**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)

The chance of admission has positive relationship with the CGPA, for each unit increase in CGPA, the chance of admission will be increase by approximately 0.1194. maybe, you have a high CGPA, you will more chance get admission.

1. 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)

Chance=0.0019 \* GRE Score +0.0029 \* TOEFL Score +0.0218 \*（ University Rating=Five） +0.0179 \* LOR+0.1194 \*CGPA +0.0251.\* Research -1.2867

GIVEN:

University Rating=4 GRE score=327 LOR=3.5 CGPA=8.2

TOEFL score =115 SOP=3 Research=0

GET:

Chance=0.0019\*327+0.0029\*115+0.0179\*3.5+0.1194\*8.2-1.2867=0.71063

1. Interpret the impact of university rating on chance of admission. (0.5 marks)

According to the linear regression result, we found that only universities with a rating of 5 are positively correlated with chance of admission, while universities with other rating levels have no relationship, this means that only universities with a rating of 5 will be bonus for admission, and there is no advantage in other university rating situations.

1. Please help HSUHK suggest what actions can be taken by their applicants to increase the chance of admission. (0.5 marks)

In order to increasing the chance of admission, you need to enhance your CGPA score and enrich research experieces. also, if you come from the top five rating of university, this situation will be increase your chance of admission. What’s more, you can provide more letter of recommendation when you submit applicantion materials, and provide your TOFEL and GRE report. All that do like this , you will more chance get this admission.

**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)

There are 8 leaf nodes when CGPA is smaller than or equal to 8.53.

1. What is the prediction of the rightmost leaf node in the tree? State the conditions and prediction result. (1 mark)

The rightmost leaf node in the tree is positive(that is ‘Yes’) result of admission, which need satisfy CGPA is lager than 8.53 and GRE score is lager than 318. and we found the node shows Yes(204/6),this means that have 210 instances follow this path,and 204 yield a ‘Yes’ result, 6 yield a ‘no’ result. Indicates that the prediction confidence in this path is very high.

1. What types of students would be admitted when university rating was considered but LOR was not considered? (0.5 marks)

If we are consider university rating and are not consider LOR, the follow type of students will be admitted:

1.University Rating = Five;

2.University Rating = Four and TOFEL Score > 106.

1. What is the overall accuracy and misclassification rate of the tree? (Convert to 2 decimal places) (0.5 marks)

Accuray= (TP+TN)/(TP+TN+FP+FN)=(89+57)/(89+57+8+16)=148/170=85.88%

misclassification rate=(FP+FN)/(TP+TN+FP+FN)=(8+16)/(89+57+8+16)

=24/170=14.12%

1. Please interpret the confusion matrix. Which prediction case is the worst to be discovered? (2.5 marks)

According to the confusion matrix, we found that:

TP=89 TN=57 FP=8 FN=16

Precision=TP/(TP+FP)=89/97≈91.75%

Recall=TP/(TP+FN)=89/105≈84.76%

F1 score=2\*[precision\*recall/(precesion+recall)]≈88.10%​

For this prediction case, High Recall is important, because students incorrectly classified an unadmitted situation, this leads to excellent students miss out a dream school chance, and it also causes universities to miss out on outstanding students.

**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)

Cluster 0 has the smallest number of customers , and it has 257 customers.

1. What are the characteristics of the customers in Cluster #0? (1 mark)

In the cluster 0, we found that customers is married male and most of them settle in medium-sized city, they are average income is about 122976$.

1. How would you describe the relationship between a person’s income and where he/she lives? (1 mark)

We can found that people’s income has positive relationship with settlement size, the large settlement size and the higher income will be. This relationship is consistent with the realization that people in small cities have lower income, but in big cities, people get more income.

1. 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)

According to the clustering result above, cluster 2 is the target customer to promote a new product. This customer is most settled in a small city and has a lower income, so we need to launch preferential packages or coupons to promote the new product sales and stimulate their consumption.

**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)

minimum support = 60%, So in this question, it needs to appear at least 3 times. According to this rule, we can get:

|  |  |  |
| --- | --- | --- |
| Itemset | Count | Support |
| A | 5 | 1.0 |
| B | 4 | 0.8 |
| C | 4 | 0.8 |
| E | 3 | 0.6 |
| A,B | 4 | 0.8 |
| A,C | 4 | 0.8 |
| A,E | 3 | 0.6 |
| B,C | 3 | 0.6 |
| B,E | 3 | 0.6 |
| A,B,C | 3 | 0.6 |
| A,B,E | 3 | 0.6 |

Minimum confidence=75% , so get this items:

|  |  |  |  |
| --- | --- | --- | --- |
| Antecedents | Consequents | Confidence | Lift |
| A | B | 4/5=0.8 | 0.8/0.8=1 |
| A | C | 4/5=0.8 | 0.8/0.8=1 |
| E | A | 3/3=1 | 1/1=1 |
| B | C | 3/4=0.75 | 0.75/0.8=0.9375 |
| B | E | 3/4=0.75 | 0.75/0.6=1.25 |
| B | A | 4/4=1 | 1/1=1 |
| C | A | 4/4=1 | 1/1=1 |
| C | B | 3/4=0.75 | 0.75/0.8=0.9375 |
| E | B | 3/3=1 | 1/0.8=1.25 |
| A,B | C | 3/4=0.75 | 0.75/0.8=0.9375 |
| A,C | B | 3/4=0.75 | 0.75/0.8=0.9375 |
| B,C | A | 3/3=1 | 1/1=1 |
| C | A,B | 3/4=0.75 | 0.75/0.8=0.9375 |
| B | A,C | 3/4=0.75 | 0.75/0.8=0.9375 |
| A,B | E | 3/4=0.75 | 0.75/0.6=1.25 |
| A,E | B | 3/3=1 | 1/0.8=1.25 |
| B,E | A | 3/3=1 | 1/1=1 |
| E | A,B | 3/3=1 | 1/0.8=1.25 |
| B | A,E | 3/4=0.75 | 0.75/0.6=1.25 |

Display the rules in descending order of their lift values:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Antecedents | Consequents | Antecedents Support | Consequents Support | Confidence | Lift |
| A,E | B | 0.6 | 0.8 | 1 | 1.25 |
| E | A,B | 0.6 | 0.8 | 1 | 1.25 |
| E | B | 0.6 | 0.8 | 1 | 1.25 |
| B | E | 0.8 | 0.6 | 0.75 | 1.25 |
| A,B | E | 0.8 | 0.6 | 0.75 | 1.25 |
| B | A,E | 0.8 | 0.6 | 0.75 | 1.25 |
| E | A | 0.6 | 1.0 | 1 | 1 |
| B,C | A | 0.6 | 1.0 | 1 | 1 |
| B,E | A | 0.6 | 1.0 | 1 | 1 |
| C | A | 0.8 | 1.0 | 1 | 1 |
| B | A | 0.8 | 1.0 | 1 | 1 |
| A | C | 1.0 | 0.8 | 0.8 | 1 |
| A | B | 1.0 | 0.8 | 0.8 | 1 |
| C | B | 0.8 | 0.8 | 0.75 | 0.9375 |
| B | C | 0.8 | 0.8 | 0.75 | 0.9375 |
| A,B | C | 0.8 | 0.8 | 0.75 | 0.9375 |
| A,C | B | 0.8 | 0.8 | 0.75 | 0.9375 |
| C | A,B | 0.8 | 0.8 | 0.75 | 0.9375 |
| B | A,C | 0.8 | 0.8 | 0.75 | 0.9375 |

1. Based on the results in Q14, which one item will you promote to maximize the sales revenue? Why? (0.5 marks)

Based on the results, promoting item E would maximize the sales revenue. Because among the items with confidence level of 1 and lift is greater than 1, E as the antecendent item that can be improve the sales of items such as ‘B’,’A,B’ ,etc .

1. Discuss the difference between rules E → A and E → B. (0.5 marks)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Antecedents | Consequents | Antecedents Support | Consequents Support | Confidence | Lift |
| E | A | 0.6 | 1.0 | 1 | 1 |
| E | B | 0.6 | 0.8 | 1 | 1.25 |

The two rules E→A and E→B have different consequents, The confidence for both rules is the same, at 1. This means that whenever E appears in a transaction, A or B will also appear with 100% confidence. However, the lift values for the two rules differ. The lift for E→A is 1, indicating that the occurrence of E and A are independent of each other. In contrast, the lift for E→B is 1.25, indicating that the occurrence of E and B are positively correlated, and that the presence of E in a transaction increases the likelihood of B also being present.

Therefore, while both rules suggest that E is a strong predictor of the presence of A or B, the rule E→B suggests a stronger association between E and B compared to E and A. This could potentially be used to inform marketing strategies, such as promoting item B alongside item E to increase sales.

**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)

Using user-based collaborative filtering to calculate the predicted rating for Max on Book 5 by Pearson correlation, similarity formulation like: The calculation results are as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Sim(Max,v) | Book1 | Book2 | Book3 | Book4 | Book5 |
| John | 3 | 0.707 | 3 | 3 | 2 | 4 | 3 |
| David | 3.4 | -0.422 | 5 | 2 | 4 | 1 | 5 |
| Helen | 3.4 | 0.513 | 5 | 2 | 1 | 5 | 4 |
| Max | 3.5 |  | 4 | 5 | 1 | 4 | ? |

So use prediction formulation like:  get the prediction results:

Pred(Max,Book5)=3.5+≈3.752≈4

Using item-based collaborative filtering to calculate the predicted rating for Max on Book 5 by Pearson correlation, The calculation results are as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Book1 | Book2 |  | Book3 | Book4 | Book5 |
| John | 3 | 3 |  | 2 | 4 | 3 |
| David | 5 | 2 |  | 4 | 1 | 5 |
| Helen | 5 | 2 |  | 1 | 5 | 4 |
| Max | 4 | 5 |  | 1 | 4 | ? |
|  | 4.25 | 3 |  | 2 | 3.5 | 4 |
| Sim(Max,v) | 0.866 | -0.5 |  | 0.655 | -0.721 |  |

So use prediction formulation like:  get the prediction results:

Pred(Max,Book5)=≈2.708≈3

1. Based on the predicted ratings in Q17, would you recommend Book5 to Max? Why?

(1 mark)

According to the predicted ratings in Q17, we know that prediction result is rounded of 4 on user-based collaborative filtering and is rounded of 3 on item-based collaborative filtering. If this rating is out of 5, I will recommend book5 to Max. for reason that:

I consider a rating of 4 to be good , suggests that Max is likely to enjoy Book 5, and thus it would be a good recommendation.

On the other hand, I considers a 3 as an average or satisfactory rating, it may still be worth recommending, but it's not as strong of a recommendation as the one suggested by the user-based approach.

In summary, It is depending on rating threshold and users preference whether recommend book5 to Max, if Max is the type of common reader who is looking for books that he is likely to rate as 3 or above, but if Max is very selective and only prefers books that he would rate a strong 4 or 5, the recommendation would be change.

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

I will be launch special package to induce Max bundle consumption. like book5 is combined with book1 to sell that price will be less than independently purchased pricing.

Also, when Max browsing the web, we can place book5 on the homepage or checkout page that users browse to increase the purchase rate of the product in an eye-catching way.

**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. Tokenize the documents into indi vidual terms, and remove stop words. Get results like:

John likes to watch movies too Mary also football games

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | John | likes | to | watch | movies | too | Mary | also | football | games |
| 0 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |

②Create a term frequency (TF) matrix that counts how often each term appears in each document.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | John | likes | to | watch | movies | too | Mary | also | football | games |
| 0 | 0.111 | 0.222 | 0.111 | 0.111 | 0.222 | 0.111 | 0.111 | 0 | 0 | 0 |
| 1 | 0 | 0.143 | 0.143 | 0.143 | 0 | 0 | 0.143 | 0.143 | 0.143 | 0.143 |

③Compute the inverse document frequency (IDF)for each term across all documents.

IDF(‘John’)=log()=1

IDF(‘likes’)=log()=0

IDF(‘to)=log()=0

IDF(‘watch’)=log()=0

IDF(‘movies’)=log()=1

IDF(‘too)=log()=1

IDF(‘Mary’)=log()=0

IDF(‘also’)=log()=1

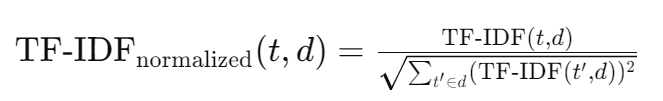
IDF(‘football’)=log()=1

IDF(‘games’)=log()=1

④Calculate the TF-IDF score for each term in each document by multiplying the TF by the IDF.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | John | likes | to | watch | movies | too | Mary | also | football | games |
| 0 | 0.111 | 0 | 0 | 0 | 0.222 | 0.111 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.143 | 0.143 | 0.143 |

⑤Normalize the TF-IDF scores for each document



Among them, the numerator refers to the TF-IDF value of each item, and the denominator refers to the L2 norm, which is the root sum of the square sum of the TF-IDF value of each item in the document. So:

L2 norm of Document 1: 0.272

L2 norm of Document 2: 0.248

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | John | likes | to | watch | movies | too | Mary | also | football | games |
| 0 | 0.408 | 0 | 0 | 0 | 0.816 | 0.408 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.577 | 0.577 | 0.577 |

1. List the top 3 tokens that are most unique to these documents. (0.5 marks)

According to normalized TF-IDF score, we know that most unique top 3 tokens in these documents are:

movies- Normalized TF-IDF score is 0.816

football- Normalized TF-IDF score is 0.577

games- Normalized TF-IDF score is 0.577

And the top 3 most unique tokens for these two documents are as follows:

Document 0

movies- Normalized TF-IDF score is 0.816

John- Normalized TF-IDF score is 0.408

too- Normalized TF-IDF score is 0.408

Document 1

also- Normalized TF-IDF score is 0.577

games - Normalized TF-IDF score is 0.577

football- Normalized TF-IDF score is 0.577

**- End -**