University of San Diego

School of Business Administration

SBA Mission Statement:

> "We develop socially responsible business leaders with a global mindset through academically rigorous, relevant, and values-based education and research."

ECON / BUS 386 (01/02): BIG DATA & BUSINESS ANALYTICS

LOCATION: OLIN HALL 330
TIMES: MONDAYS & WEDNESDAYS, 1:00-2:20PM (01) / 2:30PM-3:50PM (02)

SPRING QUARTER 2020
SYLLABUS

Instructor: Dr. Steve Levkoff, PhD, CAP®
E-mail: slevkoff@sandiego.edu
Instructor Webpage: http://stevelevkoff.com
Course Webpage: http://ole.sandiego.edu
Offices: Olin Hall Room 234
Office Hours: Mondays & Wednesdays, 12pm-1pm, open door, and by appointment

Course Description: Analytics is the process of transforming data into insight in order to make better informed decisions. Understanding and interpreting data has become an even more integral part of understanding social interactions and behavior since the advent of big data and automated extraction. Accordingly, this lab-style course will provide a solid foundation for understanding data science and analytics problems in the context of modern big data methodology, philosophy, and application to business problems. Topics include, but are not limited to, database & repository management; scripting & automation; scraping, cleaning, and harmonizing data; exploratory analysis & data visualizations, documentation & reproducibility, ethics & client interactions, practical machine learning algorithms (ranging from multiple linear regression to neural networks and support vector machines), optimization, regularization, generalization, and validation. By the end of the course, you will be able to extract, clean, and harmonize data to use in a predictive algorithm that you will be able to build yourself as part of a data product application.

Course Objectives: The primary objective of the course is to develop understanding of modeling methodologies and applications in analytics problem solving. Upon completion of this course, students should be able to:

- Manage a basic GitHub data repository
- Extract, clean, harmonize, and rescale a data set.
- Conduct an exploratory analysis of the data from which to gain basic insight
- Formulate an important question based on exploratory analysis
- Build and test a predictive model using validation and regularization
- Build a data product that makes client and server side calculations using the predictive model in an attempt to resolve the question proposed

Student Learning Outcomes: This course satisfies the requirement of an undergraduate Core class in Social & Behavioral Inquiry (SBI). The goals and learning outcomes of a SBI class are:

Goal 1: Inquiry: Students will use a disciplinary toolkit of theories and methods to analyze claims and develop informed judgments by empirical analysis of data.

Learning Outcomes:

- Framing a problem and identifying stakeholders, relevant data needs, and appropriate tools.
- Articulate and compare methodologies as they apply to data science and analytics.
- Analyze data using descriptive, prescriptive, and predictive methods
- Evaluate the quality, objectivity, and credibility of evidence in analyzing claims using data.
- State a conclusion that is a logical extrapolation from the inquiry process.

Goal 2: Application: Students will apply the tools of analytics in evaluating an issue of relevance.

Learning Outcomes:

- Apply the discipline-specific inquiry process to analyze a new set of events/fact patterns representing real world problems or issues

Prerequisites: In order to be enrolled in this course, you should have satisfactorily completed ITMG 100, some statistics (ECON 216 OR 217), and some calculus (MATH 130 OR MATH 150). Some experience scripting/programming in Python, MatLab, or R (or any object oriented/data programming) will be useful (the course uses R to stage much of the applied foundation), but not required as the beginning of the course will lay these foundations in out-of-class lab exercises.

Readings:

Required:

- [1] The Art of Data Science, Roger D. Peng and Elizabeth Matsui, Leanpub.
- [2] R Programming for Data Science, Roger D. Peng, Leanpub.
- [3] The Elements of Data Analytic Style, Jeff Leek, Leanpub.
- [4] Exploratory Data Analysis with R, Roger D. Peng, Leanpub.
- [5] Report Writing for Data Science in R, Roger D. Peng, Leanpub.
- [6] Statistical Inference for Data Science, Brian Caffo, Leanpub.
- [7] Developing Data Products in R, Brian Caffo, Leanpub.

All of the *above* readings can be found online *for free* (or you can donate) at https://leanpub.com.

Recommended / Advanced:

- [8] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. A free resource and excellent core text that can be found here: http://statweb.stanford.edu/~tibs/ElemStatLearn/.
- [9] Learning from Data, Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. More technical treatment of the learning problem. Theoretical discussion of overfitting, the theory of generalization, and the linear model. Of importance are the associated ebook chapters (found online with purchase of the book) on Neural Netoworks, Support Vector Machines, and Similarity Based Methods.
- [10] Econometrics, Jeffy R. Wooldridge. A good "cook book" approach to introductory econometrics and regression analysis.

Software Packages and Repositories:

Required:

- [1] R, a statistical computing package that can be downloaded here: https://www.r-project.org/.
- [2] RStudio, An integrated development environment (IDE) for R, an open source statistical computing software package. You can download the most recent version (compatible for Mac and Windows) for free at www.rstudio.com. All other packages and extensions can be loaded seamlessly from within the IDE.

Computers in the Olin Hall lab have RStudio and R loaded on them.

- [3] Git (or GitBash), A software configuration management system (SCM) that will be used to implement version control and database management with your github account. You can download the command line interface shell here for free: https://git-scm.com/downloads.
- [4] GitHub, an online repository system that maintains version control of branched repositories. You will need to start an account (free of charge, unless you want to upgrade and have private/locked repositories or increase memory storage) that you will maintain and post your completed project analysis to for assessment. https://github.com/join?source=header-home

Swirl Lessons: Swirl, a series of open source (free) R packages, is a self-instructional programming tutorial that is run through the RStudio IDE. The SWIRL package exercises function to quickly train you in RStudio and in the R programming syntax. Completion of SWIRL assignments for the R Programming series are sent via email notification to the instructor. You can find the instructions for installing the swirl packages here: http://swirlstats.com/students.html. The five packages that will be used in this course are R Programming, Getting and Cleaning Data, Exploratory Data Analysis, Statistical Inference, and Regression Models. The swirl course repository can be found here: https://github.com/swirldev/swirl_courses#swirl-courses. There may also be several other benchmark submissions that will be required to be posted to your GitHub repository and shared with the instructor for grading aside from the swirl exercises (see projects below). The R Programming E SWIRL modules (15 of them) are due for completion on 2/9/20. It shouldn't take more than a few hours to complete these modules to ensure you have a basic understanding of R. I will post instructions for installation as well as instructions for submission that must be followed *extremely carefully*.

Problem Sets: While not a formal part of the course grade, problem sets will be assigned regularly to provide examples for conducting data analysis utilizing R and RStudio. Detailed solutions will be provided along with the R code chunks used to generate the results.

Prediction Project: There will be one group project throughout the course. The project will entail two predictive tasks: a regression task (continuous output variable) and a classification task (discrete output variable). Your group will need to find and clean datasets appropriate for each (they should be different datasets for different tasks), propose predictive models, and validate the performance of these models to choose the best model. Your group will submit an R Markdown report summarizing your findings and submit a presentation that may be presented towards the end of the course (time permitting).

OLE Access (course webpage): It is your responsibility to make sure you are enrolled in the online course (OLE) and to routinely check it and your email for announcements and to access newly distributed material.

Examinations: There will be three exams given – two in-class midterm exams scheduled for 2/24/20 and 4/6/20. The exams will encompass both conceptual and The final examination is scheduled by the registrar for 5/18/20 (section 02) and 5/20/20 (section 01) from 2-4pm. Assume all exams are cumulative (but not necessarily uniformly so).

Grading:		
	SWIRL (R Programming)	15%
	Prediction Project (Report / Presentation)	25%
	Midterm Exam 1	15%
	Midterm Exam 2	20%
	Final Exam	25%
	Total	100%

The course is graded on a relative curve (as is any college course). In particular, students will all be ranked from highest to lowest course score according to your final course grade calculated from the raw exam score weighting above. Letter grade assignments will depend on your percentile ranking in the class and a subjective assessment by the instructor in borderline cases (say, if there was marked improvement).

In the past, a student could typically guarantee themselves some type of A by ranking in the top 25% of students in the course and some ty pe of a B by ranking in the top 60%. Typically, the median score for the course curves to a B/B-.

Absences & Attendance: Any exam or quiz missed for a *legitimate, university approved* reason may be made up at the discretion of the instructor (this may include an oral evaluation as an alternative to taking a written exam or a re-weighting of the exams in the grade calculation shown above). You will receive a zero on any exam or quiz missed without a legitimate reason.

Supplemental Material & Slides: Throughout the course, the instructor may post supplemental readings and slides via OLE. These materials are meant to be used in addition to the lecture and are not to be used as a substitute for going to lecture or reading the textbook. The instructor reserves the right to remove access to this material if he feels that it has adversely affected attendance in the lecture.

Classroom Decorum & Email: To avoid distracting others in the classroom, please arrive on time and do not leave early unless given prior permission. When class is in session, please respect others in the room and refrain from sending or receiving phone calls, pages, or text messages. Please be sure audible signals are turned off before class begins. Please restrict the use of email to the minimally necessary volume and put your full name at the end of email messages and the course name and number in the subject heading. Questions regarding course policies will be directed to the syllabus (if applicable). All other general questions are welcome!

Statement of Academic Integrity: Integrity of scholarship is essential for an academic community. The University expects that both faculty and students will honor this principle and in so doing protect the validity of University intellectual work. For students, this means that all academic work will be done by the individual to whom it is assigned, without unauthorized aid of any kind.

Examination Policies: Consistent with the University's mission to preserve academic integrity, there are several policies and procedures that must be adhered to by students during exams.

1) In order to be allowed into the exam, students must have:
    - A BLUE or BLACK PEN (NO PENCILS!!! If you use pencil, you forfeit your chance for a re-grade.)
    - A BASIC or SCIENTIFIC calculator (no graphing calculators, cell phones, or other mobile devices unless given prior approval by the instructor)
    - Your USD student ID
2) During the exam, the following policies will be enforced:
    - Your seat will be randomized for each exam. When you enter the lecture on exam day, find your name and assigned seat number on the projector and quickly and quietly sit. Once everyone is seated, the exam will be handed out.
    - NO BATHROOM BREAKS (for exams <1.5 hours). Be sure to use the restrooms before the exam begins. Exams are less than an hour and a half long! You can make it!!! For longer exams, bathroom breaks will only be allowed (one at a time) during the first half of the exam duration.
    - No hats, hoodies, or sunglasses during the exam.
    - Turn cell phones off during the exam and leave them in your bag.
3) Violations of academic integrity will not be tolerated. For this course in particular, violations include, but are not limited to <u>anything that may be perceived as the following actions</u>:
    - looking at or copying from other students' exams
    - writing answers after time has been called

- talking during an exam while exams are still out
- looking at notes during an exam
- taking the wrong version of an exam
- removing an exam from the examination room
- removing pages from an exam
- falsifying identification or an exam book during or after the exam
- sitting in the wrong seat during an exam (if applicable)
- using an unapproved device/item during an exam (ie: programmable calculator, cell phone, etc. - see above list)

Violation (or perceived violation) of any of the abovementioned policies will be enforced via zero tolerance and referred to the student conduct process, so don't do anything that would even come close to something that an observer would potentially interpret as academic dishonesty. NO EXCEPTIONS.

Tentative Schedule of Topics (Subject to Change):

Part I: Getting Started: Programming, Tools, Simulation, and Cleaning Data (1-3 weeks)

- Business Problem and Analytics Problem Framing
- Predictive vs. Prescriptive and Descriptive Analytics
- What is Learning?
- Elements of a Predictive Problem
  - Supervised vs. Unsupervised Learning
  - Model Estimation as an Optimization Problem
- Version Control Software & Statistical Packages
- Pushing and Cloning to GitHub account
- Importing Data in RStudio
- Descriptive Statistics & Exploratory Analysis in R
  - Plotting Packages
- Simulation and Random Sampling in in R

Part II: Predictive Analytics I: Linear Models (3-4 weeks)

- Linear Regression Revisited
  - Linear Regression in Matrix Form
    - Some Linear Algebra
  - Review: Assumptions and Properties of the Model
  - Review: Hypothesis Testing About Parameter Values
  - Nonlinear Transformations
  - Heteroskedasticity and Multicollinearity
  - Analytics Solution: 1-step learning and the pseudo-inverse

Part III: Model Selection & Performance Analysis (1-2 weeks)

- In-sample vs. Out-of-sample error
- The Bias-Variance Tradeoff: The Statistical Perspective
- Generalization Error: A Machine Learning Perspective
- Validation
  - Cross-Validation, K-fold, LOO/Jackknife
- Regularization
  - RIDGE, LASSO, etc.

Part IV: Machine Learning & Iterative Methods (3-4 weeks)

  - Bisection Algorithm
  - Gradient Descent
    - Stochastic, Mini-batch, Full-batch
- Logistic Regression for Classification
  - As an application of Gradient Descent
  - Theory: Cross-entropy Error Minimization
  - Application: Logistic Regression in R
- Quick Look: Neural Networks (classification)
  - Theory: Forward & Backward Propagation
- Quick Look: Support Vector Machines (classification)
  - Theory: Hard & Soft Margin
  - Theory: Nonlinear Kernel Transformations
  - Application: SVM in R
- Quick Look: CART
  - Resampling, Bootstrapping, and Random Forests
  - Applications: CART and Random Forests in R
- Quick Look: Unsupervised Learning
  - Clustering: K-means, X means, Hierarchical
  - Anomaly Detection
  - Similarity Based Methods