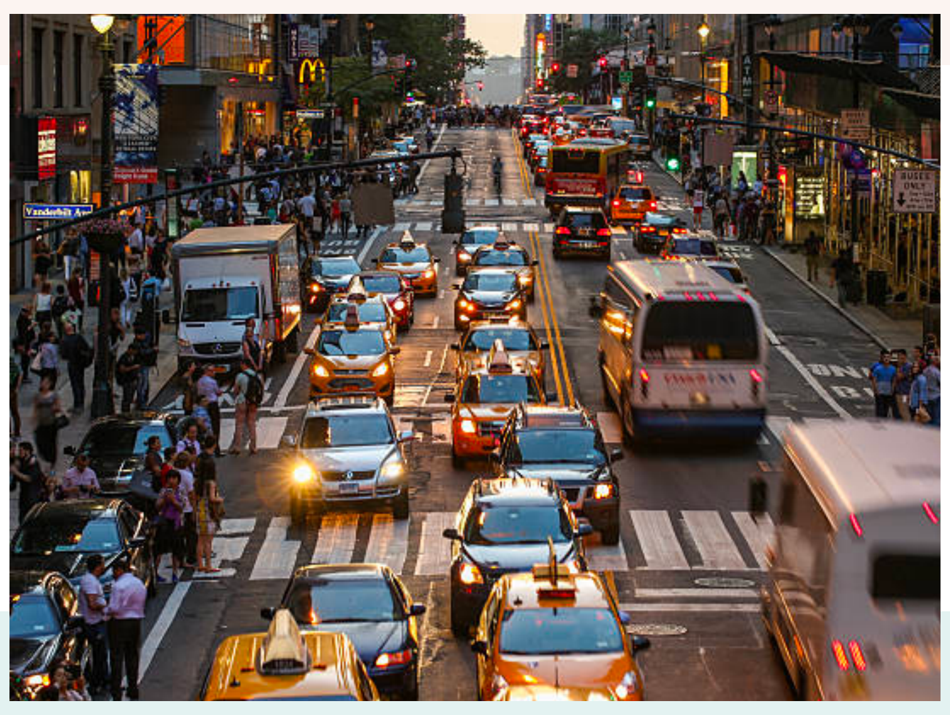
Predicting Parking Violations and Optimizing Parking Systems

Srajan Dube, Katherine Lang, Nilesh Manivannan

Problem Definition

The deficiencies of urban parking systems have only been exacerbated by increased traffic and expansion of metropolitan areas. Good parking systems are essential for creating livable and sustainable cities for a multitude of reasons. First, **traffic congestion** increases as the time it takes for drivers to search for parking increases, leading to unwanted emissions and poor air quality. It also worsens **accessibility** for residents and visitors alike, to businesses and public spaces. Third, insufficient parking systems also cause an increase in **unsafe parking practices.** Practices such as double parking or stopping in no-parking zones can cause accidents, affect pedestrian safety, and hinder emergency services and public transportation. Overall, it creates a **negative social impact,** and little is being done to remedy these issues. We propose a machine learning-based approach that leverages historical parking data to examine where parking violations occur in New York City and connecting this data to existing zoning laws to make recommendations to future parking systems and ultimately improve quality of life for urban residential and visiting drivers.

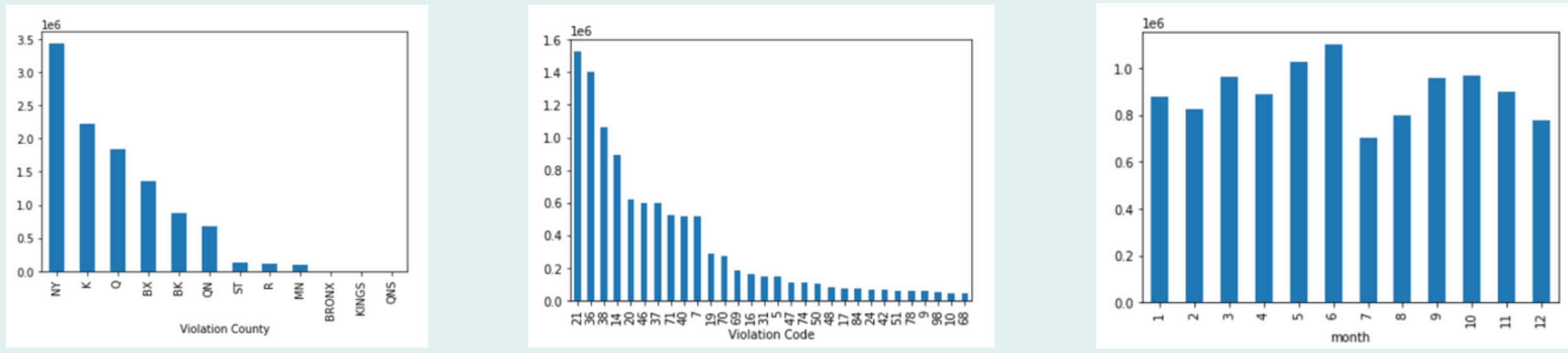


Approaches

Data Cleaning

For our first step, we wanted to correct and remove errors from our data sets before applying any algorithms. We began with data profiling, examining the structure, information, and consistency of each file to make sure it would be effective for our intended uses. Then, we resolved any invalid and outlier data that could significantly affect our predictions. We utilized the Pandas library to pinpoint entries with null values and compute averages to identify and outliers. Thus, we were able to account for large events or gatherings that may have skewed parking violation statistics on a particular day or street. We finally grouped our dataset by Date and Precinct to perform the necessary operations separately.

The graphs below depict parking tickets by county, violation code, and month , respectively.



ARIMA Regression

We decided to use ARIMA as one of our approaches because of its effectiveness and flexibility working with time series data, and its ability to accurately forecast future values based on past observations. ARIMA's functionality was also easy to interpret. ARIMA, which stands for AutoRegressive Integrated Moving Average consists of 3 components.

1. Autoregression (AR)
2. Integration (I)
3. Moving Average (MA)

The **autoregressive** component uses past values of the time series to forecast other values. For **integration**, the model differences the data at consecutive values to stabilize mean and variance over time. **Moving average** then uses these values to establish the relationship between observation and past errors.

Gradient Boosted Decision Trees

Gradient Boosted Decision Trees (GDBT) is also an effective machine learning algorithm with high accuracy and flexibility. This method also offers advantages for handling time series data. GDBT adapts to non-linear relationships, which can be frequent in time series data, and so it was able to address the relationships present in our data. Also, it is less sensitive to noisy data. Despite initial data cleaning, there could still have been some outliers present. However, due to GBDT's capabilities, the model still performed consistently well. The algorithm works by iteratively adding decision trees to a model, the following way:

1. The model is first initialized with a single decision tree based on the target variable for the entire data set.
2. The model makes predictions based on the training data and computes errors
3. The errors are used to train a new decision tree to correct these errors. The new tree is then added to the model.
4. Steps 2 and 3 are then repeated until performance stops significantly improving.
5. Final predictions are made based on all the trees in the model.

This technique also makes use of several tunable hyperparameters, including learning rate, number of trees, depth of trees, sampling rate, and regularization.

VAR Model

The VAR model, or the Vector Autoregression model is a common technique used for time series forecasting. The model uses a system of linear equations, with each representing a variable. The model estimates the coefficients of these linear equations using **maximum likelihood estimation.** It is then trained and can be used to forecast particular variables using the set of linear equations. The VAR model presents several advantages over other forecasting models. Its ability to handle multiple variables and identify relationships among all of them makes it a robust and flexible technique for generating short-term predictions.

Data

Dataset	Description	Source	Size	Data Attributes
NYC Parking Tickets	Parking tickets issued in NYC over 4 years	Kaggle, collected by New York City Department of Finance	2GB, 42M rows	summons number, the violator's plate ID, registration state, plate type, issue date, violation code, vehicle body type, vehicle make, issuing agency, and street code
NYC Parking Meters	Locational data of NYC parking meters	NYC Open Data	3MB, 16K rows	BOROUGH, ON_STREET, FROM_STREET, TO_STREET, METER_HOUR

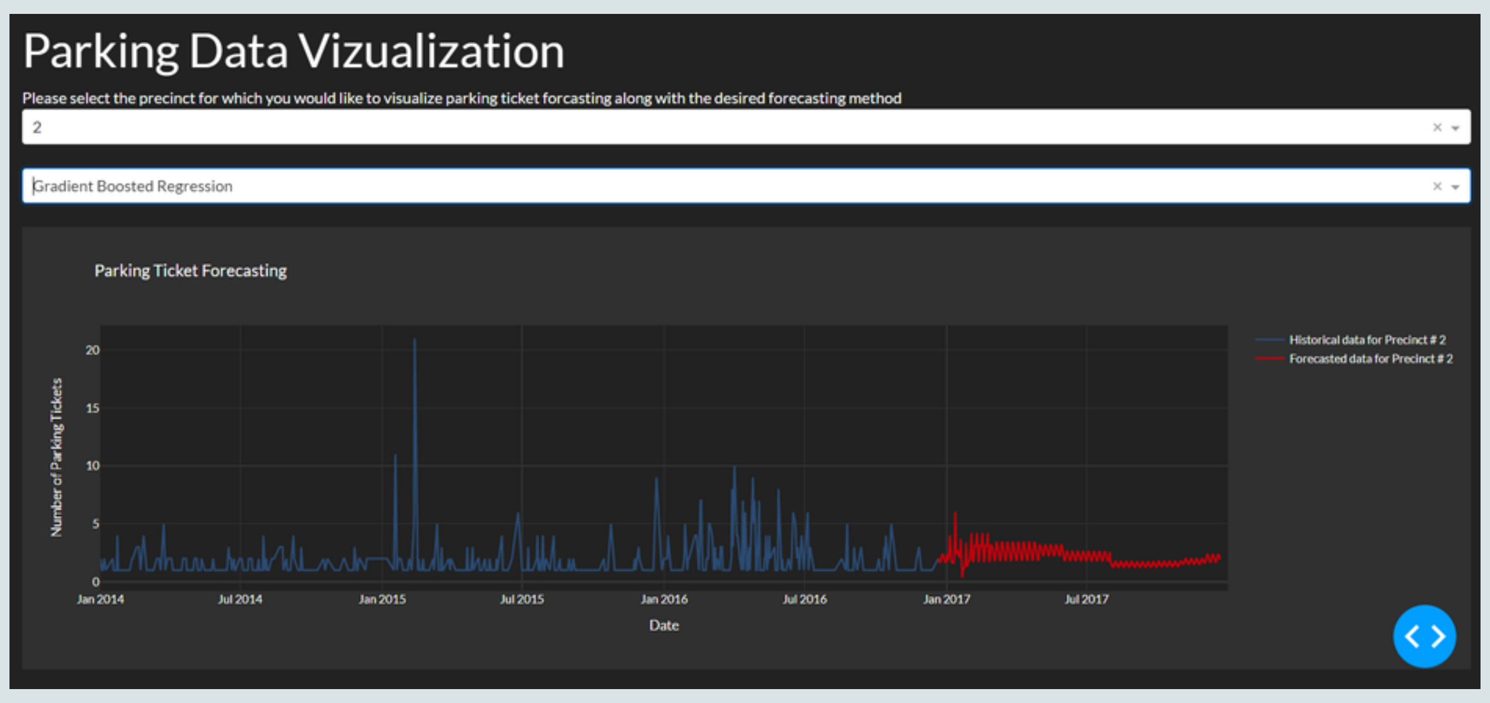
Experiment and Results

Evaluation

As part of the evaluation method, we used a technique called **hold-out validation.** Evaluating the models, demonstrated that although there are few instances where the models' predictions match the actual values, they do manage to capture the overall patterns. This is because when there are sudden changes in the actual values, the predicted values also show similar changes around the same time, although not with the exact same magnitude. Therefore, while the model predictions may not be reliable for determining exact values, they can be helpful in detecting trends and predicting the general direction in which the values will change.

Results

To better visualize the results achieved from our machine learning models, we created a GUI that displays predictions for each model.



Parking Data Vizualization

Please select the precinct for which you would like to visualize parking ticket forcasting along with the desired forecasting method

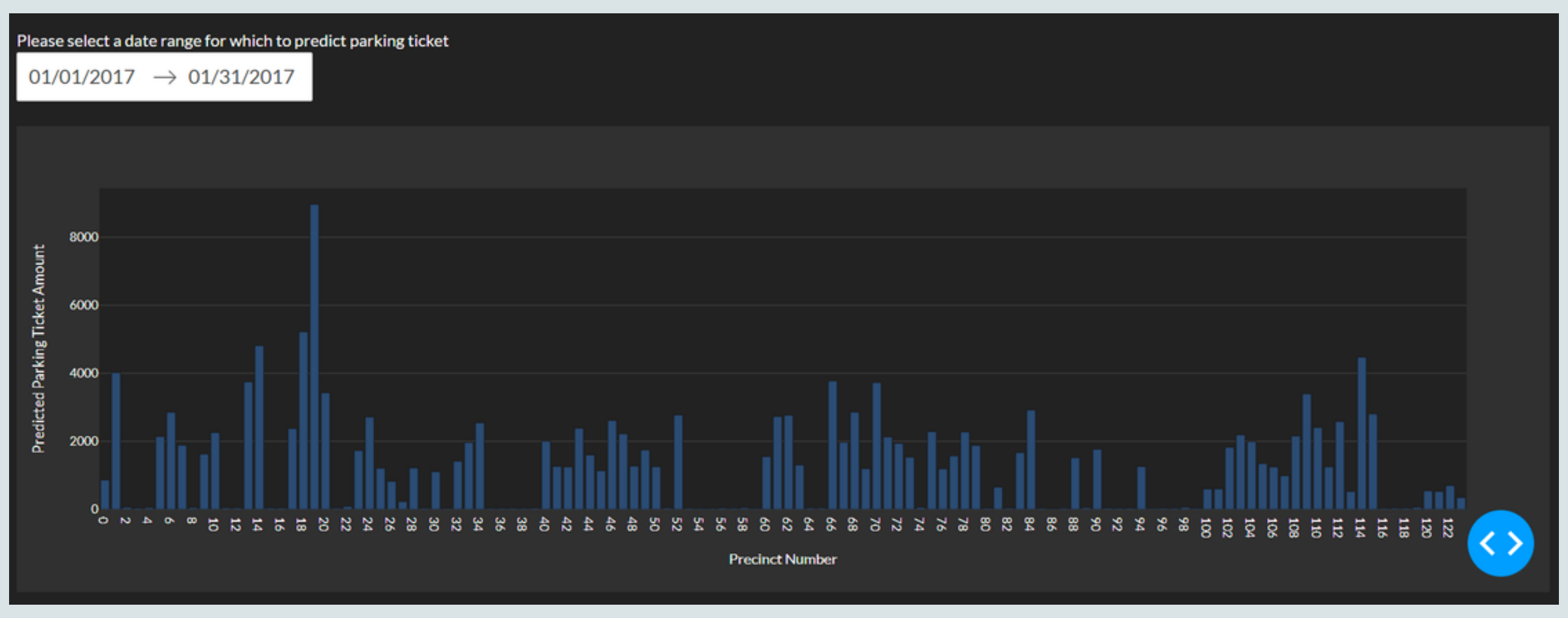
2

Gradient Boosted Regression

Parking Ticket Forecasting
Number of Parking Tickets
20
15
10
5
0
Jan 2014
Jul 2014
Jan 2015
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Date
Historical data for Precinct # 2
Forecasted data for Precinct # 2

The user is able to input the precinct for which they want to view data. They also have the option of viewing predictions using different algorithms.

Given precinct, they can also manually select the dates for which predictions are shown

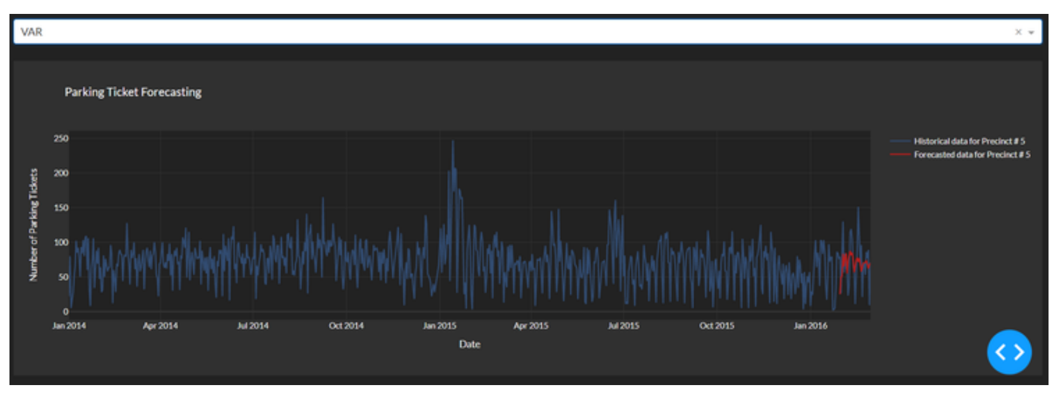


Please select a date range for which to predict parking ticket

01/01/2017 → 01/31/2017

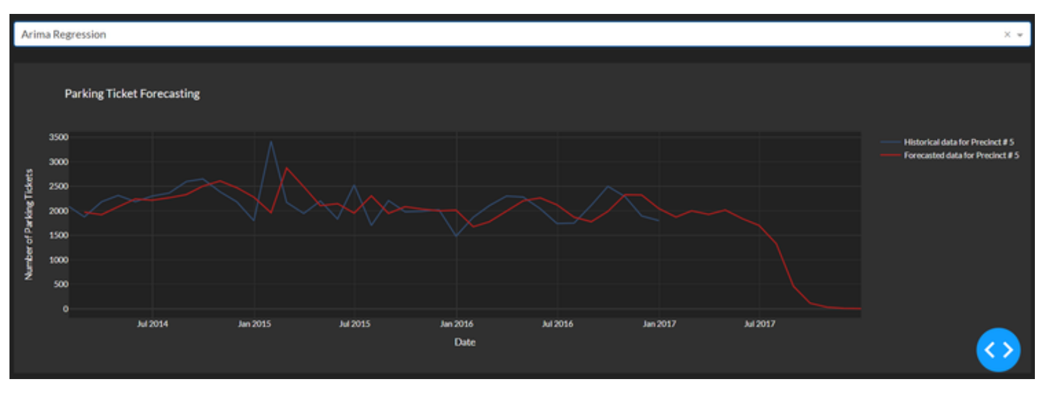
Using the information from our visualizations, we were able to note several trends about parking violations in New York City. For example, we see high volumes of issued tickets in precincts 18 and 19, which serve Manhattan, a densely populated borough with many tourist attractions. There is also limited parking, so these factors combined contribute to the high number of parking tickets. We also see high numbers at precincts 114-115, which serve Queens, the largest borough in NYC, which also contains the LaGuardia Airport, which could also contribute to increased rates of illegal parking practices. .

Of the three models we used, Gradient Boosted Regression worked the best. With its ability to accommodate for more volatile and nonlinear relationships, it was clearly the best choice for the parking data we worked with. VAR was promising for short term predictions but failed over longer ranges. Even in shorter ranges, it clearly performed worse than GBR. Finally, ARIMA performed fairly decently but failed to capture some of the trends seen throughout the data. It provided better data overall though when considering month-wide aggregation.



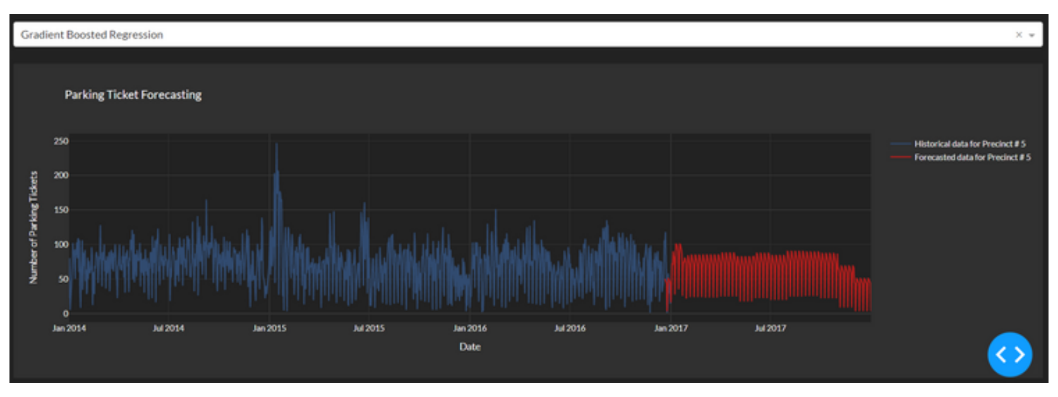
VAR

Parking Ticket Forecasting
Number of Parking Tickets
250
200
150
100
50
0
Jan 2014
Apr 2014
Jul 2014
Oct 2014
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Date
Historical data for Precinct # 5
Forecasted data for Precinct # 5



Arima Regression

Parking Ticket Forecasting
Number of Parking Tickets
3500
3000
2500
2000
1500
1000
500
0
Jul 2014
Jan 2015
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Date
Historical data for Precinct # 5
Forecasted data for Precinct # 5



Gradient Boosted Regression

Parking Ticket Forecasting
Number of Parking Tickets
250
200
150
100
50
0
Jan 2014
Jul 2014
Jan 2015
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Date
Historical data for Precinct # 5
Forecasted data for Precinct # 5

The graphs show predictions made using each machine learning technique: VAR (top left), ARIMA regression (above), and GDBT (left).

Conclusions

Through this project, we learned about various machine learning algorithms used for time-series forecasting. The **VAR model** performs strongly with multivariate data, but yielded worse predictions. **ARIMA,** on the other hand, served well for univariate time-series relationships, and was easily adaptable to our project's purposes. **Gradient Boosted Decision Trees** demonstrated the most proficiency in producing accurate results, due to its ability to model complex and nuanced relationships.

There are several avenues for potential future work we would like to explore.

1. **Incorporating real-time data:** This could involve gathering data from sensors, parking meters, or parking systems to refine our models and allow them to better predict future violations.
2. **Gathering weather data:** weather can have a huge impact on parking data, so by considering it as another variable in our models, we may yield more accurate results.
3. **Testing feasibility:** This data could be used to mitigate parking-related traffic congestion. In order for this to work, we would need to incorporate existing zoning laws and collaborate with city planning officials to consider common parking practices in construction of future parking structures and systems.

CSE 8803 Introduction to Urban Computing ***B. Aditya Prakash, Spring 2023***