University of Missouri

STAT 8330

# Binary Prediction of Smoker Status using Bio-Signals

**Souvik Bag**

14421334

## Introduction

Tobacco epidemic is one of the biggest public health threat that world has ever seen. Tobacco kills over 8 million people every year worldwide. Smoking stands as one of the primary avenues of tobacco usage. The data on smoking holds particular significance for government bodies and public health agencies. Within this project, our aim is to categorize the smoking status of adults by leveraging various biosignals, including factors like age, height, weight, cholesterol levels, and more.

## 1 Exploratory data analysis

The training data has 23 predictor variables and the binary response variable "smoking" status where "1" means the person smokes and "0" otherwise.

First we convert all the qualitative variables "smoking", "hearing left", "hearing right", "Urine protein" and "dental caries" to factors. We also discard the "id" variable as it is of no interest in terms of data analysis.

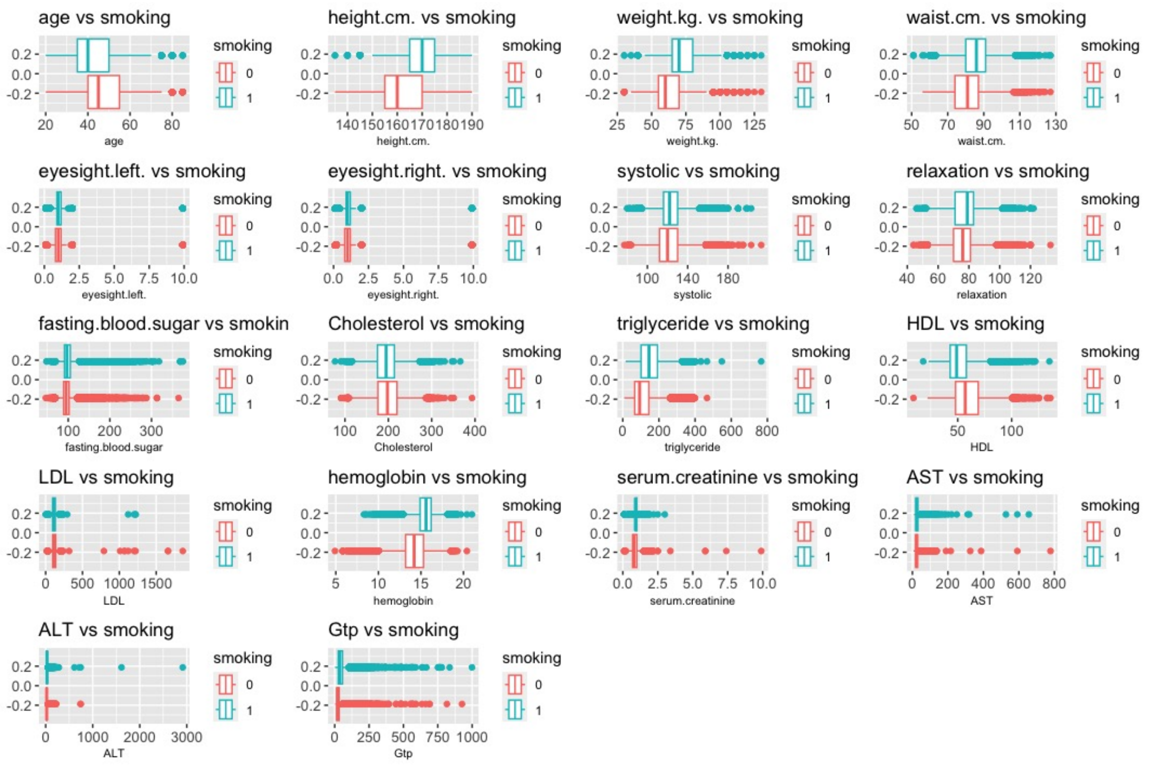


Figure 1: Boxplot of quantitative variables

Next we plot the boxplot for all the quantitative variables for different smoking status. From Figure (1) we can see that some variables like age, height, weight, waist size, triglyceride, HDL and hemoglobin have different distribution for different smoking status indicating that they might be useful for our prediction.

Next we check if our data suffers from multicollinearity problem or not. We use a corplot to show the correlation coefficient between all the quantitative variables.

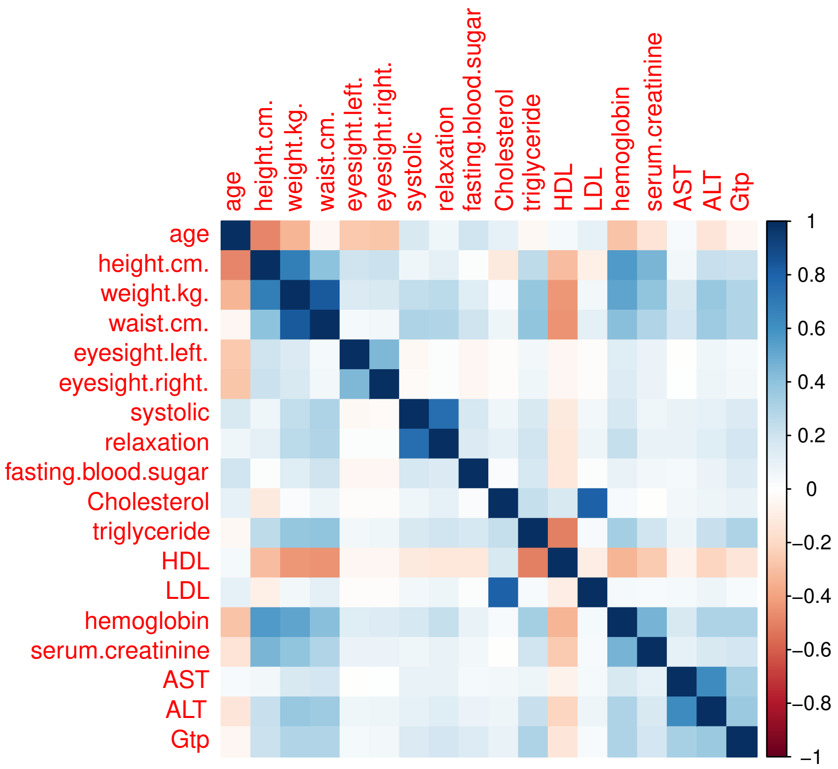


Figure 2: Boxplot of quantitative variables

From Figure (2) we can see that there are several instances of multicollinearity. For example, weight and waist size are highly correleated. Cholesterol level and LDL level are highly correlated. Meanwhile triglyceride and HDL are highly negatively correleated.

## 2 Model fitting

As a natural choice of model fit, first we fit a logistic regression model with all the variables in it.

Before we fit our model, we further create a partition of our train data set into a training and testing data. Then we fit the logistic regression model on that training set.

$$logit(\hat{p}) = -20.60 - 0.0008395 \times Age + 0.08899 \times Height - 0.02281 \times Weight + \ldots + 0.3626 \times Dental.caries1$$

From the estimated coefficients and their corresponding p-values we can see that variables such as age, eye sight left, hearing conditions etc are not statistically significant.

And we get an AUC value of 0.847 on our test data set which is quite good for our basse model.

Now we will try to improve over this model.

Because from figure (2) we can see that our predictor variables are correlated, we use ridge and Lasso regularization to mitigate this issue.

We create two data matrix for training and testing data. And fit a lasso model using "glmnet" package. We use cross validation for choosing the optimal value for our hyperparameter $\lambda$.

Surprisingly with ridge logistic regression we get AUC value of 0.828 which is lower than that of logistic regression.

Next we try our hands on non-parametric techniques such as random forest.

First we fit a random forest with 100 trees. We get a AUC of 0.87 which is a significant improvement over the logistic regression model.

The variable importance plot (3) shows that hemoglobin level, height, Gtp and triglyceride level are some of the most important variables in splitting the trees.

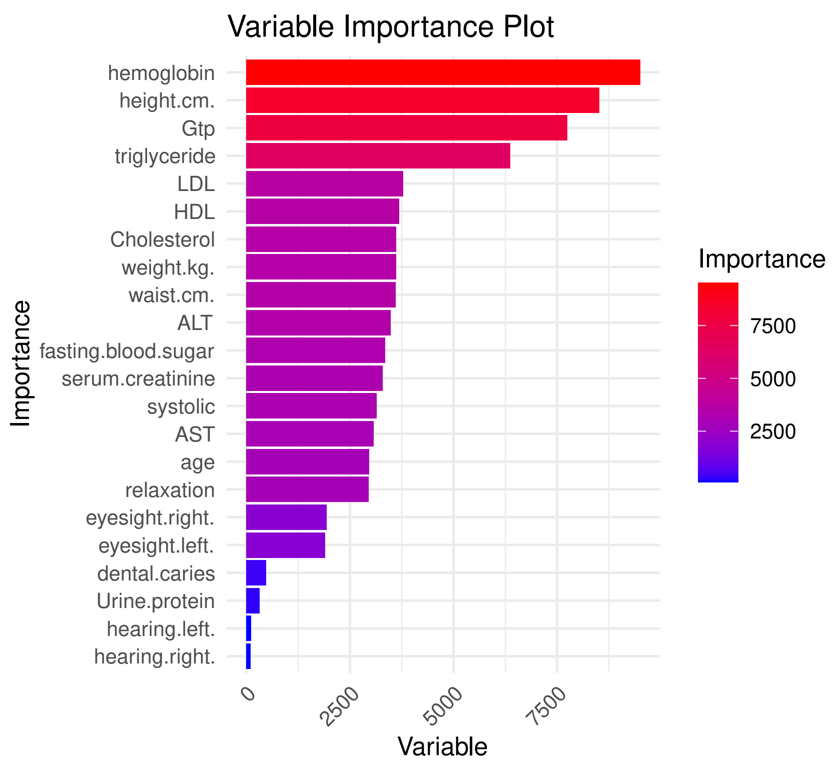


Figure 3: Variable importance plot

Now we will try to ensemble different trees using extreme gradient boosting.

We used the "xgboost" library to fit the model. In order to use the library, we first need to convert our training data into a matrix and convert our factor variables to dummy variables. Then we specify the parameters of the model with multiclass classification problem and multiclass logarithmic loss function.

We further played with different values of $\eta$, $\gamma$ and "max.depth" reduce the overffitting.

Figure (4) shows the training and testing log loss at different iteration levels.

We also fitted a neural network with two hidden layers. We acheived a AUC score of 0.857.

These two i.e. xgboost and neural network are my best fitted model.

## 2.1 Model description of two best models

In our best performing model i.e. xgboost, we implemented 500 rounds, value of $\eta$ parameter was choosen to be 0.09 to mitigate overfitting, max.depth parameter was choosen to be 7 and $\gamma$ is 2. We acheived a AUC score of 0.881 on our test data and a kaggle score of 0.87131 (5).

In our second best model we implemented a neural network with two hidden layers, 50 input units per hidden layer, "RELU" activation function and dropout fraction of 0.4 at both the layer. It yields a kaggle score of 0.85661 (6).

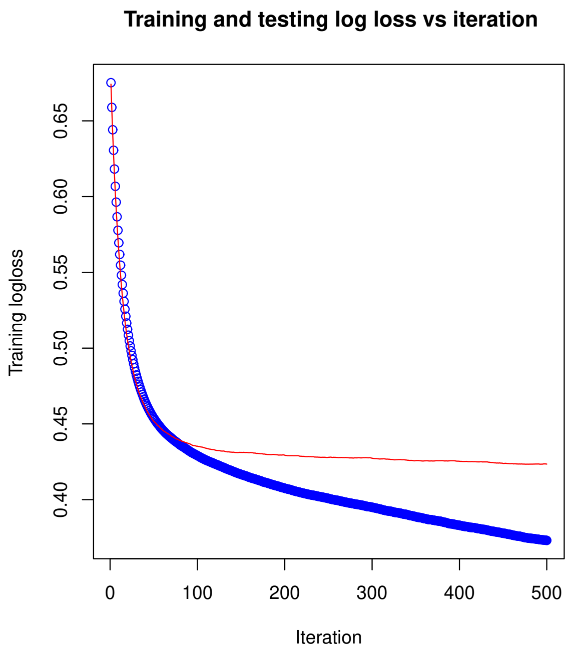
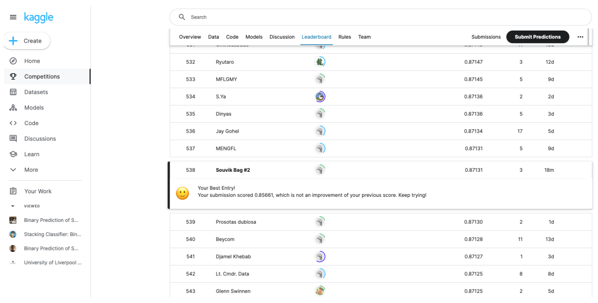


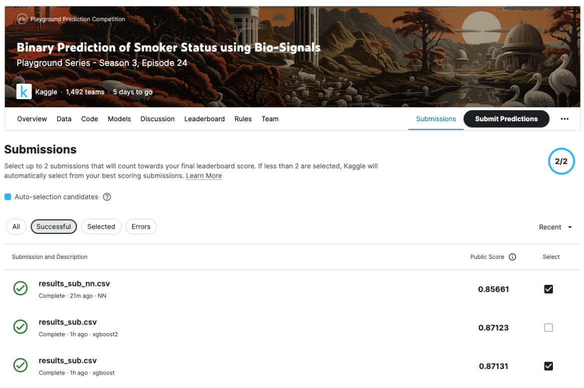
Figure 4: Xgboost training and test log loss



| | | | | |
|---|---|---|---|---|
| 532 | Ryutaro | 0.87147 | 3 | 12d |
| 533 | MFLGMY | 0.87145 | 5 | 9d |
| 534 | S.Ya | 0.87136 | 2 | 2d |
| 535 | Dinyas | 0.87136 | 5 | 3d |
| 536 | Jay Gohel | 0.87134 | 17 | 5d |
| 537 | MENGFL | 0.87131 | 5 | 9d |
| 538 | **Souvik Bag #2** | 0.87131 | 3 | 18m |
| 539 | Prosotas dubiosa | 0.87130 | 2 | 1d |
| 540 | Beycom | 0.87128 | 11 | 13d |
| 541 | Djamel Khebab | 0.87127 | 1 | 3d |
| 542 | Lt. Cmdr. Data | 0.87125 | 8 | 8d |
| 543 | Glenn Swinnen | 0.87125 | 2 | 5d |

Your Best Entry!
Your submission scored 0.85661, which is not an improvement of your previous score. Keep trying!

Figure 5: Xgboost AUC score submission


Playground Prediction Competition
Binary Prediction of Smoker Status using Bio-Signals
Playground Series - Season 3, Episode 24
Kaggle · 1,492 teams · 5 days to go
Overview
Data
Code
Models
Discussion
Leaderboard
Rules
Team
Submissions
Submit Predictions
2/2
Auto-selection candidates
All
Successful
Selected
Errors
Recent

Submissions

Select up to 2 submissions that will count towards your final leaderboard score. If less than 2 are selected, Kaggle will automatically select from your best scoring submissions. Learn More

| Submission and Description | Public Score | Select |
|---|---|---|
| results_sub_nn.csv<br>Complete · 21m ago · NN | 0.85661 | ✓ |
| results_sub.csv<br>Complete · 1h ago · xgboost2 | 0.87123 | ☐ |
| results_sub.csv<br>Complete · 1h ago · xgboost | 0.87131 | ✓ |

Figure 6: Neural Network AUC score submission

## 3 Limitations & future work

The main limitations of our work in terms of neural network training is that we could explore a more complicated model with high number of parameters and hidden layers. Even though it will require very high computational power but it might yield better results. Otherwise extreme gradient boosting performs best for this dataset and the kaggle submission results are also preety similar in this case.

## 4 References

- ChatGPT
- ISLR2 - Daniela Witten, Gareth M. James, Trevor Hastie, Robert Tibshiran