Project 1 - Bank Customer Churn Prediction

Creation of various machine learning classification models for bank customer churn prediction

Alessandro Sciorilli

Customer churn refers to the phenomenon in which customers decide to terminate their relationship with a business, specifically in this context, a bank. The rationales behind such decisions can be diverse: they may include poor customer service, shifts in financial needs, loss of trust in or damaged reputation of the bank, as well as attractive offers from competitors. Another contributing factor could be the practice known as "bank bonus hunting," a widely adopted personal finance strategy involving individuals opening and closing multiple bank accounts to capitalize on sign-up bonuses.

Customer churn is a very critical concern for banks and other businesses, since retaining existing customers is often more cost-effective than acquiring new ones. By identifying customers who are at a higher risk of churning, the bank can take pre-emptive measures to engage with them, understand their concerns, and offer them tailored solutions to meet their needs and enhance their experience.

The objective of this project is to develop and test multiple Machine-Learning Models that aim to predict, based on various features, whether a customer will choose to leave the bank or not. The label (Y) in the dataset is represented by the variable **"Exit"**. This variable takes on the value **"Yes"** for customers who have left the bank and **"No"** for customers who haven't. As a result, a binary classification task is carried out. The list of X features is as follows:

- **X_1 – CreditScore.** Numerical representation of a customer's credit creditworthiness. Customers with varying credit scores might have different financial profiles that could influence their decisions regarding continued service usage.
- **X_2 – Geography.** Geographic location might influence customer behavior. Factors such as economic conditions, competition, and cultural differences could play a role in whether customers decide to churn.
- **X_3 – Gender.** Gender might have some influence on churn, as certain genders might have different preferences, needs, or reasons for discontinuing a service.
- **X_4 – Age.** Age can play a significant role in churn. Younger customers might churn due to changing life circumstances, while older customers might leave due to retirement or changing needs.
- **X_5 – Tenure.** Tenure is the length of time a customer has been with the institution, is a strong indicator of churn. Newer customers might be more likely to churn than long-term ones who have established relationships.
- **X_6 - Balance:** A higher balance might indicate financial stability, reducing the likelihood of churn. However, extremely high balances could also indicate a lack of engagement with the institution's services.

- **X_7 – NumOfProducts.** Customers with more products might be more engaged, but they could also be more susceptible to churn if they encounter issues with multiple products.
- **X_8 – HasCrCard.** Having a credit card might not be a strong indicator of churn by itself, but it could be considered along with other features to understand the overall customer profile.
- **X_9 – IsActiveMember. I**nactive members are more likely to churn since they are not actively engaging with the institution's offerings.
- **X_{10} – EstimatedSalary.** A higher estimated salary might suggest greater financial stability and reduce the likelihood of churn. However, other factors can influence churn decisions.

The Machine-Learning algorithms tested are**: K-Nearest Neighbors, Naive Bayes, Support Vector Machines, Decision Tree, Random Forest, K-means (clustering)** and **Logistic Regression.**

After importing the dataset into a Pandas DataFrame, I proceed to remove the columns **"RowNumber", "CustomerId",** and **"Surname"**. Since these columns consist of unique identifiers, they do not contribute relevant information for building machine learning models. Additionally, to enhance their interpretability, I convert the values of the "Exited" label from numerical to categorical, converting **0** to **"No"** and **1** to **"Yes"**. While inspecting the data for missing values, I discover that all rows are adequately populated; the dataset does not contain any missing data. Upon examining the data types within the dataframe, I observe two categorical variables in the format of string: **"Geography"** and **"Gender"**. The variable **"Georgraphy"** contains the categories **"France", "Germany"** and **"Spain",** while the variable **"Gender"** contains the category **"Male"** and **"Female"**.

In my data analysis, I present various data visualizations **(Plots from 1 to 9).** A key insight emerges: the largest proportion of exited customers, constituting **40%** of the total, resides in **Germany**. In contrast, **Spain** only accounts for **20.3%** of the exiting customers. Going deeper into the data, it becomes evident that more females leave the bank compared to males, in absolute numbers. Notably, customers who hold a **Credit Card** exhibit a higher tendency to leave compared to those who do not. Moreover, the majority of departing customers are **inactive members** of the bank. When assessing age distribution, a clear trend emerges: those who leave are generally **older** than those who decide to remain. Additionally, customers who exit the bank demonstrate a slightly **higher average bank account balance**, while maintaining a slightly **lower credit score**. In terms of **tenure**, the correlation with the number of exits lacks a definitive pattern. However, a trend emerges when observing customers associated with the bank for less than a year or for a decade or more – both segments display lower exit rates. The impact of the **number of products** purchased becomes evident when studying exit trends. Notably, customers who purchase only a single product before exiting outnumber exits among those who acquire 2 or more products by over threefold.

To explore potential correlations, I analyze the correlation matrix. Interestingly, the matrix reveals a general absence of significant correlations between the independent variables, indicating no significant concerns about multicollinearity. However, an intriguing observation arises: a robust

positive correlation of approximately **40%** between the variable **"Balance"** and the country of Germany. Conversely, negative correlations of **-23%** and **-13%** emerge between **“Balance”** and the countries of France and Spain, respectively. This suggests that customers in Germany tend to maintain larger account balances compared to their counterparts in France and Spain.

As second step, I start preparing my data for the development of the machine learning models. I perform **One Hot encoding** on the variables **“Gender”** and **“Geography”**, thus converting them into a numerical format. This conversion is necessary for a successful application of the machine learning algorithms.

Given that the number of observations for people who haven't exited the bank (Exit = **"No"**) is significantly larger compared to those who have (Exit = **"Yes"**), the dataset is unbalanced towards the **"No"** class. To avoid biases in my estimation, I rebalance the dataset by setting a sample size of **2,000** observations for both the **"Yes"** and **"No"** categories. After making this adjustment, I separate the features (**X**) from the labels (**y**), storing each in separate arrays. Next, I split the dataset into training and test sets using a **70/30** ratio. This means that each of my machine learning algorithms (except K-means) is trained on **70%** of the data and tested on the remaining **30%**. As the final step before running the machine learning model, I rescale my data using the standardization method. This ensures that data with different magnitudes are normalized and reduces the impact of outliers.

To construct my initial machine learning model, the **K-Nearest Neighbor (KNN)**, I begin by plotting the accuracy corresponding to different values of **k**. This aids me in identifying the optimal value for k within my dataset, with **19** emerging as the most suitable choice due to the accuracy reaching its pinnacle at this point. Subsequently, I proceed to train the KNN model using my dataset. To ensure a more robust performance evaluation and to mitigate overfitting, I engage in **5-Fold Cross-validation**. I then generate and visualize a confusion matrix as well as a classification report to comprehensively evaluate the model's performance. This identical procedure is followed for all other machine learning models. When approaching the **Naïve Bayes Model**, I opt for a **Bernoulli distribution** over the traditional Gaussian distribution. This decision is rooted in the Bernoulli distribution's propensity for superior performance in binary classification scenarios. For the **Support Vector Machine**, I undertake training and testing utilizing four distinct kernel types: Linear, Radial Basis Function (RBF), Polynomial, and Sigmoid (Hyperbolic Tangent). Notably, the **Polynomial Kernel** emerges as the most accurate, achieving a score of **77.6%.**

After executing **training and testing** procedures for the **Decision Tree Model** using the **CART** algorithm, I delve into identifying the pivotal feature determining customer churn (the root of the tree). This is achieved by calculating the **Gini impurity** for each feature. **Age** is recognized as the most influential feature impacting a customer's decision to leave the bank, followed by **Balance** and **Credit Score**. However, a contrasting feature ranking arises when assessing feature importance through the **Random Forest Model**. While **Age** remains at first position, **Estimated Salary** and **Account Balance** ascend to the second and third positions respectively, relegating **Credit Score** to the fourth position. After running the **Logistic Regression** Model, I start

performing **Clustering**. I plot the **Distortion Score Elbow** to pinpoint the optimal number of clusters (the point at which SSSE experiences the most rapid decline). Following the **K-Means** clustering and a corresponding visualization, I proceed to compute the **purity score** utilizing various distance metrics such as **Euclidean** and **Manhattan**. The purity scores range from a minimum of **51.7%** (**Canberra**) to a maximum of **57.28%** (**chi-square distance**).

After comparing the classification reports, it is evident that the **Random Forest** machine learning model exhibits the best overall performance. The model achieves an **accuracy** of **0.76**, signifying its ability to accurately predict the correct class for approximately **76%** of all instances. In terms of **precision**, the **"No"** class yields a value of **0.75**, indicating that when the model predicts an instance as **"No,"** it is accurate **75%** of the time. On the other hand, the **precision** for the **"Yes"** class stands at **0.77**, suggesting that the model's predictions for **"Yes"** are accurate **77%** of the time. Additionally, both **recall** and **F1** score demonstrate similar levels of performance. These metrics show that the model is somewhat consistent for both classes, but it only predicts correctly in **75%**, providing only a moderate level of reliability.

Following closely is the **Support Vector Machine** with the **Polynomial Kernel**, which is the second-best performing model. Although its values for accuracy, precision, recall, and F1 are slightly inferior, they closely resemble those of the Random Forest model. The model reporting the weakest performances is the **Decision Tree** model, which accurately predicts the class for only **69%** of all instances, followed by the **Logistic Regression**, with an accuracy score of **70%.**

As a final step, I export the pickles for both the Random Forest Classifier (the best performer) and the scaler for deployment.

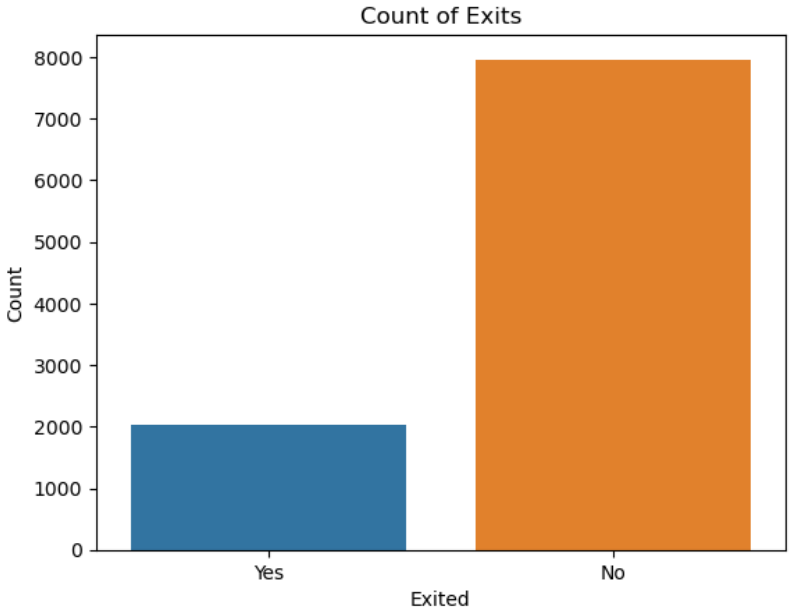
In conclusion, this project successfully developed and evaluated various machine learning models to predict customer churn in a banking context. Through comprehensive analysis of customer attributes and behaviors, including factors like geography, age, balance, and product usage, the project identified key insights to inform customer retention strategies. Among the models assessed, the Random Forest exhibited the highest accuracy at 76%, followed closely by the Support Vector Machine with the Polynomial Kernel. his model holds value for marketing and banking customer analysts seeking to predict potential customer discontinuation based on specific characteristics. By inputting customer details into the deployed model, users can obtain a probability-based prediction, serving as a foundational tool to formulate effective retention strategies.

Access to the deployed model is available at the following link:
https://asciorilli.pythonanywhere.com/Bank_Customer_Churn_Prediction

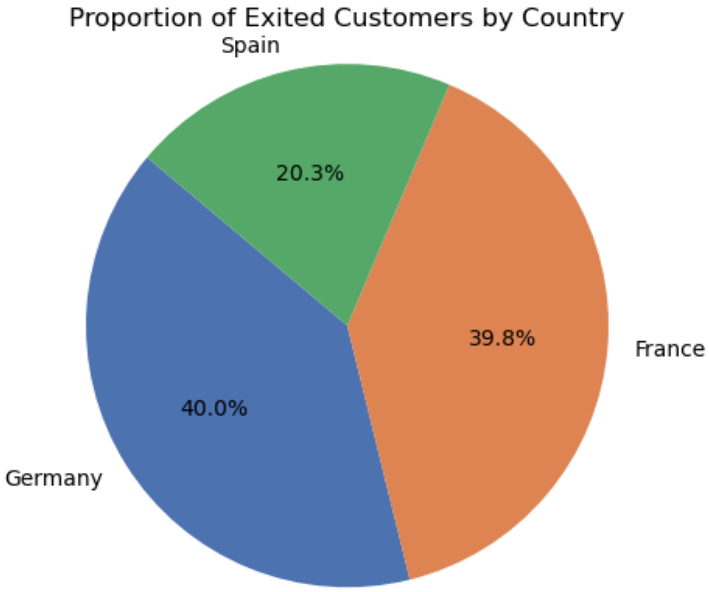
The Dataset is available at the following link:
https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction?select=Churn_Modelling.csv

Appendix

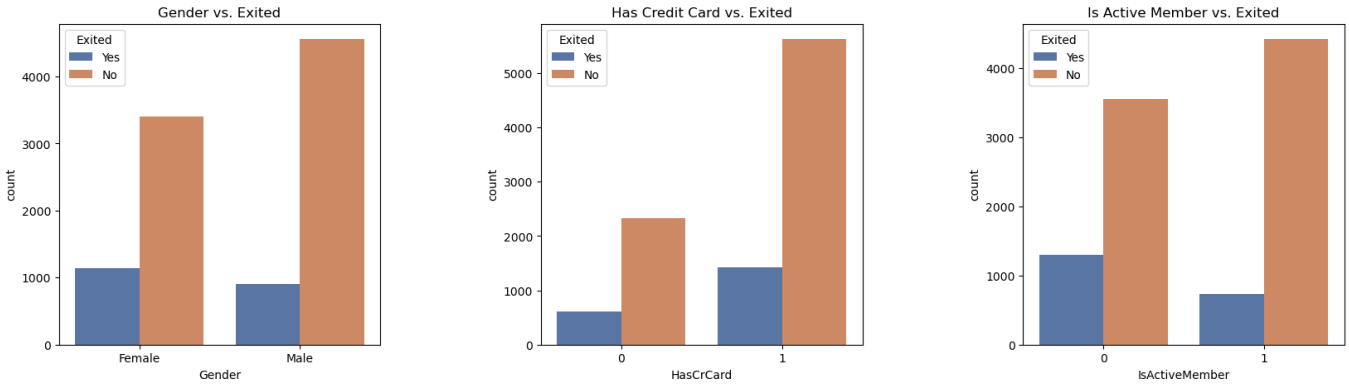
Plot 1 - Count of Exits



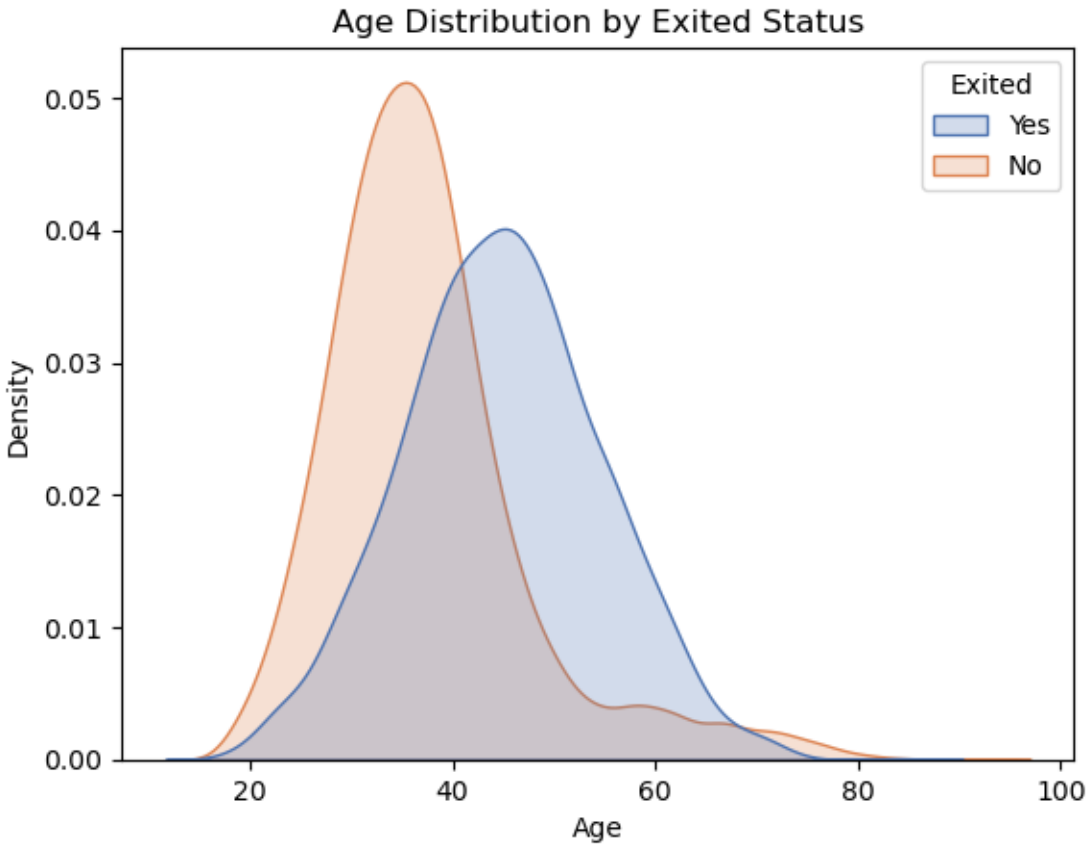
Plot 2 - Proportion of Exited Customers by Country



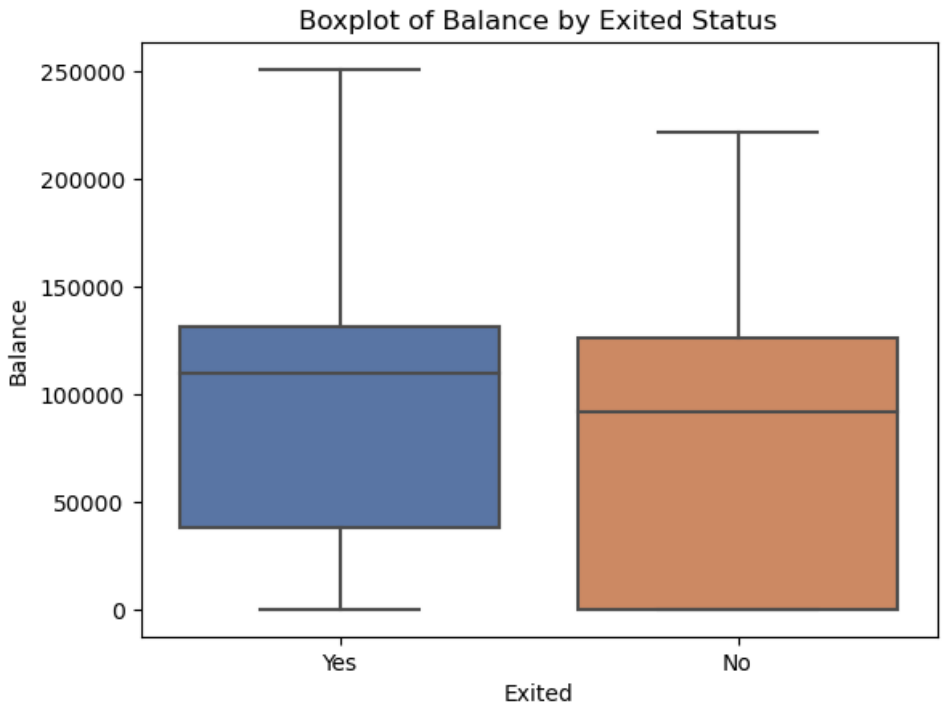
Plot 3 - Proportion of Exited Customers by Gender, Credit Card Possession and Active Member Status



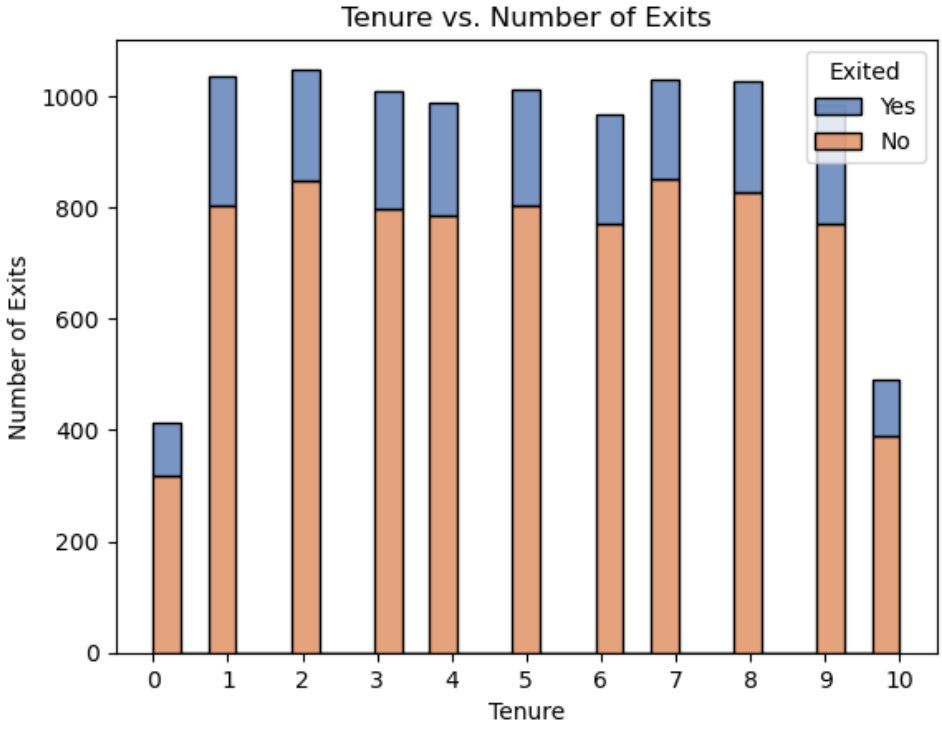
Plot 4 - Age Distribution by Exited Status



Plot 5 - Boxplot of Balance by Exited Status

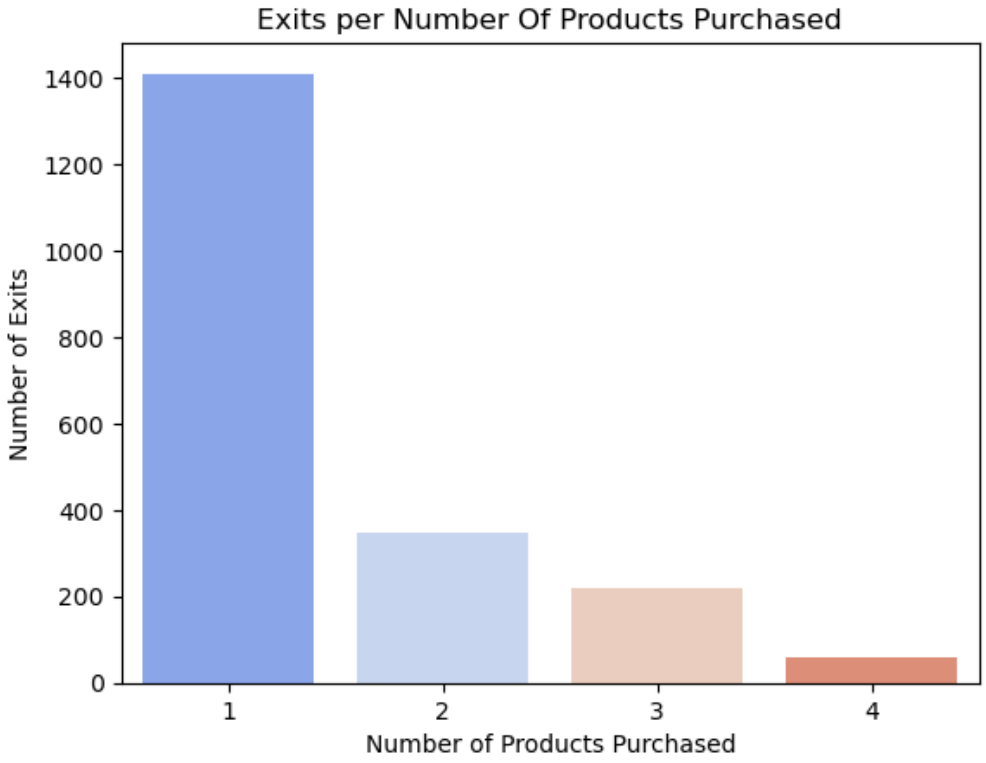


Plot 6 - Tenure vs. Number of Exits

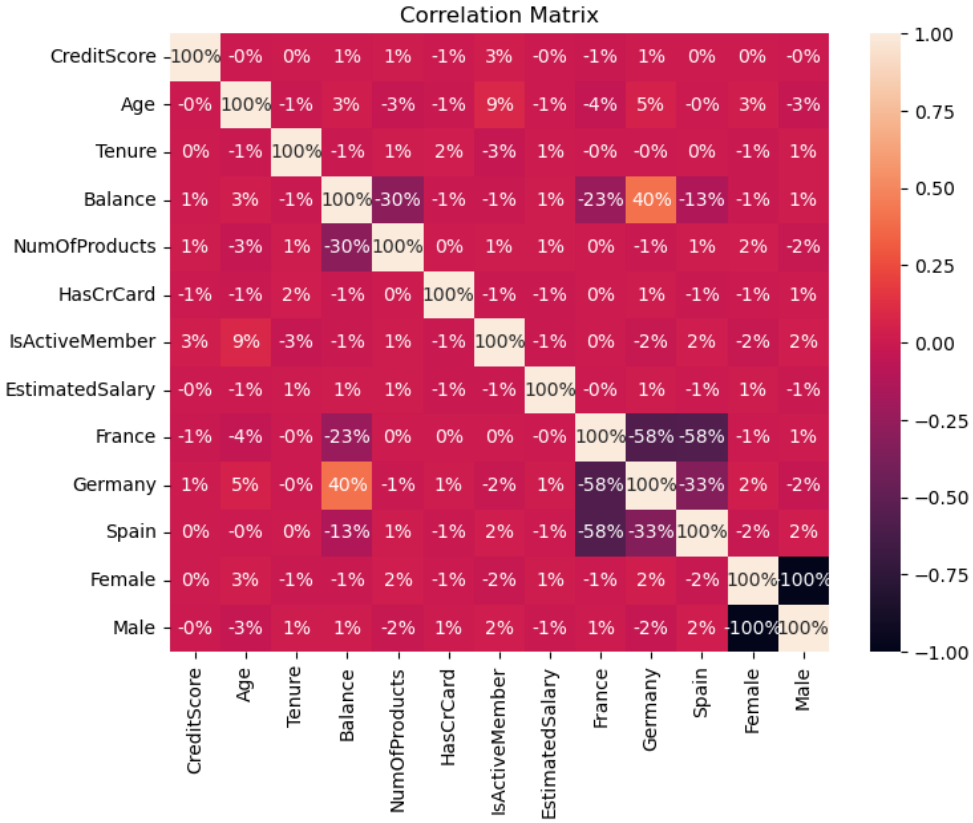


Plot 7 - Violin Plot of CreditScore by Exited

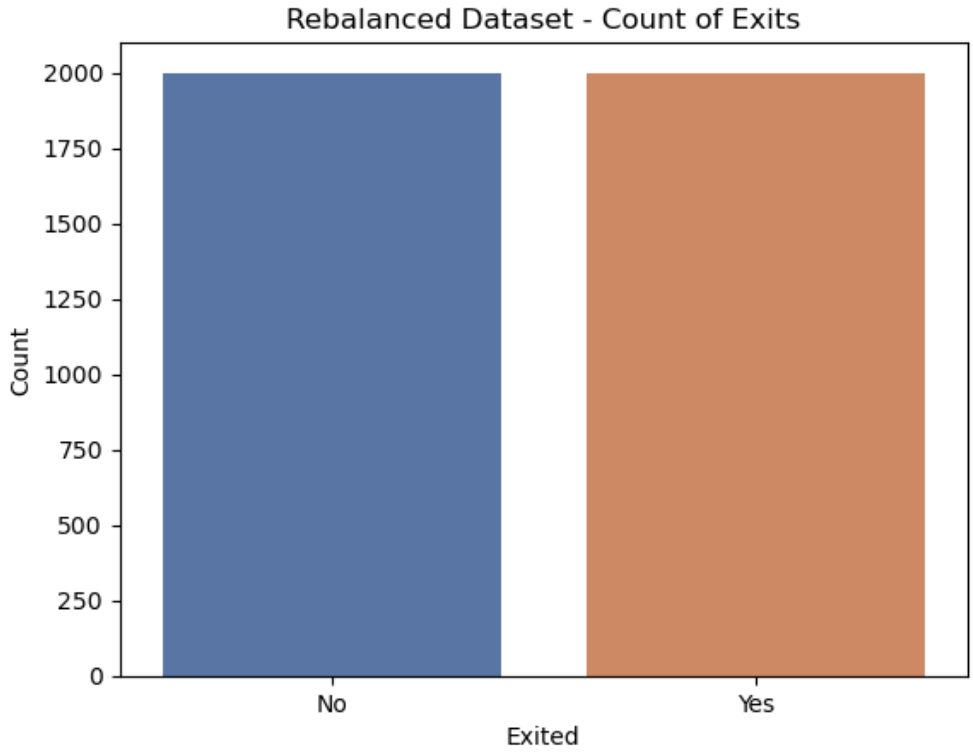
Plot 8 - Exits per Number Of Products Purchased



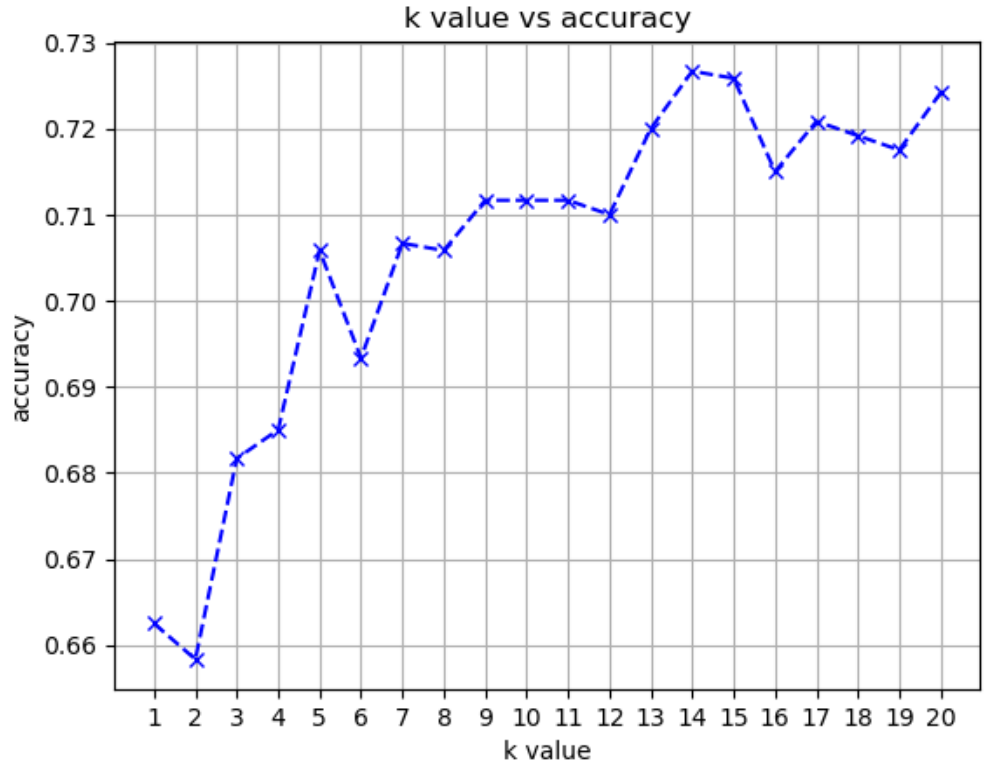
Plot 9 - Correlation Matrix



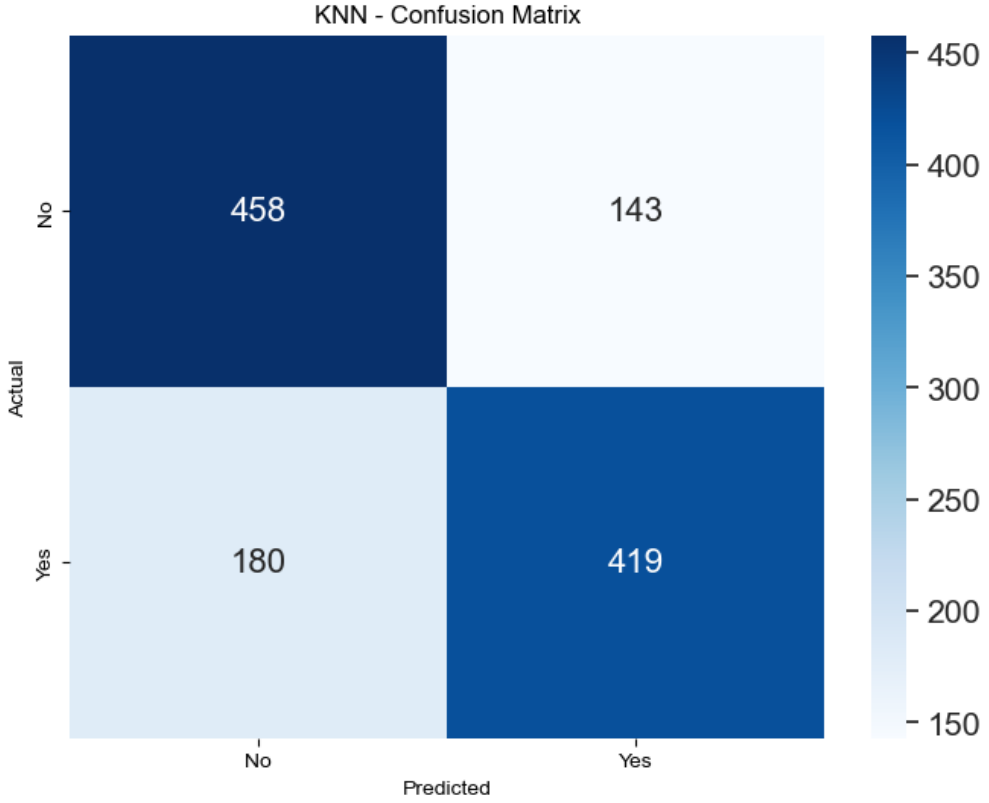
Plot 10 - Count of Exits with Rebalanced Dataset



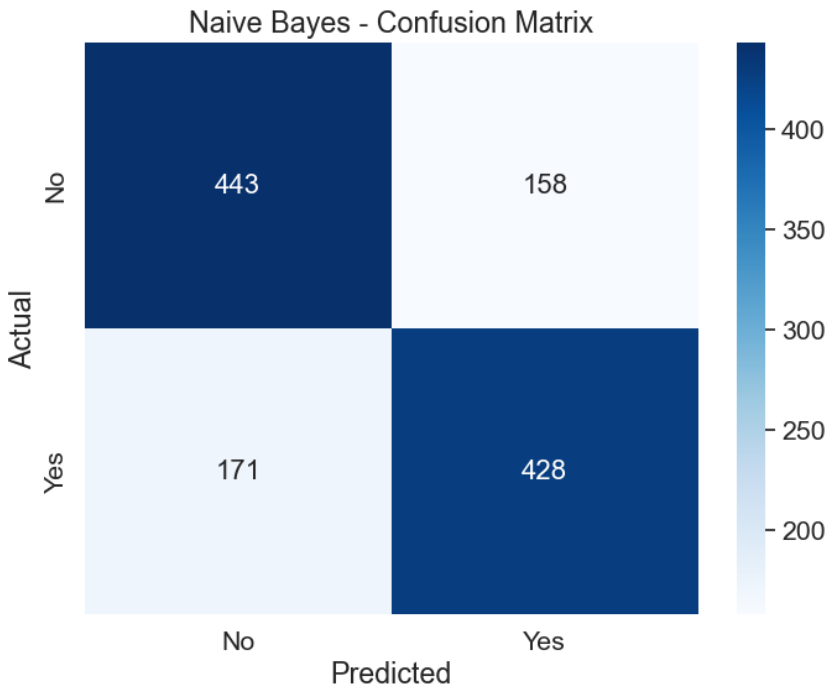
Plot 11 - K-Value vs Accuracy



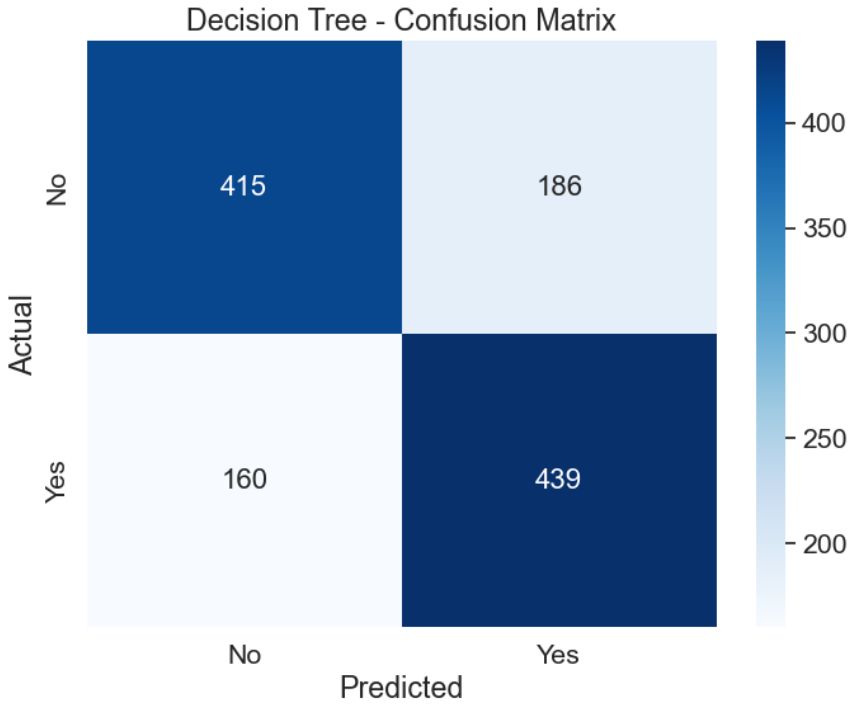
Plot 12 - Confusion Matrix for KNN



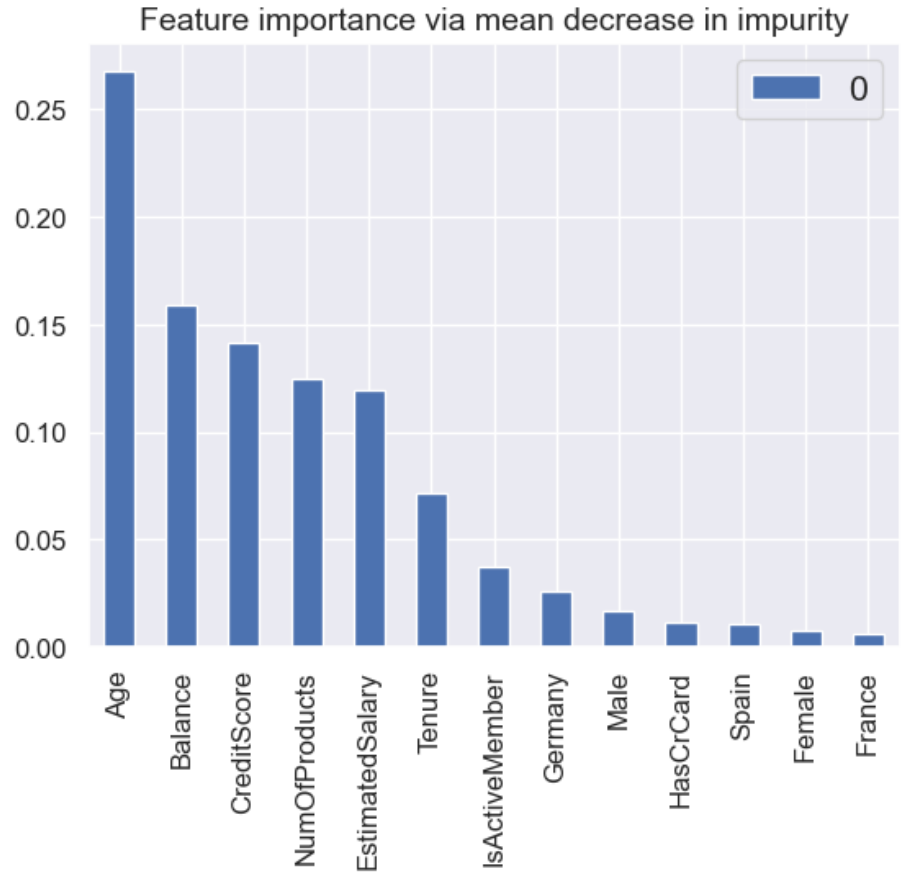
Plot 13 - Confusion Matrix for Naive Bayes



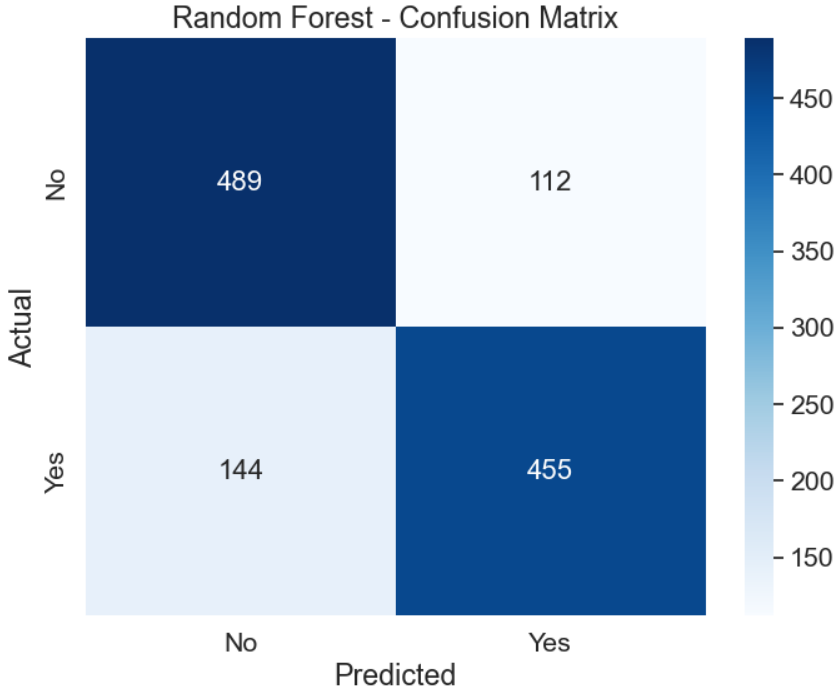
Plot 14 - Confusion Matrix for Decision Tree



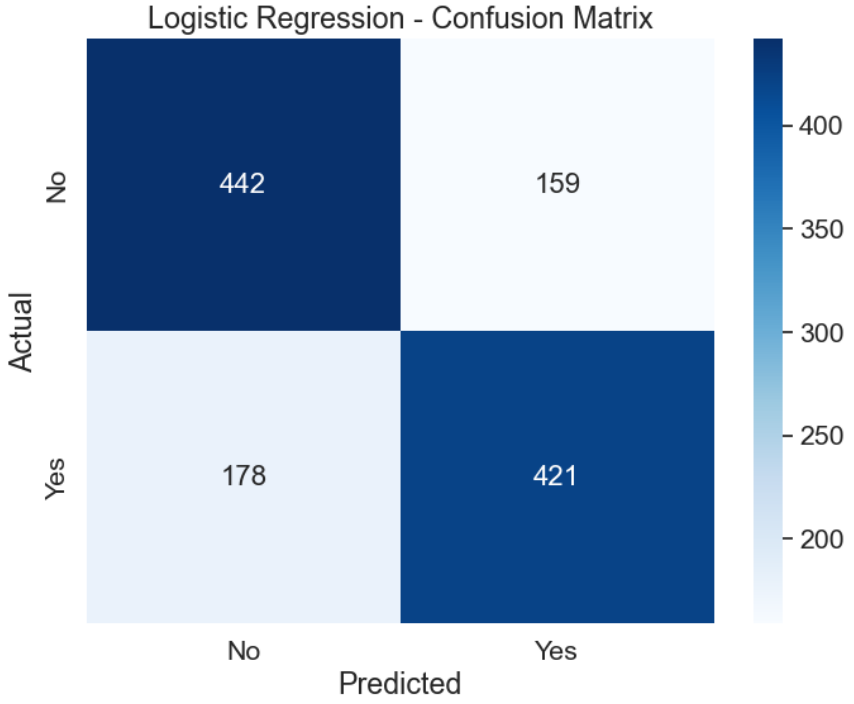
Plot 15 - Bar Plot Depicting Feature Importance



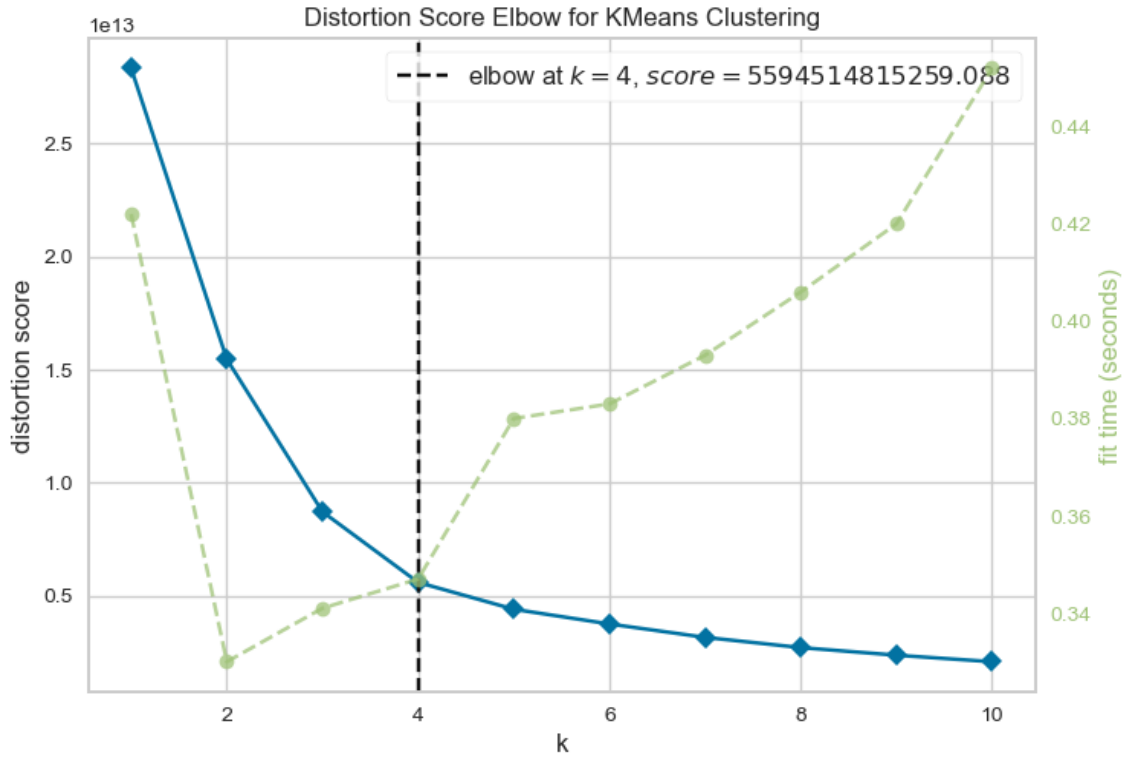
Plot 16 - Confusion Matrix for the Random Forest



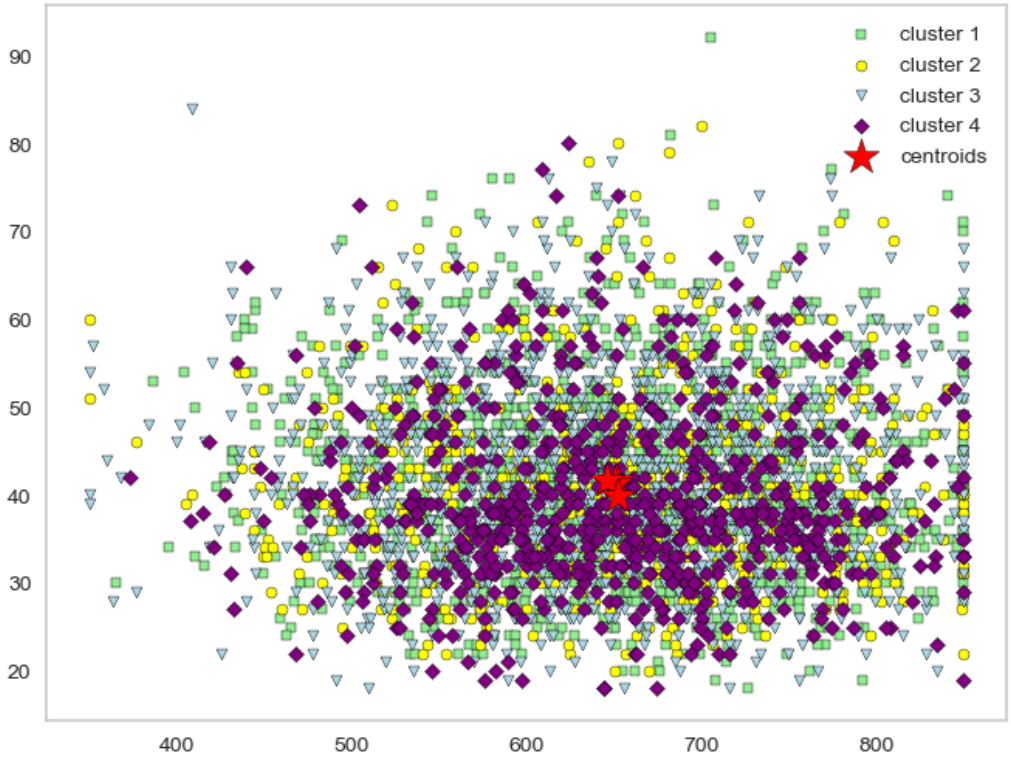
Plot 17 - Confusion Matrix for Logistic Regression



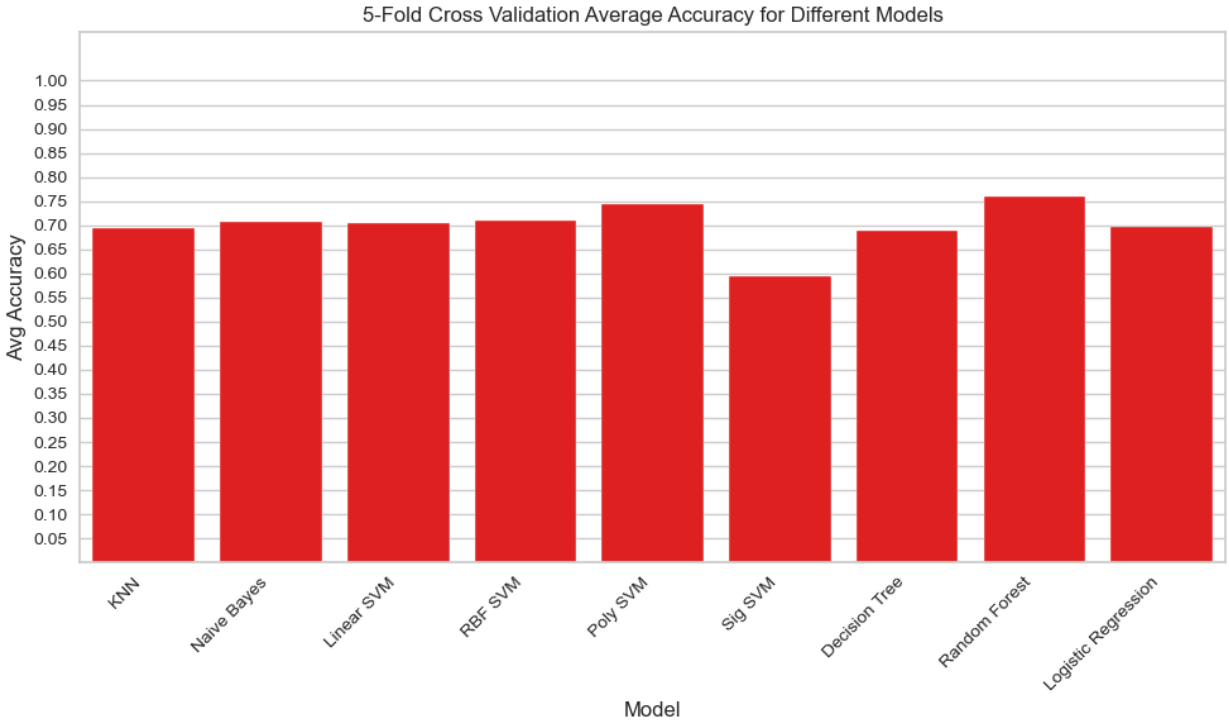
Plot 18 - Distortion Score Elbow for KMeans Clustering



Plot 19 - Clusters Visualization



Plot 20 - 5-Fold Cross Validation Average Accuracy for Different Models



Plot 21 - Model vs Weighted Avg Of Various Metrics