DS-GA 1001 Introduction to Data Science - Drug Killer

Ling Ho (lh1663), Ria Pinjani (rp3084),

Shizhan Gong (sg5722), Yakun Wang (yw3918)

December 8, 2018

1 Business Understanding

Cigarette smoking has been experiencing a sharp drop amongst American teenagers, however vaping and marijuana use are now more common, according to a national survey of adolescent drug use. The 2017 report, sponsored by the federal governments National Institute on Drug Abuse and administered by the University of Michigan, shows that nearly 24 percent of students in all of the 43,703 surveyed high school students said they have used marijuana over the past year, a rate that has stayed relatively stable in recent years (NYTimes). Over the past two decades, there are significant studies which establishes links between cannabis use and first episode psychotic outcomes including other related mental and physical deterioration. Twenty-two percent of U.S. states have already legalized or are in the process of legalizing cannabis. The increasing prevalence leads to an increase in the percentage of tetrahydrocannabinol (THC), the most active ingredient in cannabis, making the drug more potent.

However, the perception that marijuana is not dangerous has been driven in part by society. The fear is that we may be seeing a start of a long-term increase in marijuana use among youth. The machine learning problem we propose is to detect marijuana usage on youth. In this study, we aim to utilize pattern detection and develop predictive models for cannabis usage with features including cannabis usage, household and academic environment, physical and mental health history and the perception of the drug (Appendix, Table 2).

By constructing such a model with cannabis usage of people under the age of 18 as the target variable and these individual chosen attributes as predictors, we can predict marijuana usage. We will be analyzing the data to find correlations between the predictor variables and the target variables in the training data.

Precisely, this model can be used by government institutions, schools, childrens hospitals, and medical research facilities. If a school with a large student population is able to predict which of their students are more likely to start using marijuana, they can focus their resources on adequate information, guidance and counseling services to this group to act as preventive measures. This is beneficial as school resources can be utilized in an efficient and effective manner.

The public health department can determine some key factors that can lead to cannabis usage, which can be alleviated from the source found from our study. Medical research facilities can use the study to enhance their research on psychotic disorders and preventions, using the knowledge about underlying issues and variables that trigger cannabis use.

2 Data Understanding: Data Collection and Extracted Features

We downloaded the datasets from University of Michigans ICPSR data repository. Our data is pulled from the National Survey on Drug Use and Health, 2012. The survey measures prevalence and correlations of drug use in each of the 50 U.S. states. There is 3120 variables in the original data source, but we have chosen to focus on 31 variables. We chose the 31 variables from each sub-category: mental health, personality, depression, academic and home environments in order to provide a representative features from each subgroup in the original dataset. The original data source also included variables for adults so we left them out to focus on youth-relevant variables such as schooling. We chose variables that were easily attainable with high response rate and non-personal. Since this is survey based data, it is highly likely that participants can game the questions or be subject to voluntary response bias. We can make it non-invasive by avoiding bias and loaded words and by starting with broad general questions and progress to specific and harder ones. We did not include variables like times moved in the past 5 years or importance of friends sharing the same religious beliefs, because it does not share relevance in the prediction of marijuana usage.

We focus on the age group 12-17 year olds to utilize these findings for preventative purposes amongst adolescence. We focus on youth to gear our study on the reasons people start using the drug at a young age in order to predict implications for the future.

3 Data Preparation

3.1 Target Variables and Feature Engineering

The target variable is the binary variable - marijuana usage. The data instance in our baseline problem is the respondents who have answered yes or not out of the 17,385 respondents in the study and their response to the corresponding variables we have chosen. The base rate (from the target variable) of respondents that have used marijuana versus those who havent is in the ratio: 3102:14283. We accounted for this imbalance as shown later in the report. Here are the initial transformations we performed:

- Subsetting our data by including respondents aged between the ages of twelve to seventeen.
- Missing values: random missing responses. There are cases where people leave the question blank or choose I dont know. We regard these kind of missing values as random missing. To account for random missing values, we replaced them with mean values for all continuous variables and mode values for all discrete variables.
- Missing values: legitimate skips. This happens because people may answer differently to other questions, thus certain questions dont apply to certain groups of people. These groups of people may share some hidden common characteristics which may affect the target variable. Therefore, for questions that are answered with legitimate skips, we conducted chi-square tests to decide whether the skipped answers should be filled with mean/mode values or should be treated as another feature. If the test result indicates that there is significant distributional difference of the target variable between the skipping group and the common group, we use a dummy variable to denote whether this value is missing or not. We can replace legitimate skip with the mode or mean of the respective variable only in the case when there are no significant results.
- Lastly, we used the train test split function from sklearn to split our dataset randomly into training and testing data sets, without overlap. This is used for training the model, doing cross-validation, and

testing the dataset used for final performance. We used 30 percent of our data set as the test data while the rest was used for training.

3.2 Descriptive Statistics

Two of the interesting descriptive statistics are shown here. The distributions of features are compared between respondents who have used Marijuana and those who have never used it. Figure 1(Left) shows the distribution of the variable YEPMJEVR (how parents would feel you trying Marijuana) amongst 12-17 year olds compared against the MJEVER (Marijuana use) target variable. We can see that despite whether you are using marijuana or not, parents would tend to disapprove of the drug. However, although the group size of Marijuana users is much smaller, the number of those with somewhat disapproval or neutral attitudes from parents are higher than that for samples that have never used Marijuana.

Figure 1(Right) shows the distribution of ages in the sample between those who use marijuana and those who dont. Out of the people that use marijuana, the older you get, the more likely you will start using. This fact indicates that youth are allured to Marijuana use continously as they grow up. These differences of distributions make our prediction possible and also help to explain the model later.

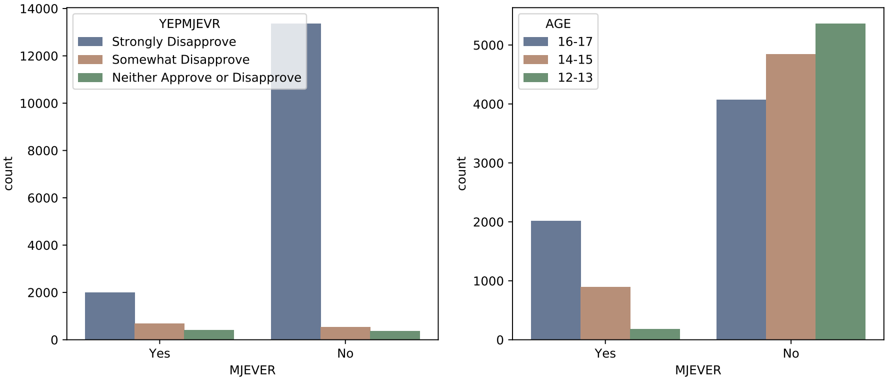


Figure 1: Histogram of MJEVER (Target Variable/Marijuana usage) plotted against YEPMJEVR(how parents would feel you trying Marijuana) (Left) & Histogram of MJEVER (Target Variable/Marijuana usage) plotted against distribution of Age (Right)

4 Modelling

4.1 Evaluation Framework

A proper evaluation framework should be decided before we started to build models and make improvements. We selected AUC as our major criterion. This is because AUC is insensitive to the base rate, which is good match to the uneven nature of our target variable. AUC is also able to evaluate the classifier comprehensively. Moreover, since our final goal is to detect the potential drug user and try to prevent them, we will stress high importance on the situation with high true positive rate (TPR). Therefore, we draw the ROC curve of each model as well. We also list the F1 score and accuracy of each algorithm for reference.

Due to the limited volume of our data, all criteria are calculated as mean value of validation set in a 5-fold cross validation in order to get stable evaluations. It should also be noted that Stratified KFold cross-validation, which can preserve the percentage of samples for each class, is used here rather than KFold cross-validation, since in an imbalanced learning problem, the original KFold might split all samples in the minority class to one single set, leading to the bias of training data and harming the performance of models.

4.2 Baseline Model

A Complement Naive Bayes model is used as our baseline here. Since our data is highly discrete with most of the variables binary, Naive Bayes is considered to be suitable. Complement Naive Bayes classifiers, instead of Multinomial Naive Bayes, is chosen because of its adjustment for imbalanced data sets. Moreover, parameter estimates for them are proved to be more stable than those for Multinomial Naive Bayes classifiers. (Rennie, Jason D. M., et al). We got an average AUC equal to 0.768, f1 score equal to 0.402 and accuracy equal to 0.649. The ROC curve is shown in Figure 2.

4.3 Improving on our Baseline Model

The ROC curve of the baseline model is not concave. We attributed that to the strong assumption of the independence of features in the Naive Bayes model. Therefore, models with more tolerance and further feature selections should be carried out. To account for the uneven nature of our target variable, for each of the training sets split by the Stratified KFold cross-validators, random oversampling was performed first to make the base rate of the training samples even or 1:1 before the modeling process. Note that the base rate is kept for the validation sets.

In general, the algorithms we used are divided into two groups: linear models and tree-based models. Linear models include logistic regression and SVM, which can give us a intuitive linear and nonlinear decision boundaries. They are usually sensitive to the input features. Therefore, we conducted feature selection through chi-square tests and t-tests. Only features highly correlated with the target variable will be included in the models. Since all the features has to be numerical in linear models, we transformed all the categorical data into dummy variables and normalized on all the features. Tree-based models can catch the complexity of the data distribution. These models can detect the importance of each feature and discard useless ones automatically. Moreover, tree based models can address categorical variables as well as numerical variables with different scales, therefore we put all the attributes in their original form in the calculation. Hyperparameter tuning is conducted through grid search and stratified 5-fold cross validation based on the AUC criterion, as described before.

4.4 Choice of Algorithms

4.4.1 Logistic Regression

Logistic regression is an algorithm suitable for a small dataset and is able to output a probabilistic interpretation. Moreover, after normalizing all the attributes, the scale of the weight can directly reflect the importance of each feature. To avoid overfitting, we tuned the hyperparameters related to the penalty terms. We used l1 and l2 penalty with varying regularization strength. For regularization strength, we initially searched through a wide rough range and gradually narrowed our search space, until we found the optimal value.

4.4.2 Support Vector Classifier

Support vector classifiers (SVC) have proven to have high accuracy even in highly complicated problems. It can give a nonlinear decision boundary through the kernel trick. It usually scales well to high dimensional data and faces less risk of overfitting. The challenges of SVC is finding a proper kernel as well as a suitable regularization term. Again we carried out grid search to find the optimal kernel and regularization strength term.

4.4.3 Random Forest

Next we decided to move on to tree models. A tree model is able to fit arbitrary complicated decision boundaries. It is insensitive to the form of the feature which minimizes the necessity of feature engineering. Moreover, tree based models can also give us feedback on feature importance. However, a single tree is very likely to overfit the data. Therefore, we choose to use ensemble methods to build several trees together. For tree models (Random Forest and Boosting Models), we were most concerned about the number of estimators, the learning rate, and parameters controlling the complexity and size of the trees. Since the number of estimators and learning rate are highly correlated, we fixed the learning rate to a reasonable value, 0.01, and tuned for the best number of estimators. Meanwhile, tree-related parameters, including the maximum depth and minimum samples split, were also tuned. We started off with a bagging algorithm, Random Forest.

4.4.4 Extreme Gradient Boosting

Next we decided to use boosting algorithms; XGBoost and GBDT. XGBoost is a popular implementation of gradient boosting. XGBoost incorporates a sparsity-aware split finding algorithm to handle different types of sparsity patterns in the data.

4.4.5 Gradient Boosting Decision Tree

GBDT has built-in mechanisms to figure out how to split categorical features and place missing values in the trees. It also features high efficiency, low memory footprint and collections of loss function.

Table 1: Results

Classifier	AUC	F1 Score	Accuracy
Logistic Regression	0.8654	0.5707	0.7975
SVC	0.8647	0.5725	0.8009
Naive Bayes	0.7680	0.4025	0.6495
Random Forest	0.8732	0.5833	0.8069
XGBoost	0.8766	0.5877	0.8108
GBDT	0.8759	0.5893	0.8128

5 Evaluation

5.1 Results

The result of each algorithm is shown in Table 1. If we argue that AUC is a good determinant for selecting an appropriate model, we could conclude that most classifiers perform equally well on our data and XGBoost is the best classifier. However as we see from the table below, XGBoost is only slightly better than the Logistic Regression Model. XGBoost is kind of like a black box model while logistic regression has an intuitive statistical explanation. Considering interpretability and model complexity being important criteria; we choose logistic regression as our final model. Finally, we retrain our model on the whole training set with the best hyper parameter (l1 penalty, C=1.2), and evaluate our final model on the test set, which give us AUC of 0.8661, f1 score of 0.5897 and accuracy of 0.7931. The confusion matrix of the test set is also shown in picture Figure 3.

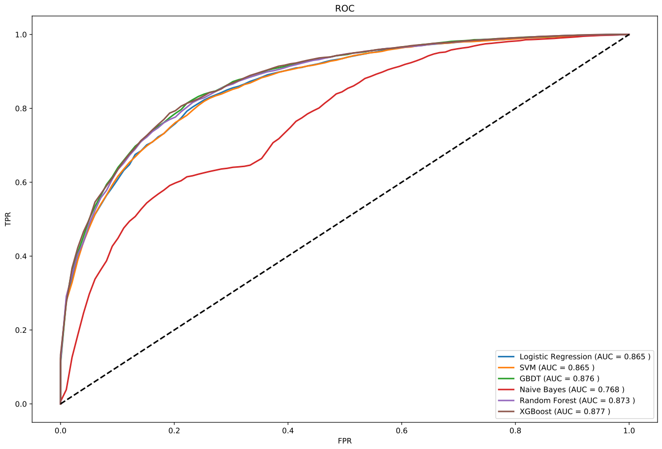


Figure 2: ROC Curves

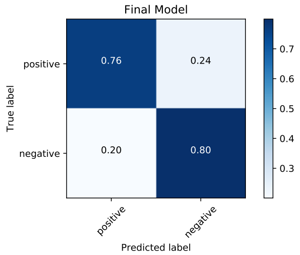


Figure 3: Confusion Matrix

5.2 Survey Reliability

The survey shows a high percentage of response. Even when there is an option to skip certain questions, 99.91 percent respondents have answered either yes or no for the target variable . It should be noted that positive individuals (answered yes) may have lied about their marijuana usage due to social pressure, however, respondents that have responded no to marijuana usage are more likely not to have gamed the survey because there is no incentive or reason for lying. Thus, respondents who have responded yes to marijuana usage in the survey will have higher reliability than those who have responded otherwise because there is no reason for he or she to have lied. Since the dataset is highly imbalanced and negative data is sufficient, the impurity of negative samples is treated as a minor problem.

5.3 Inference

One of the advantages of logistic regression is that it can give an explicit expression of the classification model, which is a function of the linear combinations of all the explanatory variables. Since we have normalized all the features, the weight of logistic regression can represent the importance of each feature directly. Though correlation between explanatory variables may have some confusing effects on the weight, we can safely conclude that at least the direction of effects would be the same for the most important features. In Figure 4, we plot the weight of each of the 31 chosen attributes in a bar plot. From the plot we can conclude that YEPMJEVR, YMDELT ,CATAG7 and YELSTGRD-4 are four of the most significant factors that affect marijuana use, which leads to the following findings:

- The more parents care about their childrens marijuana use, the less likely children will start use marijuana. Schools should communicate with parents if their children show signs of drug use, academic disability or any sort of odd behavior.

- Depression is one of the key reasons for initial marijuana use. Social programs and schools should cater more resources to the mental health of teenagers especially at the adolescence stage. Counseling, mental-health programs, anti-bullying campaigns are examples of effective solutions.

- Older teenagers are more likely to use marijuana than younger teenagers. As expected, older teenagers learn more about this recreational drug through media and society, which does not do a good job of portraying the danger of the drug, making it more prevalent. As of 2017, it is the third most popular drug in America behind tobacco and alcohol. Again, schools and parents have the responsibility to educate teenagers about the risks of drug abuse in general and the detriment of peer pressure and following media trends.

- Teenagers ages 12-17 using marijuana are more likely to have a bad academic performance. Though the levels of marijuana intake and reactions can vary from person to person, marijuana is tied to poorer school performance because students are more likely to exhibit reduced neural connectivity in regions responsible for memory, learning and inhibitions. Thus, educating youth about these physical and mental consequences are important.

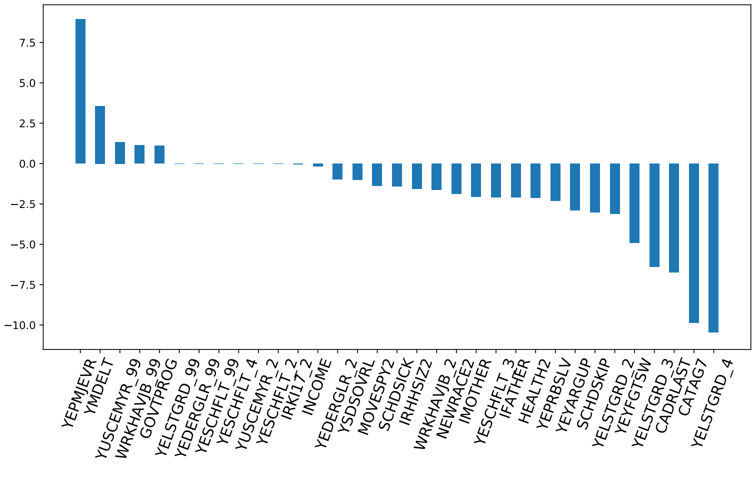
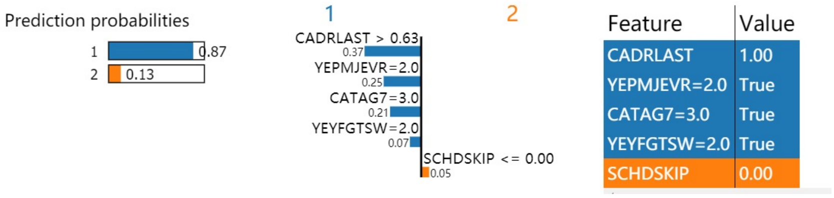


Figure 4: Feature Importance

6 Deployment



Feature	Value
CADRLAST	1.00
YEPMJEVR=2.0	True
CATAG7=3.0	True
YEYFGTSW=2.0	True
SCHDSKIP	0.00

Figure 5: LIME Package Illustration

We can use the LIME (Locally Interpretable Model-Agnostic Explanations) package to provide local model interpretability (Brown, 2018). In the Figure 5 below we are explaining individual predictions for the logistic regression classifier using the LIME package.

The figure above tells us that the test sets ith prediction is 0.87. SCHDSKIP (How many days missed school from skipping) provides us with the most negative valuation and the other features provide a positive valuation in the prediction.

By providing local model interpretability, we can enable schools and policy makers to learn about the factors that encourage certain individuals to engage in marijuana usage and make decisions. For instance, a school could look at all individuals with prediction probabilities greater than 0.75. By looking at these individuals and the features that provided positive valuation to marijuana usage, they could make suitable decisions about mental health programs, counseling, etc.

We can also give students an annual survey with the same questions and conduct this on the national level. We may find that certain school or geographic communities have higher rate of potential marijuana usage in youth. Thus, this helps us target specific groups of people to instigate an intervention. Feedback from schools, communities, mental health groups, associations, and parents on a bi-annually would be useful. Some things to consider before deploying this particular finding for public use is to acknowledge that not all marijuana usage is detrimental. In other words, not all marijuana smoking, vaping or consuming should necessarily be brought to attention. This could lead to overstating the seriousness of the issue. One risk is schools may totally ban marijuana use and consumption which could backfire and encourage students to use the drug more. Since marijuana is the third most prevalent drug in the U.S., the responsibility of adults and learning environments is to communicate the risks and drug perception clearly.

Appendices

A Variable Definition

Table 2: Variable Definitions

Variable Code	Corresponding Definition
MJEVER	Ever used Marijuana
CADRLAST’	Number of drinks the last time drank in past 30 days
IRHHSIZ2	Number of persons in household
IRKI17_2	Number of kids aged < 18 in household
IRHH65_2	Number of persons in household aged > 65
IMOTHER	Mother in household
IFATHER	Father in household
GOVTPROG	Participated in one or more government assist programs
INCOME	Total family income record
YMDELT	Lifetime major depression episode
YSDSOVRL	Max severity level of MDE role impairment
ANYHLTI2	Covered by any health insurance
CATAG7	Age category recode (7 levels)
HEALTH2	Overall health recode
IRSEX	Imputation revised gender
NEWRACE2	Race/hispanicity recode (7 levels)
MOVESPY2	Number of times moved past 12 months
SCHENRL	Now enrolled in any school
SCHDSICK	How many days missed school from sick
SCHDSKIP	How many days missed school from skipping
WRKHAVJB	Did you have a job or business
YESCHFLT	How you felt about going to school
YELSTGRD	Grades for last semester
YEYFGTSW	Gotten into a serious fight at school/work
YEYARGUP	Argued/had a fight with at least one of your parents
YEPMJEVR	How parents would feel about youth trying Marijuana
YEPRBSLV	Problem solving/communication skills/self esteem
YEDERGLR	Films/lectures/discus/info about drg/alc in class
YUMHCRYR	During the past 12 months, did you receive treatment or counseling at a mental health clinic or center your behavior or emotions that were not caused by alcohol or drugs?
YUTPSTYR	During the past 12 months, did you receive treatment or counseling from a private therapist, psychologist, psychiatrist, social worker, or counselor for emotional or behavioral problems that were not caused by alcohol or drugs?
YUSCEMYR	At any time during the past 12 months, did you attend a school for students with emotional or problems?

B Github Link

https://github.com/670973787/marijuana-usage-detection

C Contribution

We were able to research and come up with a machine learning project all together and separated the work between all of us in the following way:

- Yakun Wang and Shizan Gong: Scrape the data, first clean-up of the data, set up of the machine learning framework (metric choice, cross- validation), model selection and hyper- parameter tuning, visualization, project write up
- Ling Ho and Ria Pinjani: Initial research, business understanding and project write up

References

[1] Rennie, Jason D. M., et al. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. people.csail.mit.edu/jrennie/papers

[2] Brown, Eric. Local Interpretable Model-Agnostic Explanations - LIME in Python. Python Data, 11 June 2018, pythondata.com/local-interpretable-model-agnostic-explanations-lime-python/.

[3] Marijuana Use Tied to Poorer School Performance. ScienceDaily, ScienceDaily, 11 May 2017, www.sciencedaily.com/releases/2017/05/170511083745.htm.

[4] Hoffman, Jan. Marijuana and Vaping Are More Popular Than Cigarettes Among Teenagers. The New York Times, The New York Times, 14 Dec. 2017, www.nytimes.com/2017/12/14/health/teen-drug-smoking.html.