Polynomial Interpolation and K-Mean Mining to Predict Crime Rates

Anish Goyal

All charts, graphs, photos, and diagrams are the product of the student researcher.

Introduction/Brief Background

Atlanta has a 108% higher crime rate than the national average. It is within the top 3% most hazardous cities in the United States with approximately 30,000 crimes annually and a 61% crime rate per capita ("Atlanta crime rates," 2019). Given that Atlanta is such a high-profile city with its large number of crimes, the police department cannot respond to each case individually without wasting excess human resources. They must be led to high-crime locations ahead of time. Effective patrols in crime-heavy areas can be established when the best prediction model is used to determine the locations with the greatest crime rates. This can be applied to society because the police force can routinely deal with crimes before they ever occur on a daily basis, as their mere appearance is often sufficient to prohibit crimes from happening. This is where machine learning comes in. Using various machine learning algorithms on released data for past crime occurrences can provide valuable insight as to how crimes are distributed geographically, and which areas are the most crime heavy. Ultimately, as long as useful data is provided into the model—assuming that the best machine learning algorithm is selected, which is what I am trying to determine—the overall rate of crime will continue to fall.

Purpose

The purpose of the machine learning program is to predict and visualize the occurrences of different crime types for each neighborhood in Atlanta for the year 2019 after learning from crimes in the years 2009-2018 and comparing which algorithms are the most accurate with the actual crime counts from 2019 using the Atlanta PD's released COBRA datasets. The accuracy of the results will be a percentage measured by taking the average of the percent differences of each of the crime types for a neighborhood against the actual outcomes. This is needed because identifying clusters of high criminal activity permits the Atlanta PD to assign optimized police patrol routes and other crime-prevention measures like street cameras and neighborhood watches in the areas that need them the most, which allows the police department to efficiently allocate human resources while preemptively stopping large amounts of crime from occurring.

Hypotheses AND Engineering Goal

- Experimental hypothesis: Decision trees provide the most accuracy when compared to other supervised and unsupervised machine learning algorithms and are easy to process and implement.
- Null hypothesis: Decision trees do not provide the most accuracy when compared to other supervised and unsupervised machine learning algorithms and are difficult to process and implement.
- Engineering goal: To determine which machine learning algorithm has the most accuracy by making a program that can preemptively determine crime intensities in specific Atlanta neighborhoods and visualizing them.

Materials

Materials (detailed list); Quantity	
2009-2018 & 2019 COBRA Data sets; 2	Python libraries (Sci-kit learn, pandas, numpy, matplotlib, seaborn); 5
Excel; 1	Atlanta Beats Shapefile; 1
Computer with python3 installed; 1	GitHub; 1
Anaconda Navigator; 1	Jupyter Notebook; 1

Statistics/Observations

- Decision trees were the third-most accurate algorithm behind random forests and Naive-Bayesian classifiers with an accuracy of 42.9% compared to 51.5% and 63.9. It is possible that overfitting may have occurred, especially in the Naïve-Bayesian classifier, as it had an extremely low RMSE value for its extremely high amount of accuracy, which means that it will have poor performance when compared with real-time data.
- Adding features to the regressors and classifiers of the random forest algorithm showed a 7.22% increase from 47% accuracy to 52% accuracy as well as tripled the RMSE value of the regressors, showing an increase in accuracy. Random forests had the largest RMSE value peaking at 88.83 after applying PCA with regressors, while linear regression had the smallest RMSE value with classifiers only at 0.13.

Data

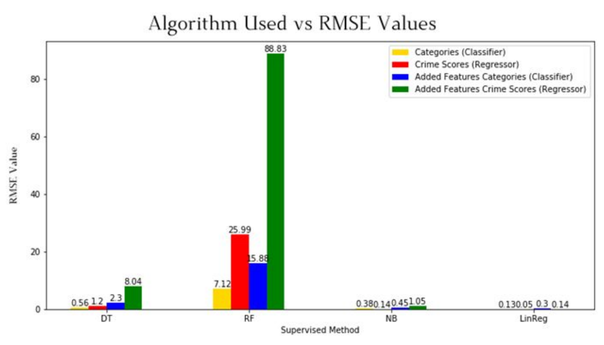
Decision Trees	Root Mean Squared Error Value (dimensionless)	Accuracy (%)
Downtown	0.56	41.24
Greenbriar	1.20	48.93
Wildwood (NPU-C)	2.30	54.77
…Georgia Tech (243)	8.04	65.31
Average	3.0123	42.904

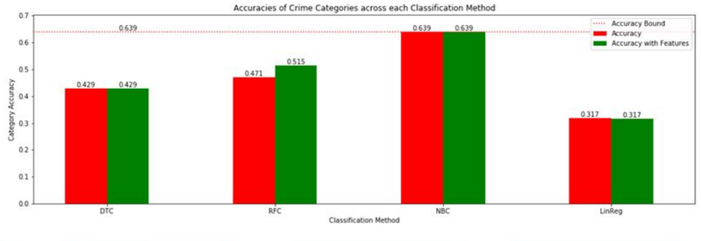
Naïve-Bayes	Root Mean Squared Error Value (dimensionless)	Accuracy (%)
Downtown	0.38	32.20
Greenbriar	0.14	81.17
Wildwood (NPU-C)	0.45	50.34
…Georgia Tech (243)	1.05	70.69
Average	0.402	63.903

Random forests	Root Mean Squared Error Value (dimensionless)	Accuracy (%)
Downtown	7.12	46.83
Greenbriar	25.99	50.21
Wildwood (NPU-C)	15.88	56.44
…Georgia Tech (243)	88.83	49.72
Average	34.411	52.500

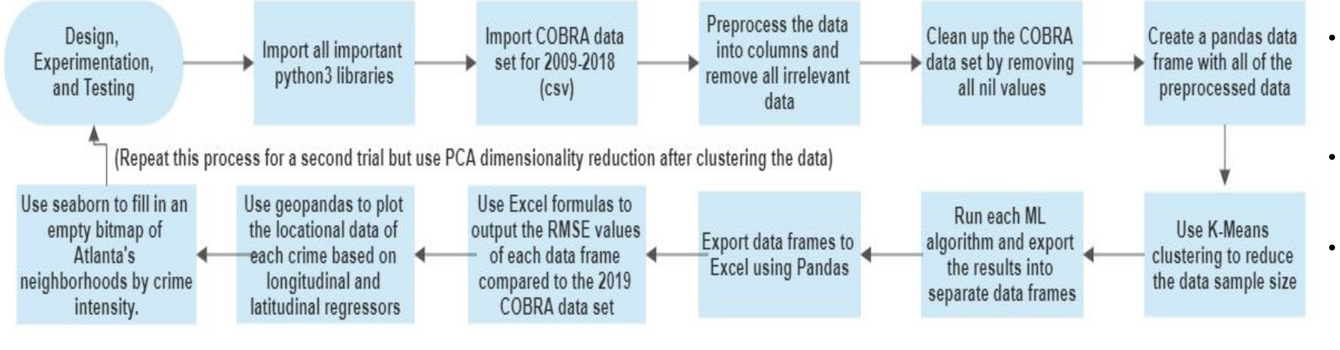
Linear Regression	Root Mean Squared Error Value (dimensionless)	Accuracy (%)
Downtown	0.13	34.24
Greenbriar	0.05	39.17
Wildwood (NPU-C)	0.3	28.28
…Georgia Tech (243)	0.14	26.88
Average	0.158	31.701

Graph





Procedure



Conclusion

In this experiment, I discovered that the supervised Naïve-Bayes classifier algorithm had the best accuracy, then random forests, and then decision trees, which is the algorithm I hypothesized would have the most accuracy in comparison to the COBRA 2019 data set; however, this was incorrect after looking at the summarized data. Decision trees were the third-most accurate algorithm behind random forests and Naive-Bayesian classifiers with an accuracy of 42.9% compared to 51.5% and 63.9% after PCA reduction, respectively. This shows that decision trees did not have the best accuracy in comparison to the other data sets.

Applications

- Implementing similar prediction models using neural networks
- Merge predictive models with socioeconomic models for each neighborhood to apply population regressors to make the data more realistic
- Knowing the best supervised and unsupervised algorithms for statistical analysis of future crimes
- Further testing is necessary to determine whether overfitting occurred for the Bayesian model.