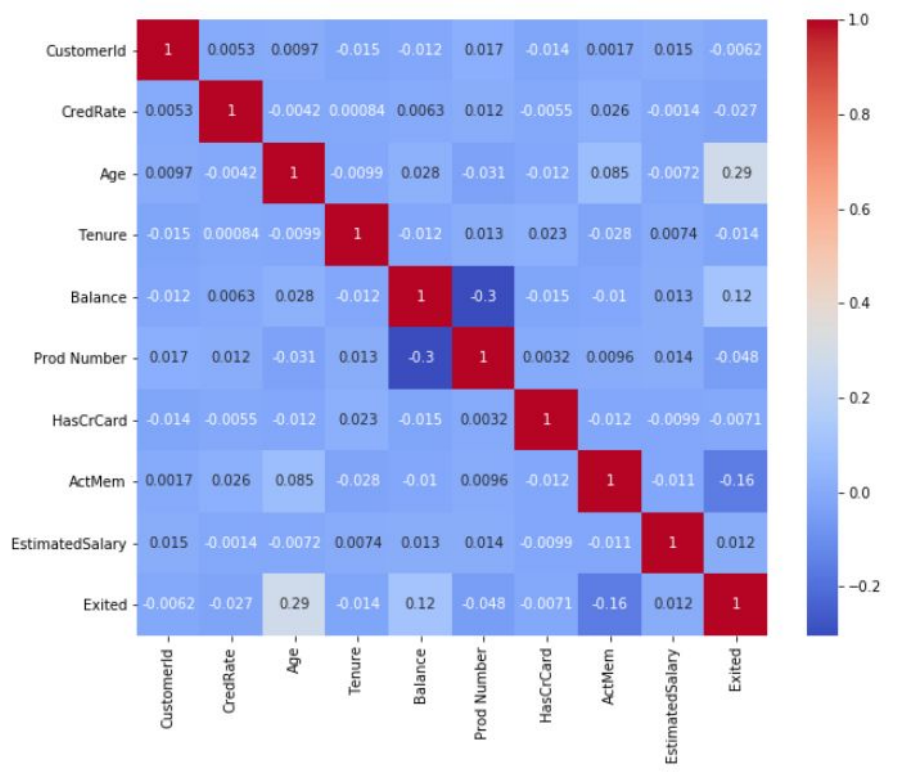
- To explore and identify the potential factors that influence/impact behind customer churn.
- Build a prediction model to help predict and classify the potential customer that may be churn.
- Analysis and provide insights to the business on how to tackle the problems and successfully increase its market share in Europe

Data understanding / Preprocessing

Key elements:

- Inspect
- Exploring the data
- Experiments of the data
- Handling missing value
- Understand the relationship of dataset

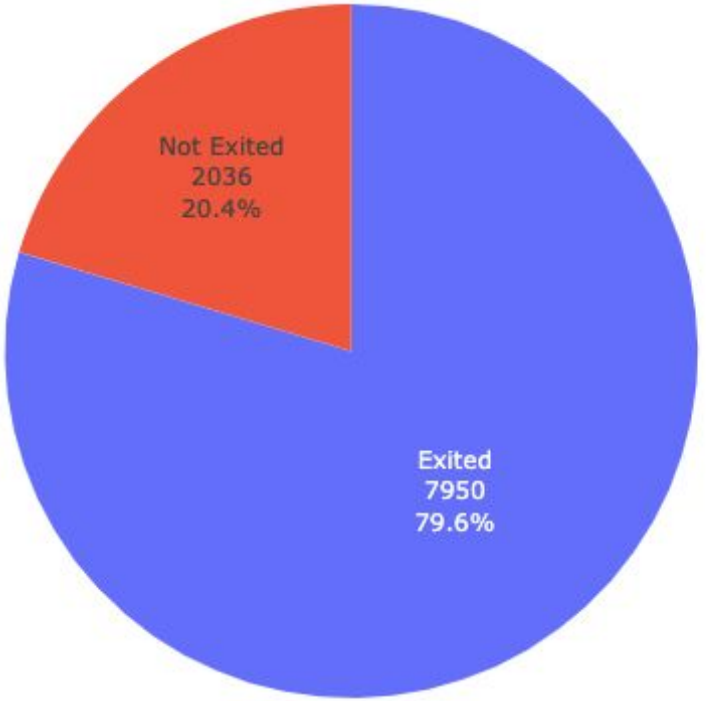
Correlation of Dataset Features



From the heatmap, surprisingly the dataset features aren't that correlated. Only "Age" has a higher number of correlation with "Exited" feature among all.

Exploratory Data Analysis

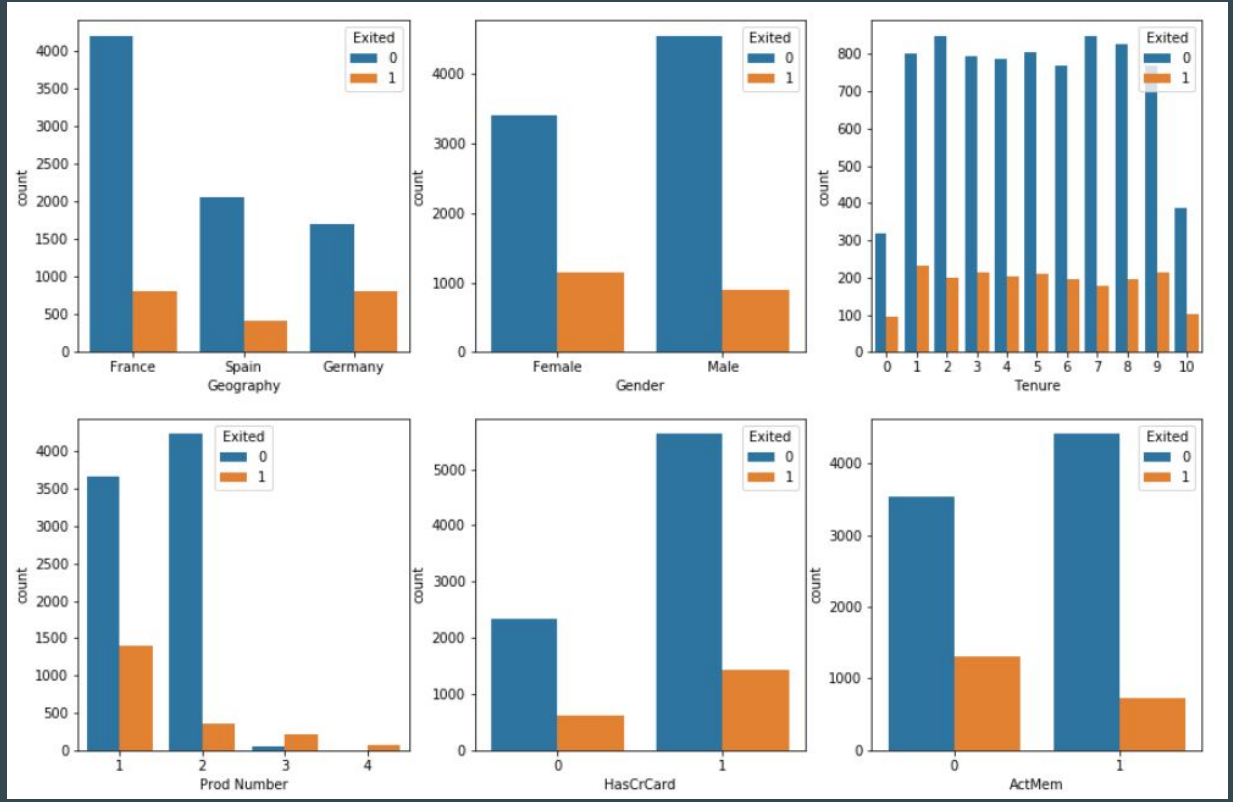
From the Pie Charts, a quick overview of the percentage of "Exited" group is high and almost close to 80% of the total number.



Features Plotting

From these 6 Features countplots, here are my hypothesis on their correlation that in relates to the customer churn.

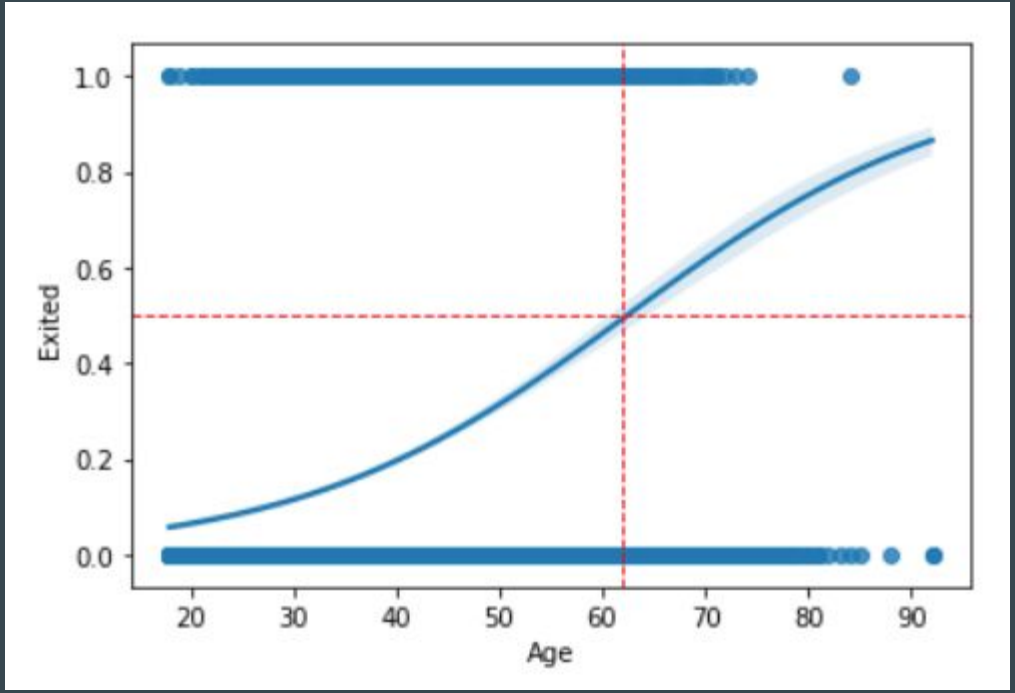
- Geography
  - Germany is smaller differences compare to France
- Gender
  - Male has the higher not exited rate and lower exited rate compared to Female
- ActMem
  - The non-Active Member exited is much higher than the Active Member



‘Age’ versus ‘Exited’

Countplot was hard to present the correlation of Age vs Exited feature so decided to use the regplot to study the data.

From this plot, it shows that when age is higher the likelihood to churn/exit. At the threshold of 0.5, customer with age of 62 and above are likely to churn



Machine Learning Model Training

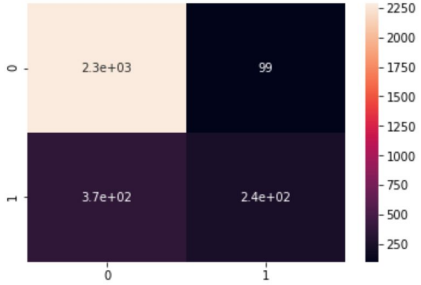
After Data Preprocessing: Encoding, Scaling and Features Engineering

Choice of Algorithms

- Logistic Regression
- ~~XG Boost~~
- Support Vector Machine
- Random Forest

Logistic Regression

Commonly use to predict the target of categorical or binary and in which case our dataset and the outcome trying to achieve suggest that this algorithm is one of the simplest to begin with.



	precision	recall	f1-score	support
0	0.86	0.96	0.91	2384
1	0.71	0.39	0.50	612
accuracy			0.84	2996
macro avg	0.78	0.68	0.71	2996
weighted avg	0.83	0.84	0.82	2996

Accuracy Score : 84.27903871829105

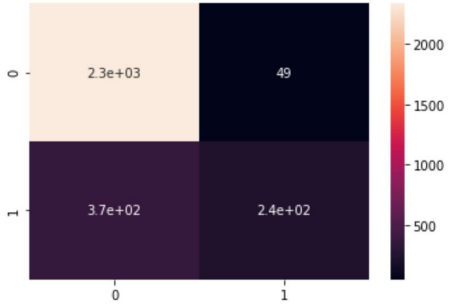
AUC Score : 67.5315008553757

*Precision and Recall will be good enough if the prediction is to focus on only True Positive.

Whereas Accuracy focus on true results, good for balanced data. AUC good for balanced measuring of positive and negative classes.*

SVM “Support Vector Machine”

This method objective is to find a hyperplane which maximized the separation between data of different classes. Suitable for what our prediction needs - (eg. Exited or Not Exited, 2 classes)



	precision	recall	f1-score	support
0	0.86	0.98	0.92	2384
1	0.83	0.39	0.53	612
accuracy			0.86	2996
macro avg	0.85	0.69	0.73	2996
weighted avg	0.86	0.86	0.84	2996

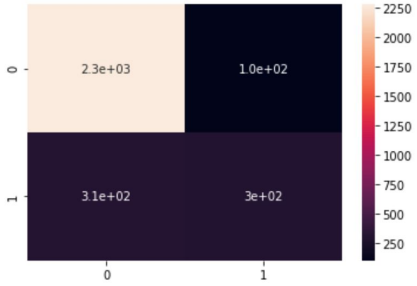
Accuracy Score : 85.94793057409879

AUC Score : 68.5801585734965

SVM evaluation results looks similar but still better than Logistic Regression at the accuracy and auc scores.. Let's look at my 3rd choice of the algorithm - Random Forest and it's prediction.

Random Forest

Random Forest gives a wide diversity that generally results in a better model. It adds additional randomness to the model, rather than searching for the most important feature it searches for the best feature among a random subset of features.



	precision	recall	f1-score	support
0	0.88	0.96	0.92	2384
1	0.74	0.49	0.59	612
accuracy			0.86	2996
macro avg	0.81	0.72	0.75	2996
weighted avg	0.85	0.86	0.85	2996

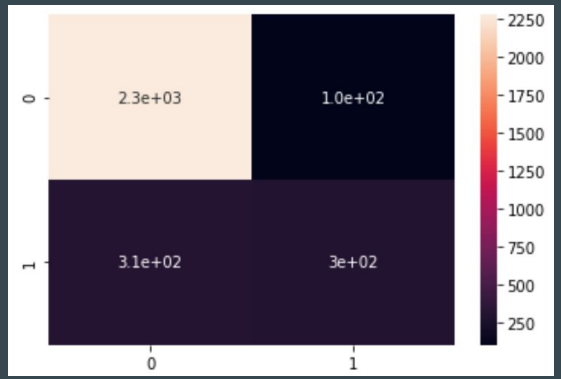
Accuracy Score : 86.11481975967958

AUC Score : 72.38932205992018

Out of 3 algorithms Random Forest did the best out of them. So I want to spent a bit more focus on see how can I make improvement to the prediction.

Parameter Tuning - Cross Validation, GridSearch CV

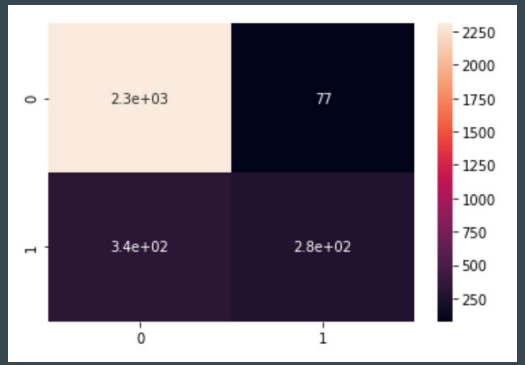
GridSearch CV gathers all the possible combinations of parameter values are evaluated and retained the best combination



Accuracy Score: 86.11481975967958

AUC Score : 72.38932205992018

Before tuning



Accuracy Score: 86.2483311081442

AUC Score : 71.01578606834232

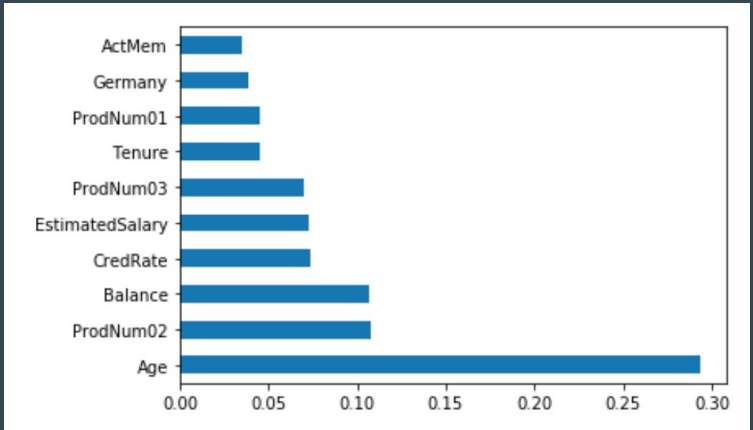
After tuning

Despite the improvement is not much but the model prediction definitely improved compare to the one without the tuning.

Feature Importance

From the right side visualizations, it helps us to understand what are the main factors of customer churn based on the dataset features.

We can see and suggest that Age, Customers have 2 Products and Balance are the top 3 most crucial factors of customer churn where their importance rates are pretty high compared to other features.



	importance
Age	0.293355
ProdNum02	0.107315
Balance	0.106610
CredRate	0.073605
EstimatedSalary	0.073015
ProdNum03	0.070101
Tenure	0.045373
ProdNum01	0.045236
Germany	0.038677
ActMem	0.034913
NotActMem	0.032519
ProdNum04	0.020389
France	0.012281
Female	0.011191
Male	0.011137
Spain	0.009153
NoCrCard	0.007669
HasCrCard	0.007460

Conclusion

- Random Forest Model is good prediction model for tackle this problem matter. But it can be improve better with some boosting methods.
- If given more times, first thing will want to try on will be different ensemble methods and boosting algorithms to refine and improve the models. Also other different ML algorithms as those I chosen here consider the basic algorithms
- Biggest takeaways
  - First ever ML model from scratch.
  - Research and reference in Kaggle how others build their model and create their notebook for better understanding.