Taylor & Francis
Taylor & Francis Group

Addiction Research & Theory

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/iart20

Applying machine learning methods to model social interactions in alcohol consumption among adolescents

Aliaksandr Amialchuk, Onur Sapci & Jon D. Elhai

To cite this article: Aliaksandr Amialchuk, Onur Sapci & Jon D. Elhai (2021) Applying machine learning methods to model social interactions in alcohol consumption among adolescents, Addiction Research & Theory, 29:5, 436-443, DOI: 10.1080/16066359.2021.1887147

To link to this article: https://doi.org/10.1080/16066359.2021.1887147

Published online: 22 Feb 2021.

ADDICTION RESEARCH & THEORY
2021, VOL. 29, NO. 5, 436–443
https://doi.org/10.1080/16066359.2021.1887147

Taylor & Francis
Taylor & Francis Group

ARTICLE

Applying machine learning methods to model social interactions in alcohol consumption among adolescents

Aliaksandr Amialchuk[a], Onur Sapci[a] and Jon D. Elhai[b,c]

[a]Department of Economics, University of Toledo, Toledo, OH, USA; [b]Department of Psychology, University of Toledo, Toledo, OH, USA; [c]Department of Psychiatry, University of Toledo Toledo, OH, USA

ABSTRACT

Background: Existing research using machine learning to investigate alcohol use among adolescents has largely neglected peer influences and tended to rely on models which selected predictors based on data availability, rather than being guided by a unifying theoretical framework. In addition, previous models of peer influence were typically estimated by using traditional regression techniques, which are known to have worse fit compared to the models estimated using machine learning methods.
Methods: Addressing these limitations, we use three machine-learning algorithms to fit a theoretical model of social interactions in alcohol consumption. The model is fit to a large, nationally representative sample of U.S. school-aged adolescents and accounts for various channels of peer influence.
Results: We find that extreme gradient boosting is the best performing algorithm in predicting alcohol consumption. After the algorithm ranks, the explanatory variables by their importance in classification, previous year drinking status, misperception about friends' drinking, and average actual drinking among friends are the most important predictors of adolescent drinking.
Conclusions: Our findings suggest that an effective intervention should focus on school peers and adolescents' perceptions about drinking norms, in addition to the history of alcohol use. Our study may also increase interest in theory-driven selection of covariates for machine-learning models.

ARTICLE HISTORY
Received 3 September 2020
Revised 3 February 2021
Accepted 3 February 2021

KEYWORDS
Alcohol; peer influence; social norms; adolescents; machine learning

1. Introduction

Reducing adolescent drinking remains a challenge despite many public policies targeting underage drinking such as minimum legal drinking age laws (Clark and Loheac 2007; Johnston et al. 2018). In 2017, the share of adolescents (12–20 years old) in total alcohol consumption was 11% in the USA (Johnston et al. 2018) and one-third of 12th-graders consumed alcohol in the past 30 days. Underage drinking bears significant costs in the form of accidents, premature mortality, productivity loss, health, and behavioral issues (DHHS 2013 2019). Most notably, auto accidents associated with drinking are the leading cause of death among adolescents between the ages of 17 and 20 (Schulenberg and Maggs 2002). Adolescence is a critical age when initiation and increase in alcohol consumption takes place; heavy drinking and alcohol-related problems peak in the late teens through early twenties (Johnston et al. 2018). In addition, initiating drinking in teen years puts people at a higher risk of alcoholism and abuse compared to those who start drinking at a later age (DHHS 2013; Strashny 2014).

Several individual, family, and neighborhood-level variables have been identified in the literature as contributing factors to adolescent alcohol use (Windle et al. 2005; Fletcher 2012). The influence of peers has been identified as an especially important factor (Clark and Loheac 2007; Fletcher 2012; Rees and Wallace 2014; Amialchuk et al. 2019). Several mechanisms of peer influence have been proposed, including (often misperceived) social norms, which make excessive alcohol use by peers appear common and acceptable (Borsari and Carey 2001; Perkins 2014).

Most of the previous studies of peer influence in substance use, including alcohol use, were based on traditional multiple regression analysis and focused on the significance of individual predictors in a model. Machine learning algorithms, on the other hand, are focused on the generalizability and reliability of prediction (or classification) of the entire model and are especially useful in the presence of a large number of predictors some of which are less important for prediction than others (Yarkoni and Westfall 2017). While machine learning is similar to traditional regression analysis, it offers an advantage of detecting patterns in example or training data that can be subsequently used to enhance predictions in other sets of data (a method called 'supervised machine learning') (LeCun et al. 2015). This feature of supervised machine learning offers an improved statistical accuracy over traditional statistical methods (Jordan and Mitchell 2015). In addition, comparing the performance of different machine learning algorithms offers insight into the nature of the prediction process, such as identifying significant low-order interactions among the predictors or

CONTACT Aliaksandr Amialchuk ✉ Aliaksandr.Amialchuk@utoledo.edu Department of Economics, University of Toledo, 2801 Bancroft St. MS 922, Toledo 43606-3390, OH, USA

other non-linearities that traditional regression models may not capture (Breiman 2001; Friedman 2001).

To our knowledge, only a few studies have used machine learning tools to investigate alcohol use among adolescents (Whelan et al. 2014; Hariharan et al. 2016; Pagnotta and Amran 2016; Palaniappan et al. 2017; Afzali et al. 2019; Sarić et al. 2019) and only one study has incorporated peer influence among predictors in a machine-learning model (Afzali et al. 2019). A major limitation of these studies is the atheoretical kitchen-sink approach to model specification, where selection of predictors was primarily based on data availability rather than guided by a unifying theoretical framework. In the absence of an underlying theoretical model, which suggests causal pathways between the variables, the empirical model fit by any statistical routine to a survey dataset is just a data-mining exercise of finding statistically significant associations among variables (Elhai and Montag 2020). Such atheoretical approach to model specification would potentially result in inclusion of predictors which are only spuriously correlated with the outcome or are themselves affected by the outcome – for example, when variables 'weekly study time,' 'number of academic failures,' 'number of school absences' are included among predictors (see, for example, Hariharan et al. 2016; Pagnotta and Amran 2016; Palaniappan et al. 2017; Sarić et al. 2019). The empirical model that results from such approach does not have a causal interpretation but only represents as a set of adjusted correlations. Even when such model shows good fit and prediction quality, it is not useful for making policy suggestions regarding how the outcome can be affected by changing the values of the explanatory variables. In addition, all previous machine-learning studies have been based on non-representative and convenience samples and only one study has attempted to test model generalizability by conducting cross-sample validation using an independent sample (Afzali et al. 2019).

Our study utilizes a large and nationally representative survey of health behaviors and school friendship networks of U.S. middle and high school adolescents with the goals of: (i) using the theory of social interactions to specify an empirical model which allows for different mechanisms of peer influence; (ii) comparing the performance of three of the most accurate machine learning algorithms (discussed below) to fit this model and to understand the importance ranking of variables in predicting alcohol consumption behavior of adolescents.

2. Theoretical framework and statistical identification of peer effects in alcohol use

Peers are thought to influence alcohol consumption through several mechanisms. Social learning theory posits that there is a combination of direct and indirect peer influences (Kandel 1985; Borsari and Carey 2001). The direct influence includes active/overt offers of alcohol such as offering a drink, encouraging, or forcing someone to have a drink. Indirect peer influence occurs when peers' actions in a given social context provide information about acceptable and appropriate behaviors, which, if followed by the individual, may lead to his/her social acceptance and approval. The forms of indirect peer influence include modeling of others' drinking, and influence through perceived social norms (Borsari and Carey 2001).

Perceptions about peer behavior were identified as some of the most important drivers of teen drinking (Neighborset al. 2007; Perkins 2014; Pedersen et al. 2017). Moreover, it is often observed that adolescents overestimate the levels of use and abuse of alcohol by their peers and inaccurately conclude that the norm is drinking heavily after observing heavy drinking by a few peers (Borsari and Carey 2001, 2003; Perkins 2014). This deviation of an individual's perception of the group norm from the actual group norm has been attributed to such phenomena as 'pluralistic ignorance' and the 'fundamental attribution error' (Prentice and Miller 1993; Borsari and Carey 2001; Halbesleben et al. 2004). Inflated perceptions of peer drinking were, in turn, found to be associated with several alcohol consumption behaviors of adolescents including drinking initiation intentions, age of first drink, current drinking status, problem drinking, heavy drinking, drunkenness, and changes in drinking over time (Oldset al. 2005; D'Amico and McCarthy 2006; Juvonen et al. 2007; Page et al. 2008; D'Amico et al. 2012; Song et al. 2012; Perkins 2014; Amialchuk et al. 2019).

Statistical identification of peer influence from observational data is usually complicated by several empirical issues, which may create a correlation between behaviors of peers that has nothing to do with behavioral peer influence. These issues were collectively referred to as 'the reflection problem' by Manski (2000), who pointed out three competing explanations for the correlation between the individual's and peers' outcomes in observational data: the correlated effects, the contextual (or exogenous) effects, and the endogenous peer effect. The correlated effects refer to the influence of common unobserved confounding characteristics faced by all individuals who are in the same peer group. Contextual effects refer to the effect of background characteristics of the peer group, such as peers' family influence. The endogenous peer effect measures how individual behavior is influenced by the behavior of his/her peers. In this analysis, we control for the correlated and contextual effects in order to help statistically identify the behavioral peer effect in adolescent drinking.

3. Method

3.1. Data

We obtained data from the in-home portion of the first two waves of the National Longitudinal Study of Adolescent to Adult Health (Add Health). This is a nationally representative survey of American adolescents in 132 schools who were in grades 7–12 in 1994. The initial sample was first interviewed in their homes in 1994 (Wave 1,20,745 respondents) with a follow-up survey in 1996 (Wave 2,14,738 respondents). A unique feature of Add Health not present in any other large dataset is that Waves 1 and 2 contain information about nominations of the respondents' close friends (up to five male and five female friends, best friends nominated first). Because almost all of these friends were part of the school survey, we are able to obtain actual alcohol use of friends from those individuals̀ direct responses. This study was approved as 'exempt' by the Institutional Review Board.

3.2. Outcome variable

Our outcome variable is the indicator variable measured in Wave 2 of the survey and is based on the question 'During the past 12 months, on how many days did you drink alcohol?' with possible responses (1) 'every day or almost every day,' (2) '3 to 5 days a week,' (3) '1 or 2 days a week,' (4) '2 or 3 days a month,' (5) 'once a month or less (3-12 times in the past 12 months),' (6) '1 or 2 days in the past 12 months,' and (7) 'never.' We collapsed the alcohol variable into a new binary variable based on if respondents chose response 1, 2, 3 or 4 (i.e. drinking more than once per month, now coded '1'), and coding a value of 'zero' otherwise.[1]

3.3. Explanatory variables

The influence of norms is captured by two variables, the actual average proportion of friend-drinkers and the normative misperception score. The actual proportion of friend-drinkers is determined from the direct responses of nominated friends and is computed as the fraction of the three first-nominated male and three first-nominated female friends who reported drinking more than once a month in Wave 1, with possible values of 0; $1/N$, $N=2,3,4,5,6$; 2/N, $N=3,4,5,6$; $3/N$, $N=4,5,6$; $4/N$, $N=5,6$; 5/6, and 1. This variable captures peer influences other than through perceived norms; such as through modeling and overt offers of alcohol. The normative misperception score is calculated as the difference between the perceived proportion of drinkers among friends as reported by the respondent and the actual proportion of friend-drinkers. This approach yields an alcohol misperception score that ranges between -1 and 1, where negative values indicate under-estimation and positive numbers indicate over estimation of the proportion of friend-drinkers, and the misperception score of zero indicates correct estimation (no perceptual bias) (Amialchuk et al. 2019).[2]

We also include the respondent's own alcohol use in the previous wave (Wave 1), which is used to control for any fixed differences among adolescents, which are attributable to drinking, as well as the addictive nature of drinking. The explanatory variables that are intended as controls for the correlated effects (Manski 2000) were measured in Wave 1 and included grade-level indicators for grades 7–12 (grade 7 being the reference category), age, gender, race (white, black, other race), Hispanic ethnicity, picture vocabulary test (PVT) score (Harris 2013), log of pretax family income, indicator of whether residing with both biological parents, whether mother and/or father has a college degree, indicator for attendance of religious services at least once a month, indicator for presence of older siblings. In order to account for household context and parental supervision, we included indicators for whether alcohol, cigarettes, or illegal drugs are easily available at home. We also include age of the adolescent when they first moved to their current location and whether the parents chose their residence because of the school district in order to better account for selection of location. Finally, we included averages of these explanatory variables among the respondents' nominated friends in order to account for the contextual effects in the identification of peer effects (Manski 2000).

3.4. Sample

Our study sample consists of respondents who completed both Wave 1 (1994) and Wave 2 (1996) of the survey ($N=14{,}738$). The data from nominated friends are linked to each respondent. The average number of friends that each respondent nominated is 2.54. 85% of all friendship nominations are from the same school as the respondent and these are the usable friendship nominations in the analysis. The sample size is restricted to individuals who were at most 20 years old (below legal minimum drinking age) in Wave 2, and nominated at least one friend. After removing observations with missing friends' data, missing data on the explanatory variables, and missing sample weights in Wave 2, the usable sample size contains 4686 respondents; this sample was used in all of our models. Add Health utilizes a multistage clustered sample design with observations having unequal probability of selection; therefore our estimations use Add Health sampling weights to make the estimates nationally representative (Chen 2014). Table 1 presents summary statistics for the sample and the variables used in the analysis.

3.5. Machine learning analysis

We used R software version 3.6.1 (R Core Team 2019) for analysis, using the R package *caret* for machine learning. In addition to using R's base package for generalized linear modeling, we used the package *xgbTree* (extreme gradient boosted regression), and *ranger* (random forest) for our machine learning algorithms. We first randomly shuffled participant data rows, using a random number seed which we fixed to allow for the same participant order upon subsequent re-analysis if necessary. We stratified the sample 70/30 in training/test allocation by the values of the dependent variable in order not to end up with a test set that has too few cases of the less prevalent category of the dependent variable. In the training sample, there were 2305 subjects endorsing 0, and 976 subjects endorsing 1. In the test sample, there were 993 subjects endorsing 0, and 412 subjects

[1] Computer-assisted personal interviewing of Add Health was used to ensure confidentiality of responses and reduce reporting bias.

[2] The paper by Amialchuk et al. (2019) introduced this normative misperception score and focused on estimating the effect of normative misperception on the use of three substances (alcohol, marijuana, and smoking) using a linear regression. The present paper focuses on using machine-learning methods to fit a social interaction model which incorporates several mechanisms of peer influence.

Table 1. Descriptive statistics, $N = 4686$.

Variable	Mean	SD	Min	Max
Drinking, W1	0.155	0.362	0	1
Drinking, W2	0.183	0.387	0	1
Drinking misperception score, W1	0.234	0.470	−1	1
Average friends' drinking, W1	0.177	0.336	0	1
Grade 8, W1	0.147	0.355	0	1
Grade 9, W1	0.174	0.379	0	1
Grade 10, W1	0.244	0.429	0	1
Grade 11, W1	0.235	0.424	0	1
Grade 12, W1	0.0630	0.243	0	1
Age, W1	14.90	1.549	11	20
Male	0.480	0.500	0	1
Black	0.182	0.386	0	1
Other race	0.0110	0.106	0	1
Hispanic	0.158	0.365	0	1
PVT score	100.5	14.16	12	135
Ln(pretax family income)	3.614	0.876	−2.303	6.907
Both biological parents present	0.578	0.494	0	1
Mother or father has college degree	0.373	0.484	0	1
Religious	0.607	0.488	0	1
Older siblings present	0.514	0.500	0	1
Alcohol easily available at home	0.299	0.458	0	1
Cigarettes easily available at home	0.303	0.460	0	1
Illegal drugs easily available at home	0.0330	0.179	0	1
Parents chose school	0.426	0.494	0	1
Years old when moved	8.113	5.590	0	19

W1: Wave 1; W2: Wave 2

endorsing 1. After training/test set allocation, we preprocessed the predictor (explanatory) variables by centering and scaling values to z scores, conducted separately for the training and external test samples as suggested by Kuhn and Johnson (2013) and Wainberg et al. (2016). Next, we conducted machine-learning analyses of the model of social interactions where alcohol consumption measured in Wave 2 is the dependent variable. The predictor variables are alcohol consumption measured in Wave 1, normative misperception scores measured in Wave 1, actual average alcohol use of friends (actual norm), and a set of individual and family-level characteristics for the respondent and his/her best friends.

We used three separate machine-learning algorithms, subsequently comparing their performance. First, we included generalized linear regression using maximum likelihood estimation and a logit link function. Next, we used random forest, a type of ensemble machine learning algorithm, in which many weaker, simpler subsets of observations and variables are randomly chosen and iteratively tested, built to form a stronger, comprehensive model (based on majority vote from the weaker learner results), reducing overfitting and variance as significant advantages. Finally, we included an additional ensemble algorithm, extreme gradient boosting (Kadiyala and Kumar 2018), also iteratively building a strong model from weaker models but correcting weak models' errors along the way. We used the *xgbTree* method of extreme gradient boosting which additionally inherently conducts variable subset selection to remove empirically non-important predictor variables.[3] Aside from generalized linear regression, ensemble algorithms are among the top-performing machine learning algorithms, conceptually (Hastie et al. 2016) and empirically (Amancio et al. 2014; Fernández-Delgado et al. 2015; Wainberg et al. 2016). We compared algorithms based on their area under the curve (AUC), accuracy, sensitivity and specificity in accurate prediction of '1' and '0' values on the dependent variable, and statistically compared them.

For each algorithm we tested, we used k-folds repeated cross-validation (RCV) to simulate validation data prior to testing our model with the external test sample, using a fixed number seed. In RCV, the training sample was randomly split (without replacement) into five folds of non-overlapping participant subsets (as suggested in Kuhn and Johnson 2013; Hastie et al. 2016), whereby four of the given folds were trained and the fifth fold served as the simulated test sample; we repeated this process so that each fold served as a simulated test sample once, resulting in a single completed cross-validation procedure. We repeated this process nine more times for a total of 50 RCVs. While not applicable to generalized linear regression, we used automatic grid search to find the best estimates for the tuning parameters in the other machine learning algorithms (for boosted regression, nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, and subsample; for random forest, mtry, splitrule, and min.node.size). Finally, we applied the final aggregated model for each algorithm to the external test sample as a further validation test.

4. Results

Table 1 presents the descriptive statistics for the variables used in our analysis. The overestimation of the proportion of best friends who drank alcohol at least once a month was by 23.4% points on average in Wave 1. Self-reported drinking rate among those friends was only 17.7%. Table 1 also shows that drinking more than once a month in the

[3] See Kuhn and Johnson (2013) and Hastie et al. (2016) for an overview of these methods, and why random forest and extreme boosting are some of the better, more accurate machine learning algorithms.

Table 2. Comparison of three machine learning-based regression algorithms for the training sample using repeated cross-validation.

Mean classification accuracy findings over repeated cross-validations in the training sample

	AUC	Accuracy	Sensitivity	Specificity
Generalized linear model	0.6541	0.6570	0.6797	0.6285
Extreme gradient boosted regression	0.8057	0.7583	0.7810	0.7297
Random forest	0.8045	0.7521	0.7792	0.7180

AUC: area under the curve, denoting the extent to which the model classified monthly *vs.* less drinking. Sensitivity is the proportion of monthly drinkers correctly classified, while specificity is the proportion of less than monthly drinkers correctly classified. Accuracy is the proportion of all correctly classified participants (correctly classified monthly drinkers, and correctly classified less than monthly drinkers) related to all participants.

Table 3. Comparison of three machine learning-based regression algorithms for the external test sample.

Classification accuracy findings in the test sample

	Accuracy	Sensitivity	Specificity
Generalized linear model	0.4484	0.2797	0.5824
Extreme gradient boosted regression	0.7359	0.6977	0.7663
Random forest	0.7203	0.6688	0.7612

previous year has increased from 15.5% to 18.3% as the average age increased from 15 to 16 years old between Waves 1 and 2.

Next, we present the machine learning results, compared across the three algorithms. Table 2 demonstrates that extreme gradient boosting was most accurate in classification, with slightly larger AUC, accuracy, sensitivity, and specificity values. Using Bonferroni-adjusted pairwise tests, extreme gradient boosting performed significantly better than generalized linear (logit) regression, and was better than random forest on specificity only. Table 3 shows that the ensemble learning algorithms generalized fairly well from the training to the test sample, without substantial worsening in classification accuracy. Extreme gradient boosting appeared slightly better performing than random forest in the test sample.

Because of the slightly better performance for extreme gradient boosting, we used this algorithm to display the relative importance of the predictor variables in classifying values on the dependent variable. Figure 1 displays variable importance, interpreted as standardized regression coefficients; we used the single-tree approach, but importance values were summed across iterations of boosting. Because of subset selection, the bottom nine predictors were automatically excluded, with their regression coefficients fixed to zero. Drinking in the previous wave, drinking misperception score and average friends' drinking were the three most important variables in classification. The other two algorithms also ranked these variables as the most important (results not displayed but available upon request); for generalized linear regression, we used the absolute value of the t-statistic for each model parameter; for random forests, we recorded out-of-bag performance, and performance across permutations of predictor variables, computing the difference between these two accuracy indices across all trees and normalized by the standard error.

5. Discussion

This article combines theory-driven selection of covariates and machine-learning techniques in order to fit a model of social interactions, which allows for the influence of social norms and normative misperceptions on alcohol consumption of adolescents. We address the problem of statistical identification of behavioral peer effects in the model by including controls for the correlated and contextual effects (Manski 2000). We fit the model of social interactions to a unique dataset containing large school networks of friends. We use classification-based machine learning, implementing algorithms that are among the most accurate and overcome important limitations inherent in traditional statistical methods. These algorithms include generalized linear model, random forest, and extreme gradient boosted regression. The estimation focused on goodness of fit and interpretability of the results by ranking the predictors by their importance. Extreme gradient boosting regression performed the best among the three algorithms in predicting alcohol consumption. After ranking the explanatory variables in terms of their importance in classification and after removing empirically non-important predictor variables, previous year drinking, misperception about friends' drinking, and average actual drinking of friends were the three most important predictors of adolescent alcohol consumption.

These results echo the findings of previous studies that found perceived and actual peer norms to be important determinants of adolescent drinking (Neighbors et al. 2007; Fletcher 2012; Song et al. 2012; Perkins 2014; Pedersen et al. 2017). However, to the best of our knowledge, this is the first study to use the theory of social interactions to select predictors for the model that is estimated using machine learning techniques. Previous studies using machine-learning analysis relied on a 'kitchen sink' approach for model specification by selecting all seemingly relevant variables available in the dataset and letting the machine learning algorithms data mine statistically important relationships and variables. Previous studies were also based on non-representative convenience samples and rarely included measures of peer influence.

Our findings should be considered in light of the following limitations. First, the dataset lacks information on friends' approval (or disapproval) of drinking, which may also play a role in predicting drinking among adolescents (Lee et al. 2007; LaBrie et al. 2010; Pedersen et al. 2017). Second, the machine learning algorithms used in this study are not designed to deal with inherent data structures such as school-level clustering of observations. Third, even though close friends can be thought as the most influential group on onѐs decisions, other peer groups such as neighbors, and parents may have their own independent influences on

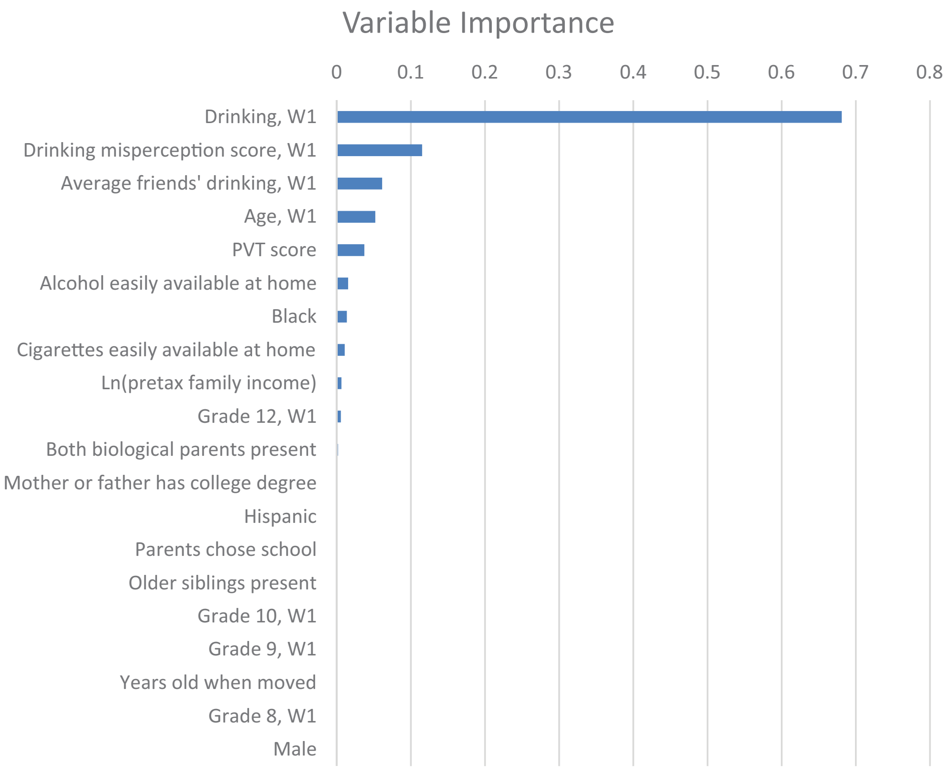

Figure 1. Variable importance.

adolescent drinking (Song et al. 2012). While the age of the data used in this analysis (1994–1996) may be viewed as a limitation, our results are still relevant today because close friends continue to play an important role in adolescent alcohol use (Perkins 2014; Pedersen et al 2017; Afzali et al. 2019)2019. Future research should investigate cross-sample generalizability of our model using an independent test sample.

Despite these limitations, the major strengths of this study are the theory-based approach to selecting covariates for machine-learning model building, and using a large nationally representative dataset of adolescents to fit the model and understand importance of the predictors. This study is the first to test different machine learning algorithms to predict alcohol use of adolescents using a causal model as opposed to a 'kitchen sink' approach to variable selection and the use of convenience samples in the previous studies. Using machine-learning techniques allows obtaining a better fitting model compared to traditional regression techniques (Jordan and Mitchell 2015).

Our approach and findings have important implications. In terms of policy, effective interventions should focus on the drinking of school peers and teens' perceptions about rates of drinking among their friends, in addition to focusing on adolescents with history of alcohol use. This study might also increase interest in theory-driven selection of covariates for machine-learning model building. Having an underlying causal model implies that interventions that change the values of important predictors in the model will produce reductions in adolescent drinking. Finally, this study suggests a possibility of development of computerized software for screening, prevention, and education of at-risk adolescents, which could lead to reduction in the economic and social costs of drinking.

Acknowledgments

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent statement

This study used secondary data analysis and no research involving human subjects was done; therefore no informed consent to participate in the study was obtained.

Disclosure statement

The authors declare that they have no conflict of interest.

Funding

The authors received no specific funding for this work.

ORCID

Onur Sapci http://orcid.org/0000-0003-3739-0439

Data availability statement

The data that support the findings of this study are available from The Carolina Population Center at The University of North Carolina – Chapel Hill (UNC) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of The Carolina Population Center at UNC.

References

Afzali MH, Sunderland M, Stewart S, Masse B, Seguin J, Newton N, Teesson M, Conrod P. 2019. Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation. Addiction. 114(4):662–671.

Amancio DR, Comin CH, Casanova D, Travieso G, Bruno OM, Rodrigues FA, Costa LDF. 2014. A systematic comparison of supervised classifiers. PLoS One. 9(4):e94137.

Amialchuk A, Ajilore G, Egan K. 2019. The influence of misperceptions about social norms on substance use among school-age adolescents. Health Econ. 28(6):736–747.

Borsari B, Carey KB. 2001. Peer influences on college drinking: a review of the research. J Subst Abuse. 13(4):391–424.

Borsari B, Carey KB. 2003. Descriptive and injunctive norms in college drinking: a meta-analytic integration. J Stud Alcohol. 64(3):331–341.

Breiman L. 2001. Random forests. Mach Learn. 45(1):5–32.

Chen P. 2014. Appropriate analysis in add health correcting for design effects & selecting weights. Chapel Hill (NC): Carolina Population Center, University of North Carolina at Chapel Hill.

Clark A, Loheac Y. 2007. "It wasn't me, it was them!" social influence in risky behavior by adolescents. J Health Econ. 26(4):763–784.

D'Amico EJ, McCarthy DM. 2006. Escalation and initiation of younger adolescents' substance use: the impact of perceived peer use. J Adolesc Health. 39(4):481–487.

D'Amico EJ, Tucker JS, Miles JNV, Zhou AJ, Shih RA, Green HD. Jr. 2012. Preventing alcohol use with a voluntary after-school program for middle school students: results from a cluster randomized controlled trial of CHOICE. Prev Sci. 13(4):415–425.

DHHS. 2013. US department of health and human services. Report to congress on the prevention and reduction of underage drinking. https://store.samhsa.gov/shin/content/PEP13-RTCUAD/PEP13-RTCUAD.pdf.

DHHS. 2019. Fact sheets - underage drinking. Division of population health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention. Page visited: June 30, 2019. Washington (DC): DHHS.

Elhai JD, Montag C. 2020. The compatibility of theoretical frameworks with machine learning analyses in psychological research. Curr Opin Psychol. 36:83–88.

Fernández-Delgado M, Cernadas E, Barro S, Amorim D. 2015. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 15:3133–3181.

Fletcher J. 2012. Peer influences on adolescent alcohol consumption: evidence using an instrumental variables/fixed effect approach. J Popul Econ. 25(4):1265–1286.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Ann Statist. 29(5):1189–1232.

Halbesleben JR, Buckley MR, Sauer ND. 2004. The role of pluralistic ignorance in perceptions of unethical behavior: an investigation of attorneys' and students' perceptions of ethical behavior. Ethics Behav. 14(1):17–30.

Hariharan B, Krithivasan R, Angel D. 2016. Prediction of secondary school students' alcohol addiction using random forest. Int J Comput Appl. 149(6):21–25.

Harris K. 2013. The add health study: design and accomplishments. Chapel Hill (NC): Carolina Population Center, University of North Carolina.

Hastie T, Tibshirani R, Friedman J. 2016. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Berlin (Germany): Springer.

Johnston LD, Miech RA, O'Malley PM, Bachman JG, Schulenberg JE, Patrick ME. 2018. Monitoring the future national survey results on drug use: 1975–2017: overview, key findings on adolescent drug use. Ann Arbor (MI): Institute for Social Research, The University of Michigan.

Jordan MI, Mitchell TM. 2015. Machine learning: trends, perspectives, and prospects. Science. 349(6245):255–260.

Juvonen J, Martino S, Ellickson P, Longshore D. 2007. But others do it! Do misperceptions of schoolmate alcohol and marijuana use predict subsequent drug use among young adolescents? J Appl Social Pyschol. 37(4):740–758.

Kadiyala A, Kumar A. 2018. Applications of python to evaluate the performance of decision tree-based boosting algorithms. Environ Prog Sustain Energy. 37(2):618–623.

Kandel DB. 1985. On process of peer influences in adolescent drug use: a developmental perspective. Adv Alcohol Subst Use. 4(3–4): 139–163.

Kuhn M, Johnson K. 2013. Applied predictive modeling. New York (NY): Springer.

LaBrie JW, Hummer JF, Neighbors C, Larimer ME. 2010. Whose opinion matters? The relationship between injunctive norms and alcohol consequences in college students. Addict Behav. 35(4):343–349.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature. 521(7553): 436–444.

Lee CM, Geisner IM, Lewis MA, Neighbors C, Larimer ME. 2007. Social motives and the interaction between descriptive and injunctive norms in college student drinking. J Stud Alcohol Drugs. 68(5): 714–721.

Manski CF. 2000. Economic analysis of social interactions. J Econ Perspect. 14(3):115–136.

Neighbors C, Lee CM, Lewis MA, Fossos N, Larimer ME. 2007. Are social norms the best predictor of outcomes among heavy-drinking college students? J Stud Alcohol Drugs. 68(4):556–565.

Olds RS, Thombs DL, Tomasek JR. 2005. Relations between normative beliefs and initiation intentions toward cigarette, alcohol and marijuana. J Adolesc Health. 37(1):75

Page RM, Ihasz F, Hantiu J, Simonek J, Klarova R. 2008. Social normative perceptions of alcohol use and episodic heavy drinking among central and eastern European adolescents. Subst Use Misuse. 43(3–4):361–373.

Pagnotta F, Amran MH. 2016. Using data mining to predict secondary school student alcohol consumption. Camerino MC, Italy: Department of Computer Science, University of Camerino.

Palaniappan S, Hameed NA, Mustapha A, Samsudin N. 2017. Classification of alcohol consumption among secondary school students. JOIV: Int J Inform Visualization. 1(4–2):224.

Pedersen ER, Osilla KC, Miles JNV, Tucker JS, Ewing BA, Shih RA, D'Amico EJ. 2017. The role of perceived injunctive alcohol norms in adolescent drinking behavior. Addict Behav. 67:1–7.

Perkins HW. 2014. Misperception is reality: the reign of error about peer risk behaviour norms among youth and young adults. In: Xenitidou M, Edmonds B, editors. The complexity of social norms, computational social sciences. Switzerland: Springer International Publishing; p. 11–36.

Prentice D, Miller D. 1993. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. J Pers Soc Psychol. 64(2):243–256.

R Core Team 2019. R: a language and environment for statistical computing. https://www.R-project.org/.

Rees C, Wallace D. 2014. The myth of conformity: adolescents and abstention from unhealthy drinking behaviors. Soc Sci Med. 108:34–45.

Sarić R, Dejan J, Edhem Č. 2019. Identification of alcohol addicts among high school students using decision tree based algorithm. CMBEBIH. 2019:459–467.

Schulenberg JE, Maggs JL. 2002. A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood. J Stud Alcohol Suppl. 14:54–70.

Song EY, Smiler AP, Wagoner KG, Wolfson M. 2012. Everyone says it's ok: adolescents' perceptions of peer, parent, and community alcohol norms, alcohol consumption, and alcohol-related consequences. Subst Use Misuse. 47(1):86–98.

Strashny A. 2014. Age of Substance Use Initiation among Treatment Admissions Aged 18 to 30. The CBHSQ Report: July 17, 2014. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.

Wainberg M, Alipanahi B, Frey BH. 2016. Are random forests truly the best classifiers? J Mach Learn Res. 17:1–5.

Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T, Barker GJ, Bokde ALW, Büchel C, Carvalho FM, et al. 2014. Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature. 512(7513):185–189.

Windle M, Mun EY, Windle RC. 2005. Adolescent-to-young adulthood heavy drinking trajectories and their prospective predictors. J Stud Alcohol. 66(3):313–322.

Yarkoni T, Westfall J. 2017. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci. 12(6):1100–1122.