Title

Machine Learning Analysis of Case Clearance in Los Angeles Violent Crimes, 2020 to 2024

ISBN

9798252407210

Author

Hong, Shu

Publication Date

2026-06-11

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Machine Learning Analysis of Case Clearance in Los Angeles Violent Crimes, 2020 to 2024

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Applied Statistics and Data Science

by

Shu Hong

2026

ABSTRACT OF THE THESIS

Machine Learning Analysis of Case Clearance
in Los Angeles Violent Crimes, 2020 to 2024

by

Shu Hong

Master of Science in Applied Statistics and Data Science
University of California, Los Angeles, 2026
Professor Yingnian Wu, Chair

This thesis explores whether the information contained in initial police reports can help predict whether a crime will be solved. I employed two different definitions of "case clearance": a broad definition, under which any form of case closure is considered a clearance; and a strict definition, which counts a case as cleared only if an arrest is made. I ran three predictive models under each of these two definitions, using data on 20,553 violent crimes recorded by the LAPD between 2020 and 2024. Among three, Gradient Boosting method achieves the best calibration and lowest Brier score. A single behavioral indicator, showing whether the victim knew the suspect, accounted for roughly a third of total feature importance. Interestingly, victim descent contributed close to zero and observed group level disparities tracked the distribution of crime types rather than any learned ethnic pattern.

The thesis of Shu Hong is approved.

Frederic Schoenberg

Vivian Lew

Yingnian Wu, Committee Chair

University of California, Los Angeles

2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Background

One of the most intuitive indicators of how well the city police force is performing is the way crimes are handled. When a crime occurs, with a report to the authorities (a 911 call), the police first open a case and initiate an investigation; if leads emerge, they work to identify and locate a suspect, and they hand the case over to the prosecution for further action. The proportion of reported cases that are ultimately marked as solved, commonly referred to as the clearance rate (Federal Bureau of Investigation, 2013). It is currently one of the most widely utilized statistical metrics within the American policing system.

Statistically speaking, the clearance rate lumps together two entirely distinct scenarios: One is when the suspect is taken into custody, while the other is when the police know who it did but can't make an arrest for various reasons. But both end up being counted as "closed cases." where they're treated as the same.

Interestingly, however, these two scenarios differ significantly in reality. The proportion of such "special clearances" varies naturally across different types of cases; for instance, in domestic violence cases, it is more common for victims to be uncooperative, whereas victims of stranger-on-stranger robberies tend to be more cooperative with the police. Consequently, even if two police departments expend roughly the same amount of effort, their final case clearance rates may appear vastly different. Therefore, using the case clearance rate as a metric to evaluate police effectiveness is somewhat problematic.

This asymmetry is not evenly distributed across different types of crimes. For example, homicide cases usually rely on an actual arrest to be considered clear. In contrast, in intimate partner violence cases, a larger share of cases is cleared through exceptional means. Sexual offenses show a different pattern. No matter how clearance is defined, their clearance rates remain low, and the rate becomes even lower when only actual arrests are counted. When these vastly different scenarios are aggregated and summarized by a single "clearance rate," it becomes difficult to accurately reflect how cases are handled. Consequently, the more critical question is this: what factors truly determine whether a case gets solved? I utilized public data released by the LAPD spanning from 2020 to 2024; after filtering, I retained only violent crimes committed against individuals with totaling 20,553 incidents. Each record contains details such as the time and location of the incident, crime type, victim information, weapon used, premises type, and the modus operandi (MO) as recorded by the police. The ultimate resolution status of each case, whether it was closed or not, will serve as the target variable I sought to predict.

There are three primary tasks. First, I specifically processed the "Modus Operandi codes" (MO codes), transforming them into features directly used for modeling, a step that has often been overlooked in many previous studies. Second, rather than relying on a single definition of "case clearance," I employed two simultaneous definitions: one broad and one strict. This approach allowed me to compare cases resolved through arrest versus those closed via other means. Third, I conducted an analysis of group level fairness to determine whether observed disparities between different demographic groups were artifacts of the predictive model itself or inherent characteristics of the underlying data.

In the following section, I will begin by introducing the dataset itself, followed by an explanation of the data cleaning procedures I implemented. Then the problem is framed as a machine learning task and the models are introduced. After that, the results are presented, including model performance, calibration, feature importance, and fairness

analysis. The final section discusses what these findings suggest about how clearance rates are currently used.

1.2 Research Questions

In this thesis, I focus on several interrelated questions.

The first question is: how can I determine whether a case will ultimately be solved, relying solely on information available now that the case is initially reported? I compare results from different definitions of what constitutes a "solved" case; utilizing the same set of features, I construct separate models to examine whether predictive performance varies depending on the specific definition employed.

Once these models are completed, I check on which information the models rely. If predictions are driven by a select feature, these features may reflect key underlying mechanisms in the process of solving criminal cases. The results presented in Chapter 4 indicate that a significant factor is whether the suspect and the victim know one another, a finding that remains consistent across the various models.

Furthermore, I value problems regarding the fairness of these models. Since the models are trained on historical data, which inherently contains pre-existing patterns, they run the risk of perpetuating these disparities. I compare the models' performance across different ethnic groups of victims. Once performance gaps appear, it becomes necessary to determine whether these discrepancies stem from the model itself, from relevant patterns, or from pre-existing differences in baseline clearance rates among the various groups.

CHAPTER 2

Data

2.1 Source and Scope

The data I used in this thesis are from the Los Angeles Police Department's (LAPD) public crime data portal(Los Angeles Police Department, 2025). The dataset is about reported incidents in Los Angeles with occurrence dates ranging from January 1, 2020, to December 26, 2024. After my data cleaning, there are 20,553 rows in the final dataset, with each row corresponding to a single reported case. In addition to a unique LAPD case number, each row contains information regarding the time of the incident, the reporting date, the LAPD geographic area and reporting district, a crime code and its textual description, a modus operandi code, the victim's age, gender, and ethnicity, the location code and description, the weapon code and description, and finally the case disposition status.

Upon examining the dataset, I found that violent crimes predominate: aggravated assault accounts for 33.0%; simple assault and battery for 16.7%; robbery for 11.0%; intimate partner violence for 12.7%; and sexual offenses for approximately 4%. Interestingly, crimes such as theft or fraud are almost absent from the dataset. Therefore, I treat this data as a subset of violent crimes rather than as a complete, representative sample.

The disposition status variable comprises five distinct values. "Adult Arrest" and "Juvenile Arrest" indicate that an actual arrest was made. "Adult Other" and "Juvenile Other" denote a special case clearance. The "Investigation" status means that investigative work is still ongoing and no outcome has yet been reached. The first four categories

generally align with the FBI's broad definition of a "closed" case. However, only the first two categories satisfy the strict definition of "arrest." Therefore, my modeling analysis uses these two distinct definitions, and the resulting outcomes will be different as well.

2.2 Assumptions and Limitations

During the EDA stage, I treated the case status within the dataset as the predictive target. The status of a case may change if new evidence emerges. Since the LAPD data does not include timestamps for updates, I was unable to directly track these changes. However, this had a negligible impact on the data, as the records I utilized were already three years old at the time of analysis.

Another problem was that the data for 2024 remained incomplete: although the volume of cases remained stable from 2020 through 2023, the data volume for 2024 was small, amounting to roughly only one-third of that seen in previous years. Therefore, I restricted my analysis to data spanning the period from 2020 to 2023.

The data also required a certain degree of cleaning. For example, a victim's age was recorded as "0," which signified that the age was unknown. I treated these entries as missing values and imputed them using the median. Additionally, the gender field was sometimes left blank or contained anomalous values; therefore, I uniformly categorized these instances as "Unknown."

However, the MO codes are not always the same. Whether they are recorded, and which ones are included, often depends on the officer. This variable therefore needs to be used with care in the analysis. The specific codes I later use appear at rates that are stable across the five study years and across LAPD areas, and I read that stability as partial evidence of reasonably consistent use. No stronger claim is warranted: some incidents that genuinely involved, say, a victim who knew the suspect may simply be missing the corresponding code because the officer did not type it in.

Moreover, the data covers only incidents that were reported at all. Crimes that

victims chose not to report are invisible. Nonreporting is known to be substantial for violent crime in general, and especially for sexual offenses. Inferences from this data apply to reported incidents, not to the underlying population of violent crime in Los Angeles.

2.3 Exploratory Data Analysis

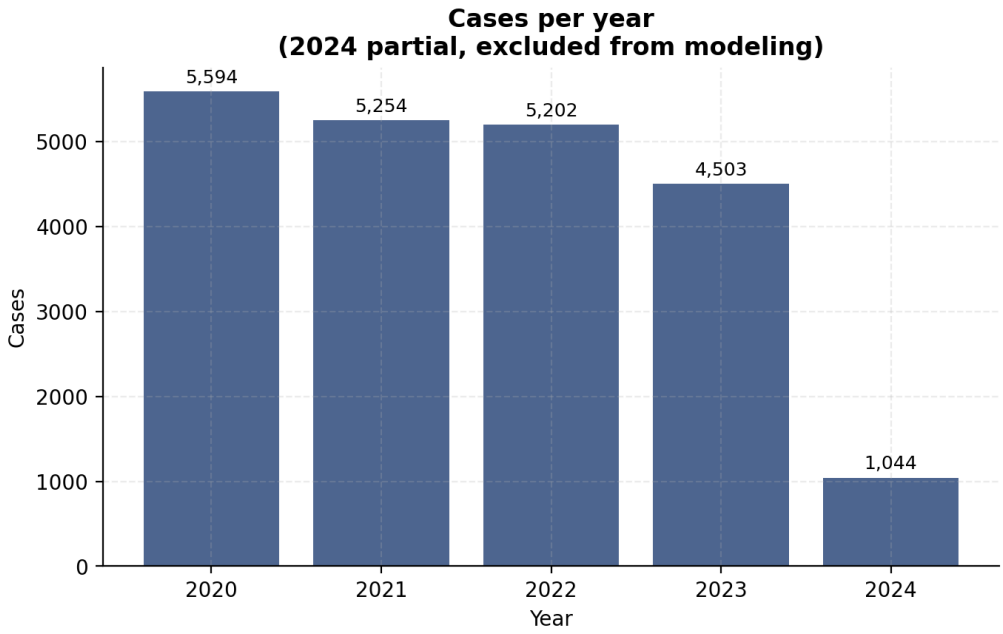

Figure 2.1: Annual case counts from 2020 through 2024.

Figure 2.1 shows cases per year in Los Angeles. From 2020 to 2022, the number of cases stays stable at around 5,300 per year. In 2023, it drops to 4,503. Then the sharp drop in 2024, down to 1,044 cases, is due to incomplete data, since records only cover the first quarter.

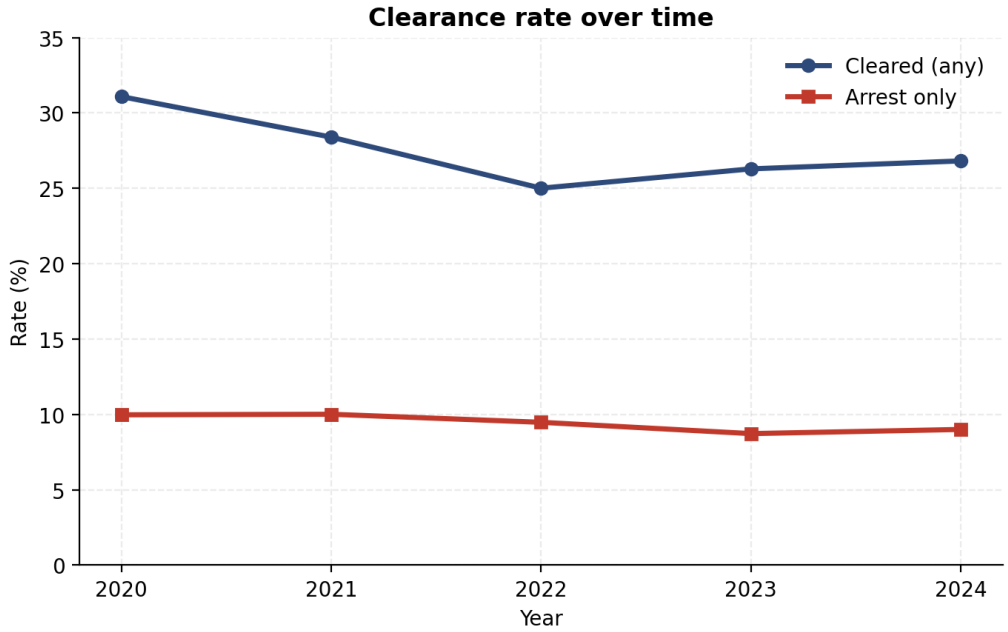

Figure 2.2: Clearance rate from 2020 through 2024.

Figure 2.2 shows that the clearance rate declines over time from 2020 to 2022, then slightly increases at the year of 2023. The arrest rate decreases as well, but more slowly, from 10.0% to around 8.7%. This shows that most of the change comes from a reduction in exceptional clearances.

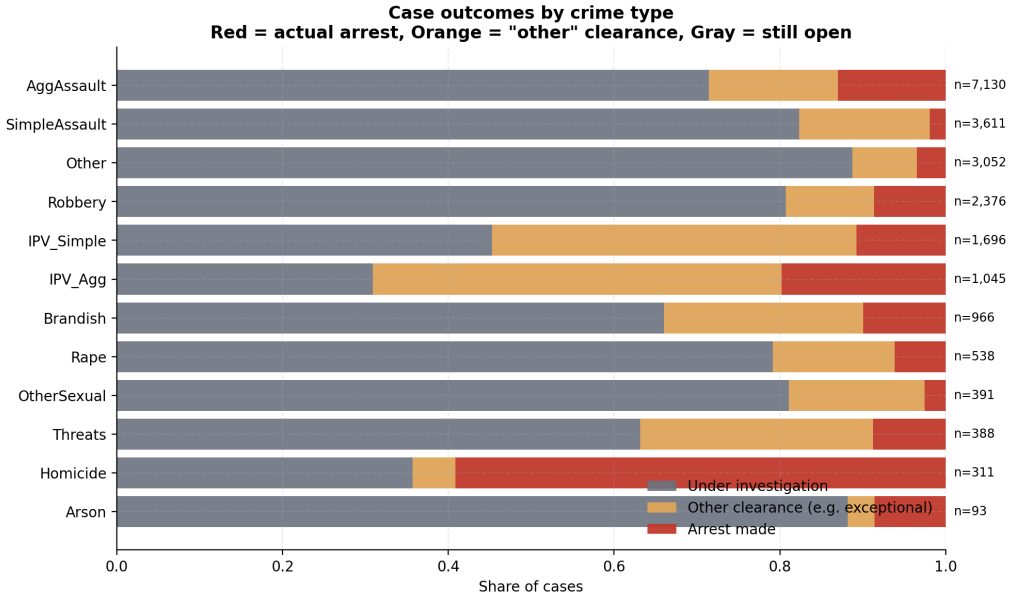

Figure 2.3: Case outcomes by crime type.

Figure 2.3 is the case that outcomes breakdown by crime types. For aggravated assault, simple assault, and robbery, arrests and exceptional clearances show up at similar rates, both around 10% to 15%. These two ways of clearing cases appear side by side within these categories. Intimate partner violence breaks the pattern in the opposite direction, with exceptional clearances at roughly two thirds of all dispositions, arrests below 20%, and the combined clearance rate pushing 70%.

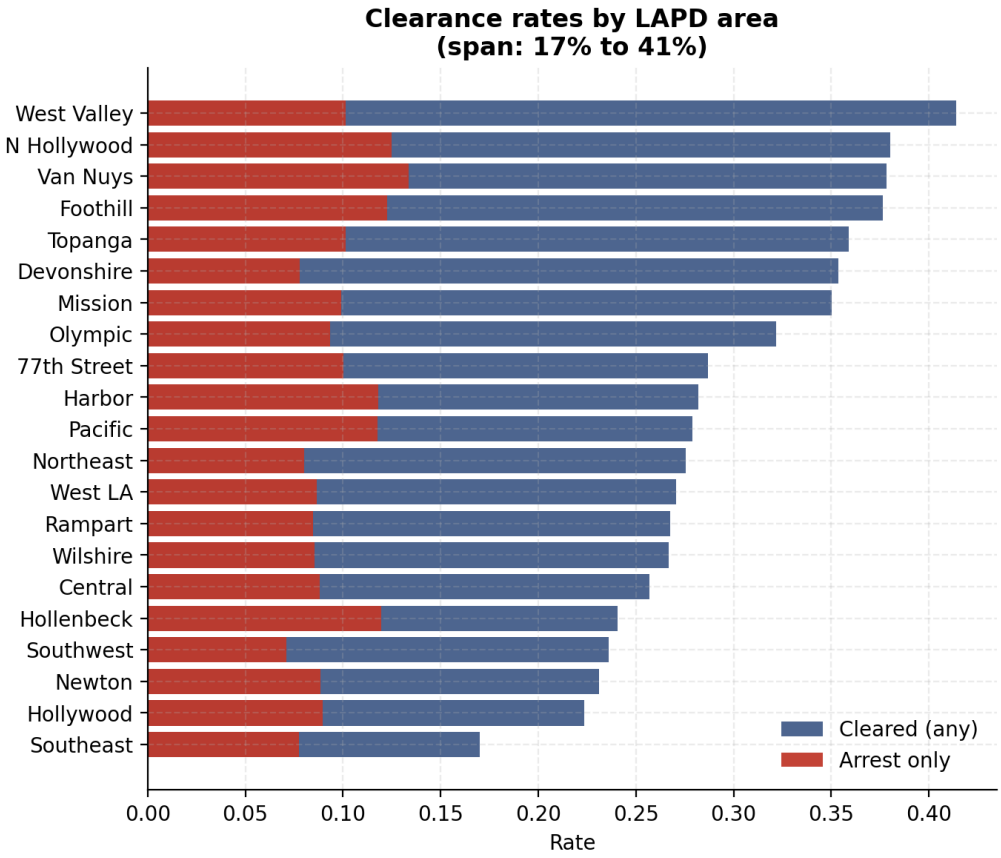

Figure 2.4: Clearance rate by LAPD area.

Figure 2.4 shows the two main dimensions of clearance heterogeneity. Arrests, measured alone, vary over a narrower band of 5.5 to 14.2 %. Central Division, which covers downtown Los Angeles, contributes the largest single case count at 5,889 (27.3 % of the corpus) and lands in the middle of the distribution at 25.7 %.

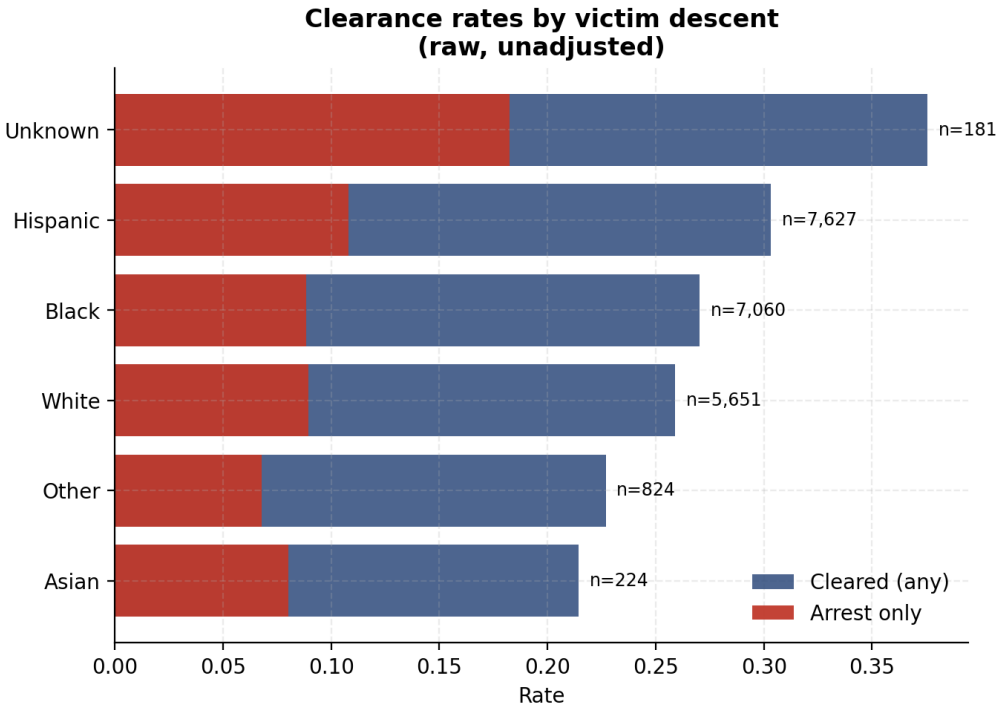

Figure 2.5: Clearance rate by victim race.

Figure 2.5 shows the victims' race. Hispanic victims, at 35.3% of the corpus, show a broad clearance rate of 29.0%; Black victims, at 32.7%, show 24.8%; White victims, at 26.2%, show 25.0%. None of these unadjusted figures constitute evidence of differential police effort. Raw rates reflect the mix of crime types each group experiences and the LAPD areas in which those crimes are concentrated.

CHAPTER 3

Methods

3.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ denote the modeling corpus, where $n = 20{,}553$ after the 2024 exclusion. Each $\mathbf{x}_i$ contains information available when the incident was reported.

Two outcome variables were created from the disposition status. For ycleared, cases with Adult Arrest, Juvenile Arrest, Adult Other, or Juvenile Other were coded as 1, while Invest Cont was coded as 0. For yarrested, only Adult Arrest and Juvenile Arrest were coded as 1; all other outcomes were coded as 0. The disposition status field was used only to create the labels. It was not included as a predictor, and neither was its text description.

I split the Incidents from 2020–2022 as the training set ((n=16,050)), and incidents from 2023 as the testing ((n=4,503)). This was done intentionally instead of using a random split. Clearance patterns can change from year to year, and a random split could allow the model to pick up information from future cases. Using an earlier period for training and a later period for testing gives a better sense of how the model would perform on new data.

3.2 Feature Construction

I construct features in three groups: categorical, numerical, and binary MO indicators. The first group is categorical features. Among them: LAPD area name (21 levels); crime category (12 grouped levels, obtained by consolidating raw crime code descrip-

tions); premises description (21 levels); weapon group (9 levels: firearm, edged, blunt, bodily force, verbal, chemical, vehicle, none recorded, other); victim sex (4 levels); victim descent group (7 levels); and a 4-level time-of-day bucket spanning morning, afternoon, evening, and overnight.

Numerical features form the second group. These include hour of occurrence, day of week, month, a weekend indicator, victim age, report delay in days (truncated at 365 to tame the long right tail), the number of modus operandi codes on the incident, and incident latitude and longitude.

Last come the six binary indicators built from the modus operandi field. Each is the output of a search for a specific four-digit code in an incident's code list.

Table 3.1: Modus operandi features used in the model.

Feature	**Code**	**Meaning**
mo_victim_known	1402	Victim knew the suspect prior to the incident
mo_multiple_susp	2004	Incident involved two or more suspects
mo_alcohol_drugs	1307	Suspect used alcohol or drugs
mo_vict_drugs	0913	Victim used alcohol or drugs
mo_firearm_used	0400	A firearm was used
mo_strongarm	0416	Bodily force was used, without a weapon

Categorical features are one-hot encoded with no baseline category dropped, producing 93 columns after encoding. For logistic regression, numerical features are standardized (scaled by their standard deviation). The tree-based models use the raw numerical values, since they are invariant to monotone transformations anyway.

3.3 Logistic Regression

Logistic regression is a supervised machine learning algorithm in data science. It is a type of classification algorithm that predicts a discrete or categorical outcome.

$$P(y = 1 \mid \mathbf{x}) = \sigma(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}. \tag{3.1}$$

Estimation proceeds by maximizing the ℓ_2-regularized log-likelihood:

$$\mathcal{L}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[y_i \log \sigma(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) + (1 - y_i) \log \Big(1 - \sigma(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \Big) \Big] - \frac{1}{2C} \, \|\boldsymbol{\beta}\|_2^2. \tag{3.2}$$

I fix the regularization constant at $C = 1.0$. Sample weights inversely proportional to class frequency enter the loss, correcting for the class imbalance that shows up in both outcomes.

3.4 Random Forest

ARandom Forest is a machine learning algorithm that uses many decision trees to make better predictions.

$$\hat{P}(y = 1 \mid \mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_t(y = 1 \mid \mathbf{x}). \tag{3.3}$$

$T = 400$ trees are fit. At each split, the tree considers $\lceil \sqrt{d} \, \rceil$ features, which works out to 10 features from the full set of 93. Trees grow to saturation, subject only to a minimum leaf size of 5, and class-balanced sample weights apply during fitting.

3.5 Gradient Boosting

Gradient boosting builds an additive ensemble of decision trees stage by stage, originally proposed by Friedman (2001). At iteration m, the prediction $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta\, h_m(\mathbf{x})$, where h_m is a tree fit to the negative gradient of the logistic loss at F_{m-1}, and η is the learning rate. After M iterations, the predicted probability is $\sigma(F_M(\mathbf{x}))$.

The model uses $M = 300$ trees, a maximum depth of 4, and a learning rate of $\eta = 0.05$. At each iteration, 85% of the data is randomly subsampled for variance reduction. These hyperparameters were fixed in advance rather than tuned on the test set, since doing so would lead to overly optimistic AUC estimates.

3.6 Dual Target Evaluation

Most quantitative studies concerning case clearance typically aggregate the two clearance methods defined by the FBI into a single, simple binary outcome. However, this approach effectively smooths over the inherent differences that exist between various types of crimes. I used a different strategy: for each model, I conducted separate evaluations under two definitions of "case closure," utilizing the exact same set of features and an identical split between the training and testing datasets.

This approach also shows some advantages. First, it compares how a model performs under two different objectives, which provides clearer insight into which specific "closure mechanism" is generating the predictive signals. Then, it shows an analysis of feature importance across both scenarios. If certain features exhibit different behaviors under the two definitions, it suggests that they play distinct roles within the different mechanisms.

The strict outcome is also closer to what most people would consider a successful enforcement outcome. Because of that, models trained on this definition are likely to be more useful in practice and more relevant for policy discussions. Model performance is usually evaluated using different metrics. Accuracy is the share of cases classified

correctly when a threshold of 0.5 is used. Since positive outcomes are rare in our data, we should use average precision to better reflect the model's performance. To provide context, accuracy is reported together with the majority-class baseline. AUC is calculated by the area under curve, evaluating how well the model can distinguish between different outcomes. Average Precision (AP) is calculated from the precision-recall curve and tends to be more sensitive to class imbalance. Because positive cases are rare, I believe AP is a more useful indicator of model performance in our settings

The Brier score,

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2, \tag{3.4}$$

captures overall probabilistic accuracy, combining both calibration and discrimination. Lower values mean predicted probabilities track observed frequencies more closely. Calibration also gets a visual check, via reliability diagrams. Probabilities are sorted into ten quantile bins, and the mean predicted probability in each bin is plotted against the observed positive rate in the bin. A perfectly calibrated model traces the identity line.

3.7 Fairness Metrics

The fairness audit reports three quantities for each victim descent group g with at least 150 incidents in the test set. With $\hat{y}_i = 1$ whenever $\hat{p}_i \geq \tau$ (τ being the classification threshold), the group-level true positive rate, false positive rate, and predicted positive rate are

$$\mathrm{TPR}_g = P(\hat{y} = 1 \mid y = 1,\, G = g), \tag{3.5}$$

$$\mathrm{FPR}_g = P(\hat{y} = 1 \mid y = 0,\, G = g), \tag{3.6}$$

$$\mathrm{PPR}_g = P(\hat{y} = 1 \mid G = g). \tag{3.7}$$

Equalized odds (Hardt et al., 2016) holds when TPR and FPR match across groups.

Demographic parity holds when PPR matches. In general, both cannot be achieved simultaneously once base rates differ across groups, which is the situation here. The audit therefore reports both sets of quantities alongside a group-stratified AUC, the latter being a threshold-free measure of discrimination. All values are reported at $\tau = 0.5$; no corrective post-processing is applied.

CHAPTER 4

Results

4.1 Model Performance

Table 4.1 reports test-set performance for each classifier on each outcome. The test set is the 4,503 incidents with occurrence year 2023.

Table 4.1: Test-set performance on 2023 incidents ($n = 4{,}503$).

Target	Model	AUC	AP	Brier
CLEARED (broad)	Logistic Regression	0.814	0.620	0.174
	Random Forest	0.823	0.632	0.158
	Gradient Boosting	0.820	0.631	0.142
ARRESTED (strict)	Logistic Regression	0.796	0.298	0.176
	Random Forest	0.803	0.306	0.102
	Gradient Boosting	0.800	0.290	0.071

All three models achieve AUCs around 0.82 for broad and 0.80 for strict outcomes. Since the models only use information available at the time of the initial report, without any later updates, this level of performance is quite strong. The differences across models are also very small. In both outcome settings, the AUC gap stays within 0.01. This suggests that model performance is mainly limited by the features rather than the choice of model.

However, when evaluated based on the Brier score, the situation appears somewhat

different. The Gradient Boosting model achieved scores of 0.142 and 0.071 for "broad outcomes" and "strict outcomes," respectively—the best performance among the three models. Therefore, I will primarily use this model in the subsequent analyses.

When I switched from the broad definition to the strict definition, the Average Precision (AP) dropped from 0.62 to 0.30. This makes sense, as the proportion of positive samples itself decreased from 0.263 to 0.087. It is easy to assume that the model's performance has deteriorated; however, this is not the correct way to interpret the results. A more reasonable approach is to assess the extent to which the model demonstrates improvement relative to the "baseline probability." It shows an improvement of roughly 2.4 times the baseline, whereas the AP for "strict outcomes" corresponds to an even greater improvement, with approximately 3.3 times the baseline. Thus, although "strict outcomes" occur less frequently, the existing features demonstrate a stronger discriminative capacity for identifying them.

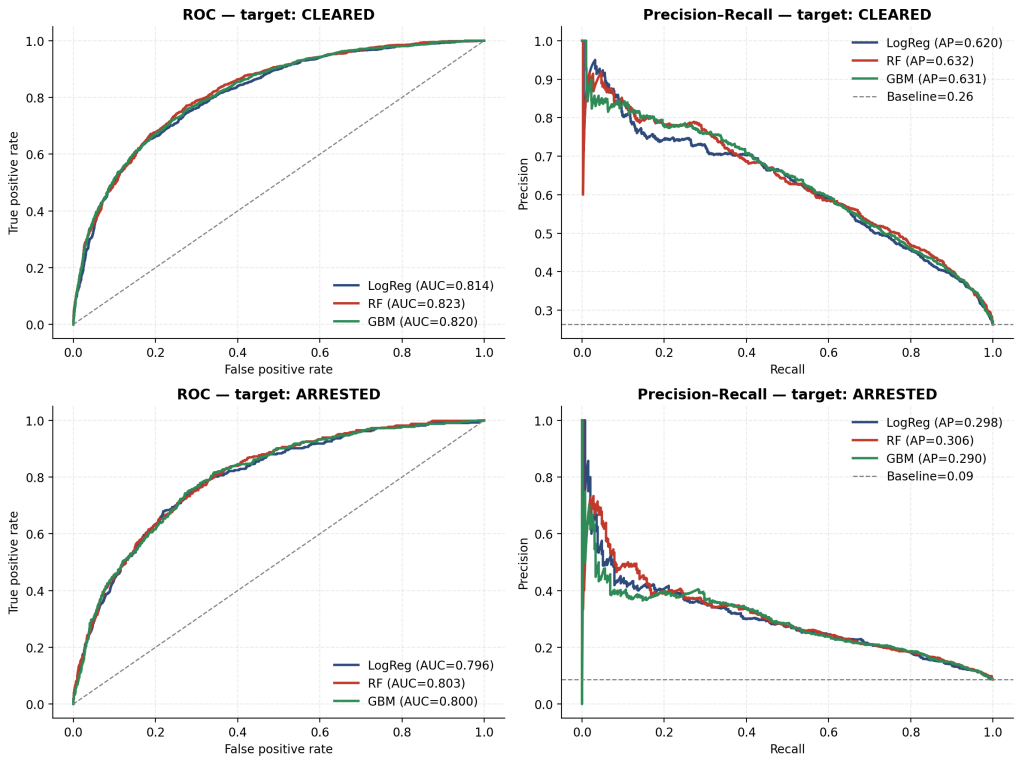

Figure 4.1: ROC and Precision–Recall curves for Cleared and Arrested outcomes.

4.2 Calibration

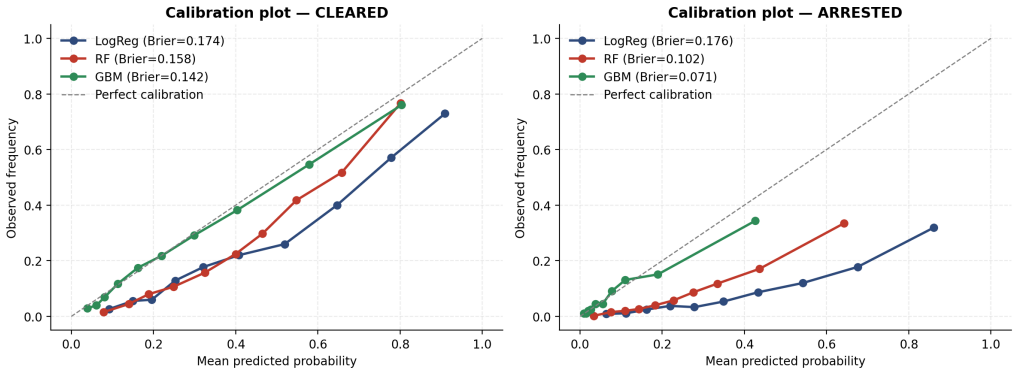

Figure 4.2: Reliability diagrams for each classifier.

Figure 4.2 shows the calibration performance of the 3 different method. Gradient boosting is the closest method to the diagonal in both outcomes and achieves the lowest Brier score. The random forest also follows the diagonal reasonably well, although some deviation appears at higher predicted probabilities in the CLEARED outcome. Logistic regression is far from the diagonal in the middle probability range and exhibits a mild S shaped pattern. Overall, the gradient boosting probabilities are well calibrated.

4.3 Feature Importance

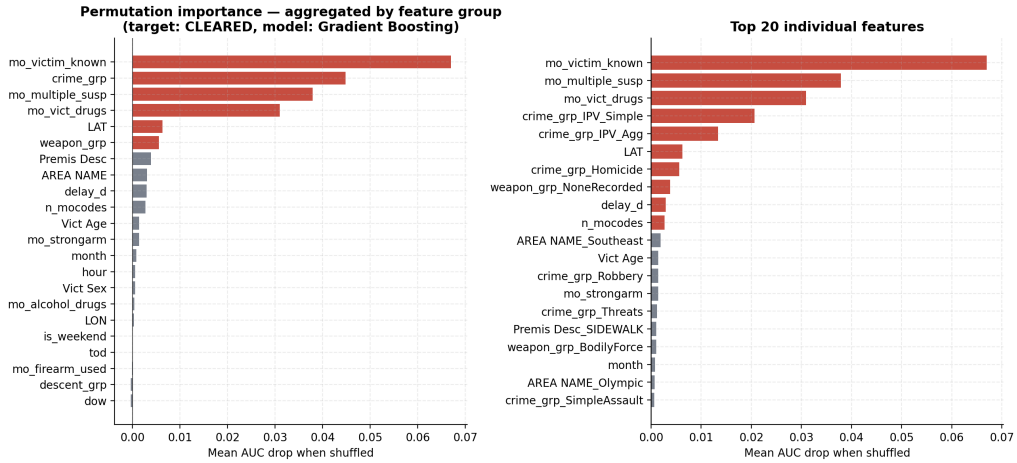

Figure 4.3: Permutation importance for the gradient boosting model on the broad outcome.

Figure 4.3 reports permutation importance for the gradient boosting model on the broad outcome. The quantity being measured is simple enough: shuffle a single feature's values at random and measure the drop in test AUC. Shuffling destroys whatever information the model had been extracting from that feature. Ten shuffles are averaged to produce each estimate.

Table 4.2: Aggregated permutation importance, gradient boosting, broad outcome.

Feature group	**AUC drop when shuffled**
mo_victim_known (MO 1402)	0.067
Crime group	0.044
mo_multiple_susp (MO 2004)	0.038
mo_vict_drugs (MO 0913)	0.031
Latitude	0.007
Weapon group	0.005
Premises description	0.004
Report delay (days)	0.003
Victim descent	≈ 0.000

Table 4.2 shows a ranking that is somewhat unexpected. The indicator for incidents in which the victim knew the suspect alone accounts for roughly one-third of the total feature importance recovered by the model. Then crime group shows an AUC drop of 0.044, with multiple suspects at an AUC drop of 0.038. These three features explain approximately 63% of the total importance.

Looking at the raw numbers, the same pattern shows up. Cases with MO 1402 clear at 64.6%, compared to 22.2% for those without it. That difference is larger than most differences across crime types once homicide is set aside, and larger than any gap across victim race groups.

4.4 Directional Effects

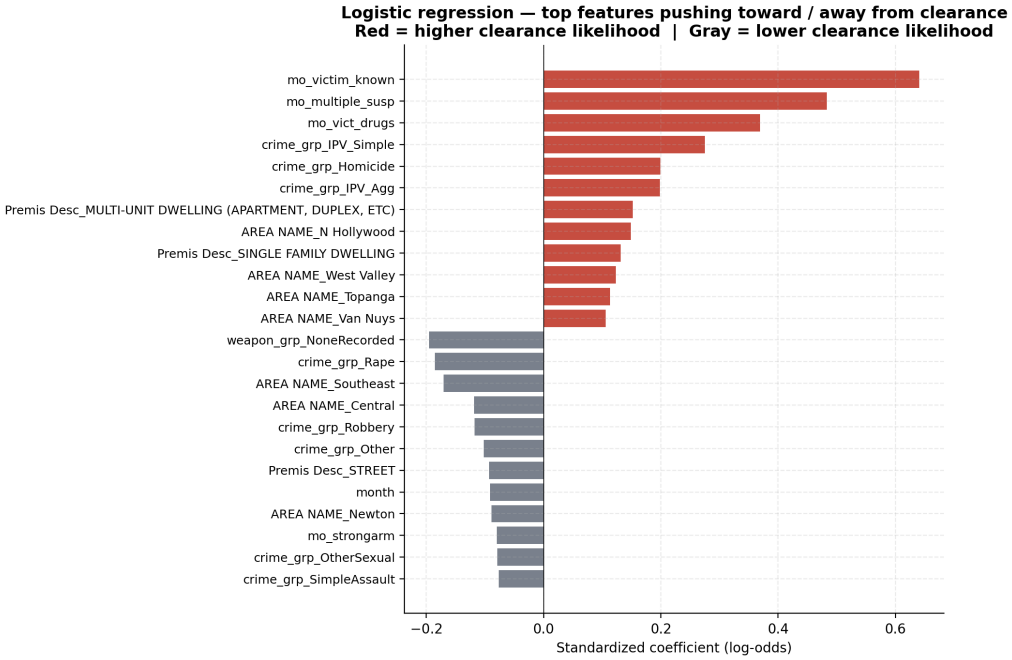

Figure 4.4: Top positive and negative standardized coefficients from the logistic regression model.

Tree-based models achieve slightly better predictive performance than logistic regression, but they do not readily support interpretation in terms of directional effects. Figure 4.4 shows the largest positive and negative standardized coefficients from the logistic regression model under the broad outcome definition. The coefficients corroborate the tree-based ranking and add sign to it. From the graph we can see that victim knew suspect is at around 0.64, multiple suspects at around 0.48, victim used drugs or alcohol at around 0.37. Intimate partner violence indicators follow at 0.28 and 0.20 for the simple and aggravated variants, respectively. Homicide contributes a positive coefficient of 0.20. Residential premises, both multi-unit and single-family, fall in the 0.13 to 0.17 range.

On the negative side, weapon_grp_NoneRecorded has the most negative coefficient at around -0.20, likely a proxy for cases with limited incident documentation. The coefficient for rape cases is -0.19. The coefficients for the Southeast and Central precincts

are −0.18 and −0.14, respectively; for robbery, it is −0.13; and for "street locations," it is −0.10. Rape stands out on the negative side. Even after controlling for precinct, time, weapon, and victim characteristics, the model still assigns lower clearance probabilities to these cases. This aligns with Figure 2.3, where rape shows a broad clearance rate of only 20.2% and an arrest rate of 5.6%.

4.5 Fairness Audit

Table 4.3 and Figure 4.5 show group-level fairness statistics for the gradient boosting model on the broad outcome at threshold $\tau = 0.5$. Four groups clear the minimum test-set sample threshold.

Table 4.3: Group-level fairness statistics at $\tau = 0.5$.

Group	n	Base rate	PPR	TPR	FPR	AUC
Hispanic	1,673	0.291	0.213	0.484	0.102	0.813
Black	1,436	0.248	0.183	0.469	0.088	0.824
White	1,140	0.250	0.155	0.428	0.064	0.812
Other	163	0.221	0.129	0.472	0.032	0.883

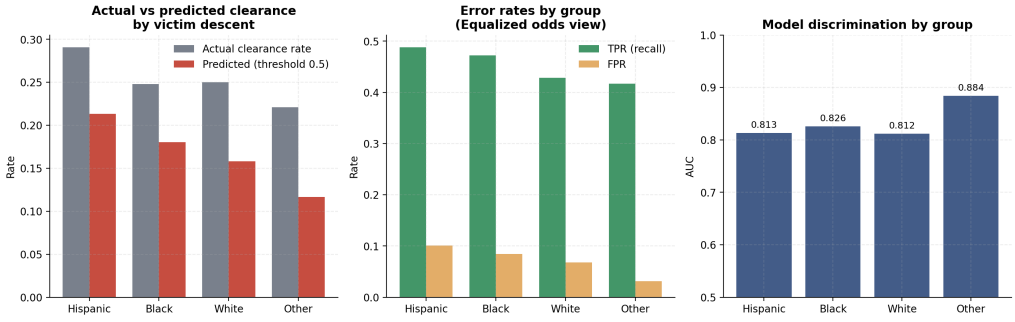

Figure 4.5: Group-level fairness diagnostics.

From Table 4.3 and Figure 4.5: first, we can see that AUC is approximately constant across the three large groups, between 0.81 and 0.82, and higher (0.88) for the smaller

Other group. Even though the victims' races are different, the model's ability to rank cleared from uncleared incidents works the same. Second, at $\tau = 0.5$ the model predicts positive at a rate below the observed base rate for every group, meaning the classifier is conservative overall. The degree of conservatism varies with group: the ratio of predicted rate to base rate is 0.73 for Hispanic incidents, 0.74 for Black, 0.62 for White, and 0.58 for Other. Both TPR and FPR move across groups, which is to say equalized odds does not hold. Hispanic shows the highest TPR (0.484) and the highest FPR (0.102). White shows the lowest on both counts (0.428 and 0.064).

Interpreting those disparities calls for some care. Victim descent was in the training feature set. Its permutation importance is approximately zero. Retraining with descent dropped from the feature set moves test AUC by less than 0.002 and barely touches the equalized odds gap. So the model is not learning to discriminate on descent; the mechanism lies elsewhere. What drives the gap is that other features in the model covary with descent. Hispanic victims in this corpus appear in incidents flagged with MO 1402 at 14.7%, against 11.3% for White victims, and MO 1402 is the single most powerful positive predictor the model has. The higher Hispanic TPR is, in effect, a mechanical consequence of the higher Hispanic rate of MO 1402 incidents. The fairness numbers are real enough. They reflect the incident distribution, not a decision rule the model learned about descent.

CHAPTER 5

Discussion

We can observe what the model is learning: one important factor is whether the victim knows the suspect. This is followed by the presence of multiple suspects and the type of crime committed. Fundamentally, they all reflect a single reality—whether the police can identify the suspect or at least obtain a reasonably clear description. This interpretation is broadly consistent with prior homicide clearance research, which has found that incidents involving non-stranger offenders are substantially more likely to be cleared (Roberts, 2007). If the victim already knows the suspect, they can typically make a direct identification. The presence of multiple suspects usually implies that there are more witnesses available to provide descriptions, and that these accounts can be cross-referenced to corroborate one another. Therefore, I believe that the type of crime is of great importance. This is because, in cases involving intimate partner assault—to cite one example—the victim and the perpetrator are known to each other. In other types of crimes, however, such as street robbery, the victim and the perpetrator are, for the most part, strangers. The unifying concept running through these scenarios is the identifiability of the suspect. Geography, weapons, time of day, victim demographics: these feed the model only as secondary inputs, and in several cases not at all.

Taken by itself, the finding is consistent with a great deal of policing folklore. Investigators have observed that in most solved cases, the suspect is already known when the investigation opens. Cases where the suspect is identifiable at report time tend to clear at much higher rates. Whether that reflects a causal mechanism, I cannot say from this data.

The second thing is about the dual-target framework. Under the broad outcome, intimate partner aggravated assault looks like one of the most cleared categories of violent crime in Los Angeles, at 69%. Under the strict outcome, the arrest rate for the same category is 19.8%. About 49% of the gap gets filled in by exceptional clearance. In the intimate partner context, exceptional clearance is dominated by victims declining to prosecute. What the broad rate measures, then, is something closer to administrative closure of the case than to any punitive outcome. In homicide cases, the difference between the broad clearance rate and the strict clearance rate is only about 5%. I believe the reason is that exceptional clearances are relatively rare in homicide cases, and most cleared cases are resolved through actual arrests.

When we put all different types of cases together and evaluate them based on a single "clearance rate," the result can be misleading. Different types of cases require different approaches. In domestic violence cases, many such cases are labeled as exceptional clearance, but this is because the victim has chosen not to pursue the matter further. A department that handles mostly stranger crime will post low clearance rates not because of weaker investigative effort, but because stranger crimes are inherently harder to clear. The FBI does not currently require agencies to report the two rates separately. It should. Doing so would make clearance statistics more informative and cross-agency comparisons a good deal more honest.

Two smaller points round out the chapter. Report delay, despite correlating with clearance in bivariate statistics, has almost no independent predictive power in the model, with a permutation importance of 0.003. The cases reported with a delay appear more difficult to resolve. Cases reported more than seven days after the incident have a clearance rate of only 18%, whereas those reported within a single day have a clearance rate of 28%. This "reported late" pattern usually occurs in sexual offense cases, which are considered harder to clear. The difference comes from the type of crime; therefore, reporting delay's effect is small in our model.

Then it comes to calibration. Overall, the probability estimates generated by the gradient boosting model are reasonably reliable. But if we apply the default threshold of 0.5, the model will appear a little conservative, predicting fewer "resolved" cases than it should. The outcome is a result of the chosen threshold; this issue could be remedied by adjusting the threshold; however, in the context of the descriptive analysis presented here, this factor has a negligible impact.

CHAPTER 6

Conclusion

6.1 Summary of Findings

I used machine learning models to analyze violent crime cases recorded by the LAPD between 2020 and 2023 in this thesis, using "whether a case is cleared" as the predictive target. I applied two definitions for this purpose: a broad definition, which counts both arrests and "exceptional clearances" as cleared cases; and a stricter definition, which considers only actual arrests as cleared cases. The best model is Gradient Boosting for our dataset, having an AUC of approximately 0.82 under the broad definition and 0.80 under the strict definition.

The results show one important feature, is whether the victim knew the suspect from MO codes. When I removed this variable from the model, the AUC dropped by 0.067.

Moreover, the model's AUC is stable across victim race groups, meaning the model ranks cases equally well regardless of group. TPR and FPR do differ across groups at a fixed threshold, but this reflects differences in the distribution of features across groups (particularly MO 1402), not a decision rule the model learned about race.

6.2 Future Research

If I had more time, there are a few things I would want to work on next.

The first is survival analysis. Right now, I treat clearance as a simple binary outcome. But it's more about how long it takes for a case to be cleared. The LAPD data doesn't include the actual clearance date, so I can't do this with the current dataset. If I could

link it to court records, then I could switch to a survival analysis. That would let me separate cases that are solved quickly from those that take much longer and see whether the same factors that predict whether a case gets solved also say something about how fast it happens.

The second is causal analysis. All of my current results are descriptive. I can say that knowing the suspect is the strongest predictor of clearance, but I can't say it actually causes a case to be solved. To get closer to that, I would need to use methods like instrumental variables or matching. That would require some kind of external variation in the relationship between the victim and the suspect.

The third thing I want to explore is to find out if my findings work on other cities. The key conclusion here is whether the suspect is known matters more than anything else. It would also be interesting to compare how big the gap is between the broad and strict clearance rates across different cities. Finally, I would want to make better use of the MO codes. In this thesis, I only used a small number of them, even though there are hundreds in the full codebook.

References

[1] Federal Bureau of Investigation. (2013). *Clearances*. Uniform Crime Reporting Program, U.S. Department of Justice.

https://ucr.fbi.gov/crime-in-the-u.s/2013/ crime-in-the-u.s.-2013/ offenses-known-to-law-enforcement/ clearances/clearancetopic_final

[2] Roberts, A. (2007). Predictors of homicide clearance by arrest: An event history analysis of NIBRS incidents. *Homicide Studies*, 11(2), 82–93.

https://doi.org/10.1177/1088767907300748

[3] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315–3323.

https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning

[4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.

https://doi.org/10.1214/AOS/1013203451