Forecasting Prison Sentences:

Plead or not Plead?

Victor Cheung

May 1, 2015

1 Introduction

Courts pass sentences with some consistency. There are minimum sentences at the federal and state level for classes of crimes, and judges will pass similar length sentences for similar offenses. Hence, as defense attorneys, we would like to construct a forecasting model that could predict the likely length of sentence for a given offense, and given the characteristics of the defendant.

In the first section, we discuss the dataset to be analyzed as well as the data generation process. In the second, we conduct a univariate analysis and briefly go over bivariate statistics. In the third, we discuss missing data in the data and how we deal with them. In the fourth section, we cover the models to be built, justifying our choice of stochastic gradient boosting, level II analysis and relative cost considerations. We then discuss tuning parameters and construct the final model to be used. In the fifth, we analyze the results from our models. In the sixth, we conclude and propose future extensions of our work.

2 Data Summary

Our dataset consists of all 65,098 offenders convicted and sentenced in our judicial district of interest in the past year. We have 9 predictors and 1 response value.

1. Nominal Sentence - Sentence given by the judge. Early release is possible.
2. Education - total number of years of education. Over 12 is post-secondary education.
3. DrugCondition - refers to any kind of drug problem ("Yes" or "No or Unknown").
4. PsychCondition - refers to any kind of psychological illness ("Yes" or "No or Unknown").
5. AlcoCondition - refers to any kind of alcohol dependence and addiction ("Yes" or "No or Unknown").
6. Sex - gender of the offender.
7. CurrentAge - age in years.
8. MaritalStatus - indicates whether the offender is married ("Yes" or "No or Unknown").
9. as.factor.Offense1 - is the most serious offense for which the offender was found guilty. This implies there may be other guilty sentence for any given offender.
10. as.factor.OffGroup1 - are groupings for the primary offense. The groups are delineated in the appendix.

We know that these data are generated by forms that staff were responsible for filling out. In the case of sentences, age or education the staff entered a numeric value or left it blank. This is likely due to their not knowing the required information. In the case of all other factors, including drugs, psychological conditions or alcohol problems, these were "Yes/No" entries that the staff checked off. Instances where these entries were left blank were either ignored by the staff or implied to mean "No". From the data entry process, we will assume that the staff were largely consistent in ignoring certain entries, and only a few of those unmarked were by mistake. In this way, the errors are close to randomly missing and should not invalidate our analysis.

Table 1: Summary Statistics for Numeric Data

Statistic	N	Mean	St. Dev.	Min	Median	Max
NominalSentence	56,110	26.907	19.812	0	20	120
CurrentAge	65,089	32.832	11.252	14	30	80

Figure 1: Summary Statistics for All Data

```
NominalSentence    Education           DrugCondition         PsychCondition
Min.   :  0.00   12     :24027   YES          :31291   YES          : 3695
1st Qu.: 12.00   Unknown:22445   NO OR UNKNOWN:33690   NO OR UNKNOWN:61372
Median : 20.00   11     : 5314   NA's         :  117   NA's         :   31
Mean   : 26.91   10     : 3928
3rd Qu.: 36.00   9      : 2202
Max.   :120.00   14     : 2014
NA's   :8988     (Other): 5168
      AlcoCondition        Sex          CurrentAge       MaritalStatus
YES          :26955   MALE   :52182   Min.   :14.00   SINGLE    :35923
NO OR UNKNOWN:38002   FEMALE :12536   1st Qu.:23.00   MARRIED   : 5863
NA's         :  141   UNKNOWN:  380   Median :30.00   DIVORCED  : 2988
                                       Mean   :32.83   SEPARATED : 1723
                                       3rd Qu.:41.00   WIDOWED   :  375
                                       Max.   :80.00   COMMON LAW:   62
                                       NA's   :9       UNKNOWN   :18164
         as.factor.Offense1. as.factor.OffGrp1.
ASSAULT-2ND DEGREE : 9386    49     :18598
THEFT-MISDEMEANOR  : 6753    13     :11045
POSSESSION CDS     : 6396    25     : 9703
POSS MARIJUANA     : 3849    61     : 5426
POSS W-I DISTR. CDS: 3695    64     : 4282
(Other)            :35017    22     : 2959
NA's               :    2    (Other):13085
```

Figure 2: Univariate Analysis - Histograms

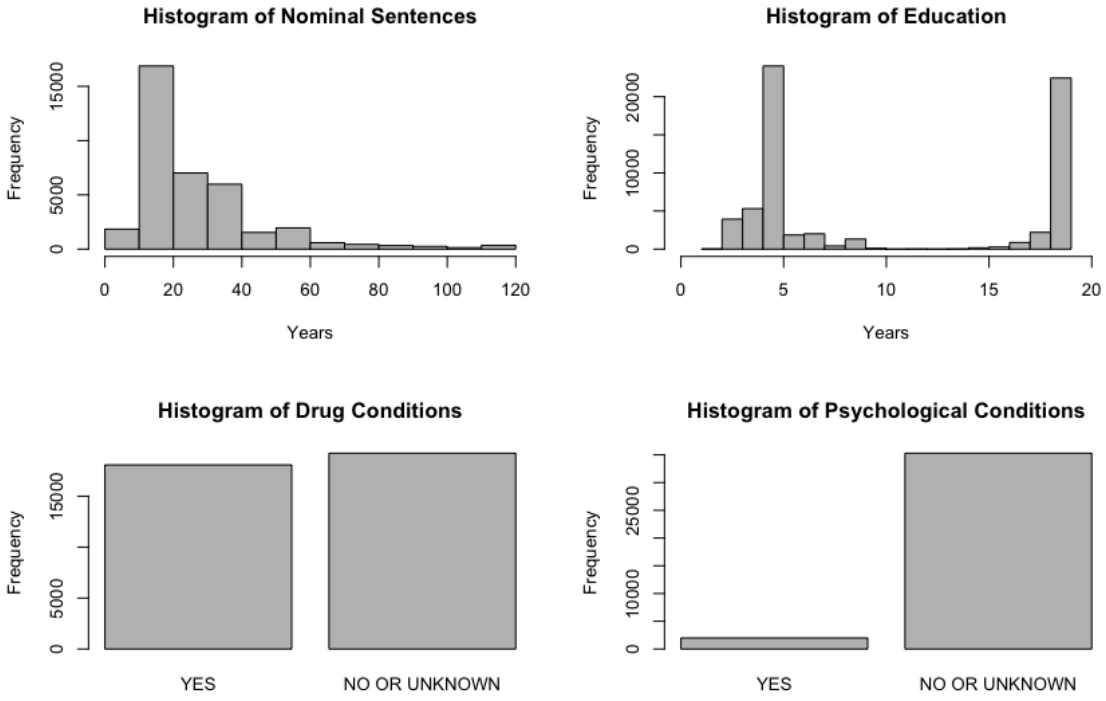


Figure 3: Univariate Analysis - Histograms

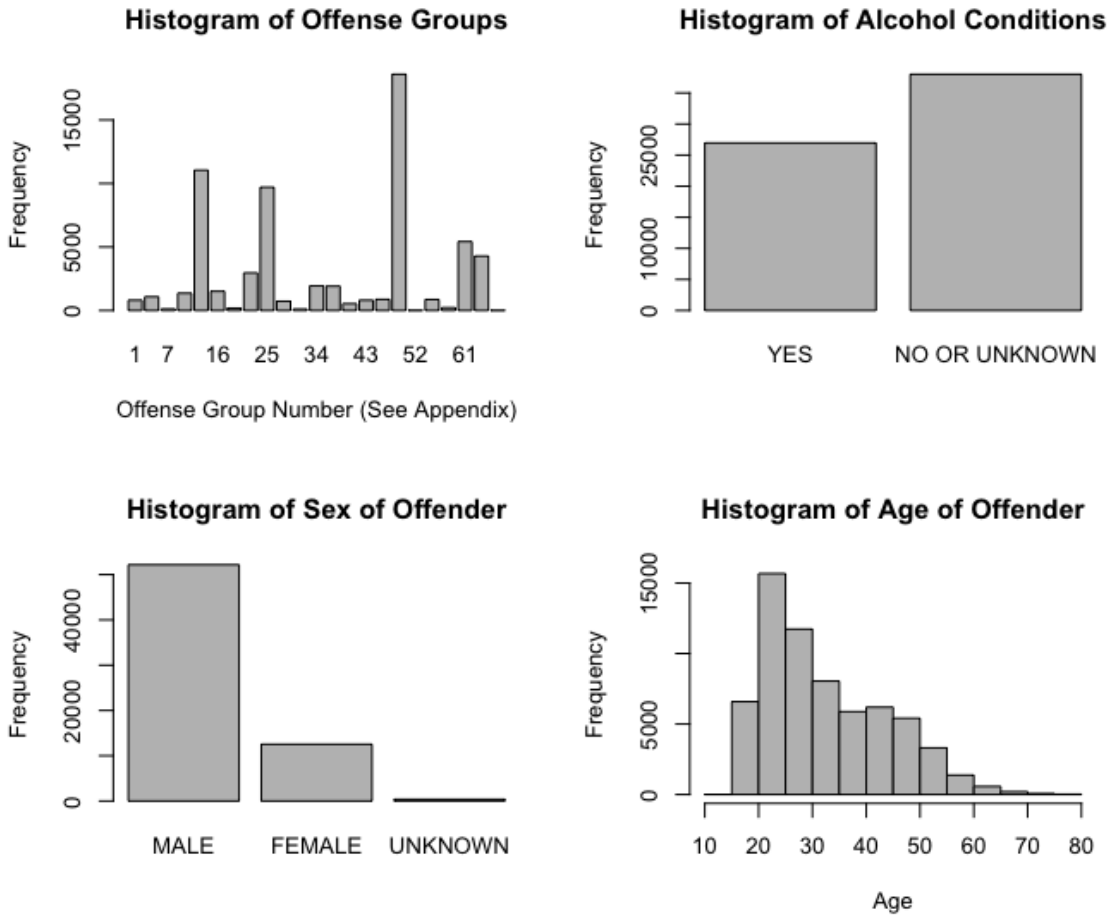


Figure 4: Univariate Analysis - Histograms

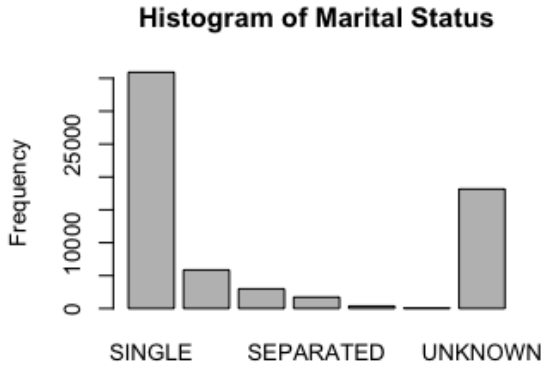
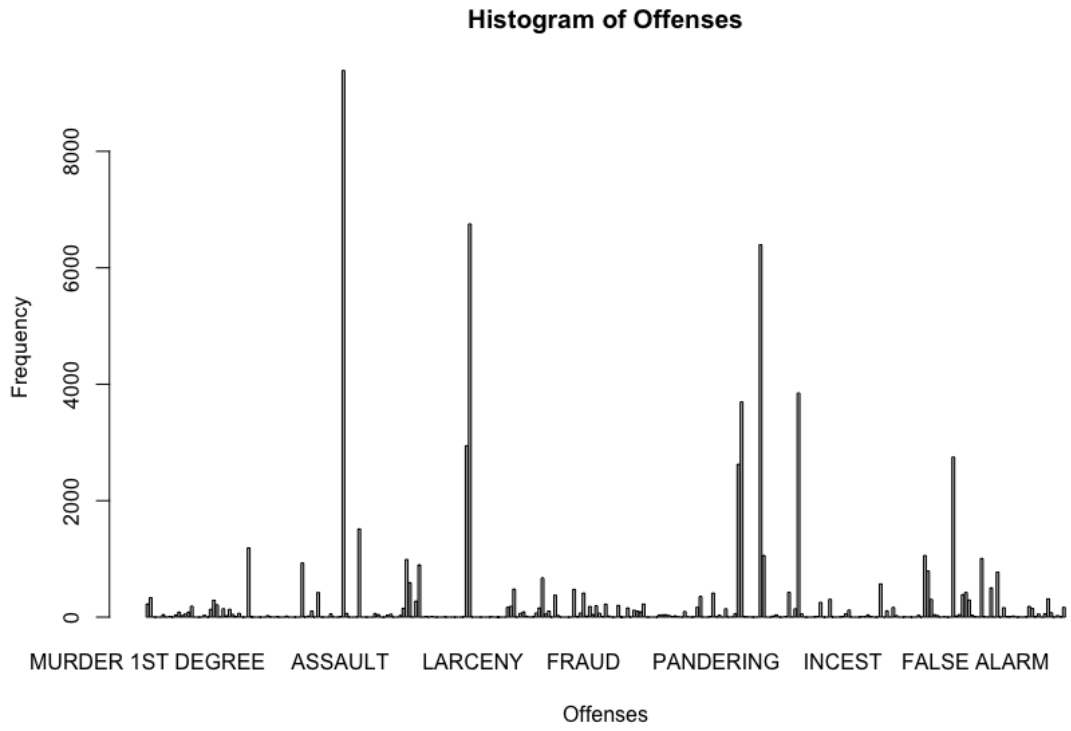


Figure 5: Univariate Analysis - Histograms



2.1 Univariate Analysis

Consider figure 1 to 3.

Nominal Sentence The distribution is highly right skewed, as seen in figure 1 and from the summary statistics. We are worried about sparsity for sentences passed on this fat right tail. We may not have enough observations there to capture the local features of the data. Furthermore, there are close to 9,000 observations with no sentencing data.

Education Education is close to bimodal. There are many with few years of education. Presumably they dropped out of elementary or middle school. There are very few who have gone to high school or college, or who have dropped out. There are close to as many people with advanced education - those with more than 16 years of education. We are worried about the sparsity of observations for those who have gone to high school and college students. Furthermore, we note that education is a factor variable, instead of a numeric one. This may present problems in model construction. Discrete levels of education may not make sense when there is little difference between an offender with 3 years of education vs 4 years of education. A better predictor might be whether an offender has completed elementary school, middle school, high school, and post-secondary education.

Drug Condition The distribution of drug conditions is close to even. However, we note that "No or Unknown" is not automatically equivalent to "No".

Psychological Condition There are very few people with recognized psychological conditions. We note that due to the stringent process by which courts acknowledge someone as having a psychological condition, as these are legitimate legal defenses, that there many be some in the "No or Unknown" category who suffer from undiagnosed or discredited psychological conditions. In any case, the relative sparsity of observations with psychological conditions may be a problem.

Offenses From the histogram, it's clear that the types of offenses are unevenly distributed.

A few offenses see large numbers of occurrences, while others are extremely rare. We expect this impact the offense groupings as well. For the offenses which are infrequently committed, we are worried about the effects of sparsity on our model. Their effects on sentences may not be fully captured by our models.

Offense Groups The approximate groupings of offenses are unevenly distributed between a few offenses that occur frequently and those that occur rarely. From the appendix, the most frequent offenses are those related to assault, possession of or intent to distribute controlled drug substances, theft, DUI and a miscellany of minor crimes such as home improvement and disturbing the peace.

Alcohol Condition There are many people with confirmed alcohol problems.

Sex The majority of offenders are male, although there are many offenders who are female as well. There is a very small minority with unknown sex. We interpret this either someone who does not identify with either sex, or due to errors in data collection.

Current Age The histogram of age for offenders is right skewed. The modal age is around 20, falling off from there till age 80. Given that very few offenders are very old (>65), we are worried about the sparsity and not capturing the effects of old age on the leniency of sentences.

Martial Status The majority of offenders are single, followed in frequency by unknown, married, divorced, separated, widowed, common law. There are very few observations for the last two. Again, sparsity here may be a concern.

2.2 Bivariate Analysis

A bivariate analysis is difficult here given the nature of the data and the number of variables that are factors. However, from figure 6 in the appendix, we can still spot a number of relationships between those variables we think are the most important. From the education vs. current age plot, we see that for any given level of education that the distribution of

ages remain largely similar. The medians occur at the same age; however, some education levels have larger IQRs and greater number of outliers. We expect this is just an artifact of the data. From the MaritalStatus vs age plot, we see that those who are married or who are divorced tend to be older, as expected. Those with unknown martial status or who are single tend to younger, which leads us to think that unknown may be the same as being single. From the offense groups vs. age plot, we see that some crimes tend to be committed by older offenders while others by younger offenders. These differences are not large. The offense group vs. current age plot show that each group is related to age in a right skewed distribution.

3 Missing Data

We note that for 9000 observations that the response, the nominal sentence, is missing. Since the focus of our model is precisely to predict these response values, any sort of data imputation here doesn't make sense, regression or otherwise. Whether or not the data is systematically, we must delete these observations from our data set prior to training. Our hope is that these data pints are missing at random - that is, when the staff were filling out the forms, that some offenders were missing sentencing data at random. We note further that the sentences of zero are not due to error - instead, these are legitimate values indicating that the sentences were already served when the offenders were awaiting trial or sentencing. From figure 6 and 7, we see that the histograms confirm that these observations are missing as if at random. The histograms are largely similar to those for the data as a whole. We will listwise delete all these observations. This still leaves us a large sample size of 55,937 observations.

Furthermore, there are a small number of observations missing data for `AlcoCondition`, `CurrentAge`, `as.factor.Offense1`. We will delete these observations as well since there are so few of them relative to the overall size of the dataset that it's unlikely to impact our model. We don't trust that CART's default imputation procedures using modal values

make sense here, especially when there are so few points. It's cleaner to simply remove them.

4 Model Building

4.1 Boosting

We choose boosted trees using stochastic gradient boosting over random forests.

Caruana and Niculescu-Mizil at Cornell University found boosted trees had the best accuracy out of sample when compared with an array of supervised learning algorithms, including bagged trees, random forests, SVM, neural nets, etc. Since we are most concerned with the forecasting performance of our model, this choice makes sense.

As well, stochastic gradient boosting presents a number of advantages shared with random forests. This and the fact that our fitted values are a linear combination of all fits provides effects similar to interpolation and local robustness. At the same time that we are capturing local features of the data, we also reduce noise by weighting all previous results to prevent overfitting. Given a large number of iterations, we find that instead of severely overfitting, we are further smoothing over noise in the data, which leads to better performance out of sample. Furthermore, boosting allows us to fit extremely well to what would otherwise be outliers. However, we note that interpolation in regression is not the same as in classification. We cannot achieve perfect homogeneity in our trees, since we will require a minimum terminal node size greater than one due to the need to compute quantiles. This implies that the benefits of interpolation and local robustness may be limited somewhat due to the inability to build extremely large trees. However, the boosting algorithm still captures local features through a focus on residual reduction; hence, local robustness still holds, so long as the number of iterations is sufficiently large.

We concede that random forests are by design concerned with out of sample predictive performance. Random forests use out of bag data to construct fitted values and produce performance metrics. No such options exist for boosted trees. At best, we use randomly

sampled (without replacement) data from the training set as a source of randomness. But performance metrics, variable importance and partial dependence are all constructed by in-sample data. Finally, consider that boosting is a margin maximizing procedure. A formal explanation for this was given for classification, but the idea holds for regression as well. Margin maximization will provide stable forecasts based on a stable model.

4.2 Level II Analysis

There is no guarantee that a boosted tree will provide accurate approximations of the true response surface. Instead, our estimation target will be a tree based approximation of that surface. Furthermore, the generalizability of our model, as measured by our estimation of the true generalization error, technically holds only for our given sample size, tuning parameters and specific settings used in boosting. However, as found in the statistical literature, boosted trees generally have the best performance out of sample when compared against other techniques, including random forests. Boosted trees also perform well in low-dimensions, which is the case here. The mathematics underlying this is not yet fully understood.

The joint probability distribution underlying the sentencing in our judicial district is likely stationary overtime. The sentencing decision, once a defendant is found guilty, is influenced by at least two factors outside of the crime committed: the judge and the law. The latter hardly changes once legislation is passed - Americans found guilty of piracy have a minimum life sentence, according to a law passed in 1790 and which still holds today. The former, judges, bring idiosyncrasies into the sentencing based on past experience, education, ideology and so on; nonetheless, judges have long tenures and most stay in their given positions for long periods of time. Hence, the unique ways in which they sentence will be part of the joint probability distribution for at least the next year or two, barring any judges retiring or being selected for higher courts. In cases where this occurs, we might expect that new judges will sentence in similar ways. Judges are often educated from the same elite institutions (Harvard, Yale, etc.) and likely hold similar ideologies.

The population from which the joint probability distribution draws are all criminal offenders who have gone to court and received a sentence in the last year in the judicial district where our own clients would be tried. We have the entire sample, at least for the past year, at our disposal. Note however that this is not a random sample from all cases that have ever been tried in this judicial district. If the defendants and their crimes in the last year differed systematically from years prior and from years in the future, then our model will not hold. Furthermore, there is no guarantee that future defendants will be drawn from this population. For example, there may be classes of crimes which did not receive sentences in the last year - we cannot forecast sentences for crimes we have not witnessed before. As well, defendants may be different in ways not measured by the information we have in terms of education, age, gender, etc. In justifying our level II analysis then, we must assume that this population of offenders holds going forward, and that future offenders will be drawn from this population, or at least from a population close to it.

4.3 Relative Cost

People are generally risk averse. For most our defendants then, it's better to plead guilty to a lesser crime and receive a shorter sentence for certain, than to risk going to trial and bear the uncertainty of convictions and of the crime. Therefore, we look unfavorably upon overly optimistic forecasts of shorter sentences. In this case, a 3-to-1 cost ratio seems justified. Underestimates of sentences are at least three times more costly than overestimates. We will target the 75% quantile.

4.4 Final Model

In tuning parameters, we are lucky that stochastic gradient boosting is sufficiently robust that different sets of parameters often provide similar performance.

There a number of tuning parameters we are concerned with. The first is the number of iterations. We will set the number of iterations to be very large at first in order to determine the convergence to some lowest residual deviance. If the fitting error increases past some

point, we will choose that minimum number of iterations that results in the lowest error. Otherwise, if the error rate effectively ceases to decrease past some threshold, we stop at that threshold.

The second is the sampling size within the training data set. For this, craft lore holds that 50% works relatively well. Since we have many thousands of iterations, there is no worries that some observations cannot contribute. As well, this even split allows us to construct robust out of bag loss (but still within training data). We will also try 0.325% and 0.625% to assess the difference in quantile loss.

The third is the learning rate, which adjusts the rate at which we adapt to residuals. Too large a learning rate risks overfitting, and lower learning rates incur greater computational time, but increases flexibility of fitting and improves stability. We will set learning rate at 0.001 and incur more time computing rather than risking overfitting.

The fourth is interaction depth, a misnomer for the maximum number of splits allowed in the tree and hence the depth of the tree. We allow for larger trees as we have a training set of over 30,000 observations. This capitalizes on that larger (more complex) based learners in boosting helps improve performance. While a more complex tree risks overfitting, the large number of trees and the linear combination taken should average out errors in a way similar to what happens in random forests.

The fifth is the minimum terminal node size. Since we are doing quantile regression, we require a terminal node that allows us to determine quantiles within that node. Hence, we choose a minimum size of 10.

In constructing our final mode, we primarily tweaked the `bag.fraction` parameter and found that it made little difference. We then tried changing the minimum observations in terminal nodes as well as the maximum allowed interaction depth. We picked our final model according to the final in-sample quantile loss calculated by method of Out-of-bag data, as well as the observed relative cost.

The quantile loss, partial dependence plots, and variable importance plots of our final model are model.

$$\text{Testing Data Observed}\alpha = \frac{1}{N}\sum I(y_i > f(x_i)) = 0.7455218$$

5 Discussion

Our final parameter specifications produces a boosted tree with a quantile loss close to 5.71. On testing data, the observed α was 0.7455, very close to the target quantile of 0.75. This was expected, since we allowed the asymmetric cost that underestimates are three times more costly than overestimates. Hence, there are three times fewer underestimates as there are overestimates, which is reflected by the fact that 75% of predictions fall above the 1-to-1 line. This can also be seen in figure 15.

From figure 10, we see the reduction in quantile loss against the number of iterations run. We calculate this from the OOB data, which provides a more reliable performance metric than the data used already for training. Nonetheless, we note that this data is still in-sample. After around 4000 iterations, the reduction in loss is is effectively 0. However, using the software provided by R in the package, we found that using the OOB method that the effective stopping point was 5867. These first 5867 trees are used to construct fitted values, as well as to assess variable importance and partial dependence.

From the variable importance plot, we see that by relative influence, the most important variables are the offense group, current age, and education. Other variables contribute minutely to the reduction in quantile loss. The first is unsurprising - the nature of the crime heavily influences the sentence being passed. However, it turns out that current age and education also influence the sentencing relatively strongly. We note also that there multicollinearity may be at play, as age is correlated with years of education. While we cannot ascribe causal effects, we might expect the young receive lighter sentences due to the juvenile justice system. We note that these results are standardized such that the relative influence of the different predictors add to 100. It does not tell us directly the increase in loss that would result if we shuffle

Consider the partial dependence plots. In interpreting them, we do not know how predictors were transformed or combined; we cannot ascribe causal relationships to the single variables. We only know that all other variables were held constant as the predictor being examined was held constant. All relationships discussed are on average.

In figure 12, we see that for sentences vary considerably and in an inexplicable way with the data. For example, those with three years of education receive lightest sentences at 29 years; however, those with seven years of education receive sentences of around 33.5 years, close to the max observed of 35. We expect that this observed input-output response is a consequence of the dataset. We struggle to come up with other explanations for it. In terms of drug use, there is little variation in the average sentence received by confirmed drug use, which drops from slightly over 34 to just below 34 years.

In figure 13, we see that those with confirmed psychological conditions receive heavier sentences by 1.2 years at 35.2 years on average, as opposed to those without such conditions. As well, those with alcohol conditions receive about 0.1 years more on average. These differences are sufficiently small relative to the range of sentences observed that they may be discarded as noise or interpreted as being non-significant.

In figure 14, we see that males receive the heaviest sentences at 35.5 years, dropping to just below 32 years for females, and rising to just below 34.5 for those of unknown sex. This might be due to the difference in the natures of the offenses committed by the three sexes. Age is very interest. There is considerable variation here. Those at below 20 years of age receive light sentences always below 30 years; however, there is a steep ascent right after 20 years of age, peaking and plateauing at 36 years for those of age about 27. This may be function of the juvenile justice system in place, which allows younger offenders greater leniency for various offenses; furthermore, judges may be more willing to show leniency for young offenders. After a certain age, the adult justice system takes effect and judges may no longer sympathize. Then sentences around 36 are the average for all adult offenders.

In figure 15, we see that marital status has a small effect on the average length of sentences. Those who are separated receive the lightest sentences at below 33.5 years, while

those with unknown marital status receive the heaviest at about 35 years. These differences however are small enough to be a consequence of the data. All other marital status receive about the same sentence. And finally, as expected, the partial dependence plot for offense groups shows the greatest variation. Offense groups 61 and 64, which are related to driving under the influence, as well as miscellaneous offenses, receive the lightest sentences at just over 20 years. Offense group 1, related to murder and homicidal manslaughter, receive the heaviest sentences at over 80 years. This is to be expected. The justice system is designed to accord punishment based on the severity of the crime; hence, the worse of the offense, the heavier the punishment. This is reflected in the plot.

From figure 16, we see that our model systematically miss predictions for certain actual sentence lengths. This is recognizable by the prominent horizontal stretches. There is some underlying factor that results in these sentences that is not captured by our model. Furthermore, for predicted sentences of around 35-39 years, there is a systematically deviation from the actual sentences, which range from 0 to 120. Again, our model is missing some information that causes the divergence in actual sentence length, and this information is especially relevant for those with predicted sentences for 35-39 years.

The out of bag data used in evaluating the optimal number of iterations to stop at; we do not have the option to retain the unsampled data to assess the performance of our model. We depend on the linear combination over many iterations to help prevent overfitting.

Furthermore, we are using offense groups, which put into equivalence classes offenses deemed close to each other in nature or seriousness of consequences or damages. This is reflected in the appendix. Nonetheless, just as different cases differ in the severity of the crime committed, within the same offense group there is considerable variation in the severity of the offense. Group 13 encompasses many types of assaults, from first degree to fourth. Each of these have different minimum sentences and average sentences. Grouping them together confounds the different effect on the sentence of these different severities. This complicates and reduces the reliability of our findings. It doesn't always make sense to forecast based on an offense group if the offense committed is low in severity relative to the

other offenses in the group. This also complicates the assessment of whether a predicted sentence is an overestimate or an underestimate, and may invalidate our opinions based on the model of whether someone should plead guilty to a lesser offense or proceed to trial.

6 Conclusion

We chose boosted trees using stochastic gradient boosting over random forests as our forecasting model. Our final model has an observed alpha of 0.7455 and uses the first 5867 iterations of the tree.

We found that the variables which contributed most to loss reduction were the offense group, age and education. Other variables, marital status, sex, and other conditions, contribute minutely to reductions in loss. From the partial dependence plots, we find that we cannot explain the relationship between sentencing and level of education. Psychological conditions, drug use, alcohol problems and marital status have little effect on sentences. Gender minutely affects sentences, with males and those of unknown receiving slightly longer sentences on average. Sentences vary considerable with age; those below 20 years of age receive lighter sentences, past which threshold sentences jump dramatically and plateau for those around 27 years of age onwards. Offense groups predictably have the largest effect on sentencing. More severe crimes with higher minimum sentences receive the most severe sentences. Our model systematically fail to produce accurate predictions for certain actual sentencing lengths; it also fails to predict accurate sentences for those offenders and offenses which it deems to warrant around 38 years.

We would be hesitant to use our results in practice; there is considerable variability in the accuracy of our predictions, and there are instances where we systematically fail to produce accurate estimates. At best, we would use our results as another source of input in guiding the defendants in deciding whether to plead to a lesser offense or go to trial.

An extension of this work could move to forecasting conditional probabilities of conviction contingent on the characteristics of the offender and the nature of the offense. This

work could contribute greatly to our efforts as attorneys. Currently, we only have the predicted sentence supposing that someone were convicted. However, the probability of conviction may be very low, and it may turn out that the expected sentence may be very low as well if we were to proceed to court. Hence, we would need a database of all cases tried, instead of the data here which only has information on offenders actually convicted.

7 Appendix

```
load("~/Dropbox/University of Pennsylvania/S2016/STAT474/Project 4 -
    Boosting/Sentence.rdata")
attach(Sentence)
library(stargazer)
library(gbm)
library(GGally)
#univariate
summary(Sentence)
par(mfrow=c(2,2))
hist(Sentence$NominalSentence, main = "Histogram of Nominal Sentences", xlab
     = "Years", ylab = "Frequency", col = "grey")

educ <- Sentence$Education
educ <- as.numeric(educ)
hist(Sentence$educ, main = "Histogram of Education", xlab = "Years", ylab =
    "Frequency", col = "grey", xlim = c(0,20), breaks = 19)

plot(Sentence$DrugCondition, main = "Histogram of Drug Conditions", ylab = "
    Frequency", col = "grey")
plot(Sentence$PsychCondition, main = "Histogram of Psychological Conditions
    ", ylab = "Frequency", col = "grey")
plot(Sentence$AlcoCondition, main = "Histogram of Alcohol Conditions", ylab
    = "Frequency", col = "grey")
plot(Sentence$Sex, main = "Histogram of Sex of Offender", ylab = "Frequency
    ", col = "grey")
hist(Sentence$CurrentAge, main = "Histogram of Age of Offender", xlab = "Age
    ", ylab = "Frequency", col = "grey")
plot(Sentence$MaritalStatus, main = "Histogram of Marital Status", ylab = "
    Frequency", col = "grey")
plot(Sentence$as.factor.OffGrp1., main = "Histogram of Offense Groups", xlab
     = "Offense Group Number (See Appendix)", ylab = "Frequency", col = "
    grey")
plot(Sentence$as.factor.Offense1., main = "Histogram of Offenses", xlab = "
    Offenses", ylab = "Frequency", col = "grey", las = 0)

Sentence_missing <- subset(Sentence, is.na(Sentence$NominalSentence))

plot(Sentence_missing$DrugCondition, main = "Histogram of Drug Conditions",
    ylab = "Frequency", col = "grey")
plot(Sentence_missing$PsychCondition, main = "Histogram of Psychological
    Conditions", ylab = "Frequency", col = "grey")
plot(Sentence_missing$AlcoCondition, main = "Histogram of Alcohol Conditions
    ", ylab = "Frequency", col = "grey")
plot(Sentence_missing$Sex, main = "Histogram of Sex of Offender", ylab = "
    Frequency", col = "grey")
hist(Sentence_missing$CurrentAge, main = "Histogram of Age of Offender",
    xlab = "Age", ylab = "Frequency", col = "grey")
plot(Sentence_missing$MaritalStatus, main = "Histogram of Marital Status",
    ylab = "Frequency", col = "grey")
plot(Sentence_missing$as.factor.OffGrp1., main = "Histogram of Offense
    Groups", xlab = "Offense Group Number (See Appendix)", ylab = "Frequency
    ", col = "grey")
```

Table 2: Summary of Offenses by Offense Groups

Group	Key Offenses
1	Murder and Homicide Manslaughter
2	Rape and Sexual Offense, Abuse, Sodomy
7	Kidnapping, False Imprisonment, Abduction
10	Robbery with deadly weapon, Carjacking
13	Assaults, Stalking, Reckless Endangerment
16	Robbery
19	Arson, Burning Personal Property, Treat of Arson
22	B&E, Burglary, Attempted Burglary
25	Theft, Thievery
28	Autotheft, Unauthorized Usage
31	Possession of Stolen Property
34	Firearms violations, possession and violence
37	Miscellaneous financial fraud, impersonation, false pretenses, police reports
40	Forgery and Counterfeiting
43	Escape, Prison Contraband, Contempt of Court, intimidation of witness obstruction of justice, aidding and abetting, attempt to flee police
46	Perverted practices, sex offender registration, sexual offense prostitution, child abuse (sexual)
49	controlled drugs offenses, incl. distribution, possession, conspiring
52	Illegal lottery, gambling
55	child abuse, domestic nonsupport, violated of protect order
58	Conspiracy, accessory after, public trust
61	Driving under influence and fleeing scene
64	CP, rogue, alcohol, animal cruelty, disturbing peace, public order obscene materials, littering, home improvement

```
plot(Sentence_missing$as.factor.Offense1., main = "Histogram of Offenses",
    xlab = "Offenses", ylab = "Frequency", col = "grey", las = 0)

#bivariate
Sentence_Bivariate <- subset(Sentence, select = c(Education, CurrentAge,
    MaritalStatus, as.factor.OffGrp1.))
Sentence_Bivariate$Education <- sort(Sentence_Bivariate$Education,
    descending = F)
ggpairs(Sentence_Bivariate)


#cleaning & data partition
Sentence_Clean <- na.omit(Sentence)
index <- sample(1:55937, 55937, replace = F)
Sentence_Clean <- Sentence_Clean[index,]
training <- Sentence_Clean[1:37292,]
testing <- Sentence_Clean[37292:55937,]

#model building
attach(training)
set.seed(123)
out1 <- gbm(NominalSentence ~ Education + DrugCondition + PsychCondition +
    AlcoCondition + Sex + CurrentAge + MaritalStatus + as.factor.OffGrp1.,
            data = training,
            distribution = list(name = "quantile", alpha = 0.75),
            n.cores = 4,
            bag.fraction = 0.5,
            shrinkage = 0.001,
            n.trees =10000,
            interaction.depth = 5,
            n.minobsinnode = 10)

par(mfrow=c(1,1))
OOBntrees1 <- gbm.perf(out1, oobag.curve = T, method = "OOB")
p1 <- predict(out1, newdata = training, n.trees = OOBntrees1, type = "link")
sum(p1 > training$NominalSentence)/length(p1)


out2 <- gbm(NominalSentence ~ Education + DrugCondition + PsychCondition +
    AlcoCondition + Sex + CurrentAge + MaritalStatus + as.factor.OffGrp1.,
            data = training,
            distribution = list(name = "quantile", alpha = 0.75),
            n.cores = 4,
            bag.fraction = 0.625,
            shrinkage = 0.001,
            n.trees =10000,
            interaction.depth = 5,
            n.minobsinnode = 10)

par(mfrow=c(1,1))
OOBntrees2 <- gbm.perf(out2, oobag.curve = T, method = "OOB")
p2 <- predict(out2, newdata = training, n.trees = OOBntrees2, type = "link")
sum(p2 > training$NominalSentence)/length(p2)

out3 <- gbm(NominalSentence ~ Education + DrugCondition + PsychCondition +
```

```
    AlcoCondition + Sex + CurrentAge + MaritalStatus + as.factor.OffGrp1.,
             data = training,
             distribution = list(name = "quantile", alpha = 0.75),
             n.cores = 4,
             bag.fraction = 0.375,
             shrinkage = 0.001,
             n.trees =10000,
             interaction.depth = 5,
             n.minobsinnode = 10)

par(mfrow=c(1,1))
OOBntrees3 <- gbm.perf(out3, oobag.curve = T, method = "OOB")
p3 <- predict(out3, newdata = training, n.trees = OOBntrees3, type = "link")
sum(p3 > training$NominalSentence)/length(p3)

out4 <- gbm(NominalSentence ~ Education + DrugCondition + PsychCondition +
    AlcoCondition + Sex + CurrentAge + MaritalStatus + as.factor.OffGrp1.,
             data = training,
             distribution = list(name = "quantile", alpha = 0.75),
             n.cores = 4,
             bag.fraction = 0.5,
             shrinkage = 0.001,
             n.trees =10000,
             interaction.depth = 9,
             n.minobsinnode = 5)

par(mfrow=c(1,1))
OOBntrees4 <- gbm.perf(out4, oobag.curve = T, method = "OOB")
p4 <- predict(out4, newdata = training, n.trees = OOBntrees4, type = "link")
sum(p4 > training$NominalSentence)/length(p4)


#final model
final <- out4
OOBntreesfinal <- gbm.perf(final, oobag.curve = T, method = "OOB")
finalpredtraining <- predict(final, newdata = training, n.trees =
    OOBntreesfinal, type = "link")
sum(finalpredtraining > training$NominalSentence)/length(finalpredtraining)

#partial dependence
par(mfrow = c(2,1))
plot(final, "Education", OOBntreesfinal, type = "link")
plot(final, "DrugCondition", OOBntreesfinal, type = "link")
plot(final, "PsychCondition", OOBntreesfinal, type = "link")
plot(final, "AlcoCondition", OOBntreesfinal, type = "link")
plot(final, "Sex", OOBntreesfinal, type = "link")
plot(final, "CurrentAge", OOBntreesfinal, type = "link")
plot(final, "MaritalStatus", OOBntreesfinal, type = "link")
plot(final, "as.factor.OffGrp1.", OOBntreesfinal, type = "link")

#variable importance
par(mfrow=c(1,1))
summary(final, method = relative.influence)
#can't use permutation.test.gbm which is similiar to that for random forest
    as it's not supported for quantile regresison
```

```
#honest performance assessment
predictions <- predict(final, newdata = testing, n.trees = OOBntreesfinal,
    type = "link")
plot(predictions, testing$NominalSentence, xlab = "Predicted Nominal
    Sentence", ylab = "Actual Nominal Sentence", main = "Predictions on
    Testing Data, alpha = 0.75")
abline(0,1,col="blue", lwd = 2)
#observed relative cost
sum(predictions > testing$NominalSentence)/length(predictions)
```

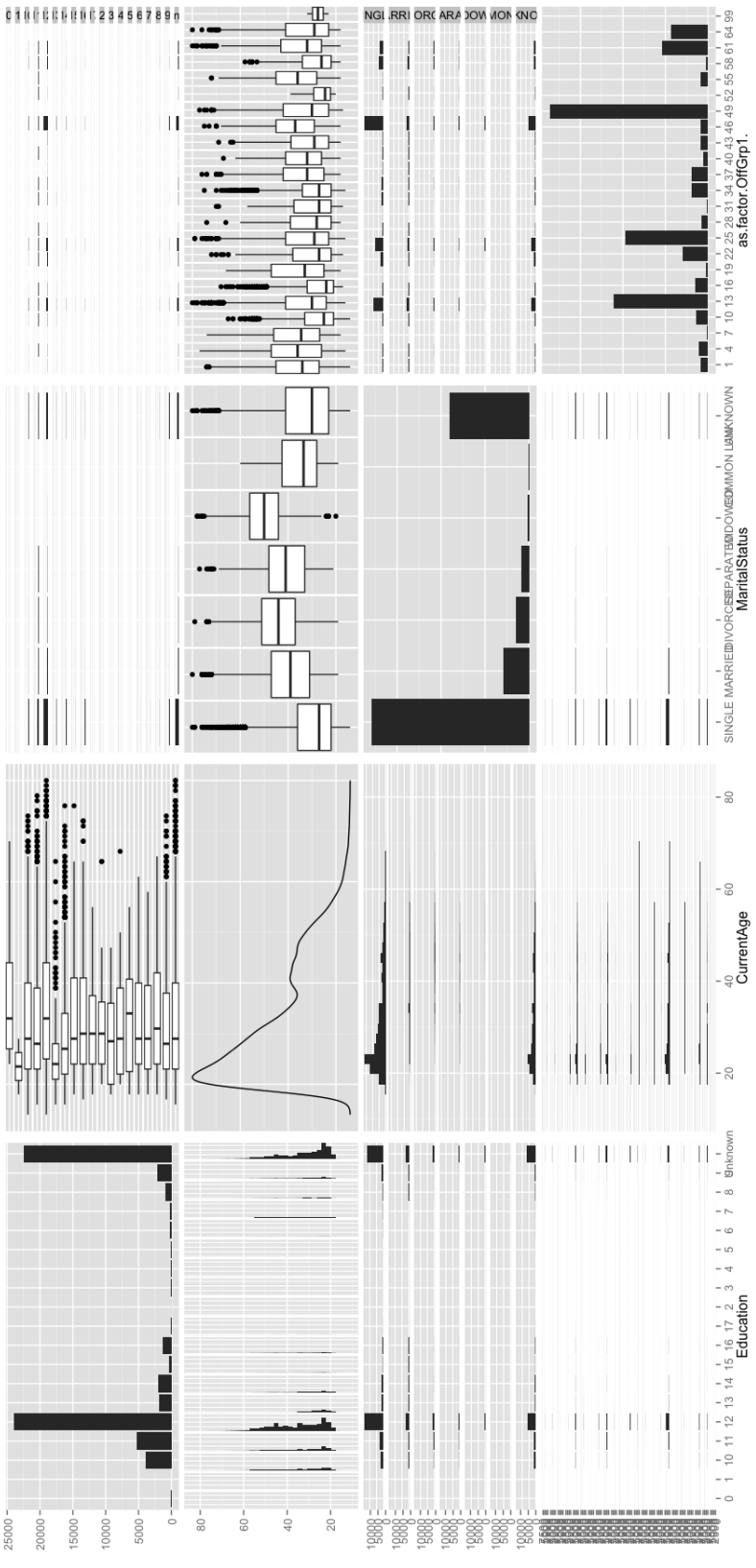


Figure 6: Property profile of the diverse library compared to the compound pool.

Figure 7: Univariate Analysis - Histograms for Observations with Missing Sentences

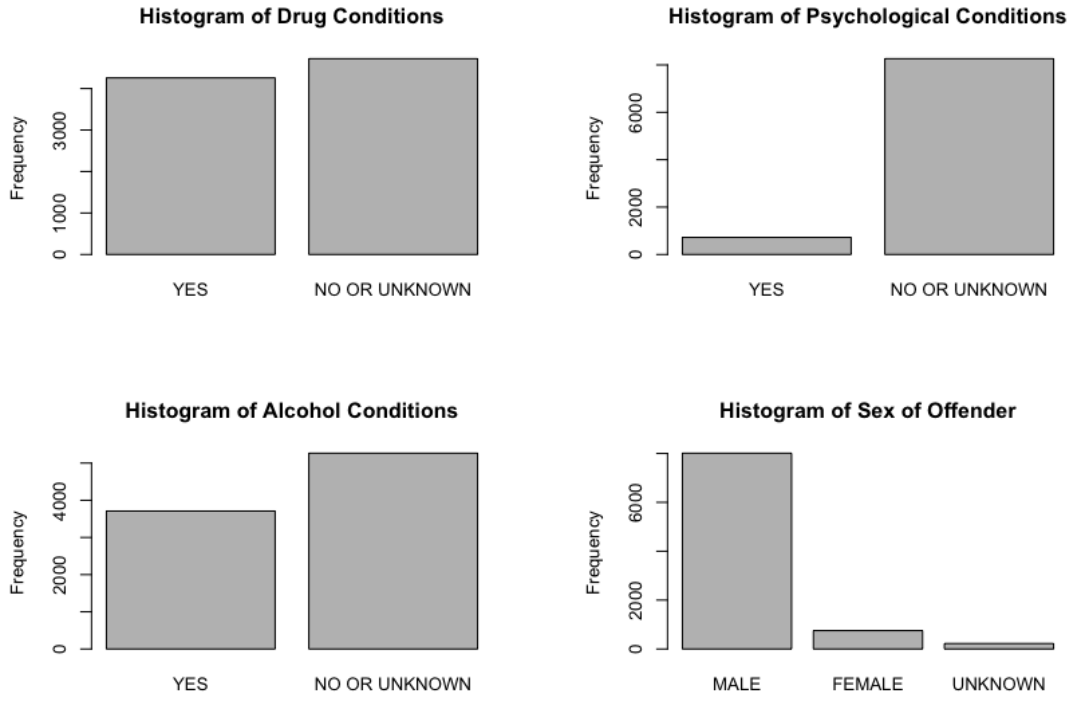


Figure 8: Univariate Analysis - Histograms for Observations with Missing Sentences

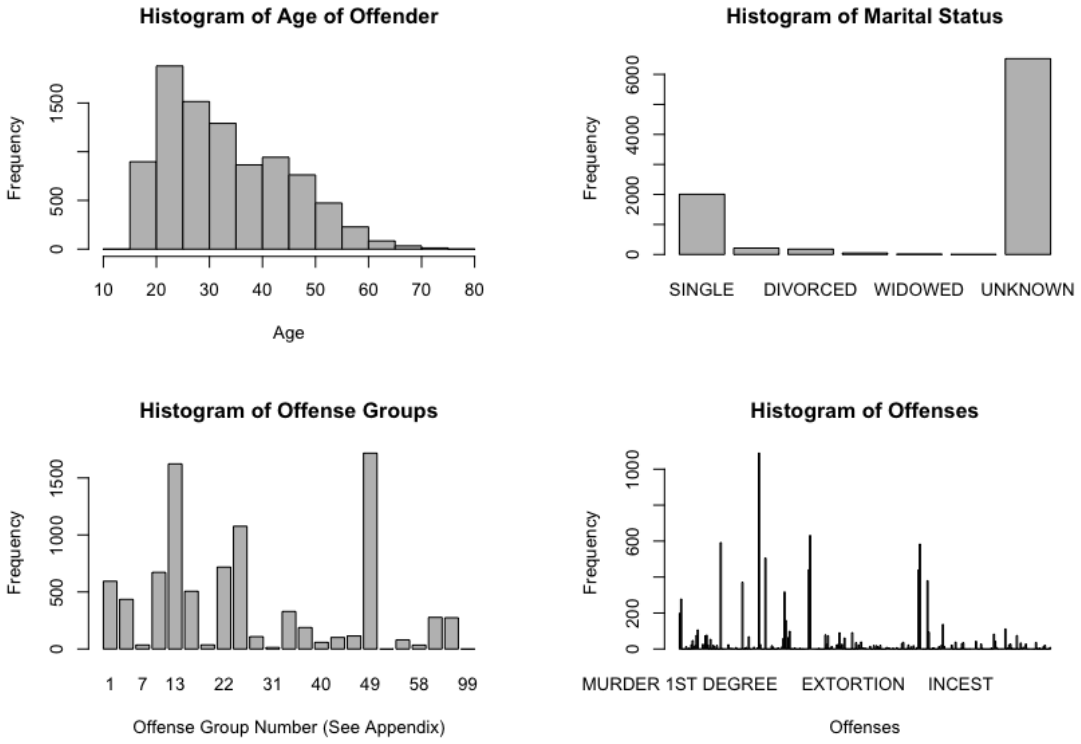


Figure 9: Final Specifications - Quantile Loss vs Iteration

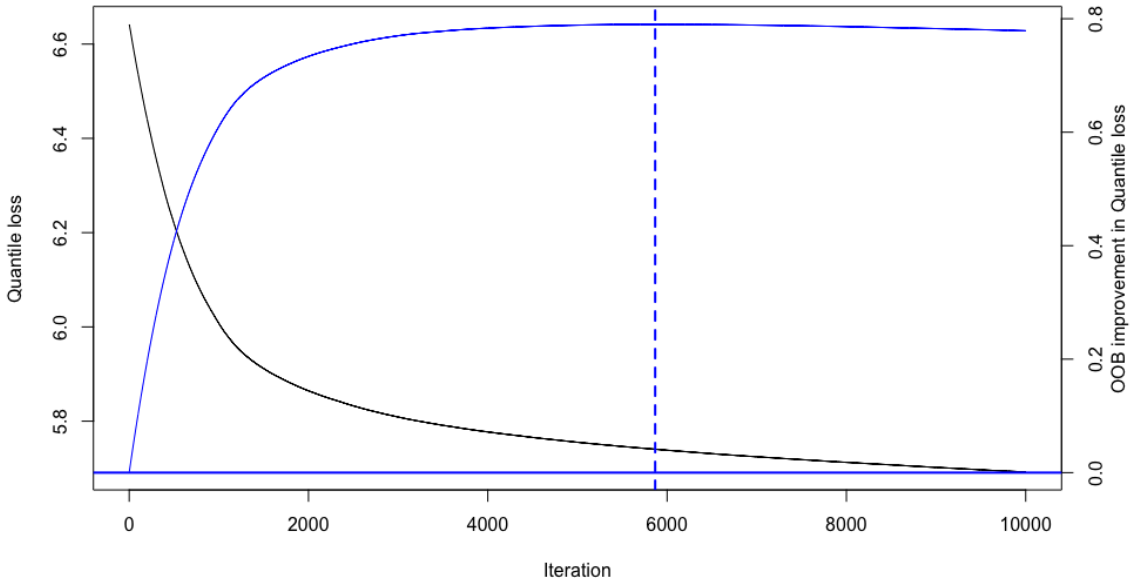


Figure 10: Final Specifications - Change in Quantile Loss vs Iteration

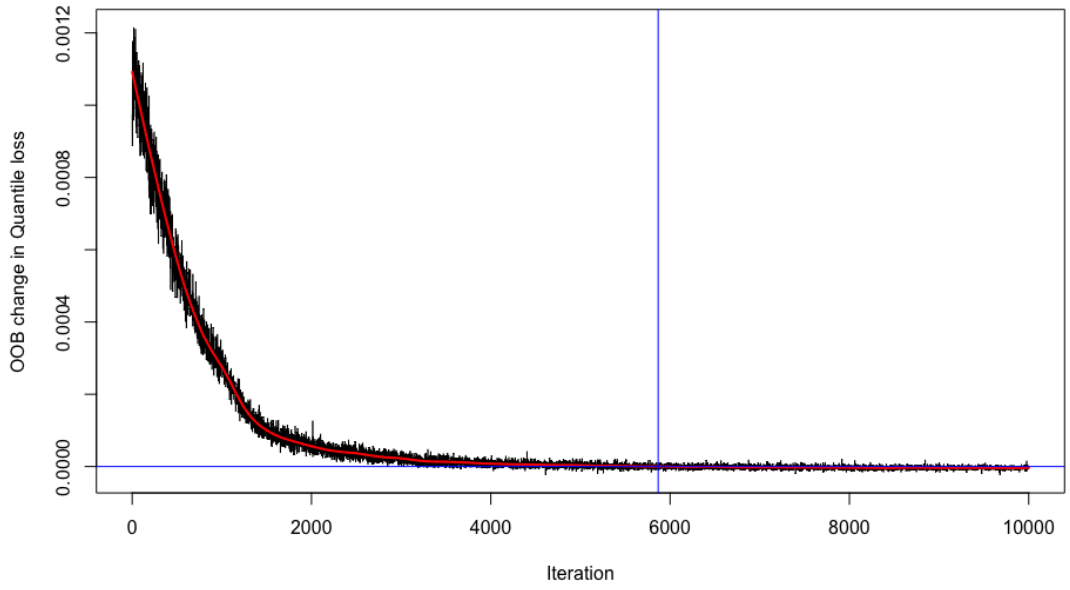
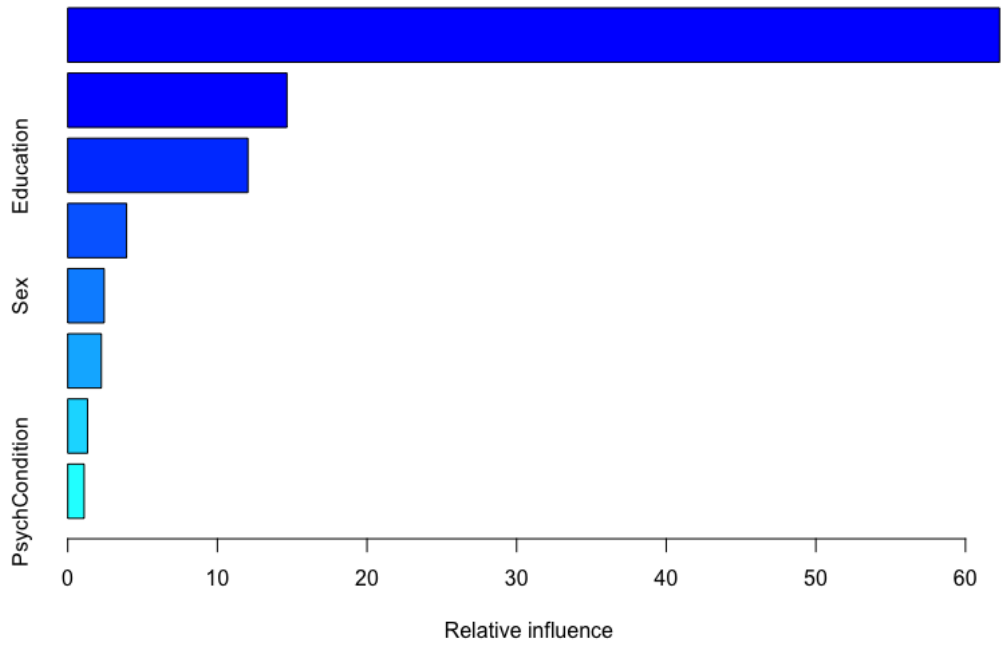


Figure 11: Final Specifications - Variable Importance Plot, Relative Influence



	var	rel.inf
as.factor.OffGrp1.	as.factor.OffGrp1.	62.277670
CurrentAge	CurrentAge	14.648917
Education	Education	12.041564
MaritalStatus	MaritalStatus	3.930343
Sex	Sex	2.435573
AlcoCondition	AlcoCondition	2.244672
DrugCondition	DrugCondition	1.324347
PsychCondition	PsychCondition	1.096913

Figure 12: Final Specifications - Partial Dependence Plots

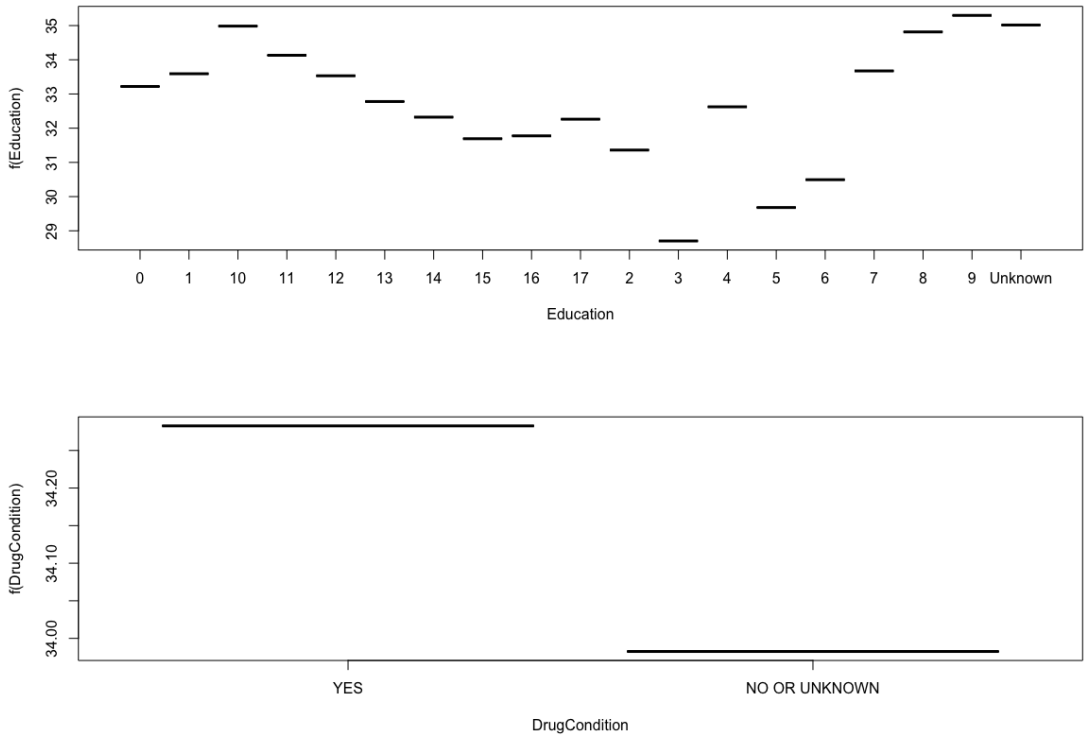


Figure 13: Final Specifications - Partial Dependence Plots

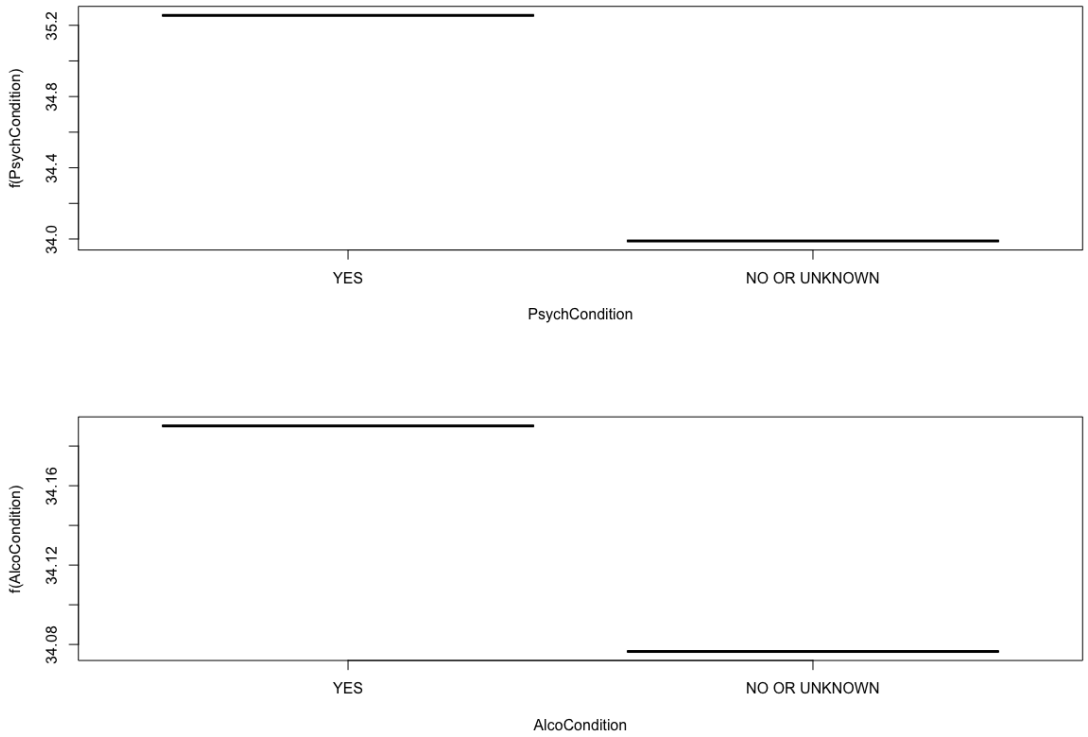
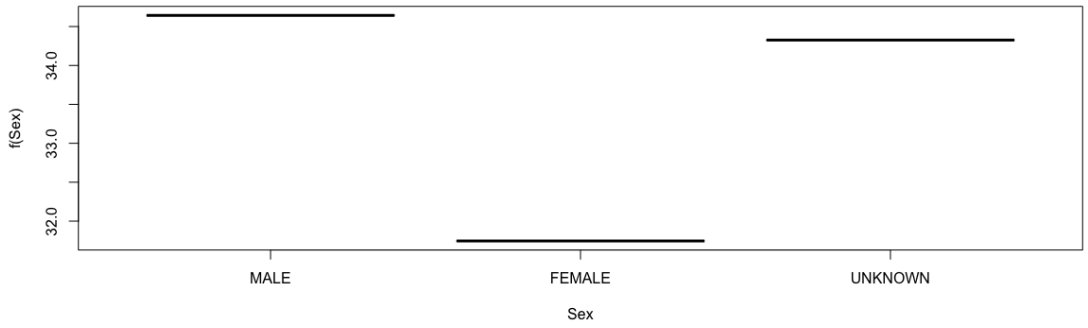


Figure 14: Final Specifications - Partial Dependence Plots



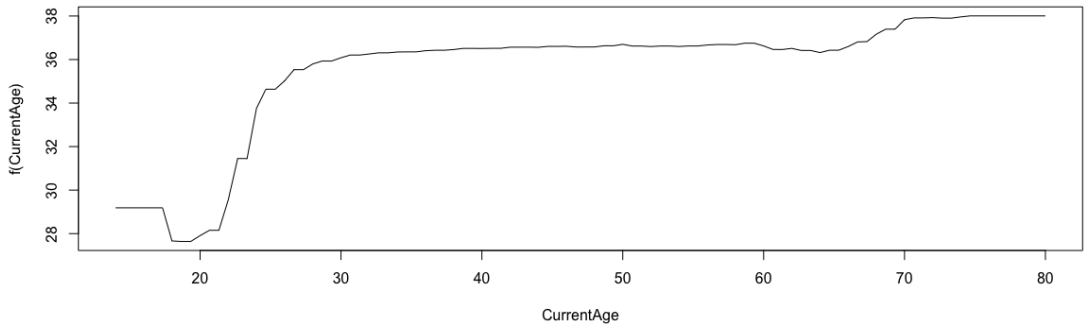
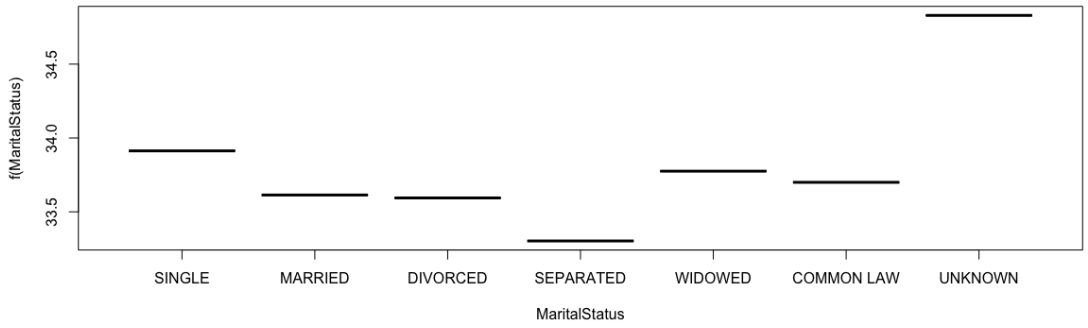


Figure 15: Final Specifications - Partial Dependence Plots



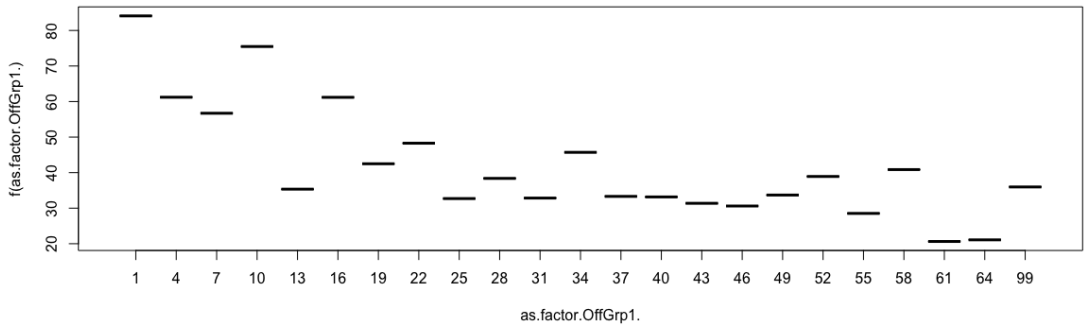
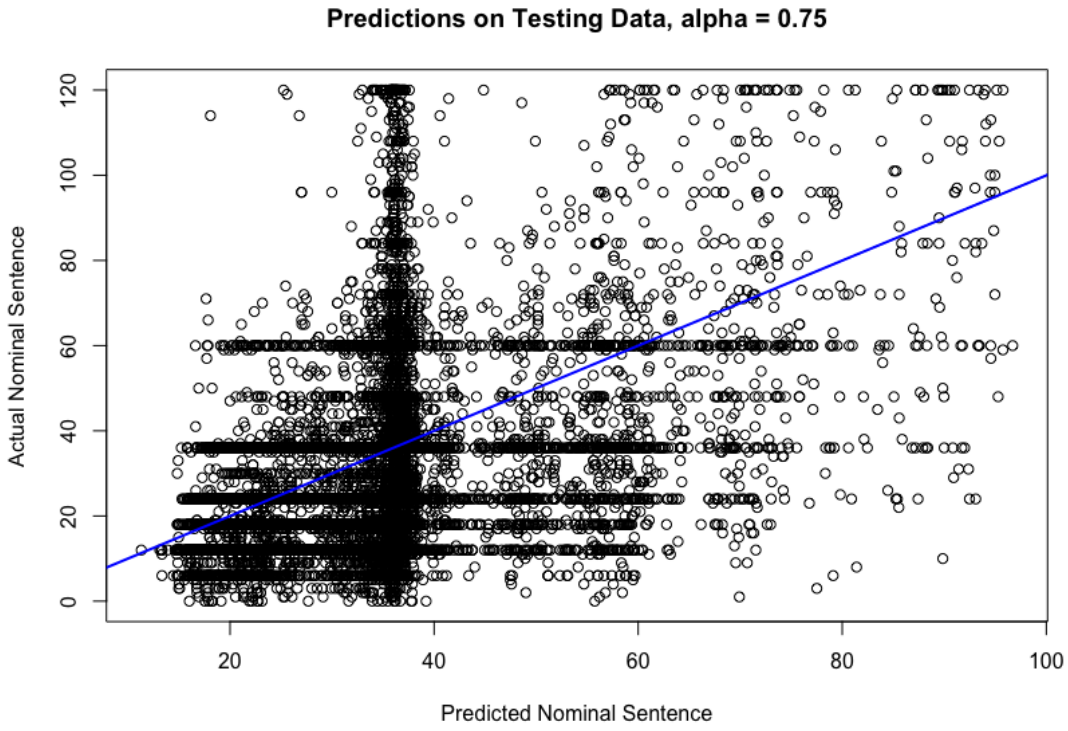


Figure 16: Final Specifications - Predictions vs Actual Sentence Length on Testing Data




```
Observed alpha = 0.7455
```