Course Syllabus

Eco 6380.701
Predictive Analytics for Economists
Spring 2017
TTh 6:30 – 7:50 pm
110 Dedman Life Sciences 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 statistical methods. An emphasis of this course will be to demonstrate how this extended toolbox can be used to improve economic and business decision-making.

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** where predictions and classifications are made using combinations of models
- How to use standard **Data Mining Packages** including XLMINER, SPSS Modeler, and R.

Evaluation of the Student:

The evaluation of the student will consist of four parts:

- Quick Quizzes (20%)
- Homework Exercises (20%)
- Mid-term Exam (30%)
- Final Exam (30%) Thursday, May 11, 6:00 – 9:00 PM in Room 110 Dedman Life Sciences Building

The **Quick Quizzes** (QQs) will consist of occasional short answer and/or multiple-choice quizzes that will be administered in the first five minutes of the class. They are meant to reinforce the concepts presented in the previous lecture. 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 that the more Quick Quizzes missed by students, the lower their scores on the mid-term and final exams. **The bottom line is that it pays to come to class!** I will be dropping your lowest QQ grade before calculating the QQ average.

The purpose of **homework exercises** is to reinforce the concepts discussed in class. They invariably will be based on computer-oriented empirical problems using XLMINER, SPSS Modeler, and R. In completing the 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 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 homework exercise score before calculating the exercise average. Additionally, I want all homework handed in in **hardcopy form** – no pdf files sent to my e-mail address or the address of my teaching assistant. If you must send in your finished homework by e-mail, a point deduction (out of 10 points) will be applied to the student's exercise. **Also, I am expecting the homework to be typed as compared to handwritten. Handwritten homework will be given a one grade point deduction for not being typed.**

The **mid-term exam** will cover the topics in the first half of the course. The **final exam** will cover only the topics covered following the mid-term exam.

Note: After 4 unexcused class absences, I reserve the right to administratively drop students from the class.

My grading scale in this course is approximately 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

Additional Details

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 TTh 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, **3rd ed**., 2016) 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 ©**. I will be giving you more instructions on how to download the add-in to your computer in class. Later in class I will provide pdf files for operating XLMiner (XLMinerUserGuide_2016.pdf and XLMinerReferenceGuide_2016.pdf). In addition you can download, **free**, another book in pdf form that I will be referring to in class: **An Introduction to Statistical Learning with Applications in R** by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (Springer, 2015). The free download of the book can be obtained at www.StatLearning.com. We will also be using the SPSS Modeler software package. Later in class I will provide pdf files of the User's Guide for this software (SPSS_Modeler 17 UsersGuide.pdf) and Algorithms Guide (Modeler17 Algorithms Guide.pdf). 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.

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.**

Important Dates to Remember:
First Day of Class: Tuesday, January 24
Spring Break: Monday – Sunday, March 13 - 19 (No Classes)
Last Day to Drop a Course: Tuesday, April 11
Last Day of Semester in this Class: Thursday, May 4

Exam Dates:
Midterm Exam – Thursday, March 23. This is the second class following spring break.
Final Exam – Thursday, May 11, 6:00 – 9:00 PM in 110 Dedman Life Sciences Building.

Some Standard Stuff You Should Know

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 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: 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. Power Point 1 "Two Cultures"

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, and Power Point 2 "Data Mining Software." These files will be posted to the students by class e-mail and will be on CANVAS.

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

References: **SPB**, Chapter 3 and Power Point 3 "Missing Obs and EDA"

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

References: **SPB**, Chapter 4, Modeler 17 Algorithms Guide.pdf on "Feature Selection Algorithm" starting on page 153, XLMinerReferenceGuide.pdf on "Feature Selection Option" on pages 77 - 101, and Power Point 19 "Principle Component Analysis".

V. Evaluation Methods for Prediction and Classification Problems

A. Prediction Measures: MAE, MSE, RMSE, MAPE, MSPE, and RMSPE
B. Application to Validation and Test Data Sets
C. Avoiding Overtraining

References: **SPB**, Chapter 5, pp. 106 – 111 and Power Point 4 "Avoiding Overtraining".

VI. 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

References: **SPB**, Chapters 6, 7, 9, 11 and Power Points 5, 6, 7, 8, and 12.

Mid-Term Exam
Approximately
Thursday, March 23

VII. Evaluation Methods for Classification Problems

A. Classification Measures: Classification (Confusion) Matrix, Accuracy Measures, Profit Curves, ROC Curves, Lift Charts, and Lift Charts
B. The Role of Over-sampling in Classification Problems

References: **SPB**, Chapter 5, pp. 112 – 137 and Power Points 9, 10, and 11 on Evaluation of Classifiers.

VIII. 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, 10, and 11.

IX. Ensemble Methods

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

Reference: **SPB**, Chapter 13.

Non-supervised Learning Techniques

X. Association Rules

A. Support and Confidence
B. The A priori Algorithm
C. The Selection of Strong Rules

Reference: **SPB**, Chapter 14.

XI. Cluster Analysis

A. Hierarchical Clustering and Dendrograms

B. Non-hierarchical Clustering – the K-means Algorithm

Reference: **SPB**, Chapters 15

XII. Text Mining

A. Preprocessing the Data
B. Singular Value Decomposition (SVD)
C. Prediction with SVD variables

Reference: **SPB**, Chapter 20.

Final Exam
Thursday, May 11
6:00 – 9:00 PM
110 Dedman Life Sciences Building
