ISM 6137: Statistical Data Mining
W 6:30pm – 10:15pm, Spring Semester 2023
BSN 231

Professor: Dr. Daniel Zantedeschi	*Office:* CIS 2066
E-mail: danielz@usf.edu	*Office Hours:* By appointment
Website:	*Office Phone:* (813) 694 1122
TA: None	*TA Hours:*

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 inappropriate, such as count data, mixed-level data, classification data, temporal data, and survival data, using R. More than learning the tools, the focus of the course will be on solving important problems in the healthcare, retail, service, and other domains from start to end, in a systematic manner, and doing so using the right statistical models. This is a "thinking" course that will require a significant amount of time, effort, and attention on your part. Prior knowledge of R is required for this class.

Prerequisite: QMB 6304 (Analytical Methods for Business). This is strictly enforced.

Learning Objectives:

To learn how to:

- Assess and handle violations of regression assumptions.
- Model count data, mixed-level data, classification data, temporal data, and survival data.
- Preprocess data and engineer features from the data for analysis.
- Implement advanced statistical models in R.
- Interpret these models meaningfully to generate useful and usable insights.
- Conceptualize, plan, conduct, and document real-world analytics project from start to end.
- Use data and analytics correctly and ethically.

Books and Materials:

Textbook: None. I will draw my personal knowledge and materials sourced from many books and online and offline sources.

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. If you are coming to the in-class sessions, bring your laptop computer every class for in-class exercises. If attending online, a dual monitor system may be helpful.

Grading:

Grade components and weights:

Individual Assignments (eight)	40%
Exams (two)	30%
Team Project (one)	20%
In-class Quizzes	10%

Tenative 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%
Please note that the instructor reserves the right to curve the distribution of grades to ensure homogeneity with concurrent and past sections. Examples and grade projections will be offered in class.

Assignments:

Assignments are due at the start of class on their assigned dates. Assignments must be submitted via Canvas, and are date-stamped when submitted. Assignments will be graded not just for correctness, but also clarity of approach and cleanliness of code. I will discuss my solutions in class, to show you the "professional" approach to solving these problems; hence late assignments are NOT acceptable. Students are encouraged to help each other on the assignments. However, assignments must represent your own individual effort. There is a fine line between "taking help" and "plagiarizing." I may cold-call you to explain your assignment in class. If you cannot explain your work, you will lose points or even get zero grade on the assignment. A repeat offense will result in an F grade for the course and a report to the Dean's Office.

In-Class Quizzes:

Short pop-quizzes (about 5 or 6 multiple-choice questions) may be held toward the end of most classes to test what you have learned from the class. The quizzes will be based on materials that we covered in that class. You do not have to prepare in advance for these quizzes, but you have to pay attention in class. Since most online students don't pay attention to class lectures (e.g., browse Internet, check Facebook or Instagram, e-mail and chat with friends, etc.) thinking that there will be a recorded lecture that they can access later, this is my strategy to keep you focused on class. The quizzes are not worth much, but can be the difference between an A and a B grade, or a B and a C grade.

Exams:

There will be two online exams during the semester, worth 10 and 20 points respectively. In each exam, you will be asked to solve one problem, and you will be graded based on not just what you did, but also on your thought process, and simplicity and elegance of code. Exams are open-book, open-notes, and open-Internet. They are just like the assignments, but with a restricted time period. Hence, the best way to prepare for an exam is to pay attention in class, completing assignments by yourself, and continuously looking for ways to improve your solutions. Exams will be similar to the "screening test" at Google or Microsoft, if you apply for a job there. Exams will be proctored using the online Proctorio software.

Please also 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 harder than, the regular quizzes.

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 solving a real-world data analytics problem. You will identify a decision-making scenario of managerial or social relevance, source appropriate data, clean and merge the data as needed, use the right models, and derive actionable insights. Several times during the semester, you will present your intermediate project work, which will allow me to keep track of your progress and give you guidance. The final project report (10-12 pages, plus appendix) is due at the end of the semester. This report 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. See project report assignment on Canvas for further details. Sample projects are posted on Canvas for your use.

In the past, I have seen teams spend so much time searching for the right project that they did not have sufficient time to do a quality job with the project. Starting early and allocating sufficient time are absolutely critical for a good project. Your project topic and data source must be approved by me before you can start working on the project. Please use authentic data sources that fully describe the data collection process, which you must discuss in your report. Kaggle or UCI dataset of unknown quality is not acceptable; such data is good for personal learning, but not for serious analysis. Please avoid classification problems as that is not the focus of this class.

Note that 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, practical, and useful. Downloading an online project, making minor modifications, and submitting as your team project 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. Please use this help. I do give extra-credit to teams who surpass my expectations, experiment with new methods and tools, and help the class learn something new. In the past, students have used my class projects to get internships and jobs. But you can only do that if your project is sufficiently good!

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 "intelligently", making sure that you have team members proficient in R, in reading and understanding technical papers, and in writing professional-quality reports. If someone in your team does not show up for team meetings or is perpetually late with their work, you have the option to fire that person by majority vote.

Class Policies:

Class Attendance and Etiquette: I do not care if you come to class or not. If you skip classes, you will struggle in the assignments and quizzes. Also, do not disrupt the class by coming to class late or leaving early. Arrive 15 minutes early rather than 5 minutes late.

Cell Phones: While in class, turn your cell phones to silent mode, and step out of the class if you must take or make an important phone call. See USF's official policy on class disruption at http://www.ugs.usf.edu/policy/DisruptionOfAcademicProcess.pdf.

Academic Integrity: I have a zero-tolerance policy for plagiarism. Plagiarism includes presenting others' work as your own, lifting materials from the Internet or other sources without attribution, and many others. Per USF policy, any such behavior will result in a zero grade for that grade component for the first offense, F grade in the class and report to the Dean's Office for the second offense, and dismissal from USF for third offense. There will be no make up for plagiarized work. See USF's official academic integrity policy at http://www.ugs.usf.edu/policy/AcademicIntegrityOfStudents.pdf.

Free Riders: Team members who regularly miss team meetings or do not contribute adequately to their team project will receive at least one full letter grade lower than the rest of their team. There will be no make-up opportunities for free riders.

Disability: Students requiring disability accommodations must register with Students with Disabilities Services (SDS) office, and e-mail me a letter from SDS specifying your accommodation needs. See USF disability policies at http://www.sds.usf.edu.

Religious Observance: If you must miss class or a test because of religious reasons, you must notify me in writing at the start of the semester, so that I can make alternate arrangements. See USF's religious observance policy at http://www.ugs.usf.edu/policy/ReligiousDays.pdf.

Food and Housing Insecurity: Students facing financial difficulty is securing a stable place to live or affording groceries are urged to visit the USF Feed-A-Bull website for assistance.

Emergency: In the event of an emergency, it may be necessary for USF to suspend normal operations. During this time, USF may opt to deliver instruction through Canvas, Elluminate, Skype, e-mail, or alternate methods. Please monitor Canvas for course-specific communication, and USF website and e-mails for important general information.

Please understand that this syllabus constitutes a **CONTRACT** between you and me. By registering for this class, you are agreeing to the terms and conditions as stated in this contract. These terms will be strictly enforced, without exceptions. If you find these terms unreasonable, you may drop the class and take it with a different professor.

Class Schedule

Week	Date	Topic	Due
1	Feb 8	Introduction & Regression Refresher	-
2	Feb 15	OLS Assumptions	A1 (Credit Rating), P1 (Names & Project Topic)
3	Mar 1	WLS/FGLS & Non-Linear Models	A2 (Online Retail Campaign)
4	Mar 8	GLM & Poisson Models	A3 (Medical Expense)
5	Mar 15	Spring Break – No class	
6	Mar 22	**MIDTERM EXAM**	A4 (Doctor's Visit), P2 (Prior Work & Variable Selection
7	Mar 29	Multi-level and Panel Data Models	
8	Apr 5	Survival Models	A5 (Big Mart Sales)
9	Apr 12	Classification Models (OUT OF TOWN)	A6 (Lung Cancer)
10	Apr 19	Time-Series Models	A7 (Telco Churn)
11	Apr 26	Simultaneous Equation Models	A8 (Pricing & Promotions Analysis for a Retail Chain), P3 (Project Presentations)
12	May 3	**FINAL EXAM**	P4 (Project Report)

Note: The syllabus is tentative and subject to change. However, all such changes will be announced in class.