UNIVERSITY OF SOUTHERN CALIFORNIA

MARSHALL SCHOOL OF BUSINESS

DATA SCIENCES AND OPERATIONS DEPARTMENT

DSO 530 – APPLIED MODERN STATISTICAL LEARNING METHODS

FALL 2018

COURSE DETAILS

Professor Dr. Robertas Gabrys
Office BRI 401-O
Email gabrys@marshall.usc.edu
Office Hours Online or in person by appointment

COURSE OBJECTIVES

Over the last two decades, we have witnessed an explosion in the availability of data. Firms routinely collect point of sales transactions, monitor operating performance throughout their supply-chain, mine website traffic, and track customer engagement. Business analytics and data are transforming modern firms, and, in some cases, disrupting entire industries. Importantly, these changes are not limited to the "back-office" or operations; every aspect of the firm - organizational structure, marketing, product design, and strategic planning – is shifting towards data-driven decision-making. With this shift comes an increased need for "data-savvy" analysts; analysts who are not necessarily data-science experts, but understand what analytics can and cannot do, how to ask the right questions, and, most importantly, how to interpret data to make better decisions.

This course aims to go far beyond the classical statistical methods, such as linear regression, that are introduced in GSBA 516, GSBA 524, GSBA 545 or any other intro to stats course. As computing power has increased over the last 20 years many new, highly computational, regression, "Statistical Learning", and "Machine Learning" methods have been developed. In particular the last decade has seen a significant expansion of the number of possible approaches. "*Since these methods are so new, the business community is generally unaware of their huge potential.*" (**This was true 10 years ago!**) This course aims to provide a very applied overview to such modern non-linear methods as *Decision Trees, Boosting, Bagging, Support Vector Machines and Neural Networks* as well as more classical linear approaches such as *Logistic Regression, Linear Discriminant Analysis, K-Means Clustering, Hierarchical Clustering* and *Nearest Neighbors*.

We will cover these approaches in the context of Marketing, Operations, Finance and other important business decisions. At the end of this course you should have a fundamental understanding of how these methods work and be able to apply them in real business situations. With the explosion of "Big Data" problems, data science has become a very hot field in many scientific areas as well as marketing, operations, finance, accounting, operations, supply chain and other business disciplines. People with data science skills are in high demand!

To this end, approximately a half of the class time is dedicated to in class labs where the students will work through the methods we have covered, on their own laptops, under the supervision of the instructor. These labs will ensure that every student has a full understanding of the practical, as well as the theoretical, aspects of each method.

Several of the approaches we will cover in this course are new even in the data science community. Hence at the end of this course you will have truly innovative and important applied skills to market and differentiate yourself with.

LEARNING OBJECTIVES

At the end of this course, you will be able to:

- Explain in your own words the key ideas behind fundamental techniques in data analytics, including prediction, classification, and clustering
- Identify new opportunities to use these techniques across business domains to guide decision-making
- Confidently apply these techniques to novel problems using R
- Formulate and communicate actionable business recommendations based upon your analysis, including its limitations
- Critically assess the validity of analytics-based recommendations in the context of specific business decisions

COURSE MATERIALS

Course Book:

We will be using 2 books in this course:

- ***Data Mining for Business Analytics in R*** (2017) by Shmueli, Bruce, Yahav, Patel, and Lichtendahl

  The book's website is www.dataminingbook.com
  The book could be purchased on Amzon:
  https://www.amazon.com/Data-Mining-Business-Analytics-Applications/dp/1118879368/ref=sr_1_3?s=books&ie=UTF8&qid=1534806707&sr=1-3&keywords=%E2%80%A2%09Data+Mining+for+Business+Analytics+in+R

- ***An Introduction to Statistical Learning with Applications in R*** by James, Witten, Hastie, and Tibshirani.

  The book's website is http://www-bcf.usc.edu/~gareth/ISL/index.html
  Also, USC has subscription to Springer, so you should be able to access the book online: http://link.springer.com/book/10.1007/978-1-4614-7138-7/page/1

Additional resource:

I would also recommend that you obtain a copy of "*An Introduction to R*" by Venables and Smith which we will use as a manual for learning R. You can either purchase the book ($13 on Amazon!) or download it for free from http://cran.r-project.org/doc/manuals/R-intro.pdf.

Statistical Software:

Clearly a statistics package is essential for such an applied course. There are very few packages that can implement all of the different approaches we will cover. Of those that can, most are extremely expensive. The one that we will use in this course is R. R has several advantages. In addition to supporting all of the statistical learning methods we will cover, it is also the package of choice for research statisticians. This means that it is at the cutting edge with respect to new methods. R is also an extremely flexible program. For example, one can use R to write ones own functions to format data or implement new procedures. Finally, R is free so you can easily use it at any company that you may end up at! R can be downloaded from http://www.r-project.org/

Rstudio is a recommended interface for the R software. It is also free, and it runs on Windows, Mac, and Linux operating systems. http://www.rstudio.org

Please install both R and Rstudio on your computers! Students are expected to bring their laptops to class for all class sessions.

EVALUATION (i.e. Grades)

In line with the applied nature of this class, a large portion of the assessment will be made through homework. There will be approximately 8 homework assignments. The homework will contain some theory questions but the majority of the material will involve implementing the different methods that we cover in class using the computer package. There will also be a presentation on a group project and two in class tests. There will be no final exam. The breakdown of grades will be:

Homework	25%
Midterm Exam (Take home)	20%
Final Project: Proposal	5%
Final Project: Presentation	10%
Final Project: Report	15%
Final Exam (In class)	25%

HOMEWORK

Students will work on weekly HW assignments individually or in teams of 2. Homework assignments will provide an opportunity for you to develop and apply your data analysis skills to various business problems. In many ways, these assignments are a good example of the kinds of analytics work you may expect to do in your first job out of Marshall.

Answer the questions that you are asked clearly and concisely. Some questions will ask for code, specific numbers and/or calculations. To receive full credits, you must show your work. In some cases, you may wish to include a chart or graph. Please make sure to format it appropriately. Your scores on each assignment will depend on the quality and clarity of your submission. Finally, there may be questions that ask for you to make business recommendations based on your insights. Persuasive arguments tend to be brief. Long-winded answers often receive poorer scores.

MIDTERM AND FINAL EXAM

Both exams will involve conceptual and technical portions. Before each exam you will be provided with a list of topics you will be tested on and a sample exam. The final exam will be cumulative.

FINAL CASE PROJECT(=PROPOSAL+10 MIN HIGH LEVEL PRESENTATION+REPORT)

Students will work in teams of 2 to analyze a data set of your choice. This assignment will require you to apply a variety of data analysis techniques you've learned throughout the semester.

Your project will involve a PROPOSAL, a 10 minute PRESENTATION to the class of your findings and constructive feedback on other team's presentations and analyses, and a write-up/REPORT. Guidlines and requirements for the final project, including grading rubrics, will be distributed later in the semester.

CASES/DATA SETS STUDIED

We will work on a lot of different cases in this class. A few examples are listed below.

Case 1. PREDICTION OF FUTURE MOVEMENTS IN THE STOCK MARKET: Recently several of the statistical learning methods we discuss in this course (such as Boosting and Bagging) have been used to predict future values of financial markets. Such methods have obvious potential economic implications. We will investigate the performance of these methods on daily movements of the S&P500. We show that, while there is a weak signal, there are clearly some non-linear patterns that we can potentially exploit to predict whether the market will increase or decrease on each given day.

Case 2. PREDICTING INSURANCE PURCHASE: This is a very large data set recording whether a given potential customer purchased insurance or not. For each customer we have a record of 80 different characteristics and we wish to predict which customers are most likely to purchase insurance. Overall, only 6% of potential customers actually buy the insurance so if we randomly choose people to target our success rate is very low. However, using the methods from this course to target specific people we raise the success rate to around 30% (a five fold improvement).

Case 3. DIRECT MARKETING: This case involves a real dataset of direct mailings to potential donors of a not-for-profit organization. We wish to predict which people are most likely to respond so that the campaign can be better targeted. However, there is an extra wrinkle to this problem because those people that are most likely to respond also tend to give the least while, among those that are unlikely to respond, those that do tend to give the most. Hence we ultimately want to predict dollar giving.

Case 4. HOUSING VALUATIONS: This case involves understanding which variables (both macro and micro) affect housing valuations. For example what is the effect of house size, lot size, neighborhood, number of bedrooms, mortgage rates etc. on the value of a property.

Case 5. MARKETING OF ORANGE JUICE: This is a detailed data set with observations from many customers purchasing one of two brands of OJ. For each transaction many variables are recorded including, prices of each brand, promotions, discounts, which store the purchase was made at etc. The aim is to build a model for predicting which type of OJ a customer will purchase and which variables have an impact on the decision.

Case 6. EMAIL SPAM: Detecting whether an email is a SPAM based on relative frequencies of the 57 most commonly occurring words.

IMPORTANT DATES

Date	Assessment
Oct 19, 6:00 - 8:00 PM	Midterm (Take Home)
Oct 31	Project Proposal
Nov 28	Project Report
Nov 27, 29	Project Presentations (In Class)
TBD	Final Exam

STUDENTS WITH DISABILITIES

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP.

Please be sure the letter is delivered to me as early in the semester as possible. DSP is located in STU 301 and is open 8:30 am - 5:00 pm, Monday through Friday. The phone number for DSP is 213 740-0776.

COURSE OUTLINE

This is course is intended to cover the following topics:

1. **Introduction to Modern Data Science Approaches**
2. **Introduction to R and Data Exploration**
3. **Assessing the Accuracy of a Data Science Method**
4. **Linear Regression**
5. **Variable Selection**
6. **Logistic Regression**
7. **Linear Discriminant Analysis**
8. **Resampling Methods**
9. **Shrinkage and Dimension Reduction Methods**
10. **Generalized Additive Models**
11. **Tree Methods**
12. **Bagging and Boosting**
13. **Support Vector Machines (SVM)**
14. **Neural Networks**
15. **Clustering Methods**
16. **Presentations**
17. **Presentations**
18. **Final Exam**