Teaching Analytics: A Demonstration of Association Discovery with SAS Enterprise Miner

Full Papers

Amir Hassan Zadeh
Wright State University
Amir.zadeh@wright.edu

Shu Schiller
Wright State University
Shu.Schiller@wright.edu

Kevin Duffy
Wright State University
Kevin.Duffy@wright.edu

Abstract

In the current age of data analytics, there has been a push for the emergence of technologies that allow for interactive analysis of extensive amounts of quickly produced, highly varied data. These technologies require people (nicknamed “data scientists”) from many business disciplines who are capable of managing and analyzing this data for use in decision making processes. In order to educate and train more of these people, there has been an increase in the teaching of analytical tools in both Management Information Systems (MIS) and Business Analytics (BA) programs. This article will describe details of an exercise on business analytics specially tailored for the Introduction to MIS or BA course. The exercise utilizes a market basket analysis for a case scenario in which point-of-sale data will be used to compare different products. The goal of the exercise is to give students hands-on experience with association mining, network analysis, recommender systems and visualization in order to facilitate higher level learning of data analytics. SAS Enterprise Miner is a full-feature stand-alone data analytics platform that will be used by the students to perform the business-centric data analysis. The main goal for this project is to educate first year business students about the importance and usefulness of data analytics without discouraging them with excessive coverage of technical software details.

Keywords: Data Analytics, Business Analytics, SAS, Hands-on Training, Teaching

Introduction

In the past ten years, the analytics field has grown at an unprecedented rate. As a part of this progress, big data platforms such as Hadoop have come to the fore (Chambers & Dinsmore, 2014). Business data scientists are now integral to the field as they are capable of understanding the sheer velocity, volume, and variety of data these platforms must handle (Russom, 2011, Abbasi et al. 2015). However, a massive shortage of these scientists is expected, based on research from McKinsey’s big data report. The US is expected to face a deficit of between 140 to 190 thousand scientists trained to analyze the data and another 1.5 million managers with the necessary skills to produce actionable business intelligence (BI) used for making effective business decisions (Manyika et al., 2011). This trend has forced many vendors including SAS, IBM, Teradata, Hortonworks, Cloudera and others to fix their big data needs by creating scalable analytics architectures. Additionally, many academic programs have made data-analytics courses a priority in the curriculum (Guthrie, 2013; Wixom et al., 2014). Even with these new initiatives, education in regards to data analytics is still lagging behind, as it is missing from many current programs (Zhao et al., 2014). One of the major challenges for business professors is to overcome student apprehension in regards to the math, statistics and computer programing facets of analytics, as well as keeping them interested in the subject matter (Zanakis & Valenzi, 1997). To overcome this issue, replacing traditional teaching methods with a "hands-on" training approach has been thought to better engage students and educate them on the usefulness and benefits of learning analytics. By implementing hands-on training, students will achieve the desired interest and understanding of business analytics. Focus on the use of cutting-edge data technologies in the educational process will also assist with these teaching goals. By transitioning to this proposed style of learning, professors can better equip their students with the skills necessary to produce meaningful work in the field of data analytics (Sigman et al., 2014).

To assist with the transition towards a more practical hands-on approach to teaching BI/BA techniques and using these tools to validate business decisions, we have outlined an exercise that will can easily be incorporated into any introductory level Management Information Systems (MIS) or Business Analytics (BA) course. In the exercise, the students will explore data analytics using SAS Enterprise Miner (EM), a proficient interactive analytics software. This exercise will expose students to the field of data analytics and give them a visual and interactive experience using high-level software in order to extract valuable decision making information from large quantities of data. The SAS EM software combines best practices in data analytics and visualization, thus helping to engage and educate the students. This platform creates accurate and detailed predictive analytics by utilizing popular data mining and analytical processes to evaluate large amounts of data efficiently.

Learning Objectives and Case Description

This exercise is structured to use some of the most interesting computing "nodes" in SAS Enterprise Miner to help students understand both the business side of market basket application and how to analyze a transaction database efficiently to extract patterns (i.e. business rules) among existing customer transactions. Specifically, the exercise is designed in such a way that the students learn how to query, transform, analyze and visualize data from a customer transaction dataset. Fig. 1 displays the workflow for this exercise. It starts with data preparation/exploration and finishes with data analytics and visualization to acquire insight from the data. A typical market basket analysis scenario involves discovering patterns in customer purchasing behavior at a retail store where items are purchased together frequently. For instance, if a customer buys bread and milk, will she also buy peanut butter? We may learn that 10% of all transactions include bread, milk and peanut butter. We may also discover that of all the transactions that include bread and milk, peanut butter is present 70% of the time. Then, the association rule "bread and milk -> peanut butter" appears to be an interesting cross-selling rule for the target customer group.

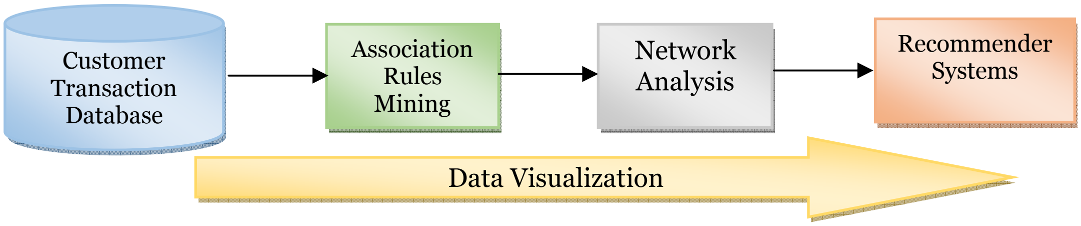


Figure 1: Exercise's Workflow

SAS EM includes a Sample library that contains a dataset named Assocs. Students are instructed to establish access to *Sampsio* library in order to bring the market basket data source to the SAS EM environment. Once the data is loaded into the SAS EM diagram, either Association Node or Market Basket Node is used for association and sequence analysis. Next, students are guided to use the SAS EM's network analysis tool, i.e. Link Analysis Node, to create various graphical representations of the association between the items. For example, the item constellation plot provides a co-purchasing graph that displays what products customers frequently purchased together. Finally, students use the recommendation systems tool embedded in the SAS EM in order to make personalized business recommendations to customers. They are required to figure out what product(s) would be recommended to a customer based on the purchase history of customers who share the same or who have similar product purchases.

In addition, the exercise implements the SAS's SEMMA analytics framework that guides students through the data mining initiative. SEMMA comprises of 5 stages: Sample, Explore, Modify, Model and Assess. The Sample phase helps users select the data source that is best-suited to meet a given business purpose. The Explore phase covers understanding the data by discovering relationships between the variables. The Modify phase includes methods for variable selection/transformation in preparation for data modeling. In fact, data preparation and preprocessing is the most crucial process of mining massive data sets and can take up to 80 percent of the time and effort involved in a project (Piramuthu, 2003; Zolbanin, Delen, & Zadeh, 2015). This process includes understanding the volume, variety and velocity of the data that must be processed (Douglas, 2001), understanding what the data represents, binding the data streams, transforming the data for storing in proper format for querying and analysis purposes, and filtering and reducing the data, among other steps. The Model phase includes various data modeling (data mining) techniques that can be applied on a prepared dataset in order to create predictive models that can provide the desired outcome. Finally, in the Assess phase, the usefulness and reliability of the candidate models are examined.

The case study addressed all of the course objectives (Table 1), but focused in particular on student objectives listed in Table 2. The complete set of course objectives is listed in Table 1.

Table 1: Course Objectives

Course Objective	Program Learning Goal
Understand the role of Information Systems in organizations	• Business Knowledge & Competency
Be able to use information systems as a resource in decision making	• Technology Skills
Understand the impact of technological change in accessing and disseminating information	• Business Knowledge & Competency
Learn how E-Commerce and E-Business have changed how we do business	• Business Knowledge & Competency
Be able to work with a database and data management tools	• Technology Skills • Analytics Skills
Perform business analytics tasks using Excel and other advanced software programs	• Technology Skills • Analytics Skills
Be able to apply analytic and computer-based techniques from science and business to analyze and interpret data	• Critical and Creative Thinking

Table 2: Key Case Objective

By the end of the case students will have:

- Gained experience working within a Data Analytics and Visualization environment;
- Applied current modeling and visualization techniques to the Market Basket Data;

By the end of the case, students will have learned how to complete all the following tasks:

- Create a data analytics project in SAS EM;
- Create a Library for use in SAS EM and Import the data source to the workbench diagram;
- Explore the customer transaction database;
- Work with data without going through complicated technical steps;
- Perform association rules mining to discover interesting patterns;
- Create visualizations of the data;
- Use link analysis node for detailed and precise insights about the co-purchase networks;
- Perform recommender systems analysis using the recommender table of SAS EM to recommend items to users;

This exercise guides students through a data analytics journey that enables them to understand different aspects of data analytics and motivates them to develop skills in this emerging area. The main objective of this hands-on exercise is to provide first-year business students a hands-on, interactive experience of multiple aspects of business analytics. In order to capture and retain students' attention and interest in the subject, the exercise is designed in such a way that students are exposed to advanced data analytics tools and techniques without delving too much into the mathematics side of the underlying algorithms.

We proceeded as follows: The instructors provided a hand-out that covered the concepts and the tools pertinent to the exercise and took the students through a step-by-step process to complete each task. Students were instructed in advance to peruse the tutorial (hand-out) and prepare prior to coming to class. During class time, we attempted to maximize higher level learning by devoting less time to introductory concepts, covering simple point-click software questions in the tutorial, and focusing instruction on more technical and practical sides of the exercise. Nevertheless, students were expected to demonstrate some level of proficiency in class. This method represented a more blended learning methodology in which an appropriate mix of face-to-face instructional methods and learning technologies were used to support planned learning and achieve subsequent learning outcomes (Lim & Morris, 2009). In the class time, the instructors spent between 35% and 70% of their class time presenting the business scenario followed by a live demonstration and a practical hands-on session. Through the hands-on class, students were supposed to replicate what have been done in the hand-out, so that they gained experience and confidence working in a data analytics environment. At this stage, instructions were explicitly provided to retain students' attention and interest. After thorough exposure to the learning environment through extensive explorations and visualizations, students were encouraged to go beyond what was covered in the hand-out, and develop their own ideas, questions and problem solving strategies applied to the market basket data. This exercise was designed to engage students in a challenging learning experience where they would become content developers.

The Exercise

To define the different phases in designing our data analytics exercise, the SAS SEMMA data mining methodology was used as a guideline. The data analytics methodology underlying this exercise consists of five major components: exposure/exploration, data preparation and preprocessing, data modeling and visualization, interpretation and elaboration that is consistent with the SEMMA framework. Fig. 2 displays the analytics workflow for this exercise. In the following, we will discuss each of the steps briefly.

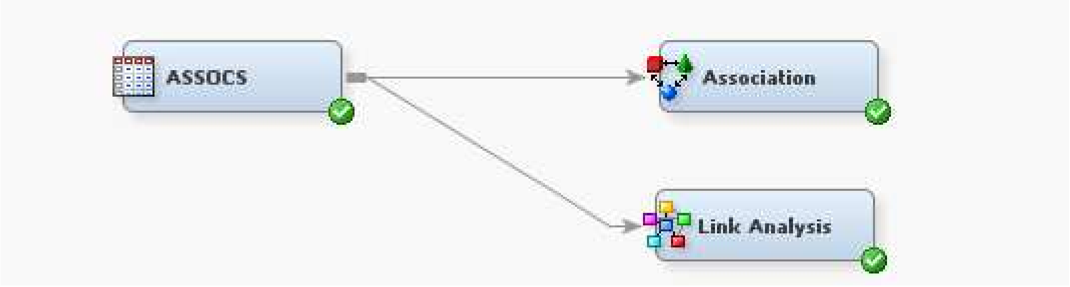


Figure 2: Analytics Workflow of the Exercise in SAS EM

Exposure, exploration and engagement

At the beginning of the exercise tutorial, the instructor should present a variety of data analytics and visualization applications (including market basket analysis) to introduce students to this unfamiliar content. Material for this activity can be found online from Tableau, SAS Visual Analytics, and/or in course textbooks. Current, interesting content related to the topic should be used to engage students. This phase of the exercise helps students learn about the importance and applications of business analytics. During this phase, students become familiar with various data analytics and visualization tools and more importantly become open-minded to using interesting data analytics tools other than Excel and Access.

Data preparation and preprocessing

During data preparation and preprocessing, students explore the potential uses of SAS EM as a powerful data analytics and visualization tool by completing and practicing a list of steps. These steps are outlined in a handout with associated screenshots. Students are asked to get their hands on different aspects of data as they perform each of the following steps: 1) create a new project in SAS EM, 2) create a workbench diagram, 3) select a data source, 4) drag and drop the associated nodes to the diagram, 5) configure the association and link analysis nodes, 6) customize the properties, 7) re-run the nodes to get customized results, and 8) preview the results. The steps can be completed all at once or in parts, as per the discretion of the instructor in accordance with the course time frame. In the next section, these steps will be discussed in detail.

In this phase, the instructor should emphasize that data preparation, data exploration, data understanding, and business-value creation constitute a significant amount of the whole knowledge discovery process in this exercise, like in most data mining initiatives. SAS EM includes a Sample library that contains a dataset named Assocs. In this dataset there are 1,001 customers who each purchased 7 items out of a possible 20 items. Students are supposed to establish access to the *Sampsio* library in order to bring the market basket data source into their SAS EM environment. By going through the data source wizard of the SAS EM, students will be introduced to the definitions and language of data mining. The wizard helps them create a data table and metadata of the dataset in order to issue data definition language (DDL) for all columns in the dataset. As a part of this phase, students will use the data source node to do a preliminary investigation to understand the structure of the data and its content. In addition, the dataset role must be set to TRANSACTION in the data preparation and preprocessing steps in order to perform association analysis.

Data Modeling, Visualization and Interpretation

In this exercise, we defined three approaches for use in our analytics initiative: association rules mining, network analysis, and recommender systems. The technology behind these analytics powerhouses has evolved over the past 20 years into a rich collection of tools that enable practitioners to perform effective analytics without having to build the entire ship from scratch.

Visualizing data by converting data and information into graphical representation is indispensable for discovering interesting patterns in the customer transaction dataset. SAS EM provides an extensive palette of visualizations, capable of supplying the appropriate graphics over data.

First, students use the Association Node to generate business rules that help to identify interesting patterns in the given customer transaction dataset. The rules are in the form of A -> B, where A is named the precedent, and B is named the consequent. Typically these association rules are produced by counting number of instances where both A and B are present in the data (Faron & Chakraborty, 2012). Rules may contain more than two items. Examples of rules are apple -> avocado (2 items), bread & milk -> peanut butter & butter (4 items). In the Association Node property panel, students can configure to customize the rule search in order to manage the rule results. The maximum items property, the minimum confidence and the minimum support for the rules need to be configured before running the association mining node. These three metrics help students identify important rules. Rules with highest confidence and support percentages are recommended. These rules can be identified in the top right corner of the statistics plot (Figure 3).

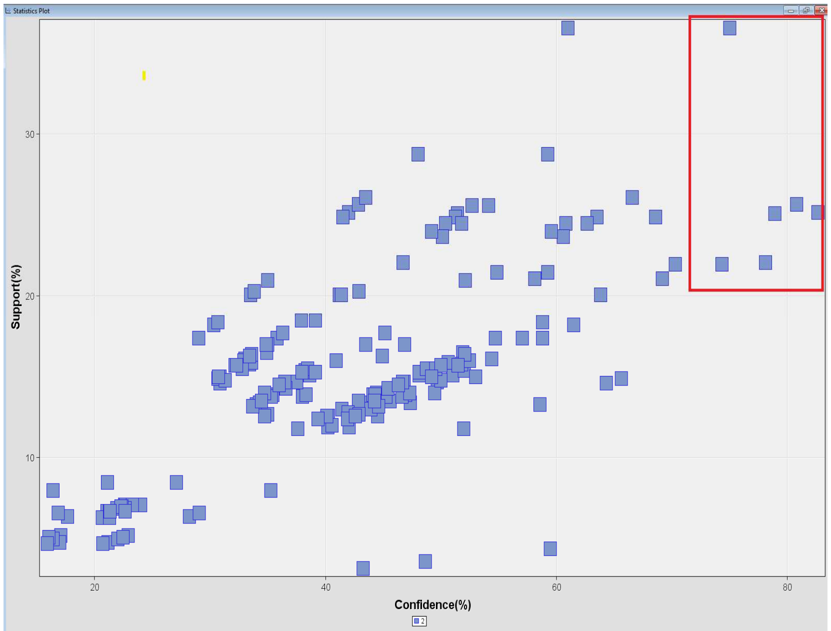
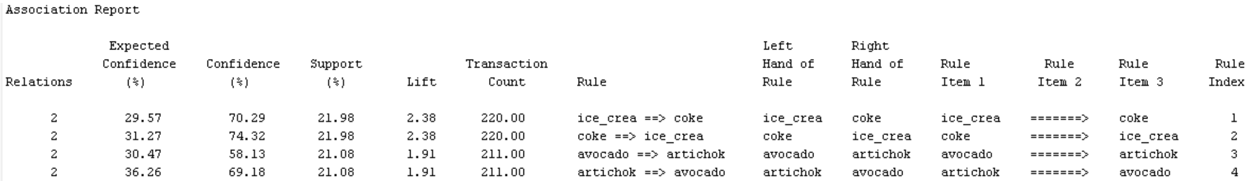


Figure 3- Statistics plot for Association Node

The Maximum Items property for the rule refers to the number of items to consider for rule generation. The Support for the rule is the joint probability that both items are in a basket, answering the question "what percent of baskets contain A and B?" The Confidence for the rule is the conditional probability that B is in the basket given that A is present. These properties are used to constrain the potential number of rules generated by screening out those that don't meet minimum benchmarks (Faron & Chakraborty, 2012). For example, as shown in Fig 4, the rule ice cream -> coke has a confidence of 70%, and a support percentage of 22%. It means that 22% percent of baskets contain ice cream and coke. It also asserts that, given ice cream in the basket, there is at least 70% confidence that coke is also purchased and comes to the basket.



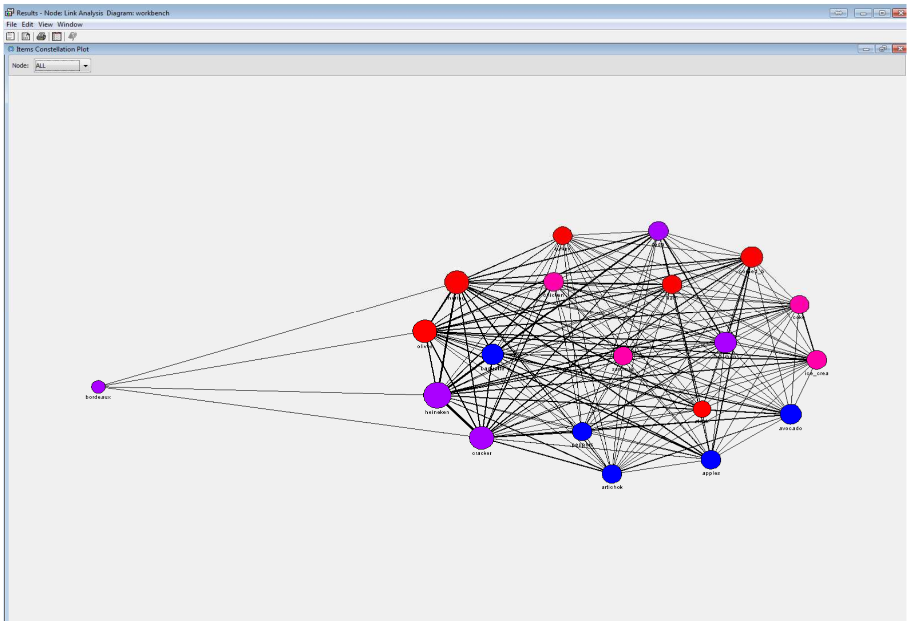
```
Association Report

           Expected                                                                    Left      Right
           Confidence  Confidence  Support          Transaction                        Hand of   Hand of   Rule      Rule      Rule      Rule
Relations    (%)         (%)         (%)      Lift     Count     Rule                  Rule      Rule      Item 1    Item 2    Item 3    Index

     2       29.57       70.29       21.98    2.38    220.00     ice_crea ==> coke     ice_crea  coke      ice_crea  =======>  coke          1
     2       31.27       74.32       21.98    2.38    220.00     coke ==> ice_crea     coke      ice_crea  coke      =======>  ice_crea      2
     2       30.47       58.13       21.08    1.91    211.00     avocado ==> artichok  avocado   artichok  avocado   =======>  artichok      3
     2       36.26       69.18       21.08    1.91    211.00     artichok ==> avocado  artichok  avocado   artichok  =======>  avocado       4
```

Figure 4- Output window

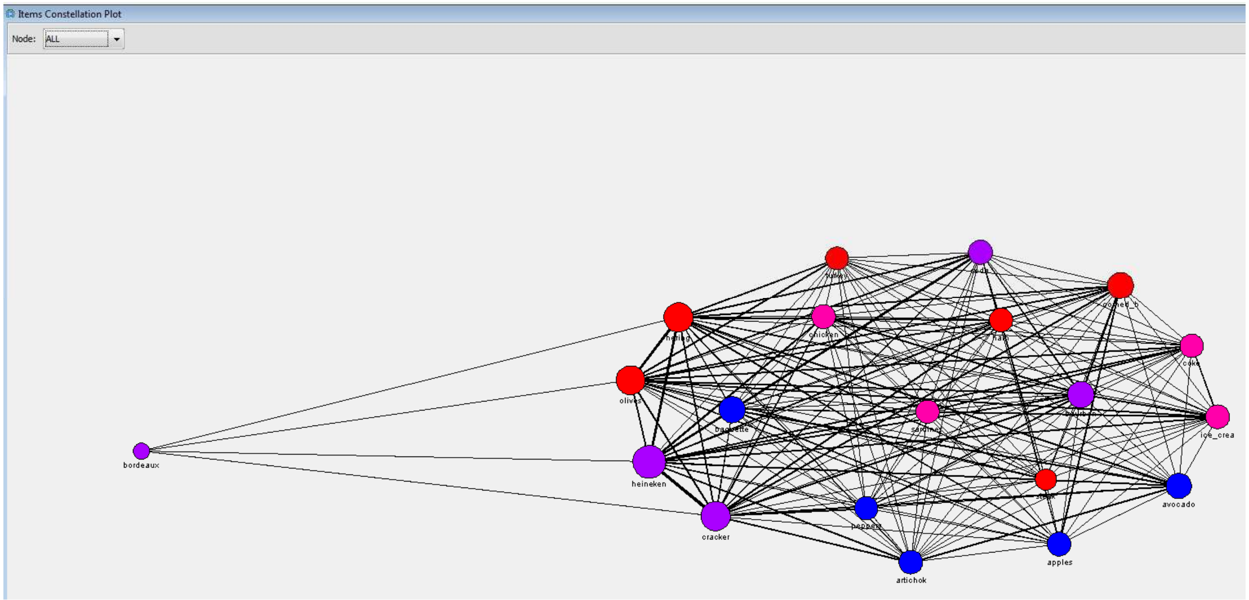
Additionally, this exercise exposes students to network analysis which is a powerful tool used to describe a system of interactions and interconnectedness between groups of objects. SAS's Link Analysis tool covers various graphical representations of the association between two or more different items. Students use the Link Analysis Node in SAS EM to further investigate the associations between the items. For example, the item constellation plot (Figure 5) can be used to display a graph representing association between all items.



Node: ALL

Figure 5- Item constellation plot

The item constellation plot can be also set to display a co-purchasing graph for a particular item. For example, Figure 6 shows a graph that represents all the items that are co-purchased with Apple. Note that the thicker the line is the stronger the association rule is.



Node: ALL

Figure 6- Item constellation plot: a co-purchasing graph view

Students use the Exploratory Plot (Figure 7) for transactional data to gain insight about each individual customer's behavior, as opposed to collective learning from all customers. For example, when students select customer from the Group list and Customer ID 10 from the Node list, they can see what different products Customer # 10 has bought. The thickness of a link represents the number of times customer # 10 has purchased that item.

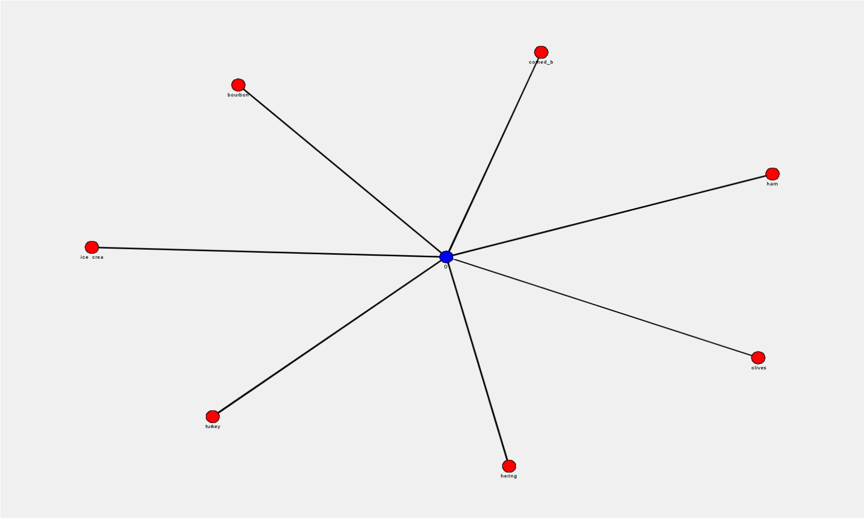
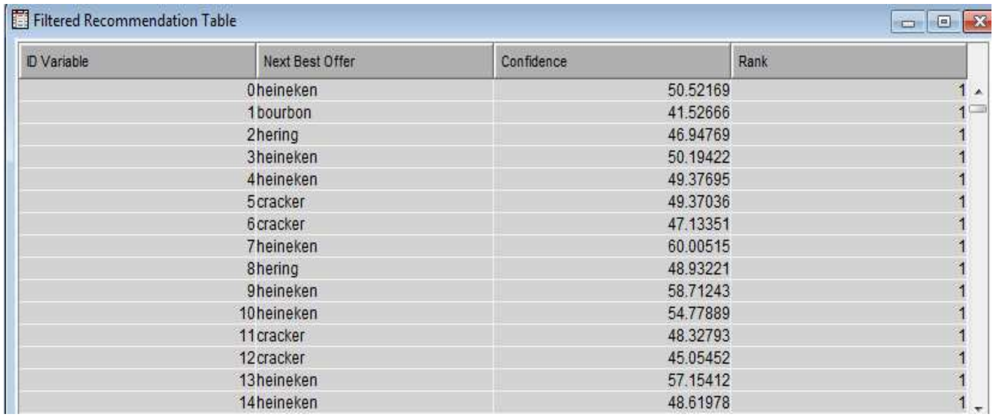


Figure 7- Exploratory Plot for individual customer's purchasing behaviors

Finally, in the last step, students are introduced to recommender systems and their applications. Recommender systems such as Amazon, Netflix, Last.fm, and Facebook have changed the way people find products, information, and even other people. The Link Analysis node in SAS Enterprise Miner can make recommendations to customers based on their transactional history data by providing a next-best-offer list. This recommendation table will tell students what product(s) should be recommended to a customer based on the purchase history of customers who share the same or who have similar product purchases. As shown in Figure 8, for example, the item "hering" [sic] should be recommended to customer #2.

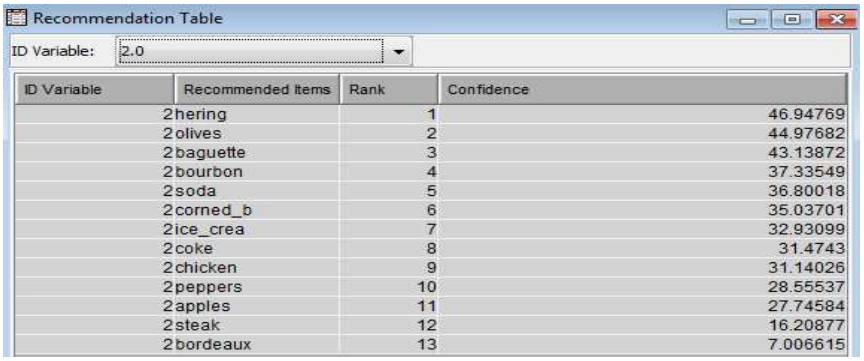


Filtered Recommendation Table

ID Variable	Next Best Offer	Confidence	Rank
0	heineken	50.52169	1
1	bourbon	41.52666	1
2	hering	46.94769	1
3	heineken	50.19422	1
4	heineken	49.37695	1
5	cracker	49.37036	1
6	cracker	47.13351	1
7	heineken	60.00515	1
8	hering	48.93221	1
9	heineken	58.71243	1
10	heineken	54.77889	1
11	cracker	48.32793	1
12	cracker	45.05452	1
13	heineken	57.15412	1
14	heineken	48.61978	1

Figure 8- Recommendation table in SAS EM

Additionally, students can create a filtered recommendation table that results from specifying criteria such as "Customer ID," "Top N" and "Minimum Confidence (%)". For example, Figure 9 shows the next-best-offer list for customer ID 2.



Recommendation Table

ID Variable: 2.0

ID Variable	Recommended Items	Rank	Confidence
2	hering	1	46.94769
2	olives	2	44.97682
2	baguette	3	43.13872
2	bourbon	4	37.33549
2	soda	5	36.80018
2	corned_b	6	35.03701
2	ice_crea	7	32.93099
2	coke	8	31.4743
2	chicken	9	31.14026
2	peppers	10	28.55537
2	apples	11	27.74584
2	steak	12	16.20877
2	bordeaux	13	7.006615

Figure 9- Next-best-offer list in SAS EM

Elaboration

After enough exposure to the SAS EM environment, students should be encouraged to go beyond what is done in the hand-out and take initiative in developing their own ideas, questions and problem-solving strategies applied to the market basket data. This process will keep students actively engaged in the learning process through a trial-and-error experimentation (Chen, Chiang, & Storey, 2012; Dupin-Bryant & Olsen, 2014). They will begin to find creative ways of interpreting the information derived from the tables and figures, which will help them develop a deeper understanding of data analytics. Also, the instructor should ask the students to share any ideas for how they could apply association mining or market basket analysis to other areas such as healthcare, sports and social media. This post-exposure discussion should nourish student curiosity and connect analytics and data mining into other aspects of their life.

Conclusion

The most significant advantage of advanced data analytics tools such as SAS EM is its ability to demonstrate and explain processes of data analytics and visualizations in multiple ways for maximized understanding by business students. With complementary instructions in statistics and computer use, students can explore the potential of data analytics to link science and business processes. The demonstrations presented in this article are designed to fulfill three main learning objectives: (1) improve students' understanding of business analytics and how it can be applied in practice, 2) perform data analytics with a new tool without exhausting students with unnecessary technical details, and 3) encourage students to learn more about business analytics in their subsequent courses. This exercise was conducted in multiple sections of MIS 3000 in spring 2016 at Wright State University. MIS 3000 (Fundamentals of Information Systems) is a core business course, required of all business majors. The assessment process was voluntary and anonymous; the response rate was 83% of enrollment (139 out of 168). Table 3 summarizes student assessments of the demonstration' effectiveness.

#	Question	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
1	Demonstrated how to preform market basket analysis using point-and-click and graphical data model to perform data analytics	%91.8	%6.1	%2.1		
2	Improved my understanding of market basket analysis using the association rule mining without delving too much into the mathematics of it	%89.3	%9.1	%1.6		
3	Demonstrated how visualization can help gain business insights and grasp trends/patterns in data quickly	%95.6	%4.4			
4	The goals of this exercise are clearly stated and consistently pursued.	%93.5	%6.5			

5	Improved my understanding of business analytics and how it can be applied in practice	%89.8	%8.4	%1.8		
6	Helped me understand what business analytics is about and what it can do	%93.8	%4.1	%2.1		
7	Was reasonable and useful	%95.4	%4.6			
8	The step-by-step handout helped me go through the exercise	%96.8	%3.2			
9	The steps described in the handout were not working					%100
10	This exercise was irrelevant for this course				%4.8	%95.2
11	I gained no new knowledge from this exercise				%3.3	%96.7
12	I like to learn more about business analytics	%95.3	%3.1	%1.6		
13	I like to do more hands-on analytics exercises like the basket analysis	%96.5	%2.2	%1.3		

The results of the students' responses to the hands-on demonstration suggest that the exercise engages students in a manner that increases their knowledge of data analytics and visualizations. After doing the exercise, students also expressed more confidence in understanding of the data analytics and curiosity to learn more. Overall the students appreciated that the ability to create knowledge from data can provide comparative advantages for those who gain early proficiency regarding this crucial technology.

REFERENCES

Abbasi, A., Surker, S. & Chiang, R. (2015). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*.

Chambers, M., & Dinsmore, T. (2014). *Advanced Analytics Methodologies: Driving Business Value with Big Data*: Pearson Education.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly, 36*(4), 1165-1188.

Douglas, L. (2001). 3d data management: Controlling data volume, velocity and variety. *Gartner. Retrieved, 6*.

Dupin-Bryant, P. A., & Olsen, D. H. (2014). Business Intelligence, Analytics And Data Visualization: A Heat Map Project Tutorial. *International Journal of Management & Information Systems (Online), 18*(3), 185.

Faron, M., & Chakraborty, G. (2012). *Easily Add Significance Testing to your market Basket Analysis in SAS Enterprise Miner*. Paper presented at the SASGlobal Forum: Paper.

Guthrie, D. (2013). The coming Big Data education revolution. *US News, 15*(08), 2013.

Lim, D. H., & Morris, M. L. (2009). Learner and Instructional Factors Influencing Learning Outcomes within a Blended Learning Environment. *Educational Technology & Society, 12*(4), 282-293.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., . . . Institute, M. G. (2011). Big data: The next frontier for innovation, competition, and productivity.

Piramuthu, S. (2003). On learning to predict web traffic. *Decision Support Systems, 35*(2), 213-229.

Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter*.

Sigman, B. P., Garr, W., Pongsajapan, R., Selvanadin, M., Bolling, K., & Marsh, G. (2014). Teaching big data: Experiences, lessons learned, and future directions. *decision line*.

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., . . . Turetken, O. (2014). The Current State of Business Intelligence in Academia: The Arrival of Big Data. *Communications of the Association for Information Systems, 34*(1), 1.

Zanakis, S. H., & Valenzi, E. R. (1997). Student anxiety and attitudes in business statistics. *Journal of Education for Business, 73*(1), 10-16.

Zhao, T., Qian, K., Lo, D., Guo, M., Bhattacharya, P., Chen, W., & Qian, Y. (2014). *Problem Solving Hands-on Labware for Teaching Big Data Cybersecurity Analysis*. Paper presented at the Proceedings of the World Congress on Engineering and Computer Science.

Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems, 74*, 150-161.