American Research Journal of Humanities and Social Sciences
ISSN (Online) : 2378-7031
Volume 4, Issue 1, 12 pages

AMERICAN RESEARCH JOURNALS
An Academic Publishing House

Research Article Open Access

Predicting Risk of Drug Use for High School Students Using Artificial Neural Network

Beichen Wang

The Wardlaw-Hartridge School, USA.
1455193088@qq.com

Abstract

Objective: This study aims to 1) examine the predictors of drug use at high school 2) build a predictive model for drug use using artificial neural network and compare its performance to logistic regression model.

Methods: Youth Risk Behavior Surveillance System (YRBSS) 2015 data were used for this study. The YRBSS was developed in 1990 to monitor priority health risk behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States.

All the participants who were eligible were randomly assigned into 2 groups: training sample and testing sample. Two models were built using training sample: artificial neural network and logistic regression. We used these two models to predict the risk of Drug Usein the testing sample. Receiver operating characteristic (ROC) were calculated and compared for these two models for their discrimination capability and a curve using predicted probability versus observed probability were plotted to demonstrate the calibration measure for these two models.

Results: About 18.1% of 8711 students were drug users, about 19.1% among the female and 17.1% among the male.

According to the logistic regression, students who had rides in a car driven by someone who is drinking were more likely to have drug use. Students who never tried cigarette smoking were less likely to use drug. Students who drank often were more likely to use drug. Student who used marijuana often were more likely to use drug. Heterosexual students were less likely to use drug. Students who slept 4 hours or less daily were more likely to use drug. Students who did not speak English well were less likely to be a drug user.

According to this neural network, the top 5 most important predictors were 'being black', Q99 (How well do you speak English), 'being Asian', Q68 (sexual orientation), Q88 (On an average school night, how many hours of sleep do you get?).

For training sample, the ROC was 0.84 for the Logistic regression and 0.88 for the artificial neural network. Artificial neural network performed better clearly. In testing sample, the ROC was 0.83 for the Logistic regression and 0.80 for the artificial neural network. Artificial neural network had worse performance.

As to calibration measure, predictions made by the neural network are (in general) less concentrated around the 45-degree line (a perfect alignment with the line would indicate an ideal perfect calibration) than those made by the Logistic model.

Conclusions: In this study, we identified several important predictors for drug use e.g., cigarette smoking, drinking, sexual orientation. This provided important information for educators as well as parents provide timely intervention. We built a predictive model using artificial neural network as well as logistic regression to provide a tool for early detection. As to performance of these two models, logistic regression and neural network had a similar discriminating capability.

Instruction

In 2014, 44.1 million Americans reported using illicit drugs over the past year. One tragic result of the widespread use of drugs and alcohol is its impact on youth[1]. In the United States alone, 7% of youth aged 12–13 took an illicit substance in the past year, while 5.6% reported drinking alcohol. Early use of drugs or alcohol has been linked to a several times greater risk of developing substance dependence, as the majority of Americans aged 18–30 admitted for substance abuse treatment initiated alcohol or drug use before the age of 18.

Monitoring the Future (MTF) survey of drug use and attitudes among American 8th, 10th, and 12th graders recently found that past-year use of illicit drugs other than marijuana continuing to decline to the lowest level in the history of the survey in all three grades—5.4 percent among 8th graders, 9.8 percent among 10th graders, and 14.3 percent among 12th graders. This is down from peak rates of 12.6 percent for 8th graders in 1995, and 18.4 percent for 10th graders in 1996, and 21.6 percent for 12th graders in 2001.[2]

This study aims to 1) examine the predictors of drug use at high school 2) build a predictive model for drug use using artificial neural network and compare its performance to logistic regression model.

Data and Methods

Data

Youth Risk Behavior Surveillance System (YRBSS) 2015 data were used for this study.

The YRBSS was developed in 1990 to monitor priority health risk behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States. These behaviors, often established during childhood and early adolescence, include

1) Behaviors that contribute to unintentional injuries and violence.

2) Sexual behaviors related to unintended pregnancy and sexually transmitted infections, including HIV infection.

3) Alcohol and other drug use.

4) Tobacco use.

5) Unhealthy dietary behaviors.

6) Inadequate physical activity.

In addition, the YRBSS monitors the prevalence of obesity and asthma and other priority health-related behaviors plus sexual identity and sex of sexual contacts. From 1991 through 2015, the YRBSS has collected data from more than 3.8 million high school students in more than 1,700 separate surveys.

Models

Artificial neural netwrok consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases is that there is actual anipulation and cross-fertilization of the data helping users makes more informed decisions.

A package called "neuralnet" in R was used to conduct neural network analysis. The package neuralnet focuses on multi-layer perceptrons (MLP, Bishop, 1995), which are well applicable when modeling functional relationships. The underlying structure of an MLP is a directed graph, i.e. it consists of vertices and directed edges, in this context called neurons and synapses. The neurons are organized in layers, which are usually fully connected by synapses. In neuralnet, a synapse can only connect to subsequent layers. The input layer consists of all covariates in separate neurons and the output layer consists of the response variables. The layers in between are referred to as hidden layers, as they are not directly observable. Input layer and hidden layers include a constant neuron relating to intercept synapses, i.e. synapses that are not directly influenced by any covariate ^{o} Neural networks are fitted to the data by learning algorithms during a training process. Neuralnet focuses on supervised learning algorithms.

The backward propagation of errors or backpropagation, is a common method of training artificial neural networks and used in conjunction with an optimization method such as gradient descent. The algorithm repeats a two phase cycle, propagation and weight update. When an input vector is presented to the network, it is propagated forward through the network, layer by layer, until it reaches the output layer. The output of the network is then compared to the desired output, using a loss function, and an error value is calculated for each of the neurons in the output layer. The error values are then propagated backwards, starting from the output, until each neuron has an associated error value which roughly represents its contribution to the original output.

We also used logistic regression models to calculate the predicted risk. Logistic regression is a part of a category of statistical models called generalized linear models, and it allows one to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a combination of these. Typically, the dependent variable is dichotomous and the independent variables are either categorical or continuous.

The logistic regression model can be expressed with the formula:

$\ln(P/P-1) = \beta0 + \beta1*X1 + \beta2*X2 + ...+ \beta n*Xn$

Model Evaluation

The two criteria to assess the quality of a classification model are discrimination and calibration . Discrimination is a measure of how well the two classes in the data set are separated; calibration determines how accurate the model probability estimated is to the true probability. To provide an unbiased estimate of a model's discrimination and calibration, these values have to be calculated from a data set not used in the model building process. Usually, a portion of the original data set, called the test or validation set, is put aside for this purpose. In small data sets, there may not be enough data items for both training and testing. In this case, the whole data set is divided into n pieces, n_1 pieces are used for training, and the last piece is the test set. This process of n-fold cross-validation builds n models; the numbers reported are the averages over all n test sets. An alternative to cross-validation is bootstrapping, a process by which training sets are sampled with replacement from the original data sets.

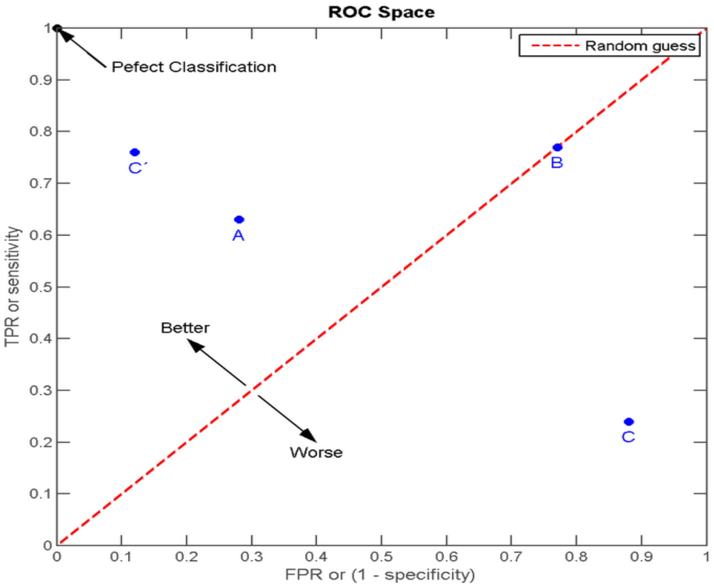
The discriminatory ability – the capacity of the model to separate cases from non-cases, with 1.0 and 0.5 meaning perfect and random discrimination, respectively– was determined using receiver operating characteristic (ROC) curve analysis. ROC curves are commonly used to summarize the diagnostic accuracy of risk models and to assess the improvements made to such models that are gained from adding other risk factors. Sensitivity, specificity, and accuracy will be also calculated and compared. For all these measures, there exist statistical tests to determine whether one model exceeds another in discrimination ability.

The contingency table can derive several evaluation "metrics" (see infobox). To draw a ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter). The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to 1 − specificity, the ROC graph is sometimes called the sensitivity vs (1 − specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space.

The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a perfect classification. A random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners (regardless of the positive and negative base rates). An intuitive example of random guessing is a decision by flipping coins. As the size of the sample increases, a random classifier's ROC point migrates towards the diagonal line. In the case of a balanced coin, it will migrate to the point (0.5, 0.5).

The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random), points below the line represent poor results (worse than random). Note that the output of a consistently poor predictor could simply be inverted to obtain a good predictor.



Calibration is a measure of how close the predictions of a given model are to the real underlying probability. Almost always, the true underlying probability is unknown and can only be estimated retrospectively by verifying the true binary outcome of the data being studied. Calibration thus measures the similarity between two different estimates of a probability. One of the ways to assess calibration is to take the difference between the average observation and the average outcome of a given group as a measure of discalibration. A more refined way to measure calibration requires dividing the sample into smaller groups sorted by predictions, calculating the sum of predictions and sum of outcomes for each group, and determining whether there are any statistically significant differences between the expected and observed numbers by a simple method.

Variables

The outcome variable isbased on Q57 (During your life, how many times have you taken a prescription drug (such as OxyContin, Percocet, Vicodin, codeine, Adderall, Ritalin, or Xanax) without a doctor's prescription) and Q58 (During your life, how many times have you used a needle to inject any illegal drug into your body). If the particpatent used any of above >=1 times, he or she was considered a drug user in this study.

Table1. *Variables used in this study*

Q1. How old are you? A. 12 years old or younger B. 13 years old C. 14 years old D. 15 years old E. 16 years old F. 17 years old G. 18 years old or older
Q2. What is your sex? A. Female B. Male
Q3. In what grade are you? A. 9th grade B. 10th grade C. 11th grade D. 12th grade E. Ungraded or other grade
Q4. Are you Hispanic or Latino? A. Yes B. No
Q5. What is your race? (Select one or more responses.) A. American Indian or Alaska Native B. Asian C. Black or African American D. Native Hawaiian or Other Pacific Islander E. White
Q6. How tall are you without your shoes on?
Q7. How much do you weigh without your shoes on?
Q10. During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had beendrinking alcohol? A. 0 times B. 1 time C. 2 or 3 times D. 4 or 5 times E. 6 or more times
Q31. Have you ever tried cigarette smoking, even one or two puffs? A. Yes B. No
Q41. During your life, on how many days have you had at least one drink of alcohol? A. 0 days B. 1 or 2 days C. 3 to 9 days D. 10 to 19 days E. 20 to 39 days F. 40 to 99 days G. 100 or more days

Q47. During your life, how many times have you used marijuana?
A. 0 times
B. 1 or 2 times
C. 3 to 9 times
D. 10 to 19 times
E. 20 to 39 times
F. 40 to 99 times
G. 100 or more times

Q60. Have you ever had sexual intercourse?
A. Yes
B. No

Q68. Which of the following best describes you?
A. Heterosexual (straight)
B. Gay or lesbian
C. Bisexual
D. Not sure

Q80. During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?
(Add up all the time you spent in any kind of physical activity that increased your heart rate and made you breathe hard some of the time.)
A. 0 days
B. 1 day
C. 2 days
D. 3 days
E. 4 days
F. 5 days
G. 6 days
H. 7 days

Q82. On an average school day, how many hours do you play video or computer games or use a computer for something that is not school work? (Count time spent on things such as Xbox, PlayStation, an iPod, an iPad or other tablet, a smartphone, YouTube, Facebook or other social networking tools, and the Internet.)
A. I do not play video or computer games or use a computer for something that is not school work
B. Less than 1 hour per day
C. 1 hour per day
D. 2 hours per day
E. 3 hours per day
F. 4 hours per day
G. 5 or more hours per day

Q84. During the past 12 months, on how many sports teams did you play? (Count any teams run by your school or community groups.)
A. 0 teams
B. 1 team
C. 2 teams
D. 3 or more teams

Q85. Have you ever been tested for HIV, the virus that causes AIDS? (Do not count tests done if you donated blood.)

A. Yes

B. No

C. Not sure

Q88. On an average school night, how many hours of sleep do you get?

A. 4 or less hours

B. 5 hours

C. 6 hours

D. 7 hours

E. 8 hours

F. 9 hours

G. 10 or more hours

Q89. During the past 12 months, how would you describe your grades in school?

A. Mostly A's

B. Mostly B's

C. Mostly C's

D. Mostly D's

E. Mostly F's

F. None of these grades

G. Not sure

Q99. How well do you speak English?

A. Very well

B. Well

C. Not well

D. Not at all

Q11. During the past 30 days, how many times did you drive a car or other vehicle when you had been drinkingalcohol?

A. I did not drive a car or other vehicle during the past 30 days

B. 0 times

C. 1 time

D. 2 or 3 times

E. 4 or 5 times

F. 6 or more times

Results

About 18.1% of 8711 students were drug users, about 19.1% among the female and 17.1% among the male.

Basically, a corrgram is a graphical representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes using visual thinning and correlation-based variable ordering. Moreover, the cells of the matrix can be shaded or colored to show the correlation value. The positive correlations are shown in blue, while the negative correlations are shown in red; the darker the hue, the greater the magnitude of the correlation.

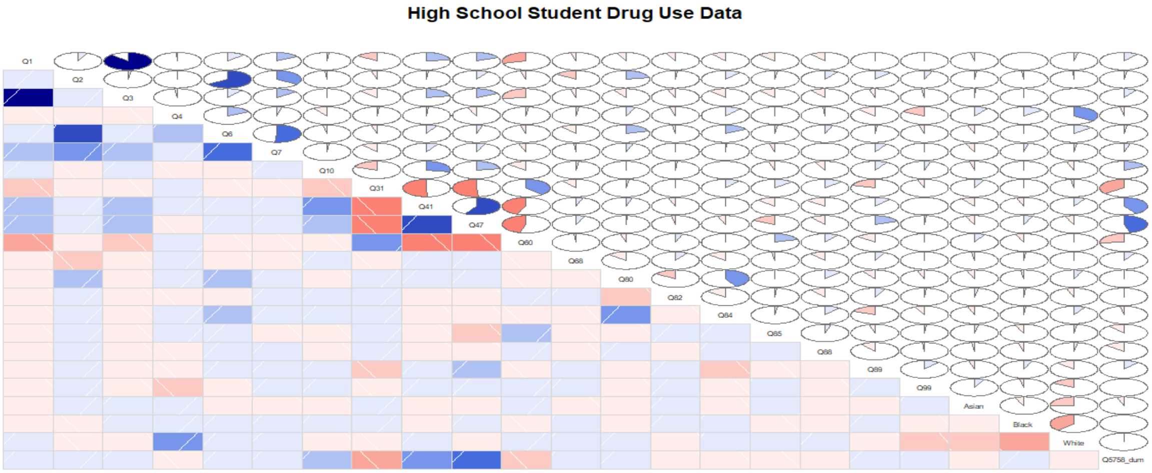


Fig1. *matrix of correlationsbetween variables*

According to the logistic regression, students who had rides in a car driven by someone who is drinking were more likely to have drug use. Students who never tried cigarette smoking were less likely to use drug. Students who drank often were more likely to use drug. Student who used marijuana often were more likely to use drug. Heterosexual students were less likely to use drug. Students who slept 4 hours or less daily were more likely to use drug. Students who did not speak English well were less likely to be a drug user.

Table 2. *Logistic Regression for Drug Useamong High School Students*

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.0467	0.80587	-2.54	0.01109	*
Q1	-0.00811	0.05967	-0.14	0.89191	
factor(Q2)2	0.01123	0.09701	0.12	0.90785	
factor(Q3)2	-0.11531	0.11808	-0.98	0.32878	
factor(Q3)3	0.02701	0.1531	0.18	0.85994	
factor(Q3)4	-0.2057	0.19829	-1.04	0.29957	
factor(Q3)5	-0.38411	1.40521	-0.27	0.78459	
factor(Q4)2	-0.01246	0.08323	-0.15	0.88097	
factor(Asian)1	-0.24721	0.22503	-1.1	0.27196	
factor(Black)1	-0.21954	0.13677	-1.61	0.10844	
factor(White)1	-0.13271	0.08735	-1.52	0.12869	
Q6	-0.0868	0.49546	-0.18	0.86093	
Q7	0.00227	0.00219	1.04	0.29975	
factor(Q10)2	0.3266	0.10853	3.01	0.00262	**
factor(Q10)3	0.23856	0.10572	2.26	0.02404	*
factor(Q10)4	0.47811	0.20211	2.37	0.018	*
factor(Q10)5	0.2009	0.17376	1.16	0.2476	
factor(Q31)2	-0.61011	0.07766	-7.86	3.90E-15	***
factor(Q41)2	0.61107	0.13653	4.48	7.60E-06	***
factor(Q41)3	1.03725	0.12802	8.1	5.40E-16	***

factor(Q41)4	1.076	0.14046	7.66	1.80E-14	***
factor(Q41)5	1.31212	0.14546	9.02	< 2e-16	***
factor(Q41)6	1.63248	0.15531	10.51	< 2e-16	***
factor(Q41)7	1.80172	0.15918	11.32	< 2e-16	***
factor(Q47)2	0.37491	0.12257	3.06	0.00222	**
factor(Q47)3	0.69603	0.11624	5.99	2.10E-09	***
factor(Q47)4	0.91035	0.13602	6.69	2.20E-11	***
factor(Q47)5	0.9428	0.14028	6.72	1.80E-11	***
factor(Q47)6	1.44834	0.14073	10.29	< 2e-16	***
factor(Q47)7	1.86718	0.11306	16.51	< 2e-16	***
factor(Q60)2	-0.14753	0.07892	-1.87	0.06156	.
factor(Q68)2	0.35806	0.22211	1.61	0.10694	
factor(Q68)3	0.3985	0.12057	3.31	0.00095	***
factor(Q68)4	0.46284	0.18759	2.47	0.01361	*
factor(Q80)2	-0.16095	0.16153	-1	0.31905	
factor(Q80)3	0.13668	0.14032	0.97	0.33003	
factor(Q80)4	0.14	0.13295	1.05	0.29233	
factor(Q80)5	0.08413	0.13799	0.61	0.5421	
factor(Q80)6	0.03941	0.1299	0.3	0.76162	
factor(Q80)7	0.12165	0.16094	0.76	0.44971	
factor(Q80)8	0.01977	0.12093	0.16	0.87014	
factor(Q82)2	0.09034	0.12093	0.75	0.45503	
factor(Q82)3	-0.18232	0.13253	-1.38	0.16894	
factor(Q82)4	-0.01765	0.11964	-0.15	0.88272	
factor(Q82)5	0.08627	0.12004	0.72	0.47234	
factor(Q82)6	0.17751	0.13266	1.34	0.18086	
factor(Q82)7	0.00391	0.10682	0.04	0.97079	
factor(Q84)2	0.0484	0.08467	0.57	0.56759	
factor(Q84)3	0.08058	0.10023	0.8	0.42143	
factor(Q84)4	0.12997	0.11873	1.09	0.27366	
factor(Q85)2	-0.11465	0.10171	-1.13	0.25965	
factor(Q85)3	-0.22178	0.15105	-1.47	0.14204	
factor(Q88)2	-0.34426	0.1397	-2.46	0.01373	*
factor(Q88)3	-0.52401	0.12677	-4.13	3.60E-05	***
factor(Q88)4	-0.62867	0.12649	-4.97	6.70E-07	***
factor(Q88)5	-0.93458	0.1365	-6.85	7.60E-12	***
factor(Q88)6	-0.83568	0.19406	-4.31	1.70E-05	***
factor(Q88)7	-0.39544	0.28661	-1.38	0.16767	
factor(Q89)2	0.09176	0.08627	1.06	0.28754	
factor(Q89)3	0.0344	0.09883	0.35	0.72783	

factor(Q89)4	0.06982	0.16356	0.43	0.66948	
factor(Q89)5	-0.07033	0.25183	-0.28	0.78002	
factor(Q89)6	0.29736	0.46545	0.64	0.52291	
factor(Q89)7	0.11844	0.20544	0.58	0.56426	
factor(Q99)2	-0.10593	0.09758	-1.09	0.2777	
factor(Q99)3	-1.01377	0.41391	-2.45	0.01432	*
factor(Q99)4	0.24685	0.53487	0.46	0.64443	

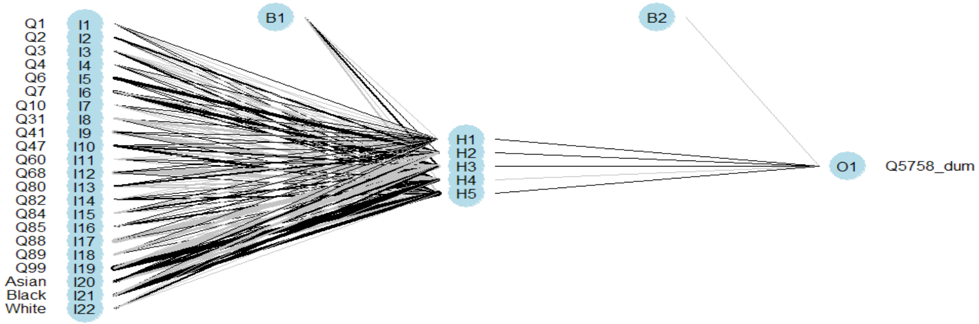


Fig2. *Artificial Neural Network in training sample*

In above plot, line thickness represents weight magnitude and line color weight sign (black = positive, grey = negative). The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. Suffice to say that the training algorithm has converged and therefore the model is ready to be used.

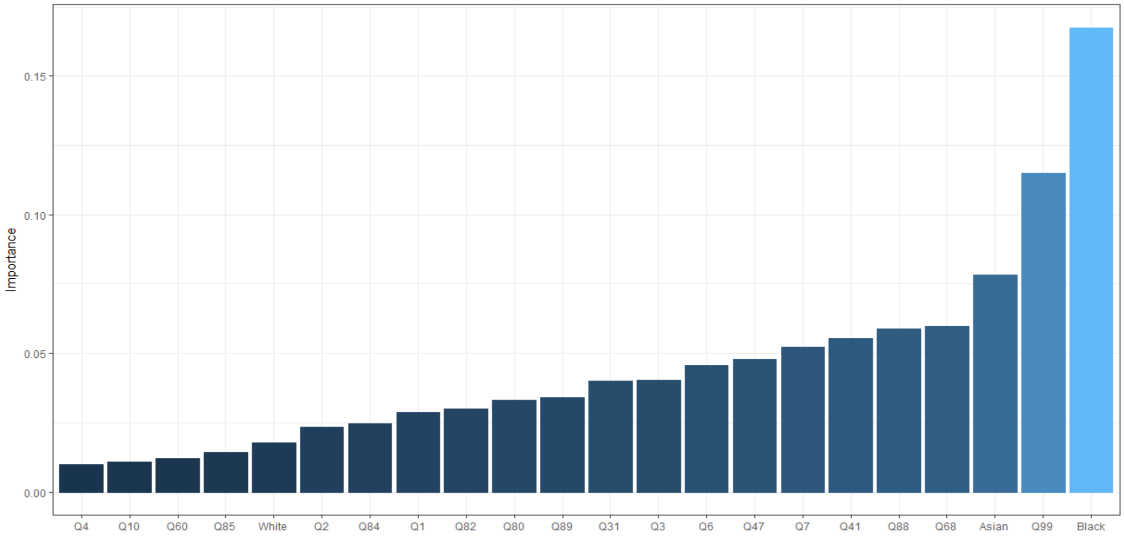


Figure 3: Variable Importance in Artificial Neural Network

According to this neural network, the top 5 most important predictors were 'being black', Q99 (How well do you speak English), 'being Asian', Q68 (sexual orientation), Q88 (On an average school night, how many hours of sleep do you get?).

For training sample, the ROC was 0.84 for the Logistic regression and 0.88 for the artificial neural network. Artificial neural network performed better clearly.In testing sample, the ROC was 0.83 for the Logistic regression and 0.80 for the artificial neural network. Artificial neural network had worse performance.

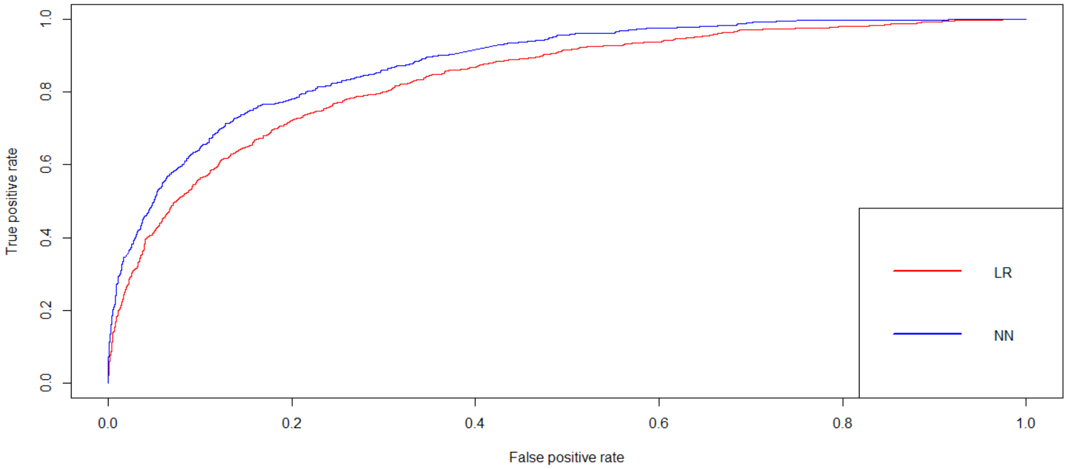


Fig4. *ROC in training sample for Logistic Regression (Red) vs Neural Network (Blue)*

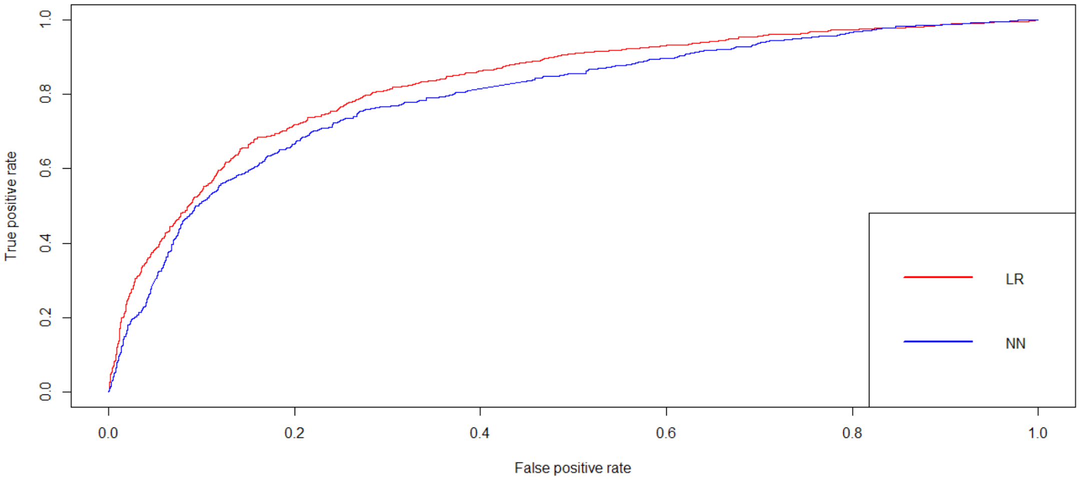


Fig5. *ROC in testing sample for Logistic Regression (Red) vs Neural Network (Blue)*

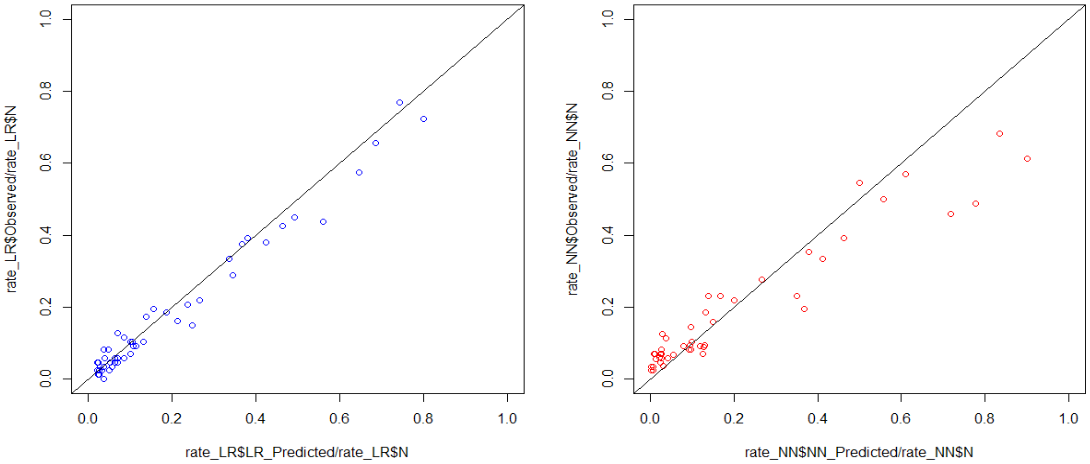


Fig6. *Predicted Probability vs. Observed Probability intesting sample for Logistic Regression (Red) vs Neural Network (Blue), sorted by predicted probability*

By visually inspecting the plot we can see that the predictions made by the neural network are (in general) less concentrated around the line (a perfect alignment with the line would indicate an ideal perfect calibration) than those made by the Logistic model.

Discussions

In this study, we identified several important predictors for drug use e.g., cigarette smoking, drinking, sexual orientation. This provided important information for educators as well as parents provide timely intervention. We built a predictive model using artificial neural network as well as logistic regression to provide a tool for early detection. As to performance of these two models, logistic regression and neural network had a similar discriminating capability.

According to the logistic regression, students who had rides in a car driven by someone who is drinking were more likely to have drug use. Peer pressure is known factor for high school drug use. [3]

Students who never tried cigarette smoking were less likely to use drug. Students who drank often were more likely to use drug. These factors were closely related to drug use in school.

Heterosexual students were less likely to use drug. Students who slept 4 hours or less daily were more likely to use drug. Students who did not speak English well were less likely to be a drug user. According to neural network, Black and Asian students were also more likely to be on drug. Minority groups in high school were more likely to have high risk behavior, for example, drug use. [4,5]

There are limitations of this study. some known factors which might predict of drug use were not available in this study, for example lack of parental supervision. Further we did not test the external validity neither for logistic regression nor for the ANN. However, we did a comprehensive split-sample validation with both strategies. Future studies could use outside data and test the performance of the outputs from these two models in this study.

A predictive model would be an extremely useful tool to detect drug use among high school students. As long as the variables included in our tool are available, the risk t could be easily predicted. Early detection and intervention could be made available for the students at high risk by either the parents or the teachers at school. As to performance of these two models, logistic regression had a similar discriminating capability and a better calibration between predicted probability and observed probability than artificial neural network.

References

1. Projectknow. http://www.projectknow.com/discover/high-school-drug-use/
2. NIH. 2016. https://www.drugabuse.gov/publications/drugfacts/monitoring-future-survey-high-school-youth-trends
3. Simons-Morton B, Farhat T. Recent Findings on Peer Group Influences on Adolescent Substance Use. The Journal of Primary Prevention. 2010;31(4):191-208. doi:10.1007/s10935-010-0220-x.
4. The Impact Of Drugs on Different Minority Groups: A Review Of The UK Literature. July 2010.
5. Factor R, Williams DR, Kawachi I. Social Resistance Framework for Understanding High-Risk Behavior Among Nondominant Minorities: Preliminary Evidence. American Journal of Public Health. 2013;103(12):2245-2251. doi:10.2105/AJPH.2013.301212.

Citation: Beichen Wang. *"Predicting Risk of Drug Use for High School Students Using Artificial Neural Network". American Research Journal of Humanities and Social Sciences, vol 4, no. 1, 2018, pp. 1-12.*