The Ohio State University
The Max M. Fisher College of Business
Department of Accounting and Management Information Systems
AMIS 7640 - Data Mining and Analytics for Business Intelligence
1.5 Credit hours, Spring Semester 2021

Instructor and contacts:

Dr. Dawit Mulugeta | (901) 517-0887 | mulugeta.14@osu.edu

Class Time and Location:

Asynchronous. Online. Lectures will be out on **Carmen** and **YouTube** on Monday & Wednesday before 4:30 PM.

Instructor's Office Hours:

Zoom meeting. Wednesday 6:00 PM to 7:00 PM and by Appointment.

Office location and phone number:

Zoom | (901) 517-0887

Course Overview

Analytics:

We are at an exciting time where the era of bigdata, machine learning, artificial intelligence, cloud computing, advanced analytics is ushering unprecedented access to and uses of data to improve our predictive power to unlock transformational changes that impact many aspects of our lives in the retail, the financial, the manufacturing, the technology, the healthcare and other industries. The growth trajectories, the potential uses, the future resource allocations, and the employment opportunities in many of the data sciences fields are quite astounding. This is truly the best time to learn and to be on the bandwagon of analytics. AMIS 7640 will focus on the art and sciences of analytics with heavy emphasis on applied uses of data and best practices of analytics to make informed decisions.

As both big and small firms alike are fine-tuning their pricing, promotion, distribution, customer-retention, risk-management, and go-to-market strategies, business leaders are increasingly expected to know or be articulate about the many facets of data and analytics. Working on analytic projects, leading analytic teams, telling a story on any project that uses data require a reasonably good understanding and working knowledge of the inner workings and contemporary topics of analytics. Critical components to these include identification of talents, tools and techniques essential to generate, to acquire, to store and to use data to enhance ROI, to increase customer satisfaction and to sustain profitable growth.

Business leaders are often required to identify, to define and to find solutions to plethora of problems. Questions such as 1) What are the market segments, potential clients, and business opportunities for targeted delivery of goods and services; 2) How to conduct competitive intelligence for differentiated positioning, value propositions and key message development on products and services; 3) How to optimize prices, promotions, product assortments, and

marketing campaigns that translate organizational objectives into successful tactics; 4) How to identify, measure, and manage risks to enhance customer retention efforts, to avoid litigations, and to implement good business practices; 5) How to forecast trends, measure purchasing behavior and extract actionable insights from diverse data; 6) How to conduct tests and experiments to measure the effectiveness of new products, services, prices and promotional offers; and 7) How to be proactive to roll out winning marketing and distribution programs to increase loyalty and enhance customer retention. The answers to all these and several other questions that focus on improving productivity and profitability, cutting costs, reducing waste, and minimizing losses are embedded in data and in analytics.

Data Mining:

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 these 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.

Course Objectives

1. Acquire a theoretical and a practical knowledge of contemporary analytics techniques and core data mining concepts.
2. Develop skills to manage, to analyze, to summarize, to report, and to present diverse data with the intent of telling compelling stories of actionable insights.
3. Gain hands-on experience in applying analytic techniques to practical real-word business problems using commercial data mining software.

Expected Outcomes

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

1. Become knowledgeable of contemporary analytic techniques and data mining processes to drive initiatives and strategies on various analytic projects.
2. Use existing data retrieval and manipulation tools and techniques to identify opportunities, to solve significant business problems and to extract actionable insights from data.

3. Fully appreciate the concept of data as a strategic resource, and understand how, when and where data mining can be used as a problem-solving technique.
4. Interpret, evaluate and describe the results of analytics and data mining on a specific business issue.

Course Prerequisites

1. Good graduate standing and completion of an introductory course in probability and statistics.
2. Recognize the importance of analytics and data mining as a powerful information generation and decision-making tools.
3. Although students will use R for data manipulation and analytics throughout the course, neither prior programming knowledge nor information technology background is required.

Text / Readings

Reference One (extensively used)
Title: Data Mining for Business Intelligence: Concepts, Techniques, and Applications, Second Edition.
Authors: Galit Shmueli, Nitin R. Patel, and Peter C. Bruce.
Publisher: John Wiley & Sons (2010).
ISBN: 9780470526828.
The book is available in digital form via the OSU library at:
https://learning.oreilly.com/library/view/data-mining-for/9780470526828/?ar

Reference Two (extensively used)
Title: Python: Data Analytics and Visualization.
Authors: Phuong Vo.T.H, Martin Czygan, Ashish Kumar, and Kirthi Raman.
Publisher: Packet Publishing (2017).
ISBN: 9781788290098.
The book is available in digital form via the OSU library at:
https://learning.oreilly.com/library/view/python-data-analytics/9781788290098/?ar

A set of articles, assignments, tutorials, data sets, lecture notes, and various supplementary materials will be 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.

Readings will be from the required text together with other supplementary materials. Some material will be covered only in the readings; others 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.

Class Format

The teaching strategy of this course will be based primarily on lectures, in-class demonstrations, assignments, and classroom discussions.

Students can participate this course through CARMEN. Students are highly encouraged to visit the course site on CARMEN (https://carmen.osu.edu) regularly and print lecture materials in advance.

Throughout the course students are expected to bring their laptop into class. Each lecture will be complemented with associated and relevant work using the programming language Python. The laptop will be used to program with Python to manipulate data, to use the graphic interface, to develop various predictive and machine learning models, and to complete assignments.

As the field of data science and analytics encompasses several disciplines and are rapidly changing students are expected to read the selected reference materials and recommended readings for each topic together with the required textbooks.

Class Participation (10% of the final grade)

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.

Homework Assignments (40% of the final grade)

In addition to the reading requirements from the text and the supplementary materials, there will be five weekly homework assignments, spaced out over the course of the 7-weeks term. Each homework assignment is worth 8% of your final grade (5 X 8%=40%). 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 classmate's general solution strategies. However, everyone should independently do and turn in his/her own work.

Homework assignments for current week will be out on Wednesday. The answers must be submitted within one week, no latter than next Wednesday at 11:59 PM.

One Final Exam (50% of the final grade)

There will be one online final exam. The exam is open-book and open-notes, and it will be held in accordance with Fisher Graduate Programs schedule during the final examination period on Monday April 26, 2021. The exam will be timed, and may take about four hours.

The exam is designed to assess each student's (a) command of factual knowledge and concepts from the course; and (b) her or his 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.

Software

Python will be the programming language of the course. The class is tailored to ensure that students with no background of Python can learn the basics of coding and do most of the tasks required in the class. However, the students must be aware that **this is NOT a Python class**. The class uses Python as a means to an end (learning analytics). The instructor will provide Python scripts and demonstrate its use in class. Students are expected to use those, and expand their learnings by going beyond.

Python is a free software widely used for data manipulation, statistical computing, machine learning, graphics, web applications and many more. It is supported by Phyton Software Foundation as well as large number of users. Python is versatile, easy and fun to learn, has extensive libraries and is known to have large user base. It is one of the most extensively used software in the analytic community worldwide. Each lecture session will be accompanied by a demonstration of Python that focuses on specific tasks related to the discussion topic. 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.

Grading Summary

Class participation (10%).

Homework assignments (40%).

Final exam (50%).

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.

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.

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 includes wearing a face mask in any indoor space and maintaining a safe physical distance at all times. Non-compliance will be warned first and disciplinary actions will be taken for repeated offenses.

Disability Services

The university strives to make all learning experiences as accessible as possible. In light of the current pandemic, students seeking to request COVID-related accommodations may do so through the university's request process, managed by Student Life Disability Services. If you anticipate or experience academic barriers based on your disability (including mental health, chronic, or temporary medical conditions), please let me know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services. After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion. SLDS contact information: slds@osu.edu; 614-292-3307; slds.osu.edu; 098 Baker Hall, 113 W. 12th Avenue.

Module	Date	Lecture Topics & Readings
1	M (03/08)	**Course Introduction** • Course objectives • Instruction components • Syllabus, schedules • Grading, homework, exams **Introduction to Python** • History, status, what it is and what it is not • Resources and focus areas • Uses, basic syntax, tools • Getting the best out of it **Readings:** *Reference Two (Module 1: Chapters 1 & 2)*
2	W (03/10)	**Analytics and Current States of Affairs** • Terms, definitions, types • Pillars of analytics • Talent, tools and techniques

		• Roles, projects and trajectories **Readings:** *Reference One (Introduction & Chapter 1)* *Reference Two (Module 1: Chapters 1 & 2)*
3	M (03/15)	**Data and Data Mining** • Definitions, types and sources • Dimension, consumption and trajectories • Point of sales, digital, panel, hierarchical, timeseries, invoice • Distribution • Data mining • Architecture, design, and governance **Readings:** *Reference One (Chapter 2)* *Reference Two (Module 1: Chapters 2 & 3; Module 2: Chapter 2)*
4	W (03/17)	**Data Processing, Extraction, Aggregation, Manipulation, Summarization** • Rescaling, aggregation, zooming, filtering • Massaging (sorting, rearranging, transposing, merging) • Normalization. transformation • Imputation • Missing values and outlier detection and management • Visualization **Readings*:*** *Reference One (Chapters 3 & 4)* *Reference Two (Module 1: Chapters 3, 4 & 7)*
5	M (03/22)	**Predictive Model Development** • Standard processes and core principles • Model development and validation • Model deployment and documentation **Readings:** *Reference One (Chapter 8)* *Reference Two (Module 2: Chapters 1 & 4)*
6	W (03/24)	**Probability Theory and Practices** • Sample spaces, events and rules • Market Basket Analysis • Product Recommendation Systems **Readings:** *Reference One (Chapter 13)* *Reference Two (Module 2: Chapter 3)*

7	M (03/29)	**Linear Regression using the Ordinary Least Squares (OLS) Method** • Definitions and case studies • Assumptions and limitations • Training and validation • Model parameters assessment • Diagnosis **Readings:** *Reference One (Chapter 6)* *Reference Two (Module 1 Chapter 8; Module 2: Chapter 5)*
	W (03/31)	**Instructional Break - no classes, offices open (2 of 2, two-day break replacing spring break) W, March 31-Th, Apr 1, 2021**
8	M (04/05)	**Logistic Regression Theory and Practices** • Definitions, assumptions • Training and validation • Parameter estimation **Readings:** *Reference One (Chapter 10)* *Reference Two (Module 2: Chapter 6)*
9	W (04/07)	**Logistic Regression – Diagnosis and Accuracy Measurements** • Applications • Diagnosis • Model parameters assessment methods **Readings:** *Reference One (Chapter 10)* *Reference Two (Module 2: Chapter 6)*
10	M (04/12)	**Experimentation, Testing and Learning** • Definitions and rules • Assumptions and basic approaches • Types and classification schemes • Uses and trajectories **Readings:** *TBD*
11	W (04/14)	**Segmentation, Decision Trees and Clustering** • Decision tree elements, rules, and steps • Purpose and advantage • Clustering types, rules and choices **Readings:** *Reference One (Chapters 7 & 14)* *Reference Two (Module 2: Chapters 7 & 8)*
12	M (04/19)	**Ensemble methods: Bagging, random forests, boosting** • Definitions and uses • Case studies and demonstrations • Bootstrapping, classification, regression, • Algorithm enhancements

		Readings: *TBD*
13	W (04/21)	**Time-Series Analysis and Forecasting** • Forecasting theory and practices • Types and uses • Demonstrations **Readings:** *TBD*
14	W (04/26)	**Final Exam**