CS-GY6053 Foundations of Data Science

PROJECT REPORT

Exploring 311 Service Request Data

Submitted by
Ayush Sethi(as11500)
Arnav Shah(ads798)

Problem Statement and Background:

Background:

311 Service Requests encompass all non-emergency requests from the city, including but not limited to noise complaints, air quality issues and reports of unsanitary conditions etc. The 311 calls in New York City (NYC) are publicly available.

311 service request dataset is available at https://data.cityofnewyork.us/Social-Services/311- Service-Requests-from-2010-to-Present/erm2-nwe9
This dataset comprises of all calls made to 311 from the year 2010-Present. The data contains more than 16M rows spread across 53 features. Size of the data is approximately 8 Gb.

53 features of the dataset include features related to
Time such as Created Date, Closed Date, Due Date, and Resolution Action Updated Date
Location specific such as Incident Zip, Incident Address, X Coordinate (State Plane), and Y Coordinate (State Plane)
Type such as Complaint Type, Agency and Descriptor
Then there are other features which are there to support specific types of requests.

Problem Statements and Motivation:

Analysis of 311 calls can be of great use for a wide variety of purposes, ranging from a rich understanding of the status of a city to the effectiveness of the government services in addressing such calls.

In our analysis, we want to answer following questions

1. What are different type of Service Requests(SRs)? Which is most/least frequent?
2. From which borough most SRs come from?
3. Which SRs peaks at what time of year or time of day?
4. How air quality issues relate to different boroughs?
5. The agencies which are more efficient in solving SRs.
6. From which type of location we get most number of complaints?

These answers will give us a better understanding of the city's issue.
Next thing we want to find out or predict is:

1. Predict time required in terms of range of days to resolve a specific complaint in a specific borough.
2. A time series analysis to forecast the volume of calls to be expected on given future date.
3. Merge the 311 dataset with the Storms dataset. Compare the average response time for complaints during a storm and otherwise.

4. To do a time series analysis of the storm data to find out the weekly or daily seasonality.

By using these answers, a city can be better prepared for a particular storm type. Policy makers can use this information to efficiently allocate resources. And the residents of the city can have a real-time sense of when their problem will be solved.

The storm events dataset is available at

https://www.ncdc.noaa.gov/stormevents/choosedates.jsp?statefips=36%2CNEW+YORK

This dataset contains features such as Location, County, Date, Type, Magnitude and few features telling about the damage caused by the event.

Methods:

For predicting time to resolve:

This is a supervised learning, classification problem. The model we build for this problem contains following:
Predictor Variables:

1. Day of Week
2. Day of Month
3. Month
4. Incident Zip
5. Descriptor

The first 3 features are stripped from the Created Date column of the 311 dataset. Descriptor column contains text data so converted each unique text value into a feature for model.

Target Variable:

Resolution Time: - First calculated the resolution time in terms of days by subtracting created date from closed date. Then divided the days into buckets such as bucket 1 for less than 2 days, bucket 2 for 2 to 6 days and bucket 3 for more than a week.

Hence our target variable represents 3 classes and we aim to classify input data into one of these classes.

So, for this problem we used classification models: Logistic Regression, Decision Tree Classifier, Random Forest Classifier.

Evaluation Metric: For evaluation of the above models we used confusion matrix**.**

For Time Series Analysis to forecast volume of calls:

For doing the time series analysis, we used Created date column and the count of complaints on each date. We used two models: traditional ARMA model and Prophet model.

Prophet is a library to build forecasting models for time series data, but instead of using the traditional way of building the model such as using ARIMA, etc., it is fitting additive regression models or known as ‘curve fitting’.

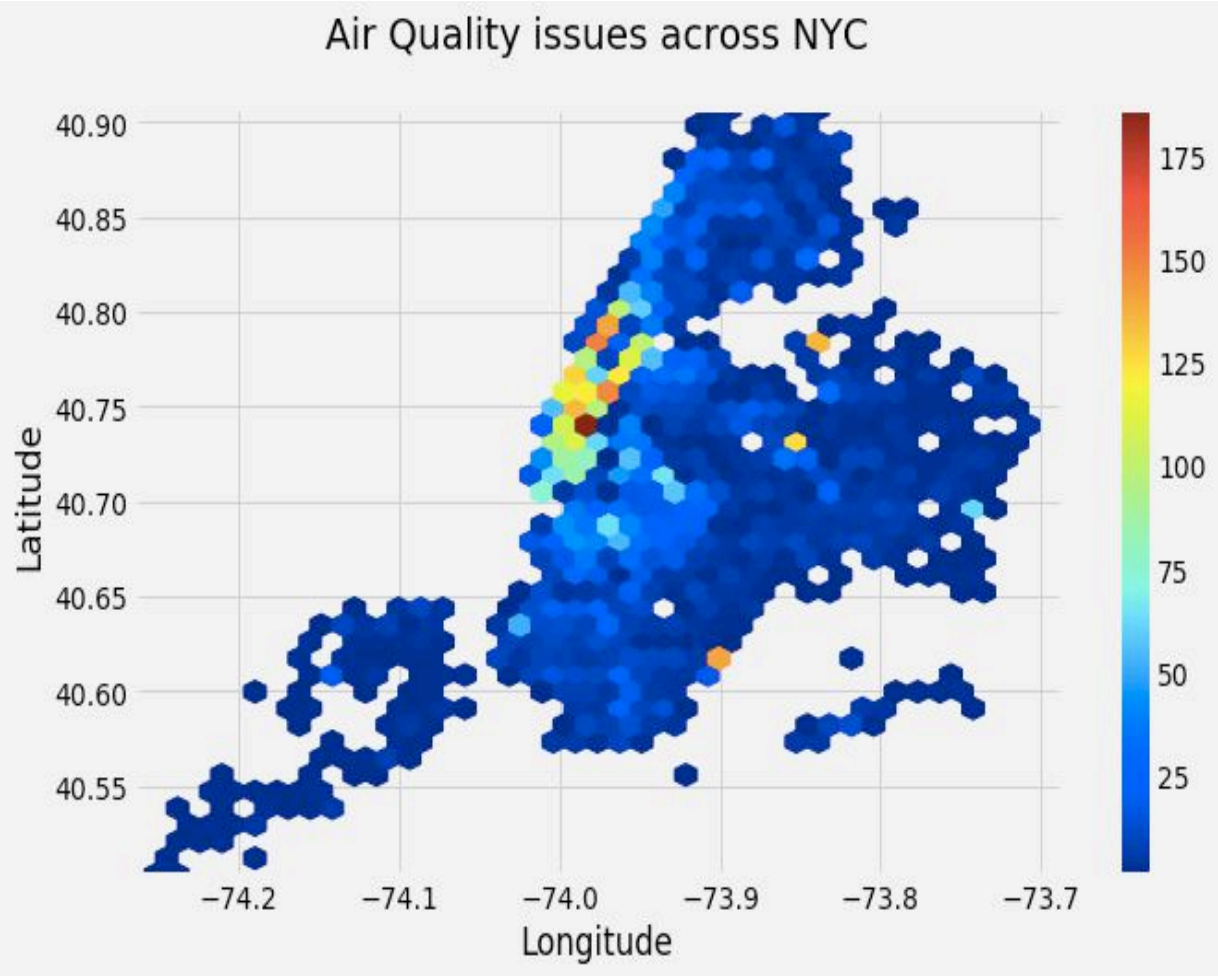
Evaluation Metric: For evaluation of ARMA model, we calculated MAE and MFE values and also the Durbin Watson Statistic. For evaluation of Prophet model, we did the goodness of fit test and calculated r2_score.

For merging Storms dataset:

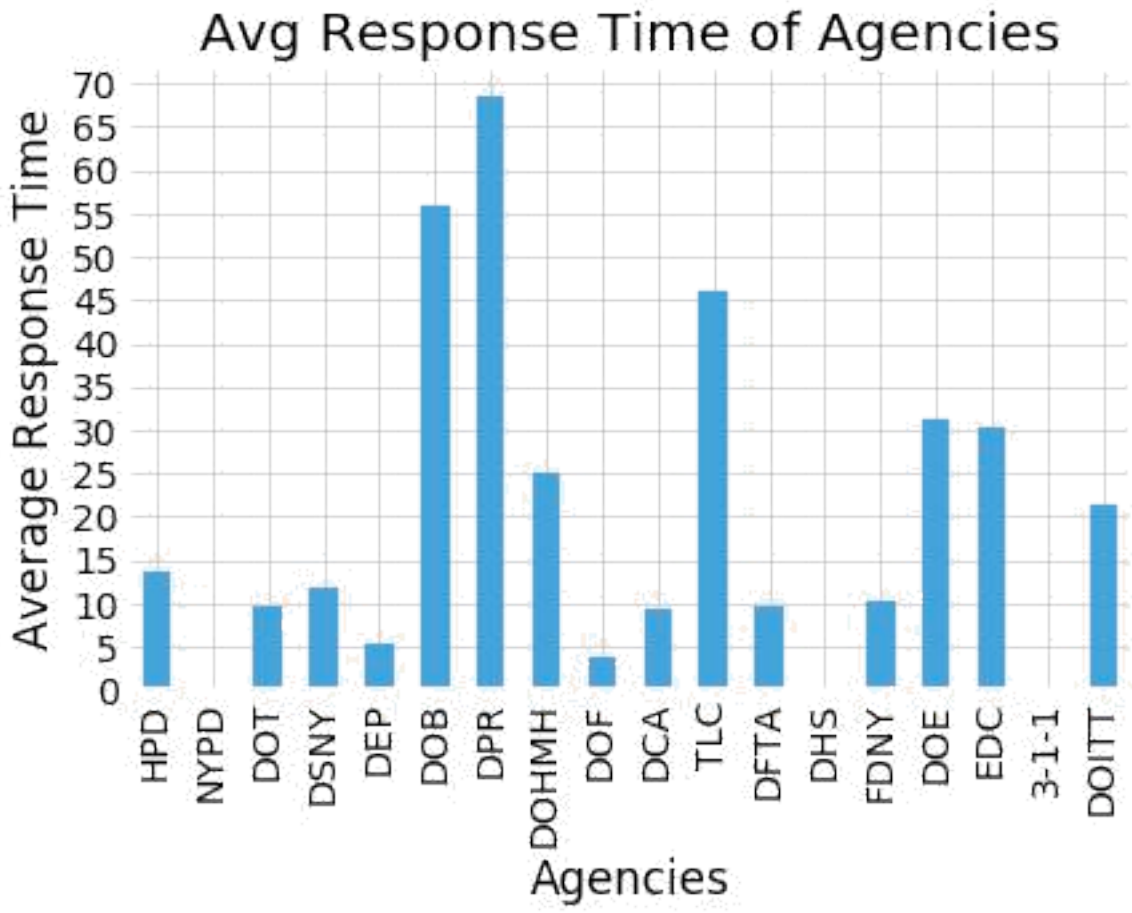
We performed an inner join on Created date column between storms dataset and 311 dataset. Performed some visual exploration and calculated weekly and daily seasonality using the Prophet model.

Results:

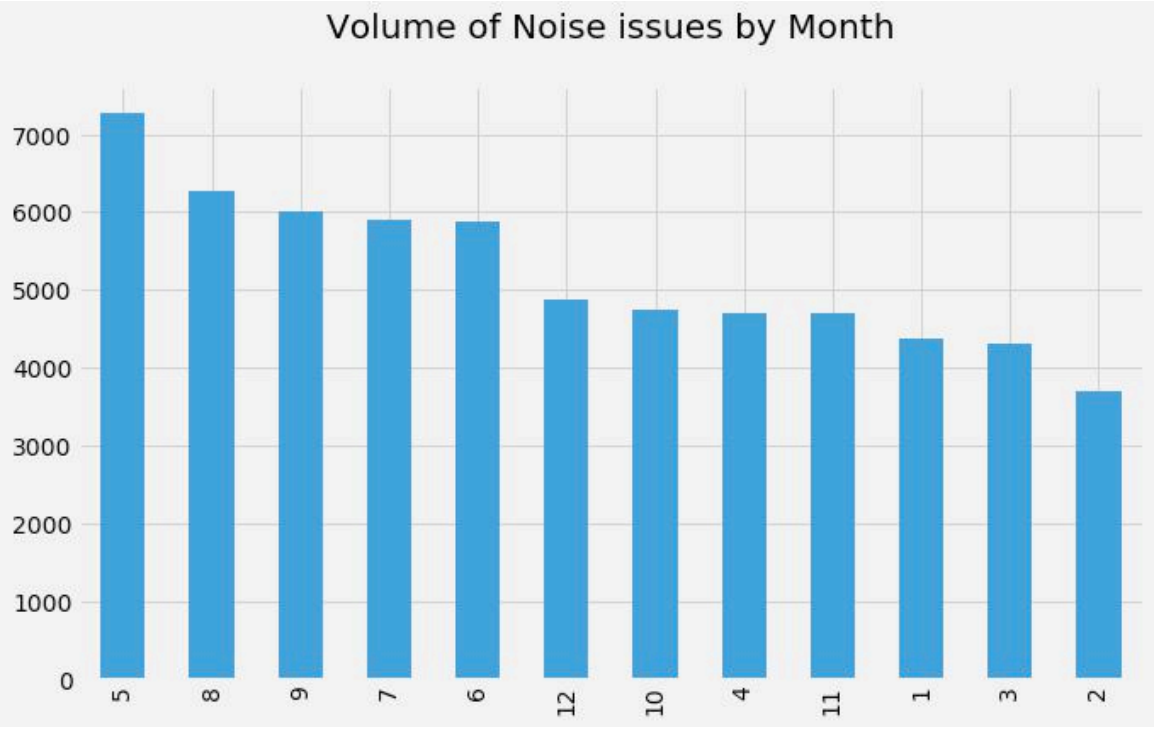
Descriptive Answers:



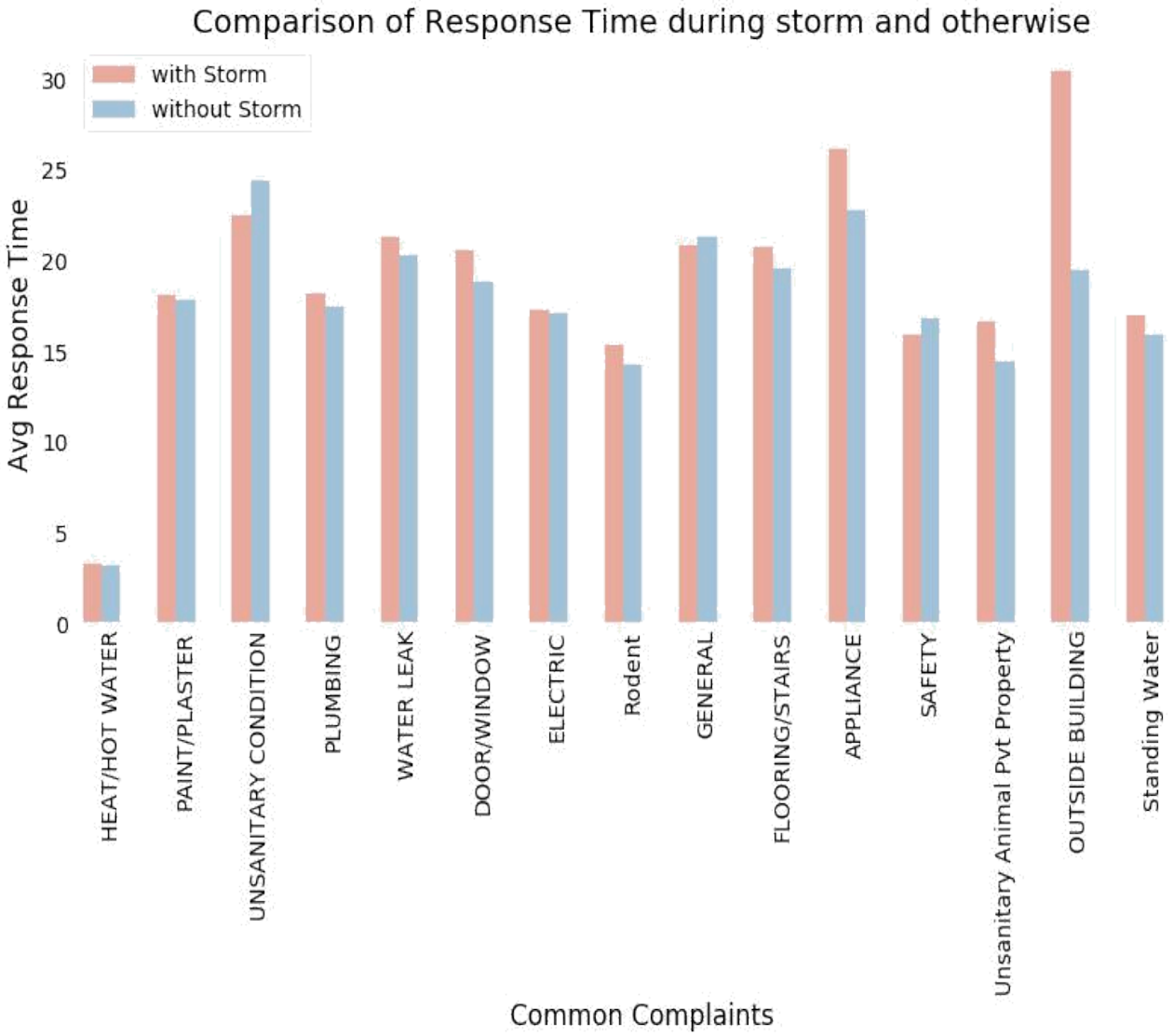
We can see that most number of air quality issues comes from Lower Manhattan.



We can see here that NYPD, DHS and 3-1-1 are the most efficient agencies among the rest. While DPR take on an average more than 2 months to solve a complaint.



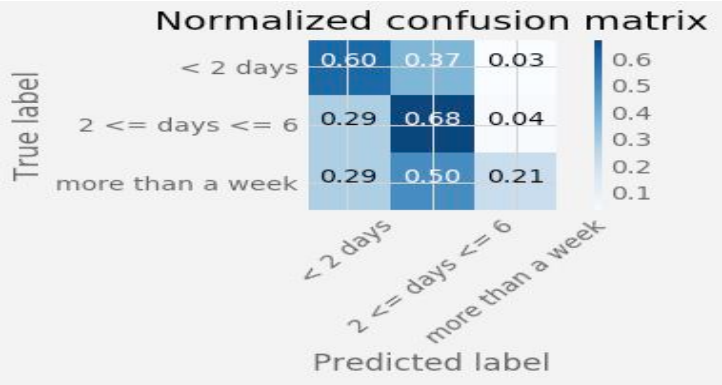
Noise Issues are most common in the summers.



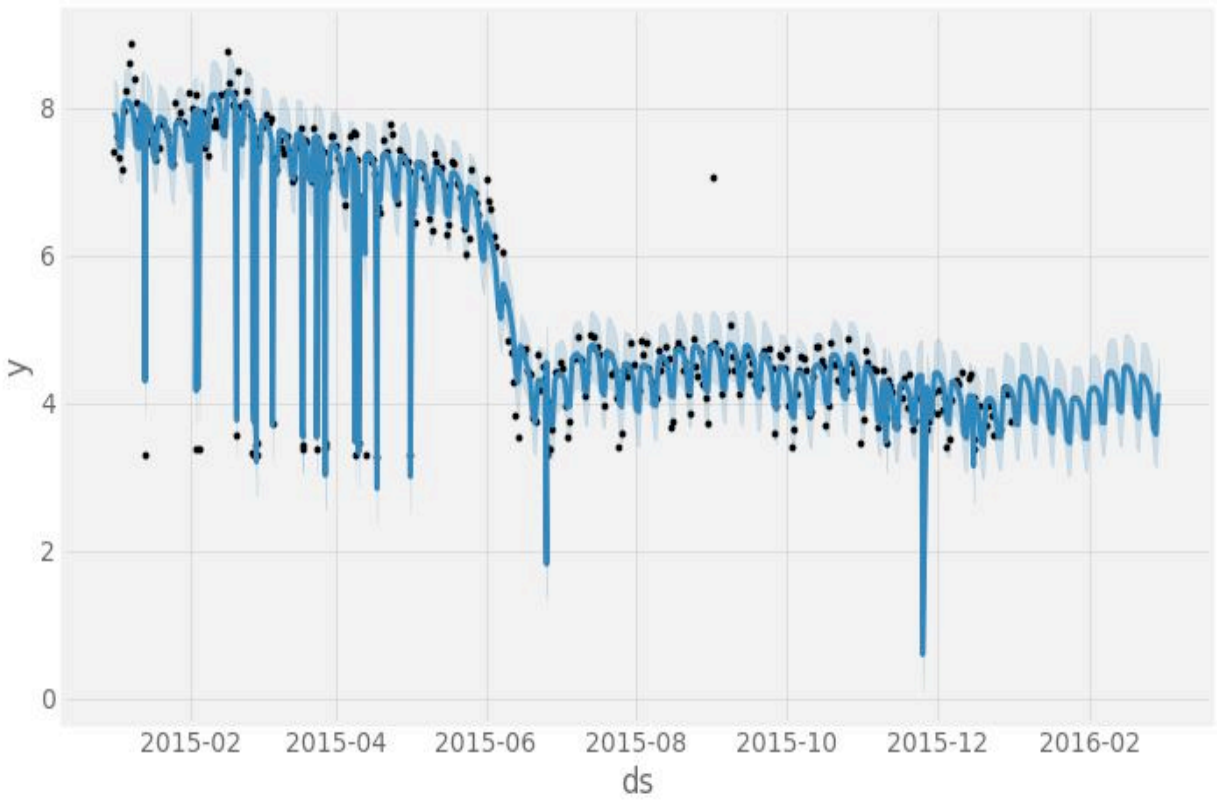
For most of the complaints, it's the expected result i.e. during storm, they take more time to resolve, but in case of unsanitary conditions, it's the reverse. Also heat issues takes almost same time in both the cases.

Models:

Predictive model for resolution time yielded an accuracy of around 62% with both Decision Tree Classifier and random forest. Confusion matrix highlights the problem with the model. The data is imbalanced. We do not have enough data points for class 3.



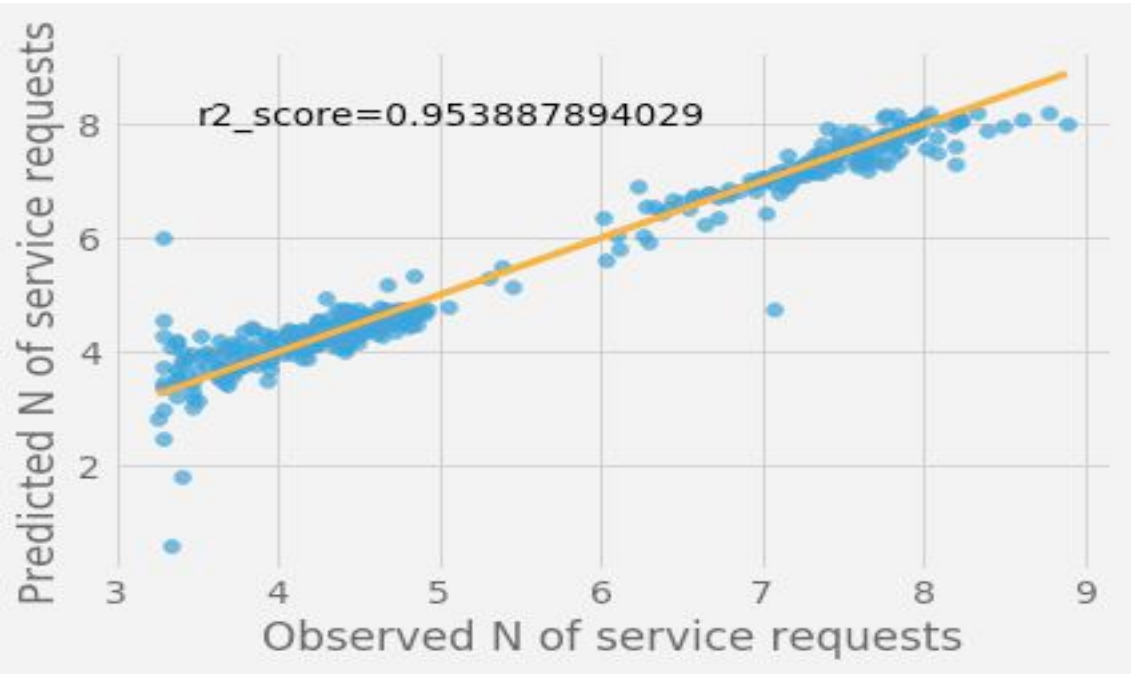
Time Series analysis to forecast volume of calls for future:



This plot is drawing the original data (black dots), the model (blue line) and the error of the forecast (shaded blue area).

We fitted ARMA(1,0) model with an MAE of 0.66 and Durbin Watson Statistic of 0.0037.

We were able to fit Prophet model with an r2_score of 0.95.



Assumptions/Limitations:

1. Dataset was very large and it was difficult to process, so filtered on many levels to reduce the data size.
2. Assumed that all the complaints were in closed state.
3. Assumed that impact of storm lasted only till the duration of storm.

Problem in Scope of Class:

The problems proposed are in the scope of the class. In addition to descriptive exploration, performed the predictive analysis for resolution time and time series analysis for forecasting the volume of complaints. Also calculated the yearly, weekly and daily seasonality of the time series data using Prophet model. Used Random Forest and Decision Tree classifiers for classification. Used appropriate evaluation metric for each model.

Changes from original proposal:

1. In the original proposal, we proposed to do a regression analysis for predicting the number of days required to resolve a complaint. But now shifted to doing classification analysis because of the large range of days in target variable.
2. In the original proposal, we gave a rather broad question of studying the impact of storms on 311 calls. Now we are able to pinpoint the questions we wanted to answer keeping in sight of our limitations.