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

ACCTMIS 7640 – Data Mining for Business Analytics

Spring Semester 2024, Term 2

Contact Information and Times:

Course Instructor: Prof. Waleed A. Muhanna (muhanna.1@osu.edu)
Phone: (614) 292-3808
Office: 420 Fisher Hall

Office Hours: Tuesdays 11:00am-Noon, 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 risk management and credit scoring, direct marketing, fraud detection, market-basket analysis of scanner data, customer relationship management, churn analysis, click-stream analysis, and personalization and recommendation systems.

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-word business problems. 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 (often, but not necessarily, large) data sets, to discover predictive and descriptive models and patterns. Data mining is related to statistics and to machine learning, but it has its own aims and scope. Statistics focuses on methods for making reliable inferences from imperfect data for the purpose of hypothesis testing, while machine learning mainly deals with methods and techniques for generating predictions, the most prevalent form of data mining. The techniques covered in this course fall into two major categories: (1) supervised learning techniques, including regression, decision trees and neural networks; and (2) unsupervised learning methods, including association rules mining and (if time permits) clustering. The application of these methods will be illustrated using modern software tools via examples, homework assignments and a group term project.

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

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

Prerequisites:

The key prerequisites consist of good graduate standing and completion of an introductory course in probability and statistics. (Assignments are best done using R, so prior exposure to and basic familiarity with R is helpful though not strictly required.)

Course Materials:

- **Primary text (DM)**: G. Shmueli, et al., *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*, or *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*,, Wiley (2018).

  - **Optional Supplementary text (APM):** Max Kuhn and Kjell Johnson, *Applied Predictive Modeling*, Springer (2013).
    - Companion APM Book Website and corresponding AppliedPredictiveModeling R package.
  - **Optional Supplementary text (ISLR)**: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer (2013).
    - Companion ISLR Book Website and corresponding ISLR R package.

- A set of articles, assignments, tutorials, data sets, lecture notes, and various supplementary materials, which will made available through the course website on Carmen. Please check the course website regularly to access newly posted materials, see when assignments are due and view reminders about the course.

Course Organization:

The course will be run as a mixture of pre-recorded and live (in-person) lectures, 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.

Mode of delivery: This class will be delivered using a traditional in-person (classroom) format. To allow more time during class to work through hands-on demonstrations, a significant number of pre-recorded lectures will be made available to you online, which you are expected to watch *before* coming to class.

Pace of online activities: This course is divided into weekly modules. Students are expected to keep pace with weekly deadlines but may schedule their efforts freely within that timeframe. Because a good portion of formalized instruction will occur asynchronously online, a higher level of self-discipline is required in order to successfully complete this course. You cannot afford to get behind. It is therefore strongly recommended that you set up a personal study schedule that specifically earmarks time when you will regularly work on the material for this course.

Credit hours and work expectations: This is a 1.5-credit-hour course. According to the standard suggested in University's Rules, specifically rule 3335-8-24, credits at Ohio State work on a 1 to 3 ratio: every 1 semester credit hour assigned to the class equates to total of 3 hours of work per week (1 hour of instruction and 2 additional study hours per week). Because this course is delivered over half a semester (7-weeks), this expectation doubles.

Assignments

In addition to the reading requirements from the text, there will be approximately 4 homework assignments, spaced out over the course of the 7-weeks term. They are designed to reinforce your understanding of the materials 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. While the homework assignments may vary in length and/or difficulty, each will be graded out of a possible 15 points. Your lowest homework score will not count toward your total assignment score in the course, so you can miss one homework without it counting toward your course grade.

Final Exam

There will be one (final) exam, which you will take in-class as a timed "quiz" on Carmen during the final examination period. The examination is 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 final exam must be taken during its scheduled time; make up exams will only be given for special and compelling

cases, in accordance with university guidelines. The precise format of the exam as well as time permitted are TBD.

Team-Based Term Project

Students will have the opportunity to put to work the tools and knowledge gained in this course, to sharpen their skills and acquire hands-on experience with real-world data sets through a term project. The project, which will be in the form of a case assignment, will be carried out in teams of 3-4 students. (If you have a project in mind that you would like to work on as an alternative, submit a written project proposal no later than the third week of classes, for my review and consideration.)

Attendance and 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 materials (readings and pre-recorded lectures). Each individual is strongly encouraged 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:

30% of the final grade will be based on graded homework assignments. The final exam will account for 40% of your grade. The group term project will account for 20% of the grade. The remaining 10% is assigned to class participation. Final grades will be based on overall class performance. Relative scaling is determined using the average grade across all students, and final course grades will be set such that this average is at a B+. No extra credit or makeup work will be offered.

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.

Restricted and Permitted Course Materials:

Use of inappropriate study materials, including previously prepared solutions to assignments and files containing tests used during previous terms, compromises the concept of equal opportunity for all students and is therefore prohibited. You may use materials that generally are available to all students provided that they maintain the spirit of the learning objectives.

Materials distributed to students via Carmen/email may be used only by students officially enrolled in ACCTMIS 7640 this term. You may not distribute any of these materials to any others at any time or be subject to disciplinary action.

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. If a student is suspected of, or reported to have committed, academic misconduct in this course, I am obligated by University Rules to report my suspicions to COAM. 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.

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. In light of the current pandemic, students seeking to request COVID-related accommodations may do so through the university's request process, managed by the Office for Disability Services.

Safety and Health Requirements:

All teaching staff and students are required to comply with and stay up to date on all University safety and health guidance, which may include wearing a face mask in any indoor space and always maintaining a safe physical distance. Non-compliance will be warned first, and disciplinary actions may be taken for repeated offenses.

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.

Weekly Module	Topic(s)	Main Readings
1. Course Introduction	• Course Introduction and Scope • Introduction to central themes of the class • Data Pre-Processing • Review of OLS Regression	R: DM §1, 6 S: APM §1, 2, 3 S: ISLR §2.1, 3.1, 3.2, 3.3
2. General Strategies: Assessing Model Performance	• Performance measures • The Problem of Over-fitting • Generalization Error and the Bias-Variance Tradeoff • Data Partitioning • Resampling Techniques	R: DM §2 S: APM §4, 5 S: ISLR §2.2, 5
3. Linear Regression and Classification Models	• Logistic Regression • Evaluating Classifiers • Linear Model Selection and Regularization	R: DM §10, 5 S: APM §6, 11, 12.2 S: ISLR §4.3, 6
4. Rule-Based (Tree) Models	• Decision Tree Induction • Preventing Overfitting • Tree Model Tuning	R: DM §8 S:APM §8.1-8.3, 14.1, 14.2 S: ISLR §8.1
5. Neural Network Models	• Introduction to Neural Nets • Building and Tuning Neural Network Models • Deep Learning • Preventing Overfitting	R: DM §11 S: APM §7.1, 13.2 S: ISLR §8.1
6. Ensemble Methods	• Model and Sample Diversity • Resampling Revisited • Boosting • Bagging • Random Forest Models	R: DM §13 S: APM §8.2-8.7, 14.2-14.6 S: ISLR §8.2, 8.3
7. Unsupervised Learning	• Association Rule Mining • Metrics: support and confidence • The Apriori algorithm • Rule Evaluation • Cluster Analysis	R: DM §14, 15
April 25 @6pm	*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 several 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 helped establish and served as academic co-director for Fisher's Specialized Master degree program in Business Analytics. He previously served 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 on the editorial boards of multiple leading academic journals.