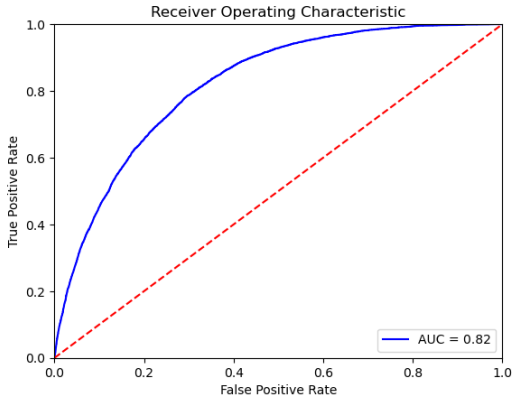
FML HW3

1. I loaded the data as a pandas data frame. I normalized the features BMI, mental health and physical health. I set the first column “diabetes” as y and the other columns as predictors. I put them into a logistic regression model and drew a ROC graph and calculated AUROC. For finding the most important predictor, I used a for loop which iterated through all the predictor columns and in each loop the column in question was dropped from the original training and testing sets and the new AUROC was recorded. I then found the smallest AUROC value.

Because BMI, mental health and physical health are not categorical data, I felt that it is best to normalize them. The smallest AUROC value is linked to the predictor that will cause the biggest drop in performance, indicating that the predictor is the most important feature.

The original AUROC using all predictors is 0.8198905022868597. The lowest AUROC is 0.80450194, which is caused by omitting the third predictor BMI.

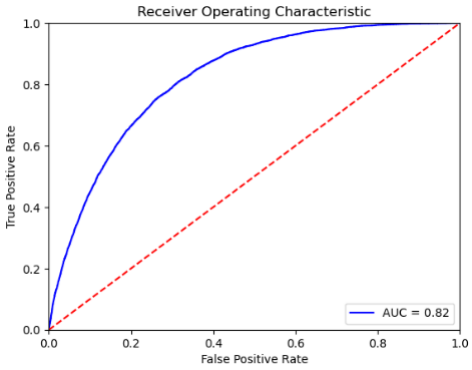


BMI is the best predictor in the logistic regression model.

2. I used a linear SVM model with C=1. The rest is the same as Q1.

The reasons are the same as Q1.

The original AUROC using all predictors is 0.8213422943230673. The lowest is 0.80458572, which is caused by omitting the third predictor BMI.

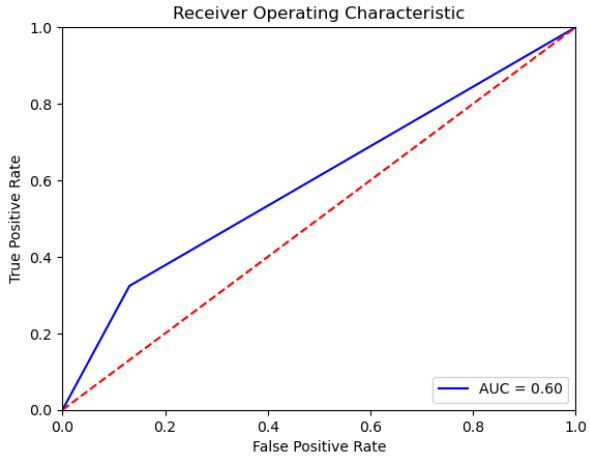


BMI is the best predictor in the SVM model.

3. I used a single decision tree with the Gini coefficient. The rest is the same as Q1.

The reasons are the same as Q1.

The original AUROC using all predictors is 0.5967340341892741. The lowest is 0.58255473, caused by omitting BMI.

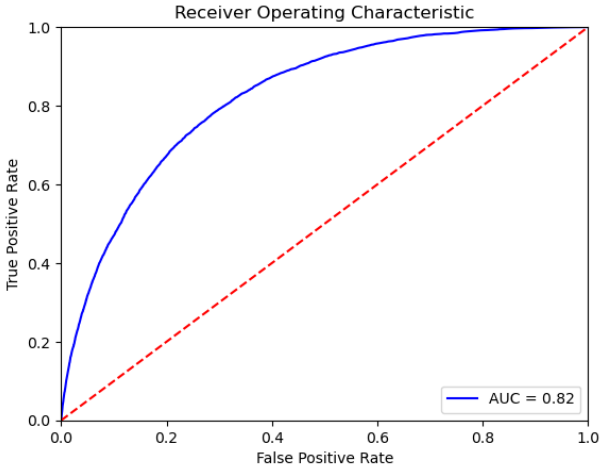


BMI is the best predictor in a single tree model, but the performance of the single tree model is bad compared to other models, indicating its status as a weak learner.

4. I used a random forest model which consists of 1000 trees. Each tree is trained on a bootstrapped dataset that is 10% the size of the full training set. Each new branch considers a random 50% of all features. The rest is the same as Q1.

The reasons are the same as Q1.

The original AUROC using all predictors is 0.8238527224589606. The lowest is 0.79901006, caused by omitting BMI.

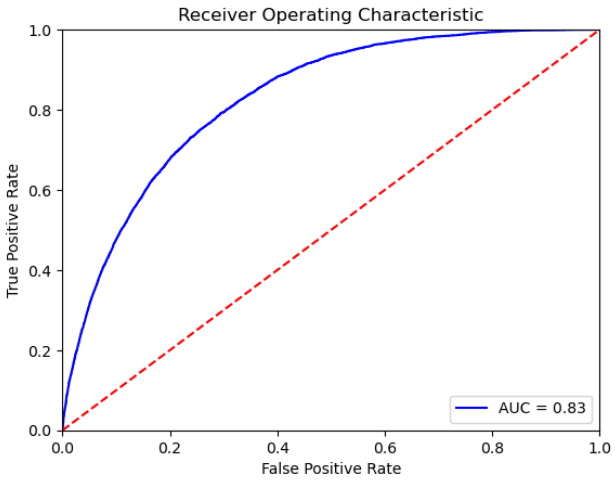


BMI is still the most important feature. The random forest performs much better than a single decision tree.

5. I used an Adaboost model with 2000 “stumps”. The rest is the same as Q1.

The reasons are the same as Q1.

The original AUROC using all predictors is 0.8278938291322622. The lowest is 0.81151376, corresponding to BMI.



BMI is the most important feature in the Adaboost model.

Extra Credits

a) Adaboost performs the best among these models.