IST 407 Report

Introduction

With New York City's reputation as a high-crime urban metropolis, we wanted to try a new approach to conducting crime analysis and prevention. Our machine-learning approach aims to investigate whether the presence and distribution of businesses within the city can play a role in predicting and preventing criminal activities. Specifically, we intend to explore the possibility of strategically opening more businesses in areas with higher crime rates, thereby addressing the issue of crime and assisting in the broader objective of rebuilding trust between the New York City Police Department (NYPD) and the communities they serve.

While supporting the NYPD remains a crucial aspect of our machine learning approach, our main commitment lies in aiding our primary stakeholders, the NYC City Planners. Tasked with overseeing project planning and development, these planners play a pivotal role in enhancing the overall prosperity and well-being of New York City. Through our project, we aim to empower city planners to formulate strategies that foster the emergence of new business opportunities. By identifying areas where business density and distribution can influence crime rate reduction, our plans take a holistic approach to fostering growth and safety within the city.

Rising crime rates within the city affect the lives of millions of NYC residents and tourists. Previous studies have delved into the demographics and geographic features associated with crime data. Our unique approach seeks to reexamine this issue from a fresh perspective and contribute to the reduction of crime rates to transform New York City into a safer and more desirable environment.

Methodology

To address the critical issue of increasing crime rates in New York City, our approach is rooted in a comprehensive analysis of both the crime and business data we acquired from NYC open data–a resource that shares NYC data with the public. Since the datasets had millions of rows, our initial focus was to reduce the size of the data we had to work with to a reasonable amount. Starting with the crime dataset, which had over 8 million rows, we performed some preliminary inspection and examination of the data to see which instances were irrelevant to solving the problem. The instances we cut include the following:

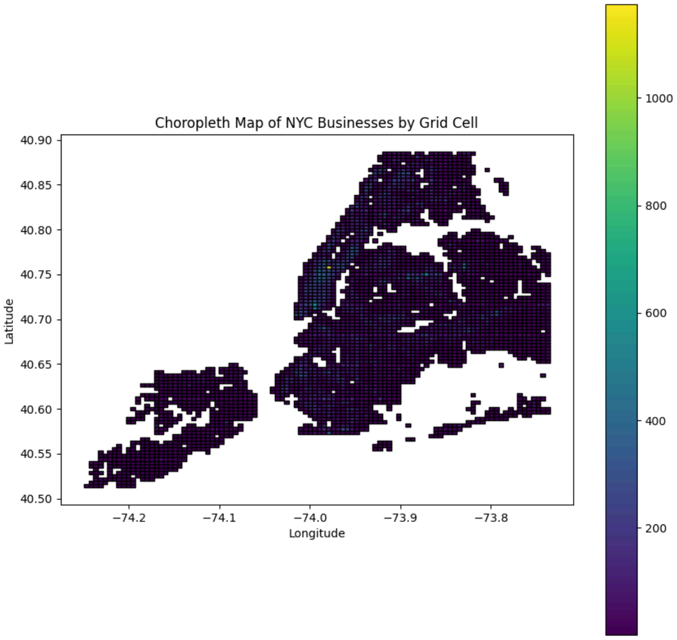
1. Crimes that were not at the level of a felony (including attempted crimes) - We wanted our analysis to focus on more serious crimes in order to provide the most value to our stakeholders
2. Crimes that had a local description of Residential - These instances were not relevant since we are focused on the relationship between crimes and businesses
3. Crimes that had no reported location - These instances added no value since our analysis is primarily focused on the location of crimes and businesses

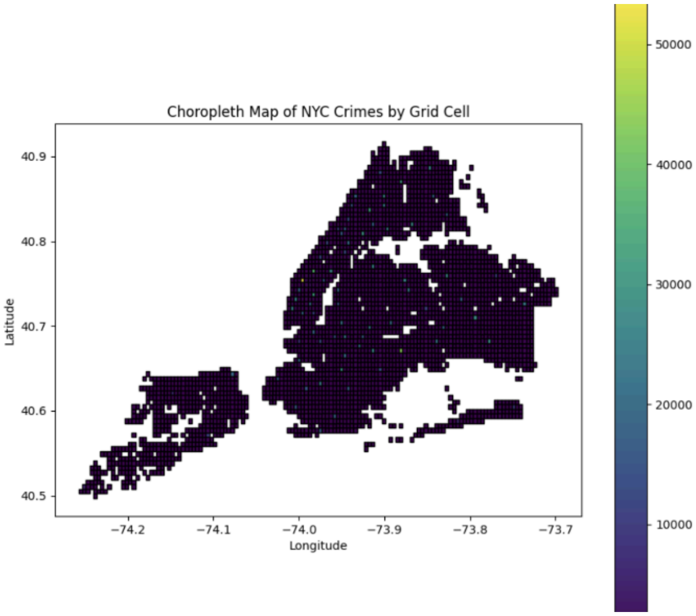
Similarly, we cut instances of businesses with an Individual license type for the NYC businesses dataset, which had over 250k rows, because none of them had location data. Next, to explore the potential relationships between business density/distribution and crime occurrences, we created a grid-based aggregation from both datasets' latitude and longitude columns. Using geospatial analysis tools like matplotlib allowed us to visualize the density and distribution of crime occurrences and business locations across the city. This approach led us to identify patterns and correlations within the data, uncovering hidden nuances that might have been overlooked in traditional analyses. By visualizing these patterns, we also provided NYC Planners with actionable insights to formulate targeted interventions in specific areas, addressing the root causes of crime and fostering community safety.

To predict potential crime hotspots and understand the underlying factors contributing to criminal activities, we ran a Random Forest classifier and looked at feature importance. We one-hot encoded our data, split it into train and test sets and fit our model accordingly. This predictive capability enables us to offer proactive solutions, empowering stakeholders to allocate resources efficiently and implement preventive measures effectively.

Results

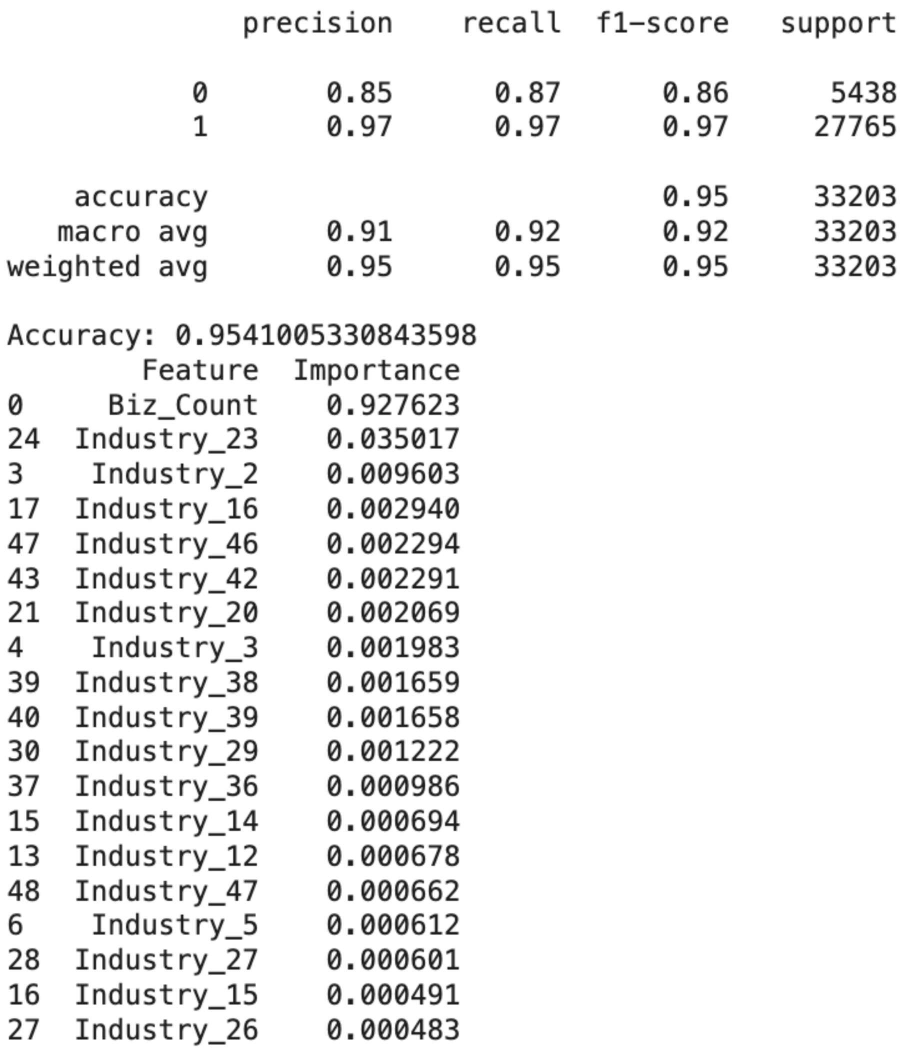
Following our data exploration and cleaning, we were able to aggregate points from both datasets onto separate grids, which were then plotted on a choropleth map of New York City. The visualization allowed us to gain insight into where the largest number of businesses are located and how crimes may subsequently be related.





After running our random forest classifier, we were able to graph out and view our feature importance. This allowed us to visualize what features are having the most effect on our predictive variable, whether or not a crime occurred. In order to create this predictive variable, we used the median of crimes committed based on the grid id calculated from our aggregated grid. Finding the median allowed us to create a threshold of the number of crimes committed as it relates to a certain area located on the grid. The median came out to 121 crimes, so our predictive variable crime occurred was measured by placing a 1 for each business that was located in an area where 121 or more crimes were committed and a 0 if there were less than 121 crimes committed. For our X variable in our random forest, we ran the model on the following features: Business count, a column recording how many businesses fell within a region on the grid, Each of the industries recorded in our business data frame, and business recorded net tangible assets.

The following image below shows the results of our random forest model.



```
              precision    recall  f1-score   support

           0       0.85      0.87      0.86      5438
           1       0.97      0.97      0.97     27765

    accuracy                           0.95     33203
   macro avg       0.91      0.92      0.92     33203
weighted avg       0.95      0.95      0.95     33203

Accuracy: 0.9541005330843598
        Feature  Importance
0     Biz_Count    0.927623
24  Industry_23    0.035017
3    Industry_2    0.009603
17  Industry_16    0.002940
47  Industry_46    0.002294
43  Industry_42    0.002291
21  Industry_20    0.002069
4    Industry_3    0.001983
39  Industry_38    0.001659
40  Industry_39    0.001658
30  Industry_29    0.001222
37  Industry_36    0.000986
15  Industry_14    0.000694
13  Industry_12    0.000678
48  Industry_47    0.000662
6    Industry_5    0.000612
28  Industry_27    0.000601
16  Industry_15    0.000491
27  Industry_26    0.000483
```

The most important feature in our model was business count. When diving into different industries, our model showed that home improvement and contractor industry businesses were the most important in predicting whether or not a crime occurred.

Conclusion & Recommendations:

We concluded that the results suggest valuable information to the NYC planners. Increasing the amount of businesses in an area of high crime can actually have a negative effect on stopping crime. It suggests that areas with a larger amount of businesses are more susceptible to crime. Thus, the information we can provide to our stakeholders is that when planning and developing projects, it is important to keep the number of businesses in a specific area in mind, as including too many businesses in a specific area may lead to more crime.

Although we did come up with a conclusion for our stakeholders, there is much more that we can do to better improve our results. This includes testing on different levels of crimes. Although felonies are more severe crimes and oftenly put citizens in more danger, they are usually not as common as other smaller level crimes. If we could have tested out different levels of crimes, we may have gotten different results, or noticed different trends that we could have reported. Another thing that we could have incorporated is exploring other features or sources of data. For example, time is most likely a significant factor and would most likely have an effect on our results. Exploring other aspects of the crimes and other outside sources could be conducted and may lead to better results and better analysis.

Acknowledgments:

This project was done in collaboration with Peter, Jess, Kunaal, Nick, & Alex. We utilized resources from our class and ChatGPT to assist with our code.

References

City of New York. (2023). NYPD Complaint Data Historic. data.cityofnewyork.us. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

City of New York. (2023). Legally Operating Businesses. data.cityofnewyork.us. https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh