CS 584-04: Machine Learning

Fall 2018 Final Examination

1. (20 points) Describe your strategies for developing the new models.
   I have chosen to use logistic regression for the region 1 and decision tree for region 2 and 3. Since at a company, explaining why the model is predicting a certain way is important for a business to decide whether to use that model or not, I thought it would be apt to use decision tree as it's very transparent. Also, visualizing and interpreting the tree is easy and this makes it easy to explain to executives or others. Decision trees make the feature importance clear and relations can be viewed easily.
   Initially, I planned to use decision tree for all 3 regions but it so happened that the it did not work well in the 1st region as the class distribution in the last node at a depth of 6 was still very uniform(it contained instances of both classes almost equally rather than a majority of one) and so opted to use logistic regression for region 1 as it is simple to understand too.

2. (20 points) Show how you selected the predictors into each new model. Your objective is to exclude variables that do not contribute to the goodness-of-fit of the new model.

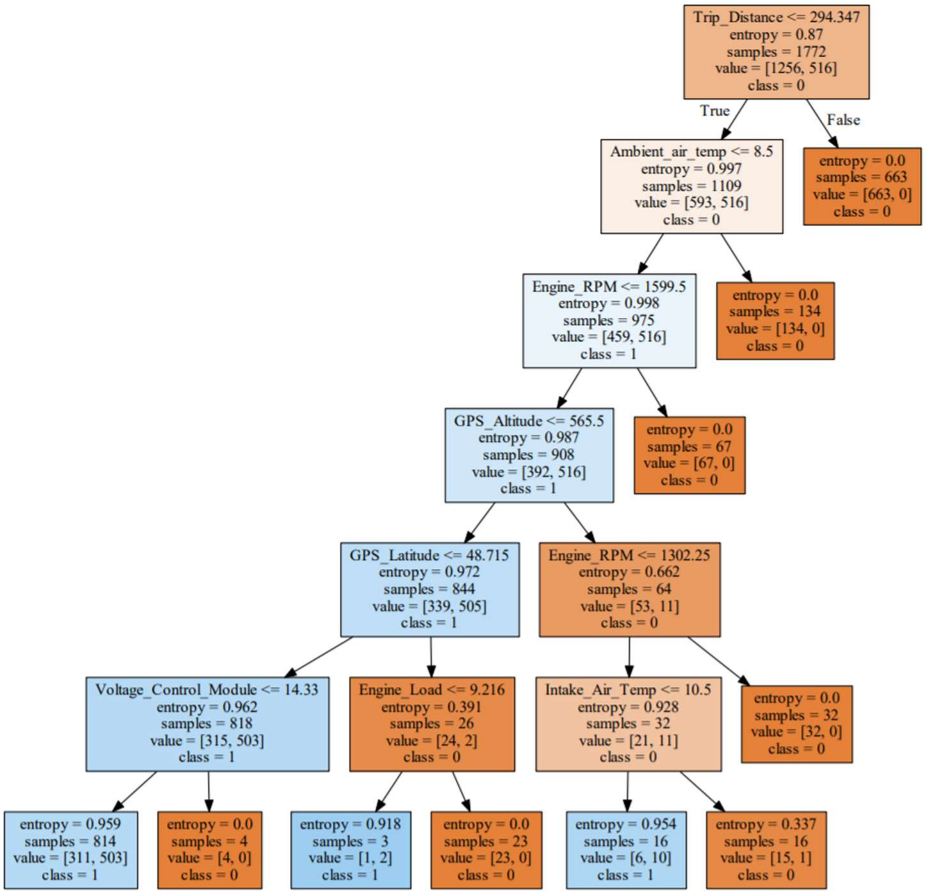
   For logistic regression the features were selected and ranked using recursive feature elimination. First, the estimator is trained on the initial set of features and the importance of each feature is obtained. Then, the least important features are pruned from current set of features and this procedure is recursively repeated until the number of features specified to select is reached. I chose to use 9 features and not too many. The output of RFE were the following features:
   Vibration, Engine_Coolant_Temp, Speed_OBD, Voltage_Control_Module, Ambient_air_temp, Engine_Oil_Temp, GPS_Longitude, GPS_Latitude, Accel_Ssor_Total

   Since decision trees themselves rank the features by the information gain and then use that feature to split upon, setting a reasonable depth ensures the best features which produce low entropy nodes after the split are used.
   Features used in region 2: Trip_Distance, Ambient_air_temp, Engine_RPM, GPS_Altitude, GPS_Latitude, Voltage_Control_Module, Engine_Load, Intake_Air_Temp

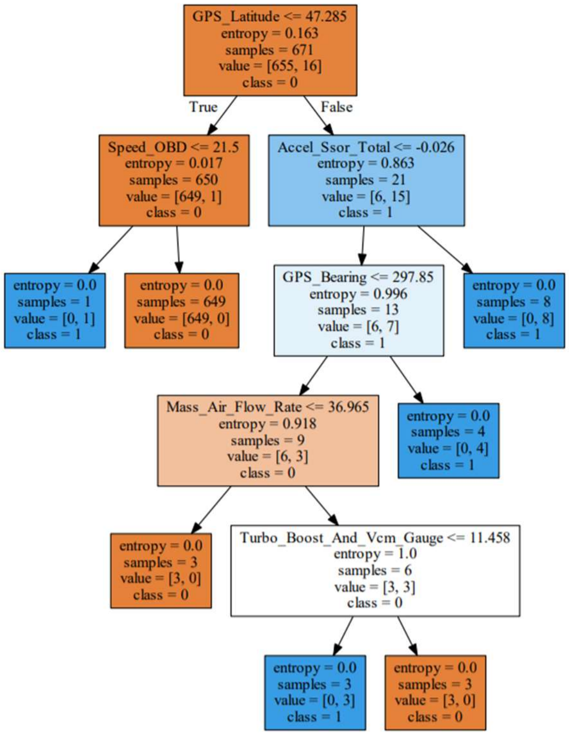
   In region 3, even though I set the max depth of the tree to be 6, the leaves are pure at depth 5 and hence, it doesn't proceed further.
   Features used in region 3: GPS_Latitude, Speed_OBD, Accel_Ssor_Total, GPS_Bearing, Mass_Air_Flow_Rate, Turbo_Boost_And_Vcm_Guage

3. (10 points) List the primary model specifications and the key model results (e.g., decision tree diagram, logistic regression, parameter estimates, support vector machine hyperplane equation, etc.)
   For regions 2 and 3, decision trees with a max_depth of 6 and the criterion of "entropy" is used

Region 2: Accuracy Score 2= 0.8741258741258742

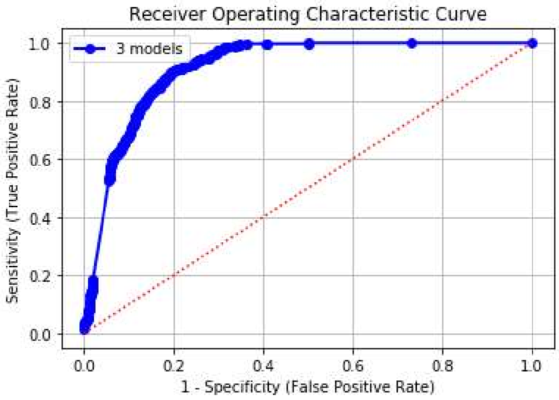


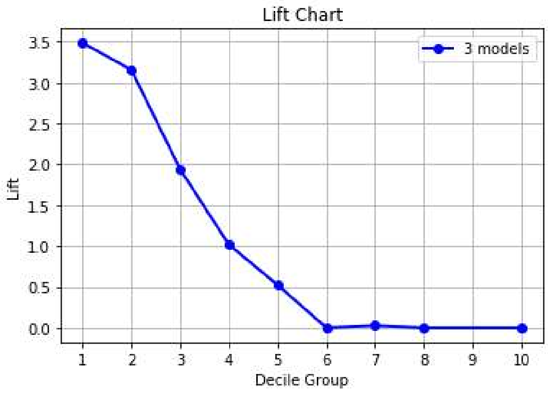
Region 3: Accuracy Score 3= 1.0

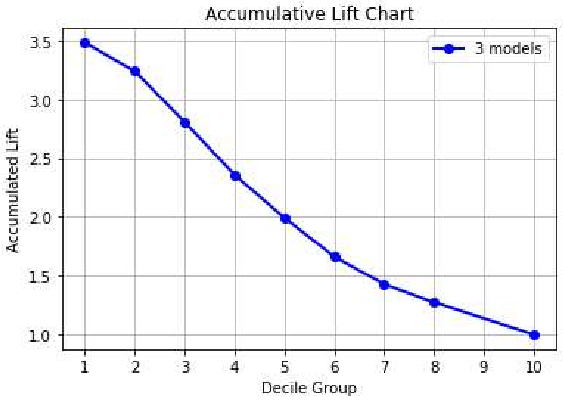


Region 1 : Logistic Regression; Accuracy Score = 0.8443067389620449
The coefficient values are :
Vibration -0.23324671
Engine_Coolant_Temp 0.26943419
Speed_OBD -0.04039195
Voltage_Control_Module 2.43774278
Ambient_air_temp -0.660558
Engine_Oil_Temp 0.0393996
GPS_Longitude -1.52197286
GPS_Latitude 0.31678666
Accel_Ssor_Total -0.26052172

4. (30 points) Show the model comparison results and list all supporting tables and charts (e.g., Area Under Curve, Root Average Squared Error, Misclassification Rate, ROC curve, Lift or Accumulated Lift curves)







Area Under Curve = 0.915815
Root Average Squared Error = 0.307828
Misclassification Rate = 0.179514

5. (20 points) Argue that you have actually found a better model than the current model.
   In the current model, the trucks are sent for maintenance only based on their region and the number of days elapsed, both of which are not strong indicators if the truck needs maintenance. However, the models I have used considers various parameters of the truck, learns this data to uses it to predict when maintenance is required. The intuition behind this is that some of the parameters indicate wear of different features of the truck and hence my model is learning when the truck is worn enough to warrant maintenance. Also my model significantly increases the area under the curve metric and lowers the RASE, while reducing the Misclassification rate on the test data by a wide margin.