CS 484: Introduction to Machine Learning

Spring Semester 2023 Assignment 4

# Question 1 (30 Points)

We provided you with the following classification tree diagram. The label variable is *Payment Type*. It has two categories, namely, *Cash* and *Credit Card*.

A picture containing qr code

Description automatically generated

1. (10 points). What is the Area Under Curve value? Please show the steps for calculating your answer.
2. (10 points). What is the Root Average Squared Error? Please show the steps for calculating your answer.
3. (10 points). Based on the above two metrics, is this model acceptable?

# Question 2 (30 points)

We trained a classification model on a binary target variable on twenty observations. The target categories are Event in ten observations and Non-Event in another ten observations. We provided you with the following table that contains the target categories and the predicted event probabilities.

| Target | Predicted Event Probability |  | Target | Predicted Event Probability |
| --- | --- | --- | --- | --- |
| Non-Event | 0.0814 |  | Event | 0.4974 |
| Non-Event | 0.1197 |  | Event | 0.6732 |
| Non-Event | 0.1969 |  | Event | 0.6744 |
| Non-Event | 0.3505 |  | Event | 0.6836 |
| Non-Event | 0.3878 |  | Event | 0.7475 |
| Non-Event | 0.3940 |  | Event | 0.7828 |
| Non-Event | 0.4828 |  | Event | 0.6342 |
| Non-Event | 0.4889 |  | Event | 0.6527 |
| Non-Event | 0.5587 |  | Event | 0.6668 |
| Non-Event | 0.6175 |  | Event | 0.5614 |

1. (10 points). Generate and display the Kolmogorov–Smirnov curve, and then determine the probability threshold that yields the highest Kolmogorov–Smirnov statistic. Please provide the exact answer.
2. (10 points). Generate and display the Precision-Recall curve, and then determine the probability threshold that yields the highest F1 Score. Please provide the exact answer.
3. (10 points). If the predicted event probability is greater than or equal to the threshold, then we will classify that observation as an Event. What are the misclassification rates for parts (a) and (b)?

# Question 3 (40 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 |

In insurance ratemaking, the ratio of the Total Claim Amount divided by the Number of Claims is called the Severity. In other words, Severity is the average claim amount per year. If a policy does not file any claims in a year, then its Severity is a missing value.

We will use the natural logarithm of Severity as our target variable. The predictors are the seven categorical predictors. 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.

1. (20 points). Train a Multi-Layer Perceptron neural network. Perform a grid search to select the most desired network structure. The maximum number of iterations is 10000. The random seed is 2023484. We will try the Hyperbolic Tangent, the Identity, and the Linear Rectifier activation functions; the number of layers is from 1 to 10; and the common number of neurons per layer is from 1 to 5. Please be reminded that the answer key of this question is prepared using sklearn version 1.2.1.

Show your grid search results in a table. The table should contain (1) the activation function type, (2) the number of layers, (3) the common number of neurons per layer, (4) the number of iterations performed (*n\_iter\_* attribute), (5) the best loss value (*best\_loss\_* attribute), (6) the root mean squared error on the testing partition, (7) the relative error on the testing partition, (8) the Pearson correlation on the testing partition, and (9) the elapsed time in seconds.

1. (10 points). Among the networks that converged, which network structure yields the lowest root mean squared error on the testing partition? In the case of ties, choose the network with a fewer total number of neurons.
2. (10 points). Among all the possible category combinations of the predictors, which combination(s) yields the lowest Severity? Similarly, which combination(s) yields the highest Severity?