SCMA 6345 Business Analytics and Data Mining

Winter-Spring 2018 - Professor L. Douglas Smith

Office: 220ESH (Express Scripts Hall)
Office Phone: 516-6108
E-mail: ldsmith@umsl.edu

Home Phone: (314) 261-5014 (STL)
Mobile: (314) 556-7742
Seabeck, WA: (206) 493-2915

I am generally in my office and available by appointment during the day M-F. You may drop in at any time. Do not hesitate to call me at home, even late in the evening. I am happy to discuss concepts and to offer help over the telephone. You may also contact me via e-mail. Even when traveling, I check e-mail frequently where practicable.

Overview of the Course

This course concentrates on converting elemental data from various sources into business intelligence. The primary focus in the course is on predictive modeling for business decisions. Using data for actual business cases, we experiment with techniques for forecasting qualitative and quantitative outcomes, clustering entities in groups, and illustrating relationships. Finally, we briefly discuss applications of text analytics, mathematical optimization and discrete-event simulation for data-based decision making.

Within broad frameworks for business analytics and data mining, we employ a rigorous process (CRISP-DM) for framing business problems and constructing, testing and validating statistical models. Students perform analysis with professional versions of commercial or open-access software (Excel, SAS, including SAS Enterprise Miner, SPSS, IBM SPSS Modeler, Tableau, KNIME, R, etc.). They work in study groups to produce presentations for class discussion.

Student teams (with up to three members) will work together on mini-cases designed to illustrate course concepts and analytical techniques. Student teams will present the results of their analysis for class discussion. They will also work on term projects involving actual business cases for sponsoring organizations. Representatives of the sponsoring organizations will present the business case to the class, offer support and feedback as work progresses, and attend the student teams' final presentations of their findings on April 21.

Goals for the Course

Students are to acquire:

- An understanding of good practices for business analytics
- A mindset for determining relevant statistical measures, considering how and why they may be related and distinguishing random versus systematic variation in data
- An understanding of the statistical underpinnings of prominent techniques for predictive analytics
- Experience in working in teams with actual business-case data provided by managers and corporate analysts
- Hands-on experience in using leading commercial software for data mining and predictive analytics
- Experience in communicating the results of analysis to a managerial audience
- Awareness of other approaches to business analytics (text analytics, mathematical optimization and computer simulation).

Course Schedule:

The course is structured as a trilogy, each involving a pair of intensive Saturday class sessions, held in SSB 134 from 9AM to 4PM (with a lunch break) as follows:

1. Applied regression analysis for business analytics and forecasting on **January 20 and February 3**
2. Decision trees and logistics modeling for categorical data analysis and forecasting on **February 24 and March 10**
3. Applications of data mining techniques to business cases on **April 7 and April 21**.

This allows a month for analysis and support on term projects involving actual business cases, for which findings will be presented by student teams on **April 21**.

E-mail communications, telephone calls and Skype conversations with the instructor are encouraged between class sessions.

Course Prerequisites:

LOGOM5300 and INFSYS 5800 or equivalents.

Course References:

Lecture notes, articles, guides for analysis of business cases and web references for self-study will be provided for each class session. The textbook from any solid introductory course in business statistics for an MBA audience should give essential background in probability and statistics.

Reference Texts (available used or new from online sources):

Multivariate Data Analysis by J.F. Hair et al. (Prentice Hall various editions since 1998) provides good background on statistical techniques for business analytics and data mining.

A Second Course in Statistics: Regression Analysis by Mendenhall, William and Terry Sincich, (Prentice Hall, various editions) is good for fundamentals of linear statistical models.

Applied Statistics and the SAS Programming Language by Ronald P. Cody and Jeffrey K. Smith, (Prentice Hall, various editions) is a good reference for the use of basic SAS for data analytics.

Computer Software:

Extensive use will be made of Excel, Base SAS, SPSS, IBM SPSS Miner, and SAS Enterprise Miner -- all of which are installed for use in UMSL computerized classrooms and in the MIS lab on 2nd Floor ESH. Students may, in accordance with terms of academic use, install these software packages on their own laptop computers for use at home. Help with downloading is available from the UMSL Technology Support Center in Lucas Hall (314-516-6034). SAS requires 20+ GB of storage and can be downloaded most easily via Ethernet connection in 211 Lucas Hall.

Approximate Grading Scheme:

Case exercises done by study teams for class discussion 20%; Exams (with mini-cases to be analyzed by students individually) 50% ; Major term project and report from study team 30%.

Grade breaks for the final weighted scores usually occur about 80 percent for B+/A-, 70 percent for C+/B-, and 60% for D+/C-.

Course Topics

1. Introduction to BI/BA/DM
    a. Major applications of business analytics
        i. Risk assessment
        ii. Target marketing
        iii. Customer relations management
    b. General frameworks (processes) for business analytics and data mining
        - CRISP-DM (Cross-Industry Standard Process for Data Mining), KDD (Knowledge Discovery in Data), SEMMA (Sample, Explore, Modify, Model, Assess)
2. Producing relevant metrics and visualizing systematic relationships
    a. Geographical distributions and heat maps
    b. Frequency distributions, charts and tables showing percentage composition, Pareto analysis
    c. Illustrating changes of measures and relationships through time
3. Cross-tabulations and classification trees (decision trees and CHAID) for investigating relationships among measures, performing exploratory multivariate analysis, and clustering entities
4. Introduction to multiple regression analysis for investigating systematic variation in continuous measures using multiple variables (continuous and categorical)
    a. Tests of statistical significance to distinguish systematic versus random variation
    b. Mini case for forecasting seasonal sales patterns
    c. Mini case for estimating business potential of retail outlets
5. Introduction to logistic regression and classification trees (decision trees) for explaining and predicting categorical outcomes
    a. Mini case for predicting credit risk with binary or multiple alternative outcomes
6. Illustration of predictive analytics using statistical models
    a. Applications of decision trees for continuous and categorical outcome measures
    b. Applications of regression modeling for predicting continuous measures
        i. Accommodating nonlinear relationships with continuous and classification variables
    c. Applications of logistic regression for predicting categorical outcomes
    d. Applications of neural networks

7. Building predictive models using data-mining tools with large datasets
    a. Partitioning data for model construction, testing and validation
    b. Integrating and transforming data for statistical modeling
        i. Combining variables into relevant metrics
        ii. Incorporating nonlinear relationships
        iii. Handling cases with missing data
    c. Avoiding over-fitting with redundant information
    d. Measuring fit and comparing predictive performance from alternative approaches
8. Discussion of challenges with data integration
    a. Changes in business entities, business environments and managerial practice
    b. Matching data from different sources
    c. Geographical attribution
    d. Time-interval attribution
    e. Missing data
    f. Errors and outliers
9. Application of the CRISP-DM and SEMMA Process to data-mining case studies and class projects
    a. Articulation of project objectives and model-selection criterion
    b. Data integration and cleaning
        i. Checks for reasonable values and consistency
    c. Model development
    d. Model testing
    e. Triangulation and comparison of results from alternative modeling approaches
10. Illustrations of text mining
    a. Applications for market research and CRM
    b. Illustrations with SPSS Miner
11. Illustrations of data-based decision support
    a. Illustrations of applied optimizing methods
    b. Illustrations of discrete-event simulation.