Supervised Learning: Classification - Final Project

Benjamin Pieczynski

September 2023

1 Main Objective

The Bank of Spaframany aims to address a recent decline in customer retention rates in Spain, France, and Germany. To achieve this, our objective is twofold: firstly, to identify the key variables contributing to customer churn, and secondly, to construct a predictive model capable of forecasting customer churn.

2 Data

Source: Churn Modelling Dataset on Kaggle by Shruti_Iyyer

Link: `https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling`

Dataset Origin: The dataset represents customer information and churn data from a bank.

Dataset Characteristics:

Size: The dataset contains a total of 10000 rows and 14 columns.

Columns: The dataset comprises 13 usable columns, these columns are:

- CustomerId: Unique identifier for each customer.
- Surname: Customer's last name.
- CreditScore: Customer's credit score.
- Geography: Customer's geographical location.

- Gender: Customer's gender.
- Age: Customer's age.
- Tenure: Number of years the customer has been with the bank.
- Balance: Customer's account balance (Euros).
- NumOfProducts: Number of bank products the customer has.
- HasCrCard: Whether the customer has a credit card (1 for yes, 0 for no).
- IsActiveMember: Whether the customer is an active bank member (1 for yes, 0 for no).
- Estimated Salary: Estimated annual salary of the customer (Euros).
- Exited: The target variable indicating whether the customer has churned (1 for churned, 0 for not churned).

Given this information we want to determine what variables are primarily responsible for whether or not a customer churns. Then we will build a machine learning model to classify whether or not a customer will churn.

3 Exploratory Data Analysis

First, I loaded the dataset '.csv' file into python. Then, I cleaned the data set by checking and removing duplicate customers and missing values (NaN). There were no duplicate or NaN values within the entire dataset. The CustomerId, Surname, and RowNumbers were removed from the dataset as these are unlikely to have any causal relationship on whether or not the customer churns. If any of these would have an impact there is insignificant data within the dataset to explain their causal relationship.

Second, I checked the data types for the remaining features. Our remaining features were floats and integers with the exception of the Geography and Gender object data types. For Geography, we have three unique locations ['France', 'Spain', 'Germany']. For Gender, we have ['Female', 'Male'] columns. Here we can choose between One-Hot encoding and Label encoding methods. For this purpose I will adapt One-Hot encoding universally due to it having less conflict with different classifier models.

Third, I considered feature scaling and binning techniques for each feature. Here, I looked at box plots, violin plots, kernel density estimation plots (KDE), and distributions of each non-integer feature. The Age feature plots indicated significant skewness (Figure 1) upon further analysis. The decision was made to log the distribution of ages. The KDE plot of ages suggests correlation between age and customer churn (Figure 2).

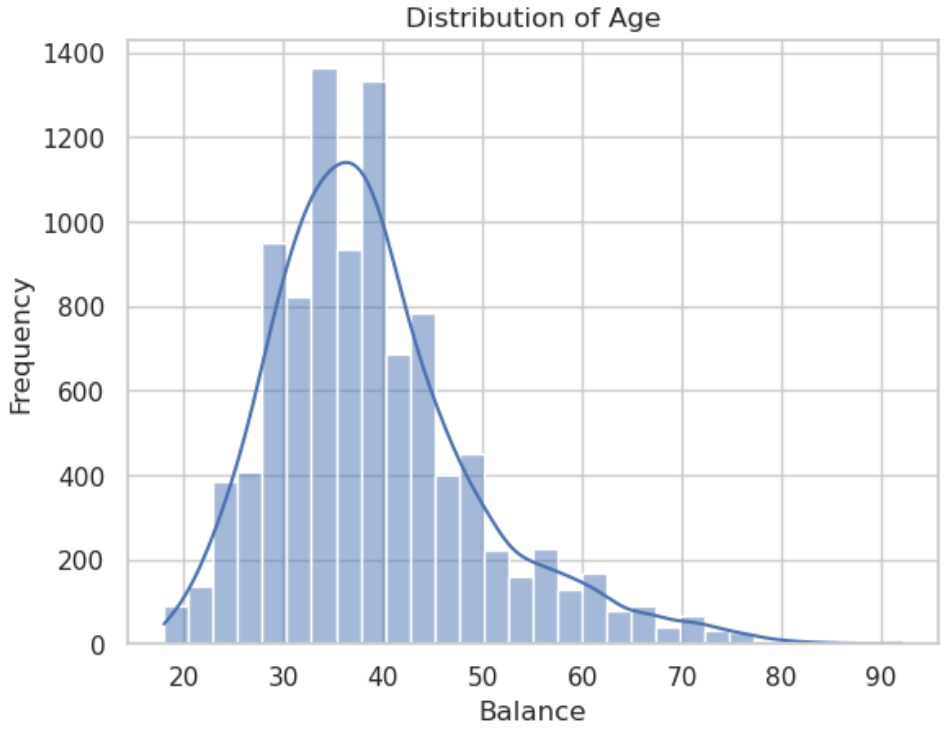

Figure 1: Distribution of customer ages prior to the log transformation.

Another noticeable trend included a correlation of customers from Germany churning more than that of Spain and France (Figure 3). Other noticeable trends included gender, account balance, and the number of products purchased. Female customers were more likely to churn and customers who purchased fewer than two products churned more than customers who purchased more products (Figure 5). Customers with a balance zero or lower churned more than customers with a higher balance (Figure 4). A list of the correlation values for each feature can be found in Figure 6.

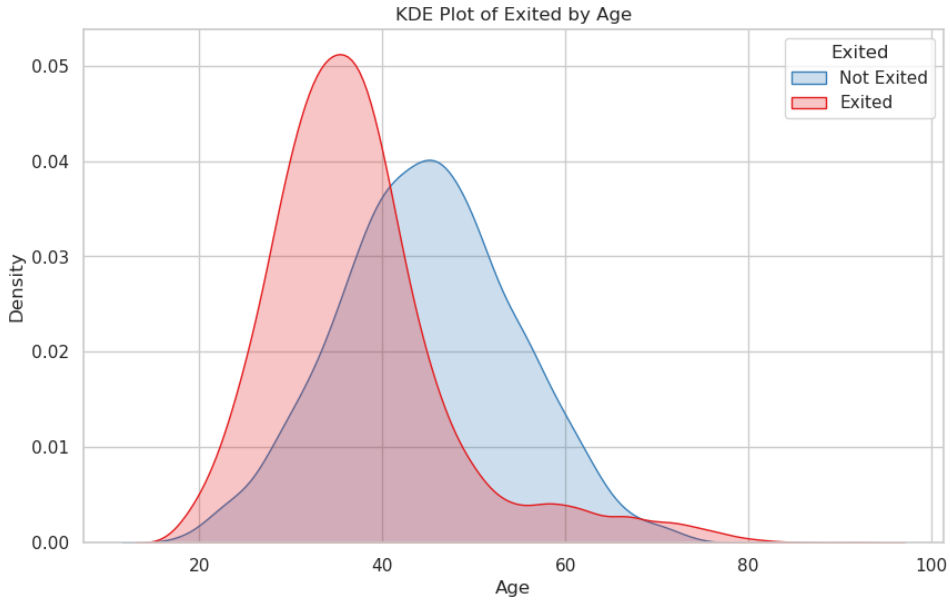

Figure 2: KDE plot of Age feature.

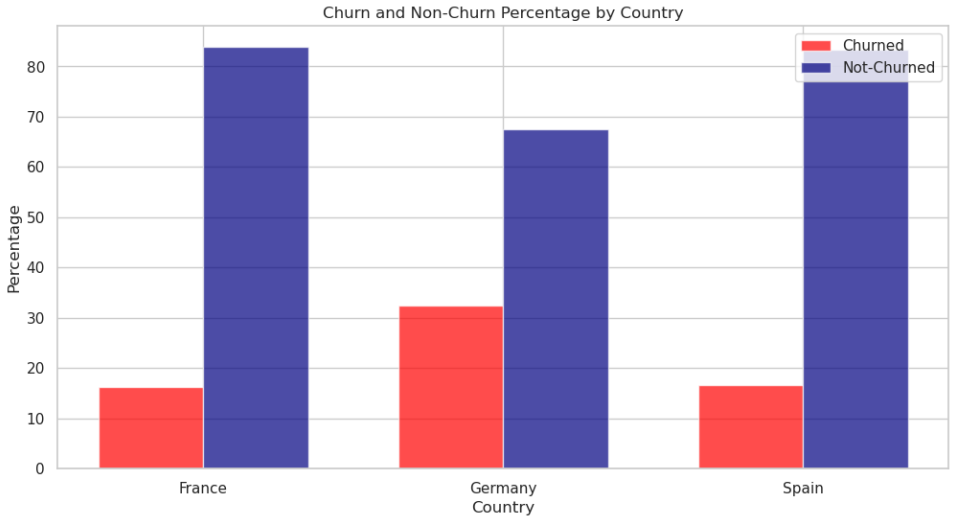

Figure 3: Customer churn percentage for each country.

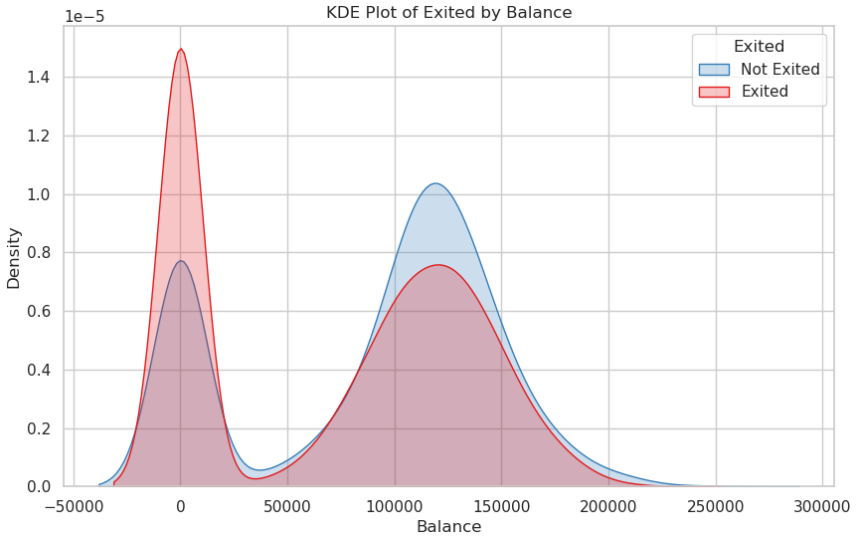

Figure 4: KDE plot of customer balance and the density of whether a customer exits (churns).

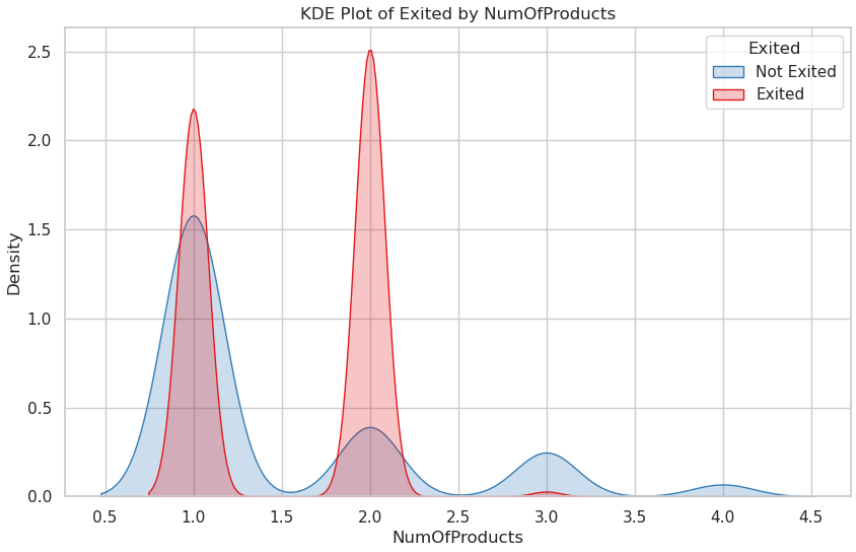

Figure 5: KDE plot of the number of products purchased and the density of whether a customer exits (churns).

Figure 6: Feature correlation with customer churn.

There were some attempts at feature engineering conducted outside of the logging of the Age feature distribution. I considered splitting the Balance and NumOfProducts features into four separate features: $0.00 or lower balance, Positive balance, 2 or more products purchased, less than 2 products purchased. Modeling with the added classes proved to have significantly less accuracy with the classifier models I used for this project. I was able to improve the models using minmax scaling for Balance, and using the initial NumOfProducts feature.

4 Model Training

I used the following classifier models for this project:

- Logistic Regression
- Random Forest
- Gradient Boosting
- Decision Tree

Considering that the data set was unbalanced, I used the Synthetic Minority Oversampling Technique (SMOTE). Initially I considered under-sampling the majority class, but the model training performed better using smote. I

also used Cross-Validation with GridSearchCV to find the best set of hyper-parameters for model training. The goal of this project is to determine which model can best predict whether or not a customer will churn. So, I consider the overall accuracy of the model, coupled with the precision, recall, and F1-score of the model's ability to predict wether or not a customer will churn (1 - churn, 0 - no churn).

Initially, I trained a standard Logistic Regression classifier without balancing the data with SMOTE. This model achieved an accuracy of 0.81, precision of 0.64, recall of 0.24, F1-score of 0.35 and the Area Under the Receiver Operating Characteristic Curve (AUC) was 0.60. Next, I trained a standard Logistic Regression Classifier using SMOTE to balance the data. The overall accuracy of the model decreased to 0.71, but the AUC of the model improved to 0.72. The model also had a precision of 0.39, a recall of 0.74 and an F1-score of 0.51, which were all improvements from the previous model.

For the rest of the models I used SMOTE along with cross-validation. For the Decision Tree Classifier, I achieved an accuracy of 0.76, precision of 0.44, recall of 0.53, F1-score of 0.48, and AUC of 0.68. Next, I moved on to ensemble methods. I started with a Random Forest Classifier, which achieved an accuracy of 0.83, precision of 0.60, recall of 0.59, F1-score of 0.59, and AUC of 0.74. The last model I trained was a Gradient Boosting classifier. This model achieved the best overall results. The model had an accuracy of 0.85, a precision of 0.67, a recall of 0.56, an F1-score of 0.61, and a AUC of 0.74.

5 Conclusion

Exploratory data analysis of the customer churn data suggests that age, geography, balance, the number of products purchased, and gender correlate highly with weather a customer will churn (exit). Multiple machine learning classifier models were trained on the dataset. The best model was constructed using a Gradient Boosting algorithm from Sci-Kit Learn. The model achieved 0.85 accuracy, and a precision of 0.67. The model had an AUC of 0.74, suggesting that the model is useful and has the ability to discriminate between whether or not a customer will churn. This study was a quick examination of possible classifier models that can be used to predict customer churn. Only four classifier models were used, which does not encompass every classifier model. It is plausible that another method can achieve better results than the models used for this analysis.