Course Syllabus

Eco 6380.701
Predictive Analytics for Economists
Spring 2016
W 6:30 – 9:20 pm
303 Lee 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, and Lasso models
- Classification models including logit/probit models, classification trees, Naïve-Bayes models, and Support Vector Machines
- Model validation by means of data partitioning
- Scoring Models on data sets with outcomes yet to be realized
- Methods of unsupervised learning – exploratory data analysis (EDA), 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 of the student consists of four parts:

- Quick Quizzes (15%)
- Exercises (15%)
- Two Mid-term Exams (25% each)
- A Term Project (20%) to be presented and turned in in PowerPoint form during the time of the scheduled final exam for this course (Wednesday, May 4, 6:30 – 9:30 PM).
The Quick Quizzes (QQs) will consist of an occasional 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 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. You should know that the homework assignments are very important in that the basic ideas covered by them invariably show up on the mid-term 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 exams will cover the topics indicated in the course outline and are not comprehensive.

The Term project: The term project will be done in groups of 4 – 5 persons per group. Its content will be determined in the first few weeks of the course. It will probably involve some separate data analysis projects coming from a recent collaboration that has been formed between our Dallas City government and SMU as part of a data-driven, “Smarter City” program. Each group will present a PowerPoint of its work during the 6:30 – 9:30 pm time slot for the Final Exam on May 4. The Term Project will be judged for content (70%) and quality of presentation (30%). The presentation time per group will be 20 minutes minimum to 30 minutes maximum.

My grading scale in this course is as follows:

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>92-100</td>
<td>A</td>
</tr>
<tr>
<td>90-91</td>
<td>A-</td>
</tr>
<tr>
<td>88-89</td>
<td>B+</td>
</tr>
<tr>
<td>82-87</td>
<td>B</td>
</tr>
<tr>
<td>80-81</td>
<td>B-</td>
</tr>
<tr>
<td>78-79</td>
<td>C+</td>
</tr>
<tr>
<td>72-77</td>
<td>C</td>
</tr>
<tr>
<td>70-71</td>
<td>C-</td>
</tr>
<tr>
<td>68-69</td>
<td>D+</td>
</tr>
<tr>
<td>62-67</td>
<td>D</td>
</tr>
<tr>
<td>60-61</td>
<td>D-</td>
</tr>
<tr>
<td>0-59</td>
<td>F</td>
</tr>
</tbody>
</table>
Classroom Website: http://faculty.smu.edu/tfomby/

Office: Room 301M, Umphrey Lee, 214-768-2559. E-mail address: tfomby@smu.edu

Office Hours: 3:00-4:30 PM MW or by appointment.

My Graduate Teaching Assistant: Igor Zhadan. His E-mail address is: izhadan@smu.edu. If you should need extra tutorials or help outside of my office hours, contact Mr. Zhadan and he will be happy to go over concepts that you may not fully understand.

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, when purchased as a new book as compared to used, 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://www.ibm.com/support/knowledgecenter/SS3RA7_16.0.0/com.ibm.spss.modeler.tutorial/clemetine/entities/clem_family_overview.htm?cp=SS3RA7_16.0.0%2F8&lang=en Access to this software package can be obtained through Apps.smu. To use SPSS Modeler on the Apps.smu system you will first need to download the Citrix Receiver. You can go to the website http://www.smu.edu/BusinessFinance/OIT/Services/AppsSMU and then, being a first time user, you will be prompted to download Citrix Receiver to your PC or laptop. Citrix Receiver provides you with “virtual” access to the SPSS Modeler software in that Citrix makes it appear that you have SPSS Modeler installed on your own computer when, in fact, it is being accessed from an SMU server on campus. After you install the Citrix Receiver on your computer, you can then logon to the Citrix Receiver by entering your student ID and personal password. Thereafter you can work on your homework assignments, etc. using SPSS Modeler.
Important Dates to Remember:

**First Day of Class:** Wednesday, January 20  
**MLK Day:** Monday, January 18 (No Class)  
**Spring Break:** Monday – Sunday, March 7 – 13 (No Classes)  
**Last Day to Drop a Course:** Wednesday, April 6  
**Last Day of Semester in this Class:** Wednesday, April 27.

**Exam Dates:**  
**First Midterm – Wednesday, March 16.** This is our first class following Spring Break, March 7 – 13. Exam will cover the topics covered in class up through Wednesday, March 2.  
**Second Midterm – Wednesday, April 27.** Exam will cover the topics covered in class since the first mid-term exam.  
**Final Exam Date:** Wednesday, May 4, 6:30 – 9:30. This is when the Term Projects are presented.

**General comments on work and class etiquette:**

In order to succeed in this class, constant work is essential. Come to class. Read all assigned readings, complete all exercises on time, and prepare for the Quick Quizzes. Don’t get behind. If there is something in class discussion or homework assignments that you don’t understand, don’t hesitate to ask me in class, after class, during office hours, or through e-mail.

Obviously, general rules of etiquette apply: cell phones are to be turned off during class and miscellaneous reading material stowed away.

**Some Standard Stuff You Should Know**

**Excused Absences for University Extracurricular Activities:**

Excused Absences for University Extracurricular Activities: Students participating in an officially sanctioned, scheduled University extracurricular activity should 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 the instructor prior to any missed scheduled examination or other missed assignment for making up the work. (University Undergraduate Catalogue)

**Disability Accommodations:** Students needing academic accommodations for a disability must first register with Disability Accommodations & Success Strategies (DASS). Students can call 214-768-1470 or visit [http://www.smu.edu/Provost/ALEC/DASS](http://www.smu.edu/Provost/ALEC/DASS) to begin the process. Once registered, students should then schedule an appointment with the professor as early in the semester as possible, present a DASS Accommodation Letter, and make appropriate arrangements. Please note that accommodations are not retroactive and require advance notice to implement.
Religious Observance:

Religious Observance: Religiously observant students wishing to be absent on holidays that require missing class should notify their professors in writing at the beginning of the semester, and should discuss with them, 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. The Distinct Purposes of Hypothesis Testing versus Prediction (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
   iv. Model Evaluation and Comparison of Models
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 and LASSO 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)
      v. Information Criteria (AIC, SBC, etc.)
   B. Penalized Regression Methods (Ridge, LASSO, Adaptive LASSO, and Elastic Net)
   C. k-Nearest Neighbors (k-NN)
   D. Regression Trees
i. CART  
ii. CHAID  
E. Neural Nets  
i. Architecture of Neural Nets  
a. Neurons  
b. Input Layer  
c. Hidden Layers  
d. Output Layer  
ii. Fitting Neural Nets: Back Propagation  
F. Comparison of the Various Methods  
G. Model Averaging (Ensemble Model)

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

First Mid-Term Exam  
Approximately  
Wednesday, March 16

VI. Classification Methods  
A. The Naïve Rule  
B. Naïve-Bayes Classifier  
C. K-Nearest Neighbors  
D. Classification Trees  
E. Neural Nets  
F. Logistic Regression  
G. Support Vector Machines (SVM)

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

Second Mid-Term Exam
   Wednesday, April 27

Term Project Presentations
   Wednesday, May 4
   6:30 – 9:30 PM
   303 LEE