Pre - Analysis Plan

Team: Cool Team

Names: Aidan Tan, Cindy Dong, Herin Seo, Joseph Choe, Allie Kim, David Bae, Benjamin Yeh, Jonah Lee

Main Question:

- Given various patterns in crime types and premises from 2020 to present, could we predict which type of crimes will be committed given premise, time, age, race, etc?
    - Additional exploration will be done on time, age, race, etc. in regards to premise.

1. What is an observation in your study?
    a. Each observation is a specific record of a crime incident that occured in Los Angeles. This includes information such as the time the incident occurred, the time the incident was reported, the type of incident, information about the victim, the current status of the report, and other relevant information.
    b. “This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy”
    c. 190326475, 2020-03-01T00:00:00.000, 2020-03-01T00:00:00.000, 2130, 07, Wilshire, 0784, 1, 510, VEHICLE - STOLEN, , 0, M, O, 101, STREET, , , AA, Adult Arrest, 510, 998, , ,1900 S LONGWOOD AVE, , 34.0375, -118.3506
2. Are you doing supervised or unsupervised learning? Classification or regression?
    a. We will be doing supervised learning using classification because our observations are qualitative. Since we are trying to do more predictive analysis as well, we want to be able to say which crimes could happen given a specific premise and other variables.
3. What models or algorithms do you plan to use in your analysis? How?
    a. KNN algorithm will be used. We will find k training examples closest to a given input and then predict the class based on the majority class of these neighbors. By plotting data based on crime type and premise, we can tell which crime will most likely be committed.
4. How will you know if your approach "works"? What does success mean?
    a. If the model accurately predicts at least X% of crime type in the data, where X is a threshold based on a specified requirement (e.g., 80% accuracy for a premise), we consider it a success.

5. What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?
    a. One weakness is that crime and criminal psychology is a much more complicated issue than just one or two variables like premise or time. The culture, demographic, socio-economic status, and more play into these factors. So the way we plan to address this is by cross tabulating a number of the other columns to ensure our model is not overly simple but able to be resilient in the face of other variables.
    b. Another weakness is that the given dataset is very large, considering it is an ongoing dataset that gets updated bimonthly. There are some columns/variables that have majority missing values which we will have to account for. We will deal with this issue at the beginning, before developing any algorithm or models when cleaning the dataset. If this approach fails, then the data was not properly cleaned and we must go back to the beginning and clean through using other methods.
    c. If our approach fails, it would show that crime prediction isn't as simple as what records show. Perhaps additional data sets would need to be analyzed alongside this one regarding different factors. It would also bring other questions to mind such as is it even possible to predict crime and how can we be sure we're not generalizing or unfairly stereotyping certain areas or peoples?

You should address the following topics in the text, as appropriate:

- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?
    - First, we will clean the dataset, handling missing values and performing data type conversions. Specifically, possibly filling missing numeric variables with the mean or median, and categorical with the mode or something like "Unknown".
    - Some variables that could be one-hot encoded would be the crime type, area, premis_desc, victim sex. Many of these variables are already encoded for us through the "cd" columns which are the corresponding codes for the values. However, we could encode victim sex as that is the one variable which is not done.
    - The features we plan to use are : time (time_occ), area (area), premise (premis_desc), victim's sex (vict_sex), victim's age (vict_age)
- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.

- One way we would communicate our results is by presenting our predicted versus actual crime types as a confusion matrix. Selecting certain observations with a specific set of variables e.g. 12:00pm, Hollywood, and victim age with what crime actually happened and what our model predicted.
- Given a specific premise, time, and area, we can find the most common crime types. This can be served as the "correct" output that the prediction will be compared to.