NYC 311 CAPSTONE REPORT

Stephen Behunin | BrainStation

stephenbehuninwork@gmail.com | linkedin.com/in/stephen-behunin

Problem Statement and Background

On average the NYC will receive 2.1 million Service Requests (SRs) through its 311-system each year. Everything from potholes to noise complaints flows into the queue of pending SRs. The city must process and resolve every SR it receives. This is in addition to dealing with emergency situations and regular maintenance/operations for the city. Naturally this means that extraordinary strain is placed on city services which are frequently stretched to the limit and need to utilize every ounce of manpower and funding available to them.

One of the fundamental challenges for successful resource allocation at this scale is consistency. Having appropriate resources available for departments to handle the flow of complaints is necessary to keep the system running smoothly, the bosses happy and the citizen's content. Unfortunately, the volume of requests received each day is far from consistent. During its slowest day on record the 311-system received 1680 SRs, which for a city of of 8.4 million seems like a reasonable figure. But on its busiest day the system received 11,735 requests. Astute observers may notice the slight 7-fold difference in daily SR volume experienced by this system. Which leads to an interesting question: *how do you effectively allocate resources, staffing and operating budgets when you may be dealing with 2,000 complaints one day and 11,000 the next?*

Business Question

The core of this project revolves on mitigating uncertainty in SR volume. While randomness is inherent to any system so complex and human dependent as city services there are still patterns within the usage of the New York 311 system. The amount of uncertainty surrounding SR volume can be dramatically reduced through understanding these patterns and using them to forecast future SR volume. Reducing the uncertainty of these numbers will help NYC better allocate its resources and effectively serve the community.

Goals of the Analysis

1. Understand the long term trends, seasonality and cycles present within the data.

- Perform EDA and timeseries decomposition to understand these components.

2. Effectively model future SR volume on a long timeframe.

- Predict SR volume by day up to a year in advance through SARIMA modeling

- This model can be used to assist in longer term budgeting and planning by providing reasonable projections of future SR volume.

3. Create a short-term forecasting tool to accurately predict SR Volume3-4 weeks ahead.

- Apply Recurrent Neural Networks a more powerful and accurate, but nearsighted, forecast of SR volume.

- This model can be used to make short term staffing decisions such as employee scheduling. And assist in effective resource allocation in a more granular but shorter time scale than the previous model.

The Data

Data for this project was retrieved from the NYC Open Data portal which contains all the publicly available data for the City of New York. The data was in .csv format and contained approximately 26 million rows and 41 columns. Each row contained a single Service Request and all the accompanying data necessary to address the complaint.

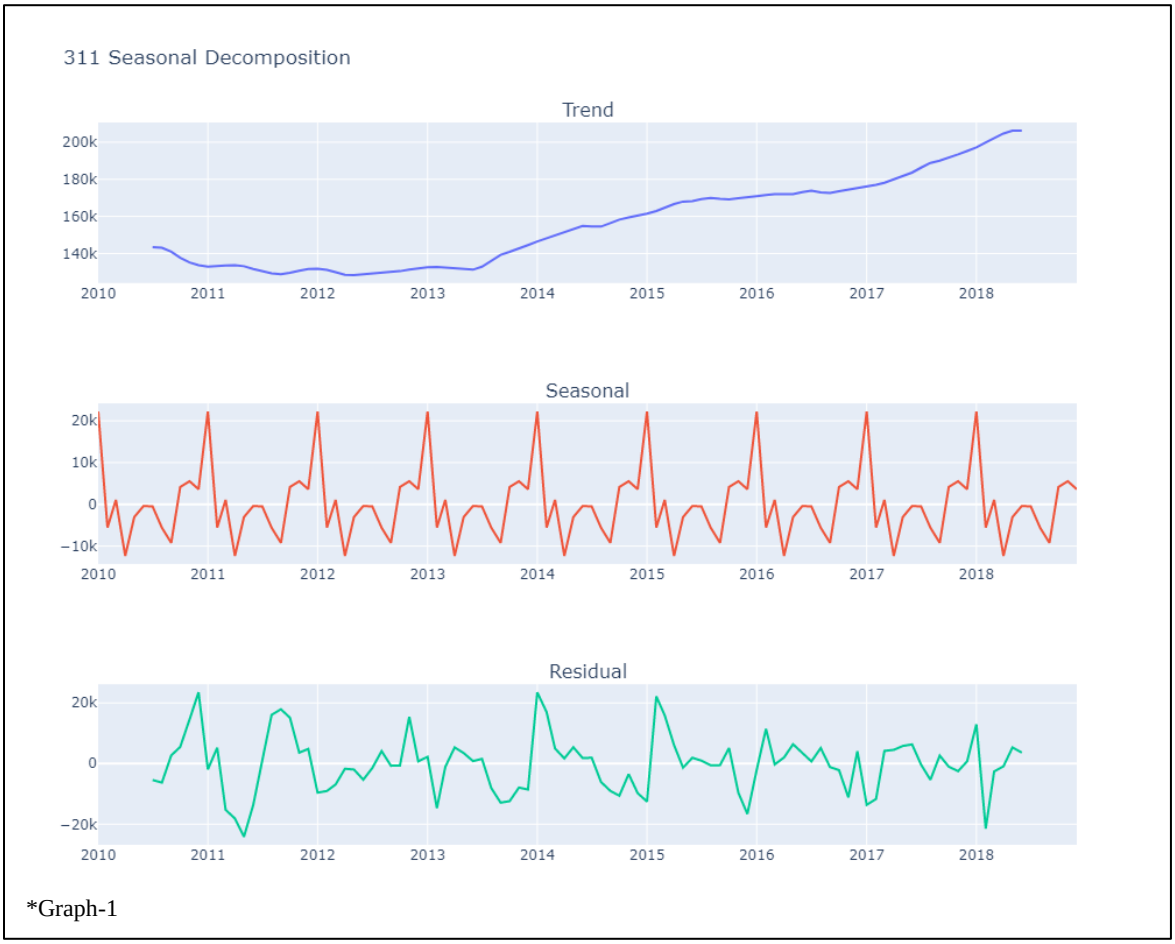
After cleaning the complaint dataset by removing rows with missing information, duplicates, and unnecessary columns the complaint dataset was reduced to 8 columns. The row count was further restricted when the dataset was limited to only SRs filed in 2010-2019. This decision was made to due to data collection issues related to the COVID-19 pandemic which effected the accuracy of information gathered in 2020 and 2021. The final row count for the complaint dataset was approximately 19.2 million rows.

The final step in the cleaning was converting the data into a timeseries format. Predicting SR volume falls under timeseries analysis as the target variable SR volume is measured against itself, with time acting as the independent variable. To put the data into timeseries format the complaints were grouped by their creation date and the total number of complaints filed on each day became the new data points. This conversion resulted in a data frame with 3652 rows and 2 columns. The dataset covered daily data for a full 10 years of the 311 system. In the RNN notebook an additional transformation was applied to the timestamp column that created two more columns with seasonality information encoded through sine and cosine. Which brought the timeseries dataset up to 4 columns.

EDA

The EDA process for this project uncovered several key insights. The most illustrative exercise in the process was performing trend-seasonal decomposition. This process breaks the data into three distinct

lines that represent different aspects of the data's variance. Trend: this data displays an initial plateau in SR volume followed by a steady increase until the end of the collected data.



*Graph-1

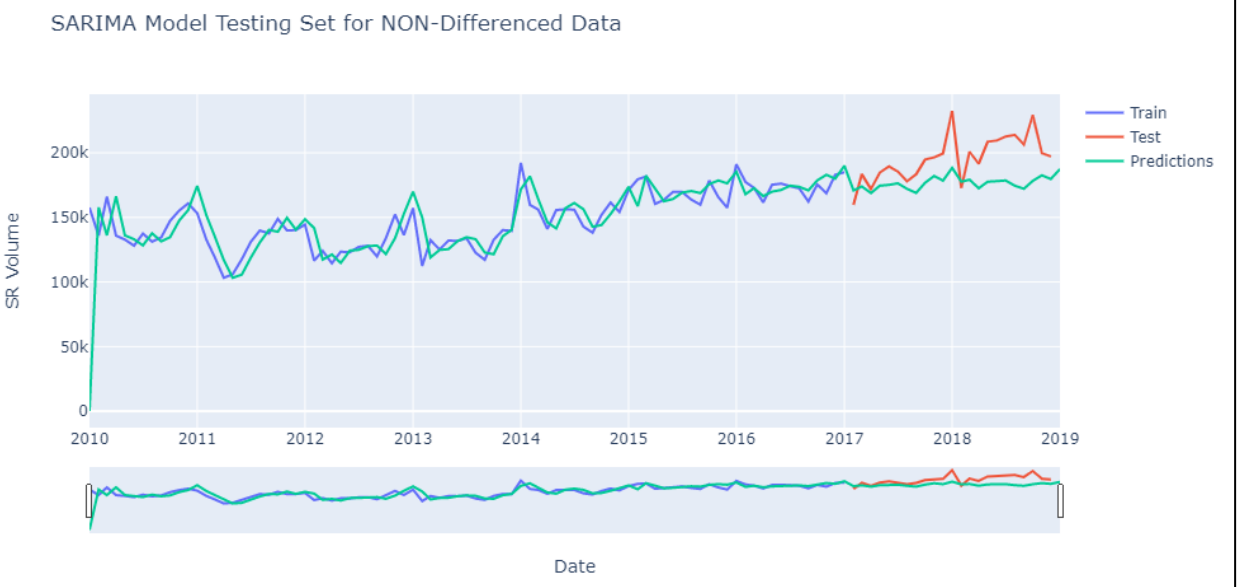
Seasonal: the dataset displays heavy seasonality with a large spike occurring in January every year, while the lowest points of the year are late spring and early fall. (Graph-1)

Residual: the residual line illustrates the remaining variance in the data not explained by the two previous lines. Some of this variance is random but a significant amount was successfully modeled later on.

Another important discovery in the EDA phase was the mechanism behind the large spike in SR volume occurring every January. Closer examination of the distribution of complaint descriptors revealed that heat related complaints surge every January and drive a much higher SR volume for that month.

Long-Term Modeling with SARIMA

The approach taken for long-term modeling was a progression of SARIMA style models. SARIMA models use trend, seasonality, and moving averages to forecast values in timeseries data. Initially a baseline was established with a mean prediction to compare against progressively more complex models.



*Graph-2

Accuracy for the models peaked with SARIMA modeling. (Graph-2) A parameter search was performed to find the final settings of the SARIMA model which resulted in a testing MAPE score of 10%compared to a baseline of 28% . This nearly 3-fold improvement meant that the average prediction made by the model was off by only 10%, a good margin for values predicted a year into the future.

Such results meet the requirements laid out in the goals of the project by providing accurate long-term forecasting of SR volume. This kind of information will allow better budgeting and resource allocation on a broad scale for the 311 system.

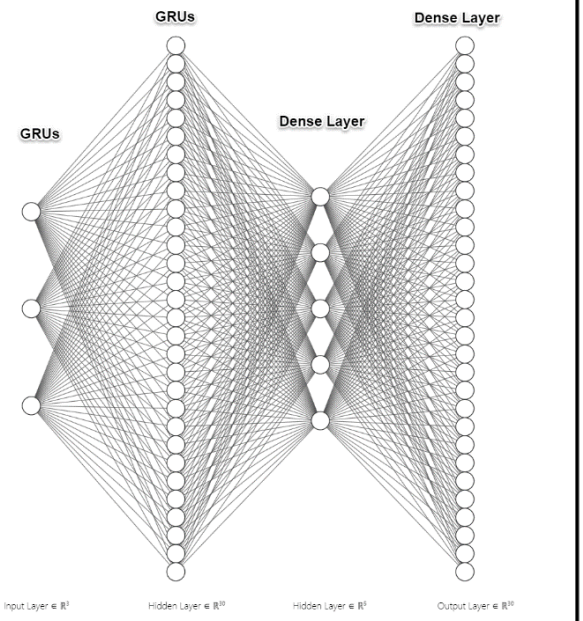
Short-Term Modeling with RNNs

To tackle the short term, 1 month, prediction task this analysis employed Recurrent Neural Networks. RNNs are a type of Neural Network that creates hidden states that are fed as inputs to the network in an effort to reduce “amnesia” within the models. This structure allows for RNNs to remember information about long sequences of data that would likely be lost in other types of Neural Networks.

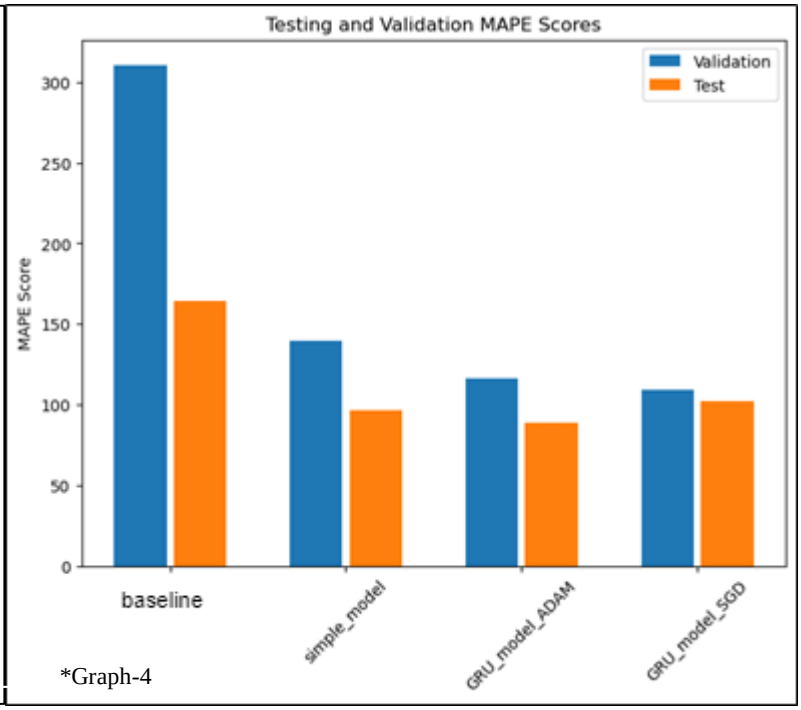
The progression path for RNN implementation was similar to the SARIMA implementation. Complexity and subsequently efficacy increased with each type of model. In the end three models were chosen for implementation along with a baseline:

- Baseline- used the input data directly as predictions.

- Simple RNN – a (1x1x1) RNN composed of simple RNN cells.



*Graph-3 **GRU Structure**



*Graph-4

- GRU with Adam – a (3x30x5x30) RNN with GRU cells and dense layers. (Graph-3)

- GRU with SGD – an identically structured RNN with SGD instead of Adam as the optimizer.

The final results of the models displayed in Graph-4 show a large difference between the baseline and all of the models for both testing and validation. Differences between the models were less dramatic with the Simple Model coming surprisingly close to the more complex models. In the end the GRU with Adam was deemed the best model as it had stable performance and the best testing scores.

Results for these models are harder to interpret than with the SARIMA models. This was due to background processes for the neural networks requiring normalization that distorts the scale of the data. But the MAPE scores of the models show that short-term forecasting as described in the goals of the analysis is fall within the model's reach. Additional work will be required to transform any predictions back into useable values but the underlying model efficacy is present.

Conclusion

Each of the goals laid out for this project was fulfilled. The EDA work performed showed strong seasonal trends in the data and explained the likely reasons behind the patterns. SARIMA modeling was successfully able to predict long-term SR value with a high degree of accuracy and model key components of the trends driving SR activity. Finally the RNN implementations in the last section demonstrated sufficient capability in short-term prediction. All of these elements combined make this project successful and these solutions viable for use in the problem space.

Next Steps

But additional work can always be done. Next steps for this project include:

- Aggregating and incorporating more/different data.
- Introducing exogenous regressors into SARIMA models.
- Reworking and tuning RNN configurations.
- Incorporating more advanced techniques for RNNs and searching a broader set of parameters.

Applying these next steps will allow for a more reliable and interpretable form of modeling.