Course Syllabus

Eco 5385.701
Predictive Analytics for Economists
Summer 2014
TTh 6:00 – 8:50 pm and Sat. 12:00 – 2:50 pm
First Day of Class: Tuesday, June 3
Last Day of Class: Tuesday, July 1
251 Maguire Building

This course is a follow-up to Eco 5350 Introductory Econometrics. Statistical methods used in engineering and computer science are introduced to complement the traditional economist’s toolbox of business and economics decision-making tools.

Purposes of Course:

There are several major purposes of this course. As the result of taking this course, the student should have an understanding of:

- The basics of **supervised learning** – prediction and classification
- **Prediction models** including multiple linear regression, artificial neural networks, regression trees, K-nearest Neighbors
- **Classification models** including logit/probit models, classification trees, Naïve-Bayes models
- **Model validation** by means of data partitioning
- **Methods of unsupervised learning** – exploratory data analysis, principal components, cluster analysis, association rules
- **Ensemble modeling**
- How to use standard **Data Mining Packages** including XLMINER and SPSS Modeler

Evaluation of the Student:

The evaluation in the class consists of four parts:
- Quick Quizzes (20%)
- Homework Exercises (20%)
- Take-home project to be turned in on Thursday, June 26 (10%)
- Mid-Term Exam (25%)
- Final Exam (25%)
The **Quick Quizzes** (QQs) will consist of a short answer and/or multiple-choice quiz that will be administered in the first five minutes of the class. It is meant to see if you have retained the information of the previous lecture and if you have done any assigned readings that I may have asked you to do. In addition to keeping the students current in the class and providing review material for the mid-term and final exams, the QQs allow me to keep track of student attendance. It has been my experience is that for each Quick Quiz a student misses before the mid-term exam the student, on average, loses 2.5% on his/her mid-term score. **The bottom line is that it pays to come to class!** I will be dropping your lowest QQ grade before calculating the QQ average.

With respect to homework exercises, students can confer with each other with respect to programming advice and discussion of basic ideas but in the final analysis each student is expected to write up his/her own homework answers and not make copies of others’ homework. Copying someone else’s homework to hand in as one’s own work is a violation of the SMU Honor Code and will be dealt with according to the rules of the SMU Honor Code. It is important to know that the homework assignments are very important in that the basic ideas covered by them invariably show up on the mid-term and final exams. If you know you are going to be missing a class on the day a homework exercise is due, hand in your homework in advance to receive full credit for your work. Any homework that is handed in late will be given a one letter grade reduction for each day of tardiness. I will be dropping the lowest exercise score before calculating the exercise average.

The **mid-term exam** will cover approximately 50% of the course material and the final exam is going to be comprehensive including to some degree the material from the mid-term exam. A comprehensive final of course forces the student to form an overall view of the course material and gives the student a chance to revisit what the student might not have fully understood before taking the mid-term. In my opinion, developing a overall view of how the pieces of this course fit together is an important accomplishment of the student.

At roughly the half-way point in class, I will assign a term project which the students are to complete individually. However, the students should feel free to consult the professor on any questions that he/she might have on the project. This project is governed by the SMU Honor Code. That is, you are to do your own work and not rely on the work of your classmates. You are to hand in the term project on Thursday, June 26.

---

**Additional Details**

**Required Textbook:** Data Mining for Business Intelligence (2nd ed., 2010) G. Shmueli, N.R. Patel, and P.C. Bruce

**Classroom Website:** [http://faculty.smu.edu/tfomby/](http://faculty.smu.edu/tfomby/)
Office: Room 301M, Umphrey Lee, 214-768-2559. E-mail address: tfomby@smu.edu

Office Hours: Tuesday and Thursday 3:30 – 5:00PM or by appointment.

Teaching Assistant: Jing Li, E-mail address: lij@smu.edu

Textbook and Computer Software:

The required textbook for this course is Data Mining for Business Intelligence by Galit Shmueli, Nitin R. Patel, and Peter C. Bruce, (Wiley, 2nd ed., 2010) hereafter referred to as SPB. This book includes complementary access to an EXCEL © add-in called XLMiner ©. In the back of the book you will find an insert that contains the license for downloading the add-in to your computer from the website www.solver.com/xlminer and using it for a six month period. This “education” version is slightly less equipped than the professional version of the add-in but it will be adequate for the work we will be doing in this course. Once you have registered your copy of the SPB textbook you will have online access to all of the datasets used as case studies in the textbook. We will also be using SPSS Modeler and the tutorial found at http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp. Access this software package can be obtained by direct download, through the Virtual Computer Lab (VCL) at https://apps.smu.edu/Citrix/AppsSMUWeb/ or on lab computers in the Economics department or other locations on campus.

Excused Absences:

Students will be excused from taking the mid-term exam or the final exam only with a note from a physician, or in the case of a death in the family, with a note from a parent or guardian. Even with an excused absence, either of these exams must eventually be taken before a course grade will be assigned to the student.

If you must miss a class due to legitimate circumstances beyond your control, be sure and contact me beforehand so that I will know of your circumstances. If excused, I will correspondingly excuse you from any QQ that is given that day. I want to emphasize that diligent attendance in this course is essential because a lot of the course material presented in class will be from my personal class notes and can’t be found in the textbook. Note: After 4 unexcused absences, I reserve the right to administratively drop students from the class.
Grading Scale:

My grading scale in this course is as follows:

- 92-100 A
- 90-91    A-
- 88-89    B+
- 82-87    B
- 80-81    B-
- 78-79    C+
- 72-77    C
- 70-71    C-
- 68-69    D+
- 62-67    D
- 60-61    D-
- 0-59     F

Excused Absences for University Extracurricular Activities:

Students participating in an officially sanctioned, scheduled University extracurricular activity will be given the opportunity to make up class assignments or other graded assignments missed as a result of their participation. It is the responsibility of the student to make arrangements with me prior to any missed scheduled examination or other missed assignment for making up the work.

Disability Accommodations:

Students needing academic accommodations for a disability must first contact Ms. Rebecca Marin, Coordinator, Services for Students with Disabilities (8-4557) to verify the disability and establish eligibility for accommodations. They should then schedule an appointment with the professor to make appropriate arrangements (See University Policy No. 2.4).

Religious Observance:

Religioustly observant students wishing to be absent on holidays that require missing class should notify me in writing at the beginning of the semester, and should discuss, in advance, acceptable ways of making up any work missed because of the absence (See University Policy No. 1.9).
Honor Code:

All SMU students are bound by the Honor Code (see SMU Student Handbook for a complete discussion of the SMU Honor Code). The code states that “any giving or receiving of aid on academic work submitted for evaluation, without the express consent of the instructor, or the toleration of such action shall constitute a breach of the Honor Code.” A violation can result in an “F” for the course and an Honor Code Violation on your transcript.

Topics

I. Introduction
   A. What is Data Mining?
   B. Terminology of Data Mining
   C. Types of Variables: Interval, Nominal (Unordered Categorical), and Ordinal (Ordered Categorical)
   D. Statistical Models vs. Machine Learning Models (Read Breiman article)
   E. Data Mining from a Process Perspective (Fig. 1.2 in SPB)
   F. Data Mining Methods Classified by Nature of the Data (Table 1.1 in SPB)

References: SPB, Chapter 1 and Breiman, Leo (2001), “Statistical Modeling: The Two Cultures,” Statistical Science, 16, 199-231. The Breiman article will be posted to the student by class e-mail.

II. Overview of the Data Mining Process
   A. Core Ideas in Data Mining
      i. Classification
      ii. Prediction
      iii. Association Rules
      iv. Data Reduction
      v. Data Exploration
      vi. Data Visualization
   B. Supervised and Unsupervised Learning
   C. The Steps in Data Mining
   D. SEMMA (SAS) and CRISP (IBM)
   E. Preliminary Steps
      i. Sampling from a Database
      ii. Pre-processing and Cleaning the Data
      iii. Partitioning the Data: Training, Validation, and Test data sets
F. Building a Model – An Example with Linear Regression

References: SPB, Chapter 2. SAS_SEMMA.pdf and CRISP_DM.pdf. These pdf files will be posted to the students by class e-mail.

III. Data Exploration and Data Refinement
   A. Data Summaries
   B. Data Visualization
   C. Treatment of Missing Observations
   D. Detection of Outliers – the Box Plot
   E. Correlation Analysis

Reference: SPB, Chapter 3.

IV. Variable Importance and Dimension Reduction
   A. Binning: Reducing the Number of Categories in Categorical Variables
   B. Principal Component Analysis of Continuous Variables
   C. Dimension Reduction using Best Subset Regression Modelling Techniques
   D. Dimension Reduction using Bivariate Association Probabilities (as in the “Feature Selection” node in SPSS Modeler), and Regression and Classification Trees

Reference: SPB, Chapter 4.

IV. Evaluation Methods for Prediction and Classification Problems
   A. Prediction Measures: MAE, MSE, RMSE, MAPE, MSPE, and RMSPE
   B. Classification Measures: Classification Matrix, ROC Curves, Lift Charts, and Lift Charts that Incorporate Costs and Benefits
   C. The Role of Over-sampling in Classification Problems

Reference: SPB, Chapter 5.

V. Prediction Methods
   A. Linear Regression: Best Subset Selection
      i. Forward Selection
      ii. Backward Selection
      iii. Step-wise Regression (Efroymson’s method)
      iv. All Subsets Regression (Cp Mallows and Adjusted R-square criteria)
   B. k-Nearest Neighbors (k-NN)
   C. Regression Trees
      i. CART
      ii. CHAID
   D. Neural Nets
      i. Architecture of Neural Nets
         a. Neurons
         b. Input Layer
c. Hidden Layers
  d. Output Layer
  ii. Fitting Neural Nets: Back Propagation
E. Comparison of the Four Methods
F. Model Averaging (Ensemble Model)

References: **SPB**, Chapters 6, 7, 9, and 11.

---

**Mid-Term Exam**
**Approximately**
**Thursday, June 19**

---

**VI. Classification Methods**
A. The Naïve Rule
B. Naïve-Bayes Classifier
C. K-Nearest Neighbors
D. Classification Trees
E. Neural Nets
F. Logistic Regression
References: **SPB**, Chapters 6, 7, 8, 9, and 10.

---

**VIII. Non-supervised Learning**
A. Association Rules
   i. Support and Confidence
   ii. The Apriori Algorithm
   iii. The Selection of Strong Rules
B. Cluster Analysis
   i. Hierarchical Methods
   ii. Optimization and the K-means Algorithm
   iii. Similarity Measures
   iv. Other Distance Measures
C. Text Mining

References: **SPB**, Chapters 13, 14, and Classroom Discussion of Text Mining.

---

**IX. Ensemble Methods**
A. Nelson and Granger-Ramanathan Methods for Continuous Targets
B. Majority Voting for Categorical Targets
C. Bagging
D. Boosting

Reference: Classroom lecture and handouts
FINAL EXAM

Tuesday, July 1, 6:00 – 9:00 PM
Room 251 Maguire Building