Indian School of Business

Business Intelligence using Data Mining

Academic Year & Term: 2011 – 12: **Term:** 6	
Instructor: Galit Shmueli **Affiliation:** ISB **Email:** Galit_Shmueli@isb.edu **Course info:** http://galitshmueli.com/content/business-intelligence-using-data-mining	**Office Hours:** 2 hours / week (depending on class schedule)
Academic Associate: TBA **Email:**	

Course Objective and Key Takeaways from the course

Students taking this course will appreciate and be able to identify the enormous opportunities that currently exist in providing business intelligence services based on data mining techniques of micro-level data capture, cleansing, validation, storage and analysis.

You are not required to have any extraordinary technical ability in Statistics or other methodological areas to benefit from the course. The material is developed and presented in an intuitive manner with the objective of making you smart consumers of this widely applicable (social science) technology. Where possible, cases will used to motivate the topic being covered.

Upon successful completion of the course, you should possess valuable practical analytical skills that will equip you with a competitive edge in almost any contemporary workplace. In particular, the knowledge acquired in this course will benefit those who plan careers in analytics, targeted marketing, predictive modeling and strategic consulting. More formally, the course will provide participants with the following skills and knowledge:

•   Be aware of the business intelligence potential of today's data rich environment

•   Gain a practical understanding of the key data mining methods of classification, prediction, data reduction and exploration

•   Know how to decide when to use which technique

•   Understand how to implement major techniques using Excel add-ins

•   Become a smart and critical consumer of data mining techniques

•   Gain the intellectual capital required to provide business analytics services

Learning Goals
In addition to the course objectives listed above, students should expect to develop the following by the end of the course:

1.   Effective Oral Communication
     Each student shall be able to communicate verbally in an organized, clear, and persuasive manner, and be a responsive listener.

     Assessment: Project Presentations

2. Critical and Integrative Thinking
   Each student shall be able to identify key issues in a business setting, develop a perspective that is supported with relevant information and integrative thinking, to draw and assess conclusions.

   Assessment: Individual Assignments (plus ungraded: in-class and online discussions)

3. Interpersonal Awareness and Working in Teams
   Each student shall demonstrate an ability to work effectively in a team, exhibiting behavior that reflects an understanding of the importance of individual roles and tasks, and the ability to manage conflict and compromise, so that team goals are achieved.

   Assessment: Team Project Report and Team Presentations (plus ungraded: team meeting with instructor)

Course Description

How can mobile companies use their customer database to predict customer churn or to personalize SMS messages for improving customer service? How can financial institutions use past loan data to predict the chance of defaulting for a new loan applicant? How can Bollywood use data on movies to predict the next box-office hit? How can charities use data from a campaign in one location to target the right people in another location? And how can politicians use databases of supporters to segment and best target each audience?

The widespread proliferation of IT influenced economic activity leaves behind a rich trail of micro-level data. Yet, most organizations are data rich but information poor. Emerging technologies such as RFID, weblogs, social networks, website usage tracking and vast amounts of online information (such as product ratings and bid histories) have the potential to reveal a lot about consumer, supplier and competitor preferences to those that have the ears (read data-mining capabilities) to listen.

We will work with real world business data, as opposed to artificial examples. We will learn the types of questions that data mining can answer and which are the appropriate data mining tools for answering different questions. The emphasis is on understanding the concepts behind a wide set of data mining techniques and their relation to specific business-intelligence situations, rather than on mastering the theoretical underpinnings of the techniques.

The course will be driven by a business intelligence project that will encapsulate the different components of the course. For examples of past course projects see http://galitshmueli.com/student-projects

An important feature of this course is the use of Excel, an environment familiar to MBA students. All required data mining algorithms (plus illustrative data sets) are provided in an Excel add-in, XLMiner. In addition, we will introduce you to TIBCO Spotfire, an industry leading data-visualization tool, and time permitting to its data mining tool.

Required Text Book

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner **by Galit Shmueli, Nitin R. Patel and Peter C. Bruce, Wiley, 2007.**

Recommended: 2nd edition of the same textbook (copies are available in the library)

Software requirements for the course

We will make extensive use of Microsoft Excel and a data mining software called XLMiner, which is an Excel add-in.

We will also use the interactive visualization tool TIBCO Spotfire. Please download your copy from http://registration.spotfire.com/eval/default_edu.asp (make sure to use your isb.edu email address). For further details see the Learning Management System.

Session-Wise Topics/Readings

Session (Date)	Reading	Topics	Deliverable
#1 Nov 28	Chp 1, 2 Chp 5	**Course Overview** (data mining vs. statistical inference; examples) **Prediction** (Linear regression for prediction; over-fitting; data partitioning; predictive metrics; Intro to XLMiner)	Email team members to AA
#2 Dec 2	Chp 5	**Prediction** Prediction vs. classification Discussion of Project	
#3 Dec 5	Chp 3 Handout: Visualization	**Data Preparation & Visualization** Intro to Spotfire **Dimension Reduction** (Reducing # categories; Principal Components Analysis)	Assignment 1 (Problem 5.4, page 89; predicting used car prices)
#4 Dec 7	Chp 12	**Unsupervised learning: Cluster Analysis** (Hierarchical Clustering; k-Means Clustering)	Assignment 2 (Problem 3.2 on p.51 using Spotfire; breakfast cereals)
#5 Dec 12	Chp 4	**Classification Goals** (Classification vs. profiling; examples) **Classification Performance** (Prediction vs. ranking; Misclassification costs)	Present graph for project data
#6 Dec 14	Chp 6	**K-nearest neighbors** **Naïve Bayes** Case: *Personal Loan Acceptance*	
#7 Dec 19	Chp 7	**Classification and Regression Trees**	Assignment 3 (Problem 6.2 on pp. 108-109; predicting injuries in car accidents)
#8 Dec 21	Chp 8	**Logistic Regression** **Generalizing to multiple classes**	Assignment 4 (Problem 7.3 on pages 135-136; predicting used car prices)
#9 Dec 26	Chp 11	**Association Rules** **Closure:** The Big Picture	
#10 Dec 28		**Team Presentations**	Project report

Evaluation Components

Deliverable	Weight	Coding Scheme (see appendix)
Four individual assignments	60%	2
Final Team Project Report	30%	1
Mid-term Team Presentation	5%	1
Final Team Presentation	5%	1

Individual Assignment Schedule

Name of the Component	Date of Submission/Deadline	Take-home or in-class	Group Assignment (Y/N)	Instructions to students on word limit/format of submission etc	Coding Scheme
Assignment #1	Session #3	Take-home	N	Hardcopy	2
Assignment #2	Session #4	Take-home	N	Hardcopy	2
Assignment #3	Session #7	Take-home	N	Hardcopy	2
Assignment #4	Session #8	Take-home	N	Hardcopy	2

About Individual Assignments:

Because the most effective way to learn this material is by "doing," the course will have four individual assignments. Please budget approximately four hours per assignment. Our expectation is that as a result of these assignments, you will observe a marked improvement in your conceptual understanding of data mining.

Assignments are to be submitted via hardcopy at the beginning of the class meeting they are due. Late assignments are accepted, but subject to a penalty that deducts 2 points (on a total of 15 points) per day they are late. For example, if an assignment due on Wednesday is not submitted via hard copy at the beginning of the class, but sent via email after the class, there will be a 2 point penalty. If, on the other hand, it is send on Thursday at 8pm, there will be a four point penalty and so on.

No extra credit work is offered under any circumstances. You are always welcome to submit assignments prior to the due date in case you have an exigency on that particular date.

Team Project

Based on their collective work experience, each group should identify, and to the extent possible, execute a business intelligence project that relies on the data mining techniques we will cover in the class.

Students should form groups of 4-5 members. Our prior experience indicates that the more diverse the group the more interesting the project and the analysis.

Project Meeting: Each team must **meet with the instructor** *once* before session 7 for a 30-min meeting. The meeting is intended to help you with the team project by discussing challenges and suggesting tips. All team members must be present at the meeting.

Team Project Schedule

Name of the Component	Date of Submission/Dead line	Take-home or in-class	Group Assignment (Y/N)	Instructions to students on word limit/format of submission etc	Coding Scheme
Send names of team members	Before Session #2	Take-home	Y		
Team presents graph of project data	Session #4	In-Class	Y	Upload PPT to LMS	1
Team meets with instructor	Before Session #7	Instructor's office	Y	Bring laptop with data + questions	
Project Presentation	Session #10	In-Class	Y	Hardcopy	1
Project Report	Midnight of Session #10	Take-Home	Y	Submit PDF by email	1

Where do we get the idea for the project?

For the most part, based on your past experience, you should be able to revisit any data driven business decision making exercise from a data mining perspective.
In the absence of this, you should be cognizant of the fact the Internet has spawned an enormous amount of online activity that increasingly permeates our lives and our society. Because of such pervasiveness, the Internet offers data-rich environments that enable new business intelligence opportunities. For instance, it is possible to collect information about how online retailers price differentiate (from sites such as bizrate.com), how airlines make pricing decisions (orbitz.com), how consumers bid in a given auction or a series of auctions (ebiz.com), the degree of interest a news piece generates (blog ranking sites)-- all in a very cost-effective and non-intrusive manner. We leave it to your collective imagination to find something interesting for which you can collect some data.

For a list of sample projects that the instructor has guided please visit http://galitshmueli.com/student-projects

What is the grading criterion?

You will be graded on your ability to demonstrate your grasp of basic data mining concepts and how they can be applied in a business context. Grades will be determined based on the in-class presentations, and your final project report. Purely conceptual projects, that do not have any empirical data component, will receive a lower grade than those projects that do have such a component.
The weights will be split between the presentations (10%) and report (30%). Projects will be peer reviewed within and across groups.

The key tasks for your team project are:
• Identify a business problem or a series of interesting questions that deal with classification, prediction or clustering
• Identify sources of data that could potentially be useful in addressing your questions
• Pre-process – clean, validate, visualize and partition your data
• Develop your model considering alternative techniques, selecting the most appropriate one in the process
• Interpret your results, and write a final report including an executive summary of your findings
• Prepare a 15-20 minute presentation for the last class meeting - 5 slides max
• Prepare a report that will serve as your term paper

What are the guidelines for the project report?

The report should be no longer than five pages in length (please use at most 12pt font size, 1.5-spacing, and 1 inch margins on all sides of the page) and accessible to any manager (not necessarily familiar with data mining). Avoid jargon. The report should have the following sections:

1. Executive summary
2. Problem description, including existing/traditional way of addressing the issue
3. Brief description of the data, its source and its key characteristics. More details can be provided in an Appendix
4. Findings from the application of data mining models
5. Conclusion

You can provide additional technical details of the model and the data in an Appendix.

Group Information

Group Size	4-5
Group Composition	Student-chosen (encouraged to be diverse)
Can groups be formed across different sections?	no

Attendance & Punctuality

Learning is an interactive process. ISB students are admitted partly based on the experiences they bring to the learning community and what they can add to class discussions. Therefore attendance is an important aspect of studying here. You have to be present in all the classes. Absence is only appropriate in cases of extreme personal illness, injury, or close family bereavement. Voluntary activities such as job interviews, business school competitions, travel plans, joyous family occasions, etc. are never valid reasons for missing any class. The faculty with the assistance of the Academic Associate will keep track of your attendance and decide on the nature and extent of penalty for any absence from the class. Penalty may include reduction in grade.

Late arrival is disruptive to the learning environment; so you have to be in class before the scheduled time. Most courses meet twice a week during the day. Normally there are no classes scheduled on Friday or in the evenings, but there are exceptions. Class and Exam schedules are posted on the PGP intranet site. Any change in the class schedule is notified in advance.

Class participation, both in an out of the classroom, is strongly encouraged. It forms an important part of the course pedagogy. Participation is a direct function of your outside experiences and the level of mastery of the material. You are expected to enhance the overall learning environment of the class by coming prepared, asking questions and bringing issues to life based on your experience.

You can further contribute by engaging in the **Online Discussion Forum**.

Online Course Management

We will use the LMS extensively for material distribution as well as for online discussions.
•   Handouts, assignments, datasets and other materials will be posted only on the LMS
•   Discussion board will be used to post Q&A for each assignment

.

Appendix I

Coding scheme for ALL course work

Code number	Individual	Group
1	Anything and everything is fine. The student can talk to anyone and use ANY resources in completing the assignment or project. However, need to have proper citations for all borrowed work (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware/ other secondary/ primary sources), including case/problem set solutions.	Anything and everything is fine. The group can talk to anyone and use ANY resources in completing the assignment or project. However, need to have proper citations for all borrowed work (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware/ other secondary/ primary sources), including case/problem set solutions.
2	The student is limited to discussing only concepts and ideas with other people. The student is NOT allowed to discuss the specifics of the case or assignment with anyone else and has to complete all work on his/her own. The student can use ANY reference material with proper citation (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware/other secondary/ primary sources), including case/problem set solutions.	The group is limited to discussing only concepts and ideas with other people. The group is NOT allowed to discuss the specifics of the case or assignment with anyone else and has to complete all work on their own. The group can use ANY reference material with proper citation (company website/ databases/ industry reports/ Google/ journals/article/ open courseware /other secondary/ primary sources), including case/problem set solutions.
3	The student is limited to discussing only concepts and ideas with other people. The student is NOT allowed to discuss the specifics of the case or assignment with anyone else and has to complete all work on his/her own. The student can use SOME reference material (except case/problem set solutions) with proper citation (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware/other secondary/ primary sources) as specified by the course instructor.	The group is limited to discussing only concepts and ideas with other people. The group is NOT allowed to discuss the specifics of the case or assignment with anyone else and has to complete all work on their own. The group can use SOME reference material (except case/problem set solutions) with proper citation (company website/ databases/ industry reports/ Google/ journals/articles/open courseware/ other secondary/ primary sources) as specified by the course instructor.
4	The student is limited to discussing only concepts and ideas with other people. The student is NOT allowed to discuss the specifics of the case or assignment with anyone else. The student **cannot** use ANY reference material. This applies to all reference material (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware /other secondary/primary sources, case/problem set solutions).	The group is limited to discussing only concepts and ideas with other people. The group is NOT allowed to discuss the specifics of the case or assignment with anyone else. The group **cannot** use ANY reference material. This applies to all reference material (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware /other secondary/primary sources, case/problem set solutions).
5	The student is NOT allowed to discuss anything with anyone. The student can use ANY reference material with proper citation (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware/other secondary/ primary sources), including case/problem set solutions.	The group is NOT allowed to discuss anything with anyone. The group can use ANY reference material with proper citation (company website/ databases/ industry reports/ Google/ journals/article/ open courseware /other secondary/ primary sources), including case/problem set solutions.
6	The student is NOT allowed to discuss anything with anyone. The student can use SOME	The group is NOT allowed to discuss anything with anyone. The group can use SOME reference material

	reference material (except case/problem set solutions) with proper citation (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware/other secondary/ primary sources) as specified by the course instructor.	(except case/problem set solutions) with proper citation (company website/ databases/ industry reports/ Google/ journals/articles/open courseware/ other secondary/ primary sources) as specified by the course instructor.
7	The student can ONLY use the information given in the case or assignment. The student is NOT allowed to discuss anything with anyone or to refer to any other material. This applies to all reference material (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware /other secondary/primary sources, case/problem set solutions).	The group can ONLY use the information given in the case or assignment. The group is NOT allowed to discuss anything with anyone or to refer to any other material. This applies to all reference material (company website/ databases/ industry reports/ Google/ journals/ articles/open courseware /other secondary/primary sources, case/problem set solutions).

As a general rule:

- Students can discuss cases and assignments with the course instructor and the Academic Associate for the course.

Required and recommended textbooks for the course and the course pack can be used to answer any individual or group assignment.