Paper 2602-2018

Using Market Basket Analysis in SAS® Enterprise Miner^{TM} to Make Student Course Enrollment Recommendations

Shawn Hall, Aaron Osei, and Jeremiah McKinley, The University of Oklahoma

ABSTRACT

Market basket analysis, an example of association rule mining or affinity analysis, is used most widely in marketing to target customers by identifying the products they purchase in combination. Discovery of existing purchase patterns allows for better product placement, targeted marketing and product recommendations. This data mining technique has also been applied to the analysis of credit card and other service purchases in fraud detection, medical insurance claims and event promotion.

In the fall of 2013, the University of Oklahoma implemented flat rate tuition, in which students enrolled in 12 or more credit hours, pay a flat rate amounting to 15 credit hours of tuition and mandatory fees. Students paying flat rate, but enrolling in fewer than 15 credit hours, may bank those unused hours for enrollment in the following summer term.

Here we consider the application of market basket analysis to student course enrollment. Market basket analysis via the Association Node in SAS® Enterprise Miner^{TM}, allows us to identify and capitalize on existing enrollment patterns, applying the resulting association rules to current fall and/or spring enrollment to fashion course enrollment recommendations for the coming summer term. Encouraging students to continue their studies through the academic year lends to increased retention, higher graduation rates and has students using banked credit hours in our flat rate tuition system that they may forgo otherwise.

INTRODUCTION

Market basket analysis, or more precisely association and sequence analyses, are data mining techniques used most often to identify products purchased in combination and are accomplished using the Association node in SAS® Enterprise Miner^{TM}. Although typically used in marketing, the simplicity of the association discovery has been used in a number of industries, including fraud detection, medical insurance claims, telecommunications and event promotion [4].

Both analyses create a list of association rules based on patterns identified in an initial transaction data set. Each rule is assigned a set of strength of association metrics to aid in identifying useful or interesting rules. Those metrics generally include support, confidence and lift. When we consider the rule that the purchase of, or enrollment in, A (antecedent) leads to the purchase of, or enrollment in, B (consequent), support indicates the probability of the two occurring together, while confidence indicates the conditional probability of B given A. Furthermore, lift, the standard in identifying actionable items, represents the confidence of the rule divided by the expected confidence. A lift greater than 1 signifies a positive correlation and provides an indication of the extent of their dependence.

The goal of this paper is to apply the pattern detection of the association analysis to previous course enrollment in order to make future summer course enrollment recommendations for currently enrolled fall and spring students and provides a how to for those interested in such an application. These course recommendations exist as one branch of a larger next best action recommendation system our office is implementing at the University of Oklahoma.

FLAT RATE TUITION AND SUMMER SESSION

In the fall of 2013, the University of Oklahoma implemented a flat rate tuition program for full time undergraduate students. The applicable tuition and fee rate is based on 15 credit hours of tuition and hourly mandatory fees for the current academic year. Those students registered in fewer than 12 credit hours pay on a per credit hour basis, while those enrolled in 12 or more credit hours pay the current flat rate. Furthermore, a student paying flat rate but enrolled in fewer than 30 credit hours for fall and spring combined, may be eligible to bank those hours and participate in the summer session incentive program at our Norman campus.

Flat rate tuition is intended to encourage students to graduate sooner, allowing them to enter the workforce and earn income earlier. This saves them one to two years of room, board, transportation and other college related expenses, and reduces student loan debt. Additionally, encouraging students to continue their studies through the academic year leads to increased retention, higher graduation rates [1] and has students using banked credit hours they may forgo otherwise.

SUMMER COURSE RECOMMENDATIONS

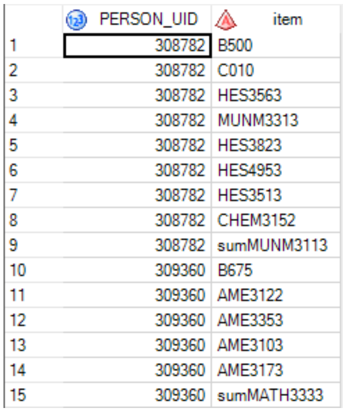
Currently we are in our third round of summer recommendations. Our first effort focused on undergraduate students enrolled in the spring, but not yet enrolled for summer courses. Review of the previous year's undergraduate enrollment for spring and summer, enabled us to identify relationships that were then applied to current student enrollment. In our second effort, recommendations were sent to all undergrads for whom an association rule applied. If a summer course recommendation could be made for a student based on either their fall or spring enrollment, after identifying patterns in fall to summer and spring to summer enrollment of the previous year, a recommendation was provided. Most recently, summer 2018 recommendations based on fall enrollment were communicated to students in December. As of this writing, summer 2018 recommendations based on spring enrollment are scheduled for communication to students in conjunction with the opening of their next enrollment window starting in late March. In all instances, students receive either one, two or three rule-based course recommendations, depending on the rules generated, followed by one general education interest course recommended by the Summer Session office. Freshmen and sophomores receive a lower division general education interest course while juniors and seniors receive an upper division course.

Students in focus groups conducted prior to the larger releases of recommendations reported general agreement in the applicability and interest in the course recommendations they received. Additionally, we have observed an overall increase in credit hour enrollment for summer courses and summer session revenue.

APPLICATION OF ASSOCIATION ANALYSIS TO STUDENT ENROLLMENT DATA

DATA PREPARATION IN SAS® ENTERPRISE GUIDE^{TM}

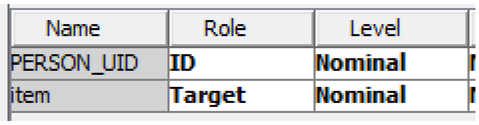
From our student enrollment database, we create a transaction data set with two variables, a unique student id and a basket item variable that includes a combination of major codes, fall or spring course codes and summer course codes (Table 1). These observations were limited to the academic year directly preceding the academic year in which we are making recommendations, the idea being to identify recent patterns in enrollment for application to current enrollment. The data set is further limited to undergraduate student enrollment in more traditional lecture style courses. PROC TRANSPOSE is helpful in converting "wide" data to "narrow" data in creating the transactional structure of this transaction data set.



	PERSON_UID	item
1	308782	B500
2	308782	C010
3	308782	HES3563
4	308782	MUNM3313
5	308782	HES3823
6	308782	HES4953
7	308782	HES3513
8	308782	CHEM3152
9	308782	sumMUNM3113
10	309360	B675
11	309360	AME3122
12	309360	AME3353
13	309360	AME3103
14	309360	AME3173
15	309360	sumMATH3333

Table 1. Transaction Data Set for Association Analysis

This data set is loaded into Enterprise Miner^{TM} as a transaction data set where the role for the unique student ID is set to “ID” and the role for the basket item is set to “Target” (Table 2).



Name	Role	Level
PERSON_UID	ID	Nominal
item	Target	Nominal

Table 2. Transaction Data Set Roles for Association Analysis

ASSOCIATION NODE IN SAS® ENTERPRISE MINER^{TM}

In the Enterprise Miner^{TM} diagram, you can use either the Input Data node or drag and drop from a previously loaded data source to add a data set. Add the Association node to the diagram and connect the two, as depicted in Figure 1.

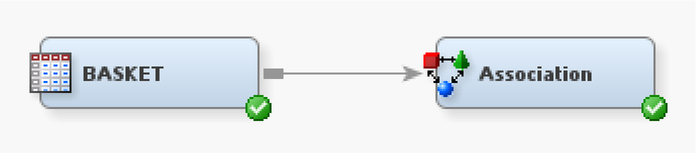
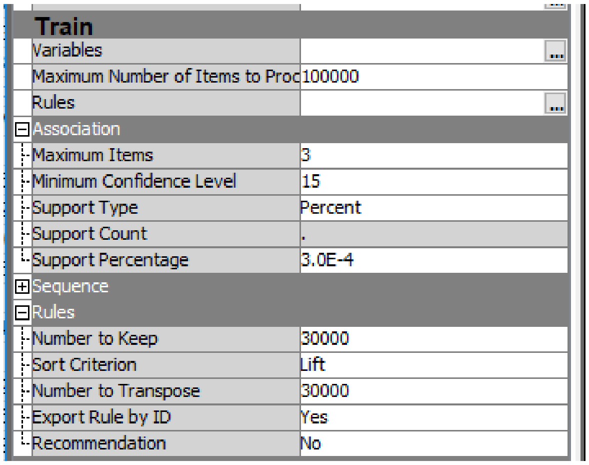


Figure 1. Association Analysis Diagram

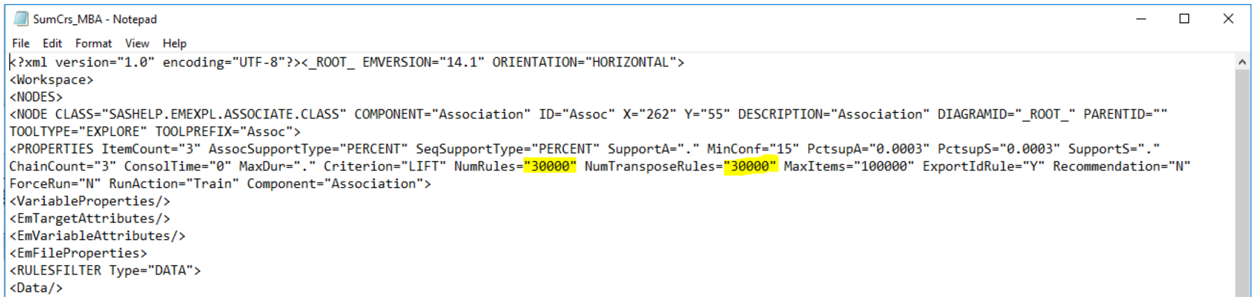
When the Association node is selected, you will see a properties menu on the left hand side of the workspace. In the Association Section set Maximum Items, Minimum Confidence Level and Support Percentage based on your goals. In the Rules section, identify Number to Keep, Number to Transpose, set Sort Criterion to “Lift” and set Export Rule by ID to “Yes” (Figure 2).



Train	
Variables	...
Maximum Number of Items to Proc	100000
Rules	...
⊟ Association	
Maximum Items	3
Minimum Confidence Level	15
Support Type	Percent
Support Count	.
Support Percentage	3.0E-4
⊞ Sequence	
⊟ Rules	
Number to Keep	30000
Sort Criterion	Lift
Number to Transpose	30000
Export Rule by ID	Yes
Recommendation	No

Figure 2. Association Analysis Parameters

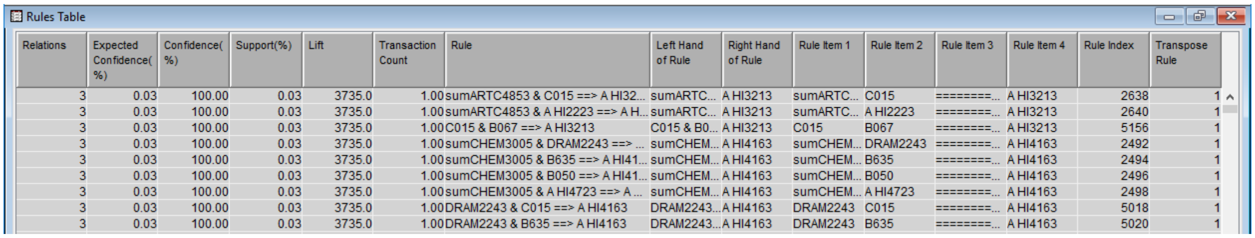
Our transaction data set, or basket, consists of approximately 1,632 items across nearly 3,735 students. Such a sparse matrix calls for lowering Minimum Confidence Level and Support Percentage below default settings. Additionally, because of our interest in summer courses specifically, we raised the Number to Keep and Number to Transpose to 30,000 to ensure we had plenty of resulting rules with a summer courses on the right hand side of the generated rules, as consequent. Adjusting Number to Transpose is not essential, unless you are scoring your current student data set within Enterprise Miner^{TM}. To edit Number to Keep and Number to Transpose beyond the default maximum settings in Enterprise Miner^{TM}, you must save your diagram as an .xml file (Figure 3), edit it in Notepad or another text editor, and import the diagram back into Enterprise Miner^{TM}.



```
<?xml version="1.0" encoding="UTF-8"?><_ROOT_ EMVERSION="14.1" ORIENTATION="HORIZONTAL">
<Workspace>
<NODES>
<NODE CLASS="SASHELP.EMEXPL.ASSOCIATE.CLASS" COMPONENT="Association" ID="Assoc" X="262" Y="55" DESCRIPTION="Association" DIAGRAMID="_ROOT_" PARENTID=""
TOOLTYPE="EXPLORE" TOOLPREFIX="Assoc">
<PROPERTIES ItemCount="3" AssocSupportType="PERCENT" SeqSupportType="PERCENT" SupportA="." MinConf="15" PctsupA="0.0003" PctsupS="0.0003" SupportS="."
ChainCount="3" ConsolTime="0" MaxDur="." Criterion="LIFT" NumRules="30000" NumTransposeRules="30000" MaxItems="100000" ExportIdRule="Y" Recommendation="N"
ForceRun="N" RunAction="Train" Component="Association">
<VariableProperties/>
<EmTargetAttributes/>
<EmVariableAttributes/>
<EmFileProperties>
<RULESFILTER Type="DATA">
<Data/>
```

Figure 3. Diagram .xml

The Rules Table, generated by the Association node, includes the strength of association metrics introduced earlier; support, confidence and lift, in addition to a number of columns addressing the antecedent and consequent of each rule (Table 3).



Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Index	Transpose Rule
3	0.03	100.00	0.03	3735.0	1.00	sumARTC4853 & C015 ==> A HI32...	sumARTC...	A HI3213	sumARTC...	C015	========...	A HI3213	2638	1
3	0.03	100.00	0.03	3735.0	1.00	sumARTC4853 & A HI2223 ==> A H...	sumARTC...	A HI3213	sumARTC...	A HI2223	========...	A HI3213	2640	1
3	0.03	100.00	0.03	3735.0	1.00	C015 & B067 ==> A HI3213	C015 & B0...	A HI3213	C015	B067	========...	A HI3213	5156	1
3	0.03	100.00	0.03	3735.0	1.00	sumCHEM3005 & DRAM2243 ==> ...	sumCHEM...	A HI4163	sumCHEM...	DRAM2243	========...	A HI4163	2492	1
3	0.03	100.00	0.03	3735.0	1.00	sumCHEM3005 & B635 ==> A HI41...	sumCHEM...	A HI4163	sumCHEM...	B635	========...	A HI4163	2494	1
3	0.03	100.00	0.03	3735.0	1.00	sumCHEM3005 & B050 ==> A HI41...	sumCHEM...	A HI4163	sumCHEM...	B050	========...	A HI4163	2496	1
3	0.03	100.00	0.03	3735.0	1.00	sumCHEM3005 & A HI4723 ==> A ...	sumCHEM...	A HI4163	sumCHEM...	A HI4723	========...	A HI4163	2498	1
3	0.03	100.00	0.03	3735.0	1.00	DRAM2243 & C015 ==> A HI4163	DRAM2243...	A HI4163	DRAM2243	C015	========...	A HI4163	5018	1
3	0.03	100.00	0.03	3735.0	1.00	DRAM2243 & B635 ==> A HI4163	DRAM2243...	A HI4163	DRAM2243	B635	========...	A HI4163	5020	1

Table 3. Association Rules Table

SCORING CURRENT STUDENT ENROLLMENT FOR SUMMER COURSE RECOMMENDATIONS

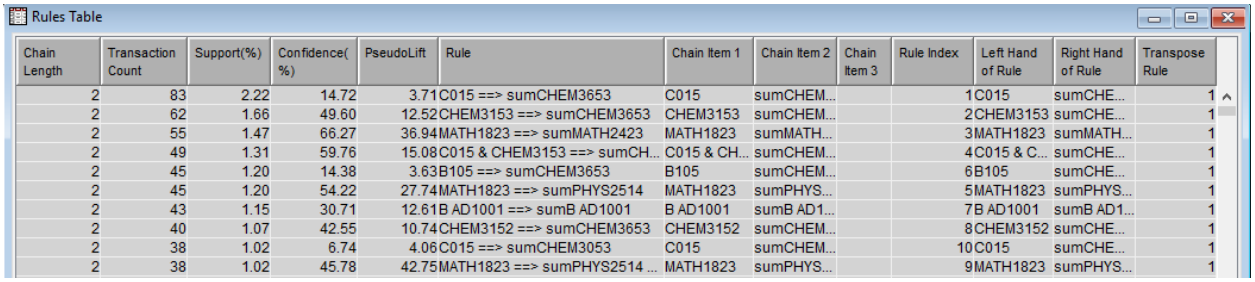
After building the association rules, you will score current student enrollment using either the Score node or by accessing the assoc_links file output by the association analysis, through Enterprise Guide™ and manually pair the antecedent in the rules with current student enrollment to identify recommendations (the rule's consequent). The decision to do either is somewhat arbitrary and you can decide what works best for you as your next step in applying the rules of the association analysis. We chose to pair the two in Enterprise Guide™ because summer enrollment is quite limited this early in the year. Without that variability in the scoring data set, the results you receive from the Score node will be quite limited.

The current student enrollment data set is constructed in the same manner as the basket data set, as a transaction data set. Again, PROC TRANSPOSE is quite helpful in this preparation. From there, join assoc_links, which contains the rules produced by the analysis, to your current enrollment data set. The result is a list of summer course recommendations by student ID.

In a more restrictive/for profit setting, where determining the statistical significance of a rule is important, SAS® provides you the opportunity to introduce significance testing in Enterprise Miner™ with the SAS Code node. Pearson's Chi square test of independence, is a standard test of independence between categorical variables, assessing whether the presence or absence of a consequent is dependent on the presence or absence of an antecedent. Faron and Chakraborty provide instruction and code for addition of the Chi square test in their 2012 publication [3]. We did not find significance testing necessary as we considered almost all rules that resulted in a summer course as a consequent.

ADDED BENEFIT OF A SEQUENCE ANALYSIS

Even when allowing for 30,000 rules in the association analysis, there are too few applicable rules with a summer course as the consequent to make much of a recommendation for the nearly 12,500 undergraduate students we hope to reach. In our search for a method to limit rules further upstream in the analysis, to those leading directly to a summer course, we discovered a special use case of the sequence analysis outlined by Xinli Bao in a 2007 SAS publication [2]. By introducing an artificial "time" variable, setting summer courses to 2 and everything else to 1, you are able to identify relationships associated with a particular basket item or category of basket items. For us, this resulted in a set of rules that all led directly to a summer course consequent (Table 4).



Rules Table

Chain Length	Transaction Count	Support(%)	Confidence(%)	PseudoLift	Rule	Chain Item 1	Chain Item 2	Chain Item 3	Rule Index	Left Hand of Rule	Right Hand of Rule	Transpose Rule
2	83	2.22	14.72	3.71	C015 ==> sumCHEM3653	C015	sumCHEM...		1	C015	sumCHE...	1
2	62	1.66	49.60	12.52	CHEM3153 ==> sumCHEM3653	CHEM3153	sumCHEM...		2	CHEM3153	sumCHE...	1
2	55	1.47	66.27	36.94	MATH1823 ==> sumMATH2423	MATH1823	sumMATH...		3	MATH1823	sumMATH...	1
2	49	1.31	59.76	15.08	C015 & CHEM3153 ==> sumCH...	C015 & CH...	sumCHEM...		4	C015 & C...	sumCHE...	1
2	45	1.20	14.38	3.63	B105 ==> sumCHEM3653	B105	sumCHEM...		6	B105	sumCHE...	1
2	45	1.20	54.22	27.74	MATH1823 ==> sumPHYS2514	MATH1823	sumPHYS...		5	MATH1823	sumPHYS...	1
2	43	1.15	30.71	12.61	B AD1001 ==> sumB AD1001	B AD1001	sumB AD1...		7	B AD1001	sumB AD1...	1
2	40	1.07	42.55	10.74	CHEM3152 ==> sumCHEM3653	CHEM3152	sumCHEM...		8	CHEM3152	sumCHE...	1
2	38	1.02	6.74	4.06	C015 ==> sumCHEM3053	C015	sumCHEM...		10	C015	sumCHE...	1
2	38	1.02	45.78	42.75	MATH1823 ==> sumPHYS2514 ...	MATH1823	sumPHYS...		9	MATH1823	sumPHYS...	1

Table 4. Sequence Rules Table

DATA PREPARATION, ANALYSIS AND SCORING

As with the association analysis, you begin with a transaction data set and link to the Association node, renamed “Sequence” (Figure 4).

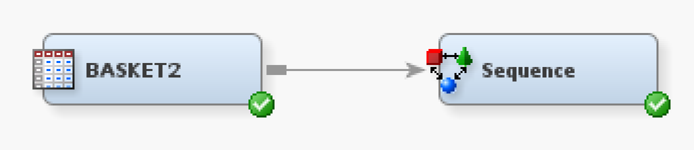
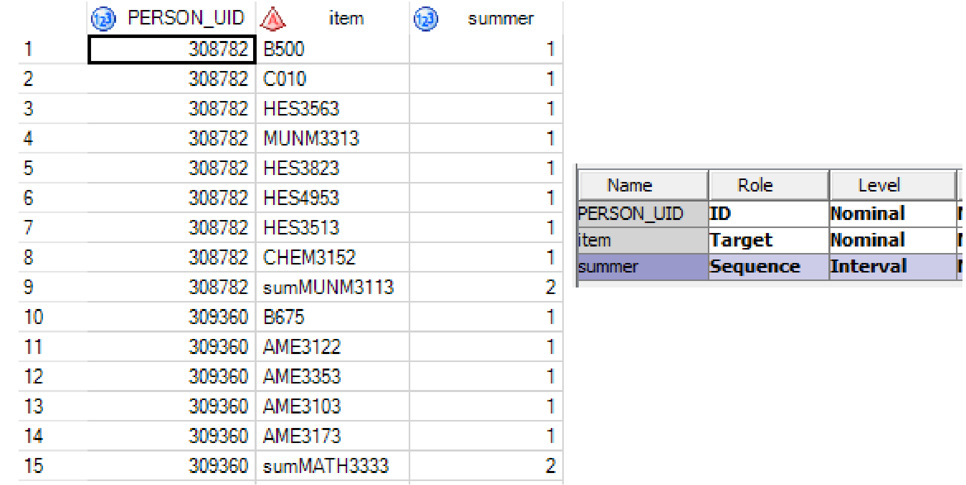


Figure 4. Sequence Analysis Diagram

In this case, your transaction data set includes a time, or sequence, variable with the role for this variable set to “Sequence” (Figure 5).

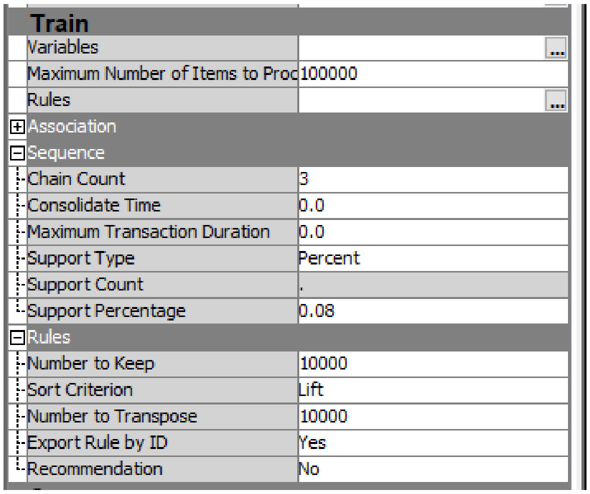


	PERSON_UID	item	summer
1	308782	B500	1
2	308782	C010	1
3	308782	HES3563	1
4	308782	MUNM3313	1
5	308782	HES3823	1
6	308782	HES4953	1
7	308782	HES3513	1
8	308782	CHEM3152	1
9	308782	sumMUNM3113	2
10	309360	B675	1
11	309360	AME3122	1
12	309360	AME3353	1
13	309360	AME3103	1
14	309360	AME3173	1
15	309360	sumMATH3333	2

Name	Role	Level
PERSON_UID	ID	Nominal
item	Target	Nominal
summer	Sequence	Interval

Figure 5. Transaction Data Set and Date Roles for Sequence Analysis

Our analysis parameters, once again more relaxed than default parameters, can be seen in Figure 6.



Train	
Variables	...
Maximum Number of Items to Proc	100000
Rules	...
⊞ Association	
⊟ Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duration	0.0
Support Type	Percent
Support Count	.
Support Percentage	0.08
⊟ Rules	
Number to Keep	10000
Sort Criterion	Lift
Number to Transpose	10000
Export Rule by ID	Yes
Recommendation	No

Figure 6. Sequence Analysis Parameters

Sequence analysis does not provide all the same strength of association metrics of the association analysis. However, Bao provides code for calculating expected confidence, lift and a z-score. A z-score here serves as an assessment of the significant difference between confidence and expected confidence, where a score greater than 2 indicates a significantly better confidence. This code can be introduced either in Enterprise Miner™ with a SAS Code node or by direct application to data in Enterprise Guide™. This method provides a good number of applicable rules without relaxing parameters as dramatically as the association analysis necessitated and requires less data cleansing before rules can be applied. As was the case with the application of rules in the association analysis, we use Enterprise Guide™ to access the assoc_links data set associated with the sequence analysis. The antecedent of the rule serves as our key for joining currently enrolled students with the consequent, or summer course recommendation.

CONSIDERATIONS

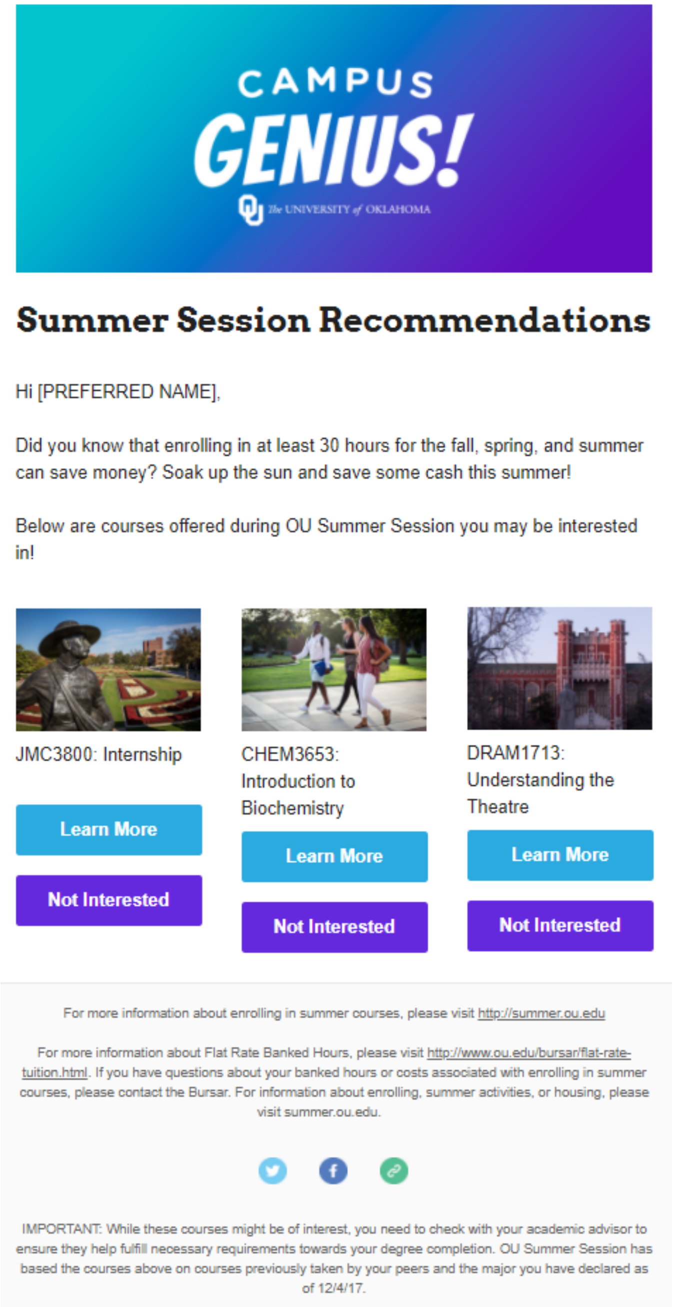
Although simple and applicable in many fields, there are a few things to consider when applying these analyses to course enrollment data. First, in most every case, the matrix of student to course is going to be quite sparse. And because those associations do not occur in high frequency, it is necessary to relax model parameters in order to identify them.

Second, recommendations are necessarily limited to more traditional lecture courses. The pre-requisite and permission structure of laboratories, discussions, independent studies, field studies, etc… is a labyrinth students master with their academic advisors/counselors and we defer to the professionals for these courses.

Third, recommendations are also limited, by nature of the analysis, to existing summer courses. More precisely, recommendations are limited to the courses of the previous summer session. If course selection from one summer to the next does not vary greatly, this becomes less of an issue. However, it is incumbent upon us to remain aware of course offerings and find ways to introduce variability back into the recommendations we make. We currently manage this with the general education interest course recommendation.

COMMUNICATION OF RECOMMENDATIONS

The Office of Business Analytics partners with Web Communications at OU to share the recommendations we build with students. Students currently receive notice of recommendations through email. Figure 7 is a sample of the communication planned for March, that includes our Campus Genius branding and an opportunity for students to provide direct feedback by selecting to view more information about each recommended course or letting us know they are not interested.



Summer Session Recommendations

Hi [PREFERRED NAME],

Did you know that enrolling in at least 30 hours for the fall, spring, and summer can save money? Soak up the sun and save some cash this summer!

Below are courses offered during OU Summer Session you may be interested in!

JMC3800: Internship

CHEM3653: Introduction to Biochemistry

DRAM1713: Understanding the Theatre

For more information about enrolling in summer courses, please visit http://summer.ou.edu

For more information about Flat Rate Banked Hours, please visit http://www.ou.edu/bursar/flat-rate-tuition.html. If you have questions about your banked hours or costs associated with enrolling in summer courses, please contact the Bursar. For information about enrolling, summer activities, or housing, please visit summer.ou.edu.

IMPORTANT: While these courses might be of interest, you need to check with your academic advisor to ensure they help fulfill necessary requirements towards your degree completion. OU Summer Session has based the courses above on courses previously taken by your peers and the major you have declared as of 12/4/17.

Figure 7. Communication with Students

Campus Genius will serve as the home, be it as an addition to our existing web portal or a downloadable web application, of our greater effort at a next best action recommendation engine. Here, University of Oklahoma students can access personalized recommendations for everything from course enrollment to membership in student organizations to calendar events and job opportunities.

CONCLUSION

Market basket analysis is a simple tool for identifying patterns in large sets of transaction data and is applicable across many fields. Here we have shown how relatively easily it is applied to student course enrollment, identifying existing enrollment patterns, from which we are able to fashion course enrollment recommendations for the coming summer term. This task becomes even more straightforward with the added capabilities of the sequence analysis.

Despite a few limitations regarding the scope of our recommendations, we have received quite favorable feedback from students. Additionally, the university has seen increased participation in summer session and growing summer session revenue.

REFERENCES

[1] Attewell, P. & Jang, S.H. 2013. “Summer Coursework and Completing College.” *Research in Higher Education Journal*, 20 (1): 117-41.

[2] Bao, X. 2007. “Mining Transaction/Order Data Using SAS® Enterprise Miner™ Association Node.” *Proceedings of the SAS® Global Forum 2007 Conference*. Orlando, FL: SAS Institute Inc. Available at http://www2.sas.com/proceedings/forum2007/132-2007.pdf.

[3] Faron, M. & Chakraborty, G. 2012. “Easily Add Significance Testing to Your Market Basket Analysis in SAS® Enterprise Miner™.” *Proceedings of SAS® Global Forum 2012 Conference*. Orlando, FL: SAS Institute Inc. Available at https://support.sas.com/resources/papers/proceedings12/204-2012.pdf.

[4] Kotsiantis, S. & Kanellopoulos, D. 2006. “Association Rules Mining: A Recent Review.” *GESTS International Transactions on Computer Science and Engineering*, 32 (1): 71-82.

[5] SAS Institute Inc. 2011. *Applied Analytics Using SAS® Enterprise Miner™ Course Notes*. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

We thank Glenn Hansen, Ph.D., Director of Office of Business Analytics, University of Oklahoma, for his initial efforts in summer course recommendations and continued and enthusiastic support of OBA.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Shawn Hall
University of Oklahoma, Office of Business Analytics, Norman OK
srsinger@ou.edu