Unveiling Patterns in NYC 311 Calls Data: Towards Smarter City Management

Asiya Shakeel

Data Science Capstone project Aug 2023

Springboard

NYC
311

1. Introduction:

The business problem addressed in this data science project is to improve urban service management and delivery in New York City by analyzing and gaining insights from the 311 calls data. The 311 calls data contains records of various service requests and complaints made by residents, businesses, and visitors across the city. By understanding the patterns and trends within this data, the project aims to provide valuable insights to key stakeholders.

The project's detailed implementation, analysis, and deliverables can be found in the GitHub repository https://github.com/asiyashakeel78/Spring-Board/tree/main/Capstone%20Project%

2. Approach:

2.1 Data Acquisition and Wrangling:

The datasets used in this project are available from the official website of the City of New York's Open Data platform. You can access the datasets at the following link:

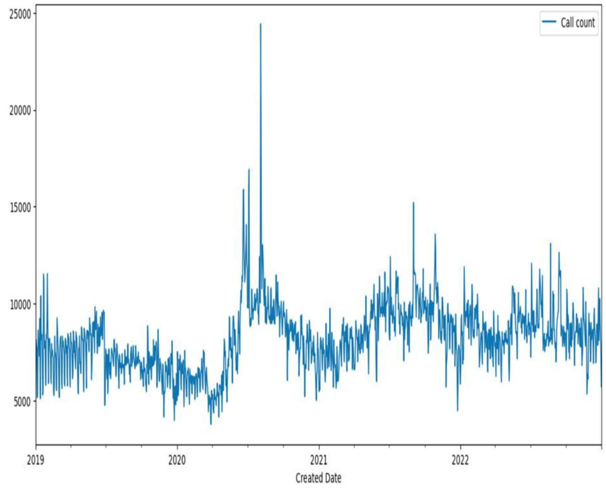
311 Service Requests from 2010 to Present | NYC Open Data

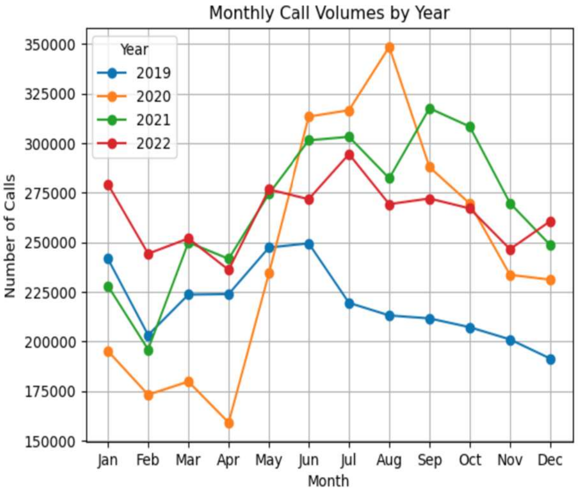
The website offers a comprehensive collection of datasets related to various aspects of the city, including 311 call data, which was used in this capstone project to analyze urban services.

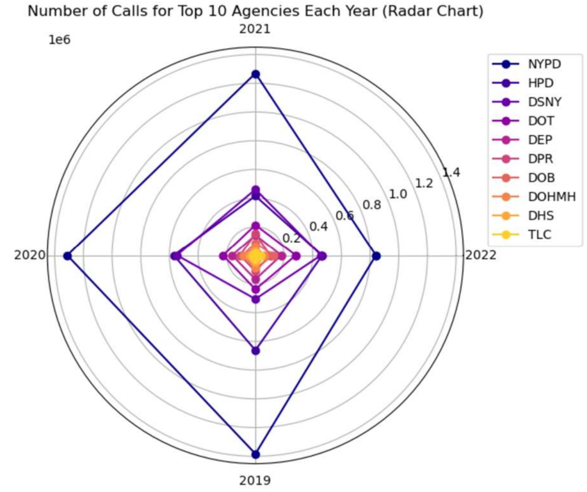
During the data wrangling phase, we handled missing values, outliers, and transformed variables as necessary to prepare the dataset for analysis. Given that this data involves a time series, it was a bit different to handle, especially when setting and treating the date column as the index. Through these data acquisition and wrangling steps, we ensured that the dataset was in a suitable format for further exploration, modeling, and time series analysis.

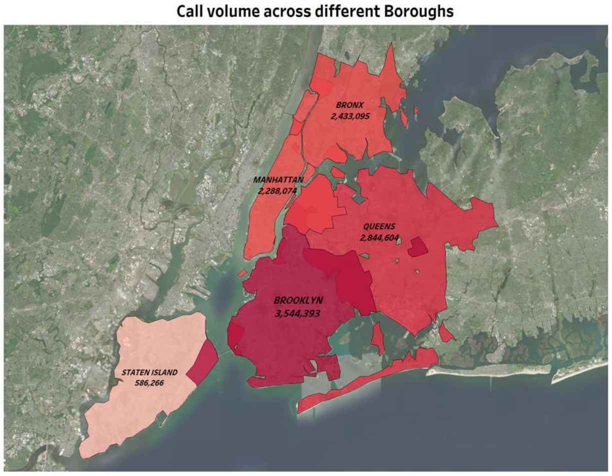
2.2 Story Telling and Inferential Statistics:

The first 2 plots reveal the daily call trend. 311 receives more than 8k calls every day. There are specific times when people make more phone calls. During the summer months of May, June, July, and August, there are more calls. However, in April, calls are fewer regardless of the year.



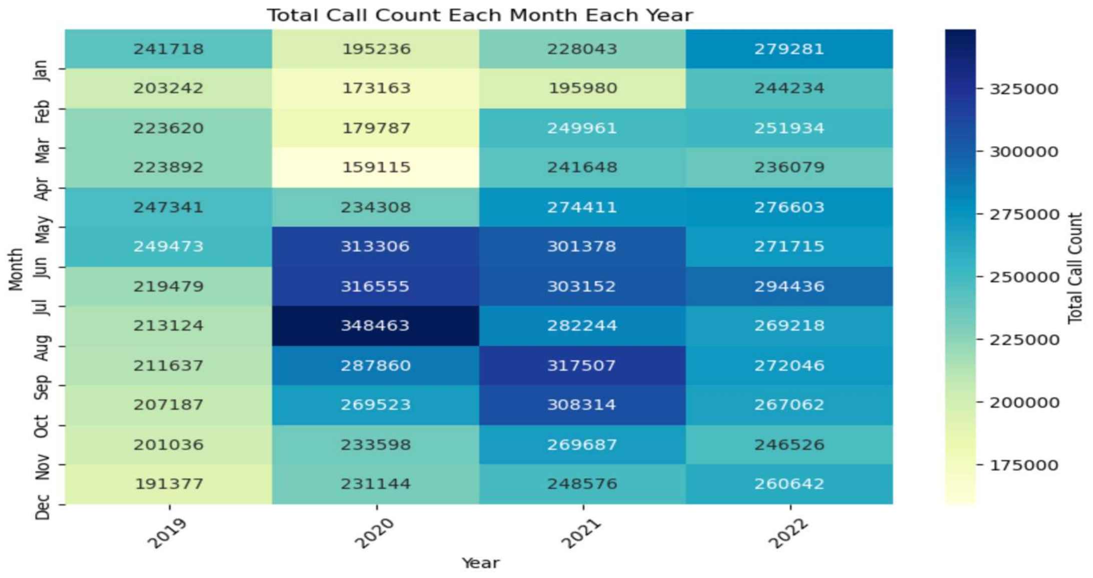



Number of Calls for Top 10 Agencies Each Year (Radar Chart)


Call volume across different Boroughs

The third plot shows that the agencies that receive the most calls are NYPD, HPD, and DSNY. Others also receive many calls, providing a complete picture.

The fourth plot leverages Tableau visualization to provide a geographical overview of call intensity. This visualization underscores that Brooklyn has the highest call volume, closely followed by Queens. Additionally, pockets of Staten Island experience elevated call volumes, while the overall trend for the borough is relatively low.



The fifth plot utilizes a heat table, making it evident that June, July, and August consistently exhibit heightened call volumes. This pattern aligns well with the anticipated increase in activities and events during the summer months.

2.3 Baseline Modeling:

We began by implementing a Linear Regression model, preceded by the application of the Box-Cox Transformer. This transformation was crucial in stabilizing the variance and bringing the dataset closer to a normal distribution. Following this, we explored the Auto ARIMA, SARIMA, and Prophet models, meticulously evaluating their performance.

Among the array of models assessed, the SARIMA model emerged as the clear frontrunner, exhibiting outstanding predictive capabilities. It achieved the lowest Root Mean Squared Error (RMSE) of 20161.45 and the most minimal Mean Absolute Percentage Error (MAPE) of 0.061. This outcome underscores the model's proficiency in capturing the inherent patterns within the data. The Auto ARIMA model closely followed, demonstrating marginally higher RMSE and MAPE values of 27,654.94 and 0.0806, respectively.

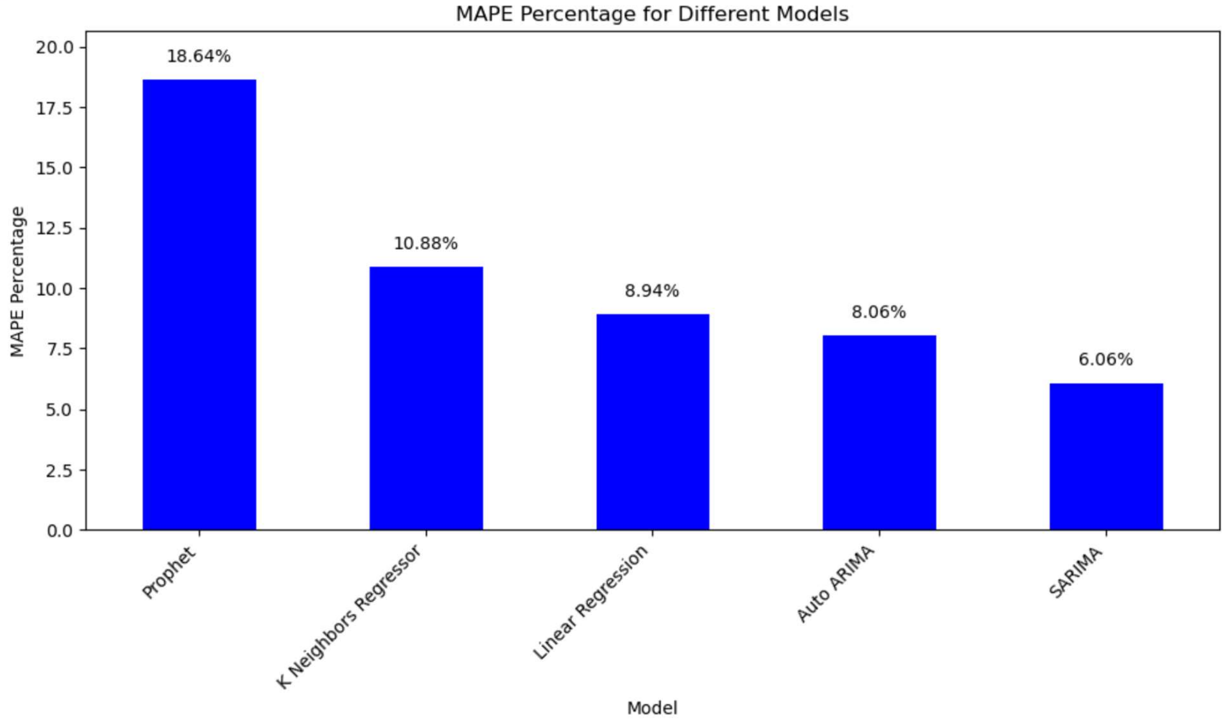
Conversely, the Linear Regression model displayed less precise forecasts, evidenced by an RMSE of 29,042.71 and a MAPE of 0.0894. The Prophet model, on the other hand, exhibited the highest level of prediction errors, reflected by its elevated RMSE of 56,315.61 and a MAPE of 0.1864.

2.4 Extended Modeling:

In the advanced modeling phase of our notebook, we employed the Seasonal Autoregressive Integrated Moving Average (SARIMA) model to project the volume of incoming calls over the upcoming 7 months (January 2023 to July 2023). Subsequently, we leveraged a sophisticated tool known as "MLjar-supervised," capable of conducting simultaneous experimentation with various models including Random Forest, Gradient Boosting, and more.

The standout outcome from this tool was (AutoML best model): "Default_NearestNeighbors." While not flawless, it emerged as the most promising candidate among all the models evaluated, yielding a prediction with a Root Mean Squared Error (RMSE) of 40212.89 and Mean Absolute Percentage Error (MAPE) of 10.88%.

3. Findings:



Interestingly, our conventional SARIMA approach continued to demonstrate its efficacy. It generated a prediction with an RMSE of approximately 20161.45 calls and showcased the lowest Mean Absolute Percentage Error (MAPE) of 6%.

The collective projections from the models assessed collectively indicate a declining trend in call volume for 2023 compared to the figures recorded in 2022.

4. Conclusions and Future Work:

As this project provides a robust foundation, there are several avenues for future exploration:

Enhanced Feature Engineering: Delve deeper into feature engineering to extract more informative attributes that might contribute to more accurate predictions.

Model Ensembling: Explore the potential of combining the strengths of different models through ensembling techniques to achieve even better predictive performance.

Incorporating External Factors: Incorporate external variables, such as public holidays, weather conditions, and special events, that could influence call volume for more accurate predictions.

Real-time Monitoring: Implement a real-time monitoring system to adapt the model to changing patterns and dynamics as new data becomes available.

Feedback Loop: Establish a feedback loop with stakeholders to continuously improve the model's performance based on real-world outcomes.

User-friendly Visualization: Develop user-friendly dashboards or visualizations to allow stakeholders to easily interpret and interact with predictions and insights.

Cross-Validation Techniques: Implement various cross-validation techniques to validate model performance and robustness.

By pursuing these avenues, we can continue to refine our predictive models and generate actionable insights that contribute to better resource allocation and decision-making.

5. Recommendations for the clients:

Based on the analysis and insights gathered from this project, here are some recommendations that can guide the clients in their decision-making process:

Resource Allocation Planning: Utilize the predictive capabilities of the SARIMA model for accurate resource allocation. The model's consistent performance in capturing call volume trends can assist in staffing and equipment planning to meet anticipated demands.

Seasonal Staffing: Given the observed trend of higher call volumes during the summer months, consider implementing seasonal staffing adjustments to efficiently manage increased call loads during peak times.

Performance Evaluation: Continuously monitor the performance of the "Default_NearestNeighbors" model from AutoML. While it exhibits promising predictive capabilities, regular assessment will help identify any changes in trends or patterns that might require adjustments.

Budget Allocation: With the decline in projected call volume for the upcoming year, 2023, clients can strategically allocate budgets by considering reduced call-related expenses while ensuring operational efficiency.

Model Integration: Integrate the SARIMA model's forecasts into the operational framework to enhance planning and decision-making. Its reliable predictions can be an asset in optimizing resource utilization.

Exploratory Analysis: Explore the factors that contribute to the decline in call volume in 2023 compared to 2022. Investigate potential reasons, such as changes in user behavior, technological advancements, or shifting external circumstances, to better understand the underlying dynamics.

Feedback Loop: Establish a mechanism for receiving feedback from front-line staff who interact with callers. Their insights can help improve the model's accuracy by incorporating qualitative aspects that data might not capture.

Data Collection: Continuously refine data collection processes to ensure data quality and relevance. High-quality data is essential for accurate model predictions.

Capacity Planning: Leverage the predictive models to guide long-term capacity planning. Accurate forecasts will aid in determining future infrastructure requirements, ensuring seamless service provision.

Implementing these recommendations will empower clients to make informed decisions, optimize their operations, and better respond to changing call volume dynamics, ultimately leading to improved customer satisfaction and operational efficiency.

6. Consulted Resources:

https://otexts.com/fpp3/

https://www.kaggle.com/learn/time-series

311 Service Requests from 2010 to Present | NYC Open Data

https://supervised.mljar.com/

https://youtu.be/3ZCTC32EqKk