Individual Report
Introduction to Data Mining
Sam SR

Introduction

The main goal of the project is to clasify crimes based on time, location, victim's sex, and age group in addition to some correlation and patter analysis. We experimented with different kinds of models in order to find the best results, which ended up being a set of decision tree classifiers which predicted the possibility of a given observation of being a felony, misdemeanor, or violation-type crime. Our results were not as accurate as we would like, but through careful feature selection and tweaking available parameters in the models, we were able to get some results. [1]

I organized the Github repo, developed the decision tree model which was ultimately used in our app, prepared the presentation, and compiled the final report. Omar did exploratory modeling and prepared the app interface. Sertan did exploratory modeling and helped prepare the final report. Our division of work was such that we all worked on different modeling algorithms in order to find what was most effective.

My portion of the work:

My biggest contribution to the project were the decision tree models we used to predict crime classification. This involved cleaning the data by encoding continuous numerical variables into scikit-learn-readable numbers, encoding ordinal categorical variables similarly, and encoding purely categorical variables with a one-hot-encoding scheme, where each possible value for an overarching feature is given it's own feature, is "feature" with a 0 for false and a 1 for true. For numerical and ordinal variables, null values were replaced with the average, for categorical variables, nulls values were replaced with a random value based on the distribution of all possible values for that feature. Some nulls were removed entirely, if there were few enough for that particular feature (order of magnitude 100 observations out of over 300,000 in the original dataset). All of the encoding was written from scratch, in order to have the most control over how it was implemented.

Results

[1] Overview borrowed from full report

In order to measure the success of our models, we took the accuracy, F1 score, and confusion matrix of each model, after performing a train-test-split. Where our accuracies were strong, our F1 scores were rather low. This means our dataset is not well balanced and our model was underfitting, we were unable to resolve in the time we had. This is reflected in our confusion matrices.

Target	**Violation**	**Felony**	**Misdemeanor**
Accuracy	84%	70%	58%
F1 Score	11%	3%	61%
Confusion Matrix	$\begin{array}{cc} 28134 & 14 \\ 5243 & 7 \end{array}$	$\begin{array}{cc} 22903 & 561 \\ 9343 & 591 \end{array}$	$\begin{array}{cc} 8497 & 6832 \\ 7034 & 11035 \end{array}$

Figure 1: Metrics of success, Accuracy, F1 Score, Confusion Matrix for the Decision Tree Models

This ended up being our most successful set of models, which is surprising given the binary nature of decision trees implemented in scikit-learn. I was careful to make sure that my preprocessing of the data did not introduce any unintended order to the variables, keeping purely categorical variables separated. I tweaked parameters like max leaf nodes and the train-test-split in addition to the specific features in order to maximize the above metrics. As I initially predicted during the exploratory data analysis portion of this project, features specifying what kind of establishment or where on the premises the reported crime took place served to improve the accuracy. The dataset was limited though, in that there were a significant amount of null values that we had to fill.

Summary and conclusions

The lack of balance in our dataset proved to be a challenge for the modeling process. My results were meaningful, and my models were able to ultimately address our initial research question: Can we predict the type of crime reported based on other features in the report? The feature selection may not have been perfect, and if we had more time I would have attempted to make the data more balanced so there was not so much underfitting. Additionally, we could have expanded our models to include data from more years of crime reports, which for the sake of time we did not explore. The accuracy we achieved, particularly on the model predicting if a report would be a violation, is a valuable result.

Code

All of my code was written by me. Some of it was inspired by lecture code from this class, but nothing was copied directly from the internet. I was unable to quickly get existing code for algorithms like one-hot-encoding to work, so I wrote one myself.