A picture containing text

Description automatically generated

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

CAT II

DSA 8502: PREDICTIVE AND OPTIMIZATION ANALYTICS

DATE: 23/01/2023 TIME: 3 HOURS

**Instructions**: Only answer 2 questions. For Question 2 and 3, you can use either R or Python, even though they are set within R environment. The CAT should be submitted not later than Friday 27th January 2023 00:00 Hours.

1. Based on the DSA 8502 coursework and its application in solving data problems:
   1. Answer the following questions about sampling:
      1. List and briefly describe 2 sampling methods that can be used to overcome poor performance from a classification algorithm due to an unbalanced data set. (2 points)
      2. For each of the sampling methods listed in your answer to part (a), briefly describe a drawback of each method (2 points)
   2. In the context of supervised and unsupervised learning:
      1. Briefly describe two differences between unsupervised and supervised learning (4 points)
      2. Give two situations in which unsupervised learning would be preferred to supervised learning (2 points)
      3. Give two situations in which supervised learning would be preferred to unsupervised learning (2 points)
   3. Consider one of the validation approaches for measuring test errors in model validation. Briefly describe one advantage and one disadvantage of this approach (4 points).
   4. A binary classifier has been fitted to a dataset and the following confusion matric has been calculated based on holdout (test) dataset.

|  | | **Predicted Class** | |
| --- | --- | --- | --- |
| Positive | Negative |
| **Actual Class** | Positive | 475 | 116 |
| Negative | 93 | 841 |

Based on this confusion matrix, calculate the following metrics:

1. Accuracy (1 point)
2. Precision (1 point)
3. Sensitivity (1 point)
4. Specificity (1 point)
5. The data set MotorcycleData2 has the following predictor variables:

Age, gender, zone (i.e., geographic territory), engine.class, car.age, bonus.class (experience rating class reflecting history of accidents), policy.duration, CreditScore (the policyholder’s insurance company assigned credit score). It has the following possible dependent variables: claims (number of claims for the policy), losses (total losses for the policy), and Claim Indicator (a binary variable indicating whether the policy has had at least one claim).

1. Produce descriptive statistics of losses (2 points)
2. What proportion of policy holders in the data have a claim? (1 point)
3. Using the random seed of 2, create a training and test sample, using 50% of the data for each. (4 points)
4. Sing the training sample, fit a tree to the dependent variable claims using the predictor variables, but excluding the other dependent variables. Set the minimum deviance parameter to 0.005 using mindev = 0.005 in your tree function. Print output from tree fit. (5 points)
5. Plot the tree and comment on the plot. (4 points)
6. Use set.seed(100007) and the cv.tree function to determine the best pruned tree, then produce the pruned tree. (4 points)
7. Your dataset contains 10,000 observations of a binary response variable and six predictor variables. The data is split into training and test sets with an 80/20 split.
8. Fit the training data with a generalized linear model with binomial family and logit link function.
   1. Use the model to make the predictions for the test set. (3 points)
   2. How many observations in the test set were predicted correctly? (3 points)
   3. What is the accuracy of the model when applied to the test set? (1 point)
9. Fit the data with k-nearest neighbor classification model, with number of nearest neighbors equal to 3. Before constructing the model, run set.seed(1).
   1. Use the model to make the predictions for the test set. For the kNN function, make sure class variable is in matrix, with code such as “train.class = as.matrix(train$Y) (3 points)
   2. How many observations in the test set were predicted correctly? (3 points)
   3. What is the accuracy of the model when applied to the test set? (1point)
10. Based on the results above, which of the two models would you recommend? Why? (6 points)