Predicting Resolved and Unresolved Criminal Cases Using Logistic Regression

Kelly Li, Raam Pravin, Xavier Lee, Zoe Lu

Abstract

Los Angeles is a major city in the United States with crimes occurring frequently, ranging from small crimes such as theft to more serious crimes such as murder. Naturally, as a city with a lot of crimes, the Los Angeles Police Department (LAPD) often needs to evaluate how they handle certain cases and how much resources to put into select crimes. This project focuses on trying to predict whether crimes at the time of data reporting for the LAPD's tracking system were marked as resolved or not using different machine learning methods as there is belief that there are underlying reasons for certain crimes being resolved sooner than others. Further research beyond this project hopes to use the best identified predicting model to look more in depth into the patterns of cases left unresolved to better aid the LAPD in deciding how they handle criminal cases.

The dataset for this project comes from the Data.Gov website, a government owned website, which contains all crimes reported by the LAPD from November 2020 to the day of download (November 26th, 2024) as data is constantly being added everyday. Methods of data cleaning and exploratory data analysis were conducted to understand the data better while parallelization was used to speed up the analysis time. To predict whether each case will be classified as resolved or unresolved, models such as logistic regression, neural network, and naive bayes were performed on multiple subsets of test and training datasets. To compare the performance of each model, the mean accuracy of each model was calculated using confusion matrices.

The results of this study found that the model that best predicts whether each crime at reporting was resolved or not was the neural network model with a mean accuracy rate of around 75%. The next best model was the predictive logistic regression model which had a mean accuracy rate of about 76%. The performance of the neural network and logistic regression model were very similar. The worst performing model was naive bayes with only about a 68% accuracy rate.

In conclusion, using a logistic regression model or a neural network model best predicted whether a case was resolved or not at the time of reporting by the LAPD. In conjunction with additional research, this project can serve as a starting point into looking at whether new crimes that occur would end up being left open by the time it's recorded into the LAPD data system. This project has many limitations that should be taken into account such as a lack of extensive knowledge in machine learning methods to conduct a full analysis and run more accurate models. Additional models can be run to potentially identify other models that have better predictive performance and further research into other important predictive features would be beneficial.

Introduction

Methods

test test 2

Results

Conclusions and Limitations

The results of this project show that the model that best predicts whether a crime at the time of LAPD's data reporting has been resolved or not was the neural network model with an average accuracy of 76% across all subsets of data split into train and test sets. Similarly, the logistic regression model yielded similar results with a mean accuracy of 75%. The worst performing model was the naive bayes model which only had about a 68% accuracy rate. Using these results, this study hopes that it can help the LAPD predict whether future criminal cases that occur will end up being resolved by the time they report the data into their tracking system. Perhaps cases that are not resolved by the LAPD by the time of their tracking system reporting need more attention and resources or are just being neglected due to the crime's severity level being low. Further research can also look into analyzing the patterns of cases that are not resolved at the time of reporting and identify similarities between these cases to better understand what types of crimes get resolved quicker than others.

This project has limitations and thus the results should not be generalized. The lack of extensive knowledge on all possible types of machine learning models means there are additional models left unexplored that could potentially be better at predicting whether cases are resolved or not. Additionally, there are better methods that can be formed to more accurately give insights and advice to the LAPD and how they choose to handle cases. For example, models predicting the length of time/ duration categories each case takes to get resolved can be explored and would also be more informative to the LAPD. More knowledge on how to implement these methods would be beneficial. Additionally, some computational challenges such as the handling of missing cases can be done better with more knowledge on missing data and efficiency of code can be further optimized.