NYC 311 - What are some of the factors that determine the number of days spent to solve non-emergency issues in New York City?

1. INTRODUCTION

3-1-1[1] is a telephone number used in the United States for citizens to get access to non-emergency municipal services. We used *NYC311* dataset to find some of the factors that affect the time spent on solving each issue that happened in New York City. What factors may result in an extremely long solving time?

The audience of this project could be the governor of New York. By analyzing the factors that have a significant impact on the length of time spent on each issue, the governor could create new policies in order to reduce the amount of time spent to solve each issue.

For example, `agency` would be an interesting variable to look at because different agencies might have different procedures in solving issues. Different procedures might result in different average of time spent in solving the issues.

I listed the top five agencies that spend on average the most days on solving an issue below:

	agency <fctr>	average_days_spent <dbl>
1	DOB	78.5801498
2	DPR	67.6886157
3	DOE	54.6442333
4	TLC	34.4797999
5	EDC	29.0380258

Table 1. The top five agencies that spent on average the most days on solving an issue

2. DATA

2.1 Data Collection

We chose to perform analysis on New York City 311 phone tracking system data collected from NYC OpenData. To extract this data, we used `nyc311`, a package that provides an interface to NYC 311 Phone Call information from NYC OpenData[2], to download data from the month we are interested in. There are three major problems occurred in the process of data cleaning.

2.2 Data Mining

2.2.1 Data Size

The very first difficulty that we encountered was to deal with size of our dataset. The original dataset for each month is around 150 MB each, and in order to get a meaningful result we wanted to study data from as many months as possible. In order to reduce the size of our datasets, we first chose the time frame of the data that we wanted to focus on. We decided to look at data that have been collected from January 2015 because it is the first month of a year. Then, we filtered out all the variables that were not essential to the question that we are interested in. We also got rid of some of the texts, since we did not cover how to deal with texts in this class. After these two steps, we were able to reduce the size of each month's dataset to about 50MB.

2.2.2 Data Type

The second issue we encountered in our project was to deal with incorrect data type. `created_date` and `resolution_action_updated_date` are two `character` variables. In order to calculate the number of days spent in solving an issue in New York City, we needed to transform `character` into `POSIXct` variable to make duration a numeric variable. In order to deal with this issue, we tried to use `as.POSIXct` function to transform the data type. However, because of the special way that the characters have been encoded, we had to use a combination of `gsub` and dplyr package to reorganize the characters before transforming the data types. We eventually were able to create four new variables, `time_spent`, `days_spent`, `hours_spent` and `long`. `long` is a binary categorical variable that returns `T` if it took more than 30 days to solve the issue and `F` if it took 30 days or less to solve the issue.

2.2.3 Categorical Variable Data Level

In order to build a tree in R, each categorical variable must have at most 32 levels. We used `dplyr` package to filter out the extra variable levels so that we can focus on the levels that occurred the most in the dataset.

3. RESULT

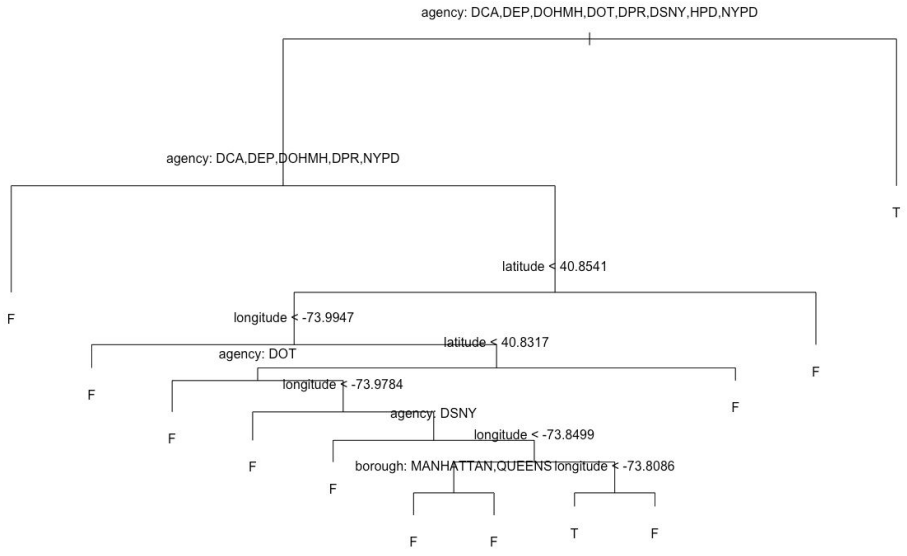

Figure 1. Tree model with 12 terminal nodes, and four variables: agency, latitude, longitude, and borough.

We started with a multiple regression model, but ended up with a tree model. We also tried to fit a logistic regression model that searches for a single linear decision boundary. However, we might have a non-linear decision boundary, so a decision tree, which partitions the nyc311 dataset into smaller subsets, fit our dataset better. Also, since most of our variables were

categorical, we decided to create a tree model to make predictions. In our tree, `agency`, the topmost decision node, is the best predictor.

4. CONCLUSION

The response variable is `long`, which gives True if the number of days spent on solving is greater than 30 days and otherwise, False. The first two branches are divided according to the agency. The agencies that had been categorized True for `long` came down to the right leaf, and otherwise, on the right leaf. The next few following branches were categorized according to the value of `longitude` and `latitude`. The specific value of `longitude` and `latitude` used to categorize the leaves were different for every branch. Lastly, `borough` was used as the last variable for the tree model.

The error rate of our tree model is about 12%. About 12% of the predictions were incorrectly classified while 88% of the predictions were correctly classified. This is a satisfactory error rate for a tree model. Through our model, we can conclude that `agency`, `latitude`, `longitude`, and `borough` were appropriate explanatory variables to use to create a model to make our predictions of `long`.

5. DISCUSSION & FUTURE WORK

There are many interesting patterns in NYC311 dataset. Steven Johnson[3] used the NYC311 dataset to write an article about the "new aroma" event which took place in New York City about 15 years ago. The city officials realized that the NYC311 phone call tracking system data actually provided clues about where the aroma was from. By analyzing the NYC311 data, we will be able to find more interesting "clues" about what is happening in New York City. For our future study, we would love to add more variables into our model and hopefully create a model with lower error rate. We would also love to analyze other variables in the NYC311 dataset in order to investigate “clues” or interesting event that happened in New York City.

6. REFERENCE

1. "3-1-1." *Wikipedia*. Wikimedia Foundation, n.d. Web. 09 Dec. 2016.

2. Baek, Jennifer. "NYC Open Data." *NYC Open Data*. The City of New York, n.d. Web. 09 Dec. 2016.

3. Johnson, Steven. "What a Hundred Million Calls to 311 Reveal About New York." *Wired*. Conde Nast, 1 Nov. 2010. Web. 09 Dec. 2016.