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, by early usage of drugs, and by education level. 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, frequent use of cocaine made respondents 19 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).

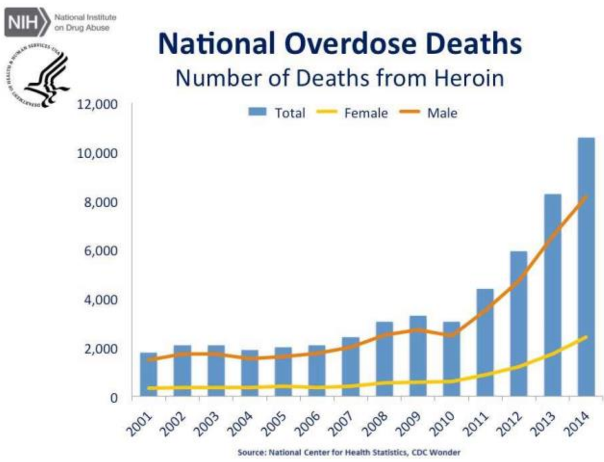


Figure 1: Heroin Overdose Deaths

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 focused on the variable HEREVER, which represents a yes or no answer to the question: "Have you ever used heroin?" We also examined similar variables corresponding to other drugs. As we can see in Fig. 2[1], 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.

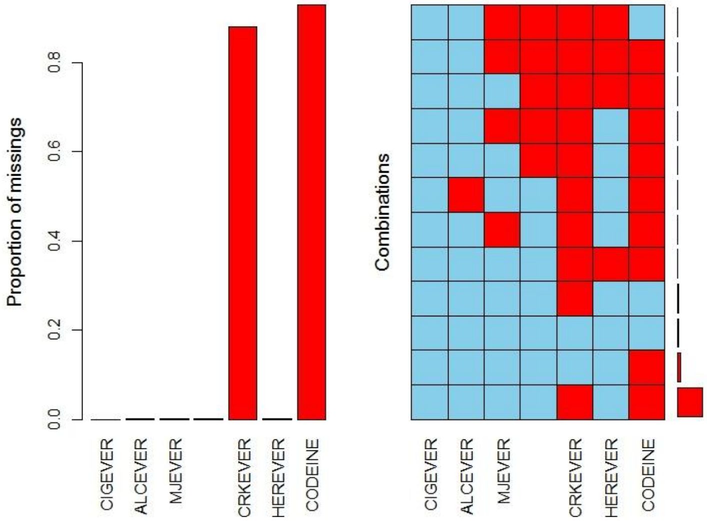


Figure 2: Example of Missing Raw Data in 2014 NSDUH

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

[1] In the "Combinations" chart in Figure 2, each row in the grid represents a combination of missing variables, and the bar chart in the right margin represents the proportion of observations with missing variables represented by that combination. The last row is the set of observations with missing data in both CRKEVER and CODEINE. That combination is the most common among observations with missing data.

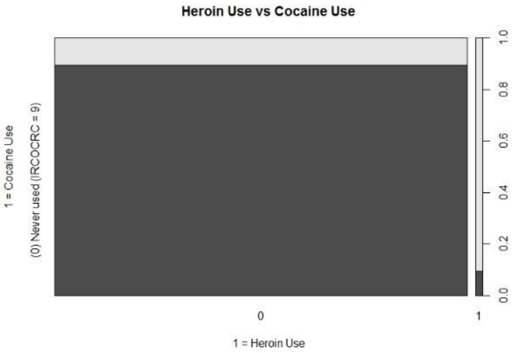


Figure 7: Cocaine Use

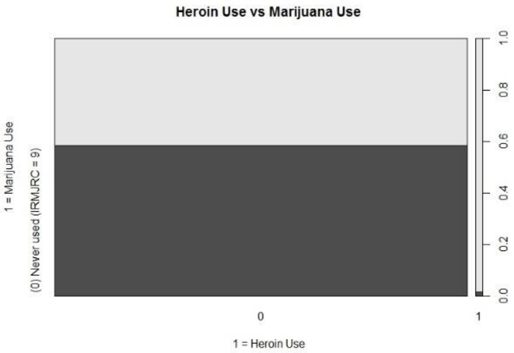


Figure 6: Marijuana Use

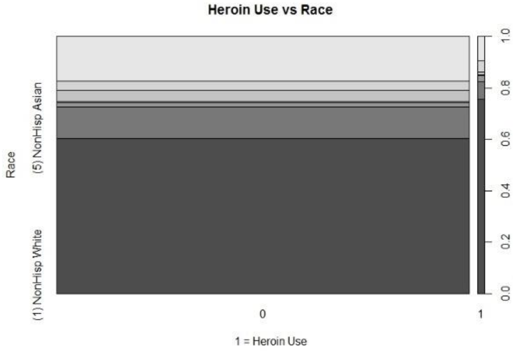


Figure 5: Distribution of Race

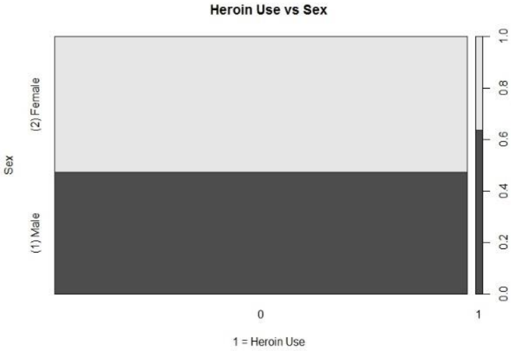


Figure 4: Distribution of Sex

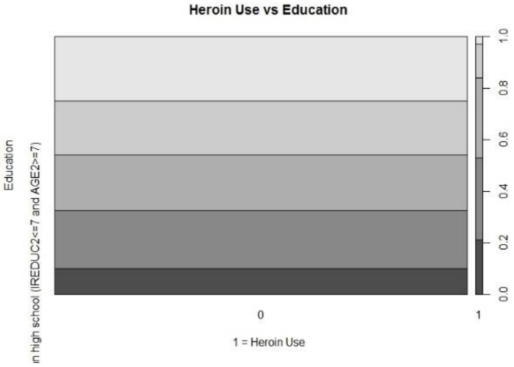


Figure 3: Distribution of Education

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 analyses 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[2]. 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.

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

[2] Throughout this discussion, we will assume that the respondents' answers were honest and accurate.

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)[3].

From these visualizations we can see some potential predictors of heroin use. It 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

[3] Figs. 3-7 have two bar charts – the left hand bar is the distribution for non-heroin users, and the right hand bar is for heroin users. The width of the bars is proportional to the respondents between the two categories.

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

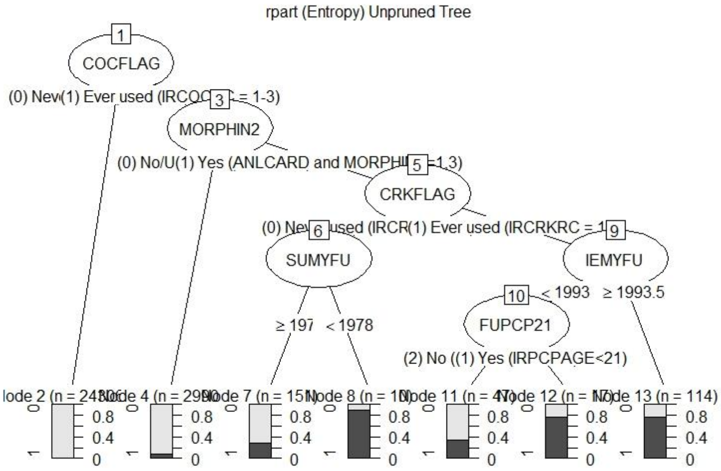


Figure 8: Unpruned CART Decision Tree

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 50% of the total observations for CART.

This training set was designed to be large enough to capture many of the small number of heroin respondents.

CART Decision Tree Analysis

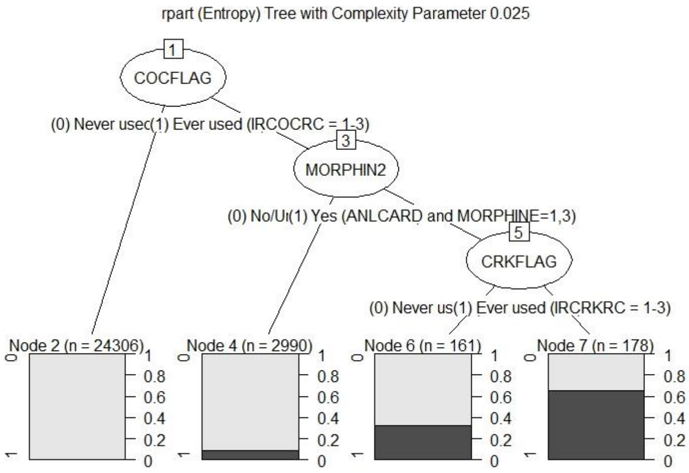

rpart (Entropy) Tree with Complexity Parameter 0.025
1
COCFLAG
(0) Never usec(1) Ever used (IRCOCRC = 1-3)
3
MORPHIN2
(0) No/Ur(1) Yes (ANLCARD and MORPHINE=1,3)
5
CRKFLAG
(0) Never us(1) Ever used (IRCRKRC = 1-3)
Node 2 (n = 24306)
Node 4 (n = 2990)
Node 6 (n = 161)
Node 7 (n = 178)
0
1
1
0.8
0.6
0.4
0.2
0

Figure 9: Pruned CART Decision Tree

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 tree that included a variable of little use: the year in which drug use first began (Fig. 8). This variable is dependent on age, so we should trim it out. 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 0.025. When we pruned our tree using this parameter, we obtained a much cleaner, interpretable tree (Fig. 9).

In this tree, we see three 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 morphine. The final split was based upon whether or not the respondent had ever used crack. We can thus surmise that if a respondent had abused cocaine, morphine, and crack, they were likely to have abused heroin as well.

To check the performance of the model, we ran it on the test dataset and ran a set of performance diagnostics[4]. 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 0.924 and kappa statistic of 0.411. 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

[4] We made use of binaryEvaluator(), an R script we created in a previous analysis.

Model Performance Comparison

Model	AUC	Optimal Selection Probability	Test Accuracy	Kappa Statistic
CART (pruned)	0.924	0.300	0.982	0.411
C5.0	0.918	0.350	0.974	0.407
AdaBoost	0.922	0.190	0.981	0.487
Logistic Regression	0.960	0.140	0.970	0.381

Figure 10: Model Performance

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.

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 0.922 and a kappa of 0.487 (Fig. 10).

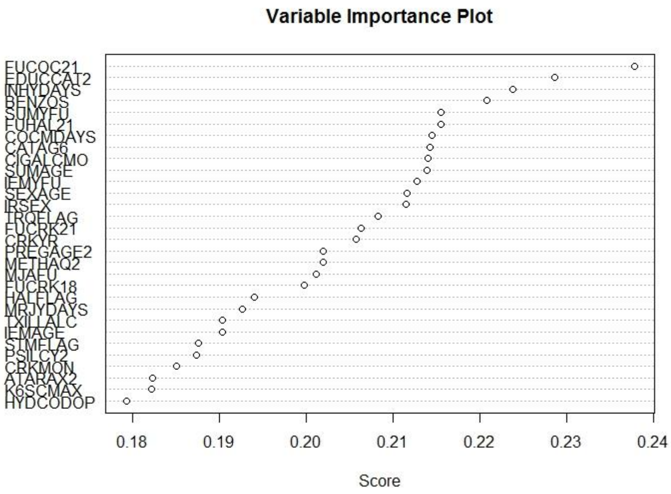


Figure 11: AdaBoost Variable Importance

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 differed from CART in the variables identified 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 regarding early drug use – whether the respondent first used crack under the age of 21, and the age of first illicit drug use. Two of the important variables are of little use to us. The calendar 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 most important variables (those with importance scores over 0.21) as predictors. 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 0.960 and a kappa statistic of 0.381. However, this model allows us to better understand the impact of individual variables on the likelihood of heroin use.

Logistic Regression Variable Summary

Variable	Coefficient	Independent Probability	Likelihood Muliplier
Intercept (base includes "Never Used")	-22.284	2.100E-10	1.000
Past Month Use of Cocaine in Days (>19)	2.941	3.976E-09	18.935
Past Month Use of Cocaine in Days (6-19)	1.452	8.970E-10	4.272
Education: Less Than High School	1.089	6.239E-10	2.971
Ever Used Tranquilizers	1.082	6.196E-10	2.951
First Used Cocaine Under 21	1.077	6.165E-10	2.936
Ever Used CPN/Methamphetamine	0.974	5.561E-10	2.649
Past Month Use of Cocaine in Days (3-5)	0.942	5.386E-10	2.565
First Used Inhalants Under 21	0.934	5.343E-10	2.545
Education: High School Graduate	0.874	5.032E-10	2.396
Education: Some College	0.847	4.898E-10	2.333
Past Month Use of Cocaine in Days (1-2)	0.777	4.567E-10	2.175
Past Month Use of Cigs and No Alcohol	0.741	4.405E-10	2.098
Male	0.450	3.293E-10	1.568
Past Month Use of Alcohol and No Cigs	-0.576	1.180E-10	0.562

Figure 12: Logistic Regression Coefficients

The variables with a high degree of statistical significance in the regression model are shown in Figure 12.[5] 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: $\log(\frac{\pi}{1+\pi}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ where π represents the probability of heroin use and β_i represents the regression coefficient corresponding to variable X_i. 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 cocaine, tranquilizers, methamphetamine, or inhalants before 21, and did not use cigarettes

[5] 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.

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 must include the coefficient for both the intercept and the variable of interest, X_i:

$$\pi_i = \frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}}$$

For example, the probability of heroin use for the base case is 2.100e-10. When we consider the case where the respondent has used cocaine more than 19 days in the past month, the probability increases to 3.976e-9. This value remains extremely low, but its 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 X_i as $L_i = \pi_i/\pi_b$ where π_i is the probability when the factor is positive and π_b is the probability of the base case. So, for the “19+ days of cocaine” case, the likelihood multiplier is 18.935. We interpret this result by saying that respondents who have used cocaine for more than 19 days in the past month are 18.935 times more likely to use heroin than the base case.

The base case education level is “user is under the age of 18”, meaning that they should still be in school. The survey did not identify underage high school drop outs. When we consider the effect of education level on heroin usage, we are focusing on respondents who are older than 18. We thus 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 cocaine, tranquilizers, and methamphetamine, also increase potential heroin use. And again, age of first drug use is a factor. For example, respondents who used cocaine before the age of 21 were about three times more likely to use heroin than the those in 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 did not graduate from high school and first used cocaine before 21 are three times more likely than those who first used cocaine earlier than 21 but did graduate from high school. 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).

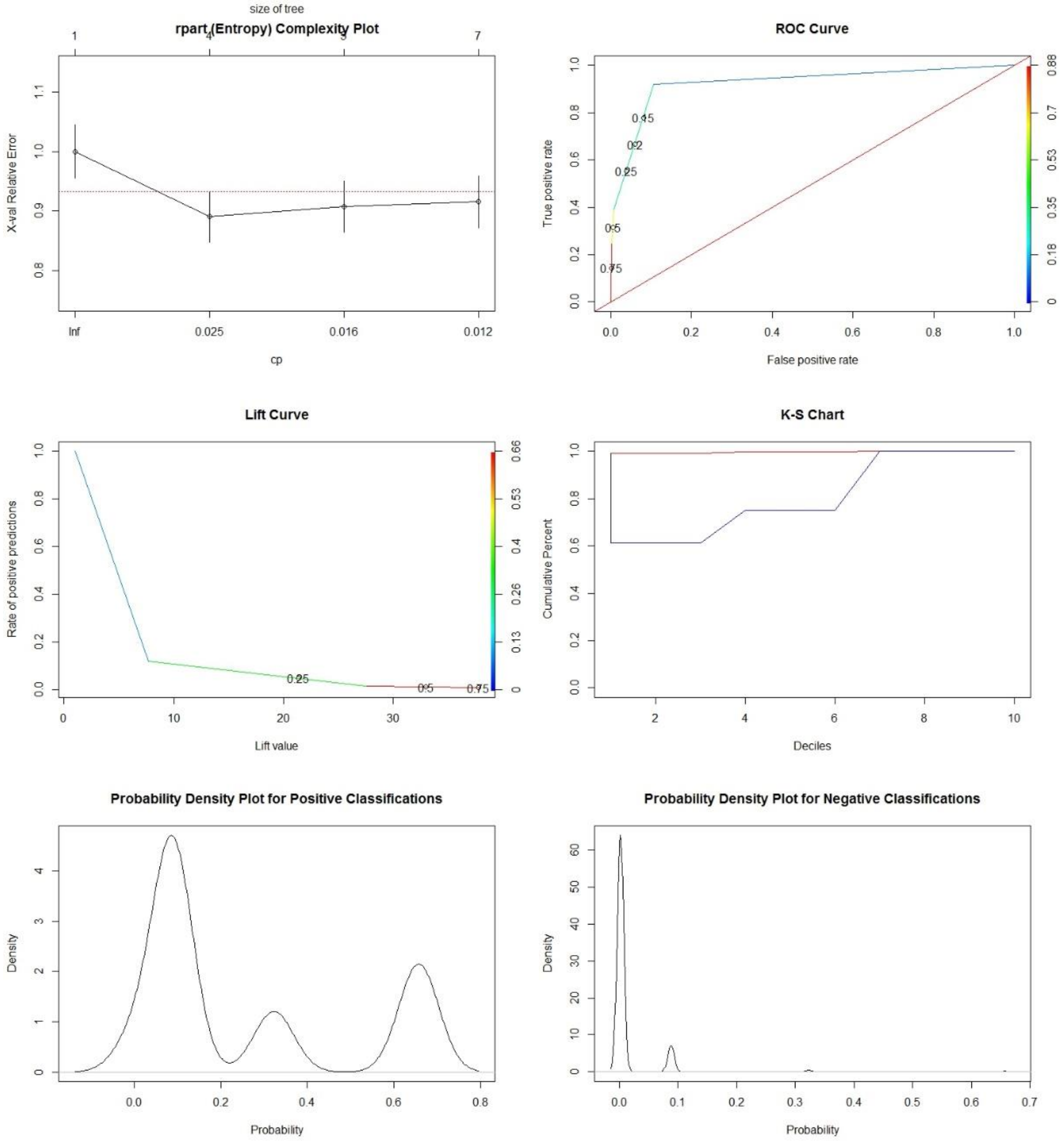
Conclusion

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 when 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 addressed 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, morphine, and crack correlated to heroin usage, as did drug abuse at an early age. When using a boosted decision tree algorithm, AdaBoost, we found that the most important variables weren't just drug use factors, early age drug use, frequency of use, and education level. Finally, our logistic regression analysis confirmed that use of other drugs, early and frequent use of cocaine, 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 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.

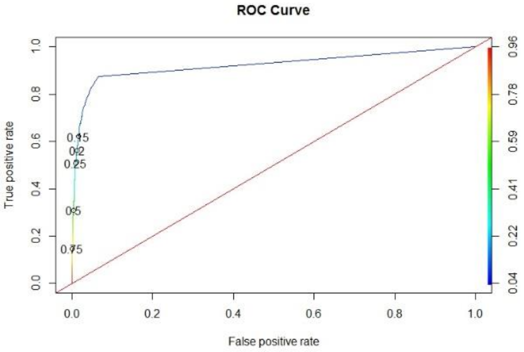
Appendix A: CART Visualization, Performance, and Output

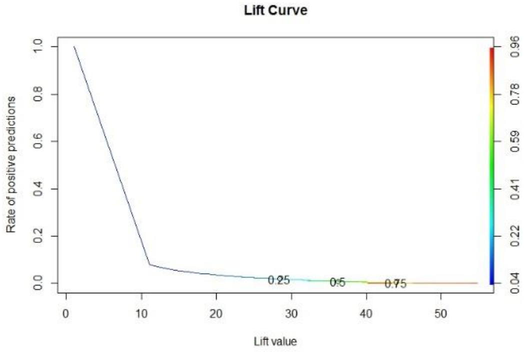


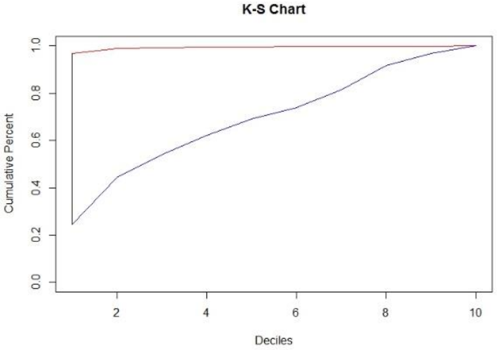
```
Call:
rpart(formula = HERFLAG ~ ., data = dfV1.train, parms = list(split = "information"))
  n= 27635

          CP nsplit rel error    xerror       xstd
1 0.03856749      0 1.0000000 1.0000000 0.04505474
2 0.01652893      3 0.8842975 0.8904959 0.04255787
3 0.01549587      4 0.8677686 0.9070248 0.04294471
4 0.01000000      6 0.8367769 0.9152893 0.04313674

Variable importance
 COCFLAG  FUCOC21  LSDFLAG  CRKFLAG  FUCOC18  PSILCY2 MORPHIN2 METHDON2  DILAUD2   SUMYFU 
DEMEROL2 
      32       19        9        9        8        8        6        1        1        1 
1 
  IEMYFU  PSYYFU2  ULTRAM2 
       1        1        1 


            Overall
CATAG6     6.689699
COCFLAG  785.233068
CRKFLAG  832.193013
FUANL21    7.269476
FUCOC21  620.192013
FUCRK21   14.659219
FUECS21    8.014528
FUPCP21   24.434407
HALFLAG  644.599428
IEMYFU    21.090518
MORPHIN2 159.007936
OTHANL   142.551278
OTHSED     4.486270
OXYCODP2 785.396181
PCPFLAG   18.618843
PERCTYL2 156.342279
PSYYFU2   11.227137
SUMYFU    15.183695
TXILALEV  19.986625


**************** PRUNED TREE OUTPUT
Fitted party:
[1] root
|   [2] COCFLAG in (0) Never used (IRCOCRC = 9): 0 (n = 24306, err = 0.2%)
|   [3] COCFLAG in (1) Ever used (IRCOCRC = 1-3)
|   |   [4] MORPHIN2 in (0) No/Unknown (Otherwise): 0 (n = 2990, err = 8.8%)
|   |   [5] MORPHIN2 in (1) Yes (ANLCARD and MORPHINE=1,3)
|   |   |   [6] CRKFLAG in (0) Never used (IRCRKRC = 9): 0 (n = 161, err = 32.3%)
|   |   |   [7] CRKFLAG in (1) Ever used (IRCRKRC = 1-3): 1 (n = 178, err = 34.3%)

Number of inner nodes:    3
Number of terminal nodes: 4

***************** BINARY EVALUATOR OUTPUT (Probability = 0.30)
$AUC
$AUC[[1]]
[1] 0.924047


$`D Statistic`
[1] 0.2418382

$`KS Statistic`
  Group   CumPct0   CumPct1       Dif
1     1 0.9921996 0.6113537 0.3808459
2     2 0.9921996 0.6113537 0.3808459
```

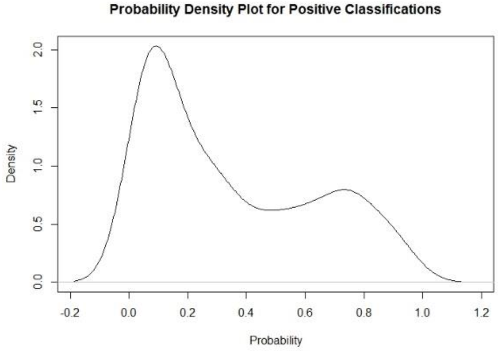
```
3     3 0.9921996 0.6113537 0.3808459

$`Confusion Matrix`
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 26966   212
         1   280   178

               Accuracy : 0.9822
                 95% CI : (0.9806, 0.9837)
    No Information Rate : 0.9859
    P-Value [Acc > NIR] : 1.000000

                  Kappa : 0.4108
 Mcnemar's Test P-Value : 0.002523

            Sensitivity : 0.9897
            Specificity : 0.4564
         Pos Pred Value : 0.9922
         Neg Pred Value : 0.3886
             Prevalence : 0.9859
         Detection Rate : 0.9758
   Detection Prevalence : 0.9834
      Balanced Accuracy : 0.7231

       'Positive' Class : 0
```

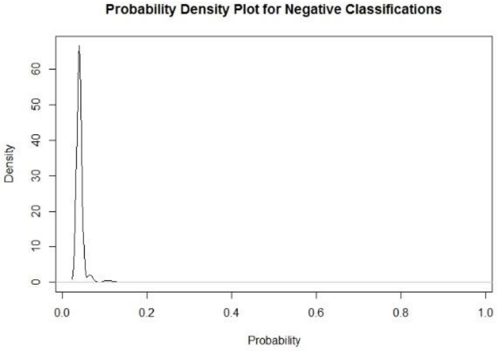
Appendix B: AdaBoost Visualization, Performance, and Output







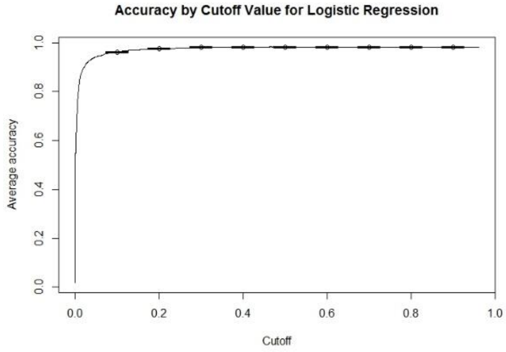


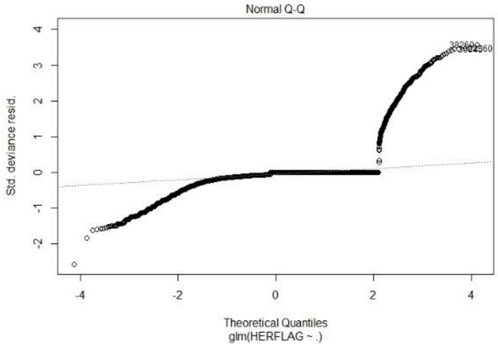


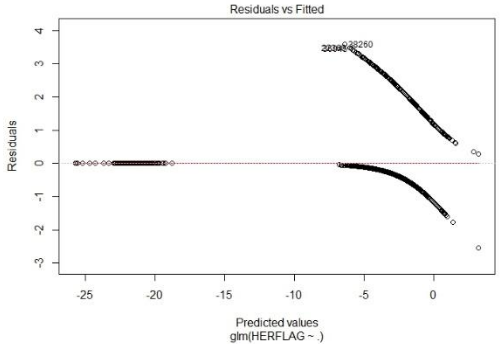
```
Call:
ada(HERFLAG ~ ., data = dfV3.train, iter = 10)

Loss: exponential Method: discrete   Iteration: 10

Final Confusion Matrix for Data:
          Final Prediction
True value     0     1
         0 27111    40
         1   270   214

Train Error: 0.011

Out-Of-Bag Error:  0.013  iteration= 10

Additional Estimates of number of iterations:

train.err1 train.kap1
         9          9

> summary(fit.dfV3)
Call:
ada(HERFLAG ~ ., data = dfV3.train, iter = 10)

Loss: exponential Method: discrete   Iteration: 10

Training Results

Accuracy: 0.989 Kappa: 0.575


************** BINARY EVALUATOR OUTPUT (PROB 0.19)
$AUC
$AUC[[1]]
[1] 0.9222888


$`D Statistic`
[1] 0.3015609

$`KS Statistic`
  Group   CumPct0   CumPct1       Dif
1     1 0.9685039 0.2445415 0.7239625

$`Confusion Matrix`
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 26840   338
         1   195   263

               Accuracy : 0.9807
                 95% CI : (0.979, 0.9823)
    No Information Rate : 0.9783
    P-Value [Acc > NIR] : 0.002323

                  Kappa : 0.487
 Mcnemar's Test P-Value : 7.714e-10

            Sensitivity : 0.9928
            Specificity : 0.4376
         Pos Pred Value : 0.9876
         Neg Pred Value : 0.5742
             Prevalence : 0.9783
         Detection Rate : 0.9712
   Detection Prevalence : 0.9834
```

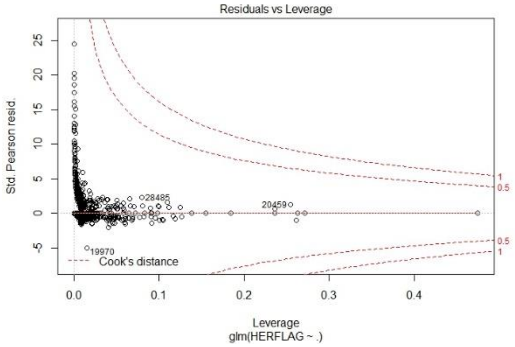
```
Balanced Accuracy : 0.7152

 'Positive' Class : 0
```

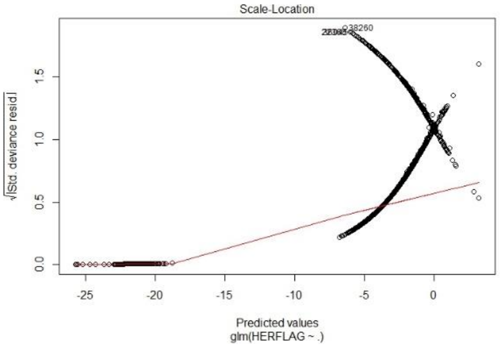
Appendix C: Logistic Reg Visualization, Performance and Output

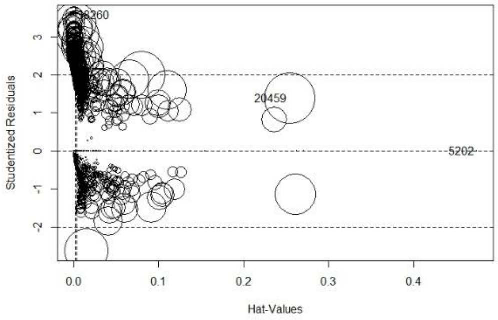


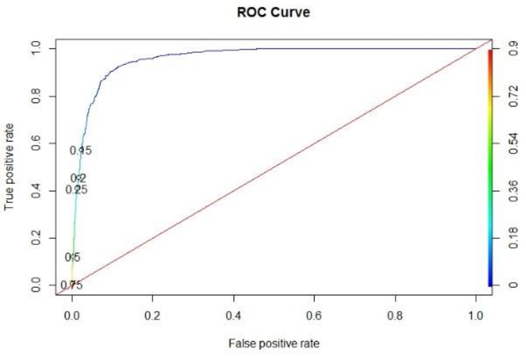


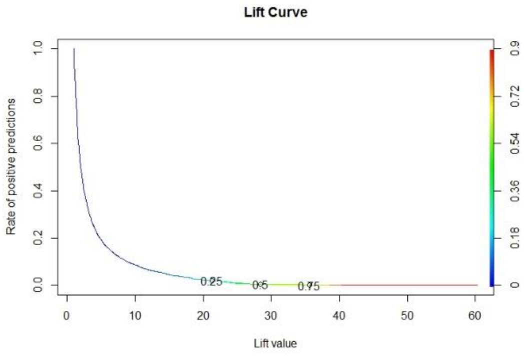


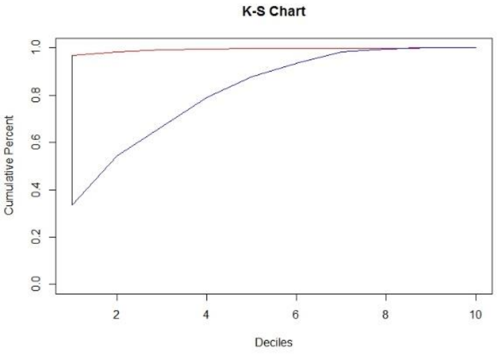


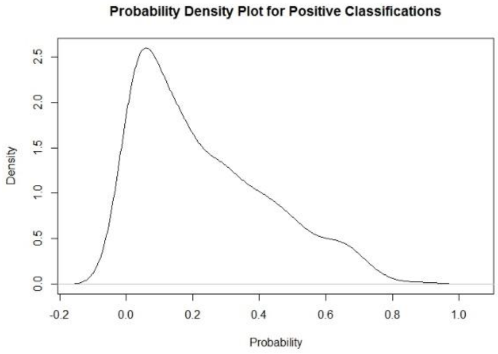


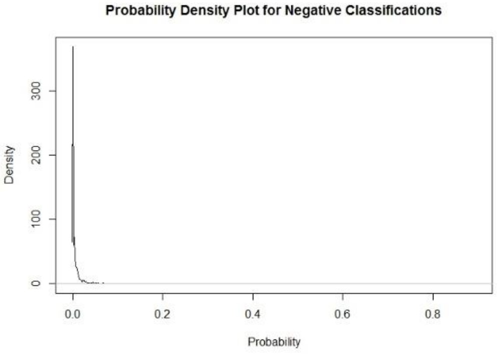



ROC Curve


Lift Curve


K-S Chart


Probability Density Plot for Positive Classifications


Probability Density Plot for Negative Classifications

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5556  -0.1057   0.0000   0.0000   3.5762  

Coefficients:
                                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)                               -22.28416  240.16493  -0.093 0.926073    
EDUCCAT2(2) Less than high school           1.08917    0.31559   3.451 0.000558 ***
EDUCCAT2(3) High school graduate            0.87361    0.30776   2.839 0.004532 ** 
EDUCCAT2(4) Some college                    0.84702    0.30946   2.737 0.006199 ** 
EDUCCAT2(5) College graduate                0.53821    0.32856   1.638 0.101403    
SUMAGE(2) Under 18                         16.27090  240.16486   0.068 0.945986    
SUMAGE(3) 18-25                            15.92886  240.16489   0.066 0.947119    
SUMAGE(4) 26-34                            16.13702  240.16553   0.067 0.946429    
SUMAGE(5) 35-49                            -0.58362 1939.18815   0.000 0.999760    
SUMAGE(6) 50-64                            -0.65557 4722.72596   0.000 0.999889    
SUMAGE(7) 65-99                            -0.83575 8890.08872   0.000 0.999925    
FUCRK181                                    0.15749    0.17574   0.896 0.370165    
CIGALCMO(2) Past Mon Use of Cig & Alc       0.07555    0.16527   0.457 0.647558    
CIGALCMO(3) Past Mon Use of Cig & No Alc    0.74149    0.18497   4.009 6.10e-05 ***
CIGALCMO(4) Past Mon Use of Alc & No Cig   -0.57586    0.18789  -3.065 0.002177 ** 
FUCOC211                                    1.07679    0.12869   8.367  < 2e-16 ***
CPNMTHFG1                                   0.97350    0.13078   7.444 9.80e-14 ***
TRQFLAG1                                    1.08195    0.40854   2.648 0.008089 ** 
FUHAL211                                    0.93445    0.13634   6.854 7.19e-12 ***
BENZOS1                                     0.46050    0.40314   1.142 0.253332    
IRSEX1                                      0.44958    0.11052   4.068 4.74e-05 ***
COCMDAYS(2) 1-2                             0.77717    0.28075   2.768 0.005636 ** 
COCMDAYS(3) 3-5                             0.94176    0.43504   2.165 0.030404 *  
COCMDAYS(4) 6-19                            1.45229    0.49045   2.961 0.003065 ** 
COCMDAYS(5) More than 19                    2.94083    0.61460   4.785 1.71e-06 ***
INHYDAYS(2) 1-11                           -0.03070    0.37651  -0.082 0.935009    
INHYDAYS(3) 12-49                           0.09278    0.68619   0.135 0.892448    
INHYDAYS(4) 50-99                         -19.63138 5043.89978  -0.004 0.996895    
INHYDAYS(5) More than 99                   -0.20586    1.00735  -0.204 0.838077    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4874.8  on 27634  degrees of freedom
Residual deviance: 2698.0  on 27606  degrees of freedom
AIC: 2756

Number of Fisher Scoring iterations: 20

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

Odds factors for glm model are:
> exp(coef(fit.lr))
                             (Intercept)        EDUCCAT2(2) Less than high school 
                            2.099473e-10                             2.971815e+00 
        EDUCCAT2(3) High school graduate                 EDUCCAT2(4) Some college 
                            2.395546e+00                             2.332697e+00 
            EDUCCAT2(5) College graduate                       SUMAGE(2) Under 18 
                            1.712936e+00                             1.165100e+07 
                         SUMAGE(3) 18-25                          SUMAGE(4) 26-34 
                            8.275936e+06                             1.019105e+07 
                         SUMAGE(5) 35-49                          SUMAGE(6) 50-64 
                            5.578749e-01                             5.191456e-01 
                         SUMAGE(7) 65-99                                 FUCRK181 
                            4.335498e-01                             1.170573e+00 
   CIGALCMO(2) Past Mon Use of Cig & Alc CIGALCMO(3) Past Mon Use of Cig & No Alc 
                            1.078482e+00                             2.099059e+00 
CIGALCMO(4) Past Mon Use of Alc & No Cig                                 FUCOC211 
                            5.622190e-01                             2.935229e+00 
                               CPNMTHFG1                                 TRQFLAG1 
                            2.647191e+00                             2.950439e+00 
                                FUHAL211                                  BENZOS1 
                            2.545800e+00                             1.584871e+00 
                                  IRSEX1                          COCMDAYS(2) 1-2 
                            1.567660e+00                             2.175308e+00 
                         COCMDAYS(3) 3-5                         COCMDAYS(4) 6-19 
                            2.564486e+00                             4.272908e+00 
                COCMDAYS(5) More than 19                         INHYDAYS(2) 1-11 
                            1.893146e+01                             9.697644e-01 
                       INHYDAYS(3) 12-49                        INHYDAYS(4) 50-99 
                            1.097218e+00                             2.979888e-09 
                INHYDAYS(5) More than 99 
                            8.139501e-01 


***************** OBSERVATION AND VARIABLE INFLUENCE FACTORS ******************
            StudRes          Hat       CookD
20459  1.3848768575 2.546507e-01 1.81032e-02
38260  3.5838977344 9.213292e-05 1.89899e-03
5202  -0.0001412278 4.749065e-01 4.07866e-10
> vif(fit.lr)
              GVIF Df GVIF^(1/(2*Df))
EDUCCAT2  1.191555  4        1.022149
SUMAGE    1.212210  6        1.016166
FUCRK18   1.388649  1        1.178409
CIGALCMO  1.187239  3        1.029018
FUCOC21   1.397727  1        1.182255
CPNMTHFG  1.461947  1        1.209110
TRQFLAG  13.234482  1        3.637923
```

```
FUHAL21   1.397185  1        1.182026
BENZOS   13.208718  1        3.634380
IRSEX     1.031408  1        1.015582
COCMDAYS  1.168726  4        1.019680
INHYDAYS  1.113510  4        1.013530


************** BINARY EVALUATOR OUTPUT ***********
> binaryEvaluator(predict.lr$y, predict.lr$fitted.values, .14)
$AUC
$AUC[[1]]
[1] 0.9604477


$`D Statistic`
[1] 0.2204424

$`KS Statistic`
  Group  CumPct0   CumPct1       Dif
1     1 0.966885 0.3340611 0.6328238

$`Confusion Matrix`
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 26538   640
         1   188   270

               Accuracy : 0.97
                 95% CI : (0.968, 0.972)
    No Information Rate : 0.9671
    P-Value [Acc > NIR] : 0.002692

                  Kappa : 0.3811
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9930
            Specificity : 0.2967
         Pos Pred Value : 0.9765
         Neg Pred Value : 0.5895
             Prevalence : 0.9671
         Detection Rate : 0.9603
   Detection Prevalence : 0.9834
      Balanced Accuracy : 0.6448

       '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.