Aman Rangapur
A20517739

CS484: Intro to Machine Learning- Assignment 5

Question-1

The Homeowner_Claim_History.xlsx contains the claim history of 27,513 homeowner policies. The following table describes the eleven columns in the HOCLAIMDATA sheet.

Suppose we want to predict the number of claims using the above features. Instead of using the reported number of claims, we put the policies into four groups according to their number of claims. The first group comprises policies without claims (i.e., zero claims). The second group is policies with exactly one claim. The third group is policies with exactly two claims. Policies with three or more claims go to the fourth group. We will use the above grouping as our target variable which has four levels.

The categorical predictors are f_aoi_tier, f_fire_alarm_type, f_marital, f_mile_fire_station, f_age_tier, f_primary_gender, and f_residence_location.

After dropping the missing target values, we will divide the observations into the training and the testing partitions. Observations whose Policy Identifier starts with the letter A, G, and Z will go to the training partition. The remaining observations go to the testing partition. As a result, your training partition should have 9155 observations and your testing partition should have 3164 observations.

Since we have sufficient computing resources, we will train multinomial logistic models for all the possible subsets of combinations of the seven categorical predictors. We will include the Intercept term in all the models. To help us select our "optimal" model, we will calculate the AIC and the BIC criteria of the Training partition, the Accuracy of the Testing partition, and the Root Average Squared Error of the Testing partition.

(a) How many policies are in each of the four groups in the Training partition? Also, in the Testing partition?

	Train Partition	Test Partition
Group1	9352	3182
Group2	5171	1774
Group3	1557	529
Group4	394	136

(b) What is the lowest AIC value on the Training partition? Also, which model produces that AIC value?

Lowest AIC: 40514.30941979058
Model: ['f_primary_age_tier', 'f_residence_location', 'f_fire_alarm_type', 'f_mile_fire_station', 'f_aoi_tier']

(c) What is the lowest BIC value on the Training partition? Also, which model produces that BIC value?

Lowest BIC: 40880.07276393123

Model: [‘f_primary_age_tier', 'f_fire_alarm_type', 'f_mile_fire_station', ‘f_aoi_tier']

(d) What is the highest Accuracy value on the Testing partition? Also, which model produces that Accuracy value?

Accuracy: 0.5479157713794586
Model: [‘f_primary_age_tier', 'f_primary_gender', 'f_marital', 'f_fire_alarm_type', ‘f_mile_fire_station’]

(e) What is the lowest Root Average Squared Error value on the Testing partition? Also, which model produces that RASE value?

RASE: 0.957034
Model: [‘f_primary_age_tier', 'f_primary_gender', 'f_marital', 'f_fire_alarm_type', ‘f_mile_fire_station']

Question- 2

The Center for Machine Learning and Intelligent Systems at the University of California, Irvine manages the Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). We will use two of the datasets in the repository for analyses, namely, the WineQuality_Train.csv for training and the WineQuality_Test.csv for testing.

The categorical target variable is *quality_grp*. It has two categories, namely, 0 and 1. The input features are *alcohol*, *citric_acid*, *free_sulfur_dioxide*, *residual_sugar*, and *sulphates*. These five input features are considered interval variables.

We will apply the Adaptive Boosting technique for training a classification tree model. The model specifications are as follows.

- **The Splitting Criterion is Entropy.**
- **The maximum tree depth is five.**
- **The initial random state value is 20230101 for the classification tree and boosting.**
- **The maximum number of Boosting iterations is 100.**
- **Stop the iteration if the classification accuracy on the Training data is greater than or equal to 0.9999999.**
- **If the observed *quality_grp* is 1, then the absolute error is 1 – Prob(*quality_grp* = 1). Otherwise, the absolute error is Prob(*quality_grp* = 1).**
- **If an observation is correctly classified, then the weight is the absolute error. Otherwise, the weight is the absolute error plus 2.**
- **If Prob(*quality_grp* = 1) > 0.2, then the predicted *quality_grp* is 1. Otherwise, the predicted *quality_grp* is 0.**

(a) What is the Misclassification Rate of the classification tree on the Training data at Iteration 0 (i.e., when all the weights are one)?

Misclassification Rate: 0.247195.

(b) How many iterations are performed to achieve convergence? Show the iteration history in a table. The table should show the iteration number, the sum of weights, and the weighted accuracy at each iteration.

The model took 19 iterations to achieve convergence.

	Iterations	Sum of Weights	Weighted Accuracy
0	0	4547.0	0.7528040466241480
1	1	3298.1008263742900	0.8398342905040090
2	2	8602.232914649530	0.9755766326177170
3	3	3614.427040090970	0.9738398884068390
4	4	8975.477397381020	0.9934014157116960
5	5	3657.1996550126000	0.9871631849937750
6	6	8699.609749421740	0.9989365506261060
7	7	3823.272844236130	0.9922962645290410
8	8	9267.338476201140	0.9989697483749740
9	9	3709.49851588819	0.9995885640214830
10	10	8922.412080883550	0.9999379963581280
11	11	3831.0501974943700	0.9991376176867860
12	12	9505.01156145901	0.9998849333752110
13	13	3884.037597497890	0.999906352933896
14	14	9365.172683484960	0.9999908714219150
15	15	3876.058957357370	0.9999983667905820
16	16	9398.973974392970	0.9999997754987900
17	17	3960.000661252520	0.999999822381173
18	18	9431.997108003680	0.9999999751425140

(c) What is the Area Under Curve on the Testing data using the final converged classification tree?

Test AUC score: 0.8383386581469648

(d) What is the Accuracy of the Testing data using the final converged classification tree?

Testing Accuracy: 0.8492307692307692

(e) Generate a grouped boxplot for the predicted probability for *quality_grp* = 1 on the Testing data. The groups are the observed *quality_grp* categories.