DSC 550

Inman, Gracie

Final Project

11/18/23

Project Summary

The Telco dataset was obtained from Kaggle. The Telco dataset was acquired in hopes to accurately predict customer churn. For companies, churn is when a customer chooses to cancel their plan and/or not renew their subscription. This means that they will no longer be a customer and customers are the source of revenue for Telco. Predicting and preventing churn will prevent loss of revenue from lost accounts. To a group of stakeholders, I would pitch this as a money saving and preventive measure. With a plan for identifying customers who will churn, this shows an active plan for customer retention. Being able to accurately predict those who would turn, could allow for targeted deals instead of customer wide and this can also prevent a substantial amount of loss profits by limiting marketing and retention deals by using targeting. Having an active plan for marketing and customer retention as well maximizing profits will show the stakeholders that investing in Telco is not a risky investment. Once it is proved to the stakeholders that Telco has active measures to reduce risk, the stakeholders will be more likely to want to buy in.

Finding a dataset and problem I wanted to solve was arguably the most difficult part. For my dataset, I wanted to solve a problem that would be applicable across multiple different fields. Searching through Kaggle, I came across the Telco dataset which could be used to identify and predict churn for this company. Churn is a problem for many fields with a customer bases and is important for the reasons described above. After deciding to use the Telco dataset, I identified the

goal was to create a model that would accurately predict churn and allow for selection of the customer to prevent churn. After finding the dataset and deciding what I wanted to achieve, I began the next step to visualize the data.

To first begin analysis, the data was loaded into a data frame then checked to make sure it loaded properly. The first visualization were pie charts to visualize turn with relation to if the customer had children or not. This was shown that 15.2% of people with children turned. While 31.7% of people without children turned. This shows a larger percentage of people without children turned versus people with children. A bar graph of the number of people who churned by contract length shows that a large majority of people who churned were on a month-to-month contract. I was then curious if the total amount of monthly charges was connected to how long the customer has been with the company. However, no clear correlation was found. Using count plots, I wanted to visualize how many people had dependents and paperless billing. I also want to see how many of the customers in the dataset have churned. The count plots showed that the dataset contains a larger number of customers who do not have dependents. This also showed that more customers have paperless billing than don't. Lastly, it was shown that a larger number of the customers have not churned than customers that churned. After visualizing the dataset using graphs. I began to transform the dataset in order to prepare for analysis and training a model.

I began data transformation by checking for missing values in the dataset. This is important because missing values can cause issues during building the model and analysis. There were no missing values throughout the entire dataset. I then checked the data frame again to visualize what else needed to be done to transform the dataset. I noticed a lot of columns that I felt were not critical to customer retention such as online security, online backup, device

protection, and tech support. These columns were not attributes I wanted to focus on in terms of customer retention and therefore they were removed from the dataset. I checked the dataset for the number of unique values in each column. This becomes important in identifying categorical variables that I will need to standardize.

There was a significant number of categorical columns that needed to be standardized. The following data were standardized for analysis: contract term, internet service, dependents, paperless billing, and churn. The columns containing yes, and no values were standardized by transforming the numbers 0 (False) and 1 (True). The contract term and internet service columns were standardized by mapping the three responses with the numbers 0 to 2. For example: month-to-month was 0, one year is 1, and two years was 2 for contract length. The column names were then replaced with more user friendly and consistent formatting. This allows for better understanding and readability of the data frame. Describe was used to get a statistical overview of the data to determine next steps. Using z-scores the tenure and monthly charges columns (the only columns that were not categorical) were checked for outliers using a threshold of both two and three. No outliers were found in the dataset. Duplicates were dropped from the dataset to eliminate bias. The shape was checked before and after and 100 rows were removed from the dataset. The dataset was highly imbalanced in terms of churn and therefore SMOTE was used to balance the data in order to obtain the most accurate analysis. The dataset was resampled in order to prepare the data for modeling.

The first model I chose was logistic regression. This model was chosen due to the large number of categorical variables within my dataset. The accuracy for the logistic regression model was 0.78 and the precision was 0.66. The next model I wanted to try was Random Forest. This is due to how Random Forest does not assume a linear relationship. This model also helps

prevent overfitting which is something I wanted to avoid. The accuracy for the Random Forrest was 0.76 and the precision was 0.60. Due to the size of my dataset, I also wanted to try the Naïve Bayes model. The accuracy for the Naïve Bayes model was 0.76 and the precision was 0.56. The last model I wanted to try was a decision tree. The decision tree models are commonly used in churn prediction due to how it splits the data into smaller groups and considers both numerical and categorical variables. The accuracy for the decision tree model was 0.72 and the precision for this model was 0.51. Based on accuracy and precision, I had previously chosen the logistic regression model as the best model for the data set and performed cross validation to confirm the accuracy and precision. However, accuracy and precision are not always the best method for choosing a model.

My initial goal with training the model was to obtain the highest accuracy and precision. However, after doing a confusion matrix for each model, it was shown that accuracy and precision are not always a good measure of how well a model is working. It is also important to consider how the model is accurate. The confusion matrix showed that my initial model choice was not the best for my problem. All of my models had decent accuracy and precision, but when looking at the confusion matrix I discovered that for my problem accuracy and precision were not accurate descriptors of how well the models were working. Looking at the model in terms of true negative, false negative, true positive, and false negative gave me a better overview of how well the model was working. While true negatives are important, for my issue the important category is true positive (how many times the model predicted a customer would turn and they did turn). This is important because I am targeting people who will turn. I also want a smaller number of people who were predicted not to turn but did turn (false negatives). False positives are important because they indicate a customer will turn and they don't. This causes wasted

money and resources on targeting this customer. However, giving deals to a loyal customer in goal of retention will not cause a customer to churn. Looking at the confusion matrix as well as accuracy and precision led me to discover that the best model for my dataset is actually Naïve Bayes and not Logistic Regression. This was due to the large number of false negatives and lower number of true positives by the Logistic Regression model than the Naïve Bayes as shown below in figure 1 by the confusion matrix. The categories of the confusion matrix are 0,0 is a true negative (top left), 0,1 is a false negative (bottom left), 1,0 is false positive (top right) and 1,1 is true positive (bottom right).

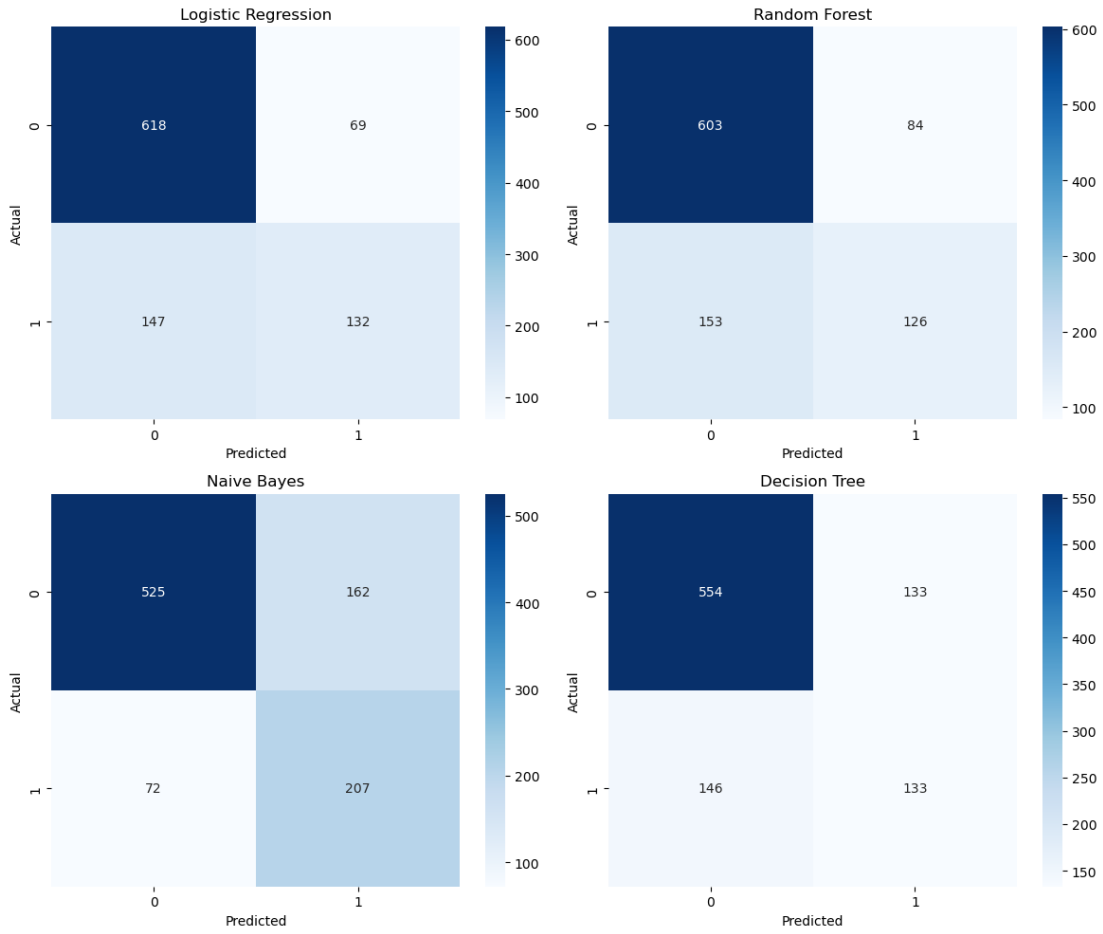

Figure 1: Confusion matrix for each model.

The Navies Bayes model tells me how accurately it can predict whether or not a customer will churn. This helps us predict within a certain certainty customer retention and how to target customers in order to prevent churn. While the Naïve Bayes seems to be a good model for the provided dataset, there are other models that exist to try. These models could be a way better fit for the dataset. In addition, the current chosen model can be changed/fine-tuned in order to better fit the dataset and improve accuracy and the number of true positive predictions. The model could also be reevaluated using the removed columns from earlier in analysis. While these columns were not something that I deemed relevant enough to consider in churn prediction, having more things to base churn off of could help improve the number of true positives and decrease the instances of false negatives and positives.

Citation:

Arighy, Reyhan. “Data Telco Customer Churn.” *Kaggle*, 29 Sept. 2023, www.kaggle.com/datasets/reyhanarighy/data-telco-customer-churn/.