CS 584 Machine Learning

Pradyot Mayank(A20405826)
Spring 2019 Midterm Test

Question 12 (25 points)

In the automobile industry, a common question is how likely a policy-holder will file a claim during the coverage period. Your task is to build two models. After evaluating and comparing the models, you will recommend the model that performs better. In order to avoid discriminating policy-holders, we will use predictors that can be verified and are related to the risk exposures of the policy-holders. The CSV file policy_2001.csv contains data about 617 policy-holders. We will use only the following variables.

Target Variable

- CLAIM_FLAG: Claim Indicator (1 = Claim Filed, 0 = Otherwise) and 1 is the event value.

Nominal Predictor

- CREDIT_SCORE_BAND: Credit Score Tier ('450 – 619', '620 – 659', '660 – 749', and '750 +')

Interval Predictors

- BLUEBOOK_1000: Blue Book Value in Thousands of Dollars (min. = 1.5, max. = 39.54)
- CUST_LOYALTY: Number of Years with Company Before Policy Date (min. = 0, max. ≈ 21)
- MVR_PTS: Motor Vehicle Record Points (min. = 0, max. = 10)
- TIF: Time-in-Force (min. = 101, max. = 107)
- TRAVTIME: Number of Miles Distance Commute to Work (min. = 5, max. ≈ 93)

Since the tools may not take the nominal predictor as is, you will first derive the dummy indicators from the nominal predictors and then use the dummy indicators in building the models. You will build the three models according to the following specifications.

Classification Tree Model

- The maximum number of depths is 5
- The splitting criterion is Entropy
- The random seed is 20190402

Logistic Model

- The optimization algorithm is the Newton-Raphson method
- The maximum number of iterations is 100
- The relative error in parameter estimates acceptable for convergence is 1E-8
- The Intercept term must be included in the model

You will divide the data into the Training and the Testing partitions. You will build and evaluate the three models using the Training partition. Later, you will recommend one model based on the evaluation and the comparison results from the Testing partition.

Data Partition

- The Training partition consists of 75% of the original observations, the remaining 25% goes to the Testing partition.
- The claim rates (i.e., the fraction of observations whose CLAIM_FLAG is 1) must be the same in both partitions.
- The random seed is 20190402.

Please answer the following questions.

a) (5 points) How many observations are in the Training and the Testing partitions?
Number of Observations in Training partitions =462
Number of Observations in Testing partitions =155

b) (5 points) What is the claim rate in the Training partition?

Number of Claims Filed for training data=133
Number of training observations=462
Claim Rate for training data:
Number of Claims Filed for training data/ Number of training observations=133/462= 0.2879.
So, Claim Rate for training data=0.2879.

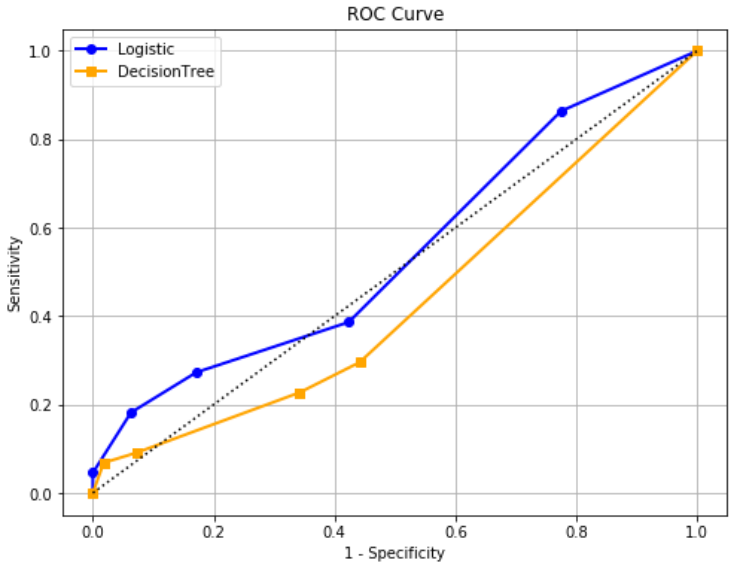
c) (5 points) Use **the claim rate in the Training partition** as the probability threshold in the misclassification rate calculation. A claim is predicted if the predicted probability of filing a claim is greater than or equal to the probability threshold. Calculate the Area Under Curve metric, the Root Average Squared Error metric, and the Misclassification Rate for both models using the Testing partition. Present your results in a table, list the metrics in the column dimension and the models in the row dimension.

Tabulated results of logistic model and decision tree with respect to Area Under Curve, Root Average Squared Error and Misclassification Rate.

	Logistic Model	Decision Tree
Area Under Curve	0.559378	0.509419
Root Average Squared Error	0.302202	0.270708
Misclassification Rate	0.440622	0.464516

d) (5 points) Calculate (but no need to display) the coordinates of the Receiver Operating Characteristic curve for both models using the Testing partition. Plot both ROC curves in the same chart but uses a different color for each curve. The chart (including the axes, the title, and the curve legends) must be properly labeled.

ROC curves in the same chart but uses a different color for each curve with properly labeled chart:



e) (5 points) Based on the evaluation and the comparison results in (c) and (d), recommend your model. You must state your reasons for your recommendation.

Based on the evaluation and the comparison results we got in (c) and (d) I will recommend Logistic Model because Area under Curve of the logit model is greater than that of Decision tree model and also misclassification rate of logistic model is less than that of decision tree model and from part (d) we can observe that the area under the logistic curve is also greater than the Decision Tree or classification tree curve.