ISE 535: Data Mining
3 Units

Day/Time: Tues/Thurs, 3:30 – 4:50PM
Location: ZHS 159

Instructor: Bruce Wilcox
Office: GER 203
Office Hours:
>   Tuesdays, 2:00PM – 3:00PM GER 203
>   Wednesdays, 2:00PM – 3:00PM GER 206
>   Thursdays, 2:00PM – 3:00PM GER 206
>   See Piazza for weekly updates

Contact Info:
Please use Piazza for all course communications
Email: brucewil@usc.edu

Teaching Assistants:
TBA

Catalog Description

Data preprocessing, data cleaning, data summarization, data visualization, and predictive modeling for classification and regression; modeling dependencies using association rules.

Course Description

Data mining is the discipline of extracting useful insights from large quantities of data. As such, the focus in this class is on inference and not on prediction (which is the topic of ISE-529).

This course is organized into three broad sections:

*   Data preprocessing, data wrangling and data cleaning to prepare data for analysis.
*   Exploratory data analysis and statistical data analysis techniques to find useful information about the data.
*   Algorithmic data mining techniques of classification, clustering, association rule mining, linear modeling for inference, and tree-based modeling for inference.

To the maximum extent possible, this course teaches the concepts by means of case studies using actual or simulated but realistic business data.

Learning Objectives and Outcomes

- Develop an advanced level of proficiency with the preprocessing, visualization, and statistical analysis of data as well as several of the primary data mining algorithmic techniques.
- Develop skills in using the R programming environment and some of its packages that are broadly used in industry by data scientists (primarily the Tidyverse packages)
- Review and re-enforce basic statistical concepts that are important in the field of data science.

Class Delivery Mode: This class will be conducted in a fully in-person mode. Lectures will be broadcast and recorded using Zoom to accommodate students who are ill and those who wish to re-view portions of the lecture after the class. The mid-term and final exams must be taken in person.

Prerequisite(s): None

Recommended Preparation: ISE 225 (Engineering Statistics I) or equivalent, working knowledge of Python

Course Notes: All course materials (PowerPoints, assigned readings, etc.) will be distributed via Blackboard.

Technological Proficiency and Hardware/Software Required

This course will utilize the R programming language and the R Studio Integrated Development Environment (IDE) which are open source and available to the students for no cost.

Required Readings and Supplementary Materials

The theoretical material in the course is drawn from the following texts:
- Scmueli, et. al., *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R,* Wiley, 2017 (DMBA)
- James, et. al., *An Introduction to Statistical Learning with Applications in R*, Springer, 2017 (ISLR)
- Bruce, et. al., Practical Statistics for Data Scientists, O'Reilly, 2020 (PSDS)

In addition, the following text will be used as our reference for R programming:
- Wickham, *R for Data Science*, O'Reilly, 2017 (RDS)

Grading Breakdown

Grading will be based on four primary components:
- 8-10 homework assignments (approx. one per week) - 50% of final grade
- Mid-term exam (in class) – 20% of final grade (covering Modules 1 – 4)
- Final exam - 30% of final grade

Grading Scale

Course final grades will be determined using the following scale

A   95-100
A-   90-94
B+   87-89
B   83-86
B-   80-82
C+   77-79
C   73-76
C-   70-72
D+   67-69
D   63-66
D-   60-62
F   59 and below

Up to two points may be added to the overall grade based on class engagement.

Assignment Submission Policy

Assignments will be posted on Blackboard and submitted for grading on GradeScope (student instructions will be provided)

Timeline and Rules for Submission

- Homework assignments will be posted on or shortly after each weekly class and will be due one week after posted
- All assignments must be submitted prior to the due date
- You get one "free" late submission per semester. After that, there is a 10% penalty for late homework submitted within 48 hours of due date
- No submissions will be accepted after 48 hours
- The lowest homework grade will be dropped

Course Communications

- All materials will be uploaded to Blackboard
- Assignments will be submitted through Gradescope
- We will use Piazza as the primary communications mechanism
    - Class announcements will be posted there, and we request that any questions you have be posted there so that other students can benefit from your question and responses from the instructors, TAs, and hopefully other students
    - Students who actively post responses to questions MAY receive extra credit (which could result in an increase by one letter grade in borderline cases)
- I will periodically post "discussion questions" on Piazza. Class engagement credit can be earned by participating in these online discussions

Course Schedule: A Weekly Breakdown

Week	W/E	Topics/Daily Activities	Assignments	References
1	8/25	**Module 1: Introduction** Introduction to Data Mining Introduction to R, RStudio, and R Markdown.	Module 1 HW Assigned	Class notes RMD DMBA Ch. 1&2
2	9/1	**Module 2: Data Preparation** Data integration, cleaning, reduction, enhancement *Tools: Tidy Data, Tidyverse, DPLYR, Tibble*	Module 1 HW Due Module 2 HW Assigned	RDS, Section 3 ("Wrangle") DMB, Ch 4
3	9/8	**Module 3: Exploratory Data Analysis (EDA)** Univariate/bivariate analysis, data quality assessment *Tools: ggplot*	Module 2 HW Due Module 3A HW Assigned	PSDS, Ch 1 RDS, Section 2 ("Explore") DMBA, Ch 3
4	9/15		Module 3A HW Due Module 3B HW Assigned	
5	9/22	**Module 4: Statistical Data Analysis** Data and sampling distributions, statistical experiments, significance testing, time series forecasting	Module 3B HW Due Module 4A HW Assigned	ISLR, Ch 2 & 13 PSDS, Ch 2 & 3
6	9/29		Module 4B HW Due Module 4B HW Assigned	
7	10/6		Module 4B HW Due Module 4C HW Assigned	
8	10/11	**Mid-Term (October 11)** No class 10/13	Module 4C HW Due	
9	10/20	**Module 5: Unsupervised Learning/Clustering**	Module 5 HW Assigned	ISLR, Ch 12, PSDS, Ch 7 DMBA, Ch 15
10	10/27	**Module 6: Classification**	Module 5 HW Due Module 6 HW Assigned	ISLR, Ch 4
11	11/3	**Module 7: Association Rule Mining**	Module 6 HW Due Module 7 HW Assigned	DMBA, Ch 14
12	11/10	**Module 8: Linear Modeling for Inference**	Module 7 HW Due Module 8A HW Assigned	ISLR, Ch 3
13	11/17		Module 8A HW Due Module 8B HW Assigned	
14	11/22	**Module 9: Tree-Based Modeling for Inference** No class 11/24	Module 8B HW Due Module 9 HW Assigned	ISLR, Ch 10
15	12/1	**Course Review**	Module 9 HW Due	
Final		**Final Exam – 12/12 2:00PM – 4:00PM**		

Notes:
- This schedule is subject to change throughout the semester. This syllabus will not be updated, but the latest schedule will always be available on Piazza.

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the Office of Equity and Diversity http://equity.usc.edu or to the Department of Public Safety http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. The Center for Women and Men http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems:

Student Health Counseling Services - (213) 740-7711 – 24/7 on call
engemannshc.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-4900 – 24/7 on call
engemannshc.usc.edu/rsvp
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) | Title IX - (213) 740-5086
equity.usc.edu, titleix.usc.edu
Information about how to get help or help a survivor of harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following protected characteristics: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations.

Bias Assessment Response and Support - (213) 740-2421
studentaffairs.usc.edu/bias-assessment-response-support
Avenue to report incidents of bias, hate crimes, and microaggressions for appropriate investigation and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy - (213) 821-4710
studentaffairs.usc.edu/ssa
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.
Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.