NY Traffic Project

Group: 7406 Project Group 215

Alex Dreo
and Bear Jordan

November 2022

1 Abstract

Recent articles suggest New York City experienced a massive population shift during the COVID-19 pandemic. The present study seeks to understand how the impact of this population shift on predictive analytics through the lens of a logistics company, like Uber or FedEx. Specifically, Bayesian parameter estimation was used on the NYC Taxi dataset in order to compare pre-pandemic and post-pandemic traffic levels. This analysis showed that indeed, traffic volume decreased over the observation period consistent with a decrease in population. Next, XGB Regression models were fit to pre- and post-pandemic data to simulate a company, like Uber, who wanted to predict the average trip speed within New York. The analysis shows that despite the decreases in traffic, the predictive models do not show significantly different predictions when trained on either pre- or post-pandemic data. These results suggest that pre-pandemic data is still a valuable resource in modeling traffic dynamics within New York City.

2 Introduction

Living with traffic is synonymous with living in New York City (NYC). However, once predictable rhythms were turned on their heads when residents packed their bags in light of the COVID-19 pandemic. With the city easing restrictions, and vaccination taking effect, the city is finally seeing residents return.

Are these residents returning to the same city they left? The present study seeks to understand the city before and after the pandemic by examining what New Yorkers know best—traffic.

2.1 Stakeholder Benefits

A change in traffic volume has significant impacts for companies in the logistics sphere. Consider a logistics company that relies on historical data to predict the time to delivery. Does their pre-pandemic data accurately reflect the present reality? Are they introducing more error into their models by including the historical data? On the other hand, if the pre-pandemic data still reflects the present reality, are they making their models worse by only training models on post-pandemic data?

As a result of this study, companies will be able to make informed decisions on whether or not they should be concerned with the impacts a change in population may cause on their predictive traffic models.

2.2 Data Sources

The research question will be addressed by analyzing the The New York Taxi dataset[1]. This dataset was commissioned by the New York Taxi and Limousine Commission (TLC) starting in 2009. The ongoing project includes fourteen years of documented trips throughout the city for both Yellow and Green cabs.

[1] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Year	Mean	Std.	Median
2018	10.15	5.07	9.00
2022	12.01	5.95	10.55

Table 1: Calculated mean traffic statistics for each borough for each year in mph.

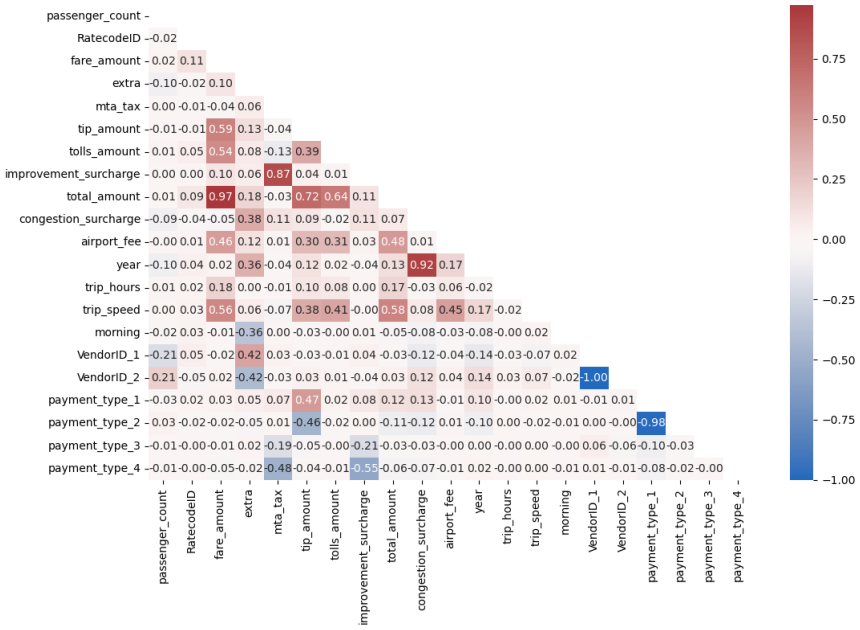

Figure 1: Correlation heatmap between variables

The dataset is subsetted between two types of cabs. Yellow cabs typically make trips in highly serviced areas like Manhattan. Green cabs were introduced in 2013 to service areas like Brooklyn and Queens that had low service levels.

The data includes "pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts" factors. This information provides a high-resolution look into New York City traffic patterns throughout all times of day and seasons. Traffic, and their changes, may also be used a proxy for many human activities.

Additionally, the Taxi Zone Lookup Table [2] is used to map location IDs to boroughs.

Examining the data, trip speed was calculated. It can be seen that both mean and median speeds are higher in the 2022 than in the equal sized 2018 dataset.

3 Modeling Overview

The present study consists of two main sections. The first section analyzes the underlying assumptions of the research hypothesis with Bayesian parameter estimation. And, the second section explores the implications of the research hypothesis.

[2] https://d37ci6vzurychx.cloudfront.net/misc/taxi+_zone_lookup.csv

3.1 Assumptions

Before diving into modeling, we will discuss simplifying assumptions and data preparation steps that are necessary for the Bayesian analysis and XGBoost modeling.

In order to compare like-in-kind data, analysis will be limited to peak traffic times—starting between 7:30am to 9:30am, and from 5:00pm-7:00pm, on non-holidays only. Only data from the Yellow Taxis will be used. As discussed in Section 2.2, the Yellow Taxis represent areas of "high-service" while the Green Taxis were introduced in order to service areas of low demand.

3.2 Data Preparation

The first step of the analysis is to clean and prepare the dataset. As discussed in the problem statement, the analysis focuses on how NYC has changed through the course of the pandemic. The authors have selected the year 2018 as representative of pre-pandemic traffic levels, and have selected 2022 as representative of post-pandemic levels.

Cloud computing is not within the scope of the current analyses. As such, the target file size per time period will be around 1GB of data. In order to meet this requirement, the analysis is limited to the month of January for each year.

In order to build models, several preprocessing steps are required, including filtering of NaNs, and removal outlier data. Outliers are individually considered and depending on their respective influence in the analysis. Some examples of outliers removed are trips with a total time length of 0 time and covered a non-zero distance. Finally, both years' datasets are further filtered down to 500,000 records each.

After resolving outliers, exploratory data analysis, correlation analysis among the features—as seen in Figure 1—feature selection, and construction of derived features (features that we create) is performed.

This was necessary to prevent the introduction of additional NaNs when average trip speed was calculated. Finally, all remaining categorical features are one-hot encoded.

4 Bayesian Analysis

In order to address this research question, simplifying assumptions must be made. First, the project must focus on a single unifying metric in order quantify "traffic." Second, the analysis focuses on "high-service" areas during peak hours.

Traffic will be defined via commute-time-per-mile. Areas whose commute-time-per-mile are high will be defined as areas of high traffic. And, areas whose commute-time-per-mile are low are defined as areas of low traffic.

4.1 Methods

As discussed above, the proposed analysis seeks to understand how traffic has changed in NYC from pre-pandemic to post-pandemic levels in NYC. This idea has an underlying assumption that traffic patters have changed at all. The first step in our modeling is to evaluate this assumption.

To explore this assumption Bayesian Parameter Estimation is used to describe the distribution of traffic within each borough for each time period. The goal here is to model the amount of traffic for a single borough in 2018 and again in 2022. With these two descriptive models, the distributions can be compared to provide detailed insight into how traffic changed over the course of the pandemic.

In addition to the simplifying assumptions and data cleaning steps listed above, outliers are removed by inspection for points that appear to be influenced by errors during the data entry process. Specifically, trips with negative traffic values (minutes per mile) are removed from the dataset. Also, some trips are recorded as taking over 10,000 minutes (>160 days) and are removed from the analysis.

As a general guide, the present analysis followed the best practices laid out by Andrew Gelman et al.'s Bayesian Workflow[3].

With the clean data, the next considerations are the priors. A log-normal prior distribution with a mean of 1.2 and a standard of deviation of 1.0 is selected based off personal experience in taxis. Practically, this distribution states that the number of minutes per mile is explicitly positive, is expected to take between zero and twenty

[3]http://www.stat.columbia.edu/ gelman/research/unpublished/Bayesian_Workflow_article.pdf

minutes per mile, and allows for positive extreme values. The prior assumptions are tested using prior-predictive checks.

The traffic data itself is simply modeled using an exponential distribution assuming traffic followed a power law. The conceptual model is as follows—on average people will only use a taxi if they expect the ride to save them more time than walking. Therefore, rides that experience less traffic should be significantly more likely than rides that experience greater traffic (which would cause people to use other means of transportation).

All Bayesian analysis is performed using the probabilistic programming package, Turing.jl, within the Julia programming language[4].

Posterior sampling is performed using Hamiltonian Monte Carlo with a leapfrog size of 0.001 and a ten leaps. Each prior and posterior distribution is sampled 1,000 times.

For each model, several diagnostic plots are produced including a probability density plot of the estimated traffic parameter for all posterior draws, a mean plot of the of estimated traffic parameter by iteration, and a trace plot of the sample value by iteration.

4.2 Results

4.2.1 Prior Predictive Checks

As the name suggests, prior predictive checks serve as a way to evaluate the influence of the selected priors. For each borough and time period, the prior predictive distributions are evaluated. As an example, the Brooklyn 2022 distribution shows that the expected number of minutes to travel one mile is greater than zero, most likely less than ten, but can be any positive value (Fig. 2). All prior predictive and posterior predictive distributions may be found in the project folder.

In summary, all prior predictive distributions showed a valid range of expected values for each borough during each time period. On average, the priors suggest that it should take around five minutes to travel a mile which is around 12 miles per hour. A reasonable rate in heavy traffic. These distributions also contain significant densities between zero and the median suggesting that lighter traffic (and faster trip speeds) are common. In addition, the prior predictive checks within a single borough show no substantial differences between the 2018 and 2022 values. However, the Bronx and Brooklyn boroughs show a slight increase in traffic moving from 2018 to 2022.

4.2.2 Posterior Predictive Checks

Posterior predictive checks offer a similar utility to prior predictive checks. However, now that the priors are conditioned on the data, they represent the reality of the environment.

Before examining these distributions, the validity of the analysis is verified by examining mean and trace plots for each combination. For a result to be considered, the mean should be stable after burn-in, and the trace plots should show some internal consistency.

For example, the posterior mean of EWR does not converge and the results are considered suspect (e.g. Fig. 3). Similar to EWR the means of Staten Island does not converge. All other boroughs are acceptable.

The trace plots for the EWR and Staten Island boroughs also do not converge (e.g. Fig. 4). All other boroughs show stable traces.

Acknowledging the suspect results for EWR and Staten Island, the posterior predictive distributions for each borough are examined. The fitted models produce distributions that show the average time to travel one mile of 5.10 minutes. The lowest time to travel one mile is the 2018 Manhattan value with a mean of 0.14 minutes, and the longest time to travel one mile is the 2018 Bronx value with a mean of 10.37 minutes.

4.2.3 Within-Borough Changes

Examining all the boroughs, the Bronx, Queens, and Staten Island all show posterior predictive traffic means in 2018 greater than their corresponding values in 2022. The Brooklyn, EWR, and Manhattan boroughs all increased, however, the differences in the posterior predictive means are minimal (Fig. 6).

The above analyses validate the present research hypothesis—traffic within each borough generally decreased over the observation period of 2018 to 2022. This analysis makes no claim

[4]https://turing.ml/stable/

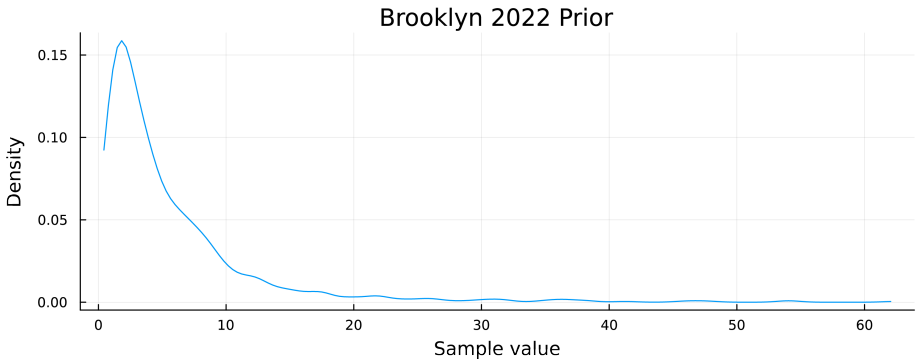

Figure 2: Prior predictive distribution of Brooklyn for 2022.

Borough	Model	Mean Traffic 2018	2022
Bronx	Prior	5.21	5.65
	Posterior	10.37	8.5
Brooklyn	Prior	5.42	5.66
	Posterior	9.02	8.93
EWR	Prior	5.46	5.46
	Posterior	2.15	2.39
Manhattan	Prior	5.69	5.69
	Posterior	0.14	0.22
Queens	Prior	5.85	5.25
	Posterior	7.64	5.58
Staten Island	Prior	5.37	5.33
	Posterior	4.49	0.92

Table 2: Calculated mean traffic statistics for each borough for each year (reported in minutes per mile).

on the causality of the decrease in traffic times. However, it is consistent with the news articles commenting on the numbers of people leaving during COVID. With the research hypothesis confirmed, future sections will explore the impact of these changes on predictive analytics for traffic models in New York.

5 XGBoost Modeling

The next model we built was a XGBRegressor or eXtreme Gradient Boosting Regression models. This regression model predicts a numerical variable, in this case average trip speed. It is an ensemble type model that uses a gradient descent algorithm over many different weaker regression trees, to return one optimal regression tree.

5.1 Methods

Two XGBRegressor models were built, one for for 2018 and 2022 respectively. The goal of these models was to predict average trip speed in each of these years, and compare the performance.

After the pre-processing steps discussed in Section 3.2 were conducted on our datasets, additional features were added. These derived features include is_morning, a boolean that indicates whether or not the trip happened in the morning rush hour period, and the many one-hot-encoded

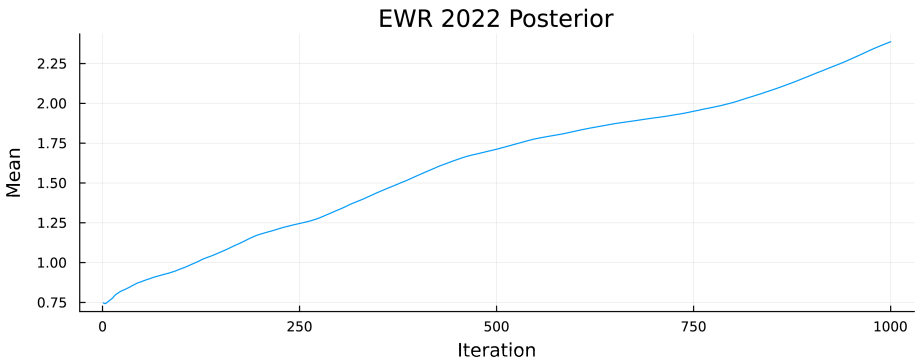

Figure 3: Posterior mean of EWR for 2022 is not stable.

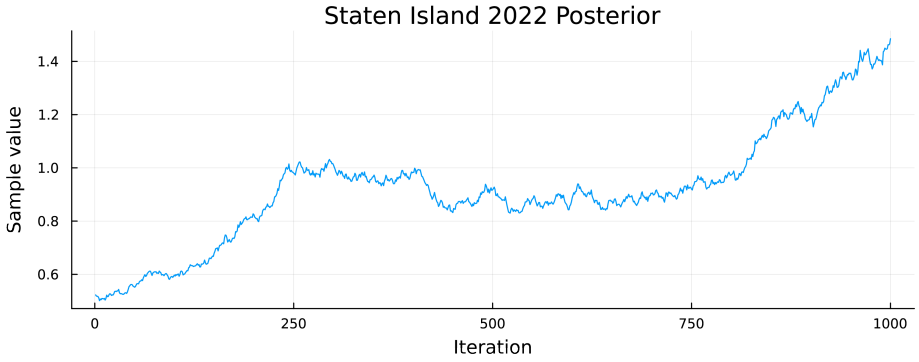

Figure 4: Posterior trace of Staten Island for 2022 is not stable.

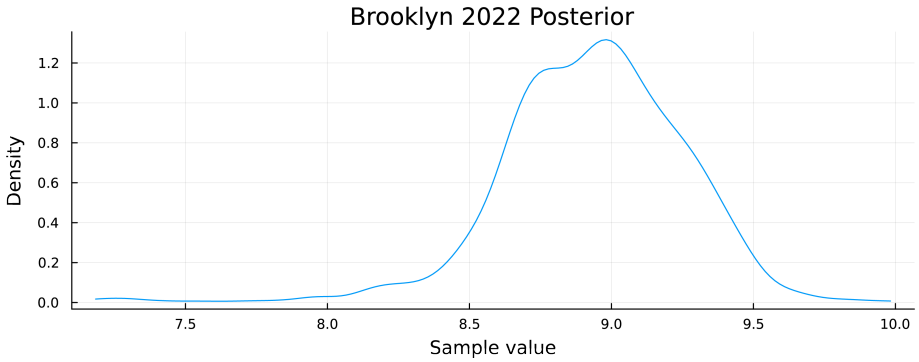

Figure 5: Posterior predictive distribution of Brooklyn for 2022.

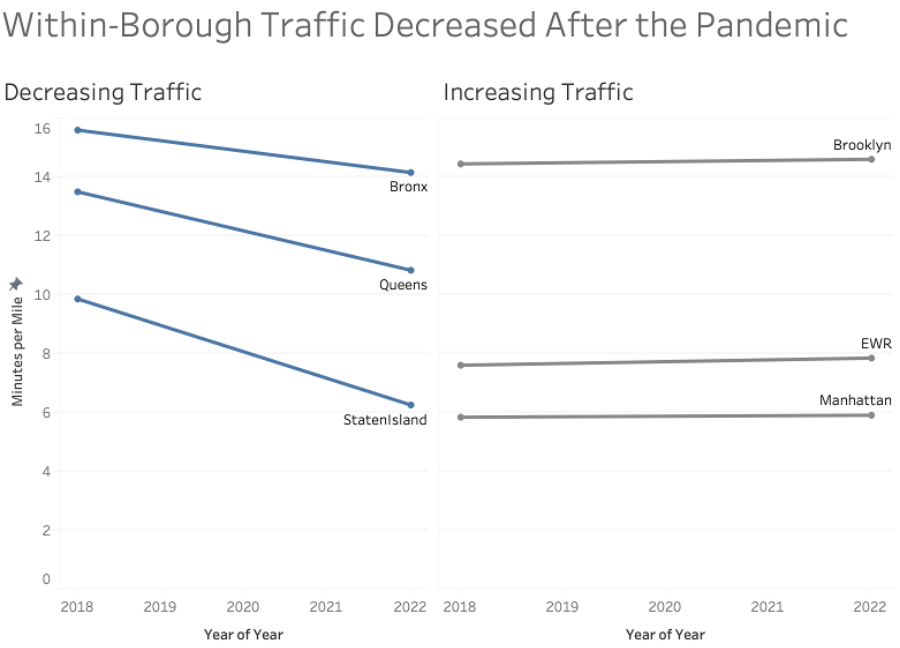

Figure 6: Changes in the amount of traffic for each borough between 2018 and 2022.

categorical variables.

After all of these features were added, feature selection to remove multicollinearity was necessary and was conducted using a LassoCV feature selection model. The final variables selected[5] were:

> 'RatecodeID', 'airport_fee', 'congestion_surcharge', 'extra', 'fare_amount', 'tip_amount', 'tolls_amount', 'total_amount', 'trip_hours'.

As these ensembles are comprised of aggregation of the 100 boosted trees each, it is difficult to visualize these models. They were both built with the same model parameters however. The full parameters can be found in Appendix Section 7.2. Some of the parameters such as max node depth of 6, were chosen to prevent over-tuning of the model onto the training set.

5.2 Results

The 2018 and 2022 models were trained, and tested on separate data sets from each of these years respectively. Then, both models' performance was calculated on the same validation set, with n=300,000. We found that the XGBRegressor models performed almost identically, with their resulting mean squared errors (MSE) of 2.68 vs 2.76 for the 2018 and 2022 models respectively. This increase of 3% in MSE is not an appreciable increase, and was not considered significantly different.

Next, the distribution of the predicted average speed was plotted with both years' models on the same validation set in Figure 7. The distribution of the difference between the predicted result and the actual average speed was also plotted in Figure 8. From these plots, we again see no significant differences between the two years' models. From Figure 8, we see the vast majority of the average speed differences were 0.

The goal of this model was to determine if there were any significant changes to average trip speed pre- or post-pandemic. Since the models' errors differed by such a small margin when applied to the same validation set, it seems likely that the same model to predict average trip speed could be used both pre- and post-Covid.

[5]Note that the one-hot-encoded versions of categorical variables, such as RateCodeID, were used.

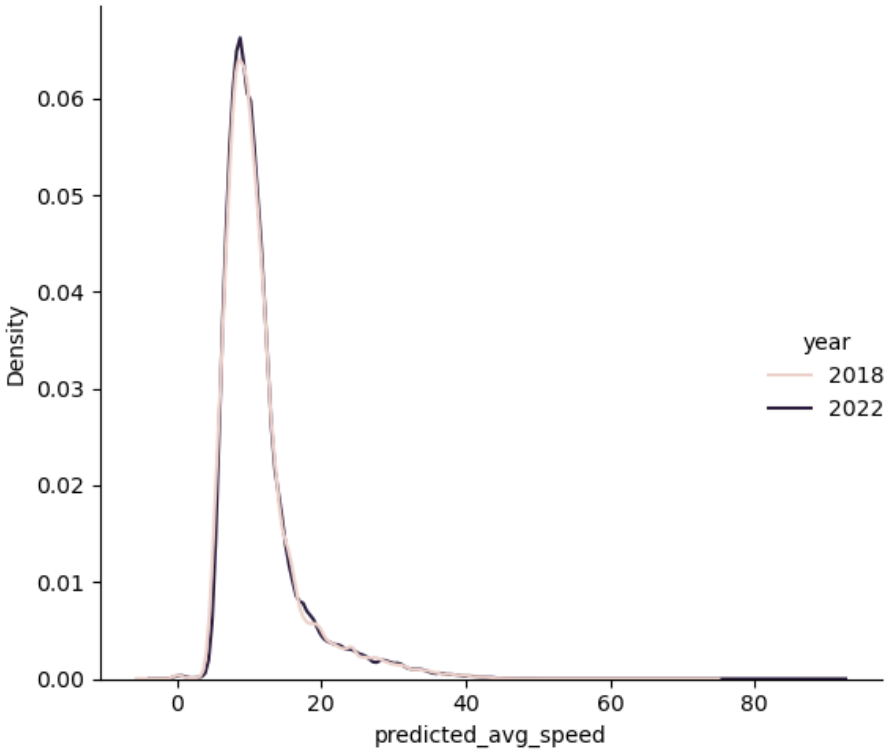

Figure 7: Distribution of predicted average speed.

6 Conclusion

The present study demonstrated that their are observable decreases in traffic volume over the course of the pandemic with Bayesian parameter estimation. Given this information, the introduction posed the question, "should companies be concerned with the impacts a change in population may cause on their predictive traffic models?" The traffic models generated in this study showed little to no differences when trained on pre- and post-pandemic data. This similarity suggests that a singular predictive model may be used to predict average trip speed pre- and post-pandemic, without a loss in accuracy, despite the change in population.

6.1 Future Work

Having different model types to compare the results of the XGBRegressor would have been beneficial. A Bayesian hierarchical regression (BHR) model in particular would have provided a helpful comparison point. BHR models can fit parameters for multiple different populations, resulting in pooled estimates. Besides just the performance comparison of these different model types, i.e. MSE, one could compare the parameters of each of the models. These parameter estimates would provide insight on the nature of our two populations considered: 2018 and 2022 average trip speeds.

6.2 Lessons Learned

6.2.1 Project

Due to the nature and complexity of the XGBRegressor ensemble models, the boosting tree model structures themselves could not be compared easily compared. This should be kept in mind when choosing a model type in the future. While these performed better than other tree models considered, it will be difficult to apply the findings of

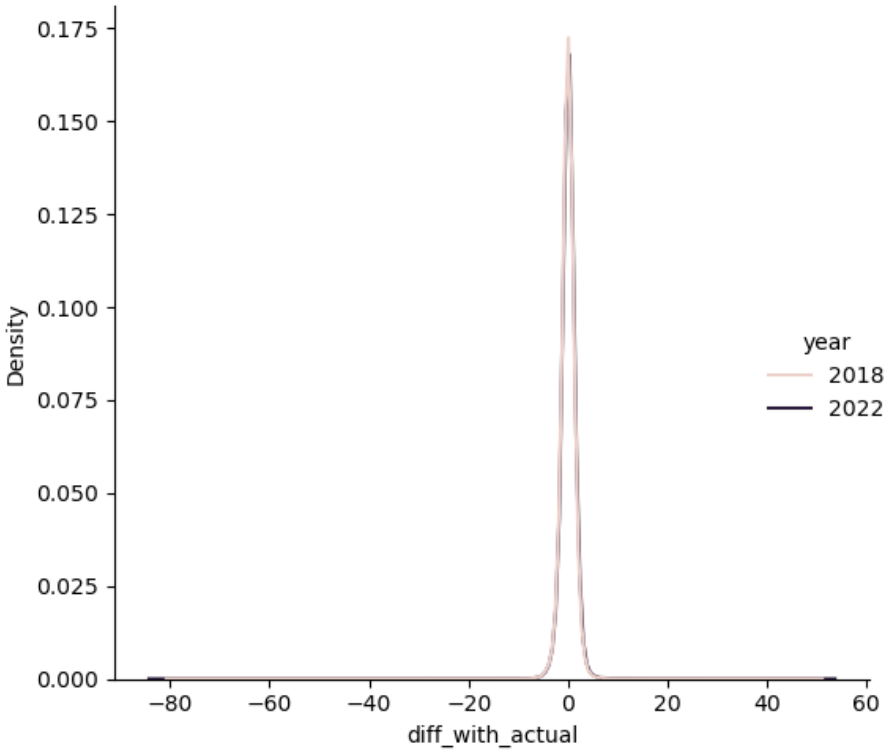

Figure 8: Distribution of difference between predicted average speed and ground truth actual speed.

an ensemble model generically to your system. If a parametric model were instead used, then tuned model parameters themselves could have been compared between the two years' models.

The Bayesian analysis was particularly difficult for us. We tried to incorporate it into coursework, but had to abandon it due to time constraints and the complexity of the setup. It was extremely rewarding actually getting the models to converge and produce reasonable results. Specifically, I finally have a better understanding of the relationship between prior and posterior distributions. For this project, we started off using a normal distribution for the prior on the exponential distribution's rate parameter. We discovered this was fundamentally flawed because the normal distribution was able to produce negative values for the rate parameter which is not allowed for the exponential distribution. The rate parameter must be positive. This was a breakthrough moment in understanding the relationship between the prior and posterior distributions while creating a Bayesian model.

6.2.2 Course

This course built on our pre-existing knowledge of concepts like cross validation, and reinforced the need for it in small datasets. We appreciated this learning in particular as it is so generally applicable to many model building processes. The kernel methods and multiple model evaluation sections stood out to us as being particularly applicable in our jobs. We strongly prefer the project based learning style of this course. Analytics is such a broad field that having space to bring our own interests and combine it with the course material really made it more interesting and engaging.

7 Appendix

7.1 Bayesian Output

See the attached Supplementary Information file for the Bayesian figures and model output.

7.2 XGBRegressor Parameters

```
{'objective': 'reg:squarederror',
 'base_score': 0.5,
 'booster': 'gbtree',
 'colsample_bylevel': 1,
 'colsample_bynode': 1,
 'colsample_bytree': 1,
 'gamma': 0,
 'gpu_id': -1,
 'interaction_constraints': '',
 'learning_rate': 0.300000012,
 'max_delta_step': 0,
 'max_depth': 6,
 'min_child_weight': 1,
 'monotone_constraints': '()',
 'n_jobs': 12,
 'num_parallel_tree': 1,
 'predictor': 'auto',
 'random_state': 0,
 'reg_alpha': 0,
 'reg_lambda': 1,
 'scale_pos_weight': 1,
 'subsample': 1,
 'tree_method': 'exact',
 'validate_parameters': 1,
 'verbosity': None}
```

8 Bibliography and Credits

References were included as footnotes in the relevant sections.