Tyler Gelfand

Diabetes Report

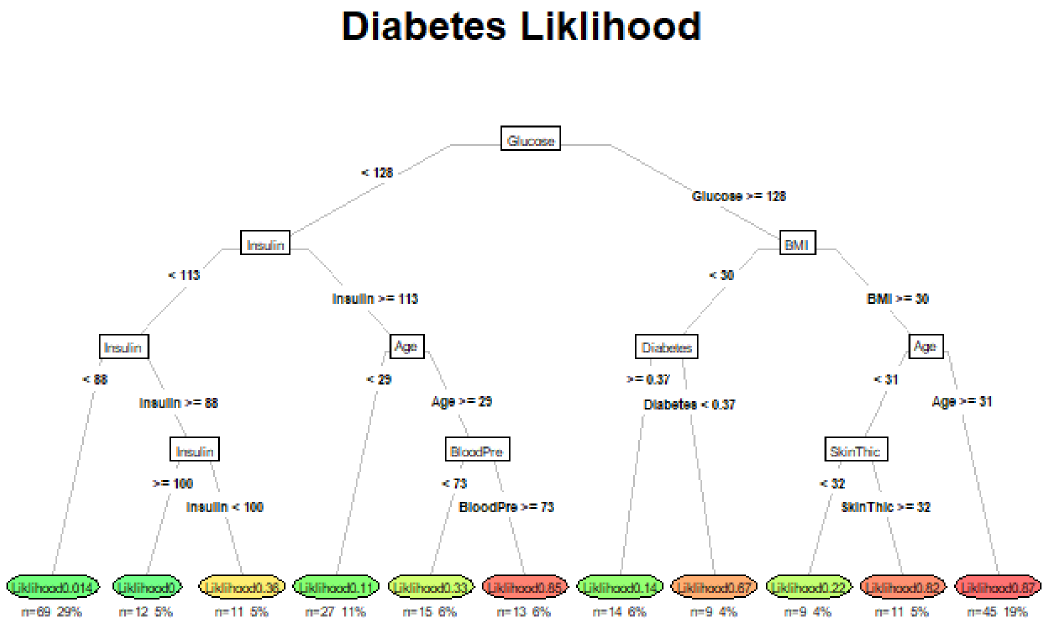
Modeling Methods

To correctly predict whether a patient is at high risk of developing gestational diabetes, 4 separate risk scoring models were run. A normal classification tree model utilizing a partitioned data set which, splits the data into a test and training set with a 70/30 split, a random forest model, an adaBoost model and an xgBoost model. These models predicted our outcome with varying success rates.

Model Performance

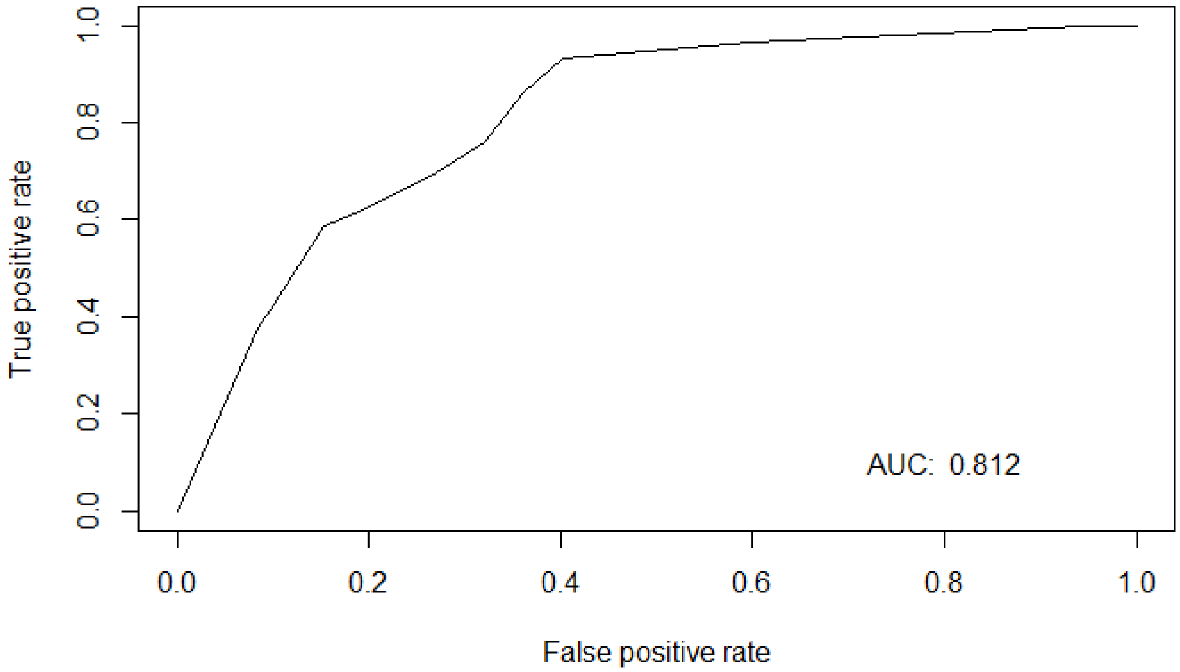
The first model we will look at is the classification tree which obtains a few insights, some do not come of surprise. The first is that those patients with high glucose levels, high BMI and those who are older are at higher risk. Additionally, those with high glucose and lower insulin levels are at the highest risk of developing gestational diabetes, which makes sense given that insulin is what regulates glucose levels.

Classification Tree



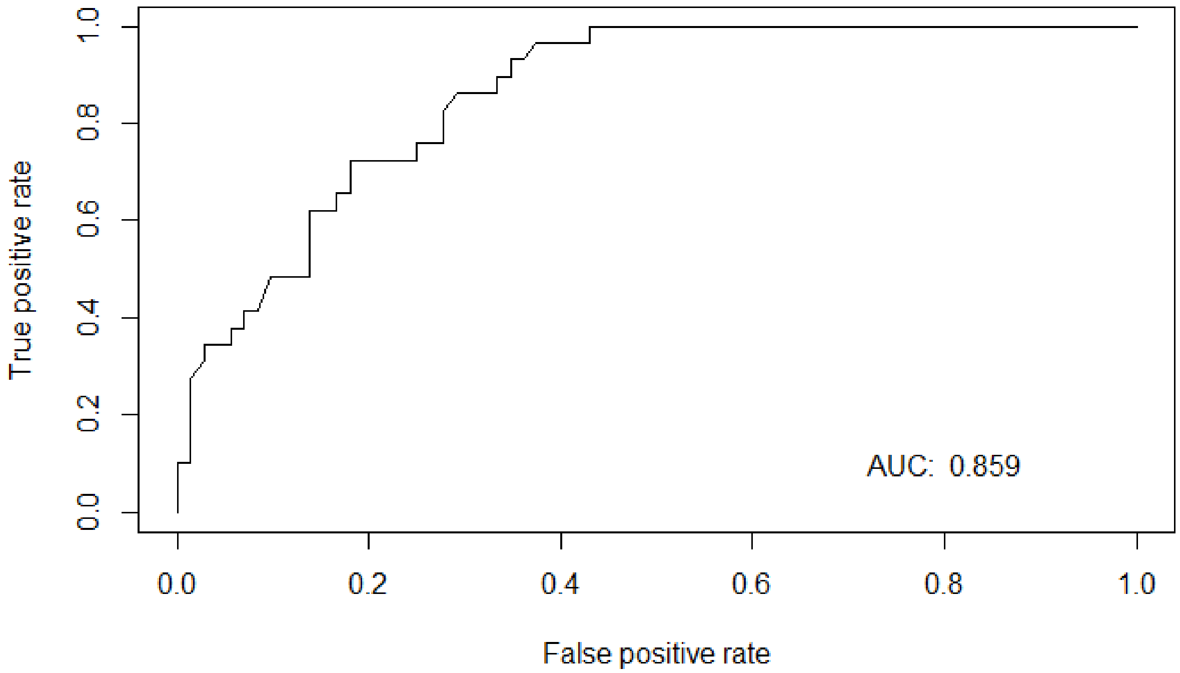
The ROC curve that was created from the classification tree has an AUC of .812 meaning that if we select 2 patients, one with gestational diabetes and one without, this model will be able to tell them apart

Classification Tree ROC Curve



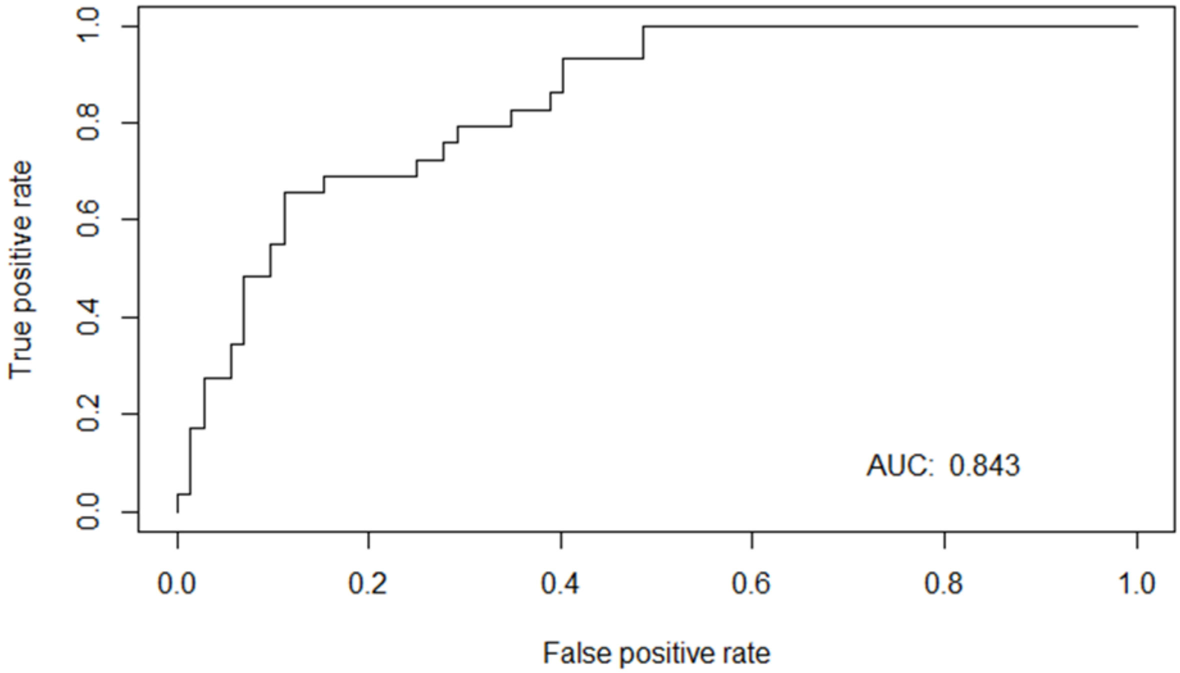
The next model we will look at is the random forest model. For this another ROC curve was generated, showing a better performance than the classification tree model. In this modeling method, AUC value is .859, making this a slightly more accurate method for determine which patients may have a higher likelihood of developing the disease.

Random Forest ROC Curve



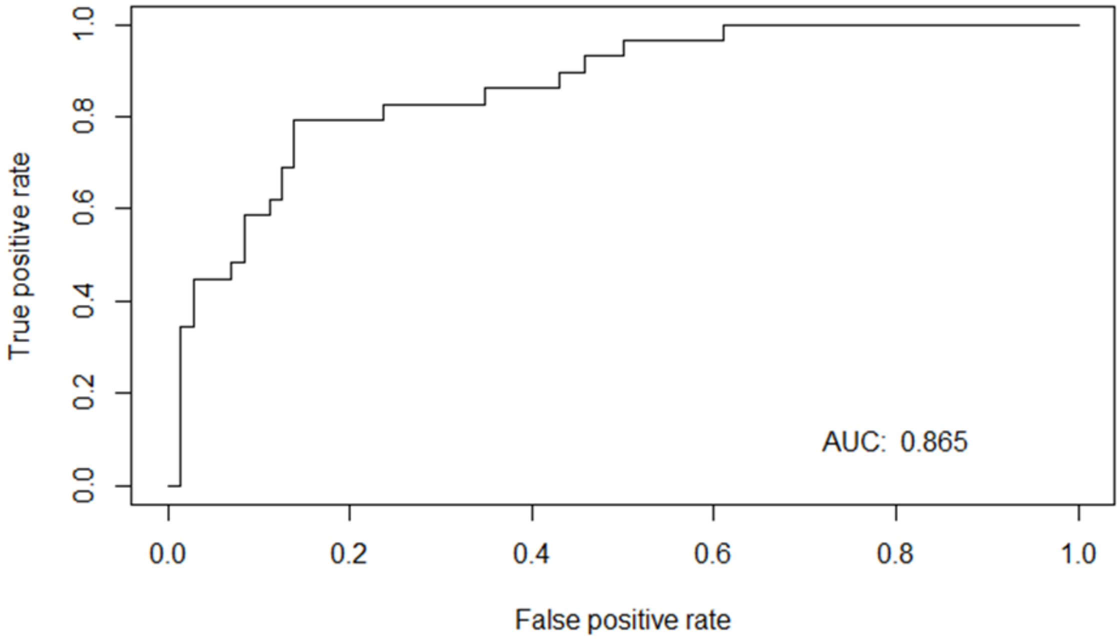
The next model is the adaBoost, which utilized the weaker classifiers of the dataset. This modeling method turned out to be less optimal than the previous method of Random Forests but still more optimal than the original classification tree. Using this method produced an ROC curve with an AUC of .843.

adaBoost ROC Curve

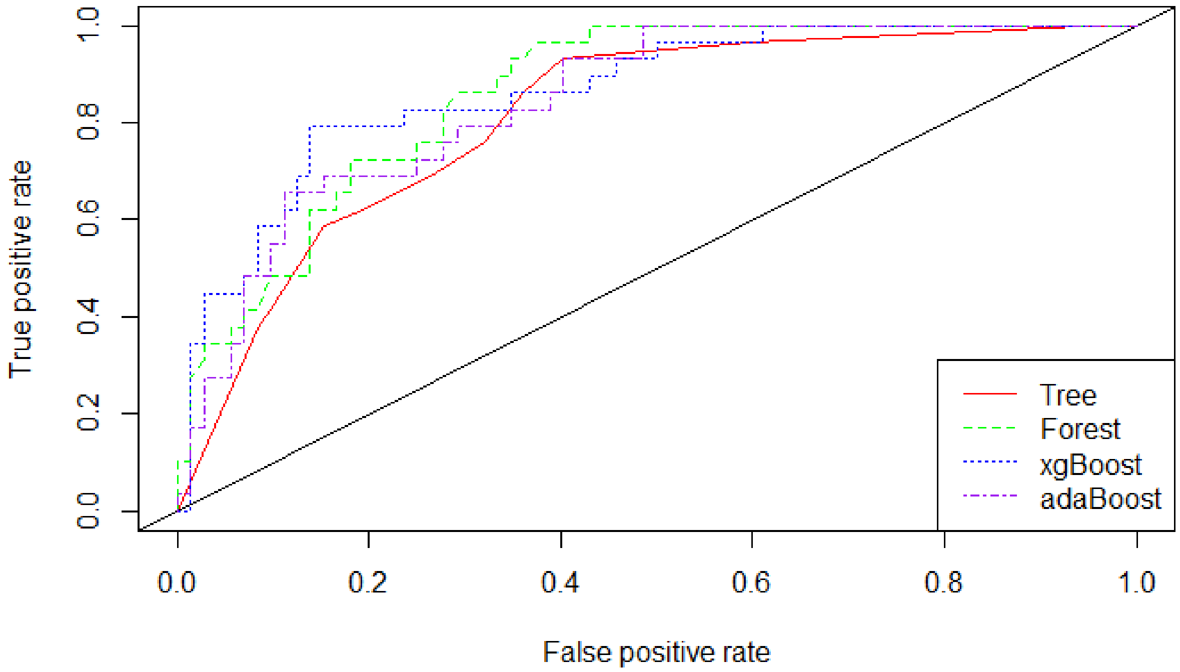


The final model that was ran was the xgboost model. This turned out to be the most effective model used. The ROC curve generated shows the AUC of .865.

xgBoost ROC Curve



The below ROC curve is an all in one view of each risk model for an easier comparison of them all. As we can see, there really is little variation in the ROC curve and how accurate these models may be. However, we need to look at the context of what we are talking about and what we are predicting. Because this is a health matter, even a 1% increase in accuracy may save countless lives. For that reason the xgBoost model is what should be used by providers, given the proper resources.



Modeling in Different Settings

When evaluating patients in a hospital setting, the recommendation is that the xgBoost model would be best suited. The hospital should have access to the best resources and computational power needed to run an xgBoost model on an extremely large dataset. The patient data can be constantly updated by physicians which will help with accuracy. As stated previously in this report, the xgBoost provided the best modeling results, which in what could be a life-or-death situation, is crucial.

On the other hand, health professionals may need to perform in-home visits for patients. There may be no EMR access and the professional will have limited computational access. In this specific instance a classification tree may be best suited. The information on the tree is more high level and easier to interpret than other model of increased computational power. It can be used to glance at quickly and obtain the desired information in a timely manner.