STAT 101C Final Project

Predicting Alcoholic Status Using Person's Vitals

Group 4 - Lec 2
Kailyn Nguyen, Teresa Bui, Kyle Wong, Andrew Arteaga

I. Abstract

In this Kaggle competition, our team was tasked with predicting the alcoholic status of a person given data on their vitals using several statistical learning models. The data set used was collected from the National Health Insurance Service in Korea. The following report details our full process: An introduction, data set cleaning/imputation, exploratory data analysis, feature selection, model construction, and a conclusion.

The final model is constructed using a boosted tree and uses all of the predictor variables. It has a Kaggle score of 0.73286 and we are ranked 11th.

II. Introduction

Alcohol use is a seemingly harmless pastime as these drinks are present in just about any adult social function. More than 85 percent of U.S. adults have had an alcoholic beverage at some point in their lives (Caron, 2023). However, alcohol use can easily turn into a dangerous and deadly addiction when it starts to cause issues in one's daily life. In South Korea, where the data is collected, approximately 50.8% of men and 26.9% of women reported binge drinking [consuming 7 or more standard drinks (7–8 g of pure alcohol) in 1 drinking session for men or 5 for women] in the past month (NLOM, 2020).

Short term health risks due to use of alcohol include: driving under the influence from poor judgment, alcohol poisoning, risky sexual behaviors, and or miscarriage or stillbirth for pregnant women (CDC, 2022). Long term health risks include: high blood pressure, cancers, weakening of immune system, learning and memory problems, mental health issues, and social life problems. By being able to properly diagnose an individual on their alcoholic status, we can resources early on to those in need of help.

The data set provided by the National Health Insurance Service in Korea contains 26 different variables of each individual’s vitals. The training set contains 70,000 observations while the testing set contains 30,000. Of the 26 variables, 20 are numerical and 6 are categorical.

III. Imputation of Missing Values

The initial imputation method was a mean/mode imputation. We used this for our initial exploratory data analysis and early model building just to get an idea of the data without having to use more extensive imputation methods. In some cases, a simple mean/mode imputation can provide extremely good results and this was no exception. We eventually noticed that this kaggle competition was coming down to accuracy differences within a fraction of a percent margin so the team concluded to try an alternative method of imputation to see if it could help improve our accuracy to compete with the rest of the class. To this end, we utilized the Hmisc package's function aregImpute() to impute the missing values. To see how this imputation method compared with our last, we pitted identical parameter models against each other, only changing the data from which they were built and seeing which yielded the better results. We found that across three different model types, the Hmisc imputation method filled in missing values better and continued on with the competition using the Hmisc imputed dataset.

IV. Data Analysis

Our team conducted various data analysis techniques before constructing a model which include the imputation of the data, an exploratory data analysis, and feature selection.

A. Cleaning the Data with Imputation

The training and testing data set had an overwhelming amount of missing data and if not handled properly it would negatively affect our model's performance. To combat this, we tried multiple imputation methods to see which would provide us with the highest accuracy score when it came to the actual model construction.

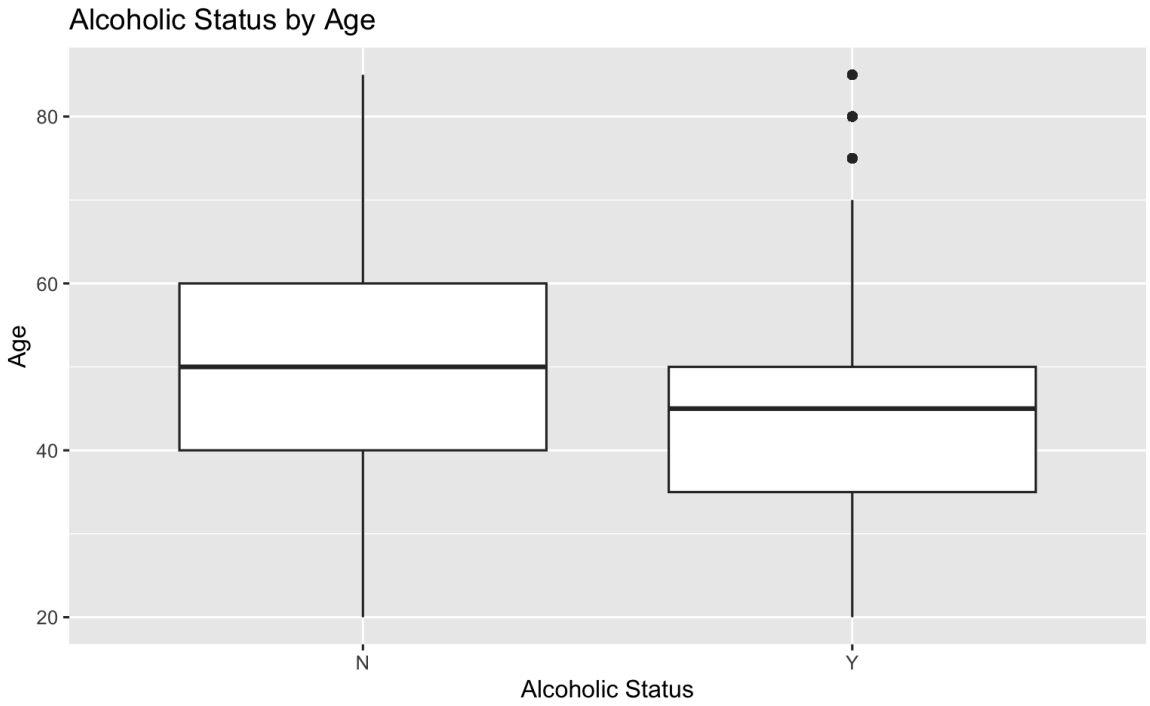
The first method of imputation we did was by replacing the missing numerical values with the mean and the missing categorical values with the mode. However, this resulted in wildly low accuracy rates ranging from 0.4 to 0.5. We suspect that this is because the mean and mode values would be highly biased and skewed, since there were many far away outliers for multiple predictor variables.

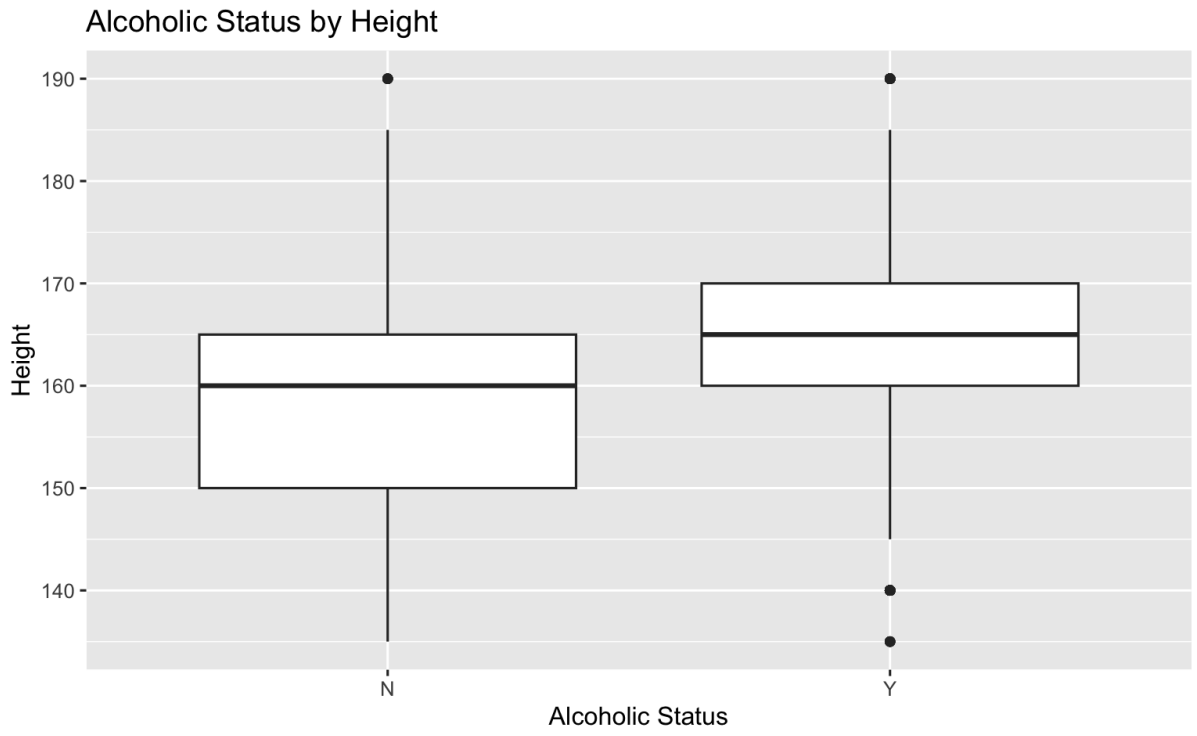
The best imputation method was by using the aregImpute() function in the HMISC package. We found that with the use of this function and keeping everything on the model but the data set constant, it improved our prediction accuracy across all models.

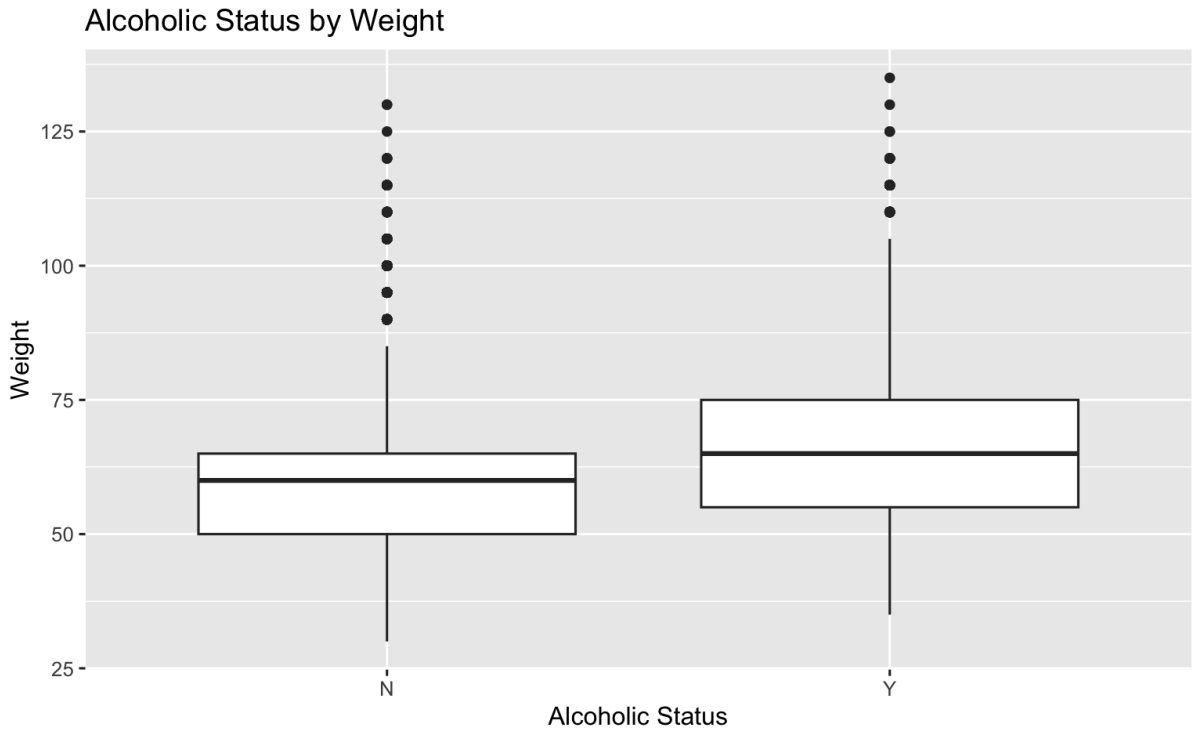
B. Exploratory Data Analysis

Numerical Variables

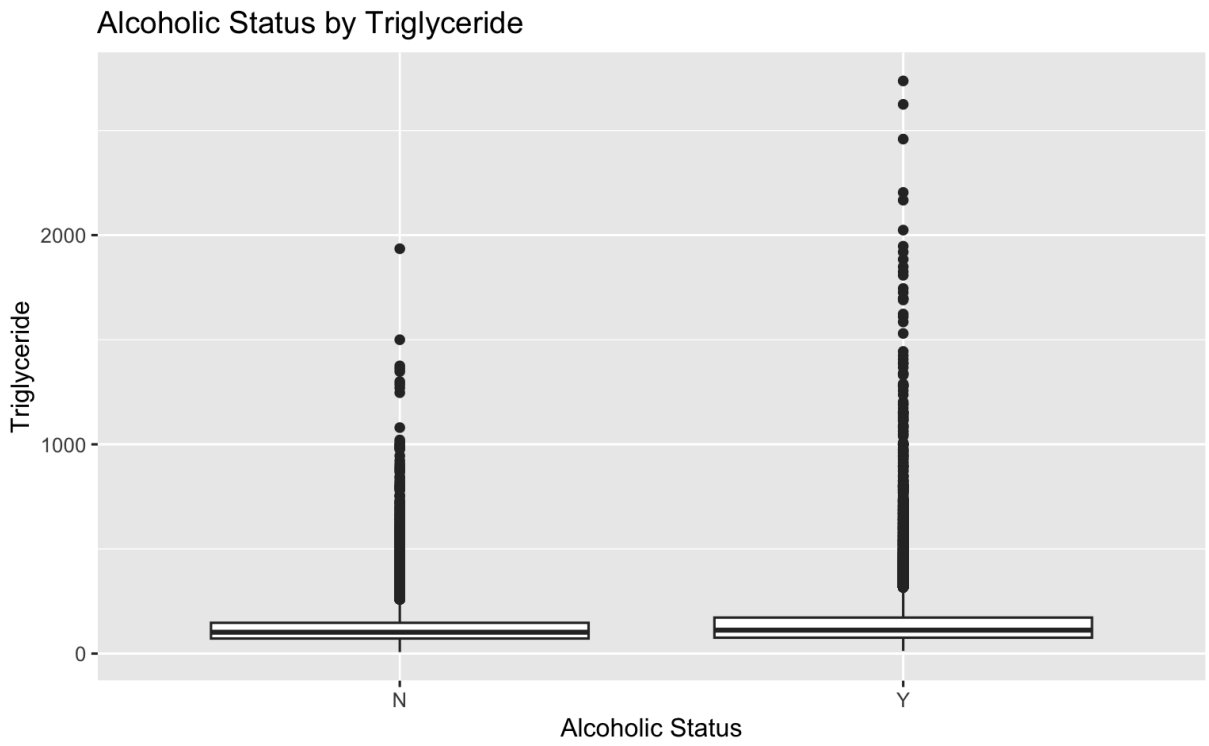
We ran boxplots on each of the numerical variables as it would help us to see the distribution of the data for each variable. Box plots with noticeable differences in mean and median could be potentially significant. We found that the variables that satisfy this were age, height and weight.







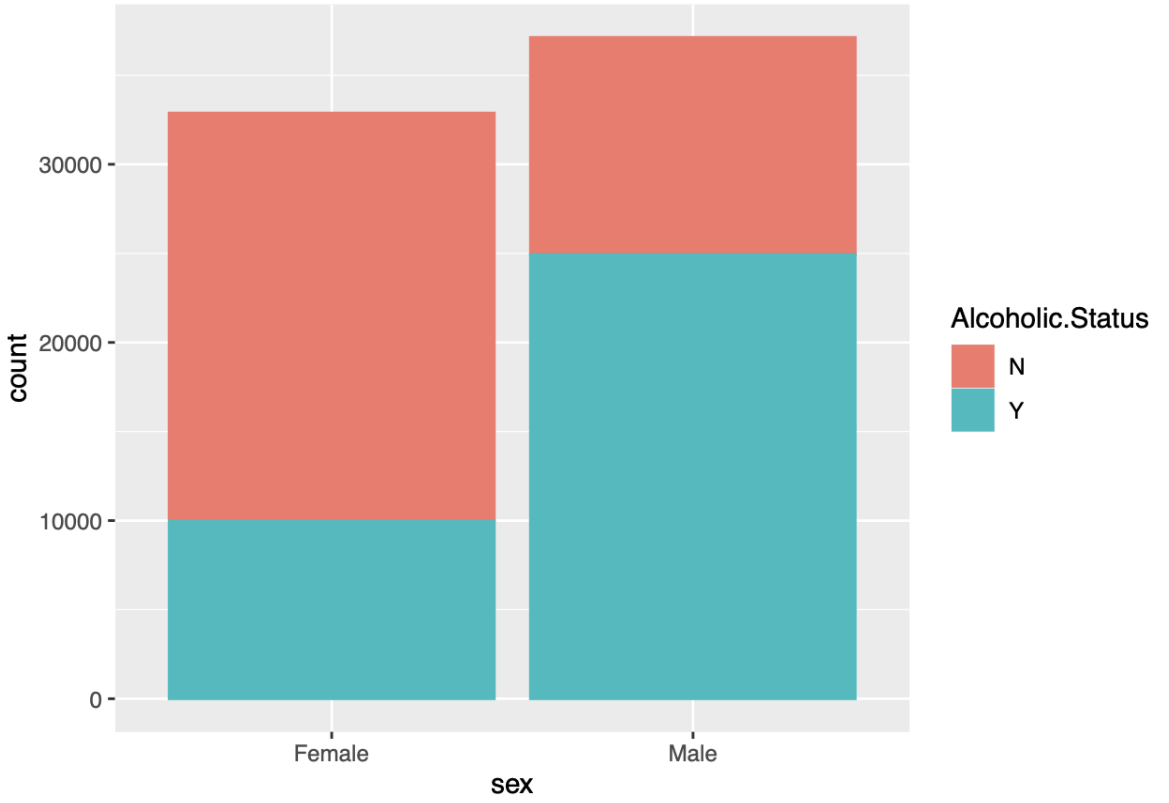
Majority of the other numerical predictors had many faraway outliers, making it really difficult to see the true distribution of the data, such as the triglyceride variable.



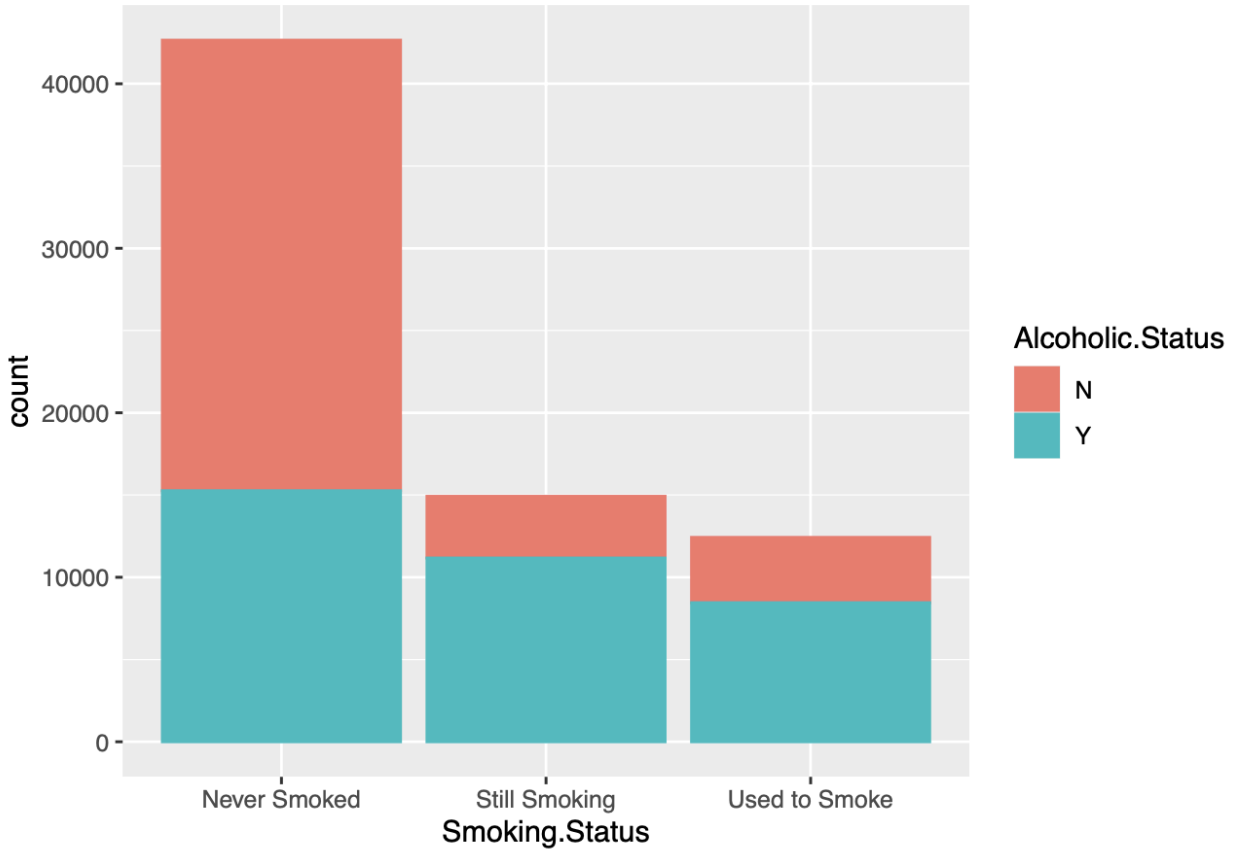
Categorical Variables

We found that stacked bar charts did the best at showing the distribution of the categorical predictors. Variables with many levels are more likely to be eliminated from the final model since it would be more difficult to interpret a multi-leveled predictor.

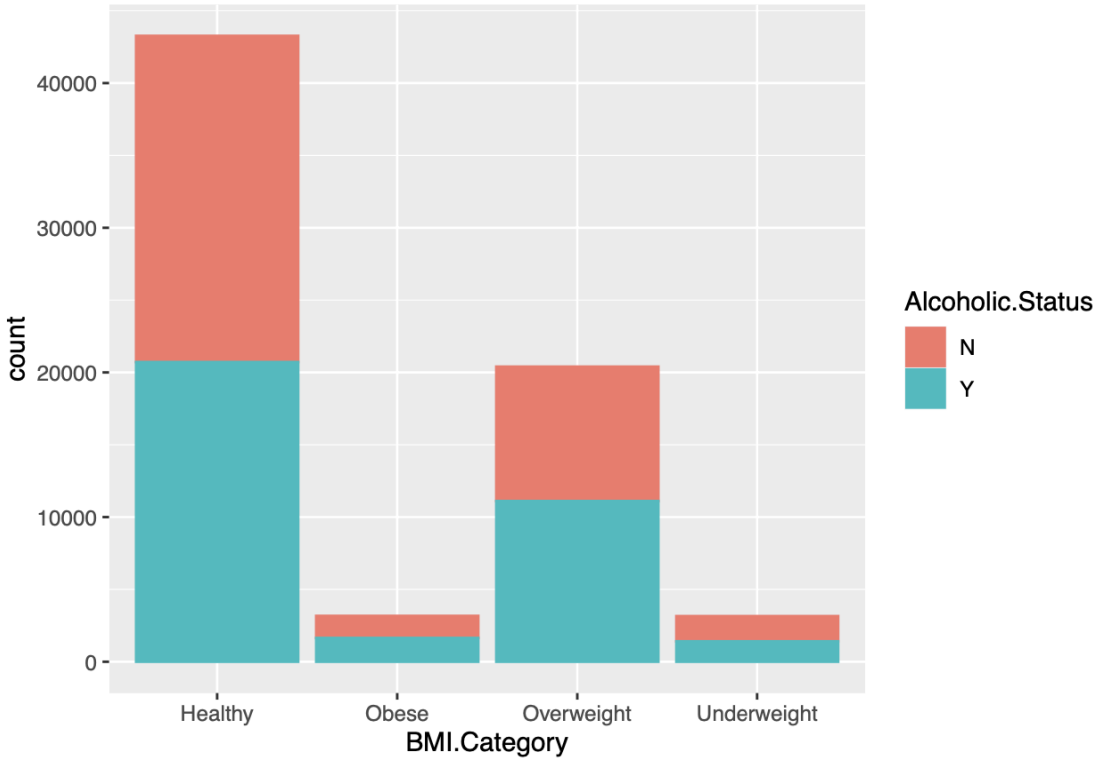
The variable sex is a potentially significant predictor as it does not have many levels and there is a clear difference between the amount of men and women and their alcoholic status.

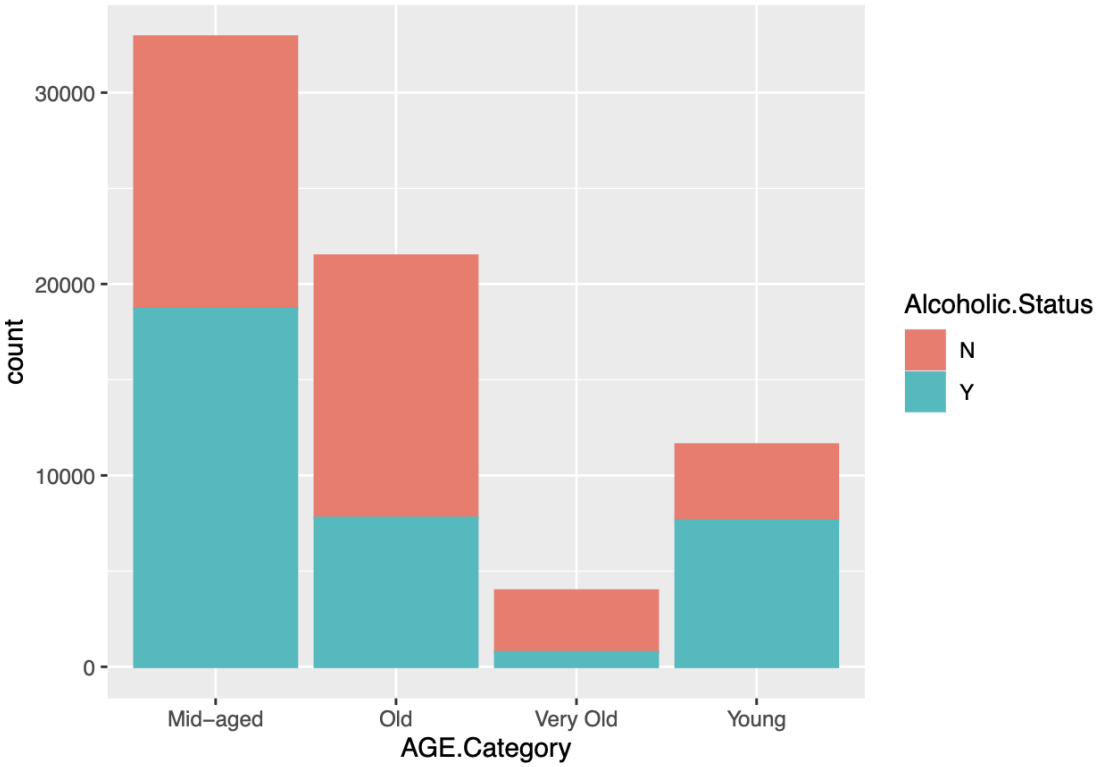


The variable Smoking.Status shows a clear difference in distribution, making it apparent that a large majority of those who responded with Never Smoked are also not alcoholics. The opposite is true for the categories Still Smoking and Used to Smoked, where the majority are alcoholics.


40000
30000
20000
10000
0
count
Never Smoked
Still Smoking
Used to Smoke
Smoking.Status
Alcoholic.Status
N
Y

The variables BMI.Category and AGE.Category are likely to be eliminated due to the amount of levels they have.

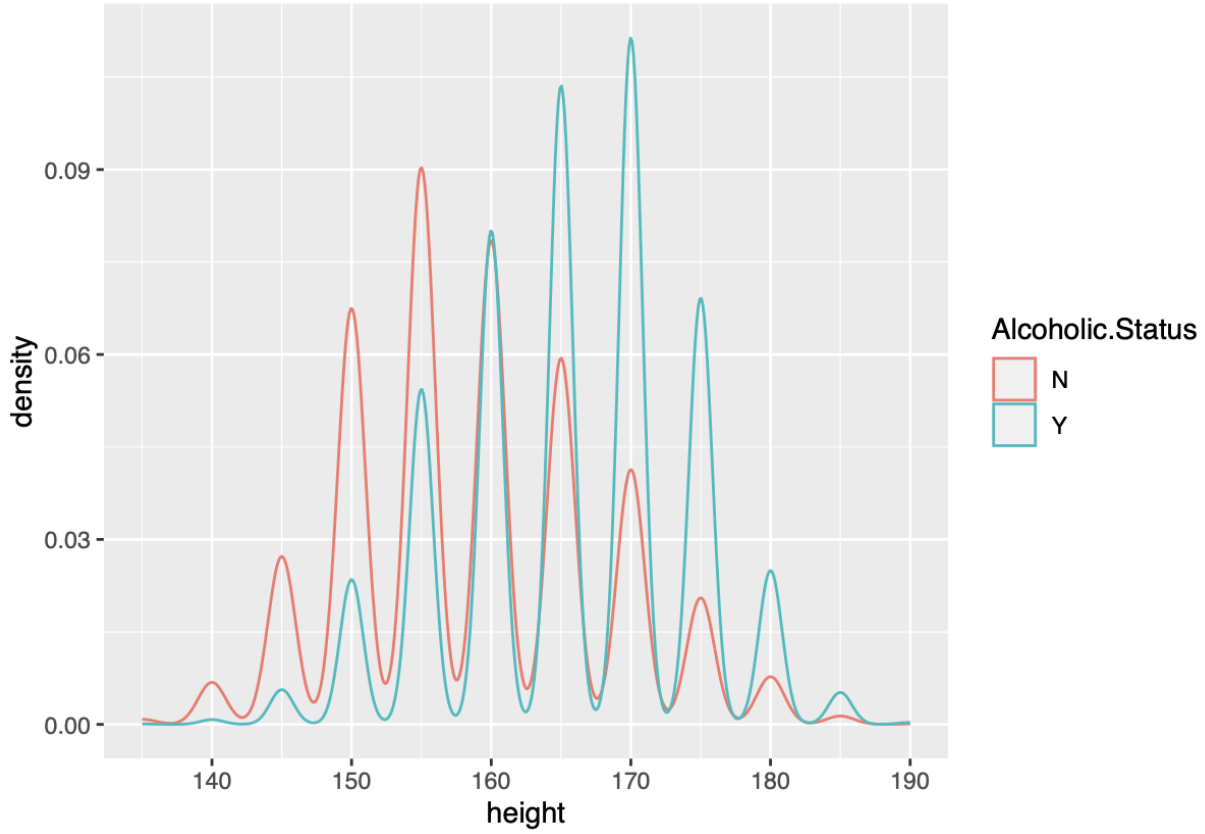


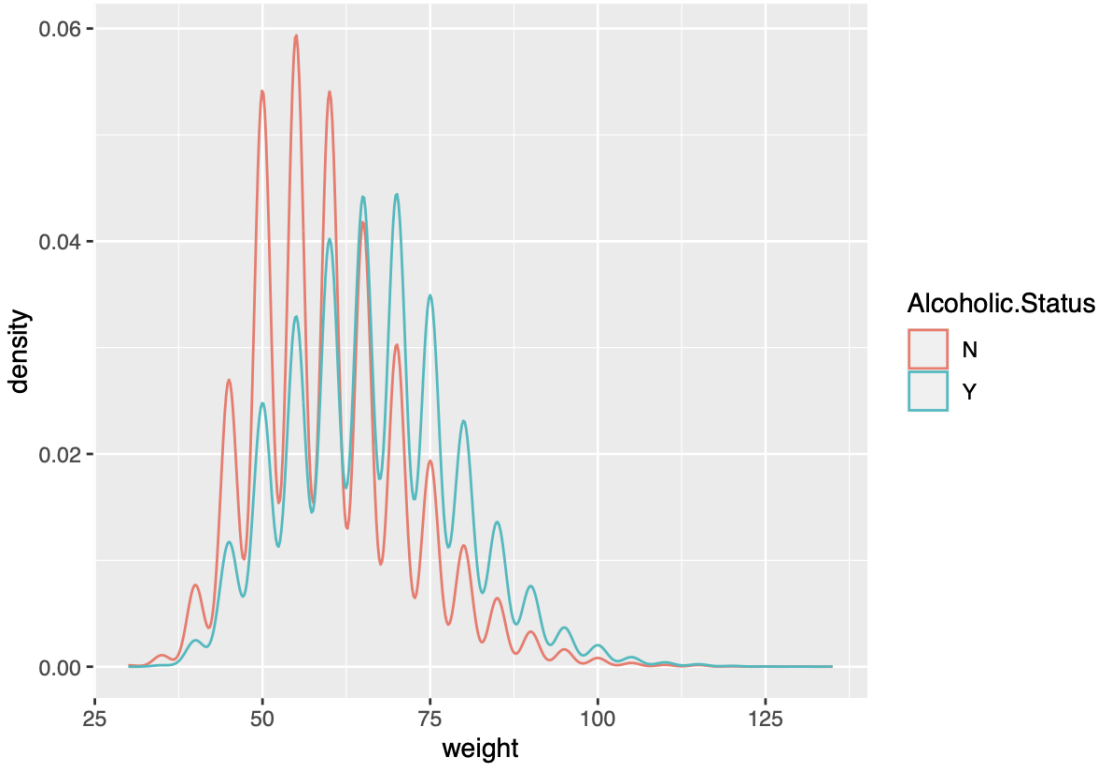


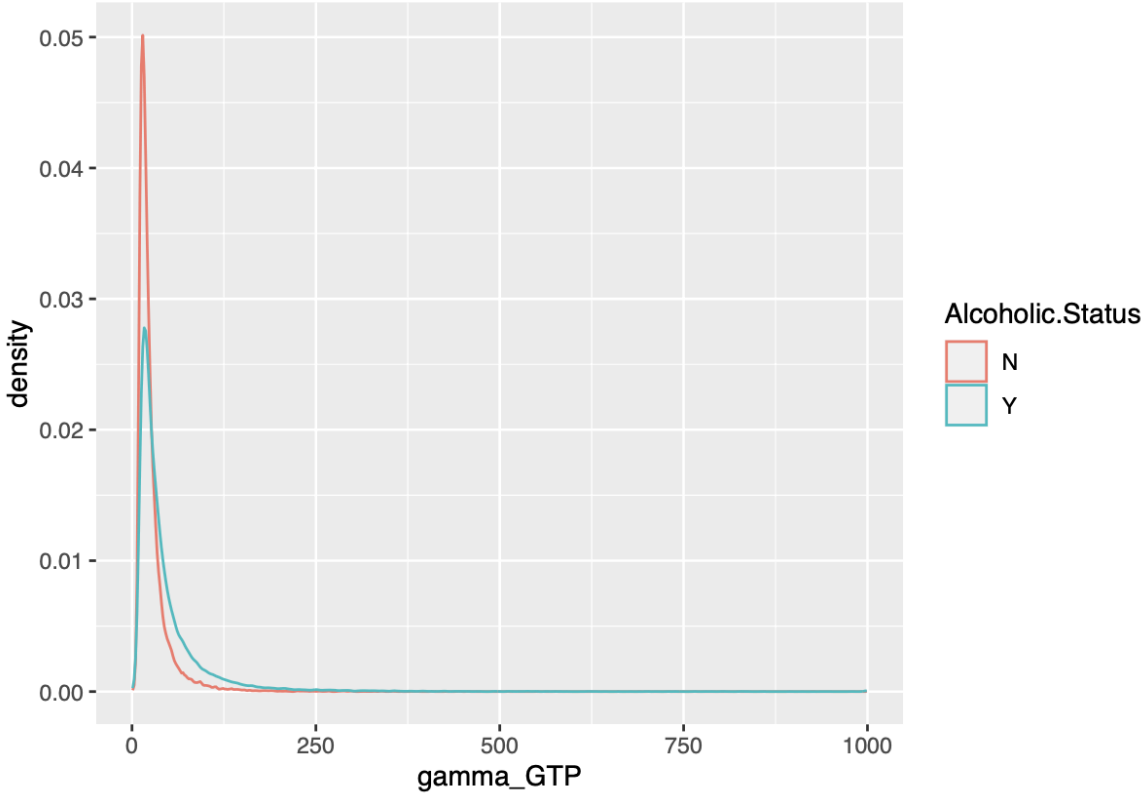
C. Feature Selection

With the information we gathered from the box plots in our EDA, we decided to conduct further testing and created density plots for each of the variables to aid us in our feature selection. Each density plot consists of the densities of each predictor plotted against the alcoholic status. Plots that show a difference in distribution are likely to be significant, while those that overlap would not.

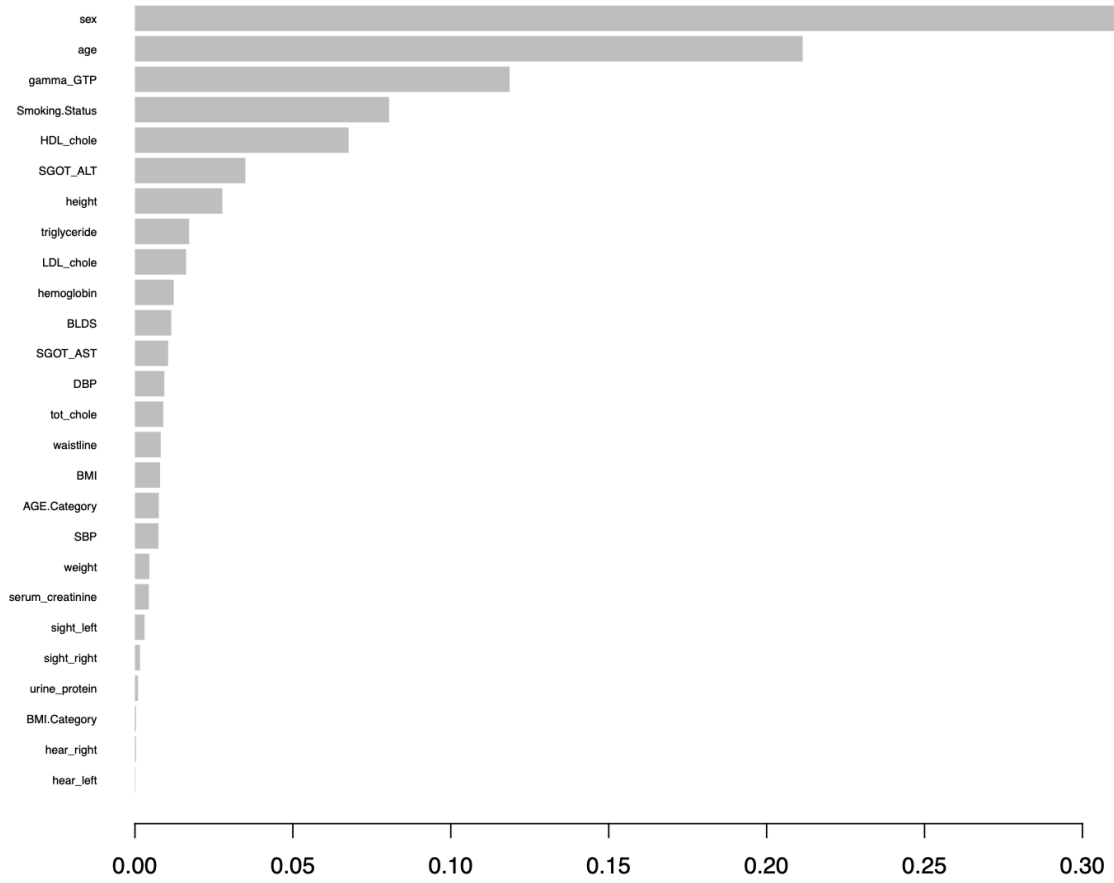
We found that the variables with the least overlap and significant differences in distribution were height, weight, and gamma_GTP.




0.06
0.04
0.02
0.00
density
25
50
75
100
125
weight
Alcoholic.Status
N
Y


0.05
0.04
0.03
0.02
0.01
0.00
density
0
250
500
750
1000
gamma_GTP
Alcoholic.Status
N
Y

Furthermore, we were able to plot a feature importance diagram using the XGBoost package. The model revealed that the most important variables used in creating the model were sex, age, gamma_GTP, Smoking.Status, and HDL_chole. Prior research suggested that gamma_GTP and someone's smoking status were usually good indicators of someone's alcoholic behavior, so this graph here reinforces those beliefs.



V. Method and Models

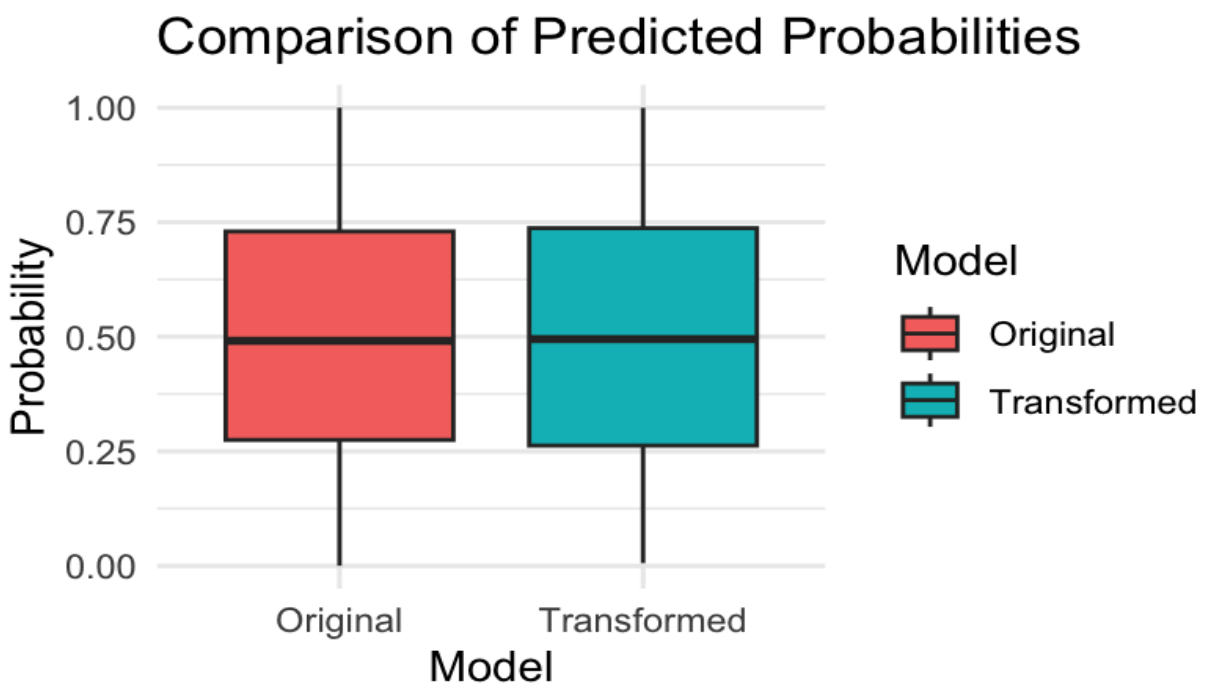
A) Boosted Tree

Our best model was created using Extreme Gradient Boosting (XGBoost). Like any tree based model, it's ability to make use of different types of predictors, which is particularly useful for our dataset. We created a very baseline xgboost model, only using cross-validation to determine the number of iterations that would yield the lowest error. After having determined this number to be 23, we used default parameters to create the first version of our model which performed exceptionally well and moved us up about 10 places in the competition. We decided to focus our attention on improving this model,

refining it further by tuning and making use of hyperparameters to get an even more accurate result. We also did some research to try and create new variables to improve the predicting power of our model. Some research showed that a ratio between AST and ALT above 1.5 is a great indicator of an individual's alcoholic behavior. Unfortunately, our findings did not agree with the articles we read because our model suffered a bit when we added in the new variable that contained the ratio. So we proceed back to tuning the model using tasks and learners from the MLR package. These would help us determine which hyperparameters would be best in reducing error. We learned in class that tree models tend to overfit data, applying these methods would help ensure that our model used the optimal parameters to yield the lowest error on our predictions.

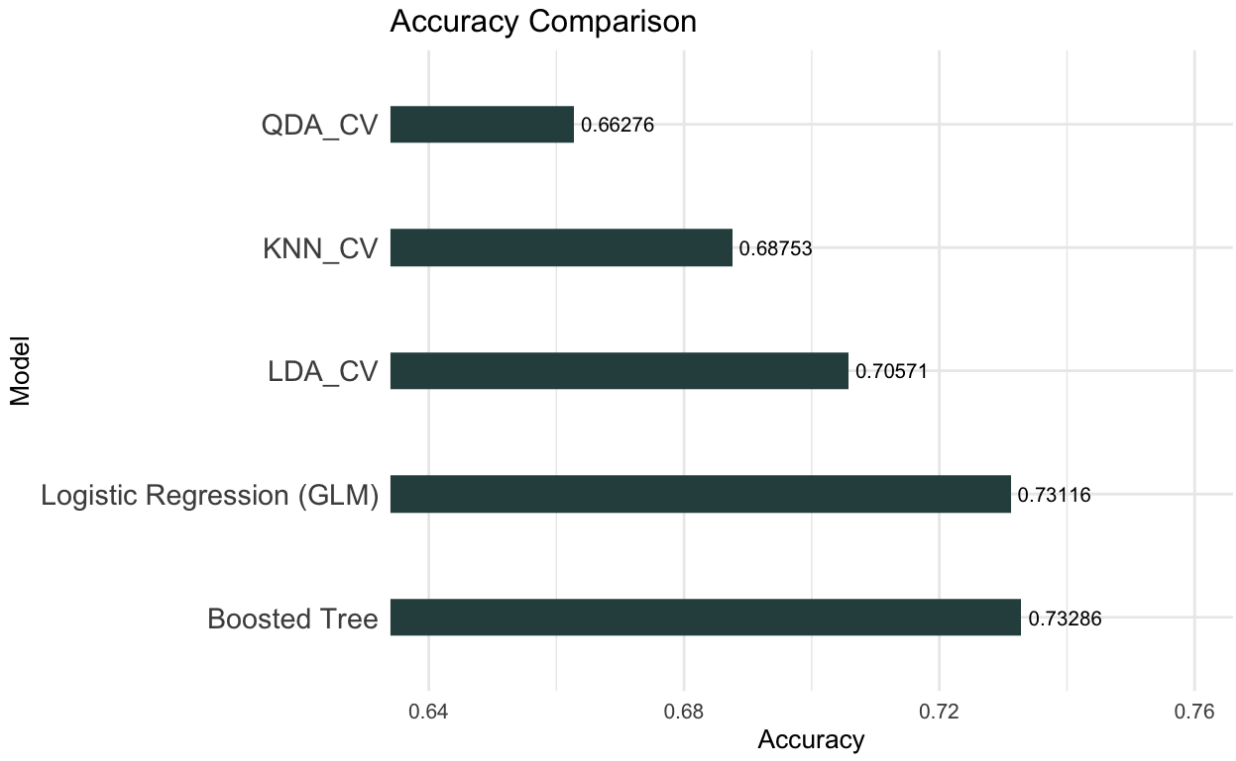
B) Logistic regression with GLM

Given that our dataset contains significantly more observations than predictors, we decided to implement a logistic regression model for our data. A logistic regression models the relationship between a binary response variable and one or more predictor variables by estimating probabilities. Thus, we deemed it is well suited for our data. On our first run with GLM, we decided to fit our model using only the numerical predictors after the imputation discussed under “Cleaning with Imputation.” which gave us an accuracy rate of 0.7104. To improve our score, we decided to fit our model with the full data and got an accuracy rate of 0.72336. Aiming for a higher accuracy rate, we decided to transform our data using log and by removing some predictors. Using our density and barplots from our EDA, we decided to remove both the quantitative and qualitative variables one-by-one. For the quantitative predictors, the insignificant variables were the ones that overlapped a lot, so we removed them separately to ensure that they are insignificant as the plots display. As for the qualitative predictors, we removed the ones with similar distributions. After doing so, we proceeded to transform the predictors that were heavily skewed. This model gave us an accuracy rate of 0.73123. Finally, after some additional research on the body vitals, we decided to add an interaction between *sex* and *gamma_GTP* which brought our accuracy rate up to 0.73166. This last accuracy rate was the highest we achieved using GLM.



C) Other Classifiers

In addition to boosted tree and logistic regression with GLM, we also created LDA, QDA, and KNN (k = 25) classifiers. Since cross-validation provides a more accurate estimate of how well a model is likely to generalize to new and unseen data, we found that implementing cross-validation in LDA, QDA, and KNN gave us better results. We chose 10 as our number of folds since having too high of a number of folds leads to heavier computations and having too low of a number might lead to a higher prediction error. Thus, our accuracies for our cross-validation classifiers were 0.70571, 0.66275, and 0.68753 respectively. Although cross-validation for these methods provided us higher accuracies than their non-cross-validation counterparts, our boosted tree and logistic regression with GLM still provided higher accuracies, with accuracies 0.73286 and 0.73116 respectively.



VI. Conclusion

With the Boosted Tree model we were able to get our highest accuracy score of 0.73286 placing us in rank 11 out of 35 teams. Our model includes all of the predictor variables, because we found that removing even insignificant variables decreased the accuracy of our predictions.

The model does come with some limitations that should be addressed. If the data is noisy, the boosted tree model may overfit and start modeling the noise. This could be solved with regularization techniques to mitigate the over-fitting. Also, boosted model trees can be computationally expensive and take a long time to run, making it a not very efficient model. And even though the model is able to run, it may be difficult to interpret the final model and its coefficients.

However, gradient boosted trees are known to make more accurate predictions when predictive performance is important. It is also extremely helpful since they work well with outliers, a common feature of our data set and an issue that heavily affected our accuracy. Though our model comes with its limitations, we found that it was the best fit for our imputation method.

VII. References

Caron. (2023). *The American Alcohol Problem: An Overlooked and Deadly Epidemic.* Caron Treatment Center. https://www.caron.org/blog/the-american-alcohol-problem#:~:text=Alcoholism%20Stats&text=More%20than%206%20percent%20of,year%20in%20the%20United%20States.

Park, J., Sohn, A., Choi, C. (2020, December). *Solitary and Social Drinking in South Korea: An Exploratory Study.* National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7752144/

Centers for Disease Control and Prevention. (2022, April 14). *Alcohol Use and Your Health.* https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm#:~:text=Long%2DTerm%20Health%20Risks,liver%20disease%2C%20and%20digestive%20problems.&text=Cancer%20of%20the%20breast%2C%20mouth,liver%2C%20colon%2C%20and%20rectum.

Saraswat, M. (n.d.). Beginners tutorial on XGBoost and parameter tuning in R tutorials & notes: Machine learning. HackerEarth. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/

BALLDIN, J. et al. (2004) HIGH AST/ALT RATIO MAY INDICATE ADVANCED ALCOHOLIC LIVER DISEASE RATHER THAN HEAVY DRINKING, Academic.oup.com. https://academic.oup.com/alcalc/article/39/4/336/139692 (Accessed: 08 December 2023).