Predicting Parking Space Violations and Optimizing Parking Systems

Srajan Dube
Georgia Institute of Technology
sdube7@gatech.edu

Katherine Lang
Georgia Institute of Technology
katherinelang@gatech.edu

Nilesh Manivannan
Georgia Institute of Technology
nmanivannan6@gatech.edu

ABSTRACT

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. We evaluate the effectiveness of the proposed system through hold-out validation and visual analysis to demonstrate its potential to significantly improve parking availability and reduce traffic congestion in urban areas.

Our WebPage

1 INTRODUCTION/MOTIVATION

The problem of parking availability in urban areas has become increasingly urgent, as populations grow and urbanization continues. There are a multitude of problems that indicate the need to improve parking systems in cities. In densely populated areas, there is typically a shortage of parking space availability. The lack of efficiency in parking space searches causes wasted time for drivers, traffic congestion, and higher rates of carbon emissions. There is also inefficiency in parking management systems. Due to outdated management and data collection methods, the available data on parking systems is inaccurate and often lacking necessary information about parking. Additionally, parking fees make city parking practices impractical for many residents and visitors. This yields many negative impacts, such as increased illegal parking practices, less accessibility for low-income residents, and negative social impacts from less access to essential services. Finally, many urban areas are afflicted by poor parking infrastructure. Without appropriate designated parking areas, conditions will be crowded and unsafe for drivers and pedestrians alike.

2 PROBLEM DEFINITION

These obstacles call for a data-driven solution to improving parking systems in urban areas. In order to combat the problems associated with poorly designed parking systems, we propose a solution with a machine-learning based approach that will incorporate existing data about parking violations into future additions or modifications to urban parking infrastructure. For this purpose, we leveraged data from New York City's issued parking tickets, collected by the New York City Department of Finance, along with locational data from NYC Open Data, a website with free public data contributed to by NYC agencies and partners. This data was then used to train models that would predict where, why, and how often these violations would occur.

3 RELATED WORK AND SURVEY

Before we began implementation of this project, it was important for us to explore the field and domain to gauge what had been done, what areas can be expanded upon, and how our work can contribute to advancements. In order to do this, we reviewed relevant works, mostly in the form of academic papers, that would help us build upon existing research and develop novel approaches to a current problem.

First, we wanted to better understand the severity of the problem, and plenty of papers were able to highlight how parking ticket distributions had a negative social impact. One piece in particular, "The Unequal Spatial Distribution of City Government Fines: The Case of Parking Tickets in Los Angeles" by Noli Brazil [2] alerted us to the serious nature of the problem. This research paper investigates the distribution of parking ticket fines in Los Angeles and the relationship between these fines and neighborhood characteristics such as race, income, and education. We found that its strengths lie in its clearly-defined research problem and its connection to public policy, as well as clear recommendations about how to remedy these issues from a social standpoint. The contents of this paper encouraged us to search for solutions that addressed the importance of detecting parking violations using our own set of skills.

One of the earlier sources we discovered, "Disaggregate Analysis of Relationships Between Commercial Vehicle Parking Citations, Parking Supply, and Parking Demand" [8] brought our attention to connections between parking violations and how this affects parking supply and demand. The paper focuses on an important topic of commercial vehicle parking demand and supply, which is a critical issue for urban transportation. The research methodology is sound and uses detailed parking data, including citations, occupancy rates, and parking supply, over a two-year period. The study provides insights into the relationship between parking citations, parking supply, and parking demand and sheds light on the impact of these factors on each other. The paper provides useful information for policymakers and transportation planners in developing effective

strategies for managing commercial vehicle parking and improving urban mobility. However, the study is limited to the city of Toronto, Canada, and may not be generalized to other cities with different transportation systems and policies. It also does not take advantage of many technological-based approaches that we will see in further works.

We knew we wanted to leverage machine learning techniques, as they would be advantageous for things that demanded predictive models. As such, we looked into works by Nikolaos Karantaglis, Nikolaos Passalis, and Anastasios Tefas who conducted research in works like "Predicting on-street parking violation rate using deep residual neural networks" [6] and "Deep Learning for On-Street Parking Violation Prediction" [5]. These papers detail the use of Deep Residual Neural Networks with impressive data augmentation systems that allow them to fix noisy data. Overall, these papers were incredibly beneficial to us as it explains many of the details behind defining and using a neural network for parking violation predictions, and we were able to use them as references to overcome certain challenges (such as the issue with noisy/incomplete data) that we encountered. A few issues we noticed in this paper included a lack of any data outlier correction as well as the use of a quite sparse and incomplete dataset. While the results were promising, we believed them to be skewed due to the nature of their data augmentation and we addressed this by using the more complete and detailed New York parking ticket dataset and making the necessary additional data augmentations. The paper also predicted parking violations in Greece as a whole and failed to identify more specific locations of those predicted violations. This was a problem, as parking behaviors varied differently across a broad region, and not necessarily only urban areas. Our goal was to extend this as well by providing more specific predictions in the New York area.

A couple other sources also explored neural networks as techniques for parking violation and on-street parking predictions. One such source was "A Comparative Analysis of Machine/Deep Learning Models for Parking Space Availability Prediction" [1]. The authors of this paper focus on solving the issue of parking space mismanagement, and the struggle of those who are unable to access free parking spaces in crowded urban areas. Using machine learning and deep learning approaches, the authors wanted to devise a resolution to help predict parking space availability. This problem is within the domain of our project, as facilitating the transfer of parking spaces between attendants contributes to efficient use of these parking spaces. One significant strength of this paper is that the researchers explore machine learning and deep learning methods, and compare them to find the best predictive model. As part of their research, they investigated neural networks, specifically the MLP neural network technique. However, they were not able to expand on the applications of their research in-depth, which is what we plan to address in ours.

Similarly, in "Multisource Data Integration and Comparative Analysis of Machine Learning Models for On-Street Parking Prediction" [4], they also adopted an MLP architecture and ran multiple iterations to determine hidden layers and neuron size. The pitfalls of this paper was that the data was gathered from Melbourne, which also has very different parking and traffic behaviors compared to urban America.

An additional concern we had with the two former works was that they limited themselves to one type of model. However, we knew we wanted to explore multiple techniques that could be compared and contrasted. Several works accomplished this, like "Predicting the spatiotemporal legality of on-street parking using open data and machine learning" [3], which used 6 different Regression models and reported both the RMSE and F1 scores to compare and analyze all the models they used. Though this paper was incredibly comprehensive, it placed a lot of emphasis on the spatial aspect, while we also wanted to bring attention to the temporal aspect. "Short-Term Prediction of Available Parking Space Based on Machine Learning Approaches" [9] used similar machine learning techniques to the methods we used for our implementation. As part of a traditional method, they also use an ARIMA model for straightforward and fast calculations. To better represent the nuances of time-related changes in parking, they also utilize a Gradient Boosting Decision Tree model (GBDT). By using results from multiple decision trees, the GBDT best method for predicting future parking-related traffic in urban areas. This technique will help us predict when and where parking violations could occur.

Several works also allowed us to consider avenues of future work and additional implications. One of the earlier works mentioned, "Multisource Data Integration and Comparative Analysis of Machine Learning Models for On-Street Parking Prediction", applied machine learning principles to parking in Melbourne. Though the research was conducted in a different city, their predictions were incredibly accurate due to the comprehensive dataset they were able to work with. This data included many external factors, for example: weather data, public transportation, and pedestrian traffic. If we were able to access this data, we could also better train our models to produce more accurate predictions. Similarly, "Smart parking pricing: A machine learning approach" [7] made recommendations to parking systems using real-time occupancy data to dynamically price parking. Though we were not able to find the data required to apply these concepts to our own research, we hope that this data could become available to improve our models.

4 PROPOSED METHOD

Currently, the process for creating zoning laws regarding parking does not take into account actual parking ticket data. However, this is a vital piece that must be taken into consideration to properly outline and define zoning laws that effectively regulate parking. By analyzing parking tickets, law makers can allocate resources better and enable laws that provide enough parking where there is demand. This would improve the urban planning for any city that accurately analyzes parking ticket data.

4.1 Data Collection Process

For data collection, we gathered information from 3 different datasets: A time-series parking ticket dataset, a parking meter location dataset, and a New York zone dataset. We started by loading in the datasets

into a Pandas dataframe. The data included many outliers and missing information. We first cross referenced the zone and parking meter location datasets to find out which zones each parking meter corresponds to. We then cross referenced the resulting dataset with the parking ticket dataset. This gave us a final time-series dataframe for the parking ticket data in New York along with location and zone information.

To deal with the outliers, we calculated the standard deviation of the number of parking tickets each month in each zone and limited data to be ±3 standard deviations from the mean. This ensures that random occurrences wouldn't affect our overall analysis and that our data would be consistent. To do this, we had to group our dataset by zone and month. We also replaced any missing data with the mean for the month and zone it corresponds to.

Finally, we set the dates as our dataframe index to make it easier for time series analysis.

4.2 Approach, Models, and Mathematical Background

After data collection and cleaning, our goal was to apply numerous state-of-the-art forecasting models to forecast future parking ticket levels. The models we used were Gradient Boosted Regression due to its ability to perform well on nonlinear datasets, ARIMA regression due to its ability to handle data with trends, seasonality, and other complex patterns, and VAR due to its ability to work well with multivariate data and calculate interdependencies. We felt that these three models were quite diverse in their benefits which would truly help us analyze the data from all angles. After training these models and generating predictions, we created a GUI to visualize our predictions along with the actual data and compared the results of the three models (our results are shared later in the document). Here we are adding a short description and mathematical background of the 3 models we used.

(1) Gradient Boosted Regression
The mathematical background necessary for gradient boosting using decision trees involves an understanding of regression analysis, decision trees, and the concept of boosting. Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to find the best-fitting line (or curve) that describes the relationship between the variables. In gradient boosting for regression, decision trees are used as the base learners, and the objective is to improve the accuracy of the predictions made by combining multiple decision trees.

Decision trees are a type of machine learning algorithm that is used for both classification and regression tasks. A decision tree consists of a series of nodes that split the data based on specific features or attributes. Each split is chosen to maximize the separation between the data points in the resulting subsets. The process of splitting continues until a stopping criterion is met, such as reaching a certain depth or having a minimum number of data points in each leaf node.

Boosting is a technique used to improve the performance of machine learning models by combining multiple weak learners to create a stronger ensemble learner. The idea behind boosting is to iteratively train weak learners on the data, with each subsequent learner being focused on the data points that the previous learners have struggled to classify correctly.

Now that we have a basic understanding of these concepts, we can delve into how gradient boosting using decision trees works. The algorithm begins with a single decision tree, which is trained on the data using a loss function such as mean squared error. The predictions made by this first tree are compared to the actual values in the training data, and the differences (or residuals) are calculated. The next decision tree is then trained on these residuals, rather than the original target variable, in an attempt to improve the predictions of the first tree. This process is repeated iteratively, with each subsequent tree trained on the residuals of the previous tree. The predictions made by each tree are then combined to create the final ensemble prediction.

To prevent overfitting, regularization techniques such as shrinkage, subsampling, and column subsampling can be used. Shrinkage involves scaling down the contributions of each tree to the final prediction, while subsampling and column subsampling involve using only a subset of the data and features for each tree.

(2) ARIMA Regression
The mathematical background necessary for ARIMA (Autoregressive Integrated Moving Average) regression involves an understanding of time series analysis, autocorrelation, and the concept of stationarity.

Time series analysis is a statistical technique used to model and analyze data that is collected over time. In time series analysis, the goal is to model the underlying pattern or trend in the data, and to use that model to make predictions about future values. Autocorrelation is a measure of how related a variable is to itself over time. If a variable is highly autocorrelated, it means that there is a strong relationship between its values at different points in time. Stationarity is a property of time series data that refers to the stability of its statistical properties over time. A stationary time series has constant mean and variance, and its autocorrelation function is constant over time.

ARIMA regression is a popular technique used in time series analysis to model non-stationary time series data. It involves three main components: autoregression, differencing, and moving average. Autoregression refers to a regression model that uses past values of the variable to predict its future values. In ARIMA, the autoregression component (AR) involves modeling the variable as a linear combination of its past values. Differencing refers to the process of subtracting the previous value of the variable from the current value. This is done to make the time series stationary, by removing any

trends or seasonality in the data. Moving average refers to a model that uses the past errors to predict future values of the variable. In ARIMA, the moving average component (MA) involves modeling the errors as a linear combination of past error terms. The 'I' in ARIMA stands for Integrated, which means that the time series is differenced a number of times until it becomes stationary.

(3) VAR Model
Vector Autoregression (VAR) models are a type of statistical model used to analyze the relationships between multiple time series variables. (An explanation of time series analysis is provided in the ARIMA background section)

In VAR models, a set of p time series variables y1, y2, ..., yp are modeled simultaneously. The VAR model assumes that each variable depends on its own past values and the past values of the other variables in the system. The model can be written as:

yt = c + A1yt-1 + A2yt-2 + ... + Apyt-p + ut

where yt is a p-dimensional vector of the current values of the time series variables, c is a p-dimensional vector of constants, A1, A2, ..., Ap are p x p matrices of coefficients, and ut is a p-dimensional vector of error terms.

To estimate the coefficients in the VAR model, various methods can be used, such as maximum likelihood estimation or Bayesian inference. Once the coefficients are estimated, the model can be used to forecast the values of the time series variables.

VAR models have several advantages over other time series models. For example, VAR models can capture the interdependencies between multiple time series variables, while AR models can only model the relationship between a single variable and its past values.

5 INITIAL FINDINGS

After an initial analysis of our dataset, we found some interesting information regarding the months, counties, and violation codes that receive the most parking tickets. Here are some of the charts we generated

Based on this rudimentary analysis, we noticed a few interesting things. Firstly, parking tickets are given out quite evenly across months with a few exceptions being a noticeable decrease between June and July. We also analyzed violation codes which showed us that the most common parking violation is "Failing to show a receipt or tag in the windshield." We were also able to see that other driving related crimes such as not obeying traffic signals is included in this dataset. Finally, we found that New York, KoreaTown, and Queens are the counties with the highest violation occurrences.

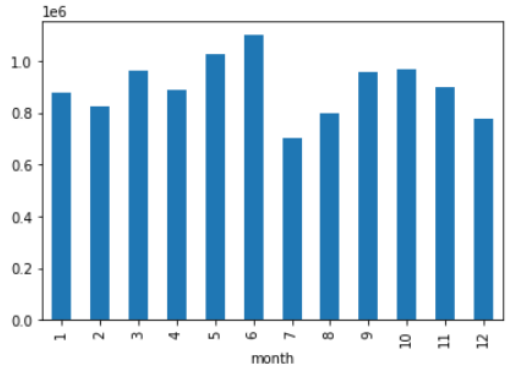

Figure 1: Parking Tickets by Month

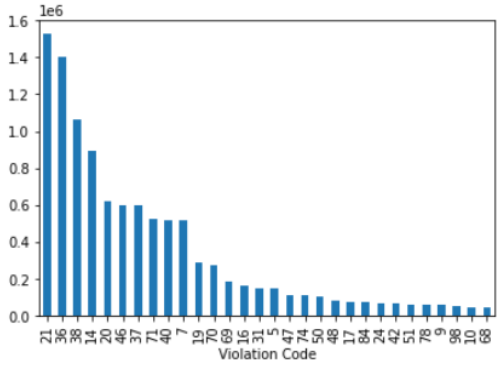

Figure 2: Parking Tickets by Violation Code

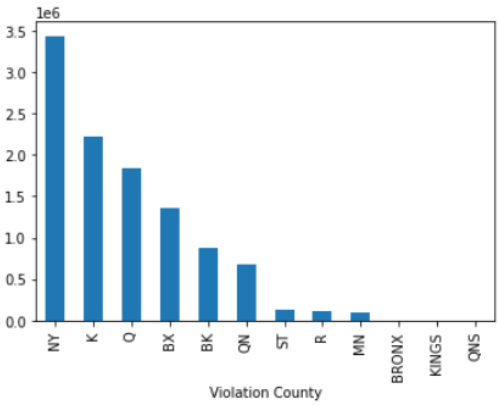

Figure 3: Parking tickets by County

Furthermore, after an initial analysis of the zoning laws in New York, we have identified a few important laws that may contribute to more parking tickets.

- No Standing Zones: In New York City, there are specific zones where standing (i.e., parking for any length of time)

is prohibited. These zones are typically designated by signage and include areas near fire hydrants, bus stops, and pedestrian crosswalks.
- Alternate Side Parking: Alternate side parking rules require drivers to move their vehicles to the other side of the street on designated days and times to allow for street cleaning. Failure to comply with alternate side parking rules can result in parking tickets and even towing.
- NYC Manhattan Core Parking: Within the core zones of Manhattan, parking is not required of any building. This results in a lower supply of parking than the existing demand. As a result, there may be more parking tickets in these zones.

6 EXPERIMENTS AND RESULTS

As part of the evaluation method, we used a technique called hold-out validation. Essentially, we trained each model on a portion of the dataset (All the data excluding the final year in the dataset), and predicted on the remaining data. We then compared our predictions to the actual data to confirm our model accuracy. 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. Due to this factor, we opted against using certain metrics like accuracy and loss as they wouldn't be able to demonstrate the true performance of our models. Small differences in the predictions that still follow the overall trends would work against the overall accuracy. We therefore visually compared the methods, similar to our approach with VAR in class.

To better visualize and compare the results achieved from our machine learning models, we created a GUI that displays predictions for each model. Here are some screenshots of the GUI to demonstrate our visualization:

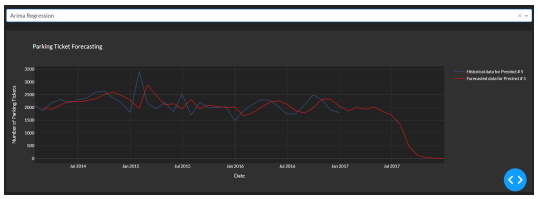

Figure 4: Forecast using ARIMA model for precinct 5

These visualizations show future predictions of the number of parking tickets in each precinct. We were unfortunately limited by the scope of the available dataset and were thus not able to work with more recent parking data. Regardless, the GUI allows users to choose a model and a precinct to predict around a year of future parking ticket data (around 3 months in the case of VAR). We also allow the user to input a future date range to visualize the precincts

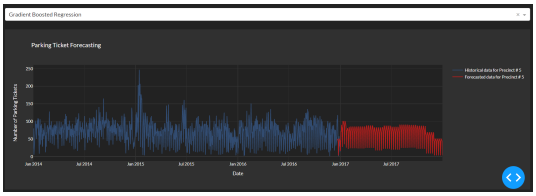

Figure 5: Forecast using Gradient Boosted Regression model for precinct 5

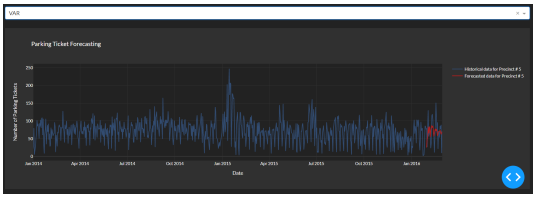

Figure 6: Forecast using VAR model for precinct 5

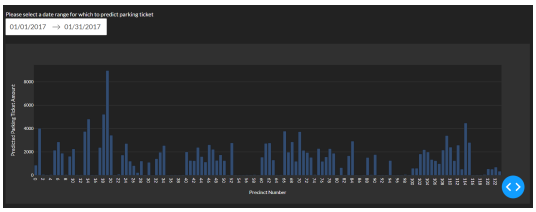

Figure 7: Parking ticket count per precinct for selected Date Range

the most number of predicted parking tickets.

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. After an initial scanning of NYC zoning laws, we identified one in particular that could contribute to this. This is the NYC Manhattan Core Parking law that states that any building in the zones do not need parking. This results in many buildings not incorporating parking into their layouts. As a result, there are less spots than the demand and an increase in illegal parking in the area. 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 clearly worked the best. It was the only model to mimic the oscillating nature of the actual dataset. This is likely due to its ability to accommodate for more volatile and nonlinear relationships. Overall, 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. No matter what we tried and how we tuned the model, the predictions would eventually reach a steady state, levelling out at a certain prediction. This isn't really effective for our use case or the data we worked with as we are looking to provide accurate long-term future predictions which just wasn't possible with the VAR model. Even in shorter ranges, it still clearly performed worse than GBR.

Finally, ARIMA performed fairly decently but failed to capture some of the trends seen throughout the data. It failed to work well with daily data but was able to model monthly aggregated data quite well which unfortunately lacks the specificity we were hoping to provide. Its ability to adapt to previously trained data and improve performance as it trained was quite impressive and generally improved the overall predictions. However, in comparison to GBR, it still fell short due to its inability to provide accurate, daily predictions.

7 CONCLUSION AND DISCUSSION

Through this project, we were able to learn about and experience working with various machine learning algorithms used for time-series forecasting. We applied VAR, ARIMA, and Gradient boosted decision trees to predict the rate of future parking violations across time and NYC boroughs. We found that each model presented its own set of advantages and disadvantages. VAR was able to capture interdependencies between multiple variables in time series data. However, it performed less optimally than the other approaches implemented. ARIMA was also flexible, simple to understand, and easily adaptable to our own purposes. It proved less effective when working with shorter periods of time, like daily patterns compared to monthly. GBDT is computationally expensive and more difficult to train compared to the other two algorithms, but it yielded the most accurate results for both shorter and longer durations of time.

The results of this project demonstrated massive potential for generating accurate predictions of parking violations. However, there are several avenues of future work we would like to explore that could further improve accuracy of the models.

- **Incorporating real-time parking data**: One pitfall of of our model design is that the data it is based on is old and cannot be constantly updated. By gathering real time data, potentially through sensor data or from parking meters, we would be able to train our models to make better short-term forecasting predictions.
- **Considering external factors**: There are several external factors that may affect patterns and trends in parking practices. For instance, inclement weather days see an increase in illegal parking. By considering external factors, like weather behavior, our models will be able to take in more variables that will yield more accurate results.
- **Testing feasibility**: One of the main purposes of our project was to be able to use the results gathered to make more informed decisions about parking infrastructure in urban areas. Part of this would include evaluating how existing New York City zoning laws affect the further development of parking systems. Since Manhattan has the most clearly-defined zoning laws, we were able to learn that buildings in this borough do not require their own parking spots, unlike other boroughs. This not only explains its high level of parking tickets issued, but also advises about what can be done to mitigate this issue in the future. Since we cannot add more parking spots, other solutions could be explored, such as making parking fees more affordable, or shared garages so businesses can still attract customers. In the future, we would like to examine more laws for other boroughs, and see how solutions could cooperate with other zoning regulations.
- **Application to other cities** New York City is one of the densest cities in America, population-wise, and is infamous for its traffic and parking issues. Many other cities are also afflicted by this issue, but not necessarily in the same magnitude or way as New York City. In the future, we would like to collect data available for other cities and apply our models to see how well they are able to predict these same values for cities with varying contributing factors. For example, we could explore options like San Francisco, Los Angeles, Atlanta, Chicago, etc. These cities all vary in density, public transportation availability, number of interstates, construction and roadwork, and general culture that could all affect parking patterns of city dwellers.

We were able to learn a lot from this project and drew a lot of useful conclusions from our results, and fortunately there are still plenty of areas to conduct future work to make more significant impacts on an issue in every urban area: parking.

REFERENCES

[1] Faraz Malik Awan, Yasir Saleem, Roberto Minerva, and Noel Crespi. 2020. A comparative analysis of machine/deep learning models for parking space availability prediction. *Sensors* 20, 1 (2020), 322.

[2] Noli Brazil. 2020. The unequal spatial distribution of city government fines: The case of parking tickets in Los Angeles. *Urban Aff. Rev. Thousand Oaks Calif* 56, 3 (May 2020), 823–856.

[3] Song Gao, Mingxiao Li, Yunlei Liang, Joseph Marks, Yuhao Kang, and Moying Li. 2019. Predicting the spatiotemporal legality of on-street parking using open data and machine learning. *Annals of GIS* 25, 4 (2019), 299–312. https://doi.org/10.1080/19475683.2019.1679882

[4] Saba Inam, Azhar Mahmood, Shaheen Khatoon, Majed Alshamari, and Nazia Nawaz. 2022. Multisource data integration and comparative analysis of machine learning models for on-street parking prediction. *Sustainability* 14, 12 (2022), 7317.

[5] Nikolaos Karantaglis, Nikolaos Passalis, and Anastasios Tefas. 2022. Deep Learning for On-Street Parking Violation Prediction. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 1–5.

[6] Nikolaos Karantaglis, Nikolaos Passalis, and Anastasios Tefas. 2022. Predicting on-street parking violation rate using deep residual neural networks. *Pattern Recognition Letters* 163 (2022), 82–91. https://doi.org/10.1016/j.patrec.2022.09.023

[7] Eran Simhon, Christopher Liao, and David Starobinski. 2017. Smart parking pricing: A machine learning approach. In *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 641–646.

[8] Adam Wenneman, Khandker M Nurul Habib, and Matthew J Roorda. 2015. Disaggregate analysis of relationships between commercial vehicle parking citations, parking supply, and parking demand. *Transp. Res. Rec.* 2478, 1 (Jan. 2015), 28–34.

[9] Xiaofei Ye, Jinfen Wang, Tao Wang, Xingchen Yan, Qiming Ye, and Jun Chen. 2020. Short-term prediction of available parking space based on machine learning approaches. *IEEE Access* 8 (2020), 174530–174541.