Boston 311 Reports Analysis

Suneh Bhatia, Nathan Lee, Ayesha Nayeem, Alexandra Noor, Jane Pan

Introduction

Boston's Open Data portal offers many city government datasets freely accessible to the public. One standout dataset, which tracks 311 Service Requests from 2011 onwards, captures over 20 million records detailing the city's non-emergency issues reported by residents. This dataset, continually updated, serves as a crucial tool for Bostonians to request municipal services, report problems, and obtain information via a dedicated 311 phone line and online platforms. Our article explores patterns and trends within this expansive dataset, applying advanced data analytics and machine learning techniques to predict which city departments will respond to specific types of requests. Through this analysis, we aim to enhance the effectiveness of resource allocation and improve civic engagement across Boston's diverse neighborhoods. We will do this through a rigorous exploration of the data and analyze trends between certain columns to draw meaningful conclusions. We also aim to create a neural network model that will accurately classify the department that will respond to a certain type of request.

Data Cleaning/Processing

We dropped several columns when cleaning our data that we found to be unimportant in our analysis. The first column we dropped was case enquiry id. There was no beneficial reason to know the id for each of the cases, as no one would know what it means. We also dropped the submitted and closed photo rows – there were around 2 to 3 million null rows meaning photos were not submitted in the 311 report. We also chose to drop the case title column because it did not provide any additional information beyond what was already included in the subject, reason, and type columns, which are more relevant to our analysis. Similarly, the closure reason column was removed since it was not necessary to address our eight base questions. The location street name and zip code columns were dropped because we already had the neighborhood column, which provides sufficient geographical context for our analysis. Several district-related columns, including precinct, pwd_district, fire_district, city_council_district, and police_district, were also removed as they were not critical for answering our research questions. We decided to exclude the latitude, longitude, and geom_4326 columns because they provided specific coordinates, which were unnecessary given that we were focusing on neighborhood-level insights rather than exact locations. Similarly, the neighbourhood services district column was deemed redundant since we already had the neighbourhood column, and the numerical district values did not add meaningful value to our analysis. Finally, we addressed the SLA_target_dt column, which contained unrealistic dates extending to the year 2062 for cases reported as early as 2011. Rows with these out-of-range dates were removed, leaving only cases with dates between 2011 and 2024. This resulted in the removal of 15,334 rows, ensuring our dataset contained only valid and realistic data points.

Initial Exploration

Now that we had our dataset cleaned we felt it important to perform an initial analysis of the requests over the last 13 years. One important trend we wanted to visualize was how the number of requests changes seasonally, year by year. This visualization is shown below:

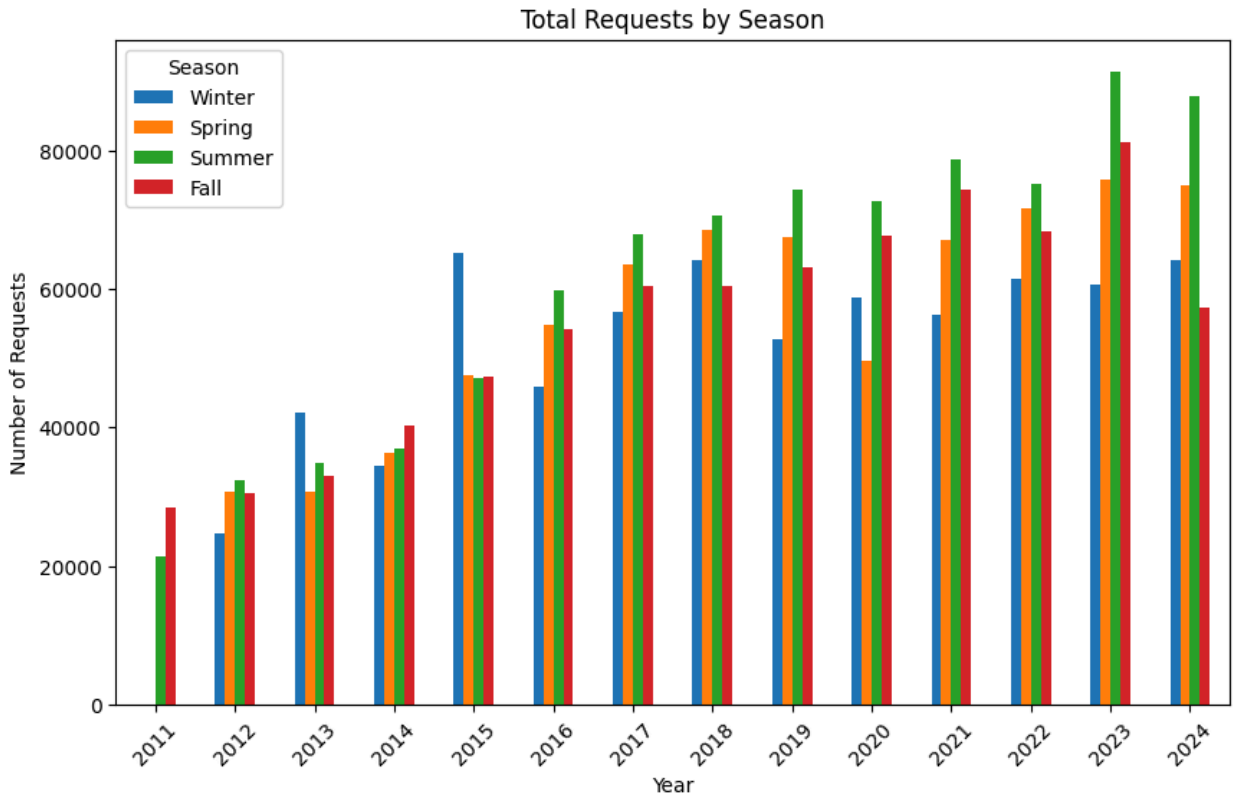

Figure 1. Seasonal changes in number of requests per year

This can give us some insightful information into the biases that may arise when using public data sets. We split the seasons by every three months, where December-February is Winter, March-May is Spring, June-August is Summer and September-November for Fall. There were no values before July of 2011 included in the dataset, so we excluded Winter and Spring for that year.

There is a general increasing trend in the number of requests, especially comparing 2024 to 2011. This is most likely as a result of the release of new technologies and applications that make submitting requests easier.

We can observe a relatively sharp spike in requests during the winter of 2015 (counted from December 2014 into January and February). This is explained by the fact that Boston experienced one of the most brutal winters it had ever seen in the winter of 2014-2015 with

108.6 inches of snowfall. In fact we confirmed this by looking at the top requests in 2015 and discovered that the 'Request for Snow Plowing' was made 30152 times and was the most frequent one.

Further we noticed drops in spring of 2020, likely due to COVID-19. The pandemic was an important consideration we knew we had to make, but we can see that the relative disparities in number of requests are not too significant during the years of COVID-19 (2020-2022).

Another important trend that we wanted to look at was the most frequent 311 complaints by department, as shown in Figure 2.

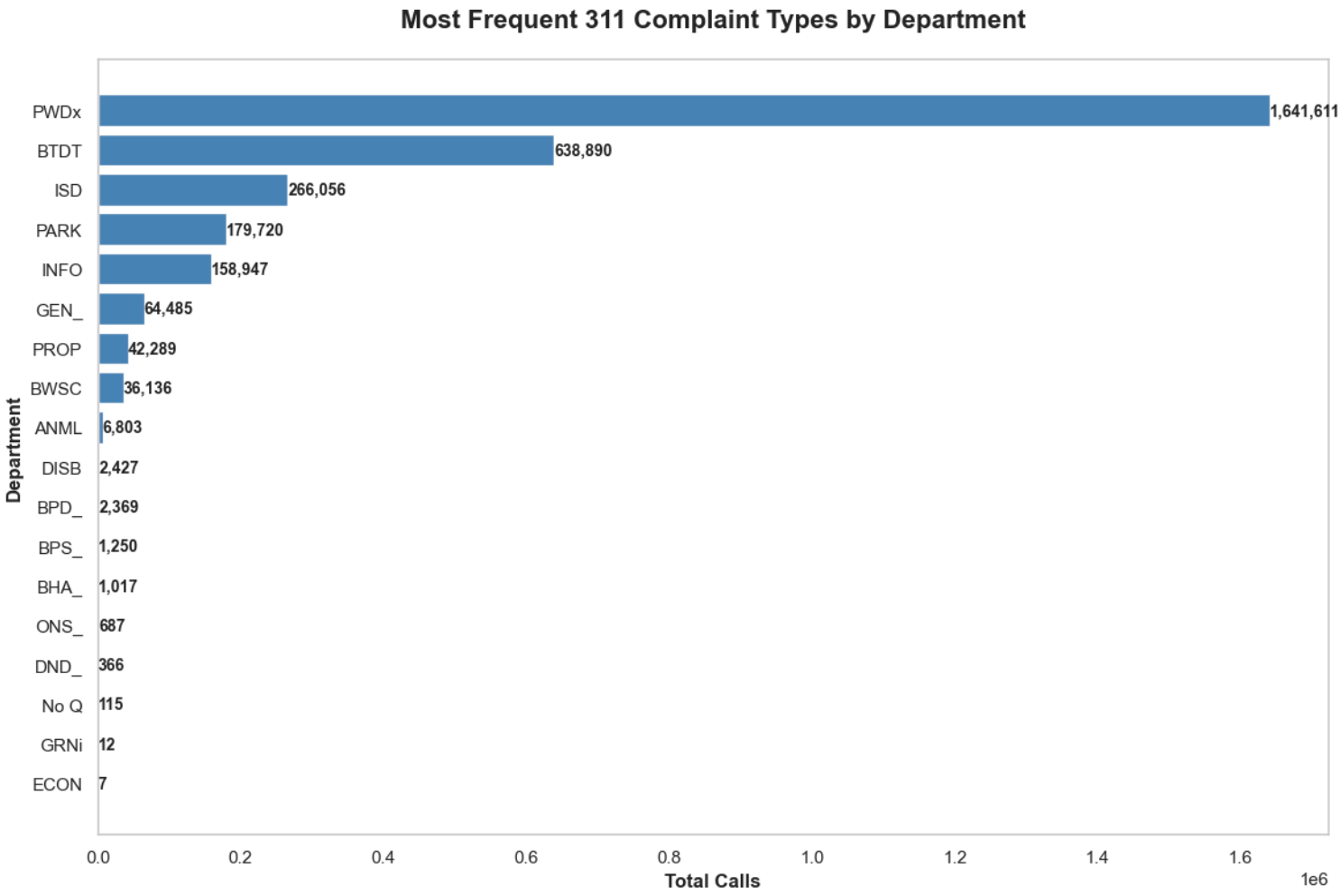

Figure 2. Most Frequent 311 Complaint Types by Department

This gives us a good insight into which subjects have the most complaints coming in from 2011 to 2024. The PWDx department has the most complaints by over one million compared to the BTDT, the department with the second highest complaints. We could attempt to predict the department that a complaint will be assigned to using a machine learning algorithm. However,

the skewed distribution of data creates an unbalanced dataset, which we need to account for in our analysis and training.

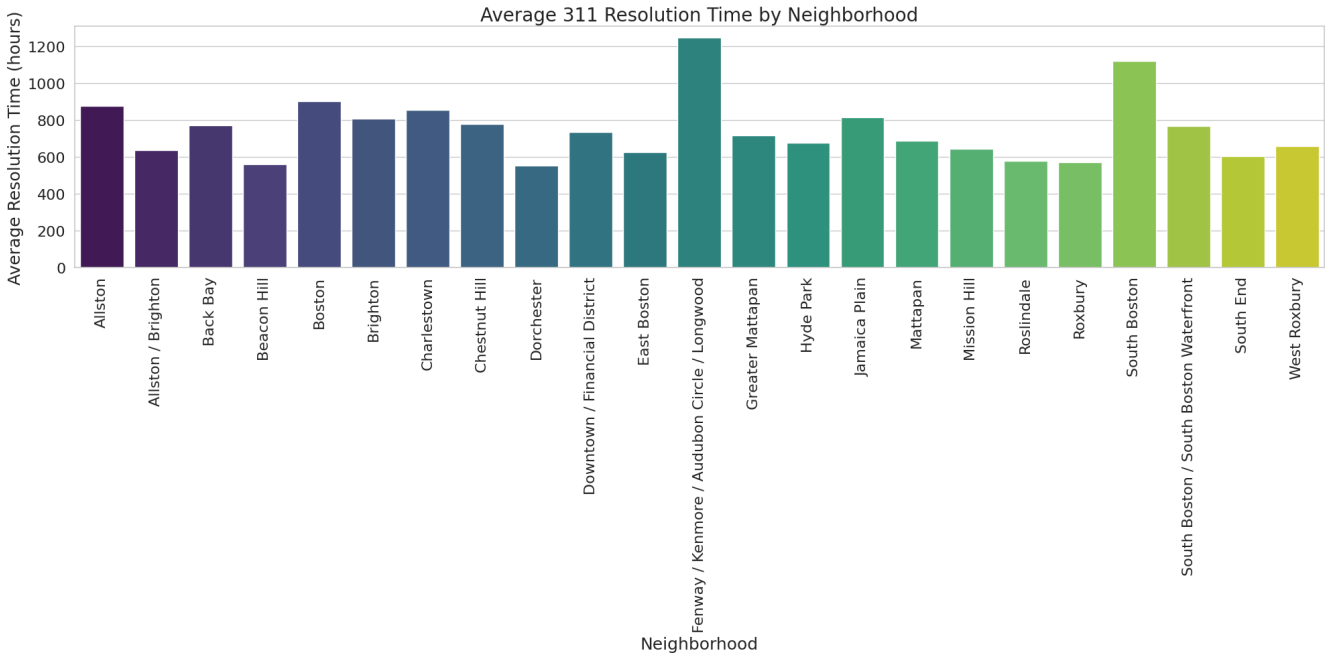

Figure 3. Average 311 resolution time across Boston neighborhood

We also looked at the average resolution time by neighborhood as illustrated by Figure 3. Notably, Fenway/Kenmore/Audubon Circle/Longwood and South Boston exhibit significantly higher resolution times compared to others. This potentially reflects higher service demands or operational inefficiencies in these areas. Conversely, neighborhoods like West Roxbury and Jamaica Plain show shorter resolution times, which might suggest more effective service delivery or lower service request volumes. Understanding these disparities can guide resource allocation and operational improvements.

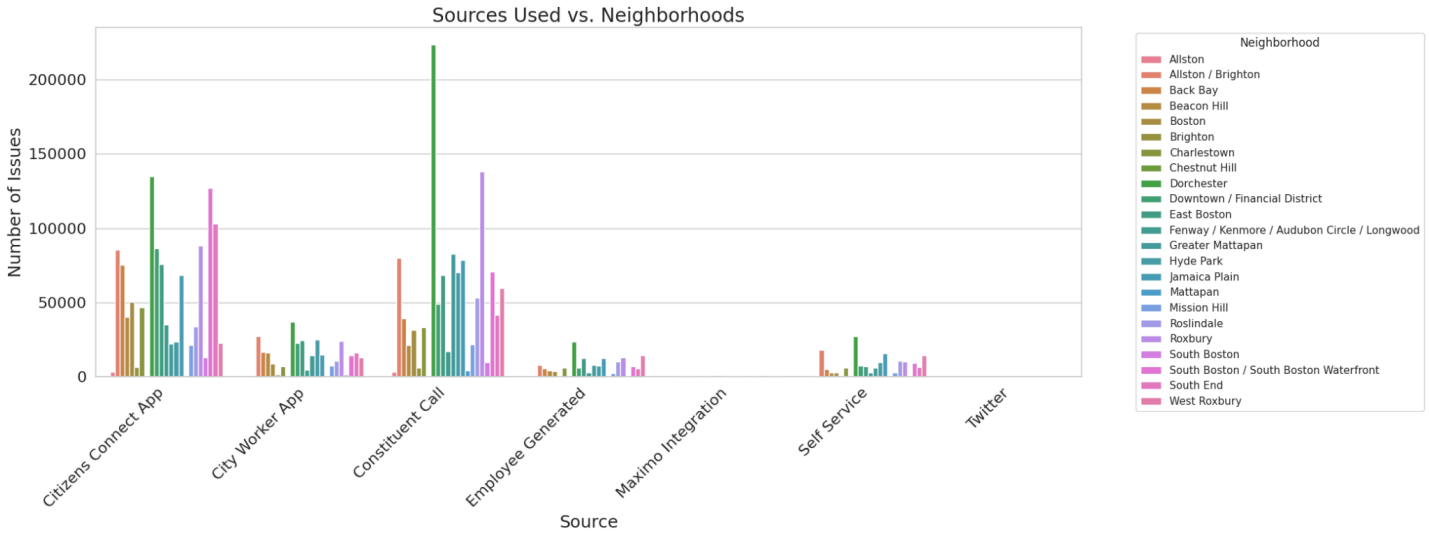

Figure 4. Distribution of sources used to report issues across Boston neighborhoods

Similarly, the sources used to issue complaints also varied across neighborhoods. The bar chart in Figure 4 compares the number of 311 issues reported across Boston neighborhoods by source type, such as the Citizens Connect App, Constituent Calls, and City Worker App. Constituent Calls and the Citizens Connect App dominate as the primary sources, reflecting their accessibility and popularity for issue reporting. Fenway/Kenmore/Audubon Circle/Longwood stands out for a high number of constituent calls, potentially indicating higher awareness or engagement. Neighborhood-specific trends may reflect differences in technology access, community engagement, or service needs. Identifying these patterns can inform outreach strategies and improve service equity.

Assumptions

We also considered a range of biases and assumptions to ensure a comprehensive and accurate analysis. For one, we acknowledged potential technical challenges and misroutings in the 311 call system, particularly with mobile app submissions, which could influence data accuracy. This understanding stemmed from insights into similar issues faced by other cities, where mobile apps for service requests like those in San Jose were present with problems that affected data reliability, as reported by San Jose Spotlight.

Public data has immense potential to empower communities and improve services by providing detailed information about urban challenges. This data can reveal disparities in service distribution or engagement across different neighborhoods, helping to address inequities and optimize city resources effectively. However, it's crucial to recognize that while data can guide better decision-making and foster community involvement, it may also not fully capture every community's unique needs or the full extent of certain urban issues, as discussed in Scholars Strategy Network. Similarly, we noted that certain complaint types or departments might be

over-represented due to reporting frequencies and departmental responsiveness, requiring careful consideration to avoid skewing the model's predictive accuracy based on dominant classes.

Another important consideration are that requests that get submitted towards the end of one year, but only closed in the following year. The minimal quantity of these cases meant we decided to include them as part of the current year (i.e. if a case opened in December 2014 only closed in January 2015, we included it as part of the 2014 data).

Visualizations + Explanations

What is the total volume of requests per year, or how many 311 requests is the city receiving per year?

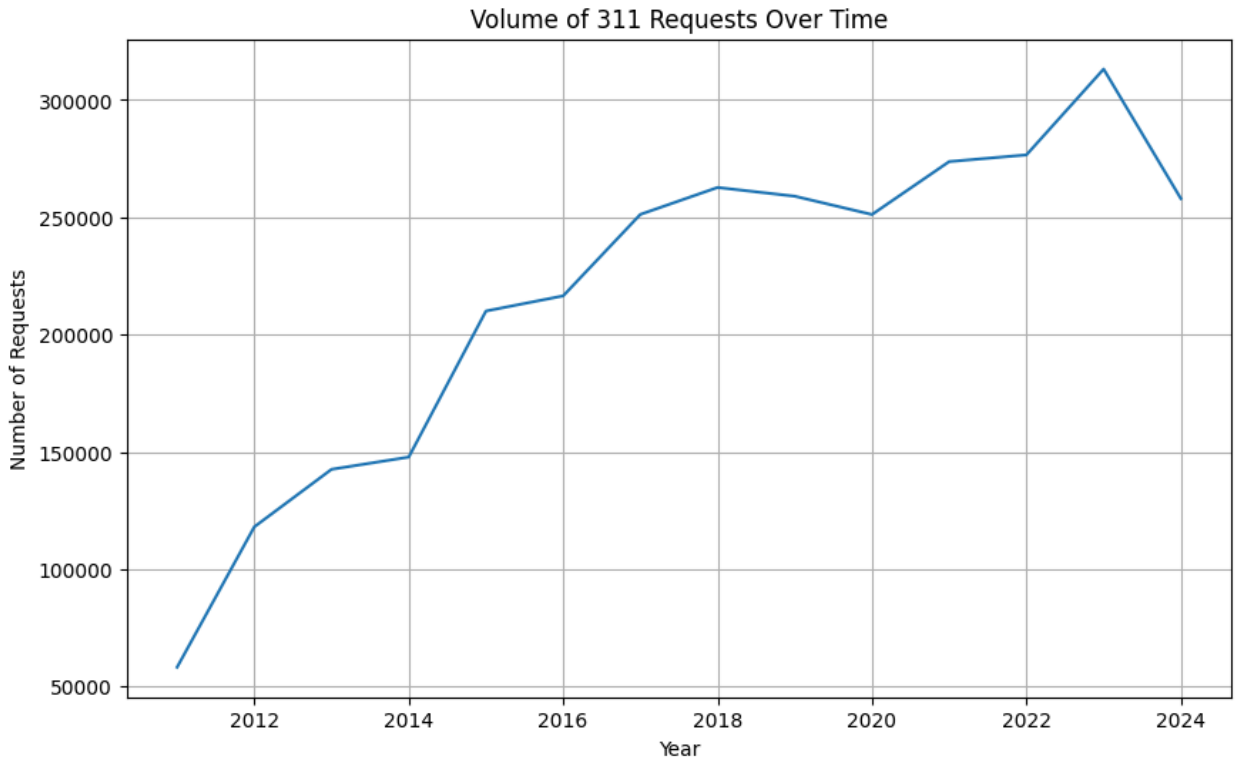

Figure 5. Total volume of request from 2012 to 2024

This visualization highlights how the usage of Boston's 311 service has evolved over the years, potentially reflecting increased community engagement, improved awareness of the service, and shifts in urban population dynamics. The steady rise in requests through 2016 may indicate growing reliance on city services or enhanced reporting mechanisms. The subsequent plateau and peak around 2023 suggest a saturation point or heightened activity, possibly due to policy changes, major events, or socioeconomic developments.

Which service requests are most common for the city overall AND by NEIGHBORHOOD and how is this changing year over year by SUBJECT (department), REASON,QUEUE?

Note: Refer to Appendix A

Due to the high volume of queues, they were separated into individual graphs, and have been inserted into a separate document (See Appendix A). Figures A.1 and A.2 show a slow increase in the amount of service requests in BTDT and PWDx, where District 02 seems to have a large jump in requests. This could be due to some underlying issue going on in District 02. However, ISD is stagnant from 2015-2024, while increasing from 2011-2015 (Figure A.3); this means that there is less of a need for inspection services throughout Boston. Furthermore, the Information Channel (INFO) seems to follow a similar trend according to Figure A.4; this may be because more people were being exposed to this channel in early years, but now further exposure seems to have halted. Additionally, Parks has seen a general increase in the amount of service requests coming in, with General comments being the largest section (Figure A.5). From a general perspective, there has been a general increase in total service requests by Neighborhood over the years from 2011-2024 according to Figure A.6; however, it seems to have halted in growth in 2018 with a small spike in requests in 2023. Going into 2024, there seems to be a drop in requests compared to 2023, which could be due to the fact that 2024 has not been finished yet, and many requests are still open. This can be reflected in Figure A.7, where the requests are sectioned by department (subject). The Public Works Department seems to have the most requests, and Transportation following shortly after. Lastly, Figures A.8 and A.9 depict the top 5 and lowest 5 reasons for service requests, which allows us to have a greater understanding of issues that are starting to arise in Boston. For example, starting 2016, Code Enforcement requests have creeped into the top 5, and are persistent to 2024. Street Cleaning has been consistent since 2013, and Enforcement & Abandoned Vehicles have been gradually increasing since 2011. Highway Maintenance and Sanitation has been consistent from 2011-2024, meaning that despite the efforts that Boston puts into handling these requests, there is always work to be done every year in those areas. On the other hand, Valet has been one of the lowest reasons for service requests in every year from 2011 to 2024, with a large volume of Current Events in 2011, 2012, 2017, and 2018.

How is the case volume changing by submission channel SOURCE?

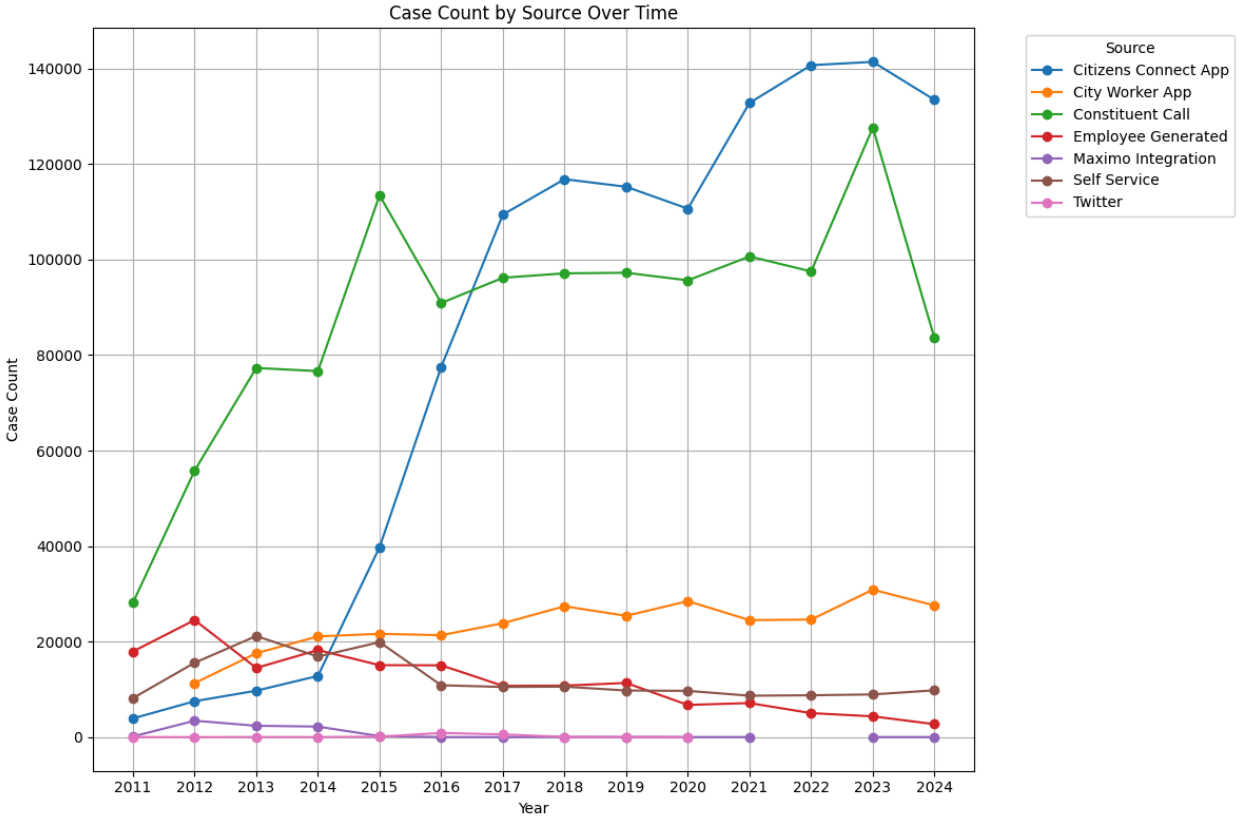

Figure 6. Total volume of request from 2012 to 2024

This trend shows a shift from traditional calls to app-based reporting over time. Constituent Calls were initially dominant, but the Citizens Connect App grew rapidly from 2014, peaking by 2023. Meanwhile, City Worker App and Employee Generated reports saw steady but limited use, reflecting specific applications. The slight decline in 2024 may suggest a plateau or the emergence of new platforms. Overall, the data indicates a shift toward digital reporting for civic issues.

What is the average # of daily contacts by year?

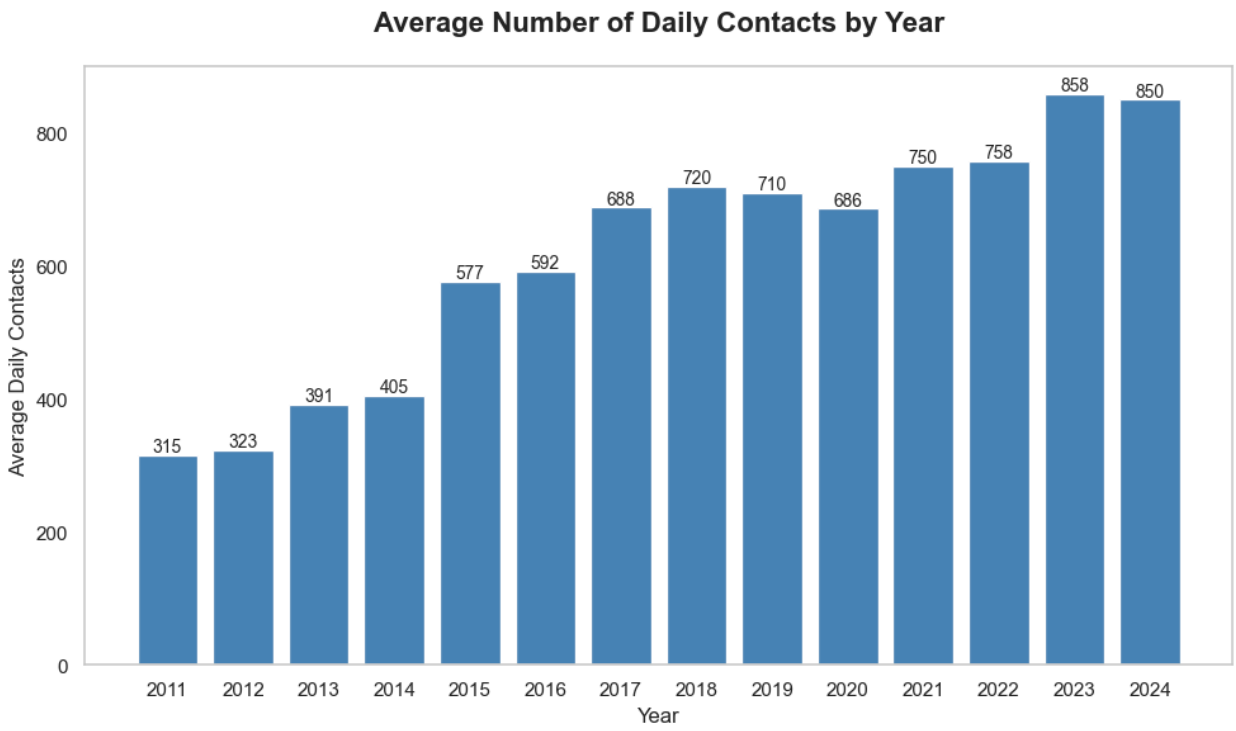

Figure 7. Average Number of Daily Contacts by Year

This visualization depicts the average number of daily contacts to the Boston 311 service over the years. We can see that there is an increasing number of engagements daily, indicating that there is a rise in awareness of Boston's 311 Service. The small dip from 2019 to 2020 suggests that there may have been major events or socioeconomic developments that might have resulted in less service requests over those years.

Volume of top 5 request types (TYPE)

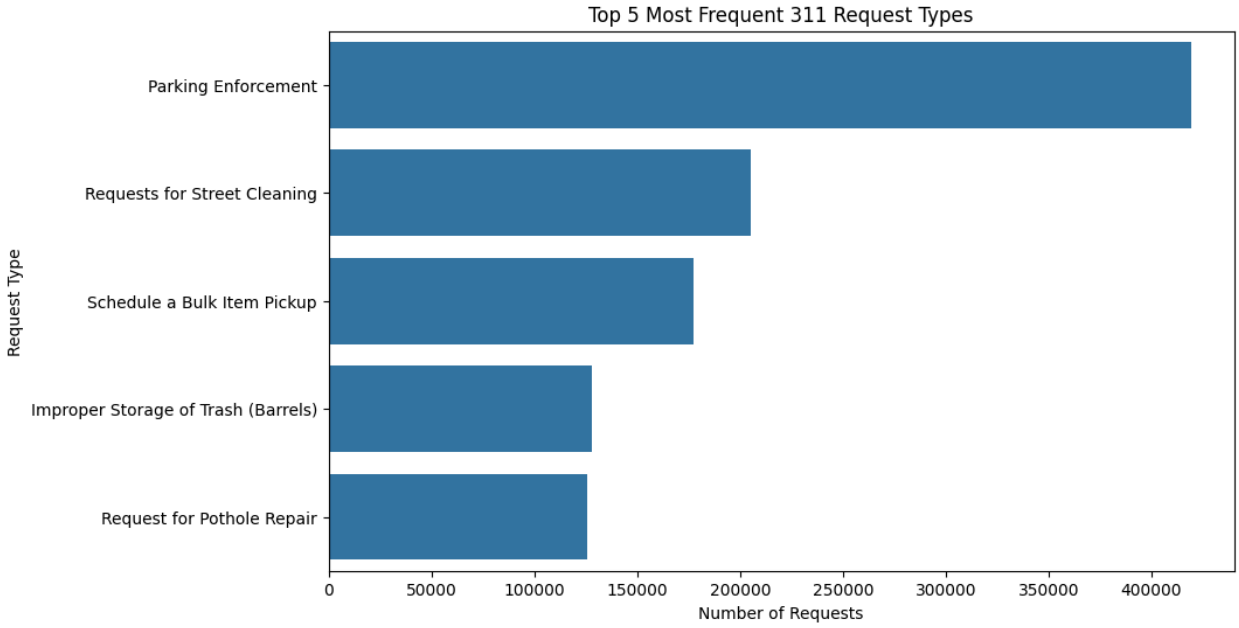

Figure 8. Top 5 most frequent 311 request Types

This bar chart shows the top five most frequent 311 service requests in Boston. Parking Enforcement leads with over 400,000 requests, reflecting widespread parking issues in a densely populated city. Street Cleaning follows at 250,000, indicating high public concern for urban cleanliness. Bulk Item Pickup (200,000) highlights the need for efficient waste disposal services. Improper Trash Storage and Pothole Repair (both ~150,000) point to challenges with waste compliance and infrastructure maintenance.

These top requests align with urban priorities: managing parking in crowded areas, maintaining clean streets, handling bulky waste efficiently, ensuring proper trash practices, and addressing road safety.

Average goal resolution time by QUEUE

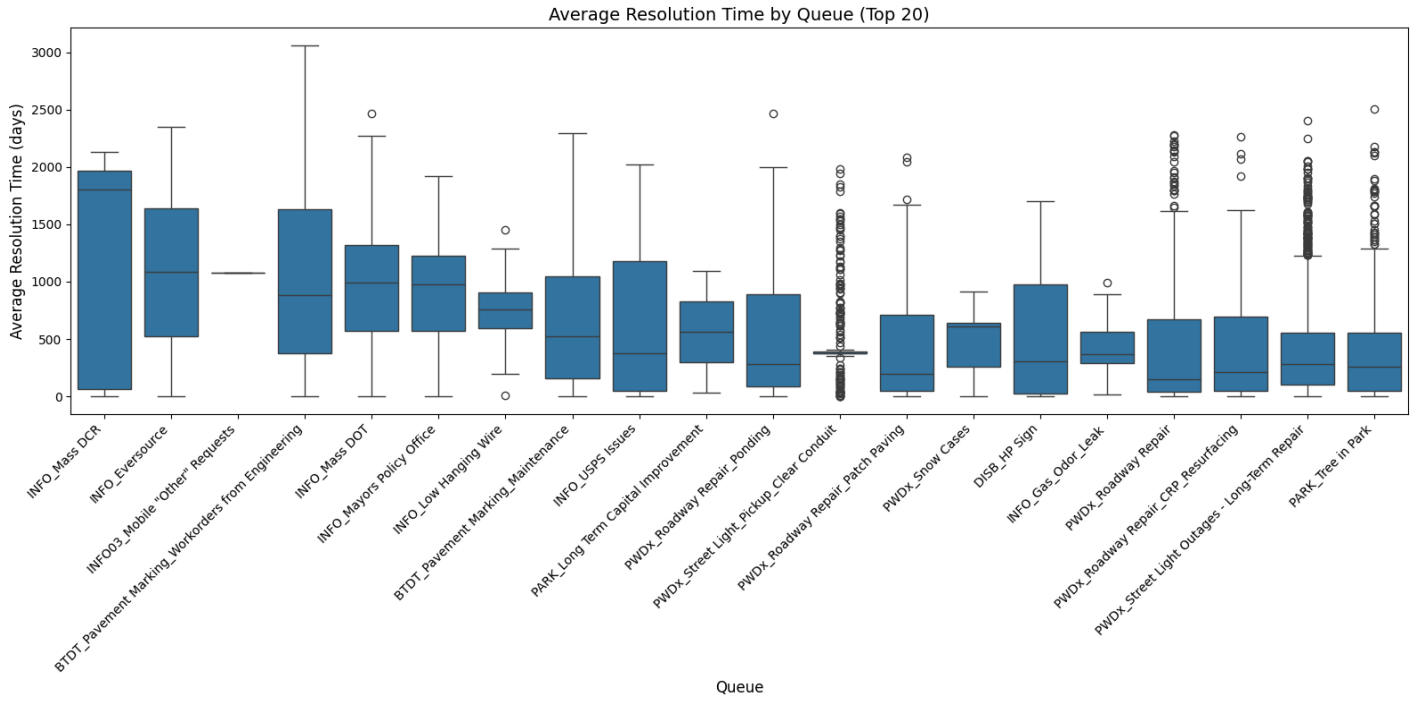

Figure 9. Average Resolution Time by Queue

This boxplot shows the distribution of resolution times for the top 20 service request queues. The plot reveals a lot of variability, with some services exhibiting wide ranges in resolution times and numerous outliers, indicating inconsistency in handling requests. Outliers in the boxplot indicate service request resolution times that deviate significantly from the typical range observed for each queue. They could indicate issues like complex service requests that take longer to resolve, or inconsistencies in service handling,

Average goal resolution time by QUEUE and NEIGHBORHOOD

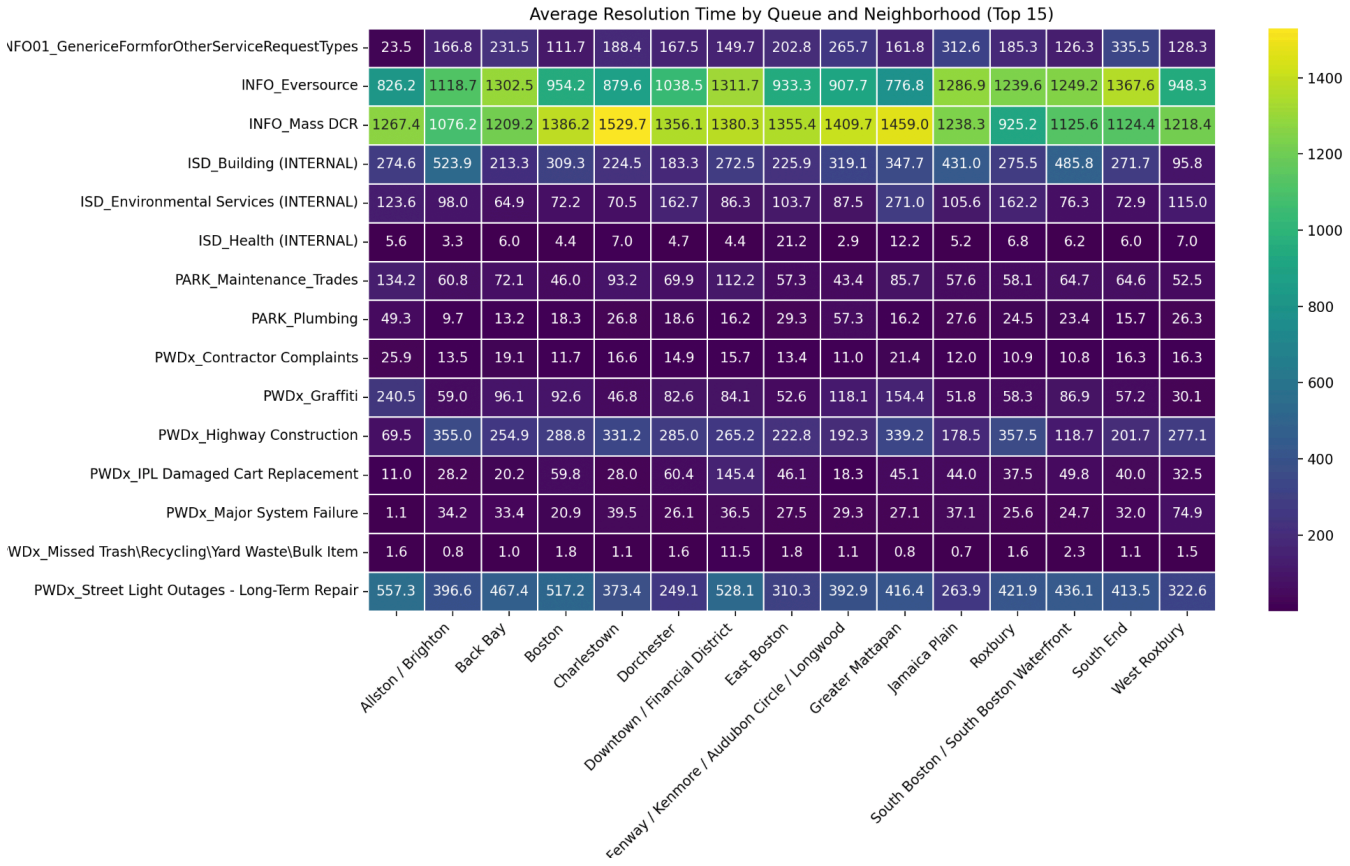

Figure 10. Average Resolution Time by Queue and Neighborhood

This heatmap visualizes the average resolution times for different service requests (queues) across neighborhoods in Boston. The empty label in the *first* neighborhood column signifies entries in the dataset where neighborhood data is missing or not recorded (e.g Unknown). We also limited the heatmap to the top 15 queues and neighborhoods, so that the visualization is less cluttered and more readable. In addition, there is a significant variability in resolution times across different queues and neighborhoods. For example, queues like "INFO_Mass DCR" and "INFO_Eversource" show higher resolution times, which suggests these issues might be more complex or depend on coordination with external agencies. On the contrary, queues such as “WDx_Missed Trash\Recycling\Yard Waste\Bulk Item” show low resolution times, often less than two days, suggesting the operational efficiency of waste management services and generally less complex nature of these tasks. Additionally, the variation in resolution times across different neighborhoods for the same service types might suggest disparities in resource allocation or differing local conditions that impact how quickly a service request can be addressed.

Percentage of service requests are closed (CLOSED_DT or CASE_STATUS) vs. unresolved (CASE_STATUS = open)

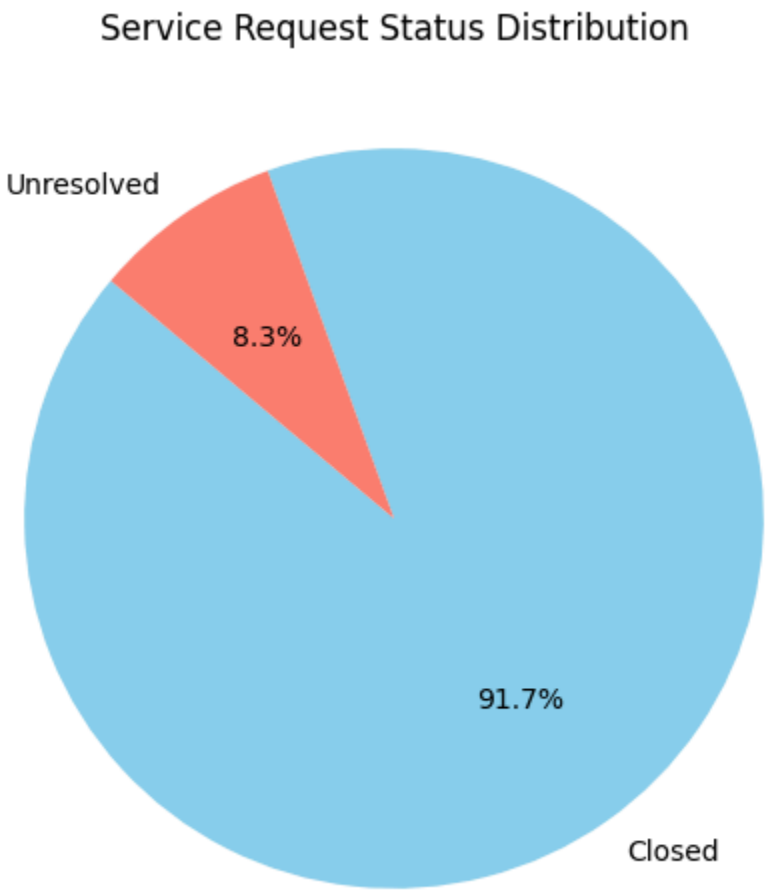

Figure 11. Distribution of Closed and Unresolved Service Requests

This plot shows the percentage of closed and unresolved cases in the data. We can see, for the most part, that the majority of cases per year are closed. However there are biases to take into account: cases that are opened towards the end of the year tend to be closed in the following year which can affect whether or not a case is deemed closed or null. This indicates a high closure rate but highlights a small proportion of unresolved cases that may need attention.

Model to predict responding departments

Our goal was to create a neural network model that could predict the responding department based on the type of request. However, there are some considerations we had to make. First, we can see that the Public Works Department (PWD) dominates the volume of calls, as illustrated in Figure 2. This is significant, as some departments respond to less than 1% of what the PWD responds to, leading to an unbalanced data set. This imbalance can bias the model toward predicting the majority class (PWD) and limit its ability to accurately identify and generalize to minority classes.

Dealing with Class Imbalances

To mitigate this challenge, we employed an oversampling strategy. By increasing the representation of underrepresented departments, the training data became more balanced. This enables our neural network to better learn patterns associated with smaller classes. The goal was to maintain a realistic data distribution while still providing the model with enough variety to learn meaningful distinctions. In addition, we considered evaluating the model on both a *balanced* test set (oversampled to reflect a more even distribution) and a *random* subset of the original data (which retains the natural proportions of each department).

Feature Engineering

Our input features for the network were from textual descriptions in the 'type' column. We used Keras' Tokenizer for tokenization, a process that converts the type descriptor into a vector of binary values, each value representing the presence/absence of words. This essentially created a bag-of-words for the model to use, providing a starting point for it to learn which terms were associated with particular departments.

An additional feature we used was the "resolution_time_hours," calculated from the difference between closed_dt and open_dt. This feature provides temporal context—certain departments may consistently close requests faster or slower, giving the model another important dimension to differentiate between classes.

After oversampling, we split the data into training, validation, and test sets. The training and validation sets came from the balanced, oversampled dataset. The validation set helps monitor the model's performance during training and tune hyperparameters (like the number of epochs) to avoid overfitting.

To measure real-world performance, we also drew a random subset from the original, unmodified dataset. This "random original subset" test set maintains the natural class imbalance and distribution. Evaluating this subset allows us to see how the model might perform under the true conditions it will face once deployed, rather than an artificially balanced scenario.

Furthermore, due to limited computing power we were unable to run our model on all three million rows. As a result, we took a random sample of 300,000 rows, ensuring that the sample was representative of the overall dataset in terms of class distribution and key features. This approach allowed us to train and evaluate the model efficiently while maintaining meaningful insights into the patterns and performance metrics of the full dataset.

Model Design

For our models design, we implemented a feedforward neural network using Keras with a Sequential model, which allowed us to build our model layer-by-layer. The input layer matched the size of our feature vector, which combined both tokenized textual information (extracted from the request "type") and a numeric feature representing resolution time. This setup ensures that the model receives all relevant data—both linguistic patterns and temporal context—at once.

This is the design of our model:

```
model = Sequential()
model.add(Dense(128, input_shape=(X_train.shape[1],), activation='relu'))
model.add(Dense(64, activation='relu'))
model.add(Dense(y_train.shape[1], activation='softmax'))

model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

To learn non-trivial relationships, we incorporated two hidden layers of neurons, each using the ReLU activation function. ReLU helps the network capture nonlinear patterns as it will enable certain neurons to activate strongly in response to particular input combinations, while remaining inactive for others (all or nothing). As a result, the model can discern subtle distinctions—such as certain word patterns that correlate with specific departments when paired with particular resolution times.

Our output layer used a softmax activation function, which transformed the final layer's raw scores into a probability distribution over all possible departments. This probability-based output provides a clear, interpretable prediction, with the highest probability indicating the most likely responding department.

For training, we chose "categorical_crossentropy" as our loss function, since it is a standard for tasks like these where we are predicting multiple classes (departments). It quantifies how far our predicted probability distribution is from the true distribution (where the correct class is represented by a 1, and all others are 0). We optimized this loss using Adam, an adaptive optimizer that adjusts the learning rate of each parameter automatically, often leading to faster convergence and more stable training compared to standard gradient descent.

We tracked accuracy as our performance metric to gauge how often the model's top prediction matched the true department. During the 50 training epochs, we monitored both training and validation metrics. The validation set—held out from training—served as a checkpoint to ensure that improvements weren't due to the model simply memorizing the training examples. If validation accuracy began to plateau or decline while training accuracy continued to rise, it would suggest overfitting. This monitoring process allowed us to confirm that our model wasn't just learning idiosyncrasies from the training data, but instead capturing meaningful patterns that generalize to previously unseen examples.

These are the parameters we used to train our model:

```
history = model.fit(
  X_train, y_train,
  epochs=50,
  validation_data=(X_val, y_val),
  verbose=1
)
```

Results

On the balanced test set, where all departments were equally represented due to oversampling, the model achieved an accuracy of 86.48%. This indicates that the model learned meaningful patterns for both majority and minority classes, even those departments that are less frequently represented in the original dataset. The relatively low loss value of 0.4047 further confirms that the model was confident and accurate in its predictions for this balanced scenario.

When evaluated on the random original subset, which maintains the real-world distribution of departments, the model achieved a slightly higher accuracy of 87.12%. This result highlights the model's robustness and its ability to generalize effectively to the imbalanced data distribution, where the Public Works Department (PWD) dominates the call volume. This allows us to safely assume no overfitting is taking place. The slightly lower loss of 0.3697 on this dataset reflects the model's confidence in making predictions for real-world scenarios. The similarity in performance between the balanced and random test sets suggests that the oversampling strategy and feature engineering—such as the inclusion of textual descriptors and resolution time—allowed the model to avoid overfitting to the balanced training data and maintain high generalization performance.

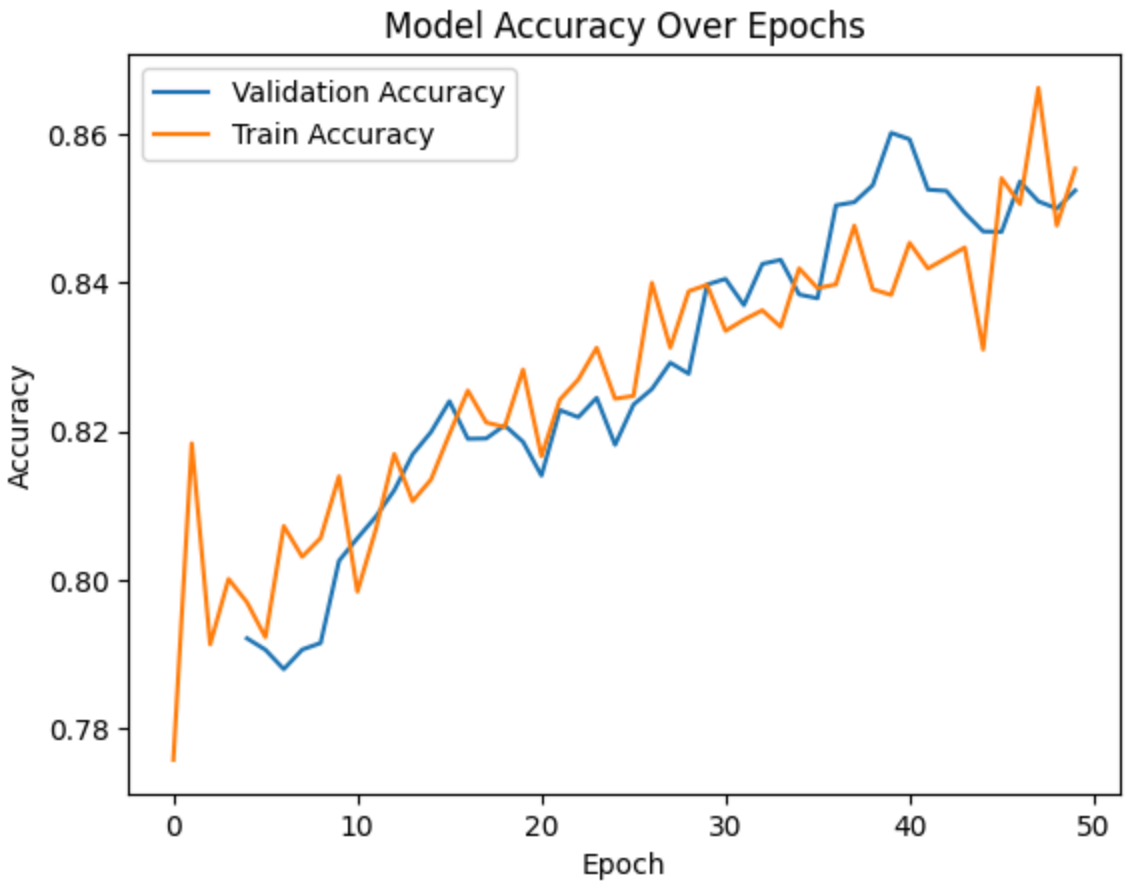

Figure 12. Accuracy graph over 50 epochs

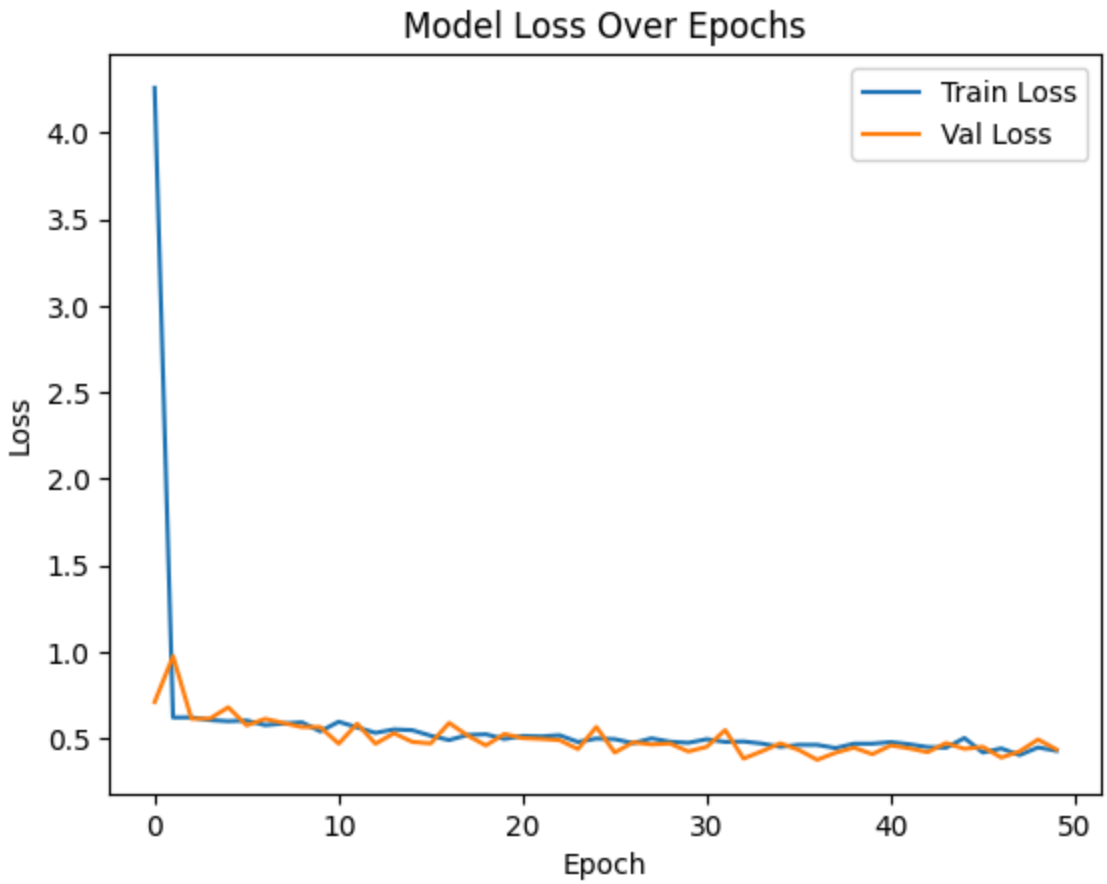

Figure 13. Loss over 50 Epochs

The training and validation accuracy and loss trends further reinforce this conclusion. Both training and validation accuracy improved steadily over the 50 epochs, while loss values were consistent, indicating effective learning without signs of overfitting. The fluctuations in validation accuracy are typical and expected in such tasks, but the overall alignment between training and validation performance underscores the model's stability and reliability.

In conclusion, the model's ability to maintain high and consistent accuracy across both balanced and real-world test sets highlights its practical utility. It successfully handles the class imbalance inherent in 311 requests while ensuring that less frequent departments are not ignored. This makes it a robust and valuable tool for streamlining departmental assignments in Boston's 311 system.

Conclusion

The results of our analysis suggest that machine learning models can play a significant role in improving the efficiency of municipal service delivery. A classifier like ours could help optimize

the assignment of requests to departments, potentially speeding up response times and thus improving the quality of services provided to residents. Moreover, these methods could be expanded to analyze patterns in service delivery, helping policymakers understand which types of requests are most common and how they are resolved (the departments that will respond etc.). Understanding these high-volume request types allows for better prioritization of resources and more strategic planning, ensuring that recurring problems are addressed proactively.

Additionally, incorporating spatial and temporal trends into the analysis could further enhance the understanding of service delivery patterns. For instance, location data can reveal neighborhoods with higher concentrations of certain issues, while timestamp data might highlight trends such as noise complaints peaking on weekends. These insights enable targeted interventions, such as scheduling additional resources during high-demand times or focusing maintenance efforts on frequently affected areas.

Beyond descriptive analytics, these methods pave the way for predictive modeling, which can forecast future service demands. For instance, historical trends in waste collection requests during holiday seasons could inform staffing and resource deployment plans for upcoming years. By combining these insights, policymakers can ensure that the city services remain effective and as the Bostonian community needs' evolve.

Appendix A

Link to Appendix