OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Identifying Associated Factors of Substance Use in Adolescents Using Machine Learning

Jocelyn Gao
jgao91892@gmail.com

ABSTRACT

Adolescent substance use remains a pressing public health concern with long-term implications for individuals' physical and mental health. Research in adolescent development suggests that family structure, school environment, and socioeconomic status influence substance use (1), yet the relationship of different factors with substance use is not fully understood. This study applies machine learning techniques to identify environmental, economic, and demographic factors associated with adolescent substance use using data from the 2023 National Survey on Drug Use and Health. It features three different machine learning models: LASSO L1 Penalized Logistic Regression (LASSO), Random Forest, and Light Gradient Boosting Machine (LightGBM), for their predictive accuracy in identifying environmental and demographic correlations for substance use among adolescents. The most accurate model methodology identified was Random Forest based on Area Under the Curve (AUC) values, Area Under the Precision-Recall Curve (AUPRC) values, and Kolmogorov-Smirnov (K-S) statistics. The leading association factors identified by Random Forest were the respondent's school attendance and the number of times the respondent moved in the past year, among others.

INTRODUCTION

Substance use among adolescents in the U.S. has become a serious social issue. It is a known risk factor for the development of neuropsychiatric and substance use disorders in adulthood (2). Drug overdoses are now the 3rd leading cause of pediatric deaths in America, after firearm-related injuries and motor vehicle accidents (3). In 2020, 1.6 million adolescents aged 12-17 and 8.2 million aged 18-25 met the diagnostic criteria for a substance use disorder (e.g. alcohol, tobacco, cannabis, and other substances) (4).

Previous studies of adolescent substance use identified many risk factors, including various biological, psychological, and psychiatric characteristics. While they examined correlations with adolescent substance use using traditional statistical methods, few studies have utilized modern machine learning models or evaluated their relative predictive performance across substances and age groups. Machine learning methods were chosen for this study because of their ability to handle multicollinearity and non-linear relationships, which are limitations commonly encountered in regression-based approaches (5). The research question asks: what environmental, economic, and demographic factors are most strongly associated with adolescent substance use and how do different machine learning models compare in their predictive performance for substance use outcomes? The goal is to address disparities and help to propose targeted solutions.

In the Literature Review section, existing research on the causes and effects of adolescent substance use is discussed. In the Methods section, the procedure is introduced; the data cleaning and processing, as well as machine learning model selection, are discussed. In the Results section, the most correlated factors and the most predictive model are analyzed.

LITERATURE REVIEW

Adolescence is a period of dynamic biological, psychological, and behavioral changes. During adolescence, developmental changes in neural circuitry of reward processing, motivation, cognitive control, and stress may contribute to vulnerability to increased levels of engagement in substance use and non-substance addictive behaviors (6). Adolescents have more dopamine receptors than adults, resulting in a heightened response to substance use. At the same time, the brain regions that control executive functioning (e.g., logical reasoning, planning, and complicated decision-making), including the prefrontal cortex and the cerebellum, remain immature as they undergo a dynamic choreography of synaptic pruning into the mid-20s (7).

Studies have found that families play vital roles in adolescents' risk for engaging in substance use (8). Family structures were found to have both positive and negative associations with substance use among adolescents. As described in one study, paternal knowledge was consistently found to be a protective factor against substance use (9). A study by Luk et al. reported a positive association of maternal psychological association towards substance use (IRR 2.41, $p < 0.05$) (9).

Past research has also found that youth from families of lower socioeconomic status are more likely to smoke. Youth from affluent families exhibit patterns of being more prone to alcohol use, heavy episodic drinking, and marijuana use (10).

Moving to a new location can be a significant change for adolescents. A move can bring challenges like feelings of loss, worries about the unknown, and shifts in parental attention, all of which may be associated with emotional and behavioral problems (11). In a study exploring the relationship between the number of geographic moves before the age of 16 and the timing of onset of substance use and progression to substance-related problems, the results showed highly significant positive relationships between moving and early initiation of illicit substances including marijuana, hallucinogens, crack/cocaine, and illicit use of prescribed drugs (12).

Although prior research has extensively documented individual biological and environmental risk factors for adolescent substance use, several limitations remain. Much of the existing literature relies on traditional regression-based approaches that examine a limited number of risk factors in isolation. As a result, these studies may struggle to capture complex, nonlinear relationships and interactions among demographic, socioeconomic, and environmental variables. Additionally, predictive modeling has not been a central study in much of the existing literature, which leaves a gap in understanding how these diverse factors can jointly impact substance use. Addressing these gaps is essential to mitigate the risk of substance use in adolescents.

This paper considers how various environmental, demographic, and socioeconomic characteristics relate to different types of substances used by adolescents by applying three machine learning methodologies: LASSO, Random Forest, and LightGBM.

LASSO is a linear regression technique that adds a penalty to the model's cost function, equal to the absolute value of the coefficients (L1 regularization). LASSO stands for least absolute shrinkage and selection operator (13). Random Forest is a supervised machine learning algorithm that builds and combines multiple individual decision trees to produce a single, more accurate, and stable prediction (14). Finally, LightGBM is a fast and efficient gradient boosting model (15) that builds ensembles of decision trees sequentially, where each new tree fixes the errors made by the previous trees. The final prediction is the weighted sum of all previous trees.

METHODS

Study Data

The machine learning analysis used data from the 2023 National Survey of Drug Use and Health (NSDUH) (16). The survey collected 67,679 interviews and included 56,705 responses in the final dataset. The survey sampled residents of households and non-institutional group quarters, using a multistage area probability sample including all 50 states and the District of Columbia. States were stratified into state dwelling regions, which were further divided into census tracts, block groups, and area segments, from which dwelling units were selected to be screened. Within each unit, up to two residents aged 12 and above were selected for an interview. The screening response rate was 24.36% and the interview response rate was 50.45%.

The original survey assessed use of cigarettes, alcohol, marijuana, cocaine, heroin, hallucinogens, inhalants, methamphetamines, pain relievers, tranquilizers, stimulants, and sedatives. This study assesses use of cigarettes, alcohol, marijuana, and inhalants because these were the substances with the highest usage rates among adolescents (see Table C.1 and Figure C.1). The other substances were not assessed as, for substances with extremely low usage prevalence, observed usage is largely driven by idiosyncratic variation rather than stable, population-level patterns. As a result, estimates of risk factors and intervention effects are statistically unstable, underpowered, and highly sensitive to measurement error. Given finite public health resources, interventions targeting such rare behaviors are unlikely to be cost-effective or scalable, and may raise ethical concerns due to the need for intrusive identification of a very small affected population. Consequently, intervention design is better focused on substances with sufficient prevalence to support robust inference and meaningful population-level impact.

Independent variables were selected from the Demographics section of the codebook. These variables encompassed various demographic characteristics. Variables were intentionally selected to be broad rather than overly specific and to ensure that each variable value had a sufficient number of responses.

Data Overview

The NSDUH dataset was first downloaded into the RStudio integrated development environment, and filtered for only the entries of respondents aged 12-20. Then the dataset was further separated into 4 age groups of interest–12-13, 14-15, 16-17, and 18-20–by selecting them using age. The separation was done because adolescents of different ages are likely to have varying levels of exposure to substances (17), which could influence model performance and interpretation. The dataset was then cleaned by deleting entries with missing responses (e.g. NAs in independent variables) and recoding categorical variables into binary variables. After cleaning, the training set for the 12-13 age group had 2,220 entries and the test set had 555 entries; the training set for the 14-15 age group had 2,714 entries and the test set had 678 entries; the training set for the 16-17 age group had 2,602 entries and the test set had 650 entries; the training set for the 18-20 age group had 2,988 entries and the test set had 747 entries. To address the potential sampling bias resulting from the 24.36% response rate, all analyses incorporated the sampling weights provided by the survey. These weights were adjusted for unequal probabilities of selection, nonresponse, and post-stratification to known population margins like sex and ethnicity.

Dependent variables were selected for their representation of familial stability (31), demographics (32), and income (10), as these factors are indicative of adolescent substance use. All variables were recoded into binary variables in order to perform the logistic regression. The outcome variables should be binary because the logistic function outputs values between 0 and 1, which aligns with the probability of one class vs another (25). In addition, recoding independent variables as binary makes the coefficient clearer to interpret, as the coefficient will indicate how the odds ratio differs between the two groups. Meanwhile, the categorical variables were converted into dummy binary variables for inclusion in the models. For each categorical variable, one reference category was omitted to avoid multicollinearity, which is a standard practice in regression analysis (18).

Beyond dummy coding, select variables were recoded to reflect conceptual groupings used in prior research. MOVSINPYR2, the number of times moved in the past year variable, was defined such that 2 or more moves in the past year was indicative of residential instability, a standard practice in previous studies (19). Therefore, 0 moves and 1 move were recoded to 0 and 2 moves and 3+ moves were recoded to 1. SEXATRACT2, the variable defining sexual attraction, was recoded into the variable ISHETERO. All original variable classes that weren't strictly attracted to the opposite sex were recoded as 0. This approach was chosen to ensure sufficient statistical power and this approach aligns with previous studies that used a similar strategy (20). Table A.1 lists all the independent variables and how they were treated or recoded.

Pearson correlation coefficient heatmaps were generated to assess the degree of linear independence among the independent variables for each of the four age groups. Exhibit B.1 displays the code used to generate the correlation heatmap, while Figures B.1-B.4 are the generated heatmaps.

The dataset for each age group was then split into a training and testing set, with an 80% training set and a 20% held-out test set.

The dependent variables were taken from the question about past consumption for each substance. Cigarette, alcohol, marijuana, and inhalant use were selected for the final assessment because they had the highest target rate of use. Table C.1 lists all the substances assessed by the NSDUH and the target rates across each of the four relevant age groups. Meanwhile, Figure C.1 displays the target rates combined for all adolescents in one bar graph.

Model methodology

This study aims to identify the key risk factors associated with substance use. Machine learning methods were chosen not solely for predictive accuracy, but because adolescent substance use is influenced by many potentially correlated demographic, economic, and environmental variables, with complex relationships. Traditional regression approaches can struggle in such settings due to multicollinearity and model misspecification (5).

Supervised learning models were selected because they are well-suited for learning from labeled data and optimizing predictive accuracy. The primary focus was selecting models that handle complex patterns in the data and generalize well beyond the training sample.

LASSO was chosen for its powerful automatic feature selection, creating simpler, more interpretable models by shrinking less important feature coefficients to zero, especially useful in high-dimensional data (many features, few observations) to prevent overfitting and build sparse models. This consideration was especially important for datasets like this one, where the outcomes being assessed are relatively low-prevalence. LASSO's regularization penalizes weak or unstable predictors, which reduces overfitting and yields more reliable variable selection when positive cases are rare. In addition, LASSO returns both direction and magnitude of feature coefficients, which is important for determining how different variables affect substance use.

Random Forest uses an ensemble of decision trees to make more accurate and robust predictions for both classification and regression tasks (21). It builds multiple decision trees on different random subsets of the data and with random subsets of features, then combines their individual predictions through majority voting (for classification) or averaging (for regression) to produce a final, more reliable output (14). In this study, majority voting is used for classification, as the model's goal is to classify observations based on the consensus of predictions across different folds of data. Random Forest was particularly well-suited to this dataset because it can accommodate a large number of correlated, mixed-type survey variables and capture nonlinear relationships.

LightGBM was selected for its speed and efficiency, driven by a histogram-based algorithm that groups features into bins, a leaf-wise tree growth strategy for faster learning, and exclusive feature bundling (EFB) to reduce computational load (15). It also has high accuracy and handles categorical features natively (15). LightGBM's ability to handle high-dimensional datasets like this one allows it to efficiently model many complex variables while remaining feasible to train across multiple outcome and age group combinations.

The models were trained separately by substance and age to account for differences in usage patterns across substances and developmental stages. Prior research suggests that the determinants of substance use differ meaningfully by both substance and age (26). Training separate models allows each model to learn substance- and age-specific associations.

In this study, stratified k-fold cross-validation for model validation was used because of the imbalanced target variable. The ideal target rate is 50% for balanced data, and only one of the subgroups exceeded that rate. Therefore, stratified k-fold cross-validation ensured that each fold maintained the same target variable value distribution as the original dataset. Exhibit D.1 displays the code that was used to create stratified 5-fold cross-validation in LASSO. In addition, hyperparameter tuning was also used to find the best hyperparameters that yielded the best model performance.

Model performance was assessed using AUC, AUPRC, and K-S statistic. AUC measures the area under the receiving operating characteristic (ROC) curve. The ROC curve plots the True Positive Rate and False Positive Rate. The greater the AUC value, the better the target prediction. An AUC value above 0.7 is generally considered a fair test, while anything below 0.7 is considered to be nonuseful (22). AUPRC summarizes precision and recall across classification thresholds and is particularly informative in imbalanced datasets, where the positive class is rare (23). Higher AUPRC values indicate better ability to correctly identify positive cases while minimizing false positives. The K-S statistic measures the maximum absolute vertical distance between two cumulative distribution functions (24). A larger K-S statistic indicates a greater difference between the distributions. Therefore, the larger the K-S statistic, the better the target prediction. Table E.1 displays ranges of K-S statistics and their corresponding strength of model predictiveness.

The model that achieved the highest average AUC, AUPRC, and K-S statistic across all folds and substance/age groups was selected as the best model.

<u>Machine learning procedure</u>

LASSO, Random Forest, and LightGBM with stratified k-fold cross-validation and hyperparameter tuning were run on each subgroup by age and substance. The use of cross-validation and tuning was vital to ensure that the model is generalizable for imbalanced datasets, mitigates overfitting too well on training data, and provides performance estimates for across many subsets of data instead of a single train/test split. Hyperparameter tuning was performed using Bayesian optimization. To ensure adequate evaluation of the minority class, multiple metrics emphasizing positive-class performance were calculated, including AUPRC, precision, recall, and F1 score. See Exhibit F.1 for more details. (These metrics complement the AUC by providing more informative performance estimates when the positive class is rare. With threshold adjustment, a decision threshold was selected that maximized the F1 score on the held-out test set to achieve the best trade-off between identifying positive cases and avoiding excessive false positives.)

Each model also incorporated cost-sensitive learning to avoid bias toward the majority (non-use of substance) class. For LASSO, inverse-prevalence sample weights were assigned so that positive cases contributed proportionally more. The Random Forest and LightGBM models incorporated identical sample weights during training. This weighting scheme penalizes misclassification of the minority class more heavily and reduces the risk of systematically under-predicting substance use.

See Exhibit F.2 for details on how each model was tuned. (LASSO was tuned using the lambda hyperparameter. For each lambda (regularization strength), glmnet fitted the model and evaluated the cross-validated deviance. It then selected the lambda with the lowest mean cross-validated error. Random Forest was tuned using the mtry, min.node.size, and sample.fraction hyperparameters with Bayesian optimization. Mtry specifies the number of variables randomly sampled as candidates at each split. Min.node.size tunes the minimum number of observations allowed in a leaf node of a decision tree, which controls the tree's depth. Sample.fraction determines the fraction of data rows to sample with replacement when building each tree. LightGBM was tuned using num_leaves, feature_fraction, bagging_fraction, min_data_in_leaf. Num_leaves controls the maximum number of leaf nodes a decision tree can have, tuning the complexity of a tree. Feature_fraction tunes the fraction of features randomly sampled for training each tree in a model. Bagging_fraction tunes the proportion of training data to be used in each boosting iteration. Min_data_in_leaf sets a minimum threshold for the number of data points in a leaf node.)

All model fitting, hyperparameter tuning, and cross-validation were conducted strictly on the training set to prevent information leakage. Performance on the held-out test set was used to provide an unbiased estimate of each model's generalizability. To further evaluate how well predicted probabilities corresponded to observed use, calibration curves were generated for each substance and age group to compare mean predicted probabilities with observed substance use rates. Exhibit G.1 displays the code used to generate a calibration curve for the LASSO model for the 12-13 year old age group.

Finally, patterns in influential independent variables were identified across different age groups and substances. The feature importance of each variable was derived using the absolute value of standardized coefficients for LASSO, mean decrease in impurity for Random Forest, and gain-based importance scores for LightGBM.

RESULTS

The predictive performance of each model generally improved with age, reflecting higher prevalence and more stable behavioral patterns among older adolescents. While use of specific drugs didn't change across age groups, drug use consistently increased. Overall, tree-based methods demonstrated superior performance relative to LASSO, highlighting the importance of modeling nonlinear relationships and interactions in this context.

There were a few patterns across age groups that were identified by multiple models. The association between substance use and residential instability tended to decrease with age, as well as the association between substance use and whether a respondent was covered by Medicaid or CHIP. The association between substance use and sexual orientation tended to increase with age.

Random Forest Results

Random Forest returns the magnitude of variable importance scores, which measure the contribution of each variable to the model's predictive power. Each score corresponds to the relative contribution of the variable, based on the variable of the most importance. The scores reflect the contribution of each variable to reducing classification error across the ensemble and should be interpreted as relative predictive relevance rather than directional or causal effects. In addition, precision, recall, F1 scores, and AUPRC scores were also generated with a threshold that maximized F1.

The hyperparameters used were mtry, min.node.size, sample.fraction, num.trees, importance, and probability. Mtry, min.node.size, and sample.fraction were tuned using Bayesian optimization. Num.trees was fixed at 500 trees in the forest. Importance was set to "impurity" so the model assessed mean decrease in impurity to measure feature importance. Probability was set to TRUE so the model outputted the predicted probability of an observation belonging to each possible class instead of just the class label. The model was trained with case weights to handle class imbalance, assigning greater weight to the minority class.

The Random Forest model with hyperparameter tuning proved to yield the best predictive results, with the highest average AUC values and K-S statistics. Most of the AUC values were greater than 0.7. The K-S statistics had a wide range of values, but all of them corresponded to fair predictiveness, and many of them corresponded to excellent predictiveness. While its AUPRC values were not the highest, the positive outcome was rare and AUPRC is sensitive to class imbalance (23). The variables with the highest importance scores are displayed below in Table 1, as Random Forest was the most predictive model.

The variable of a respondent's school attendance had the highest association with substance use. This association was especially evident among older age groups. The increasing importance of school attendance and residential instability in older age groups may reflect greater autonomy and exposure to risk environments as adolescents age. Whether the respondent moved twice or more in the past year and whether the respondent was heterosexual both also highly associated with substance use. For a few age groups and substances, there weren't enough positive variable responses for classification, which are denoted by N/A.

Exhibit H.1, Table H.1, and Exhibit H.2 display the classification code, metrics values, and the code for the metrics evaluation, respectively.

	12-13	14-15	16-17	18-20
cig	ISNATAM_new 0.93 MOVSINPYR2_new 0.87 EDUSCHLGO_new 0.72 CAIDCHIP_new 0.67 ISMETRO_new 0.61	MOVSINPYR2_new 2.00 ISHETERO_new 1.95 EDUSCHLGO_new 1.73 ISMIXED_new 1.44 ISWHITE_new 1.42	EDUSCHLGO_new 3.15 ISHETERO_new 3.02 ISWHITE_new 2.90 IMOTHER_new 2.87 ISMETRO_new 2.43	EDUSCHLGO_new 29.10 ISWHITE_new 9.05 TWENTYK_less 8.72 ISMETRO_new 7.02 MOVSINPYR2_new 6.96
alc	IRSEX_new 12.80 IFATHER_new 10.37 GOVTPROG_new 9.86 CAIDCHIP_new 9.35 ISWHITE_new 9.30	IRSEX_new 3.66 ISHETERO_new 3.46 MOVSINPYR2_new 3.02 ISWHITE_new 2.71 EDUSCHLGO_new 2.67	ISAFRAM_new 6.02 IRSEX_new 5.96 ISWHITE_new 5.02 TWENTYK_less 4.48 IFATHER_new 4.06	EDUSCHLGO_new 43.99 TWENTYK_less 11.59 ISWHITE_new 11.45 IRSEX_new 10.65 CAIDCHIP_new 9.25
mrj	EDUSCHLGO_new 1.68 ISNATAM_new 1.67 CAIDCHIP_new 1.66 SEVENTYFIVEK_less 1.40 MOVSINPYR2_new 1.36	ISHETERO_new 2.98 IFATHER_new 2.97 IRSEX_new 2.51 GOVTPROG_new 2.43 CAIDCHIP_new 2.36	ISHETERO_new 5.97 IMOTHER_new 4.25 IFATHER_new 3.67 EDUSCHLGO_new 3.62 ISMETRO_new 2.86	EDUSCHLGO_new 45.79 ISHETERO_new 10.24 TWENTYK_less 9.00 IRSEX_new 7.02 ISWHITE_new 6.77
inh	MOVSINPYR2_new 2.57 IRSEX_new 2.56 CAIDCHIP_new 2.49 GOVTPROG_new 2.48 IFATHER_new 2.43	GOVTPROG_new 16.19 CAIDCHIP_new 15.10 IRSEX_new 14.86 IFATHER_new 14.16 ISMETRO_new 11.79	ISHETERO_new 1.58 ISNATHI_new 1.21 ISNATAM_new 1.04 EDUSCHLGO_new 1.03 GOVTPROG_new 0.99	ISHETERO_new 3.563 CAIDCHIP_new 1.735 MOVSINPYR2_new 1.550 GOVTPROG_new 1.546 EDUSCHLGO_new 1.477

Table 1

<u>LASSO Results</u>

LASSO was run next to determine the direction of effect sizes for variables. As the L1 penalty shrinks coefficients toward zero, the estimated weights reflect the model's variable-selection mechanism rather than true effect magnitudes. Therefore, inference is based on predictive performance rather than coefficient strength. Variables selected should be interpreted as improving out-of-sample classification rather than as estimates of effect size or causal influence.

All independent variables were standardized to have a mean of 0 and standard deviation of 1 prior to model fitting. As penalization is scale-dependent, standardization ensures comparable penalization across variables. Separate models were fit for each age group and substance to identify the most influential independent variables. The penalty parameter λ was selected using cross validation tuning, choosing the λ that minimized cross-validated deviance ("lambda.min"). Precision, recall, F1 scores, and AUPRC were generated with a threshold that maximized the F1 score.

Most of the AUC values, including confidence intervals, were below 0.7, indicating the limited reliability of this model. In addition, some of the K-S statistics were below 0.2, indicating the poor predictiveness of this model on a few of the age groups and substances. The relatively weak predictive performance likely reflects the complexity of adolescent substance use behavior, which may involve nonlinear relationships and interactions that linear models cannot capture. As a result, while LASSO is strong in terms of interpretability and variable selection, it may sacrifice predictive accuracy in this setting.

The variables that were found to have the most association were whether the respondent was currently attending school, whether they had moved twice or more in the past year, and whether they were of Native American or African American descent. For a few age groups and substances, there either weren't enough positive variable classes for regression or LASSO penalized all the variable coefficients to 0, which are denoted by N/A.

Exhibit G.2, Table G.1, Table G.2, and Exhibit G.3 display the classification code, metrics values for each age and substance, variables with the highest association with substance use, and the code for the metrics evaluation, respectively.

<u>LightGBM Results</u>

LightGBM returns feature importance values that measure how much the feature improved model accuracy (15). It grows trees and splits while choosing the split with the greatest reduction in error. This error reduction is measured by the gain, which is returned as feature importance. In addition, precision, recall, F1 scores, and AUPRC were also generated with a threshold that maximized F1.

The hyperparameters used were nrounds, num_leaves, min_data_in_leaf, bagging_fraction, feature_fraction, max_depth, learning_rate, boosting type, and class imbalance handling. Num_leaves, min_data_in_leaf, bagging_fraction, and feature_fraction were tuned using Bayesian optimization. Nrounds was set to 500 with early stopping allowed. Learning_rate was fixed at 0.05. Boosting type was set to "gbdt" (gradient boosting decision tree). Class imbalance was handled with scale_pos_weight which assigns greater weight to the minority class.

Most of the AUC values were below 0.7, indicating the limited reliability of this model. In addition, some of the K-S statistics were below 0.2, indicating the poor predictiveness of this model on a few of the age groups and substances. Despite its flexibility, LightGBM did not consistently outperform Random Forest in this study. This may be due to the relatively modest sample sizes within each age–substance subgroup and the high degree of class imbalance, which can limit the benefits of boosting-based methods.

The variable that was found to have the most association with substance use was whether the respondent was covered by Medicaid or CHIP. Medicaid/CHIP coverage emerged as a highly associated variable in several models, potentially reflecting broader socioeconomic vulnerability. Other highly associated variables included whether the respondent was currently attending school, whether or not the respondent was heterosexual, and the respondent's sex at birth. For a few age groups and substances, there weren't enough positive variable responses for classification, which are denoted by N/A.

Exhibit I.1, Table I.1, Table I.2, and Exhibit I.2 display the classification code, metrics values, variables with the highest association with substance use, and the code for the metrics evaluation, respectively.

Discussion

Variables correlated with adolescent substance use

The top three variables identified by Random Forest were the respondent's school attendance, the number of times the respondent moved in the past year, and the respondent's sexual orientation.

The variables identified in this study are consistent with existing findings reported on adolescent substance use. This model indicated that low school attendance is correlated with substance use. This is consistent with an association between school membership and low risk for smoking, drinking, and cannabis use (27). From a developmental and social-control perspective, school attendance may serve as a proxy for structured daily routines, adult supervision, peer norms, and access to institutional support. Adolescents who are disengaged from school may have greater exposure to unstructured social environments, increased stress, and fewer protective social bonds, all of which are associated with elevated substance use risk (28). Importantly, school attendance should not be interpreted as a causal mechanism, but rather as an indicator of broader social integration.

The model also found a correlation between a respondent having moved twice or more in the past year with substance use. This is in line with highly significant positive relationships between early geographic relocation and use of illicit substances, such as marijuana and hallucinogens, as well as illicit use of prescription drugs (12). Residential instability may disrupt peer relationships, reduce continuity of social support, and increase exposure to stressors associated with housing insecurity. These disruptions may contribute to substance use as a coping mechanism or through increased exposure to new peer networks where substance use is more prevalent (12). Mobility may therefore function as a marker of broader socioeconomic and family instability rather than an independent risk factor.

Finally, the model found a correlation between sexual orientation and substance use. This is consistent with developmental disparities in substance use for sexual and gender minority adolescents compared with heterosexual and cisgender adolescents. These disparities were present by age 12 and persisted to age 18 and older (29). Importantly, sexual orientation itself should not be interpreted as a risk factor; rather, it likely captures exposure to structural and psychosocial stressors that increase vulnerability to substance use.

Predictive performance and variable importance varied across age groups, suggesting meaningful developmental differences in substance use risk. In younger adolescents, model performance was generally lower, likely reflecting lower rates of substance use and more limited behavioral differentiation. In contrast, predictive accuracy improved in older age groups, where substance use becomes more prevalent and socio-environmental factors such as school engagement and residential instability may exert stronger influence.

From a public health perspective, the predictive factors identified in this study highlight opportunities for population-level intervention. Variables such as school attendance and residential mobility are observable and potentially actionable within educational and community systems. Rather than targeting individuals based on immutable characteristics, these findings suggest that strengthening school engagement and providing additional support to highly mobile adolescents may reduce substance use risk at the population level.

<u>Model verification</u>

To verify that Random Forest was the model with the best predictive power, the AUC values, AUPRC values, and K-S statistics for each model were pooled into one average value per model for comparison. Table J.1 displays the average AUC, AUPRC, and K-S statistic per model (among all substances and age groups) to compare.

Limitations and assumptions

One limitation of the dataset was the absence of quantitative behavioral and psychological variables measuring peer influence, trauma history, or mental health measures, which are important factors in substance use but were not assessed in this survey.

The responses may have been influenced by certain factors that may have resulted in inaccuracies. Cultural and/or social bias may have played a part, leading respondents to answer based on what they perceived to be cultural/social norms, instead of the truth. In addition, there may have been nonresponse bias, as respondents may not have wanted to share their true beliefs and, as a result, abstained from responding. The interview response rate was 24.36%, indicating potential sampling bias. Finally, respondents may have forgotten about past substance use or been unsure about their answers to certain demographic questions.

It was assumed that the survey answers constituted a random sample of American adolescents and adults. Similarly, after cleaning the dataset and selecting only the answers of respondents aged 12-20, it was assumed that this smaller dataset was representative of all American adolescents. In addition, it was assumed that all respondents were given enough privacy to answer independently and without pressure from others.

Due to the imbalanced nature of the dataset, accuracy was not used to assess model performance. Accuracy measures the fraction of correctly classified samples. However, for an imbalanced dataset, the model can predict the majority class every time and still get high accuracy because most of the responses belong to the majority class anyway. Therefore, accuracy doesn't reflect model performance on the minority class well, especially for imbalanced datasets (30). The formula for accuracy is displayed in Exhibit K.1.

A further limitation related to class imbalance is model instability for substances with extremely low prevalence rates. In these cases, models may become highly sensitive to small changes in the training data, leading to unstable estimates of performance metrics and feature importance rankings. As a result, findings for very low-prevalence substances should be interpreted cautiously and viewed as exploratory rather than definitive.

Separate models were trained for each substance within each age group. No formal hypothesis testing was conducted; therefore, traditional multi-comparison corrections (e.g. Bonferroni or FDR) are not directly applicable. Instead, model comparisons rely on out-of-sample predictive performance, which is not affected by multiple hypothesis testing.

All independent variables were binarized or collapsed into broad categories to ensure sufficient sample sizes within each class and to facilitate model training. However, this process introduced important limitations. Complex, multidimensional constructs such as race and sexual orientation were reduced to simplified binary indicators, which may have obscured intra-group differences. This simplification could have potentially overstated the apparent importance of these variables by masking structural or contextual factors correlated with them.

Finally, there are ethical considerations inherent in applying predictive machine learning models to substance use among minors. Although this study is intended to identify population-level risk patterns rather than to predict individual behavior, models that associate demographic or social characteristics with substance use outcomes risk stigmatization if misapplied. Feature importance rankings may be misconstrued as causal or deterministic. For this reason, results should be interpreted as tools for informing public health understanding and prevention strategies, not for individual-level screening.

CONCLUSION

This study compared three machine learning models for their predictive accuracy in identifying adolescent substance use. This study also identified the most influential risk factors of adolescent substance use. The Random Forest model performed the best, compared to LASSO and LightGBM. Hyperparameter tuning was used for each model to enhance its predictive accuracy. Stratified k-fold cross-validation was used to create sample folds with proportions representative of the whole dataset, due to theL1 imbalanced nature of the original dataset. Area Under the Curve (AUC) values, Area Under the Precision-Recall Curve (AUPRC) values, and Kolmogorov-Smirnov (K-S) statistics were used to evaluate the predictive accuracy of each model. The top factors returned by Random Forest were school attendance, geographic relocation, and sexual orientation.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. Nick Pretnar, for his advice and time spent helping me improve my project.

REFERENCES

(1): Swadi, H. (1999). *Individual risk factors for adolescent substance use*. Drug and Alcohol Dependence, 55(3), 209-224. https://doi.org/10.1016/S0376-8716(99)00017-4

(2): Steinfeld, M.R., Torregrossa, M.M. (2023). *Consequences of adolescent drug use*. Translational psychiatry, 13(1), 313. https://doi.org/10.1038/s41398-023-02590-4

(3): Goldstick, J.E., Cunningham, R.M., Carter, P.M. (2022). *Current Causes of Death in Children and Adolescents in the United States*. New England Journal of Medicine, 386(21), 1955-1956. https://doi.org/10.1056/NEJMc2201761

(4): Substance Abuse and Mental Health Services Administration. (2019). *Key Substance Use and Mental Health Indicators in the United States: Results from the 2019 National Survey on Drug Use and Health*. https://www.samhsa.gov/data/sites/default/files/reports/rpt29393/2019NSDUHFFRPDFWHTML/2019NSDUHFFR1PDFW090120.pdf

(5): Hastie, T., Tibshirani, R.,, Friedman, J. (2001). *The Elements of Statistical Learning*. https://doi.org/10.1111/j.1541-0420.2010.01516.x

(6): Hammond, C.J., Mayes, L.C., Potenza, M.N. (2015). *Neurobiology of Adolescent Substance Use and Addictive Behaviors: Prevention and Treatment Implications*. Adolescent medicine: state of the art reviews, 25(1), 15-32. https://pmc.ncbi.nlm.nih.gov/articles/PMC4446977/

(7): Simon, K.M., Levy, S.J., Bukstein, O.G. (2024). *Adolescent Substance Use Disorders*. New England Journal of Medicine, 1(6). https://doi.org/10.1056/EVIDra2200051

(8): Martínez-Loredo, V., Fernández-Artamendi, S., Weidberg, S., Pericot, I., López-Núñez, C., Fernández-Hermida, J. R., Secades, R. (2016). *Parenting styles and alcohol use among adolescents: A longitudinal study*. European Journal of Investigation in Health, Psychology and Education, 6(1), 27-36. https://doi.org/10.3390/ejihpe6010003

(9): Luk, J. W., King, K.M., McCarty, C.A., McCauley, E., Vander Stoep, A. (2017). *Prospective Effects of Parenting on Substance Use and Problems Across Asian/Pacific Islander and European American Youth: Tests of Moderated Mediation*. Journal of Studies on Alcohol and Drugs, 78(4), 521-530. https://doi.org/10.15288/jsad.2017.78.521

(10): Patrick, M.E., Wightman, P., Schoeni, R.F., Schulenberg, J.E. (2012). *Socioeconomic Status and Substance Use Among Young Adults: A Comparison Across Constructs and Drugs*. Journal of Studies on Alcohol and Drugs, 73(5), 772-782. https://doi.org/10.15288/jsad.2012.73.772

(11): Lin, K., Twisk, J.W.R., Huang, H. (2012). *Longitudinal Impact of Frequent Geographic Relocation from Adolescence to Adulthood on Psychosocial Stress and Vital Exhaustion at Ages 32 and 42 Years: The Amsterdam Growth and Health Longitudinal Study*. Journal of epidemiology, 22(5), 469–476. https://doi.org/10.2188/jea.je20110141

(12): DeWit, D.J. (1998). *Frequent childhood geographic relocation: Its impact on drug use initiation and the development of alcohol and other drug-related problems among adolescents and young adults*. Addictive Behaviors, 23(5), 623-634. https://doi.org/10.1016/S0306-4603(98)00023-9

(13): Tibshirani, R. (1996). *Regression Shrinkage and Selection Via the Lasso*. Journal of the Royal Statistical Society: Series B (Methodological). 58(1), 267–288. https://academic.oup.com/jrsssb/article/58/1/267/7027929

(14): Breiman, L. (2001). *Random Forests*. Machine Learning. 45, 5–32. https://doi.org/10.1023/A:1010933404324

(15): Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

(16): Substance Abuse and Mental Health Services Administration. (2023). *2023 National Survey on Drug Use and Health Releases*. https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/national-releases/2023

(17): Novacek, J., Raskin, R., Hogan, R. (1991). *Why do adolescents use drugs? Age, sex, and user differences*. J Youth Adolescence 20, 475–492. https://doi.org/10.1007/BF01540632

(18): Hardy, M. A. (1993). *Regression with dummy variables*. Quantitative applications in the social sciences, No. 93. SAGE Publications. https://us.sagepub.com/sites/default/files/upm-binaries/21120_Chapter_7.pdf?utm_source=chatgpt.com

(19): Glasheen, C., Forman-Hoffman, V., Hedden, S., Ridenour, T., Wang, J., Porter, J. (2019). *Residential transience among U.S. adolescents: association with depression and mental health treatment*. Epidemiology and Psychiatric Sciences, 28(6), 682-691. https://doi.org/10.1017/S2045796018000823

(20): Jun, H., Webb-Morgan, M., Felner, J.K., Wisdom, J.P., Haley, S.J., Austin, S.B., Katuska, L.M., Corliss, H.L. (2019). *Sexual orientation and gender identity disparities in substance use disorders during young adulthood in a United States longitudinal cohort*. Drug and Alcohol Dependence, 205, 107619. https://doi.org/10.1016/j.drugalcdep.2019.107619

(21): Shaik, A.B., Srinivasan, S. (2018). *A Brief Survey on Random Forest Ensembles in Classification Model*. International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, 56. https://doi.org/10.1007/978-981-13-2354-6_27

(22): Carter, J.V., Pan, J., Rai, S.N., Galandiuk, S. (2016). *ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves*. Surgery, 159(6), 1638-1645. https://www.surgjournal.com/article/S0039-6060(16)00066-0/fulltext

(23): McDermott, M.B., Zhang, H., Hansen, L.H., Angelotti, G., Gallifant, J. (2024). *A Closer Look at AUROC and AUPRC under Class Imbalance*. 38th Conference on Neural Information Processing Systems, 44102-44163. https://doi.org/10.52202/079017-1400

(24): Kirkman, T.W. (1996). *Kolmogorov-Smirnov Test*. http://www.physics.csbsju.edu/stats/KS-test.html

(25): Harris, J.K. (2021). *Primer on binary logistic regression*. Family medicine and community health. https://doi.org/10.1136/fmch-2021-001290

(26): Chen, P., Jacobson, K.C. (2012). *Developmental trajectories of substance use from early adolescence to young adulthood: gender and racial/ethnic differences*. The Journal of adolescent health : official publication of the Society for Adolescent Medicine, 50(2), 154–163. https://doi.org/10.1016/j.jadohealth.2011.05.013

(27): Gaete, J., Rojas, G., Fritsch, R., Araya, R. (2018). *Association between School Membership and Substance Use among Adolescents*. Frontiers in Psychiatry, 9. https://doi.org/10.3389/fpsyt.2018.00025

(28): Bond, L., Butler, H., Thomas, L., Carlin, J., Glover, S., Bowes, G., Patton, G. (2007). *Social and School Connectedness in Early Secondary School as Predictors of Late Teenage Substance Use, Mental Health, and Academic Outcomes*. Journal of Adolescent Health, 40(4), 9-18. https://doi.org/10.1016/j.jadohealth.2006.10.013

(29): Fish, J.N., Bishop, M.D., Russell, S.T. (2021). *Developmental Differences in Sexual Orientation and Gender Identity-Related Substance Use Disparities: Findings from Population-Based Data*. Journal of Adolescent Health, 68(6), 1162-1169. https://doi.org/10.1016/j.jadohealth.2020.10.023

(30): Fawcett, T. (2006). *An introduction to ROC analysis*. Pattern Recognition Letters, 27(7), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

(31): Wills, T.A., Yaeger, A.M. (2003). *Family Factors and Adolescent Substance Use: Models and Mechanisms*. Current Directions in Psychological Science, 12(6), 222-226. https://doi.org/10.1046/j.0963-7214.2003.01266.x

(32): McDermott, M.J., Drescher, C.F., Smitherman, T.A., Tull, M.T., Heiden, L., Damon, J.D., Hight, T.L., Young, J. (2013). *Prevalence and Sociodemographic Correlates of Lifetime Substance Use among a Rural and Diverse Sample of Adolescents*. Substance Use, 34(4), 371-380. https://doi.org/10.1080/08897077.2013.776000

Appendix A - dependent variable treatments

Variable name	Variable responses	Variable definition	Variable usage	New variable name (N/A if inapplicable)
AGE3	1 = 12-13 years old 2 = 14-15 years old 3 = 16-17 years old 4 = 18-20 years old	Recode - final edited age Recoded variables are variables created using one or more edited or imputed (variables that have missing data replaced with nonmissing values using statistical imputation procedures) source variables. These variables are often the variables used in final analysis. The recoded variables in this dataset will be preceded with the code “RC”.	Divided all responses into one of four AGE3 values; evaluated substance use separately by age group	N/A
MOVSINPYR2	0 = 0 times 1 = 1 time 2 = 2 times 3 = 3+ times	Number of times moved in past year - recoded	Binarized the variable: 0, 1 -> 0 2, 3 -> 1	MOVSINPYR2_new
COUTYP4	1 = large metro 2 = small metro 3 = nonmetro	County metro/nonmetro status	Binarized the variable: 3 -> 0 1, 2 -> 1	ISMETRO_new
IRSEX	1 = male 2 = female	Sex at birth - imputation revised Missing values for this question were not permitted. The variable has the prefix IR for consistency with surveys prior to 2002 where missing values were permitted.	Binarized the variable: 1 -> 0 2 -> 1	IRSEX_new

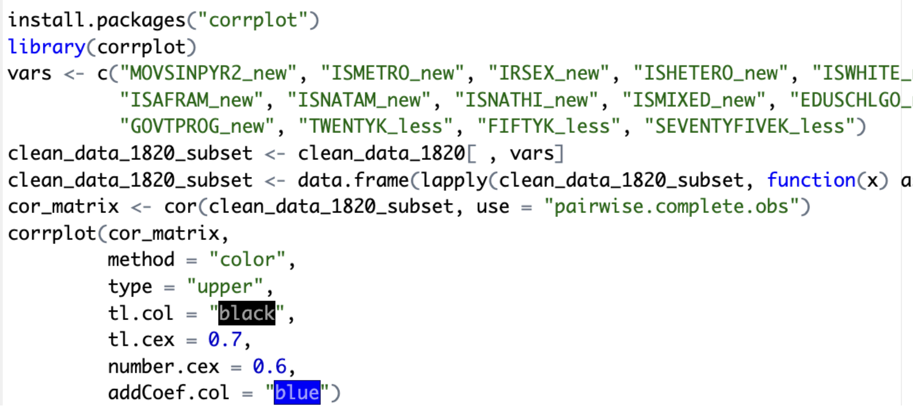
SEXATRACT2	1 = only attracted to opposite sex 2 = mostly attracted to opposite sex 3 = equally attracted to males and females 4 = mostly attracted to same sex 5 = only attracted to same sex 6 = I am not sure	Sexual attraction	Binarized the variable: 2, 3, 4, 5, 6 -> 0 1 -> 1	ISHETERO_new
NEWRACE2	1 = NonHisp white 2 = NonHisp black/Afr Am 3 = NonHisp native Am/AK native 4 = NonHisp native HI/other Pac Isl 5 = NonHisp	RC - race/hispanicity recode (7 levels)	Separated the variable into 6 binary dummy variables	ISWHITE_new, ISAFRAM_new, ISNATAM_new, ISNATHI_new, ISASIAN_new, ISMIXED_new

	Asian 6 = NonHisp more than one race 7 = Hispanic			
EDUSCHLGO	1 = yes 2 = no	Now going to school	Binarized the variable: 2 -> 0 1 -> 1	EDUSCHLGO_new
IMOTHER	1 = respondent is 12-17, mother in household 2 = respondent is 12-17, no mother in household 4 = respondent is 18 or older	RC - mother in household	Binarized the variable: 2 -> 0 1 -> 1 Removed this variable for the 18-20 dataset	IMOTHER_new
IFATHER	1 = respondent is 12-17, father in household 2 = respondent is 12-17, no father in household 4 =	RC - father in household	Binarized the variable: 2 -> 0 1 -> 1 Removed this variable for the 18-20 dataset	IFATHER_new

	respondent is 18 or older			
CAIDCHIP	1 = yes 2 = no	Covered by Medicaid/CHIP	Binarized the variable: 2 -> 0 1 -> 1	CAIDCHIP_new
GOVTPROG	1 = yes 2 = no	RC - participated in one or more government assistance programs	Binarized the variable: 2 -> 0 1 -> 1	GOVTPROG_new
INCOME	1 = less than $20,000 2 = $20,000 - $49,999 3 = $50,000 - $74,999 4 = $75,000 or more	RC - total family income recode	Separated the variable into 3 binary dummy variables	TWENTYK_less, FIFTYK_less, SEVENTYFIVEK_less

Table A.1

Appendix B - Pearson correlation coefficient heatmap information



```r
install.packages("corrplot")
library(corrplot)
vars <- c("MOVSINPYR2_new", "ISMETRO_new", "IRSEX_new", "ISHETERO_new", "ISWHITE_new", "ISASIAN_new",
          "ISAFRAM_new", "ISNATAM_new", "ISNATHI_new", "ISMIXED_new", "EDUSCHLGO_new", "CAIDCHIP_new",
          "GOVTPROG_new", "TWENTYK_less", "FIFTYK_less", "SEVENTYFIVEK_less")
clean_data_1820_subset <- clean_data_1820[ , vars]
clean_data_1820_subset <- data.frame(lapply(clean_data_1820_subset, function(x) as.numeric(as.character(x))))
cor_matrix <- cor(clean_data_1820_subset, use = "pairwise.complete.obs")
corrplot(cor_matrix,
         method = "color",
         type = "upper",
         tl.col = "black",
         tl.cex = 0.7,
         number.cex = 0.6,
         addCoef.col = "blue")
```

Exhibit B.1

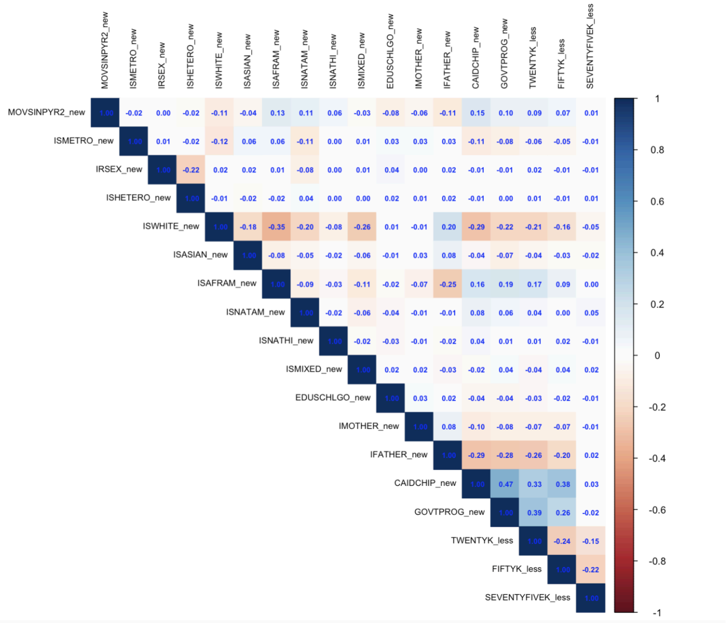


Figure B.1

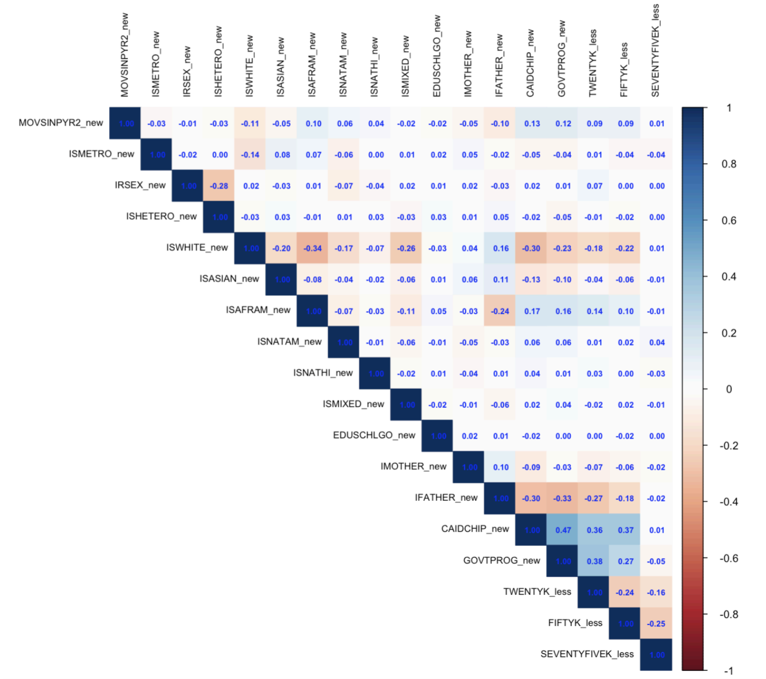


Figure B.2

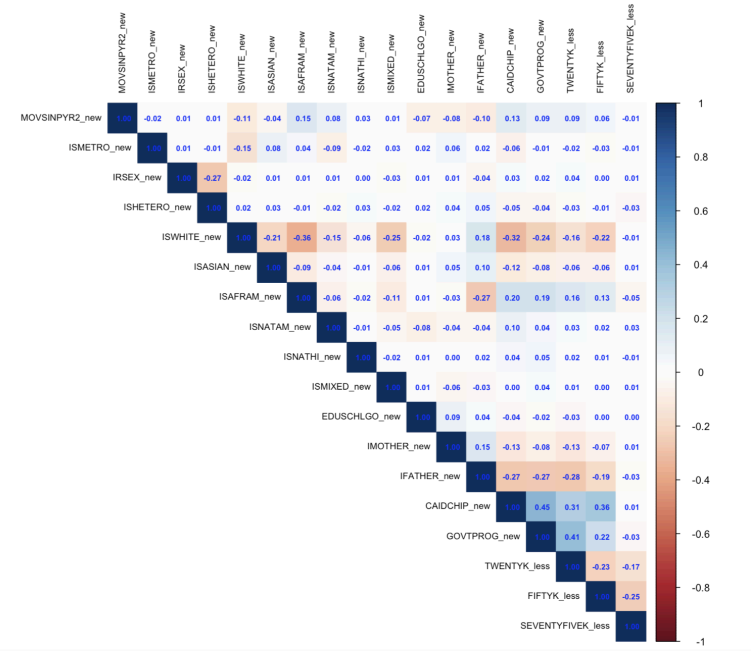


Figure B.3

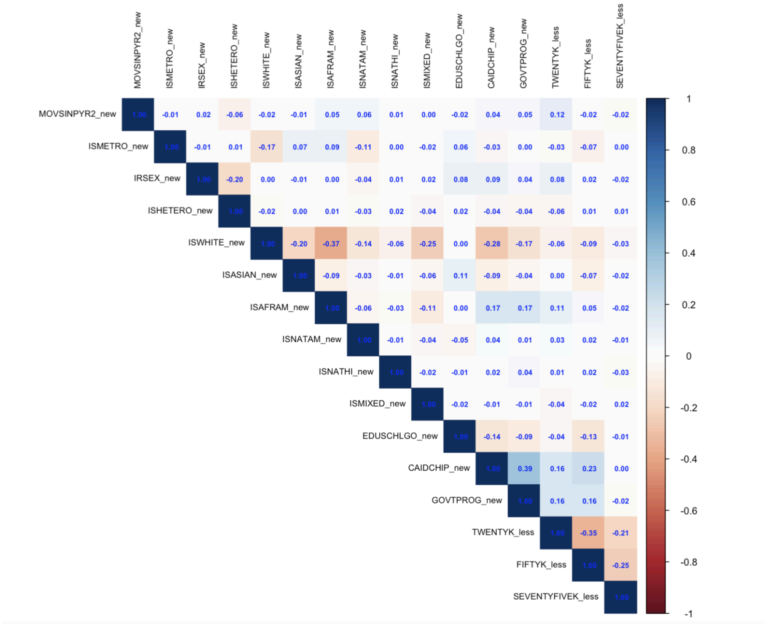


Figure B.4

Appendix C - original substances and their target rates

Substance	Age	# of users	Total #	Target rate
Cigarettes	12-13	87	2,775	3.1%
	14-15	234	3,392	6.9%
	16-17	407	3,252	12.5%
	18-20	900	3,735	24.1%
Alcohol	12-13	234	2,775	8.4%
	14-15	725	3,392	21.4%
	16-17	1,254	3,252	38.6%
	18-20	2,127	3,735	56.9%
Marijuana	12-13	104	2,775	3.7%
	14-15	470	3,392	13.9%

	16-17	893	3,252	27.5%
	18-20	1,542	3,735	41.3%
Cocaine	12-13	1	2,775	0.0%
	14-15	6	3,392	0.2%
	16-17	24	3,252	0.7%
	18-20	129	3,735	3.5%
Heroin	12-13	0	2,775	0.0%
	14-15	0	3,392	0.0%
	16-17	2	3,252	0.0%
	18-20	13	3,735	0.3%
Hallucinogens	12-13	18	2,775	0.6%
	14-15	64	3,392	1.9%
	16-17	138	3,252	4.2%
	18-20	376	3,735	10.1%
Inhalants	12-13	179	2,775	6.5%
	14-15	256	3,392	7.5%
	16-17	248	3,252	7.6%
	18-20	252	3,735	6.7%
Methamphetamines	12-13	0	2,775	0.0%
	14-15	5	3,392	0.1%
	16-17	15	3,252	0.5%
	18-20	41	3,735	1.1%
Pain relievers	12-13	70	2,775	2.5%

	14-15	106	3,392	3.1%
	16-17	133	3.252	4.1%
	18-20	157	3,735	4.2%
Tranquilizers	12-13	6	2,775	0.2%
	14-15	26	3,392	0.8%
	16-17	47	3,252	1.4%
	18-20	70	3,735	1.9%
Stimulants	12-13	14	2,775	0.5%
	14-15	42	3,392	1.2%
	16-17	52	3,252	1.6%
	18-20	106	3,735	2.8%
Sedatives	12-13	5	2,775	0.2%
	14-15	7	3,392	0.2%
	16-17	19	3,252	0.6%
	18-20	24	3,735	0.6%

Table C.1

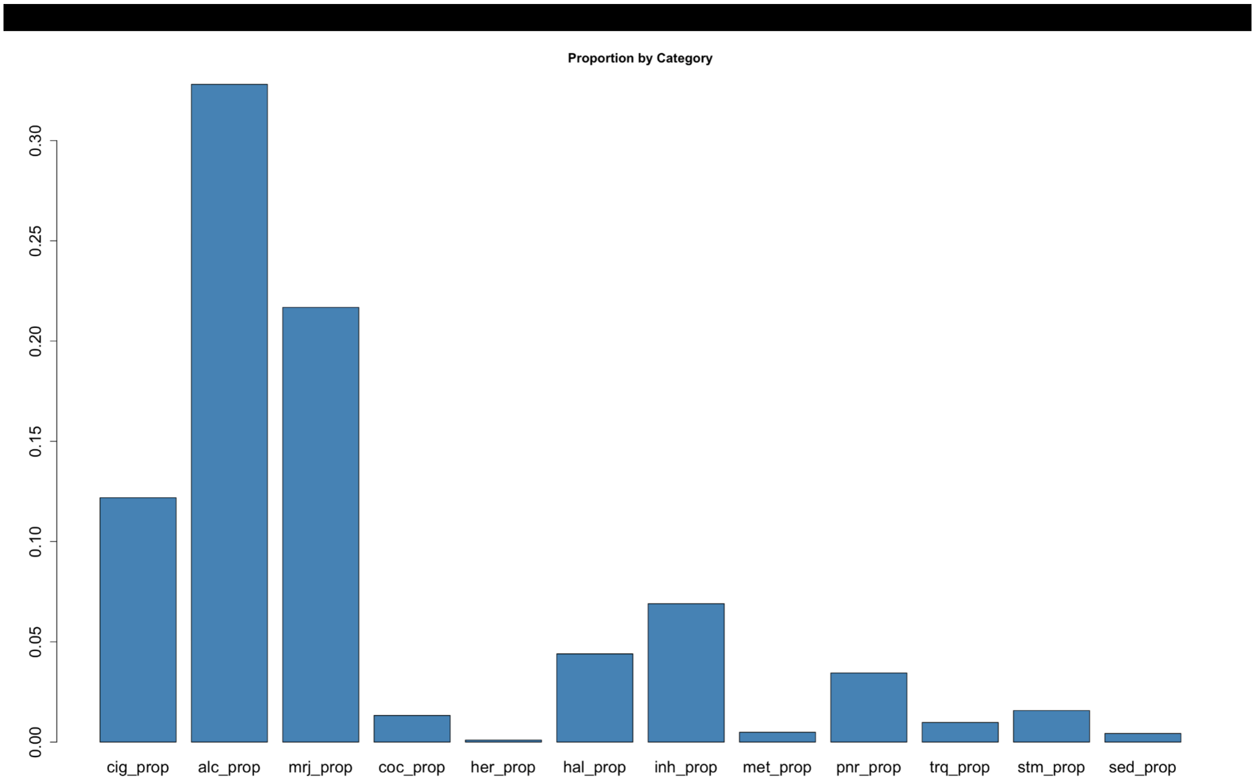
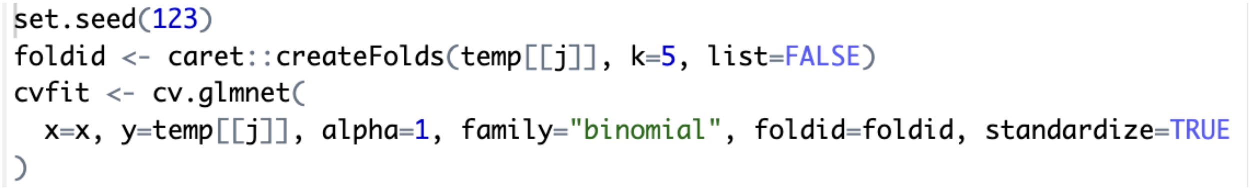


Figure C.1

Appendix D - LASSO stratified 5-fold cross validation code



```
set.seed(123)
foldid <- caret::createFolds(temp[[j]], k=5, list=FALSE)
cvfit <- cv.glmnet(
  x=x, y=temp[[j]], alpha=1, family="binomial", foldid=foldid, standardize=TRUE
)
```

Exhibit D.1

Appendix E - Kolmogorov-Smirnov statistic ranges

Kolmogorov-Smirnov statistic	Model predictiveness
K-S < 0.2	poor
0.2 < K-S < 0.4	fair

0.4 < K-S < 0.6	good
KS > 0.6	excellent

Table E.1

Appendix F - detailed machine learning explanation

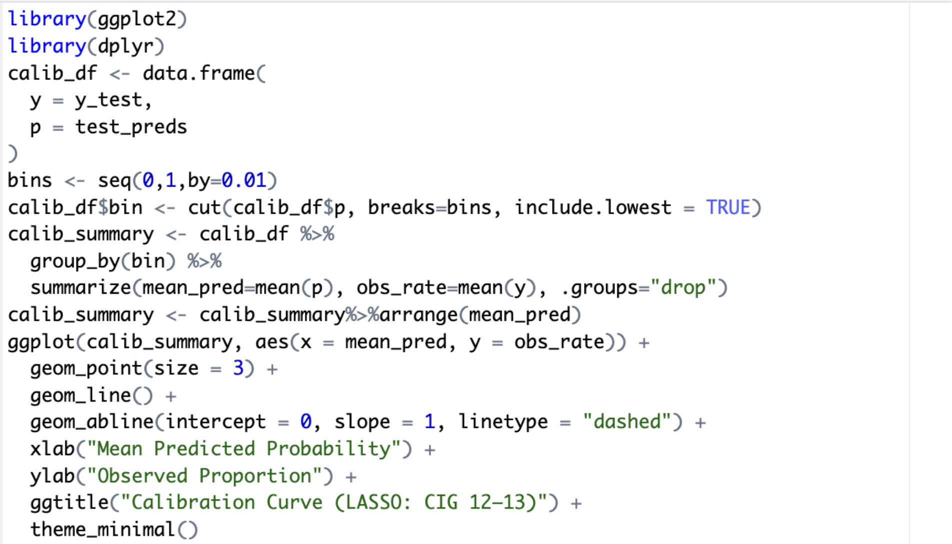
AUPRC, precision, recall, and F1 score complement the AUC by providing more informative performance estimates when the positive class is rare. With threshold adjustment, a decision threshold was selected that maximized the F1 score on the held-out test set to achieve the best trade-off between identifying positive cases and avoiding excessive false positives.

Exhibit F.1

LASSO was tuned using the lambda hyperparameter. For each lambda (regularization strength), glmnet fitted the model and evaluated the cross-validated deviance. It then selected the lambda with the lowest mean cross-validated error. Random Forest was tuned using the mtry, min.node.size, and sample.fraction hyperparameters with Bayesian optimization. Mtry specifies the number of variables randomly sampled as candidates at each split. Min.node.size tunes the minimum number of observations allowed in a leaf node of a decision tree, which controls the tree's depth. Sample.fraction determines the fraction of data rows to sample with replacement when building each tree. LightGBM was tuned using num_leaves, feature_fraction, bagging_fraction, min_data_in_leaf. Num_leaves controls the maximum number of leaf nodes a decision tree can have, tuning the complexity of a tree. Feature_fraction tunes the fraction of features randomly sampled for training each tree in a model. Bagging_fraction tunes the proportion of training data to be used in each boosting iteration. Min_data_in_leaf sets a minimum threshold for the number of data points in a leaf node.

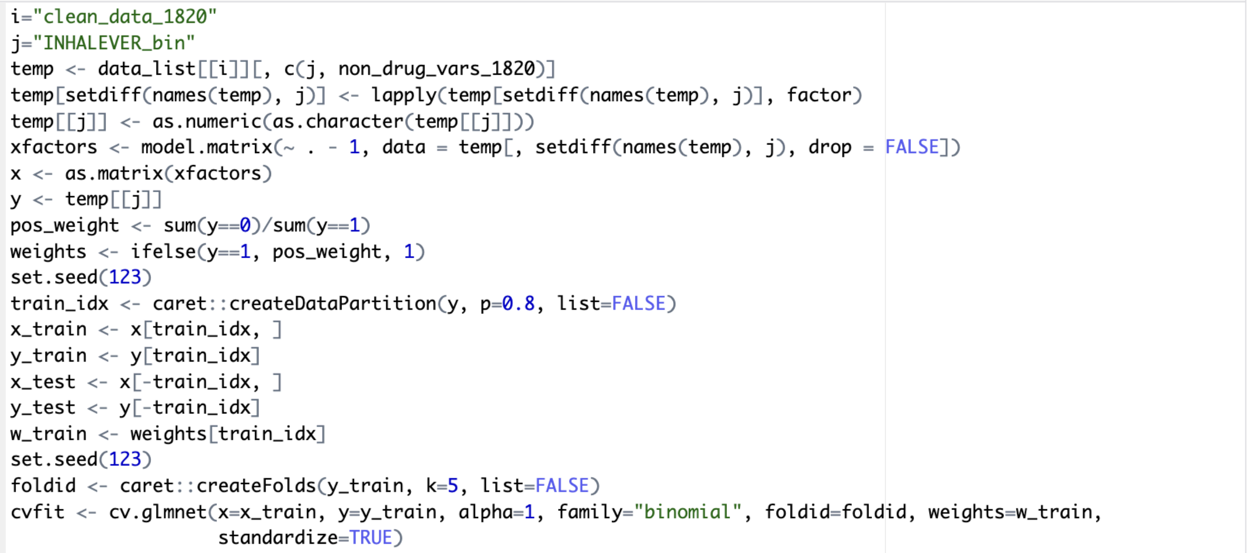
Exhibit F.2

Appendix G - LASSO results



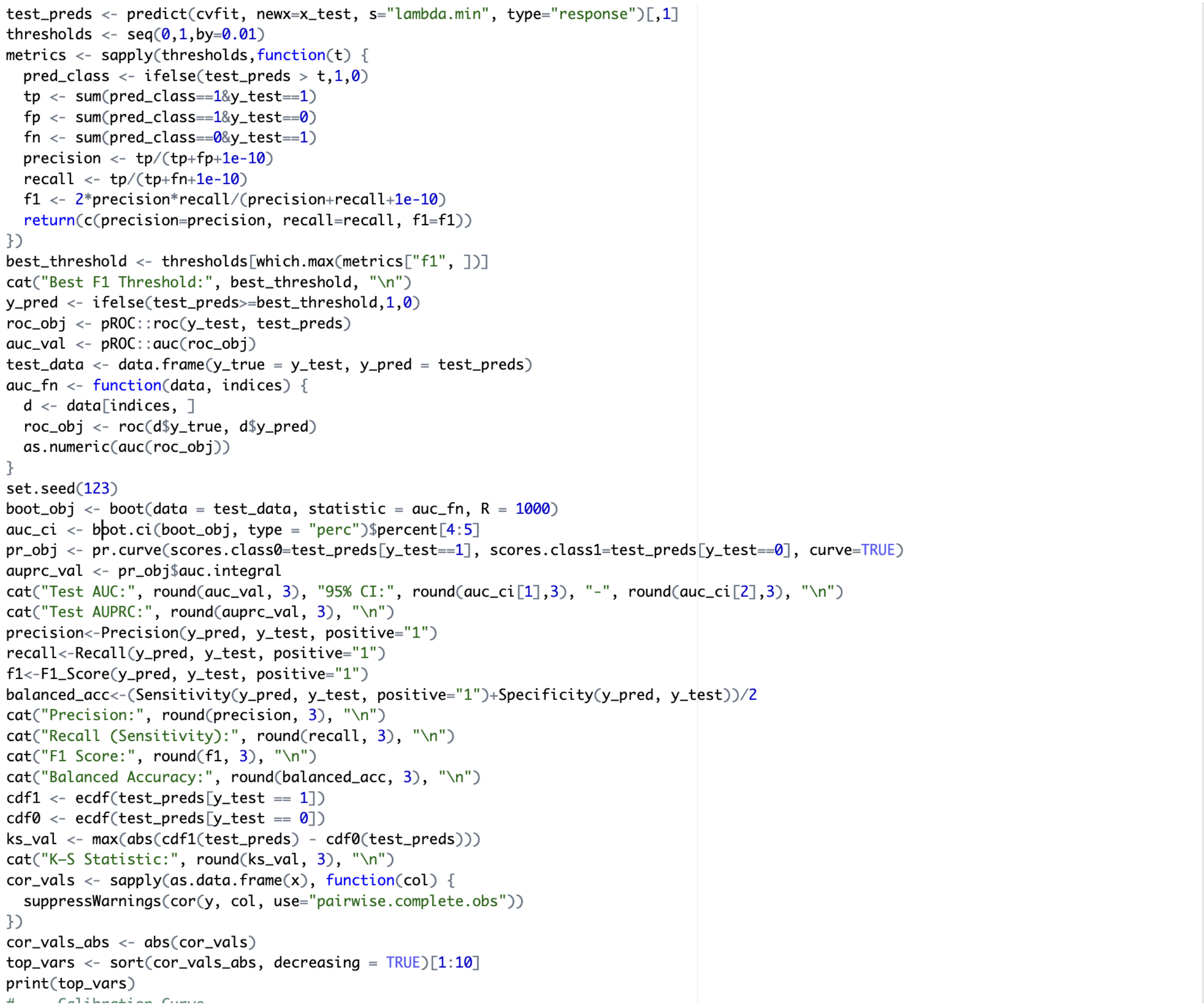
```
library(ggplot2)
library(dplyr)
calib_df <- data.frame(
  y = y_test,
  p = test_preds
)
bins <- seq(0,1,by=0.01)
calib_df$bin <- cut(calib_df$p, breaks=bins, include.lowest = TRUE)
calib_summary <- calib_df %>%
  group_by(bin) %>%
  summarize(mean_pred=mean(p), obs_rate=mean(y), .groups="drop")
calib_summary <- calib_summary%>%arrange(mean_pred)
ggplot(calib_summary, aes(x = mean_pred, y = obs_rate)) +
  geom_point(size = 3) +
  geom_line() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  xlab("Mean Predicted Probability") +
  ylab("Observed Proportion") +
  ggtitle("Calibration Curve (LASSO: CIG 12-13)") +
  theme_minimal()
```

Exhibit G.1



```
i="clean_data_1820"
j="INHALEVER_bin"
temp <- data_list[[i]][, c(j, non_drug_vars_1820)]
temp[setdiff(names(temp), j)] <- lapply(temp[setdiff(names(temp), j)], factor)
temp[[j]] <- as.numeric(as.character(temp[[j]]))
xfactors <- model.matrix(~ . - 1, data = temp[, setdiff(names(temp), j), drop = FALSE])
x <- as.matrix(xfactors)
y <- temp[[j]]
pos_weight <- sum(y==0)/sum(y==1)
weights <- ifelse(y==1, pos_weight, 1)
set.seed(123)
train_idx <- caret::createDataPartition(y, p=0.8, list=FALSE)
x_train <- x[train_idx, ]
y_train <- y[train_idx]
x_test <- x[-train_idx, ]
y_test <- y[-train_idx]
w_train <- weights[train_idx]
set.seed(123)
foldid <- caret::createFolds(y_train, k=5, list=FALSE)
cvfit <- cv.glmnet(x=x_train, y=y_train, alpha=1, family="binomial", foldid=foldid, weights=w_train,
                   standardize=TRUE)
```

Exhibit G.2



```
test_preds <- predict(cvfit, newx=x_test, s="lambda.min", type="response")[,1]
thresholds <- seq(0,1,by=0.01)
metrics <- sapply(thresholds,function(t) {
  pred_class <- ifelse(test_preds > t,1,0)
  tp <- sum(pred_class==1&y_test==1)
  fp <- sum(pred_class==1&y_test==0)
  fn <- sum(pred_class==0&y_test==1)
  precision <- tp/(tp+fp+1e-10)
  recall <- tp/(tp+fn+1e-10)
  f1 <- 2*precision*recall/(precision+recall+1e-10)
  return(c(precision=precision, recall=recall, f1=f1))
})
best_threshold <- thresholds[which.max(metrics["f1", ])]
cat("Best F1 Threshold:", best_threshold, "\n")
y_pred <- ifelse(test_preds>=best_threshold,1,0)
roc_obj <- pROC::roc(y_test, test_preds)
auc_val <- pROC::auc(roc_obj)
test_data <- data.frame(y_true = y_test, y_pred = test_preds)
auc_fn <- function(data, indices) {
  d <- data[indices, ]
  roc_obj <- roc(d$y_true, d$y_pred)
  as.numeric(auc(roc_obj))
}
set.seed(123)
boot_obj <- boot(data = test_data, statistic = auc_fn, R = 1000)
auc_ci <- boot.ci(boot_obj, type = "perc")$percent[4:5]
pr_obj <- pr.curve(scores.class0=test_preds[y_test==1], scores.class1=test_preds[y_test==0], curve=TRUE)
auprc_val <- pr_obj$auc.integral
cat("Test AUC:", round(auc_val, 3), "95% CI:", round(auc_ci[1],3), "-", round(auc_ci[2],3), "\n")
cat("Test AUPRC:", round(auprc_val, 3), "\n")
precision<-Precision(y_pred, y_test, positive="1")
recall<-Recall(y_pred, y_test, positive="1")
f1<-F1_Score(y_pred, y_test, positive="1")
balanced_acc<-(Sensitivity(y_pred, y_test, positive="1")+Specificity(y_pred, y_test))/2
cat("Precision:", round(precision, 3), "\n")
cat("Recall (Sensitivity):", round(recall, 3), "\n")
cat("F1 Score:", round(f1, 3), "\n")
cat("Balanced Accuracy:", round(balanced_acc, 3), "\n")
cdf1 <- ecdf(test_preds[y_test == 1])
cdf0 <- ecdf(test_preds[y_test == 0])
ks_val <- max(abs(cdf1(test_preds) - cdf0(test_preds)))
cat("K-S Statistic:", round(ks_val, 3), "\n")
cor_vals <- sapply(as.data.frame(x), function(col) {
  suppressWarnings(cor(y, col, use="pairwise.complete.obs"))
})
cor_vals_abs <- abs(cor_vals)
top_vars <- sort(cor_vals_abs, decreasing = TRUE)[1:10]
print(top_vars)
#     Calibration Curve
```

Exhibit G.3

	12-13			14-15			16-17			18-20		
	AUC	AUPRC	K-S	AUC	AUPRC	K-S	AUC	AUPRC	K-S	AUC	AUPRC	K-S
cig	0.753 95% CI: 0.631 - 0.853	0.076	0.419	0.696 95% CI: 0.603 - 0.791	0.112	0.338	0.638 95% CI: 0.566 - 0.704	0.236	0.232	0.663 95% CI: 0.627 - 0.695	0.239	0.261

alc	0.631 95% CI: 0.552 - 0.718	0.17	0.157	0.597 95% CI: 0.538 - 0.646	0.305	0.152	0.606 95% CI: 0.562 - 0.649	0.508	0.184	0.6 95% CI: 0.576 - 0.623	0.424	0.17
mrj	0.685 95% CI: 0.585 - 0.785	0.071	0.34	0.582 95% CI: 0.457 - 0.648	0.183	0.211	0.614 95% CI: 0.568 - 0.661	0.406	0.165	0.609 95% CI: 0.583 - 0.636	0.334	0.163
inh	0.444 95% CI: 0.401 - 0.653	0.054	N/A	0.605 95% CI: 0.513 - 0.686	0.099	0.131	0.59 95% CI: 0.507 - 0.67	0.124	0.16	0.523 95% CI: 0.482 - 0.565	0.083	0.23

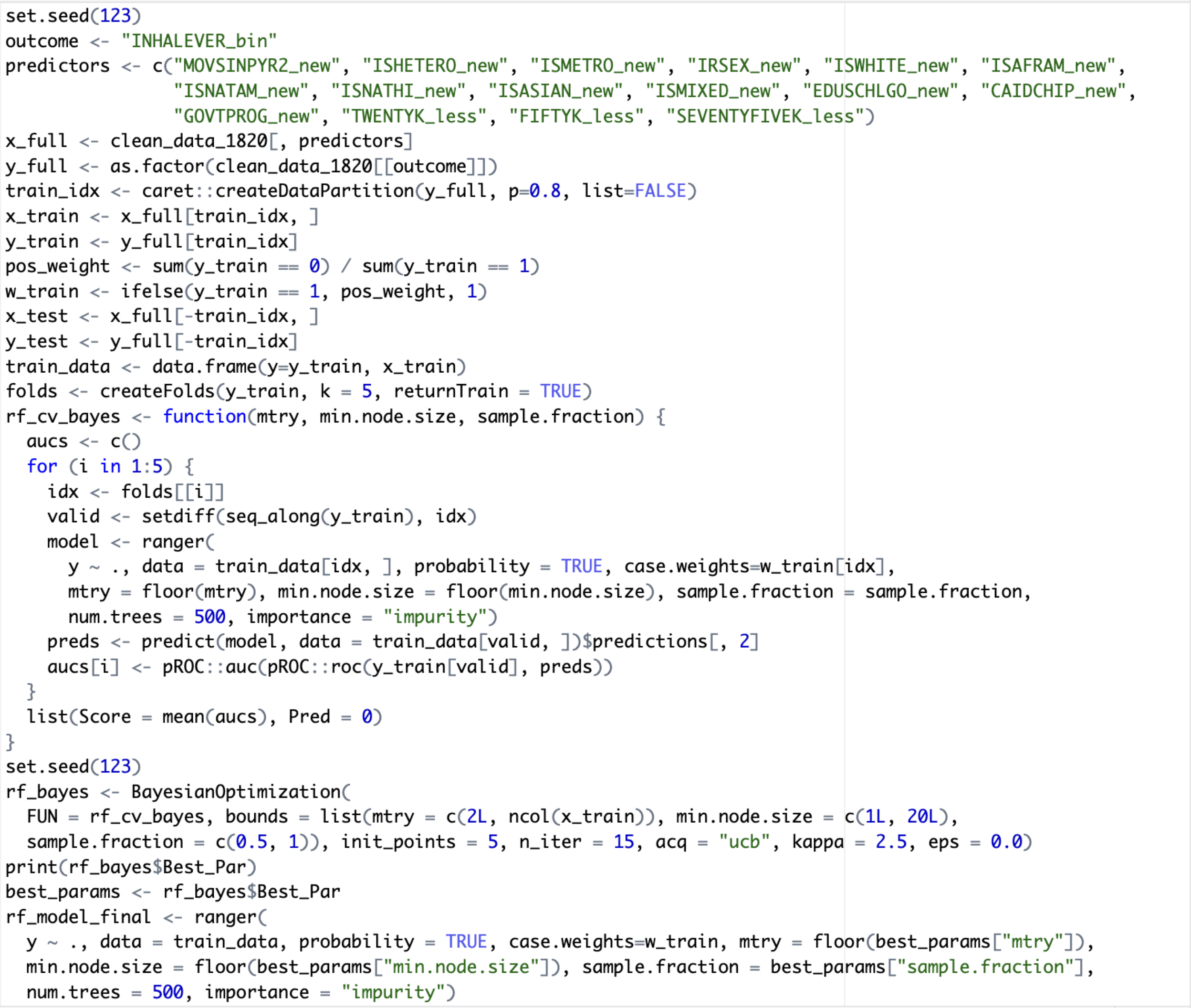
Table G.1

	12-13	14-15	16-17	18-20
cig	CAIDCHIP_new 0.95 ISNATAM_new 0.80 MOVSINPYR2_new -0.59 ISMIXED_new 0.58 EDUSCHLGO_new -0.42	ISMIXED_new 0.78 MOVSINPYR2_new -0.72 ISHETERO_new -0.64 ISNATAM_new 0.56 EDUSCHLGO_new -0.53	ISAFRAM_new -0.597 ISASIAN_new -0.464 ISHETERO_new -0.457 EDUSCHLGO_new -0.456 ISNATAM_new 0.386	ISNATHI_new -1.09 ISAFRAM_new -0.77 ISWHITE_new 0.60 TWENTYK_less 0.55 ISNATAM_new 0.48
alc	EDUSCHLGO_new -0.57 ISMETRO_new -0.28 ISAFRAM_new -0.19 CAIDCHIP_new 0.11 ISMIXED_new 0.10	ISASIAN_new -0.48 MOVSINPYR2_new -0.28 IRSEX_new 0.25 ISHETERO_new -0.180 ISAFRAM_new -0.176	ISNATHI_new -0.77 TWENTYK_less -0.73 ISASIAN_new -0.68 ISAFRAM_new -0.66 IFATHER_new -0.44	ISNATHI_new -0.824 CAIDCHIP_new -0.406 ISASIAN_new -0.364 ISNATAM_new -0.356 ISAFRAM_new -0.290
mrj	CAIDCHIP_new 0.91 ISASIAN_new -0.86	ISASIAN_new -0.507 IFATHER_new -0.328	ISASIAN_new -0.83 ISNATHI_new -0.75	ISASIAN_new -0.75 ISNATHI_new -0.70

	ISNATAM_new 0.73 EDUSCHLGO_new -0.71 SEVENTYFIVEK_less 0.64	ISHETERO_new -0.313 IRSEX_new 0.292 IMOTHER_new -0.287	ISHETERO_new -0.45 IMOTHER_new -0.39 TWENTYK_less -0.37	EDUSCHLGO_new -0.34 ISHETERO_new -0.32 ISWHITE_new 0.19
inh	N/A	ISMIXED_new 0.322 ISHETERO_new -0.318 ISWHITE_new 0.133 EDUSCHLGO_new -0.079 IRSEX_new 0.055	ISNATHI_new 1.72 ISNATAM_new 0.58 ISHETERO_new -0.49 FIFTYK_less -0.20 ISMETRO_new 0.19	ISNATHI_new -1.13 ISHETERO_new -0.72 ISAFRAM_new -0.36 ISASIAN_new 0.35 ISWHITE_new 0.33

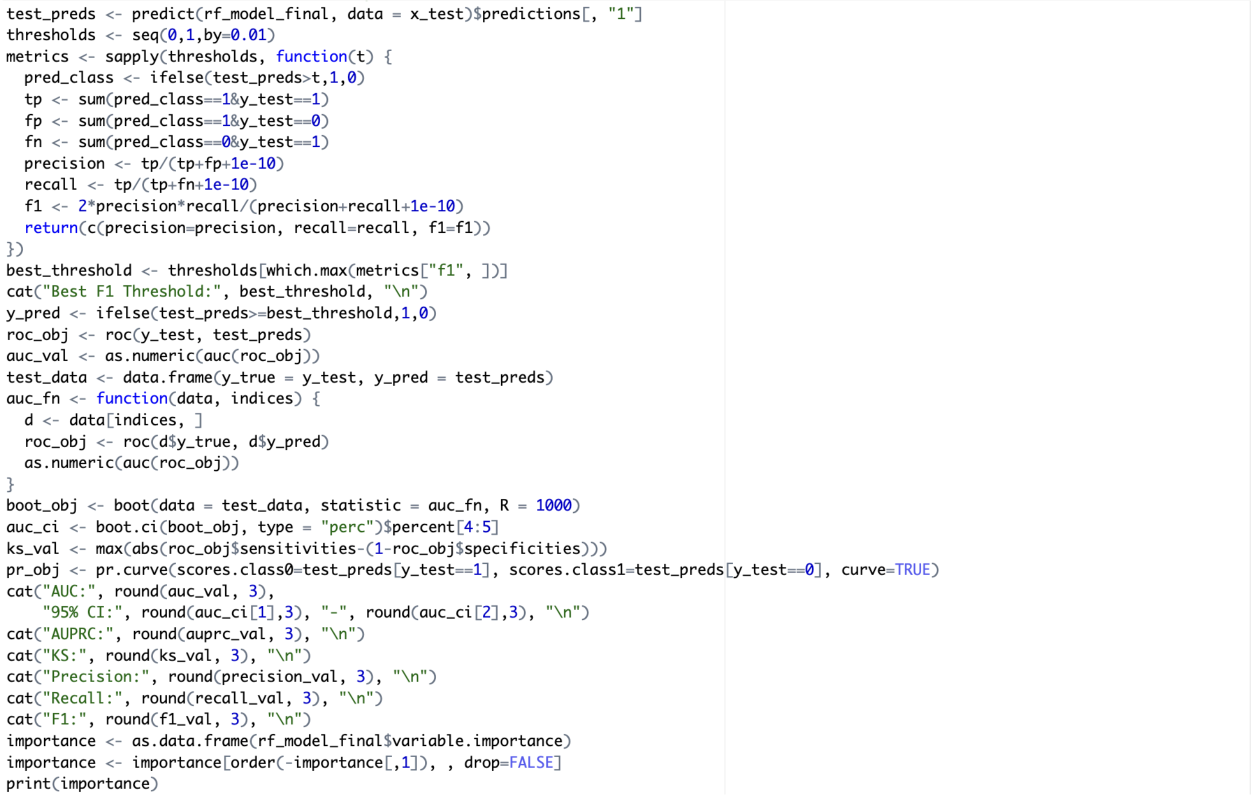
Table G.2

Appendix H - Random Forest results



```
set.seed(123)
outcome <- "INHALEVER_bin"
predictors <- c("MOVSINPYR2_new", "ISHETERO_new", "ISMETRO_new", "IRSEX_new", "ISWHITE_new", "ISAFRAM_new",
                "ISNATAM_new", "ISNATHI_new", "ISASIAN_new", "ISMIXED_new", "EDUSCHLGO_new", "CAIDCHIP_new",
                "GOVTPROG_new", "TWENTYK_less", "FIFTYK_less", "SEVENTYFIVEK_less")
x_full <- clean_data_1820[, predictors]
y_full <- as.factor(clean_data_1820[[outcome]])
train_idx <- caret::createDataPartition(y_full, p=0.8, list=FALSE)
x_train <- x_full[train_idx, ]
y_train <- y_full[train_idx]
pos_weight <- sum(y_train == 0) / sum(y_train == 1)
w_train <- ifelse(y_train == 1, pos_weight, 1)
x_test <- x_full[-train_idx, ]
y_test <- y_full[-train_idx]
train_data <- data.frame(y=y_train, x_train)
folds <- createFolds(y_train, k = 5, returnTrain = TRUE)
rf_cv_bayes <- function(mtry, min.node.size, sample.fraction) {
  aucs <- c()
  for (i in 1:5) {
    idx <- folds[[i]]
    valid <- setdiff(seq_along(y_train), idx)
    model <- ranger(
      y ~ ., data = train_data[idx, ], probability = TRUE, case.weights=w_train[idx],
      mtry = floor(mtry), min.node.size = floor(min.node.size), sample.fraction = sample.fraction,
      num.trees = 500, importance = "impurity")
    preds <- predict(model, data = train_data[valid, ])$predictions[, 2]
    aucs[i] <- pROC::auc(pROC::roc(y_train[valid], preds))
  }
  list(Score = mean(aucs), Pred = 0)
}
set.seed(123)
rf_bayes <- BayesianOptimization(
  FUN = rf_cv_bayes, bounds = list(mtry = c(2L, ncol(x_train)), min.node.size = c(1L, 20L),
  sample.fraction = c(0.5, 1)), init_points = 5, n_iter = 15, acq = "ucb", kappa = 2.5, eps = 0.0)
print(rf_bayes$Best_Par)
best_params <- rf_bayes$Best_Par
rf_model_final <- ranger(
  y ~ ., data = train_data, probability = TRUE, case.weights=w_train, mtry = floor(best_params["mtry"]),
  min.node.size = floor(best_params["min.node.size"]), sample.fraction = best_params["sample.fraction"],
  num.trees = 500, importance = "impurity")
```

Exhibit H.1



```
test_preds <- predict(rf_model_final, data = x_test)$predictions[, "1"]
thresholds <- seq(0,1,by=0.01)
metrics <- sapply(thresholds, function(t) {
  pred_class <- ifelse(test_preds>t,1,0)
  tp <- sum(pred_class==1&y_test==1)
  fp <- sum(pred_class==1&y_test==0)
  fn <- sum(pred_class==0&y_test==1)
  precision <- tp/(tp+fp+1e-10)
  recall <- tp/(tp+fn+1e-10)
  f1 <- 2*precision*recall/(precision+recall+1e-10)
  return(c(precision=precision, recall=recall, f1=f1))
})
best_threshold <- thresholds[which.max(metrics["f1", ])]
cat("Best F1 Threshold:", best_threshold, "\n")
y_pred <- ifelse(test_preds>=best_threshold,1,0)
roc_obj <- roc(y_test, test_preds)
auc_val <- as.numeric(auc(roc_obj))
test_data <- data.frame(y_true = y_test, y_pred = test_preds)
auc_fn <- function(data, indices) {
  d <- data[indices, ]
  roc_obj <- roc(d$y_true, d$y_pred)
  as.numeric(auc(roc_obj))
}
boot_obj <- boot(data = test_data, statistic = auc_fn, R = 1000)
auc_ci <- boot.ci(boot_obj, type = "perc")$percent[4:5]
ks_val <- max(abs(roc_obj$sensitivities-(1-roc_obj$specificities)))
pr_obj <- pr.curve(scores.class0=test_preds[y_test==1], scores.class1=test_preds[y_test==0], curve=TRUE)
cat("AUC:", round(auc_val, 3),
    "95% CI:", round(auc_ci[1],3), "-", round(auc_ci[2],3), "\n")
cat("AUPRC:", round(auprc_val, 3), "\n")
cat("KS:", round(ks_val, 3), "\n")
cat("Precision:", round(precision_val, 3), "\n")
cat("Recall:", round(recall_val, 3), "\n")
cat("F1:", round(f1_val, 3), "\n")
importance <- as.data.frame(rf_model_final$variable.importance)
importance <- importance[order(-importance[,1]), , drop=FALSE]
print(importance)
```

Exhibit H.2

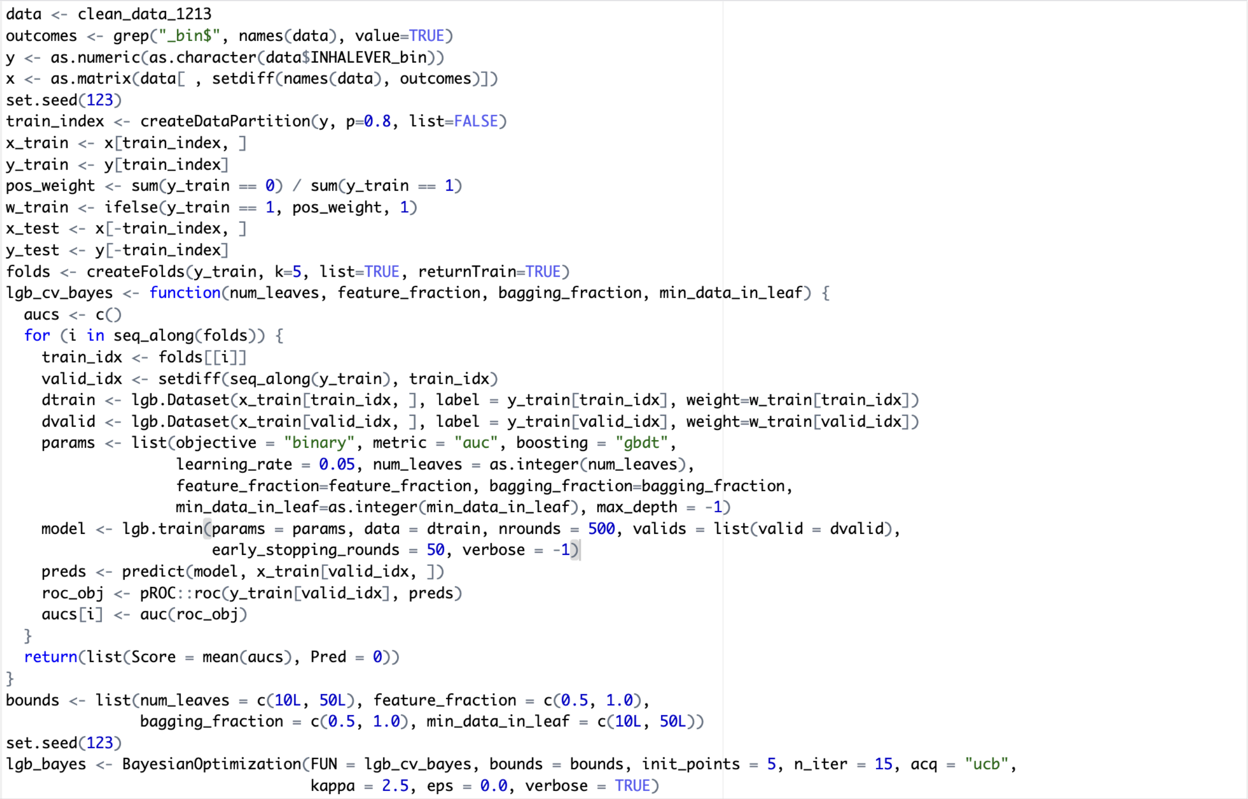
	12-13			14-15			16-17			18-20		
	AUC	AUPRC	K-S	AUC	AUPRC	K-S	AUC	AUPRC	K-S	AUC	AUPRC	K-S
cig	0.666 95% CI: 0.583 - 0.798	0.095	0.62	0.778 95% CI: 0.69 - 0.861	0.108	0.494	0.69 95% CI: 0.602 - 0.761	0.278	0.346	0.776 95% CI: 0.738 - 0.81	0.243	0.337
alc	0.706 95% CI: 0.666 -	0.116	0.59	0.687 95% CI: 0.674 -	0.248	0.294	0.722 95% CI: 0.68 -	0.497	0.304	0.727 95% CI: 0.705 -	0.457	0.216

	0.802			0.767			0.767			0.751		
mrj	0.754 95% CI: 0.66 - 0.852	0.05	0.59	0.683 95% CI: 0.551 - 0.749	0.191	0.349	0.656 95% CI: 0.561 - 0.707	0.331	0.283	0.736 95% CI: 0.711 - 0.761	0.354	0.257
inh	0.617 95% CI: 0.574 - 0.712	0.066	0.518	0.597 95% CI: 0.565 - 0.699	0.112	0.566	0.755 95% CI: 0.683 - 0.843	0.09	0.39	0.732 95% CI: 0.687 - 0.774	0.081	0.243

\

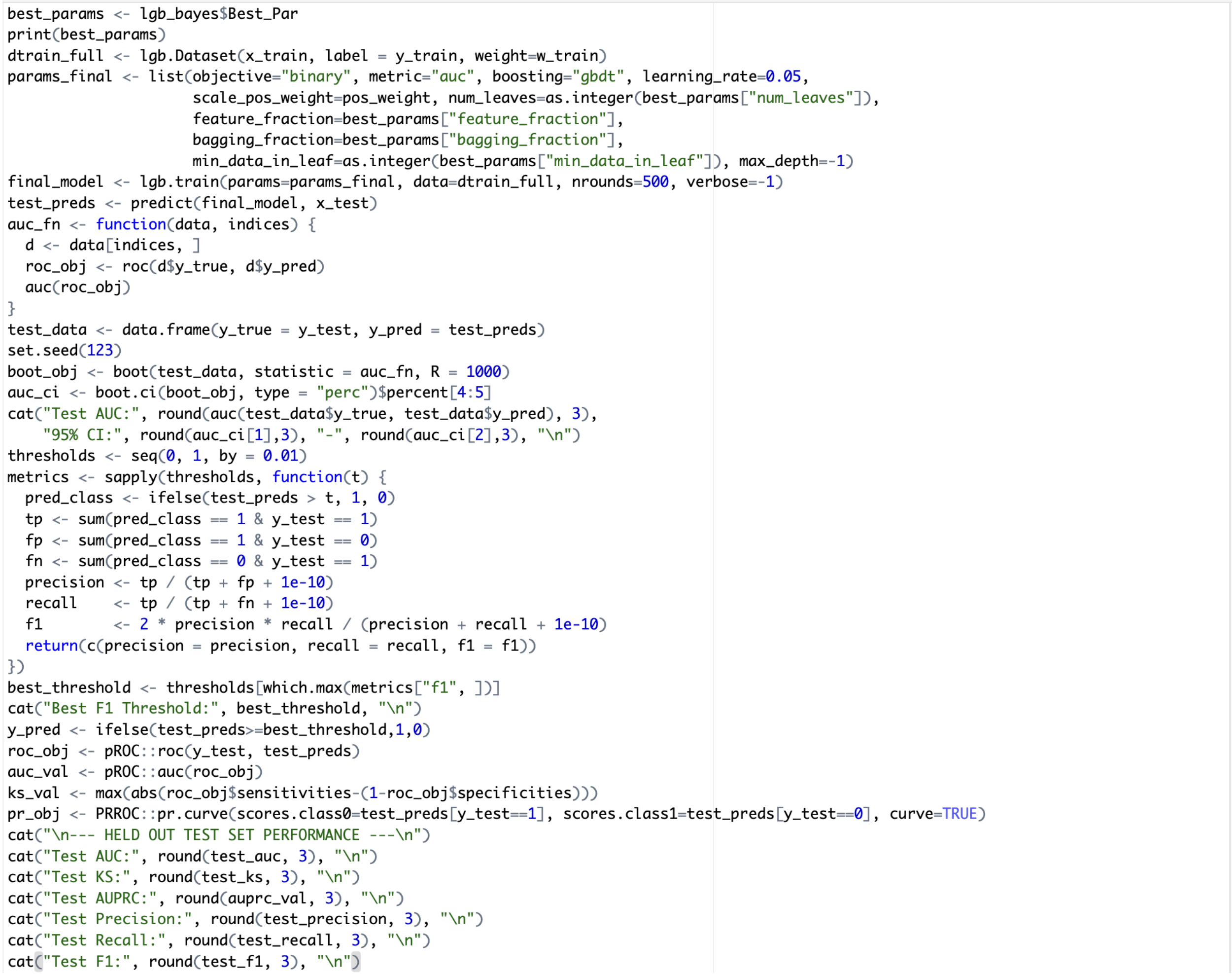
Table H.1

Appendix I - LightGBM results



```r
data <- clean_data_1213
outcomes <- grep("_bin$", names(data), value=TRUE)
y <- as.numeric(as.character(data$INHALEVER_bin))
x <- as.matrix(data[ , setdiff(names(data), outcomes)])
set.seed(123)
train_index <- createDataPartition(y, p=0.8, list=FALSE)
x_train <- x[train_index, ]
y_train <- y[train_index]
pos_weight <- sum(y_train == 0) / sum(y_train == 1)
w_train <- ifelse(y_train == 1, pos_weight, 1)
x_test <- x[-train_index, ]
y_test <- y[-train_index]
folds <- createFolds(y_train, k=5, list=TRUE, returnTrain=TRUE)
lgb_cv_bayes <- function(num_leaves, feature_fraction, bagging_fraction, min_data_in_leaf) {
  aucs <- c()
  for (i in seq_along(folds)) {
    train_idx <- folds[[i]]
    valid_idx <- setdiff(seq_along(y_train), train_idx)
    dtrain <- lgb.Dataset(x_train[train_idx, ], label = y_train[train_idx], weight=w_train[train_idx])
    dvalid <- lgb.Dataset(x_train[valid_idx, ], label = y_train[valid_idx], weight=w_train[valid_idx])
    params <- list(objective = "binary", metric = "auc", boosting = "gbdt",
                   learning_rate = 0.05, num_leaves = as.integer(num_leaves),
                   feature_fraction=feature_fraction, bagging_fraction=bagging_fraction,
                   min_data_in_leaf=as.integer(min_data_in_leaf), max_depth = -1)
    model <- lgb.train(params = params, data = dtrain, nrounds = 500, valids = list(valid = dvalid),
                       early_stopping_rounds = 50, verbose = -1)
    preds <- predict(model, x_train[valid_idx, ])
    roc_obj <- pROC::roc(y_train[valid_idx], preds)
    aucs[i] <- auc(roc_obj)
  }
  return(list(Score = mean(aucs), Pred = 0))
}
bounds <- list(num_leaves = c(10L, 50L), feature_fraction = c(0.5, 1.0),
               bagging_fraction = c(0.5, 1.0), min_data_in_leaf = c(10L, 50L))
set.seed(123)
lgb_bayes <- BayesianOptimization(FUN = lgb_cv_bayes, bounds = bounds, init_points = 5, n_iter = 15, acq = "ucb",
                                  kappa = 2.5, eps = 0.0, verbose = TRUE)
```

Exhibit I.1



```r
best_params <- lgb_bayes$Best_Par
print(best_params)
dtrain_full <- lgb.Dataset(x_train, label = y_train, weight=w_train)
params_final <- list(objective="binary", metric="auc", boosting="gbdt", learning_rate=0.05,
                     scale_pos_weight=pos_weight, num_leaves=as.integer(best_params["num_leaves"]),
                     feature_fraction=best_params["feature_fraction"],
                     bagging_fraction=best_params["bagging_fraction"],
                     min_data_in_leaf=as.integer(best_params["min_data_in_leaf"]), max_depth=-1)
final_model <- lgb.train(params=params_final, data=dtrain_full, nrounds=500, verbose=-1)
test_preds <- predict(final_model, x_test)
auc_fn <- function(data, indices) {
  d <- data[indices, ]
  roc_obj <- roc(d$y_true, d$y_pred)
  auc(roc_obj)
}
test_data <- data.frame(y_true = y_test, y_pred = test_preds)
set.seed(123)
boot_obj <- boot(test_data, statistic = auc_fn, R = 1000)
auc_ci <- boot.ci(boot_obj, type = "perc")$percent[4:5]
cat("Test AUC:", round(auc(test_data$y_true, test_data$y_pred), 3),
    "95% CI:", round(auc_ci[1],3), "-", round(auc_ci[2],3), "\n")
thresholds <- seq(0, 1, by = 0.01)
metrics <- sapply(thresholds, function(t) {
  pred_class <- ifelse(test_preds > t, 1, 0)
  tp <- sum(pred_class == 1 & y_test == 1)
  fp <- sum(pred_class == 1 & y_test == 0)
  fn <- sum(pred_class == 0 & y_test == 1)
  precision <- tp / (tp + fp + 1e-10)
  recall    <- tp / (tp + fn + 1e-10)
  f1        <- 2 * precision * recall / (precision + recall + 1e-10)
  return(c(precision = precision, recall = recall, f1 = f1))
})
best_threshold <- thresholds[which.max(metrics["f1", ])]
cat("Best F1 Threshold:", best_threshold, "\n")
y_pred <- ifelse(test_preds>=best_threshold,1,0)
roc_obj <- pROC::roc(y_test, test_preds)
auc_val <- pROC::auc(roc_obj)
ks_val <- max(abs(roc_obj$sensitivities-(1-roc_obj$specificities)))
pr_obj <- PRROC::pr.curve(scores.class0=test_preds[y_test==1], scores.class1=test_preds[y_test==0], curve=TRUE)
cat("\n--- HELD OUT TEST SET PERFORMANCE ---\n")
cat("Test AUC:", round(test_auc, 3), "\n")
cat("Test KS:", round(test_ks, 3), "\n")
cat("Test AUPRC:", round(auprc_val, 3), "\n")
cat("Test Precision:", round(test_precision, 3), "\n")
cat("Test Recall:", round(test_recall, 3), "\n")
cat("Test F1:", round(test_f1, 3), "\n")
```

Exhibit I.2

	12-13			14-15			16-17			18-20		
	AUC	AUPRC	K-S	AUC	AUPRC	K-S	AUC	AUPRC	K-S	AUC	AUPRC	K-S
cig	0.694 95% CI: 0.581 - 0.798	0.06	0.44	0.572 95% CI: 0.466 - 0.674	0.11	0.342	0.57 95% CI: 0.539 - 0.706	0.231	0.249	0.701 95% CI: 0.672 - 0.731	0.297	0.279

alc	0.515 95% CI: 0.467 - 0.604	0.094	0.199	0.64 95% CI: 0.59 - 0.698	0.257	0.155	0.622 95% CI: 0.582 - 0.666	0.491	0.215	0.727 95% CI: 0.706 - 0.746	0.555	0.17
mrj	0.505 95% CI: 0.441 - 0.632	0.036	0.354	0.572 95% CI: 0.461 - 0.635	0.178	0.233	0.575 95% CI: 0.484 - 0.623	0.369	0.195	0.735 95% CI: 0.712 - 0.756	0.413	0.167
inh	0.595 95% CI: 0.556 - 0.734	0.066	0.166	0.562 95% CI: 0.468 - 0.629	0.078	0.127	0.507 95% CI: 0.47 - 0.6	0.098	0.156	0.639 95% CI: 0.599 - 0.68	0.088	0.236

Table I.1

	12-13	14-15	16-17	18-20
cig	CAIDCHIP_new 0.32 MOVSINPYR2_new 0.14 GOVTPROG_new 0.11 EDUSCHLGO_new 0.08 ISMETRO_new 0.06	CAIDCHIP_new 0.19 ISHETERO_new 0.17 MOVSINPYR2_new 0.11 EDUSCHLGO_new 0.10 GOVTPROG_new 0.07	ISHETERO_new 0.15 ISAFRAM_new 0.14 EDUSCHLGO_new 0.12 CAIDCHIP_new 0.11 IMOTHER_new 0.08	ISAFRAM_new 0.19 EDUSCHLGO_new 0.16 ISWHITE_new 0.13 TWENTYK_less 0.10 ISMETRO_new 0.08
alc	EDUSCHLGO_new 0.31 CAIDCHIP_new 0.13 SEVENTYFIVEK_less 0.18 IMOTHER_new 0.09 ISMETRO_new 0.08	IRSEX_new 0.187 ISHETERO_new 0.109 ISASIAN_new 0.102 EDUSCHLGO_new 0.098 IFATHER_new 0.088	ISAFRAM_new 0.1510 IRSEX_new 0.1435 TWENTYK_less 0.0904 CAIDCHIP_new 0.0774 ISASIAN_new 0.0772	CAIDCHIP_new 0.27 ISWHITE_new 0.26 ISAFRAM_new 0.10 TWENTYK_less 0.07 MOVSINPYR2_new 0.06
mrj	CAIDCHIP_new 0.467 IFATHER_new 0.117	CAIDCHIP_new 0.175 IFATHER_new 0.134	ISHETERO_new 0.17 ISASIAN_new 0.11	EDUSCHLGO_new 0.214 ISASIAN_new 0.173

	EDUSCHLGO_new 0.073 IRSEX_new 0.067 SEVENTYFIVEK_less 0.064	IRSEX_new 0.125 ISHETERO_new 0.119 ISASIAN_new 0.093	IRSEX_new 0.09 TWENTYK_less 0.08 CAIDCHIP_new 0.07	ISHETERO_new 0.164 TWENTYK_less 0.076 ISWHITE_new 0.075
inh	IRSEX_new 0.20 GOVTPROG_new 0.14 ISHETERO_new 0.13 ISAFRAM_new 0.11 CAIDCHIP_new 0.10	ISHETERO_new 0.232 ISWHITE_new 0.133 IFATHER_new 0.127 IRSEX_new 0.097 ISMIXED_new 0.087	ISHETERO_new 0.233 ISMETRO_new 0.204 FIFTYK_less 0.083 IMOTHER_new 0.079 ISNATHI_new 0.078	ISHETERO_new 0.24 IRSEX_new 0.16 ISAFRAM_new 0.12 GOVTPROG_new 0.10 TWENTYK_less 0.07

Table I.2

Appendix J - average metrics results

	LASSO	Random Forest	LightGBM
AUC	0.615	0.705	0.608
K-S	0.221	0.4	0.230
AUPRC	0.214	0.208	0.214

Table J.1

Appendix K - accuracy formula

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Exhibit K.1