DIEGO MARTIN CRESPO A20432558

Question 1 (100 points)

You are a data scientist in a trucking company in Europe. The truck company operates their trucks in three regions. The regulations require the company to maintain the trucks in good operating conditions. At the same time, the company wants to minimize the downtime due to taking the trucks offline for maintenance. Therefore, you are asked to develop business rules that help you determine if a truck is due for maintenance.

You have access to historical data about the trucks in your company. The data contain 9,398 observations which you have divided into the training and the testing partitions. The training partition is made available in the **fleet_train.csv** file and it has 7,504 observations. The testing partition is made available in the **fleet_monitor_notscored_2.csv** file and it has 1,894 observations.

Both partitions contain the following variables for modeling.

Target Variable

1. Maintenance_flag: 1 = Offline for maintenance, 0 = Otherwise. The event value is 1.

Nominal Predictor

1. Region: 1, 2, and 3

Interval Predictors

25

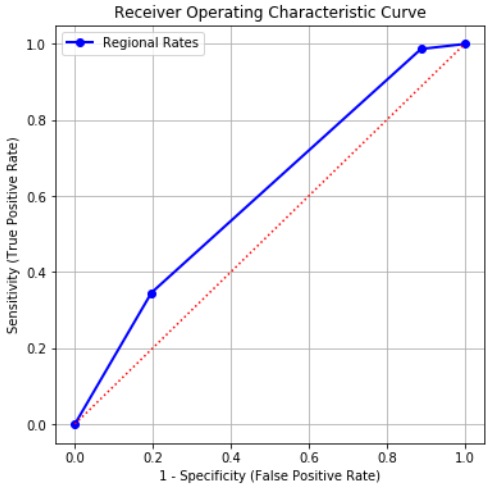
Currently, a truck is taken offline for maintenance according to the region it belongs to. A truck in Region 1 is maintained once every five days. A truck in Region 2 is maintained once every three days. A truck in Region 3 is maintained once every fifty days. This maintenance schedule is equivalent to this business rule.

Region	Probability that Maintenance_flag = 1
1	1/5
2	1/3
3	1/50

When applying this model to the testing partition, we obtained the following model metrics.

1. Area Under Curve = 0.60587637
2. Root Average Squared Error = 0.39477599
3. Misclassification Rate = 0.28880676 (the threshold value is 0.20469083)

The Receiver Operating Characteristic curve for the testing partition is



You will prepare a report for your new models. Your report must contain the following five sections.

1. (20 points) Describe your strategies for developing the new models.

First, the train and test datasets are cleaned. That is to say all the columns that are not attributes are removed. Both train and test datasets are grouped into the different regions for future treatment (1, 2 and 3). The models to be studied for each of the 3 regions are: Gradient Boosting Classifier, Logistic, classification tree and SVC.

We start by training and testing each of the 3 selected model for each of the 3 regions. We suppose that it may not converge for some cases, so a study about the predictors should be done. Therefore, first we try to train each model with all the attributes and then we will optimize them by removing the attributes that do not contribute well to the model.

To choose the proper predictors we will use the p_value between the different attributes divided by regions and groped by target. Also, the plotbox of each attribute and the target will be studied with the distribution of the target values with each attribute. We will remove the attributes with higher p_value of worst distribution. Except for the tree classifier than we will train it with all and then see the attributes with lower entropy and just use those.

We will calculate the AUC, RMSE and Misclassification Rate for all the three regions using the different models and by selecting the adequate predictors for each model and region we will try to optimize these values as much as possible.

Finally, we will choose the best model for each region and plot the ROC and Lift curves for the testing partition and make the conclusions.

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 the logistic model we take the attributes with lower p_value

By studying the p_vlaues for each region and the distribution of each attribute with the target using the plotbox. These are the lowest p_values for each region. It has to be considered that firstly it was done starting by leaving the latest 10 attributes with the lowest p_value. Then one by one it was removed. Also, for some cases it was considered the plotbox because even though the latest p_values where considered the model did not converge, so by testing finally we got the following attributes for each model and region that leave us to get a better rate of RMSE, ACU and Misclassification Rate. The VSC model did not converge for any of the attributes even by leaving the two with lowest p_value. Therefore, it was removed and it was not considered for the problem statement.

These are the selected attributes for each of the models and regions:

LOGISTIC	REGION 1	REGION 2	REGION 3
Attributes	'Vibration', 'Engine_RPM', 'Speed_OBD', 'Ambient_air_temp', 'Speed_GPS'	'Engine_Oil_Temp', 'Trip_Distance', 'Trip_Time_journey', 'GPS_Altitude', 'Engine_Coolant_Temp'	'Intake_Air_Temp', 'Speed_GPS'

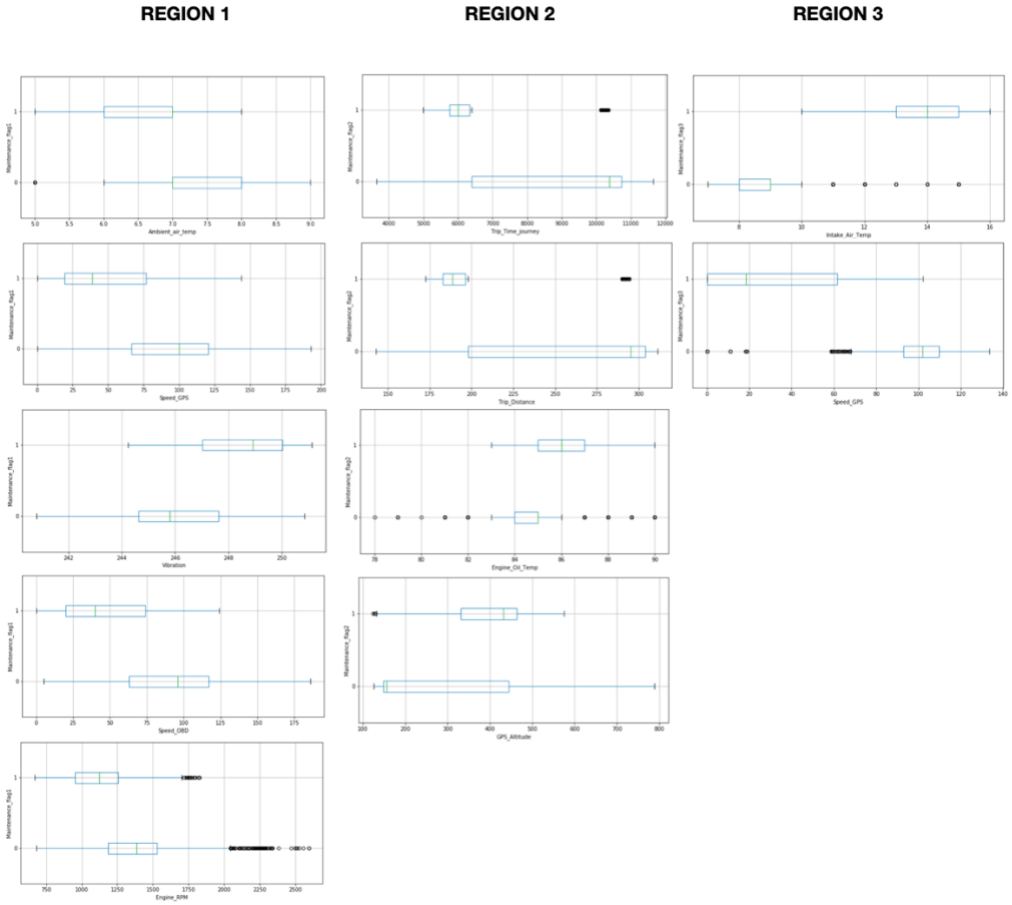
GBC	REGION 1	REGION 2	REGION 3
Attributes	'Vibration', 'Engine_RPM', 'Speed_OBD', 'Ambient_air_temp', 'Speed_GPS', 'Vehicle_speed_sensor', 'Throttle_Pos_Manifold', 'Mass_Air_Flow_Rate'	'Engine_Oil_Temp', 'Trip_Distance', 'Trip_Time_journey', 'GPS_Altitude', 'Engine_Coolant_Temp'	'Intake_Air_Temp', 'Speed_GPS'

TREE CLASSIFIER	REGION 1	REGION 2	REGION 3
Attributes	'Vibration', 'Engine_RPM', 'Speed_OBD', 'Ambient_air_temp', 'Speed_GPS', 'Vehicle_speed_sensor', 'Throttle_Pos_Manifold', 'Mass_Air_Flow_Rate'	'Engine_Oil_Temp', 'Trip_Distance', 'Trip_Time_journey', 'GPS_Altitude', 'Engine_Coolant_Temp'	'Intake_Air_Temp', 'Speed_GPS'

Considered lowest p_values ordered:

Region 1	P_values	Region 2	P_values	Region 3	P_values
Mass_Air_Flow_Rate	2.0811160e-78	Engine_Coolant_Temp	3.7892588e-32	CO2_in_g_per_km_Inst	1.5307844e-36
Throttle_Pos_Manifold	3.5618871e-85	Engine_Oil_Temp	6.7815377e-38	Litres_Per_100km_Inst	6.9795585e-36
Ambient_air_temp	2.8843171e-109	GPS_Altitude	4.1916136e-40	Speed_GPS	5.9703999e-38
Engine_RPM	6.4383176e-127	Trip_Time_journey	1.3687703e-77	Intake_Air_Temp	2.5282490e-44
Vibration	1.1528007e-180	Trip_Distance	3.5506361e-78		
Vehicle_speed_sensor	1.1528007e-180				
Speed_OBD	1.1528007e-180				
Speed_GPS	1.6965744e-184				

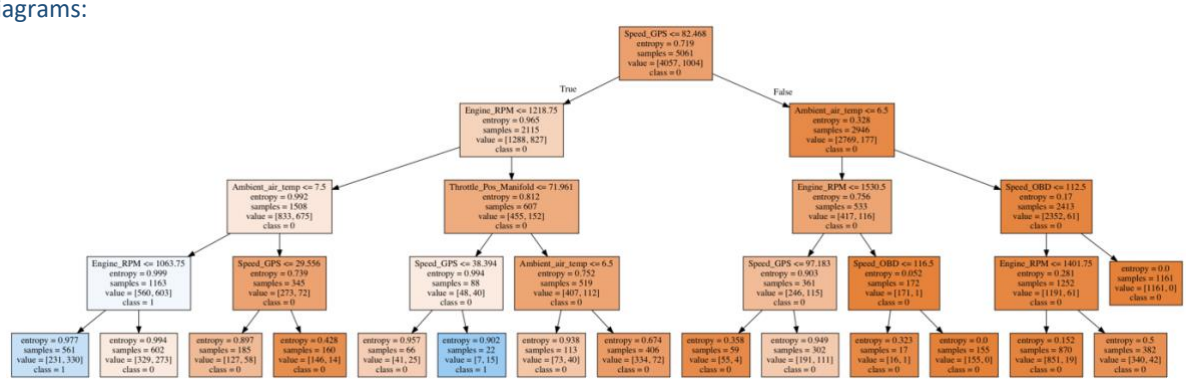
And some of the best distributed attributes using the plotbox util:



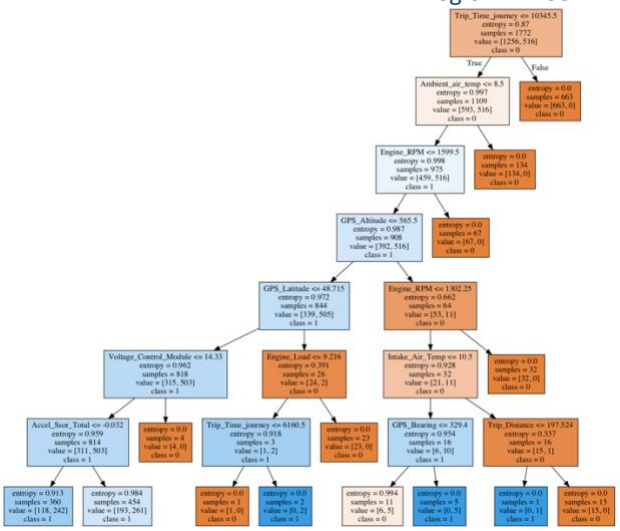
3. (10 points) List the primary model specifications and the key model results (e.g., decision tree diagram, logistic regression, parameter estimates, support vector machine hyperplane equation, etc.).

Logistic	GBC	Tree Classifier
Method: newton Max iterations: 100 Tolerance: 1e-8	Criterion loss: deviance Criterion: MSE Nº estimators: 1500 Maximum leaf nodes: 5	Criteria: entropy Maximum depth: 4,4,2 (region1, region2, region3) Random state: 20181010

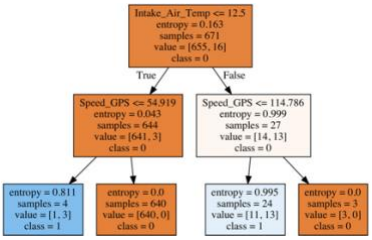
Tree Diagrams:



Region 1 Tree



Region 2 Tree



Region 3 Tree

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)

To improve: Misclassification Rate = 0.28880676, RMSE = 0.39477599, AUC = 0.60587637

Green: Improve the initial model Red: worse model

Region 1:

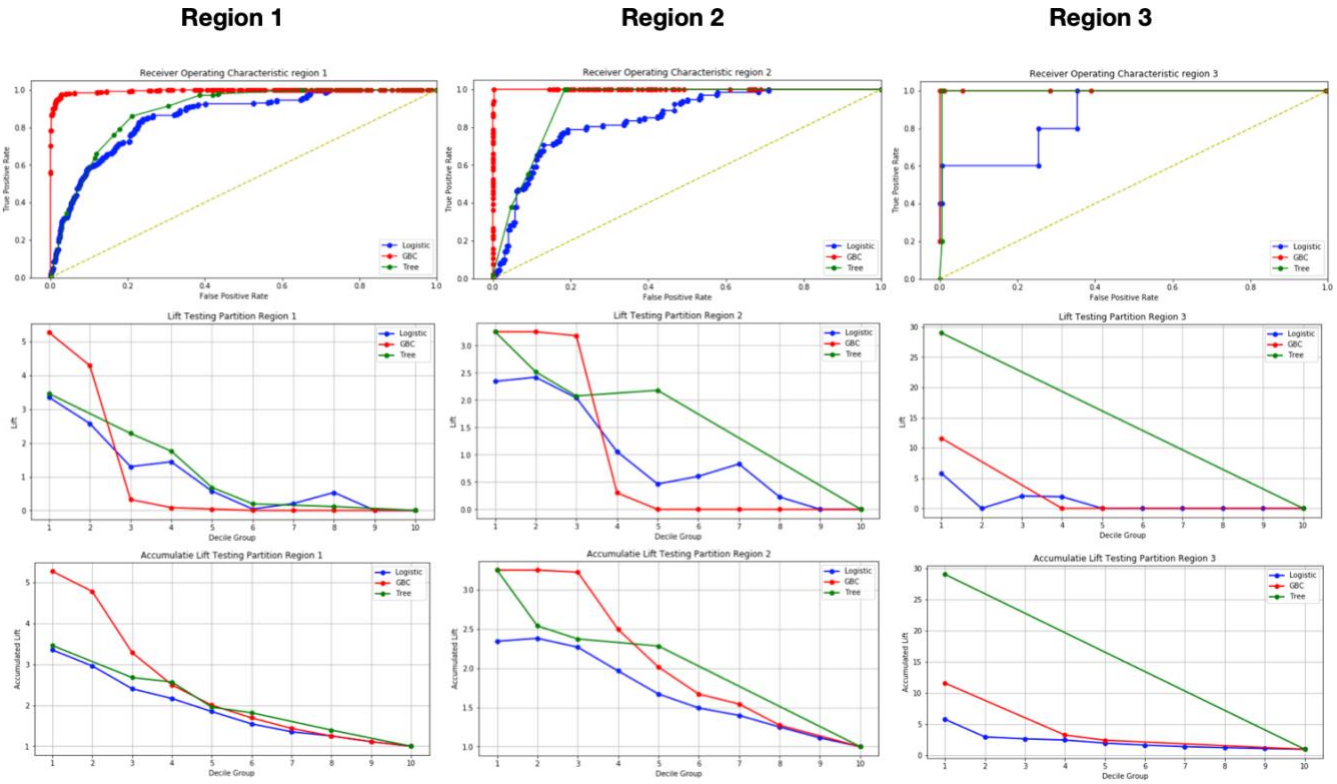
Logistic	GMC	Tree
Misclasification1= 0.21920991 RMSE1= 0.397513573 AUC1= 0.8545869590666094	Misclasification1= 0.08365608 RMSE1= 0.164653475 AUC1= 0.99468139	Misclasification1= 0.198295895 RMSE1= 0.1587916344 AUC1= 0.8854665002

Region 2:

Logistic	GMC	Tree
Misclasification2= 0.2004662 RMSE2= 0.45802861 AUC2= 0.84514335271	Misclasification2= 0.030303030 RMSE2= 0.0482804549 AUC2= 0.9997704315	Misclasification2= 0.125874125 RMSE2= 0.13053613 AUC2= 0.9117309458

Region 3:

Logistic	GMC	Tree
Misclasification3= 0.06896551 RMSE3= 0.15161960 AUC3= 0.876923076	Misclasification3= 0.0 RMSE3= 0.0 AUC3= 1.0	Misclasification3= 0.005747126 RMSE3= 0.00574712644 AUC3= 0.9946745562



5. (20 points) Argue that you have actually found a better model than the current model.

If we compare the results the ACU, RMSE and Misclassification Rate of each region for each model and compare with the initial one. The initial model can be improved using the models in green color of part 4 tables. That is to say that for region 1 and region 2 it can be used either the GMC or the Tree models because they improve the rates. And for region 3 we can use either of the three models.

In order to take the best approach and take the best model of the different possibilities a study of the different values and charts will be done. The following table shows the best advantages of each model in each region.

	Region 1	**Region 2**	**Region 3**
AUC and ROC chart	The GMC has the highest value of AUC. Therefore, the best ROC.	The GMC has the highest value of AUC.	The GMC has the highest value of AUC. However, the tree behaves very similar.
RMSE	The Tree has the lowest RMSE. However, the GMC is not bad as well has a close value to the Tree.	The GMC has the lowest RMSE.	The GMC has the lowest RMSE. However, the tree has also a very low value.
Misclassification Rate	The GMC has the lowest Misclassification rate.	The GMC has the lowest Misclassification rate.	The GMC has the lowest Misclassification rate. The tree has also really low and close value
Lift and Accumulated Lift charts	The GMC has the greatest values in the first deciles in both charts.	The GMC has the greatest values in the first deciles in both charts.	The Three has the greatest values in the first deciles in both charts.

In conclusion we choose the GMC model which produces the best model for every of the 3 regions in terms of AUC, RMSE, Misclassification Rate and Roc Chart. However, in terms of only improve the initial model and get faster results. We could use the model Tree Classifier for all 3 regions. Or other option could be, the three for the first two and the logistic for the region 3.