NYC Crime Prediction Final Project Proposal

Denis Gostev, Mary Krakowski, Moriyah Schick

2. Project Description.

Our project seeks to predict the type of crime of a given crime complaint for different areas of New York City based on historical city crime complaint data. There are three different categories of crime we seek to predict: felonies, misdemeanors, and violations. Since all of our data is labeled with the category of crime this is a supervised machine learning task.

3. Team members' roles.

- Mary
  - Exploratory Data Analysis. Feature selection
  - Build Decision Tree Model
  - SVM binary and multi classification
  - Explainability
  - Final selection and analysis of the best model
- Moriyah
  - Exploratory Data Analysis. Resolve concern of bias in crime prediction
  - Utilize dimensionality reduction on our dataset
  - Implement Logistic Regression model for prediction
  - Build ANN model for prediction
  - Final selection and analysis of the best model
- Denis
  - Exploratory Data Analysis. Missing data clean up.
  - Feature engineering
  - Development of KNN Algorithm for prediction
  - Clustering
  - Final selection and analysis of the best model

4. CSCI 740 related topics (questions and solution approaches) each individual will contribute to the project.

Exploratory Data Analysis Explore the data, engineer new features if needed, identify redundant features, remove or replace missing data, remove outliers and irrelevant observations, visualize the data.

Feature Engineering We plan to convert some available data into new categorical features to improve performance of our models. For instance, we plan to create a 'time of day'

feature and convert the time of incident occurrence into one of the categories such as ‘morning’, ‘afternoon’, ‘night’. Similarly, we plan to introduce a day of the week, month or week of the year. We also plan to explore additional resources for supplemental meaningful information to bring novelty and accuracy to our algorithms. Such data may be population density in a given region, holiday calendar, time of sunrise/sunset, and weather information related to precipitation.

Dimensionality Reduction Dimensionality reduction is a helpful tool in machine learning that maps high dimensional data to a lower dimension, reducing noise in the dataset and amplifying helpful features.

Explainability Being able to explain how our models work and what their decisions are based on is important, especially in a case like this based on crime complaint data. Understanding how features are being used and interpreted inside a model will allow us to catch biases that can affect our results.

Logistic Regression A relevant machine learning algorithm often used for classification. This will be one of the machine learning algorithms implemented to predict crime. It will be used as a multiclass classifier to predict which type of crime the given crime is most likely to be. Hyperparameter tuning, as learned about in class lectures, will be necessary to ensure the best performance.

Artificial Neural Network An ANN is a machine learning model that mimics that synapses in the brain by connecting layers of artificial neurons that learn and pass on information to the next layer until the output layer is reached. ANNs can be very helpful in finding hidden patterns and classifying multi-class data.

Decision Tree A decision tree algorithm is a type of supervised learning to find a target value using if-else conditions. Using a decision tree allows us to use only the most relevant features that will lead to more accurate predictions.

KNN We plan to use Nearest Neighbors as one of the algorithms to identify a type of crime using proximity to the location of the crime and similarity in other categorical features. We plan to fine tune hyperparameters such as number of neighbors and distance method calculations to identify the best predicting performance of the model.

Clustering We plan to implement clustering algorithm(s) to identify crime hotspots and find previously unsuspected feature relationships. We hope this will provide greater insights into data and guide us in the direction of bettering accuracy of our predictive algorithms.

SVM A support vector machine will be used to identify crime hotspots. We also want to explore the use of a multi-classifier SVM to test it against our other models predicting the type of crime.

5. Dataset description.

The first dataset we are using is a list of complaints made to the NYPD across the 5 boroughs and includes over 413k accounts. Complaints include specific information about location, time, and the response to the complaint. It also includes demographic features of the suspect and target involved in the crime. The dataset contains records for the year 2020. We will also be using a second dataset containing historical information from 2006 to 2019. We will be training our models on historical data from past years and testing our models on more recent data.

There are multiple challenging parts to our project that we will each work on solving. The dataset we are using is a real world dataset and thus faces the challenge of missing data. For instance, 'suspect age group' feature is unknown or missing in nearly 40% of observations. We may consider imputing the data with the most frequently occurring value for a particular type of crime. Similarly, description of location (LOC_OF_OCCUR_DESC) is missing value in 16% of observations. We may consider populating the values based on other location features or eliminate this feature altogether if it has high correlation with other location features.

Another challenge we must resolve before training our models is how to deal with bias in crime prediction. One concern relating to bias is that crime prediction models may compound systemic biases from humans already endemic to the criminal justice system. Some critics of predictive policing argue that any prediction based on historical data is in itself biased as it reflects the inequalities of the past on to future events (Mayson, 2019). The Los Angeles Police Department ran a controlled experiment over the course of over 1 year in which predictive algorithms placed some police patrols while others were placed by criminologists. Brantingham, Valasik, and Mohler analyzed the results of the experiment and found that there was no significant bias reflected in the proportion of arrests by racial-ethnic groups between the groups of police patrols placed by the algorithm and those placed by criminologists (Brantingham, 2018). We use this result as proof that this type of unbiased predictive analysis is possible.

6. Project timeline.

Week 1 (March 15 - 21) Prepare the proposal. Start Exploratory Data Analysis.
Week 2 (March 22 - 28) Evaluate proposal feedback, adjust project if needed. Continue EDA.
Week 3 (March 29 - April 4) - Spring Recess
Week 4 (April 5 - 11) - Develop competing algorithms.

Week 5 (April 12 - 18) - Develop competing algorithms.
Week 6 (April 19 - 25) - Develop competing algorithms
Week 7 (April 26 - May 2) - Finish algorithms. Evaluate results between algorithms. Compare results against existing published works.
Week 8 (May 3 -9) - Finalize conclusion. Prepare the presentation.
Week 9 (May 10 -16) May 15th Final presentation

7. Plan for the final demo.

We plan to show how each of our models work (KNN, Decision Tree, Logistic Regression) and compare the accuracy between them. We will display the performance metrics such as accuracy, precision, recall, F1 scores, confusion matrices, and ROC curves for each model. We will describe the positives and drawbacks of each model.

8. Plan to evaluate the project.

We will train and test our models on the same data so we will thus be able to compare our models and evaluate our results. We will determine which model is the best one based on metrics created for our demo and compare it to the baseline found in Chainey et al. in which different categories of crimes are predicted using London crime data. The results are calculated in terms of the Predictive Accuracy Index (PAI), an area-standardized measure of accuracy so we too will calculate the PAI scores of our model in addition to other metrics in order to compare our results.

Bibliography

Brantingham, P. Jeffrey, et al. "Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial." *Statistics and Public Policy*, vol. 5, no. 1, 2018, pp. 1-6. *Taylor & Francis*, https://doi.org/10.1080/2330443X.2018.1438940. Accessed 15 March 2021.

Chainey, S., Tompson, L. & Uhlig, S. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Secur J* 21, 4–28 (2008). https://doi.org/10.1057/palgrave.sj.8350066

Mayson, Sandra G. "Bias In, Bias Out." *The Yale Law Journal*, vol. 128, no. 8, 2019, pp. 2218-2300. *The Yale Law Journal*, https://www.yalelawjournal.org/article/bias-in-bias-out. Accessed 15 March 2021.