311 Service Requests Maximum Resource Utilization and Resolution Time Prediction

A Project Report

For

CS-GY 6053

Foundation of Data Science

By Prof Rumi Chunara

Group Members:

Nikhil Soni (ns6062)
Sumedh Sandeep Parvatikar (sp7479)
Kaartikeya Panjwani (kp3291)

PROBLEM STATEMENT

The departments responsible for complaints from 311-service requests can often delay the response due to unavailability of resources. Hence, the primary objective of our project is to maximize resource utilization and predict resolution time.

- To identify various types of service requests by complaint types and location (to analyze the trends).
- Assess the seasonal variations (Winter, Summer, Autumn, Spring) of complaints to observe any unusual spikes indicating systemic issues.
- To create 2 models, the Regression model to predict the resolution time of a complaint and the Classification model to predict whether extra resources are needed to be dispatched for that complaint at that time or not.(considering additional data from ZipCode and Holiday datasets).

TARGET VARIABLE(S)

1. Resolution Time (For regression): Number of days it takes to resolve a complaint

2. Additional resources required (For Classification): This is a binary variable which indicated if there is a need for dispatchment of additional resources for that complaint or not.

BACKGROUND

The 311 service is now a vital conduit for citizens to report problems and concerns to local authorities in the busy urban environment of New York City. The resolution time, or the amount of time that passes between the start of a service request and its completion, is one important factor that has a direct impact on how effective this service is. The main goal of this project is to create a machine learning model that can forecast if a service request's resolution time will be longer than the typical departmental resolution time.

This project's main objective is to use machine learning algorithms to develop a predictive model that will improve the city's capacity for efficient resource management and allocation. The model focuses on forecasting whether the resolution time will surpass the departmental average in order to offer practical insights for response strategy optimization and prioritization. By being proactive, the city can ensure that issues reported by citizens are resolved more quickly and with more efficiency and effectiveness.

Feature engineering, model training, and evaluation with historical data are all part of the methodology. We will investigate a range of machine learning algorithms, including ensemble techniques and regression models, to determine the best method for precisely projecting resolution times. The success of the project depends on its capacity to develop a strong predictive model and to understand the model's results in a way that allows local officials to implement the suggestions.

In a nutshell this data science project is an innovative endeavour that aims to maximize the speed at which NYC 311 service requests are resolved by utilizing machine learning. The

initiative aims to enhance civic involvement, resource allocation, and overall efficiency in meeting the different needs of New York City residents.

DATASETS

Our 311-Service Requests dataset, which contains 34 million service request records from 2010 to the present, forms the foundation of our data science initiative. After a thorough pre-processing step, the dataset—which originally contained 41 features—was reduced to 27 essential columns while preserving the most important data, including the type of request, its location, timestamps, and the estimated resolution time. Important details such as the type of request, the location, and the dates of the case's opening and closing are included in every entry of the dataset. The time it takes to resolve each case is quantified by means of the newly introduced variable called "resolution time," which is computed by taking the difference between the start date and the closed date.

This dataset provides a thorough archive of problems that citizens have reported, providing insightful information on the difficulties that citizens of New York City encounter. The ZipCode and Holiday databases supplement this main dataset. Spatial dimensions are included in the ZipCode dataset, which offers more information on the population, zip codes, and local characteristics. The Holiday dataset, on the other hand, provides a temporal context through the cataloguing of federal holidays for the years 2010 through 2023. Combining these information yields a multifaceted, complex picture of service demands that includes textual, numerical, category, spatial, and temporal data types.

A thorough pre-processing step was conducted- to prepare the datasets for effective analysis. The depth of data necessary for significant insights is preserved while ensuring computing efficiency through this technique. The combination of these datasets not only provides the foundation for our analysis, but it also helps us move closer to creating a resolution time prediction model. This project has the potential to advance scholarly knowledge of the dynamics of civic involvement as well as offer useful advice for maximizing resource allocation and service delivery in New York City's intricate metropolitan environment.

Name of Column	Data Type	Data Characteristic
POPULATION	float64	Numeric
Agency	Text	Categorical
Complaint Type	Text	Categorical
Location Type	Text	Categorical
Address Type	Text	Categorical
City	Text	Categorical
Community Board	Text	Categorical
BBL	float64	Numeric
Borough	Text	Categorical
Open Data Channel Type	Text	Categorical
Latitude	float64	Spatial
Longitude	float64	Spatial
Resolution Time	float64	Numeric
Created_Date_Year	int64	Numeric

Created_Date_Month	int64	Numeric
Created_Date_Day	int64	Numeric
Created_Date_Hour	int64	Numeric
Created_Date_Minute	int64	Numeric
Created_Date_Second	int64	Numeric

DATA SOURCES

a) 311-Service Requests: https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

b) ZipCode Data : https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u/data?no_mobile=tr

c) Holiday Data :
i) 2010 - 2020 : https://data.world/jennifer-v/us-holiday-dates-2010-2020
ii) 2021-2023 : https://www.generalblue.com/calendar/usa/us-holidays-2022

DATA PREPROCESSING

The stage of data preprocessing was essential in streamlining the unprocessed datasets and guaranteeing a clear and rich basis for our research that followed. The primary procedures used to improve the caliber and applicability of the data are outlined in the phases that follow:

i) Column deletion for NaN values:
To streamline the dataset, columns with a majority of NaN values were systematically removed. This initial cleaning step aimed to eliminate variables with insufficient data, focusing on preserving the most informative attributes.

ii) Filtering Rows Without Closure Dates:
Rows absent In order to highlight the significance of closed cases for examining resolution timelines, closure dates were removed. By taking this step, it was made sure that only comprehensive and pertinent records were given further thought.

iii) NaN Value Imputations:
NaN values were strategically and carefully replaced to maintain data integrity. 'Address Unknown' was assigned to NaN records in columns where this label already existed, ensuring consistency. Similarly, 'UKNOWN' replaced NaN values in the city column, aligning with the dataset's existing conventions.

iv) Department Filtering Based on Pareto's Principle:
Applying Pareto's principle, the focus narrowed to the top 5 departments that collectively represented 80% of the data. This strategic filtering allowed for a more concentrated and impactful analysis, concentrating efforts on the most significant contributors.

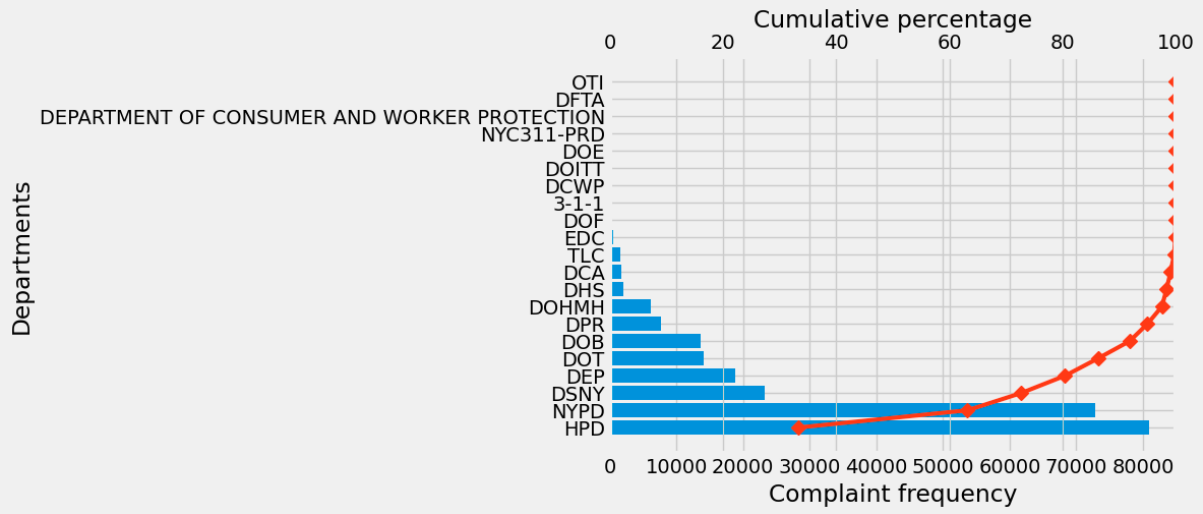

Fig. 1: Pareto's Chart of Departments vs Number of Complaints

v) Exclusion of Records with 0 Population and NaN Values:
Exclusion of records with zero populations and NaN values from the study improved the quality of the data by getting rid of entries that contained incomplete or unnecessary information.

vi) Exclusion of Records with Area <= 0 and NaN Values:
A similar approach was applied to records with an area less than or equal to zero, ensuring that only meaningful spatial data contributed to the subsequent phases of analysis.

vii) Dataset Merging:
The ZipCode and 311 datasets were seamlessly merged, combining spatial details with service request records to enrich the dataset comprehensively.

viii) Irrelevant Column Removal in Data Reduction Phase:
In order to further streamline the dataset and preserve just the most important aspects for analysis, unnecessary columns were methodically removed throughout the data reduction step.

FEATURE ENGINEERING

The feature engineering stage was critical in identifying target variables and boosting the dataset's predictive potential as we worked to develop strong classification and regression models for our NYC 311 service request research.

i) Regression Model Target Variable - Resolution Time:
The primary goal for our regression model was to predict the resolution time of complaints. This target variable was derived by computing the time difference between the Closed Date and Created Date attributes, providing a quantifiable measure of the duration taken to resolve each service request.

ii) Classification Model Target Variable - Additional Resources Required:
For the classification model, a new class label, Additional Resources Required, was introduced. This binary label (1 or 0) was based on department-wise average resolution times.

If the resolution time for a complaint exceeded the departmental average, the record was labeled as 1; otherwise, it was labeled as 0. This approach aimed to identify cases requiring additional resources, facilitating a targeted and proactive response.

iii) Temporal Feature Enhancement:

The Created Date attribute was split into day, month, year, and weekday components. The original Created Date attribute was then dropped. This temporal decomposition enables the model to consider finer-grained time-related patterns and dependencies.

iv) Outlier Removal:

Resolution time values greater than the 90th percentile were recognized as outliers and eliminated in order to increase the robustness of the model. By ensuring that the model was not overly influenced by extreme values, this phase helped to produce predictions that were more trustworthy.

Together, these feature engineering processes gave the dataset an intricate and predictive form, which prepared the way for the development of efficient regression and classification models. The feature engineering stage greatly aids in the project's overall objective of streamlining civic service resolution procedures by extracting major target variables and enhancing the dataset with temporal components.

DATA ANALYSIS

i) We start by converting the target variable Resolution Time into hours from seconds by dividing it by 3600.

ii) Next, we calculated the number of complaints in each borough and found out the following result.

Borough	**Number of Complaints**
BROOKLYN	73604
QUEENS	53083
BRONX	50372
MANHATTAN	47395
STATEN ISLAND	10861

Table 1: Number of Complaints by Boroughs

iii) For determining the most frequent complaint types, we plot a bar chart for the same as follows.

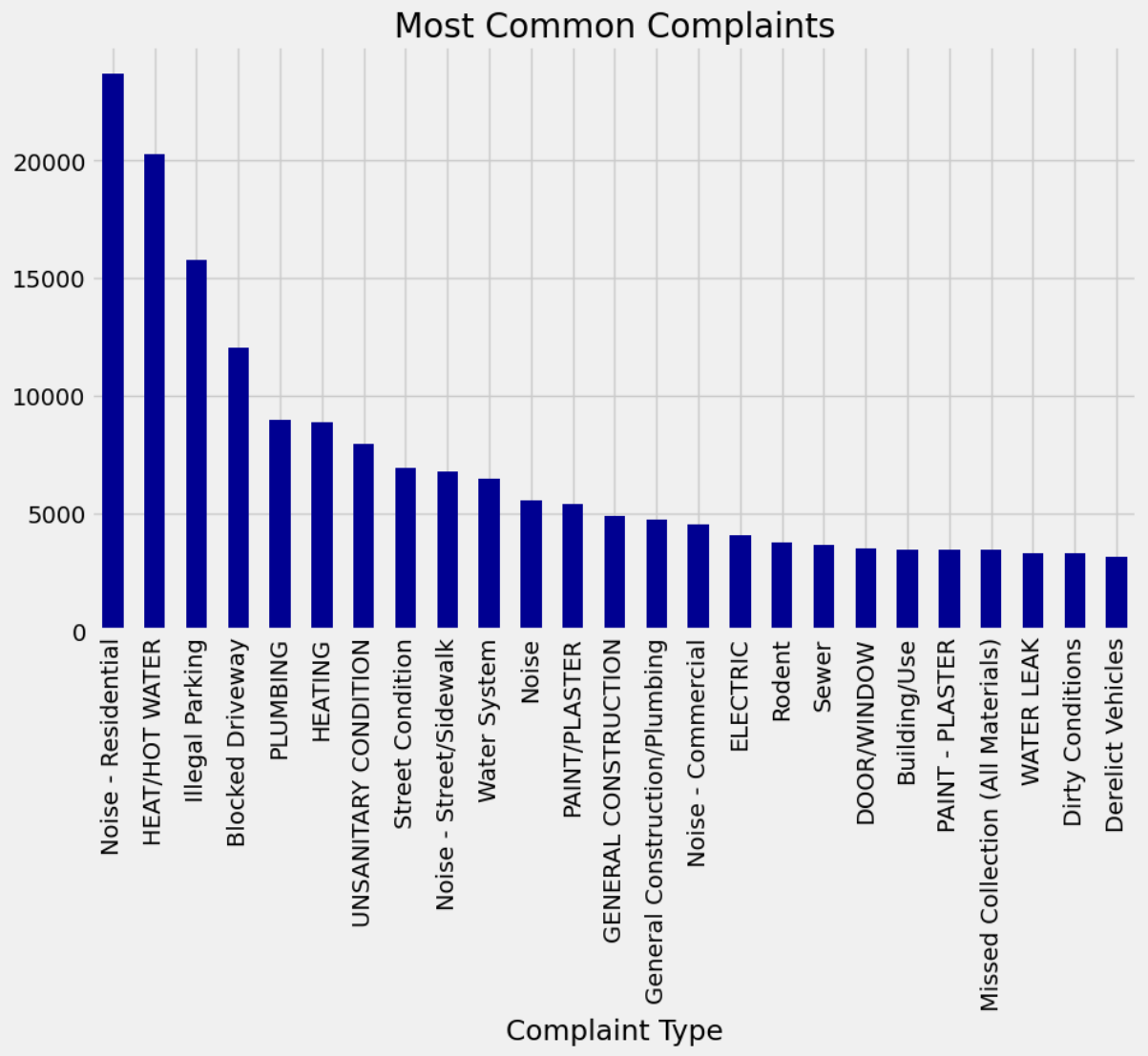

Fig. 2: Most Common Complaint Types

iv) Next, to visualise the concentration of complaints across New York City, we plot a heat map

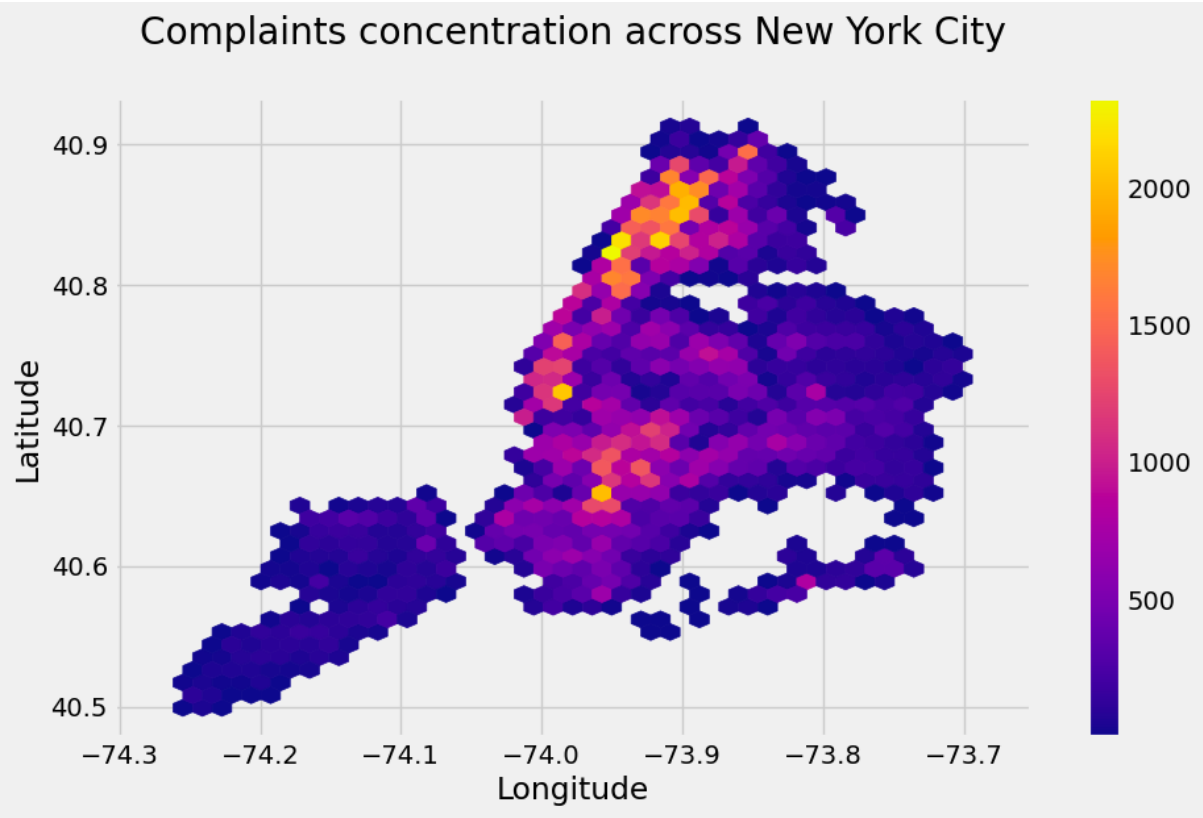

Fig. 3: Complaints Concentration across NYC

v) We then plot the pie chart of top 10 Complaint Types

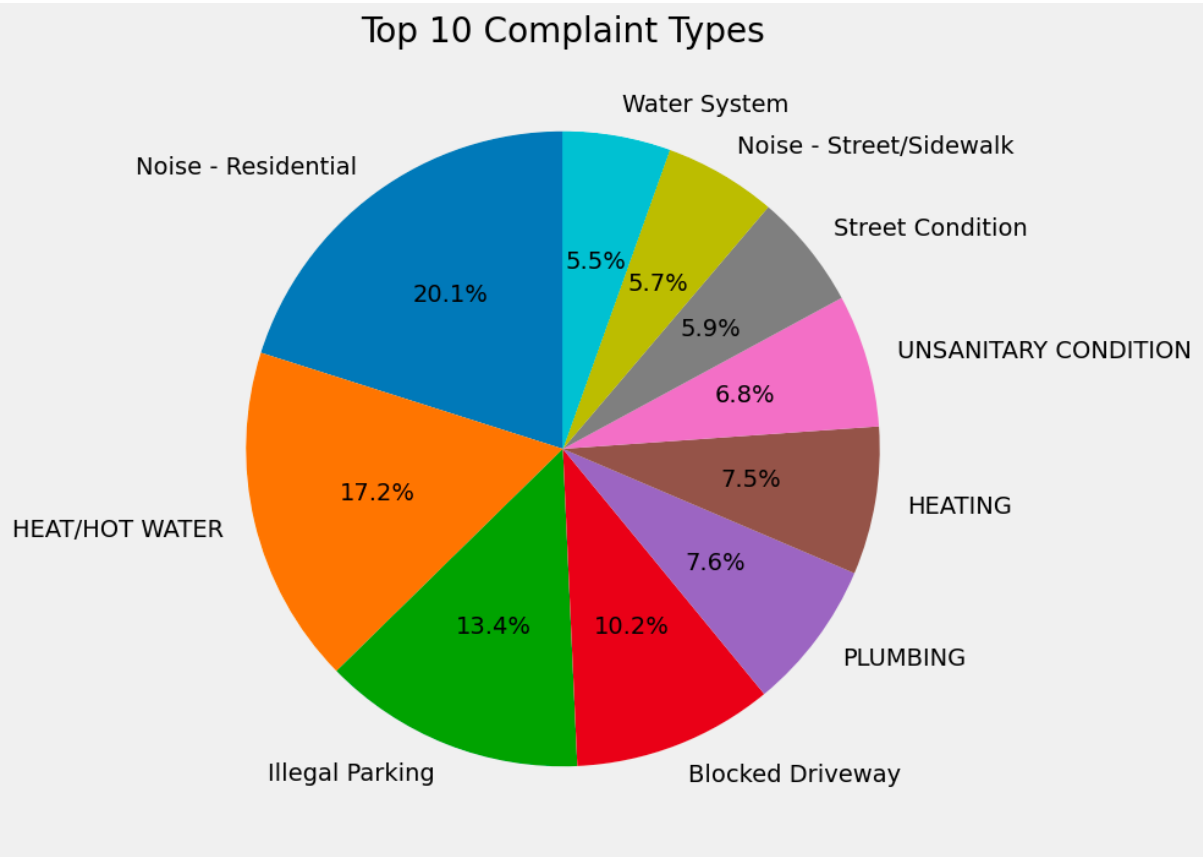

Fig. 4: Top 10 Complaints across NYC

vi) Next, we aim to find seasonal variations in our data. We make a new dataframe which contains the data from 2020 to 2022. This data is taken from the sample to present an illustration explaining the seasonal trends of complaints in a time period of 3 years. We create a new column 'Season' which contains the season the complaint is recorded in. We determine this by assigning a season to the record according the complaint's created month.

vii) Next, we plot the bar plot of complaint counts by seasons.

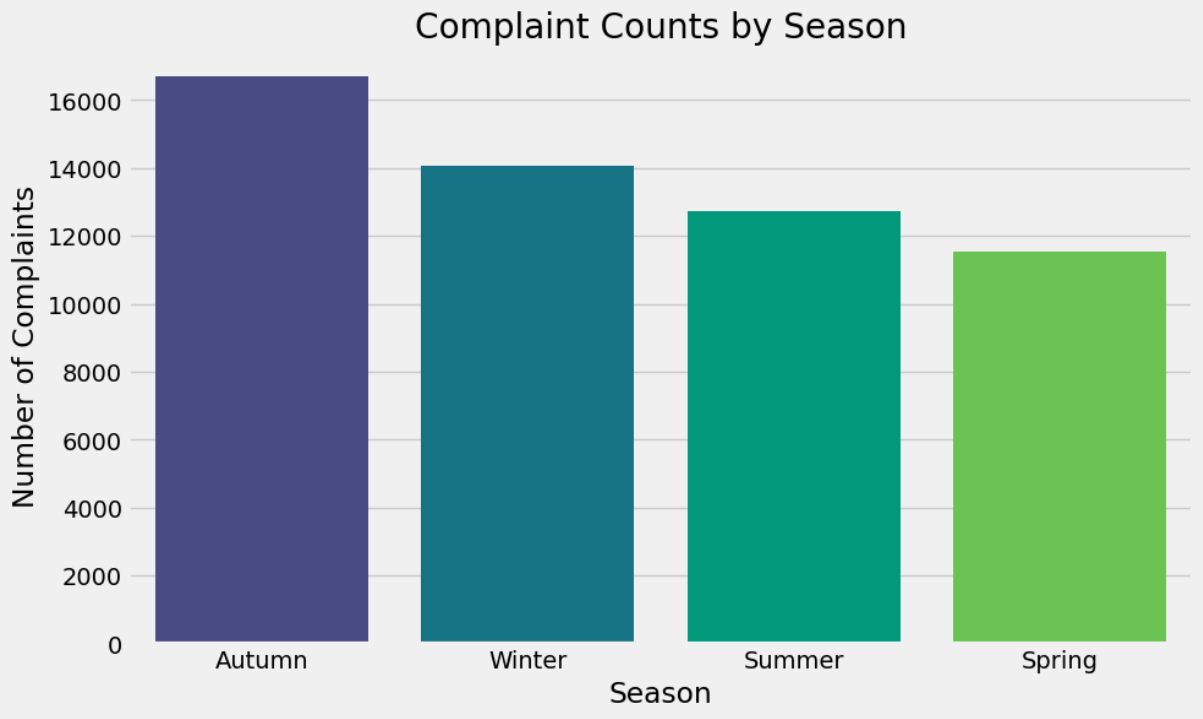

Fig. 5: Complaint Counts across NYC by seasons in the years 2020, 2021 and 2022

viii) As we can infer from the plot, the maximum complaints reported from 2020 to 2022 in the sample are in the Autumn season.

ix) Further, we divided our data borough-wise by creating a dataframe for each borough and within each borough we calculate delays in resolution time for different departments.

x) Next, we Dividing the data into 2 parts (less than and greater than 80th percentile) to filter out delays and plot a stacked bar plot for the same.

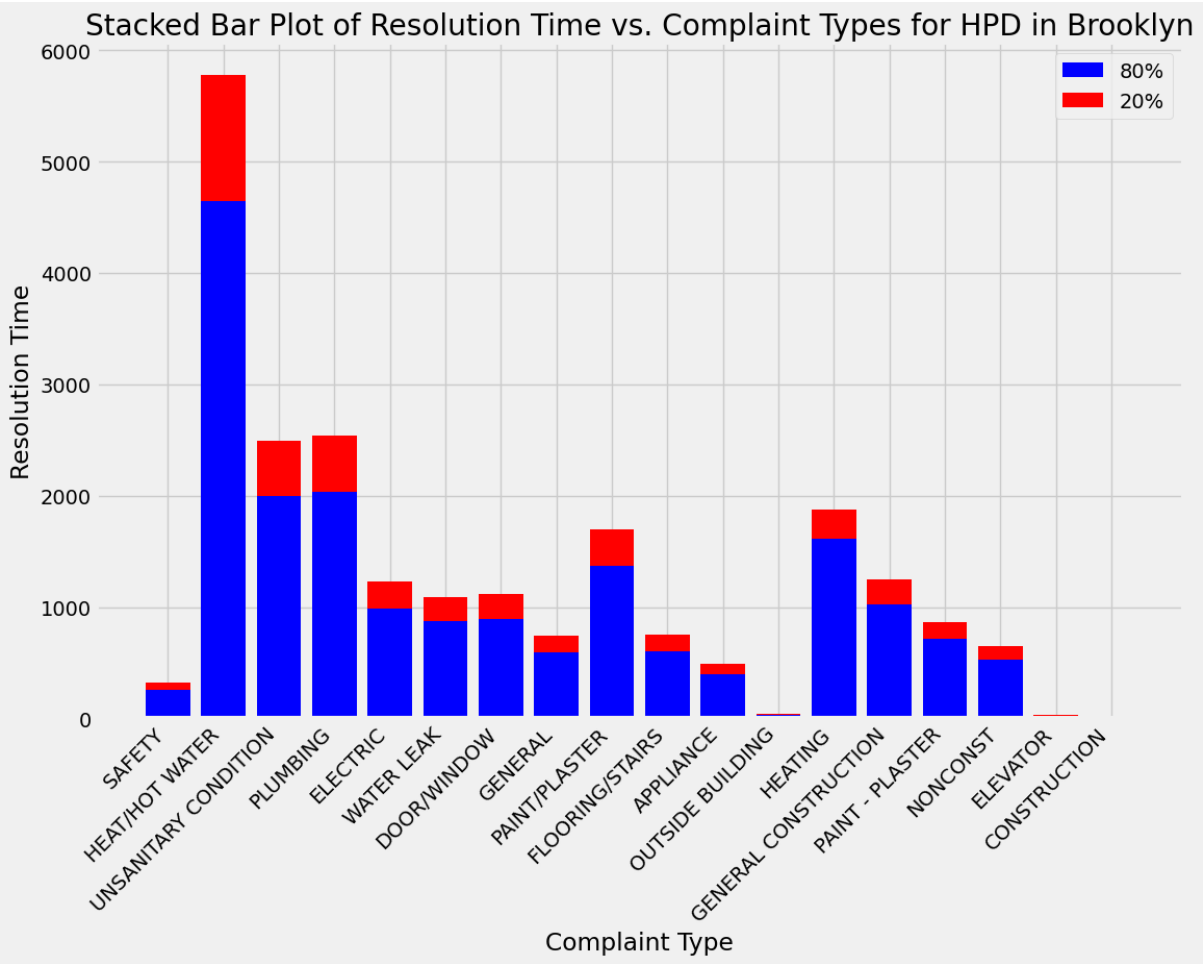

Fig. 6: Resolution Time vs Complaint Types for department HPD in Brooklyn

xi) Next, we calculate the delay scores for each department in each borough which is defined by:

$$delay\ score = 80th\ percentile\ of\ resolution\ time\ values \div total\ resolution\ time\ values$$

xii) Higher the delay score, greater is the resolution time for departments. After extensively calculating delay scores for all boroughs for their top 5 departments, we plot a bar chart for a single department, New York Police Department.

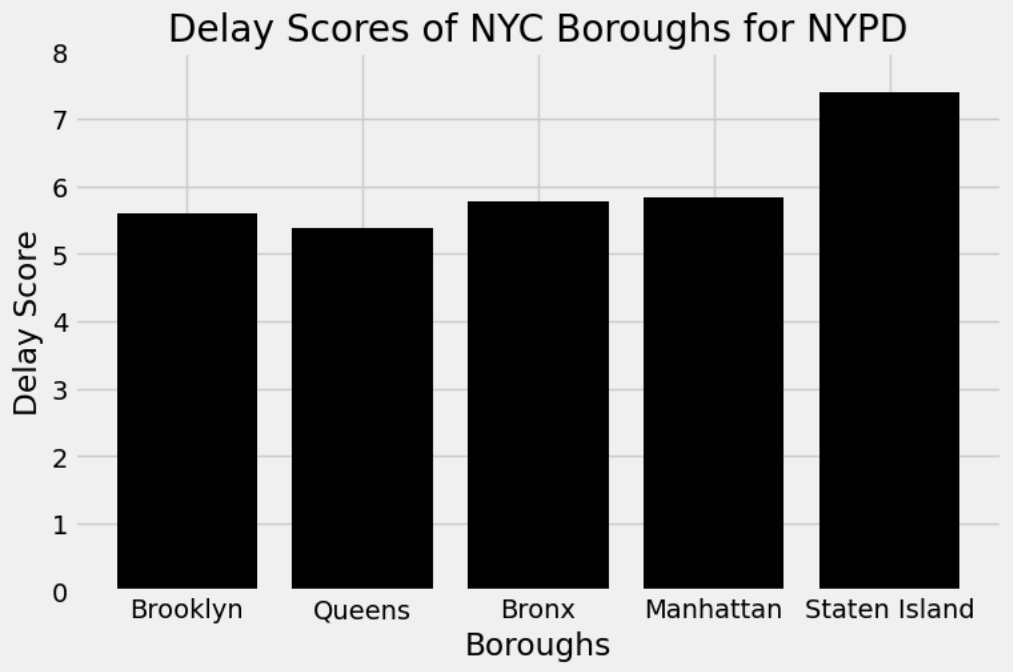

Fig. 7: Delay Scores of New York City Boroughs for NYPD

MODEL SELECTION

We've Covered two modelling techniques names regression and classification.

i) For regression, we've used the following models:

a) Ridge Regression: Because ridge regression can manage multicollinearity among the predictor variables, it is very useful for datasets such as 311 service requests. Ridge regression introduces a regularization component that penalizes large coefficients, which helps minimize the problem of high collinearity in the case of 311 service calls, where multiple factors may be coupled or correlated. By stabilizing the regression coefficients and preventing overfitting, this regularization term—which is managed by the hyperparameter alpha—creates more reliable and comprehensible models. Ridge regression helps to improve the model's generalization performance in scenarios where the dataset shows high dimensionality and interdependence between features, which makes it an effective tool for deciphering and forecasting intricate patterns in 311 service request data.

b) ElasticNet: As ElasticNet regression can handle high-dimensional data with possibly linked attributes, it is useful for datasets such as 311 service requests. ElasticNet combines L1 and L2 regularization in the context of 311 service requests, where multiple variables may impact the result. This enables it to handle multicollinearity and execute feature selection efficiently. This reduces the chance of overfitting and aids in identifying the most pertinent variables influencing the results while working with various service request data. ElasticNet's flexible parameter tweaking also enables users to modify the ratio of L1 to L2 regularization, making it a useful tool for striking a balance between robustness and model sparsity for handling the intricate 311 service request datasets.

c) Random Forest Regressor: Considering its versatility and capacity to handle intricate, non-linear relationships, the Random Forest Regressor is very useful for datasets like

as 311 service requests. It can also capture interactions between different characteristics. Random Forest performs exceptionally well at modelling complex patterns in the data when it comes to 311 service requests, where a variety of factors may impact response times or service outcomes. Furthermore, by merging numerous decision trees, its ensemble nature improves generalizability and prediction accuracy while reducing overfitting. This is especially important for 311 service request datasets, which frequently display a variety of dynamic patterns driven by time, location, and other service-related aspects.

d) XG Boost: XGBoost, a gradient boosting implementation, is especially useful for datasets like 311 service requests because of its ability to handle complex, non-linear relationships and capture intricate patterns within the data. By creating decision trees one after the other and tuning them to reduce prediction errors, it performs exceptionally well in predictive modelling tasks. Given the complex and dynamic nature of the relationships between different features and the service request outcome in the context of 311 requests, XGBoost's ensemble approach and flexibility make it an excellent choice for pattern recognition, factor identification, and highly accurate service request outcome prediction. In addition, XGBoost's ability to manage huge datasets and regularization strategies to avoid overfitting enhance its efficacy in handling the varied and dynamic characteristics of 311 service request data.

e) GB Regression: When it comes to integrating different characteristics and their relationships in the context of 311 service requests, where multiple factors may impact response times or resolution rates, GB Regression performs exceptionally well. It adapts to the complexities of the dataset by building a series of weak learners one after the other, fixing each other's mistakes. This adaptability is essential in situations where the target variable (response time) and input parameters (such as location or request type) may not have a straight line relationship. In the dynamic and complex setting of 311 service requests, GB Regression optimizes resource allocation and response tactics by utilizing ensemble learning and boosting to increase forecast accuracy.

ii) For Classification, we've used the following models:

a) Decision Trees: Decision trees' interpretability, simplicity, and capacity to handle both numerical and categorical data make them useful for assessing datasets such as 311 service requests. Decision Trees are an efficient way to divide data in the context of 311 service requests depending on pertinent variables like location, request type, time, and other pertinent criteria. A variety of factors may contribute to different sorts of service needs. Furthermore, decision trees can capture intricate interactions between variables, offering transparent and easily available insights into the critical elements determining the form of service requests. This is important for enhancing public services and the distribution of resources.

b) Random Forest: Random Forest's capacity to handle complicated, high-dimensional data with a wide range of attributes makes it especially useful for datasets such as 311 service requests. Random Forest does a great job of capturing complex linkages and interactions among characteristics in the context of 311 service requests, where several factors may influence the categorization of requests. Because of its ensemble method, which combines predictions from several decision trees, it is more resilient

and less prone to overfitting, which makes it ideal for noisy and diverse datasets like those seen in urban service requests. Moreover, Random Forest offers feature importance measurements that help identify important aspects impacting the outcomes of service requests. This is helpful for comprehending trends and allocating resources optimally in urban management systems.

c) GB Classifier: Gradient Boosting (GB) classifiers are advantageous for a dataset such as 311 service requests because of their high-dimensional feature spaces and ability to handle intricate, non-linear relationships. GB classifiers are particularly good at identifying complex patterns and producing precise predictions in the context of 311 service calls, where a variety of parameters (such as location, issue kind, and prior trends) may influence the nature of the request. These models are strong at managing heterogeneous data sets, supporting both category and numerical characteristics that are frequently included in datasets containing service requests. In the context of urban services like 311 requests, GB models are useful tools for decision-making and resource allocation because of the interpretability of feature importance, which further helps in understanding the major factors impacting service requests.

d) KNN Classifier: KNN can efficiently utilize the geographic data linked to each 311 service request, as location might be really important in comprehending and resolving difficulties like infrastructure issues or demands for public services. Based on previous patterns of comparable incidents in the area, KNN can identify and forecast the type of a current service request by comparing its spatial coordinates to those of its nearest neighbours. Finding patterns in service requests within particular neighbourhoods or regions can be very helpful in this regard, as it allows for more targeted and effective resource allocation for prompt and successful municipal replies.

Evaluation Metrics:

- **R2 Score:** It is commonly used for regression models and signifies the fit of the models. It also allows for the comparison of different models for the same dataset.
- **Accuracy:** Accuracy is used to evaluate the performance of a classification model. It is a measure of how well the model correctly predicts the class labels of the instances in the dataset and is easy to understand and interpret without requiring complex calculations.
- **Precision:** We have used precision to evaluate the performance of our classification mode. Precision is defined as the number of true positive predictions divided by the sum of true positives and false positives.
- **Recall:** We have used recall to identify whether the model correctly predicts the positive samples
- **F1 Score:** F1 score provides a balance between the two, making it a suitable metric when there is a need for a trade-off between precision and recall. It provides a single value that summarizes the performance of a model in binary classification. It simplifies the evaluation process and is especially useful when comparing different models or tuning hyperparameters.

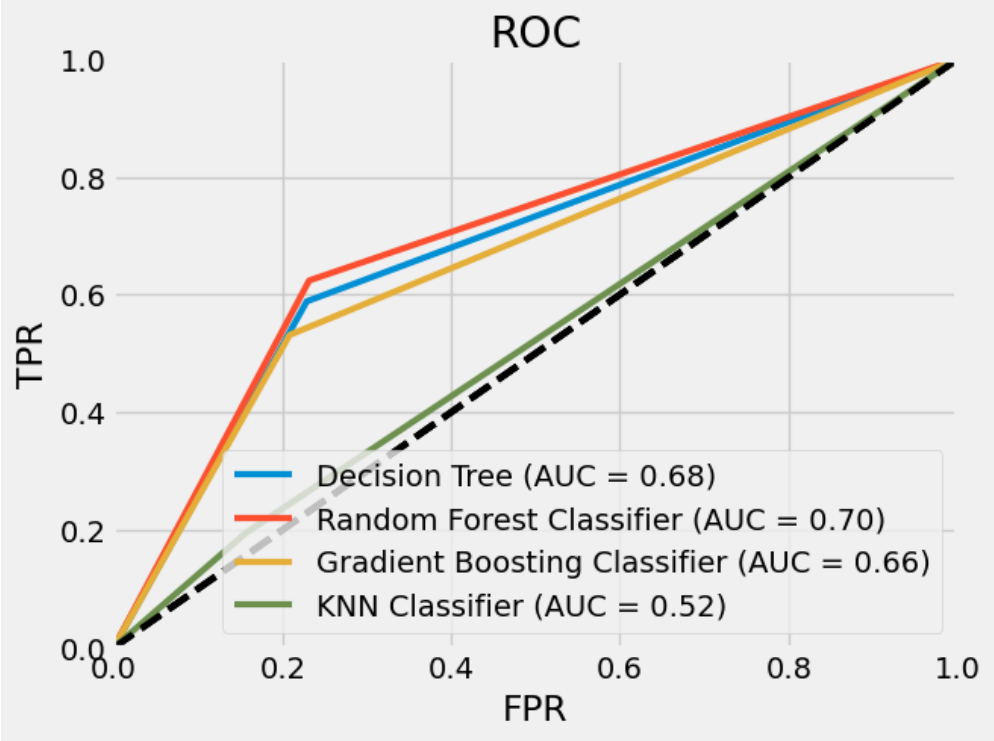

Fig 8: AUC-ROC Curve for classification models (Choppy ROC curve because of skewed data)

ASSUMPTIONS

- All complaints are reported with equal probability.
- Assuming negligible change in the population over the course of 10 years for the joining of ZipCode dataset.
- The columns with mostly NaN values do not have much significance
- Assuming that the departments accurately keep track of their records.
- The resolution time values above the 80th percentile are assumed as delays.

LIMITATIONS

- The scope of our analysis is limited to top 5 departments.
- We have limited our project to indicate resource requirement and have not quantified it.

FUTURE SCOPE

- The future scope of this project can include quantifying the amount of resources required.
- Time series prediction.

DAVIATION FROM PROPOSAL:

- Instead of classifying if the response time is within 24 hours or not, we classified a delay if time taken was greater than the average time taken by the respective department to cater to that complain. So that we could better classify the need of additional resource requirement.
- Updated Assumptions.
- Calculation of Delay Score.
- We adhered to the guiding of Pareto's principle.

INDIVDUAL CONTRIBUTION:

- **Nikhil Soni:**
  1)Problem Statement formulation and roadmap creation
  2)Data Sampling and Validation (Validating trends on the large dataset in chunks)
  3)Data integration (zipcode and holidays)
  4)Regression Modeling and tuning

- **Sumedh Parvatikar:**
  1) Data cleaning
  2) Data reduction
  3) Classification
  4) Result interpretation and hyper parameter tuning
- **Kaartikeya Panjwani:**
  1) Data Selection
  2) Complaints analysis and plots
  3) Department delay score analysis
  4) Season and holiday analysis