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

BUSMGT 3332—Predictive Analytics

Autumn Semester 2016

Contact Information:

Instructor: Prof. Waleed A. Muhanna (muhanna.1@osu.edu)
Phone: (614) 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—but certainly not all—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 data sets—both small and large. 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 predictive analytics concepts and techniques, the most prevalent form of data mining; and (2) to provide you with hands-on experience in applying these techniques to practical real-word business problems using commercial 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 and evaluate results.

The techniques covered in this course fall into three major categories: (1) supervised learning techniques, including regression, decision trees and neural networks; (2) unsupervised learning

methods, including association rules mining, principal components analysis, and clustering; and (3) time-series forecasting. Students learn how to interpret and evaluate the quality of the predictive models produced, and how they might be able to combine different models to obtain results that may improve on the results that an individual model can produce on its own. The application of various 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. Describe different methods for predictive analytics;
3. Select an appropriate predictive analytics technique for a specific problem; and
4. Build, estimate, interpret and evaluate the results of prediction models.

Prerequisites:

The official prerequisites for the course are Econ 2001.01 or equiv. and Econ 2002.01 or equiv. and Math 1152 or equiv. and Stat 3202 or equiv. and CSE 2111 or equiv.

Course Materials:

- Textbook: *Data Mining for Business Intelligence*, 2nd Edition, by Galit Shmueli, Nitin R. Patel, and Peter C. Bruce (Wiley: 2010). The textbook is available "free of charge" in digital form via the OSU library (if you are off campus, you will need to provide OSU credentials to pass thru the library's proxy server) at:
  https://proquest-safaribooksonline-com.proxy.lib.ohio-state.edu/book/databases/business-intelligence/9780470526828
- A set of articles, assignments, tutorials, data sets, lecture notes, and various supplementary materials which will made available through the course website on Carmen.
- Software: XLMiner (see relevant section below)

Course Organization:

The course will be run as a mixture of lectures, in-class demonstrations, assignments, and classroom discussions. Readings will be from the required text together with other supplementary materials. Some material will be covered only in the 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 5-6 homework assignments, spaced out over the semester. 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 among students on homework and lab assignments is permitted. You may 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 seventh week of classes on **Tuesday, October 11th**. The second exam will be held during the final examination period on **Friday, December 9th (@ 8am)**. 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 3-4 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 case assignment will be made available as an alternative to team proposed projects. Teams will submit a written project proposal partway through the term, followed by a written report and, if time permits, brief class presentation on the project during the last class meeting.

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. (If you buy a new copy of the textbook, your copy should include a 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). Alternatively, I will provide you with a special Textbook Code and Course Code that will enable you to download the software, and use it throughout the term with a 140-day license.

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 with related topics. Your class

participation grade will reflect my judgment of the quality and quantity of your contributions during the entire term.

Cold calling: On occasion, I will make "cold calls". This is not intended to put you on the spot but to encourage class discussion and participation.

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.

Standards of Integrity and Conduct:

Academic integrity is essential to maintaining an environment that fosters excellence in teaching, research, and other educational and scholarly activities. Each student in this course is expected to be familiar with and abide by the principles and standards set forth in The Ohio State University's code of student conduct and code of academic conduct. You can view these documents or download pdf versions at:
http://studentaffairs.osu.edu/resource_csc.asp
http://www.gradsch.ohio-state.edu/academic-and-research-misconduct.html

It is also expected that each student will behave in a manner that is consistent with the Fisher Honor Statement, which reads as follows:

> *As a member of the Fisher College of Business Community, I am personally committed to the highest standards of behavior. Honesty and integrity are the foundations from which I will measure my actions. I will hold myself accountable to adhere to these standards. As a future leader in the community and business environment, I pledge to live by these principles and celebrate those who share these ideals.*

Students with Disabilities:

Any student who feels s/he may need an accommodation based on the impact of a disability should contact me privately to discuss your specific needs. I rely on the Office for Disability Services for assistance in verifying the need for accommodations and developing accommodation strategies. If you have special needs and have not previously contacted the Office for Disability Services, I encourage you to do so.

Tentative Course Schedule:

The following schedule gives the general plan for the course; changes may be made at my discretion but are designed to optimize the quality and flow of the content. 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	*Course Introduction* • Overview/goals of predictive modeling • Myths about data mining and predictive modeling • The Data Mining/Predictive Modeling process Readings: • Big Data: The Management Revolution, *HBR*, 10/12. • TB: Chapters 1 & 2
Session 2 & 3	*Data Extraction and Manipulation* • The Relational Data Model and Relational DBMS • Enterprise Reporting • Relational Algebra • SQL: The Relational Query Language
Session 4 & 5	*OLAP and Multidimensional Data Analysis* • Datawarehousing and Multidimensional Databases • Data Quality • Summarization and Data Cubes • OLAP Tools and Pivot Tables Readings: • (Check course web site) • "An Introduction to OLAP Multidimensional Terminology and Technology" (PDF)
Session 6 & 7	*Data Exploration and Dimension Reduction* • Data Summarization and Visualization • Correlation Analysis • Principal Component Analysis Readings: • TB: Chapters 3 & 4
Sessions 8 & 9	*Predictive Modeling Using Regression* • Review of OLS Regression • Logistic Regression • Model Evaluation and Interpretation

	Readings: • TB: Chapters 6 & 10
Sessions 10 & 11	*Evaluating a Model's Predictive Performance* • Over-fitting and Under-fitting • Performance Metrics • Lift Charts • ROC curves • Determining cutoffs for classification Readings: • TB: Chapters 5
Sessions 12 & 13	*Predictive Modeling Using Decision Trees* • Decision Tree induction • Model Evaluation and Interpretation Readings: • TB: Chapters 9 & 5
Session 14	*Predictive Modeling Using Neural Networks* • Introduction to Neural Networks • Neural Networks vs. Regression Readings: • TB: Chapter 11
Session 15 (10/11)	***Midterm Exam***
Session 16	*Ensemble Methods* • Condorcet Jury Theorem • Sample diversity: Bagging (bootstrap resampling) and Boosting Methods • Model diversity
Sessions 17 & 18	*Association & Market-Basket Analysis* • Frequent Itemset and Association Rule Mining • Pattern evaluation • Sequential patterns Readings: • TB: Chapter 13 • R. Agrawal and R. Srikant, "*Fast Algorithms for Mining Association Rules*," Proc. 20th Int. Conf. Very Large Data Bases (VLDB), 1994. (only skim)
Sessions 19 & 20	*Cluster Analysis* • Segmentation and Personalization • The K-means algorithm • Hierarchical (Agglomerative) Clustering

	• Cluster Validation and Interpretation Readings: • TB: Chapter 14
Sessions 21 & 22	*Time-Series Analysis and Forecasting: An introduction* • Principles and risks of forecasting • Getting to know the data • Visualizing time series • Time Series Components • Assessing forecast quality Readings: TB: Chapter 15
Sessions 23 & 24	*Time-Series Forecasting: the simplest models* • The mean (constant) model • Linear Trend Model • The random walk model • The random walk model with drift
Session 24 & 25	*Averaging and Smoothing Models* • Moving Average models • Exponential Smoothing Models • Combination of smoothing and seasonal adjustment Readings: TB: Chapter 17
Session 26 & 27	*Regression-based Forecasting* • Stationarity and differencing • Models with Trends and Seasonality • Lagged variables • Other explanatory/causal variables • Creating forecasts and Model testing Readings: TB: Chapter 16
Sessions 28 & 29	*Generalized framework for time-series forecasting: The ARIMA models* • Autocorrelation and partial autocorrelations functions • Identifying the order of differencing • Identifying the number of AR and MA terms • ARIMA models with regressors
(F 12/9)	***Final Exam***

About The Instructor:

Dr. Waleed A. Muhanna is Professor of Accounting & Management Information Systems at the Fisher College of Business, The Ohio State University. He received his undergraduate degree in computer science from the University of Tulsa, and holds a master's degree in computer science and doctorate in management information systems from the University of Wisconsin—Madison. Dr. Muhanna's teaching and consulting activities span a number of areas, with particular emphasis on e-commerce, data management and mining, and information systems strategy. Professor Muhanna's current research focuses on IT strategy, data analytics, assessing the business value of information technology, and understanding the impact of information technology, including the Internet, on organizations and markets. His other research interests include trust and reputation online, e-commerce strategy, model and database management systems, and system performance modeling and evaluation. Professor Muhanna has published numerous articles in scholarly journals, including *Management Science, MIS Quarterly, Strategic Management Journal, Decision Sciences, the Journal of Information Systems, the International Journal of Accounting Information Systems, ACM Transactions on Computer Systems, IEEE Transactions on Software Engineering, Communications of the ACM, Decision Support Systems, Information & Management, European Journal of Operational Research, Computers in Human Behavior,* and the *Annals of Operations Research.* Dr. Muhanna's teaching and consulting activities span a number of areas, with particular emphasis on e-commerce, data management and mining, business analytics, internet entrepreneurship, and information systems strategy. He recently completed a 3-year term as Chairperson of the Department of Accounting & Management Information Systems at the Fisher College of Business, and prior to that as the Director of the Ph.D. Program in Accounting & MIS. He also previously served as Vice-Chair of INFORMS' Information Systems Society and serves on the editorial boards of multiple leading academic journals.