# Predicting Resolved and Unresolved Criminal Cases Using Logistic Regression

Kelly Li, Raam Pravin, Xavier Lee, Zoe Lu

**Project Repository Link:** https://github.com/625group/625finalproject/tree/main"

## 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(Cleaning/Parallelizing)

**Data Description:**
The raw data included 2,093,455 observations of 28 variables. The predictors of interest included Area (representing a geographic area in Los Angeles referring to any 21 community police station locations), Part 1 or 2 (referring to whether LAPD reported the crime to the FBI or not), Crime Code (a code referring to the crime that was committed), Victim Age (age of the victim of crime at time of crime), Victim Sex (sex of victim of crime), Premis (encoded values representing the mode of structure, vehicle, or location where crime took place), Weapon Used(encoded values representing what type of weapon was used in commission of crime), Victim Descent Description(Race/Ethnicity of victim of crime), Time Occurred(time of occurrence of crime in 24 hour military time), Date Occurred(date of occurrence of crime in MM/DD/YYYY format), and Date Reported(date of report of crime in MM/DD/YYYY format). The outcome variable was Status(referring to the status of the case).

**Data Re-coding:**
***Predictors***
Based on the variables of Date Occurred and Date Reported a new variable was encoded named date_occur_report_difference which was a continuous numeric value representing the number of days of difference there was between Date Occurred and Date Reported. A new variable named time_occurr_cat was encoded which represented the time the crime occurred in military time in one of four categories: Morning, Afternoon, Evening, and Night.

***Outcome***
Basic EDA revealed that in there was a small number of cases whose statuses were recorded as UNK(unclear) these values were set to null. A new binary outcome variable named Legal_Action which represented whether a case was resolved or not was coded based on the values of the variable Status; if the value of this column was Adult Arrest, Adult Other, Juv Arrest, or Juv Other then Legal_Action was set to 1. Otherwise, if the value of Status was Investigation Continuing then Legal_Action was set to 0 for all corresponding rows.

***Data Subsetting***
EDA revealed that a vast majority of the criminal incidents committed were one of fifty of the most common crimes, took place in one of the fifty most common premises, and occurred with the use of the ten most common weapons used in crime(with the largest category being no weapon used in crime). Criminal incidents that were included in the rest of the analysis were classified as one of the fifty most common crimes, occurred in one of the fifty most common premises for committing a crime, and occurred with the use of one of the ten most commonly used weapons.

***Data Cleaning***
EDA revealed that there were ages as low as 0 and as high as 120 for Victim Age, subsequently rows with Victim Ages less than or equal to 0 or higher than 100 were included. EDA also revealed that a very small minority of Victim Sex values were something other than Male or Female, these rows were also not included in the analysis. The final clean and re-coded dataset included 1,528,645 rows of 11 columns. A train and test data set were made from a 80/20 split from this final dataset.

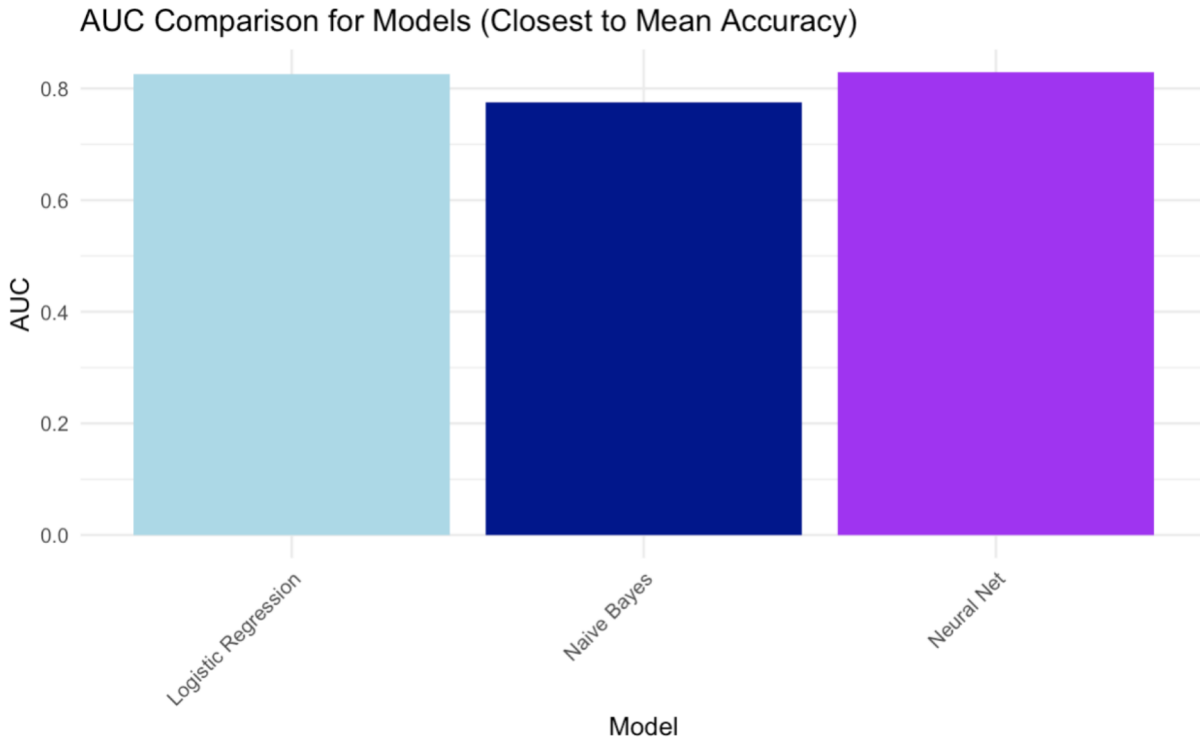
***Computation:***
Since the dataset was very big and the goal of the analysis was to compare the performance of multiple algorithms(Logistic Regression, Naive Bayes, Neural Net) in predicting a binary outcome of whether a case had been resolved or not; the train and test data sets were equally partitioned into 31 smaller datasets. EDA revealed a concerning class imbalance in the outcome variable as the vast majority of cases had not been resolved so all 31 train datasets were bootstrapped until there was about a 50/50 frequency in the binary outcome variable. Each of the 31 resulting datasets had around 40,000 rows. Since the time needed to apply each of our classification algorithms would have been exceedingly long, parallel processing was used in order to speed up computation time. To evaluate the performance of each of the three algorithms the following steps were taken: 1. A gridsearch was applied using the "extra" 31st dataset in order to obtain the

optimal parameters for each algorithm 2. Using the optimal parameters each algorithm was created using its corresponding train dataset and then tested using its external corresponding test dataset. Evaluation metrics such as accuracy, AUC(area under curve), and confusion matrices were stored in a list for each model. 3. Using the vector of 30 accuracies from the 30 models created a 95% confidence interval was created to compare performance of all three models. The confusion matrix of the model with the closest accuracy to the mean accuracy for each algorithm was also returned.

## Results

To determine the best model we assessed the accuracy, sensitivity, specificity, AUC, and average run time of the models that were closest to the mean accuracy rates. Our logistic regression model had the fastest average run time of 4 minutes. It has an accuracy of 75.23%, sensitivity of 74.47%, specificity of 77.59%, and an AUC of 0.8264. The neural network was the slowest model with an average run time of 16 minutes. It has an accuracy of 75.89%, sensitivity of 90.24%, specificity of 51.40%, and an AUC of 0.8285. Lastly, our naive bayes model had an average run time of 8 minutes. It has an accuracy of 70.26%, sensitivity of 67.18%, specificity of 79.64%, and an AUC of 0.7751.

It is clear from our results that our naive bayes is the weakest model of the three; it is neither the fastest model, nor excels at any metric we were assessing besides specificity. The similar AUC scores between our logistic regression model and neural network suggests that they are comparable in terms of classification ability. If classification is the bottom line, then logistic regression would prove to be the better model due to a significantly lower average run time. However, the two models drastically differ in their sensitivity and specificity rates as mentioned above. If correctly predicting a suspect will get caught after committing a crime has more value than correctly predicting a suspect won't get caught, then the neural network could be the best option.



## 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.