

Predicting Smokers

Andrew Ross

2023-11-08

Introduction

“Do you smoke?” is a common question everyone gets asked when they have a doctor’s appointment. Some lie while others tell the truth. If the patient lies, it doesn’t help them obtain quality care. If someone lies about smoking, it can lead to other altercations for their health. It would be nice to have a quantitative scientific answer rather than listening or seeing into someones lungs. Here we present two models (random forest and boosting) to help predict someone smokes based on many typical vitals taken at the doctors. We used other models that will not be shown in detail, but we will show how they performed.

Data

Feature Engineering and EDA

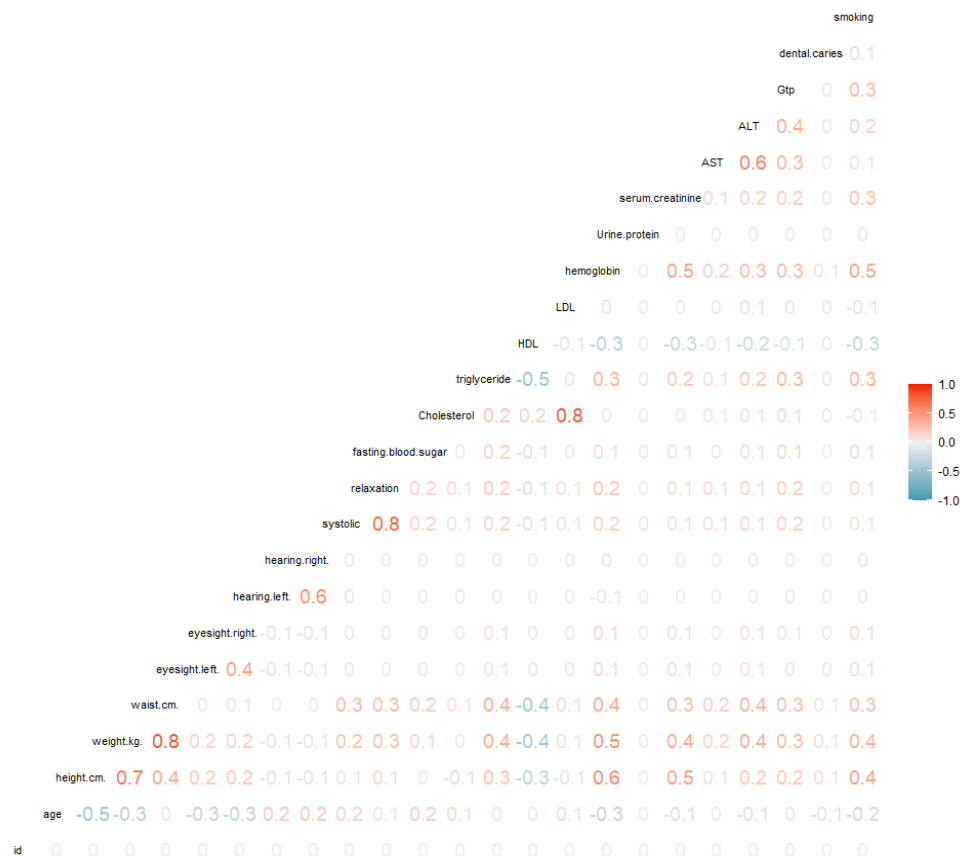
Below is a glimpse of the data, so we can understand the information.

feature	num_missing
id	0
age	0
height.cm.	0
weight.kg.	0
waist.cm.	0
eyesight.left.	0
eyesight.right.	0
hearing.left.	0
hearing.right.	0
systolic	0
relaxation	0
fasting.blood.sugar	0
Cholesterol	0
triglyceride	0
HDL	0

feature	num_missing
LDL	0
hemoglobin	0
Urine.protein	0
serum.creatinine	0
AST	0
ALT	0
Gtp	0
dental.caries	0
smoking	0

We can see we have 24 features and a sample size of 159256 patients. The entire data set is complete. Each variable in the data set is quantitative or factors such as whether or not someone smokes (1 = yes, 0 = no). We notice the data has multiple vitals as well as if someone can see or hear. There's quite a bit of information related to blood.

We can see the relationship between all variables in the plot below:



Smoking is not highly correlated with any feature. For those that have high correlation, weight and height, systolic and relaxation, cholesterol and LDL are correlated pairs.

We use the relationship of weight and height to create a new variable called body mass index (BMI). BMI screens for weight categories that may lead to health problems. The formula for BMI is as follows:

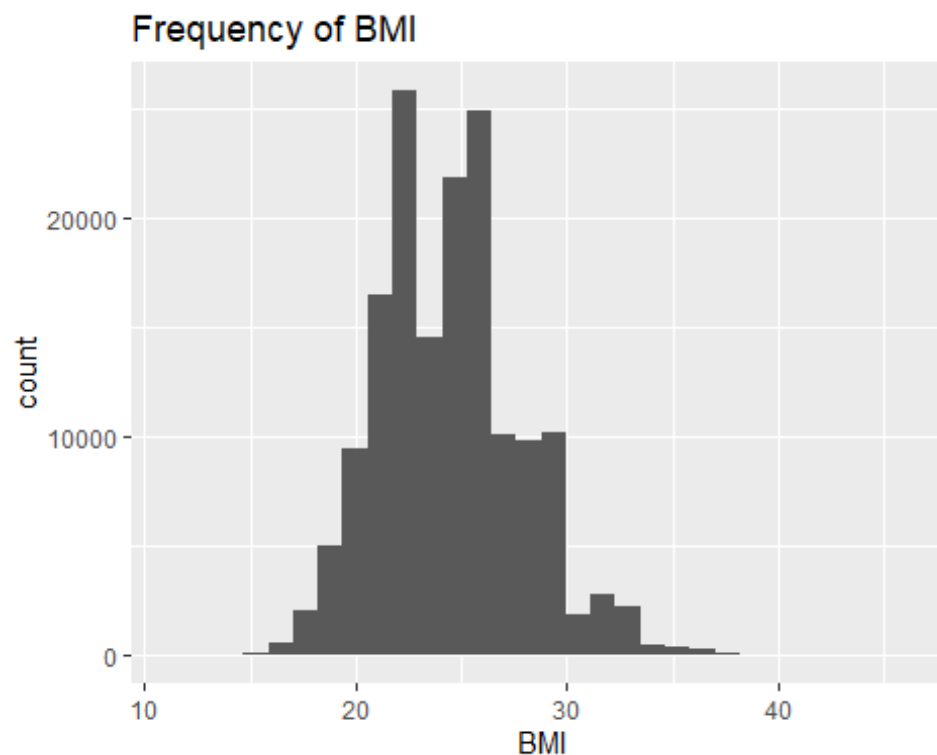
$$BMI = weight_{kg} / (height_m^2)$$

We were given height in centimeters which was properly converted to meters.

Here we present two graphs: We create a histogram of body mass index (BMI) and the other being smoking (our predictor).

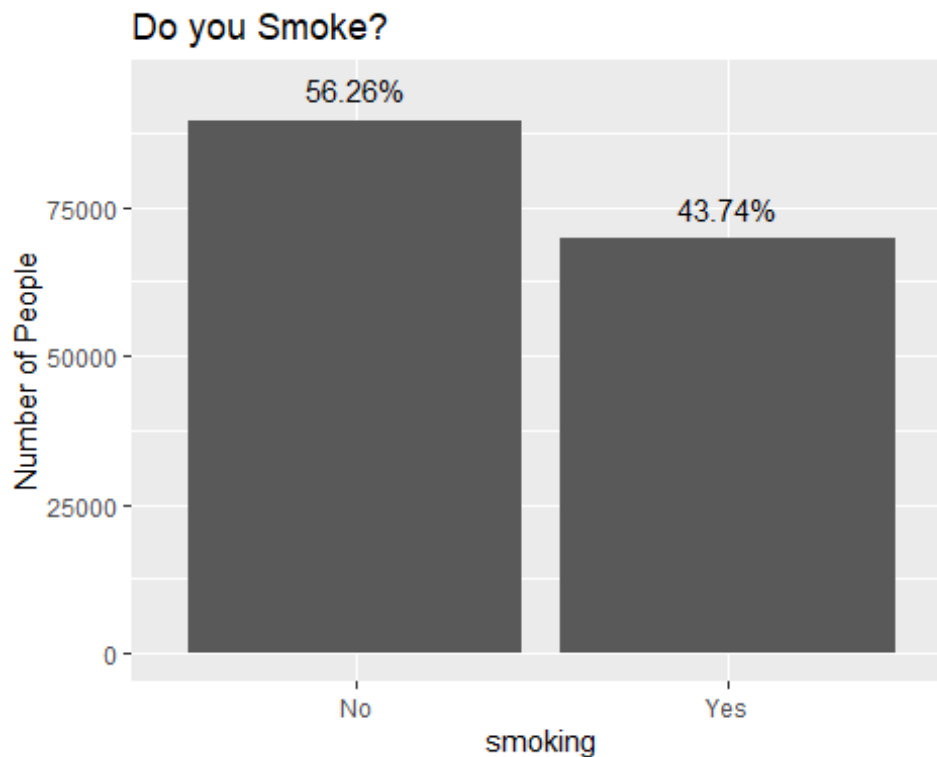
The reason for constructing this variable is based off of the correlation between height, weight, and waist. Below is a correlation matrix for BMI, weight, height, and waist.

	BMI	height.cm.	weight.kg.	waist.cm.
BMI	1.0000000	0.1676777	0.8275171	0.8180234
height.cm.	0.1676777	1.0000000	0.6866450	0.4094996
weight.kg.	0.8275171	0.6866450	1.0000000	0.8302078
waist.cm.	0.8180234	0.4094996	0.8302078	1.0000000



BMI is a function of height and weight as it will be correlated with them. We can see BMI is correlated with waist as well. The rest of the variables are correlated with each other.

The graph below shows the proportion of smokers vs nonsmokers. The data is almost split evenly.

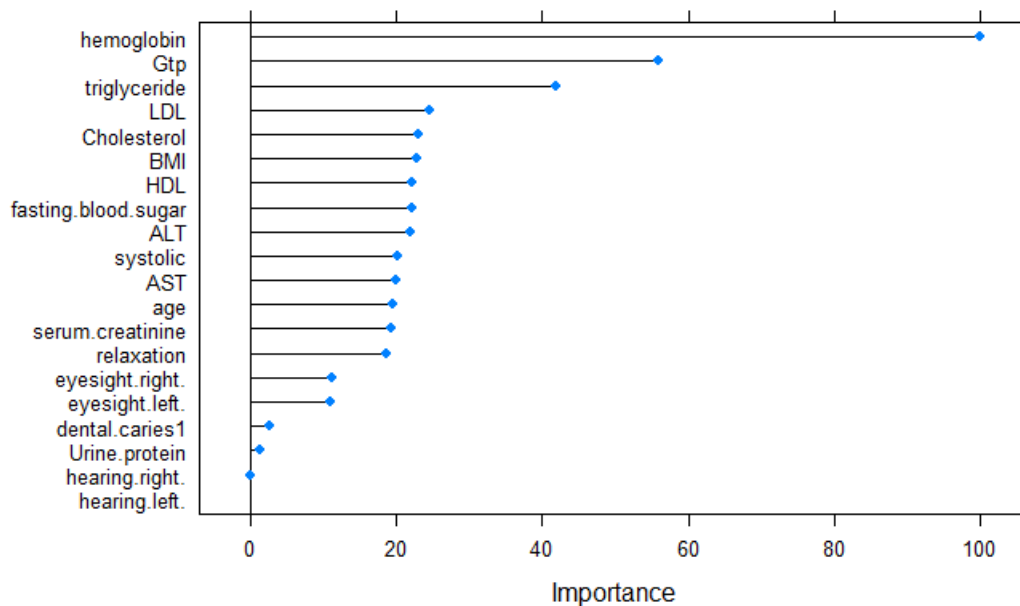


Model Building

We use a 70-30 training and test split because the goal will be to get the highest area under the ROC curve, and Kaggle does not provide a 'smoking' column in the test set for us to use the AUC function. The models we considered were logistic regression, Random forest, XGTree boost, MARS, Linear and quadratic discriminant analysis, and K-Nearest Neighbors. We will speak in more detail about XBTre boost and Random forest, since these two models performed the best while the results from the others will only show the AUC metric.

First, let us look at random forest. Using 10 fold cross validation, we find the models performs the best on the test data set when there are 11 variables considered at each split. We show the variable importance below:

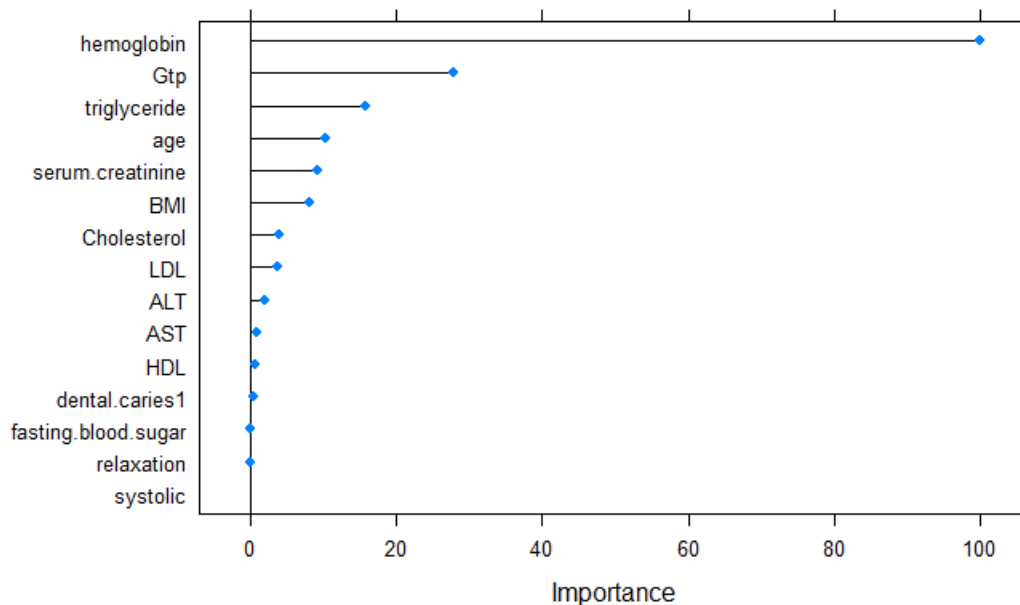
Random Forest



One could remove dental.varies, urine protein, and below as they won't change much of the prediction accuracy for the testing set. The final accuracy for the model is 84.81% for area under the ROC curve.

Boosting

Now we implement a XBTtree boosting model. For this model, we did remove eyesight, hearing, and urine protein. They were not important just like in random forest. We use 10 fold cross validation to find the optimal number for eta to be 0.4, which represents a weight for tree values. The number of rounds is 150 and max depth is 3. Below is a plot of the variable importance:



Any value below BMI may not be important for this model to predict smoking. The final accuracy for this model is 85.99% area under the ROC curve.

To compare the two models, we notice hemoglobin is the greatest predictor and every variable after does not live up to hemoglobin's potential. After creating BMI as a variable, it is nice to see that it is a top predictor for these models.

MARS

LDA and QDA

KNN

Conclusion

Below is a table showing the AUC for each model:


Model	AUC
KNN	80.33
LDA	82.04
QDA	79.74
Logistic	82.36
Random Forest	84.81
Boosting	85.99
MARS	83.68

As stated previously, boosting and random forest had the highest AUC percentage. The other models were not far away from these two and can still be reliable options.

The last thing I want to bring to attention is that we performed other feature engineering that wasn't successful and will be left out for this report. One example we tried is for each variable that is related to blood (systolic, HDL, LDL, etc). We can bin those variables into classes. For example, high, normal, or low blood pressure. The models performed worse with those classes than the quantitative values. We could speak to subject matter experts to find a better way to group variables, or we could use a categorical boosting model. Furthermore, blood results differ between men and female, so it would be interesting to see which patients are at-birth male and female in future analysis. There are other variable related measurements such as AST/ALT.

Finally, a screenshot of my kaggle score:

YOUR RECENT SUBMISSION



predictions.csv
Submitted by Philip Ross · Submitted 32 seconds ago

Score: 0.86399
Private score:

[↓ Jump to your leaderboard position](#)