EXPLAINING HEROIN USAGE IN THE UNITED STATES: DATA MODELLING USING THE NSDUH 2014 DATASET

By

MATTHEW J. BEATTIE

University of Oklahoma

Norman, OK

December 2016

Research Project for:

DSA5103-995 Intelligent Data Analytics

# Executive Summary

In the last several years, heroin use in the United States has been growing at a rapid pace. It has grown to a point where the general media has sounded a grave tone of alarm. In January 2016, the New York Times pointed out that “deaths from drug overdoses have jumped in nearly every county across the United States, driven largely by an explosion in addiction to prescription painkillers and heroin.” (Park, 2016)

In this analysis, we seek to explore what patterns can explain heroin use in the United States. While there are many potential factors that could explain heroin usage, we chose to focus on those that could be explored in a time tested, well designed and documented study. In particular, we chose to examine the 2014 National Study for Drug Use and Health (NSDUH), a dataset containing 55,271 observations with 3,148 variables. We then opted to model heroin use as an outcome from a combination of predictors within the dataset.

We limited our consideration to several areas, including drug usage patterns, mental health, and demographics. This still left us with 405 variables to investigate. Because we were seeking an explanation of usage, and because the number of potential variables is so large, we concentrated our work on dataset preparation, decision tree analysis via CART and AdaBoost methods, and logistic regression modelling. The outcome of the decision tree analyses showed that heroin usage could be explained to a great extent by usage of other drugs, in particular cocaine and prescription opioids, early usage of those drugs, and by prior treatment for drug and alcohol abuse. We then validated this outcome more explicitly by performing a logistic regression, using heroin usage as the dependent variable and a set of factors highlighted by boosted decision tree analysis as predictors. Our regression analysis showed that certain factors greatly increased the likelihood of heroin use. For example, prior use of hallucinogens made respondents five times more likely to be categorized at heroin users.

For further study, we recommend looking at non-survey data, such as datasets from treatment center admission. This data does not rely on responder bias and can be used to map trends in heroin usage as it grows and correlate it to regional demographic information.

# Problem Description and Background

While drug use has been an ongoing problem in the United States, recent years have seen an explosion in more dangerous drugs, such as prescription opioids and heroin. The impact of this increase has been drastic growth in the number of drug poisoning deaths since 1999. In 1999, there were a total of 16,849 drug deaths in the U.S., and in 2014 that number had become 47,055. Perhaps even more alarmingly, the rate of growth in drug deaths was steady and showed no signs of declining. (Rossen LM, 2016).

Several categories of drugs account for this increase, including prescription opioids and benzodiazepines. However, heroin deaths have seen the sharpest increase in the last four years, as shown in Fig. 1. (National Institute on Drug Abuse, 2015). For that reason, we have chosen to try to explain heroin usage as an outcome of a set of predictors. In order to do so, we used the 2014 National Study on Drug Use and Health (NSDUH).

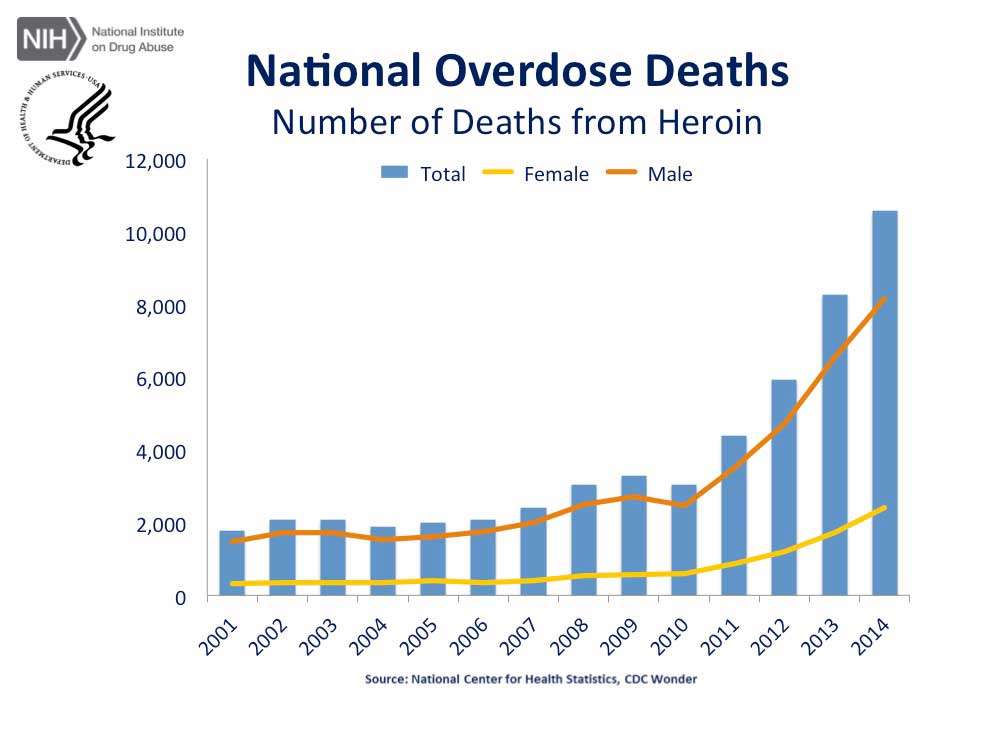


Figure 1: Heroin Overdose Deaths

The NSDUH is a study that began in 1980, and is intended to measure the prevalence and correlates of drug use in the United States. The study is conducted via a survey which is sponsored by the Center for Behavioral Health Statistics and Quality, a component of the Substance Abuse and Mental Health Services Administration (SAMHSA). The survey is conducted by RTI International. The 2014 study and an archive of previous studies is made available to the public at <http://www.datafiles.samhsa.gov>.

The study consists of 55,271 observations with 3,148 variables, and is available in several formats, including a pre-constructed R dataset. The design of the study has been refined over its history and has been built to represent virtually all census blocks of the United States. The principal weakness of the study is that it is based upon voluntary information from its participants. For our purposes, the biggest problem presented by this method is in the prevalence of missing data for many of the questions on the survey. As we will see in the analysis, SAMHSA has solved this problem for us through complicated imputation techniques involving modelling and manual data correction. Consequently, most of the data engineering we performed was on preliminary variable selection.

# Exploratory Data Analysis

As mentioned earlier, the NSDUH data set is large and complicated. Fortunately, it is published with an 888 page codebook, the National Survey on Drug Use and Health, 2014 Codebook, published by the Inter-University Consortium for Political and Social Research. The survey is broken up into several sections. The core section contains answers to questions regarding the use and frequency of use of specific drugs. Other sections include questions regarding drug treatment, mental health, mental health treatment, income and insurance, employment, and other demographics.

## Missingness

The primary challenge in working with the NSDUH data is missingness. Raw data are represented by direct answers from respondents to specific questions, and are subject to a high degree of skipped questions, refusals to answer, and other situations. There are 582 variables in the core raw data section. We can choose a subset of these, of likely interest to us, to depict the degree of missing data in this section. In this case, we chose the variable we with to model, HEREVER, which represents whether a user has ever used heroin, along with similar variables corresponding to other drugs. As we can see, over 80% of the responses for crack and codeine abuse are missing. We could make the reasonable assumption that a missing response represents a “no” or a skipped question. Alternatively, we could use imputation to fill in the missing data, but by doing so we would ignore better information that arises from responses to other questions in the survey.

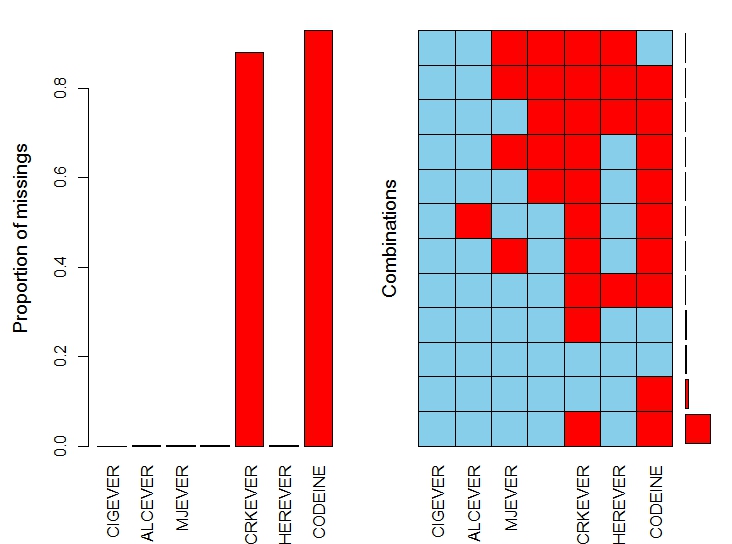


Figure 2: Example of Missing Raw Data in 2014 NSDUH

Fortunately, SAMHSA has developed a thorough methodology for eliminating missingness. In 1999, they developed an imputation method specific to the study – predictive mean neighborhood. This method combines model-assisted and nearest-neighbor hotdeck methods. SAMHSA recommends the use of the resulting imputed variables for multivariate analyses. (SAMHSA, 2014).

## Basic Analysis and Visualization

Since we are interested in the population of heroin users, we can conduct some basic analysis to see if there are any simple discernable patterns in the data that would assist us in starting our analysis. The first thing we note is that there are only 942 respondents who have ever used heroin[[1]](#footnote-1). This is a very small portion of the total population of respondents, and because it is small, we will need to be careful to draw only statistically valid conclusions.

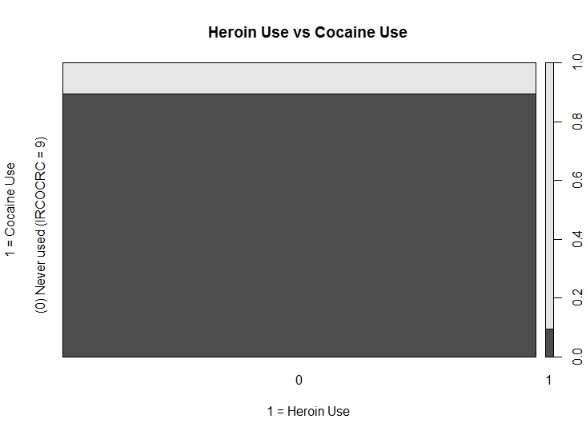


Figure 3: Cocaine Use

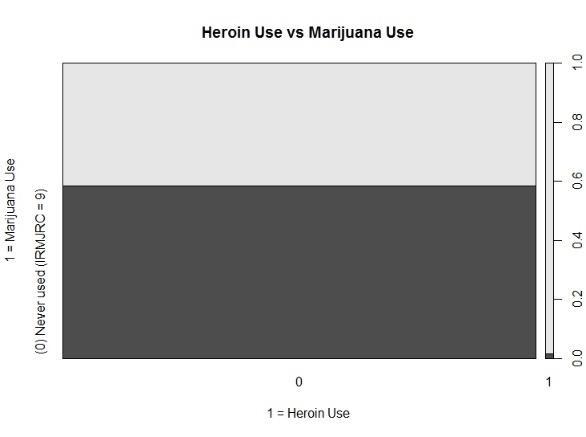


Figure 4: Marijuana Use

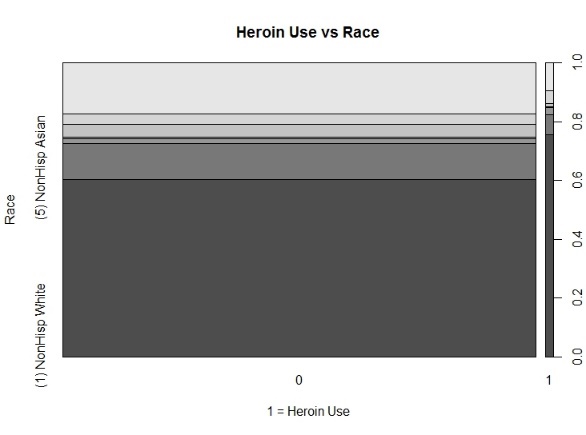


Figure 5: Distribution of Race

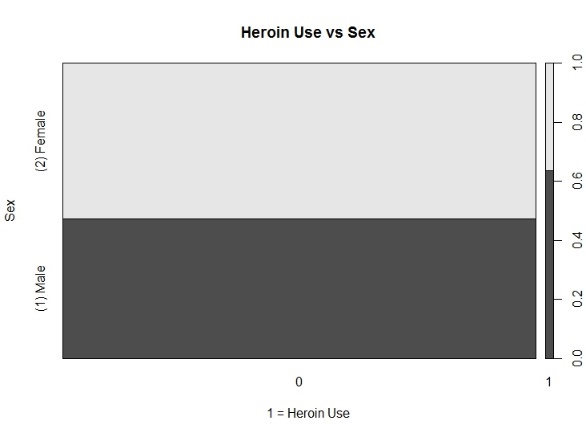


Figure 6: Distribution of Sex

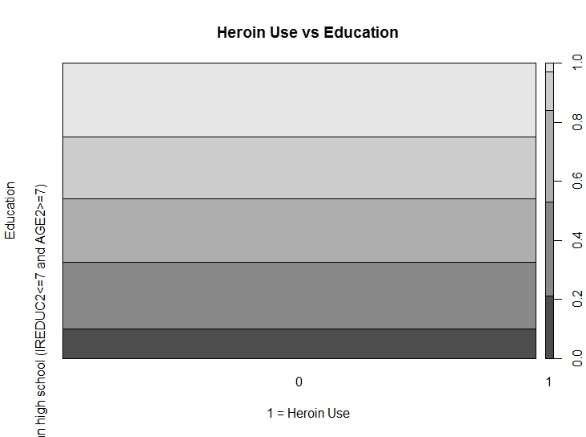


Figure 7: Distribution of Education

Next, we shall investigate the data to see if there are any obvious differences in either demographics or other drug usage between heroin users and non-heroin users. We chose to look at several characteristics: sex, marital status, education level, cocaine usage, and marijuana usage. In each case, we did indeed see a difference in distribution of each of the predictors for heroin users (Figures 3-7).

From these visualizations we can see some potential predictors of heroin use. In appears that heroin users are more likely to be male, white (non-Hispanic), and their education level deviates slightly from that of non-heroin users. However, we can see a stronger potential correlation with the use of other drugs. Heroin users appear more likely to have used marijuana, and they appear to be strongly more likely to have used cocaine. It is important to note that while these distribution differences may seem promising, and that these variables may be good predictors for heroin use, there are a total of 3,148 variables, and a more robust means of selecting potential important factors is necessary.

# Analysis Plan

This analysis seeks to explain, not predict, potential heroin use. We are not concerned with predicting, to a high degree of accuracy, whether or not an individual is likely to use heroin in their lifetime. Instead, we are trying to identify some patterns that may indicate some commonality of background among heroin users. In other words, we would like to determine which of the 3,148 possible predictors are most influential in identifying a likely heroin user. Because this is our goal, and because of the large number of variables in the NSDUH dataset, we elected to use two categorization methods in our analysis: decision trees and logistic regression.

We decided that a simple decision tree model would be a good way to explain heroin usage. We opted to use the CART method of decision tree analysis using information gain as a node selection criteria. With such a large number of variables, we anticipate that CART will yield a very complex tree. To reduce that tree to an interpretable model, we will add penalization for large trees via the use of the complexity parameter. Our resulting tree should be one that is small enough to be of value and highlights variables that contribute most to classification. For this first portion of our analysis, we are not using tree ensembles, which while more accurate than CART, preclude our ability to easily understand the tree model.

The small fraction of heroin users in the NSDUH dataset suggests a weakness in the CART approach. CART is very sensitive to observation selection, and very different trees can arise from consideration of different sets of data. For this reason, the second portion of our analysis will be done via logistic regression, an approach which is more stable. Rather than conduct a regression on the entire set of variables, which would be computationally complex and result in coefficients of little interest, we will constrain the predictors of the model to those most likely to be of importance. To select these predictors, we will use a boosted tree classification model – the AdaBoost algorithm. This model, a tree ensemble, allows us to identify which variables are most influential in classification. Because we are not concerned with using the boosted tree to explain heroin usage, we are not concerned with the complexity of its output. Instead, we will use the variables identified by AdaBoost as predictors for heroin use in the logistic regression procedure.

As mentioned earlier, we are using the imputed and corrected variables in the NSDUH dataset, which allows us to avoid dealing with the high degree of missingness in the raw data questions. For validation, we will conduct a brief survey of peer reviewed literature on heroin usage.

# Results and Validation

## Data Engineering and Preparation

While there are 3,148 total variables in the NSDUH dataset, we can reduce the number of ones to consider by narrowing our scope. First, we can ignore any nonimputed or noncorrected raw variables, as those have already been transformed into more manageable features. Next we constrained our investigation to look for a relationship between heroin use and several areas of interest. We chose to use variables related to:

* Tobacco, alcohol, and drug use
* Substance dependence and abuse – we eliminated any variables from this section that directly correlated to heroin dependence
* Substance treatment, such as stays in rehabilitation centers
* Mental health treatment
* Mental health and suicidal tendency
* Demographics, including sex, race, education level, etc.
* Employment status

This left us with 405 variables. When running our decision tree analyses, we found other variables that also measured heroin use, such as “first heroin use under the age of 18”. Additionally, several variables categorized age in different ways, so we included only one of those. Our final dataset included 387 total variables and 55,271 observations. We then created a training dataset equal to 20% of the total observations for CART. Our training dataset size selection was based in part by performance of the algorithms. For our AdaBoost and logistic regression analyses, we increased the training set to 30% of the observations.

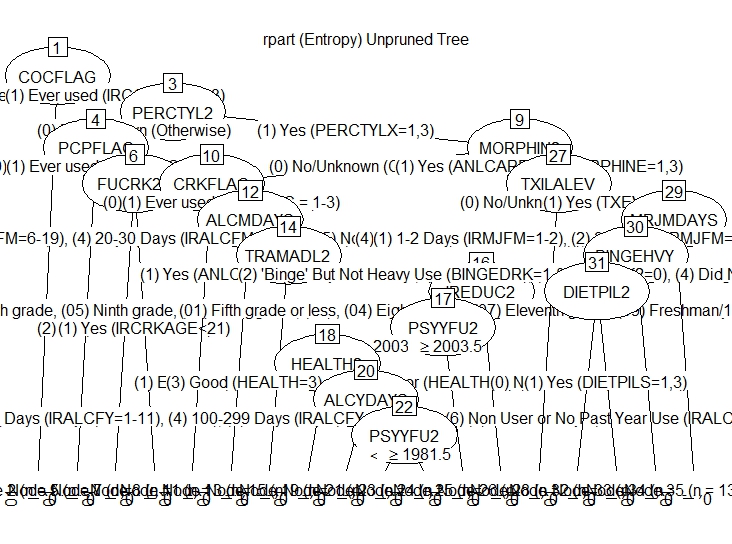


Figure 8: Unpruned CART Decision Tree

## CART Decision Tree Analysis

We ran a decision tree analysis on the training set using rpart(). We chose to use information gain as the criterion for feature selection at the nodes. When we did so, CART produced a complex tree that was not useful for intrepretation (Fig. 8). Therefore, we elected to rerun the model using a complexity parameter. As shown in the complexity parameter plot (See Appendix), the optimal complexity parameter was 0.0178. When we pruned our tree using this parameter, we obtained a much cleaner, interpretable tree (Fig. 9).

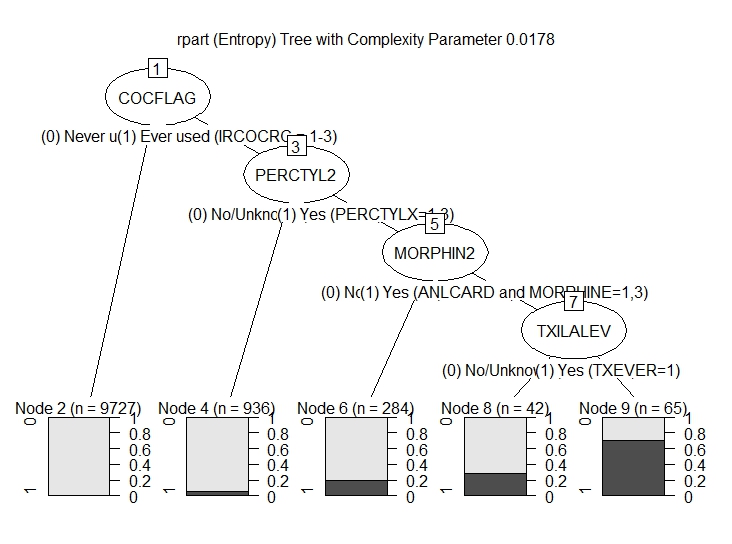


Figure 9: Pruned CART Decision Tree

In this tree we see four binary nodes that lead to categorization of heroin users. The first split is whether or not the respondent had ever used cocaine. If the answer to that question was “no”, the respondent was very unlikely to be a heroin user. Within the category of cocaine usage, the next split was based upon whether or not the respondent had ever used percoset, percodan or tylox – all prescription drugs that are commonly abused. The next split was on morphine usage, and at this point we start to see a higher proportion of heroin users. The final split was based upon whether or not the respondent had ever received treatment for drug and/or alcohol abuse. We can thus surmise that if a respondent had abused cocaine, prescription drugs, and morphine, and had received prior treatment for substance abuse, they were likely to have abused heroin as well.

While this outcome is relatively simple, we note that the most important variables in the unpruned tree were somewhat different. In this case, cocaine use was again the most important variable, but whether or not a user first used cocaine under the age of 21 was the second most important variable (See Appendix, rpart output). We will see this pattern again in the AdaBoost and logistic regression analyses.

To check the performance of the model, we ran it on the test dataset and ran a set of performance diagnostics[[2]](#footnote-2). Our primary means of evaluating the performance our models was the area under the receiver operating characteristic curve. For simplicity, we refer to these parameters as the AUC (area under curve), and ROC curve (Fig. 10). As with all of our models, the test accuracy of CART was very high. This is in part due to the large number of negative heroin use observations. Because our algorithms will in most cases predict “no heroin usage”, we should expect accuracy to be high and of little value in evaluating our models. For this reason, AUC and the kappa statistic are our primary performance criteria. The CART model performed well, with an AUC of 0.926 and kappa statistic of 0.405. However, we found the CART model to be very sensitive to training data selection. If we changed the size of the training set even a small amount, we saw some differences in variable importance. For this reason, we ran another simple tree analysis using the C5.0 algorithm. We did not obtain significantly different results with this method, nor did we see an increase in performance. Because C5.0 does not have robust visualization methods, we chose to keep our CART output as our baseline decision tree.



Figure 10: Model Performance

## AdaBoost Tree Analysis

AdaBoost, a tree ensemble method, creates a more stable outcome than CART. AdaBoost is a boosting algorithm that creates a tree based upon a set of equally weighted weak classifiers. It then increases the weights associated with incorrect classifications and recomputes the tree. The final tree is a majority vote combination of the classifiers. We can fix the number of iterations of the algorithm, and in this case we chose to limit AdaBoost to ten iterations. The AdaBoost algorithm also performed well, with an AUC of 0.908 and a kappa of 0.446 (Fig. 10).

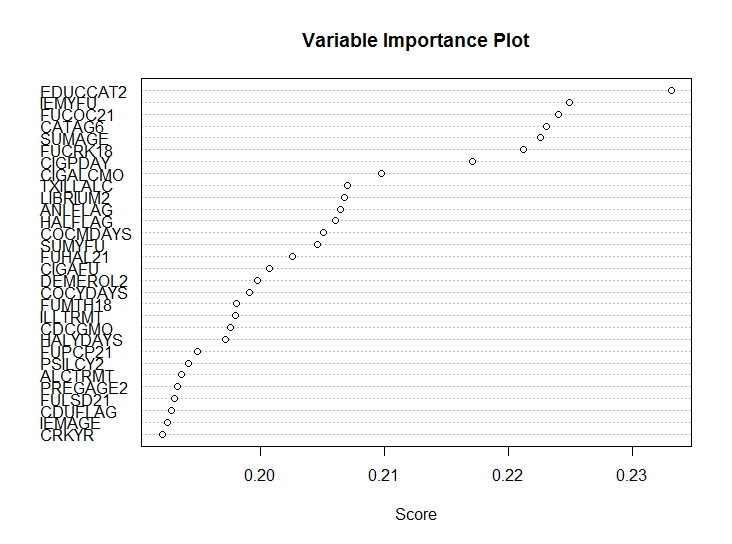


Figure 11: AdaBoost Variable Importance

Because it is a tree ensemble method, AdaBoost does not produce a simple binary tree from which we can explain heroin usage. Instead, it focuses on prediction accuracy. However, it does show us which variables were most important in the final tree. This list of variables (Fig. 11) is then available to us for use in logistic regression. AdaBoost identified different variables than CART as important. In particular, we see the importance of education level is now very high. As with CART, we also see the importance of first cocaine use under the age of 21. We also see other variables that regarding early drug use – whether or not the respondent first used crack under the age of 18, and the age of first illicit drug use. Two of the important variables are of little use to us. The year of a respondent’s first drug use (IEMYFU) is dependent on the age of the person, so we can discard it. Additionally, we chose to ignore the age of the respondent (CATAG6) because as age increased, the likelihood of heroin use increased as well. This is intuitive because an older respondent simply had more time over which heroin use could have occurred.

## Logistic Regression

For our final model, we performed a logistic regression using heroin use as a dependent variable and the eleven most important variables (excluding IEMYFU and CATAG6) as predictors. In order to build the model, we factorized age of first drug use from a number into six categories. This results in a model that consists entirely of factor predictors. We then recast the category variables in such a way as to make the base cases for dummy variable creation “No Use” cases. The logistic regression model performed similarly to the tree analyses (Fig. 10), with an AUC of 0.963 and a kappa statistic of 0.404. However, this model allows us to better understand the impact of individual variables on the likelihood of heroin use.



Figure 12: Logistic Regression Coefficients

The variables with a high degree of statistical significance in the regression model are shown in Figure 12.[[3]](#footnote-3) In this table, we show the coefficient for each variable in the logit model. Recall that these coefficients correspond to the logarithm of the odds of heroin usage: where represents the probability of heroin use and represents the regression coefficient corresponding to variable . The independent probability column represents the likelihood of the base case with the presence of the variable in question. The likelihood multiplier is the ratio of the independent probability of the variable to the base case (intercept coefficient only).

## Interpreting the Logistic Regression Results

If we consider the base case where all factor variables are not present, or zero, we are left with only the intercept coefficient. It is important to note that due to the way dummy variables are cast, this case includes the situation where the respondent is between 12-17 years old, never used an illicit drug, does not use cigarettes on a daily basis, and did not use cigarettes or alcohol within the past month. Unsurprisingly, the probability of heroin usage in this case is near zero. When we calculate the probability of heroin use in any of the cases where a factor is present, we are including the coefficient for both the intercept and the variable of interest, :

For example, the probability of heroin use for the base case is 1.486e-10. When we consider the case where the respondent has used hallucinogens, the probability increases to 7.536e-10. This value remains extremely low, but it’s size is dominated by the effect of the regression intercept. To better understand the impact of a variable, we can calculate a “likelihood multiplier” for variable as where is the probability when the factor is positive and is the probability of the base case. So for the “used hallucinogens” case, the likelihood multiplier is 5.073. We interpret this result by saying that respondents who have used hallucinogens are 5.073 times more likely to use heroin than the base case.

When we look at Figure 11, we see that education *increases* the probability of heroin use. This is because the base case includes respondents who are under 18, and should still be in school. As we have noted earlier, the older a respondent is (up to a point), the more likely they are to have used heroin. So instead we consider the differences within the education category and see that respondents with less than a high school education are more likely to use heroin than those with more education. We see that using other drugs, especially Librium (a prescription sedative), increases potential heroin use. And again, age of first drug use is again a factor, at least in the case where respondents who used cocaine before the age of 21 were about three times more likely to use heroin than the base case.

The logistic model for explaining heroin use is extremely robust. We can calculate likelihood multipliers for any combination of variables, and we can compare changes in likelihood from one set of positive factors to another. For example, we can say that respondents who used hallucinogens and first used cocaine before 21 are three times more likely that those who did not use cocaine before 21. We could also add back factors we did not consider in this model due to our AdaBoost analysis to look for other patterns.

## Validation

Surprisingly, a simple survey of academic literature for studies using the NSDUH dataset did not uncover an analysis using data science techniques. However, there are many studies which show results similar to what we have discussed, especially regarding the increase in likelihood of heroin use due to other drug use. Fergusson shows that early use of marijuana leads to use of other illicit drugs (Fergusson, Boden, & Horwood, July 2008). Lynskey found similar results (Lynskey, Andrew C. Heath, Bucholz, & al, January 2003). In a specialized study, Hartel found that methadone patients who used cocaine were more likely to resume heroin use (Hartel, et al., January 1995).

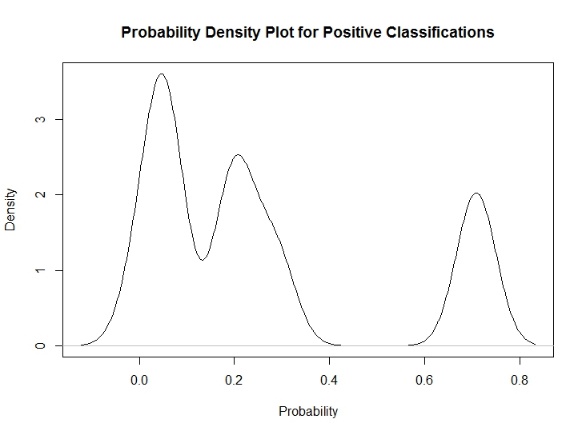
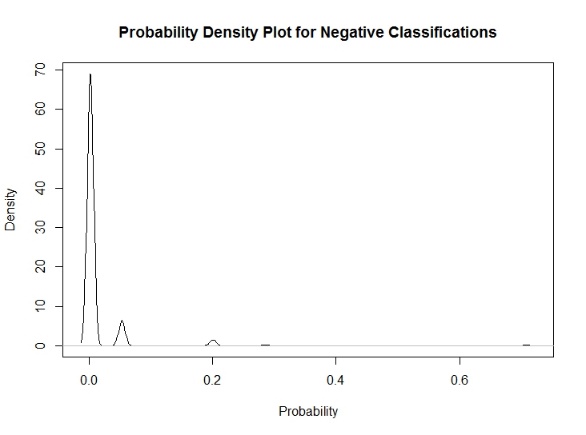
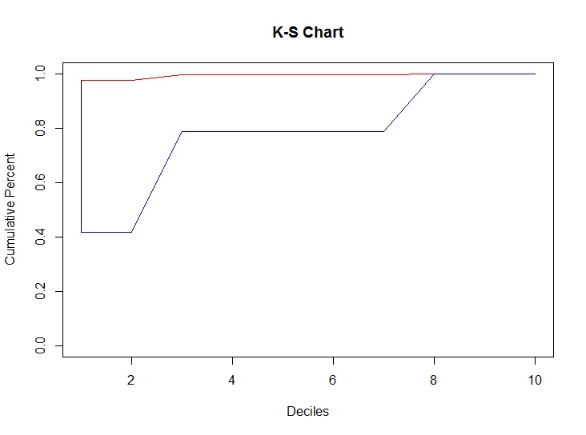
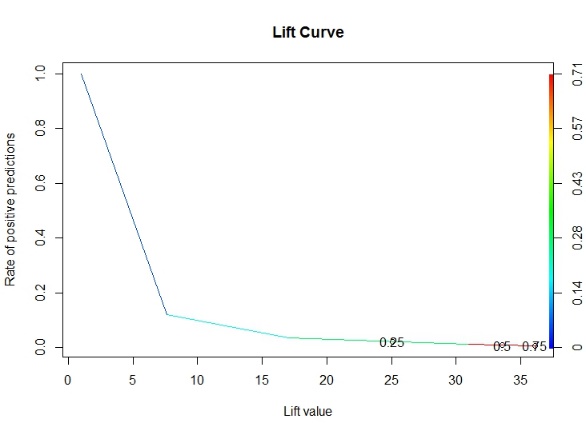
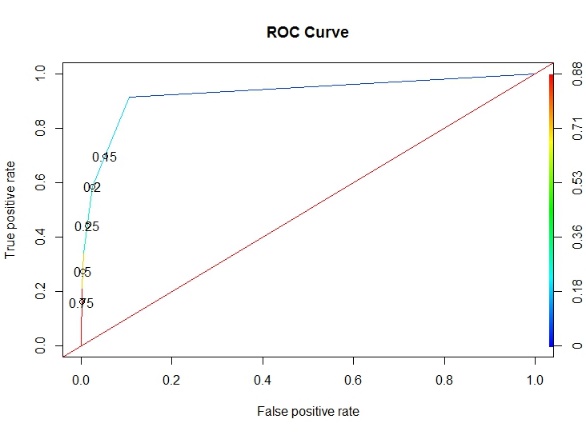
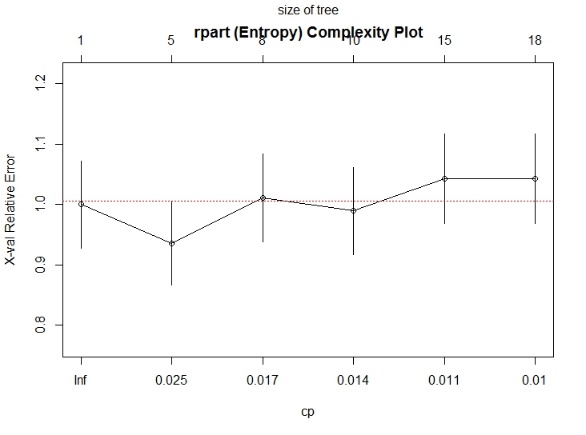
# Conclusion

The NSDUH dataset is an extremely useful tool for researchers, and has been published in such a way as to make it easy to analyze, especially using R. The dataset is also conditioned to relieve the researcher of some of the more difficult aspects of large dataset examination. In particular, missingness, a common feature of self-respondent surveys, has been alleviated through imputation and correction. However, the dataset is quite large and has an enormous number of variables, and researchers must use sophisticated techniques to find statistically valid conclusions from the study.

We used both decision tree analyses and logistic regression to explain factors that contribute to heroin use. In all cases, our models performed well, with high levels of accuracy and area under the receiver operating characteristic curve. The CART decision tree model showed that usage of other drugs, specifically cocaine, prescription opiates, and morphine, correlated to heroin usage, as did prior treatment for drug and/or alcohol abuse. When using a boosted decision tree algorithm, AdaBoost, we found that the most important variables weren’t just a factor of drug use, but highlighted drug use at an early age. Finally, our logistic regression analysis confirmed that use of other drugs, early use of cocaine or crack, and low education levels all increased the probability of heroin use. The logistic regression also pointed to an extreme impact due to early illicit drug use, but the lack of statistical significance of those factors precluded us from including them in the conclusions of this study.

We recommend further data mining of the NSDUH dataset in order to refine our main conclusion, which is that early drug use and the use of certain drug categories help explain heroin use. Such data mining studies can also be used to guide field work, saving researchers valuable time by better predicting potential areas of significance.

# Appendix A: CART Visualization, Performance, and Output



Call:

rpart(formula = HERFLAG ~ ., data = dfV1.train, parms = list(split = "information"))

n= 11054

CP nsplit rel error xerror xstd

1 0.03590426 0 1.0000000 1.0000000 0.07230964

2 0.01773050 4 0.8563830 0.9361702 0.07000246

3 0.01595745 7 0.8031915 1.0106383 0.07268656

4 0.01196809 9 0.7712766 0.9893617 0.07193060

5 0.01063830 14 0.6914894 1.0425532 0.07380493

6 0.01000000 17 0.6542553 1.0425532 0.07380493

Variable importance

COCFLAG FUCOC21 LSDFLAG CRKFLAG FUCOC18 HALFLAG PERCTYL2 OXYCODP2 VICOLOR2 OTHANL HYDCODOP

23 14 7 7 6 6 5 5 2 2 2

OXYFLAG MORPHIN2 PCPFLAG FUCRK21 IREDUC2 FUPCP21 ALCMDAYS TXILALEV PSYYFU2 TRAMADL2 IEMYFU

2 2 1 1 1 1 1 1 1 1 1

FUCRK18 FUPCP18 BINGEHVY MRJMDAYS

1 1 1 1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* PRUNED TREE OUTPUT

Fitted party:

[1] root

| [2] COCFLAG in (0) Never used (IRCOCRC = 9): 0 (n = 9727, err = 0.2%)

| [3] COCFLAG in (1) Ever used (IRCOCRC = 1-3)

| | [4] PERCTYL2 in (0) No/Unknown (Otherwise): 0 (n = 936, err = 5.3%)

| | [5] PERCTYL2 in (1) Yes (PERCTYLX=1,3)

| | | [6] MORPHIN2 in (0) No/Unknown (Otherwise): 0 (n = 284, err = 20.1%)

| | | [7] MORPHIN2 in (1) Yes (ANLCARD and MORPHINE=1,3)

| | | | [8] TXILALEV in (0) No/Unknown (Otherwise): 0 (n = 42, err = 28.6%)

| | | | [9] TXILALEV in (1) Yes (TXEVER=1): 1 (n = 65, err = 29.2%)

Number of inner nodes: 4

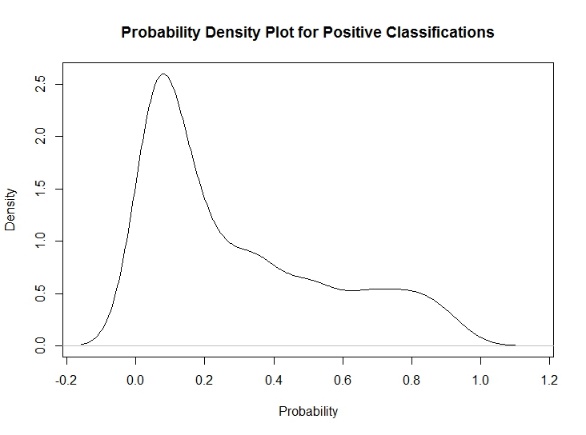
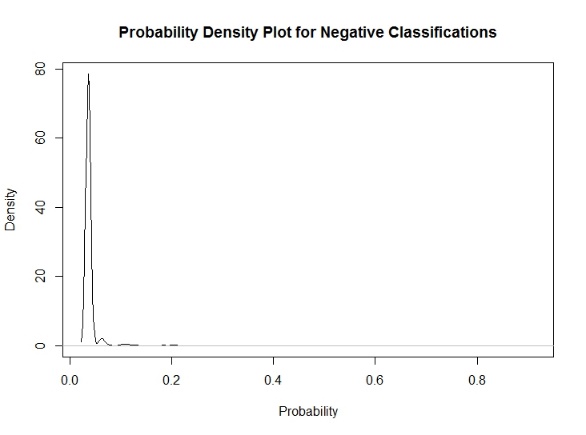
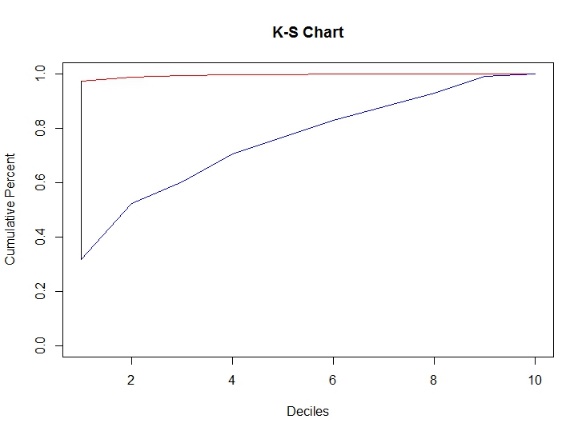
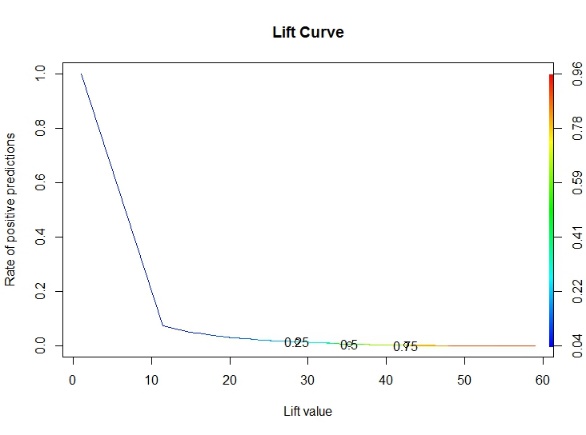
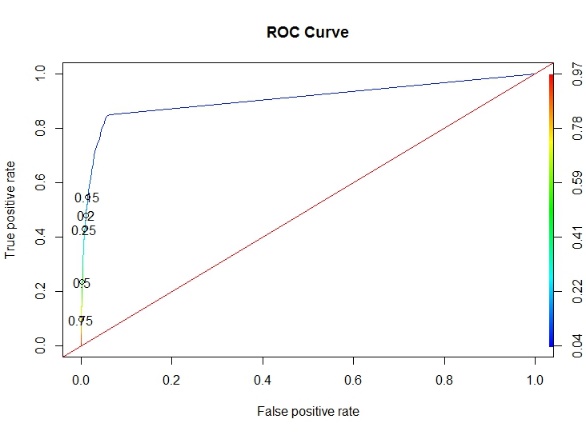
Number of terminal nodes: 5

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* BINARY EVALUATOR OUTPUT

$AUC

$AUC[[1]]

# Appendix B: AdaBoost Visualization, Performance, and Output



Call:

ada(HERFLAG ~ ., data = dfV3.train, iter = 10)

Loss: exponential Method: discrete Iteration: 10

Final Confusion Matrix for Data:

Final Prediction

True value 0 1

0 16281 13

1 165 122

Train Error: 0.011

Out-Of-Bag Error: 0.012 iteration= 10

Additional Estimates of number of iterations:

train.err1 train.kap1

9 10

Loss: exponential Method: discrete Iteration: 10

Training Results

Accuracy: 0.989 Kappa: 0.573

\*\*\*\*\*\*\*\*\*\*\*\*\*\* BINARY EVALUATOR OUTPUT

$AUC

$AUC[[1]]

[1] 0.9084735

$`D Statistic`

[1] 0.2507476

$`KS Statistic`

Group CumPct0 CumPct1 Dif

1 1 0.9730248 0.3160305 0.6569943

$`Confusion Matrix`

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 37483 552

1 298 357

Accuracy : 0.978

95% CI : (0.9765, 0.9795)

No Information Rate : 0.9765

P-Value [Acc > NIR] : 0.02389

Kappa : 0.4456

Mcnemar's Test P-Value : < 2e-16

Sensitivity : 0.9921

Specificity : 0.3927

Pos Pred Value : 0.9855

Neg Pred Value : 0.5450

Prevalence : 0.9765

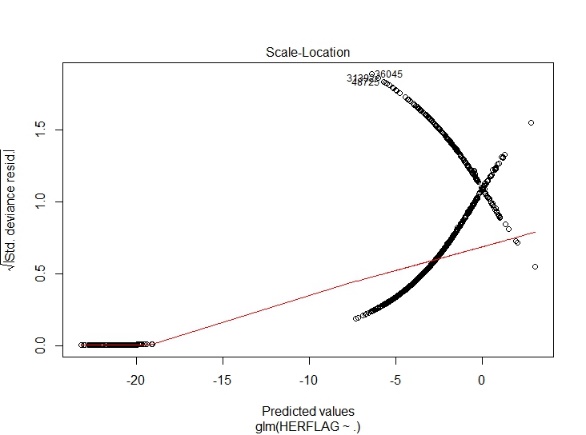
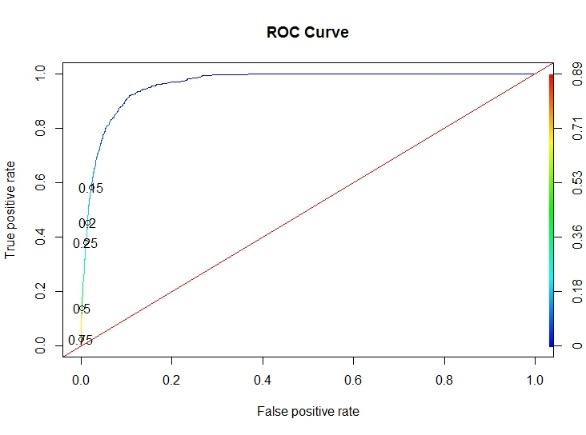
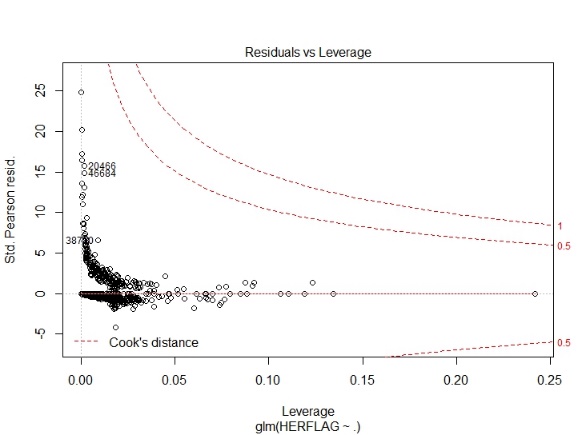
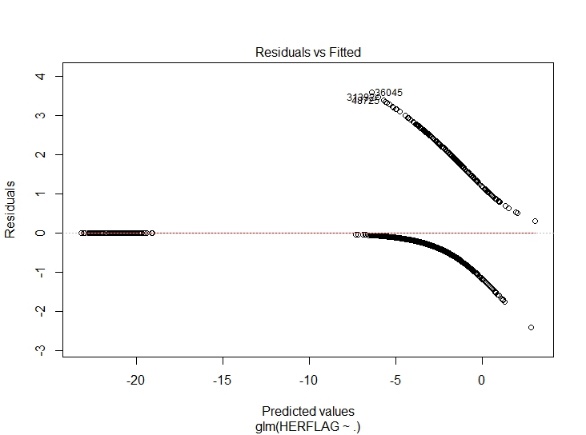
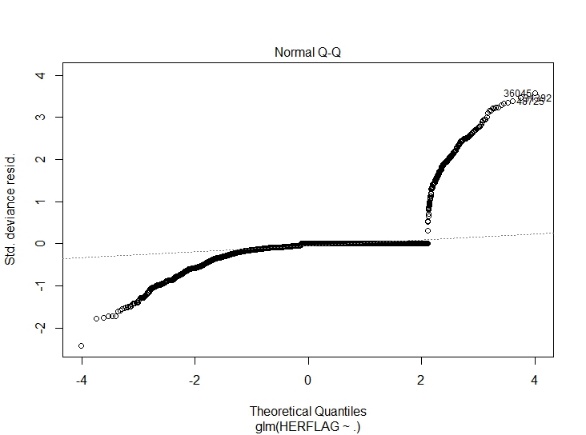
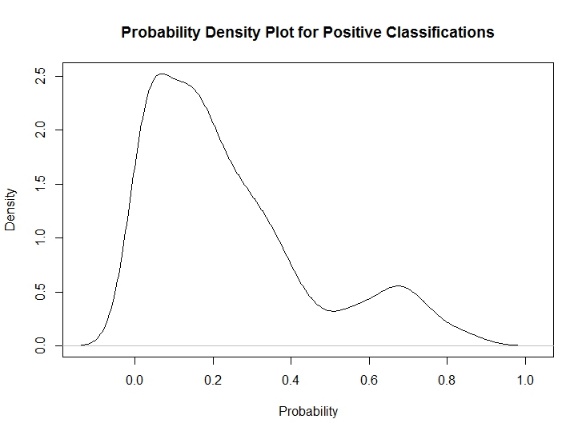
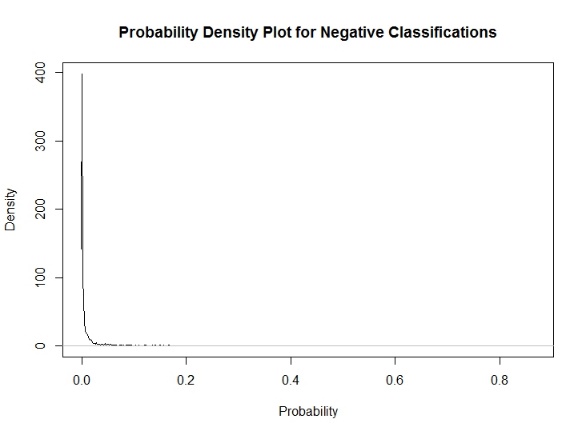
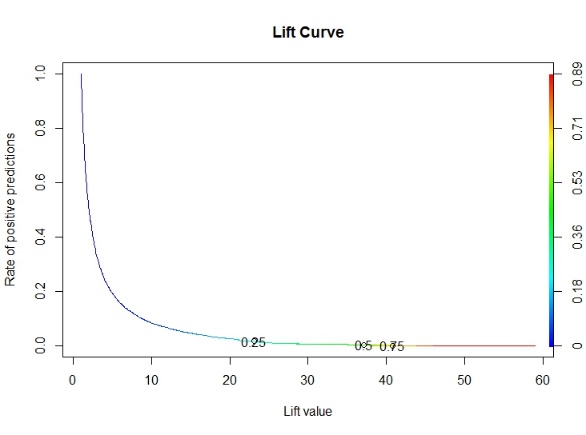
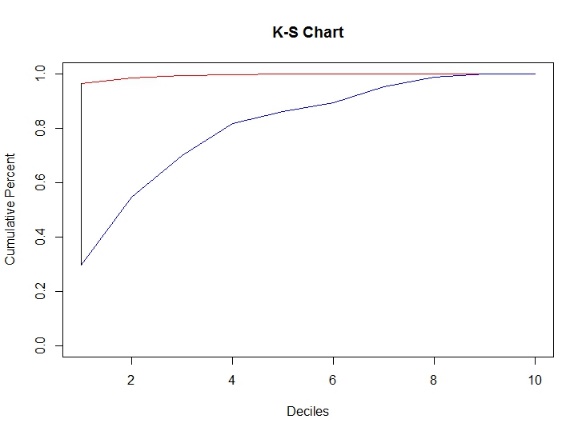
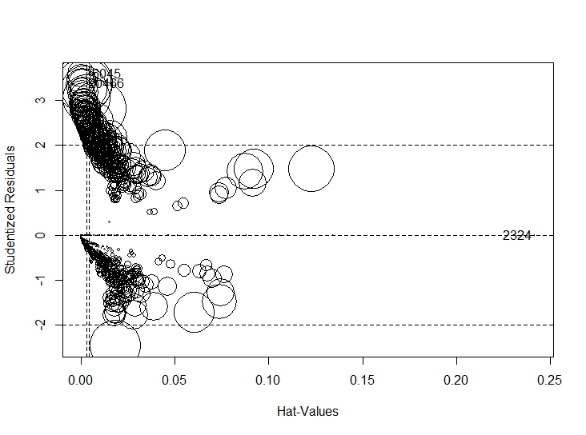
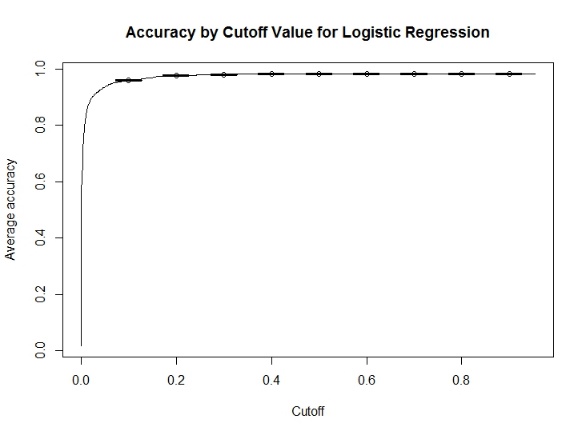
Detection Rate : 0.9688

Detection Prevalence : 0.9831

Balanced Accuracy : 0.6924

'Positive' Class : 0

# Appendix C: Logistic Regression Visualization, Performance and Output



Call:

glm(formula = HERFLAG ~ ., family = "binomial", data = dfImp.train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.3970 -0.0952 0.0000 0.0000 3.5841

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -22.6301 306.5130 -0.074 0.941145

EDUCCAT2(2) Less than high school 1.7907 0.5076 3.528 0.000419 \*\*\*

EDUCCAT2(3) High school graduate 1.5543 0.4983 3.119 0.001815 \*\*

EDUCCAT2(4) Some college 1.6278 0.4996 3.258 0.001120 \*\*

EDUCCAT2(5) College graduate 1.3788 0.5195 2.654 0.007958 \*\*

SUMAGE(2) Under 18 15.8577 306.5128 0.052 0.958739

SUMAGE(3) 18-25 15.3783 306.5129 0.050 0.959985

SUMAGE(4) 26-34 16.1577 306.5133 0.053 0.957959

SUMAGE(5) 35-49 -0.9174 2449.1966 0.000 0.999701

SUMAGE(6) 50-64 -0.8336 6095.1701 0.000 0.999891

SUMAGE(7) 65-99 -1.1358 9506.5148 0.000 0.999905

FUCRK181 0.4691 0.2083 2.252 0.024316 \*

CIGPDAY(2) Fewer than 6 0.4740 0.2997 1.582 0.113681

CIGPDAY(3) 6-15 0.6087 0.2339 2.602 0.009256 \*\*

CIGPDAY(4) 26 or More 0.9671 0.3537 2.734 0.006254 \*\*

CIGPDAY(5) Not Reported 0.8576 0.2354 3.644 0.000269 \*\*\*

CIGALCMO(2) Past Mon Use of Cig & Alc -0.4230 0.2472 -1.711 0.087116 .

CIGALCMO(3) Past Mon Use of Cig & No Alc -0.1438 0.2885 -0.499 0.618078

CIGALCMO(4) Past Mon Use of Alc & No Cig -0.5482 0.2373 -2.311 0.020859 \*

FUCOC211 1.0973 0.1573 6.974 3.07e-12 \*\*\*

TXILLALC1 1.2883 0.1875 6.870 6.41e-12 \*\*\*

LIBRIUM21 1.8509 0.5285 3.502 0.000461 \*\*\*

ANLFLAG1 1.2383 0.1623 7.630 2.36e-14 \*\*\*

HALFLAG1 1.6239 0.1930 8.414 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2897.5 on 16580 degrees of freedom

Residual deviance: 1618.3 on 16557 degrees of freedom

AIC: 1666.3

Number of Fisher Scoring iterations: 20

> cat("Odds factors for glm model are:")

Odds factors for glm model are:

> exp(coef(fit.lr))

(Intercept) EDUCCAT2(2) Less than high school

1.485554e-10 5.993592e+00

EDUCCAT2(3) High school graduate EDUCCAT2(4) Some college

4.731662e+00 5.092608e+00

EDUCCAT2(5) College graduate SUMAGE(2) Under 18

3.970014e+00 7.707478e+06

SUMAGE(3) 18-25 SUMAGE(4) 26-34

4.772186e+06 1.040390e+07

SUMAGE(5) 35-49 SUMAGE(6) 50-64

3.995530e-01 4.344995e-01

SUMAGE(7) 65-99 FUCRK181

3.211756e-01 1.598568e+00

CIGPDAY(2) Fewer than 6 CIGPDAY(3) 6-15

1.606483e+00 1.838113e+00

CIGPDAY(4) 26 or More CIGPDAY(5) Not Reported

2.630335e+00 2.357517e+00

CIGALCMO(2) Past Mon Use of Cig & Alc CIGALCMO(3) Past Mon Use of Cig & No Alc

6.550888e-01 8.660453e-01

CIGALCMO(4) Past Mon Use of Alc & No Cig FUCOC211

5.779723e-01 2.996062e+00

TXILLALC1 LIBRIUM21

3.626783e+00 6.365388e+00

ANLFLAG1 HALFLAG1

3.449904e+00 5.072753e+00

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* OBSERVATION AND VARIABLE INFLUENCE FACTORS \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

StudRes Hat CookD

2324 -4.783603e-05 0.2421087933 1.73265e-11

36045 3.596920e+00 0.0001499016 3.84079e-03

20466 3.383948e+00 0.0016449298 1.71076e-02

> vif(fit.lr)

GVIF Df GVIF^(1/(2\*Df))

EDUCCAT2 1.192565 4 1.022257

SUMAGE 1.186666 6 1.014365

FUCRK18 1.085626 1 1.041934

CIGPDAY 2.025690 4 1.092249

CIGALCMO 2.153950 3 1.136421

FUCOC21 1.277222 1 1.130142

TXILLALC 1.061605 1 1.030342

LIBRIUM2 1.017008 1 1.008468

ANLFLAG 1.090831 1 1.044429

HALFLAG 1.233110 1 1.110455

\*\*\*\*\*\*\*\*\*\*\*\*\*\* BINARY EVALUATOR OUTPUT \*\*\*\*\*\*\*\*\*\*\*

$AUC

$AUC[[1]]

[1] 0.9627096

$`D Statistic`

[1] 0.2274903

$`KS Statistic`

Group CumPct0 CumPct1 Dif

1 1 0.9643223 0.2961832 0.6681391

$`Confusion Matrix`

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 37430 605

1 324 331

Accuracy : 0.976

95% CI : (0.9744, 0.9775)

No Information Rate : 0.9758

P-Value [Acc > NIR] : 0.4168

Kappa : 0.4042

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.9914

Specificity : 0.3536

Pos Pred Value : 0.9841

Neg Pred Value : 0.5053

Prevalence : 0.9758

Detection Rate : 0.9674

Detection Prevalence : 0.9831

Balanced Accuracy : 0.6725

'Positive' Class : 0

# References

Fergusson, D. M., Boden, J. M., & Horwood, L. J. (July 2008). The developmental antecedents of illicit drug use: Evidence from a 25-year longitudinal study. *Drug and Alcohol Dependence*, 165-177.

Hartel, D. M., Schoenbaum, E. E., Selwyn, P. A., Kline, J., Davenny, K., Klein, R. S., & Friedland, G. H. (January 1995). Heroin use during methadone maintenance treatment: the importance of methadone dose and cocaine use. *American Journal of Public Health: Vol. 85, No. 1,*, 83-88.

Lynskey, M. T., Andrew C. Heath, A. C., Bucholz, K. K., & al, e. (January 2003). Escalation of Drug Use in Early-Onset Cannabis Users vs Co-twin Controls. *JAMA*, 427-433.

National Institute on Drug Abuse. (2015, December). *Overdose Death Rates*. Retrieved from https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates

Park, H. a. (2016, January 19). *How the Epidemic of Drug Overdose Deaths Ripple Across America*. Retrieved from The New York Times: http://www.nytimes.com/interactive/2016/01/07/us/drug-overdose-deaths-in-the-us.html?\_r=0

Rossen LM, B. B. (2016, January 19). *Drug poisoning mortality: United States, 1999–2014*. Retrieved from National Center for Health Statistics Data Visualization Gallery: http://blogs.cdc.gov/nchs-data-visualization/drug-poisoning-mortality/

SAMHSA. (2014). *National Survey on Drug Use and Health, 2014 Codebook.* Ann Arbor, Michigan: Inter-University Consortium for Political and Science Research.

1. Throughout this discussion, we will assume that the respondents’ answers were honest and accurate. [↑](#footnote-ref-1)
2. We made use of binaryEvaluator(), an R script we created in a previous analysis. [↑](#footnote-ref-2)
3. While not statistically significant, the coefficients in the model for age of first illicit drug use were extremely large. This suggests a need for further analysis to definitively discount or show linkage between early drug use and heroin usage. [↑](#footnote-ref-3)