Teaching Data Mining in the Business School: Experience from Three Continents

Galit Shmuéli GALIT_SHMUELI@ISB.EDU
Indian School of Business, Gachibowli, Hyderabad, 500 032, India and
Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, USA

Carlos Soares CSOARES@FEP.UP.PT
INESC TEC/Faculdade de Economia/EGP-UPBS, Universidade do Porto, Portugal

Abstract

The number of data mining courses in Business Schools is growing. Due to the differences in the profiles and interests of business school students compared to students in computer science departments, data mining courses must be organized and delivered in different ways. Based on experience of teaching data mining courses at business schools in India, Portugal and USA, we discuss those differences in this paper.

1. Introduction

In the last decade there has been an increasing integration of data mining (DM) and predictive analytics in the business world, often referred to as Business Analytics (Linoff & Berry, 2011; Shmueli et al., 2010; Soares & Ghani, 2010). This evolution has spurred the need for business professionals who have an understanding of DM. To address this need, many reputable business schools (b-schools) worldwide have introduced data mining electives into the MBA and undergraduate programs, and recently there has even been a trend to create a Business Analytics program or specialization (examples include NYU and University of Connecticut among others). The demand for data mining courses in a business context has created a tremendous opportunity but also a complex challenge for instructors. While some instructors are knowledgeable in statistics, operations research or econometrics, they often lack knowledge in data mining. Another big challenge is the diverse and usually non-technical audience. MBA students in the three schools that the authors have taught have very diverse backgrounds, both in terms of their undergraduate education and their workplace experience. (e.g., management, economics, engineering, life sciences, social sciences, etc.). Finally, a main challenge is to re-package data mining to create a course that is business-oriented. This means less focus on the algorithmic and technical aspects of DM, and more focus on the application issues.

The authors of this paper have extensive experience in teaching DM courses to non-technical students in business schools in India, Portugal and the USA. Despite the geographical distance between the courses, we found out that similar answers were found to address the same issues. Additionally, we found that these courses are quite different from traditional DM courses offered at Computer Science (CS) departments. In this paper, we discuss those differences. Both authors designed new data mining courses for MBA (or other business graduate) students and have taught such courses for several years. In this paper, we discuss our experience, and the main points that differentiate a DM course in a business school from a course offered in a more technical program (computer science, statistics, etc.). The remainder of the paper is organized as follows. Section 2 provides an overview of business school courses and students. Section 3 presents the course objectives, and Section 4 describes the content covered in a typical course. Section 5 focuses on course delivery and the different course components and deliverables. A discussion is presented in Section 6.

2. Data Mining Courses at Business Schools

DM courses are typically offered as part of MBA programmes, and less frequently as part of an undergraduate business degree. Of the five courses described here,

Appearing in *Proceedings of the Teaching Machine Learning workshop at ICML 2012*, Edinburgh, Scotland, UK, 2012.

all are offered as an MBA elective, with the exception of the Database Marketing course, which is part of an M.Sc. programme (this is a lower cost program with a larger proportion of students without professional experience). MBA programs at UMD include both part-time and full-time programs. In the part-time program at UMD as well as in the programs at UPorto, students are mostly young professionals who work during the day and take classes in the evenings or weekends. The full time program students take 2 years off from their job for studies. MBA students tend to have diverse backgrounds (management, economics, engineering, social sciences, etc.). Students are typically highly motivated and especially those who self-select to take the DM elective.

The five courses range in length from a mini-semesters (5-7 weeks) to a full semester (14 weeks), with 2-3 hours of lectures per week. Class size ranges between 15-40. On these dimensions, the b-school courses are similar to CS courses.

The most important difference is the audience, both in terms of their professional profile and their background. Students attending DM courses at CS departments are mostly computer scientists and have no professional experience. This is a significant difference and, in fact, it motivates the different approach to organizing and delivering DM courses at b-schools.

3. Course Objectives

All five courses described here share the general goal of introducing business students to Data Mining from the managerial perspective. In particular, the main course objectives are that a student will be able:

- to identify opportunities for using data mining
- to use a particular tool to address those opportunities
- to understand how to properly interpret and evaluate the models and results
- to recognize the challenges involved in a DM project

These objectives are directly aligned with managerial job requirements. Given these objectives, it is clear that the course emphasis is on familiarization with the main ideas and concepts of DM and especially of the connection between the business problem and context and the use of DM.

The idea is to de-mystify DM and create managers who are reasonably conversant in DM, who can then communicate effectively with the technical people as well as convey technical results to upper management. Additionally, they should be able to carry out small DM projects on their own and, as their experience grows and with further self-learning, even medium-sized projects. This is particularly important in small companies, in which the students of these courses will take on multiple roles.

4. Course Content

Our experiences are based on designing and teaching five courses:

1. Business Intelligence Using Data Mining (BIDM) at *Indian School of Business*, since 2010. This course covers supervised and unsupervised learning methods for cross-sectional data, with an emphasis on data visualization, the data mining process, and integration into the business context (namely, in terms of problem definition and results validation).

2. Business Forecasting (BSFC) at *Indian School of Business*, since 2011. This course covers predictive analytics methods (regression, neural nets, smoothing, etc.) for time series. Aside from models, the course emphasizes the forecasting process, from problem definition and data collection to deployment and communication with stakeholders.

3. DM and BI (DMBI) at *EGP-University of Porto Business School*, since 2008. This course covers supervised and unsupervised learning methods for cross-sectional data, with an emphasis on the data mining process, data preparation and integration into the business context (namely, in terms of problem definition and results validation).

4. Database Marketing (DBM) at *Faculty of Economics of Porto*, since 2006.[1] This course covers supervised and unsupervised learning methods for cross-sectional data and, time allowing, the combination of DM techniques with optimization methods (e.g., for optimization of multiple campaigns) The emphasis is on the data mining process, data preparation and integration into the business context (namely, in terms of problem definition and results validation).

5. Data Mining for Business at *Smith School of Business* (UMD), 2004-2010. This course combines supervised and unsupervised learning methods mostly for cross-sectional data, but with a short

[1] The first three years together with Alpio Jorge.

module on forecasting time series.[2] The emphasis is on data visualization, the data mining process, and integration into the business context (namely, in terms of problem definition and results validation).

In all courses, the focus is on the concepts, rather than the algorithmic, statistical and mathematical aspects of the methods. The goal is to make sure that students have a rough idea of how the methods work, and more importantly, when each method is appropriate and what are its strengths and weaknesses. Predictive evaluation is a crucial component, introducing students to the concept of holdout data and the danger of over-fitting.

5. Course Delivery

Learning in the course is based on several components: in-class lectures, out-of-class communication, projects and in some cases individual assignments. We note that none of our courses has a final exam. Instead, the key component of each of our courses is a *team project*, where teams use real data to address a business problem.

5.1. Team Project

Students are expected to find their own data for the project. Students in part-time programs tend to have access to data from their workplace. This is important because it exposes students to the issues involved in real DM projects, including data availability and quality. The student who provides the data has the important role of the domain expert, leading the problem definition process and validating the outputs. This leads to very diverse and interesting projects, such as "Forecasting Monthly Tourism to Sikkim, India", "Predicting Delays in the Operating Room", "Segmentation of customers and market basket analysis in the pharmaceutical market" and "customer segmentation, targeting and sales prediction in a heavy industry".

Other data sources are the Internet (various websites provide rich data), data from other courses, data from course alumni and from faculty. Students must frame the business problem, and then use the data to address that problem. At the end of the course, each team gives a 10-15 minute project presentation, and submits a professional report. Examples of these deliverables at UMD and ISB can be found at `http://galitshmueli.com/student-projects`.

[2] Due to administrative constraints, in 2011 this course was converted to a mini-semester course that excludes time series forecasting.

Because most courses are 5-7 weeks long, it is essential to get the project going as early as possible. However, it is very hard for students to find the right data in the beginning of the course because they still don't know what kinds of problems they can solve with DM. In the case of students getting data from their workplaces is to direct them to sales/marketing/CRM data. One of the authors uses an early deliverable scheduled in week 2, where each team must present one good chart of their data. Additionally, while it is usually not a problem to keep the students motivated to work on the project, ensuring that they focus on the overall project goals is sometimes difficult. For instance, students might spend too much time trying different algorithms to optimize the results; or, particularly in the case of data from their workplaces, they may get stuck on data quality issues, which despite being crucial in a real application scenario, are not critical for these courses. One approach to guide the progress of the projects is to ask the team specific questions for each one and then tell students to focus on answering those questions and nothing else. After they solve a first set of questions on all the subjects covered, they are asked to individually do one or two more iterations on each one of those subjects. They can clean data better, prepare data in a different way, tune parameters and try new algorithms. This has multiple goals, including 1) to illustrate the iterative nature of the DM process, 2) to consolidate knowledge, which is hard to do in the first iteration and 3) to promote the understanding of the materials by all members of the group and not just the most active ones.

5.2. Lectures

Each course consists of in-class lectures, each 2-3 hours long. After a few introductory classes that introduce the idea of data partitioning, cleaning, etc., a typical lecture covers one-two tasks (e.g., clustering). One of the authors has found it useful to introduce a modeling task as early as possible in the course, to give students a feel for what can be achieved with DM. Additionally, this makes it easier to motivate the need for a DM methodology and the other steps of the knowledge discovery process (e.g., data preparation and evaluation).

A lesson starts with a motivating real-life example, followed by a data example, to which the methods are applied. The methods are then introduced at the conceptual level, with focus on how to execute the methods with software, how to interpret the output, and a discussion of their pros and cons. For example, at UPorto, the k-means method is explained in the following steps:

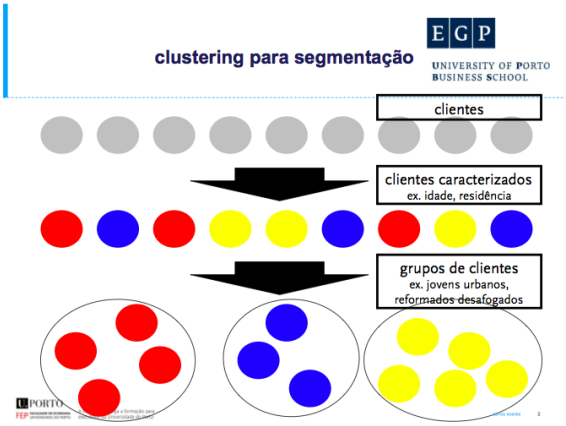

Figure 1. The slide motivating the use of clustering for market segmentation.

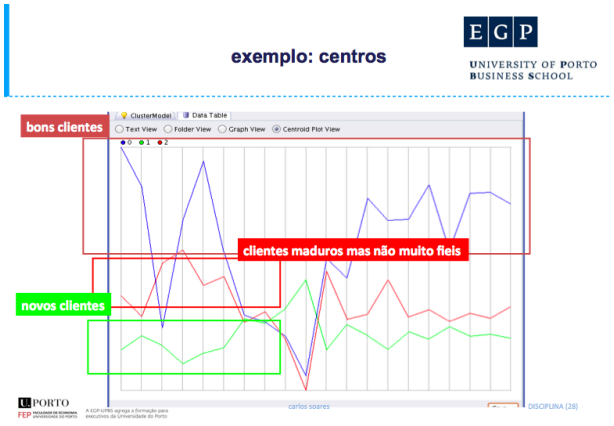

Figure 2. The slide with clustering results.

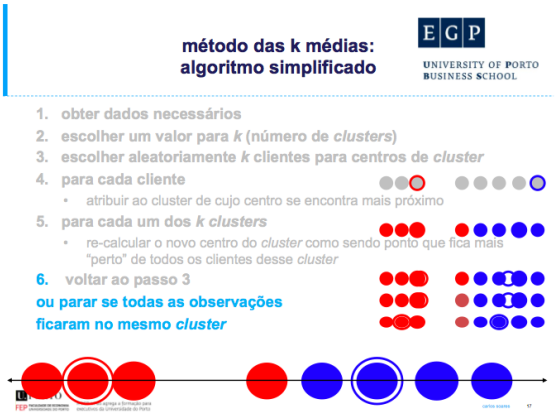

Figure 3. One of the slides explaining the K-means algorithm.

1. Motivation of how clustering can be used for market segmentation, supported with examples from real applications (Figure 1).

2. Hands-on example, consisting of the description of a data set, the execution of a clustering experiment using a drag-and-click tool (e.g., RapidMiner) and interpretation of the results from a marketing perspective (Figure 2).

3. Visual explanation of the clustering algorithm on a very simple, univariate example which is, nevertheless, sufficient to introduce the basic concepts (Figure 3).

4. Discussion of some of the characteristics of the algorithm (e.g., its stochastic nature) and their implications from a business perspective (e.g., due to the stochastic nature of the algorithm, two executions may result in different sets of clusters, which may reduce the credibility of the data mining project).

An illustration of the teaching style at UMD can be found in a YouTube video at `http://youtu.be/8QA54Xf2rd0`. The topic covered in this video is logistic regression. Teaching at ISB is illustrated by the two following videos, covering linear regression. The first one (`http://youtu.be/XXGkgM0R27M`[3]) gives an idea of how topics are motivated and the interaction with the students. The audience constitutes students who previously attended a course on Statistics and one of the goals is to help them transition to the DM mindset. The second video (`http://youtu.be/BHy3Nt78Jyw`) shows an in-class demonstration of the use of software for data partitioning and evaluating the predictive power of models. This video also illustrates the types of questions that students raise.

At ISB and UPorto, some lectures have a hands-on component where a case study is introduced, and students are requested to approach small modeling tasks on their laptops, followed by in-class discussion.

5.3. Out-of-Class

In addition to face-to-face class time, both instructors supplement the course with online interactions through email, skype, online discussion boards and/or Facebook groups. The online communication channel gives students support in-between lectures and allows faster learning. It also helps troubleshoot small computational and data preparation issues that might arise. Finally, students can share not only Q&A but

[3]Starting at minute 5.

also relevant news articles and other interesting business analytics links.

5.4. Software and Textbooks

Because of the diversity in technical skills, the lack of programming knowledge and job expectations, instructors of DM courses at b-schools tend to prefer software that does not require programming. Excel and Drag-and-click software are common choices. At ISB and UMD the main software is XLMiner (an Excel add-on), and Spotfire Miner (drag-and-click) is also introduced. At UPorto, Excel and Rapid Miner are used.

The focus on data mining within the business context with application to business problems, rather than the algorithmic focus, calls for textbooks that have the same flavor. Textbooks meant for CS students are too technical and insufficiently business-focused for our courses (e.g., (Witten & Frank, 2005; Hastie et al., 2001; Hand et al., 2001; Han & Kamber, 2000; Shmueli, 2011)). The second author resorted to writing a textbook for the business audience (Shmueli et al., 2010), which since its first edition in 2006 has become popular in b-schools. The Business Forecasting course also uses a self-authored text (Shmueli, 2012). At UPorto, a combination of three textbooks is used (Delmater & Hancock, 2001; Linoff & Berry, 2011; Parr Rud, 2001).

6. Results and Conclusions

At all three b-schools, while the data mining course is one of the more difficult courses in terms of its demands, students see its value, and our courses have become more and more popular over the years, with more students self-selecting to take them.

The feedback that we have been receiving from students is that the data mining course has given them an important advantage both in getting hired (there is a large demand for business experts with understanding of business analytics), as well as on the job. The ability to converse in the DM language and to see the value in it gives a large advantage in todays environment.

At the Indian School of Business, a current effort by one of the authors is to create ties with the local industry, and have them share data for student projects. Such ties can benefit the course participants and the industry collaborators, who are just beginning to venture into the analytics zone. Such ties can also help with hiring and further disseminating DM knowledge into business.

A critical challenge in teaching DM courses in b-schools is the mini-semester structure which means that courses are typically 6-7 weeks. For course-based projects, this is a real challenge and requires the instructor to make sure teams start working on the project as early as possible. Team management in itself can be challenging, requiring good interpersonal communication skills with individuals and teams as well as conflict resolution skills. Also, when working with data from their workplaces, students have great expectations for the projects, which are sometimes shared by their bosses. For several reasons, especially lack of time, these expectations are never met which may cause anxiety and frustration during the execution of the project. Therefore, the management of expectations by the instructor is essential for the success of these courses. An important part of expectation management is to make sure that students frame a specific, potentially useful problem. While initially they are eager to answer many questions, within the timeframe of the course the project scope must be very limited.

One direction that we have explored but has proven challenging is that of guest lectures from industry. While a good speaker can motivate and highlight the usefulness of DM in business, the short durations of courses in b-schools mean that a guest lecture requires sacrificing a precious lecture. One possibility is to hold out-of-class, and even online, guest lectures.

Lastly, while both authors come from a technical background (computer science and statistics), there is now a growing number of faculty from other disciplines who are taking on teaching DM courses in the b-schools. For these instructors, a serious challenge is mastering the material, the DM approach (which differs markedly from econometrics, operations research, and classic statistics), the hands-on teaching style, and software. Recommendations for such instructors include joining online professional peer groups (e.g., `http://www.facebook.com/groups/DMteaching`), co-teaching with an experienced teacher, and taking self-paced online courses such as those on statistics.com.

Acknowledgments

The authors gratefully acknowledge colleagues who helped design, co-teach or give feedback on the courses described in this paper and the students who helped to improve them with their participation and feedback.

References

Delmater, Rhonda and Hancock, Monte. *Data mining explained: a manager's guide to customer-centric business intelligence*. Digital Press, Newton, MA, USA, 2001. ISBN 1-55558-231-1.

Han, Jiawei and Kamber, Micheline. *Data mining: concepts and techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000. ISBN 1-55860-489-8.

Hand, D, Mannila, H, and Smyth, P. *Principles of Data Mining*. MIT Press, 2001.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome H. *The Elements of Statistical Learning*. Springer, 2001. ISBN 0387952845.

Linoff, Gordon S. and Berry, Michael J. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley, 3rd edition, 2011. ISBN 978-0-470-65093-6. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470650931.html.

Parr Rud, Olivia. *Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management*. Wiley, 2001. ISBN 978-0-471-43751-2. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471437514,descCd-authorInfo.html.

Shmueli, Galit. *Practical Time Series Forecasting: A Hands-On Guide*. CreateSpace, 2011. ISBN 1460977637. URL http://galitshmueli.com/practical-time-series-forecasting-book.

Shmueli, Galit, Patel, Nitin R., and Bruce, Peter C. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. Wiley, 2nd edition, 2010. ISBN 978-0-470-52682-8. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-EHEP002378.html.

Soares, Carlos and Ghani, Rayid. *Data Mining for Business Applications*, volume 218 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2010. URL http://10.255.0.115/pub/2010/SG10.

Witten, Ian H and Frank, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. ISBN 0120884070.