ISM 6137: Statistical Data Mining
Sec 001, CRN 19668, M 8:30a-12:15p, BSN 230
Sec 902, CRN 22648, M 6:30p-10:15p, BSN 2304
Spring Semester 2023

Professor: Dr. Anol Bhattacherjee	*Office:* CIS 2065
E-mail: ABhatt@usf.edu	*Office Hours:* M 5-6 pm or by appointment
Website: http://ab2020.weebly.com	*Office Phone:* (813) 974-6760
Teaching Assistant: Enock Chemochek	*TA E-mail:* echemochek@usf.edu

Course Overview:

This is the second course in a two-course sequence on statistical modeling of data. It will examine advanced statistical models for handling different kinds of data where linear regression is not applicable, such as count data, mixed-level data, classification data, survival data, and temporal data, using the R language. More than learning specific tools, the focus of the course will be on thinking about, structuring, and solving industry-grade problems business analytics problems in retail, service, healthcare, and other sectors. This is a "thinking" course that will require a significant amount of time, effort, and attention on your part. Prior knowledge of linear models and R are required for this class.

Learning Objectives:

To learn how to:

- Frame, structure, and systematically complex business analytics problems.
- Plan, conduct, and document real-world analytics project from start to end.
- Model count, mixed-level, classification, temporal, survival, and other types of complex data that don't lend themselves to simple linear regression.
- Preprocess data and engineer features from the data before analysis.
- Interpret these models meaningfully to generate useful and usable insights.
- Examine if our models are robust and will hold up in the real world.
- Use data and analytics correctly and ethically.

Course Prerequisite:

The prerequisite for this course is QMB 6304 (Analytical Methods for Business). Students must be comfortable with building different kinds of linear models in R prior to this class.

Grading:

Grade components and weights:

Individual Assignments (eight)	50%
Exams (two)	35%
Team Project (one)	15%

Grading scale: A+: Over 97%, A: 93-97%, A-: 90-93%, B+: 87-90%, B: 83-87%, B-: 80-83%, C+: 77-80%, C: 73-77%, C-: 70-73%, D+: 67-70%, D: 63-67%, D-: 60-63%, F: Below 60%

Books and Materials:

Textbook: None. Materials for this class has been sourced from many books and online/offline sources as well as the professor's personal knowledge of statistical modeling.

Readings, slides, data sets, and other materials: Available on Canvas.

Hardware/software: A Windows or Mac laptop. Download and install: (1) the R software from http://cran.r-project.org and (2) R-Studio from https://www.rstudio.com. Bring your laptop computer EVERY class for in-class exercises.

Assignments:

Assignments are due at the start of class on their assigned dates. Assignments must be submitted on Canvas and are date-stamped when submitted. Late assignments are NOT acceptable after I give out my solutions in class. You are welcome to help each other with the assignments, but each assignment must represent your own effort. Copying answers from each other or from prior students of this class will be viewed as "plagiarism" and will earn you a zero grade on that assignment for the first offense, F grade for the class for the second offense, and dismissal from the MS-BAIS program for the third offense.

Exams:

There will be two exams during the semester. Each exam will ask you to solve one problem, similar to the assignments, and you will be graded based on not just what you did, but also on your thought process, approach to solving the problem, and simplicity and clarity of your code. Exams are open-book, open-notes, and open-Internet and similar to "screening tests" conducted by companies like Google or Microsoft. If you attended all classes and paid attention in class, completed all assignments on your own, and continuously try to improve your own solutions, you should have no problems with the exams.

Please do NOT miss any exam without a documented medical or family emergency. A "documented emergency" must be accompanied with appropriate documentation such as a doctor's note. Make-up exams will be different from, and potentially harder than, the regular exams.

Team Project:

This term project, to be done in teams of 3-4 students, is the heart of this class. In this project, you will apply the statistical concepts and methods you learned in class to solve a real-world "serious" data analytics problem. Examples of team projects may include: (1) strategies to reduce police violence in America, (2) strategies to reduce customer churn in an industry, (3) strategies to reduce America's mental health crisis. You will identify a decision-making scenario of business or social relevance, source appropriate data, clean and merge the data as needed, use the right models, and derive actionable insights. Please avoid classification problems as that is not the focus of this class.

Your project report (10-12 pages, plus appendix) must include the following sections: (1) executive summary; (2) problem definition & significance; (3) prior literature; (4) data source & preparation; (5) variable choice; (6) descriptive analysis & data visualizations; (7) data modeling; (8) quality checks; (9) actionable recommendations; and (10) references; and (11) appendix (with R code). The project report must be sufficiently detailed, include appropriate graphics, and be of professional quality. Intermediate project deliverables are due at different points in time during the semester and the final project report is due at the end of the semester. See project report assignment on Canvas for further details. Sample projects are posted on Canvas.

Note that data sourcing and cleaning are very much integral and graded components of this project. A Kaggle or UCI dataset of unknown quality is unacceptable, and will get you zero points on data sourcing/cleaning sections of the project. Kaggle data is okay for personal learning, but not for serious analysis. Please use authentic data sources that fully describe the data collection process, which you must

discuss in your report. Your project topic and data sources must be preapproved by me before you can start working on your project.

Your professor will compare your projects with similar projects on the Internet/Github, run a plagiarism check, and ask you to explain your work. Feel free to browse online projects and learn from them, but your work must be sufficiently original and useful. Downloading and submitting an online project, with minor modifications, will be considered "plagiarism" and get the entire project team a zero grade.

The professor is available to help you with your project throughout the semester. Use this free help. Please also use this project as an opportunity to showcase what you learned inside and outside class. A well-done project can help you get internships and jobs. I do give extra-credit to teams who surpass my expectations, experiment with new methods and tools, and help the class learn something new.

E-mail me the names of your team members and a tentative topic for your project before the second day of class. Choose your team members "wisely", making sure that you have team members proficient in the domain knowledge, in reading and understanding technical papers, and in writing professional-quality reports. If someone in your team does not deliver or is late with their assigned work or rarely comes to team meetings, you have the option to fire that person by majority vote.

Class Policies:

Since this a business class, I expect a certain level of **business professionalism** in this class. This means coming to class on time, not making excuses, not plagiarizing assignments, timely completion of assigned work, willingness to learn things on your own, managing your team project like a professional work project with weekly scrums and reviews, etc. Lack of professionalism will be penalized appropriately.

For USF policies and procedures regarding academic integrity, academic grievance, COVID-19, disability access, religious observances, sexual misconduct, harassment, and academic continuity, please visit https://www.usf.edu/provost/faculty/core-syllabus-policy-statements.aspx

Class Schedule

Week	Date	Topic	Due
1	Feb 6	Introduction & Regression Refresher	-
2	Feb 13	OLS Assumptions	A1: Credit Rating, P1: Names & Project Topic
3	Feb 20	WLS/FGLS & Non-Linear Models	A2: Hunters Green Home Sales
4	~~Feb 27~~	GLM & Poisson Models	A3: Medical Expense
	Mar 06	No Class – Professor out of town	
	Mar 13	No Class – Spring Break	
5	Mar 20	Multi-level and Panel Data Models	A4: Online Retail Campaign
6	Mar 27	**MIDTERM EXAM**	P2: Prior Work & Variable Selection
7	Apr 03	Survival Models	A5: Big Mart Sales
8	Apr 10	Classification Models	A6: Lung Cancer
9	Apr 17	Time-Series Models	A7: Telco Churn
10	Apr 24	Special Topics: Tobit Regression, Beta Regression, Simultaneous Equation Models	A8: Pricing & Promotions Analysis for a Retail Chain, P3: Project Presentations
11	May 01	**FINAL EXAM**	P4: Project Report