Classifying Opioid Prescription Using Machine Learning Techniques*

CAPP 30254—Machine Learning for Public Policy

Wesley Janson
wrjanson@uchicago.edu

Matt Kaufmann
mkaufmann1@uchicago.edu

Piper Kurtz
kurtzp@uchicago.edu

Angela The
angelathe@uchicago.edu

Eujene Yum
eujeneyum@uchicago.edu

May 15, 2022

*Working title

1 Data Exploration

We have pulled data from the Medical Expenditure Panel Survey (MEPS), a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. We specifically use the following three data sets:

1. Prescribed Medicine (PMEDS): This file contains information on what kind of medication a person was prescribed along with whether the prescription was due to an injury. This is where we pull information on whether a person has taken opioids or not.

2. Medical Conditions (MC): The MC file is a condition-level dataset that contains information on the medical conditions of each individual.

3. Household full-year (HC): This file contains information on demographics, health status, access to care, employment, quality of care, health insurance, and other socio-demographic information.

1.1 Data Cleaning & Basic Exploratory Statistics

Each dataset was individually cleaned and aggregated before merging. The Prescribed Medicine dataset originally contained one row for each prescription per individual (identified by DUPERSID). For our purposes, we collapsed all prescriptions by DUPERSID and created variables for total prescriptions, number of opioid prescriptions, number of non-opioid prescriptions, and an indicator variable for if there were opioids prescribed. On aggregate, $\approx$3.4% of prescriptions were for opioids and $\approx$17% of users with prescriptions were prescribed opioids.

Next, we collapsed the medical conditions dataset at the DUPERSID level, aggregating the number of medical conditions for each individual. We also added an indicator variable that specifies whether or not any conditions an individual had was due to injury. Across 2014 to 2019, we

have 143,267 observations at the individual-year level on 101,680 unique individuals (the difference likely caused by some individuals being interviewed more than once between 2014-2019). Approximately 25.3% of individuals in a given year with a medical condition had at least one condition caused by injury. Additionally, the number of conditions is skewed right with an average of 4.57 conditions per person in a year (and a median of 3).

Finally, the household full-year dataset is merged with the medical data. It contains a total of 195,810 observations from 2014 to 2019. After cleaning the dataset to retain the columns we want and dropping NAN values, there are a total of 65,834 rows and accounts for each DUPERSID in the Prescribed Medicine dataset. This dataset will provide many of the features going into our model such as demographics, whether the individual has access to health care, employment, health status, etc.

To merge the three datasets mentioned above, we created a unique ID that combines DUPERSID (unique ID in each annual data file) and YEAR. This allows us to treat each instance of a surveyed individual separately and essentially create more data points to run our algorithms on.

Our final dataset has 65,834 observations, containing 29 potential features and a column indicating whether the patient was prescribed opioids.[I]

2 Baseline Models

We have been able to estimate two baseline approaches thus far—logistic regression and a decision tree. The results from both are promising, but leave room for improvement. However, before fitting any models, our first step was to ensure relevant variables did not suffer from multicollinearity by empirically testing.

Using variance inflation factor analysis (VIF), a widely accepted off-the-shelf test for multicollinearity among variables, we identify variables that show a strong correlation with others that

[I]The exhaustive list of variables in the final dataset can be seen in Appendix A.3.

could contaminate estimation. The idea behind VIF is simple–calculate the ratio of the model variance of estimating some feature variable i in a model that includes multiple other features by the variance of a model constructed using only the i^{th} feature. This provides us with a simple representation of how much the variance of an estimated regression coefficient is increased due to collinearity. A feature variable with a VIF > 5 is considered highly collinear, and its inclusion in a model would contaminate estimation. We make an algorithm that performs VIF analysis on each potential feature variable. From there, if there is at least one variable with a VIF > 5, we eliminate it from our list of potential features. This process is repeated until no variable has a VIF exceeding 5. We found that only two variables suffered from multicollinearity, leaving us with 26 feature variables for our models.

Upon completing collinear variable elimination through VIF, we moved on to model construction. To predict if an individual will receive an opioid prescription or not we developed two separate binary classification models, a logistic regression model, and a decision tree, using the variables remaining after the VIF removal. After separating into training and testing data, using `sklearn`'s logistic regression package and the `lbfgs` solver, we found our initial linear regression model did not converge on the training set, which could lead to high variance within our model. The model did eventually converge with a large enough epoch count, but given the size of the data set and to limit run time on future, more complex models, we chose to scale our data to match a standard normal Gaussian distribution, centered at zero with unit variance ($X \sim \mathcal{N}(0,\, 1)$). Separating the data again and conducting another logistic regression on the normalized data, using the $L2$ regularization parameter, we found that our model converged and predicted opioid prescriptions on the training set with 82.2% accuracy and 82.4% accuracy on the testing set. Following this, we constructed a basic decision tree on the standardized training data and found that the test set performed with 74.3% accuracy.

3 Next Steps

We have two specific machine learning algorithms that we would like to test out next: neural networks and random forest. Looking at the results of our initial models, we are starting from a good point. Our models perform with relative accuracy, the logistic regression outperforming the decision tree by an approximate 8% margin. However, both models are valuable as we plan on developing them further and could find different results as they become more refined. For our logistic regression, our next step is to move from a single-layer network into a multi-layer network. It is unlikely that a single-layer neural network would be able to linearly separate the data on opioid prescription, so our theory is that with a 3-layer model, two hidden and one output, we should expect to find the global minimum of the loss function, finding a linearly separable network.. This model will also be necessary as we continue to reconsider variable inclusion in the model; increasing the dimensions of the attributes will necessitate a multi-layer network. We'll also institute a decaying learning rate. For our decision tree, we aim to develop a random forest model. Ensemble learning will hopefully help us increase the accuracy of our decision tree, although it will decrease interpretability. We are not yet sure of the appropriate number of trees to include in the model but will attempt to minimize it (while not compromising accuracy) to limit overfitting.

Additionally, there are a few elements we plan to add to the regressions in general. We plan on connecting additional healthcare data from MEPS, to identify if an individual's healthcare provider has any impact on their likelihood of being prescribed opioids. We also will separate the data this time into training, development, and testing, to try and optimize the accuracy of our models. Finally, we can add various graphical representations of our model results (a classification matrix, and possibly an AUROC figure) that may further help display our results.

A Appendix

A.1 Logistic Regression Calculation

Logistic regression works to fit a linear model to data with a binary response variable.

$$P(Y = 0|X') = \frac{1}{1 + exp[\beta_0 + \sum_j \beta_j x]}$$

$$P(Y = 1|X') = \frac{exp[\beta_0 + \sum_j \beta_j x]}{1 + exp[\beta_0 + \sum_j \beta_j x]}$$

Where Y is the outcome variable of whether one was prescribed an opioid and X' is a vector of j feature variables. We set out to minimize the negative log likelihood function with $L2$ regularization

$$\text{NLL}_{L_2}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left[Y_i \log \sigma(\boldsymbol{\beta}^T \mathbf{x}^{(i)}) + (1 - Y_i) \log(1 - \sigma(\boldsymbol{\beta}^T \mathbf{x}^{(i)})) \right] + \lambda \sum_{k=1}^{j} \beta_k^2$$

A.2 Decision Tree Splitting

Using the decision tree model, we split to maximize information gain (IG) at every node. This uses the concept of $entropy$, denoted as $H(X)$, which can be described as measuring the randomness of the information being processed at a node, where p_c is the fraction of examples in class c.

$$H(X) = -\sum_c p_c \; log_2(p_c)$$

$$IG(X_{p,i}) = H(X_p) - \frac{|X_{i,left}|}{|X_p|} H(X_{left}) - \frac{|X_{i,right}|}{|X_p|} H(X_{right})$$

A.3 Relevant Variables

Table 1: Variables in Final Data

Variable	Description
$OPIOID\_PRESCRIBED\_AT\_ALL^*$	Binary variable whether patient was prescribed at least one opioid.
$YEAR$	Year survey was taken.
$REGION\_YEAR$	Census Region of residency at the end of year survey is taken.
$AGELAST$	Person's age last time eligible for survey.
SEX	Sex.
$RACETHX$	Race/Ethnicity.
$MARRY\_YEARX$	Marital status at end of calendar year survey was taken.
$EDUCYR$	Years of education attained when first entered MEPS.
$BORNUSA$	Binary variable whether person was born in the US.
$FOODST\_YEAR$	Binary variable of whether anyone received food stamps in the past year.
$TTLP\_YEARX$	Person's total income.
$FAMINC\_YEAR$	Family's total income.
$POVCAT\_YEAR$	Family income as percent of poverty line - categorical.
$POVLEV\_YEAR$	Family income as percent of poverty line - continuous.
$WAGEP\_YEARX$	Person's wage income.
$DIVDP\_YEARX$	Person's dividend income.
$SALEP\_YEARX$	Person's sales income.
$PENSP\_YEARX$	Person's pension income.
$PUBP\_YEARX$	Person's public assistance.
$ADHDADDX$	Binary variable whether person is diagnosed with ADHD.
$TRIMA\_YEARX$	Covered by TRICARE/CHAMPVA in March of survey year
$MCRMA\_YEAR$	Covered by Medicare in survey year.
$MCDMA\_YEAR$	Coverage by Medicaid or SCHIP in survey year.
$ACTDTY$	Binary variable whether military full-time active duty.
$RTHLTH$	Perceived health status.
$MNHLTH$	Perceived mental health status.
$EMPST$	Employment status.
$NON\_OPIOID\_PRESCRIPTIONS$	Number of non-opioid prescriptions.
$NUM\_CONDITIONS$	Number of medical conditions.
$INJURY$	Binary variable whether any condition is a result from an injury sustained.

$*$ Variable of interest in models.