Final Report for Data Analytics Project

By: Ahmad Sohail | 64011748

Project Title: *Predicting Customer Churn based on several factors.*

Introduction:

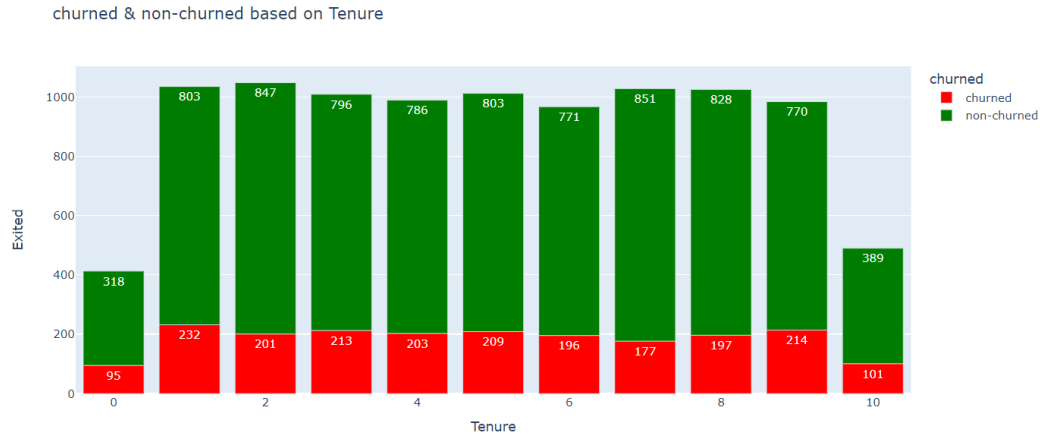
In the ever changing landscape of the banking industry, the capability to predict customer behavior is a pivotal concern. The industry faces the continuous challenge of discerning patterns in customer actions, particularly the critical issue of customer attrition, commonly referred to as churn. My objective was to dissect this complex problem by leveraging data analytics to anticipate whether customers are likely to discontinue their services.
The significance of churn prediction transcends mere statistics; it represents an opportunity for banks to proactively address and mitigate customer departures. By understanding the factors that contribute to customer churn, from card types to account standings, banks can craft targeted strategies. These may encompass tailored promotions, bespoke communication, and refined service offerings, all aimed at bolstering customer retention and positive feedback.

Data Overview:
In this project, I delved into a dataset comprising roughly 10,000 records and 18 variables, detailing various customer characteristics for a banking institution. During the initial analysis, I identified several columns that were of no analytical value, such as unique customer identifiers. These columns were discarded to streamline the dataset, as outlined in both our preliminary report and presentation.
Central to our investigation is the 'Exited' column, which serves as the target variable for this binary classification problem. It is coded in binary format, with '1' indicating a customer who has left the bank and '0' signifying retention. This aligns with previous classroom discussions where we tackled issues of dimensionality and clustering.
Furthermore, the dataset presented variables of mixed types. Non-numeric features such as card type, gender, and geography required encoding to quantitative values to fit the predictive modeling process. The initial assessment of the dataset reaffirmed the complexities often encountered in real-world data, necessitating a comprehensive approach to preprocessing for subsequent modeling.

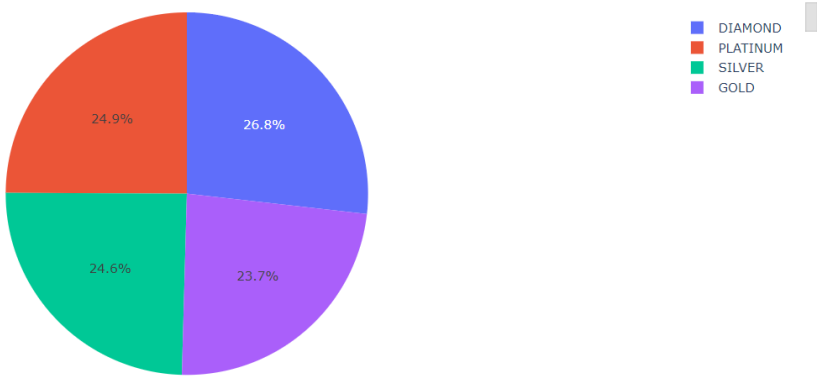
Data Preprocessing:
In the data preprocessing phase, my focus was to ensure the dataset's readiness for modeling. It was a relief to find no missing values, which often complicate the preprocessing stage. However, there were features that necessitated removal for their irrelevance to the churn prediction, namely 'RowNumber', 'CustomerId', and 'Surname'.
I then turned my attention to encoding categorical variables. The 'Card Type' variable, with categories such as Diamond, Gold, Silver, and Platinum, along with 'Gender' and 'Geography', were transformed from categorical to numerical representations, conducive to our analytical models. Geography was particularly interesting as it was confined to three countries: France, Spain, and Germany, hence a straightforward encoding.

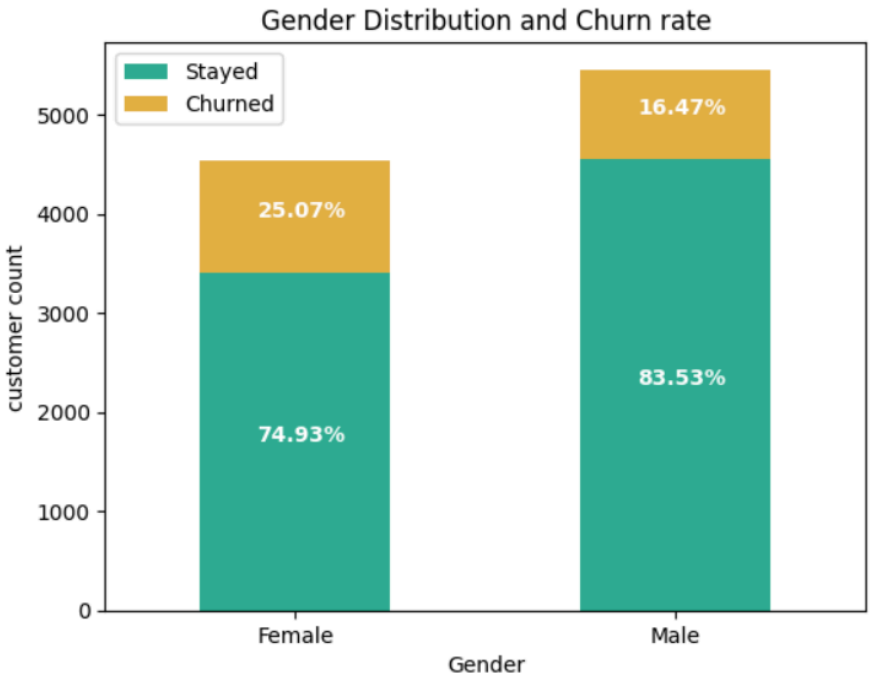
Post-evaluation, I engaged in fine-tuning the models. This involved employing techniques like SMOTE for balancing the classes and tweaking parameters for optimal performance. An interesting discovery was the 'Complain' feature's influence, which initially seemed to skew the model towards overfitting by making it memorize rather than learn. Removing it enhanced the model's ability to generalize and made it much more realistic and an actual model rather than a preprocessed machine running algorithm.

Exploratory Data Analysis (EDA):
The EDA was instrumental in discerning the predictive power of various features. Visual analysis brought forth the prominence of variables like 'Age', 'Tenure', and 'Balance' over 'Gender' or 'Number of Products'. It was evident that certain factors had a pronounced impact on a customer's likelihood to churn. Below are a few visualizations that I used which cleared out the room for more hindsight and helping find areas of concern:

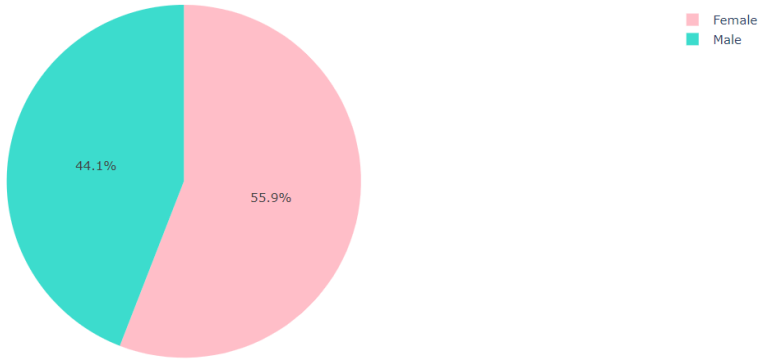


Churned percentage based on Card Type



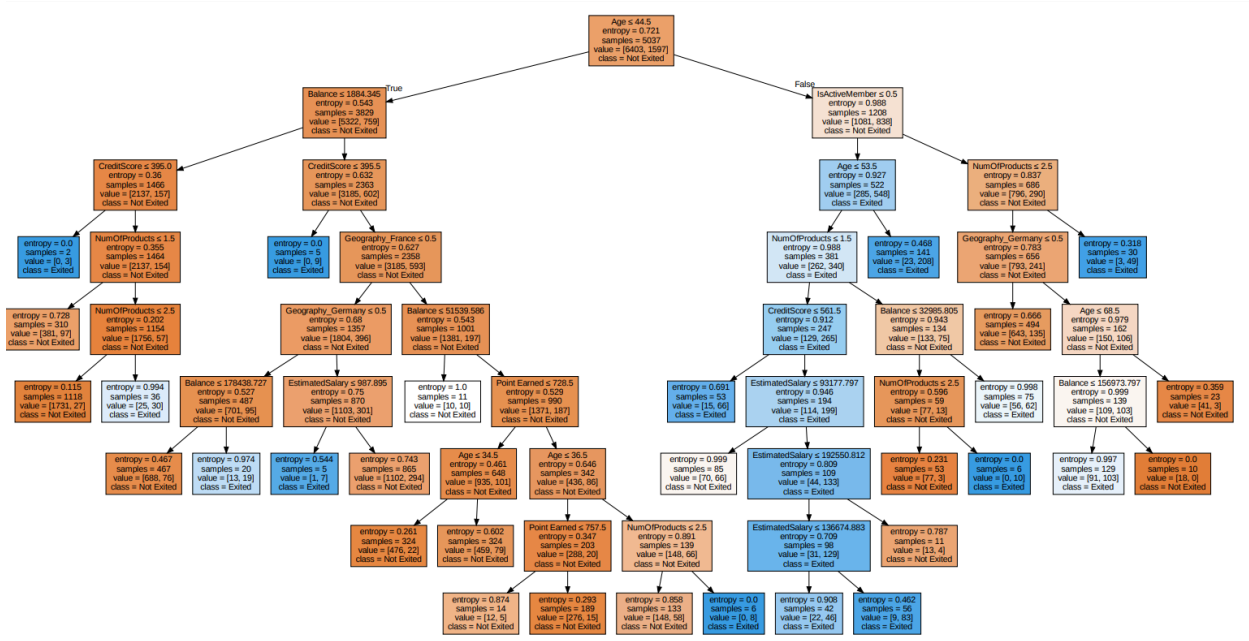


Churned percentage based on Gender

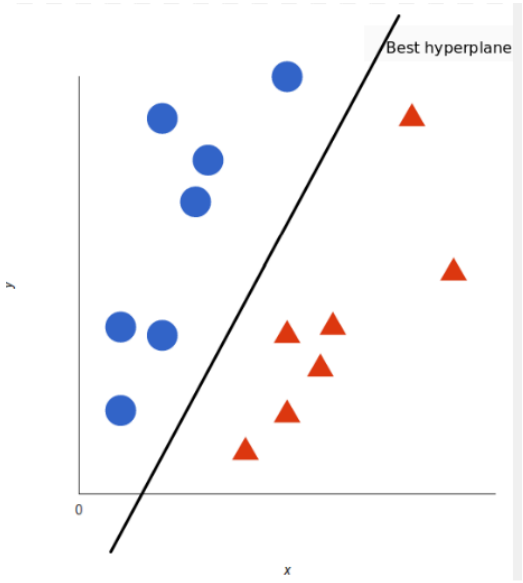


Modeling:

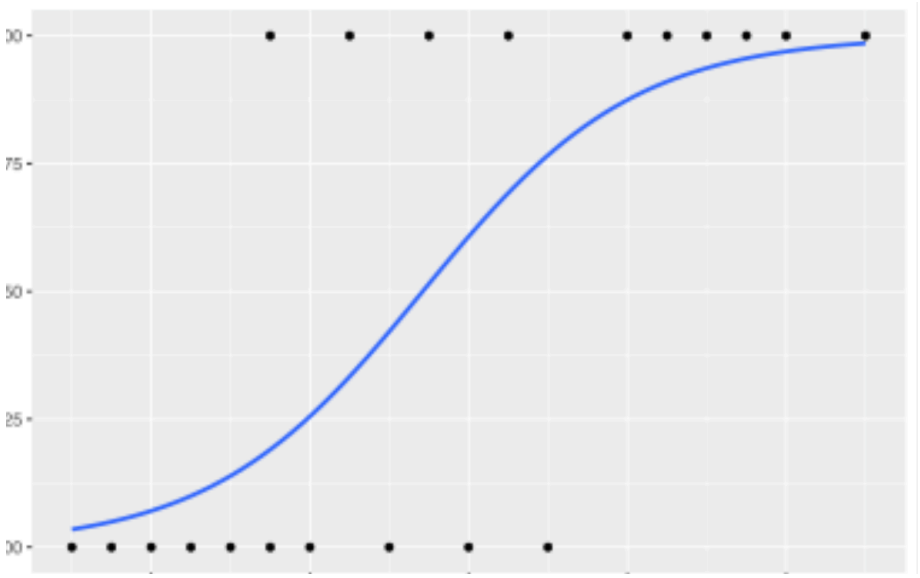
In my approach to modeling, I employed three models: Random Forest, Logistic Regression, and SVM. The selection of these models was based on their relevance to binary classification challenges and recommendations from experienced seniors. Each model was rigorously tuned, exploring various hyperparameters, test sizes, and feature selections to optimize performance. I specifically focused on the balance between recall and precision, aiming for a model that performs well across both majority and minority classes.



Entropy Graph



SVM



Logistic regression

Results:
The iterative modeling process yielded a spectrum of accuracies. Initially, a modest 69%. That was the case because I fit the dataset with a missing column. I, in fact, didn't take the original author's dataset. A subsequent naive model fitting surprisingly suggested a 99% accuracy, an immediate red flag for overfitting, corroborated by an improbable 100% accuracy in another iteration. These two findings were received by the correct author's dataset but the issue lied in what each column represented. After adjustments, particularly the removal of the 'Complain' feature, the accuracy settled into a more credible range of 80-90%. My logistic regression model, after fine-tuning, demonstrated particularly promising results, achieving a high precision and F1 score.

Discussion & Conclusion:
Reflecting on the project, I am confident that the objectives were met with success. The deployment of multiple models and their validation against real-world standards have paved the way for practical application. The final models not only predict churn effectively but also offer insights into the subtleties of customer behavior, paving the way for strategic interventions. I got the appropriate dataset, built the models, tested them, evaluated them and displayed my results in a concise manner. In my opinion it works really well. This project, though challenging, stands as a testament to the effective application of data science in addressing critical business questions.