BUSINESS DATA MINING (IDS 572)

HOMEWORK 3
DUE DATE: FRIDAY, NOVEMBER 11 AT 11:59 PM

- You should submit an electronic pdf or word file along with your Rmd file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.
- One submission is sufficient for the entire group.

This assignment is based on the case study "Predicting Earning Manipulations by Indian Firms using Machine Learning Algorithms." Please answer the following questions as detailed as possible. Please do not forget EDA.

(1) Do you think the Beneish model developed in 1999 will still be relevant to Indian data?

(2) The number of manipulators is usually much less than non-manipulators (in the accompanying spreadsheet, the percentage of manipulators is less than 4% in the complete data). What kind of modeling problems can one expect when cases in one class are much lower than the other class in a binary classification problem? How can one handle these problems? – To answer this question: Watch "balancing data in R" video uploaded under "R documents".

(3) Use a sample data (220 cases including 39 manipulators) and develop a logistic regression model that can be used by MCA Technologies Private Limited for predicting probability of earnings manipulation.

(4) What measure do you use to evaluate the performance of your logistic regression model? How does your model perform on the training and test datasets?

(5) What is the best probability threshold that can be used to assign instances to different classes? Write two functions that receive the output of the ROC performance function and return the best probability thresholds using the distance to (0,1) and Youden's approach respectively.

(6) Based on the models developed in questions 4 and 5, suggest a M-score (Manipulator score) that can be used by regulators to identify potential manipulators.

(7) Develop a decision tree model. What insights do you obtain from the tree model?

(8) Develop a logistic regression model using the complete data set (1200 non-manipulators and 39 manipulators), compare the results with the previous logistic regression model and comment on differences.