P233340 Kwok Tsz Yi

Part 1

1. CGPA has a 0.1194 positive correlation with the chance of admission
2. Predictive chance:

0.0019 \* 327 + 0.0029 \* 115 + 3\*0 + 0.0179 \* 3.5 + 8.2 \* 0.1194 + 0 – 1.2867

= 0.70983

1. Only university having a rating with FIVE has a 0.0218 positive correlation with the chance of admission, it is the third important variable that affecting the admission rate
2. Improving the CGPA, practise more research to gain research experience and receive as many letters of recommendations from professors

Part 2

1. 12
2. Condition: CGPA > 8.53, GRE Score > 318, the predicted result is: accept
3. Student from Universities having a rating of 5, or student from universities having a rating of four and TOEFL score higher than 106
4. Overall accuracy: (57 + 89)/ (57 + 89 + 16 + 8) = 0.86, misclassification rate: (16+8)/ (57 + 89 + 16 + 8) = 0.14
5. Precision: 89/ (89 + 8) = 0.91, Recall: 89/ (89 + 16) = 0.85, as per the precision and recall, this model has a sightly better performance on precision than recall. For the case of admission prediction, precision is more important than recall, as the consequence of making false positive predict of student admission (student predicted to be accepted into the program but turns out not accept) is more serious than making false negative (student predicted to be not accepted into the program but turns out be accepted), the previous case give a greater disappointment to the student, while the latter giving surprise. Therefore, the worse case is: making false positive prediction, and the model having higher precision rate is more important than recall.

Part 3

1. Cluster 0, it has only 257 customers
2. These customers are: married male, having the second highest average income among other clusters, and most of them probably live in a medium-sized city
3. There’s a positive correlation between customer’s income and the size of their hometown: the higher the income, the higher chance that the customer lives in a larger city
4. Cluster 2, as cluster 2 is the only cluster having single female customers. As in cluster 2, they are more sensitive to the product price (the lowest average income among other cluster groups) and most of them live in small city, we can launch marketing campaign in small cities providing promotion there.

Part 4

1. This is the association rules:

A black screen with white text

Description automatically generated

1. Item E, as it appears in all of the association rules having the highest lift, as per Q14, which implies that more customers tend to purchase item E with others.
2. Association rules of E > A and E > B have the same support and confidence, however the lift of E > B is higher than E > A for 0.25, as support of A is 5/5 = 1 while support of B is 4/5 = 0.8, per the lift formula list of (E>B) will be larger than (E>A) for: 1/0.8 – 1/1 = 0.25.

Part 15:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | 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 | ? |

1. User-based collaborative filtering:

Find the mean rating by person:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Book1 | Book2 | Book3 | Book4 | Book5 | **Mean** |
| John | 3 | 3 | 2 | 4 | 3 | **3** |
| David | 5 | 2 | 4 | 1 | 5 | **3** |
| Helen | 5 | 2 | 1 | 5 | 4 | **3.25** |
| Max\* | 4 | 5 | 1 | 4 | ? | **3.5** |

\*For Max rating, I exclude Book5 from calculating the mean

Find the similarities between Max and other people, exclude Book 5:

Formula for Pearson correlation:

A black background with white text

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Name | Mean | Similarity with Max |
| John | 3 | 0.7071067811865476 |
| David | 3 | -0.42163702135578396 |
| Helen | 3.25 | 0.5134360308102703 |

Select user with similarity over 0.5 as threshold, hence only John and Helen opinions will be considered.

Predicted Rating for Max on Book5 via user-based collaborative filtering:

A black and blue text

Description automatically generated

3.5 + (0.707\*(3-3) + 0.513 \* (4-3.25))/(0.707 + 0.513) = **3.82**

Item-based collaborative filtering:

Find the similarities between Book 5 and other items.

Formula for Cosine similarity:

A black text on a white background

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Book1 | Book2 | Book3 | Book4 |
| John | 3 | 3 | 2 | 4 |
| David | 5 | 2 | 4 | 1 |
| Helen | 5 | 2 | 1 | 5 |
| Max | 4 | 5 | 1 | 4 |
| Similarity with Book5\* | 0.99422059 | 0.92609236 | 0.9258201 | 0.80740619 |

\*Exclude Max’s rate

Select item with top 2 similarity with Book5, hence Book1 and Book2 are selected.

Predicted Rating for Max on Book5 via item-based collaborative filtering:

A black text on a white background

Description automatically generated

R = (5 \* 0.926 + 4 \* 0.994)/(0.926 + 0.994) = 4.48

1. Yes, I will recommend, as the average predicted rating by Max is: (4.48 + 3.82)/2 = 4.15, which is higher than 2.5, therefore Max is probably like reading Book5.
2. I would state that Book5 will probably be interested by Max, as the user having similar preference with him rate a high score with Book5, and the genre of this book is similar to Max’s preference.

Part 20

1. Normalized TF-IDF

A screenshot of a black screen

Description automatically generated

1. Most unique tokens as per the TF-IDF, i.e. tokens having the highest value:

A black screen with white text

Description automatically generated