Saving Lives by Detecting Opioid Prescribers

Context

Drug overdose and opioid-involved deaths continue to increase in the United States. Opioids are powerful pain relievers that disrupt transmission of pain signals through the nervous system. According to the Centers for Disease Control, more than six out of ten overdose deaths involve an opioid. Unfortunately, the number has been in perpetual increase during the past 15 years (1), highlighting a growing addiction problem. As to avoid systematic opioid prescriptions as pain relievers, it would be interesting to identify unnecessary prescriptions, limiting therefore opioid addiction rates and overdose deaths.

Our aim behind this challenge is to use predictive modelling to identify significant opioid prescriptions. For that we will be using data (2) with prescription records for 250 common opioid and non-opioid drugs written by 25.000 professionals in 2014.

Material and Method

Presenting the data

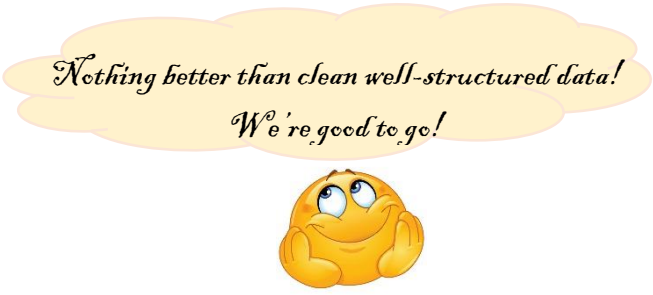
The dataset we used is a subset of a larger one, published by (2), containing 25.000 prescription instances. For each instance, we have prescriber-related characteristics like his gender, his specialty and his state of origin. The data also contains 250 numerical variables related to drug names, where for each prescriber, we indicate the number of times his prescriptions included these drugs. Instances are labeled using a binary variable that takes 1 whenever the individual prescribed opiate drugs more than 10 times in the year. The main data is in *prescriber-info.csv*. There is also *opioids.csv* that contains the names of all opioid drugs included in the data and *overdoses.csv* that contains information on opioid-related drug overdose fatalities.

Preprocessing the data

As we all know, machine learning is only as effective as the data that drives it. In other words, if you want to implement effective machine learning, you need to pay attention to data quality. For that, we had to preprocess and clean our data before starting modeling.

Preprocessing steps included:

- Modifying Opioid names-formation to match the corresponding attributes in the main data file.
- Cleaning up Categorical variables (identifying aberrant categories, correcting typos and cleaning the abbreviations)
- Factorizing categorical features (like Gender, State, Specialty)
- Eliminating opioid attributes. This step is essential in the context of our objective. Since we aim at detecting opioid prescribers, opioid-related attributes shall be removed, otherwise there would be a data leakage and we would be cheating!

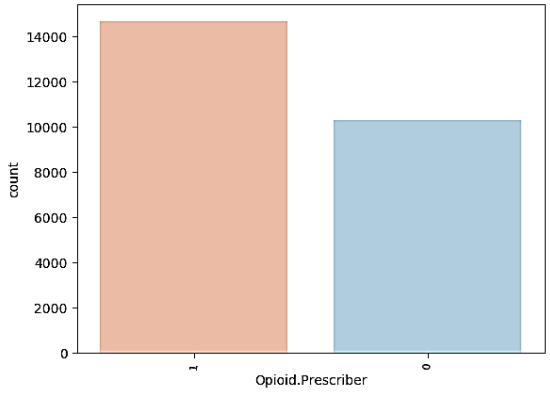


Predictive Modeling

Once our data is ready, we moved to classification. For that, we used a Random Forest Classifier with parameter tuning using 10-fold Cross-Validation. Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. It is one of the most accurate learning algorithms available. What's also good about RF is that it runs efficiently on large databases and can handle multiple input variables.

We noted that our data has balanced

To evaluate the model's performance, we are going to use the *Balanced Accuracy* score, which is the average of the correct proportions of each class individually and it is more suitable for imbalanced test set than the regular accuracy. In our case, we do not have imbalanced classes (0: 41.3% ; 1: 58.7%), thus using either metrics is Ok.



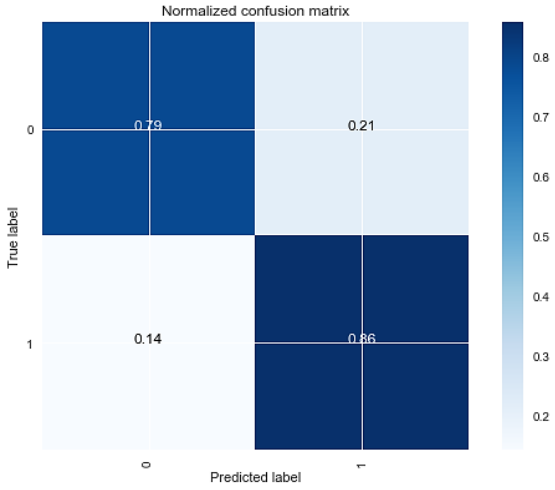
Preliminary Results

In our experiment, we fitted a Random Forest Classifier with the following optimal hyper-parameters, resulting from the 10-folk cross validation:

Hyper-parameter	Value
n_estimators	100
min_samples_leaf	10
criterion	'entropy'

For the rest of the algorithm's hyper-parameters we used the default sklearn's values.

The model resulted in the confusion matrix showed in the figure below and a balanced accuracy score of 0.82.



It seems that we did a good job, we obtained a good accuracy just by considering a doctor's specialty and non-opioid prescription trends, allowing us to predict how likely it is that he prescribes a significant quantity of opiates.

As we previously mentioned, one of the pros to use Random Forests is the ability to assess features importance. We noted that some features had zero and close-to-zero importance (like ABILIFY and ACYCLOVIR). Thus, possible improvements of our model can be performed through paying more attention to features selected for the modeling.

We shall also note that, within the context of detecting significant opioid prescribers, it would be wiser to care more about precision than accuracy, since a false positive would lead to harm "innocent" professionals. In our case, our precision was 0.8, but we can surely do better!