An algorithm to predict the how often drugs will be use within nine months of parole from previous drug related factors.

Final report prepared for the Data Analysis and Interpretation Specialization.

September 2016

Introduction to the research question.

This project aims to find out whether there is a relationship between those whom partake in certain drug related activities prior to being incarcerated and how many drug related activities they will participate in during their 9 month parole period.

The reason for investigating parolee's drug related activities is to see whether it is possible to predict early if certain parolees are likely to partake heavily in drug related activities after their release from incarceration and to hopefully be able to work with those whom are more likely to be heavily influenced.

Methods.

Sample.

The capstone project takes its data from the Step'n'Out study using a randomized selection of 450 drug-involved parolees, N= 450.
Follow-up at 3-and 9-months of their parole will assess primary outcomes of rearrests, crime and drug use. If collaborative behavioral management is effective, its wider adoption could improve the outcomes of community reentry of drug-involved ex-offenders.

A sample will be taken from the overall dataset by sub setting out candidates that committed illegal drugs, possession of drugs, drug manufacturing and drug sale in the 30 days to 6 months prior to being incarcerated.

Since the data required contains miscellaneous and incorrect data, such as column C3CM6M (or Number of times a parolee committed illegal drugs in the past 6 months) has cells indicating a parolee has committed illegals drugs 999 times which is obviously incorrect. It has been decided that these unrealistic values or any NaN values will be replaced by a zero value. This allows the data to be cleaner and also increases the number of values that can be used in my analysis.

Measures.

Below is a list of the variables used. They are all drug related activities committed either 30 days or 6 months prior to being arrested.

C3CM30	# Times Committed Illegal Drugs 30 Days Prior to arrest
C3CM6M	# Times Committed Illegal Drugs 6 Mths Prior to arrest
C5CM30	# Times Committed Possession of Drugs 30 Days Prior to arrest
C5CM6M	# Times Committed Possession of Drug Past 6 Mths Prior to arrest
C6CM30	# Times Committed Drug Manufacture 30 Days Prior to arrest
C6CM6M	# Times Committed Drug Manufacture 6 Mths Prior to arrest
C7CM30	# Times Committed Drug Sale 30 Days Prior to arrest
C7CM6M	# Times Committed Drug Sale 6 Mths Prior to arrest

Below is the response/target variable.

Dataset Variable Name	**Dataset Variable Description**
ANYDRUGS	Any Illegal Drug Use (exc tobacco and alcohol) during parole

All the variables are quantitative and any NaN, miscellaneous or incorrect data was replaced with a zero value to indicate that nothing was committed over that variable period.

Analysis.

To begin with, a summary of the data will be made. This will be used to quickly validate the data.

A bivariate analysis of all the individual predictors vs the response variable will be plotted to visually check for any noticeably patterns. Alongside this, the associated pearson coefficient and r-squared values will also be calculated to further analyse for correlation.

Lasso regression was chosen to identify the best predictors to use for the algorithm. The lasso regression will be trained from a training set made from 60% of the sample and tested on remaining 40% of the data. All variables will be standardized with a mean of 0 and standard deviation of 1. Cross validation will be done using k-fold cross validation using 10 folds. The change in the cross validation mean squared error rate at each step is used to identify the best subset of predictor variables. Predictive accuracy is assessed by

determining the mean squared error rate of the training data prediction algorithm when applied to observations in the test data set.

Results.

Descriptive statistics for the predictive variables used.

	COMMITTED_DRUGS_PAST_30_DAYS	COMMITTED_DRUGS_PAST_6_MTHS	DRUG_POSSESION_PAST_30_DAYS
count	441	441	441
mean	33.968254	264.231293	26.596372
std	32.874954	336.033692	31.289708
min	0	0	0
25%	3	30	0
50%	30	180	30
75%	30	180	30
max	99	999	99

	DRUG_POSSESION_PAST_6_MTHS	MADE_DRUGS_PAST_30_DAYS	MADE_DRUGS_PAST_6_MTHS
count	441	441	441
mean	195.76644	3.444444	28.943311
std	297.577618	15.439095	148.301379
min	0	0	0
25%	0	0	0
50%	100	0	0
75%	180	0	0
max	999	99	999

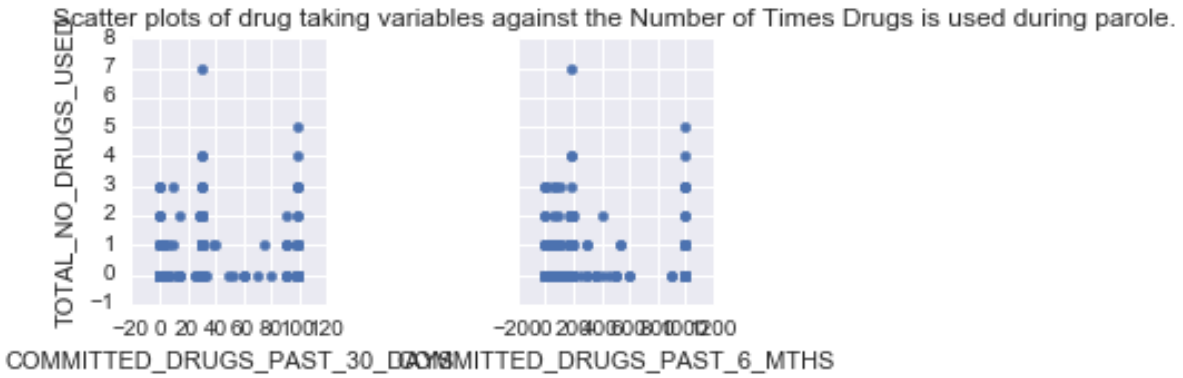
	SOLD_DRUGS_PAST_30_DAYS	SOLD_DRUGS_PAST_6_MTHS	TOTAL_NO_DRUGS_USED
count	441	441	441
mean	23.47619	187.938776	0.712018
std	33.052713	317.264621	1.064325
min	0	0	0
25%	0	0	0

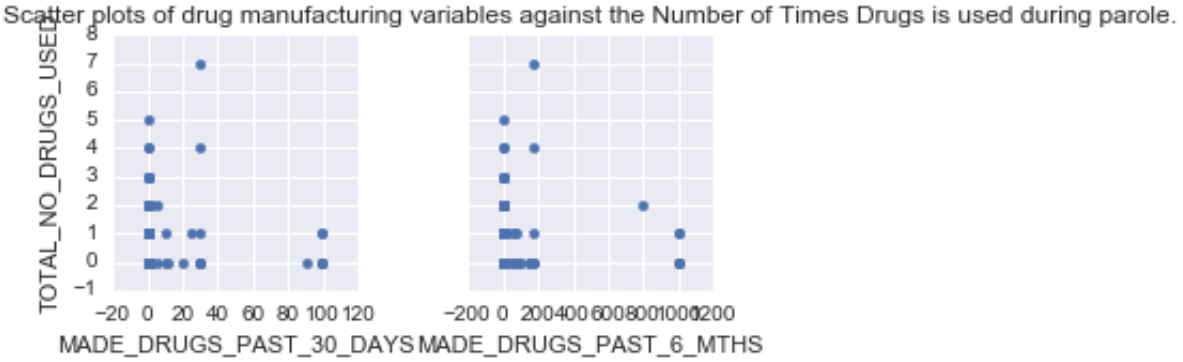
50%	2	20	0
75%	30	180	1
max	99	999	9

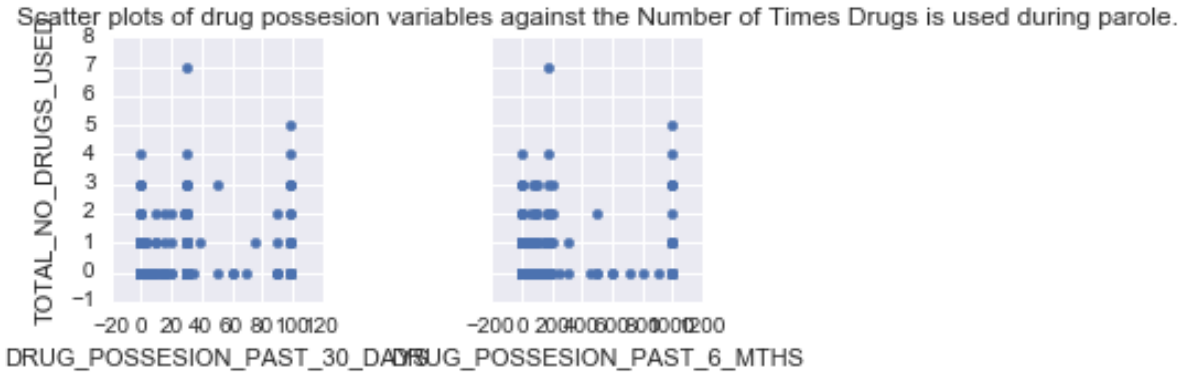
The table above shows some simple statistics for the “Total Number of Drugs Used” in the 9 months after parole and the predictive variables used for our analysis.
The average number of times a parolee used drugs with in 9 months of parole is 0.712 times.
There are 441 clients, N=441 in the sample and there are 8 predictive variables.

Bivariate Analysis.

Below are scatterplots of the predictors used plotted against the target variable of “Total number of times a parolee used drugs 9 months after being released”. These plots were created from the training dataset.







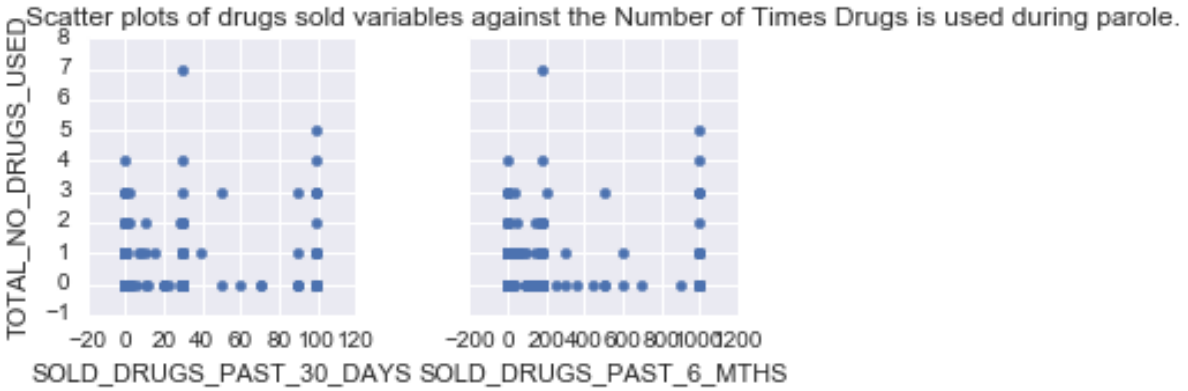


Table below shows the Pearson coefficient and R-squared for each predictor and the target variable.

Pearson Coefficient COMMITTED_DRUGS_PAST_30_DAYS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = 0.0786605556941	RSquared = 0.0061874830221
Pearson Coefficient COMMITTED_DRUGS_PAST_6_MTHS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = 0.0852711708807	RSquared = 0.00727117258336
Pearson Coefficient DRUG_POSSESION_PAST_30_DAYS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = 0.100670802646	RSquared = 0.0101346105053
Pearson Coefficient DRUG_POSSESION_PAST_6_MTHS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = 0.0906437989703	RSquared = 0.00821629829176
Pearson Coefficient MADE_DRUGS_PAST_30_DAYS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = -0.02355028744	RSquared = 0.000554616038508
Pearson Coefficient MADE_DRUGS_PAST_6_MTHS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = -0.0074667692343	RSquared = 5.57526427983e-05
Pearson Coefficient SOLD_DRUGS_PAST_30_DAYS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = 0.0639299158849	RSquared = 0.00408703414505
Pearson Coefficient SOLD_DRUGS_PAST_6_MTHS vs TOTAL_NO_DRUGS_USED and Rsquared	
Pearson Correlation = 0.0621968297472	RSquared = 0.00386844563061

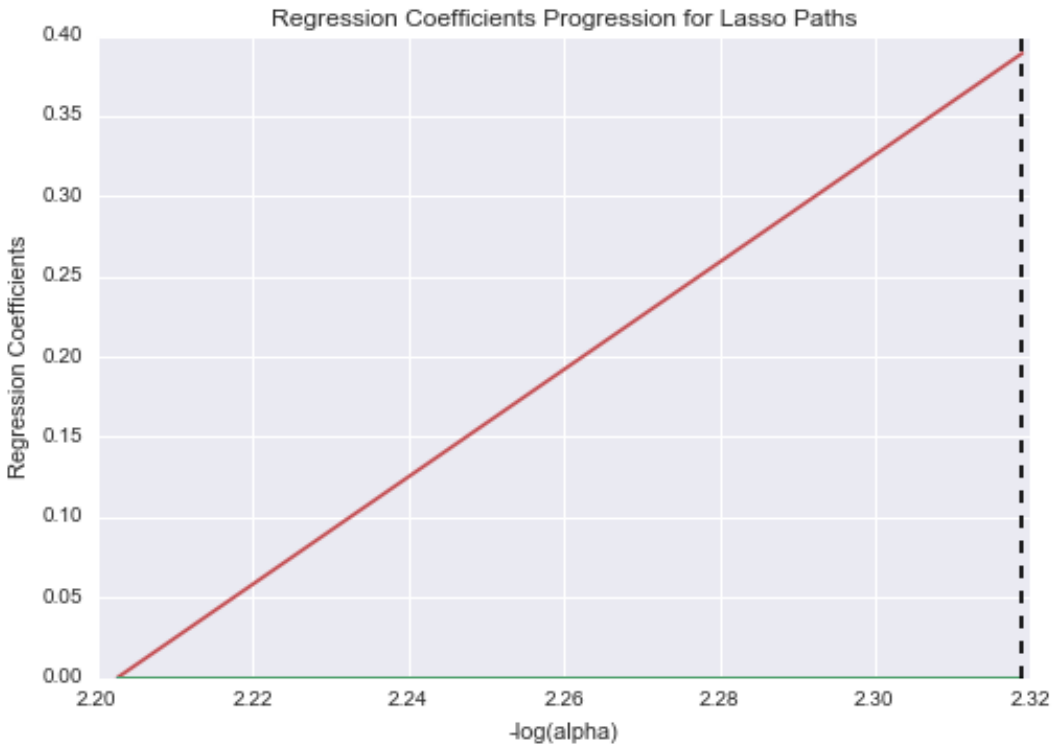
Statistical Model/Multivariate analysis.

LASSO regression was used to analysis the multivariate data. K-fold cross validation with 10 random folds was used on the training set to choose the final statistical model. In this case the first fold will be used as the validation set with the other 9 folds used to build the model.

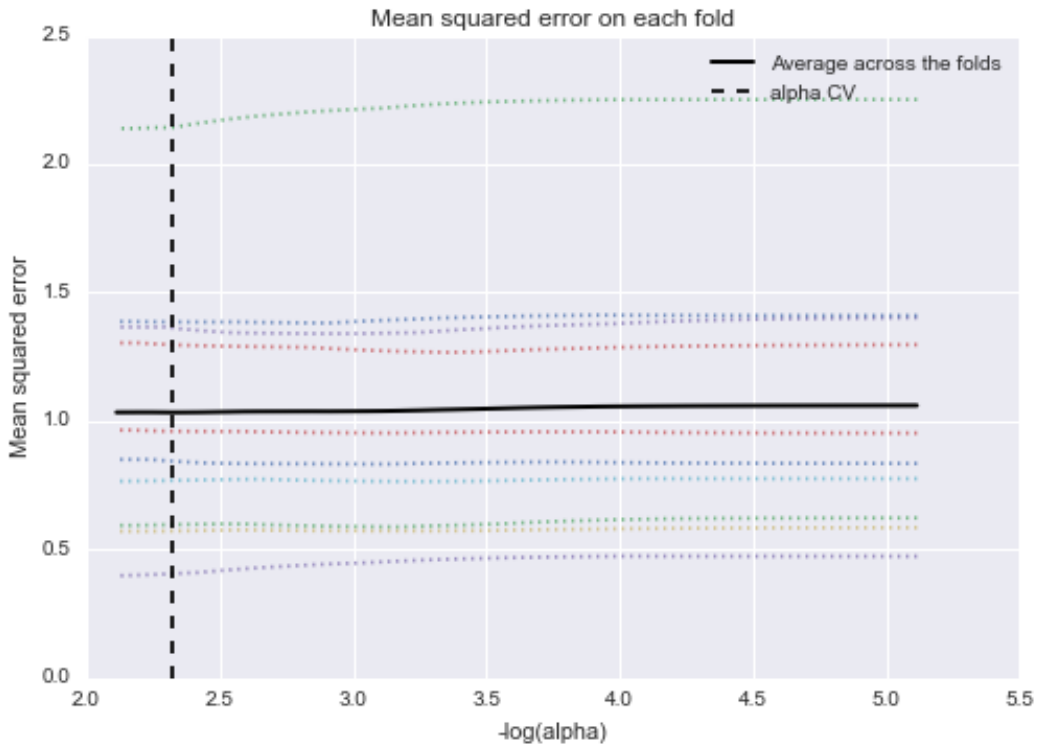
Below is a table of predictors that will be selected for the model. Any non-zero regression coefficient will not be used in the final model.

Predictors	**Regression Coefficient**
COMMITTED_DRUGS_PAST_30_DAYS	0
COMMITTED_DRUGS_PAST_6_MTHS	0
DRUG_POSSESION_PAST_30_DAYS	0.022702019
DRUG_POSSESION_PAST_6_MTHS	0
MADE_DRUGS_PAST_30_DAYS	0
MADE_DRUGS_PAST_6_MTHS	0
SOLD_DRUGS_PAST_30_DAYS	0
SOLD_DRUGS_PAST_6_MTHS	0

The plot below shows the relative importance of the predictors through each of the selection processes and how the regression coefficient changes along each step as a new variable enters the model.



Plot of MSE for each fold in the validation process.



Below is the Mean Squared Error and R-squared for the model based on the training and test data.

training data MSE	1.020428499
test data MSE	1.301845764
training data R-square	0.004203519
test data R-square	-0.024813456

Conclusion:

The aim of the project is to be able to predict how often a parolee incarcerated through certain drug related activities will be participating in any drug related activities during their 9 month parole period. From the lasso regression done above, the results have been pretty poor and the predictors used and algorithm created would not be very efficient at future prediction.

To begin with, 8 predictors were chosen to help create our algorithm. After the data was cleaned there were N=441 parolees in our sample. A bivariate analysis or scatterplot of each predictor was plotted against the target variable. Each variable pair not just had a scatterplot done but also their associated pearson correlation and r-squared value calculated. The best result came from “Drug Possession over last 30 days” prior to arrest as a predictor versus the target of “Total drugs used” during parole. The pearson correlation was 0.1 with an r-squared value of 0.01. A pearson correlation of 0.1 really does not show any correlation with only 0.01 of the variance explained through the same predictor which is also a pretty terrible value.

The regression coefficient calculated from the lasso regression shows that there is only 1 predictor worthy of using in the final model. The only predictor used for the model is also not surprisingly ”Drug possession in the past 30 days” with a coefficient of 0.0227 which is also a relatively poor value in strength for the final model.

The plot of Mean Squared Error for each fold in the lasso regression shows that at least the predictors show the same shape in each fold however they do not converge and thus don’t really show any real relevance for our final model.

The MSE for the test dataset shows an increase from the training dataset meaning the model failed to predict as accurately through initial training of the model with respectively poor values in their corresponding r-squared values.

Due to the poor results the implications of the project are not really clear. The predictors are clearly not the best for the target results. Other predictors would need to be looked into. Also, in the data management part, the erroneous values were substituted for zero values and this may place too much weight towards zero values for some predictors.

Going forward, better predictors need to be chosen. A slight change in target variable may also be needed since the mean outcome for the target variable chosen (Any Drug Use during parole) was 0.71 which really doesn't mean very much. Better data management would also aid in potentially creating better sample data for training the algorithm.