BUDT733: Data Mining for Business

Syllabus for DC01, Fall 2009

Instructor: Professor Galit Shmuéli
My office: 4361 Van Munching Hall
E-mail: gshmueli@rhsmith.umd.edu
Tel: (301) 405-9679
Office hours: Before class, by appointment.

Teaching Assistant: Brad Greenwood
E-mail: qiulsilgas@gmail.com
Office hours: by email

Overview

"Data Mining for Business" is a second level course in managerial data analysis and data mining. The emphasis is on understanding the application of a wide range of modern techniques to specific decision-making situations, rather than on mastering the theoretical underpinnings of the techniques. 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. The course covers methods that are aimed at prediction, forecasting, classification, and clustering. It also introduces cutting edge interactive data-visualization tools, as well as data reduction techniques.

Required Textbook

1. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner,* by Shmueli, Patel, and Bruce, 1st edition, John Wiley & Sons (ISBN: 0-470-08485-5). Accompanying datasets are available from www.dataminingbook.com.

2. An additional set of chapters on Time Series Forecasting will be supplied via Blackboard.

Required Software (see Blackboard Course Documents > Software)

1. We will make extensive use of Microsoft Excel and the datamining software add-in called *XLMiner*. It is available through the Portal (https://portal.rhsmith.umd.edu/login), in the College Park labs, and a free 6-month copy comes with your textbook (see sleeve in inner back cover for download instructions and license key).
2. We will also use the interactive visualization tool *Spotfire*. Please download your copy from http://registration.spotfire.com/eval/default_edu.asp (make sure to use your *umd.edu* email address).
3. Another data-mining software that we will see is *SAS Enterprise Miner*. It is available through the Smith Portal

Course Website

The course website is hosted in the 'Blackboard' course environment (http://bb.rhsmith.umd.edu/).
- ❑ The website will contain files for you to browse, download, or print (data files, instructions, assignments, solutions, class examples, etc.).
- ❑ A university LDAP id and password are required to access Blackboard courses. For more info see https://ldap.umd.edu/cgi-bin/chpwd.
- ❑ Please maintain your current e-mail address in Testudo as Blackboard uses only this address to send course related e-mail. (Since e-mail addresses are imported from Testudo into Blackboard, it is not possible to update e-mail addresses from within Blackboard).

Clickers

We will be using clicker technology for interactive learning. Each student will receive a keypad (a remote-like device) which is used for answering questions in class. It allows confidential responses, yet enables everyone to participate and express themselves, and we all get immediate feedback.
- ❑ Once you get your clicker, register it ASAP at your my.umd.edu page. You will see the clicker registration "portlet" under the Academics and Testudo tab. Click either the "Register clicker" or "Update clicker" link, and complete the form using the 6 character device ID located on the back of the keypad.

❑ During our last class you must return the keypad in order to receive a final grade in the course.

Course Objectives

A wide array of skills is required in arriving at informed managerial decisions. Among these are analytical and quantitative skills. This course seeks to develop these two important skills. More formally, the goals of this course are:
- ❑ To introduce 2nd level statistical and data mining techniques
- ❑ To introduce exploratory analytics and visualization
- ❑ To understand the limitations of various techniques
- ❑ To decide *when* to use *which* technique
- ❑ To implement major techniques using Excel add-ins
- ❑ To improve spreadsheet skills
- ❑ To become smart/skeptical consumers of statistical techniques
- ❑ To enable interaction with personnel specializing in analytics

Specific topics covered in this course include: a review of regression analysis (with emphasis on prediction), time series forecasting, contingency tables, classification methods (discriminant analysis and logistic regression), cluster analysis, and principal components analysis.

Grading Policy

Your final letter grade for the semester will be determined by several components according to the following system of weights:

Individual Assignments (5)	35 %
Team Assignments (2)	20 %
Quizzes (2)	20 %
Team project report	10 %
Team Presentation/Poster Session	10 %
Attendance/ Participation	5 %
	--------
	100 %

Individual Assignments
- ❑ Approximately 5 short exercises will be assigned over the course of the semester. These assignments are to be done individually. There will be a thread on Blackboard's discussion board for Q&A.
- ❑ Each assignment will consist of a straightforward application of a particular data analytic technique to a decision-making situation. You may be asked to interpret the results via some specific and directed questions.
- ❑ You will have one week to complete each assignment. It is estimated that each assignment can be completed in approximately 2 hours.
- ❑ Assignments *hardcopies* will be collected at the beginning of class. Upload your Word and Excel (or other) files to Blackboard. Late submissions will receive *zero* points.

Team Assignments
- ❑ There will be 2 team assignments due. These assignments are to be done in teams of 4-5 students. Please send me an email with your group members' names. I will assign remaining students into groups.
- ❑ Each team assignment will consist of the application of a set of analytical methods. Additionally, you may be asked to provide relatively broad insights from the results of your analysis.
- ❑ At the completion of each team assignment, *each* team member should be thoroughly familiar with *each* step of the various procedures involved in the project. You should be prepared to explain in detail each of the various steps that you went through.
- ❑ When submitting a team assignment, each team member is required to fill out and submit a *peer evaluation form*, downloadable from Blackboard. After the peer evaluation data has been aggregated, I will scan it for evidence of any significant problem with regard to any individual student. Students identified as such will be individually contacted by me and will receive a lower grade for that assignment.
    - o If you feel that there is a serious problem with the participation of one of your team members, please contact me immediately. If informed well in advance, I can try to take remedial action to improve the functioning of the team for the ongoing assignment. The formal peer evaluation process can only improve the functioning of the team for *future* assignments.

- The due dates for these team assignments are listed in the course schedule. Hardcopies of the assignment and peer evaluations will be collected at the *beginning* of class on the due date.

Quizzes
- Two short quizzes are scheduled. (Please see attached course schedule.) *There will be no make-up quizzes.*
- Each quiz will be timed at approximately 60 minutes.
- There is no final exam in this course.

Team Project, Report, and Presentation
- Each team of 4-5 students will work on a data analysis problem involving real business data. The project will focus on classification methods, and will be carried out throughout the semester.
- To assist you in choosing a feasible problem and dataset, each team should obtain an "OK" from me regarding the scope and nature of the dataset and intended analysis. In order to get this OK, you should submit a short descriptive write-up by **Sept 29**. Guidelines for this document will be given later.
- Each group will meet with me once for 20-30 minutes before class to discuss their progress on the project and get some feedback. We will schedule these meetings a few weeks into the semester.
- Each team will present one good graph of their data on **Oct 27**. Please post the graph to BB by 16:00 on that day.
- A presentation session is scheduled for the last day of class. This is a team activity.
- You are encouraged to be creative in designing your presentation. A typical presentation has about 10-12 PowerPoint slides. Examples from previous semesters can be found at http://www.rhsmith.umd.edu/faculty/gshmueli/web/html/student-projects.html
- A final professional report will be submitted. Guidelines for this report will be given at a later date.

Attendance
- You are expected to be *on time* for each class. Attendance will be recorded using keypads. The only exception to this rule is if you have sought *and* received permission for your absence by writing to me well in advance of your planned absence.
- If you miss a class, it is your responsibility to keep abreast of all the information, academic as well as administrative, which was conveyed in that class.
- Class participation includes asking questions as well as offering insightful comments.
- As part of the participation requirement, you must post at least one comment to an entry on http://blog.bzst.com.

Course Pre-Requisites
While there are no official pre-requisites for this course, it is expected that all students are familiar with elementary probability and statistics, up to and including the basics of linear regression.

Other Course Policies
- **Academic Integrity:** The University's *Code of Academic Integrity* is designed to ensure that the principles of academic honesty and integrity are upheld. All students are expected to adhere to this Code. The Smith School does not tolerate academic dishonesty. All acts of academic dishonesty will be dealt with in accordance with the provisions of this code. For more information on the Code of Academic Integrity, please visit *http://www.inform.umd.edu/CampusInfo/Departments/JPO/AcInteg/code_acinteg2a.html*. In particular, you should neither give nor receive assistance from anyone in doing the *individual assignments* and quizzes.
- **Special Needs:** If you have a disability and/or special needs, you should bring this to my attention as soon as possible, but not later than the second week of class.

Data Mining for Business (BUDT733): Tentative Schedule

Date	Reading	Topics	Deliverable
1-Sep	Chp 1, 2 Chp 5	Course Overview Linear Regression for Prediction	Email team members
8-Sep	Chp 5	Linear Regression for Prediction	
15-Sep	Chp on Time Series	Time Series Analysis: Regression-based Methods & Autocorrelation	Indivd. Assignment 1
22-Sep	Chp on Time Series	Time Series Analysis: Smoothing methods	Indivd. Assignment 2
29-Sep	Chp on Time Series	Time Series Analysis: AR and econometric models	Project Proposal
6-Oct		**Quiz 1** Data preparation, cleaning, & exploration	
13-Oct	Chp 3.7 Chp 4	Principal Components Analysis Intro to Classification Methods	Team Assignment 1
20-Oct	Chp 4	Intro – cont., Evaluation Metrics Contingency Tables *Time Series Assignment Feedback*	
27-Oct	Chp 6	K-nearest neighbors, Naïve Bayes	Present graph for project
3-Nov	Chp 7 Chp 8	Classification and Regression Trees Logistic Regression	Indivd. Assignment 3
10-Nov	Chp 8	Logistic Regression *SAS EM Tutorial*	Indivd. Assignment 4
17-Nov	Chp 10	Discriminant Analysis Generalizing to multiple classes	
24-Nov	Chp 12	Cluster Analysis: hierarchical clustering and k-means clustering	Team Assignment 2
1-Dec		**Quiz 2** *Logistic Regression Assignment Feedback* Closure: The Big Picture	Indivd. Assignment 5
8-Dec		**Presentations**	Project report

All readings refer to chapters from *Data Mining for Business Intelligence* by Shmueli, Patel, and Bruce or the additional time series chapters.