**MIS419 Business Data Mining**

**Assignment 2**

**Name: Fahad Khalid Alofani ID: 437104121**

**Name: Faisal Abdullah Almahmood ID: 437104470**

**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.

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.



**-GMAT:** There is a relationship, people who like DM will most likely to have high GMAT.

****

**-Bachelor:** There are relationships between the majors and LikeDM, people who major’s arts & science tend to dislike DM. However, people who major’s engineering tend to like DM more.

****

**-Quant**: There is a relationship,students who likes Quant will most likely to like DM.



**-Stats:** There is a relationship, student with score 4 & 5 in stats will most likely to like DM.

****

**-HBO:** There is a reverse relationship, student who likeDM most likely will dislike HBO.

****

**-Acct:** There is no relationship.



**-E-comm:** There is a relationship, students who didn’t major’s E-comm most likely will dislikes DM.

1. [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 (2000) and uncheck “apply pruning” and “apply prepruning,” then answer the following:
   1. What is the depth of the tree?

17.

* 1. How many leaves (decision nodes) does it have?

240.

* 1. What is the accuracy of the **testing** data, and why is it not 100%?

The accuracy of the testing data is 88.29%. It’s not 100% because in testing data it’s like the future we don’t know the outcome. Also, the training data accuracy is 100% because the model is overfitting. So, when we apply the model on the testing data the accuracy will decrease.

1. [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?

Training: 93.92%

Testing: 89.57%

* 1. Provide two strongest If-Then rules from this decision tree. Please explain why these rules are chosen.

if Bachelor = A and GMAT ≤ 702 and GMAT ≤ 640.500 then F (400 / 0).

if Bachelor = S and GMAT ≤ 620.500 and E-comm = F then F (255 / 0).

Because they both have the highest ratio of total.

1. [15 points] Try to 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.

**Model3 Training:**

A screenshot of a cell phone

Description automatically generated

**Model3 Testing:**

A screenshot of a cell phone

Description automatically generated

**Model3 Settings:**

A screenshot of a cell phone

Description automatically generated

**Model4 Training:**

A screenshot of a cell phone

Description automatically generated

**Model4 Testing:**

A screenshot of a cell phone

Description automatically generated

**Model4 Settings:**

A screenshot of a cell phone

Description automatically generated

1. [10 points] Use the models developed above to compare between their performance by filling the table provided. Which model is the best? Why?

|  |  |  |  |
| --- | --- | --- | --- |
| **Configuration** | **Training** | **Testing** | **Difference** |
| Model 1. First (Overfitting) from b) | 100% | 88.29% | 11.71 |
| Model 2. from c) | 93.92% | 89.57% | 3.72 |
| Model 3. from d) | 90.31% | 90.00% | 0.31 |
| Model 4. from d) | 93.38% | 89.71% | 3.67 |
|  |  |  |  |

Model 3 is the best, because it has the highest testing performance and lowest difference.

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?