Predicting Service Category of ITS Tickets

by Ethan March

Table of Contents

Introduction

The goal of this project was to create a model that can predict service category of tickets in the Northeastern ITS ticketing system. The motivation for this project was described in the original proposal and has been included below for reference.

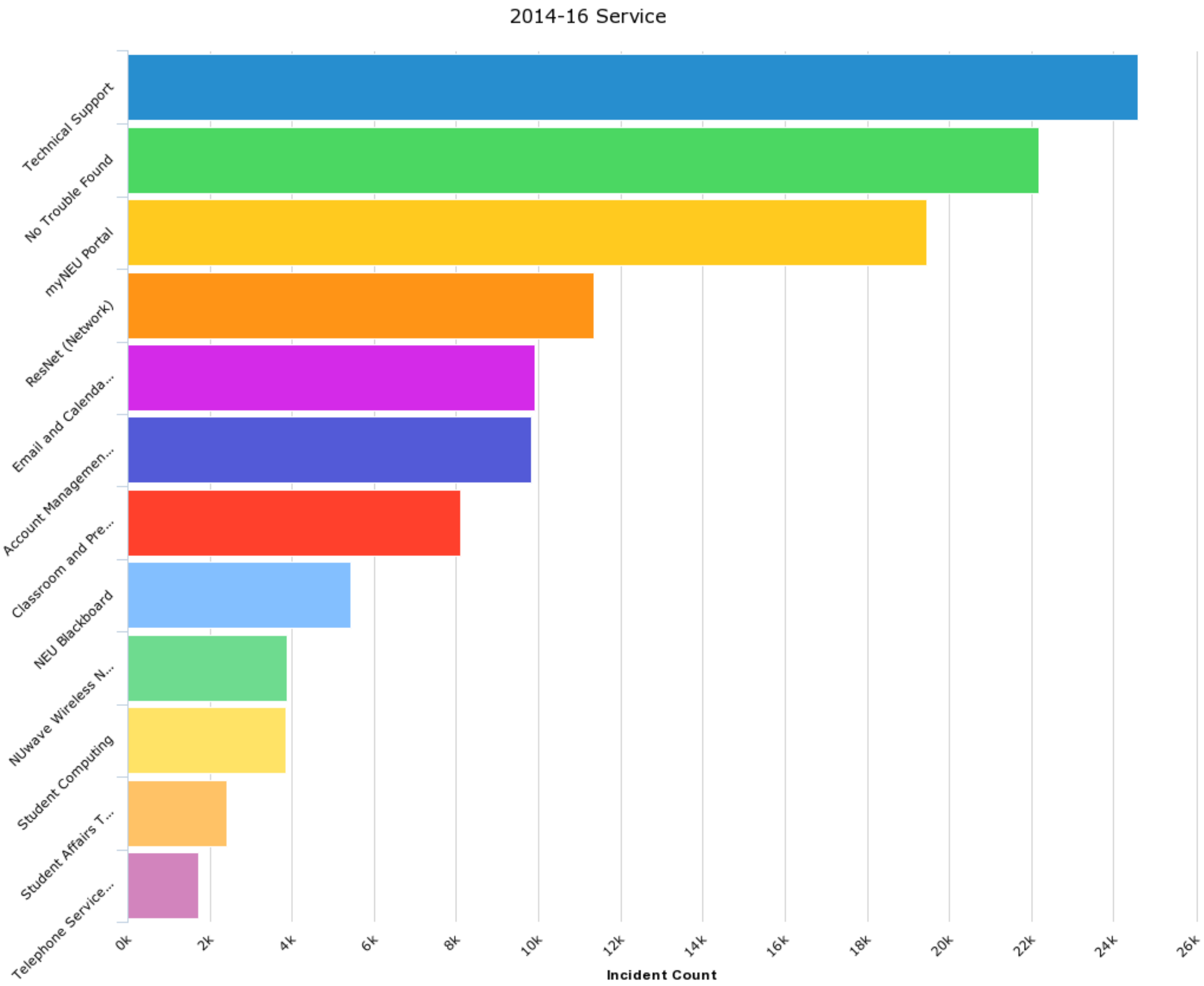
Working part time at ITS, one of the most common issues with our internal operations is the mischaracterization of tickets. This can often lead to problems being assigned to incorrect groups and frustration among the many different teams functioning under the ITS umbrella. A model that would be able to correctly predict the service category of tickets could eliminate the human element in ticket characterization which is where errors occur.

Data

The data for this project was taken from Northeastern ITS's ServiceNow instance. Through this online tool users can view and download old tickets. For my project, I used the "Reporting" feature of ServiceNow to export data from 2014-16. The tool only allowed for the export of up to 10,000 rows at a time so I broke the exports down into months.

Once the data had been exported to CSV files I wrote the *combine-data.R* script to import all the data into R and combine it all into one data frame. Once the data had been consolidated I renamed the columns of the data frame as the names from the ServiceNow exports were not intuitive. The chart below shows the twelve most common service categories and how many occurrences there were of each.

Finally, I established a connection to a MongoDB database and inserted the combined data into a collection using the *mongolite* library for R.



Model

In my original proposal, I stated that I would most likely be using a logistic regression model due to the categorical nature of the data and the type of problem I was trying to solve. A problem I ran into early in the construction of my model was the problem of dealing with text data. It seemed that the most important factor in determining service category was the written description from each ticket. Based on this, I chose to create a document classification model using a naïve Bayes classifier. The model takes “Description” as its sole feature and then classifies it into a service category.

The first step in creating the model was retrieving the data from the database. As part of this process, I only retrieved data for the tickets that had non-empty descriptions. I also created a data frame that contained only “Service” and “Description” as those were the only two columns that would be used for the model.

The data was split into training and test sets (70/30) by selecting a random sample from the total data set for the training set and the taking the rest of the data for the test set. After this division, samples of 10% of each of the training and test sets were taken to be used in the creation of the model. This was done to decrease the computation time of training and testing.

After training the model was used to predict the service categories of the test data. These predictions were then compared to the actual service categories and the percentage of correct predictions was calculated. Initially the model performed quite poorly, achieving an accuracy of only 1-3%. However, after adjusting the model to account for the presence and absence of words in descriptions instead of their frequencies, I created a model that consistently predicts around 48% of the service categories for the test set.

Conclusion

The final model that I ended up with can consistently predict the correct service category for around 48% of tickets. While this is far from perfect, it is a significant improvement from the initial model I developed and given the variation in ticket writing styles, I was surprised that even this level of accuracy was achieved. The main reason I chose to use a naïve Bayes model was the resources available for learning how to create one and the relative quickness with which they can be trained. Going forward I may attempt to adjust the type of model used to achieve more accurate predictions on this data.