

The Ohio State University
The Max M. Fisher College of Business
Department of Accounting and Management Information Systems

AMIS 894.31 – Data Mining for Business Intelligence

Autumn Quarter 2009

Contact Information:

Instructor: Prof. Waleed A. Muhanna (muhanna.1@osu.edu)
Phone: 292-3808; Fax: 292-2118
Office: 420 Fisher Hall
Office Hours: TR 1:00-2:00, and by appointment.

Course Overview:

Advances in information technologies and the increased digitization of business have led to an explosive growth in the amount of structured and unstructured data collected and stored in databases and other electronic repositories. Much of this data comes from operational business software (e.g., finance/accounting applications, Enterprise Resource Management (ERP), Customer Relationship Management (CRM), workflow and document management systems, surveillance and monitoring systems, and Web logs) and is often archived into vast data warehouses to become part of corporate memory. The result of this massive accumulation of data is that organizations have become *data-rich yet still knowledge-poor*. What can be learned from these mountains of data to improve decisions? How can an organization leverage its massive data warehouses for strategic advantage? A large number of methods with roots in statistics, informational retrieval and machine learning have been developed to address the issue of knowledge extraction from very large data sets. The term "data-mining" refers to this collection of methods. These methods have broad applications; they have been successfully applied in areas as diverse as market-basket analysis of scanner data, customer relationship management, churn analysis, direct marketing, fraud detection, click-stream analysis, personalization and recommendation systems, risk management and credit scoring.

The key objectives of this course are two-fold: (1) to provide you with a theoretical and practical understanding of core data mining concepts and techniques; and (2) to provide you with hands-on experience in applying these techniques to practical real-world business problems using commercial data mining software. As an applied course, the emphasis will be less on the inner working of each method and more on when and how to use each technique and how to interpret the results.

Broadly defined, data mining is the process of selection, exploration and modeling of very large data sets, in order to discover predictive and descriptive models and patterns. It encompasses both top-down (confirmatory or hypothesis driven) analysis using traditional statistical techniques and bottom-up (exploratory) analysis using database and machine learning techniques to discover regularities, relations, or “local structure/patterns” that are at first unknown. (The confirmative tools of top-down analysis can then be used to confirm the discoveries and evaluate the quality of decisions based on those discoveries.) In keeping with this broad definition, topics and related methods discussed include information retrieval and enterprise reporting, classification, predictive modeling, clustering, association and sequence analysis, and anomaly detection. The application of these functions and methods will be illustrated using modern software tools via examples, homework assignments and group term projects.

Upon completion of this course, students should be able to:

1. Fully appreciate the concept of data as a strategic resource;
2. Use existing database query tools for data/information extraction and enterprise reporting;
3. Understand how and when data mining can be used as a problem-solving technique;
4. Describe different methods of data mining;
5. Select an appropriate data mining technique for a specific problem;
6. Use existing data mining software to mine a prepared data set; and
7. Interpret and evaluate the results of data mining.

Prerequisites:

The course is specifically designed with MBA/MACC students as the intended target audience. The key prerequisites consist of good graduate standing and completion of an introductory course in probability and statistics. A review of basic concepts will be provided in class. Assignments do not involve programming, per se, and no prior IT experience or training is assumed.

Course Materials:

- **Textbook:** *Data Mining for Business Intelligence*, by Galit Shmueli, Nitin R. Patel, and Peter C. Bruce (Wiley: 2007). (Please be sure to buy a new copy or that your used copy includes a valid license to XLMiner.)
- A set of articles, assignments, tutorials, data sets, lecture notes, and various supplementary materials which will be made available through the course homepage at: <http://fisher.osu.edu/~wmuhanna/amis894X.html>

Course Organization:

The course will be run as a mixture of lectures, in-class demonstrations, assignments, and student presentations. Readings will generally be taken from the required text, together with tutorials, notes, and article supplements from the literature. Some material will be covered only in readings; other will be covered only in lecture which may depart from the text in either content or order. To maximize learning, classroom discussion and the amount of time spent on different topics will be adjusted according to the background and interests of the students.

Assignments

In addition to the reading requirements from the text and the supplementary materials, there will be six (6) homework assignments, spaced out over the course of the quarter. They are designed to reinforce your understanding of the topics covered. Assignments are to be handed in on or before the class period of the due date. No late work is accepted. A limited amount of cooperation on homework and lab assignments is permitted. You may, for example, discuss with classmates general solution strategies. However, everyone should independently do and turn in his/her own work.

Exams

There will be two in-class exams: a midterm and a final. The first is scheduled after the fifth week of classes on Tuesday, November 3. The second exam will be held during the final examination period on Tuesday, December 8. The second exam is not explicitly cumulative. The examinations are designed to assess each student's (a) command of factual knowledge and concepts from the course; and (b) his or her ability to integrate and generalize these concepts and principles and apply them to new situations. The format of both exams will primarily be problems and short essay questions. Exams must be taken at their scheduled times; make up exams will only be given for truly special and compelling cases, in accordance with University guidelines.

Team-Based Term Project

Students will have the opportunity to further sharpen their skills and acquire hands-on experience with practical databases and real data mining problems through a term project. The projects will be carried out in teams of 2-4 students and involve the use of DM software. Although I am generally open to suggestions, each project will normally involve the selection, design, and performance of a data mining plan using a public data set (such as those provided by the SAS Institute or in the UCI KDD Archive (<http://kdd.ics.uci.edu/>) or a non-proprietary data set available through private student contacts. A detailed handout about the project will be made available online at the beginning of the course, along with a list of suggested projects. Teams will submit a written project proposal partway through the quarter, followed by a written report and brief class presentation on the project during the last week of classes.

Software

The methods discussed in this class are computationally intensive and non-trivial; they cannot be performed using Excel. Fortunately, these methods have matured enough to the point where they are now implemented in commercial software. We will use Microsoft Access to familiarize you with relational query language SQL, the industry standard for data extraction, summarization and enterprise reporting. XLMiner, an EXCEL © add-in, will be introduced in class and used by students to do assignments and solve business problems using data mining techniques. Your required textbook includes complementary 6-month license to XLMiner. In the back of the book you will find an insert that contains the license for downloading the add-in. Once you have registered your copy of the textbook you will have online access to all of the datasets used as case studies in the textbook.

In addition to XLMiner, through a special agreement with SAS, free web-based access to SAS Enterprise Miner will be provided to students for evaluation and possible use in this class. While we cannot spend much valuable class time discussing the specifics of each package, certain features of each software tool will be demonstrated in class and I am available to help you with any questions you may have. Student will primarily learn the use of software through basic online tutorials and course assignments.

Participation

A portion of the final grade will be based on your class attendance and active participation, elements that are crucial to the success of class meetings. Attendance refers to punctual attendance. Your fellow students and I will expect you to come fully prepared to answer questions and discuss the assigned readings. Each individual is expected to actively and constructively contribute to class discussions. Good contributions transcend assigned readings and are inspired, timely, analytical, and relevant to the topics discussed. Students can also earn participation credit by drawing attention to related development, information and resources dealing. Your class participation grade will reflect my judgment of the quality and quantity of your contributions during the entire quarter.

Evaluation:

35% of the final grade will be based on graded homework assignments. The exams (100 points each) will each account for 20% of your grade. The group term project will account for 15% of the grade. The remaining 10% is assigned to class participation. Final grades will be based on overall class performance.

Feedback and Continuous Improvement:

Students are strongly encouraged to visit with me in my office and/or use e-mail to ask questions, to share suggestions about any aspect of the course, or to clear up possible points of confusion. I will use your feedback to continuously improve and fine-tune the coverage levels and the teaching/learning processes. Please note that I may not always be able to make all of the changes suggested, but I will do my best to accommodate your suggestions.

Tentative Course Schedule:

The following schedule gives the general plan for the course; deviations may be necessary. The course web site gives the dynamic picture and is an integral part of the class; please make sure to check it on a regular basis.

Session & Date	Topics and Required Readings
Session 1 (R 9/24)	<p><i>Course Introduction</i></p> <ul style="list-style-type: none"> • Overview/goals of data mining (DM) and knowledge discovery (KD) • Myths about data mining • The Data Mining process <p>Readings:</p> <ul style="list-style-type: none"> • <u>“Competing on Analytics”</u> <i>HBR</i>, 1/06 • Case: <u>“Diamonds in the Data Mine”</u> <i>HBR</i>, 5/03 (PDF) • <u>“A golden vein”</u> <i>The Economist</i>, 1/04. (PDF) • <u>“How Verizon cut Customer churn”</u> <i>Financial Express</i>, 10/03. (PDF) • TB: Chapters 1&2
Sessions 2 & 3 (T 9/29 & R 10/1)	<p><i>Data Extraction and Manipulation</i></p> <ul style="list-style-type: none"> • The Relational Data Model and Relational DBMS • Enterprise Reporting • Relational Algebra • SQL: The Relational Query Language <p>Readings:</p> <ul style="list-style-type: none"> • TB: Section 2.1 <p>Assignments:</p> <ul style="list-style-type: none"> • <u>Tutorial #1: Formulating queries in SQL</u> • <u>Assignment 1</u>
Sessions 4 & 5 (T 10/6 & R 10/8)	<p><i>OLAP and Multidimensional Data Analysis</i></p> <ul style="list-style-type: none"> • Datawarehousing and Multidimensional Databases • Data Quality • Summarization and Data Cubes • OLAP Tools and Pivot Tables <p>Readings:</p> <ul style="list-style-type: none"> • <u>“Datawarehousing: Managerial Considerations”</u> <i>CACM</i>, 9/98. (PDF) • <u>“An Introduction to OLAP Multidimensional Terminology and Technology”</u> (PDF)

Session 6 (T 10/13)	<i>Introduction to XLMiner and SAS Enterprise Miner</i>
Session 7 (R 10/15)	<p><i>Data Exploration and Dimension Reduction</i></p> <ul style="list-style-type: none"> • Data Summarization and Visualization • Correlation Analysis • Principal Component Analysis <p>Readings:</p> <ul style="list-style-type: none"> • TB: Chapter 3 • <u>Assignment 2</u>
Sessions 8 & 9 (T 10/20 & R 10/22)	<p><i>Association & Market-Basket Analysis</i></p> <ul style="list-style-type: none"> • Frequent Itemset and Association Rule Mining • Pattern evaluation (subjective and objective interestingness measures) • Sequential patterns <p>Readings:</p> <ul style="list-style-type: none"> • TB: Chapter 11 • R. Agrawal and R. Srikant, “<i>Fast Algorithms for Mining Association Rules</i>,” Proc. 20th Int. Conf. Very Large Data Bases (VLDB), 1994. (PDF; only skim) <p>Assignments:</p> <ul style="list-style-type: none"> • <u>Tutorial #3</u>: Using XLMiner for Association Rule Mining • <u>Assignment 3</u>
Sessions 10 & 11 (T 10/27 & R 10/29)	<p><i>Cluster Analysis</i></p> <ul style="list-style-type: none"> • Segmentation and Personalization • The K-means algorithm • Hierarchical (Agglomerative) Clustering • Cluster Validation and Interpretation <p>Readings:</p> <ul style="list-style-type: none"> • TB: Chapter 12 <p>Assignments:</p> <ul style="list-style-type: none"> • <u>Tutorial #4</u>: Finding clusters with XLMiner • <u>Assignment 4</u>
Session 12 (T 11/3)	<i>Midterm Exam</i>
Sessions 13 & 14 (R 11/5 & T 11/10)	<p><i>Classification and Predictive Modeling</i></p> <ul style="list-style-type: none"> • Decision Tree induction • Evaluation <p>Readings:</p>

	<ul style="list-style-type: none"> • TB: Chapters 6 & 7 <p>Assignments:</p> <ul style="list-style-type: none"> • <u>Tutorial #5</u>: Using XLMiner to build decision trees • <u>Assignment 5</u>
Sessions 15 & 16 (R 11/12 & T 11/17)	<p><i>Predictive Modeling Using Regression</i></p> <ul style="list-style-type: none"> • Review of OLS Regression • Logistic Regression • Model Evaluation and Interpretation <p>Readings:</p> <ul style="list-style-type: none"> • TB: Chapters 5 & 8 <p>Assignments:</p> <ul style="list-style-type: none"> • <u>Tutorial #6</u>: Regression • <u>Assignment 6</u>
Session 17 (R 11/19)	<p><i>Predictive Modeling Using Neural Networks</i></p> <ul style="list-style-type: none"> • Introduction to Neural Networks • Neural Networks vs. Regression • Model Evaluation <p>Readings:</p> <ul style="list-style-type: none"> • TB: Chapters 4 & 9 <p>Assignments:</p> <ul style="list-style-type: none"> • <u>Tutorial #7</u>: Creating Neural Network models
Session 18 (T 11/24)	<p><i>Anomaly Detection</i></p> <ul style="list-style-type: none"> • Statistical Approaches • Proximity-based and Density Based Outlier detection • Clustering-based Techniques
Session 19 (T 12/1)	<p><i>Text Mining</i></p> <ul style="list-style-type: none"> • Issues in Text Mining • Analysis of Document Collections
Session 20 (R 12/3)	Project Presentations
Session 21 (T 12/8)	<i>Final Exam</i>