

MIS522 Business Data Mining

Assignment 2

Name: _____ ID: _____

Name: _____ ID: _____

Problem 1. Download the dataset `datamining.xlsx` from LMS. This dataset contains 2,000 cases. This dataset is to be used to predict whether a person in an MIS program will like a data mining course or not. The fields for each of the 2000 records are as below:

- GMAT: GMAT score of a student
- Bachelor: Field of BS degree (A: Arts, S: Science, E: Engineering)
- Quant, Stats, HBO, Acct: Course rating of the student for each of the courses from 1 (lowest) to 5 (highest)
- E-comm: Flag that is T if student intends to specialize in e-commerce, F otherwise
- Datamine: Course rating of the student for Data Mining
- LikeDM: Flag that is T if course rating for Data Mining is 4 or 5; F otherwise (*note that this attribute is derived from “Datamine” attribute, so you should eliminate “Datamine” from exploration and modeling*).

Using RapidMiner, please answer the following questions. A sample process is provided as a starter.

- Use the entire data (`datamining.xlsx`) and explore the relationship between LikeDM and each individual field. What effect does each field seem to have on LikeDM? You can use scatterplots and histograms to explore the relationships and show only what seems to be important relations. (using excel or any preferred visualization tool)
- Split the data into 65% for training and 35% for testing using Split Validation operator. Click on the operator and change random seed value to “12345”. Create a Decision Tree (Modeling→Predictive→Tree→Decision Tree) and make sure to get 100% accuracy on **training** data. To do this, set criterion to gini index, set the tree depth to high number (2000) and uncheck “apply pruning” and “apply prepruning,” (**Model 1**)

then answer the following:

- What is the depth of the tree?
- How many leaves (decision nodes) does it have?

3. What is the accuracy of the **testing** data, and why is it not 100%?
- c) Now change the settings of the decision tree model as follows, then answer the questions:
- Click on the decision tree and choose criterion to “information_gain” and set maximum depth to 8.
 - Check “apply prepruning” and set minimal gain to 0.01, minimal leaf size to 2, and minimal size for split to 4 (leave other options as is). **(Model 2)**
1. What is the accuracy of training and testing?
- d) Use the models developed above to compare between their performance by creating any of the following two charts : gain chart, lift chart, and response chart.
1. Now, which model is the best? Why?