MSDS 422 Final Project - Customer Churning Prediction

Zachary Cimel & Jerry Zhao

Dataset: https://www.kaggle.com/competitions/playground-series-s4e1/data

Github: https://github.com/RamenNoodleJerry/My-Projects/tree/main/MSDS%20422%20Project/Final%20Project

Executive Summary

Customer data can be used to identify whether someone has exited, whether that means something like unsubscribing or moving to a different service. By using the Bank Customer Churn dataset we tuned four classification models to obtain the most accurate results when predicting customer “Exited” categories. Initial EDA revealed that 11 features were quantitative and 3 were qualitative. The data was clean with no null values in any of the columns. Plotting the correlations of the quantitative data revealed that balance and age had the strongest correlation to whether or not a customer exited whereas several products and whether the customer was active showed the least correlation.

To Do:

- Explain any other EDA we complete
- Explain encoding if we do it for qualitative data
- Feature engineering
- List the models we use (can’t be done for the midpoint)
- Explain the results (can’t be done for the midpoint)
- Conclusion- what do we recommend (can’t be done for the midpoint)

Problem Statement/Research Objectives

For businesses, analyzing Customer Lifetime Value (CLV) and ways to retain customers is essential to creating a profitable and sustainable business. Continually acquiring new

customers is critical, but successful businesses must know when to instead maximize the value of existing ones and focus on customer retention. In general, it is 5 times more expensive to acquire a new customer than it is to retain an existing one (Source). By exploring customer data, companies and organizations can benefit from a model that can categorize customers as likely to exit in some way i.e. unsubscribe, quit, or move to a competitor. They can use these predictions as a guideline for which customers to focus retention efforts. Furthermore, classification models can help identify which features of a customer correlate to exiting thus allowing them to optimize their retention plan.

By using the Bank Customer Churn dataset we hope to be able to create a model that can categorize and predict whether or not a customer has exited based on various features such as credit score, balance, and salary. Significant exploratory data analysis along with hyperparameter tuning will allow us to identify important features of customers that we can recommend to banks to focus on to retain current customers and maximize CLV. After using the training and test sets to build and optimize our models we will be able to report on what these features are and deliver a tool that can be used to analyze a customer base and provide insights to banks on ways to reduce customer churn, a key action to increase profitability.

Exploratory Data Analysis

We start by checking missing values and outliers of the data. Fortunately, the data is clean enough with no outliers across all variables. We checked from boxplots (Figure 1), there are some outliers across “number of product“, “credit score“, “age“, and “has credit card flag” variables. These outliers are natural variations and should not be dropped. We will keep those outliers. We also find some irrelevant variables like customer ID and surname, which should be dropped and not used in model training.

Next, we check the correlation heat map (Figure 2) of numeric variables and customer churn flag. From the plot, we can see variables like the “number of products“ and

“isactivemember“ are negatively correlated, and “age” and “balance” are positively correlated. This helps us confirm features we used are likely to be useful.

We want to confirm the distribution of dependent variables, the churn flag and see if there is an imbalance. Checking the histogram (Figure 3), we can see there is indeed imbalanced distribution of positive and negative. There is only around 20% of churning cases. We may need to deal with the imbalance later otherwise models like Logistic Regression and Support Vector Machine (SVM) would produce biased results. We then check the two-way boxplot of categorical variables (Figure 4), and see Germany and female are more likely to churn.

Data Preparation/Feature Engineering

During the preparation stage, we will first drop irrelevant features like customerids and surname. Following on, we will keep outliers and split the dataset into 80% train and 20% test. Next, We encode every categorical feature into multiple boolean dummy variables using Sklearn one-hot encoding. As an example, geography location will be encoded into a binary True/False dummy based on the value. If a sample is in German, a flag of True will be assigned to Geography_German column, while the other Geography columns like Geography_Spain will be False. This allows us to include categorical features in model training. Finally, since the value of some features like estimated salary are significantly higher than others, when training linear / non-linear classifiers like Logistic regression and SVM we will use a standard scaler to make sure all features are on the same scale. This will drastically reduce these models’ bias. For tree-based models like decision tree classifiers, we will not apply standard scaling as they are not based on absolute value of features.

Figure 1:

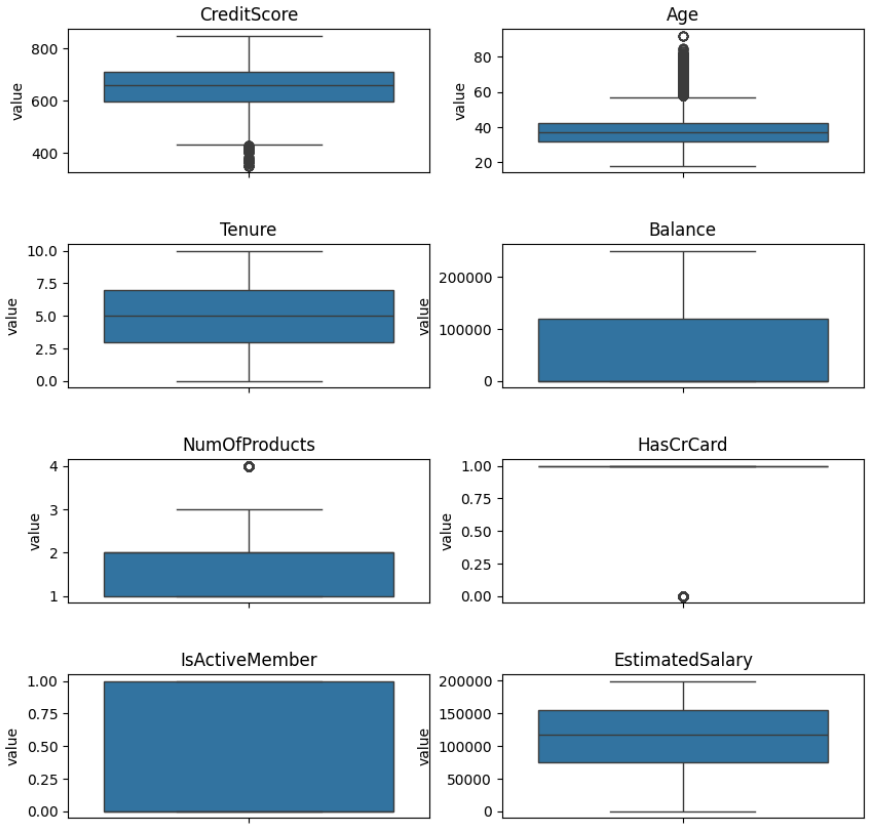

Figure 2:

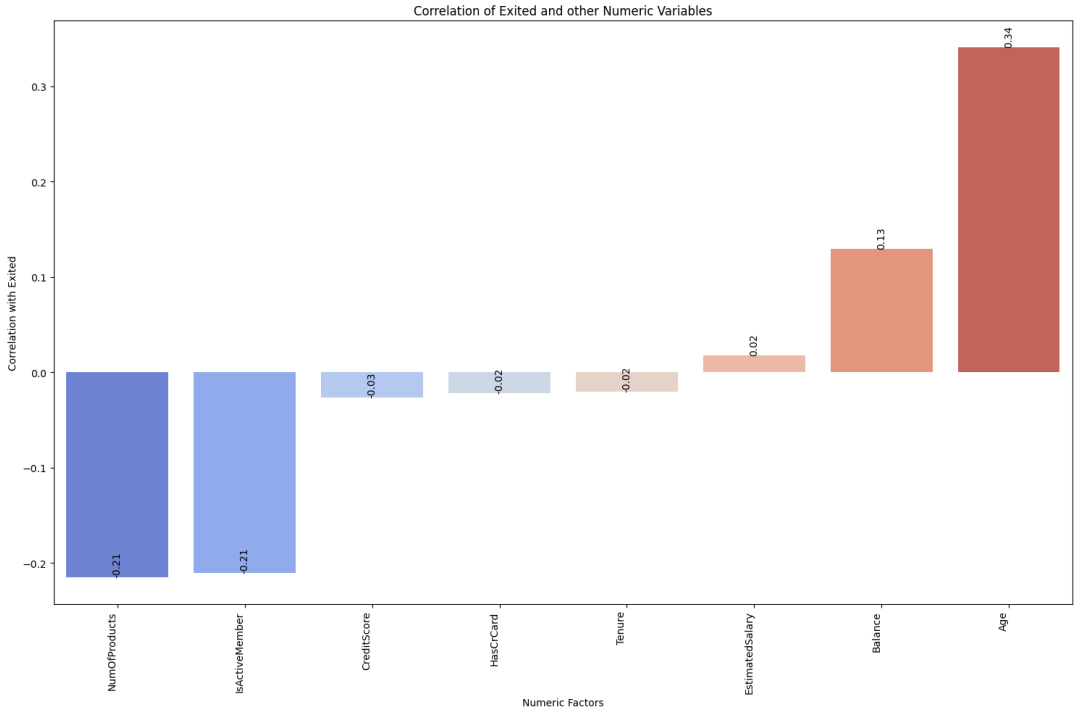

Figure 3:

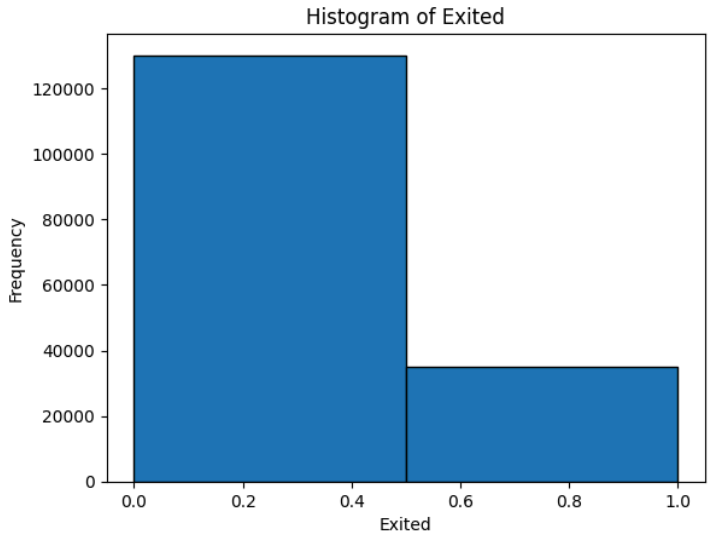

Figure 4: