Team 7

Zachary Qian, Renzo Tanaka-Wong, Kyle Icban, Manu Kumar

Demographic Demarcations of 'Misuse' of Non-Prescription Substances in the US

The 2019 United States surveys provided by the Rocky Mountain Poison and Drug Safety center detailed demographic information in tandem with drug usage patterns of the 30,000 individuals that participated. Team 7's primary goals were to identify significant demographic factors that could predict an individual's potential to misuse a non-prescription substance in the United States, as well as assess the efficacy of these factors.

The definition of an umbrella term as generic as misuse is a long winded discussion covering policy and morality. This is clearly not in the scope of a 48 hour competition to discuss, hence we rely on the government's existing categorizations of substances to find a sufficient jumping off point. The United States DEA describes **Schedule 1** drugs as substances that either have: high potential for abuse, no currently accepted medical treatment use or a lack of accepted safety for use under medical supervision. We decided to settle on the usage of Schedule 1 drugs in an individual's lifetime as the definition of misuse for our model. This decision was partly taken in order to use as much filled data for our target variable and avoid significant sparsity (ie. low case/control ratio).

Using a dataset provided by the DEA, we identified schedule 1 drugs listed in the survey and the associated tickers to look out for. From the dataset of 2019 US surveys, we created a feature matrix of demographic information and a target variable vector denoting usage of concerned drug tickers. These secondary data were used to make a binary classifier to determine whether an individual has, in their lifetime, taken schedule 1 drug or not. Because much of the raw dataset's features were binary flags, we collapsed certain features together in order to reduce the number of trivial features. For example, we created a categorical "Race" feature that encompassed all six of the binary race categories (White, Black, Asian, American Indian, Pacific Islander, Other) into a single feature column.

We implemented three classical machine learning models: logistic regression, decision trees, and extreme gradient boosted trees. After using gridsearch and k-fold Cross Validation for hyperparameter tuning, we achieved an 83% accuracy and an AUC score 0.73. We implemented recursive feature elimination, as well as manually taking the top features from our model and constructed another logistic regression model with those seven features again. Based on our model, we concluded that our most important covariates were **gender, age group, handicaps due to medical conditions, current cigarette smoking habits, employment in the healthcare industry, possession of private health insurance, and overnight hospital visits in the last year**. We achieved similar accuracy and AUC score (83% and 0.72, respectively). For our decision tree model, we used a max depth of four and achieved accuracy of 83% and an AUC of 0.71. Lastly, our XGBoost model achieved the highest accuracy of 84% and AUC of 0.75

It is evident that in comparison to social and economic demographics, there is a much more rigorous correlation of drug misuse with conditions that demarcate the overall health and constitution of an individual. The strongest observed pattern is a negative correlation associated with tobacco usage, indicative of previous gateway drug addiction due to how the TOB_LIFE category is defined.