DTI 5126: Fundamentals for Applied Data Science

Summer 2021

Assignment 4

Submission Deadline: 14th July 2021 on Brightspace.

This assignment should be completed individually. Upon completion, present your result (including the answers generated or plots) as a single PDF report. **The PDF report must be submitted as a separate file.** Where applicable, submit the R source codes used to generate your results as a zip folder **excluding the PDF report.**

Part A: Association Rules

I. Given a simple transactional database X: Using the threshold values support = 25% and confidence = 60%,
   a) Find all frequent itemsets in database X;
   b) Find strong association rules for database X;
   c) Analyze misleading associations for the rule set obtained in (b).

X:	TID	Items
	T01	A, B, C, D
	T02	A, C, D, F
	T03	C, D, E, G, A
	T04	A, D, F, B
	T05	B, C, G
	T06	D, F, G
	T07	A, B, G
	T08	C, D, F, G

II. A store is interested in determining the associations between items purchased from its Departments. The store chose to conduct a market basket analysis of specific items purchased to analyze customer's buying behavior. You are hereby provided with a file '*transactions.csv*' containing information for transactions made over the past 3 months.
   a) Generate a plot of the top 10 transactions
   b) Generate association rules using minimum support of 0.002, minimum confidence of 0.20, and maximum length of 3. Display the rules, sorted by descending lift value.
   c) Select the rule from Q1 with the greatest lift. Compare this rule with the highest lift rule for maximum length of 2.
      i) Which rule has the better lift?
      ii) Which rule has the greater support?
      iii) If you were a marketing manager, and could fund only one of these rules, which would it be, and why?

Part B: Course Recommender System using Collaborative Filtering

The Institute for Statistics Education at Statistics.com asks students to rate a variety of aspects of a course as soon as the student completes it. The Institute is contemplating instituting a recommendation system that would provide students with recommendations for additional courses as soon as they submit their rating for a completed course. Consider the excerpt from student ratings of online statistics courses shown in the Table 14.16, and the problem of what to recommend to student E.N.

1) First consider a user-based collaborative filter. This requires computing correlations between all student pairs. For which students is it possible to compute correlations with E.N.? Compute them.
2) Based on the single nearest student to E.N., which single course should we recommend to E.N.? Explain why.
3) Use R to compute the cosine similarity between users.
4) Based on the cosine similarities of the nearest students to E.N., which course should be recommended to E.N.?
5) Apply item-based collaborative filtering to this dataset (using R) and based on the results, recommend a course to E.N.

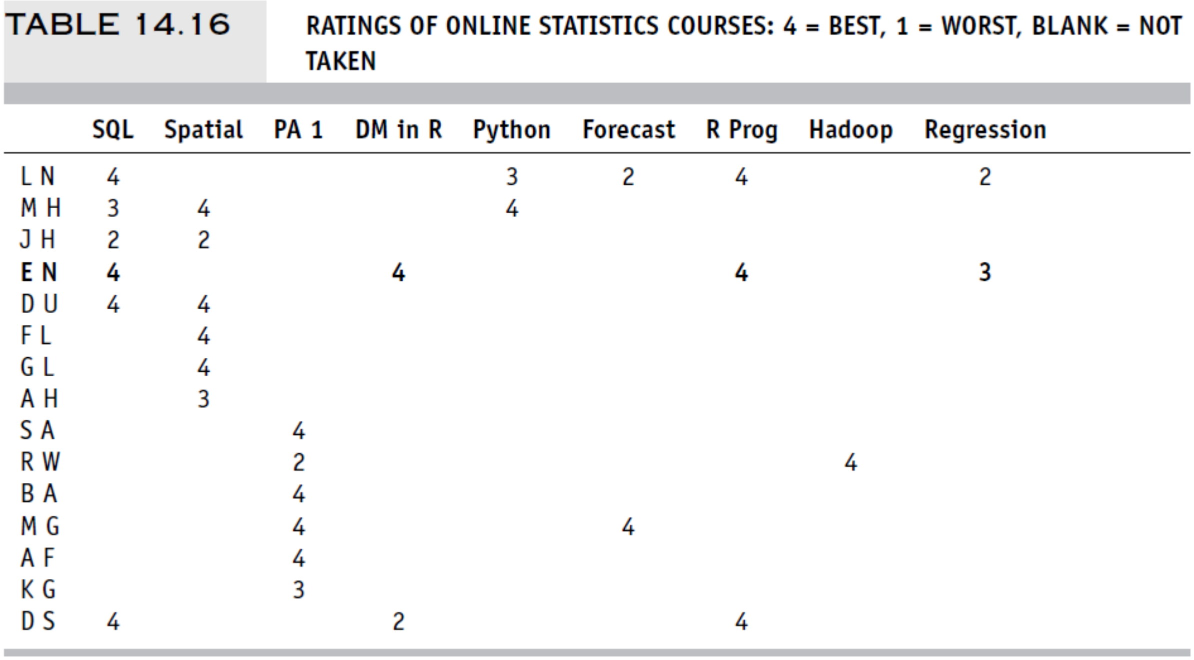

TABLE 14.16 **RATINGS OF ONLINE STATISTICS COURSES: 4 = BEST, 1 = WORST, BLANK = NOT TAKEN**

	SQL	Spatial	PA 1	DM in R	Python	Forecast	R Prog	Hadoop	Regression
LN	4				3	2	4		2
MH	3	4			4				
JH	2	2							
EN	**4**			**4**			**4**		**3**
DU	4	4							
FL		4							
GL		4							
AH		3							
SA			4						
RW			2					4	
BA			4						
MG			4			4			
AF			4						
KG			3						
DS	4			2			4		