MBA786M: Alert Models in Finance

Submission deadline: Sunday, 15 September 2024, 11:59 PM

This project asks you to determine the credit quality of a bank's customer. You are encouraged to use the techniques introduced in this module or any other you think is applicable to the dataset. You will be assessed on your ability to critically evaluate the methods used and report your findings in a coherent way.

This project counts for 30% of your overall grade. You should work in groups (5 students per group) and submit a single project report for your group. The maximum word count for the report is 3,000. Please email me (parvati@iitk.ac.in) your group project report and a zip file containing your program code by Sunday, 15 September.

Data

The data consists of variables that inform the credit worthiness of a bank customer. The dependent variable, *Class***,** is binary and differentiates customers, on their observed credit performance, as either *Good* or *Bad***.**

The independent variables consists of: *checking account status, duration, credit history, purpose of the loan, amount of the loan, savings accounts or bonds, employment duration, Instalment rate in percentage of disposable income, personal information, other debtors/guarantors, residence duration, property, age, other instalment plans, housing, number of existing credits, job information, number of people being liable to provide maintenance for, telephone, and foreign worker status.*

Project Tasks

1. (a) Fit a logistic regression model on the dataset. Choose a probability of default threshold of 20%, 35%, and 50%, to assign an observation to the *Bad* class. Compute a confusion matrix for each of the models. How do the True Positive and False Positive rates vary over these models? Which model would you choose?

   (b) Divide the dataset into training (70%) and test (30%) sets and repeat the above question and report the performance of these models on the test set.

(c) Plot the ROC for a logistic model on a graph and compute the AUC. Explain the information conveyed by the ROC and the AUC metrics. [7 marks]

2. (a) Fit classification tree, bagging and random forest models on the dataset and comment on the performance of these models. Do you think we are overestimating the performance of these models by fitting them on to the whole dataset? If so, state your reasons.

(b) Split the dataset in two parts: training (70%) and test sets (30%). Fit the models on the training dataset and evaluate their performance on the test set. Which model would you choose and why?

(c) For the best model chosen, rank and plot predictors according to their predictive power.

(d) How do these models perform compared to the model in question 1? [7 marks]

3. (a) Standardize your predictors and fit KNN classifier with K equal to 1, 3, 5, 10, respectively. Evaluate the performance of these models on the test set.

(b) How do these models perform compared to the tree-based models in question 2 and logistic model of question 1? [7 marks]

4. (a) Fit at least one other binary classifier (e.g., a linear probability model or a Support Vector Machine classifier) to the dataset. Describe its performance relative to the classifiers highlighted above.

(b) Is your training dataset balanced? Comment on the drawbacks of fitting a Statistical Learning technique on an unbalanced dataset. Can confusion matrix be a useful performance metric for this problem? Can you think of / identify a technique to address this concern? If so, why do you think that the method(s) could work?

Hint: This question has not been discussed by way of a formal teaching section on the module. It is up to each student group to search for a systematic understanding and solution to the phenomenon of imbalanced data. [9 marks]