Bank Customer Churn Prediction

By Karen Amanya

Beginning

1.0 **Overview**

What is customer churn?

Customer churn is the rate at which customers stop using a business's product or stop paying for their services. A high customer churn rate indicates that a company is losing a lot of customers and can be directly linked to the product and service in question.

It is important for any business to track the number of customers they are losing and gaining as this is crucial to the revenue and growth of the business. Businesses should focus on this metric and find out why customers are leaving and how to prevent this from happening.

1.1 Business and Data Understanding

The dataset used is obtained from a European bank that is operational in three countries ; Spain, France and Germany and contains data on its customers.

Customer churn has become a major problem among financial institutions given that their revenue stream is directly related to the amount of customers that subscribe and pay for their services and products therefore earning them revenue. A bank with a high churn rate loses many subscribers which results in lower growth rates and this has an even bigger impact on sales and profits while an institution having a low churn rate is an indication that it can retain its customers.

A major strategy of many financial institutions while trying to maintain their growth trajectory is to acquire new customers. While this is an effective approach, it is not as important as retaining the customers that the business already has and this is because acquiring a new customer costs way more than retaining an existing one.

It is important for a business to be able to identify customers at risk of churning and come up with strategies that will maximize the likelihood of the customer staying.This is quite difficult especially for a large banking institution with many different types of customers, to know why a customer is cancelling their subscription to its services because of their different behaviours and preferences.

In the current digital environment where it is very easy for a customer to switch from one banking institution to another , customer churn prediction will allow a bank to identify when a customer is about to leave and act proactively to reverse a potential customer churn and mitigate revenue losses.

This project aims to help a bank or similar banking institutions with :

1. Customer retention

2. Increase their profits

3. Improve customer experience

4. Optimization of their products and services.

5. Gain insights on why they are losing customers

Modeling

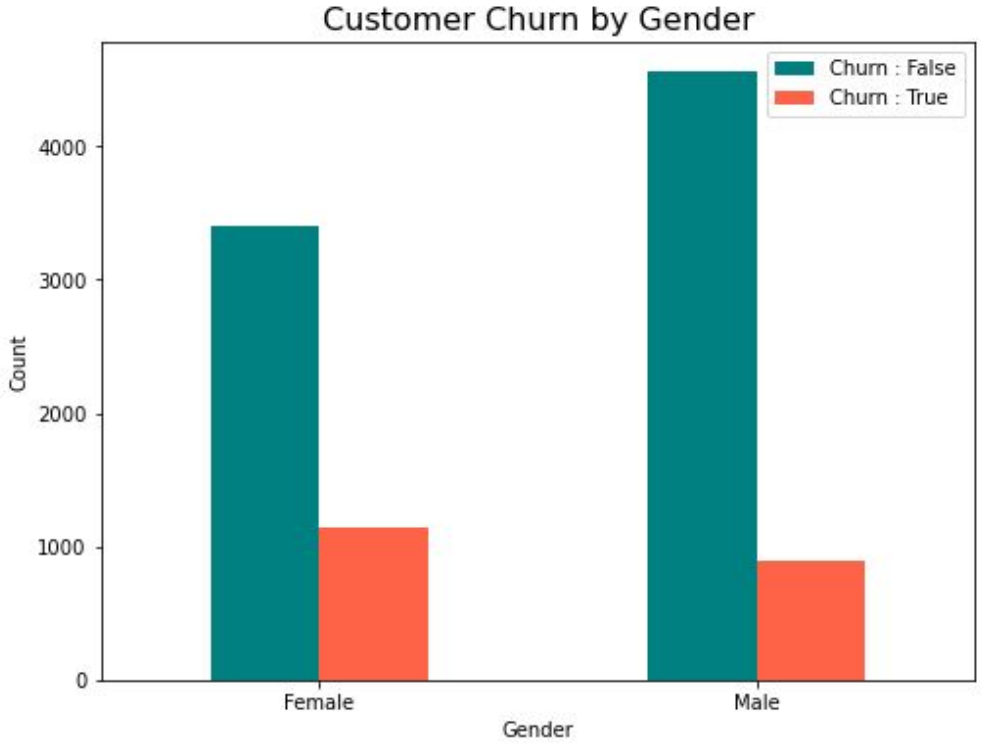
This project employs 3 classification algorithms , with each being an attempt to improve on the previous model's performance.

The first step was to evaluate and clean the data with the goals being:

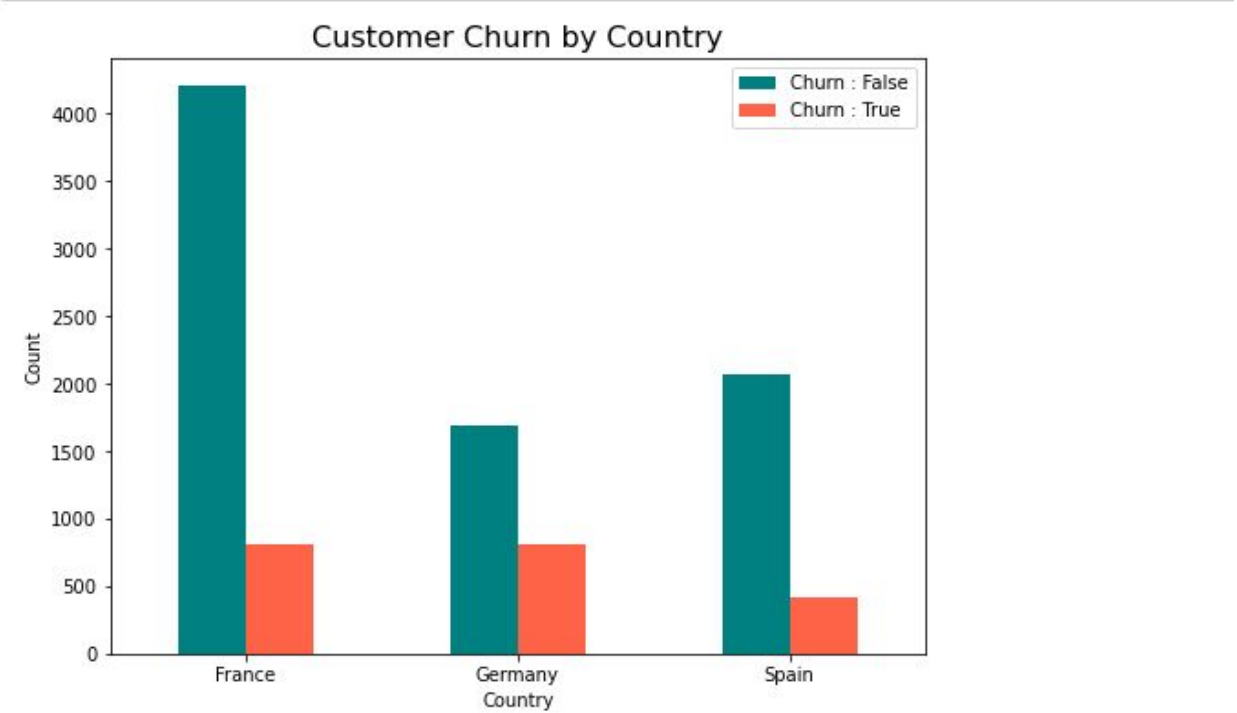
1. To find out the size of the data - number of rows and columns.
2. Check whether there were any missing data or duplicated rows - there were none.
3. To gain basic insights on the composition of the dataset - it contains a mixture of categorical and numeric variables with most being numerical.
4. Check whether the columns were in their correct data types - numbers stored as numbers e.t.c

The second step was to perform some basic exploratory data analysis to gain more insight on the data. First I looked into the categorical variables to learn the composition of the dataset based on these features:

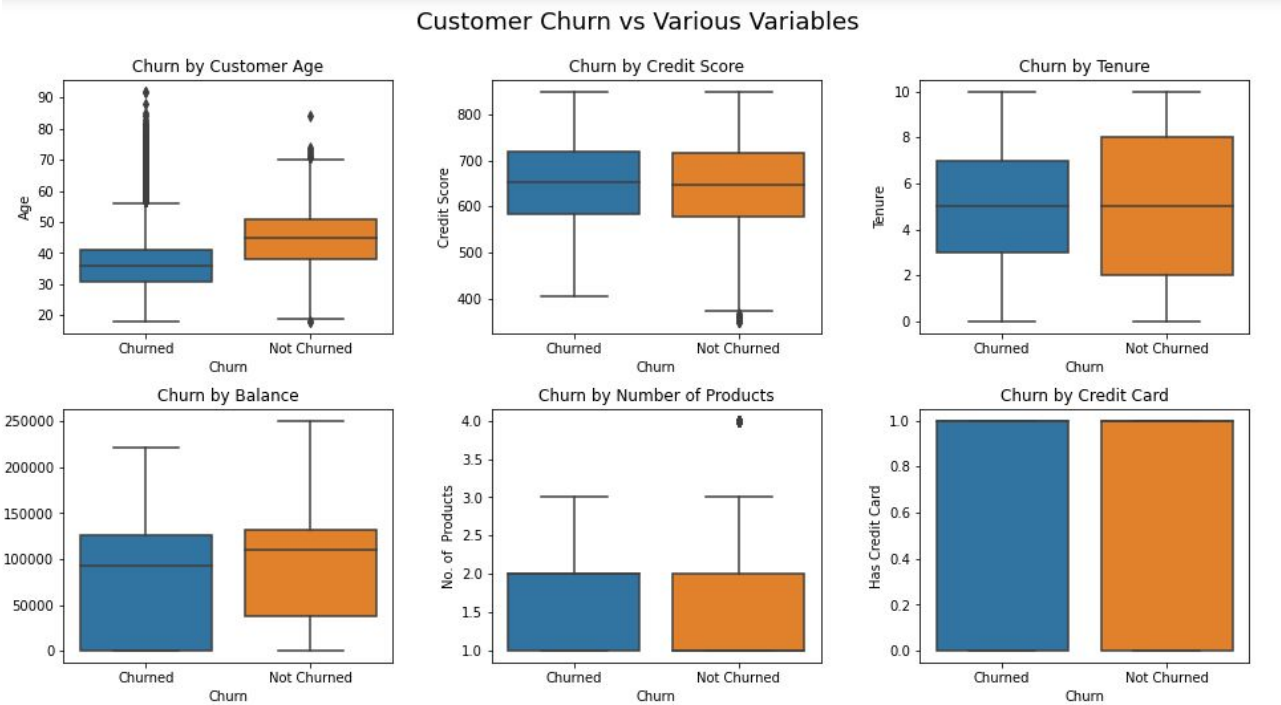
1. Gender - there were more observations of male customers than female and Female customers were more likely to churn compared to male customers.



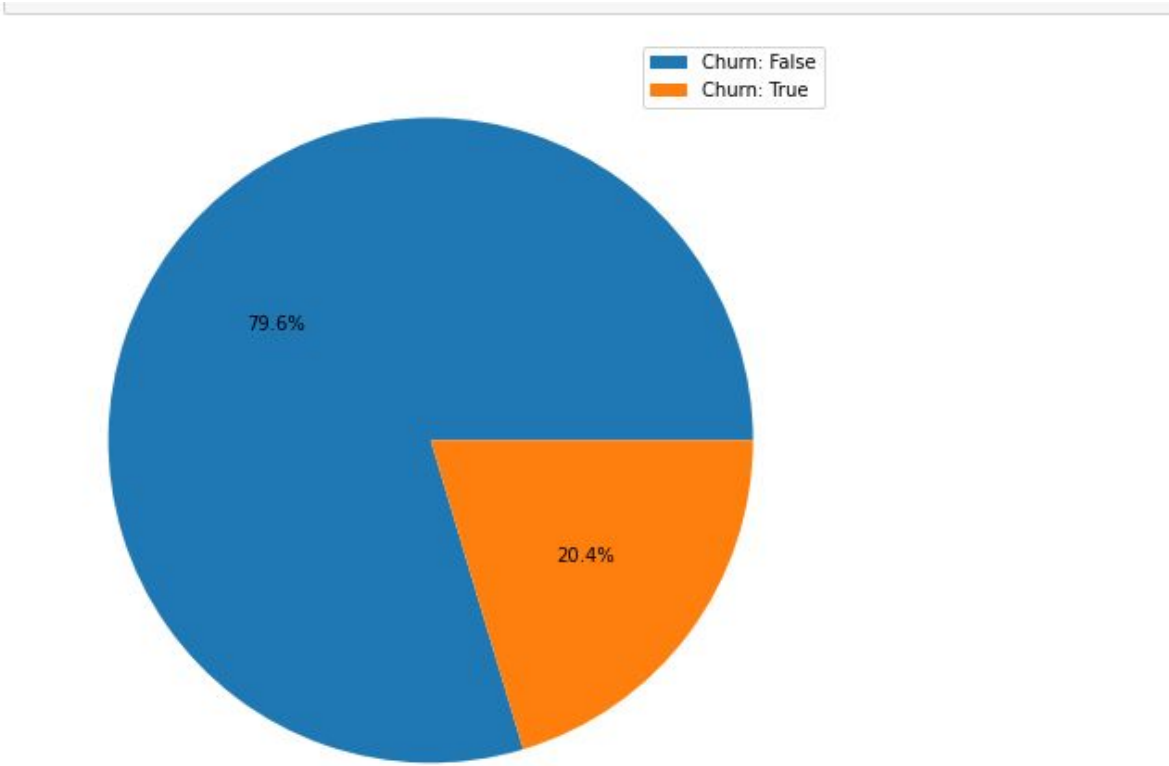
2. Country - most of the bank’s customers were from France with France and Germany having the highest churn rate.



A Visualization of Churn against the Numeric Features



A Visualization of the Target Variable



Fitting the Models

The next step was to build classification models with the goal being to achieve a well performing model that will be able to predict whether a customer will churn.

Model 1 : Logistic Regression

The first model was a logistic regression model. It employs very simple techniques in modeling and is mainly used as a baseline to help better process the data for future models to improve performance.

After building the model , predictions are made on both the training and validation data and the scores of the various metrics are recorded. This model did not perform very well and had an F1 score of 47%.

Model 2 : Random Forest Classifier

The second model was built using more complex techniques. It employes scaling to tone down the variance in ranges in the numerical features such as age, balance and estimated salary. It also employs encoding to create binary categories of the categorical variables to enhance model learning.

This model employs use of GridSearchCv which is technique used to find the optimal parameters that perform best with the model.To cater for the class imbalance problem observed in the target variable the model uses SMOTE which is short for ‘synthetic minority oversampling technique’ which is a common oversampling method that aims to balance the class distribution by randomly increasing minority class examples by replicating them.

Predictions were then made on the validation data and scores were computed. The f1 score for this model was 60% on the churned class and 90 % on the Unchurned class which is still quite low . This was a good improvement to the base model but the performance was till low on one class.

Final model - XG Boost

The final model used an XGBoost classifier but before the model was fit, more transformations were done on the data. A new column containing the credit rating was created using the customers credit rating as per the system below:

300 - 400 : Bad

630 - 689 : Fair

690 - 719 : Good

720 - 850 : Excellent

Another column indicating True or False based on whether the customer’s balance was 0.

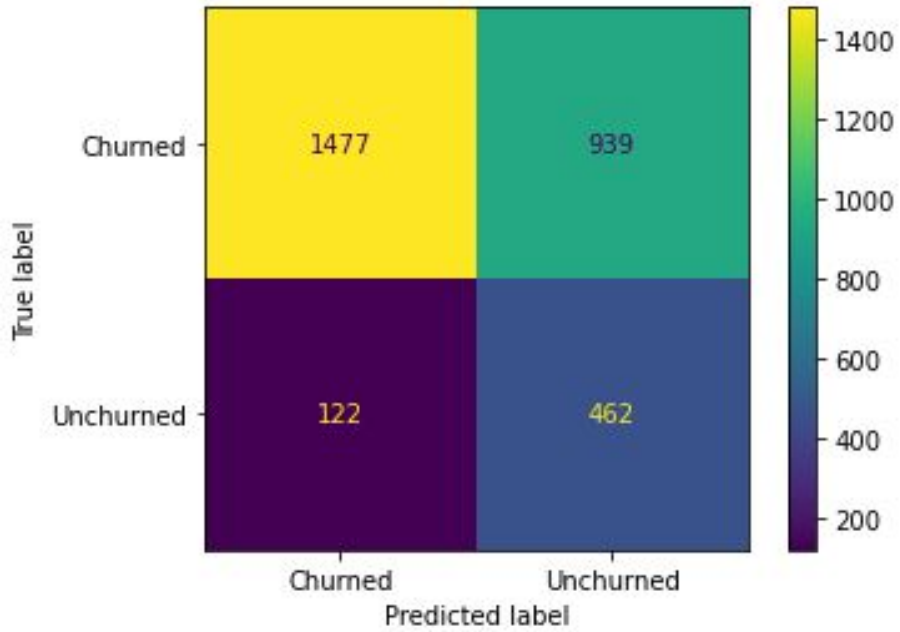
This model also included more hyperparameter tuning to boost model performance.

Evaluation

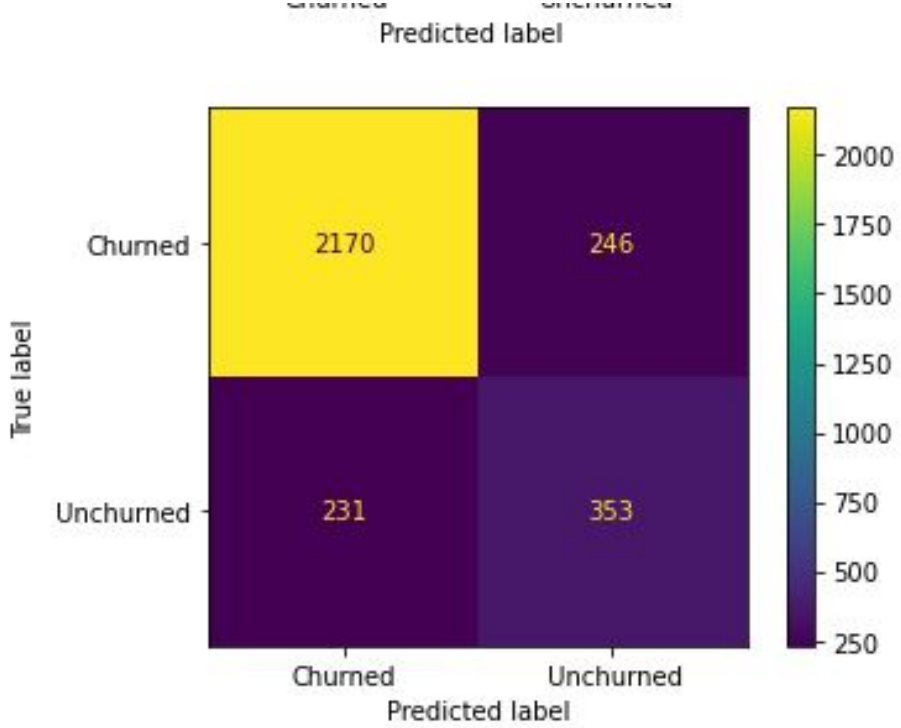
The models scores were as below:

	Model	Accuracy	Precision	Recall	F1 SCore
0	LogisticRegression(C=1000000000000.0, fit_inte...	65.0	33.0	79.0	47.0
0	GridSearchCV(cv=3,\n estimator=Pip...	84.0	59.0	60.0	60.0
0	GridSearchCV(estimator=Pipeline(steps=[('colum...	85.0	64.0	56.0	60.0

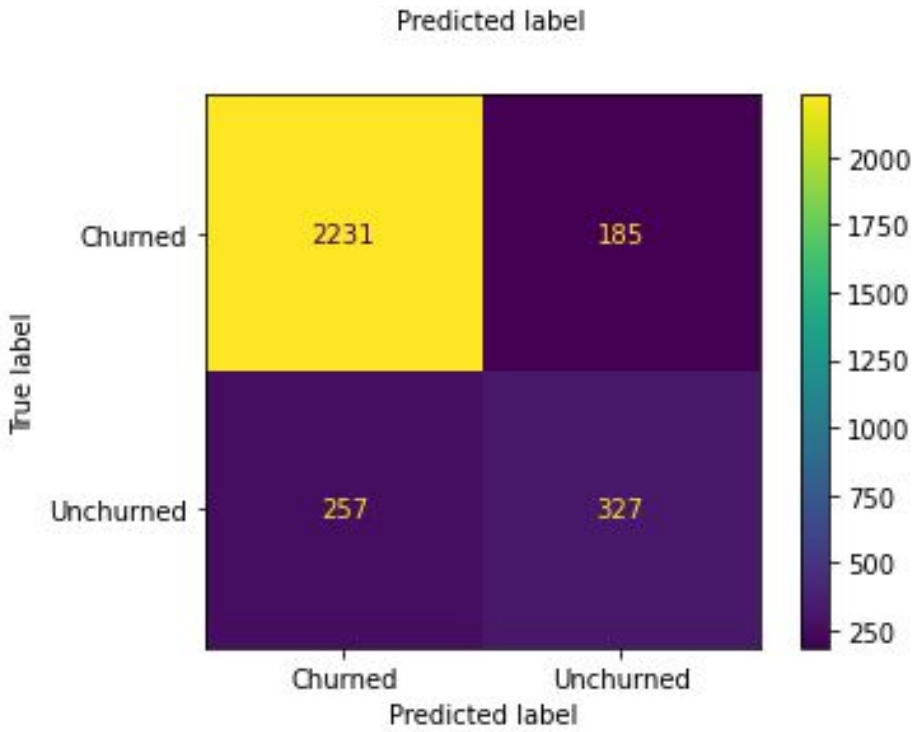
Classification Matrix of the Linear Regression Model



Classification Matrix of the Random Forest Classifier



Classification Matrix of the XGBoost Classifier



1. All three models performed better in predicting the Unchurned class than they did in predicting the Churned.
2. This is attributable to the heavily imbalanced dataset, even after employing SMOTE to cater for the imbalance, the variance in the class predictions was still significant.
3. The best performing model was the XGBoost which had the highest accuracy score and tied with the Random Forest classifier on the f1 score.
4. All three models were able to predict the unchurned class well but failed to achieve good scores in predicting the Unchurned class
5. The most important features to the models prediction on whether or not a customer is churned were balance and the number of products they were subscribed to.

Recommendations

1. ***Identify customers with a high risk of churn***

A bank or financial institution could use this model to identify customers at risk of churning . The models indicated that the most significant features were age and balance. Such institutions should actively track customer account balances and develop a minimum threshold that will act as an alarm to act on it. In addition the institution could come up with customer segments based on age and employ different tactics to better serve them or develop products suitable for the different ages.

2. ***Employ Customer engagement tactics***

The business , after identifying the customers at risk of churning should employ different tactics to improve the customer's experience . Even if some leave, they could still be a source of valuable insight on customer engagement with their products and services. The institution could then use these insights to better tailor their products for future and current customers.

3. ***Come up with segments based on age***

The project showed a positive correlation of age with churn. Age was also one of the important features of the model's performance. The bank could use this as a foundation for developing products and services. Having customer groups will help to better serve each customer based on their preferences and needs. Customer satisfaction is key to customer retention as happy customers don't leave.

Next Steps

The next step for this project is a single case study of a business particularly one in the financial services industry. The approach would be to first gather data and resample to ensure a balance in the target class. The model's performances showed just how important it is to have an even;y balanced dataset.

Thank You,

Karen.