Stats 101C Final Project:

Predicting Alcoholic Status Using a Person's Vitals

By Camden Weber, Carolyn Wang, Daniel Mendelevitch, & Elaine Fan

Stats 101C, Lecture 2

Abstract

In this project, we aim to predict whether or not someone is alcoholic from various vital tests and attributes, as accurately as possible. However, the data provided contains noise and has a significant amount of missing values, which makes it challenging to apply modeling techniques out-of-the-box. We approach this challenge in three different ways: imputation, model selection, and variable selection. For each, we perform experiments to determine the optimal strategy or modeling technique by training models and evaluating their performance on the testing data on the Kaggle Leaderboard. Throughout this project, we optimized for both model performance and model simplicity. We arrive at the conclusion that a complex imputation combined with a simple hyperparameter-tuned Gradient Boosting Machine (GBM) is the best model suited for this task.

I. Introduction

Alcoholism is a widespread issue in the United States: about one half of American adults have a family history with alcohol addiction and an estimated 88,000 Americans die from alcohol-related causes annually (National Institute on Alcohol Abuse and Alcoholism). Alcoholism also has many associated risks, both short term and long term. Short term, excessive alcohol intake can lead to injuries and violent behavior due to lack of coordination and brain impairment when under the influence. High alcohol intake can also lead to alcohol poisoning. Long term, alcoholism has been linked to high blood pressure, high cholesterol level, weakening of the immune system, and mental health and social impairment issues (CDC).

To predict whether an individual is likely to be an alcoholic, various factors must be considered. One factor to contemplate is gender, Typically, men tend to exhibit higher levels of alcohol consumption compared to women. In 2016, global consumption statistics indicated that 54% of males and 32% of females aged 15 and older consumed alcohol, and alcohol was a contributing factor to roughly 2.3 million deaths for men and 0.7 million deaths for women, illustrating the clear gender gap when it comes to alcohol consumption (Grant). However, while there are clear gender disparities, differences among gender are narrowing in recent years because of a shift in alcohol consumption behavior in women (Grant). Age is another factor that may influence an individual's likelihood in becoming an alcoholic. In general, alcohol use tends to start in the teens and early 20s, peak in middle and late 20s, and slow down by 30s, though this pattern is not uniform across all individuals (Addiction Center). Psychological issues are also closely related to alcoholism: More than 40% of bipolar sufferers abuse or are dependent on alcohol, and approximately 20% of depression sufferers abuse or are dependent on alcohol (Addiction Center).

However, research has consistently shown that genetics play the most pivotal role in determining an individual's predisposition to alcoholism, surpassing the influence of any other single factor (Addiction Center). Biological children of alcoholics are more likely to become alcoholics, regardless of whether they are raised by alcoholic parents or not. Therefore, examining one's genetic history and makeup has been proven to be most important in determining alcoholic status.

In this project, our objective is to predict an individual's alcoholic status from a dataset that encompasses a wide range of predictors, including age, smoking status, gender, BMI, and cholesterol levels. After imputing our data effectively which has missing values across all predictors, we can identify the most important predictors through statistical techniques and determine if any should be excluded. Based on our background research, we believe key variables that will be especially important to our model will include gender, age, cholesterol, blood pressure, and other variables that indicate genetic history or background with alcoholism. After imputation and feature selection, we can proceed to testing different classification techniques, from the simple logistic regression approach to complex ensemble methods such as Random Forest, GBM, and stacking. This comprehensive analysis will enable us to find the most effective model to accurately predict alcoholic status given a person's vitals and other crucial information.

II. Exploratory Data Analysis

a. Missing Values

Our initial analysis of the dataset focused on the prevalence of missing values. We found that about 6.67% or 130776/70000 of the training data was missing. For the testing data, about 6.918% or 56035/30000 was missing. The AGE.category had the highest percentage of missing data in both datasets, 11.88% and 11.73% respectively. All of the other predictors in both datasets had around 7% of missing data. We address how we imputed this missing data in Part III section a, "Imputation Method".

Percentage of Missing Data by Variable

Variable	Percent Missing (Training)	Percent Missing (Testing)
age	6.967143	7.103333
height	7.058571	7.063333
waistline	7.057143	7.183333

weight	7.102857	7.003333
SGOT_AST	6.981429	6.933333
SGOT_ALT	6.990000	7.076667
gamma_GTP	7.087143	6.853333
sex	7.088571	7.296667
DBP	6.992857	7.056667
SBP	7.027143	7.013333
BLDS	6.887143	7.050000
tot_chole	6.948571	7.203333
HDL_chole	6.880000	7.026667
LDL_chole	7.020000	7.003333
triglyceride	6.967143	6.680000
urine_protein	6.998571	6.830000
serum_creatinine	6.924286	6.924286
hear_left	6.904286	6.904286
hear_right	6.981429	6.981429
BMI.Category	6.962857	6.962857
AGE.Category	11.875714	11.875714
Smoking.Status	6.970000	6.970000

Table 1: Percent of missing values for all variables in the training and testing dataset

b. Class Balance

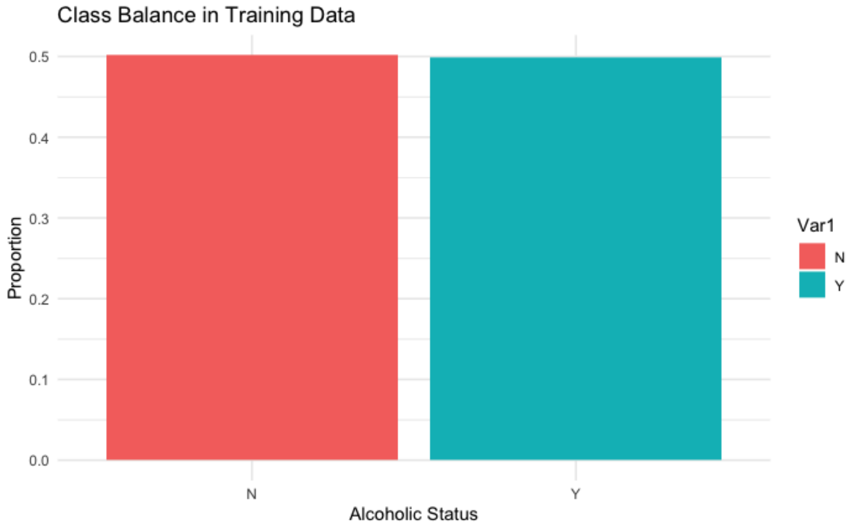

Figure 1: Class balance of our predictor variable. We can see that there is an approximate 50/50 split between the Y and N classes, which means any classifier we train should be at least 50% accurate.

We analyze the class balance of our dataset before starting our modeling efforts. As seen in Figure 1, we observe that the class balance is about 50/50, that is, there is approximately an equal amount of alcoholics and non-alcoholics in our dataset. This is good, since our models are less likely to predict all Y or all N than if the classes were more imbalanced. The maximum error rate would be roughly 50% if we “just say no” or “just say yes.”

c. Examining the Predictors

Bar Plots for Categorical Variables

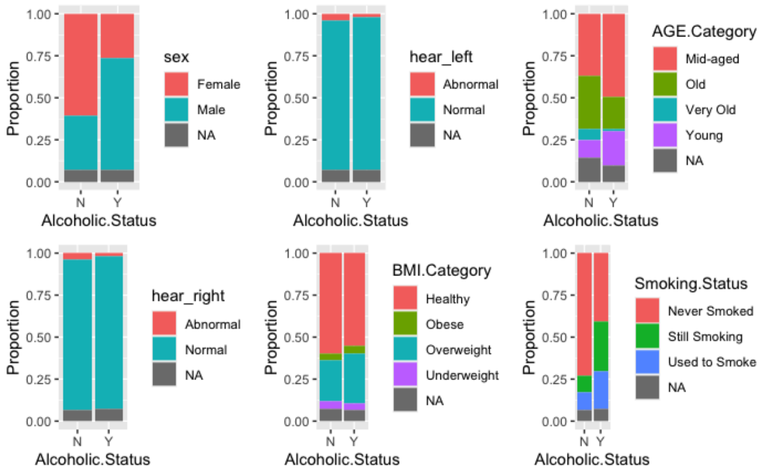

Figure 2: Stacked bar plots showing the distribution of various categorical predictors in our dataset.

As is seen in Figure 2, certain categorical variables seem to be more informative than others. For example, looking at the barplot for the sex variable, the proportion of females in the Y category is much around 25%, which is far lower than the >50% observed in the N category. Therefore, we suspect sex will be an important predictor for any future classifiers. Smoking.Status also splits the classes relatively well; there is a significantly smaller proportion of alcoholics that never smoked when compared to the non-alcoholic. Other predictors like hear_left or BMI.Category seem to be less effective, and are likely to not be very important predictors for our classifiers.

Density Plots for Numeric Variables

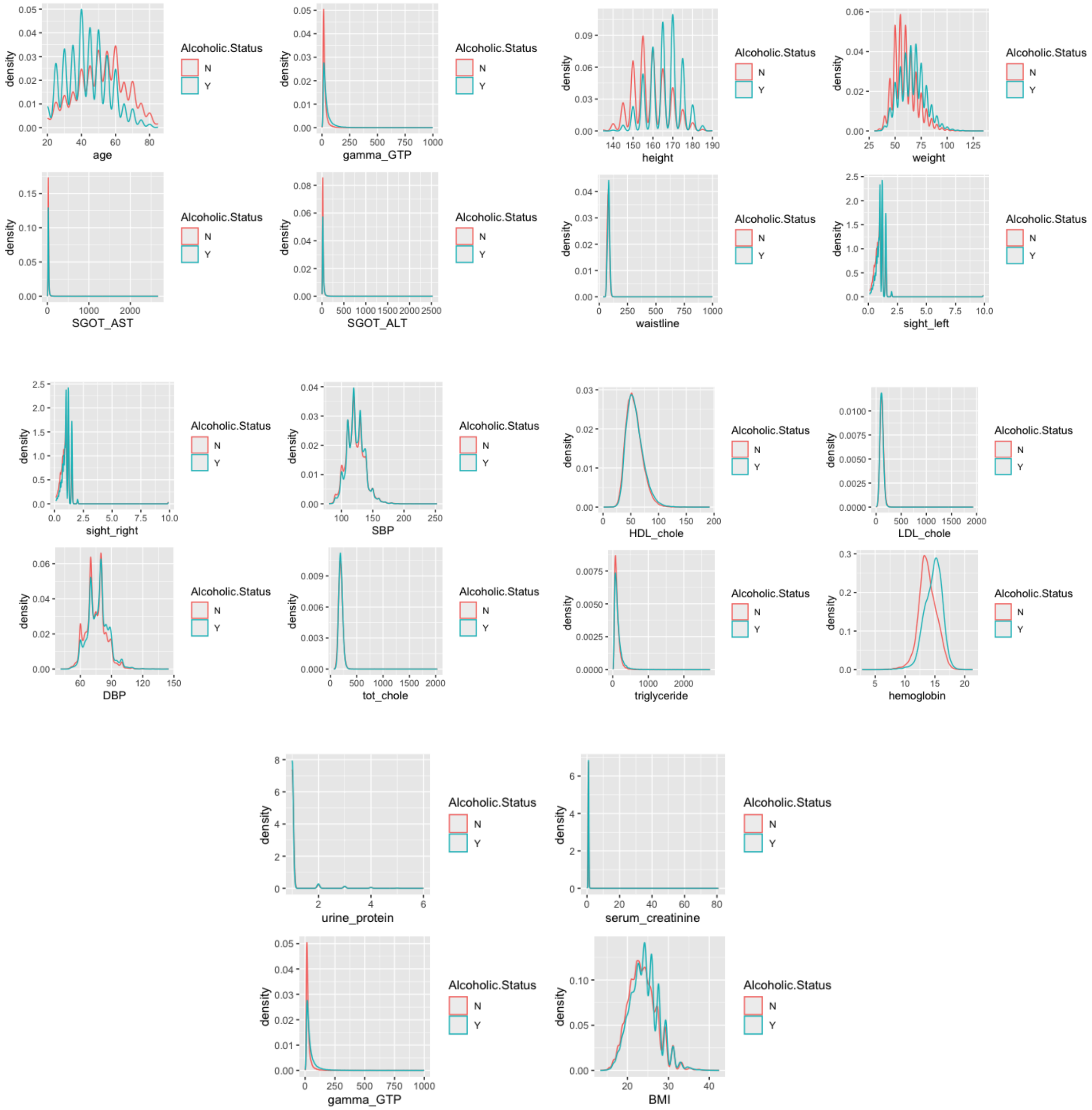

Figure 3: Density plots for all numeric variables in our dataset.

Examining the density plots of each predictors, there are several key observations to be made. First, most of the predictors do not follow a normal distribution. This means a model such as LDA, which assumes the data follows a Gaussian distribution, is likely unsuitable. When considering logistic regression, some type of variable transformation might be needed due to extreme data. In general, the density plots indicate choosing a model with limited model assumptions increases accuracy. These density plots also illustrate which predictors have the most separation between the classes of alcoholic status. Age, height, hemoglobin, and weight, seem to have some of the best separation between classes, which is displayed by the difference between the red and blue density curves, making them optimal predictors. We can also see differences by class in the following boxplots.

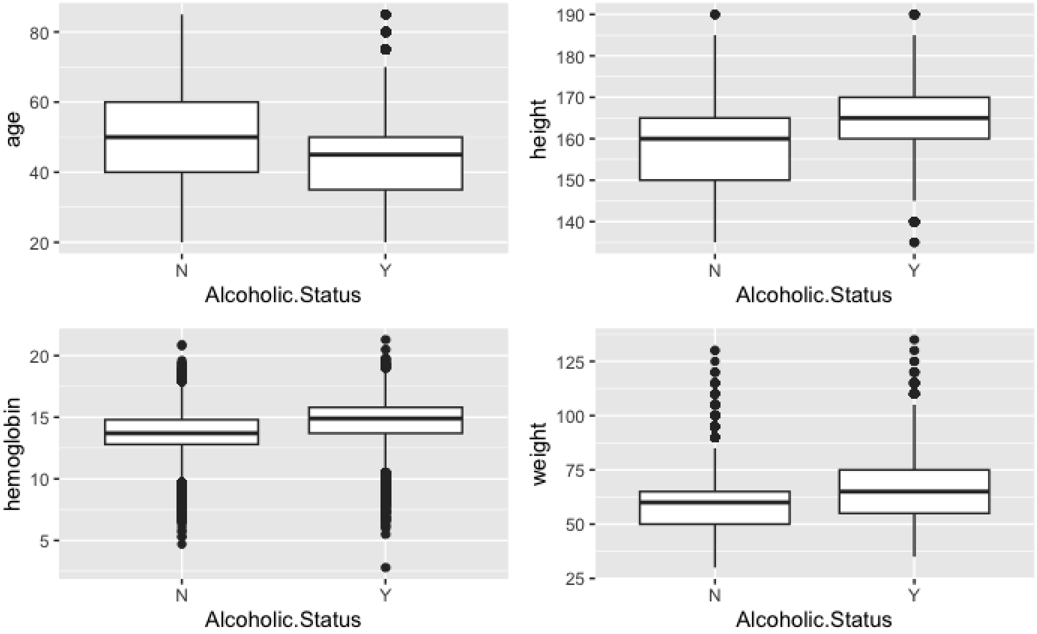

Figure 4: Box plots for four significant numeric predictors: age, height, hemoglobin, and weight.

III. Methods and Models

a. Imputation Method

For our project, we tested two primary methods of data imputation: Method 1 used the Mice package and Method 2 imputed numeric values using the Scikit Learn Iterative Imputer.

For Method 1, we used the MICE package in R. MICE, short for Multivariate Imputation by Chained Equations, is an R package designed for handling missing values. It employs a two-step approach using the mice() function to build a model and the complete() function to generate completed data. The mice() function creates multiple complete copies of the original data frame (df), each featuring different imputations of the missing data. The complete() function allows users to retrieve these datasets, with the default being the first imputation. MICE uses the imputation method of predictive mean matching which involves predicting missing values by calculating the mean of observed values from similar cases, introducing a level of randomness by selecting one of the observed values to account for uncertainty.

Method 2 was a bit more involved. It uses a combination of two imputers: a Scikit-Learn IterativeImputer for the numeric data and mode (i.e. most-frequent) imputation for the categorical data. The IterativeImputer fills missing values with an initial guess, and then iteratively updates these missing values by fitting a model on all other features to predict missing values in a given feature. It keeps updating these missing values until either a max number of iterations is reached, or the imputer updates are small enough in magnitude.

To determine the best imputation method, we fit two Gradient Boosting Machines: one on data imputed with MICE, and one with data imputed with IterativeImputer. We assumed GBM would be one of the best performing classification methods which is why we chose to test our imputation methods on this model. We then determined which imputation method to use based on which GBM performed better. To make this test as fair and accurate as possible, we split the training set into a smaller training and separate validation set, and performed hyperparameter tuning on the GBM for each imputation method to maximize their performance on that validation set via. cross-validation.

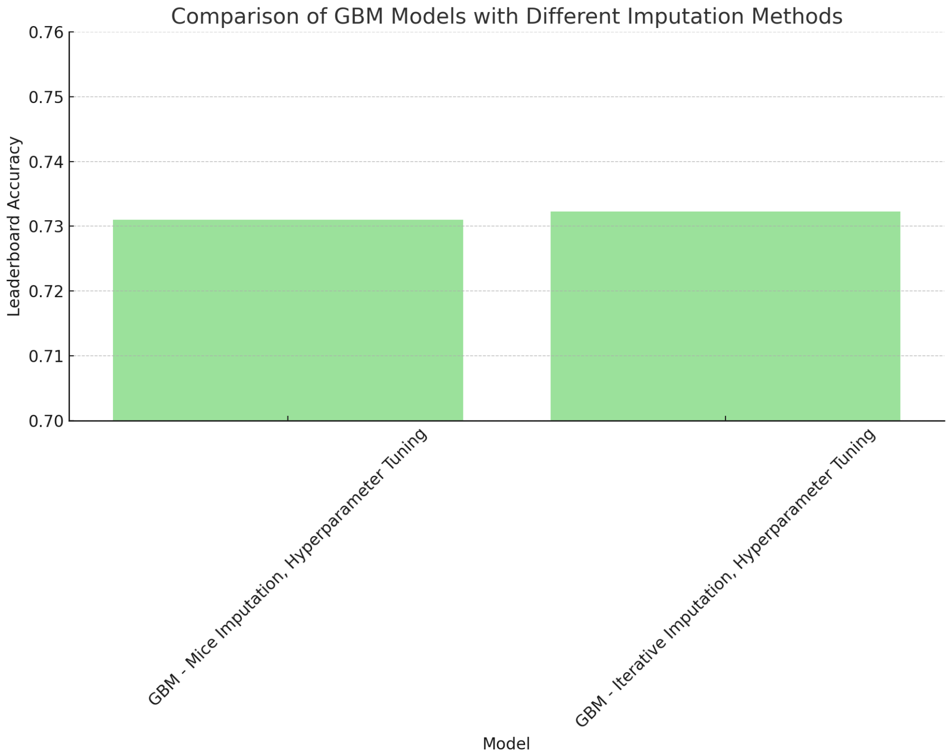

Figure 5: The GBM fit on MICE Imputation got a testing accuracy of 0.7304 whereas the GBM fit on Iterative Imputation got a testing accuracy of 0.73226. We opted to use Iterative Imputation as our data imputation strategy for the rest of this report.

As we can from Figure 3, Iterative Imputation outperformed MICE Imputation so for the rest of our experiments we used the IterativeImputer method. However, since the testing accuracy was similar between these two imputation methods, it would be unsurprising for MICE to outperform Iterative Imputation on some models. Thus, when comparing all models (see Figure 9), we included some of our models using the MICE imputation method as well.

b. Model Selection

After determining our imputation method as iterative imputation, the next step was to determine which model to use. In order to evaluate this, we trained a few different models on the full dataset and evaluated on the test data. As with our imputation experiment, we found optimal hyperparameters by doing cross validation on the train/val split used for the imputation experiments.

We chose to delve into Random Forest, Gradient Boosting Machine, and Logistic Regression as potential models. Random Forest and GBM are both methods involving classification trees. Random Forest is an ensemble learning method that constructs a multitude of uncorrelated decision trees during training and outputs the mode of the classes. GBM is an ensemble "slow-learning" machine learning algorithm that builds a predictive model in a more slow and gradual fashion, combining the predictions of multiple weak decision trees to improve overall accuracy. Finally, Logistic Regression is a rather simple statistical model used for binary classification. It models the probability of a binary outcome by fitting the probability of success to a logistic curve, making it suitable for problems where the dependent variable is categorical with two classes.

We opted to not use LDA or QDA because of the lack of normality in predictors, a key assumption of these discriminant analysis methods. We also did not use KNN because many of the categorical variables had high importance and this method involves calculating distances, making it difficult to logically include categorical variables in these models. We did not implement singular classification trees, knowing Random Forest and GBM, which use multiple decision trees, would greatly outperform it. If given more time, we would have implemented SVM or clustering, but we did not get the opportunity to fully experiment with these methods.

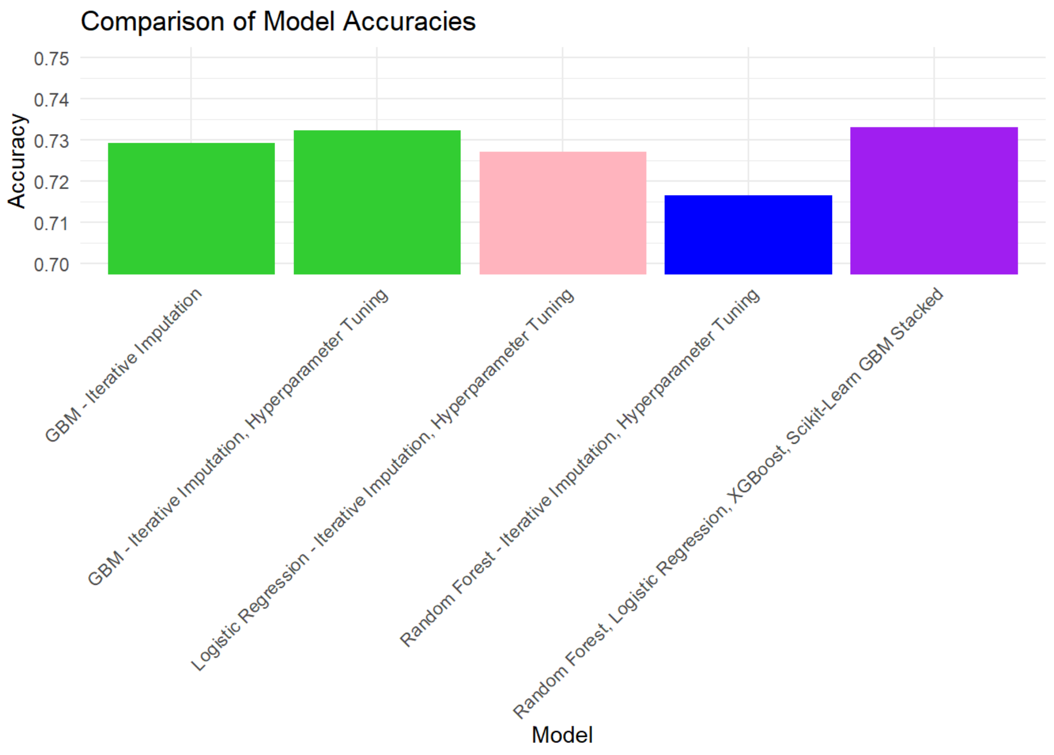

Figure 6: Testing accuracy of various models trained on imputed data with tuned hyperparameters

Figure 6 shows that the GBM with hyperparameter tuning outperforms Random Forest and Logistic Regression. As one final experiment, we tried to perform an ensemble of all of these models using stacking. That is to say, for each base model, we fit it to the data (with tuned hyperparameters), then predict a likelihood for each row in the training and testing dataset. We then further train a logistic regression to predict the ground truth labels from the likelihoods assigned by each model. We use this logistic regression to make final predictions about our unseen testing data. This was our best performing model of all, achieving a test accuracy of 73.3%. However, this model is much more expensive than the GBM, and only achieves a small <1% performance increase. Therefore, while the stacked model is technically our highest performing one, we recommend using the GBM due to its relative simplicity. It is important to note that logistic regression also performed surprisingly well, despite being the simplest algorithm.

c. Variable Selection

Another experiment we ran to potentially improve performance was variable selection. In most data science contexts, full datasets are often plagued by issues such as multicollinearity or otherwise redundant/non-informative variables. In an attempt to make our model better-performant and simpler, we ran a small experiment comparing the performance of a full model vs. a partial model.

More specifically, we computed feature importances from a GBM trained on the full dataset, used this to only select the most important variables, and then fit a new GBM on the smaller dataset and compared their performance.

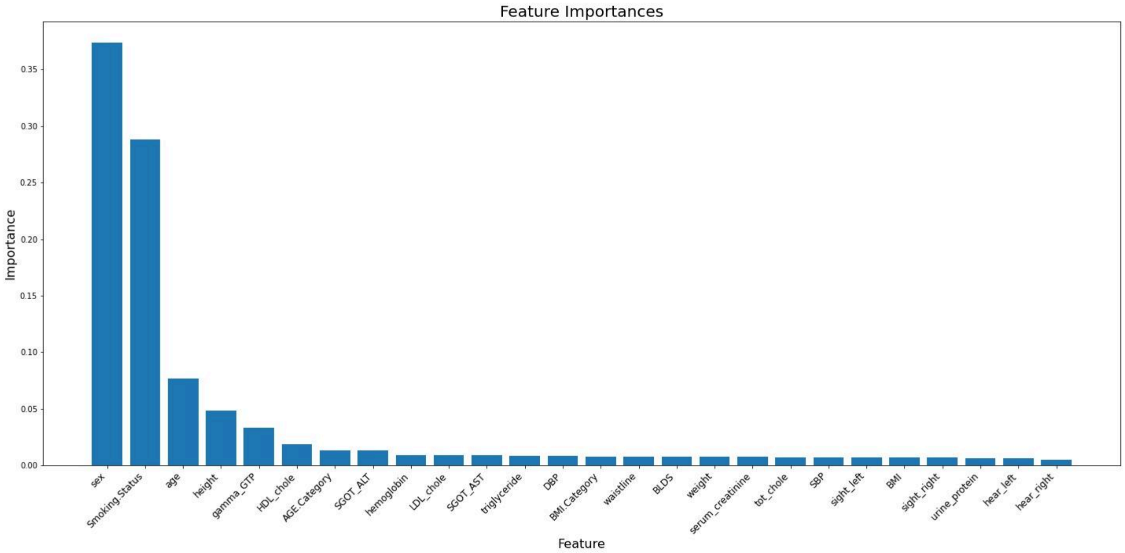

Figure 7: Feature importances according to a GBM fit on our dataset. Most of the features in the dataset have very low importance.

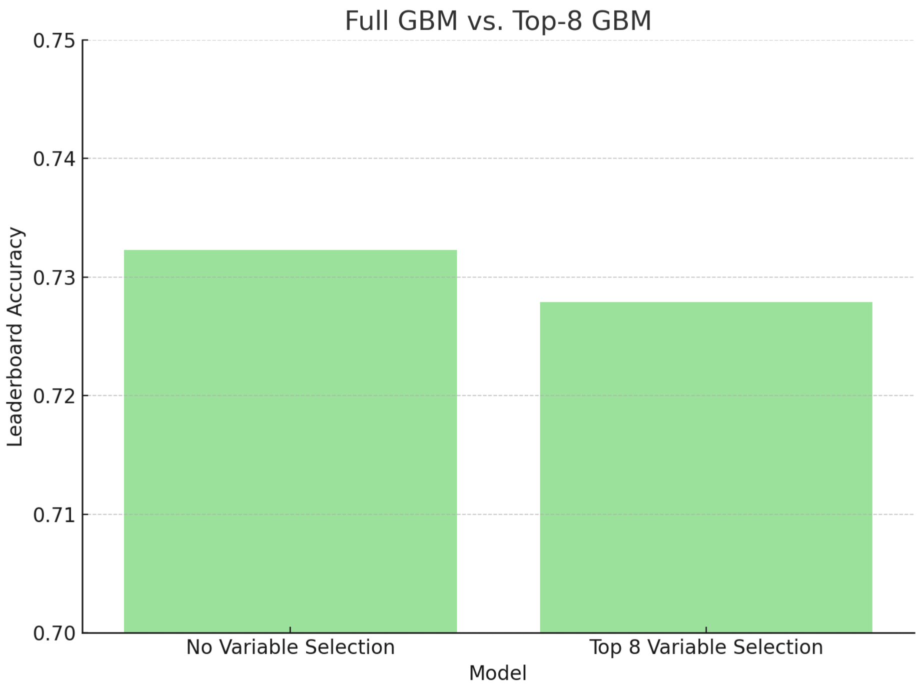

Figure 8: Performance of a GBM trained on the top 8 most important features vs. a GBM trained on the full dataset. Although the 8-feature GBM is simpler, it clearly underperforms the GBM trained on the full dataset. Therefore, we opt to use the full GBM as our final recommended model.

Sadly, the top-8 GBM underperforms somewhat significantly compared to the full model. We therefore opt to use the full model over the simpler one in this case, despite greater complexity in the full model.

Below, we show the performance of thirteen of our top performing models we trained side-by-side in the figure below. The models are color coded by model type and feature different imputation styles. It is interesting to see that the MICE imputation actually outperformed the IterativeImputation for Random Forest and Logistical Performance, but it ultimately produced a lower testing error rate than IterativeImputation for GBM, which was the model type we decided to use as our final and best performing model. This illustrates how valuable it is to test different imputation styles on different model types and the importance imputation method plays. As we can see, a simple imputation of mean and mode did not perform as well as more advanced imputation models.

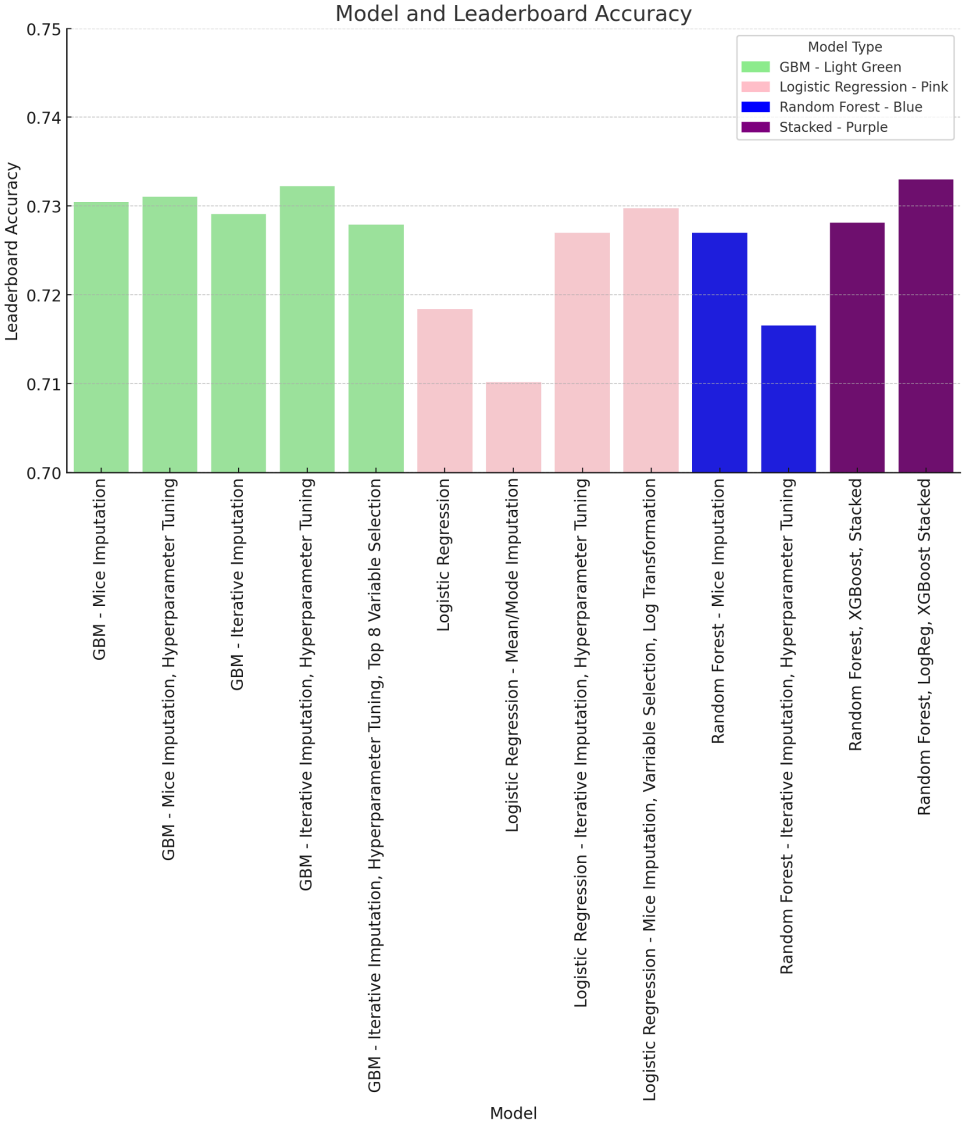

Figure 9: Comparison of performance of all models we trained. The best performing model is the stacked model on the far right, achieving a leaderboard accuracy of 0.733.

Model	Leaderboard Accuracy
GBM - Mice Imputation	0.73046
GBM - Mice Imputation, Hyperparameter Tuning	0.73103
GBM - Iterative Imputation	0.7291
GBM - Iterative Imputation, Hyperparameter Tuning	0.73226
GBM - Iterative Imputation, Hyperparameter Tuning, Top 8 Variable Selection	0.7279
Logistic Regression	0.7184
Logistic Regression - Mean/Mode Imputation	0.71013
Logistic Regression - Iterative Imputation, Hyperparameter Tuning	0.727
Logistic Regression - Mice Imputation, Variable Selection, Log Transformation	0.72723
Random Forest - Mice Imputation	0.727
Random Forest - Iterative Imputation, Hyperparameter Tuning	0.71656
Random Forest, XGBoost, Stacked	0.7281
Random Forest, Logistic Regression, XGBoost, Scikit-Learn GBM Stacked	0.733

Table 2: Leaderboard performance of the models we experimented with. High performing ones are bolded.

IV. Limitations and Discussion

a. Imputation method complexity

The imputation method we used was the Scikit-Learn IterativeImputer. While better-performing than MICE, this IterativeImputer method is expensive, since it requires fitting many models on our dataset. On a much larger dataset with more rows and columns, this cost could be prohibitive. One straightforward modification we can implement is changing to an imputation method that utilizes the mean for numerical variables and mode for categorical variables though this would result in a hit to our accuracy.

b. Model Limitations

Depending on the problem at hand, sometimes a simple model is needed. Some reasons for this would be the cost of collecting data, reducing the computing time/power needed, and it is far easier to interpret. Our main models used for experimentation were quite complex: GBMs are more computationally intensive than some simpler models like logistic regression, but also do not assume linearity and can work on much more complex data. In this context, we determined this certain model is more suitable. However, if the data exhibited a more linear structure, a logistic regression would be preferred. Our most accurate model was the stacking technique mentioned earlier. While accurate, it was complex because it required multiple models: GBM, logistic regression, and random forest. It also assumed a linear relationship between the predicted likelihoods, which is not guaranteed. However, we found that logistic regression was still suitable as this final ensemble estimator.

c. Simple Model Selection

As a final experiment, we aimed to find the simplest possible model that still performed well. Logistic regression is the simplest of the models that we attempted and is fairly easy to interpret. We started prepping the data for the model by using the MICE imputation method. Then we used backwards BIC that resulted in 12 variables. Finally, we used log transformations on variables that were very skewed. We used logistic regression on this data and it resulted in an accuracy of 72.723%. This simple method was only slightly less accurate than our best performing model, while being much less complex. Although our recommendation is to use the GBM model, this finding suggests that logistic regression is a simpler solution, and is a feasible approach to high accuracy.

V. Conclusion

Among the various models explored, the best-performing model, balancing accuracy and simplicity, was the Gradient Boosting Machine (GBM) with tuned hyperparameters applied to all variables, utilizing the iterative imputation method. This achieved a testing accuracy rate of 73.226%. Another high-accuracy model was a stacked ensemble comprising GBM, logistic regression, and random forest, using iteratively imputed and hyperparameter tuned data. This achieved a slightly higher accuracy of 73.300%, though it was much more complex. For simplicity, the logistic regression with MICE imputation (method 2) with only a subset of the predictors and log transformations on highly skewed predictors achieved a commendable 72.723% accuracy.

One of the key takeaways from the project is the significance of data imputation over the choice of the model, as many models exhibited similar performance. Data imputation was what helped certain models gain the edge, though the best data imputation was not uniform across all model types. Additionally, contrary to expectations, variable selection through GBM implementation resulted in reduced accuracy. This illustrates how model selection may result in a hit to accuracy despite decreasing the complexity of a model. We also found that stacking proved to be an effective technique for accuracy improvement, despite its computational intensity. Beyond the methods discussed in the paper, the team iterated through various models, including neural networks, k-nearest neighbors, linear models, simple ensemble methods, and PCA.

Valuable lessons learned from the project include the importance of the imputation technique, the effectiveness of a trial-and-error approach, and the necessity of exploring diverse models and problem-solving strategies. This experience highlighted the utility of online resources, especially Google, and emphasized the significance of effective team communication in generating new ideas, synthesizing ideas, and enhancing overall model accuracy. In such a project which requires creativity, divergent thinking, and where there is no “correct answer,” trial and error and collaborative thinking was crucial for success.

VI. Sources

Almohalwas, Akram. “Kaggle Competition Fall 2023”. November 1, 2023.

Addiction Center. "Alcoholism Causes & Risk Factors." https://www.addictioncenter.com/alcohol/alcoholism-causes-risk-factors/.

Centers for Disease Control and Prevention. "Alcohol Use and Your Health." CDC, https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm.

Grant, Bridget F., et al. "Gender Differences in the Epidemiology of Alcohol Use and Related Harms in the United States." Alcohol Research: Current Reviews, vol. 40, no. 2, 2019, https://arcr.niaaa.nih.gov/volume/40/2/gender-differences-epidemiology-alcohol-use-and-related-harms-united-states.

National Institute on Alcohol Abuse and Alcoholism. "Alcohol Facts and Statistics." NIAAA, https://www.niaaa.nih.gov/sites/default/files/AlcoholFactsAndStats.pdf.