Capstone Project 2

Quality Measurement of NYC 311 Service Requests

Jeffrey Z. Ma
31 Dec 2019

Problem Statement

When people require non-emergency municipal services, the number to call is 311. In New York, all the data about 311 calls received in New York City is available online. A potential problem that can be addressed is the number of late incident resolutions. For each incident that gets reported, 311 creates a service request and notifies the appropriate agency with the request. A timeframe known as a service level agreement is created based on the type of request made, which informs the customer how much time it will take to respond to the request. Agencies are required to report their success resolving 311 service requests within the timeframe presented to customers.

In this project, a machine learning model will be built that will predict if a new call that is placed in NYC will be resolved on time or not. This information would be useful to various NYC public agencies in order to improve their performance and address the needs of the people in a more timely manner because the model will identify factors that contribute to tardiness of service request resolutions. This information can also help these agencies to make a better estimate of how long it will take to resolve various service requests.

Data Wrangling

The dataset that was used is the “NYC Open Data 311 Service Requests from 2010 to Present”. The data was acquired by using the Socrata Open Data API (SODA) which provides programmatic access to the dataset. The original dataset contains 22.1 million rows and 41 columns, a total of 906.1 million data points. To make this training project more manageable in Python as well as more relevant to today’s status, a 50,000 row subset of the data from January 1, 2019 to January 15, 2019 was taken. Because subsets of the data were taken as well as the feed being updated daily, different versions of the dataset could be saved based on the date of access. In the future, the models built on the different days of access could have their performance evaluated separately and compared.

Once the dataset was read into Python, cleaning and transformation of the data was required to prepare the dataset for analysis. With 41 columns in the original dataset, we need to review, analyze, and hopefully reduce the number of features so that the model can be built and interpreted properly.

First of all, columns containing redundant information were dropped. For instance, “longitude” and “latitude” describes the location of the incident, but so do the following columns: “intersection 1”, “intersection 2”, “cross street 1”, “cross street 2”, “location zip”, “location address”, “location”, “bridge highway name”, “road ramp”, “bridge highway direction”, “bridge highway segment”, “x coordinate”, “y coordinate”, “landmark”, “street name”, “incident zip”, and “incident address”.

Secondly, rows containing null values were dropped.

A new feature called ‘late’ was created by comparing the”closed_date” of a service request with its estimated “due_date”. For those rows where the closed date was greater than the due date, a label of “True” was applied. For those rows that failed the argument above, a “False” label was applied. These labels were then multiplied by “1” in order to get Boolean values of “0” or “1”, so that we could use the “late” feature as the target variable. It was found that approximately 16.36% of all the rows in the dataset were labeled as “late”.

To further clean the data, with the consideration that the day of the week the call was placed on could contain useful information, a variable “created_date” was converted to datetime and the day of the week extracted.

Additionally certain features contained levels not desired for predictive modeling. For instance, there were six unique values for boroughs in the dataset, when there are 5 boroughs in reality. Thus all rows where the borough was unspecified were dropped. For the “open data channel type” feature, rows that had other and unspecified responses were discarded as well.

After converting categorical variables to dummy variables through one hot encoding, a total of 850 columns were produced. To improve modeling efficiency and ensure the information is manageable, a mapping dictionary was created that transformed 67 different “complaint_types” into 8 groups. Those groups being: “water”, “plants”, “animals”, “cleaning”, “human complaint”, “maintenance”, “noise”, and “traffic”. The same method was applied to the “location_types” feature where 53 unique location types was converted into 3 groups of “business”, “residential”, and “public”.

Finally, since the variable late is a combination of the features “closed_date” and “due_date”, both those features were dropped from the dataframe. Thus the final variables selected to go into the predictive model were:

Agency: which New York City agency was assigned to take care of the incident
Borough: which New York City borough the incident took place in
Latitude: latitude of the address where the incident took place
Longitude: longitude of the address where the incident took place
Open_data_channel_type: how the incident was reported
Late: whether or not the incident was resolved on time or not
Created_day_of_week: what day of the week the incident was reported
Complaint Grouping: the 8 different types of complaints
Location_Grouping: the 3 different types of location groups

Exploratory Analysis
With data preprocessing finished, exploratory analysis was conducted to better understand the data. Some bar charts were created to look at counts of the data by feature variables in order to look at the spread of the data. Some examples are shown Figs. 1,2, 3, and 4 below:

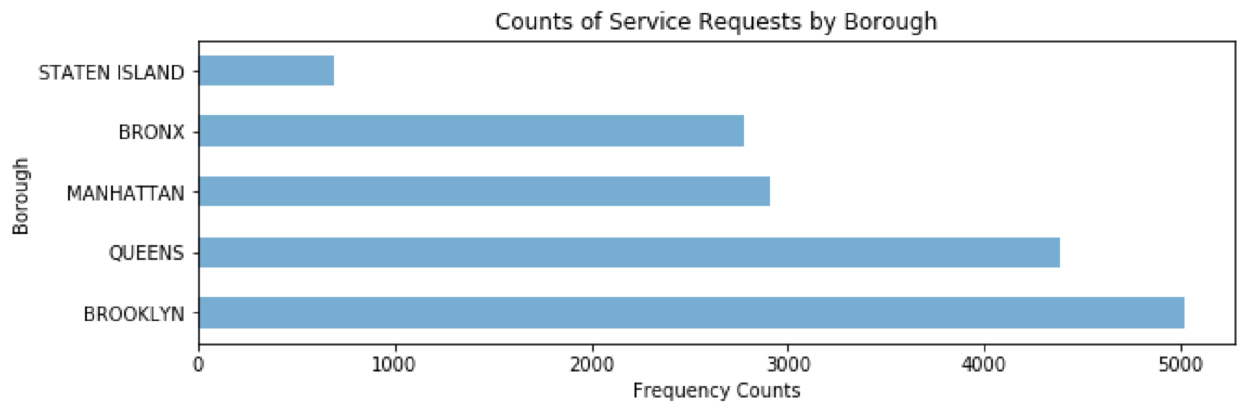

Fig 1. Counts of service requests by New York City Borough

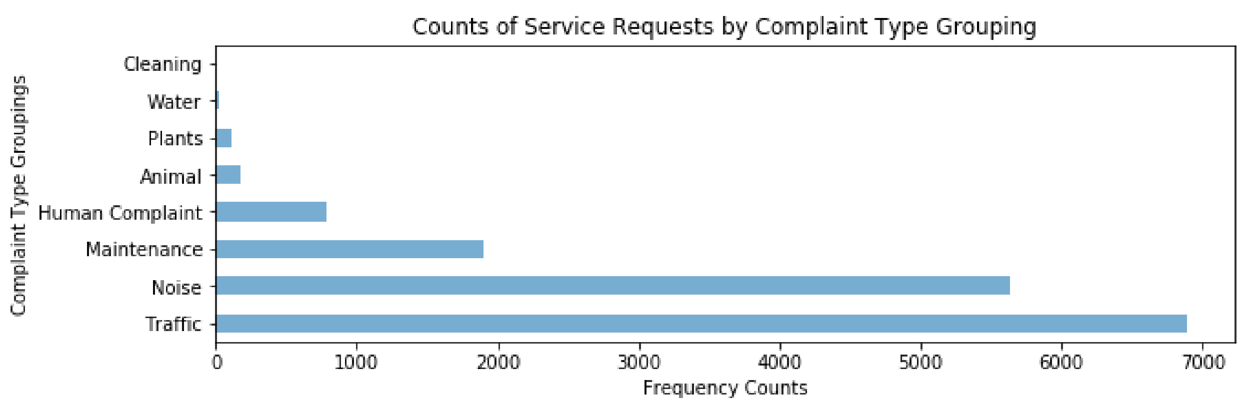

Fig 2. Counts of service requests by type of complaint

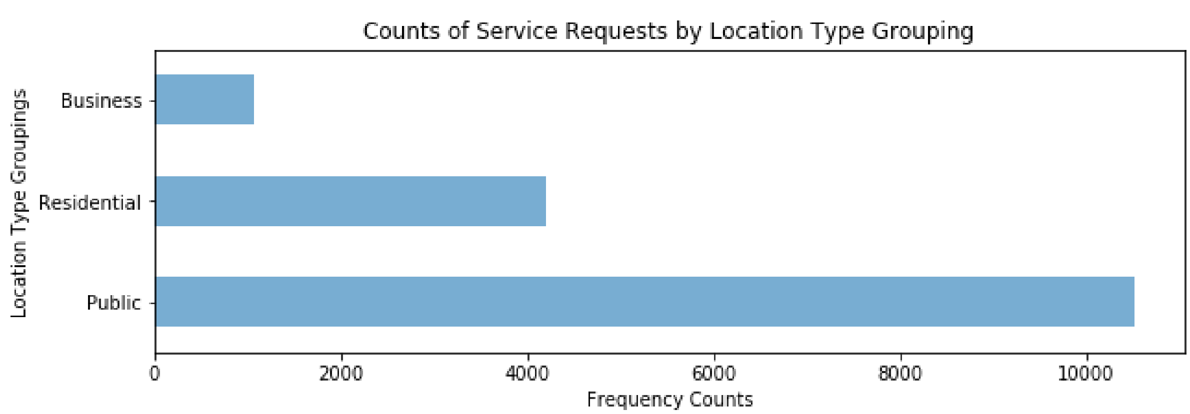

Fig 3. Counts of service requests by type of location

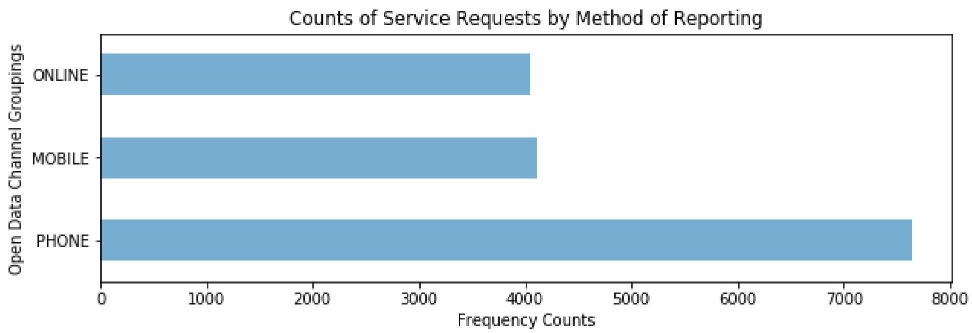

Fig 4. Counts of service requests by method of reporting

Dummy variables were then created to allow the use of machine learning algorithms from sci-kit learn. The dummy variables created were then placed in a correlation matrix to show correlation coefficients between variables as seen in Fig.5 below:

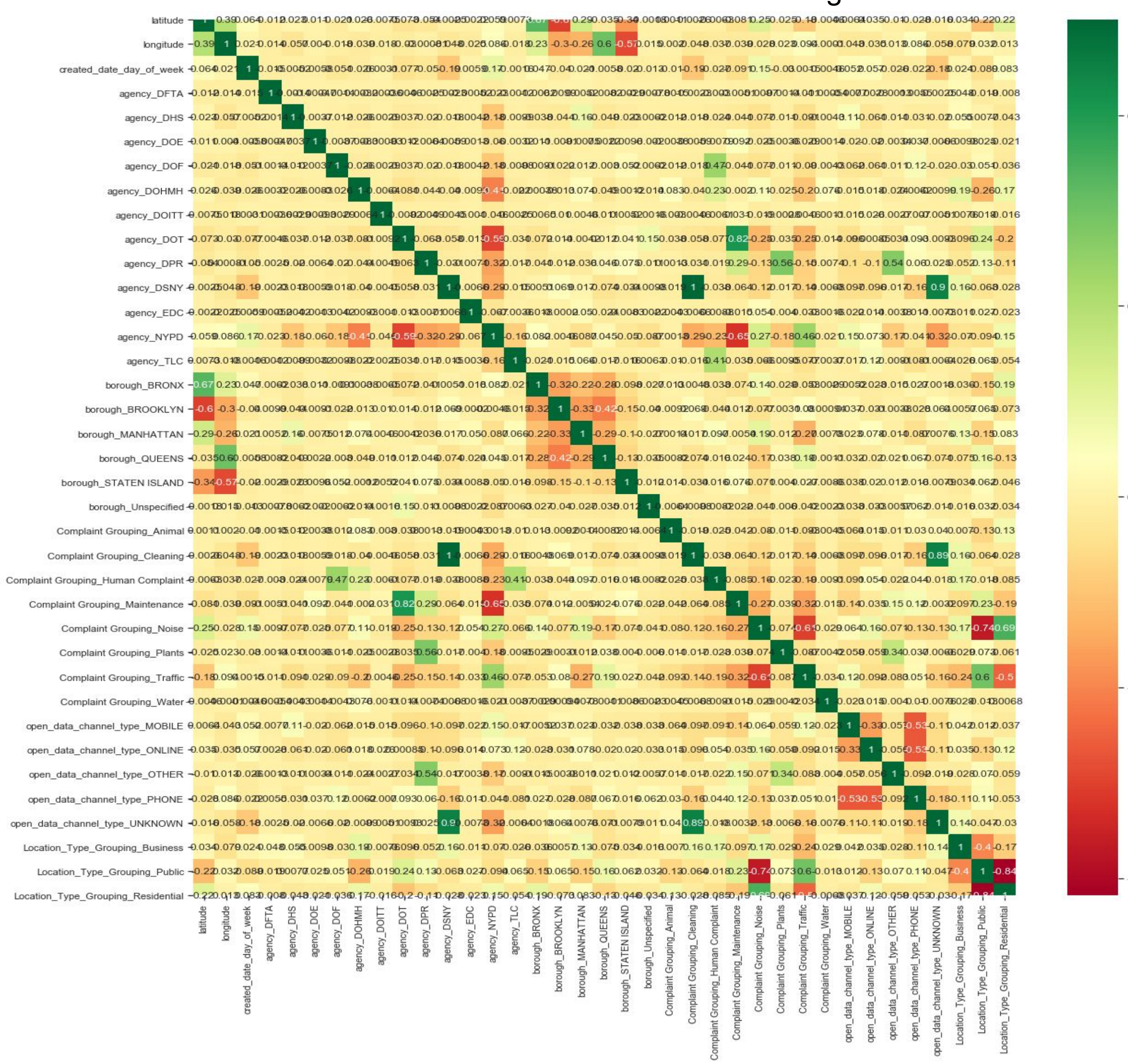

Fig 5. Correlation matrix showing correlations between variables

Although a few of the features seem to be correlated, overall there is not much correlation between the features, i.e., there are independent features. Now the dataset is ready for predictive modeling.

Predictive Modeling

The target variable selected was the 'late' column which indicates that an incident report was closed after the estimated due date. The features that will contribute to the model are the other columns of the dataframe. The first machine learning algorithm applied to the dataset was the random forest regressor. As this is a classification problem, this algorithm is applicable. After splitting the dataset into a training and test set, the random forest regressor was fit to the data to create a model. When the score of the model was reviewed, only about 27% of the variability in the data was explained by the model. This score is on the lower end, and implies the model is not strong.

A logistic regression model was then applied to the data, but this time we wanted to select for the more important features. Recursive feature elimination for logistic regression was conducted to select the features to build the model. Out of the original 37 features, recursive feature elimination recommended dropping 17 features to leave 20 columns to build the model with.

Using these 20 features, a logistic regression model was built on the training set. The performance of the model was then evaluated on the test set. Accuracy of the model improved to 87%. The confusion matrix of predicted versus actual values is as follows:

[[6726   56]
 [989 129]]

Next, an AUC (area under curve) - ROC (receiver operating characteristic) curve was graphed as shown in Fig. 6:

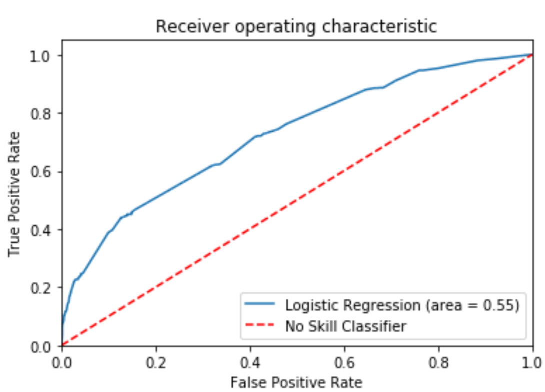

Fig.6 AUC -ROC curve: ROC is a probability curve and AUC represents degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The dotted line represents a no-skill classifier that cannot discriminate between the classes and predicts labels randomly or as a constant class.

The area under the curve was reported as 0.55, and the f1 score was 0.20. With a large AUC, the model is strong at predicting late calls as late.

Finally, a precision recall curve was graphed in comparison to a no-skill classifier showing the improvement in predictions the model made (Fig. 7).

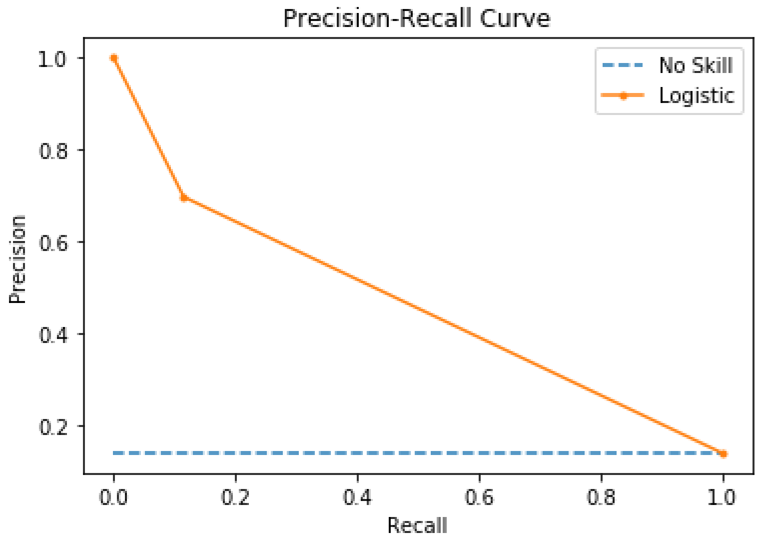

Fig. 7 Precision is the ratio of the number of true positives divided by the sum of the true positive and false positives. It describes how good a model is at predicting the positive class, in this case if a service request is resolved 'late' or not. Recall is sensitivity, or the number of true positives divided by the sum of the true positives and the false negatives. A precision-recall curve is a plot of the precision versus recall for different thresholds like the ROC curve, and again the dotted line represents a no-skill classifier that cannot discriminate between the classes and predicts labels randomly or as a constant class. For our dataset, the curve depicted above represents a skillful model because the curve bows towards (1,1) (a model with perfect skill) above the flatline of no skill.

After evaluating the performance of the model, a table was constructed containing the coefficients and odds ratios of the features as shown in Fig 8 below.

features	coefficients	odds ratio
object	float64	float64
agency_DFTA	-0.4890501427947647	0.6132085782311333
agency_DHS	0.7905214250876688	2.204545632016986
agency_DOE	0.7771162692006365	2.175190548481822
agency_DOHMH	-0.6124181891947023	0.5420385312540468
agency_DOITT	-0.1976192736550583	0.8206822490184139
agency_DOT	-1.1494340475693037	0.3168160214476185
agency_EDC	1.6669727864811212	5.296111046878813
agency_NYPD	-0.8540228941578774	0.42569894085552573
agency_TLC	2.1387588088341096	8.488894746206027
borough_BRONX	0.6556384958614397	1.926372105408051
borough_BROOKLYN	-0.4115813203032325	0.6626016358375854
borough_MANHATTAN	-0.30382587293139574	0.7379893591870111
borough_STATEN ISLAND	0.4763546105937086	1.6101939065477158
Complaint Grouping_Animal	-0.2509088880773264	0.7780932619025418
Complaint Grouping_Human Complaint	0.9912242857268226	2.6945313300373996
Complaint Grouping_Maintenance	-0.4340824675104814	0.6478588259777599
Complaint Grouping_Noise	-0.7471392687660418	0.4737198012095078
Complaint Grouping_Plants	0.3615000433971755	1.4354810842801455
Complaint Grouping_Water	0.31667372023654555	1.3725546620470068
Location_Type_Grouping_Business	0.7035743904383929	2.0209635253553206

Fig. 8 Table showing the coefficients and odds ratios of the features of the dataframe after running the logistic regression model. The acronyms of the agencies are defined in Appendix A.

Conclusion and Recommendations

Based on this study, the following are concluded:

- Based on the dataset used, approximately 16% of all incidents called into New York City 311 are resolved after the estimated time of resolution, a key performance index (KPI) of 84% with 100% as the perfect target

- The model built on the training set had 87% accuracy in predicting the correct labels on the true set, as can be seen from the confusion matrix.

[[6726   56]
[989 129]]

- Again with this dataset, incidents reported in Queens had little effect on the outcome of the prediction, but incidents reported in the other boroughs of Bronx, Brooklyn, Manhattan, and Staten Island did with 0 .66, -0.41, -0.30, and 0.48 coefficients respectively.
- Using the exponential function on the coefficients, we can compare the increase in probability of lateness for the affected boroughs as well. For the Bronx: EXP(0.66)= 1.93, which implies that a call originating in Staten Island has a 193% increased likelihood of being labeled as late according to the model, if all other features are held constant. The same method can be applied to the Brooklyn, Manhattan, and Staten Island boroughs.
- For Brooklyn: EXP(-0.41)= 0.66, implying that there is a 34% decrease in likelihood of being labeled as late according to the model if other features are held constant.
- For Manhattan: EXP(-0.30)= 0.74, implying that there is a 26% decrease in likelihood of being labeled as late according to the model if other features are held constant.
- For Staten Island: EXP(0.48)= 1.61,implying that there is a 161% increase in likelihood of being labeled as late according to the model if other features are held constant.
- From the model it appears that service requests originating in Brooklyn and New York are handled in a more timely manner than requests originating in Staten Island and the Bronx. Perhaps more effort needs to be made to address issues in Staten Island and the Bronx.
- Calls handled by the NYPD had a 57% decrease in the likelihood of being labeled as late, EXP(-0.85)=0.43, demonstrating excellence performance for the NYPD!
- The largest odds ratio seen belonged to calls handled by the Taxi and Limousine Commission with a 849% increase in probability of a call being labeled as late according to the model, EXP(2.14)=8.49. This seems like an outlier, but it is possible that the agency is understaffed and requires more help to fulfill its responsibilities, or the methodology of estimating incident resolving times needs to be revisited and improved.
- The feature with the lowest odds ratio belong to calls handled by the New York Department of Transportation. These cases were 68% less likely to result in a label of “late” according to the model, EXP(-1.15)=0.32. With 5,060 employees, this agency seems to be addressing the concerns of the public well.
- Although discarding 17 of 37 features during recursive feature elimination while building the model seems excessive; when the model is run on all 37 features together, no change in the performance of the model is seen.

Acknowledgements:

- I would like to thank my mentor Nathan Sutton for all his help and advice throughout the Springboard program, especially the capstone projects.
- I would also like to thank Springboard management and subject matter experts for their help.

Appendix A: Acronyms of NYC Agencies dealing with Service Requests

'agency_DFTA' = Department for the Aging
'agency_DHS' = Department of Homeless Services
'agency_DOE' = Department of Education
'agency_DOF' = Department of Finance
'agency_DOHMH' = Department of Health and Mental Hygiene
'agency_DOITT' = Department of Information Technology and Telecommunications
'agency_DOT' Department of Transportation
'agency_DPR' = Department of Parks and Recreation
'agency_EDC' = Economic Development Corporation
'agency_NYPD' = New York Police Department
'agency_TLC' = Taxi and Limousine Commission