Shima Abdulla
INST 414
December 16th, 2020

Final Project

The dataset I utilized for the scope of this project was found on Kaggle at the link below. The dataset itself was uploaded by user Barun Kumar. That stated the source of the information by which the dataset was comprised was not mentioned. The stated the dataset itself contains information regarding customers at a telecommunication company. The dataset itself is comprised of 3334 rows of data and eleven rows. My target was the Churn variable. Churn is depicted as a binary variable where 0 represents a customer who did not cancel their service whilst 1 was someone who did. Through this project I aim to determine which algorithm would give the most accurate prediction of possible churn based on the features described. This stated I had ten features that I examined described below:

- CustServCalls: Number of calls to customer service
- OverageFee: Largest overcharge fee from the past year
- MonthlyCharge: Average monthly bill
- DataUsage: Number of gigabytes of data used in a month
- DayMins: Average number of monthly daytime minutes used
- DayCalls: Average number of daytime calls
- AccountWeeks: Number of weeks the account was or has been active for
- ContractRenewal: If the customer has renewed their contract in the past
- Data Plan: If a customer has a data plan or not
- Roaming: Average number of roaming minutes

The first algorithm test that I preformed was the Logistic Regression. I chose this test because my target variable is a discrete variable. After performing this my accuracy was .087. I next performed a decision tree Classification and determined the accuracy to be 0.89. Next, I performed an XGBoost algorithm test and determined the accuracy to be 0.91. Lastly, I performed the Keras test and determined the accuracy to be 0.85. In simple terms accuracy represents how close a value is to the true value. Thus, when comparing all the models it is clear to see that the XGBoost algorithm performed the best as the predictive values with the x_test where closest to the true values of the y_test. This stated the difference between all for test is not wide. In fact, to improve the accuracy of the lowest performing algorithm namely the keras algorithm one could include more Epoch. Additionally, the precision value was also determined the precision describes the ratio of true positives to total predicted positives. In our case all models always have higher numbers for the zero Boolean which meant a customer has not renewed.

Additionally, I wanted to look at which features may be better indicators for determining churn. This is best done by looking at the correlation rate and the graphs. Overall, the number of customer service calls variable had the highest correlation rate at 0.208 indicating that there is a slightly positive relationship between churn and CustServCalls. In contrast Contract renewal have the lowest negative correlation indicating that there may be an indirect relationship between the two variables. The roam variable had a correlation value close to zero indicting no relationship.