New York City Crime Trends Prediction for Each Crime Classification in Each Borough

Brittany DiFede
Dept. of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, United States of America
bdifede@stevens.edu

Hanqing Liu
Dept. of Software Engineering
Stevens Institute of Technology
Hoboken, United States of America
hliu69@stevens.edu

***Abstract*—Due to the high population in New York City from not only its residents but also its tourists, it is important to ensure the safety of those who are living and traveling throughout New York City. In order to help reduce crime, as well as find ways to keep the public safe it is important to know how crime trends can change throughout each of New York City's boroughs over time. The team studied three different algorithms and applied these algorithms to New York City crime data in order to predict how crime rates for each major crime classification would change over time throughout each borough.**

I. Introduction

A. Background

Statistics in crime are important as they provide crime results that are systematic and quantitative. Having such statistics available serves as a resource to help find ways to reduce crime, improve procedures and training for fighting crimes, as well as provide the public with transparency regarding the crimes that are occurring in their areas. Additionally, these statistics can have an even greater impact by allowing for predictions to be made about potential crimes that could occur in the future allowing for an increase in safety precautions to be made. [10]

New York City is loved by many, and due to this is a home, as well as a vacation spot for millions making New York City very densely populated as seen in "Fig. 1". However, with such a large population New York City has become a major area in which crime occurs as seen in "Fig. 2". [9]

Every borough of New York City faces crime, however the amount of crime in each of these boroughs may vary. The crimes that can occur fall into three levels of classifications: felony, misdemeanor, and violations. [3] What separates crimes into these classifications is determined by the severity of the crime. [12] Therefore, it is important for people to be provided with the knowledge on the rates of these crimes in the area that they are in. Knowing the likelihood of what could occur, allows for individuals to know what to be careful of and what safety precautions to consider. Additionally, it allows for precautions and actions to be made to reduce these crime rates so that the overall safety of the public can increase. [10]

B. Problem Statement

The ability to make accurate predictions on future crimes is important for the public. Although New York City is a wonderful place for attractions, it is not free of crime and everyone living or visiting the city has the right to know the likelihood of potential types of crime they could be facing when traveling to different districts of New York City. These predictions will allow for people to have the knowledge regarding how crime levels can change in an area over time, thus allowing them to know what type of crime may be likely to occur in an area in the future. Additionally, it allows for law enforcement to utilize these trends to put safety measures in place to prevent these likely crimes from occurring. The team will utilize crime trends and statistics of crime location throughout New York City to provide predictions on the trends

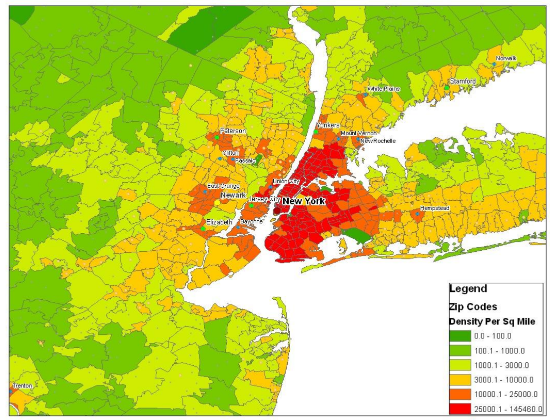


Fig. 1. New York City Population Density

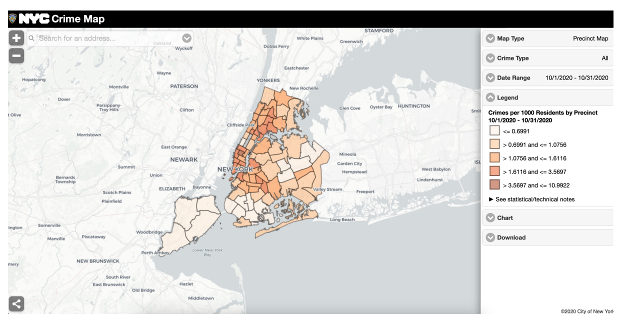


Fig. 2. Crime in New York City

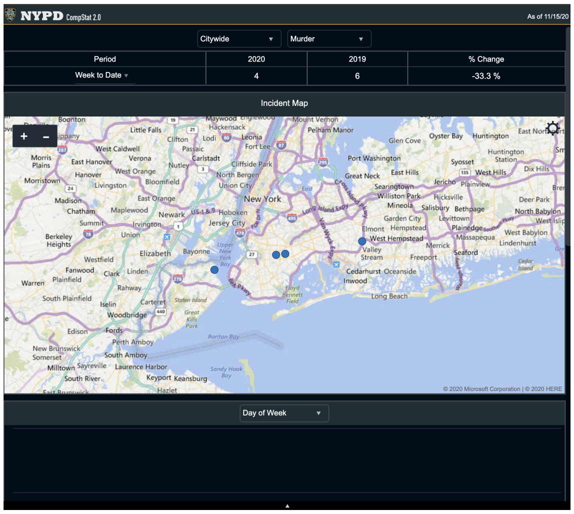


Fig. 3. CompStat Map

of crime within the three major crime classifications over time for each borough. [2]

II. Related Work

A. NYPD CompStat

There are various applications and studies that have worked with crime statistics. One of the largest applications used is the NYPD's CompStat Portal. The NYPD's COMPStat department created and maintains this portal as a way to provide the most up-to-date and historical crime-related statistics. This application serves as a management tool that the NYPD uses to notify the public and government on what is going on throughout the city, as well as a way to help them reduce crime in areas in which there are high levels of crime. [11] This portal allows for a map visualization of where any of the seven major crime categories have occurred throughout the different boroughs and precincts to be viewed. This map visualization can be broken down into various time categories and compares the crime levels for the indicated constraints for the current year to the previous year and displays the percent change over the two years. Additionally, this portal displays various graphs and charts to further narrow down the statistics. The different visualizations of the portal can be seen in "Fig. 3", "Fig. 4", "Fig. 5", and "Fig. 6". [5]

As seen from the NYPD's CompStat portal, this application focuses on past and current crime data trends rather than predicting future trends. Thus, looking at crimes in terms of what is currently happening rather than what could potentially happen in the future. The NYPD is able to utilize this data in order to determine protocols to take to reduce these crimes, which proves how useful crime statistic applications are for public safety.

B. Jackson State University Study

Another major study in the area of crime was conducted by Lawrence McClendon and Natarajan Meghanathan from

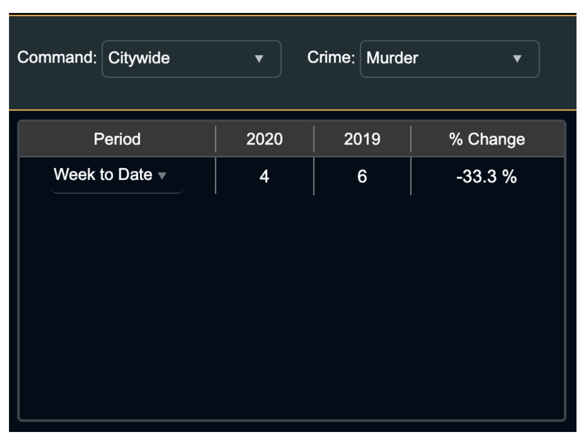
Fig. 4. CompStat Day of Week Graph

Fig. 5. CompStat Daily Graph

Jackson State University. In their study they implemented linear regression, additive regression, and decision dump algorithms to study violent crime patterns in Mississippi. This study utilized data regarding the number of crimes categorized as murder, rape, robbery, and assault within the different cities of Mississippi in order to estimate the crime totals for these categories. The study used the crime numbers per 100K to predict the projected total for the whole city. The results of this can be seen in "Fig. 7". [6]

Their conclusion to the study was that the linear regression algorithm worked best and gave the most accurate results. This was determined by comparing the error values of each algorithm which can be seen in "Fig. 8"

As seen from this study, crime statistics were utilized to make predictions. However, this study focused on estimating the current crime rate in a particular area based upon a crime


Command: Citywide
Crime: Murder

Period	2020	2019	% Change
Week to Date	4	6	-33.3 %

Fig. 6. CompStat Year Comparison Table

Table 12: WEKA Crime Totals

	Murder	Rape	Robbery	Assault
Mean per 100K	6	36	62	178
Projected Crime Totals	179	1,076	2,485	4,635

Table 13: neighborhoodscout.com Crime Totals

	Murder	Rape	Robbery	Assault
Mean per 100K	7	31	81	156
Estimated Crime Totals	195	930	2,409	4,680

Fig. 7. Jackson State Study Results

Table 11: Results for Violent Crimes per 100K of Population [Total Number of Instances - 1994]

Algorithm	*Correlation Coefficient*	*Mean Absolute Error*	*Root Mean Squared Error*	*Relative Absolute Error*	*Root Relative Squared Error*
Linear Regression Model	1	0.004	0.006	0.0009%	0.001%
Additive Regression Model	0.97	116	168	26%	27%
Decision Stump Model	0.78	276	379	62%	62%

Fig. 8. Error Results for Jackson State Study

rate of a smaller area, rather than predicting crime trends for the future. Additionally, this study predicted specific crimes rather than crime classifications. [6]

Overall during the team's research for the relevant work done in crime statistics, the team found that many crime studies use predictions for current trends rather than focusing on future trends. Additionally, the breakdown of borough and crime classification was not something that has been done for New York City. Therefore, the team felt that exploring this would be a great way for future trends to be predicted and utilized to increase the safety of those in New York.

III. Solutions

A. Data

The data used for this specific study was obtained from NYC OpenData, which is an organization that makes the wealth of public data generated by various New York City agencies and other city organizations available for public use. [8] The specific data-set utilized by the team was based upon historical NYPD Complaint Data. This data was made available to the public and included all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department. This contained the documentation of the crimes that were committed and the data related to each crime, such as the date of occurrence, the precinct in which the incident occurred, etc. The overall data-set included 6,983,207 over the span from 1970 until 2020. The team decided to narrow the records down and only include the records from 2010 on as these crimes would be more relevant and reflect the most recent population rates in different areas of the city, as well as the most to date laws and protocols that are in place. In order to do this, the team filtered the data from the NYC OpenData source to include only the relevant information after and including 01/01/2010. The team felt that this data would give an accurate presentation of the crime trends from 2010 until now, as well as help to provide an accurate prediction of what the crime rates in the future will be. This produced the data that was utilized.

There was pre-processing that had to take place in order to narrow down the data to exactly what was needed and would be utilized throughout the study. The data-set that only included the relevant time frame for the scope of the study, contained 35 relevant columns for each of the documented crimes. However, for the purpose of this study only date, crime classification, and borough were used. These columns were utilized to create a data frame that only included the information that was needed. Additionally, all rows that were empty were also dropped as only relevant crime data wanted to be used for the scope of the study. This information was the only information needed in order to predict the crime rate for each crime classification in each borough over time. Once the data was cleaned into the desired data-frame, it was ready to be utilized for each algorithm.

B. Algorithms

The three algorithms that the team decided were best to use for the study were linear regression, a neural network, and a decision tree.

Linear Regression is a machine learning algorithm that is supervised and is able to predict an output for continuous data. This method is often used for forecasting as it allows for target values to be predicted based upon predictor values that are independent. [4] This method finds the relationship between the independent variable and the dependent variable and is useful to predict a value based upon the value of another variable. [18] To find this relationship, this method uses the simple linear regression equation shown in "Eq. (1)".

$$y = b_0 + b_1 * x_1 + ... + b_n * x_n, \quad i = 1, ..., n \qquad (1)$$

The team decided to use linear regression to model the crime trend since linear regression works well with continuous data and the crime data used had a continuous time interval. Additionally, due to the wide use of regression analysis for predicting and forecasting future outcomes, the team felt that linear regression would be able to provide them with the future crime predictions that were wanted, as well as be able to help display the relationship between the different crime classifications in each borough over time.

A neural network refers to a series of algorithms that have the ability to identify relationships within a data-set through a process that simulates how the human brain functions. It conducts its performance based upon a set of connected nodes that resemble neurons in the brain. A diagram of neural networks can be seen in "Fig. 9". [14]

For the purpose of this study a neural network was used for regression. Neural networks have the ability to be reducible to regression models and have the ability to provide much more prediction power for a regression problem compared to a traditional regression approach. [13] Therefore, the team felt

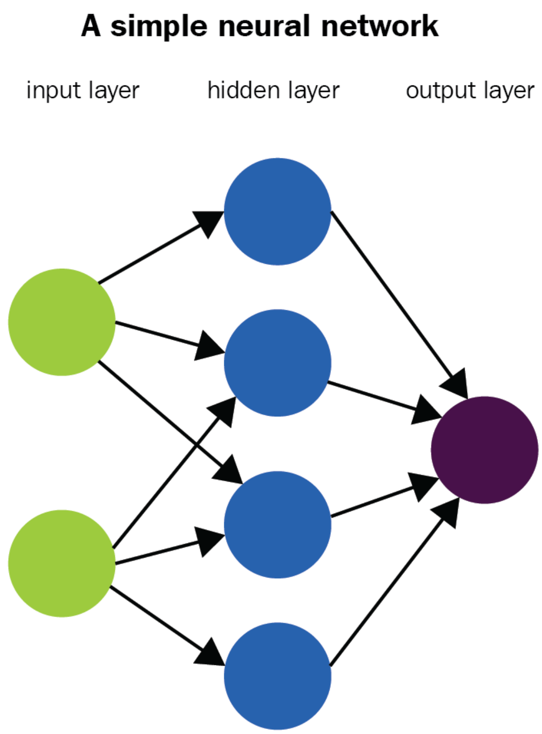


Fig. 9. Neural Network

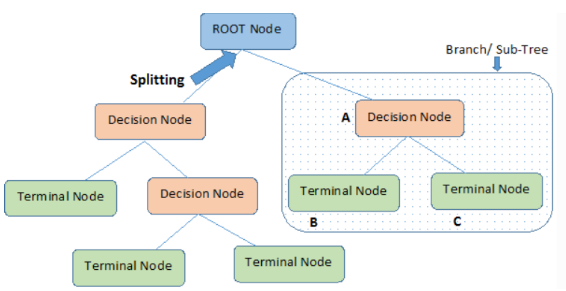


Fig. 10. Decision Tree

that utilizing a neural network to help predict future crime trends would be extremely powerful and provide accurate prediction results.

Decision trees can be used to visually display decisions through predictive modeling where observations are used in order to make conclusions about a target value. These trees utilize a set of rules that split the data in order to find the target value. This tree is made up of a root node, decision nodes, and terminal nodes. The root nodes represents the entire data-set that will get divided, the decision nodes are the nodes where the data-set gets split further based on conditions, and the terminal node is the last node along the branch of the tree that will no longer be split. [1] This structure is shown in "Fig. 10".

Decision trees can be used for regression since a decision tree is able to build a regression model in the form of a tree like structure. [1] It does this by taking the data-set and continuously breaking it down so that smaller subsets are created through decision nodes. These nodes test the features and split the data based upon different conditions. These decision trees have the ability to take continuous values for the target variable, making them great for regression. [17] Due to this, the team felt that a decision tree would be a great method to use for regression to predict the future trends of crime.

C. Implementation Process

1) Pre-Processing: Before any of the algorithms could be implemented for the data, the team had to perform data pre-processing. The steps for the pre-processing were the same for each borough and occured before each algorithm got be implemented for that borough. These steps are as followed: the specific borough was obtained, date conversions were calculated, daily crime rates were obtained for each classification, yearly crime rates were obtained for each classification, the data was split into independent and dependent variables, and the data for each classification was split into training and testing. In order to display the crime trend over time for each classification for each of the boroughs, the team felt that it was best to display each borough separate. The team felt that this was best as each borough would have their own crime trends for the three major crime classifications: felony, misdemeanor, and violation. Therefore, the pre-processing steps had to be applied to each classification for each borough so that each algorithm could be used to predict the trend for each borough separately.

Going into the above pre-processing steps into greater detail, the data was first filtered for the specific borough being predicted. Once that was done, the date was converted to date time and then mapped to ordinal values in order to convert the date to categorical values so that it was easier to find a relationship with. Once that was done, the unique dates were found in order to determine what the x-axis should be. After that was completed, dictionaries were created for each crime classification that would store the crime count for each unique date value of the x-axis. Each date was a key and the value for each of these keys would hold the corresponding crime count. Then, the team went through each row of the file for each classification and for that classification counted the number of crimes within that classification for each day. From there, the team had to take the daily crime rates and add them up so that the total crime counts for each classification for each year was obtained. This was done by creating dictionaries that stored the year as the key and had the crime count as they key values. For each value in the dictionary that held daily counts, if the year belonged to the same year as the key in the year dictionary, the crime count was added. Once the crime count per year was obtained for each classification, the data was ready to be split.

For the splitting of the data, it first had to be split into the independent and dependent variables. Due to the date being continuous, it was assigned as the independent variable for each classification and the classifications themselves were assigned as the dependent variables. Once that was completed,

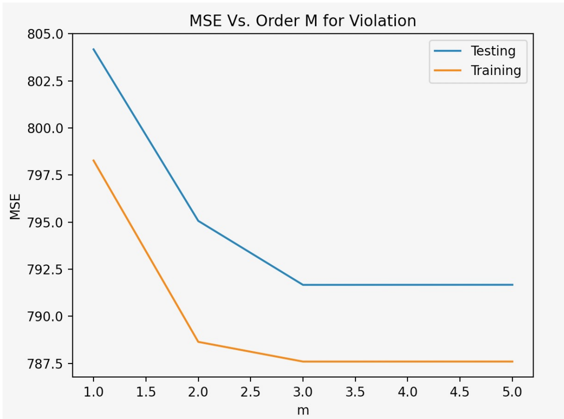


Fig. 11. m value for Violation

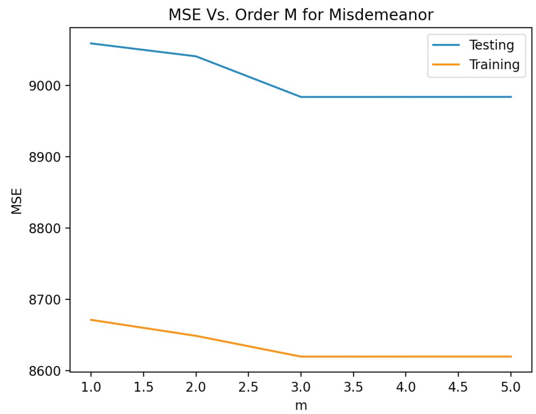


Fig. 12. m value for Misdemeanor

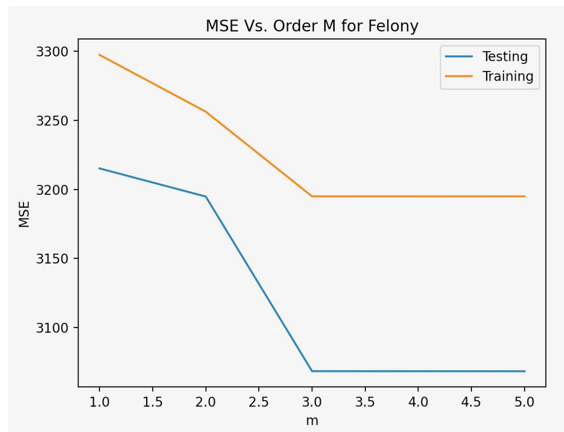


Fig. 13. m value for Felony

the data was split for each classification into the training and testing data. It was split by using 80 percent of the data for training and the other 20 percent of the data for testing. Once the data was split for each classifications within each borough, the models were ready to be implemented.

2) Linear Regression: The first machine learning algorithm that the team implemented was linear regression. Before linear regression was applied for each borough, the parameters were tuned so that each borough could have the same parameters applied and so that the parameters used would produce the best output that would have the least error with no over-fitting or under-fitting. In order to tune the parameters the whole cleaned data-set was used that contained all of the boroughs. This tuning was done in order to find the optimal m value that would produce the lowest MSE value. This was done by fitting the violation data, felony data, and misdemeanor data all with various values of m to determine which m value produced the lowest error for all three classifications.

For Violation, the optimal m value found was 3 as at this value the MSE is the lowest for both the training and testing data and then remains the same for all m values after as seen in “Fig. 11”. For Misdemeanor, the optimal m value found was 3 as at this value the MSE is the lowest for both the training and testing data and this MSE remains the same for all m values after as seen in “Fig. 12”. For Felony, the optimal m found was 3 as at this value the MSE is lowest and then remains the same for all m values after for both the training and testing data as seen in “Fig. 13”. Therefore, all three classifications have the best value of m at 3 for linear regression.

Once the best m value was found for all three classifications, the process for predicting the crime trends began. The training and testing data found from the pre-processing was able to be fit with the linear regression algorithm in order to find the trend line that would correspond with the classification. This was done by utilizing the polyfit function as fitting the data into this function allowed for the coefficients for the polynomial of degree n that best fit the data to be returned. Once this was done, poly1d was used to find the one-dimensional polynomial class associated with the polynomial that was found to best fit the data. Once the model was obtained, it was able to be used to predict the testing set and determine outcomes.

The above steps for linear regression were repeated for each classification and the three trend lines for the classifications were plotted on a graph so their trends over time could be seen for each of the perspective boroughs. Each of these graphs for each borough can be seen in “Fig. 14”, “Fig. 15”, “Fig. 16”, “Fig. 17”, “Fig. 18”.

Based upon the results of these graphs, the team was able to conclude that linear regression worked well on the data-set in terms of predicting crime trends. As seen the trend for each classification in each borough varied and by having these trends it will allow for protocols and safety measures to be put into place in order to reduce the amount of crime that may be expected in that area based upon how the trend line is moving as the years progress. Therefore, overall linear regression produced accurate trend curves that show the crime trends over time. This would allow for the use of these trends to predict future crimes.

3) Neural Network: The next machine learning algorithm that was implemented was a neural network. Just like linear regression, the parameters for the neural network had to be

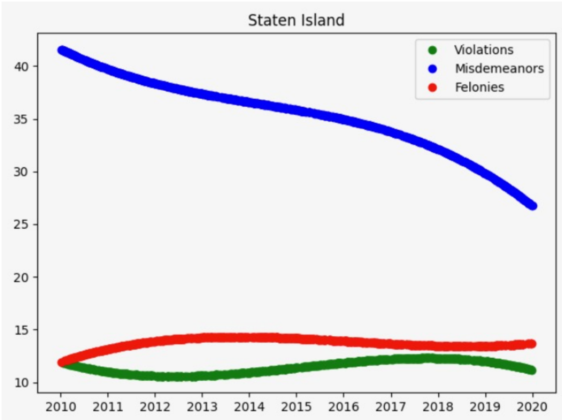


Fig. 14. Linear Regression: Staten Island Crime Trends

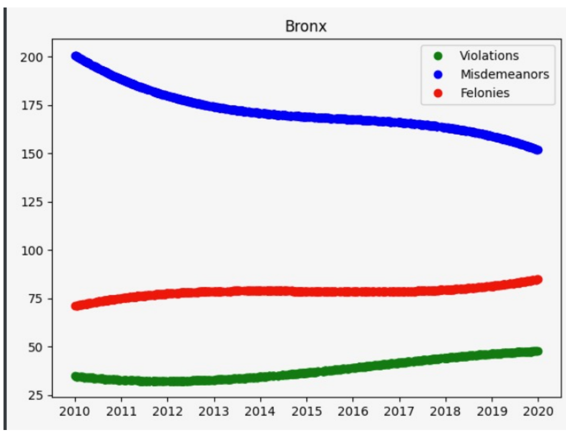


Fig. 15. Linear Regression: Bronx Crime Trends

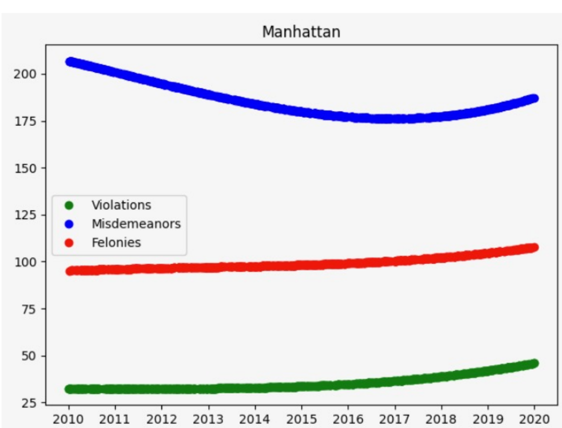


Fig. 16. Linear Regression: Manhattan Crime Trends

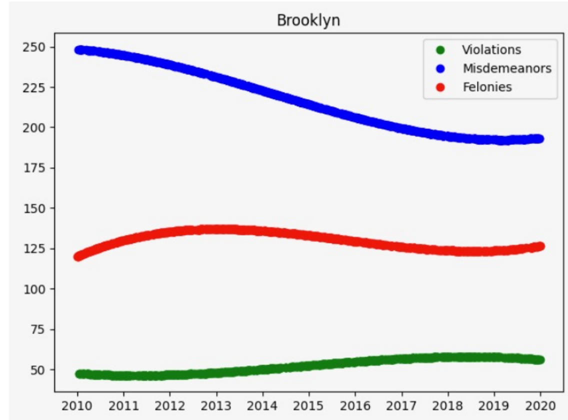


Fig. 17. Linear Regression: Brooklyn Crime Trends

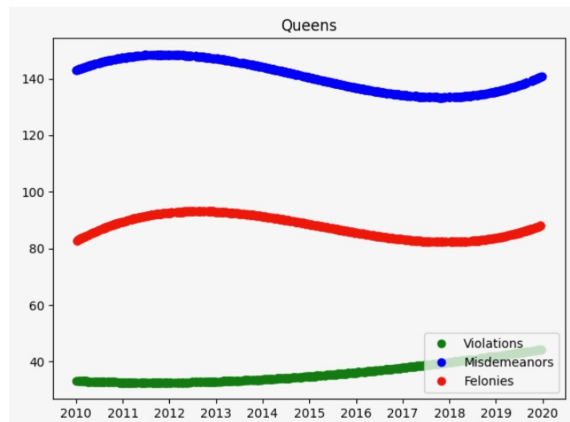


Fig. 18. Linear Regression: Queens Crime Trends

tuned so that each borough could have the same parameters applied for each classification. Additionally, this was done to ensure that the parameters used to apply the algorithm would be the parameters that would produce the least amount of error and therefore the most accurate results. The parameters that the team decided to use and therefore tune were learning rate, iterations, and the hidden layer amount.

The learning rate was the first parameter tuned in order to see what learning rate would produce the lowest MSE value. The learning rate is a hyper-parameter for a neural network that controls the amount that the model should change by in response to the predicted error every time the weights of the model get updated. Therefore, this parameter had to ensured that it was not too low or too high. If the value was too small the process of training would be too long, but if it was too large it could result in a training process that was unstable.

For violation, the best learning rate value was determined to be 0.05 since at this learning rate both the training and testing data had the lowest MSE value and then this MSE remained the same for the learning rate of 0.06 as seen in “Fig. 19”. For misdemeanor, the best learning rate was also determined to be 0.05 since at this learning rate the training data was at its lowest MSE and the testing data had the lowest MSE at this

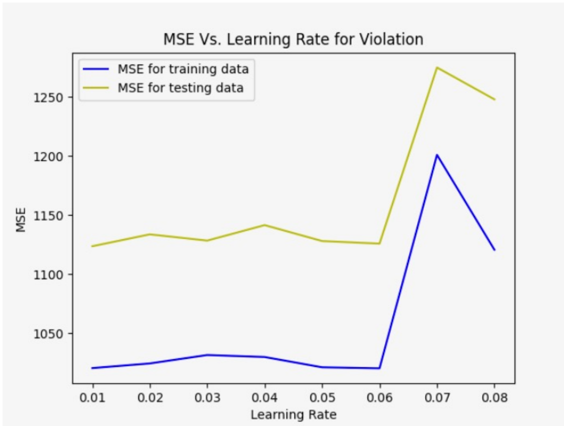


Fig. 19. Learning Rates for Violation

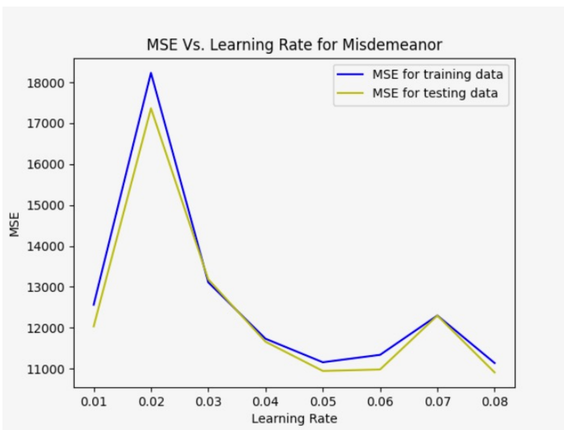


Fig. 20. Learning Rates for Misdemeanor

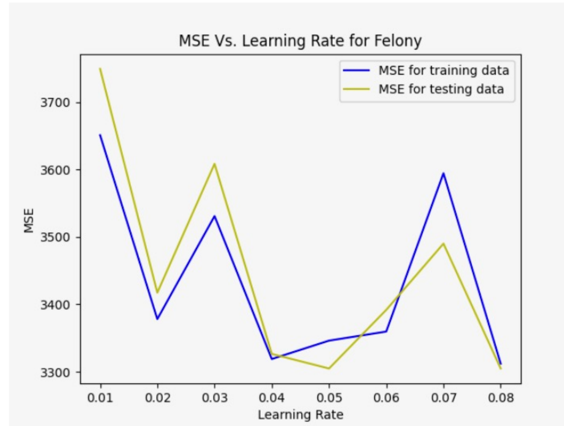


Fig. 21. Learning Rates for Felony

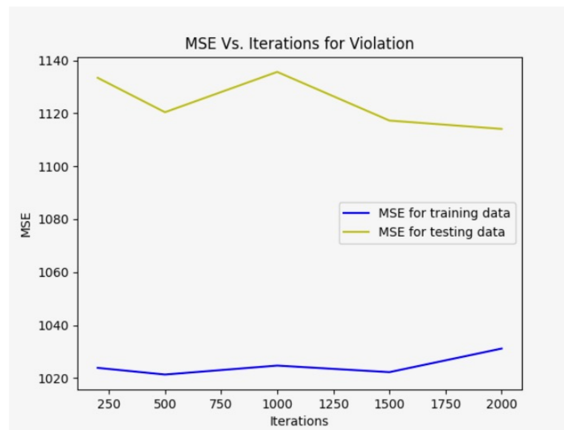


Fig. 22. Iterations for Violation

value and it remained the same for the learning rate of 0.06 as seen from “Fig. 20”. For felony, the best learning rate was also determined to be 0.05. As seen in “Fig. 21”, the training data had the lowest MSE value when the learning rate was 0.04 and the testing data had the lowest MSE value when the learning rate was 0.05. Therefore, in order to determine the best overall learning rate for felony, both of these learning rates had to considered. Ultimately, the team decided that 0.05 would be the best learning rate since it was one of the two leaning rates since this produced a lower MSE than that of 0.04 when 0.04 caused the lowest MSE value and additionally this value was the best for the other classifications as well. Therefore, all three classifications were determined to have the best learning rate at 0.05.

The second parameter that had to be tuned for the neural network was the iterations in order to determine the amount of iterations that would produce the most accurate output for each classification. The iteration indicated the amount of times a batch of the data would pass through the algorithm. Therefore, it had to ensured that this was not set too high where it would be passed through too many times causing over-fitting, as well as ensure that it was not too low that it wouldn’t be passed through enough causing not enough learning.

For violation, it was determined that both the training and testing data had the lowest MSE value for 500 iterations as seen in “Fig. 22”. Therefore, this was chosen as the best amount of iterations for the violation classification. For misdemeanor, it was determined that the best number of iterations would be at 200 because at 250 as seen in “Fig. 23”, the iterations produce the lowest MSE. However, below 250 the MSE value was continuing to decrease thus indicating that below 250 iterations would produce the lowest MSE value for both the training and testing data-sets. Therefore, the team determined an iteration value of 200 to be the best iteration value that would produce the least amount of error. For felony, it was also determined that the best iteration value would be at 200. Felony was the same as misdemeanor where below 250 iterations seemed to produce the lowest MSE value as seen in “Fig. 24”, thus causing 200 iterations to be picked as the best iteration value for felony too. The testing data for this seemed to also be low at 1500 iterations but due to the downward curve before 250 the team stuck with the decision of 200 iterations.

The last parameter that was tuned was the number of hidden layers. Hidden layers are the layers that are between the input

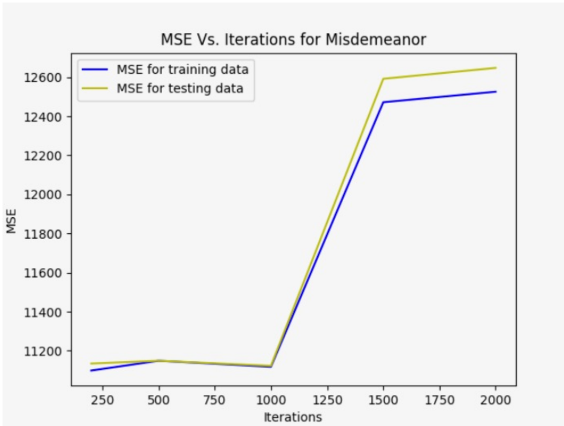


Fig. 23. Iterations for Misdemeanor

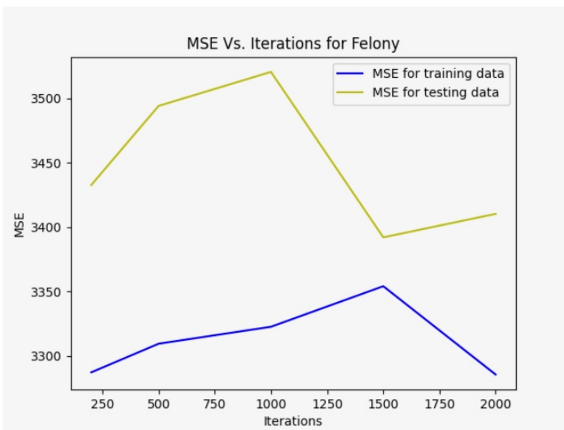


Fig. 24. Iterations for Felony

and output layer. An amount of hidden layers that is too small can cause under-fitting, leading to the neural network not being trained well enough. Two many layers can lead to over-fitting, causing the model to be trained too well and not perform well on the testing data. Thus the right amount of layers had to be chosen.

For violation, as seen in “Fig. 25”, the MSE for hidden layers less than 100 for both the training and testing data steadily declined indicating that the lowest MSE value would be less than these values. At 60 hidden layers, the lowest MSE was produced but due to the steady decline it was observed that below this value would produce an even lower MSE. Therefore, 40 was determined as the most optimal hidden layer amount for both the training and testing data. As seen in “Fig. 26”, misdemeanor was the same as violation with the lowest MSE of hidden layers at 60 and less for both the training and testing data, therefore the hidden layer value was determined to be 40. For felony, the lowest MSE value for the testing data was at a hidden layer value of 100 and the lowest MSE for the training data was at a hidden layer value of 160 as seen in “Fig. 27”. However, when trying to perform with this number of hidden layers it was computationally

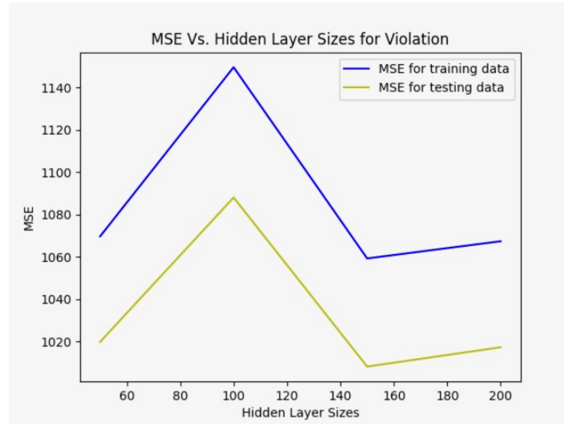


Fig. 25. Hidden Layers for Violation

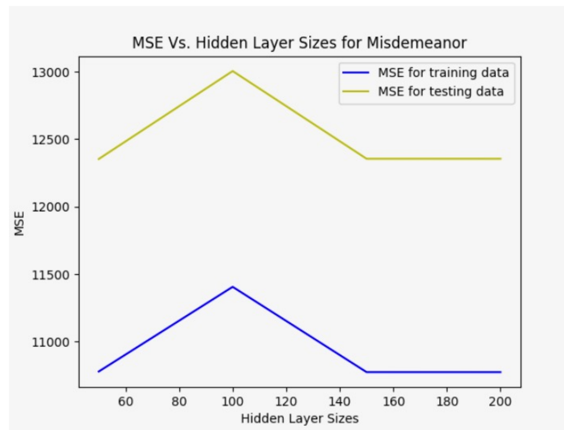


Fig. 26. Hidden Layers for Misdemeanor

difficult. Therefore, since both of the other classifications had a hidden layer value of 40, 40 was also chosen for the felony classification since it was much easier to compute. The team knew this may cause some error for this classification, however hoped that with a smaller complexity it would reduce the error that could occur.

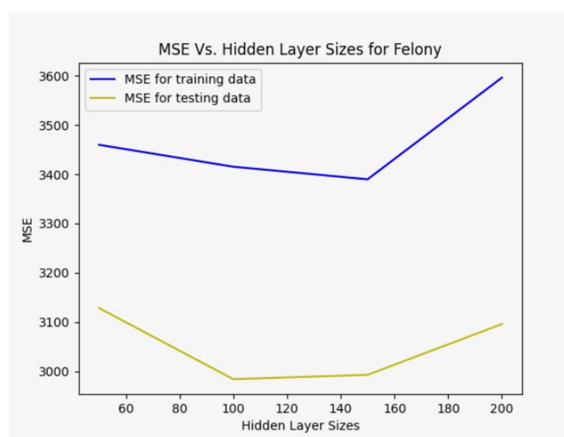


Fig. 27. Hidden Layers for Felony

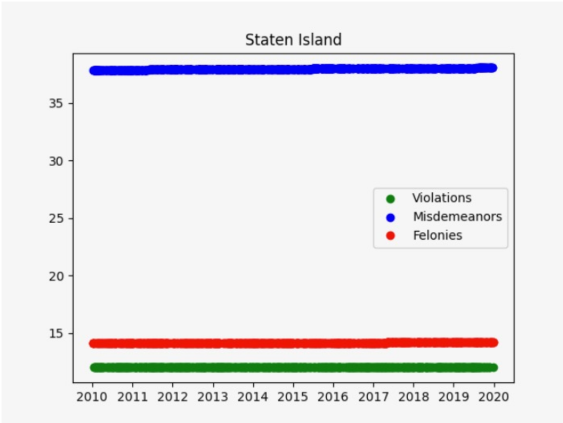


Fig. 28. Neural Network: Staten Island Crime Trends

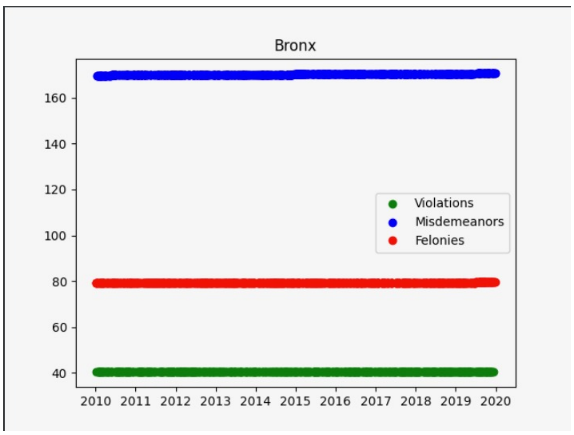


Fig. 29. Neural Network: Bronx Crime Trends

Once the parameters were tuned, the same implementation of the algorithm was repeated for each borough just like how it was for linear regression. In order to implement neural network regression analysis, the MLPRegressor library from sklearn was utilized. This model is a multi-layer perceptron model regressor that optimizes the squared-loss using LBFGS or stochastic gradient descent. For the purpose of this implementation, the stochastic gradient based optimizer was used since it was the optimizer that the team was most familiar with.

As seen in “Fig. 28”, “Fig. 29”, “Fig. 30”, “Fig. 31”, and “Fig. 22”, the curves produced by the model almost look straight with very subtle changes, indicating that this algorithm did not perform very well. Even with tuned and adjusted parameters, the neural network did not seem to change much for the curves producing curves that were not of great accuracy. There, it was concluded that this algorithm did not predict extremely accurate trends and would not be the best algorithm for predicting the future trends needed.

4) Decision Tree: The last algorithm that was implemented was the decision tree. Just like the other two algorithms that were implemented, the decision tree algorithm had to have parameters tuned as well in order to ensure that the most

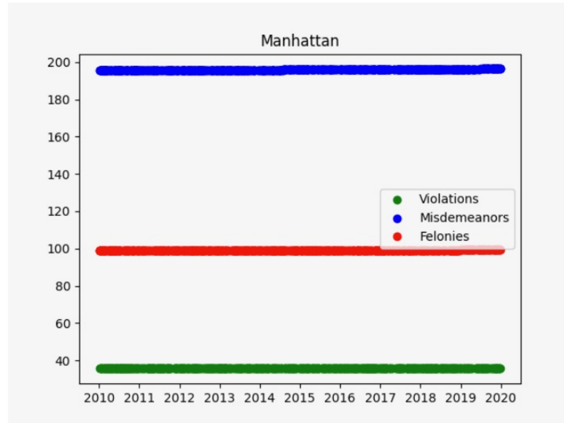


Fig. 30. Neural Network: Manhattan Crime Trends

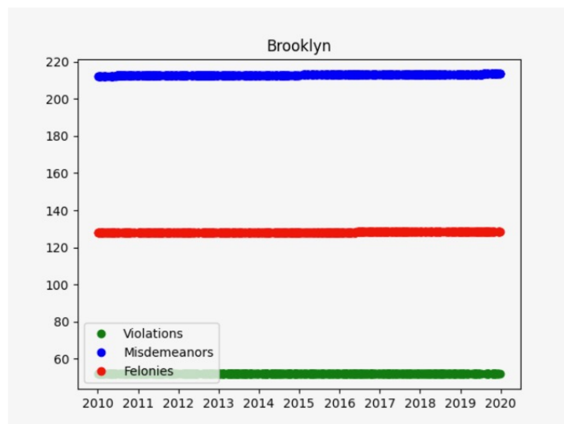


Fig. 31. Neural Network: Brooklyn Crime Trends

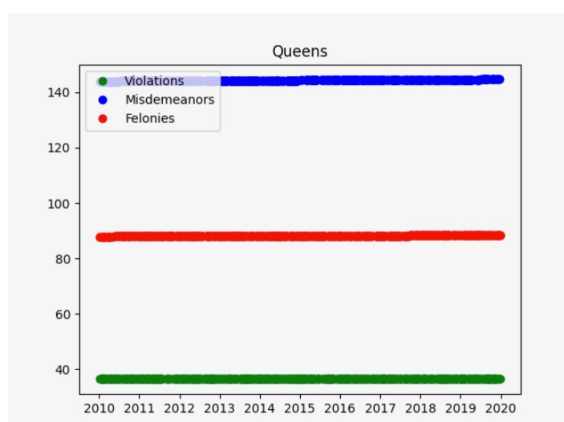


Fig. 32. Neural Network: Queens Crime Trends

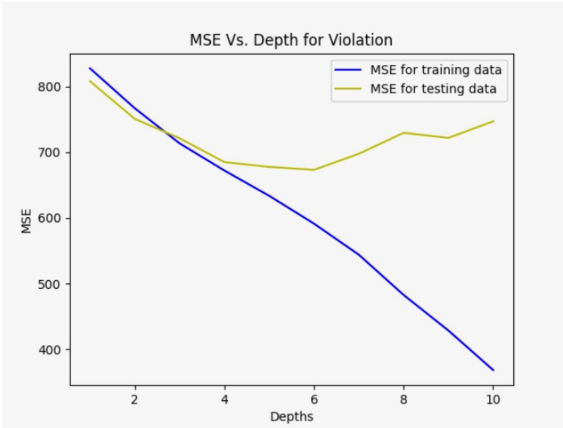


Fig. 33. Depths for Violation

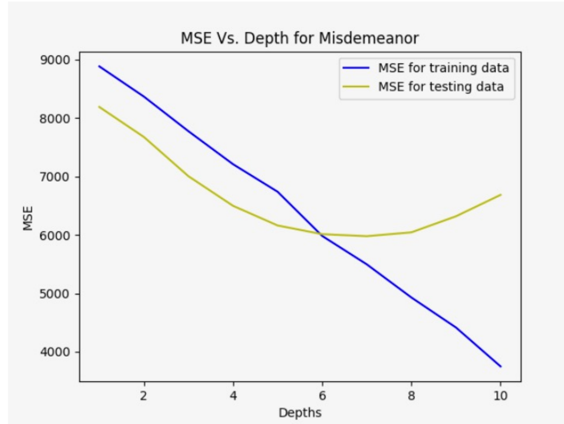


Fig. 34. Depths for Misdemeanor

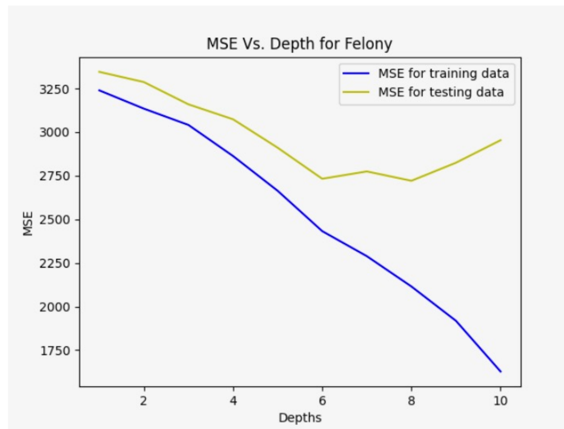


Fig. 35. Depths for Felony

optimal parameters were used that would produce the lowest error and thus the best output. The parameters that the team had decided to apply to the algorithm and therefore tune were depth and leaf nodes. The team had decided to use these two parameters as they felt that these would have the greatest impact on how the algorithm performed.

The first parameter that was tuned was the depth. Usually, the deeper that the tree grows the more complex the model will be. This will cause more splits within the decision tree allowing for more information about the data to be captured. However, if the depth is too great there will be over-fitting because the model will fit perfect for the training data but will not be able to perform well on the test data. Therefore, the team wanted to tune this parameter so that the best depth would be chosen that would allow for the training data to be fit as best as possible, while allowing for the test data to be generalized well with no over-fitting.

For violation, the best value for depth was determined to be 6 since at this depth value, the testing data had the lowest MSE value and the training data at this value was continuing to get lower as seen in “Fig. 33”. Therefore, to satisfy both the training and testing data a depth of 6 was concluded to be best. If the value at the lowest MSE for training was picked, this would not satisfy the testing data and thus produce over-fitting. For Misdemeanor, the best value was chosen as 6 since at this value the testing data was at its lowest and the training data was continuing to get lower as seen in “Fig. 34”. Therefore, again a depth of 6 was determined to satisfy both parameters without causing over-fitting. For felony, the best depth was determined to be 6 since again the testing data had produced the lowest MSE for depths of 6 and 8 and the MSE for the training data continued to get lower at this depth value as seen in “Fig. 35”. Due to the testing data having the same lowest MSE value produced at depths of 6 and 8, the depth of 6 was concluded as the best depth since it remained the same for 8. Additionally, a depth of 6 satisfied both the training and testing data without producing over-fitting. Therefore, all three classifications had the best depth value determined to be 6.

The second parameter that was tuned was the amount of leaf nodes. The same as with the depth parameter, the team wanted to ensure that the right amount of leaf nodes would be chosen that would allow for no over-fitting to occur. The leaf nodes represent the classification or decisions, therefore the team wanted to ensure that not too many decisions or classifications were being made that allowed for the model on the training data to be a perfect fit but then perform poorly on the test data thus the model being over fit. Therefore, the team tuned this parameter to get the leaf nodes that would be most optimal for both the training and test data.

For violation, the best value for the leaf nodes was determined to be 40 since at this value the testing data produced the lowest MSE and the training data at this value continued to get lower as seen in “Fig. 36”. If the leaf node value that produced the lowest value for the training data was chosen, this would cause over-fitting due to the training data being fit too well causing the model to be over-fit and not perform well on the testing data. Therefore, to satisfy both the training and testing data without being over-fit, the leaf node value was determined to be 40. For misdemeanor, a leaf node value between 30 and 40 produced the lowest MSE value for the testing data, while for the training data MSE values continued to decrease between these values. Therefore, a leaf node value of 40

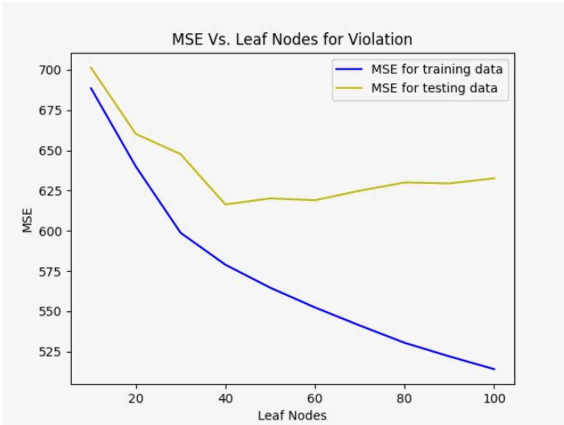


Fig. 36. Leaf Nodes for Violation

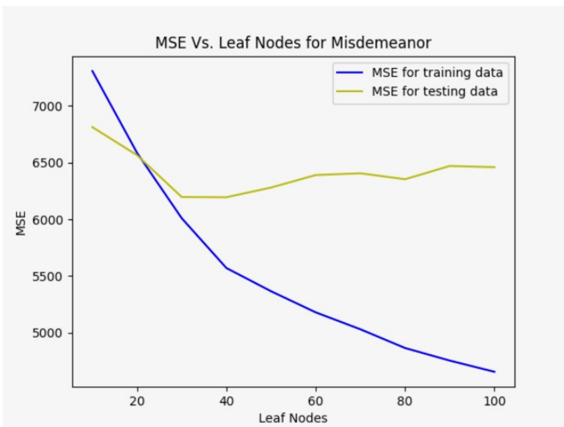


Fig. 37. Leaf Nodes for Misdemeanor

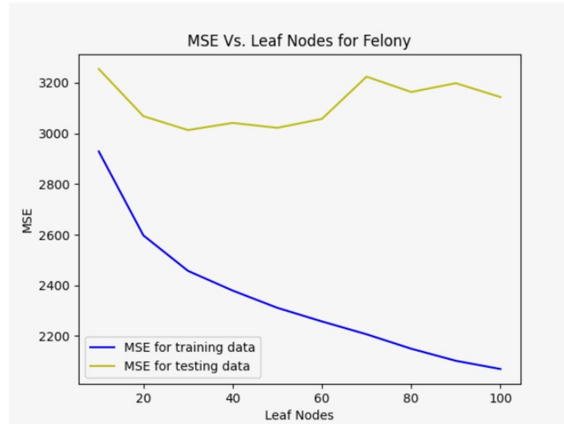


Fig. 38. Leaf Nodes for Felony

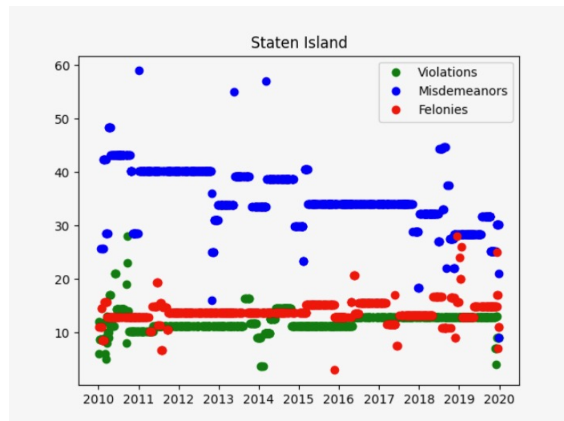


Fig. 39. Decision Tree: Staten Island Crime Trends

was determined as the best leaf node value since similarly to violation, at 40 the misdemeanor's testing data was at its lowest MSE and its training data continued to decrease for values of MSE as seen in "Fig. 37". Additionally, at this value the MSE for the training data was lower than it was at a leaf node value of 30, thus causing the team to decide 40 as the best leaf node value rather than 30. This value again allowed for both the training and testing data to be satisfied without producing over-fitting. Lastly, for felony the testing data was at its lowest MSE between 30 and 50, while the training data continued to get lower as seen in "Fig. 38". Therefore, the best MSE value was determined to be 40 since this was in the range of the lowest MSE for the testing data while producing a lower MSE for testing without causing an over-fit of the model. Overall, all three classifications had the best leaf node value of 40 that would produce the best results without an over-fit.

Once the parameters were tuned for the decision tree, the model was able to be implemented. In order to perform decision tree regression analysis, the sklearn decision tree regression library was utilized. This library allowed for a 1D regression with a decision tree to be performed by using the decision tree to fit a sine curve with the addition of a noisy observation. With this, the decision tree learns local linear regression approximating the sine curve. Once the model was fit and trained, it was able to be utilized to make predictions of the training data.

The above steps for decision tree regression analysis were repeated for each classification and the three trend lines for the classifications were plotted on a map so their trends over time could be seen for each of the perspective boroughs. Each of these graphs for each borough can be seen in "Fig. 39", "Fig. 40", "Fig. 41", "Fig. 42", "Fig. 43".

Overall, as seen from the graphs each classification in each borough contained pretty accurate trends indicating that the algorithm performed well. However, even though the predictions looked accurate the trends were not as smooth as that of the linear regression curves. Additionally, it shows that the parameters used, the depth and leaf nodes, were set well since the decision tree didn't learn too fine of details in the training data causing an over-fit. Therefore, since no under-fitting or over-fitting seemed to occur and the trends looked like they were portrayed accurately, the team was able to conclude that the decision tree regression worked well on the crime data-set to predict the trends, thus allowing for future crime predictions to be made.

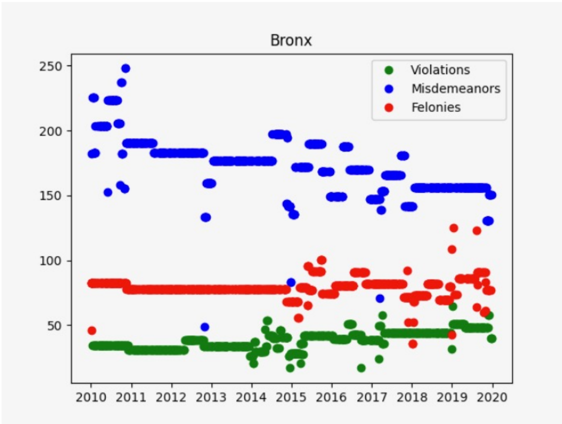


Fig. 40. Decision Tree: Bronx Crime Trends

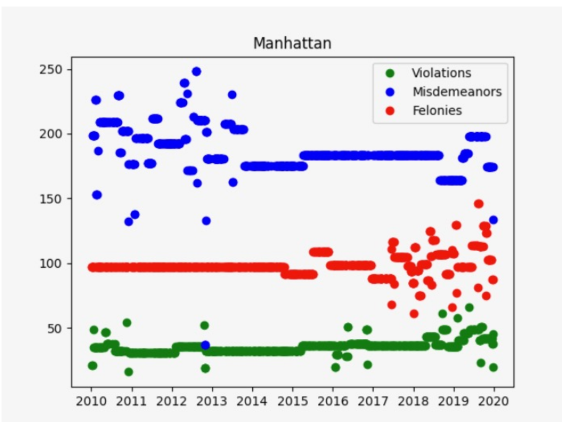


Fig. 41. Decision Tree: Manhattan Crime Trends

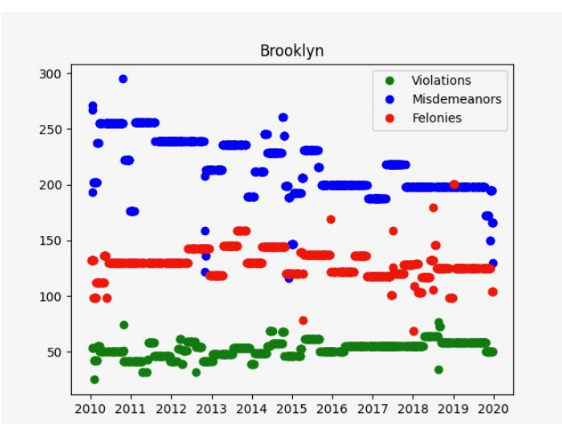


Fig. 42. Decision Tree: Brooklyn Crime Trends

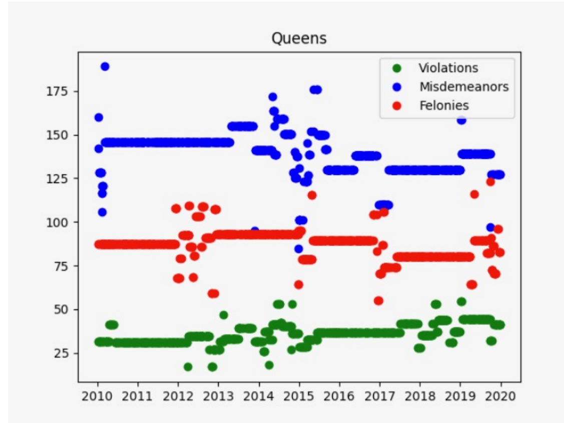


Fig. 43. Decision Tree: Queens Crime Trends

IV. Comparison

A. Accuracy and Computational Cost

In order to determine which of the algorithms performed the best and would be best suited to use for predicting future crime trends, the accuracy and computational cost of each algorithm for each classification were computed.

First, the computational cost was determined. Among the three models, linear regression was the model that had the lowest computational cost due to its lowest computational complexity of $O(nm^2)$. In this complexity n represents the sample count and m represents the feature count. The decision tree implementation has the median computational cost in between linear regression and neural network, with a computational complexity of $O(mn^2log(m))$. Just like with linear regression, in this complexity n represents the sample count and m represents the feature count. The most computationally heavy model utilized was the neural network with a computational complexity of $O(nmoih^k)$. In this complexity, o represented the output neuron count, k represented the number of hidden layers, h represented the amount of neurons in each layer, and i represented the number of iterations. Since the data used by the team had 6.9 million entries, the most time consuming area of the study was in the pre-processing that was done, since this required reading the data from a hard disk to memory, getting rid of empty data, and sorting these entries by report date. Overall in terms of computational complexity, linear regression was the simplest to use.

After the computational cost was calculated, the MSE for each algorithm was calculated. It was determined that the decision tree had the lowest MSE for all three classifications for both the training and testing set, followed by linear regression, and then neural network. At the start of the study, the team had predicted that the linear regression algorithm was going to produce the lowest MSE, however the decision tree ended up producing the lowest MSE and thus the most accurate results. The decision tree had the most accurate trends and with this, the trends outliers can also be seen on the graph indicating any value which may have been unexpectedly low or high in comparison to the other years. This made the trend look less smooth, however this reduction in smoothness was actually more accurate to the true results. Due to this, the decision tree would be the best algorithm to use to predict future crime trends thus providing the most accurate predictions for what trends will be. Even though, it was a bit more computationally heavy than linear regression, the team felt that due to its higher accuracy it was better to use despite the slight increase in computational cost. Linear regression had curves that were similar to that of decision tree showing that it was also accurate just not as accurate. The neural network however had the highest MSE and its trends did not look as accurate as the other algorithms, thus indicating that this would be the worst algorithm of the three implemented to use to predict future crime trends.

TABLE I
Computational Complexity

	LR	DT	NN
Complexity	$O(nm^2)$	$O(mn^2log(m))$	$O(nmoih^k)$

TABLE II
MSE

	LR	DT	NN
Violation Train	787	607	1058
Violation Test	792	672	984
Misdemeanor Train	8778	6023	11165
Misdemeanor Test	8356	6757	10733
Felony Train	3291	2518	3236
Felony Test	3012	2884	3628

B. Advantages/Disadvantages

There are advantages as well as disadvantages to all of the algorithms used. These advantages and disadvantages cause one algorithm to be used over another.

Linear regression has many advantages that cause it to be a great algorithm to use. Linear regression is simple to implement and due to that it is easier to interpret its output coefficients. Its simplistic implementation makes it much easier to use and get predictions. Additionally, when the relationship between the independent and dependent variables have a relationship that is linear, it is the best to use since it becomes less complex to use than the other algorithms. This was true in terms of the team's implementation as linear regression was very easy to use and not complex. Additionally, even though linear regression is susceptible to over-fitting just like many machine learning algorithms, this over-fitting can be easily avoided by using dimension reduction techniques as well as cross-validation. For the team's implementation they were able to reduce the dimensions by removing columns and also determined the best order to use, thus avoiding over-fitting, ensuring that the algorithm was implemented correctly. [19]

Even though linear regression has many advantages to its implementation, it also has many disadvantages that cause other algorithms to be used over it. One disadvantage is that often outliers in the data can cause huge effects on the regression and boundaries in this technique are linear. This could have been a problem within the data as if there were outliers in the crime numbers, there could have been distorted trends. Additionally, with this algorithm it is assumed that the relationship between the independent and dependent variables is linear. Thus, it assumes that there is a relationship that is a straight line between the variables, therefore assuming independence among the attributes. This disadvantage did not seem to be a problem for the team's implementation as the time variable increased so it was always moving in one direction and could not move backwards. Therefore, the relationship did correlate with a straight line just the trend increased and decreased. Additionally, linear regression examines the relationship between the mean of the independent and dependent variables, thus linear regression may not be a full description of the variables relationship due to the mean not being a full description of a single variable. [19]

Neural networks have many advantages that cause it to be used over other machine learning algorithms. This algorithm requires less formal and statistical training thus making the training process a bit quicker. Additionally, this algorithm can detect relationships between independent and dependent variables that are complex and linear. Due to the amount of data used, this was a beneficial advantage in the case of any outlier variables being present that may have not been linear as well as the complex nature of using such a large dataset. This algorithm also has the ability to detect the possible interactions between the predictor variables which was helpful for the team's implementation as it would determine any of the interactions that may have been occurring between the predictor variables used. [20]

Neural networks also have disadvantages that cause them to be less likely to use. One disadvantage is the computational burden of this algorithm. This was seen from the team's implementation that the neural network took much more computational power than that of the other algorithms used. Additionally, this algorithm is very prone to over-fitting. This was seen by the implementation as the team's trend produced by neural network was not as accurate as that of the other algorithms, which may have been a possible cause of over-fitting even though the parameters were tuned to avoid this. Additionally, neural networks "black box" nature is another huge disadvantage as it allows for it to be unknown of why the neural network came up with a certain output. This did prove true for the team as the trend did not match that of the other algorithms and it was unclear as to why. [20]

Decision tree regression has many advantages that cause it to be used. It has a clear visualization, is simple, and easy to understand. This proved true with this implementation as it was easier for the team to implement and understand. Additionally, due to the clear visualization it makes it easier to see the relationship. Additionally, it is able to handle both variables that are continuous and categorical which proved useful to this implementation as the time variable was continuous and the classifications were categorical. Additionally, no feature scaling is required since the approach is more rules based which made it easier to implement in the case of the data utilized. Decision tree algorithms are also robust to outliers which was beneficial as if there were outliers in the data the decision tree was able to handle them well. There is also a smaller training period for this algorithm, which allowed it to perform quickly. [21]

Decision tree regression also has disadvantages that lead it to not to not be used. One disadvantage is that over-fitting is likely to occur. This was seen with the team's implementation as they had to ensure that the parameters were tuned well so that over-fitting did not occur. If the parameters were not tuned well and over-fitting did occur, it would lead to incorrect predictions. Additionally, decision trees are likely to have a high variance due to the likelihood of over-fitting. However, since the team was careful not to over-fit, it would prevent this high variance from occurring. Additionally, this algorithm can be unstable when a new data point gets added due to the nodes having to be recalculated. However, since new data points were not added this disadvantage was avoided. This algorithm is also likely to be affected by noise which then causes predictions that are incorrect and may not be suitable for large data-sets since this can lead to a complex tree and therefore over-fitting. Even though the team tuned the parameters to avoid over-fitting, since the data-set was large there is still a possibility of over-fitting to have occurred. [21]

V. Future Research Directions

There are three future research directions that were identified to further improve the algorithms and expand the study. These directions include utilizing data reduction techniques, breaking the classification down even further, and breaking the boroughs down into precincts.

Utilizing data reduction techniques would prove beneficial as the team wanted to reduce the amount of columns used so that only necessary columns would be utilized. The team had done this manually by dropping and removing the columns that they had believed to be unnecessary for the purpose of their study. However, utilizing a data reduction technique such as PCA or ensemble trees would allow for the dimensions of the data to be reduced by transforming the larger data-set of variables into a smaller data-set that still contained most of the information from the original data-set. By the team manually removing columns themselves, they may have gotten rid of columns that could have improved their algorithm's accuracy and using one of the techniques would have allowed for other columns that could have improved accuracy to be determined and therefore kept. Additionally, the use of data reduction techniques could have allowed the team to increase efficiency and performance and reduce the computational cost by reducing the data to a level that would have allowed for more accurate and efficient results. Therefore, this is a direction that the research should follow in the future.

Breaking down the classifications even further into the specific crimes within these classifications would have allowed for the study to be expanded. This expansion would have allowed for more detailed and granular trends to be determined that would have provided the public with rates of specific crimes rather than just the crime classification. With how the study performs now, it provides the public with trends of the future classifications within their borough, however this does not show what particular crimes within the classifications that one might face. Thus, people may be taking precautions for all crimes within an area that a specific crime classification is high, however what may be causing the crime classification to be high may not be all crimes within that classifications but rather only a few. These few would be the crimes that the people should take precautions for. Therefore, expanding the study to contain this breakdown would allow the precautions that people need to take to be more clear.

Lastly, the boroughs could have been broken down into precincts, which would have provided a more precise location for crime and thus more accurate crime trends for specific areas rather than the more generalized borough. Similarly to how the classifications are broken down currently, the study only covers trends in a broad area. However, this does not necessarily mean that the crime classification rate is high in all precincts within that borough. A few precincts within that borough may be high, thus causing the overall borough predictions for that classification to be high. This breakdown would provide the public with more detailed predictions on the areas in which crime is the most prevalent, as well as what areas will be most prevalent in the future. This would therefore, allow people to have a better indication of what areas within a borough they need to take precautions for crime in rather than thinking that the crime in all areas of the borough precautionary measures are needed for.

VI. Conclusion

Overall, there was much concluded from this study conducted. The main thing that was concluded was that machine learning algorithms are successful in predicting future crime trends. The implementations showed the ability of machine learning algorithms to take in crime data and make predictions, thus proving the team's thesis that crime trends can be predicted within the three major crime classifications over time for each borough. This in turn allows for the trends to be utilized to see how the trend will continue for future years, therefore making future predictions on crime.

Additionally, it was concluded that crime predictions are important as seen through the research that was conducted for this study. Knowing crime trends has allowed law enforcement to prevent crime. Crime trends also vary per location so it is important to utilize crime trend data to know how severe a crime rate can be in a particular area.

It was also concluded that decision tree regression worked best for crime trend predictions as it produced the lowest MSE values for all the classifications in both training and testing data, thus proving that its trends were the most accurate. Linear regression also worked well as its trends were similar to that of the decision tree model, however this algorithm had a greater MSE so was not as accurate as that of the decision tree. Lastly, neural networks did not perform well for crime trend predictions since they produced the most error with the most inaccurate trends.

Overall, this study helped conclude that machine learning algorithms are a good way to predict future crime trends and thus provide transparency to the public, as well as help to reduce crime in particular areas of high predicted crime.

VII. References

References

[1] Drakos, Georgios. "Decision Tree Regressor Explained in Depth."*GDCoder*, GDCoder, 5 Feb. 2020, gdcoder.com/decision-tree-regressor-explained-in-depth/.

[2] E, Amina. "Using Machine Learning for Crime Prediction." *Technology and Operations Management*, 13 Nov. 2018, digital.hbs.edu/platform-rctom/submission/using-machine-learning-for-crime-prediction/.

[3] FindLaw. "Classifications of Crimes." *Findlaw*, Thomson Reuters, 2020, criminal.findlaw.com/criminal-law-basics/classifications-of-crimes.html.

[4] Gandhi, Rohith. "Introduction to Machine Learning Algorithms: Linear Regression." *Medium*, Towards Data Science, 28 May 2018, towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a.

[5] Harvard Kennedy School. "Compstat: A Crime Reduction Management Tool." *Compstat: A Crime Reduction Management Tool — Government Innovators Network*, www.innovations.harvard.edu/compstat-crime-reduction-management-tool.

[6] McClendon, Lawrence. "(PDF) Using Machine Learning Algorithms to Analyze Crime Data." *ResearchGate*, Mar. 2015, www.researchgate.net/publication/275220711_Using_Machine_Learning_Algorithms_to_Analyze_Crime_Data.

[7] New York City Police Department. "Crime Statistics." *Crime Statistics* - NYPD, www1.nyc.gov/site/nypd/stats/crime-statistics/crime-statistics-landing.page.

[8] NYC OpenData. "NYPD Complaint Data Historic." *NYC OPen-Data*, 2020, data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/data.

[9] NYPD. "NYC Crime Map." *NYC.gov*, maps.nyc.gov/crime/.

[10] NYPD. "NYPD Announces Citywide Crime Statistics for September 2020." *The Official Website of the City of New York*, 2 Oct. 2020, www1.nyc.gov/site/nypd/news/pr1002/nypd-citywide-crime-statistics-september-2020.

[11] NYPD. "NYPD CompStat 2.0." *Compstat.nypdonline.org*, compstat.nypdonline.org/2e5c3f4b-85c1-4635-83c6-22b27fe7c75c/view/90.

[12] PatrickAdmin. "Differences Between a Violation, a Misdemeanor, and a Felony." *Law Office of Patrick V. Parrotta*, 4 Nov. 2015, patrickparrottalaw.com/differences-between-a-violation-a-misdemeanor-and-a-felony/.

[13] Saporito, Gerry. "What Is a Perceptron?" *Medium*, Towards Data Science, 17 Sept. 2019, towardsdatascience.com/what-is-a-perceptron-210a50190c3b.

[14] Shukla, Pratik. "Main Types of Neural Networks and Its Applications-Tutorial." *Medium*, Towards AI, 28 Aug. 2020, medium.com/towards-artificial-intelligence/main-types-of-neural-networks-and-its-applications-tutorial-734480d7ec8e.

[15] Tiwari, Ritika. "Regression vs Classification in Machine Learning: What Is The Difference?" *Springboard Blog*, 17 June 2020, in.springboard.com/blog/regression-vs-classification-in-machine-learning/.

[16] Vadapalli, Sricharan. "Hands-on DevOps." *O'Reilly Online Learning*, Packt Publishing, 2020, www.oreilly.com/library/view/hands-on-devops/9781788471183/6c5143e3-f9ad-4571-82e2-986a34bbcfdd.xhtml.

[17] Yadav, Prince. "Decision Tree in Machine Learning." *Medium*, Towards Data Science, 23 Sept. 2019, towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96.

[18] Ye, Andre. "5 Regression Algorithms You Need to Know-Theory Implementation." *Medium*, Analytics Vidhya, 18 Nov. 2020, medium.com/analytics-vidhya/5-regression-algorithms-you-need-to-know-theory-implementation-37993382122d.

[19] AmiyaRanjanRout, "ML – Advantages and Disadvantages of Linear Regression" *GeeksforGeeks*, 06 Mar. 2020, https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/

[20] Jack V.Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes", *ScienceDirect*, 7 Sept. 1995, https://www.sciencedirect.com/science/article/abs/pii/S0895435696000029

[21] Naresh Kumar, "Advantages and Disadvantages of Decision Trees in Machine Learning", The Professional Point, 24 Feb. 2019, http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of.html