Homework 3

IE 7275 Data Mining in Engineering

Task 1: Tutorial

- Practice R models presented in "R Code for Textbook Examples in Chap 5.pdf." For your convenience, the data sets referenced in the document are included the Homework 3 folder.

Problem 1: Performance metrics [25 points]

The table below shows a small set of predictive model validation results for a classification model, with both propensity and the actual class membership. Consider 1 represents the class of interest.

a. Calculate error rates, sensitivity, and specificity using cutoffs of 0.0, 0.25, 0.5, 0.75, and 1.0. Construct ROC using Excel. Calculate AUROC and comment on the performance of the classification model.

b. Calculate F1 Score and comment on the performance of the classification model.

c. Calculate Matthews Correlation Coefficient and comment on the performance of the classification model.

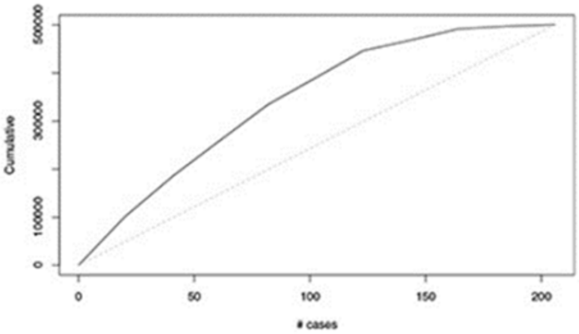
d. Create a decile-wise lift chart in R and comment on the performance of the classification model.

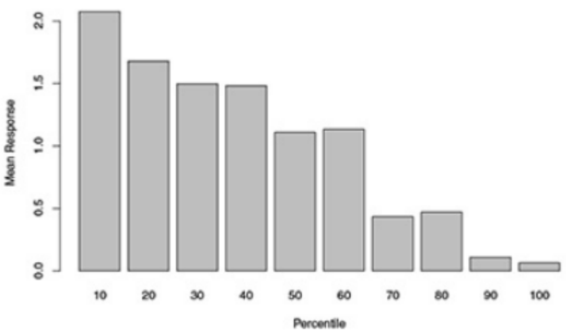
Propensity	Actual Class
0.03	0
0.52	0
0.38	0
0.82	1
0.33	0
0.42	0
0.55	1
0.59	0
0.09	0
0.21	0
0.43	0
0.04	0
0.08	0
0.13	0
0.01	0
0.79	1

0.42	0
0.29	0
0.08	0
0.02	0

Problem 2: Performance metrics [25 points]

A firm that sells software services has been piloting a new product and has records of 500 customers who have either bought the services or decided not to. The target value is the estimated profit from each sale (excluding sales costs). The global mean is $2128. However, the cost of the sales effort is not cheap—the company figures it comes to $2500 for each of the 500 customers (whether they buy or not). The firm developed a predictive model in hopes of being able to identify the top spenders in the future. The lift and decile charts for the validation set are shown in below.





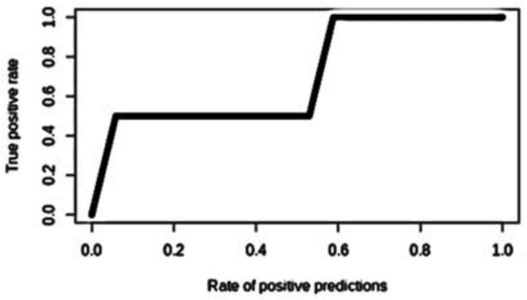
a. If the company begins working with a new set of 1000 leads to sell the same services, similar to the 500 in the pilot study, without any use of predictive modeling to target sales efforts, what is the estimated profit?
b. If the firm wants the average profit on each sale to at least double the sales effort cost, and applies an appropriate cutoff with this predictive model to a new set of 1000 leads, how far down the new list of 1000 should it proceed (how many deciles)?
c. Still considering the new list of 1000 leads, if the company applies this predictive model with a lower cutoff of $2500, how far should it proceed down the ranked leads, in terms of deciles?
d. Why use this two-stage process for predicting sales—why not simply develop a model for predicting profit for the 1000 new leads?

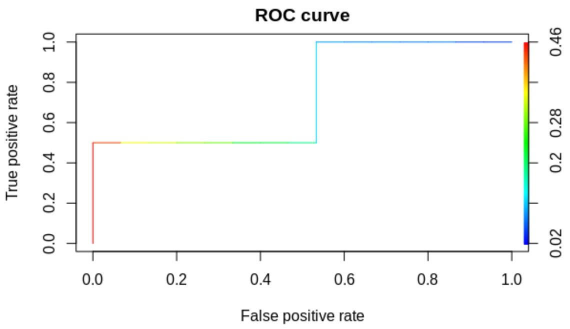
Problem 3: ROC and Lift Chart [25points]

Open ROC_LiftChart_1.Rmd file in R-Studio, and then run each block of code individually. The model and the data input source have already been implemented in the code. To complete the problem, add necessary code to perform the following tasks.

- Divide dataset into training and testing datasets in 70:30 ratio
- Using the model provided, classify the cases in test dataset.
- Draw a ROC and a Lift Chart for the test dataset.

After you finalize your code, generate the report as a pdf document. Your charts should look something like to the following:

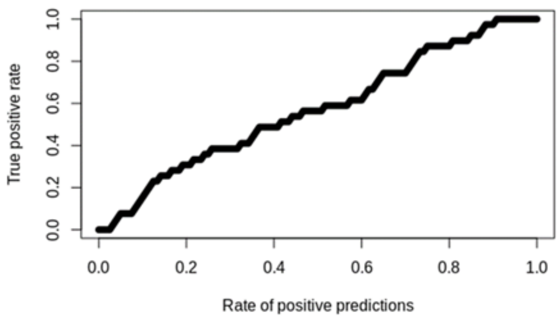


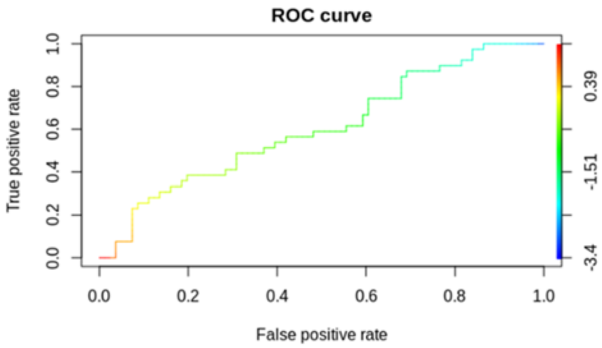


Problem 4: ROC and Lift Chart [25 points]

Repeat Problem 1 with ROC_LiftChart_2.Rmd

Your charts should look something like to the following:





Files Included in the Folder:

Homework 3.docx
R Code for Textbook Examples in Chap 5.pdf
ROC_LiftChart_1.Rmd
ROC_LiftChart_2.Rmd