Checkpoint 4 Findings

Question 1

We sought to answer the question, “How many home invasion allegations can we expect in 2021 if there is no policy change?” and predicted a value of 18.

Observations
From our initial graph of the data, we can see a fairly right-skewed distribution of the data. Visually, there doesn’t seem to be a linear trend in the number of home invasion allegations over time. Instead, there appears to be a large peak around the years of 2000 to 2005. After that, there is a steady decrease in the number of these types of allegations. Thus, our predicted value of 18 home invasion allegations for the year of 2021 does seem logical. This trend is optimistic, assuming there is no external reason for this drop, which may very well be the case. The reason for the sharp drop followed by a steady decrease begs the question of if there was a policy change, like a law change or additional required training, that could have influenced these outcomes. To our group’s knowledge, there was no such change, however, we will continue to investigate.

Disadvantages of this approach
Though we have 2840 total home invasion allegations, we had to aggregate this data by year in order to create our model as a result of the question being asked. Though there is much variation being captured within each year’s value, this reduced our dataset to a length of 22, as that is the number of years we have data for. Likely for this reason, we have a high MSE value of 13266.64.

Another disadvantage is that we likely have outliers (using domain knowledge to determine this), but two different statistical approaches did not result in the identification of any outliers, therefore we did not exclude them for this analysis. Technically speaking, they are not outliers, however, the fact that both 1997 and 1998 show one singular home invasion allegation during that year suggests that there might be something else going on. Perhaps less data was collected in general during those years, or the labeling of the location type of an allegation might not have been implemented consistently until later years. The same logic applies to why there might be only four such allegations for the year 2018. Perhaps the database at large doesn’t yet contain many allegations for the year 2018 because of a lag in data transfer.

Question 2

Coming soon!

Question 3

We sought to answer the question "Can we predict which officers will be implicated in Home Invasion complaints?"

Procedure

The relevant information here was obtained using the SQL queries present in c4_q3.sql and includes information about officers who either were or were not involved in at least one home invasion complaint.

From there, this data was combined into a single Pandas dataframe with a column for home_invasion, where a 0 signified no participation in a home invasion complaint and a 1 signified participation in such a complaint.

In order to use the data in a machine learning model, some of the categorical data needed to be converted into numerical data. Specifically, officer race, gender, and whether or not they are active had distinct non-numerical labels. Because there is no logical progression of these things, (Black is not < or > Hispanic), we used One Hot Encoding for these categories. The other dependent variables included were birthyear, honorable mention count, trr count, sustained count, civilian compliment count, and discipline count. These variables were already numerical and did not seem to significantly overlap conceptually.

From there, this data was scaled and fit to a LinearSVC model with default settings. The accuracy and confusion matrix of the model are outputted.

Other attempts

You can view other work in the jupyter notebook "Question 3 - Predicting Officers in Home Invasion Complaints". We tried several models, including GaussianNB, KNeighborsClassifier, MultinomialNB, BernoulliNB, LogisticRegression, SGDClassifier, SVC, and NuSVC. The only one that performed as well as LinearSVC was the Stochastic Gradient Descent classifier, but nearly all models performed similarly with the exception of K-Nearest Neighbors which underperformed.

We also used GridSearchCV to search for the optimal parameters. It turned out that the default was actually the most effective, so no changes were made on that front.

Results

This model does better than chance with predicting which officers are involved in Home Invasion complaints, but not much better. This model (and indeed all models tested) does poorly with accurately identifying officers who were involved in Home Invasions, with more false positives than true positives. The model's accuracy is .608.