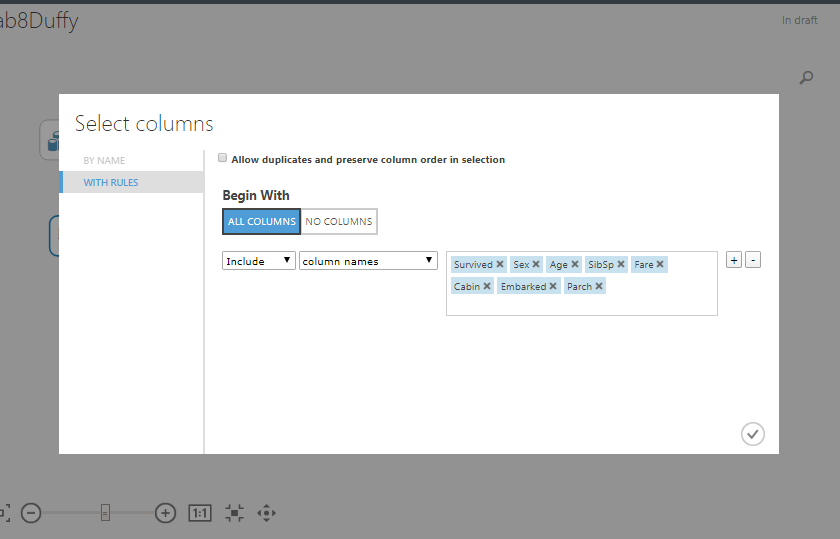
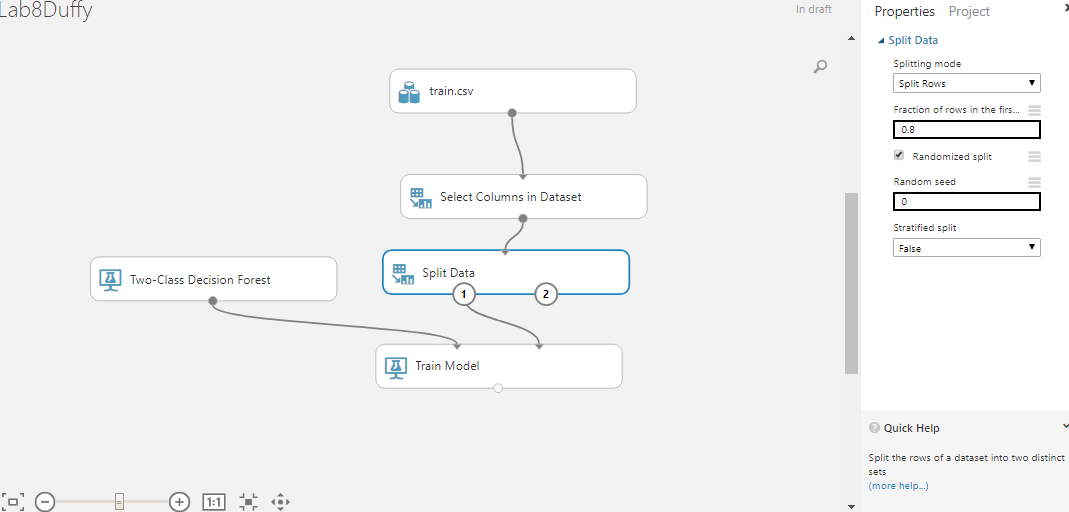
Toptim Kelly Duffy

GEOG 4990

Lab 8 Machine Learning

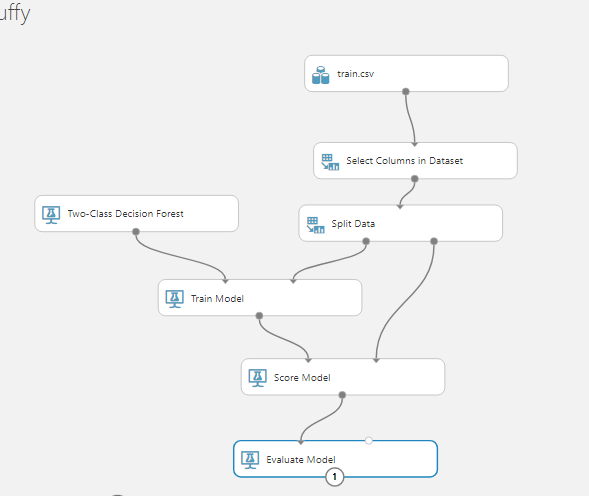


The first step was adding the train.csv data. This data gave the names, age, sex, etc. of the passengers. The columns I selected for model training were, ‘Survived, Sex, Age, SibSp, Fare, Cabin, Embarked and Parch.’ I think it would be interesting to see if there is a correlation between survival, age, and fare. It is likely those that paid more received better seating and attention since they could be considered celebrities.

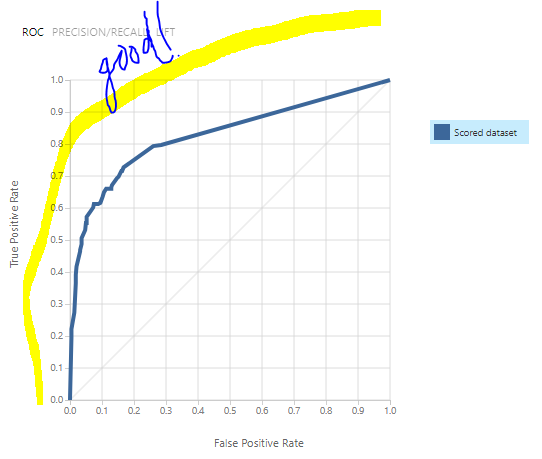


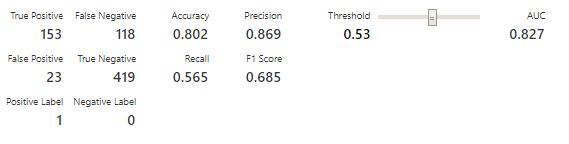
Next, under data transformation, select split data.The data was split 0.2. We split the rows of the data set into two distinct sets. In this case, the two distinct sets are those in the next step, ‘train model.’ ‘Train Model’ is set to the ‘Survived’ column.

Then, add the ‘two-class decision forest’ module. This classification is best suited for this scenario because decision forest creates decision trees on the data samples then gets the prediction from each of them and then selects the solution or the best predictor linked to survival.

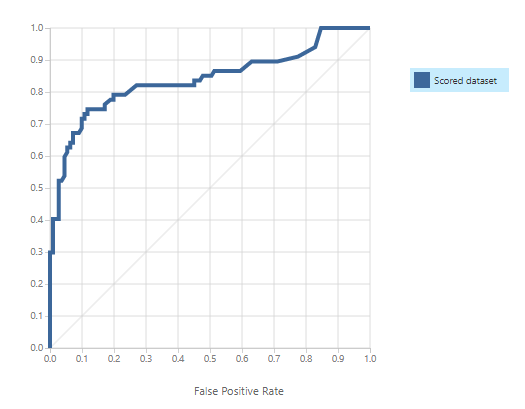


The last step is to add ‘Score Model’ and ‘Evaluate Model.’ Note that ‘score model’ is connected to both ‘split data’ and ‘train model.’ If it is not connected to both of these, you will receive an error message stating that the input port data is not connected. Score will train the model and evaluate model will display the results and visualize them.



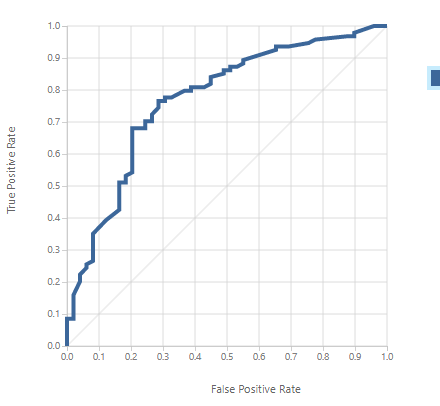


The result. The AUC is closer to 1, so the model is better at making predictions. In the result it showed that there were higher true positives and true negatives. Ultimately the model did a good job at predicting survival.

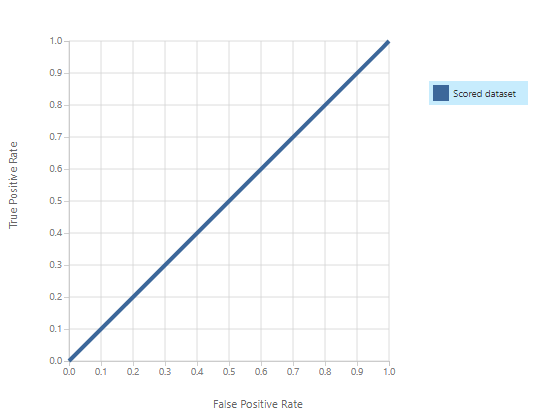


I did another run, except I split the data at 0.8 and I only selected the columns survived, sex, age, and fare. The AUC still reaches towards 1, but is a bit wonky. True Negative and True Positive are still the largest amount in this model as well.

**Q1**: What if you replace with other algorithms – e.g., Two-class boosted decision trees? Which algorithm can produce better results?



This graph is with the ‘two-class boosted decision’ classification. There is not a dramatic change in the results. This could substitute decision forest and still act as a good predictor.



This graph is with ‘Two-class Bayes Point Machine’ classification.

It is really bad. It had a very high ‘false positive’ value meaning the model predicted a passenger would survive, but actually did not survive. This algorithm is poor for making predictions in this example.

**Q2**: Despite changing the number of decision trees, the results do not dramatically change.

