Norwegian School of Economics
Bergen, Spring 2021

Correcting Witness Reports through Machine Learning

An empirical study of machine learning applied to incident reports

Petter Fredrik Hemnes

Supervisor: Floris Tobias Zoutman

Master thesis in Economics and Business Administration

Major: Economic Analysis

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

In this thesis we investigate the possibility of using machine learning models to correct witness testimony. Using data from the National Incident-Based Reporting System, we build a model to predict the race of offenders on data of arrests and compare the model predictions to that of witness guesses in non-arrest incidents. We find that witness reports are erroneous in 16.17% of the incidents, and that the error in witness reports lead to an expected yearly police cost of $8.2 million dollars for the crimes: burglary, robbery, assault, rape, and homicides. We suggest several ways the machine learning model can be used to correct witness reports. First, the model prediction can be used directly to correct reports. For instance, values can be imputed for unknown offenders, and the labels where there is a disagreement between the model and witness guesses can be replaced with model predictions. We find that witness error can be reduced to 8.77% if all labels are replaced with model predictions, saving $4.5 million in yearly police cost. An alternative to be considered is combining witness guesses with model predictions to improve predictive accuracy. The model predictions can also be used indirectly to correct reports as an alarm tool to identify the possibility of error. The reports which are labelled likely to be erroneous can then in turn be investigated by humans. Finally, the model can be used to correct the confidence of the eyewitness identification, by i) comparing the eyewitness prediction to a continuous prediction made by an accurate model, or ii) to quantify the amount of expected error in the testimony.

Table of contents

Tables

Figures

Equations

I Introduction

The account of an eyewitness is often used as evidence to uncover the truth of a crime. Research suggests that witness testimony is the single most important factor leading to wrongful convictions within the legal system (Horvath, 2009). One historical study has found that 45% of wrongful convictions are due to erroneous eyewitness testimonies (Borchard, 1932), with some sources claiming the same rate to be as high as 72% (New England Innocence Project, 2021). In congruence with these findings, decades of research on the predictive reliability of witness testimonies have revealed that testimonies can be erroneous due to several factors. Studies from the field of psychology reveal that factors pertaining to the crime and the witness, such as stress, the presence of weapons (and even hats), own-age, and own-race/cross-race can impact a witness' ability to correctly identify an offender (National Research Council, 2014). Studies also show that bias may be introduced after the incident in the way testimonies are gathered. For instance, the structure of the questioning, the line-up, how offenders are presented, and investigator's bias may impact testimonies. In addition to these factors, it is easy for a suspect to change external characteristic after a crime or disguise themselves when they commit crimes, which can severely affect the accuracy witness's testimony (Cutler, Penrod, & Martens, 1987).

Erroneous witness testimonies are costly to society as they mislead investigations and have been shown to lead to wrongful convictions. Although it is difficult to quantify cost to society, estimates of the expected yearly police cost can work as an anchor. Using estimates for the number of eyewitness cases, the average cost of policing for different crimes, and our estimate for witness error, we find the yearly police cost of erroneous reports for five different felonies. We find that the yearly expected cost due to witness error for burglary, robbery, assault, rape, and homicides is collectively 8.2 million USD. This is not accounting for the wrongful arrests and wrongful convictions.

Although it is well-established that eyewitness identification is unreliable, can lead to poor legal outcomes, and be costly to society, there has been little research on how to correct testimonies. Instead, research has focused largely on prevention in introducing biases and errors during the gathering and application of witness testimony. In 2014, the National Academy of Sciences appointed a scientific committee to review the research on eyewitness identification and provide recommendations on how to strengthen the value of eyewitness identification evidence in court. The committee recommended to improve training of law enforcement officers, develop standardized lines of questioning, videotape the process, and encouraged further research. These measures are likely to have an effect in preventing errors, however, it ignores many of the errors introduced by the witnesses themselves. The committee also recommended increased use of statistical tools and quantitative research but did not give specific guidelines on methods and data collection. The lack of application of statistical tools and quantitative methods in improving the accuracy of witness testimonies is the motivation for this thesis.

Machine learning (ML) is a method of automating data analysis to discover patterns without explicit programming. The method encompasses several algorithms and has quickly become a popular tool to do predictions in various sectors and industries**.** If we can illustrate the value of a machine learning model by producing precise estimates of error in eyewitness identification, it would be an ideal base for further research and discussion in using this method for correcting witness reports. The purpose of this paper is therefore to question whether machine learning can be used to correct witness testimony through identification of an offender. Accordingly, the problem statement for this thesis is:

Can machine learning be used to correct witness reports?

In answering the problem statement, we have made assumptions that needs to be clarified. First, we assume that actual arrests can be used as a proxy of true crime. That is, arrests correctly reflect what crimes are committed and, most importantly, the characteristics of criminals who commit the crimes. Using this assumption, we can train a machine learning model on actual arrests and assert that it represents true crime. Second, we assume that characteristics of offenders in non-arrests (i.e., witness incidents) are guessed by eyewitnesses. Using this assumption, we define the discrepancies in non-arrests reports to be due to witness error. Third, we assume that arrests and non-arrest incidents are identical in nature; arrests and non-arrests are drawn from the same distribution. This assumption allows us to generalize the performance and results from our machine learning model on witness guesses.

We focus our analysis on the witness error in labelling race, specifically for black and white perpetrators. Using a machine learning model trained on actual arrests, we produce predictions for the race of the criminals in witness incidents. Discrete model predictions are given as either 0 or 1, where 0 represents the offender being black and 1 represents the offender being white. Witness guesses are coded in the same way. The witness error is defined and estimated as the average difference between model predictions and witness guesses.

The model trained in this thesis can distinguish well between offenders of different races based on the features used, and there is significant disagreement between the model and witness guesses. The witness error is estimated to be 16.17% for the crimes studied. We suggest ways for which this ML model, or others like it, can be used to correct witness reports. For instance, witness guesses can be replaced in entirety by discrete predictions made by a machine learning model, or model predictions and witnesses guesses can be combined. Furthermore, the continuous model predictions can be used to gauge testimony confidence and guide resource allocation in law enforcement.

The remainder of the paper is organized as follows. In Section II we give an overview over existing research and methods. In Section III the data used for this study is described, as well as the data cleaning process and variable selection. In Section IV, the conceptual framework for estimating witness error is presented, and in Section V theory on classification problems along with theory of the methods for analysis is outlined. In Section VI, the main results and the robustness of our results are reported. In Section VII we provide suggestions for applications of the model including an estimate of the cost associated with witness error. In Section VIII we describe the limitations of the analysis and the interpretation of the results. We conclude in Section IX.

II Literature Review

This thesis is related to several strands of literature. First, it contributes to the policy literature on eyewitness testimony and the evidence of their unreliability. Elizabeth Loftus (1978) found that human memory is malleable, making eyewitness testimony unreliable. In Loftus et al., (1978), the researchers presented 1242 subjects with slides depicting a single auto-pedestrian accident. The subjects were then exposed to information that was either consistent, misleading, or irrelevant. The subjects that were presented with misleading information produced less accurate responding on recognition tests. Researchers had discovered unreliability in testimonies prior to the study by Loftus et al. (1978). For instance, Johnson and Scott (1976) first determined that the presence of a weapon may negatively influence eyewitness memory for an event. This effect, known as weapon focus, has been well studied. One example is Hope & Wright (2007), in which subjects were shown a slideshow of a simulated event while attending to a secondary task. In the simulated event, a target was shown holding an object which differed depending on the participant group. Participants in the weapon group had the poorest performance on recognition tests for the target's appearance. However, an analysis of the weapon focus literature show inconsistencies in findings (Fawcett et al., 2013), with slight evidence for weapon focus in actual crimes, and a slightly larger effect in laboratory studies. In a paper even earlier than Johnson and Scott's paper on weapon focus, Feingold (1914) stated that humans perceive individuals of a different race to look alike, making it difficult for people to distinguish between faces of different races. This effect has later been labelled as own-race bias or cross-race bias. As with weapon focus, own race bias is well-studied, but research is more conclusive towards own-race bias having a pronounced negative effect on accuracy, with one analysis claiming that cross-racial misidentifications were present in 42 percent of the cases in which an erroneous eyewitness identification was made (Grimsley, 2012). Most recently, own-race bias was illustrated in Wong et al., (2020), where a group of university students of different races was asked to remember the pictures of faces of individuals of different races. The subjects were shown the pictures two times: First, in a learning phase, and a second time, in a recognition phase. In the recognition phase, the subjects had to recall whether they had seen the face before (yes/no) with an additional option to label the face as known from before the study. All races amongst the subjects exhibited higher accuracy in recognizing faces from their own racial group. Furthermore, the own-race bias was not significantly reduced from (self-reported) interracial contact, indicating that exposure to other races does not significantly reduce the bias. In addition to weapon focus and own-race bias, many other effects have been studied (e.g., the effect of exposure duration, and the effect of retention interval) and successfully replicated in later times

(Fawcett et al., 2013; Palmer et al., 2013; Loftus & Hoffman 1989; Horvath, 2009; Kapardis, 1997).

Most studies in eyewitness testimony research are laboratory-based experiments as opposed to field studies (Kapardis, 1997). This is criticized in several papers (Yuille, 1986; Bruck & Ceci, 1995), as controlled research may not generalize to real world contexts, and legislation should not be based upon one research method. Our study of witness error uses second-hand data on actual incidents and shows that witness error is significant in a wide variety of crimes for a general feature, race. We also briefly revisit concepts of weapon focus and cross-race bias. We find evidence to support a theory that individuals are better at identifying individuals of their own race rather than individuals of a different race, however we do not find evidence to support the claim that the presence of weapons introduces witness error.

The thesis is also related to the literature on the identifying of misclassification in data (Sabzevari et al., 2018; Brodley et. Al. 1999; Wietman 1986). Most closely related is Brodley et al. (1999) which applies the idea of using a set of classifiers trained on one part of the data to test if instances in the remaining part of the data are mislabelled. We generate an ensemble classifier using cleaned data (true crime) and use the classifier to predict labels for unfiltered data (witness testimony) in order to magnify the error rate. If the model prediction and a witness guess do not match, we identify it as witness error.

III Data

Our objective is to increase the probability that the characteristics of a criminal are labelled correctly at the time the incident is reported. As such, it seems pertinent to use data associated with incident reports to build our model. In this section, we present the data we will use to train our model. We use second-hand data provided by FBI through NIBRS. This data has a tradition of being used by law enforcement and researchers to gather a detailed picture of crime, including data on offenses, suspected offenders and arrestees, and victims.

A. Background on NIBRS

The National Incident-Based Reporting System, or NIBRS for short, was created in 1980. The NIBRS can be viewed as the latest contribution to a 90-year effort of providing informative crime statistics to the public and law enforcement, and has the specific mission of contextualizing crime by providing higher levels of data specificity. The data has been made available to researchers and numerous studies have been published using the data. For instance, Addington (2006) which uses the data to evaluate predictors for clearances of murders, or D'Alessio et al. (2002) which uses the data to investigate the relationship between racial threat and interracial and intraracial violent crimes. From 2015 to 2021, the NIBRS have transitioned into becoming the national crime data collection program, further adding to the robustness of the data. It is expected that 75% of law enforcement agencies, serving 80% of US population have moved to NIBRS by 2021 (FBI, 2020). The high level of specificity, the quantity of the data and the robustness of the data makes NIBRS suitable for the purpose of this thesis.

B. Type of Crime

In NIBRS, crime is separated into three categories: (i) Crimes against persons (CAP), (ii) crimes against property, (iii) or crimes against society. However, incidents may be in more than one category as up to ten crimes can be committed within one incident. We have chosen to use the first offense recorded as representation for the crime as it is the most serious offense for ~70% of the incidents and have chosen to focus our analysis on crimes against persons where property was also lost. This allows us to add property variables (such as the type of property loss and the value of that property) to our analysis while looking at crimes that often involve variables that are believed to cause bias in witness testimony (weapons, force, bias motivation). In addition, relationships between the victim and the offender are recorded exclusively for crimes against persons[1]. This allows us to filter out the incidents where it is reported that the victim has prior knowledge of the characteristics of the offender. For the incidents where there is more than one offender, we use only the relationship to offender 1, as we cannot perfectly link information on offenders to arrestees. In addition, some relationships are unreported, even though this field is

[1] Relationship to offender was originally added to track domestic violence.

specified as mandatory for violent crimes in the NIBRS data guidelines. We assume that missing entries are equivalent to the offender being a stranger.

The data for this study was obtained from the 2015 records in the NIBRS database. In that year, a total of 103240 offenders of a CAP with a property component were reported (in NIBRS), and 89% of incidents involved an offender that was reported as either black or white. Among all incidents, 26781 are non-arrests – incidents in which the offender has not been arrested – and the remaining 76459 have an arrest associated with the incident. To gain an overview over the characteristics of the crimes that are being studied, we consider some relevant summary statistics. From Table III.II, the number of crimes committed between black (44%) and white people are similar with a slight majority for whites (45%). Most incidents involve two offenses committed (56%), a single offender (39%) and two victims (31%). The offenders range from juveniles to elderly, with an average age of 23. The same can be said for victims, however the average age was higher (31). Both offenders (76%) and victims (60%) are primarily male. Most of the incidents are assault offenses[2] (82%), followed by kidnapping (13%). The hotspots where the incidents occurred most frequently were residences (30%), highways and roads (18%), and parking lots (6%). A weapon was used in 84% of the cases, and the most common weapon was a handgun which was used in 35% of incidents where a gun was involved.

Table III.I Summary statistics of data

		Mean	Count
Offender	White	0.44	45274
	Black	0.45	46493
	Male	0.76	78646
	Age	23	*NA*
Victim	White	0.63	65480
	Black	0.28	28691
	Male	0.60	61740
	Age	31	*NA*
Incident	Weapon	0.84	86722
	Assault	0.82	84037

[2] Includes aggravated assault (30%), simple assault (38%), intimidation (13%).

Kidnapping	0.13	13338
Rape	0.02	2095
Residences	0.30	30771
Roads[1]	0.18	18224
Parking lots[2]	0.06	6566
Two offenses	0.56	58037
Single offender	0.39	40530
Two victims	0.31	32238

[1] Highways/Roads/Alley/Street/Sidewalk

[2] Parking/Drop Lot/Garage

C. Data Sets

We construct two data sets: One for arrests, and one for witness guesses. The data set for witness guesses are generated by identifying offender segments for which there is no corresponding arrest segment. In other words, the witness incidents are non-arrest incidents. For fluidity we will use varying names for these data sets, but the words ‘arrest’ and ‘witness/non-arrest’ is always used to distinguish the two. The data on actual arrests is used to train and validate the model, whereas the data on witness guesses is used to evaluate the witness-error.

D. Data Cleaning

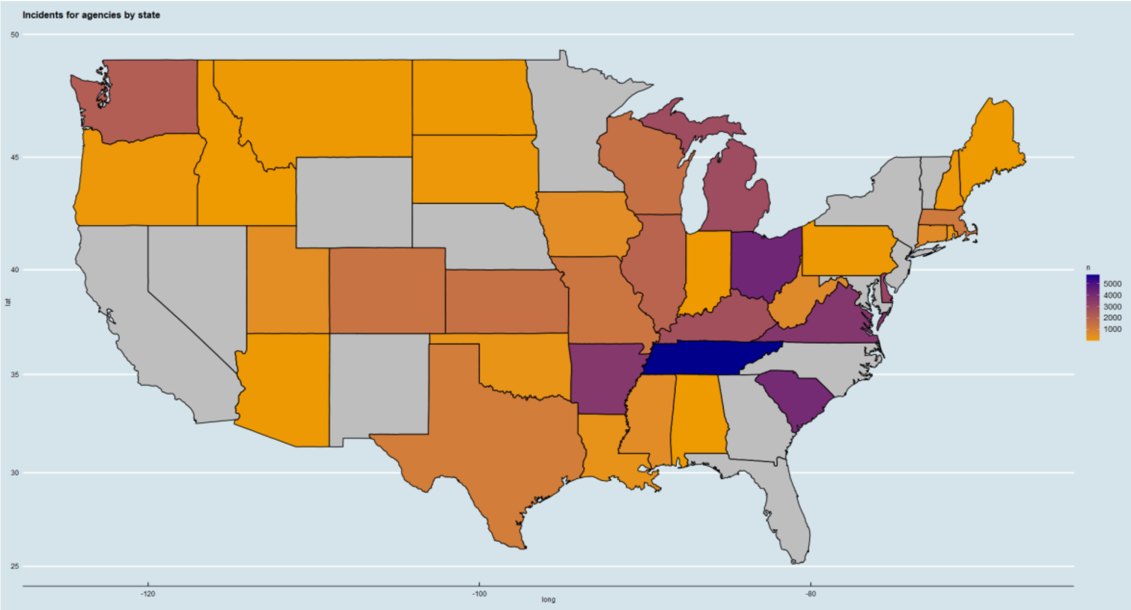
The NIBRS database separates information on an incident in five segments: arrestee, victim, offense, offender, and administrative. Out of the five segments in the NIBRS data, four are used to construct the arrest data set. We create a data set on arrests by merging the segments: arrestee, victim, offense, and administrative. We construct the witness data in a similar way as the arrest data, but instead of using data on arrests, we use data on offenders. An offender’s traits are reported as identical to that of a corresponding arrestee, suggesting that the offender data is changed when an arrest has been made to match the characteristics of the arrestee. To support the inference, we note that offender age and arrestees age are identical in all observations for which an arrest is reported. The likelihood that offender data is edited to correspond to arrest data after an arrest is made seems more probable than the witnesses perfectly predicting the age of offender at the time of reporting in all incidents. Post-arrest editing means we cannot use the differences in arrest data and offender data directly to discern the error witnesses make in labelling characteristics of a criminal. Instead, we separate the incidents by whether an arrest has been made. The witness data consists of incidents for which the offender has not been arrested and so his or her characteristics remain uncertain, and the labels are assumed to represents witness guesses. There is a caveat to the method in that the true value of the offender's race remains unknown for non-arrestees. For our later inferences on witness error to be valid we must assume that the distribution of incidents where an arrest has been made and the distribution of incidents where an arrest has not yet been made are the same.

For our dataset on actual arrest and witness guesses to be comparable it is important that the variables contained are symmetric. If the variables are the same, the model can be used to predict labels in the witness data and the difference in prediction accuracy can be used to approximate witness error (section IV.C). For all categorical variables in our data, we identify the intersect of categories between our data set and filter out the observations which are not in the intersection. For example, the variable *location* takes on the value of 42 in some incidents in our witness data, signifying that the incident took place at a camp or a campground. If there were no incidents in our arrest data which took place at a camp or a campground, we omit all campground incidents from the witness data. Some categories are labelled unknown in the witness data, such as the race of the offender and the race of the victim. This could be due to a number of reasons, one of them being that the victim is unsure of the offender's characteristics and therefore does not wish to label them. We omit these categories as we only want to look at cases for which a clear prediction was made by a witness. In addition, some observations are left completely empty. For variables such as race, NIBRS guidelines specify that is it mandatory to report a value, and that it should be reported within three categories. We infer that incidents with unlabelled data for obligatory fields are misreported. As we do not want to endogenously affect our assumption that arrests reflect true crime by imputing data, we omit the misreported incidents from our data sets.

E. Dimensionality

Optimally we want to use all data available in NIBRS. However, as we omit missing data, variable selection became crucial to maintain enough observations to train a model capable of generating accurate predictions. In addition, we decided to limit the number of categories for some categorical variables due to computational feasibility. One example of omitted variables are levels of the agency identifier. The agency identifier is especially useful because it includes information on geographic location in our analysis. As such it works as a fixed effects estimator for numerous variables such as demographics and regional wealth. For the crimes we want to analyse, incidents are reported in more than 5000 unique agencies. We remove agencies for which there are fewer than 1000 incidents reported, resulting in 144 agency identifiers. The resulting incidents to be considered grouped by agencies by state can be seen in Figure III.I. Most of the incidents used in our analysis take place in Tennessee, South Carolina, and Ohio.

Figure III.I Incidents for agencies by state



The final data - identical in dimensions for the two data sets – contains 18 independent variables of which 7 are factor variables, 4 are an indicator variable, and the remaining 7 are numeric variables. Three of the variables relate to the property crime segment, five are general offense variables, three are specific to violent crime, and three are administrative variables - such as the number of victims and offenders involved in the incident and the date and time the incident took place. The variable for hour is inherently cyclical[3] and was sine and cosine transformed to reflect this. Using a two-dimensional transformation, hour is made to swing back and forth as a cyclical variable should, and the distance between 23 and 00, will be the same as the distance between 1

[3] For hourly data reported in military time, the distance between 24 and 1 are the same as 1 and 2. If hour is coded as a numeric variable, the distance between 24 and 1 will be 23.

and 2[4]. A visualization of time as a cyclical variable is provided in Appendix A.I. Although a two-dimensional transformation could negatively impact the distance-based and tree-based algorithms, in comparing different models, we found that the model performed better when hour was coded as a two-dimensional cyclical variable as opposed to a categorical variable. In addition to cyclical time, datetime is added to represent the linear flow of time. The flattened (factor to indicator transformed) data, consists of 240 variables. The total amount of observations after filtering the unknowns and missing entries is 11599 for arrests, and 16732 for witness incidents. The ratio of predictors to observation is about 1/48, that is, there are 48 observations per predictor considered. The dependent variable, race, is divided into four categories in NIBRS (Hispanic is recorded as an ethnicity): *Black*, *White*, *American Indian/Native*, or *Asian/Pacific Island*. However, our algorithms will require a binary outcome, and so we must choose two out of the four to be used in our analysis. We use white and black as they are the two majorities represented in the data. Appendix A.III shows a complete overview over the included variables.

IV Conceptual Framework

In this section, we present the concepts, assumptions, and the qualitative framework underlying the analysis. We present how witness reports are used in criminal investigations, the qualitative definition of witness error in this thesis, and the conceptual method used to identify witness error.

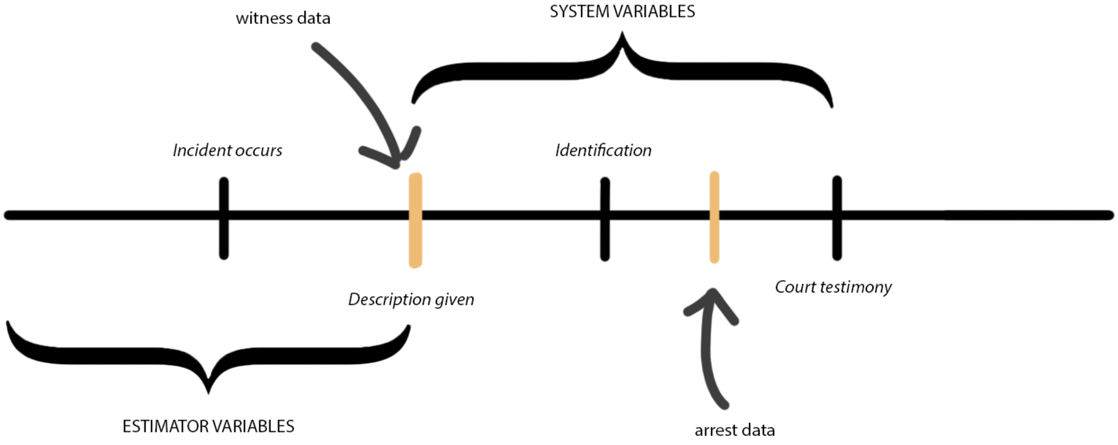
A. Stages of witness testimony

The timeline for how witness testimony is generated and given can be viewed in four stages. First, a bystander or a victim is witness to a crime. Second, police obtain a description of the offender from the victim (Clifford & Davies, 1989). Third, witnesses are used to identify the perpetrator from the potential suspects. Fourth, an eyewitness is asked to testify in court. A testimony does not have to go through all the stages, and sometimes it is not used beyond the first stage. Procedures may also differ according to jurisdiction and between countries.

[4] The cyclical relationship the periodic functions sine and cosine produce together can be demonstrated by plotting the values on a unit-circle.

Factors that negatively influence the accuracy of the testimony can be introduced in all stages. In literature, it is often differentiated between estimator variables and system variables (Wells, 1978). Estimator variables are those which occur at the time of the event (or prior to the events e.g., prejudices) and cannot be controlled for by the legal system. Conversely, system variables are defined as the variables that occur after the incident takes place. In Figure IV.I we provide an overview over the timeline for testimony, when estimator and system variables occur, and the time points for our data. As specified in the data section (III.C) we use non-arrests as witness data, and so the witness data encapsulates the process up until after the first stage. The arrest data is recorded between stage three and four, and the conviction verdict is unknown.

Figure IV.I Stages in the eyewitness process, error and our data time points



B. Conceptual method

At the time an incident is reported (the timepoint of our witness data), enforcement agencies have details on the victim, the crime, and the offender. Some of the information given by the victim to the police can be ascertained with relative certainty. For example, there is little uncertainty involved in the details about the victim as they can be verified through legal identification. Furthermore, by establishing a timeline, uncertainty around the location and hour of the crime can be reduced. However, without hard evidence such as video or photography, the characteristics of the offender remains uncertain as it is given solely by witness recollection. This creates numerous problems as witness recollection is malleable and may have been contaminated due to environmental factors or their own biases. We use the certain information associated with a victim and a crime to predict the uncertain characteristics of an offender. If the model is accurate, the disagreement between a witness guess and the model indicates that the offender has likely been misclassified. This method is motivated by Brodley & Fiedl (1999) where a set of classifiers formed from training data is used to test whether instances in the remaining part of the data are mislabelled.

In example, we look at incident WZ-ZOQC4B0W5 from agency IL1010400. The incident was a case of aggravated assault which took place at 15:00 on the 8th of March 2015, in Illinois. The incident occurred at an auto dealership, the perpetrator was reported using an automatic handgun, and the offender also damaged commercial structures worth $2000. The witness labelled the offender as a black male of unknown age. From the arrest data we know that for an aggravated assault at an auto dealership in Illinois where an automatic handgun is used, the perpetrator is likely to be white, and this is what the model predicts. In fact, for incident WZ-ZOQC4B0W5 the model predicts that the perpetrator is white with an 82% probability. Therefore, there is a high likelihood that the witness has mislabelled the criminal. We classify this case as an erroneous report. In estimating the incidence of witness error, we average the sum of all such disagreements between the model prediction and witness guesses.

We expect to find a small but significant average error in witness cases. We expect the error to be small because studies show that witness testimony is often reliable if uncontaminated by the legal system (Wixted, Mickles, & Fisher, 2018). As, our witness data time point is largely before application in the legal system, we expect reports to be subject only to estimator variables. Our overarching prediction of the results can be summarized as,

PREDICTION:

Witness error is small, but significant at the time of reporting.

For our inferences to be valid we make assumptions about arrests and witness data. First, we assume that arrests reflect the true crime rates. In other words, we assume that an arrestee is guilty. This approximates the truth, at least legally, as the conviction rate is around 90% overall and 70% for felonies (United States Department of Justice, 2012). If arrests reflect true crime rates, then a model trained using arrest data will approximate the true relationship between features of the crime and characteristics of a criminal. Second, we assume that the labelling of an offender in non-arrest cases are done by witnesses. Using this assumption, we assert that the differences we observe are due to witness mistakes. Third, we assume that arrests and non-arrests (witness cases) are drawn from the same distribution. In other words, we assume that there is no difference in the nature of the crimes between our two data sets. This assumption is necessary as we use incidents for which an arrest has not been made to uncover the error made by witnesses. If the samples are drawn from different distributions, the model predictions do not generalize to the witness sample, and disagreements between model predictions and witness guesses can be due to different reasons than witness error.

V Machine Learning Method

The primary question to be answered in this thesis is if machine learning can be used to correct witness reports by analysing disagreement between a prediction model and a witness prediction. To answer the question, we train an ensemble model to classify the race of an offender and compare the performance of our model to witness guesses. In this section, we present the method used to build this ensemble model. First, we present machine learning theory on classification problems and show how to evaluate the performance of a ML model. Subsequently, we discuss complications in approximating the relationship between a target variable to feature variables and suggest cross-validation and ensemble learning to overcome this challenge. Finally, we present our method of choice, the Super Learner.

A. Classification

In a machine learning classification problem, the objective is to use a feature vector x and a qualitative response Y to build a function $f(x)$ that takes x as input and predicts a corresponding value for Y. For our purpose, the feature vector consists of characteristics of the crime and the victim, whereas the response is a characteristic of the criminal, namely race. In the case of a classification problem, the predictions are first generated as a continuous value, typically by functions that force a value between 0 and 1, and so it can be interpreted as a probability. To turn the probabilities into a class prediction (black or white), a cut-off must be used. A class prediction can be formulated as,

$$\hat{Y}_{class} = \begin{cases} 0 & \hat{Y}_{\%} < c \\ 1 & \hat{Y}_{\%} \geq c \end{cases}, \qquad \textbf{V.I}$$

where $\hat{Y}$ is used to denote predictions, and c is a cut-off, for example 0.5. If the prediction is less (greater or equal) than 0.5, the class prediction is 0 (1).

When used a classification model is built for prediction, the focus is not on the causal relationship between the feature variables and the response variable, but instead on the accuracy, or conversely the error, of the predictions the model produces. The loss function for measuring errors between Y and $\hat{f}(X)$, denoted by L can take many forms. Some typical choices are the squared error and absolute error,

$$L = \begin{cases} \left[Y - \hat{f}(X)\right]^2 \\ |Y - \hat{f}(X)| \end{cases} \qquad \textbf{V.II}$$

To optimize the model predictions the loss function is minimized, and the best model is the model which has the least amount of error. Conversely, if formulated as a maximization problem the best model is the model which has the highest accuracy. For classification problems, the Area Under the ROC Curve (AUC) is a frequent metric for measuring model performance. Intuitively, maximizing the AUC may also lead to favourable results. As AUC is a non-differential function, a nonlinear optimizer must be used if AUC is to be maximized. Using nonlinear optimization could be problematic both in terms of finding optima and in terms of computational burden. If possible, however, maximizing the AUC for binary classifiers is shown to lead to good results (LeDell, Laan, & Peterson, 2016).

The class prediction cut-offs can also be optimized. If correct classification is equally important between the groups, the optimal cut-off is the one that separates the group such that accuracy is maximized. In our analysis we consider misclassification of black offenders equally important as misclassification of white offenders. The Youden index J is a metric that can be used to evaluate the cut-offs. The optimal cut-off is the cut-off that corresponds to the highest Youden value (Ruopp et al., 2008). In other words, the optimal c is such that,

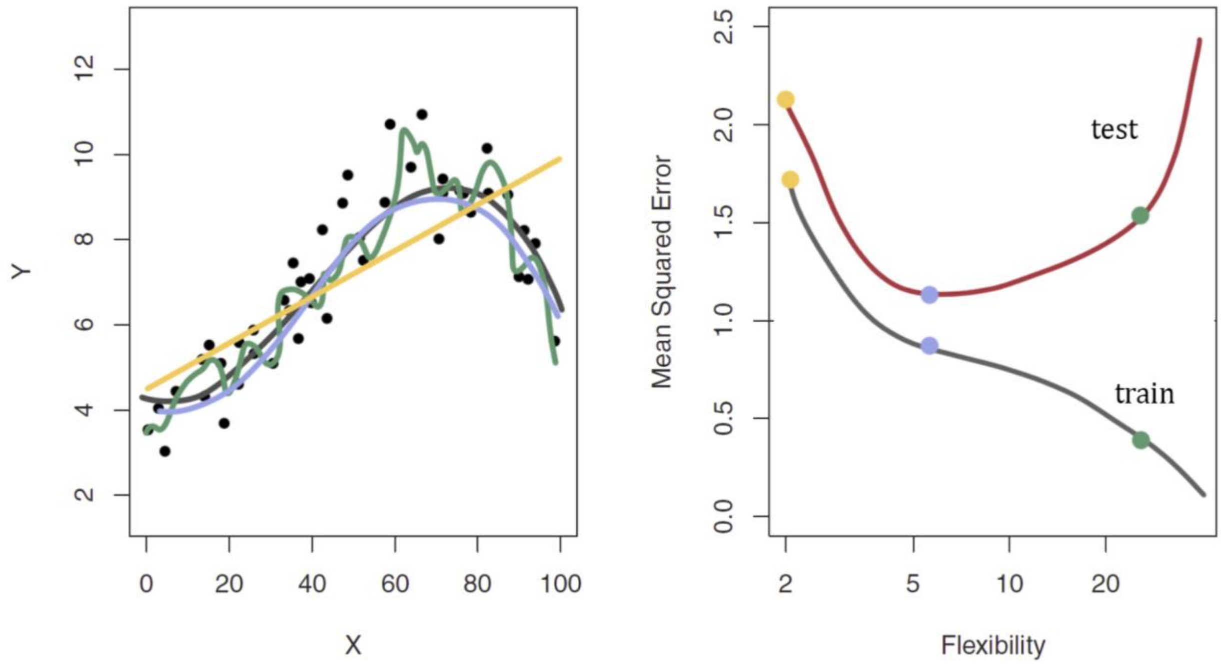
$$J_{max} = \max_{t}[sensitivity(c) + specificity(c) - 1] \qquad \textbf{V.III}$$

B. Overfitting

The goal of a classification model is to form a generalization from a data set of labelled training instances such that the prediction accuracy for unobserved instances is maximized. However, literature and studies show that using the same data to both train and evaluate the performance can be misleading for this purpose as it typically leads to *overfitting* (Gareth, Witten, Hastie, & Tibshirani, 2015). A model is said to be overfit when it is tuned too finely to the noise present in the training set and unable to generalize to new observations. In other words, the given model yields small error calculated by the training set (train error), but large error when calculated using new data (test error).

An example adapted from Gareth et al. (2015) presented here in Figure V.I perfectly illustrates the problem of overfitting. The data points are simulated from the function f given in black with added white noise. We have three competing models which approximates f named after their color: $green$, $yellow$, $blue$. The model which performs best in terms of training error is $green$. From panel A (left), we see that $green$ is complex and provides a good fit to the training data, however it does not approximate f well and will provide poor predictions on new observations – the training error is small, but the test error is large. The $green$ model is overfit to the data. On the other side of the spectrum is $yellow$, where the model overgeneralizes to a linear function when the true underlying relationship is not linear. This model is underfit to the data. Finally, the goldilocks solution is $blue$ which best approximates f and correspondingly has the lowest test error and a proportionate training error.

Figure V.I An example of overfitting and underfitting



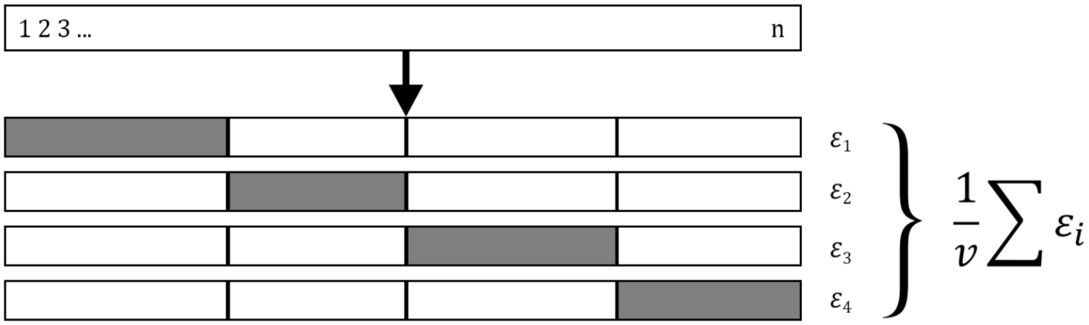
In general, increased model complexity yields increased chance of overfitting (Hastie, Tibshirani, & Friedman, 2008). As we will be using a particularly complex algorithm with hundreds of variables, we should be employing methods to avoid overfitting.

C. Cross validation

To avoid overfitting the data, cross-validation (CV) can be used. Cross validation is one of the most widely used methods for estimating prediction error (Hastie, Tibshirani, & Friedman, 2008), of which v-fold cv is probably the most common method. In V-fold CV, out of sample error is estimated by repeatedly resampling training data into different groups for fitting the model and testing the model. More specifically, in V-fold CV, the data is split into v equal-sized folds. The model is fit on the $v - 1$ folds, and the fold-specific error ε is calculated using the held-out fold. The procedure is repeated v times, where each time a different fold is held-out. A visualization of the V-fold CV process is provided in Figure V.II.

Finally, the out of sample error, or the cross-validated risk (Vapinik, 2000), is estimated by averaging the fold-specific error from each iteration. When comparing models, the best model is the one that minimizes this cross-validated error. The cv risk is also a good estimate for the prediction error and can be used to interpret the accuracy of the model. In this thesis, however, cross validation will only be used to compare the performance of models within the ensemble. To evaluate the accuracy of the final ensemble model we will use a randomly sampled independent test set.

Figure V.II Diagrammatic representation of 4-fold cross validation:
On the left we see the four iterations of the cv where the data is split into four folds in each iteration. Each time a model is refit the held-out fold is different, until all the folds have been used as the hold-out fold. In grey are the held-out folds used create calculate the iterations error rate, in white are the folds used to fit the model, and on the right are the error rates produced in each fold. The error rates are then summed, and the sum is divided by v to find the cross-validated risk.

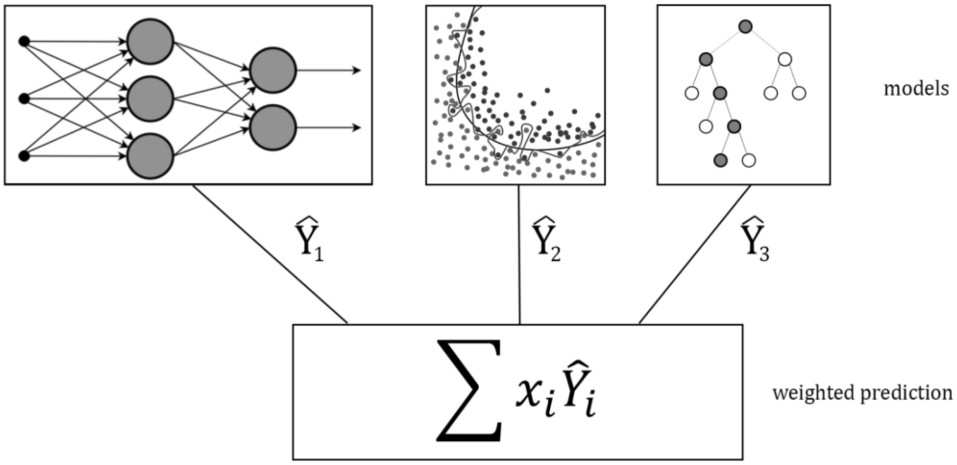


Choosing the number of folds to be used in V-fold CV is a question of computational feasibility, and a bias-variance trade-off. With v equal to n, the computational burden would be significant, and the prediction error estimate may be subject to high variance as the data used to train every model would be close to identical to one another. Using only 2 folds would be computationally inexpensive, but the error may be biased as the number of observations used to train the model is limited. In our data set we have $> 10\ 000,$ observations, and so a small number of v may still not lead to a biased estimate of prediction error. In addition, the large number of observations makes a large number for v particularly computationally expensive. Generally, a number between five and ten folds is recommended (Gareth, Witten, Hastie, & Tibshirani, 2015).

D. Ensemble learning

Many algorithms can be used to train a prediction model, however for any given set of data it is not known what algorithms yields the best model. To overcome the challenge of arbitrarily selecting one algorithm, and selecting away other algorithms, ensemble learning can be employed. Ensemble Learning is the method of combining the information from many models by averaging or weighing the numerical predictions of each model or using the most common observations between models (Gremmell, 2018). Studies have shown that ensemble learning often performs better than any individual model (Polley & van der Laan, 2010), and it works especially well in the cases where there is disagreement between the models. The improved predictive qualities often come at the expense of interpretability as the model gets increasingly complex. However, model interpretability is not important for the analysis conducted in this paper.

Figure V.III An illustration of a weighted ensemble model



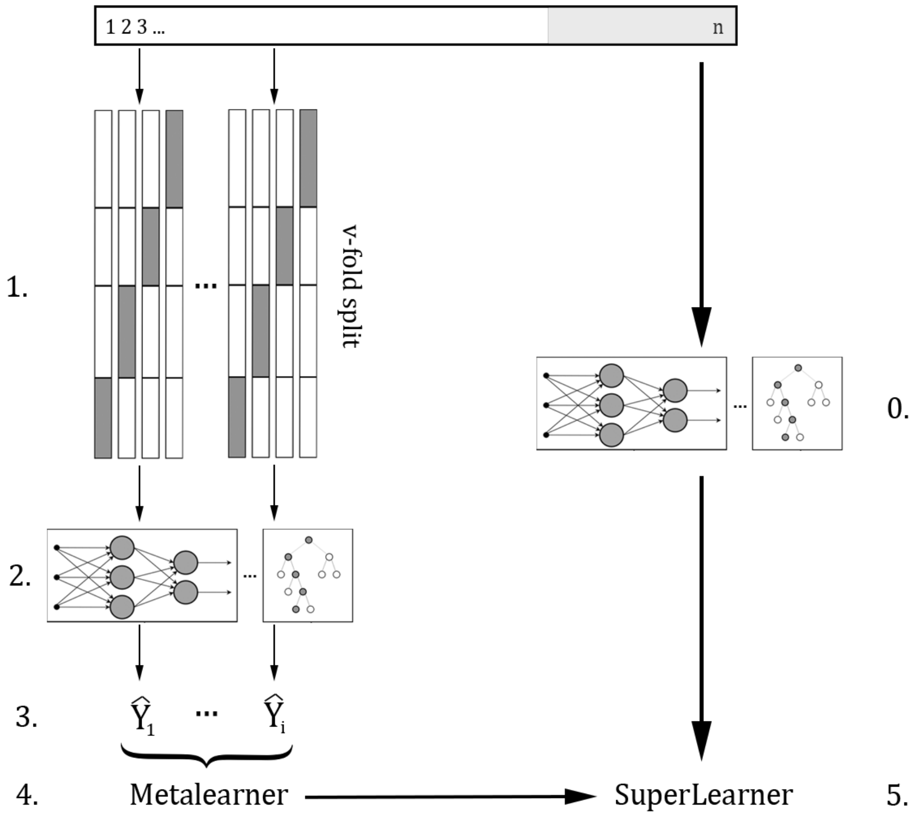
An ensemble model can be created manually by fitting multiple models and combining the results and resampling techniques such as cross validation can be used to find the optimal combination weigh. In this paper, however, we will be creating the ensemble model automatically using a method called Super Learner.

E. Super Learner

Super Learner (SL) is an automation method for finding efficient weights for an ensemble of algorithms, as well as removing models that do not improve predictive power. The SL algorithm is shown to be an *asymptotically optimal framework,* meaning for large inputs it performs at worst a constant factor worse than the best possible algorithm. Furthermore, SL has been found to be robust even in cases of small datasets (Polley & van der Laan, 2010), and over-fitting is controlled for even when the number of algorithms used in the ensemble is large. Moreover, SL is free and programmatically easy to use.

To generate predictions on data, the Super Learner algorithm goes through five processes. First, (0) the Super Learner fits all the candidate learners on the full data provided. The (1) data is then split into v folds for cross validation, and each candidate learner is refit v times for v iterations as described in section V.C (CV). Predictions are generated and stored for each fold (2). The predictions from the v folds are then stacked (3) and passed to a meta-learning algorithm (4) which is used to find the optimal weigh of each learner that minimizes the cross validated risk associated with our loss function of interest. Finally, (5) the Super Learner predictions for the full data set is created by using the weights from step 1-4 on the predictions from step 0. The Super Learner algorithm is visualized in Figure V.IV.

Figure V.IV Flow Diagram for Super Learner (adapted from "Super Learner", by E.C. Polley 2010, p. 59)



We specify three inputs: 1) The algorithms (learners), 2) a meta learner to be used for weighing the candidate learners, and 3) the number of folds to be used for V-fold cross validation. At the time of writing, Super Learner includes forty-two prediction algorithms. Polley & van der Laan (2010) recommends limiting the number algorithms, and to choose candidates based on diversity in functions. The greater the diversity of methods, the greater the ability of the ensemble to approximate the true prediction function. For our study, we choose ten algorithms diverse algorithms, shown in Table V.II. Next, we specify AUC as the metric to be maximized by the meta learner. AUC is often the metric of choice for binary classification problems, and AUC has also been empirically shown to lead to high performing models for classification problems in Super Learner (LeDell, Laan, & Peterson, 2016). Finally, we decide on ten as the number of folds to be used in cross-validation. Ten folds provides a good compromise between the bias-variance trade-off for large sample sizes, and is recommended in literature (Gareth et.al, 2015; Kuhn & Johnson, 2013).

Table V.I Candidate learners

Learner	**Description**
Extratrees	extra Trees
mean	arithmetic mean
knn	k-nearest neighbor
bayesGLM	bayesian generalized linear model
glmnet	elastic net
xgboost	extreme gradient boost
ranger	ranger (fast random forest)
ksvm	kernel support vector machine
ipredbagg	bagging for classification
rpart	recursive partitioning and regression trees
nnet	neural network

In addition to finding the optimal weight of candidate learners, we can also use Super Learner to tune hyperparameters of our algorithms. We tune hyper parameters for the two tree methods (*extratrees* and *ranger*), and extreme gradient boost. To optimize model hyperparameters, we create different variants of each model with customized hyperparameters. For the tree methods we specify 2^5 configurations using different values for maximum leaf nodes and the number of features that are randomly chosen within each tree node. For *xgboost* we configurate 3^3 models using the three hyperparameters: maximum number of trees, maximum depth, and the shrinkage.

In effect we have 99[5] candidate models although we only have 11 candidate algorithms. We optimize hyperparameters in isolation due to computational restraints and use the optimal parameters for the models in the final consideration of the ensemble.

To evaluate the ensemble model accuracy, we use a randomly sampled test set independent of the data used in the Super Learner model. In this way we can achieve an unbiased evaluation of the final model fit. In practice, there is no general rule on how to choose the size of the training and test partitions (Hastie, Tibshirani, & Friedman, 2008), however it is typical to use 2/4 of the data for training, ¼ for validation and ¼ for testing. As such, we randomly sample and pass ¾ of the data to Super Learner to be used for training the ensemble model and use the remaining ¼ for testing.

F. Estimating witness error

We train the Super Learner algorithm to solve our classification problem. That is, we train the model to estimate the relationship between a criminal's race, Y and characteristics of the crime and the victim x, given by function $f(x)$, subject to error ε. The model finds the expression $\hat{f}(x)$ that approximates the true function and produces as similar outputs as possible to what we observe for race given the predictors x. The estimated relationship is,

$$Y \approx \hat{f}(x) + \varepsilon \tag{V.IV}$$

In contrast, witness guesses can be thought of as correct, but subject to error from estimator variables and system variables and the same white noise. We define witness guesses as $\tilde{Y}$,

$$\tilde{Y} = Y + \delta, \tag{V.V}$$

where, δ is the sum of the effect of the variables that negatively impact eyewitness identification, or simply, the error associated with the witness guess. As mentioned in Section IV.B, we can then isolate witness error and model error by averaging difference between the model predictions and the witness guesses.

$$\frac{1}{k}\sum \widehat{Y_k} - \widetilde{Y_k} = \bar{\delta} + \bar{\varepsilon} \tag{V.VI}$$

[5] (trees) $2^5 + 2^5 +$ (xgboost) $3^3 +$ (remaining models) $11 - 3 = \ 67$

The error could be small or large, depending on how well the model predicts the race of an offender. It is likely that the error term will not be zero, as the characteristics of the crime and the victim are not sufficient to perfectly predict the race of the offender. To correct the discrepancy, we subtract the estimated out-of-bag error rate $\hat{\varepsilon}$ to find the estimated witness error. We define the estimated witness error $\hat{\delta}$ as,

$$\hat{\delta} = \bar{\delta} + \bar{\varepsilon} - \hat{\varepsilon} \qquad \text{V.VII}$$

We will use $1 - ACC$ (from the confusion matrix generated from the hold-out set [Table VI.II]) as $\hat{\varepsilon}$ instead of AUC, as additional error may be introduced in converting continuous prediction to discrete predictions.

VI Results

In this section, we present the result from building the model and comparing model predictions to witness guesses. First, we present baseline results and performance of the model on arrest incidents. Second, we present the results from comparing the predictions generated by the model to witness guesses and investigate when the model and witnesses disagrees. Finally, we briefly present results on model generalization to other characteristics of an offender and robustness checks.

A. Baseline Results

Using the described data and method we created a prediction model for race of an offender. The response variable used for the model was the dummy variable for offender being white, with the discrete prediction being 0 if the offender is classified as black and 1 if the offender is classified as white. The optimal combination of candidate learners, that is, the composition of our ensemble model, is shown in Table VI.I. Out of the 11 candidate learners, five are used. The most important model is *ranger*, with over half of the weight (0.5529), followed by *extratrees* (0.21). The *nnet* model and a simple mean are tied for third with weights 0.1051, and a small contribution also comes from the extreme gradient boost algorithm with a weight of 0.0267.

Table VI.I Ensemble model composition

Learner	CV-risk	Coefficient	Used
extratrees	0.0614	0.2100	Yes
mean	0.5154	0.1051	Yes
knn	0.2073	0.0000	No
bayesGLM	0.1172	0.0000	No
glmnet	0.1174	0.0000	No
xgboost	0.5592	0.0267	No
ranger	0.3613	0.5529	Yes
ksvm	0.3254	0.0000	No
ipredbagg	0.1994	0.0000	No
rpart	0.2233	0.0000	No
nnet	0.5154	0.1051	Yes

We measure the performance of the ensemble by calculating the AUC statistic. From Figure VI.I we can see that the AUC for the model is 0.96 suggesting that the model is excellent at distinguishing between the black and white offenders. The probabilistic predictions are turned into class predictions using a cut-off of 0.617. The cut-off was decided on based on optimal Youden index, as sensitivity and specificity are equally important.

Figure VI.I Model AUC and Cut-off

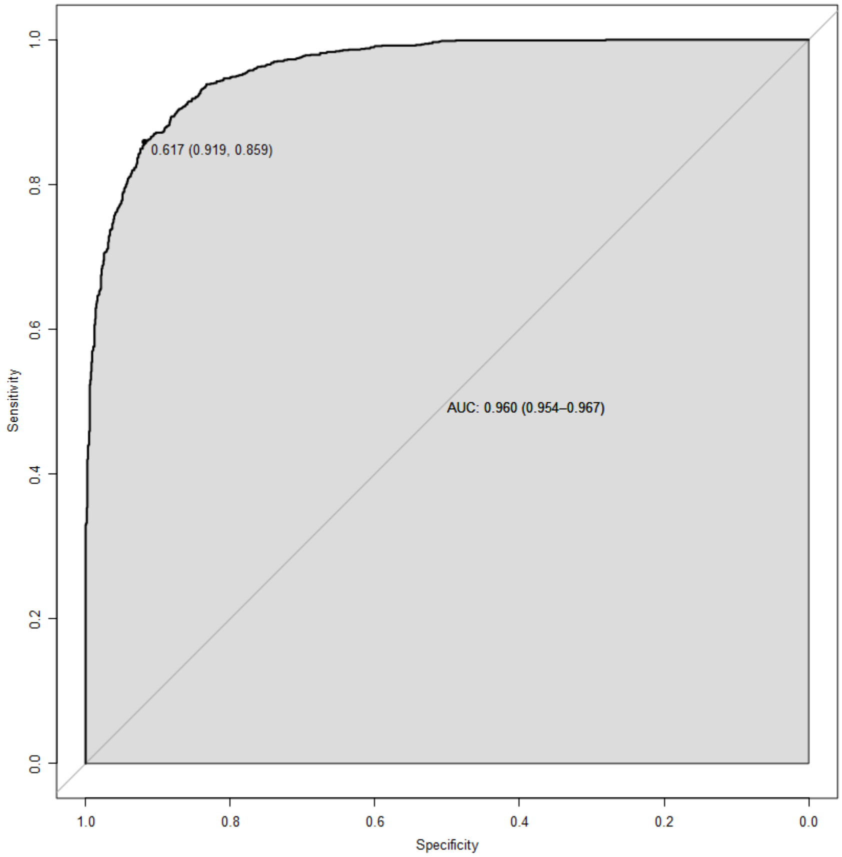


Table VI.III provides an overview of the class predictions and the true labels from the held-out test set. In accordance with a large AUC statistic, we find that the classes separate well and there are few false negatives and false positives. The overall accuracy of the model is 0.9133, seen in Panel B. This implies that for a hypothetical CAP incident with a property crime component, the probability that the offender will be correctly labelled by the model is 91.33%. The confidence interval for the accuracy at a 5% confidence level is $(0.9011, 0.9245)$, and so we can expect the accuracy of the model to be this interval in 95% of the cases should the model be retrained. The model significantly outperforms a simple average prediction as the no information rate is about 50%. The sensitivity, or the true positive rate, is 89.06% meaning that the white offenders are correctly labelled as white 89.06% of the time. As additional evidence that the accuracy is not an artifact of the sample, we use nested cross validation to find another estimate of the accuracy and AUC. We find that the accuracy from the nested CV is consistent with those from the hold-out approach.

Table VI.II Confusion matrix and accuracy metrics

Panel A: Confusion Matrix

Correct responses are marked in blue and incorrect responses are marked in orange.

		Observed	
		Black	White
Predicted	Black	1018	76
	White	125	1100

Panel B: Accuracy metrics

Y	**ACC**	**Sensitivity**	**Specificity**	**NIR**
Race	0.9133	0.8906	0.9354	0.5071

B. Variable Importance

As with other models of high complexity, the Super Learner model presented in this thesis does not output interpretable results. The Super Learner package does not provide any way to chart variable importance either. Breiman (2001) suggests that a permutation method can be used to assert variable importance in these cases. In particular, the importance of a predictor can be measured by permuting its values in the training data and observe the drop in some performance metric (Greenwell & Boehmke, 2020). As with the original model we use AUC as the performance metric. The importance is then defined to be the decrease in AUC when the feature is randomly shuffled. It should be noted that the permutation method for variable importance can be misleading, especially in cases of multicollinearity between features (Hooker & Mentch, 2019). Our data does not have particularly high levels of multicollinearity, with 99% of the features having a correlation coefficient of less than 0.1. The highest correlation coefficient (0.74) is between the feature for *Personal Weapons* (hands, feet, etc.) and *Simple Assault*. As such we deem the results from permutation to be reliable. The permutation method introduces randomness and therefore should be run more than once and averaged. We permute each feature five times using the held-out testing set and average the results. We use the testing set to highlight which features contribute to the generalization power of the model. In Table VI.IV we show the top 15 most important variables from this method.

Table VI.III Variable importance

#	Predictor	Importance
1	$Victim_{black}$	0.036
2	$Victim_{white}$	0.013
3	$Location_{Residence}$	0.013
4	$Victim\ segments$	0.012
5	$Agency_{TNMPD0000}$	0.011
6	$Offender\ segments$	0.010
7	$Date$	0.010
8	$Offense\_segments$	0.009
9	$Agency_{TN0190100}$	0.008
10	$Victim_{age}$	0.008
11	$Victim_{male}$	0.008
12	$Weapon_{Handgun}$	0.007
13	$Victim_{resident}$	0.006
14	$Location_{Department\ store/Discount\ store}$	0.005
15	$Location_{Highway/Road/Alley/Street/Sidewalk}$	0.005

Together the fifteen variables give a holistic view of the drivers of the racial model and how it distinguishes between offenders of different races. Five of the variables pertain to the victim, five to the location (two are agency identifiers and three are crime location), and four are administrative variables, more specifically: the date, and number of offenders, victim and offense segments recorded for the incident. The last remaining variable is the weapon identifier for handguns. Even though the model predicts race quite well, the importance of even the most important variables is relatively small. The most important variable attributes 0.036 AUC out of the total 0.96 AUC for the final model. This means that each variable included in the model contributes a small amount to the result.

The characteristics of the victim seem to be the best predictors for the offender’s race. On aggregate the victim variables among the most important variables have a score of 0.071, that is about 7% of the total variation the importance may be different if we looked at change in model performance when the variables were removed together rather than in isolation.

The most important variable, is the dummy for the victim being black, followed by the dummy for the victim being white. One explanation for this may be that the American cities is very segregated (Frey, 2015). This means that the victim and offender are likely to be of the same race simply by virtue of where they live. Dummies for victims of different races than black or white are absent from the most important variables, supporting this explanation. It would be interesting to see if the highest performing predictors included other races of the victim if a multinomial analysis of race were used.

Although no one variable is very important, we were surprised to find that date ranked among the most important variables. Econometrically speaking, date is often used as a proxy for an immeasurable variable, or simply data you cannot easily obtain, but that is correlated with time. Because the time frame of the data was limited to one year, we did not expect there to be significant time variant effects, but the importance of date suggests otherwise.

C. Arrests vs Witness incidents

In Table VI.V we report the results from comparing the prediction model to the witness guesses. We find clear evidence of witness error. Row 1 of Table VI.V shows that the witness error is 16.17%. This implies that on average 16.17 % of witness reports misclassifies white offenders as black or conversely black offenders as white. Using a paired t-test we find that the difference in accuracies between the arrest and witness accuracies are statistically significant at a 1% level. We note that normally, using a paired t-test to test for significant differences between two classifiers can be fallacious as the assumption of independence between the samples is violated (Diettrich, 1998). In our case, witness data has not been used to train the model and so independence between the two samples from which the accuracy is derived should not be violated. The magnitude of the error is not in line with our expectations. From our prediction in section IV.B we expected that the witness error would be small but significant. The error is especially big as we consider only one facet for which the report could be erroneous.

Out of all the witness cases, the model predicted that 79% of the crimes were committed by black offenders, and 21% were committed by white offenders. The witness guessed proportions were 75% black and 25% white. This indicates that witnesses more often mislabel black offenders as white. In congruence, the highest amount of disagreement between model and witness guesses is when the model predicts that the offender is black, but the witness has labelled the offender as white. There are 523 cases for which the model predicts black, and the witness has labelled the offender as white, and only 182 cases for which the model predicts that the witness is white, and the witness has labelled the offender as black.

When separated by race of the victim, we find that white victims identify offenders as black in 59% of the cases, and as white in 41% of the cases. Black victims identify offenders as black in 92% of the cases, and white in only 8% of the cases. In contrast, the model predictions stay consistent before and after grouping by the race of victims. By model predictions, whites are victims to black offenders in 78% of cases, and victim to white offenders in 22% of cases. Black victims are associated with 80% black offenders and 20% white offenders. In other words, black victims tend to overreport the perpetrator's race as black, and white victims tend overreport offenders' race as white.

To further break down disagreement between model predictions and witness guesses, we split the results into five groups by factors that may impact the witness's ability to recognize an offender. The first three groups target variables that has been proven to negatively affect eyewitness identification. Namely, the groups target estimator variables for conditions that affect visibility, presence of a threat (weapon) and common or different race or ethnicity (cross-race bias) on witness testimony. In addition, we investigate the groups for age as research has found that accuracy can be lower for children (Shapiro & Penrod 1986; Parker et al., 1986) and accuracy is consistently lower for elderly witnesses (Memon et al., 2003; Wilcock et al., 2007). Finally, we split the predictions by the gender of victims, as some research has found differences in accuracy between genders (Areh, 2011). However, for the gender effect there are large amounts of inconsistency in studies (Horvath, 2009).

In row 2-3, we separate incidents by day and night. We use this distinction as a proxy for variables that affect visibility, such as sunlight; a witness will have a harder time evaluating the race of a criminal as daylight fades. We define day as the hours between sunrise and sunset, and conversely night as the time between sunset and sunrise. Sunrise and sunset are calculated based on the state geodata and the date of the incident. The witness data has a slightly larger proportion of incidents happening at night (63%) than the arrest data (56%). However, the witness error is seemingly unaffected by the day and night distinction.

Row 4-5 shows the difference when incidents are separated into groups where victims and the offender are of the same races or different races. This is motivated by research on cross-race bias, a reliable phenomenon across racial groups where unfamiliar faces from other races are misremembered more often than own-race faces. For this split of the data, we find that the witness error is substantially different and lower for same races than for different races. This suggests that witnesses are more accurate when identifying offenders of the same race as themselves. However, the accuracy of the model also decreases significantly when subject to the race data split, which is problematic as the difference between noise and witness error becomes less discernible. Witness error is bigger by 19 percentage points for same races than for different races, and the model accuracy decreases by 14 percentage points. As the witness error increases more than model accuracy decreases, we interpret the increase in witness error to be in accordance with the notion of cross-race bias.

Row 6-7 shows the witness error when we separate crimes into weapon or no weapon, where no weapon also includes personal weapons such as fists. A multitude of laboratory experimental studies have shown that when a weapon is involved in an incident, a witness is less likely to remember the face or other characteristics of the offender (National Research Council, 2014). This effect, known as weapon focus, is perhaps the best-known error in witness testimony (Horvath, 2009). However, the results do not generalize well to actual incidents (Pike et al., 2002). As with day and night there seems to be no difference in witness error between cases of weapons and no weapons in our data. This is consistent with findings that weapon focus does not significantly impair accuracy in actual crimes, and that weapon focus is more pronounced in laboratory experiments (Fawcett et al., 2013).

In row 8-10 we separate data by ages of the victim. We define a minor as below 14 years old, an adult as between 14 and 64 years old, and an elderly as above 64 years old. Surprisingly, we see that the witness error is significantly smaller in cases where the victim was a child. The witness error is also lower in cases where the victim is elderly. This is contradictory to research findings, that witness testimonies from elderly are consistently less accurate than those of younger witnesses. However, the small sample sizes of minor and elderly witnesses may have biased the results.

Lastly, we present the results split by gender in row 11 and 12. The difference between the genders is small, with females providing slightly better predictions than men. Most research in this area has either found that females have higher accuracy or that there is no difference between the genders (Horvath, 2009; Areh, 2011).

Table VI.IV Model vs. Witness guesses in full data and different subsets

	Y	α	ω	$error$	n_α	n_ω
1	race	0.9133	0.7516	0.1617	2319	2319
2	day	0.9061	0.7399	0.1622	996	929
3	night	0.9196	0.7586	0.161	1323	1390
4	same race	0.9526	0.8609	**0.0917**	1598	1548
5	different races	0.8187	0.5359	0.2828	721	771
6	weapon	0.9261	0.7395	0.1866	1391	1511
7	no weapon	0.8954	0.7054	0.19	928	808
8	Minor	0.9167	0.8063	0.1104	228	191
9	Adult	0.9193	0.6877	0.2316	2032	2081
10	Elderly	0.9278	0.7143	0.2135	97	70
11	Male	0.9088	0.6802	0.2286	1447	1310
12	Female	0.9232	0.7166	0.2066	872	1009

D. Generalization

The goal of this thesis is to answer the general applicability of machine learning in correcting witness reports. As such, we are interested in how the results from the racial model generalize to other characteristics of an offender. We have access to two other characteristics of offenders from the data, namely gender and age. We estimate two new models using gender and age as the target variables and compare the model accuracies to the witnesses guesses for gender and age. The same 11 algorithms are used, albeit hyper parameter optimization is not used due to parsimony. For the age model, squared error is used as the loss function, as the dependent variable is continuous, and we present the 1 – MAPE calculated using the held out set as the measure of accuracy. We exclude 0-values of age, for the same reasons that we omit incidents for which offender's race and gender is unknown. In row 1 of Table VI.VI, we see the results of the racial model which here serves as our benchmark for comparison. In row 2, we see the difference for the gender model. The gender model suffers substantially in accuracy, suggesting the information contained in an incident report is not as well suited to predict gender as it is suited to predict race. The witness error is slightly lower (14.92%), meaning witnesses misclassify gender more infrequently than they misclassify the race of an offender. The witness error for gender is also statistically significant when a paired t-test is used. In row 3, we show the accuracies and witness error for the age model. As with the gender model, the accuracy of the age model is worse than the race model. The witness error is less than it is for both gender and race, but still large at 13.25%.

In row 4, we present a robustness check of the models. We build a classifier to distinguish between the sex of victims in arrests. For this variable, there should be no difference between the arrest and witness data as the variable is known in both data sets – the sex of the victim is certain even at the time of reporting. We find that there is no significant difference between the model predictions and the sex of the victim reported by witnesses.

Table VI.V Accuracies of model and witness predictions and estimated witness error

$\boldsymbol{Y}$	**Arrests** $(\boldsymbol{\alpha})$	**Witnesses** $(\boldsymbol{\omega})$	**Error** $(\boldsymbol{\alpha - \omega})$
Race	0.9138	0.7516	0.1622
Gender	0.8344	0.6852	0.1492
Age	0.8618	0.7293	0.1325
$\mathbf{Sex_{victim}}$	0.6224	0.5886	0.0338

The results indicate that machine learning may be used to discover witness error for several characteristics of an offender. Combining the results from multiple models or creating a multivariate response model for overall witness error seems like a possibility to consider in future research.

E. Predicting the race of unknown perpetrators

In the data section (III.D) we mentioned that races for some offenders were labelled as unknown or are missing. In our data set, there are 3256 non-arrest incidents for which the other data is intact, but the race is unknown. Using the race model, we predict the races of these offenders. We find that the 2683 (82%) of the offenders are black, and 573 (18%) are white. This differs from the proportions in the arrest data, for which the crimes are evenly distributed among the races. However, the proportions are about the same as in labelled witness incidents. Given that our model is correct, the proportions for all incidents are 66% black and 34% white, as opposed to the proportions of arrests which are 50% black and 50% white. In other words, most offenders in CAP crimes with property crime component are black, but it is not reflected in arrests being made. The implication being that white offenders are more likely to get arrested than black offenders. The findings are summarized in Table VI.VI.

Table VI.VI Complete crime[1] incidents by race:
Arrests are taken as truth, and labels are predicted by the model for witness data. The proportions are given in parenthesis next to the number.

Y	Arrests	$\text{Witness}_{\text{guessed}}$	$\text{Witness}_{\text{unknown}}$	All incidents
Black	5771 (0.50)	12487 (0.75)	2683 (0.82)	20941 (0.66)
White	5828 (0.50)	4245 (0.25)	573 (0.18)	10646 (0.34)

[1] CAP crimes with additional property theft

F. True crime

We have assumed that arrest represents true crime, and witness guesses are erroneous. It would be interesting to see how our results shift, if we instead assume that witness labels reflect true crime and arrests are erroneous. Although witness testimony is erroneous, studies also show that law enforcement treat offenders differently according to their race (Roland & Fryer, 2019). As a tangential analysis, we flip the assumption that arrests reflect true crime, such that arrests are assumed erroneous and witness incidents represents true crime. Based on the flipped assumption, the disagreement between model predictions and witness labels can be viewed as evidence for erroneous arrests. A model trained on witness incidents then reveal that 30% of arrests are erroneous. We also find that the proportions of the races even more shifted towards black offenders for all incidents, with 78% of incidents being committed by black offenders and 22% by white offenders. The implication remains that white offenders are more likely to get arrested than black offenders.

Table VI.VII Complete crime[1] incidents by race with switched assumptions
Witness labels are taken as truth, and labels are predicted by the model for arrest data and imputed for unknown offenders in witness data. The proportions are given in parenthesis next to the number.

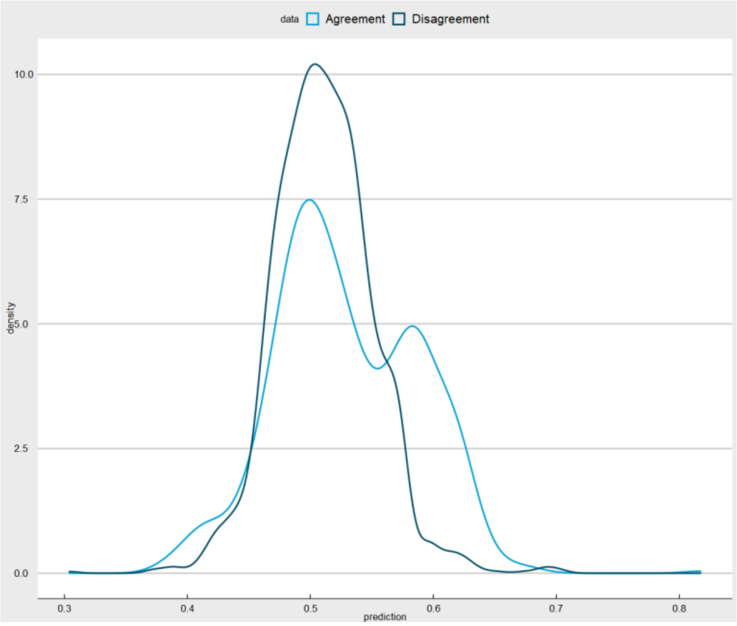
Y	**Arrests**	$\mathbf{Witness_{guessed}}$	$\mathbf{Witness_{unknown}}$	**All incidents**
Black	9136 (0.79)	12487 (0.75)	3138 (0.96)	24788 (0.78)
White	2436 (0.21)	4245 (0.25)	118 (0.04)	6799 (0.22)

[1] CAP crimes with additional property theft

G. Degree of disagreement

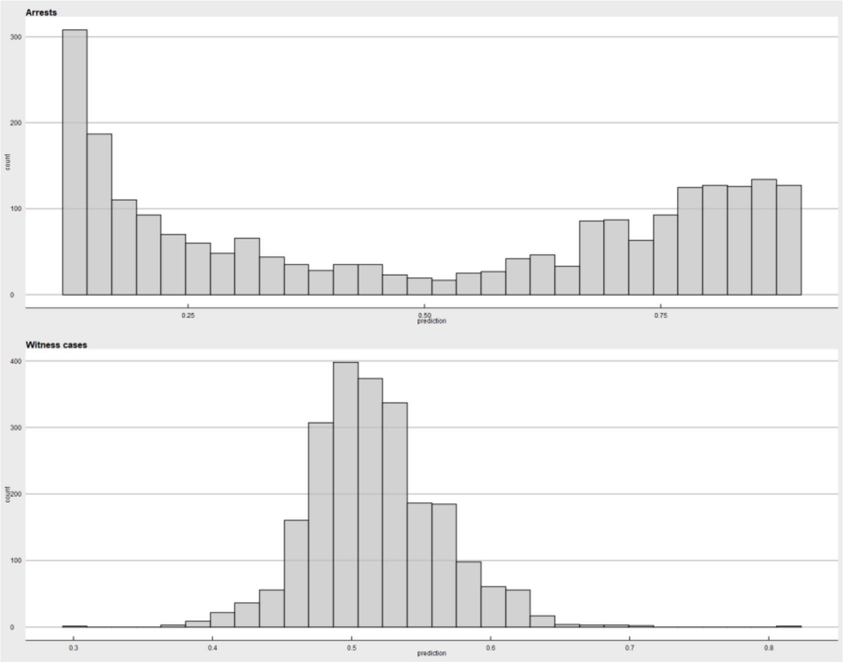
To understand the degree of disagreement within the estimated witness error, or the certainty of the estimated witness error, we look at how agreement and disagreement between the model and witness guesses are distributed. Most of the agreements and disagreements between model and witness guesses happens around a continuous prediction of 0.5. Interpreted as a probability, this is where the model predicts that the offender is 50% likely to be white, and conversely 50% likely to be black. In other words, around the 0.5 point is where the model is the most uncertain of the offender's race. As such, we expect disagreements to be centred around 0.5. However, agreement should be highest for the cases closer to 1 and 0. As agreements are also centred around 0.5, albeit with longer tails, most cases are marginally close to being labelled differently and categorized as correct labels as opposed to erroneous labels.

Figure VI.II Disagreement and agreement plot



In Figure VI.III we present two panels that contextualizes the results from Figure VI.II. Both panels provide the count of the continuous predictions for arrests and for witness data. Arrests are shown in the top panel, and witness case predictions are shown in the bottom panel. From the arrest predictions, we see that the model produces mostly high or low predictions, with few predictions in the grey area around 0.5. The distribution resembles an inverse bell-curve. In contrast, the predictions for witness cases take on a bell-curved form, where the plurality of predictions are around 0.5. This indicates that witness cases, that is non-arrests, are cases where the characteristic of the offender are more uncertain than for arrests cases. This is also reflected in the reported age of the offenders in witness cases. Age is often reported as 0 meaning the age of the offender is completely unknown.

Figure VI.III Continuous predictions from arrests and witness cases



H. Distribution of arrests and witness incidents

A possible concern is that the witness and arrest joint distributions are different. That is,

$$P_{arrest}(X) \neq P_{witness}(X)$$

We assumed that witness and arrest incidents had identical distributions in section IV.B. If the distributions are identical, we assume that it is randomness that separate arrests and non-arrests. In other words, that police catch some criminals and do not catch other criminals due to randomness. However, it may be that the offenders who get away with crime, commit crimes in different ways. Perhaps they are better at concealing themselves, target less populated areas, or are better at charting escape plans. If non-arrestees are different from arrestees, the estimated relationship between the predictors and the response variable is wrong. That is, the function we estimated to find the relationship between the characteristics of the crime and the criminal's race does not apply to non-arrestees. It follows that the disagreement between predictions and witness guesses cannot be asserted as witness error.

Interestingly, we can use the same methods that we used to build the racial model to test the assumption that the distributions are equal (Mu, Ding, & Tao, 2013). Meaning, we can generate an ensemble model for separating the data. If the classifier is successful in separating the data, we have found evidence that the joint distribution of our data sets may be different. To train the model we use the candidate learners from Table V.II, trained on the merged held-out arrest set and the witness data. In the merged data set we create a dummy variable indicating which data set the observation originated from. The dummy variable takes on a value of 1 if the observations is an arrest and 0 if it is a witness incident. We use the dummy variable as the target variable. Witness incidents are more likely to have missing observations of features (e.g., unknown age of the victim) than the arrest data, which could create a pattern of differentiation between the two data sets. Therefore, we remove all observations which have a missing feature from both sets.

We find that the classifier can separate between arrests and witness incidents with a 97% accuracy, where the no information rate is 64%. The accuracy is very high, indicating that arrests and witness incidents joint distribution is different and that our assumption of identical distribution is violated.

That the classifier can separate arrests from non-arrests well does not necessarily invalidate our inferences about error; the data can be separable and still generalizable between the two groups. For instance, non-arrest crimes are more likely to occur in the later part of the year – there is less time for the police to catch the criminal before reporting. This factor contributes to separating the data but does not necessarily impact the relationship between the crime and the criminals. In separating arrests from non-arrests, the classifier finds that some variables are especially important, namely the number of offenders, victims and offenses recorded in the incident. For non-arrests, there are fewer offenders, offenses, and fewer victims per incident. In addition, property loss tends to be higher, although property theft is less frequent for non-arrest and property seized is more frequent. Furthermore, there proportion of incidents happening at night are bigger for non-arrests than for arrests. To test if these differences drive the performance of the classifier, such that crimes are otherwise similar between arrests and non-arrest, we train the model again and omit these variables. We find that the accuracy of the classifier, even when the top 10 most important variables from the original classifier are omitted, separate between arrests and non-arrests with an 86% accuracy. This is evidence that many features are distributed differently in arrests than in non-arrests, and that criminals may be different between the groups. As such, the estimates should be verified in reproduced research before applied.

VII Estimating the cost of witness error

To understand the implication witness error has on society, and how it should be prioritized in policy, cost should be considered. The average incident cost for a crime should reflect what the government is willing to pay for resolving a crime. If the crime goes unresolved the resources are wasted. In addition, an erroneous report may result in a wrongful arrest and even a wrongful conviction. However, as we cannot, using the Super Learner model and our data, link the witness-error directly to the number of wrongful arrests, we reserve our estimates to the cost of police inability to find a perpetrator. We estimate the police cost associated with erroneous reports to be,

$$TC = \sum n_i \varepsilon \times j_i, \qquad \textbf{VII.I}$$

where, TC is the total police cost, n_i is the number of incidents for crime i where witnesses are a primary source of evidence, ε is the rate of witness error, j is the cost of policing cost for crime i. In other words, we estimate the cost of pursuing a mislabeled offender.

We estimate the combined yearly cost for five crimes: burglary, rape, assault[6], robbery, and homicide. These are representative of the crimes we used to build our estimate. Estimates suggest that eyewitness cases - cases in which the only critical evidence were eyewitnesses - constitute about 3% of yearly felony cases in the US (Goldstein, Chance, & Schneller, 1989). Farrington and Lambert (1993) found that eyewitness descriptions led to arrests in 2-15% of burglary and violence cases in England. We use a rate of 3% and multiply the rate with numbers of arrest by crime (N) to find the number of eyewitness cases (n) by crime. We use data from the Federal Bureau of Investigation to find the reported number of arrests by crime (FBI, 2021). The most recent statistics are from 2016.

We multiply the eyewitness cases by error rate to find the expected number of incidents where police pursue an erroneous report. The inferred erroneous pursue rate is 0.4%[7]. In the

[6] Assault includes aggravated assault, simple assault, and intimidation.

[7] $3\% \times 16\%$

absence of comparable rates in literature, we use the rate of wrongful convictions as a comparison. This is because wrongful convictions are often due to erroneous reports (Horvath, 2009). Compared to rates from other papers, our estimate of erroneous pursue rate is conservative. According to the Innocence Project, the wrongful conviction rate is between 2.3% and 5%, and a paper by Samuel Gross et. al. (2014) made what they believed to be a conservative claim of 4.1%.

To estimate the unit cost of investigating a crime, we average the inflation adjusted estimates from two papers. First, is a well-cited report by Miller et al., (1996) published by the National Institute of Justice. In the paper by Miller et al., (1996) police costs were derived from surveys and published statistics on the cost of police and emergency response. Second, we use numbers from a research reports by Heeks et al., (2018)[8] published by the Home Office (UK). This paper is of UK crimes and not US crimes, but it is much more recent. It also uses survey data to estimate police costs and include overhead costs in the estimate. Although UK police cost may not be representative of US police cost, we add the estimate as we could find no recent estimates from the US, and the Miller paper is old. Table VI.VI contains the estimates for each crime.

Table VII.I Accuracies of model and witness predictions and estimated witness error

Crime	N	$n \times \varepsilon$	Miller et al.	Heeks et al.	j_i	c_i
Homicide	11788	57	$2382	$17662	$10022	$567,069
Assault[9]	1462785	7021	110	1669	890	6,245,507
Rape	23632	113	68	9392	4730	536,540
Burglary	207325	995	238	783	783	508,029
Robbery	95754	460	238	1491	865	397,341

Using equation (VI.1) we find that the yearly expected cost of witness error by the five crimes on the police cost is $8.25 million. By differentiating the function with respect to ε, we find that the

[8] Estimates from Heeks et al. (2018) are average police costs and not unit costs, as it is calculated using all crimes (reported and recorded) rather than just police recorded crimes.

[9] We equate *Violence with injury* from Heeks (2018) with the US definition of assault. **This is not entirely correct as assault includes intimidation, however the majority of assault cases in our data were of a violent nature.**

cost spared from reducing the witness error by 1% is $515,905. In other words, a government should be willing to pay $515,905 every year to reduce witness error by 1% in that year.

VIII Correcting witness testimonies

With high potential costs to society and potentially devastating consequences to an individual, governments should devote more resources to reducing witness error. In this section we give our recommendations for how witness error or the cost of witness error can be reduced using the racial model.

A. Model application

We can imagine three ways of using a model as the one presented in this thesis to correct testimony directly: i) the discrete predictions from the model can replace the witness prediction in an incident report, ii) witness guesses can be included as a candidate learner in the ensemble model, or iii) the disagreement between the discrete prediction and witness guess can be used to identify possible mislabels and encourage further human intervention.

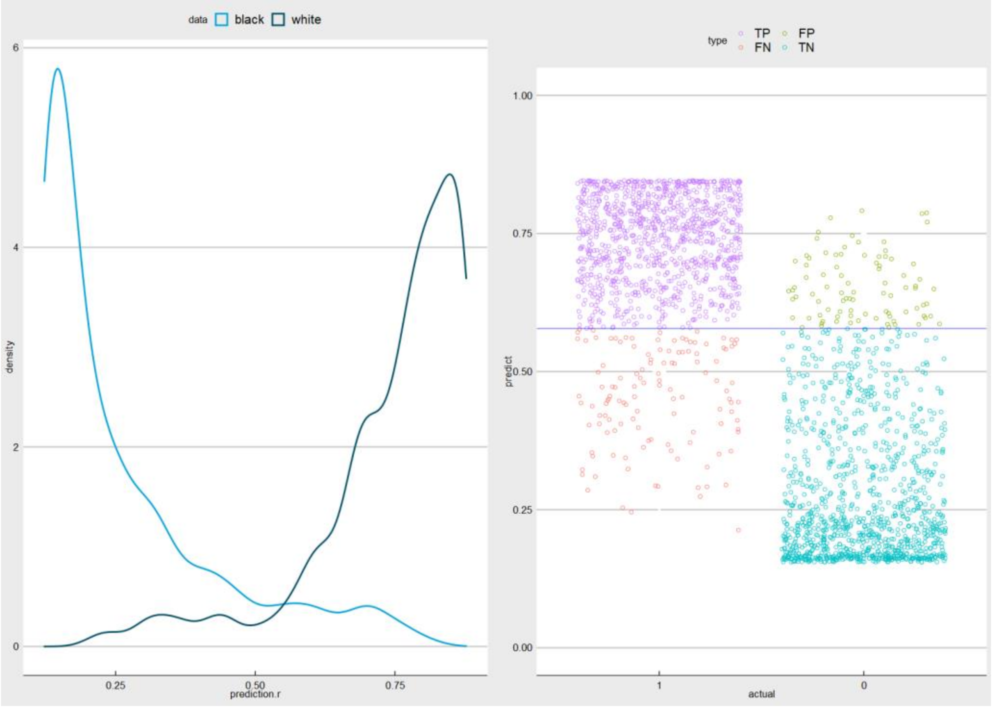
In section VI.A we showed that the racial model is 91.33% accurate in predicting the race of an offender, using only characteristics of the crime and victims in an incident. The witness error was found to be 16.17% for the same data. This implies that using only the model, the overall labelling error can be reduced to 8.77% by simply replacing all witness guesses with the corresponding class prediction generated by the model. This implies an annual police cost reduction of $4.5 million using the estimate in section VI.H. The model can also be used here to impute unknown labels of offenders, as we do in section V.E. This may further decrease the cost of investigation or lead to meaningful increases in clearance rates. One drawback of imputing values and replacing witness guesses directly is that we cannot justify the results as the model is not interpretable. This may cause moral quarrel when applied in a judicial system. Another problem is that error may not be uniformly important. A reduction in the error associated with homicides may be more important than a proportionate reduction in error rate associated with robberies. As we have not studied for which crimes the model outperforms witness guesses, replacing the witness guesses may adversely affect some crime types. A third challenge is that we cannot distinguish incidents which are exceptions from those which are noise. These challenges can in part be overcome to with increased data specificity, allowing for more accurate and specific models.

Other than increasing the quality of the data, the probability of labelling exceptions as error can be reduced by using model predictions as suggestions rather than conclusions. In section V.D, ensemble learning was discussed as a way to improve predictive accuracy and include information and benefits of multiple algorithms. A witness prediction could be considered as a candidate learner in this process as it may contain information that the other learners do not have. For example, witness predictions are typically based on contemporary experience rather than historical data. As such, witness prediction is more robust to concept drifts than the statistical model. Combining witness testimony and a prediction model can minimize the weaknesses of the two in isolation and give favourable results. Adding the witness predictions to the ensemble itself may improve final predictive accuracy and make the final predictions more robust to outliers.

Finally, predictions can be used as an alarm tool. Recent publication in misclassification analysis suggests an approach to identifying mislabelled data that requires human interaction (Ekambaram, Goldgof, & Hall, 2017). This approach assumes that error can be spotted by humans if attention is drawn to the observation. In other words, the prediction model can be used to identify the possibility of error, and then rejected or confirmed at the discretion of individuals in the legal system.

Another way of correcting witness testimony through ML may be to use the probabilistic prediction as a gauge of confidence. The racial model produces discrete predictions as well as continuous predictions. Indeed, this is true for all classification models, although some methods come prebuilt with threshold rules for transforming continuous predictions into discrete predictions. Until now we have only discussed the discrete predictions and their application in determining witness error. However, the continuous prediction for an incident can be useful as well as it can be used to assess the confidence with which a class prediction is made. For a well-calibrated prediction model, the confidence of a discrete prediction increases as the continuous prediction increases (Kuhn & Johnson, 2013). In other words, we are more certain that an offender labelled as white is white if the prediction is closer to 1 than to 0.5, or lower than 0.5. Conversely, it is more certain that an offender is black if the probabilistic prediction takes on a low value, such as 0.1, rather than a high value. This is also the case for the model developed in this thesis. In Figure 2.2, left panel, we plot the occurrences of black and white offenders by predicted probability with density on the y-axis and the predicted probability on the x-axis. The functions together take on what resembles an inverse gaussian distribution, where most black offenders correspond with small probabilities, and most white offenders correspond with high probabilities. In the right panel of the same Figure, we illustrate how the clusters of incidents are divided into discrete predictions based on the chosen cut-off.

Figure VIII.I Continuous predictions and predictions separated by cut-off (arrests)

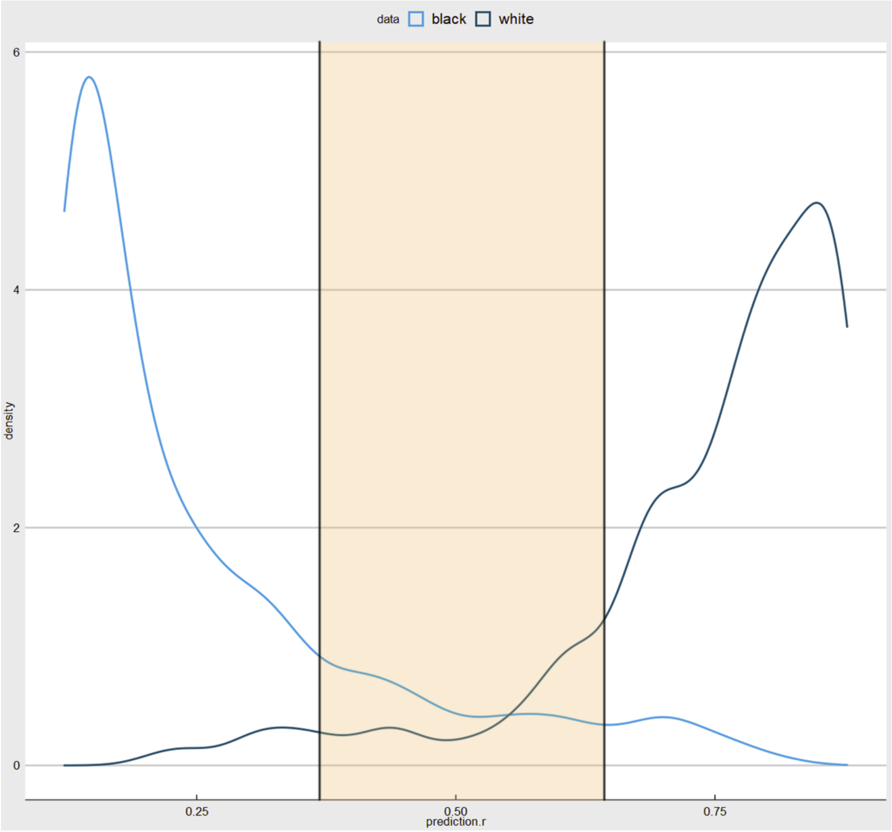


Research shows that jurors rely on witnesses' confidence to infer how accurate their testimony is (Cutler et al., 1990). According to a paper by Gary Wells (1998) reported confidence by the witness is suggested as the most powerful determinant of judged accuracy. The actual relationship between witness confidence and predictive accuracy is a well-studied phenomenon, but the evidence is conflicting (Horvath, 2009). Similarly, studies show that even judges' have limited awareness of factors that lead to witness error (Magnussen & Wise, 2008). This is especially concerning as judges play a critical role in delivering verdicts. In cases where a judge is the ultimate decision-maker the impact of the limited knowledge is self-explanatory. In addition to these cases, it is also the judge who takes on the role of informing a jury in some criminal trials. The trial judge can instruct jurors on the factors that may result in erroneous identification (National Research Council, 2014). As such, judge's knowledge is paramount to apply knowledge about witness error. The probability generated by a classifier is clearly linked to predictive accuracy and therefore should do at least as well as an estimation made by a witness. This means that the continuous predictions can be supplied to decision-makers as a quantitative suggestion of confidence either in conjunction with a qualitative estimation made by individuals, or, in place of the qualitative estimates. In turn, this should lead to fewer erroneous convictions, and curtail the effect on judges limited knowledge on wrongful convictions. Optimally, a quantitative measure of confidence is supplied in conjunction with increased knowledge about eyewitness error in the courts. However, there is a limit to how detailed instructions on eyewitness identification can be in courts. As argued in Brodes v. State, 279 Ga. 435, 439 & n.6, specific instruction about eyewitness identification is an inappropriate judicial comment. A quantitative estimate may also skirt this notion as it is tangible and impartial.

An alternative to gauging testimony confidence directly could be to develop a model where the focus is on the expected introduced error rather than trying to predict an outcome as we have done here. In other words, a model for which the output is an estimate of how erroneous testimony may be under the specific conditions. This would allow judges and juries to exercise more discretion in gauging testimony confidence than the former alternative.

The implementation of a quantitative measure of confidence can also be further linked to judicial values as *Reasonable doubt* - a legal standard of proof required to validate a criminal conviction - through equivocal zones. An equivocal zone is an approach to improving performance for a classification model that allows for the class to be labelled as unknown or indeterminate in ranges where the probabilistic prediction is very uncertain. From literature it an equivocal zone should be defined as $0.5 \pm z$, for balanced two-class problems (Kuhn & Johnson, 2013). Using our model as an example, an equivocal zone could be implemented in the ranges of 0.5 ± 0.15 ([0.35,0.65]). From Table VII.II the lines for black and white and intersecting and overlapping and so a prediction in this range unlikely to lead to accurate results.

Figure VIII.II Equivocal Zone



Finally, the predicted class probabilities can also be used to assist law enforcement in allocating their resources more effectively. Borrowing from expected utility theory and actuarial analysis, the certainty of an offender’s characteristic can be combined with the cost of investigating suspects to be used to determine if pursuing the investigation is in the police’s interest. Alternatively, the probability can be used to determine how the police should spend resources pursuing different suspect profiles.

IX Conclusion

Can machine learning be used to correct witness testimony? Unlike most studies about erroneous witness testimonies, this thesis attempts to correct the errors of the witness proactively through statistical methods. We have illustrated how a machine learning model can be used to correct witness testimony. A machine learning model was trained on actual arrests to predict race of an offender. The model exhibited high accuracy using a subset of the available information from incident reports. By layering the predictions from this model onto witness guesses of offender race we were able to uncover error in a testimony. A conservative estimate of the economic benefit of reducing the error was made. A 1% decrease in witness error corresponded with $515,905 saved in police costs. Several ways of applying the model to correct witness testimony was proposed: 1) erroneous witness guesses can be replaced with model predictions, 2) model predictions and witness guesses can be combined to reduce error, or 3) model predictions can be used to identify possibility of error and the need for additional human scrutiny. In addition, we have made suggestions for how probabilistic predictions can be used to provide a measure of confidence for a testimony, or to guide decisions about resource use in law enforcement. While in this thesis we have focused on one characteristic, race, we have showed that the results may generalize to other characteristics such as gender and age. We hope that follow-up work will pursue the creation and application of more overarching models for witness error and better classify the cases when witnesses mislabel offenders in actual incidents.

References

Abadie, M., & Camos, V. (2019). False memory at short and long term. *Journal of Experimental psychology*, 1312-1334.

Addington, L. A. (2006). Using National Incident-Based Reporting System Murder Data to Evaluate Clearance Predictors: A Research Note. *Homocide Studies*.

Areh, I. (2011). Gender-related differences in eyewitness testimony. *Personality and Individual Differences*, 559-563.

Borchard, E. M. (1932). *Convicting the Innocent.* Garden City, New York: Yale University Press.

Breiman. (2001). Random forest. *Machine Learning, 45,* 5-32.

Brodley, C. E., & Friedl, M. A. (1999). Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research*, 131-167.

Bruck, M., & Ceci, S. (1995). Amicus brief of the case of New Jersey v. Margaret Kelly Michaels presented by committee of concerned social scientists. *Psychology, Public Policy and Law*, 272-322.

Clifford, B., & Davies, G. (1989). Procedures for obtaining identification evidence. In D. Raskin, *Psychological methods in criminal investigation and evidence* (pp. 47-95). Springer Publishing Company.

Cutler, B. L., Penrod, S., & Martens, T. (1987). The reliability of eyewitness identification: The role of system and estimator variables. *Law and Human Behavior, 11*, 233-258.

D'Alessio, S. J., Stolzenberg, L., & Eitle, D. (2002, (31,3)). The effect of racial threat on interracial and intraracial crimes. *Social Science Research*, 392-408.

Delisi, M., Kosloski, A., Sween, M., Hachmeister, E., Moore, M., & Drury, A. (2010). Muder by numbers: monetary costs imposed by a sample of homocide offenders. *The Journal of Forensic Psychiatry & Psychology*, 501-513.

Diettrich, T. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 1895-1923.

Ekambaram, R., Goldgof, D. B., & Hall, L. O. (2017). Finding Label Noise Examples in Large Scale Datasets. *IEEE International Conference on Sysstems, Man and Cybernetics* (pp. 2420-2424). IEEE.

Fawcett, J. M., Russell, E. J., Peace, K. A., & Christie, J. (2013). Of guns and geese: a meta-analytic review of the 'weapon focus' literature. *Psychology, Crime & Law, 19*, 35-66.

FBI. (2020, November 25). *Five things to know about NIBRS*. Retrieved from FBI News: https://www.fbi.gov/news/stories/five-things-to-know-about-nibrs-112520

FBI. (2021, April 28). *Arrest Data - Reported Number of Arrests by Crime*. Retrieved from Crime Data Explorer: https://crime-data-explorer.fr.cloud.gov/#

Feingold, G. A. (1914). Influence of Environment on Identification of Persons and Things. *Crim. L. & Crimonology, 39*.

Frey, W. H. (2015, December 8). *Census shows modest declines in black-white segregation*. Retrieved from Brookings Institution: https://www.brookings.edu/blog/the-avenue/2015/12/08/census-shows-modest-declines-in-black-white-segregation/

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R.* London: Springer.

Goldstein, A. G., Chance, J. E., & Schneller, G. R. (1989). Frequency of eyewitness identification in criminal cases: a survey of prosecutors. *Bulletin of the Psychonomic Society*, 71-74.

Greenwell, B. M., & Boehmke, B. C. (2020). Variable Importance Plots - an intrudction to the vip package. *The R Journal vol. XX/YY, AAAA*.

Gremmell, D. (2018, February 20). *Ensemble Learning in R with SuperLearner*. Retrieved from datacamp: https://www.datacamp.com/community/tutorials/ensemble-r-machine-learning

Grimsley, E. (2012). *What Wrongful Convictions Teach us About Racial Inequality.* Innocence Project.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). Model Assessment and Selection. In T. Hastie, R. Tibshirani, & J. Friedman, *The Elements of Statistical Learning: Data mining, Inference, and Prediction* (pp. 219-261). California: Springer.

Heeks, M., Reed, S., Tafsiri, M., & Prince, S. (2018). *The economic and social cost of crime: Research report 99.* Home Office.

Hooker, G., & Mentch, L. (2019). Please Stop Permuting Features - An Explanation and Alternatives. *arXiv:1905.03151 [stat.ME]*. Retrieved from https://arxiv.org/pdf/1905.03151.pdf

Horvath, M. A. (2009). Eyewitness Evidence. In S. Tong, R. P. Bryant, & M. A. Horvath, *Understanding Criminal Investigation* (pp. 93-114). West Sussex: Wiley-Blackwell.

Johnson, C., & Scott, B. (1976). Eyewitness testimony and suspect identification as a function of arousal, sex or witness and scheduling of interrogation. *American Psychological Association Annual Meeting.*

Kapardis, A. (1997). *Psychology and law: A critical introduction.* Cambridge University Press.

Kuhn, M., & Johnson, K. (2013). Measuring Performance in Classification Models. In *Applied Predictive Modeling* (p. 249). New York: Springer.

LeDell, E., Laan, M. J., & Peterson, M. (2016). AUC-Maximizing Ensembles through Metalearning. *International Journal of Biostatistsics*, 203-218.

Loftus, E., & Hoffman, H. (1989). Misinformation and memory: the creation of new memories. *Journal of Experimental Psychology: General, 118*, 100-104.

Magnussen, S., & Wise, R. A. (2008). What judges know about eyewitness testimony: A comparison of Norwegian and US judges. *Psychology Crime and Law*, 177-188.

Memon, A., Bartlett, J., Rose, R., & Gray, C. (2003). The Aging Eyewitness: Effects of Age on Face, Delay and Source-Memory Ability. *The Journals of Geontology*, 338-345.

Miller, T. R., Cohen, M. A., & Wiersema, B. (1996). *Victim Costs and Consequences: A New Look.* Rockville: National Institute of Justice.

Mu, Y., Ding, W., & Tao, D. (2013). Local discriminative distance metrics ensemble learning. *Pattern Recognition, 46, 8*, 2337-2349.

National Research Council. (2014). Identifying the Culprit: Assessing Eye Witness Identification. Washington D.C.: The National Academies Press.

New England Innocence Project. (2021, February 9). *New England Innocence Project*. Retrieved from Eyewitness Misidentification: https://www.newenglandinnocence.org/eyewitness-misidentification

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). *The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention.* Journal of Experimental Psychology Applied.

Parker, J. F., Haverfield, E., & Baker-Thomas, S. (1986). Eyewitness Testimony of Children. *Journal of Applied Social Psychology*, 287-302.

Pike, Graham, Brace, & Nicola. (2002). *The visual identification of suspects: procedures and practice.* Policing and Reudcing Crime Unit, Home Office Research, Development and Statistics Directorate.

Polley, E. C., & van der Laan, M. J. (2010). *Super Learner.* California: University of California, Berkley.

Roland, G., & Fryer, J. (2019). An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy*, 1210-1261.

Roupp, M. D., Perkins, N. J., Whitcomb, B. W., & Schisterman, E. F. (2008). Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. *Biometrical Journal*, 419-430.

Sabzevari, M., Martinez-Munox, G., & Suarez, A. (2018). A two-stage ensemble method for the detection of class-label noise. *Neurocomputing 275*, 2374-2383.

Shapiro, P., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100(2)*, 139-156.

Tolsma, J., Blaauw, J., & Grotenhuis, M. t. (2012). When do people report crime to the police? Results from a factorial survey design in the Netherlands. *Journal of Experimental Criminology*, 117-134.

United States Department of Justice. (2012). *United States Attorneys' Annual Statistical Report.* Department of Justice.

US Department of Justice. (2021, April 28). *Arrests by offense, age, and race*. Retrieved from Statistical Briefing Book: https://www.ojjdp.gov/ojstatbb/crime/ucr.asp?table_in=2

Vapinik, V. (2000). *Principles of Risk Minimization for Learning Theory.* Holmdel: AT&T Bell Laboratories.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36(12)*, 1546-1557.

Wilcock, R., Bull, R., & Vrij, A. (2007). Are old witnesses always poorer witnesses? Identification accuracy, context reinstatement, own-age bias. . *Psychology, Crime & Law*, 305-316.

Wixted, J., Mickles, T., & Fisher, L. (2018). Rethinking the reliability of eyewitness memory. *American Psychological Association*, 324-335.

Wong, H. K., Stephen, I. D., & Keeble, D. R. (2020). The Own-Race Bias for Face Recognition in a Multiracial Society. *Frontiers in Psychology*, 208.

Yoon, S. (2015). *Why do Victims not Repor?: The influence of police and criminal justice cynicism on the dark figure of crime.* New York: City University of New York.

Yuille, J. (1986). Meaningful research in the police context. In J. C. Yuille, *Police Selection and Training: The Role of Psychology* (pp. 225-43). Dordrecht: Springer.

Appendix

Table A.I Incident hour in training data

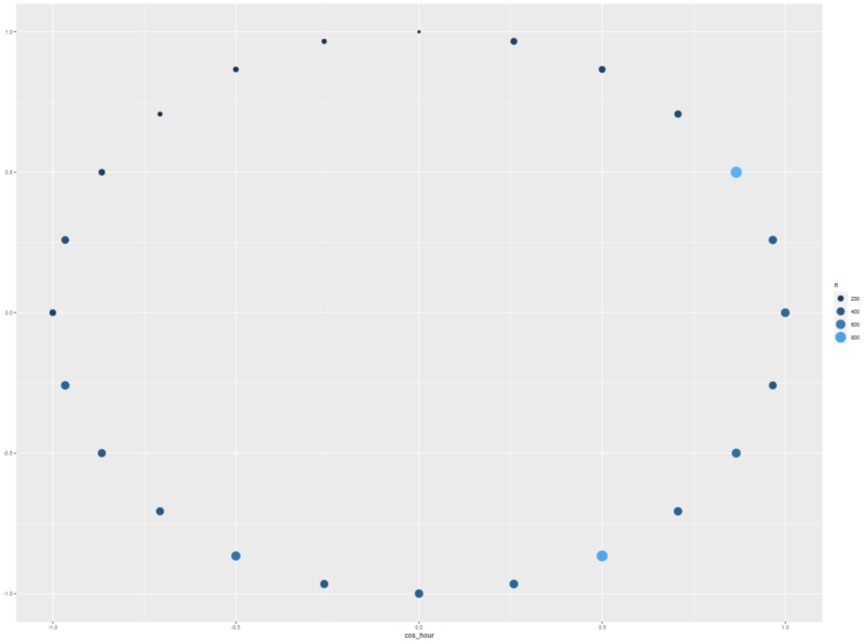


Table A.II Incident by date in full data

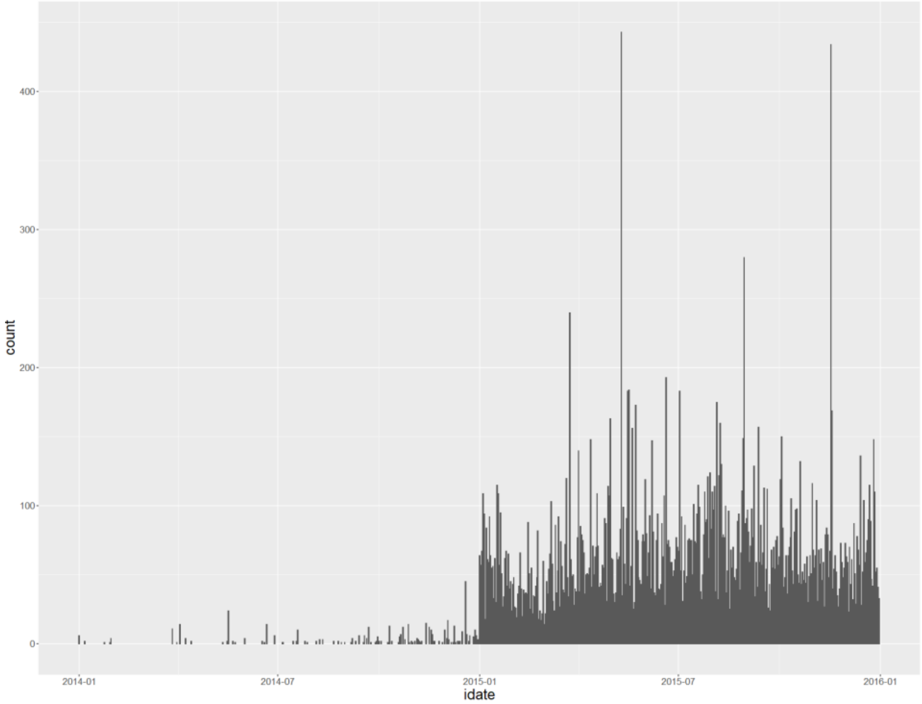


Table A.III Overview of the features used to train the racial model

Variable	Description	Type	Levels	Mean (arrests)	Mean (witness)
ori	Agency Identifier	Categorical	144	--	--
date	Date (Y%M%D%)	Numerical	--	--	--
vage	Victim age	Numerical	--	30.9	31.23
vsex	Victim male	Dummy	2	0.6322	0.5677
vrace	Victim race	Categorical	4	--	--
vresident	Victim is a resident	Dummy	2	0.8239	0.8668
ptype	Property type	Categorical	8	--	--
booty	Amount of property lost	Numerical	--	893	884.6
location	Crime location description	Categorical	40	--	--
code	Primary offense code	Categorical	10	--	--
completed	Crime was completed	Dummy	2	0.9978	0.9965
wtype	Weapon type	Categorical	19	--	--
bias	Bias motivation	Categorical	10	--	--
offense_seg	Number of offenses	Numerical	--	3.019	2.393
victim_seg	Numbers of victims	Numerical	--	3.835	3.001
offender_seg	Number of offenders	Numerical	--	2.648	2.626
unknown_booty	Unknown booty	Dummy	2	0.136	0.1143
hour	Incident hour	Numerical	2	12.37	12.31