CS 484: Introduction to Machine Learning

Spring Semester 2023 Assignment 5

Question 1 (50 points)

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.

Name	Description	Categories
policy	Policy Identifier	
exposure	Duration a Policy is Exposed to Risk Measured in Portion of a Year	
num_claims	Number of Claims in a Year	
amt_claims	Total Claim Amount in a Year	
f_primary_age_tier	Age Tier of Primary Insured	< 21, 21 - 27, 28 - 37, 38 - 60, > 60
f_primary_gender	Gender of Primary Insured	Female, Male
f_marital	Marital Status of Primary Insured	Not Married, Married, Un-Married
f_residence_location	Location of Residence Property	Urban, Suburban, Rural
f_fire_alarm_type	Fire Alarm Type	None, Standalone, Alarm Service
f_mile_fire_station	Distance to Nearest Fire Station	< 1 mile, 1 - 5 miles, 6 - 10 miles, > 10 miles
f_aoi_tier	Amount of Insurance Tier	< 100K, 100K - 350K, 351K - 600K, 601K - 1M, > 1M

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

(b) (10 points) What is the lowest AIC value on the Training partition? Also, which model produces that AIC value?

(c) (10 points) What is the lowest BIC value on the Training partition? Also, which model produces that BIC value?

(d) (10 points) What is the highest Accuracy value on the Testing partition? Also, which model produces that Accuracy value?

(e) (10 points) What is the lowest Root Average Squared Error value on the Testing partition? Also, which model produces that RASE value?

Question 2 (50 points)

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

(b) (10 points) 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.

(c) (10 points) What is the Area Under Curve on the Testing data using the final converged classification tree?

(d) (10 points) What is the Accuracy of the Testing data using the final converged classification tree?

(e) (10 points) Generate a grouped boxplot for the predicted probability for *quality_grp* = 1 on the Testing data. The groups are the observed *quality_grp* categories.