A COMPARISON OF CLASSIFICATION METHODS IN PREDICTING THE PRESENCE OF DNA PROFILES IN SEXUAL ASSAULT KITS

Derek J. Heckman

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

May 2018

Committee:

Jim Albert, Advisor

Arkajyoti Roy

Jon Sprague

ABSTRACT

Jim Albert, Advisor

In 2014 Ohio began the Sexual Assault Kit Testing Initiative with the goal of analyzing all previously untested sexual assault kits (SAKs). Approximately 13,900 previously untested SAKs were sent for forensic analysis. Of these SAKs, a sample of 2,500 was drawn for statistical analysis. The goal was to gain some general information about the SAKs as well as to answer a variety of specific questions in the hopes of producing cost-saving measures in the future. Questions considered were those such as: which forensic samples most consistently produce Combined DNA Index System (CODIS)-eligible DNA profiles, what factors predict whether or not a kit will contain a DNA profile foreign to the victim, as well as others. The results of the initiative were published in Kerka, Heckman, Maddox, Sprague, & Albert (2018).

This thesis expands upon the work in the aforementioned article. In the article, a logistic regression model was constructed to predict whether or not an SAK would contain a CODIS-eligible DNA profile. It was estimated to have a misclassification rate of 34.2%. This thesis compares three other models to the logistic regression model to determine if any improvements in performance can be made. The models tested were decision trees, bagged trees and random forests. The decision tree had an estimated misclassification rate of 29.7%, thus offering a moderate improvement over the logistic regression model.

In addition, the same models were compared for their ability to predict which SAKs would contain duplicate DNA profiles across multiple forensic samples (vaginal sample, anal sample, etc). No model was able to do a satisfactory job of predicting this response.

This thesis is dedicated to all victims of sexual assault. I hope it makes some small contribution to rectifying the injustice you have suffered.

ACKNOWLEDGMENTS

I would like to acknowledge Jaimie Kerka for her help in providing much of the background information for this thesis, as well as the countless hours she spent working with me on the Sexual Assault Kit Testing Initiative. I would like to thank Dr. Jon Sprague for all the time and effort he spent working with Jaimie and I. His patience and dedication are much appreciated. I would like to thank Dr. Jim Albert for agreeing to advise this thesis. His advice and feedback have been tremendously beneficial. Also, I would like to thank the authors of *An Introduction to Statistical Learning*, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, for allowing me to use one of the figures from their text. Lastly, thanks to the Laura and John Arnold Foundation for providing the funding for the initiative.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1 INTRODUCTION

The backlog of untested sexual assault kits (SAKs) in many states and cities has become an issue of growing alarm. In 2008, over 10,000 untested SAKs were found in the Los Angeles Police and Sheriff's Departments (Peterson, Johnson, Herz, Graziano, & Oehler, 2012). Maryland revealed that police departments across the state had over 3,700 untested SAKs (The Baltimore Sun, 2017). As of 2017, Nevada had 8,000 untested SAKs, Arizona had 6,000, and Florida had 13,000 ("Where the Backlog Exists," 2017). Other locations have similar backlogs. Many states and cities don't even know how many they have.

The federal government has taken notice. In April 2017, a group of lawmakers launched the Bipartisan Task Force to End Sexual Violence. The task force was created to address issues such as K-12 education, college campus safety, the rape kit backlog, and law enforcement training, among other things ("Bipartisan Group," 2017). Even popular actors such as Law and Order SVU's Mariska Hargitay have been raising public awareness about the issue. In June 2017, Hargitay appeared before the aforementioned task force as a witness ("Bipartisan Task Force," 2017).

Like many other states and cities, Ohio had its own backlog of untested SAKs. In Ohio's case, the backlog numbered in the tens of thousands, with cases dating from the late 80's to the early 2010's. In 2014 Ohio began the Sexual Assault Kit Testing Initiative in order to address the issue. Through the initiative, all untested SAKs were submitted for forensic analysis. Approximately 13,900 kits were analyzed. From the population of 13,900, a sample of 2,500 was taken to look for patterns that might allow for more efficient processing in the future. The significant conclusions of the work were summarized in Kerka, Heckman, Maddox, Sprague, &

Albert (2018). This thesis hopes to expand upon the work of the paper, in ways soon to be described.

The Sexual Assault Kit Testing Initiative has been successful in identifying perpetrators of sexual assaults as well as linking cases involving unknown assailants through the utilization of the Combined DNA Index System (CODIS) database. CODIS is used to store the DNA profiles of perpetrators of criminal activity. When DNA evidence is found at a crime scene (or on a victim's body in the case of sexual assault), if it meets the standards of entry into the CODIS database, it will be uploaded and potentially linked to an identical DNA profile that already exists in the database. If the perpetrator DNA profile and identity has previously been loaded into CODIS, uploading the DNA profile again will match (hit) to a person of known identity. If the perpetrator has left DNA at previous crime scenes but never been apprehended his profile may appear in CODIS but there will be no identity associated with it.

When a person becomes the victim of sexual assault, he/she goes to the hospital to have an SAK collected by a Sexual Assault Nurse Examiner (SANE). The SANE collects a variety of samples from the victim. Some of the most commonly collected samples consist of vaginal swabs, anal swabs, oral swabs, pubic hair standards and combings, and swabs and cuttings from the underwear worn during the sexual assault. Once these samples are collected they are placed in an SAK, which is then transferred to the local police department where a decision is made as to whether or not it should be submitted to the laboratory for testing. If an SAK is submitted for DNA testing, a variety of forensic samples contained within the SAK undergo processing to determine if a suspect's DNA profile can be identified from the evidence. Forensic samples are initially sent through DNA extraction where the DNA contained within the sample is isolated and purified. The commonly low-level amounts of forensic DNA are then amplified by polymerase chain reaction (PCR) to increase the number of copies of the DNA. The samples ultimately undergo capillary electrophoresis which generates data containing the DNA profiles from each individual forensic sample. From there, DNA analysts examine the DNA profiles to determine if a DNA profile foreign to the victim of the sexual assault is present. These foreign DNA profiles can then potentially be uploaded into CODIS or compared to a suspect DNA standard. If multiple forensic samples produce CODIS-eligible DNA profiles of the same individual, only the highest quality profile is uploaded. Across multiple previous studies that examined DNA testing results of forensic samples, DNA profiles were uploaded into CODIS in between 25.0% and 49.0% of all SAKs examined. Of those DNA profiles that were uploaded into CODIS, a "hit"—a DNA profile match to either a named individual or a forensic sample in an unsolved case—was seen in approximately 48.9% to 58.0% of the uploads (Wells, Campbell, & Franklin, 2016).

The reason for the backlog of SAKs is that law enforcement doesn't submit every kit they obtain for forensic analysis. There are a variety of reasons for this. One is that most jurisdictions do not have written guidelines for testing SAKs. Many decisions are left to the leadership of the department, and if the department is understaffed or does not place a high value on sexual assault, the kits may be neglected. In addition, there is little training on how to handle sexual assault. When a victim is traumatized after the attack, he/she may not clearly recall the details of the event. When a victim cannot clearly articulate what happened, this harms her perceived credibility, inclining law enforcement to drop the case. It is also possible they hold the victim responsible for what happened, further reducing her credibility. Also, many jurisdictions don't test SAKs when the identity of the assailant is known, because they don't believe the results of the testing will improve prosecution of the case. In addition, many crime labs simply don't have the funding or personnel to properly test all kits. On average, testing one kit costs between $1000 and $1500, which at times can strain resources (“Why the Backlog Exists,” 2017).

As previously stated, the Sexual Assault Kit Testing Initiative was Ohio’s effort to address the SAK backlog in the state. Kerka et al., (2018) summarized the efforts of the project. One of the primary goals of the project was to determine which forensic samples (vaginal, anal, etc) tend to yield DNA profiles that are CODIS-eligible. On the basis of the results, the paper recommended certain guidelines for future SAK testing. In addition, logistic regression was used to estimate the probability that a particular kit would contain a CODIS-eligible DNA profile. This thesis expands upon that work by comparing the performance of several different statistical methods for their ability to predict the presence of a CODIS-eligible DNA profile in a kit. Furthermore, it also tries to identify factors that predict which kits contain duplicate DNA profiles across different forensic samples (for instance, if the same profile will be found in both the vaginal and anal sample). This thesis proposes several different methods and compares them on the basis of their predictive ability and ease of interpretation. The methods compared are: logistic regression, decision trees, bagged trees, and random forests.

One of the limitations of this work is that some of the DNA profiles obtained may have been those of consent partners rather than those of perpetrators. At the time of SAK collection victims are asked whether they had recent consensual sex. If they report yes, it is the job of the local law enforcement agency to obtain a DNA sample from the consent partner, called an elimination standard, in order to ensure the consent partner’s profile is not uploaded into CODIS. Unfortunately, these standards are rarely collected around the time of the initial investigation, and as a result some of the DNA profiles uploaded are those of non-criminals.

Another limitation is that the SAKs in the sample were collected from assault scenarios that varied dramatically. Some victims were vaginally assaulted. Others were vaginally and anally assaulted. Others were forced to perform oral sex on the assailant. This information might have been incorporated into the methods, but unfortunately the type of assault the victim endured was one of the most frequently missing pieces of information from the kit. In the sample of 2,500 kits, whether a vaginal or anal assault occurred was missing almost 30% of the time. As a consequence, such information was not utilized.

A third limitation is that single-assailant and multiple-assailant cases were not separated. Both kinds of cases were used for the purposes of building the statistical models. One of the response variables of interest was whether or not at least one CODIS-eligible DNA profile was uploaded. It did not distinguish between one, two, three, or four profiles being uploaded. If there are significant differences between the two scenarios there may be a loss in the accuracy of the model predictions.

At the time of publication, no other work had previously been done to build statistical models to predict whether or not an SAK would contain a CODIS-eligible DNA profile or duplicate DNA profiles. Much work has been done in related areas, however. Fallik & Wells (2014) suggest that submitting SAKs for analysis long after initial collection is unlikely to have an impact on the investigation and prosecution of the case. Gray-Eurom, Seaberg, and Wears (2001) built a logistic regression model to predict whether or not a case would result in successful prosecution resulting in a conviction. Several cities have done studies on backlogged SAKs and looked at the rates of CODIS uploads and CODIS hits, such as Houston, Detroit, Los Angeles and New Orleans (National Institute of Justice, 2016). These projects did not use statistical modeling to predict which kits would contain CODIS-eligible DNA profiles though. As such, this work hopes to make a unique contribution to the field.

The next chapter gives a summary of the variables in the dataset. Chapter 3 is an exploratory analysis of the predictor variables. Chapter 4 gives an overview and examples of the different methods used in the thesis. Chapters 5 and 6 compare the ability of the four methods (logistic regression, decision trees, bagged trees and random forests) to predict the presence of CODIS-eligible DNA profiles and duplicate DNA profiles respectively. Chapter 7 makes some concluding remarks.

All statistical models were built using the statistical software R (R Core Team, 2017). The 'ggplot2' package was used to make many of the graphs in this thesis (Wickham, 2009). The 'dplyr' package was used for easy manipulation of the data (Wickham, Francois, Henry, & Muller, 2017). The 'rpart' package was used to fit decision trees and make graphical displays (Therneau, Atkinson, & Ripley, 2017). The 'step' and 'glm' functions were used to fit the logistic regression models (R Core Team 2017). The 'randomForest' package was used to fit the bagged tree and random forest models (Liaw & Wiener, 2002). Appendix A contains a full list of packages and functions used.

CHAPTER 2 DATA DESCRIPTION

2.1 Data Collected

Over 13,900 sexual assault kits were submitted to the Ohio Bureau of Criminal Investigation from over 294 unique law enforcement departments in the state of Ohio under the Sexual Assault Kit Testing Initiative. The results of the analyses were compiled into case files. From the population of 13,900 kits, a convenience sample of 2,500 SAKs was drawn for statistical analysis. The cases were chosen at convenience because the Ohio Bureau of Criminal Investigation already had several thousand case files associated with SAKs in its possession and opted to use those as opposed to ordering files from storage. The data from these 2,500 SAKs was manually entered into a spreadsheet for analysis. The following is a partial list of information recorded (excluding some of the more technical forensic details):

- The date of offense
- The date of SAK collection
- The date of SAK submission to laboratory for testing
- The type of assault (vaginal, anal, oral, or other)
- The age of the victim
- Whether or not foreign DNA was obtained from particular forensic samples, and the quality of that DNA
- Whether or not DNA from a particular forensic sample was uploaded into CODIS
- Whether or not the profile uploaded to CODIS produced a ‘hit’
- Whether the local law enforcement agency obtained a ‘hit confirmation standard’ from the individual identified by CODIS to confirm the identity of the perpetrator
- Number of assailants reported by the victim

- The number of distinct DNA profiles obtained
- The number of duplicate DNA profiles obtained
- The number of forensic samples run in laboratory testing
- Whether or not the victim reported having consensual sex in the last 72-96 hours

The gender of the victim was not recorded, so we don't know exactly how many of the victims were male. However, penile samples are only collected for male victims, so the number of penile samples can be used to estimate the percentage of men in the sample. The number of penile samples collected suggest that at least 3% of the victims were male.

2.2 Selection of Predictor Variables

For the purposes of building the statistical models, decisions had to be made about which variables would be considered for inclusion in the models. Every variable that could have had an impact on the response was considered unless there was a specific reason to exclude it. Recall that the two responses of interest were whether an SAK contained at least one CODIS-eligible DNA profile and whether an SAK contained duplicate DNA profiles across multiple forensic samples. Since both responses deal with the results of forensic analysis, predictor variables that were appropriate for one were also appropriate for the other. As a result, discussion of which variables were considered for the models is condensed into one discussion.

One of the biggest reasons a variable may have been excluded from consideration is that information about it was missing in the SAK case file. Table 2.1 shows the percent of the cases with certain information missing. In a small fraction of the cases the date of offense was unknown, while in almost 10% of the cases the date of kit collection was unknown. In almost 30% of the cases it was unknown whether the victim suffered a vaginal or anal assault (not

Information	Missing (%)
Did Law Enforcement Obtain 'Hit Confirmation Standard'	65.3
Did Victim Report Other Kind of Assault	29.9
Did Victim Report Oral Assault	27.8
Did Victim Report Anal Assault	27.6
Did Victim Report Vaginal Assault	24.9
Did Victim Report Recent Consensual Sex	18.8
Number of Assailants Reported by the Victim	15.5
Date of Kit Collection	9.6
Victim Age	6.0
Date of Offense	3.0
Date of Kit Submission to Laboratory	0
Was Particular Forensic Sample Analyzed	0
Was DNA Obtained from a Sample	0
Was DNA Profile from Sample Uploaded into CODIS	0
Did Uploaded Profile Produce a 'Hit'	0
Number of Distinct DNA Profiles Obtained	0
Number of Duplicate DNA Profiles Obtained	0
Number of Forensic Samples Run in Testing	0

Table 2.1: Percent of Cases with Information Missing.

shown in table). It was likely unknown whether or not law enforcement obtained a hit confirmation standard. In about 15% of the cases the number of assailants reported by the victim was unknown, and in 19% of the cases whether the victim had consensual sex was unknown.

There are multiple methods for dealing with missing data. One such method is complete case analysis. In complete case analysis, all observations with any missing information are removed. Another method is mean imputation. In this method, the average of the non-missing values is imputed for all missing values of the predictor. A third option is single imputation. In single imputation, a model is created to predict the missing values. For instance, a linear regression model might be created to predict missing values for the number of assailants reported by the victim. Another, slightly more sophisticated method is multiple imputation. Multiple imputation takes the predicted value from single imputation and adds a random error term to it in order to reproduce the variability in the original data set. Each missing value is imputed in this way. But rather than just imputing each missing value once, multiple imputation imputes each missing value *m* times, resulting in *m* fully imputed data sets. Regression models are then built from each of these *m* data sets, and the coefficient estimates from each of the models are combined to create the coefficients of the final model. For more information please see UCLA (n.d.). For a complete treatment of the subject, please see Molenberghs, Fitzmaurice, Kenward, Tsiatis, & Verbeke (2014).

For this thesis, we chose to opt for single imputation by linear regression for the quantitative variables. Two variables we attempted to impute were not directly listed in the table above but could be calculated from the table. These were the number of days to kit collection and the number of years from kit collection to kit submission to the laboratory for analysis. Two linear regression models were constructed with each of these variables as the response. All other variables in the dataset were used as predictors, except whether an SAK contained a CODIS-eligible DNA profile and whether it contained duplicate DNA profiles. For both models, the value of R^2 was no greater than 10%. This means that no more than 10% of the variability in the response could be explained by the predictor variables. This is an extremely small figure, and as a result imputation by regression was abandoned. Imputation was not attempted for the qualitative variables due to a lack of sensible predictor variables. As a result, we decided to use complete case analysis, and all observations with any missing data were removed.

Next a decision had to be made as to which variables would be excluded due to having too much missing data. We decided that the variables relating to the type of assault the victim suffered were missing too frequently to be worth including in the models. If they were included, over 30% of the cases would have to have been removed.

Still, many decisions had to be made about other variables as well. One variable considered was whether the victim suffered a single-male or multiple-male assault. The problem with including this variable is that it is difficult to determine which kind of assault the victim actually suffered. A case might be considered multiple-male if the victim reported multiple assailants in the assault. The problem with using this to label a case 'multiple-male' is that often the victim doesn't remember the exact details of the assault (if she was drugged or beaten), and may misremember the number of assailants. As a consequence, whether the assault was single or multiple-male was not considered as a predictor variable.

Several variables weren't considered because they bore no logical relation to the response or they were not chronological (such as using number of distinct DNA profiles obtained to predict whether at least one profile was obtained). The final set of predictors considered for the models were: days to kit collection (a combination of date of offense and date of kit collection), years to kit submission to laboratory (a combination of date of kit collection and date of kit submission to laboratory), victim age, number of forensic samples run in laboratory testing, the occurrence of consensual sex, and ten categorical variables corresponding to whether each of the top ten most common samples was analyzed in a particular case (vaginal sample, anal sample, pubic hair comb swab, oral sample, underwear crotch swab, underwear crotch cutting, pubic hair standard swab, underwear non-crotch cutting, external vaginal area sample, and neck/ear swab). For the sake of clarity, the variables 'days to kit collection' and 'years to kit submission to laboratory' are shown on a timeline in Figure 2.1.

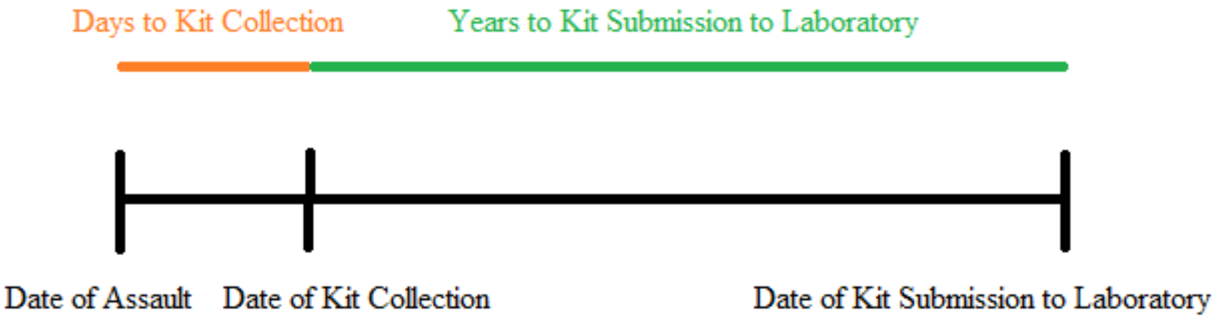
Days to Kit Collection
Years to Kit Submission to Laboratory
Date of Assault
Date of Kit Collection
Date of Kit Submission to Laboratory

Figure 2.1: Timeline of SAK submission process.

There were 538 cases with information missing about one or more of the predictor variables. As a result, these cases were removed for the purposes of building the models. This left 1,962 cases with complete information. In addition, several kits were removed because they were outliers. Two were removed because they were collected more than two weeks after the assault. Kits submitted to the laboratory 15 or more years after collection were also removed due to there being so few of them. When modeling which kits contained duplicate DNA profiles, kits with only one sample analyzed were removed. This left 1,937 cases for modeling which kits contained a CODIS-eligible DNA profile, and 1,924 cases for modeling which contained duplicate profiles.

CHAPTER 3 EXPLORATORY ANALYSIS OF PREDICTOR VARIABLES

In this chapter, we'll focus on exploring the distributions of, and relationships between, the predictor variables. Figure 3.1 shows, in three separate panes, the distribution of the number of days to kit collection, the number of years to kit submission to laboratory, and the age of the victim. Most kits are collected the same day as the assault (57.7%); only 4.4% of kits are collected three or more days after the assault. Only a few kits spent less than three years in storage. The median victim age was 22. Days to kit collection and victim age are both clearly right-skewed, while years to kit submission to laboratory is more symmetric. For the purposes of the models in Chapters 5 and 6, age is treated as a categorical variable with ranges 0-7, 8-15, 16-22, 23-50 and 51 plus.

Figure 3.2 from top to bottom shows the percent of victims who had recent consensual sex before the assault, the distribution of the number of samples run in laboratory testing, and the percent of the time the top ten most common samples were analyzed. Twenty-seven percent of victims had recent consensual sex before they were assaulted. The distribution of the number of samples run is fairly symmetric, with nearly half of the kits having four or five samples tested.

Figure 3.3 shows a violin plot of the distribution of days to kit collection and the number of samples run, both separated by victim age. The left panel helps us determine if certain age groups waited longer to have their kits collected. Violin plots are constructed such that each "violin" has an equal area. The width of the violin is relative to the total size of the group. If this wasn't the case, the violin for those age 23-50 would be much larger than for those age 0-7 because there were many more victims age 23-50. Looking at the plot, it appears those age 16-22 and 23-50 are most likely to have their kit collected on the same day as the assault, because the width of their violins at zero days is largest. This might be because those of adult age are able to

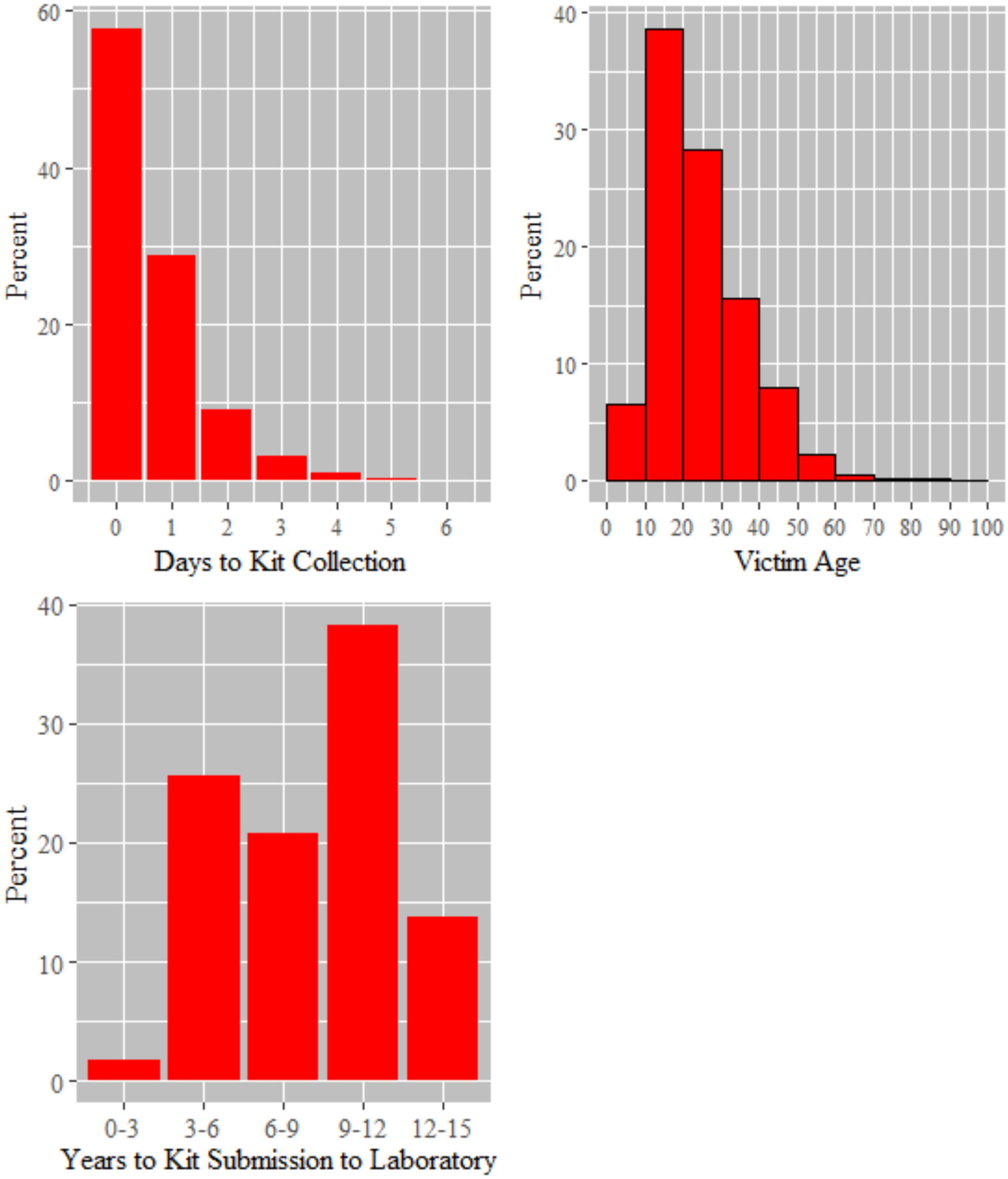

Figure 3.1: *Top Left*: Distribution of days to kit collection. *Top Right*: Distribution of victim age. *Bottom Left*: Distribution of years to kit submission to laboratory.

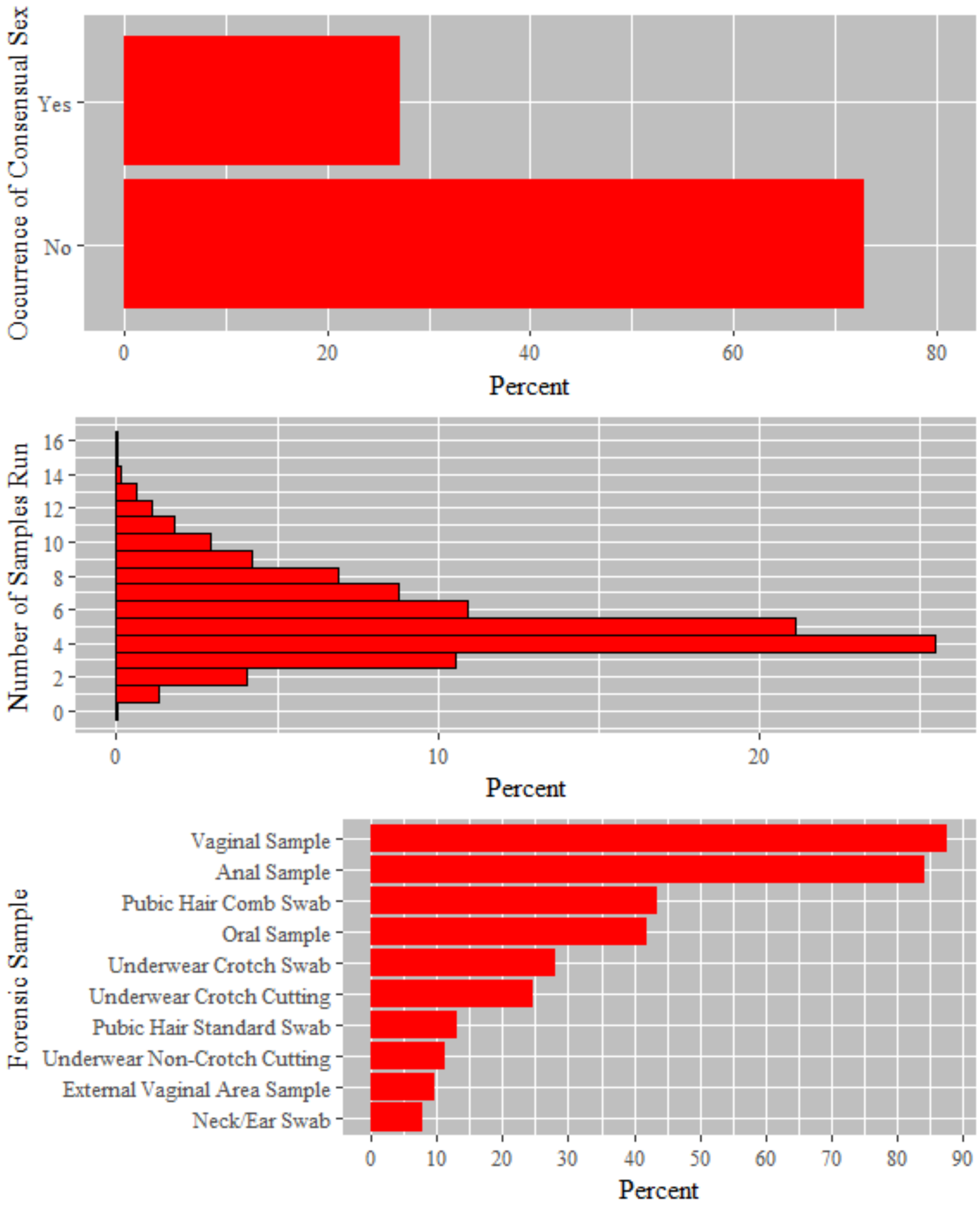

Figure 3.2: *Top*: Percent of victims that had recent consensual sex. *Middle*: Distribution of the number of samples run. *Bottom*: Percent of time common forensic samples were analyzed. The vaginal and anal samples were analyzed most frequently at 87.8% and 84.3%

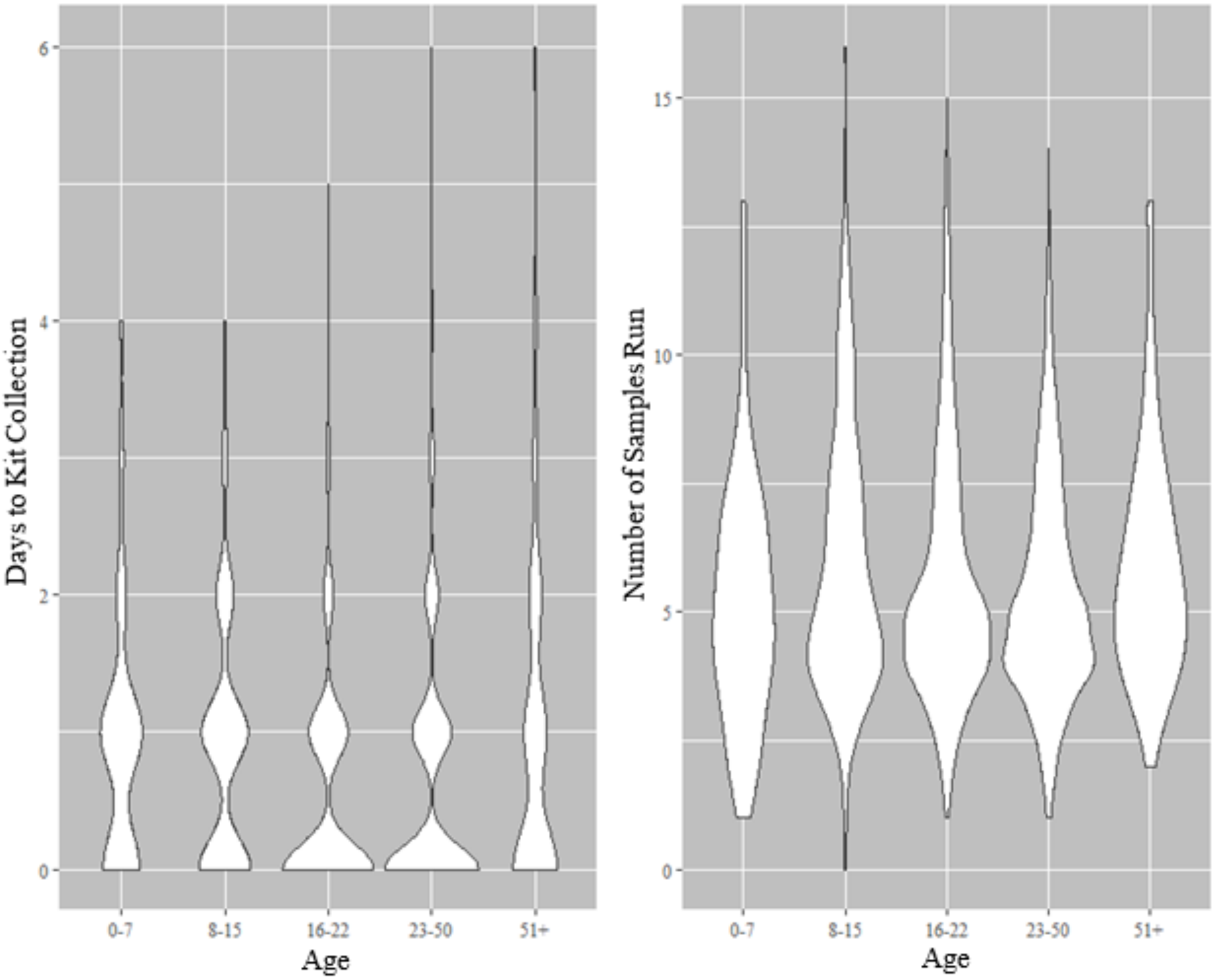

Figure 3.3: *Left*: The number of days to kit collection separated by victim age. *Right*: The number of samples run separated by victim age.

take themselves to the hospital to have the SAK collected while those of a younger age may be unable to take themselves or be uncertain of what to do in the wake of an attack. It appears those age 51+ show the greatest variability as to when they have their kit collected, as suggested by the fact that their violin is of the most uniform width.

The right panel suggests that the variability in the number of samples run is fairly uniform across different age groups. The age of the victim doesn't seem to have an impact on the number of samples run.

Figure 3.4 shows the percent of victims who had recent consensual sex by age category. As one might expect, those between the ages of 16-22 and 23-50 most frequently reported having consensual sex before the assault. In Chapter 4 we'll turn to the methods considered for the analysis.

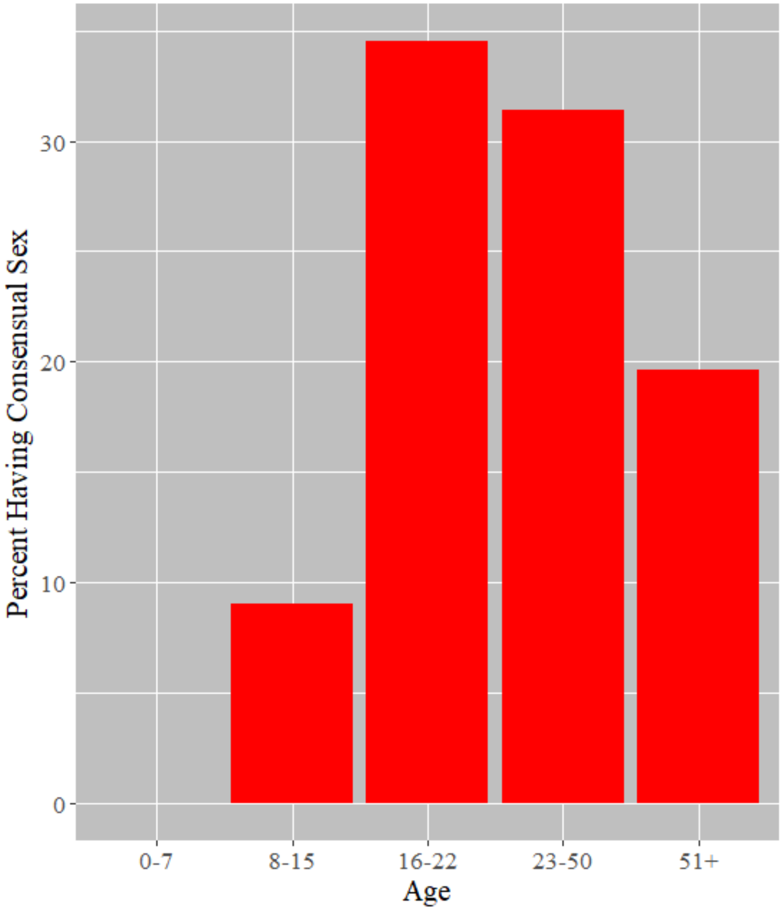

Figure 3.4: Percent of victims having consensual sex separated by victim age.

CHAPTER 4 CHOICE OF METHODS FOR MODEL-BUILDING

4.1 What is a Classification Problem?

For this thesis there are two response variables we are interested in predicting: whether an SAK contains at least one CODIS-eligible DNA profile, and whether an SAK contains duplicate DNA profiles across multiple forensic samples. These are qualitative, rather than quantitative, responses. As such, the methods that are typically used for quantitative responses, such as linear regression, will not work in this circumstance. A specific name is given to problems where the goal is to predict a qualitative response: classification problems. The goal of classification is to classify a particular observation to one of K classes. For both of the response variables we are interested in, K = 2 (either at least one profile was found or not, either a kit contained duplicate profiles or not). One of the biggest differences between classification methods and methods with quantitative responses is that typically a classification method will assign an observation a probability between 0 and 1 of belonging to a particular class. Generally, the observation is assigned to the class to which it has the greatest probability of belonging. Quantitative responses, on the other hand, can generally be any real-valued number.

4.2 Measuring the Performance of a Classification Model

There are three common ways to measure the performance of a classification model: the misclassification rate, expected loss, and cross-entropy, also known as normalized negative log-likelihood. The misclassification rate is simply the proportion of observations the model misclassifies. The expected loss takes into account the possibility that some mistakes are costlier than others. For instance, when diagnosing a disease, such as cancer, it would be costlier to misdiagnose someone who does in fact have cancer than to misdiagnose someone who doesn’t have cancer. Another scenario in which expected loss is appropriate is when a bank is deciding who to lend credit to. For the bank, lending credit to someone who defaults is costlier than failing to lend credit to someone who wouldn't default. With expected loss, different weights can be placed on different errors to account for the size of the error. Cross-entropy takes into account the certainty with which the model made the prediction. For instance, if a model estimated that a particular observation had a 90% probability of belonging to the wrong class, this would be punished more heavily than if the model had estimated only 60%. While each of these metrics is appropriate in different scenarios, for this thesis we'll use the misclassification rate as the metric of model performance. For a slightly more in-depth treatment of the subject please see Shalizi (2009).

4.3 Choice of Classification Methods

There are many different methods of classification. Common methods include K-nearest neighbors, logistic regression, discriminant analysis and decision trees. All were considered for potential use in this thesis. Both K-nearest neighbors and discriminant analysis were ultimately rejected because they are unable to handle qualitative predictor variables (such as the occurrence of consensual sex). However, both logistic regression and decision trees are able to handle qualitative predictor variables. As such, these methods were included for comparison. In addition, 'bagging' and 'random forests' were used as well. Bagging is a method that can be applied to other statistical methods in order to improve performance. It is commonly applied to decision trees, and that's how we'll use it here as well. Decision trees on which bagging is performed will be referred to as 'bagged trees' throughout the thesis. Random forests are a slight tweak on bagged trees. They are also designed to improve performance. The next few sections will describe the four methods in greater detail.

4.4 Logistic Regression Description and Example

Logistic regression takes a set of predictor variables associated with an observation as inputs and estimates the probability that the observation belongs to a particular class. Generally an observation is assigned to the class with greatest probability. Hastie, Tibshirani, & Friedman (2016) offer a description of the method. Let N denote the total number of observations in the training set, and y_i, $i = 1 \ldots N$, be the class to which observation i belongs. Let X be the N x p matrix of predictor variables, where p is the number of predictors. A single row in this matrix, denoted x_i, contains the predictor variables for a single observation. To denote the estimated probability that an observation has a positive response, we'll use $p(X) = Pr\ (y = 1 \mid x)$. In this context, both containing a CODIS-eligible DNA profile and containing duplicate profiles are considered positive responses. The probability an observation has a negative response (doesn't contain a CODIS-eligible DNA profile or duplicate profiles) is $Pr\ (y = 0 \mid x)$, or $1 - p(X)$.

Logistic regression assumes that the response variable has two classes. In this case, the model takes the form

$$log \frac{p(X)}{1-p(X)} = \beta_0 + \beta^T x \qquad 4.4.1$$

where β_0 is the intercept and β^T is a vector of coefficients of length p. As previously stated, $p(X)$ is the estimated probability of obtaining a CODIS eligible DNA profile. The ratio $\frac{p(X)}{1-p(X)}$ is called the "odds". For instance, $\frac{p(X)}{1-p(X)}$ could be the odds in favor of obtaining a CODIS-eligible DNA profile. An odds of 1 means both obtaining and failing to obtain a profile are equally likely. An odds greater than 1 means obtaining a profile is more likely, while an odds less than 1 means failing to obtain a profile is more likely. Taking the natural logarithm of the odds produces the log-odds, or "logit". The logit is a linear function of the predictor variables.

By manipulating *4.4.1*, one can solve for $p(X)$:

$$p(X) = \frac{e^{\beta_0+\beta^T x}}{1+ e^{\beta_0+\beta^T x}} \qquad 4.4.2$$

The benefit of *4.4.2* is that one can directly estimate the probability of a particular outcome by plugging in values for the predictor variables. Of course the ultimate goal of classification is to assign an observation to a class. A natural way to do this is to assign observations with an estimated probability of 50% or greater to one class (obtained a CODIS-eligible DNA profile, for instance), and those with an estimated probability of less than 50% to the other (no DNA profile), although other boundaries may be chosen. On the logit scale, observations with a logit of 0 or greater would be assigned to the first class, and those with a logit of less than 0 would be assigned to the other class.

The model is fit by maximizing the value of the likelihood function through maximum likelihood estimation. The likelihood function is displayed in Equation *4.4.3*. The likelihood function is a function of the parameters β_0 and β. Maximum likelihood estimation chooses values for β_0 and β that maximize *4.4.3*. The value $p(x_i)$ is the same as that displayed in Equation *4.4.2*. It is the probability that an observation belongs to a particular class (in our case, contains a CODIS-eligible DNA profile). While we could substitute *4.4.2* into *4.4.3,* we won't for simplicity of display. The value y_i is the true class of the observation (1 if it contains a DNA profile, 0 otherwise). In words, maximum likelihood estimation is maximizing the probability that each observation belongs to the class it does in fact belong to, multiplied by one another. Coefficient values that maximize *4.4.3* are selected. A method of iteratively reweighted least-squares is used to find the exact coefficient values. Details can be found in Hastie et al., (2016).

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \qquad 4.4.3$$

Before we discuss the variable selection process, we'll briefly describe how models can be compared to one another. One such way is on the basis of the Akaike Information Criterion (AIC). The AIC is computed as follows:

$$AIC = -2\,ln(L_{max}) + 2(p + 1) \qquad 4.4.4$$

In the above equation, p represents the number of predictor variables and L_{max} refers to the maximum value of the likelihood function for a particular model. Models with smaller AIC's are preferable to ones with larger AIC's. The equation can be broken into two components. The first component, $-2\,ln(L_{max})$, only decreases as more predictor variables are added to the model. The first component will decrease if an added variable is able to improve the predictions of the model. For instance, if it is able to increase the estimated probabilities from 60% to 80%, and those are predictions for the class the observations do indeed belong to, the first component will decrease. The second component, $2(p + 1)$, only increases as predictor variables are added to the model. The second component can be thought of as a 'penalty term'. The idea is that a predictor variable should be added as long as it decreases the first component by more than the value of the penalty term. If it cannot, the AIC will increase.

With an understanding of the AIC, we can now proceed to variable selection. There are two primary ways to select variables for a model. One is called 'best subsets'. The idea behind best subsets is to compare the performance of models with different numbers of predictor variables. To do this, the best 2-3 models with each number of predictor variables from 1 to p (the total number of predictor variables) are chosen for comparison based on some criteria, such as AIC. The complexity of the model is weighed against its predictive ability. Simpler models (ones with fewer variables) are preferred once additional predictor variables are unable to help the model make more accurate predictions. The other method is called the 'one-variable-at-a-time' approach. There are three such methods: forward selection, backward selection, and stepwise selection. For this thesis we'll use forward selection. Forward selection begins by adopting a model with no predictor variables (the 'null model'). Next, each of the *p* predictor variables are considered for entry into the model. The variable that would decrease the AIC by the largest amount is added to the model. Next, each of the *p*-1 remaining predictor variables are considered for entry into the model. The variable that would decrease the AIC by the largest amount is again added to the model. This process continues until there is no variable that would decrease the AIC.

With a fuller understanding of the method, we'll now proceed to a simple example. Suppose we want to model the probability that a basketball player makes a particular shot. Suppose that a variable selection process has been carried out, and there are two relevant predictor variables: the player's distance from the net (in feet), and whether or not he's playing at home. The first predictor is quantitative while the second is qualitative (home/away). Equations *4.4.5* and *4.4.6* show the model on both logit and probability scales respectively.

$$log\frac{p(X)}{1-p(X)} = 2 - 0.15Distance + 0.5Home \qquad 4.4.5$$

$$p(X) = \frac{e^{(2 - 0.15Distance + 0.5Home)}}{1+e^{(2-0.15Distance+0.5Home)}} \qquad 4.4.6$$

The variable Distance assumes values between 0-30 and Home assumes a value of 1 if playing at home and 0 if away. Probabilities can be calculated by assigning values to the variables. For instance, if a player is 10 feet away and is playing at home, the probability he makes a shot is $\frac{e^{(2\,-\,0.15*10\,+\,0.5*1)}}{1+e^{(2-0.15*10+0.5*1)}} = 0.731$. The number '2' in the right-hand side of both equations is called the intercept. The intercept determines the estimate of the probability when all variables are 0. In this case, the intercept is estimating the probability that a player makes a shot from 0 feet away, playing an away game. The probability he makes the shot is $\frac{e^2}{1+e^2} = 0.881.$ In this case it makes sense for the variables to assume a value of 0, but in some situations it does not. In those situations, the intercept should not be interpreted.

Figure 4.1 displays a graph of Equation *4.4.6.* The top curve corresponds to when the variable Home assumes a value of 1, and the bottom curve corresponds to when it assumes a value of 0. We can clearly see that the probability of making a shot changes as a function of distance for both home and away games. The curves are steepest at the point corresponding to probability 0.5. For the case of home games, this corresponds to a distance of 16.7 on the x-axis. This implies that the largest changes in probability are going to correspond to changes in distance around the vicinity of 16.7. For instance, assuming a home game, increasing the shot distance from 16 to 17 corresponds to a decrease in probability of 0.525 – 0.488 = 0.037. Changing the shot distance from 26 to 27 corresponds to a decrease of 0.198 – 0.175 = 0.023, a smaller change than before.

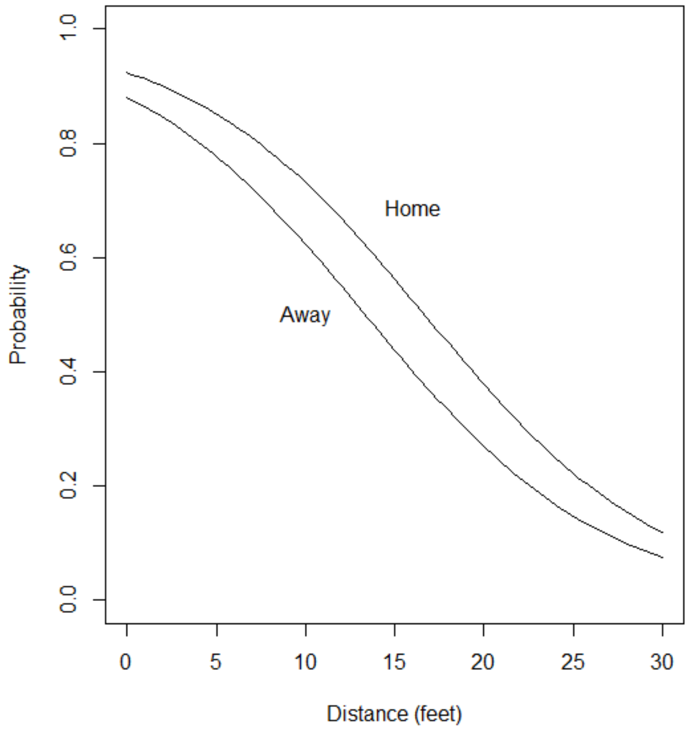

Figure 4.1: Probability of making a shot as a function of distance from the net. The top curve shows the probability when playing at home and the bottom when playing away.

4.5 Decision Tree Description and Example

Decision trees come in two types: regression trees and classification trees. Regression trees handle quantitative responses while classification trees handle qualitative responses. Just like logistic regression, classification trees can be used to assign an observation to one of multiple classes. To make a prediction using a decision tree, a series of questions are answered about an observation. Once enough information is collected, the observation will eventually land in a 'terminal node' on the tree, at which point a prediction is made as to the class of the observation. We'll use a simple example to demonstrate how they work, and then proceed to how they're fit.

In 2015, the website Kaggle.com released a dataset containing information about the individuals who lived and perished in the sinking of the Titanic in 1912 (Titanic: Machine Learning from Disaster, 2015). The dataset contains details about each passenger such as their age, gender, ticket class, number of siblings and parents aboard, as well as whether or not they survived. Figure 4.2 shows a decision tree crafted from the dataset. To make a prediction as to whether or not a passenger survived, we must answer a series of questions beginning at the top of the tree. The first is whether the person is male or female. If the person is female, move right and land in a 'terminal node'. A terminal node is a location on the tree where a final prediction is made (indicated by rectangular boxes in the tree). All other locations are simply called nodes (indicated by diamond-shaped boxes). For this terminal node, the prediction is 'Survived'. There are two numbers below the prediction. The first is the proportion of observations in that node that fall in the class 'Survived'. The second is the percentage of all observations that land in that node. So, of all female passengers, 73% of them survived, and 36% of all passengers were female. Suppose we want to make another prediction for a male passenger. First we move left, and then determine if his age was greater or less than 9.5. If it was greater, we move left again and predict that he died. From the information in that node we see that 61% of passengers on the Titanic were men older than 9.5, and only 17% of them survived.

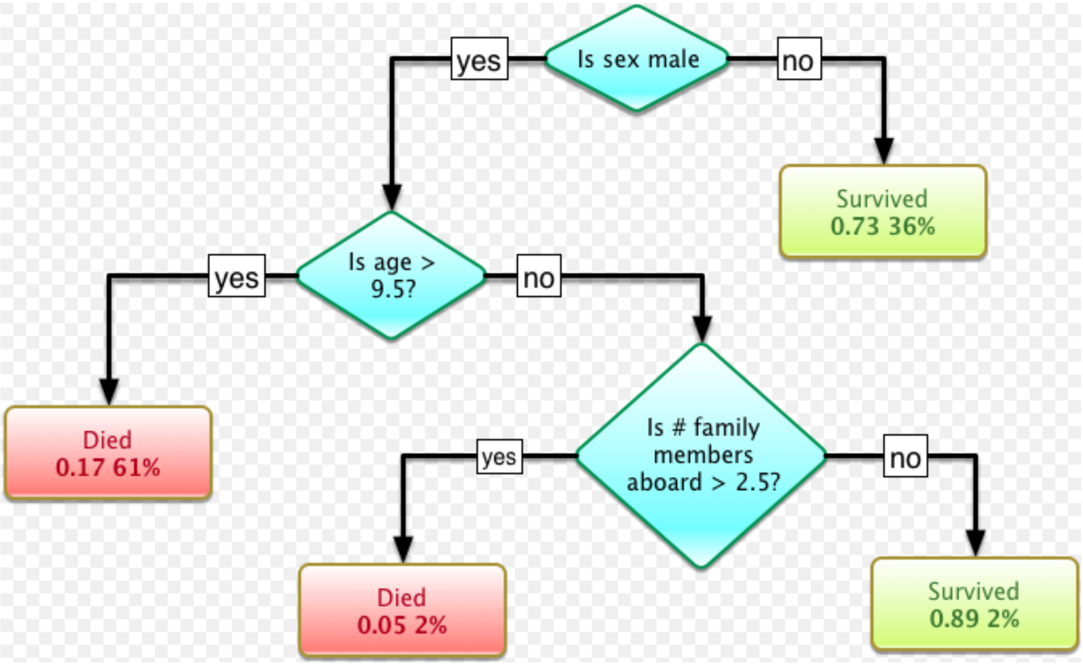

Figure 4.2: Titanic Survivor Decision Tree (Wikipedia, 2017). Retrieved from https://en.wikipedia.org/wiki/Decision_tree_learning#/media/File:Titanic_Survival_Decison_Tree_SVG.png.

There are many great resources on decision trees, such as Shalizi, (2009) and Hastie et al., (2016). Atkinson & Therneau, (2017) are useful because they describe the fitting process in the context of the function we later use to fit the tree, rpart. Next we'll proceed to how decision trees are fit.

The idea behind decision trees is to divide the predictor space (i.e. the range of all predictor variables) into regions (denoted R_m). Theoretically there are many ways of doing this, but the most common is using a method called recursive binary splitting. Recursive binary splitting begins with all observations in the same region. Observations are then split over a single variable in such a way that 'node impurity' (to be defined) within the two resulting regions is reduced as much as possible. We can illustrate this process using the SAK data. For example, the first split in our data might be over whether an SAK was collected within two days or not. Figure

4.3 offers a visual display of this process. There are now two regions after the first split. Next, each of the two is split in such a way that node impurity is further reduced as much as possible. For example, those SAKs collected within two days might be divided based on whether the victim had consensual sex or not. The SAKs collected after two days might be divided based on whether or not the victim is age 23-50. Now with four regions, each is again split in such a way that node impurity is reduced as much as possible. This process continues until some stopping rule is reached, such as when a region has fewer than a certain number of observations (say, 10).

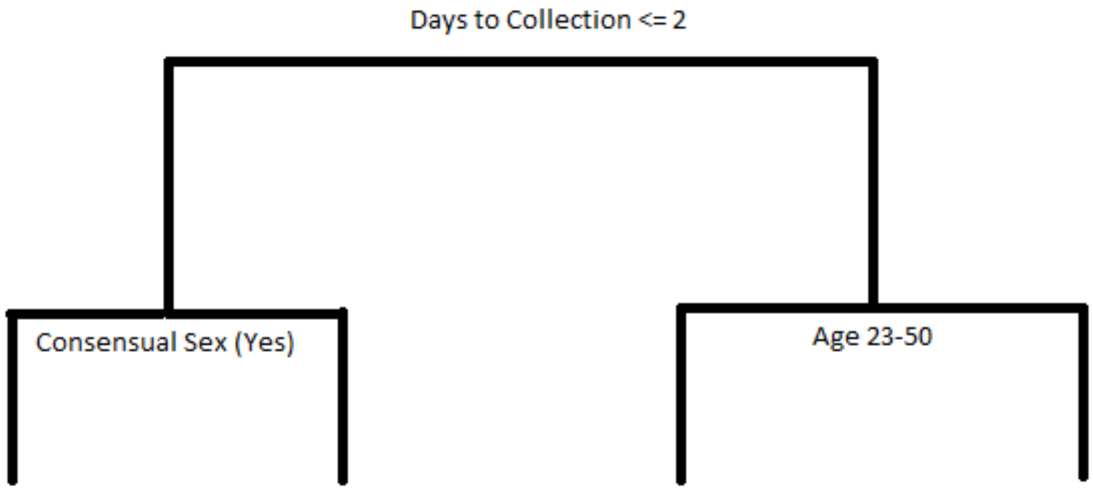

Figure 4.3: Decision Tree in Construction.

When dividing the regions, the goal is to reduce what's called 'node impurity'. There are multiple ways to measure node impurity, but on this occasion we'll use the Gini index. The Gini index is given as

$$\sum_{k=1}^{2} \hat{p}_{mk}(1 - \hat{p}_{mk}) \qquad 4.5.1$$

where $\hat{p}_{mk}$ is the proportion of observations in the mth region that belong to the kth class. In our case, since there are only two classes this can be reduced to *4.5.2*. Depending on the situation we are modeling, we can let k=1 represent “contains CODIS-eligible DNA profile” or “contains duplicate DNA profiles”.

$$2\hat{p}_{m1}(1 - \hat{p}_{m1}) \qquad 4.5.2$$

The Gini index is similar to the concept of variability in the sense that a small Gini index corresponds to low variability within a node. For categorical data, variability is considered low when all or nearly all observations are of the same class. Thus, variability within a node is lowest when $\hat{p}_{m1}$ is close to 0 or 1. The Gini index will also be lowest when $\hat{p}_{m1}$ is close to 0 or 1.

One may wonder why we use the Gini index in the first place. A natural suggestion for a criterion on which to split the regions is the misclassification rate. One could simply divide the observations in such a way that the number of misclassified observations is reduced as much as possible. Although this has intuitive appeal, it has a serious drawback. To illustrate this, suppose there are 100 observations in node A, and the misclassification rate in the node is 0.25. Suppose node A is split into two nodes, A_1 and A_2, each with 50 observations. It is possible that the misclassification rate of A_1 is 0 and the misclassification rate of A_2 is 0.5, and that overall no fewer observations are misclassified. This would be considered a useful split even though the overall misclassification rate wasn’t reduced. Please refer to Atkinson & Therneau (2017) for more information on the subject.

Now we can formally state how recursive binary splitting works to divide the regions. Recursive binary splitting works to minimize *4.5.3*, where M is the total number of regions. The value of M continues to increase until a stopping condition is met. The values being summed are easily recognized from *4.5.2.*

$$\sum_{m=1}^{M} 2\hat{p}_{m1}(1 - \hat{p}_{m1}) \qquad 4.5.3$$

Once a tree is built, there is usually one more step to arrive at a final model: pruning. Pruning is essentially the reverse process of building the tree, whereby terminal nodes are actually removed from the tree. The reason we do this is that the original tree may have overfit the data. That is, we may have found a tree that does a good job classifying observations that were used to train it, but is unlikely to do a good job classifying observations that weren't. Trees constructed in the initial building phase may have as many as 100 or more terminal nodes. The goal of pruning is to reduce the number of terminal nodes to a more manageable number without sacrificing model performance.

Pruning creates a 'subtree' of the original tree. When pruning a tree, nodes must be removed from the bottom up; that is, at a given point in time, only those splits that resulted in exactly two terminal nodes can be undone. As with building the tree, we need a criterion to decide which nodes to prune. Typically, if the goal of the model is to minimize the misclassification rate, nodes are removed in such a way as to minimize the proportion of misclassified observations. Those nodes that lead to the smallest increase in misclassification rate are collapsed first (removing nodes in general increases the misclassification rate, whereas adding nodes decreases it). It can be difficult to know how many terminal nodes the final tree should contain, so it is useful to have a way to estimate how well trees with different numbers of terminal nodes perform on real data. Leave-one-out cross-validation (LOOCV) is useful for this purpose. Leave-one-out cross-validation is a kind of 'resampling method'. Such methods were devised because they allow us to obtain information about the performance of a model without having to collect additional observations. In leave-one-out cross-validation, *n* statistical models are created, where *n* is the number of observations in the data set. Each of the models is fit using all observations from the data except one. Each observation is excluded from exactly one model. Then, each model makes a prediction for the class of the observation that was not used to train the model. This results in *n* predictions. The proportion of incorrect predictions is the LOOCV error rate.

But how exactly does LOOCV work in this context? The idea is to minimize the misclassification rate while also assuming there is a penalty associated with adding additional nodes to the tree. For every additional node added to the tree, we'll assume there is a penalty α associated with it. The penalty α can also be referred to as a complexity parameter (cp). The result is that we're trying to minimize the following:

$$C_\alpha(T) = M(T) + \ \alpha|T| \qquad 4.5.4$$

In the above equation, $C_\alpha(T)$ represents the cost of a tree *T* for a pre-defined value of α. The goal is to find a tree *T* with minimal cost for a given α-value. The term *M*(*T*) represents the misclassification rate of tree *T*, and |*T*| represents the number of terminal nodes in the tree.

As α varies, the tree that minimizes *4.5.4* varies as well. In practice, cross-validation is used to find the optimal value of α. In this case we'll use LOOCV. The process works as follows:

1. Select a set of values for α.
2. Use LOOCV to build a set of *n* trees for a particular α-value. Specifically, use all observations except one to build a set of *n* models. Each observation is excluded from exactly one model. Each of these *n* trees are constructed so as to minimize

4.5.4. Each of these *n* trees then make a prediction for the single observation that wasn't used to train the tree. The proportion of misclassified observations is the LOOCV error rate.

3. Repeat for every value of α selected in Step 1.
4. Use the results of LOOCV to select a final value of α.
5. Prune the original tree to minimize *4.5.4* based on that value of α.

More details on how values for α are selected can be found in Atkinson and Therneau (2017). In the next section we'll discuss bagged trees.

4.6 Bagged Trees Description

Bagging is a method that can be applied to a variety of statistical methods in order to improve performance. It is commonly used in the context of decision trees, and that's how we'll apply it in this thesis. The word 'bagging' is a blend of *bootstrap* and *aggregation*. Bootstrapping is another kind of resampling method, like LOOCV, that allows additional information to be gained from the original dataset without the need to obtain additional observations. A bootstrap sample is one in which *n* observations are drawn with replacement from the original dataset. Many observations will appear multiple times in a bootstrap sample, while many won't appear at all. To create a bagged decision tree simply take multiple bootstrapped samples from the original dataset and fit an unpruned decision tree to each one (meaning each tree will likely have many terminal nodes). This set of decision trees is the bagged decision tree. Generally, at least 100 bootstrap samples are taken, and thus at least 100 unpruned trees are constructed. While taking 100 bootstrap samples is sufficient for good performance in most situations, there is no drawback to taking more (James, Witten, Hastie & Tibshirani, 2015).

To make a prediction for an observation, each of the bagged trees makes a prediction as to which class the observation belongs. The class that wins a majority of the votes is the final predicted class. Although each individual tree will have high variance, the variance of predictions averaged over multiple trees is much less. This can be motivated intuitively by the fact that a given random variable X can have variance σ^2, but the average of a random sample, $\bar{X}$, has variability σ^2/n. In a similar fashion, bagging also reduces the variability of predictions, and thus improves accuracy. Bagging only works to reduce variability to a point, though, because the trees built from bootstrapped samples are correlated with one another since the bootstrapped samples all come from the same dataset.

One of the disadvantages of bagged trees is that there is no way to display them graphically apart from graphing the many individual trees. The primary benefit of bagged trees is their ability to improve predictive performance over a regular decision tree (James et al., 2015). Next we'll discuss random forests.

4.7 Random Forests Description

In addition to bagging, there is yet another method with the potential to improve the predictive performance of the decision tree. This method is called random forests. It is identical to bagging except for one tweak: when building the trees from the bootstrap samples, at each split in the tree, only a subset of the predictors is considered on which to make the split. For example, if there are 16 predictor variables, 4 might be chosen at random on which to make the split. The justification for doing this is that it makes the trees less correlated with one another. With bagging, if there is a single strong predictor it is likely that every tree in the bootstrapped samples makes the first split on this predictor. The result is that the trees will look very similar to one another. Using random forests prevents this by considering only some of the predictors at each split.

The rationale for making the trees less correlated is that averaging highly correlated values does little to improve predictive performance (James et al., 2015). To motivate this idea intuitively, suppose you wanted to predict a person's weight based on their height in inches. It would be easy to build a linear regression model to do so. Then suppose you are given a second predictor variable: each person's height in centimeters. This information would be perfectly correlated with the first predictor, height in inches, and would thus be redundant. Adding this variable would not add any additional information to the model, and would not improve model performance.

4.8 Pros and Cons of the Methods

Each classification method has pros and cons. Generally no single method will perform the best in all situations. Logistic regression has the advantages of being relatively easy to understand and having a long history of use. Decision trees are also easy to understand, can easily handle a large number of predictor variables, and have a nice visual display. The main disadvantage is that they tend not to have as great a predictive ability as other classification methods (James et al., 2015). The advantage of using bagging and random forests is that these methods frequently leads to improved predictive performance. The disadvantage is that the original method loses its interpretability, thus making it difficult to communicate to others how predictions are made. Both bagging and random forests have no simple graphical display. Typically, random forests will outperform bagged trees. In addition, different methods make different assumptions about the shape of the 'decision boundary', discussed next.

4.9 Decision Boundaries

The decision boundary is the location in the predictor space that divides predictions of one class from predictions of another. It is easiest to understand graphically. Figure 4.4 shows an example where there are two predictor variables, x_1 and x_2, and two classes for the response variable, yellow and green. A distinction can be made between the true decision boundary and the estimated decision boundary. The true decision boundary shows the ideal way to make predictions for the observations based on your stated goals, such as minimizing the misclassification rate. The true decision boundary in the figure is represented by the shape created by the adjacent colors. In the top row this shape is a line, and in the bottom it is two perpendicular half-lines.

The estimated decision boundary is shown in black. This shows the way a particular model assigns observations to classes. The top left image shows a logistic regression model's decision boundary. Logistic regression models produce linear decision boundaries, and when the true decision boundary is linear they perform very well. In this case it matches the true decision boundary. The top right image shows a decision tree's attempt to model a linear decision boundary. It performs less well because decision trees cannot produce linear decision boundaries. Rather, they divide the predictor space into rectangular regions, as the figure shows.

The bottom row shows a situation where the true decision boundary is non-linear. The bottom left image shows a logistic regression model's decision boundary. Clearly the logistic regression model performs very poorly. The decision tree bottom right, on the other hand, performs very well.

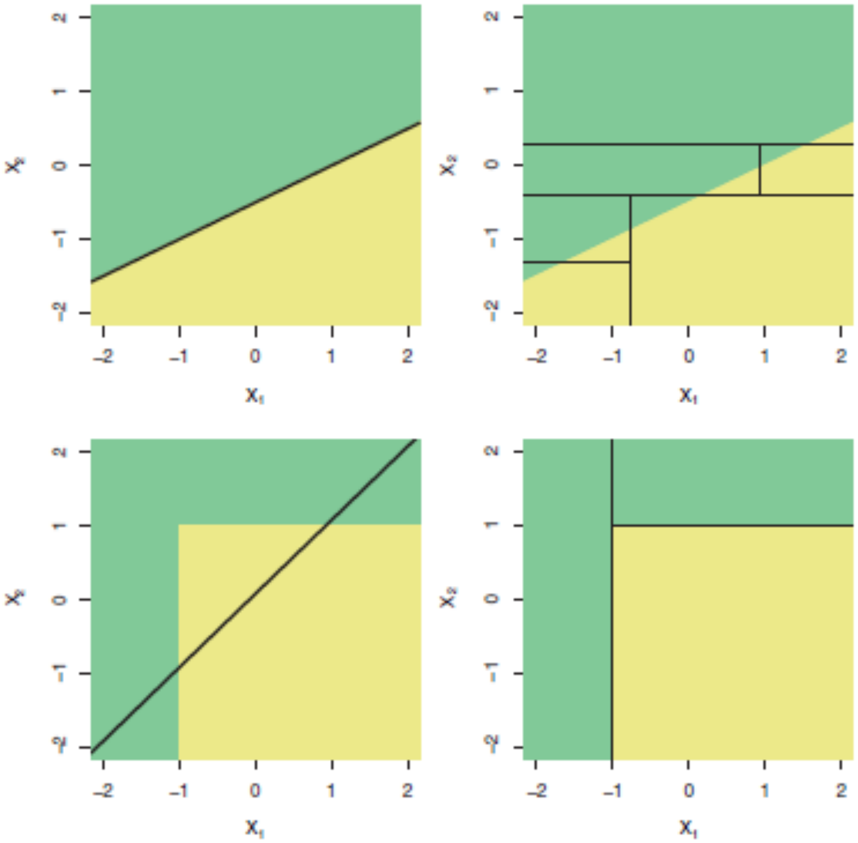

Figure 4.4: *Top Row*: A two-dimensional classification example in which the true decision boundary is linear, and is indicated by the shape created by the shaded regions. A classical approach that assumes a linear boundary (left) will outperform a decision tree that performs splits parallel to the axes (right). *Bottom Row*: Here the true decision boundary is non-linear. A linear model is unable to capture the true decision boundary (left), whereas a decision tree is successful (right). Image obtained from James et al., (2015). Reprinted with permission of the authors.

We can revisit our logistic regression model example from Section 4.4 and analyze its decision boundary. It can be shown that if the goal is to minimize the misclassification rate, observations should be assigned to the class to which they have the highest probability of belonging (James et al., 2015). In the case of a binary response, this means an observation should be assigned to the class to which it has a greater than 50% chance of belonging. In order to find the decision boundary, one simply needs to set $p(X) = 0.5$ in Equation *4.3.2* and solve for the variable Distance. Alternatively, one could set $\log\left(\frac{p(X)}{1-p(X)}\right) = 0$ in Equation *4.3.1* and solve for Distance, since a value of 0.5 on the probability scale corresponds to a value of 0 on the logit scale. The decision boundary will change based on the value of Home. Figure 4.5 shows the decision boundary for this example. The top and bottom images show the decision boundary when playing away and home respectively. In this case the decision boundary is represented by a point rather than a line. When playing away, the decision boundary corresponds to a value of Distance = 13.3, and when playing at home it corresponds to Distance = 16.7.

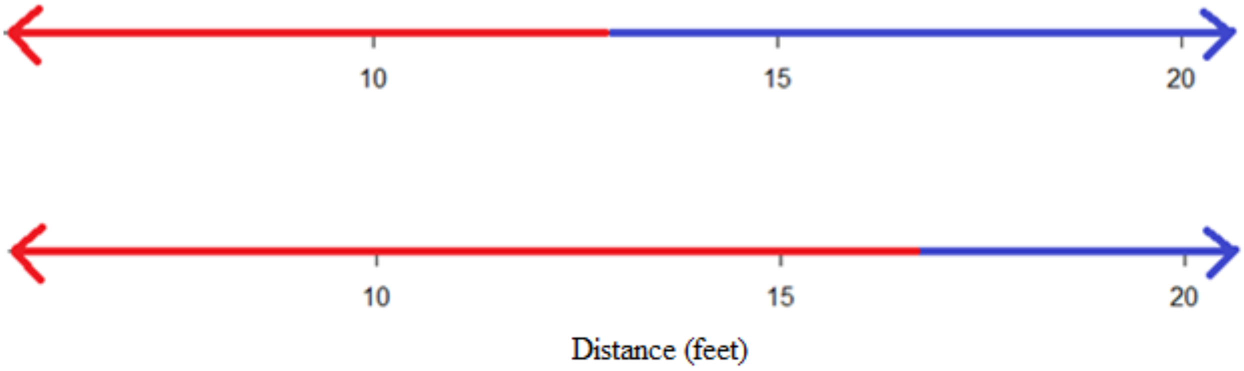

Figure 4.5: *Top*: Decision boundary when playing away. *Bottom*: Decision boundary when playing at home. Shots made from distances falling in the red shaded area are predicted to be made, while those in blue are predicted to miss.

Figure 4.6 shows the graph of this example with vertical lines drawn at the decision boundaries of 13.3 and 16.7. These lines both end when they hit the curve at probability 0.5.

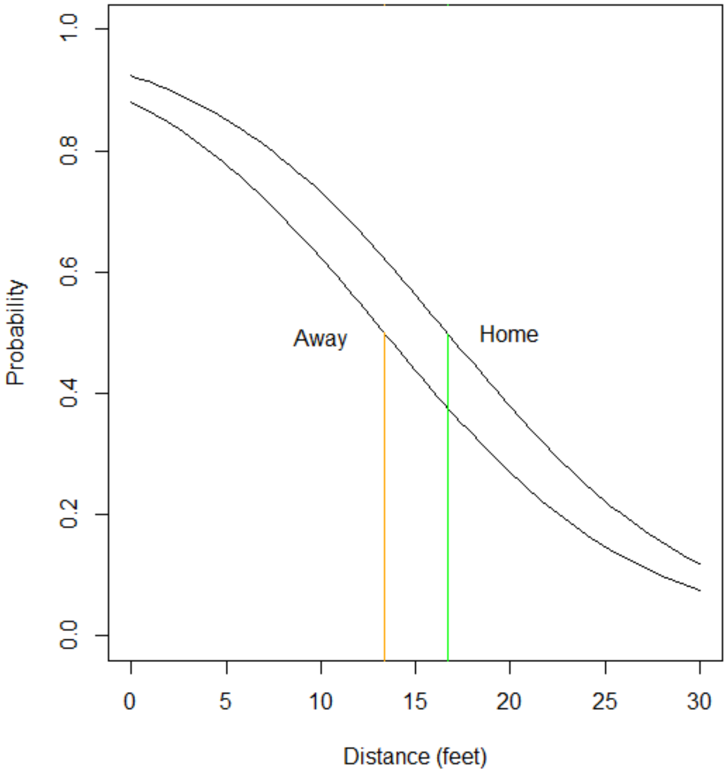

Figure 4.6: Decision boundaries are shown to correspond to probabilities of 0.5.

Bagging and random forests do not produce decision boundaries that can be displayed in the form of Figures 4.4 and 4.5. They can improve the predictive accuracy regardless of the underlying decision boundary.

CHAPTER 5 PREDICTING THE PRESENCE OF A CODIS-ELIGIBLE DNA PROFILE

5.1 Introduction

One of the goals of the Sexual Assault Kit Testing Initiative was to determine the factors that are associated with an SAK containing a CODIS-eligible DNA profile. Kerka et al., (2018) proposed a logistic regression model to estimate probability that a kit would contain such a profile. In this section several different classification methods will be compared to determine their performance relative to the logistic model presented in the paper, and reprinted here. This chapter begins by looking at the individual relationship between the response variable and the predictor variables, followed by fitting a logistic regression model, a decision tree, a bagged tree, and a random forest, and ending with a comparison of the methods.

5.2 Examining Relevant Variables

The first step was to determine how strong the relationship was between the individual predictor variables and the response. Figure 5.1 shows the relationship between the response and four predictor variables (starting in the top left and moving clockwise): days to kit collection, occurrence of consensual sex, number of samples run, and years to kit submission to laboratory. The y-axis in each of the four graphs is the percent of kits in the sample that contained a CODIS-eligible DNA profile. The top left graph shows a clear relationship between the number of days to kit collection and obtaining a profile. The longer the victim waits, the less likely it is that the kit will contain a DNA profile.

The top right pane shows the impact of consensual sex. A greater proportion of SAKs from victims who reported consensual sex contained at least one CODIS-eligible DNA profile than from those who did not.

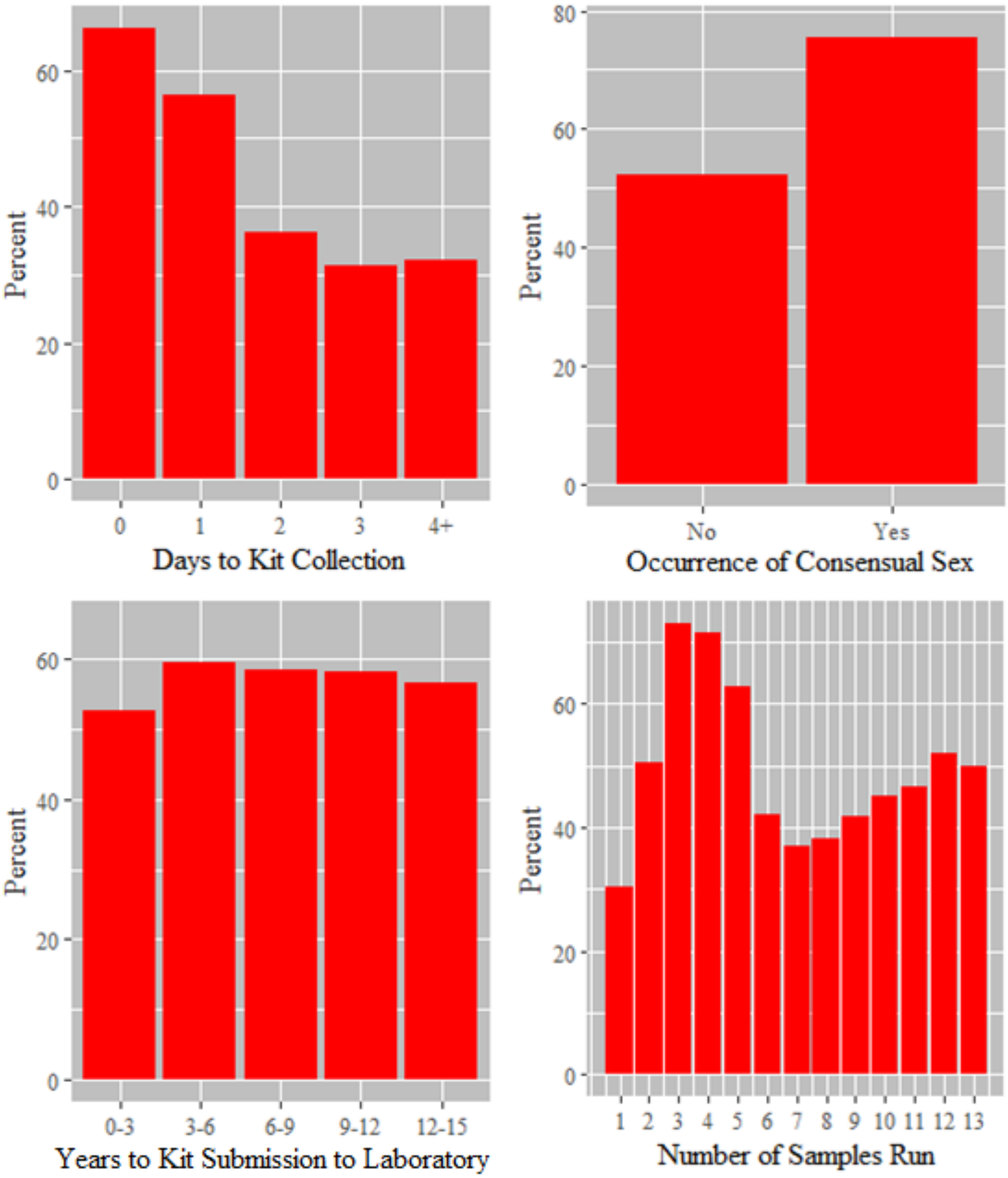

Figure 5.1: The y-axes display the percent of kits containing at least one CODIS-eligible DNA profile. The x-axes display days to kit collection, the occurrence of consensual sex, years to kit submission to laboratory and number of samples run.

The bottom right pane shows the impact of the number of samples run. Intuitively, one would expect that SAKs with more samples run would more frequently contain a profile. The figure shows a more complicated story. We see that the percentage of SAKs with at least one profile appears to increase starting at one sample and then plateaus at 3-4. It begins to drop starting at five. This seems to suggest that if 3-4 samples are run, it is likely that a profile will be found. If more than five samples need to be run there is less certainty that DNA will be found. Although this may be rather counterintuitive, there is an explanation for it in terms of the laboratory testing procedure. If the victim reports one perpetrator and no consensual sex, the laboratory will only test 3-5 samples. Since there is only one sexual partner, it is relatively easy to extract a DNA profile from the samples. If the victim reports having recent consensual sex or multiple assailants, the laboratory will typically test more than five samples. However, the presence of DNA from multiple individuals makes it difficult to isolate the DNA profile of any one person. Thus, it is slightly less likely that a DNA profile will be found.

The bottom left pane shows the impact of years to kit submission to laboratory. For the ranges 3-6, 6-9, 9-12, and 12-15, the percent of kits with CODIS-eligible DNA profiles drops as the number of years to submission increases.

The relevance of victim age was also explored. Table 5.1 shows the percent of the time kits from victims of various age ranges contained CODIS-eligible DNA profiles. Victims of age 0-7 had a surprisingly low rate of 5.3%. It is unclear why this is the case. One possibility is that young children on average wait longer to have their kits collected, possibly due to their inability to communicate. Indeed, there is some truth to this explanation. In the sample, the average wait time to kit collection for all victims was 0.63 days. For children, the average wait time was 0.88 days. Other research has shown a similar trend (Peterson et al., 2012). Another point worth noting is that the percentage of the time the date of assault was missing or unknown was higher for children than for all other age groups, also shown in Table 5.1. This suggests that the average wait time to kit collection for children may be even higher than 0.88, although this is speculative. More research will have to be done to determine what accounts for the low proportion of profiles from child victims.

Age Range of Victim	Sample Size	Percent of Kits Containing CODIS Eligible DNA Profile	Percent of Time Assault Date was Missing/Unknown
0-7	114	5.3	20.18
8-15	361	44.9	2.49
16-22	771	65.4	1.43
23-50	1031	63.8	2.13
51+	73	35.6	8.22

Table 5.1: Percent of Kits Containing CODIS-Eligible DNA By Victim Age.

Figure 5.2 shows how the response varied based on whether or not certain forensic samples were analyzed. A forensic sample is a potentially useful predictor if there’s a large difference in the percent of SAKs that contain a CODIS-eligible DNA profile when it is and is not analyzed. For instance, for the vaginal sample, a DNA profile was found a greater percent of the time when it was analyzed than when it wasn’t. Thus, the vaginal sample may be a useful predictor. Likewise with the neck/ear swab. The pubic hair comb swab and the underwear non-crotch cutting may also be useful, although, if they were analyzed, they indicate that a profile was less likely to be found. As a note, this should not be misinterpreted to mean these samples shouldn’t be collected. It simply means the cases where they weren’t collected were more likely to contain a DNA profile.

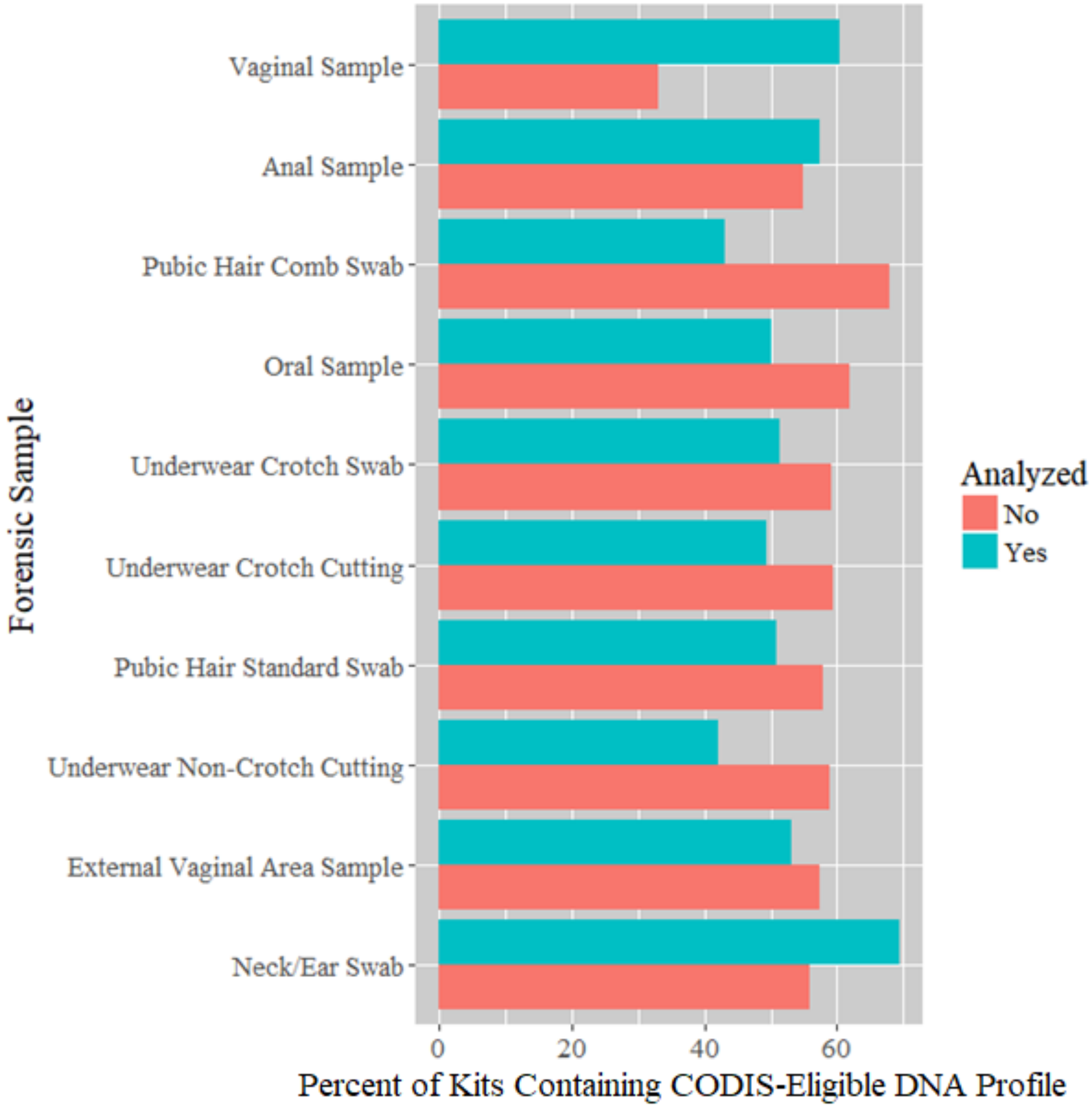

Figure 5.2: For each sample, the cases were separated based on whether or not it was analyzed. The percent of kits containing a CODIS-eligible DNA profile is shown.

5.3 Logistic Regression

The ‘step’ and ‘glm’ functions in R were used to perform forward selection and model fitting (R Core Team, 2017). The variables considered for entry into the model were: days to kit collection, years to kit submission to laboratory, consensual sex (yes/no), number of samples run, age of victim (categorical with classes 0-7,8-15,16-22,23-50,51+) and ten categorical variables corresponding to whether or not the top ten most commonly analyzed forensic samples were analyzed in a particular case (vaginal sample, anal sample, pubic hair comb swab, oral sample, underwear crotch swab, underwear crotch cutting, pubic hair standard swab, underwear non-crotch cutting, external vaginal area sample, and neck/ear swab). Table 5.2 shows the various steps in the model selection process, along with the AIC values at each step. In the first step the model contained 13 variables, with all but underwear non-crotch cutting significant at the 0.05 level. It contained all the age variables and six of ten of the forensic samples. While most of the forensic samples were significant at the 0.05 level, there were theoretical reasons why they should be removed from the model. The original purpose of building the logistic regression model was to estimate the probability that an SAK would contain a CODIS-eligible DNA profile. The idea was that a lay-person, possibly not well-educated in statistics, may use the model for this purpose. From his/her perspective, it may not make sense to ask about only six samples, particularly in light of the fact that there are over 100 that can be collected. The anal sample, collected almost 85% of the time, would not even be included in the model. The vaginal sample was included, but it may not make sense to ask if the vaginal sample was collected if, for instance, the victim was a male. The answer doesn’t seem to be ‘no’, but rather it seems the question doesn’t apply. If all of the top ten most commonly collected samples were highly

First Step:			
Coefficients	Estimate	Std Error	P-value
Intercept	-2.05	0.43	0.000
Days to Collection	-0.44	0.06	0.000
Age 8-15	2.46	0.41	0.000
Age 16-22	2.9	0.4	0.000
Age 23-50	2.95	0.39	0.000
Age 51+	2.46	0.5	0.000
Consensual Sex	0.91	0.13	0.000
Samples Run	-0.11	0.04	0.002
Vaginal Sample	1.39	0.2	0.000
Pubic Hair Comb Swab	-1.24	0.13	0.000
Oral Sample	-0.52	0.12	0.000
Underwear Crotch Cutting	-0.3	0.15	0.042
Underwear Non-Crotch Cutting	0.29	0.2	0.134
Neck/Ear Swab	0.55	0.2	0.008
AIC = 2090			
Second Step:			
Coefficients	Estimate	Std Error	P-value
Intercept	-0.38	0.42	0.372
Days to Collection	-0.43	0.06	0.000
Years to Submission	-0.06	0.02	0.001
Age 8-15	2.49	0.4	0.000
Age 16-22	2.88	0.39	0.000
Age 23-50	2.79	0.39	0.000
Age 51+	2.24	0.49	0.000
Consensual Sex	0.84	0.12	0.000
Samples Run	-0.24	0.02	0.000
AIC = 2261			
Third Step:			
Coefficients	Estimate	Std Error	P-value
Intercept	-1.65	0.4	0.000
Days to Collection	-0.42	0.06	0.000
Years to Submission	-0.04	0.02	0.008
Age 8-15	2.23	0.39	0.000
Age 16-22	2.71	0.38	0.000
Age 23-50	2.7	0.38	0.000
Age 51+	2.03	0.47	0.000
Consensual Sex	0.74	0.12	0.000
AIC = 2376			

Table 5.2: Steps in the logistic regression model selection process.

significant, a better case could be made for including them. In order to avoid confusion, though, the forensic samples were removed from the model.

In the second step, forward selection was rerun considering all variables except the forensic samples. We can see that the AIC increased from 2090 to 2261 between the first and second step. This is a moderate cost, but having a simpler model is consistent with the rest of our model-building goals. In this model, the years to kit submission makes an appearance. All eight variables were significant at the 0.05 level. The large p-value for the intercept suggests that the intercept may be zero, although this is of little inherent significance. More importantly we want to verify that the coefficients of the variables are not zero, since, if they were, those variables would have no impact.

In this model, the coefficient estimates appear sensible except that the variable 'samples run' has a negative coefficient. This implies that running more samples decreases the probability of obtaining a CODIS-eligible DNA profile. By this logic, the probability of obtaining a profile is maximized by running only one sample. This is an illogical result, and a strong reason to exclude it from the model. In the third and final step, the number of samples run was removed.

In the final model, all coefficients were significant at the 0.05 level. Again the AIC increased, this time from 2261 to 2376. However, this was necessary to ensure our model didn't yield illogical results. Years to kit submission to laboratory had the largest p-value of all predictor variables. This is consistent with the exploratory work, as it appeared that it was a relatively weak predictor. Nonetheless, it was still significant at the 0.05 level.

Next we'll examine the final fitted model more closely. The model is represented in Equations *5.3.5* and *5.3.6,* on probability and logit scales respectively.

$$p(X) = \frac{e^{(-1.65-0.42D-0.04Y+2.23A_{8-15}+2.71A_{16-22}+2.70A_{23-50}+2.03A_{\geq 51}+0.74C)}}{1+e^{(-1.65-0.42D-0.04Y+2.23A_{8-15}+2.71A_{16-22}+2.70A_{23-50}+2.03A_{\geq 51}+0.74C)}} \qquad 5.3.5$$

$$log\left(\frac{p(X)}{1-p(X)}\right) = -1.65 - 0.42D - 0.04Y + 2.23A_{8-15} + 2.71A_{16-22} + 2.70A_{23-50} + 2.03A_{\geq 51} + 0.74C \qquad 5.3.6$$

The variables in the equations were abbreviated using the first letter of their name in Table 5.2. In the equations at most one age variable assumes a value of one (if a victim is in that age category). If all age variables are set to zero this implies the victim is aged 0-7. The variable consensual sex takes on a value of one if the victim had consensual sex and zero otherwise.

Although the logit form displayed in *5.3.6* is often useful, we'll restrict our interpretation to the probability scale of *5.3.5.* The intercept of -1.65 gives us the probability when all variables are zero- that is, zero days to kit collection, zero years to kit submission, age 0-7 and no consensual sex. This probability is $\frac{e^{-1.65}}{1+e^{-1.65}}$ = 0.161. The primary reason this probability is so low is because victims age 0-7 had an extraordinarily low rate of CODIS uploads (see Table 5.1). They also, of course, hadn't had consensual sex.

Figure 5.3 shows how the probability of obtaining a profile varies as a function of days to kit collection. In order to make a graphical display, values had to be chosen for certain variables. The top graph is broken down by whether or not the victim had consensual sex, and assumes a value of 8.5 years to kit submission to laboratory and a victim aged 23-50. The bottom graph is broken down by victim age and assumes no consensual sex and 8.5 years to kit submission to laboratory. To reiterate a point from Chapter 4, the slope of the curve is steepest where the probability is 0.5. In the top figure, for instance, this corresponds to 1.5 days for a victim who hasn't had consensual sex. In this instance, if a victim waits two days instead of one to have the kit collected, the probability drops by 0.104 that it will contain a profile. If the victim waits six days instead of five the probability only drops by 0.058. As a matter of fact, there is a convenient

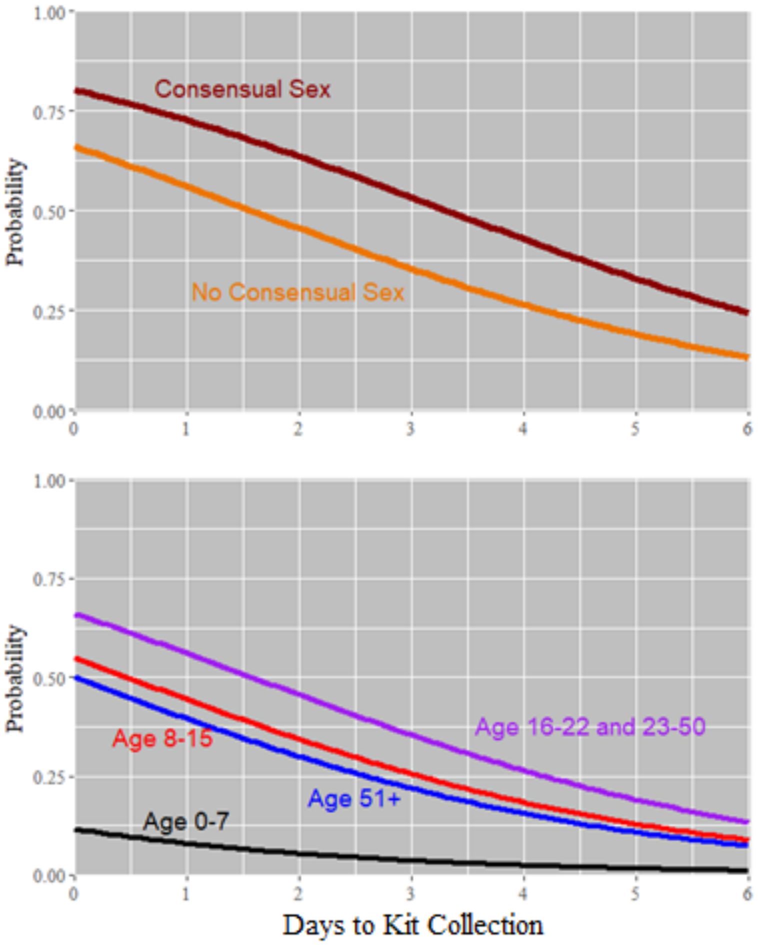

Figure 5.3: *Top*: Probability of obtaining a CODIS-eligible DNA profile plotted against days to kit collection. It is broken down by whether or not the victim had consensual sex, and assumes 8.5 years to kit submission to laboratory and a victim aged 23-50. *Bottom*: Same plot, broken down by victim age. Assumes no consensual sex.

rule describing approximately how much the probability will change by changing one variable one unit. Gelman and Hill (2007) describe the "divide-by-four" rule, which serves exactly this purpose. If you take a coefficient and divide it by four, that's approximately how much the probability will change by changing a variable by one unit at the steepest point on the curve (emphasis on *steepest point*). The coefficient for days to kit collection is -0.42. Dividing by four makes -0.105, which is remarkably close to the 0.104 above. Of course, this rule doesn't apply to categorical variables. It also becomes less accurate the further you move away from the steepest point.

In the bottom graph we can see that kits from those age 16-22 and 23-50 have the greatest probability of producing a profile. Those age 0-7 have at most a chance of 0.125.

Figure 5.4 is similar to 5.3 except it plots years to kit submission to laboratory on the x-axis. The top graph is again broken down by consensual sex and assumes zero days to kit collection and a victim aged 23-50. The bottom graph is broken down by victim age and assumes zero days to kit collection and no consensual sex. The curves in Figure 5.4 are much less steep than the curves in 5.3, indicating that the number of days to kit collection has a greater impact on the response. We can again divide the coefficient for years to kit submission by four to get an estimate of how much the probability changes. The probability will change by about -0.04/4 = -0.01 for every additional year. Next we'll proceed to build a decision tree.

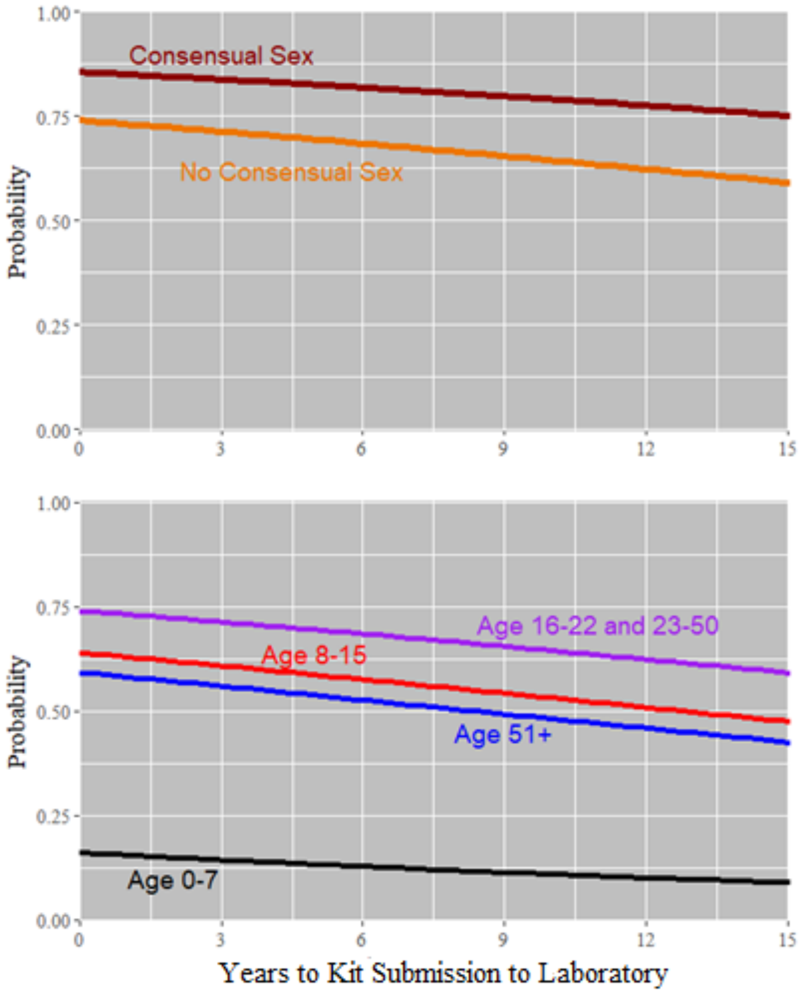

Figure 5.4: *Top*: Probability of obtaining a CODIS-eligible DNA profile plotted against years to kit submission to laboratory. It is broken down by whether or not the victim had consensual sex, and assumes zero days to kit collection and a victim aged 23-50. *Bottom*: Same plot, broken down by victim age. Assumes no consensual sex.

5.4 Decision Tree

The ten individual forensic samples were excluded from the decision tree due to the same concerns as with the logistic regression model. The number of samples run, however, was included. The reason for including the number of samples run is that decision trees are not linear classification methods (recall Figure 4.4), and thus are not necessarily going to produce illogical results the way the logistic regression model did.

The ‘rpart’ package in R was used to fit the tree (Therneau et al., 2017). The function ‘rpart’, contained within the rpart package, was coded to stop dividing the regions when they had 10 or fewer observations. The gini index was used as the measure of node impurity. The full tree contained 168 terminal nodes. Since this is far too many terminal nodes, the next step was to prune the tree.

As you may recall from Section 4.5, trees are pruned with the help of a complexity parameter α. Leave-one-out cross-validation (LOOCV) was used to find the optimal value of α. Figure 5.5 shows the misclassification rate for different values of α as estimated by LOOCV. The lower x-axis of the figure shows different values of the complexity parameter α. The upper x-axis shows the corresponding number of terminal nodes of the tree associated with that α-value. Note that the x-axis is not linear. The y-axis displays the relative error rate. The error rate is relative to that of the one-node tree. A tree with one node can only make one prediction for all observations, so naturally it will predict that all observations are of the majority class. In this case, it will predict that all SAKs contain a CODIS-eligible DNA profile, and it is correct 60% of the time because 60% of the observations in the dataset belong to that class. The error rate of the one-node tree is scaled to be one on the y-axis. As the figure shows, the LOOCV error rate initially declines as nodes are added. After a point, though, it generally begins to increase.

LOOCV seems to suggest that a five-node tree is appropriate. The five-node tree is displayed in Figure 5.6.

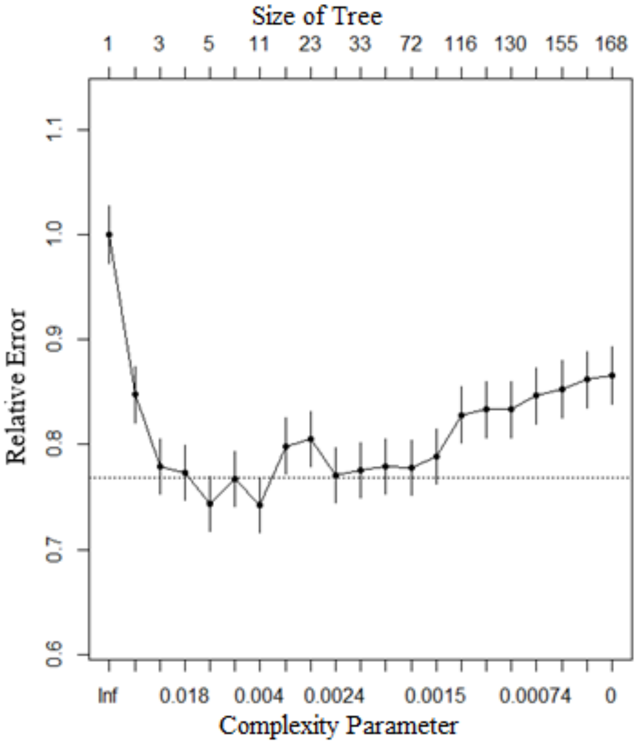

Figure 5.5: Estimated misclassification rate from leave-one-out cross validation.

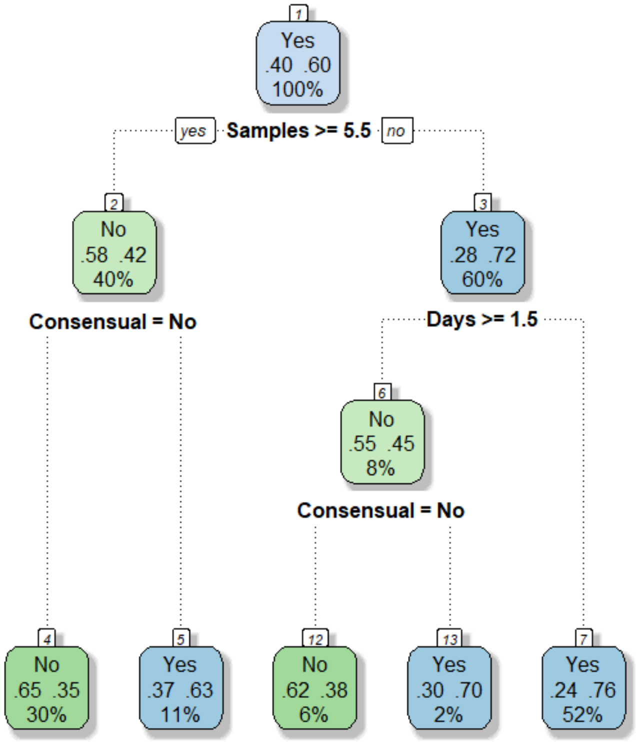

Figure 5.6: Five-node decision tree.

The five-node decision-tree contains only three variables: consensual sex, number of samples run, and days to kit collection. Interestingly, the very first split was over the number of samples run, a variable excluded entirely from the logistic regression model. The tree decided that splitting SAKs into two groups was appropriate based on whether five or fewer samples were collected or six or greater were collected. This is interesting because this corresponds to an actual difference in laboratory testing procedure. Typically, if a victim does not report consensual sex and had only one assailant, the laboratory will run five or fewer samples. If the victim reported consensual sex or multiple assailants the laboratory will typically run six or more samples. Notably absent from this model are victim age and years to kit submission to laboratory. Another striking fact about the model is its simplicity. Five terminal nodes is quite small. This makes for an interesting comparison to the logistic regression model.

As with the decision tree discussed in Section 4.5, if the split condition is met, move left. Otherwise, move right. For instance, if six or more samples were collected we would move left. If the victim had consensual sex we would move right, landing in node 5. We see that 11% of all observations fall in this node. That is, 11% of all SAKs had six or more samples run and had a victim who had consensual sex. Of these SAKs, 63% contained at least one CODIS-eligible DNA profile and 37% did not. The final prediction for this node is “Yes”. For making predictions about future SAKs, 63% and 37% can also be interpreted as the estimated probability of obtaining a profile and failing to obtain a profile respectively for SAKs that land in that node.

5.5 Bagged Trees

As with the decision tree, the number of samples run was included while the individual forensic samples were excluded. For most purposes taking 100 bootstrap samples is sufficient for good performance. Just to be safe, we took 500, since there’s no drawback to taking more. The model was built using the ‘randomForest’ package in R (Liaw & Wiener, 2002).

As previously mentioned, one of the disadvantages of bagged trees is that there is very little to show in terms of graphical output. There is, however, still a way to assess the importance of the variables. One can assess the importance of the variables in the following way: for each split in the tree, add the values of the gini index for the two resulting nodes; then sum the values corresponding to each variable over which a split was made; finally, the total value for each variable is averaged over all 500 bagged trees. Figure 5.7 displays this graphically. A large value on the x-axis indicates an important variable. Strangely, the variable years to kit submission to laboratory appears to be by far the most important variable. The reason for this is that the bagged decision trees are unpruned, meaning many of them contain well over 100 terminal nodes. For these trees many splits are made at the bottom over years to kit submission to laboratory. The individual bagged trees wildly overfit the model by splitting too many times over this variable. The result is an inflated sense of the true importance of the variable. For this reason measurements of variable importance in bagged trees can be quite unreliable. Excluding years, though, the number of samples run appears to be most important. Next we'll proceed to fit a random forest.

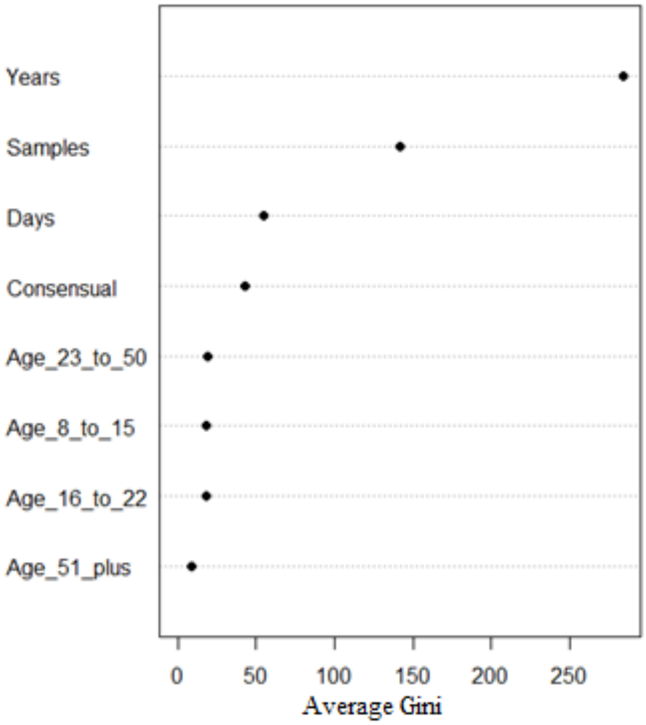

Figure 5.7: Variable importance measure of bagged tree.

5.6 Random Forest

As you may recall from Section 4.7, random forests are identical to bagged trees except for one tweak: when building the tree, at each split only a subset of the predictors is considered on which to make the split. For instance, since there are eight predictors in our model, a random set of three might be chosen at each split. The purpose is to make the trees less correlated.

The randomForest package was used to fit a random forest as well (Liaw & Wiener, 2002). Just like with the bagged tree we took 500 bootstrap samples and built 500 unpruned decision trees, this time considering only three of the predictors at each split. Just as with the bagged tree there is little to display in terms of graphical output. The importance of the variables, however, can still be assessed the same way as in bagged trees. Figure 5.8 shows the variable importance plot for the random forest model. We see that considering only three of the eight variables at each split was sufficient to reign in the unrelenting splitting over years to kit submission that cursed the bagged model. Besides the years to kit submission to laboratory, the variables largely retained the same order of importance as in the bagged model. The number of samples was most important, followed by days to kit collection, consensual sex, and victim age.

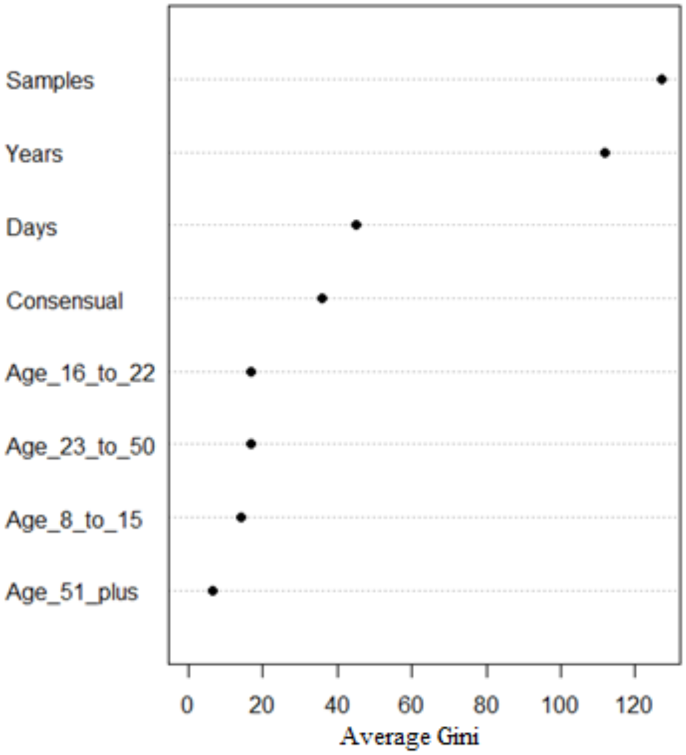

Figure 5.8: Variable importance measure of random forest.

5.7 Comparing the Models

There are multiple ways to test the performance of statistical models. One way is to build a model using the dataset and then obtain additional observations ('test data') on which to test the model. However, it may not be desirable to do this if the cost of obtaining test data is substantial. Alternatively, the dataset may be split into two parts: a training set on which to build the model, and a test set on which to test its performance. This method is called the validation set approach. The drawback of this method is that there are now fewer observations on which to train the model. To avoid this problem other methods have been developed. Cross-validation was designed to estimate the test error rate without losing observations from the training set. For this reason cross-validation is an appealing option. We already encountered leave-one-out cross-validation (LOOCV) in section 5.4. There is another kind of cross-validation, k-fold cross-validation, which is also commonly used to evaluate models. We chose to use LOOCV over k-fold cross-validation because of a specific reason relating to bagged trees and random forests. The reason is that bagged trees and random forests are easy to evaluate by a metric that is very similar to LOOCV.

Recall that bootstrap samples are taken with replacement. It is easy to show that each sample makes use of only around two-thirds of the observations in the original dataset. Thus, each sample is missing around one-third of the original observations. This implies that, for any given observation, approximately one-third of the bootstrap samples do not contain that observation. In our case, taking 500 bootstrap samples means approximately 167 of the decision trees built from the samples were not trained using a given observation. We are able to make use of this fact to estimate the error rate on real data. To do this we take the first observation in the original dataset and let each of the ~167 models that weren't trained using that observation predict its class. Out of these ~167 predictions, the class with the most votes is the final prediction. Of course, this prediction is either correct or incorrect. This is repeated for each observation in the training set, and the proportion of misclassified observations is called the out-of-bag (OOB) error rate. Interestingly, it turns out this error rate is nearly identical to the LOOCV error rate (James et al., 2015). As a result, we'll substitute the OOB error rate in place of the LOOCV error rate for the bagged tree and random forest.

Table 5.3 shows the estimated error rates for the four models. Interestingly a five-node decision tree seems to offer a moderate improvement over logistic regression. This suggests that the relationship between the response and the predictors may be better approximated by a non-linear model. The decision tree actually performed worse when bagging and random forests were used. The random forest still performed better than the bagged tree, which suggests it is still preferable to bagging.

Theoretically bagging and random forests should improve the performance of the decision tree. Since they didn't, further exploration was in order. One unexpected result of the analysis was the importance of years to kit submission to laboratory in the bagged tree and random forest. Interestingly, when this variable is removed from the pool of potential predictor variables entirely the bagged tree and random forest models actually improve. Table 5.4 shows the OOB error rate when years to kit submission to laboratory is not considered for the models. The OOB error rate drops by about 0.02 from the corresponding values in Table 5.3. This is quite an unusual result. One would expect that adding more predictor variables to the pool of potential candidates could only improve the performance of a model. In this case it appears years to kit submission to laboratory actually misled the models and made them worse.

	LOOCV Error Rate	OOB Error Rate
Logistic Regression	0.342	-
Decision Tree (five nodes)	0.297	-
Bagged Tree	-	0.316
Random Forest	-	0.306

Table 5.3: Comparison of estimated error rates of the four methods.

	OOB Error
Bagged Tree	0.293
Random Forest	0.289

Table 5.4: OOB error when years to kit submission to laboratory is removed from the pool of potential predictor variables.

The final question is which model to choose. It appears the five-node tree offers several benefits over the logistic regression model. In addition to having a lower estimated error rate, the decision tree is also simpler. It makes use of only three variables, whereas the logistic regression model required four. There is no reason to even consider the bagged tree or the random forest with years to kit submission to laboratory included. They have higher estimated error rates than the five-node tree and offer no additional benefits. The random forest with years to kit submission to laboratory excluded is an option if the only goal is prediction. However, given the benefits of having a graphical display, it seems the five-node decision tree is the best option.

CHAPTER 6 PREDICTING THE PRESENCE OF DUPLICATE DNA PROFILES

6.1 Introduction

In addition to predicting whether or not an SAK contains a CODIS-eligible DNA profile, it is also useful for cost-saving purposes to determine which kits contain identical DNA profiles across multiple forensic samples (vaginal sample, anal sample, etc). If the kits that contain duplicate profiles can be predicted with a high level of accuracy, it may be possible to develop an alternative protocol that involves analyzing fewer samples, thus saving costs.

In the sample of 2,500 SAKs, 11.4% contained duplicate DNA profiles (duplicate implies two or more). Of course the kits that contained no DNA profiles at all had no chance of containing duplicate profiles. About 57% of the kits contained at least one DNA profile, and 43% did not. Of the kits that contained at least one profile, 20% contained duplicate profiles.

As with chapter 5, this chapter will begin by examining the individual relationship between the predictor variables and the response.

6.2 Examining Relevant Variables

Figure 6.1 shows the relationship between the percent of kits that contain duplicate DNA profiles and five predictor variables: days to kit collection, the occurrence of consensual sex, years to kit submission to laboratory, number of samples run and victim age. As the top left graph shows, as the days to kit collection increased, the proportion of kits that contained a profile decreased. No kits collected five or six days after the assault contained duplicate profiles.

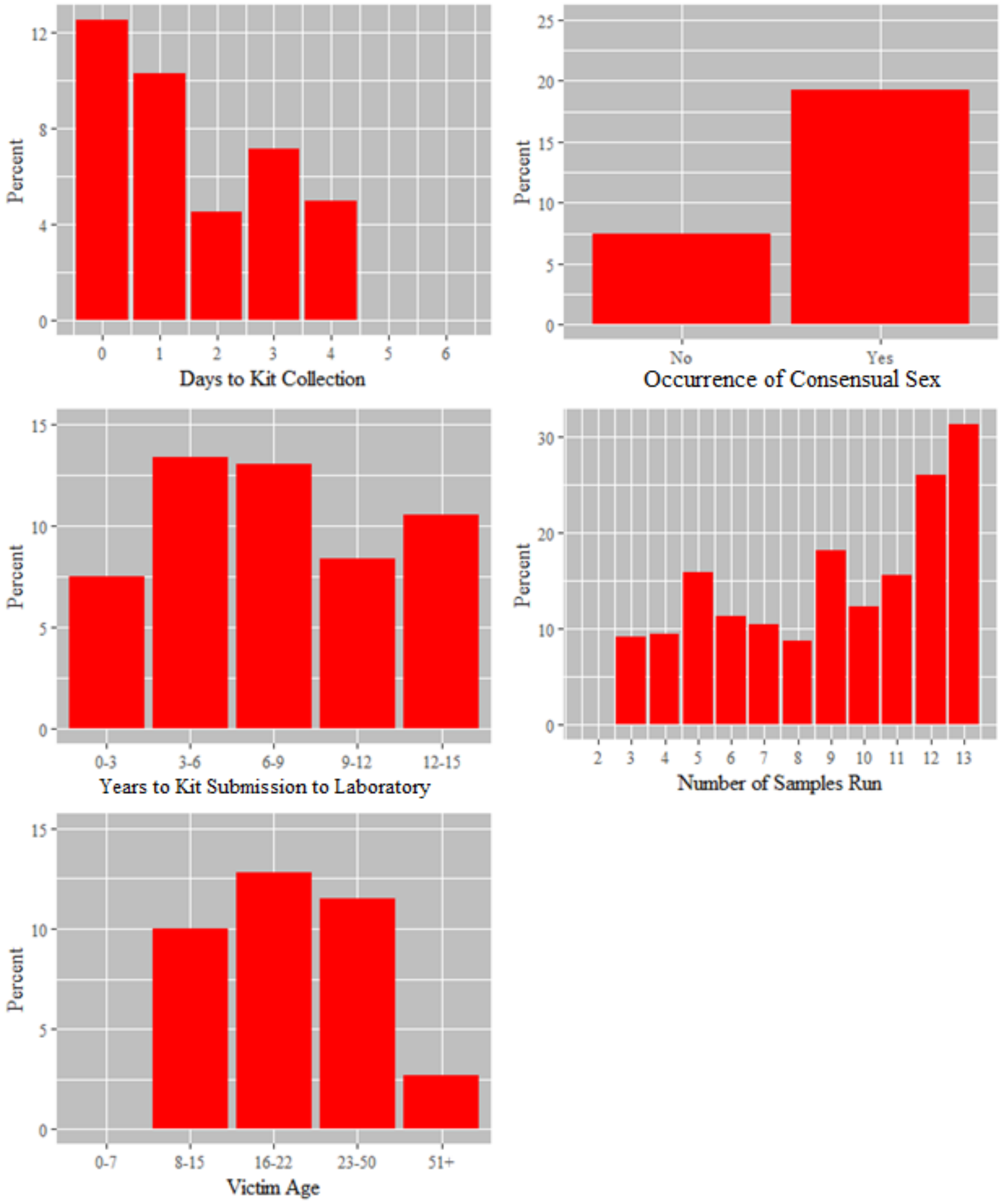

Figure 6.1: The y-axes display the percent of kits that contain duplicate DNA profiles. The x-axes display days to kit collection, the occurrence of consensual sex, years to kit submission to laboratory, number of samples run and victim age.

The occurrence of consensual sex had a large impact on the response: 19.2% of SAKs from victims who had consensual sex produced duplicate DNA profiles while only 7.4% of SAKs from victims who didn’t have consensual sex did.

The relationship between years to kit submission to laboratory and the response didn’t appear very strong. A kit may have been slightly less likely to contain duplicate profiles if it spent longer time in storage.

The number of samples run did appear to have an impact. The more samples that were run the more likely it was that a kit would contain duplicate profiles. Over 25% of kits that had 12 or more samples run contained duplicate DNA profiles.

The age of the victim also appeared relevant. No kits from those age 0-7 and only a few from those age 51+ contained duplicate profiles. It may not be surprising that SAKs from child victims didn’t contain duplicate DNA profiles since only 5% of them contained a single DNA profile to begin with.

Figure 6.2 shows the impact of analyzing certain forensic samples. The blue bar shows the percent of the time kits contained duplicate profile when that particular sample was analyzed. The red bar shows the percent of time the kits contained duplicate profiles when that sample was not analyzed. A large gap in the percent of kits that contain duplicate profiles when a sample was analyzed versus when it was not analyzed indicate a potentially useful predictor. It appears that analyzing the vaginal sample, external vaginal area sample and the neck/ear swab could be useful in predicting the presence of duplicate profiles. For every sample, the percent of kits with duplicate profiles was greater when that sample was analyzed than when it wasn’t.

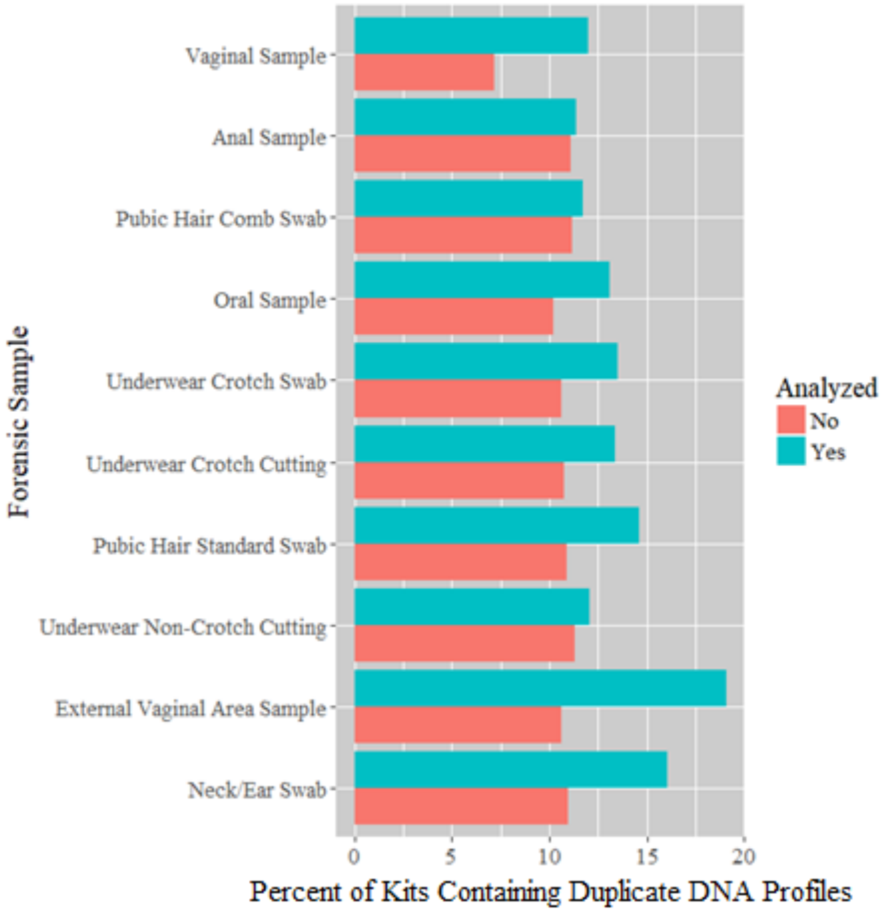

Figure 6.2: For each sample, the cases were separated based on whether or not it was analyzed. The percent of kits containing duplicate profiles is shown.

6.3 Logistic Regression

Eighteen variables were considered for the logistic regression model (four of which corresponded to victim age). Again the 'step' and 'glm' functions were used. Table 6.1 shows the steps in the model selection process.

In the first step, forward selection kept two variables that weren't significant at the 0.05 level: age 51+ and vaginal sample analyzed. It didn't make sense from a practical perspective to keep only one of the four age variables. For this reason a second model was fit with Age 51+ excluded.

The AIC only increased by 0.8 between the first and second steps. In the second step, the model had two variables that weren't significant at the 0.05 level: years to kit submission to laboratory and vaginal sample analyzed. Since vaginal sample analyzed had the largest p-value, it was considered for removal first. The model without this variable had an AIC of 1253.5, which was identical to that of the second model. As such, there was no harm in removing that variable from the model.

In the third step, years to kit submission to laboratory was just barely insignificant at the 0.05 level. However, there was a more serious problem: the coefficient for pubic hair comb swab analyzed was negative. The model was suggesting that analyzing the pubic hair comb swab actually decreased the probability of obtaining duplicate profiles. However, this contradicts our intuition that running more samples should increase the probability of obtaining duplicate profiles, and it also contradicts Figure 6.2 which suggests that the coefficient should in fact be positive. The reason this occurred was that in the total sample of 2,500 observations from which Figure 6.2 was crafted, a greater proportion of kits contained duplicate profiles when the pubic hair comb swab was analyzed than when it wasn't. However, after removing outliers, kits with

First Step:			
Coefficients	Estimate	Std Error	P-value
Intercept	-3.21	0.44	0.000
Days to Kit Collection	-0.28	0.1	0.007
Years to Kit Submission	-0.05	0.03	0.049
Age 51+	-1.05	0.73	0.153
Number of Samples Run	0.19	0.04	0.000
Consensual Sex	0.99	0.15	0.000
Pubic Hair Comb Swab Analyzed	-0.53	0.18	0.003
Vaginal Sample Analyzed	0.47	0.34	0.16
AIC = 1252.7			
Second Step:			
Coefficients	Estimate	Std Error	P-value
Intercept	-3.22	0.44	0.000
Days to Kit Collection	-0.29	0.1	0.006
Years to Kit Submission	-0.05	0.03	0.055
Number of Samples Run	0.19	0.04	0.000
Consensual Sex	0.99	0.15	0.000
Pubic Hair Comb Swab Analyzed	-0.54	0.18	0.003
Vaginal Sample Analyzed	0.46	0.34	0.18
AIC = 1253.5			
Third Step:			
Coefficients	Estimate	Std Error	P-value
Intercept	-2.82	0.32	0.000
Days to Kit Collection	-0.29	0.1	0.006
Years to Kit Submission	-0.05	0.03	0.057
Number of Samples Run	0.19	0.04	0.000
Consensual Sex	1.01	0.15	0.000
Pubic Hair Comb Swab Analyzed	-0.54	0.18	0.003
AIC = 1253.5			
Fourth Step (Final):			
Coefficients	Estimate	Std Error	P-value
Intercept	-2.57	0.3	0.000
Days to Kit Collection	-0.3	0.1	0.003
Years to Kit Submission	-0.07	0.03	0.006
Number of Samples Run	0.14	0.03	0.000
Consensual Sex	1	0.3	0.000
AIC = 1260.7			

Table 6.1: Steps in the logistic regression model selection process.

only one sample run, and observations that contained missing data, only 1,924 observations were left to train the model. Out of these 1,924 observations, when the pubic hair comb swab was analyzed a smaller proportion of kits contained duplicate profiles. Out of a concern that the negative coefficient in the table does not represent the true relationship between the variable and the response, pubic hair comb swab analyzed was removed from the final model, leaving four variables remaining.

The four variables in the final model were: days to kit collection, years to kit submission to laboratory, number of samples run, and consensual sex. All variables were significant at the 0.05 level. The AIC jumped from 1253.5 to 1260.7 between the third and final steps. This is not a very large increase considering the benefits of this model, such as having all significant predictors and not having coefficients that produce counterintuitive results. The probability and logit equations of this model are shown in *6.3.1* and *6.3.2* respectively.

$$p(X) = \frac{e^{(-2.57-0.3D-0.07Y+0.14S+1C)}}{1+\ e^{(-2.57-0.3D-0.07Y+0.14S+1C)}} \qquad 6.3.1$$

$$log\left(\frac{p(X)}{1-p(X)}\right) = \ -2.57 - 0.3D - 0.07Y + 0.14S + 1C \qquad 6.3.2$$

The variables are abbreviated with the first letter of their name in Table 6.2 except the number of samples run, which is abbreviated ‘S’. Unlike in the logistic regression model of Section 5.3, the intercept of this model has no practical meaning because it makes no sense to assign a value of zero to the number of samples run. The probability of obtaining duplicate profiles when no samples are run is obviously zero, so the model will produce patently wrong results in this case.

Figure 6.3 offers a graphical display of Equation *6.3.1* with values plugged in for certain variables. The top left image shows the probability of obtaining duplicate profiles plotted against days to kit collection. It assumes 8.5 years to submission to laboratory and five samples run. As we learned from Chapter 5, a convenient way to assess how the probability changes is to use the divide-by-four rule. Unfortunately, this rule isn't going to be very accurate because the rule only applies to x-values around probability 0.5. For instance, the rule tells us to divide the coefficient of days to kit collection by four. This would yield -0.075. But the curve doesn't decrease by 0.075 for every additional day. Going from zero to one day with consensual sex decreases the probability by only 0.042.

The top right graph shows the probability plotted against years to kit submission to laboratory. Similar to the top left graph, the divide-by-four rule won't be highly accurate since no location on the plot attains a probability of 0.5.

The divide-by-four rule is more appropriate for the bottom left graph. The bottom left graph plots probability against the number of samples run. The consensual sex curve nearly attains a probability of 0.5. The divide-by-four rule suggests that for every additional sample run the probability of obtaining duplicate profiles increases by 0.14/4 = 0.035. The actual increase in probability going from 12 to 13 samples is 0.033. Next we'll consider fitting a decision tree.

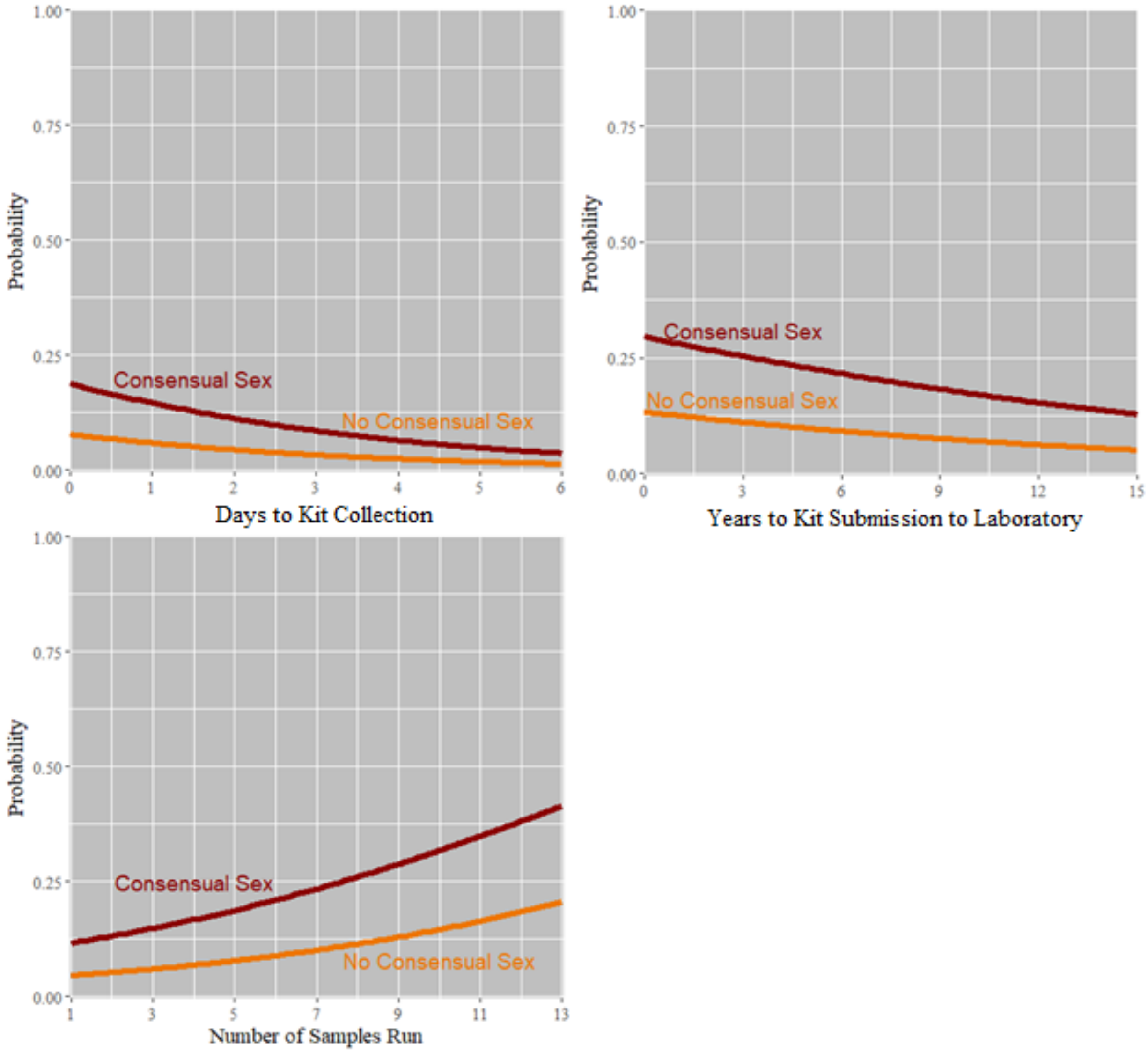

Figure 6.3: *Top Left*: Probability of a kit containing duplicate DNA profiles plotted against days to kit collection. Assumes 8.5 years to kit submission to laboratory and five samples run. *Top Right*: Probability plotted against years to kit submission to laboratory. Assumes zero days to kit collection and five samples run. *Bottom left*: Probability plotted against the number of samples run. Assumes zero days to kit collection and 8.5 years to kit submission to laboratory.

6.4 Decision Tree

Again the rpart package in R was used to fit a decision tree to the data. The pubic hair comb swab was excluded as a predictor variable since it led to illogical results in the logistic regression model. The initial tree contained 55 terminal nodes. LOOCV was used to determine the optimal number of terminal nodes. Figure 6.4 shows the LOOCV error rates for trees of varying sizes. As before, the y-axis shows the error rate relative to that of the one-node tree. In this case, the error rate of the one-node tree is 0.11. This is the proportion of observations one would misclassify if one simply predicted “No” for every observation in the dataset. One may recall from earlier that 11.4% of SAKs in the sample contained duplicate profiles. The figure 0.11 differs slightly from the number cited earlier because it was computed after observations were removed for containing missing data and other reasons.

As Figure 6.4 shows, adding additional nodes to the tree doesn’t actually appear to lower the estimated error rate. It appears that the single-node tree performs the best of all trees. As such, there is little reason to pursue a decision tree model. Next we’ll see how bagging performs.

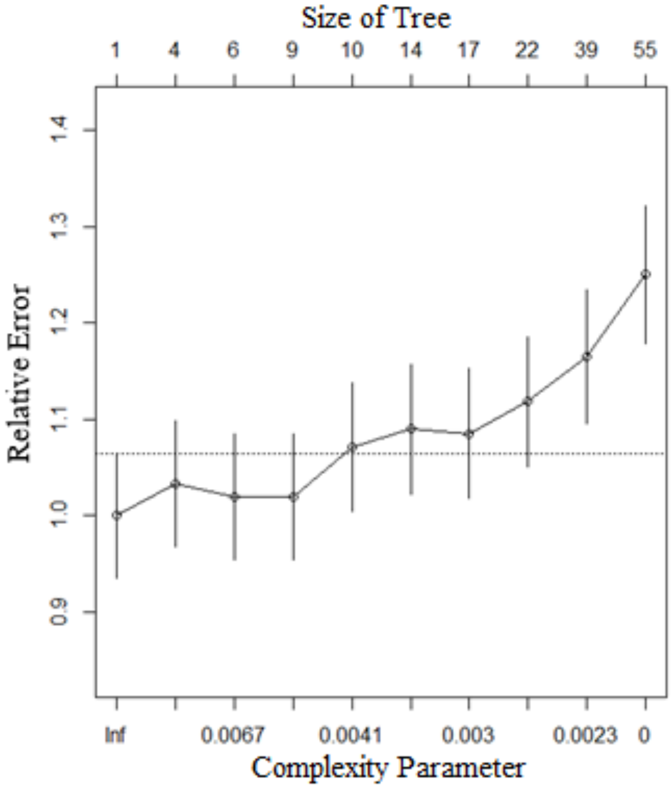

Figure 6.4: Estimated misclassification rate from leave-one-out cross validation.

6.5 Bagged Trees

A bagged tree was created in an attempt to beat the 'null error rate' of 0.11. The randomForest package was used to fit the model. The pubic hair comb swab was excluded from the bagged tree as well. Again, 500 unpruned trees were created from bootstrap samples. The out-of-bag (OOB) error rate for the bagged trees was 0.1138. As with the decision tree, the bagged tree fails to improve upon the null error rate. It appears bagging is a fruitless method in this context.

6.6 Random Forest

A random forest was also used in an attempt to beat the null error rate of 0.11. The pubic hair comb swab was excluded as a predictor. Again 500 trees were constructed, this time considering only 4 of the 17 predictors at each split. The randomForest package was used to fit the model. The model had an OOB error rate of 0.1086. It is encouraging to see that the model beat the null error rate, although by a highly unimpressive margin. In the next section we'll see how the models compare to the logistic regression model.

6.7 Comparing the Models

Table 6.2 displays the estimated test error rate for each of the four models. A decision tree was not crafted at all since no model stood a chance of improving on the null error rate of 0.11. When crafting the decision tree, we were alerted early to the fact that no tree would beat the null error rate since LOOCV is used in the model selection process. On the other hand, LOOCV is not used in the model selection process for logistic regression. As such, we were only alerted after the model was created to the fact that it didn't beat the null error rate. The only model that did beat the null error rate was the random forest. It did so by an unimpressive 0.0014. Since all models performed so poorly, there is little value in using any of them to predict the presence of duplicate DNA profiles in a kit. Next we'll turn to chapter 7 for closing remarks.

	LOOCV Error Rate	OOB Error Rate
Logistic Regression	0.1101	-
Decision Tree	(No model selected)	-
Decision Tree (Bagged)	-	0.1138
Decision Tree (Random Forest)	-	0.1086

Table 6.2: Estimated Error Rate in Predicting Duplicate Profiles.

CHAPTER 7 CONCLUSION

Four statistical methods were compared for their ability to predict two response variables: whether an SAK contained a CODIS-eligible DNA profile, and whether an SAK contained duplicate DNA profiles across multiple forensic samples. Concerning the first response variable, a five-node decision tree seemed to be the most appropriate model. It has a simple graphical display and was estimated to classify only 0.008 more observations incorrectly than the random forest. The logistic regression model unfortunately wasn't able to model the response as effectively as the decision tree. It was expected to classify 0.045 more observations incorrectly than the decision tree.

One of the most interesting findings of the analysis was that the bagged tree and random forest actually performed worse when the predictor years to kit submission to laboratory was included. One would think no harm could come of including additional predictor variables in the pool of potential predictor variables. This illustrates the importance of doing further analysis when an unexpected result comes up. In this case we were able to pinpoint the problem fairly quickly due to the fact that the graph of variable importance showed a strange result: years to kit submission to laboratory was purporting to be by far the most important variable. We were skeptical of this result because it was one of the weaker variables in the logistic regression model. After removing it from the analysis altogether we realized it had been misleading the models. Sure enough, after it was removed, model performance improved.

As of now it appears the best statistical model will be able to guess correctly about 70% of the time whether an SAK contains a DNA profile or not. One may wonder if experts would be able to guess correctly more than 70% of the time. I think it's reasonable that they may be able to. While 70% is high, there is certainly a lot of room for improvement. It's also possible that the combination of an expert and a statistical model would be able to improve on the 70% figure. This may be an area worth investigating in future research.

As stated in the introduction, one of the limitations of these models is that they don't incorporate any information about the nature of the assault, i.e. vaginal assault, anal assault, etc. More work could be done to determine how the models might change after incorporating information about assault type. The type of assault could be included as a predictor variable, or entirely different statistical models could be built for different scenarios. In addition, these models don't incorporate information about the number of assailants in the assault. The reason for this is the difficulty in determining the actual number of assailants, as noted in the introduction. Future work could try to incorporate this information into the statistical models as well.

As for predicting which kits contain duplicate DNA profiles, as of right now statistical models are largely unable to beat the 'null' model of simply predicting "No" for each kit. The best model, the random forest, was estimated to classify only 0.0014 more observations correctly than the null model. Every other model was expected to perform worse than the null model. It seems more high-quality predictor variables are needed to accurately predict this response.

Finally, the models could have been improved if there simply had not been so much missing data. Sexual assault nurses and others may not have anticipated the data would be used in a future statistical analysis and in some cases may have been careless about properly documenting the details. It is impossible to know exactly how much the missing data has affected the models, but hopefully future studies will be able to correct for any biases the models have. Hopefully this work can be used as a starting point for future statistical modeling of sexual assault kits.

REFERENCES

Atkinson, E., & Therneau, T. (2017, March 12). *An Introduction to Recursive Partitioning Using the RPART Routines.* Retrieved October 20, 2017, from https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Bipartisan Group of Lawmakers Launch Task Force to End Sexual Violence. (2017, April 5). Retrieved November 7, 2017, from https://kuster.house.gov/media-center/press-releases/bipartisan-group-of-lawmakers-launch-task-force-to-end-sexual-violence

Bipartisan Task Force to End Sexual Violence Discusses Rape Kit Backlog, Access to Sexual Assault Nurse Examiners. ((2017, June). Washington, D.C. Retrieved from https://meehan.house.gov/media-center/press-releases/bipartisan-task-force-to-end-sexual-violence-discusses-rape-kit-backlog

Fallik, S., & Wells, W. (2014). Testing Previously Unsubmitted Sexual Assault Kits. *Criminal Justic Policy Review, 26(6)*, 598-619.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* New York: Cambridge.

Gray-Eurom, K., Seaberg, D., & Wears, R. (2002). The Prosecution of Sexual Assault Cases: Correlation with Forensic Evidence. *Annals of Emergency Medicine, 39*(1), 39-46.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Kerka, J., Heckman, D., Maddox, L., Sprague, J., & Albert, J. (2018). Statistical modeling of the case information from the Ohio Attorney General's Sexual Assault Kit Testing Initiative. *Journal of Forensic Science*.

Liaw, A., & Wiener, J. (2002). Classification and Regression by randomForest. *2*, 18-22. Retrieved from http://CRAN.R-project.org/doc/Rnews/

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., & Verbeke, G. (2014). *Handbook of Missing Data Methodology.* Boca Raton: Chapman and Hall.

National Institute of Justice. (2016, March 18). *Untested Evidence in Sexual Assault Cases*. Retrieved October 5, 2017, from https://www.nij.gov/topics/law-enforcement/investigations/sexual-assault/Pages/untested-sexual-assault.aspx

Peterson, J., Johnson, D., Herz, D., Graziano, L., & Oehler, T. (2012, June). Retrieved October 10, 2017, from https://www.ncjrs.gov/pdffiles1/nij/grants/238500.pdf

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Shalizi, C. (2009, November 6). *Classification and Regression Trees.* Retrieved October 1, 2017, from Carnegie Melon Univsersity: http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf

The Baltimore Sun. (2017, January 3). Attorney General's report calls for statewide standards on rape kits. Retrieved from http://www.baltimoresun.com/news/maryland/crime/bs-md-rape-kits-report-20170103-story.html

Therneau, T., Atkinson, B., & Ripley, B. (2017). rpart: Recursive Partitioning and Regression Trees. Retrieved from https://CRAN.R-project.org/package=rpart

Titanic: Machine Learning From Disaster. (2015). Retrieved October 18, 2017, from Kaggle: https://www.kaggle.com/c/titanic

UCLA. (n.d.). *Multiple Imputation in SAS Part 1*. Retrieved October 29, 2017, from https://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1/

Wells, W., Campbell, B., & Franklin, C. (2016). *Unsubmitted Sexual Assault Kits in Houston, TX: Case Characteristics, Forensic Testing Results, and the Investiagtion of CODIS Hits, Final Report.*

Where the Backlog Exists and What's Happening to End It. (n.d.). Retrieved November 7, 2017, from End the Backlog: http://endthebacklog.org/backlog/where-backlog-exists-and-whats-happening-end-it

Why the Backlog Exists. (n.d.). Retrieved November 7, 2017, from End the Backlog: http://endthebacklog.org/backlog/why-backlog-exists

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., Francois, R., Henry, L., & Muller, K. (2017). dplyr: A Grammar of Data Manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr

Wikipedia. (2017, September 9). Decision Tree Learning. Wikimedia Foundation. Retrieved from https://en.wikipedia.org/wiki/Decision_tree_learning#/media/File:Titanic_Survival_Decison_Tree_SVG.png

APPENDIX A LIST OF R PACKAGES AND FUNCTIONS USED

Function	Package	Description
glm	stats	Fits generalized linear models such as logistic regression.
step	stats	Performs forward, backward, and stepwise selection for generalized linear models based on AIC.
rpart	rpart	Fits regression and classification trees using recursive binary splitting.
plotcp	rpart	Creates a graph of leave-one-out cross-validation results.
fancyRpartPlot	rpart	Creates a graphical display of a decision tree fit by rpart.
-	ggplot2	Graphics package for R.
randomForest	randomForest	Fits bagged trees and random forests.
VarImpPlot	randomForest	Creates a variable importance plot for random forests/bagged trees.
-	dplyr	Allows for easy manipulation of data frames.