**The Hang Seng University of Hong Kong**

**COM6101 Marketing Analytics with Machine Learning**

**Assignment 2**

**Due: 14 April 2024 (Sun) 23:59**

**Part I. Prediction (2.5 marks)**

HSUHK is predicting the admission chance for their MSc DSAI applicants based on the following variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Type | Category | Description |
| GRE Score | Independent | Numerical | GRE score: *out of 340* |
| TOEFL Score | Independent | Numerical | TOEFL score: *out of 120* |
| University Rating | Independent | Categorical | Applicant’s university rating: *One, Two, Three, Four or Five* |
| SOP | Independent | Numerical | Statement of Purpose: *out of 5.0* |
| LOR | Independent | Numerical | Letter of Recommendation: *out of 5.0* |
| CGPA | Independent | Numerical | Undergraduate Cumulative Grade Point Average: *out of 10.0* |
| Research | Independent | Categorical | Research experience: *0 (No) or 1 (Yes)* |
| Chance | Dependent | Numerical | Chance of admission: *0.0 – 1.0* |

The linear regression results are as below:

文本, 信件

描述已自动生成

1. Interpret the relationship between CGPA and chance of admission. (0.5 marks)
2. What would be the predicted chance of admission for an applicant who graduated from a class four university with GRE score of 327, TOEFL score of 115, SOP of 3.0, LOR of 3.5, CGPA of 8.2 and nil research experience? (Show your steps to score the marks) (1 mark)
3. Interpret the impact of university rating on chance of admission. (0.5 marks)
4. Please help HSUHK suggest what actions can be taken by their applicants to increase the chance of admission. (0.5 marks)

**Part II. Classification (5 marks)**

Based on the HSUHK admission case in Part I (assume that the chance of admission is now either *Yes* or *No*). The decision tree results are displayed as follows:

=== Tree Visualization ===

Diagram

Description automatically generated

======================= Confusion Matrix ========================

表格

描述已自动生成

1. How many leaf nodes does the tree have when CGPA is smaller than or equal to 8.53? (0.5 marks)
2. What is the prediction of the rightmost leaf node in the tree? State the conditions and prediction result. (1 mark)
3. What types of students would be admitted when university rating was considered but LOR was not considered? (0.5 marks)
4. What is the overall accuracy and misclassification rate of the tree? (Convert to 2 decimal places) (0.5 marks)
5. Please interpret the confusion matrix. Which prediction case is the worst to be discovered? (2.5 marks)

**Part III. Customer Segmentation (3.5 marks)**

You are the manager of HSUHK online store and have some basic data about your 2000 customers, such as gender, marital status, annual income, and size of the city they come from as follows:

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| Sex | Categorical | Customer’s gender:  0: male  1: female |
| Marital Status | Categorical | Customer’s marital status:  0: single  1: married |
| Income | Numerical | Annual income in $ |
| Settlement Size | Categorical | City size where the citizen is living  0: small city  1: medium-sized city  2: large city |

The clustering analysis result is displayed as follows:

Table

Description automatically generated

1. Which cluster has the smallest number of customers? How many customers does it have? (0.5 mark)
2. What are the characteristics of the customers in Cluster #0? (1 mark)
3. How would you describe the relationship between a person’s income and where he/she lives? (1 mark)
4. If a new product is targeted to existing single female customers, which cluster should you targeted? How would you promote it according to the clustering result above? (1 mark)

**Part IV. Association Rules Mining (3 marks)**

The HSUHK online store has the following customer transactions:

|  |  |
| --- | --- |
| Transaction ID | Items |
| T1 | A, B, C, E |
| T2 | A, B, C, D, E |
| T3 | A, B, C, G |
| T4 | A, C, F |
| T5 | A, B, D, E, F |

1. Suppose the minimum support and minimum confidence are 60% and 75%, respectively. Show all valid association rules with their support, confidence and lift values using Apriori algorithm (display the rules in descending order of their lift values). (2 marks)
2. Based on the results in Q14, which one item will you promote to maximize the sales revenue? Why? (0.5 marks)
3. Discuss the difference between rules E → A and E → B. (0.5 marks)

**Part V. Recommendation Engine (3.5 marks)**

The HSUHK online store has the following customer book ratings:

Table

Description automatically generated

1. Calculate the predicted rating for Max on Book5 using user-based and item-based collaborative filtering (with Pearson correlation and all users/items considered with similarity > 0). (Show your steps to score the marks) (2 marks)
2. Based on the predicted ratings in Q17, would you recommend Book5 to Max? Why?

(1 mark)

1. If you should recommend Book5 to Max, what would be your sales or marketing strategy? Please discuss one strategy. (0.5 marks)

**Part VI. Sentiment Analysis (2.5 marks)**

There are two documents:

John likes to watch movies. Mary likes movies too.

Mary also likes to watch football games.

1. Show the normalized TF-IDF matrix. (2 marks)
2. List the top 3 tokens that are most unique to these documents. (0.5 marks)

**Submission Guidelines**

* Please type your answers in a NEW Word document. Make your answers as simple as possible. Save your answers as a PDF document.
* Make the file name of your PDF document as *<your full name>*\_*<your student ID>*\_COM6101\_Assignment\_2.pdf. e.g., CHAN\_TAI\_MAN\_p220000\_COM6101\_Assignment\_2.pdf.
* Submit your PDF document to Moodle. There will be a score penalty for any late submissions.
* Plagiarism is always strictly prohibited and will result in severe disciplinary actions.

**- End -**