Benefits of Using Real Data Sets to Instruct Business Students in Data Mining Techniques

Dr. Kathleen Campbell Garwood

Abstract:

When presented in a business class setting, the concepts and practices of data mining can often present complex challenges as time considerations and course sequencing prevent full discussions of prevailing mathematical principles lying behind each analysis. Students struggle with appropriate analysis choice. Over the course of four semesters, data has been collected on multiple classes considering the use of real, messy, and varying data sets which immerse students in hands-on problem solving. In the first iteration one "real" data set was given and three contrived. Each semester moving forward an additional "real" data set replaced a current existing textbook data set. Generally, each student team uses a different approach. Yet, when applied properly (with logic and fundamental statistical assumptions), similar results are obtained. The number of clear, actionable conclusions given with understandable visualization added measurable value. The study initiated using schooling survey data from Bolivia, allowing students to use data mining techniques to identify the most impoverished students in high schools in different regions of the country. Additional data sources were used, including entrepreneurial Amazon inventory data with branding and marketing issues; an electric and light company optimizing fan displays across multiple venues while minimizing inventory; and also a supply chain company looking to get information from large troves of available data. Immersing the students in real-world problems and encouraging them to clean and understand the data before applying data mining tools and visualization allows them to become confident problem solvers, a challenge faced in traditional testing. Students clearly value these initiatives.

1. Introduction

Statistics is at the heart of multiple disciplines. Though the name seems to be continually evolving (e.g., decision and system sciences to business intelligence and analytics), students are still expected to have a basis in understanding the data that is in front of them (the shape, spread, and distribution) and the questions that should be asked in order to apply the best analysis technique possible. The idea of using real data is not new. The modification is that besides being real, the data needs measurable cleaning and organizing, and is not internet-accessible. Meanwhile, the students are often on a very limited timetable of either one hour or 2 days depending on the depth of the data and the presentations expected.

Moore and Roberts (1989), initiated work on the topic of the merits of teaching "data-driven" courses where the goal was to lead the class to identify questions and analyses that arise from an interesting set of data in hopes of stimulating the discussion and development of statistical techniques. The examples from this paper are discussed at length by Lock (1990). However a limit of the 1990's was that with one computer at the front of the room, the class needed to work together to guide the conclusions drawn from

the data. Current technological advances, such as laptops, have advanced the concept of using real data. Case studies have become the new textbook. According to Salmons (2014), case studies offer teachers and their students the opportunity to reflect critically on the conduct of an actual research project, and to ask questions regarding how such research might have been designed differently or what worked well or less well in a particular research scenario. But, like a textbook problem, the data is often available online and in some cases helpful hints or solutions are available as well. It is for this reason that the data provided in this course change with each derivation and are supplied by various alumni and network sources with no preconditioned answers. This allows the student groups to experience a more realistic work scenario.

2. Technique being studied versus the traditional method

In this study, the data collected on the students considers the output and effective identification of critical information by each student team. In a time series fashion it looks to identify changes in student outcomes as they work with real world data manipulation and analysis within the classroom setting over the course of one semester. Students are challenged to question, clean, organize, visualize and interpret data within a very short timeframe and need to be prepared to present interim conclusions immediately. Class performance, number of clear, clean graphics, and presentation skills are the metric being used to assess if any changes in class expectations occurred by adding this type of learning module into a traditional Introduction to Data Mining class.

Traditionally, students in analytics are introduced to a variety of complex topics and ideas. As such, a typical teaching style follows a sequence: class discussion and practice followed by homework assignments and the use of larger team projects followed by an in-class presentation (which takes up significant class time). This style works well for topics the students should be memorizing or using on a daily basis. This method allows for students with full notebooks and good memorization to do well and readily give back what the teacher has specifically shared with them during class, in assignments, and through readings. It has purpose but it will not create useful analysts who can think on their feet and realize the challenges of real world messy data. Specifically, the methodology described herein seeks to develop analytical creative thinking and the ability to draw connections among different data points in a short timeframe.

3. Project Methodology

This class utilizes the traditional model for 70% of the classroom time. However, for the remaining 30% of class time, the students are given raw data sets and expected to make clean and clear reports in about an hour. The data, with missing variables and often large (thousands of rows with many columns) challenges the students to use not only what they learn in class, but also what they have gained through other experiences (internships, other classes, readings) to try to find meaning within the data. Each data source is generally obtained from either a not-for-profit organization or SJU alumni. The students are often given a short list suggesting feasible directions to start. Like real data, the list often asks for things that the provided data cannot truly answer, leaving the students to be

creative or accept that they can only get so far. Five to eight minutes at the end of each class are used for the professor to randomly call on two groups to present their findings. Often, the person who provided the data is available on campus or via Skype to answer any initial data questions the students might have once they dive into the project, and also at the end of the class to view and comment on the presentations. All final PowerPoint materials submitted by students are shared with the source of the data. One or two larger projects allow the students a full class to work on the data (as well as two days outside the classroom). In these instances all of the groups would then present their findings.

4. Fe y Alegria:Bolivia, an example

Fe y Alegria:Bolivia (FyA:B) is a Jesuit-sponsored organization present throughout Latin America, that started in Bolivia in 1966 and currently runs over 400 schools and educational centers in that country. The motto and belief system within FyA:B is that their work begins where the pavement ends as they serve the most impoverished students wherever they are present. This particular project's objective is to identify the most disadvantaged students within a given school to help provide extra training before graduation, specifically with job-related skills, to allow them to be placed upon graduation and help break the poverty cycle. However, with no access to income information and rare interaction with parents and families once students reach middle school – assessing which students could most benefit from help becomes complex. FyA:B obtains student surveys yearly to help each school collect a variety of information. Though the purpose of the surveys is broad educational data collection, a local pioneer in education, Miguel Marca, wondered if within these surveys there might be clues as to which students are most impoverished. For this two-part in-class project, the data mining students work as consultants over a month-long period.

PHASE 1: The survey results are shared with the Saint Joseph's students and they immediately initiate a data investigation with a Skype question-and-answer session (with translator). This initial session addresses any data integrity issues, clears up translations of specific questions, resolves any ambiguity that might exist in answer scale used, and examines any other initial data cleaning issues the class might find. In one class period, the students work through the initial analysis to come up with insights as to the survey output and to identify which variables might be pertinent to pinpointing the most impoverished students in the surveyed population. The students also engage in ANOVA analyses that might give immediate insights as to differences in responses across grade, gender, or school. The students only work for one hour before submitting initial analysis.

PHASE 2: Over the ensuing two weeks, the students increase their familiarity with the data and, in parallel, continue learning advanced data mining techniques. The data is parsed and each student is given only a small sample of the survey respondents to manipulate through several techniques as a homework assignment. This step familiarizes each student with the survey while also forcing them to recognize that different conclusions can be found among different schools or even among different classes within a school (a quick lesson in sampling). During these two weeks, the class covers a variety

of topics including: Analysis of Variance (ANOVA), data reduction in the form of Factor Analysis\Principle Component Analysis (FA\PCA) and provides initial measure of realistic student groupings in cluster analysis.

PHASE 3: At the conclusion of the two-week time period, the students are more familiar with the survey questions as well as the locations of the schools being studied, including a better understanding of conditions in Bolivia. The class works to come to an agreement as to how the data should be cleaned and organized so that when the project is next administered, everyone works with the same cleaned data set. Students have one hour to work through several of the techniques in class, where they can ask more questions of the professor or of the facilitator of the data (Miguel Marca). They then have two days outside of class to clean up and prepare presentations explaining how each group has gone about finding the most impoverished students from the survey. In the next class period the students present live via Skype to Miguel Marca (the FyA:B coordinator) as well as other FyA:B facilitators. The student consultants provide meaningful insights as to who has been identified as the most impoverished. A basic outline of the data and deliverables reached can be seen in the following figures. Figure 1 is a sample of the Bolivian data set as presented to the students on the first day. Notice the lack of detail of the questions (as seen in the names of the column headers) and the missing data in some of the collected survey fields.

Data Example: Bolivia

School	Class	SEXO	NIVEL	Q1	Q2	Q3	Q4
SAGRADA FAMILIA	4to A	HOMBRE	CUARTO	3	4	2	2
JOSE MARIA VELAZ	4to B	MUJER	CUARTO	3	4	3	1
LUIS ESPINAL CAMPS	5to C	HOMBRE	QUINTO	4	3	3	1
LUIS ESPINAL CAMPS	5to C	MUJER	QUINTO	3	2	2	2
JOSE MARIA VELAZ	5to A	MUJER	QUINTO	3	1	3	2
JOSE MARIA VELAZ	5to B	HOMBRE	QUINTO	2	2		
FRAY VICENTE BERNEDO B	4to B	MUJER	CUARTO	2	2	3	1
FRAY VICENTE BERNEDO B	5to A	HOMBRE	QUINTO	2	2	2	2
FRAY VICENTE BERNEDO B	5to A	HOMBRE	QUINTO	2	4	3	1
FRAY VICENTE BERNEDO B	5to A	MUJER	QUINTO	1	1	1	1
LOYOLA DE FE Y ALEGRIA B	4to A	HOMBRE	CUARTO	3	3	4	1

Figure 1: Example of the initial uncleaned data provided to the students.

Figure 2 is an example of the use of the cluster analysis technique which is one of the methods used by many teams to establish a parsing of the impoverished students into groups that may be considered more or less impoverished by features identified in the survey. While the cluster analysis may be properly done, the visual and output are overwhelming and contain jargon and statistical data tables but no clear concise insights. Figure 2 comes from the first interaction of this experiment. The reality is that it is impossible for someone who has not been immersed in the data to get any sort of reasonable or logical conclusion from the graphic. Possibly an acceptable starting point, this should not be a conclusion graphic as it is unreasonable to expect most viewers to follow said conclusions.

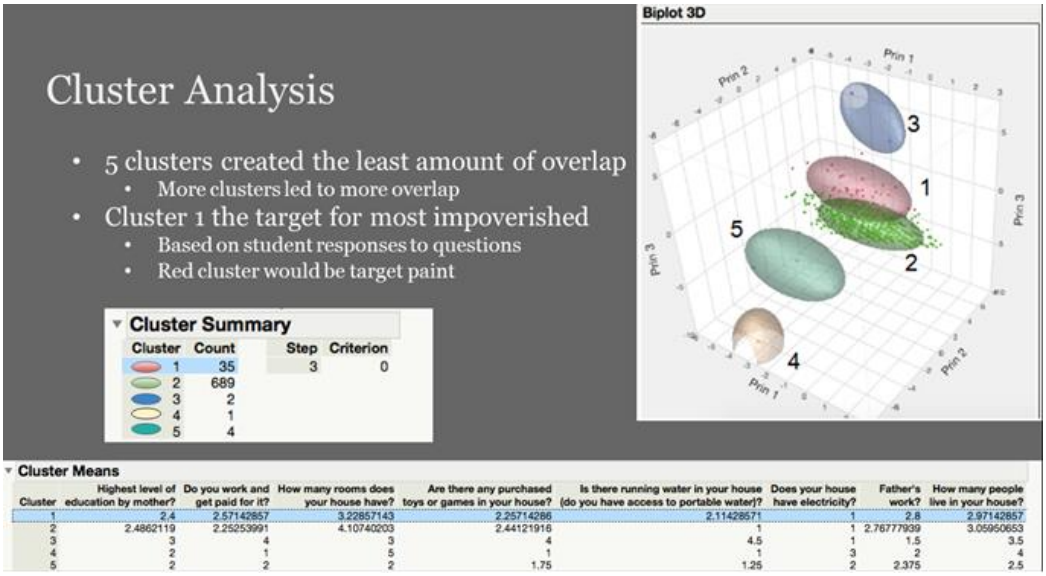

Figure 2: Typical data mining output. Uninterpretable by layman and unclear.

Meanwhile, Figure 3, taken from the third iteration, shows how students have advanced their skills in interpretation and discussion. NOTE: Each iteration looked at new surveys recently collected in Bolivia. In Figure 3, there is evidence that students have begun to recognize and share the important findings from the cluster analysis. They used the cluster analysis as a jump-off point but then went through the data to identify which schools more of the impoverished students came from in order to isolate a location for larger pockets of identifiable poverty. This type of graphic became more prevalent in the second iteration of collection and analysis.

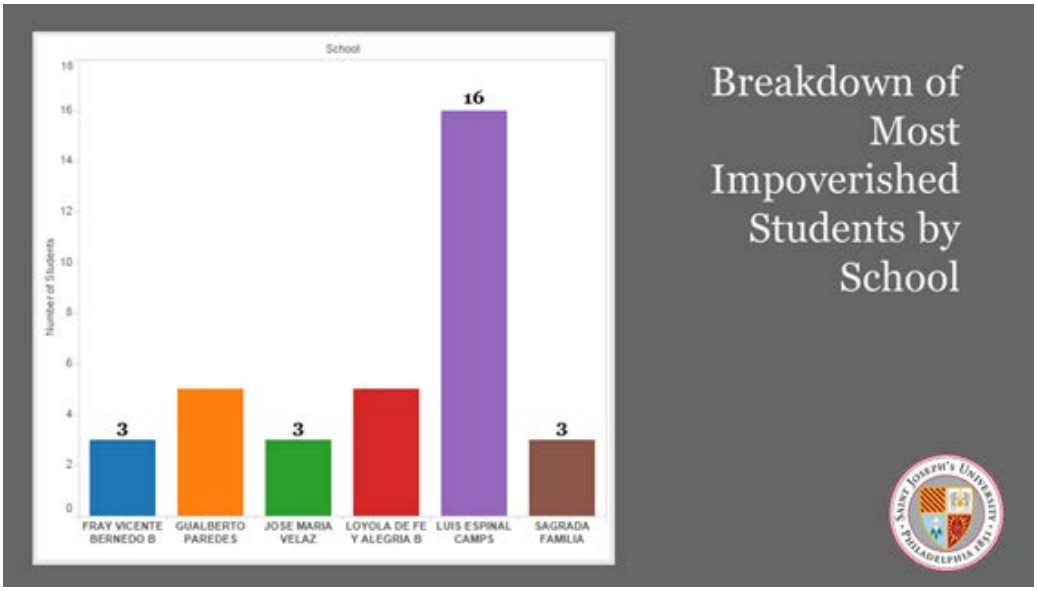

Figure 3: Clearer results identifying the schools in which most impoverished students were identified.

Figure 4 looks at a graph provided in the third iteration of the project. Please note that the student names (which were provided to Miguel) have been coded for the student's privacy. This graphic as well as similar visuals that drilled down into the data allow for identification of both the neediest students in a specific school as well as anomalies within findings, and provide the end user with insights and input for moving forward in identifying the most impoverished students.

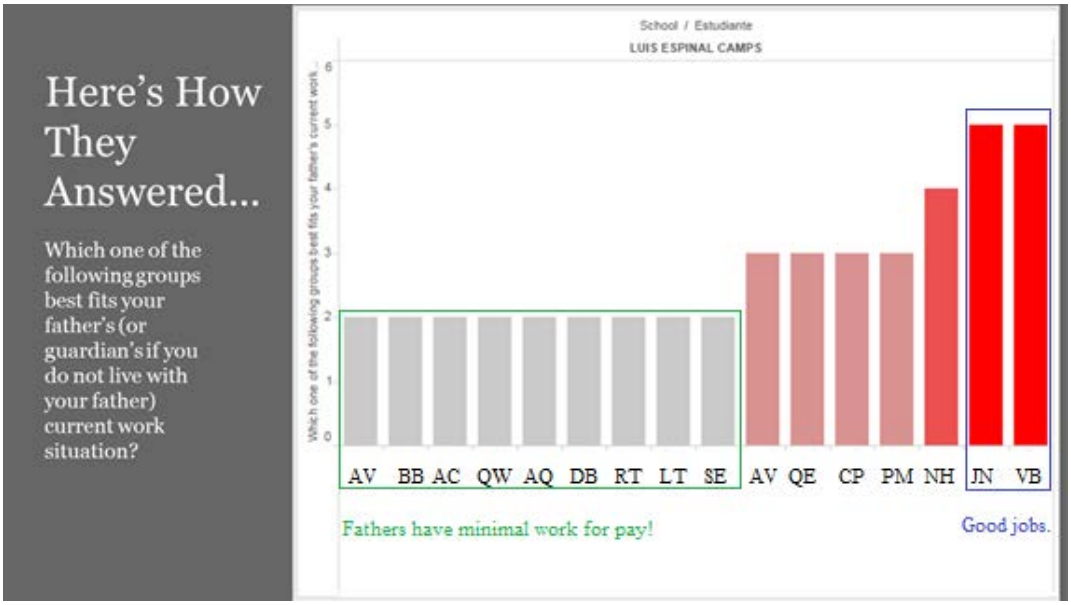

Figure 4: In-depth and logically drilled down results have become commonplace.

In the final visual, Figure 5 highlights how many usable visuals were provided by each group by semester. The grey scale blocks represent the project output during the first half of the semester while the red boxes consider the project output during the second half of the semester. There is evidence of growth and strength in students' ability to quickly analyze, visualize, and clearly interpret data results as the projects have evolved from Fall 2015 through Spring 2017 (see Figure 5).

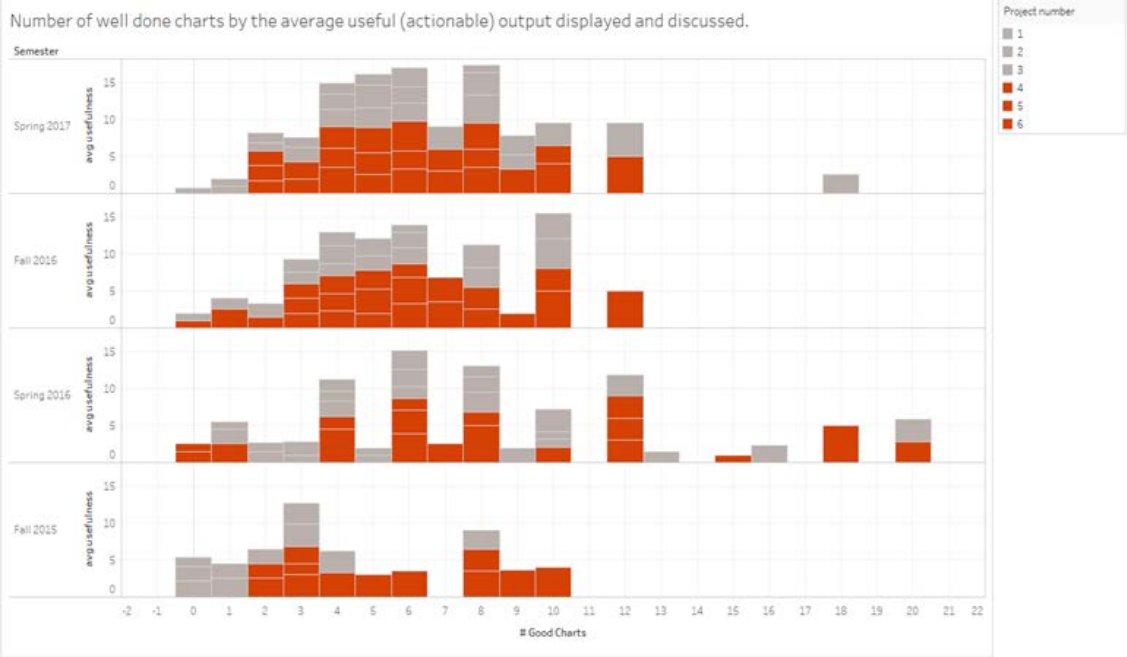

Figure 5: Increase in the number of meaningful graphs by semester.

5. Assessment of learning outcomes and conclusion

After four semesters of using projects within the semester from a variety of sources the following issues have been addressed:

- Students are encouraged to get out of the "bubble" - life isn't perfect and neither is data. Students learn not to rely solely on basic analysis, i.e., to remember assumptions, the techniques of data cleaning, and the importance of understanding the data and the questions. Students are encouraged to ask questions because the perspective of those they are working for is usually very different than their own.

- Students are challenged to use their own skills (often liking math) for the good of others and not just to prepare for post-graduation job.

- Students look deeper into data without direction or guidance and again ask questions. They have shown continued growth in trusting their own creativity to see and share stories hidden within the data.

- Students develop the ability to work with varying sizes of data, decide the best practices to analyze, interpret, and predict with limited time.

Teaching analytics of any form requires faculty to dive into a variety of statistical tests and topics that need both discussion and practice. Student ability to create, understand and read output is a continued challenge. However, as the field of data science grows, students also need to be able to explain their insights, to find patterns, and to share insights non-technically within a short timetable. This combination allows for both

methods to be applied while allowing for topics like data cleaning, data organization, and outlier treatment to be studied in practice throughout the semester.

Bibliography

Lock, R H (1990) Some Favourite Data Sets : Using the Computer on Real Data in Class. Web accessed (June, 2017) DOI: https://iase-web.org/documents/papers/icots3/BOOK2/B3-14.pdf

Moore, T L and Roberts. R A (1989) Statistics at liberal arts colleges. The American Statistician 40, 80-85.

Salmons J (2014) How to Use Cases in Research Methods Teaching: An Author and Editor's View. Sage Research Methods Cases. Web accessed (June, 2017) DOI: http://dx.doi.org/10.4135/978144627305014534935