CS 657 Mining Massive Datasets
Project Report
NYC PARKING TICKET ANALYSIS AND PREDICTION

Instructor : Daniel Barbara

Team Members:

Yahswanth Reddy Gundepally - G01391064

Pushyami Bhagavathula - G01356145

Motivation:

If you live in New York City, you could find yourself in the following challenging situation:

You stopped momentarily in front of your office building, to get a cup of coffee from a sidewalk café. A parking ticket warrior approached your vehicle, raised her scanner, and shot an invisible beam of light at your license plate. The woman issued you a parking ticket in New York City, and that beam of light cost you $115.

According to data, New Yorkers drink 6.7 times more coffee than the national average, and the average price customers pay for a cup of coffee is $2.99, however considering the above, we may have paid far more than a few dollars for the coffee.

This has always been one of the hot topics on the web, thus we have chosen to discover the parking ticket trends and how to avoid ticket violations in day-to-day life.

Questions:

In this project, we would like to address the analysis of the patterns of violation in parking tickets with the latest dataset for the years 2019-2023. We also focused on the prediction of the location of the violation in the Newyork City depending on various factors like the day, time, issuer code, violation code, etc. After a deep dive into the dataset, our questions extended more to the characteristics of parking issued each year.

1. What are the most commonly issued parking ticket types?
2. What was the distribution of parking ticket rates over the five years (2019-2023), and was there a trend for the ticket rate to change during a day and a week?

3. What kind of vehicles has been violated and have been issued in each year comparitively?
4. How was the distribution of the parking tickets geographically in and around NYC? Do they have similar patterns across all five neighborhoods?
5. Finally, by looking at these patterns can we predict the probable location of violation in NYC given a few attributes like time of the day and issuer agency code?

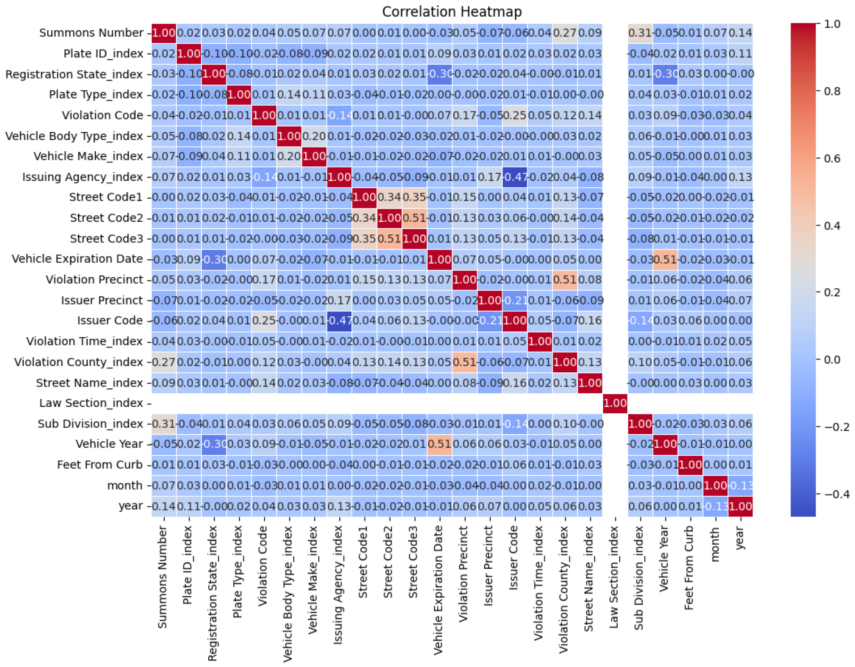
Dataset:

We have used the NYC parking tickets dataset from Kaggle. This dataset contains 8 million violations in 5 years. The data is provided by the NYC Department of Finance(DOF) where each record is issued as a violation. 51 columns included information about the summon number, issue date, violation code, violation county, violation location, etc.

Data Preprocessing and Cleaning:

Firstly, to analyze the violation data and visualization, we focussed on getting the data of multiple years together and filtering the unwanted columns. Below are the different steps we followed to clean the dataset and preprocess it.

1. After getting the data from multiple years together we first dropped the unwanted, null, and duplicate columns in the data frame.
2. Removing duplicate summon numbers: We have filtered the unique summon numbers and removed all the repeated rows.
3. Redundant Feature Elimination: The features that were not important were removed. We did this analysis by plotting a heat map and eliminating the least correlated feature with the violation_location column. Some of them were:



4. Issue Date Extraction:

   Since the issue date is an important feature we have converted the integers to DateTime format using **to_date()** and extracted three different columns containing which day of the week, month, and year using the **date_format** function for better analysis.

5. Cleaning the column Violation_location:

   Firstly, all the nan value rows in the violation location(target column) are dropped. Later we have grouped differently named similar labels of locations in NYC to have defined categories of target variables.

```
replacement_mappings = {
    'King': 'K',
    'kings':'K',
    'KINGS':'K',
    'K F':'K',
    'KING': 'K',
    'Ks':'K',

    'Qns': 'Q',
    'QN':'Q',
    'QNS':'Q',
    'Qunees': 'Q',

    'Bronx': 'B',
    'BX':'B',
    'Bron': 'B',
    'BK':'B',

    'Rich': 'R',
    'RICH':'R',

    'MN': 'NY',
}
```

6. Encoding the string labels to numerical values:

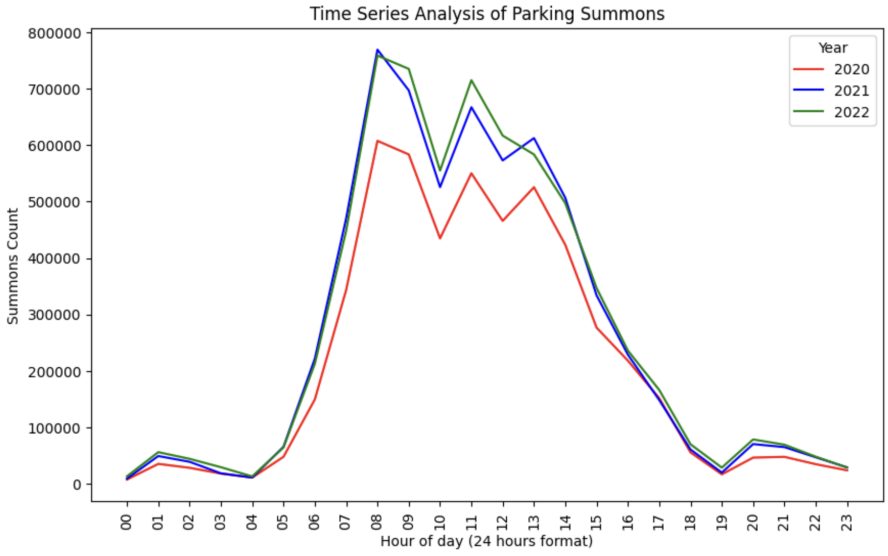
   Finally, all the string features in each column are encoded to integers mapped in the form of a matrix using one-hot encoding for categorizing features containing more than two categories whereas Stringindexer is used to categorize binary categories in a feature.

Exploratory Analysis:

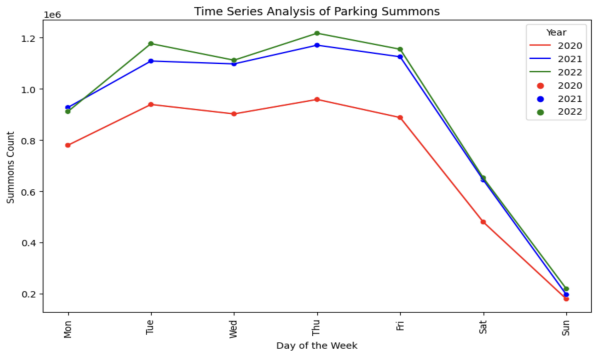
To questions mentioned in the introduction, we have analyzed the data and have brought out some valuable insights. Below is the analysis:

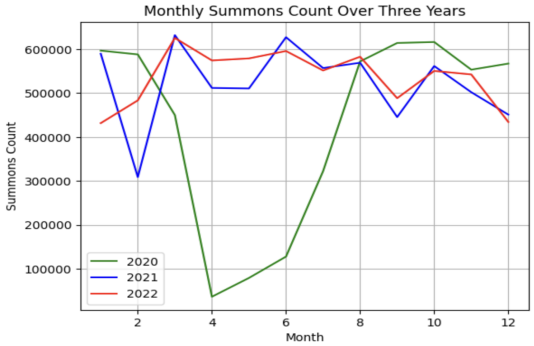
The most commonly issued parking ticket types:

Parking Ticket Rates Vs Time: Tickets are predominantly issued during the daytime, specifically between 6 am and 4 pm, indicating a concentration of enforcement activity during regular business hours. This suggests a focus on monitoring and penalizing violations that occur throughout the typical workday, aligning with increased vehicular and pedestrian traffic. The higher issuance during these hours underscores the strategic timing of enforcement efforts to address violations during peak activity periods.



Parking Ticket Rates Vs Day of the week: During weekends, the issuance of tickets is noticeably lower compared to weekdays, suggesting a reduced level of enforcement or fewer violations during non-working days. Additionally, around mid-2020, there is a significant decrease in ticket issuance, likely attributable to the COVID-19 lockdown measures, which impacted overall mobility and contributed to a decline in observed violations.





Vehicle Type Vs Number of Violations: The analysis of vehicle counts by plate type reveals consistent patterns over the years. In 2020, the most common plate types were PAS (passenger) and COM (commercial), with 3,737,989 and 916,145 counts, respectively. In 2021, PAS maintained its dominance, recording 4,712,108 counts, followed by COM with 990,626 counts. The trend continued in 2022, with PAS again leading with 4,896,667 counts, emphasizing its prevalence in the observed vehicle data. These insights provide a comprehensive understanding of the distribution of plate types across the specified years.

Vechiles count by plate type in year : 2020

Plate Type	count
PAS	3737989
COM	916145
OMT	209456
SRF	56678
OMS	50341
APP	45222
LMB	19249

only showing top 7 rows

Vechiles count by plate type in year : 2021

Plate Type	count
PAS	4712108
COM	990626
OMT	202339
SRF	75541
OMS	60986
LMB	60326
APP	46710

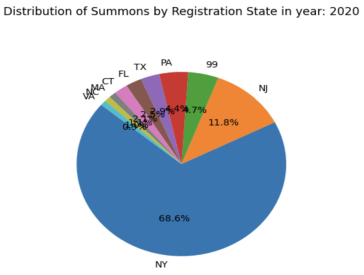
only showing top 7 rows

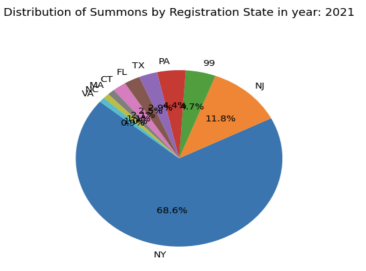
Vechiles count by plate type in year : 2022

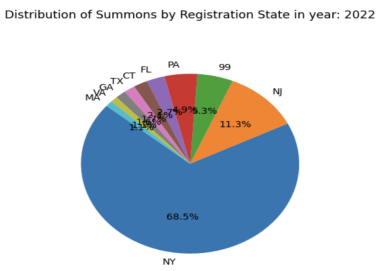
Plate Type	count
PAS	4896667
COM	994115
OMT	199208
SRF	85119
OMS	63173
APP	50912
LMB	41177

only showing top 7 rows

Distribution of tickets in different states: The majority of ticketed vehicles in New York are registered within the state, indicating a higher occurrence among local residents. Conversely, there are fewer tickets issued to vehicles registered in affluent counties such as Richmond (Staten Island), Kings (Brooklyn), Queens, and the Bronx. This observation suggests a localized pattern of traffic violations within the city, with a notable decrease in ticketed vehicles from surrounding and potentially wealthier boroughs.







Prediction Models and Performance Evaluation:

From the above analysis, we might have a clear idea of what are the important features and affecting features for predicting violation location.
For prediction, we use the violation location likely with the categorical features that are converted into numerical data using StringIndexer and Onehot encoder.

While we have tried scaling the features between 0-1 using StandardScalar, it has not given better performance when compared with the ones that are not scaled.

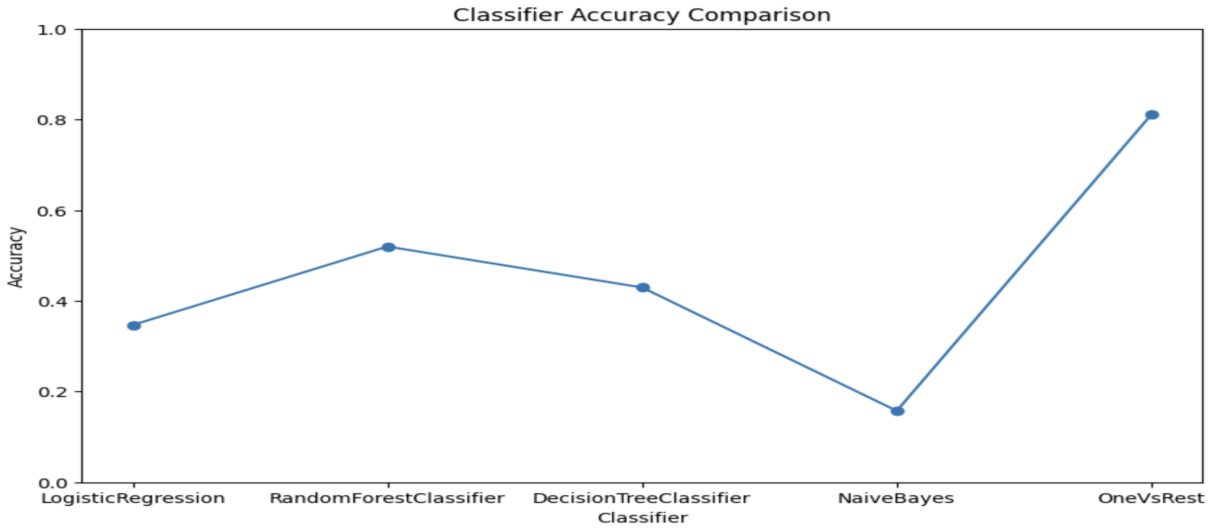
Later using VectorAssembler helped in putting all the features together and facilitating a model input for the training of the data.
We later used a pipeline to streamline the transformed data that is the assembled features, feature vector, and model classifier.

Model Training and Selection:

We have used two models to predict the locations of violations of parking.

1. **Random Forest Classifier**: Utilizes an ensemble of decision trees to make predictions, providing robustness and accuracy through aggregating multiple models.
2. **Decision Tree Classifier**: Builds a tree-like model by recursively partitioning the dataset based on features, making it interpretable and suitable for capturing complex decision boundaries.
3. **Naive Bayes Classifier**: A probabilistic model based on Bayes' theorem, assumes independence between features, making it efficient and effective for text classification and simple yet powerful for multiclass problems.
4. **OneVsRest with GBTClassifier**: One-vs-Rest with GBT is an ensemble learning method for multi-class classification, training binary classifiers for each class. The GBTClassifier, within a OneVsRest framework, extends gradient boosting to effectively handle multiple classes in PySpark.

Using PySpark MLlib library inbuilt methods we have tried building the model. There are also other methods that we would commonly use to train the model i.e. **fit** method and to extract the predictions we use **transform** method.



Conclusion :

In our detailed data analysis, we have successfully employed various classifiers, with the One vs Rest approach using the GBT classifier emerging as a particularly effective model, achieving an accuracy of 80%. This highlights its robustness in predicting violation locations. Our findings also reveal distinct temporal patterns, emphasizing increased vigilance during morning hours and a comparatively lenient approach to ticketing on weekends. These insights offer a nuanced understanding of the dynamics surrounding violation occurrences in NYC. This analysis equips stakeholders with actionable information for targeted interventions, enabling more informed urban governance strategies and resource allocation. Overall, our study not only provides valuable predictions but also contributes to a comprehensive understanding of the factors influencing violation patterns in the city.