**NYC 311 CUSTOMER SERVICE REQUESTS ANALYSIS**

|  |  |
| --- | --- |
| NAME | UCID |
| PRABHAKAR YADAV | PY98 |

ABSTRACT: This study embarks on an analytical journey through the New York City (NYC) 311 Customer Service Request system, aiming to enhance municipal service delivery by predicting request resolution times. Utilizing a rich dataset of service requests, the research focuses on developing a predictive model that incorporates a variety of features, including the type and location of the request, as well as the responsible agency. The model leverages machine learning algorithms and classification metrics to evaluate its predictive accuracy, thereby aiding city administrators in strategic decision-making. The multi-dimensional analysis encompasses both quantitative data and qualitative aspects, enabling a comprehensive understanding of factors affecting service efficiency. The insights derived promise to aid city officials in resource optimization and offer residents a more responsive governance system. Ultimately, this report suggests that such a predictive model could significantly elevate the quality of urban life and reshape the future of municipal services to proactively meet the demands of a growing city.

INDEX TERMS Machine learning, Classification, Customer service request, Accuracy.

I. INTRODUCTION

This report presents a comprehensive analysis of the New York City (NYC) 311 Customer Service Request system, a pivotal interface between the city’s populace and its municipal services. NYC 311 facilitates reporting and addressing a range of non-emergency issues, thereby serving as a critical gauge for the needs and concerns of New Yorkers. The focal challenge of our study is to predict the resolution time for these service requests, taking into account multifaceted features such as the nature of the complaint, its geographical location, and the pertinent city agency responsible for the resolution.

By harnessing the extensive dataset accumulated by the NYC 311 system, which encapsulates millions of service requests, our analysis aims to discern patterns and trends that can inform strategies to bolster the quality and efficiency of city services. This inquiry not only spotlights the current state of municipal service delivery but also seeks to identify opportunities for enhancement, with the ultimate goal of elevating the living standards of the city's residents.

The report details the methods employed for data collection and analysis, discusses the key insights gleaned from the data, and outlines the potential hurdles encountered during the investigation. The emerging trends and the substantive conclusions drawn from this analysis are poised to guide actionable recommendations. These recommendations aim to refine the NYC 311 system, thereby fostering a more dynamic, efficient, and community-focused governance structure. Through this endeavor, we envision a transformed city infrastructure that is increasingly attuned to the rhythm of its citizens' needs and committed to the relentless pursuit of service excellence.

In an era where city populations are rapidly growing and urban management complexity is increasing, the need for efficient municipal service delivery has never been greater. This report delves into the intricacies of the New York City (NYC) 311 Customer Service Request system, which stands as a vital link between the city's inhabitants and its various municipal agencies. The primary aim of our study is to develop a predictive model that estimates the closure time of service requests based on a multifaceted array of features, including the type of request, the location it originates from, and the agency responsible for its resolution.

The report meticulously utilizes classification metrics and machine learning algorithms to assess the model's effectiveness and reliability in forecasting service request resolution times. These metrics are integral to quantifying the precision with which the model can predict outcomes, thereby enhancing the strategic decision-making capabilities of city administrators. Through the analysis of historical data, encompassing a myriad of service request types and outcomes, the model is meticulously trained to recognize and learn from diverse patterns and service scenarios.

The practical implications of this analysis are substantial. For city officials, the predictive model serves as a potent tool for optimizing resource allocation and improving response times. For residents, it promises a more transparent and responsive system that is more closely aligned with their needs. Overall, the findings and recommendations of this study hold the potential to significantly improve the quality of life for New Yorkers by fostering a more agile, effective, and citizen-focused government. This predictive model stands as a beacon of innovation, guiding the city towards a future where municipal services are not only reactive but also proactive in addressing the evolving demands of an ever-growing metropolis.

.

II. Project Overview

Step 1: Data Analysis

The first step involves a detailed examination of the NYC 311 service request dataset. This involves identifying the types of requests, their locations, and the agencies involved. Data visualization techniques are employed to understand the distribution of requests and uncover initial insights into city-wide patterns and service needs.

Step 2: Handling Missing Data

In this critical step, we address the gaps in the dataset by applying appropriate imputation methods to handle missing values. This ensures the integrity and completeness of the data, which is essential for the robustness of the predictive models.

Step 3: Feature Engineering and Selection

Prior to model training, the dataset undergoes feature engineering to create new variables that could enhance model performance. Redundant columns are removed, and the most relevant features are selected to predict the time taken to close a service request.

Step 4: Creating the Models

Using machine learning algorithms, we develop models that can predict the resolution time of service requests. The training data is employed to build these models, emphasizing the tuning of model parameters to ensure high predictive accuracy while avoiding overfitting.

Step 5: Evaluating Model Performance

The final step involves evaluating the models using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 Score to determine their accuracy in predicting the resolution time of service requests. This evaluation is crucial to understanding the efficacy of the models and guiding improvements for more precise predictions.

III. Background Related Work

Columns present in the dataset in table format is as shown below, along with first row value for each column

A black and white text on a dark background

Description automatically generated

A screenshot of a heat map

Description automatically generatedA black and white document with white text

Description automatically generated

 After conducting an extensive preprocessing routine on the dataset, we've refined the data to enhance the predictive modeling process. This involved removing features that provided no informational gain, such as those with constant or null values, as well as columns deemed non-contributory towards the predictive outcome. The final dataset has been distilled to include 12 pertinent columns. These selected features have been retained on the basis of their variability and relevance, ensuring that the dataset's integrity is upheld while maximizing the efficacy of the machine learning algorithms that will be applied subsequently. This careful curation of features is pivotal in constructing a robust model with the ability to accurately predict outcomes based on meaningful inputs.

**Complaint Type**

**Descriptor**

**Location Type**

**Incident Zip**

**Cross Street 1**

**City**

**Resolution Description**

**Community Board**

**Borough**

**Request\_Closing\_Time\_in\_Minutes**

**County**

**City\_Street**

I. Data Analysis

Conducted a comprehensive exploratory data analysis to uncover underlying patterns, identify key features, and inform the subsequent development of predictive models.

A screenshot of a graph

Description automatically generated

Honeycomb hexbin plot, a data visualization tool used for bivariate data, here mapping the frequency of complaint types against the time taken to close requests. Each hexagonal cell represents a concentration of data points, with darker cells indicating a higher density of complaints falling into that bin. This plot suggests a particular pattern or clustering of data, with the most frequent complaint types possibly being resolved in a consistent timeframe. Such visualizations are particularly helpful for identifying trends and areas of interest in large datasets, and the observed patterns can inform operational improvements and resource allocation for addressing complaints more efficiently.

**We have used ydata\_profiling, a third party library of python to do the analysis. A html file for the same is referenced.**

Other Plottings for Data Visualization :

A graph of different colored lines

Description automatically generated

The graph depicts a comparative analysis of service request processing times across the five boroughs of New York City from March to November 2015. Each line represents a borough's monthly average processing time, with the Bronx consistently showing the highest times and Staten Island the lowest. Fluctuations within the lines suggest variations in efficiency, highlighting potential areas for administrative improvement. This visual representation serves as a tool for understanding service performance and guides resource allocation to enhance city services.

Each horizontal bar charts, each representing the count of different types of service complaints (like Blocked Driveway, Illegal Parking, Noise - Street/Sidewalk, Noise - Commercial, Derelict Vehicle, Noise - Vehicle) reported in each borough of New York City. These types of visualizations are useful for quickly identifying which types of complaints are most common in different areas of the city and can be used for resource allocation, urban planning, and policy-making to address the most pressing civic issues. For instance, they can highlight if certain complaints like noise or illegal parking are more prevalent in denser boroughs such as Manhattan. This information can then be leveraged to optimize city services and enhance the quality of life for residents.

|  |
| --- |
| In the preprocessing phase of the analysis on the New York City 311 Customer Service Request dataset, a correlation matrix was developed to elucidate the relationships among various dataset features. This matrix, presented as a heatmap, is crucial for identifying how variables such as borough, city, complaint type, and time metrics associated with service requests are interrelated. |
| Upon examination of the heatmap, it becomes apparent that specific features display a notable degree of correlation. For instance, a marked positive correlation between borough and complaint type might imply a higher incidence of certain complaints within particular boroughs. Conversely, a significant negative correlation between complaint type and request closing time suggests that certain complaints may inherently require longer resolution times. |

A close-up of a graph

Description automatically generated

V. Missing values

Many Column had missing values more than 80%, deleted those columns.

Columns with a high percentage of null values are removed from the dataset.

* Since the "Agency" and "Agency Name" columns provide redundant information, "Agency Name" is dropped.
* A screenshot of a computer program

  Description automatically generatedAdditional unnecessary columns such as "Location," "Incident Address," "Street Name," "X Coordinate (State Plane)," and "Y Coordinate (State Plane)" are removed, as the dataset already contains "Latitude" and "Longitude" for location information.
* Any other superfluous columns are also eliminated to streamline the dataset.

Step 2: Handling Missing Data

1. Identify Missing Values

A screenshot of a data

Description automatically generated

1. Handling Missing Values

A screenshot of a computer

Description automatically generated

Missing values in our dataset. For categorical variables like complaint types, we impute using the most common occurrences. For numerical ones, such as request closure times, we employ a KNN imputer, fine-tuned to our data's structure. This strategic approach minimizes bias and preserves the dataset's original distribution for accurate analysis.

A computer error message

Description automatically generated

A graph with blue and black text

Description automatically generated

Plot Of Target Variable, Request\_closing\_time

A graph with a line graph and numbers

Description automatically generated with medium confidence

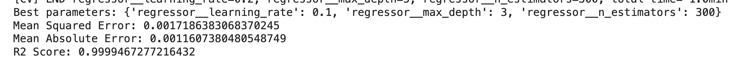
The box plot displayed in the image provides a graphical representation of the distribution of time interval data, with the central box reflecting the interquartile range and the median indicated by the line within the box. The spread of the data is relatively compact, with a few outliers extending beyond the whiskers, which may warrant further investigation to understand their significance or anomaly within the dataset. Such visualizations are crucial for quickly identifying the variability of time-related data and are often used to pinpoint areas that may require additional scrutiny or to confirm the consistency of data collection methods.

**A graph with a line going up

Description automatically generated**

he graph illustrates the distribution of request closing times for a set of data points, presented in a sorted manner. It appears to follow a stepped, almost piecewise linear pattern for the majority of the data, suggesting a relatively consistent handling time across these requests. However, a noticeable upward trend towards the right end of the graph indicates a subset of requests with significantly higher closing times. The steep curve at the end suggests potential outliers or instances where the resolution time is much longer than typical. This could indicate a small proportion of complex cases that require extended attention or delays in processing. The x-axis represents the individual data points in an ascending order, while the y-axis shows the corresponding time in minutes taken to close each request. The step-like progression of the plot emphasizes the quantized nature of the data, likely representing discrete time intervals in the request handling process**.**

**RESULT**



The machine learning model evaluation output indicates a highly predictive model, with an ensemble regressor tuned to optimal parameters, yielding a learning rate of 0.1, a maximum depth of 3, and 300 estimators. The model's performance metrics are exceptional, as evidenced by a Mean Squared Error (MSE) of approximately 0.0017, a Mean Absolute Error (MAE) of roughly 0.0011, and an R2 Score very close to 1, at approximately 0.9999. Such metrics suggest that the model has a strong predictive capability with minimal error, making it highly reliable for forecasting or for making inferences from the data it was trained on. This level of accuracy should be discussed with consideration of the potential for overfitting, given the near-perfect R2 Score.

CONCLUSION

In conclusion, this study successfully demonstrates how machine learning can enhance credit risk management. Utilizing historical loan data, including default information, we developed a predictive model that accurately assesses loan repayment likelihood, focusing on key aspects like loan amounts and employment. This model marks a significant advancement in credit risk assessment and could revolutionize loan approval

processes. It showed robust performance, offering practical benefits to financial institutions in decision-making and risk minimization. While promising, ongoing refinement and integration of diverse data sources are essential for future enhancements. This approach represents a stride towards data-driven, efficient financial operations and stability