**MIS532 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, answer the following questions. A sample process is provided as a starter.

1. [20 points] 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.
2. [25 points] 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 (e.g. 2000) and uncheck both “apply pruning” and “apply prepruning,” then answer the following:
   1. What is the depth of the tree?
   2. How many leaves (decision nodes) does it have?
   3. What is the accuracy of the **testing** data, and why is it not 100%?
3. [15 points] 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).
   1. What is the accuracy of training and testing? Do you see an improvement in the model? How?
   2. Provide two strongest If-Then rules from this decision tree. Please explain why these rules are chosen.
4. [15 points] Try to further improve the performance of the decision tree model by changing the decision tree parameters (you can change tree depth, type of criterion, or minimal size of leaf or split). What is the performance of the tree you created (both training and testing) and what have you changed in the tree settings? Produce at least two (2) different models.
5. [10 points] Use the models developed above to compare between their performance by filling the table provided. Which model is the best, and why?

|  |  |  |  |
| --- | --- | --- | --- |
| **Configuration** | **Training** | **Testing** | **Difference** |
| Model 1. First (Overfitting) from b) |  |  |  |
| Model 2. from c) |  |  |  |
| Model 3. from d) |  |  |  |
| Model 4. from d) |  |  |  |
|  |  |  |  |

1. [15 points] Use the models developed above to compare between their performance by creating gain chart, lift chart, and response chart.
   1. Now, which model is the best? Why?
   2. How can this model (the best model) be used by school officials?