# Understanding the Opioid Crisis: A Machine Learning Approach

Ningyuan Huang, Amy Lei, Zacharie Martin and In-Ho Yi

***Abstract*—The opioid crisis is an unprecedented epidemic. National surveys have been conducted to better understand the drug abuse phenomenon. Critical insights can be uncovered by performing feature selection related to drug abuse behavior, so as to obtain a good understanding of a user‘s behavioral patterns, and build a model to predict the likelihood of opioid abuse such that healthcare providers, and families of at-risk users can take preventive measures.**

## I. Business Understanding

Since the prescription of opioids in the United States became more common in the 1990s, there has been an unfortunate rise in the abuse of these drugs. According to the National Institute on Drug Abuse, “Opioids are a class of drugs that include the illegal drug heroin, synthetic opioids such as fentanyl, and pain relievers available legally by prescription, such as oxycodone (OxyContin®), hydrocodone (Vicodin®), codeine, morphine, and many others.” In 2015, approximately 2 million people in the United States abused prescription opioid pain relievers. Today, more than 115 people die every day from opioid overdose in the U.S. and the estimated economic burden of prescription opioid misuse is $78.5 million every year [1][2]. While the U.S. Department of Health and Human Services has launched initiatives to combat opioid abuse, we have the opportunity to tackle this problem from a data science perspective - how can we foresee a drug user’s risk for opioid abuse given demographic information, drug use history, and behavioral patterns? By data mining the results of the 2015 National Survey on Drug Use and Health (NSDUH-2015), we can identify the most important features that indicate opioid abuse and build a model that predicts the likelihood of opioid abuse for future drug users.

Our data mining solution is useful on several levels. It achieves social good by providing the public with a robust understanding of the factors that contribute to opioid abuse, and helps families and community members identify risky behavioral patterns in opioid users. In the research community, our solution is a resource for substance and drug abuse researchers and clinicians. In an economic context, our solution can provide valuable insight for healthcare providers and payers. Since our model identifies risk factors related to opioid abuse, key stakeholders can leverage this information to instill preventive measures so that opioid users who are at high risk for abuse do not reach that stage. In turn, this will help healthcare providers and payers curtail the economic costs for treating opioid addiction and dependence.

## II. Data Understanding

### A. Data collection and Extracted Features

The original NSDUH-2015 dataset contains responses from 68,073 individuals who identify as the non-institutionalized population of the United States aged 12 or older. The dataset available for public use, however, is treated with a statistical disclosure limitation method that filters out certain responses. Therefore, the dataset we retrieve from the Substance Abuse and Mental Health Archive as a tab-separated values file contains responses from 57,146 individuals [3]. Each survey response, or data instance contains answers to questions about demographics, the use of illicit drugs, alcohol, tobacco, and mental health. There are 2,678 features that portray this information. To account for the natural variability in responses and missing values, however, these features capture the above information in three different ways:

- Respondent‘s answer to the question
- Statistically imputed answer to the question
- Recoded answer to the question

Unfortunately, we are unable to easily distinguish among each of the above categories by just looking at the feature name (e.g. a recoded feature for cigarette use does not have the prefix or suffix RC in its name). On the other hand, missing values are not simply indicated by NA or a null character; instead they are coded by a range of meta values. For these reasons we have clean the dataset in a way that takes care of these limitations. See section *Data Preparation and Exploratory Analysis* for further explanation.

### *B. Survey Data Bias: Data fidelity and consistency*

It is important to note that survey data may contain dishonest or incomplete information. This intrinsic bias in our dataset cannot be entirely controlled. However, we carefully examine whether any sample bias exist among the survey responses. As the Codebook explains, “Demographic Domains Forced to Match Their Respective U.S. Census Bureau Population Estimates through the Public Use File Weight Calibration Process,” which shows the survey results have been upsampled or downsampled to match with the U.S. census population demographic distribution. Thus, the public file dataset represents de-biased results for different demographic groups. The detailed changes and effects of de-biased result can be found in codebook section “Public Use File Estimates and Standard Errors.” The codebook also states that “to reduce bias and improve prediction, a recoded age variable and additional mental health-related variables (i.e., serious thoughts of suicide in the past year and having a past year MDE) were added in the 2012 model” [4]. This suggests that recoded variables are preferred when doing feature selection or model comparison for a more ‘unbiased’ algorithm.

### *C. Data Leakage*

A judicious analysis of the NSDUH-2015 dataset requires careful treatment of data leakage due to information about the target variable that is potentially represented by many of the features. For this reason we opt for both a quantitative approach – using classical measures from information theory and statistics – and thoughtful consideration of important features as determined by our systematic analysis and domain knowledge.

The target variable for this analysis is opioid abuse, which is a categorical binary variable with classes 0 and 1 encoded as UDPYOPI. Combing through the descriptions of each variable – with over 2,500 variables this is no trivial feat – we see that many of them are essentially our target variable in disguise. For example, the feature UDPYHRPNR is “heroin and/or pain reliever [Opioid] abuse within the past year”. This is data leakage. Other variables are not as straightforward. For example, the feature TXEVRRCVD1 indicates whether or not the respondent “ever received alcohol or drug treatment.” In some instances, inclusion of this feature would constitute data leakage if the respondent received treatment for opioid abuse; for other respondents, this may be an important indicator of whether the individual is at risk for opioid abuse. This is how our systematic approach comes into play [4].

Using mutual information and the p-values we derive from contingency tables for each feature vis-à-vis the target variable, we measure the likelihood that inclusion of each feature leads to data leakage (see *Feature Selection*. For features with a high likelihood of producing data leakage, as determined by our statistical and information theoretic methods, we treat them as containing *a priori* information; i.e., we train separate models for each class of the suspected leakage feature(s). In effect, the model is conditioned on the class(es) of the potential leakage feature(s) and then evaluated. Our approach allows us to discriminate between models with and without potential leaky features – the emphasis here is to develop a qualitative understanding of the relationship between the features and target variable. Our methodology is consistent with literature on the subject [5].

## III. Data Preparation and Exploratory Analysis

### A. Target Variable Imbalanced Classes

Our binary target variable has imbalanced classes, with only 1% (603) of the total observations corresponding to opioid abuse. We investigate the effect of an imbalanced class structure in feature selection and prediction models using downsampling techniques. Namely, we keep all 603 observations for the abuse class and randomly sample 603 observations from the non-abuse class, resulting in a smaller balanced dataset. For feature selection, we compare the feature importance score produced by a decision tree model using the original dataset versus the down-sampled dataset. Results show that while the nominal values vary, the ranking of feature importance does not change. Since our iterative mutual information algorithm is a rank-based feature selection approach (see Section IV-A1), we can guarantee robust results. For modeling, we evaluate the performance of random forest using the original dataset versus the down-sampled data. Results show that the down-sampled model achieves a near perfect F1 score (0.99) without much hyperparameter tuning. However, since the number of features in the down-sampled dataset (4,712) is significantly greater than the number of observations (1,206) in this dataset, we account for overfitting by using the original dataset for such modeling.

### B. Data Cleanup

The dataset at hand presents two significant challenges in terms of adequate data preparation. As aforementioned, there is no specific naming convention for the different types of features, and the survey uses incoherent techniques to record missing or incomplete responses, which leads to many ‘outlying‘ values. Therefore, we analyze the structure of the data and come up with two heuristics.

The survey provides an index table of all the features and their respective categories. Therefore, we are able to classify the features according to the order of the index table.

The following heuristic is used to separate concrete answers from metadata or encoded values. Those metadata start with 9s and end with numbers in the range of 80-99, as discussed in the codebook [3]. For example, the variable ‘AGE WHEN FIRST USED MARIJUANA/HASHISH’ (encoded as ‘MJAGE’) has legal answers from range 1-78, and metadata 985, 991, 994, 997, 998 as bad data, never used marijuana, dont know, refused to answer, blank. Without proper normalization of these features, extreme metadata values will skew our analysis and produce false results.

By looking at the maximum value for each column, we routinely deduce the meta-value range for each column. Similar to one-hot encoding, we create dummmy variables for each of the meta values (e.g. __85, __91, __95 etc), or one dummy for all metadata, as such metadata is usually the cause for missing data in that particular row.

To treat missing values, we create dummy variables (__NA) for each feature that contains missing values. Then we calculate the average among concrete values and use that average per column to replace NA and meta values.

Date values are in the format *year, month pair* and are consistently named __YLU/__MLU *(year/month* of last use) or __YFU/__MFU *(year/month of first use)*. These values are combined to produce months since Jan 2000 by the formula (YEAR $-$ 2000)$\times$12+MONTH. The same processing techniques are applied to meta values and missing values.

### *C. Encoding Format for Efficient Loading/Saving*

Another challenge is the very wide nature of the dataset. The TSV file takes about 10 minutes to load or save, curtailing our ability to produce fast iterations. This is addressed by using a columnar compression format, Parquet [6]. Parquet uses run-length encoding to compress the data and stores it one column at a time. Given most of our data shape is an enum type, with the majority of it containing missing or meta values, this encoding scheme is ideal (see [7]).

This encoding format pays off handsomely; not only is the size of the dataset reduced from 385 MB to 28 MB, but the loading/saving time is also reduced to less than a minute.

## IV. Modeling and Evaluation

### *A. Feature Selection*

Our first goal is to identify key features in predicting opioid abuse through feature selection. Generally, feature selection methods are classified into three major categories: wrapper method, filter method and embedded method [8]. Due to the large number of sparse features in our dataset, we employ methods that are computationally fast and easy for interpretation. We demonstrate successful results from conditional mutual information feature selection (filter method) and regularized tree-based model (embedded method). It is important to first use the filter method in order to screen out important features, and then evaluate these features based on our domain knowledge in order to eliminate the possibility of data leakage. The embedded method is used next to further identify other suspicious features, as discussed in Section IV-A1.

*1) Iterative Mutual Information Selection:* As most of the features are categorical, mutual information is a natural selection heuristic. The naive algorithm tends to rank features based on mutual information between each individual feature and the target variable, but this approach is likely to introduce redundancy when the features are dependent on each other [9]. Thus, we propose a novel iterative conditional mutual information feature selection algorithm (CMI-FS), where a similar idea has been discussed in [9]. The algorithm iteratively selects features that maximize their mutual information with the target variable conditionally to any other feature already selected. The conditional mutual information is calculated by conditional entropy and entropy among features:

$$\begin{aligned} I(X,Y|Z) &= H(X|Z) + H(Y|Z) - H(X,Y|Z) \\ &= H(X|Z) + [H(Z) + H(Y|Z)] - [H(X,Y|Z) + H(Z)] \\ &= H(X|Z) + H(Y,Z) - H(X,Y,Z) \end{aligned} \tag{1}$$

The algorithm is outline below. It iterates until K features are selected.

1) *Initialization:*
   Set $X = X_1, X_2, ..., X_n$. Set $S = \emptyset$
2) *First pass:*
   Compute the pairwise mutual information between $X_i \in X$ and $Y$, select the feature with highest mutual information named $X_1^*$;
   $X = X \setminus X_1^*$, $S = X_1^*$
3) *Iterative feature selection:*
   a) Compute conditional entropy $H(X_i|X)$, entropy $H(Y|X)$, and $H(X_i, Y, X)$;
   b) Calculate the conditional mutual information using formula (1);
   c) Select the feature with highest conditional mutual information named $X_2^*$;

d) $X = X \setminus X_2^*$,

$$S = \{X_1^*, X_2^*\}$$

Considering that the features belong to 6 major categories, where some of them are potential data leakage features, we run CMI-FS independently on each subset of features and compare the selected features with the (conditional) mutual information scores. Key results are shown in Table VI of the Appendix.

Note that features in the first three categories, which are related to use of other drugs, tend to have high (conditional) mutual information. This indicates potential data leakage. The presence of heroin use seems to be highly related to opioid abuse, and other similar features' importances significantly drop when controlling on heroin features. This finding provides an important *a priori* and reveals a perfect data leakage feature, UDPYHRPN. As shown in the Table I, all respondents who reported opioid abuse also reported heroin or pain reliever dependence.

TABLE I
PERFECT DATA LEAKAGE: HEROINE USAGE FEATURE 'UDPYHRPNY' VS TARGET VARIABLE 'UDPYOPI'

| UDPYHRPNR | UDPYOPI=0 | UDPYOPI=1 |
|---|---|---|
| 1 = Heroin only | 0 | 87 |
| 2 = Pain reliever only | 0 | 452 |
| 3 = Heroin and pain reliever | 0 | 64 |
| 4 = Neither | 56543 | 0 |

*2) Embedded Feature Selection:* In order to detect other data leakage features, we fit a random forest model using the reduced set of features that are produced by the above iterative feature selection method. The columns we dropped are shown in Table II.

In addition, columns that have no values, identification columns, and analytic weight columns are all dropped from the dataset. After this initial cleanup, we have 4,712 columns.

TABLE II
LIST OF DROPPED COLUMNS

| Feature | Name |
|---|---|
| PNRNMREC | RC - MOST RECENT PAIN RELIEVER MISUSE (RECODE) |
| PNRNM30D | USED PAIN RLVR NOT DIRECTED BY DR PAST 30 DAYS |
| PNRNMLIF | EVER USED PAIN RELIEVER NOT DIRECTED BY DR |
| HERREC | TIME SINCE LAST USED HEROIN |
| HERPNRYR | RC-HEROIN USE AND/OR PAIN RELIEVER MISUSE - PAST YEAR |
| PNRNMWUD | RC-PAIN RELIEVER MISUSER WITH USE DISORDER - PAST YEAR |
| IRHERYFU | HEROIN YEAR OF FIRST USE - IMPUTATION REVISED |
| IRCRKRC | CRACK RECENCY - IMPUTATION REVISED |
| UDPYHRPNR | RC-HEROIN AND/OR PAIN RELIEVER DEPENDENCE OR ABUSE - PST YR |

To measure and evaluate the performance of the model, a learning curve is used with sample sizes ranging from 500 to 40,000 and a cross-validation set of five. The F1 score is used for evaluation, given the heavily skewed nature of the dataset.

Our initial random forest of max_depth=5 and n_estimators=20 produces the learning curve in Figure 1.

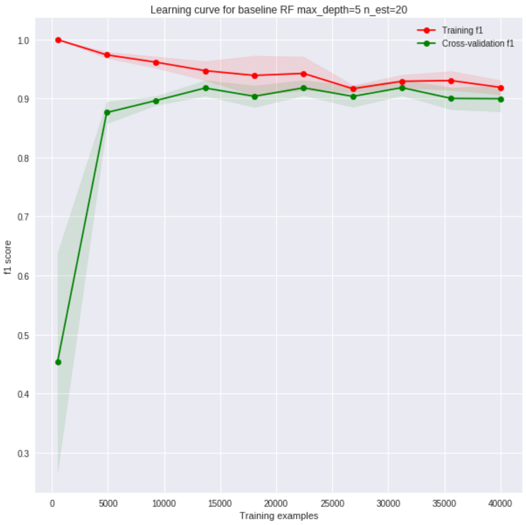


Fig. 1. Learning curve for baseline before further pruning

We see the typical symptom of high bias, as the training and cross-validation curves are close to each other. Therefore, we loosen the regularization effect by increasing the max_depth to 10 as illustrated in Figure 2.

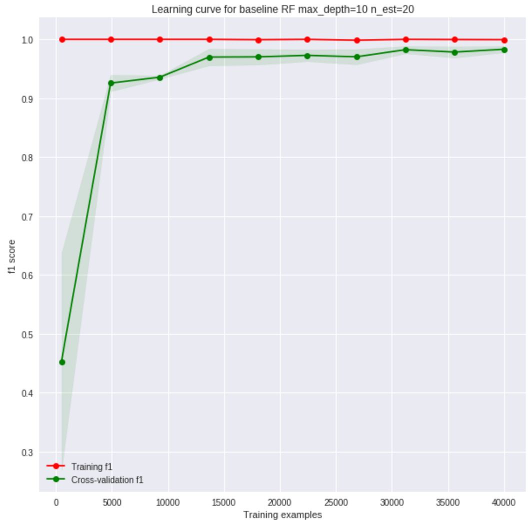


Fig. 2. Learning curve with max_depth=10 before further pruning

Now we have the near-perfect score for the training graph, and therefore we do not expect to see further improvement from the current performance. Seeing a very high score in general (around 0.98-0.99), we suspect there are still lots of data leakage features and first examine the ten most important features produced in the model, as shown in Table III.

Since almost all the prominent features are either related to pain reliever or heroin use, we aggressively dropped all columns that contain either ‘HER‘ or ‘PNR‘, assuming that these are data leakage features. After dropping these columns, we are left with 4,406 columns.

With this further reduced dataset, we obtain the following results from random forest in Figure 3. This serves as our baseline model and will be further discussed in Section IV-B.

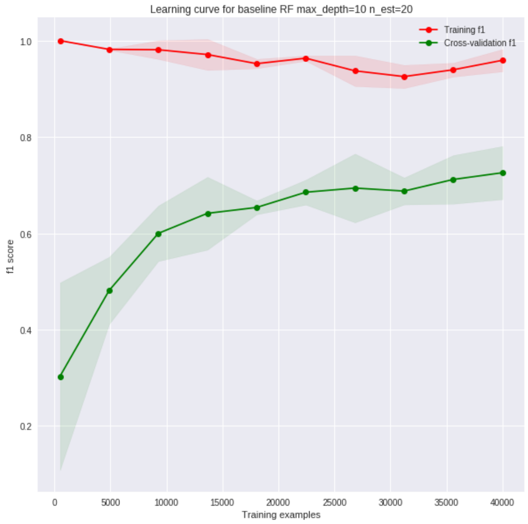


Fig. 3. Learning curve with max_depth=10 after further pruning

*3) Chi-squared Test:* To cross-check the results of the iterative mutual information approach above, we use a statistical approach - the Chi-squared test of independence - to determine if there is a significant relationship between categorical features and the target variable. This is also an interactive approach where a Chi-squared test is run for each feature with the target variable. Results indicate that 1,949 features have an associated with the target variable (p-value $<$ 0.01). Interestingly, all the features selected by CMI-FS have significant p-values from the Chi-squared test. See Table VII for the top 20 features with the lowest p-values.

As expected, most of the other significant features are related to different kinds of opioid use, including oxycodone, morphine, and fentanyl. These results provide more insight into data leakage issues.

*4) K-Modes Algorithm:* We also use an unsupervised learning approach to gain additional insight into our dataset. Our initial hypothesis is that there is a significant difference in the demographic distribution of opioid abusers and non-abusers. To test our hypothesis we employ the k-modes algorithm [10]. The k-modes algorithm is similar to the k-means algorithm, but is specifically designed for categorical variables. Therefore, it uses modes instead of means for clustering. Our results confirm that there are indeed significant demographic differences between the two classes. We see further evidence of this clustering in our 2-dimensional PCA plot in Figure 11.

TABLE III
TEN MOST IMPORTANT FEATURES BEFORE FURTHER PRUNING

| Feature | Gini imp. | Name |
|---|---|---|
| DEPNDPYPNR | 0.05821 | RC-PAIN RELIEVER DEPENDENCE - PAST YEAR |
| PNRLWD3SX | 0.03717 | HAD 3+ PN RLVR WITHDRAWAL SYMPTOMS PST 12 MOS |
| PNRLEMCTD_META | 0.03574 | CONTD TO USE PN RLVR DESPITE EMOT PRBS (Meta includes ”did not use”) |
| ABODHER | 0.02969 | RC-HEROIN DEPENDENCE OR ABUSE - PAST YEAR |
| PNRLWDSMT | 0.02729 | HAD 3+ PN RLVR WDRAW SYM SAME TIME PST 12 MOS |
| UDPYPNR | 0.02458 | RC-PAIN RELIEVER DEPENDENCE OR ABUSE - PAST YEAR |
| DEPNDPYILL | 0.02404 | RC-ILLICIT DRUG DEPENDENCE - PAST YEAR |
| ABUSEPYPNR | 0.02309 | RC-PAIN RELIEVER ABUSE - PAST YEAR |
| PNRLEMCTD | 0.02182 | CONTD TO USE PN RLVR DESPITE EMOT PRBS |
| PNRNMWOUD | 0.02117 | RC-PAIN RELIEVER MISUSER WITHOUT USE DISORDER - PAST YEAR |

In Section II-C we mention the importance of qualitative understanding in machine learning. Our data mining project tackles a complex real-world dataset – making interpretability, arguably, even more relevant. The benefit of our “bottom-up” (models built with highly restricted domain-knowledge based feature selection) analysis is that the results are easily to understand. For example, the k-modes algorithm shows significant differences between the classes of the target variable in employment status, income, location, and race. Opioid abusers are less likely to have achieved education levels beyond high school, are more likely to have lower income, are more likely to live in a rural area, and are more likely to be white men. These results are consistent with other findings. We have built an interactive dashboard to allow users who are interested in analyzing some of these demographic risk factors. See the *Appendix* for a link to this dashboard [11][12].

### B. Results and Prediction

*1) Tree-based Model:* Recall from Section IV-A2 that we take Figure 3 as our baseline model. The learning curve is indicative of high variance and moderate bias.

First, we tackle the bias problem. We loosen the regularization effect by increasing the max depth and number of estimators. A summary of the results from further parameter tuning is shown in Table IV.

For all such runs, the training F1 score is consistently at or very close to 1.0. Notice that increasing the number of estimators generally results in better performance even when the training score is already at 1.0. This shows the ensemble model’s ability to further boost the result in a high variance situation without incurring the usual bias-variance trade-off.

We also attempt to limit max_features to 10 with the hope that this will boost areas where dominant features negatively impact the model. However, the result turns out to be far below the benchmark level.

TABLE IV
SUMMARY OF FURTHER PARAMETER TUNING

| Settings | n=22444 | n=40000 |
|---|---|---|
| max_depth=20,n_est=20 | 0.7364 | 0.7854 |
| max_depth=20,n_est=50 | 0.7756 | 0.8191 |
| max_depth=50,n_est=50 | 0.7611 | 0.8134 |
| max_depth=50,n_est=100 | 0.7854 | 0.8368 |
| max_feat=10,n_est=100 | 0.2179 | 0.2409 |

Our best random forest result uses 100 estimators and an unlimited tree depth, as shown in Figure 4.

In general, the model portrays the high variance problem. The learning curve is not yet leveling off on the extreme end of the graph, which suggests that the performance will improve if we can gather more data.

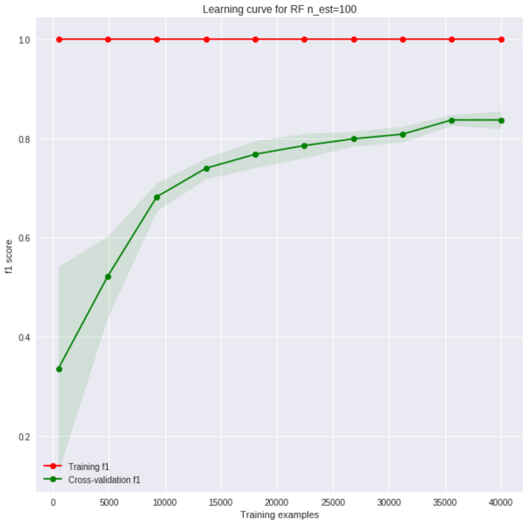


Fig. 4. Learning curve with n_est=100

With regard to the bias-variance trade-off, we try to further regularize the model. This results in lower variance, but an increase in bias. This indicates an overall ‘unfavorable’ result. Therefore, the learning curve diagnostics suggest that obtaining more data will be a good investment in improving the model. Again, however, we would need more respondents, which may be a costly process. It is therefore advisable to look at the utility function of spending resources in procuring the data.

*2) Interpretation of the Model:* A random forest of max_depth=10 and n_estimators=20 is fit with all the observations of the reduced dataset.

The 10 most important features from the model and their Gini importance values are given in Table V.

This shows that even after aggressive pruning, most of the important features are still related to drug use. Interestingly, the most prominent marijuana-related feature is DEPNDMRJ (RC-MARIJUANA DEPENDENCE - PAST YEAR) with a Gini importance value of 0.004085, which is below that of many features related to drug use.

For binary classification, logistic regression and random forest are two popular algorithms. Note that logistic regression is not suitable for our dataset due to the high number of non-linear features and the high variance introduced by the mix of range data (age) and categorical data (encoded metadata). Therefore, we prefer tree-based models and random forest as our final prediction model. We further improve its performance by examining the training and testing F1 scores as well as tuning the hyperparameters.

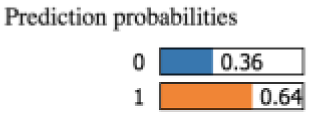


Fig. 5. An instance of the probabilities related to opioid abuse

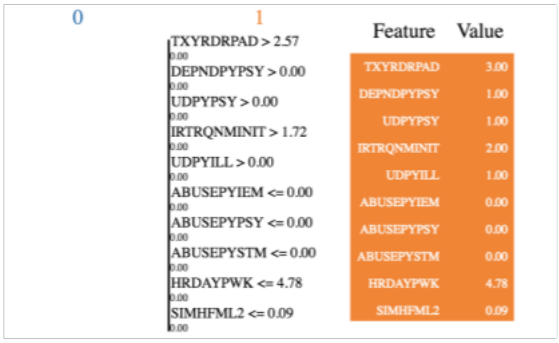


Fig. 6. An instance of the top ten most significant features and values for opioid abuse

We also utilize LIME to interpret our random forest model on the user instance basis. Below is an instance of a predicted opioid abuser. Factors that relate to opioid abuse include receiving treatment of alcohol and drug use in a private doctor’s office (TXYRDRPAD = 3), psychotherapeutic dependence (DEPNDPYPSY = 1) and psychotherapeutic dependence or abuse (UDPYPSY = 1). This suggests that the interaction between an opioid user and healthcare provider plays a role in the user’s trajectory to opioid abuse, which is an implication that resonates with findings published in the New England Journal of Medicine [13].

TABLE V
TEN MOST IMPORTANT FEATURES IN THE MODEL

| Feature | Gini imp | Name |
|---|---|---|
| UDPYPSY | 0.03779 | RC-PSYCHOTHERAPEUTIC DEPENDENCE OR ABUSE - PAST YEAR |
| DEPNDPYPSY | 0.03773 | RC-PSYCHOTHERAPEUTIC DEPENDENCE - PAST YEAR |
| UDPYIEM | 0.02849 | RC-ILLICIT DRUG OTHER THAN MRJ DEP OR ABUSE - PAST YEAR |
| TXYRNDILL | 0.02197 | RC-NEEDED TREATMENT FOR ILLICIT DRUG USE - PST YR |
| DEPNDPYILL | 0.02038 | RC-ILLICIT DRUG DEPENDENCE - PAST YEAR |
| OXCOPDPYMU | 0.01987 | RC-OXYCODONE PRODUCTS - PAST YEAR MISUSE |
| UDPYILL | 0.01919 | RC-ILLICIT DRUG DEPENDENCE OR ABUSE - PAST YEAR |
| DEPNDPYIEM | 0.01178 | RC-ILLICIT DRUG OTHER THAN MARIJUANA DEP - PAST YEAR |
| OXYCPDPYMU | 0.01104 | RC-EDITED OXYCODONE PRODUCTS - PAST YEAR MISUSE |
| HRNDLREC | 0.01013 | TIME SINCE LAST USED NEEDLE TO INJECT HEROIN |

## V. Deployment

### A. Policy Framework: Causal Inference

At the policy level, understanding the factors that contribute to the opioid crisis is of the utmost importance. Policy makers must decide on how to allocate finite resources. More accurate casual information can make the difference between the ability to ameliorate conditions and a worsening crisis. Here is where the machinery of causal inference comes into play. Causal inference is the name given to the mathematical tools used to infer causal effects, causes of effects, and direct/indirect effects [14]. Our results suggest the importance of the so-called drug environment variables (Figure 5 - where these variables are represented as ”confounding factors”). Drug environment variables include: how easy it is for an individual to purchase opioids, access to opioid users within the respondents’ social network, etc. This potentially contrasts with earlier work [12] [13] that suggests worsening medium-term economic prospects are to blame for the crisis (Figure 6).

We propose testing these casual models for future research. Results of this research may indicate an optimal balance of resource allocation between limiting access to opioids and improving medium-term economic prospects.

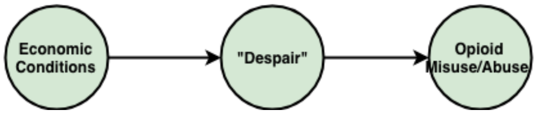


Fig. 7. Casual graph suggested by (Case and Deaton 2017; Deaton 2017) (Stiglitz 2015; Meara and Skinner 2015; Pierce and Schott 2016)

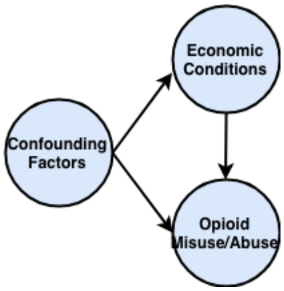


Fig. 8. Our hypothesized casual graph

### B. Infrastructure

The following lists the necessary infrastructure pieces to deploy our model:

- **Cloud storage** service, such as AWS S3, GCP CloudStore, or on-premise hosted HDFS cluster
- **Private container repository** service to store applications such as pipelines, trainers and application servers
- **Continuous integration/Deployment** service to rebuild container images upon committing new code or the availability of a new serialized trained model image
- **Declarative infrastructure deployment** (such as TerraForm [15]) will be used to update running docker

images in the pipelines, trainers and the server

- **Orchestration framework** such as Apache Airflow [16] will be used to execute the workflow

Schematic infrastructure and data flow diagram is shown in Figure 9.

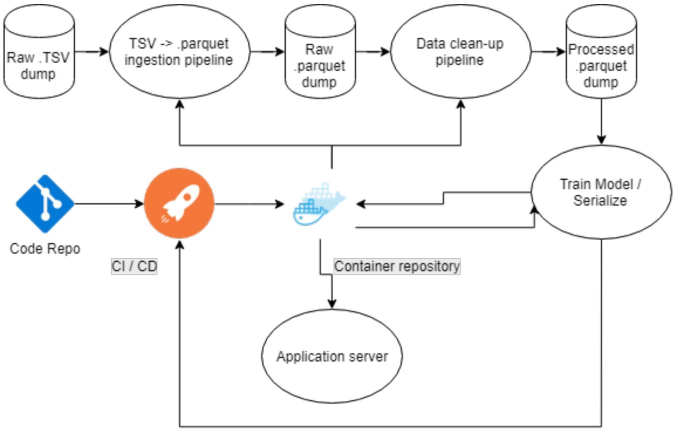


Fig. 9. High-level infrastructure design

*1) Data Update Workflow:* Data update will happen in the following steps:

1) When new survey data is available, the data is uploaded to a cloud storage location. This can be done manually or can be automated.
2) Airflow will sense the availability of the new dataset, and will kickoff the ingestion pipeline.
3) When the ingestion is complete, Airflow will then run a clean-up pipeline from the container repository.
4) Airflow will then run a model trainer/serializer, create a new model, build a new container, publish to the repository, and triggers a new build from CI
5) CI will build a new application server image, combining with the latest serialized model image, and publishes the new application server to the repository
6) Finally, CI deploys the new application server seamlessly via TerraForm.

*2) Code Change Workflow:* When the new code is checked in, the CI system will rebuild images and publish them to the repository. Airflow will then sense the availability of the new image and kickoff appropriate process to update the dataset.

*3) Developer Sandbox for Fast Iteration:* To enable fast iteration and agile development, all the datasets generated in the process will be available for the developers to download to the local machine, or query directly via SparkSQL [17]. The fact that the datasets are all in Parquet format makes it easy to enable data exploration at scale. Developers will be able to reproduce all the pipeline executions either by running the container image or by running pipelines built from the source code. All the data produced by the pipelines will be timestamped and will remain immutable to allow diffing between versions and postmortem analyses.

### *C. Ethical Considerations*

It is important to keep in mind the ethical implications of our model, as it relies on sensitive data such as a respondent's substance and drug abuse history. While our model suggests certain risk factors related to opioid abuse, the actual story of an opioid user's trajectory is dependent on other external conditions. Human behavior itself is inherently difficult to predict, so we recognize that our solution may raise concern for certain stakeholders, however, our overall goal is to provide users with a better *understanding* of the opioid crisis.

## References

[1] National Institute on Drug Abuse, "Opioid overdose crisis," 2018, [Online; accessed 7-December-2018]. [Online]. Available: https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis

[2] Centers for Disease Control and Prevention, "Opioid overdose," 2017, [Online; accessed 7-December-2018]. [Online]. Available: https://www.cdc.gov/drugoverdose/data/prescribing.html

[3] Substance Abuse and Mental Health Data Archive, "National survey on drug use and health 2015," Rockville, MD, USA, 2015, [Online; accessed 7-December-2018].

[Online]. Available: https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2015-nsduh-2015-ds0001-nid16894

[4] ——, "2015 national survey on drug use and health public use file codebook," Rockville, MD, USA, 2015, [Online; accessed 7-December-2018]. [Online]. Available: https://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/NSDUH-2015/NSDUH-2015-datasets/NSDUH-2015-DS0001/NSDUH-2015-DS0001-info/NSDUH-2015-DS0001-info-codebook.pdf

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 1135–1144. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939778

[6] Apache Software Foundation, "Apache parquet," 2018, [Online; accessed 7-December-2018]. [Online]. Available: http://parquet.apache.org/

[7] M. Kleppmann, *Designing Data-Intensive Applications*. O'Reilly, 2017.

[8] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, 2015. [Online]. Available: https://bib.irb.hr/datoteka/763354.MIPRO_2015_JovicBrkicBogunovic.pdf

[9] J. Novovičová, P. Somol, M. Haindl, and P. Pudil, "Conditional mutual information based feature selection for classification task," in *Progress in Pattern Recognition, Image Analysis and Applications*, L. Rueda, D. Mery, and J. Kittler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 417–426. [Online]. Available: http://staff.utia.cas.cz/novovic/files/CIARP07_NSHP.pdf

[10] Z. He, S. Deng, and X. Xu, "Approximation algorithms for k-modes clustering," in *Computational Intelligence*, D.-S. Huang, K. Li, and G. W. Irwin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 296–302. [Online]. Available: https://arxiv.org/pdf/cs/0603120.pdf

[11] E. Meara and J. Skinner, "Losing ground at midlife in america," *Proceedings of the National Academy of Sciences*, vol. 112, no. 49, pp. 15 006–15 007, 2015. [Online]. Available: https://www.pnas.org/content/112/49/15006

[12] A. Case and A. Deaton, "Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century," *Proceedings of the National Academy of Sciences*, vol. 112, no. 49, pp. 15 078–15 083, 2015. [Online]. Available: https://www.pnas.org/content/112/49/15078

[13] M. L. Barnett, A. R. Olenski, and A. B. Jena, "Opioid-prescribing patterns of emergency physicians and risk of long-term use," *New England Journal of Medicine*, vol. 376, no. 7, pp. 663–673, 2017, pMID: 28199807. [Online]. Available: https://doi.org/10.1056/NEJMsa1610524

[14] J. Pearl, "Causal inference in statistics: An overview," *Statist. Surv.*, vol. 3, pp. 96–146, 2009. [Online]. Available: https://doi.org/10.1214/09-SS057

[15] HashiCorp, "Terraform by hashicorp," 2018, [Online; accessed 7-December-2018]. [Online]. Available: https://www.terraform.io/

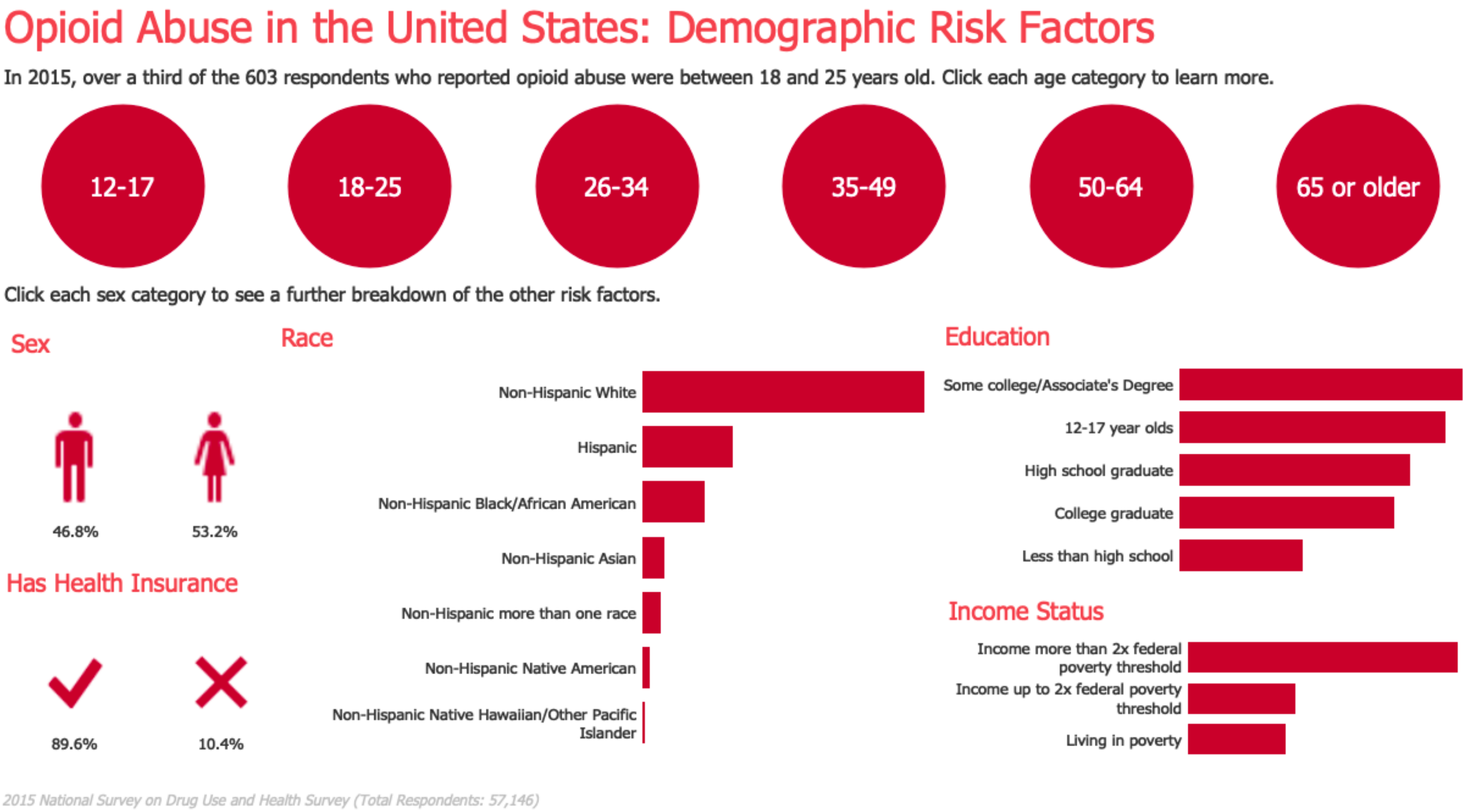
[16] Apache Incubator, "Apache airflow," 2018, [Online; accessed 7-December-2018]. [Online]. Available: https://airflow.apache.org/

[17] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia, "Spark sql: Relational data processing in spark," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 1383–1394. [Online]. Available: http://doi.acm.org/10.1145/2723372.2742797

## CONTRIBUTIONS

- **In-Ho Yi**: Data Cleanup, Random Forest hyperparameter tuning, Gini Importance-based Interpretation, Deployment Infrastructure
- **Zacharie Martin**: Data Leakage, Causal Inference Analysis, Mutual Information Analysis, Feature Selection - K-modes and Clustering
- **Amy Lei**: Business Understanding, Data Understanding, Feature Selection - Chi-squared test, Demographics Dashboard, Ethical Considerations
- **Ningyuan Huang**: Data Understanding, Target Variable Imbalanced Classes, Feature Selection - Iterative Mutual Information, Random Forest model, Interpretation of the Model

## APPENDIX



Opioid Abuse in the United States: Demographic Risk Factors

In 2015, over a third of the 603 respondents who reported opioid abuse were between 18 and 25 years old. Click each age category to learn more.

Click each sex category to see a further breakdown of the other risk factors.

2015 National Survey on Drug Use and Health Survey (Total Respondents: 57,146)

12-17
18-25
26-34
35-49
50-64
65 or older
Sex
46.8%
53.2%
Has Health Insurance
✓
✗
89.6%
10.4%
Race
Non-Hispanic White
Hispanic
Non-Hispanic Black/African American
Non-Hispanic Asian
Non-Hispanic more than one race
Non-Hispanic Native American
Non-Hispanic Native Hawaiian/Other Pacific Islander
Education
Some college/Associate's Degree
12-17 year olds
High school graduate
College graduate
Less than high school
Income Status
Income more than 2x federal poverty threshold
Income up to 2x federal poverty threshold
Living in poverty

Fig. 10. Demographics Dashboard

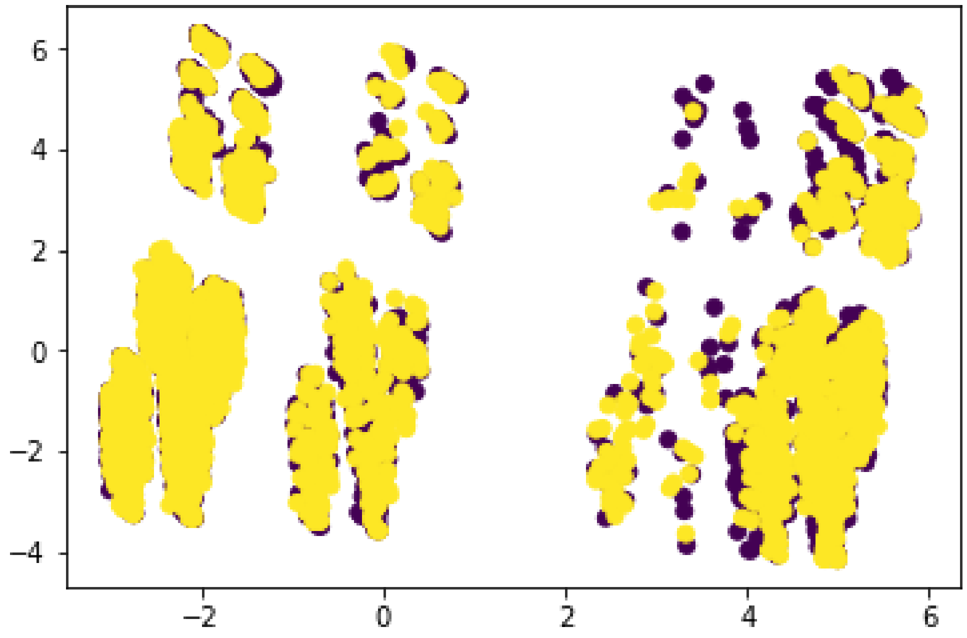


Fig. 11. 2 Dimensional PCA Results

TABLE VI
Mutual Information Test Results

| Category | Features | MI Value | Name |
|---|---|---|---|
| SASUS | PNRNMREC | 0.55 | RC - MOST RECENT PAIN RELIEVER MISUSE (RECODE) |
| SASUS | PNRNM30D | 0.36 | USED PAIN RLVR NOT DIRECTED BY DR PAST 30 DAYS |
| SASUS | PNRNMLIF | 0.35 | EVER USED PAIN RELIEVER NOT DIRECTED BY DR |
| SASUS | HERREC | 0.066 | TIME SINCE LAST USED HEROIN |
| SASUS | ALCDAYS | 0.022 | # DAYS HAD ONE OR MORE DRINKS PAST 30 DAYS |
| SASUS | ALCDAYS | 0.022 | # DAYS HAD ONE OR MORE DRINKS PAST 30 DAYS |
| ISU | HERPNRYR | 0.61 | RC-HEROIN USE AND/OR PAIN RELIEVER MISUSE - PAST YEAR |
| ISU | PNRNMWUD | 0.053 | RC-PAIN RELIEVER MISUSER WITH USE DISORDER - PAST YEAR |
| ISU | IRHERYFU | 0.024 | HEROIN YEAR OF FIRST USE - IMPUTATION REVISED |
| ISU | IRCRKRC | 0.0023 | CRACK RECENCY - IMPUTATION REVISED |
| ISU | SRCCLFRTRQ | 0.0016 | RC-COLLAPSED SOURCE OF FRND/REL TRANQLZRS FOR LAST MISUSE |
| OSAS | UDPYHRPNR | 0.69 | RC-HEROIN AND/OR PAIN RELIEVER DEPENDENCE OR ABUSE - PST YR |
| OSAS | RSKMRJWK | 4.44E-16 | RISK SMOKING MARIJUANA ONCE OR TWICE A WEEK |
| OSAS | DEPENDSED | 1.33E-15 | PRESCRIPTION SEDATIVE DEPENDENCE IN THE PAST YEAR |
| OSAS | ALCLIMIT | 1.33E-15 | SET LIMITS ON ALCOHOL USE PAST 12 MONTHS |
| OSAS | RSKYFQDGR | 8.88E-16 | GET A REAL KICK OUT OF DOING DANGEROUS THINGS |
| II | AGE2 | 0.049 | RECODE - FINAL EDITED AGE |
| DEMO | WRKSTATWK2 | 0.053 | WORK SITUATION IN PAST WEEK - RECODE |
| DEMO | IREDUHIGHST2 | 0.058 | RC-EDUCATION CATEGORIES |
| DEMO | SEXRACE | 0.15 | RC-COMBINED GENDER BY RACE INDICATOR |
| DEMO | IRFAMIN3 | 0.21 | RECODE - IMP.REVISED - TOT FAM INCOME |
| DEMO | IRHHSIZ2 | 0.13 | RECODE - IMPUTATION-REVISED # PERSONS IN HH |
| GEO | AIIND102 | 0.016 | AMER INDIAN AREA INDICATOR |
| GEO | COUTYP4 | 0.0035 | COUNTY METRO/NONMETRO STATUS (2013 3-LEVEL) |

TABLE VII
CHI-SQUARED TEST RESULTS

| Feature | P-value | Name |
|---|---|---|
| CIGREGNM | 0 | SMOKE SAME NUMBER OF CIGS FROM DAY TO DAY |
| OXCOPDAPYU | 0 | RC-OXYCODONE PRODUCTS - ANY PAST YEAR USE |
| OXYCPDAPYU | 0 | RC-OXYCODONE PRODUCTS - ANY PAST YEAR USE |
| OXCNANYYR2 | 0 | RC-EDITED OXYCONTIN - ANY PAST YEAR USE |
| MORPPDAPYU | 0 | RC-MORPHINE PRODUCTS - ANY PAST YEAR USE |
| FENTPDAPYU | 0 | RC-FENTANYL PRODUCTS - ANY PAST YEAR USE |
| BUPRPDAPYU | 0 | RC-BUPRENORPHINE PRODUCTS - ANY PAST YEAR USE |
| OXYMPDAPYU | 0 | RC-OXYMORPHONE PRODUCTS - ANY PAST YEAR USE |
| BKDRUNK | 0 | ARRSTD & BOOKED FOR DRUNKENNESS PAST 12 MONTHS |
| HYDMPDAPYU | 0 | RC-HYDROMORPHONE PRODUCTS - ANY PAST YEAR USE |
| MTDNPDAPYU | 0 | RC-METHADONE PRODUCTS - ANY PAST YEAR USE |
| ADWRLSIN | 0 | WHEN PRBLMS WORST LOSE INTRST IN ENJOYABLE THINGS |
| BKDRVINF | 0 | ARRSTD & BOOKED FOR DUI PAST 12 MONTHS |
| ALPRPDAPYU | 0 | RC-ALPRAZOLEM PRODUCTS - ANY PAST YEAR USE |
| BKARSON | 0 | ARRSTD & BOOKED FOR ARSON PAST 12 MONTHS |
| CLONPDAPYU | 0 | RC-CLONAZEP AM PRODUCTS - ANY PAST YEAR USE |
| DIAZPDAPYU | 0 | RC-DIAZEPAM PRODUCTS - ANY PAST YEAR USE |
| BKROB | 0 | ARRSTD & BOOKED FOR ROBBERY PAST 12 MONTHS |
| SOMAPDAPYU | 0 | RC-SOMA PRODUCTS - ANY PAST YEAR USE |
| BKSMASLT | 0 | ARRSTD & BOOKED FOR OTHER ASSAULT PAST 12 MONTHS |