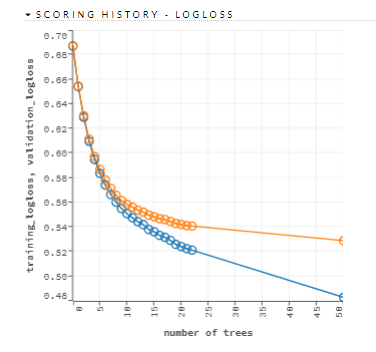
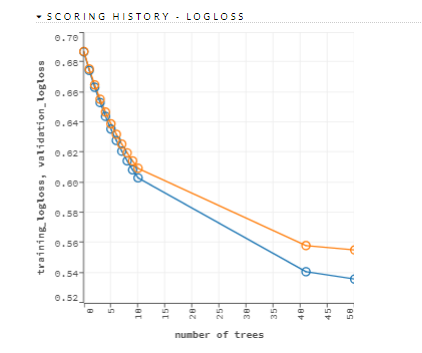
Use case 1 :

Airlines Delay Dataset :

The dataset consists of 20 years of data , i.e., 1987-2007 taken from U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) that tracks the on-time performance of domestic flights operated by large air carriers.

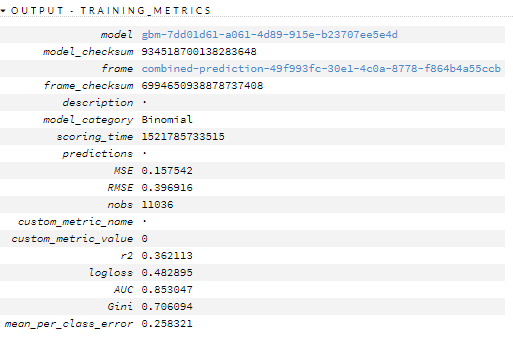
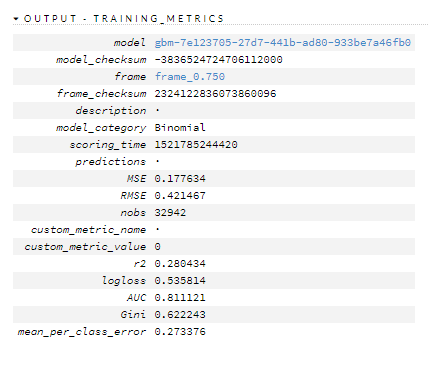
GBM model of H2O.ai is used to predict the delay in the arrival of the airlines and their causes.

Definition of GBM : Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Scoring History : Represents the error rate of the model as it is built.  


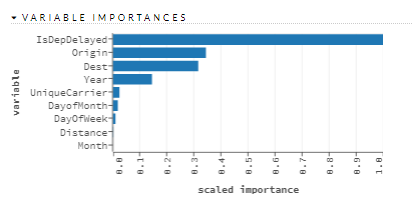
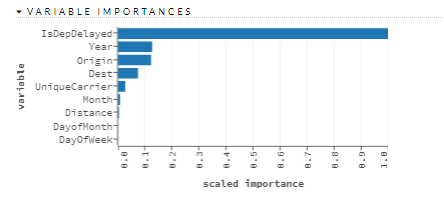
The figure on the left represents the scoring history of the data is when the GBM model was run for 1st time and the one on the right, represents the scoring history of the data after running the model for the 4th time.

The blue line indicates the training logloss and the orange line indicates the validation logloss. We can clearly observe that the model improves when more number of trees are formed in each of the figure given above. There is also a lot of difference in the 1st and 2nd figure as the model in the right figure has fitted better as compared to the one on the left.

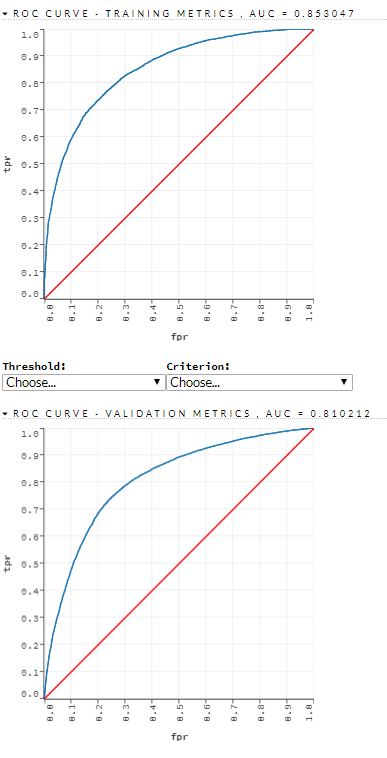
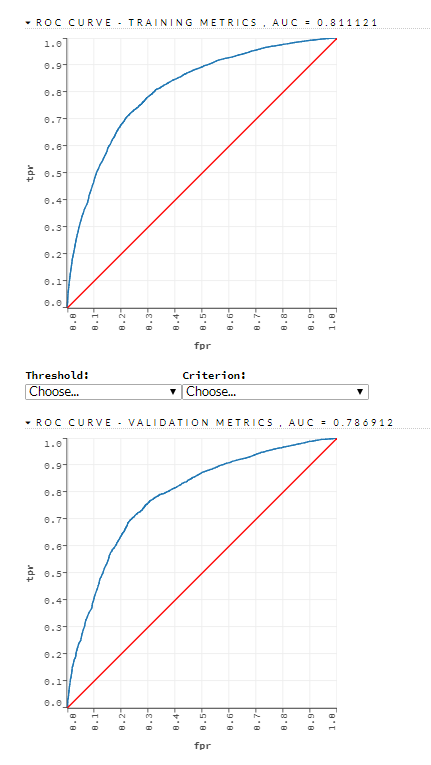


Here, in the Training Metrics output we can clearly see the r2, AUC and Gini index value has increased and the logloss(error rate) has decreased by 7%. This is definitely not the best-of-its-kind model as the r2 value is only 36.2%. But the AUC value is really good,I.e, 85.35%, which shows the model is highly predictable and as the Gini index is also high,I.e., 70.6%, it shows the airlines delay timings and causes are almost equally distributed.

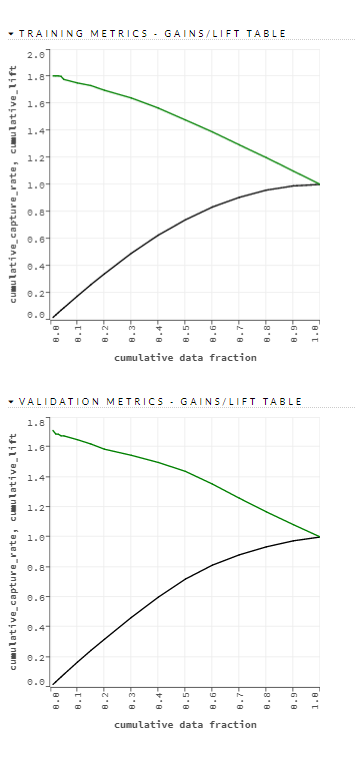
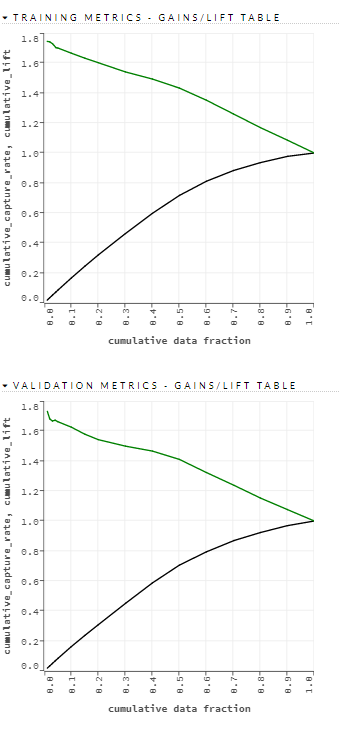
I have increased the learn rate, number of trees and nbins parameters of the model, each time while running the model, to get a better fit.



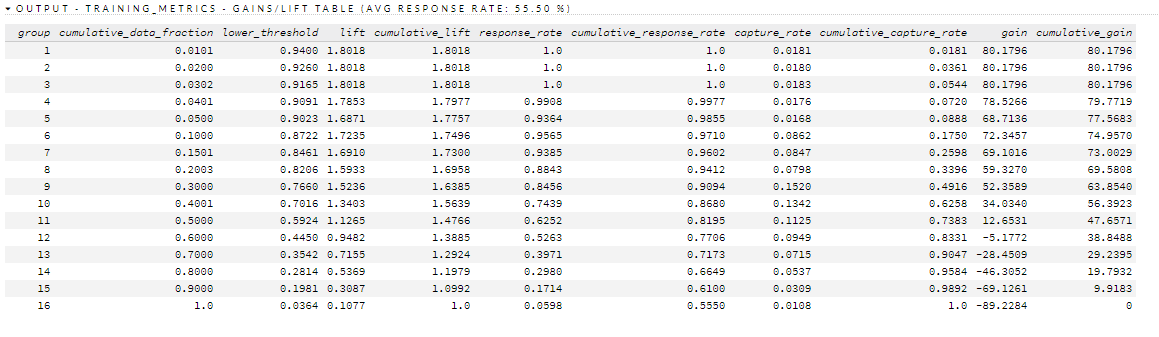
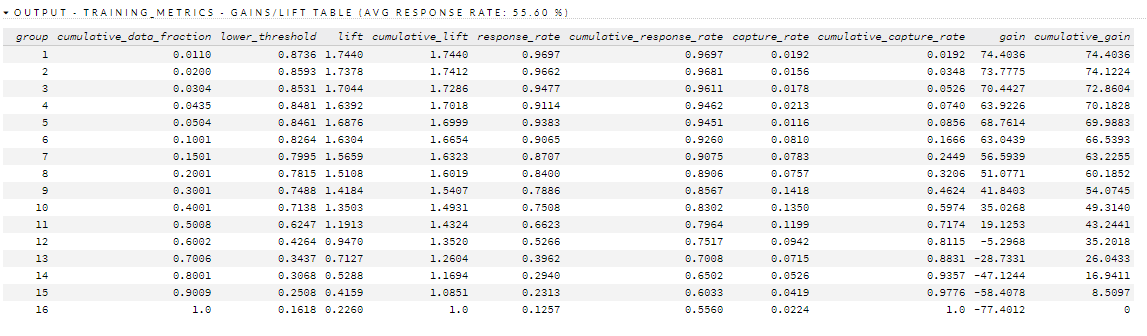
The variable importances can be used for implementing feature engineering as its showing us which are the predictor columns that are more dependent on the response column.



The ROC curve gives us the area under the curve(AUC). More the AUC, better is the model. It gives us the confusion matrix of actual vs predicted results, containing probability of true positives and number of false positives at each point .



The above figures represent the gain table, whose data is as follows :



In the above gain tables, H2O divides the dataset into 16 groups. More is the lift and variance between the gain more likely is the group to get selected. For example, here, Groups 1 to 10 have a good lift rate, which indicates airlines under these groups are more likely to be delayed.

Gain table are more prominent for direct marketing where it shows which customers are more likely to stay and be valued and which customers can be ignored.