CGU School of Social Science, Policy & Evaluation
CGU Department of Politics and Government
PP487 Applied Data Analysis: Machine Learning and Data Mining for Social Scientists
T, W, and F: 4:00-7:50 PM. Harper 10. Summer, 2019

Professor: Javier M. Rodríguez
javier.rodriguez@cgu.edu
Office: McManus 226
Phone: 909.621.8695
Office hours: Mondays and Thursdays from 2:30 PM to 4:00 PM, or by appointment
Office: McManus 224 (or at the Inequality and Policy research Center (IPRC))

Teaching Assistant: Matthew Gomies
matthew.gomies@cgu.edu
Office hours: Mondays from 4:00 to 6:00 PM, or by appointment
Office: McManus 225 (Inequality and Policy Research Center (IPRC))
TA Session: Thursdays from 5:00 to 6:30 pm (Harper 10)

Course description

In the past decades, machine learning revolutionized the way scientists across disciplines and industries analyze data, unearth hidden patterns, and apply statistical tools to solve social problems. This course introduces students to the most commonly applied machine learning, data mining, and statistical pattern recognition techniques in academic and public/private research settings. It offers a practical know-how approach while focusing on the techniques, methods, and the statistics supporting them. Some of the topics covered in the course include supervised machine learning (e.g., parametric/non-parametric algorithms), unsupervised machine learning (e.g., clustering), and performance, cross-validation, and regularization theory. Because this is not a coding/programming course, students will learn how to apply algorithms to social science data, create and evaluate data clusters, and perform predictive analytics through a variety of statistical software packages (Stata, R, and Python) and already-written code specifications in various languages. This course offers a great opportunity for social scientists to acquire advanced data science skills and application tools

Class format, course requirements, and grading

This course involves hands-on data analysis. Our philosophy is learning-by-doing. Each class will be typically divided into lecture and practice, or a combination of both. This means that we will spend a great deal of time running analyses ourselves in almost every class. Because machine learning is widely applied in academic, government, and industry settings, we will use many real-world examples across disciplines and settings and their respective applications. Because this is a course on applied machine learning methods, hands-on analyses represent the lion's share of the total course grade. Students will be evaluated via three mini-projects (45%), a midterm exam (20%), and a maxi-project (35%).

1. Mini-projects (45%—15% each)
    - Each mini-project consists of a data analytic exercise that will help students solidify basic concepts and learn how to apply the methods described during lectures.
    - For each mini-project, students will write a short report of roughly 3 pages. Students are free to collaborate with other students on these mini-projects, but they will have to submit their own write-up for credit.
        - No write-up should look similar to another one; students should make their own, independent analysis and interpretation of the problem under observation.
    - Students will deliver a 10-minute Power Point presentation of their mini-project results, which will be followed by a 10-minute Q&A section and feedback.

    NOTE: For due dates and presentation dates see our schedule below.

2. Midterm exam (20%)

    This is an open-book exam, in the computer lab on August 9, 2019. This midterm exam will include multiple choice and open-ended questions, both conceptual and methodological. It will also ask the student to run statistical analyses, generate visualizations, and interpret results. This midterm examination will be heavily grounded on lectures and mini-projects.

3. Maxi-project (35%)
    - Students will be given a research scenario and they will write a short, journal-style research paper. The maxi-project is due on August 23, 2019.
    - Students will deliver a maxi-project *proposal* presentation (approximately a 10-minute presentation followed by a 10-minute Q&A section and feedback).
    - Students will write a research paper of roughly 5 pages—excluding tables, figures, and references. The research paper will be focused on the methods implemented, interpretation of results, and discussion of findings. Students are free to collaborate with other students on the maxi-project, but they will have to submit their own write-up for credit. No Maxi-project write-up should look similar to another one.

Required texts, course requirements, and assignments

All reading assignments are available either online or through our library system. Please make sure you are *logged into your library account* when accessing the books listed below.

Since this is an intensive 4-week course, there is not much time for reading. However, that is different from no reading. I expect you will come to class after having read at least some of the ideas about a given week's topics outlined in any of the following books:

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2017), *The elements of statistical learning*. This book is available for free online. You can access it here.

Bali, Raghav, Dipanjan Sarkar, Brett Lantz, and Cory Lermeister. 2016. *R: Unleash machine learning techniques.* Available as e-book through our library system.

Gollapudi, Sunila. 2016. *Practical machine learning.* Available as e-book through our library system.

Lantz, Brett. 2013. *Machine learning with R.* Available as e-book through our library system.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of machine learning.* Available as e-book through our library system.

Coelho, Luis Pedro and Willi Richert. 2015. *Building machine learning systems with Python*. *Second edition.* Available as e-book through our library system.

Here is a brief summary of the course timeline:

Date	Topic	Assignment due
July 30 Class1	- Class introduction. - Generalized linear models (GLM) - Maximum Likelihood Estimation - Introduction to Logistic Regression.	N/A.
July 31 Class2	- GLM, explanatory models for academic research - Residual analysis - Diagnostics: leverage, deviance, Cook's Distance. - *Applications* of logistic regression: Prediction and as a classifier algorithm. - Exercise: Predict recessions.	N/A.
Aug 2 Class3	- *Applications* of logistic regression: - Inverse probability weighting for missing data. - Propensity scores (and matching) - Heteroskedasticity in logistic models.	N/A.

	- Underlying assumptions of logistic models. - Interactions in logistic models.	
Aug 6 Class4	- *Applications* of logistic regression (continuation) - Balancing [very] skewed distributions. - Introduction to Multinomial logistic regression as a multiclass classifier. - Seemingly unrelated bivariate probit regression.	- Mini-project 1, 3-page report. - Power Point presentation of Mini-project 1 results.
Aug 7 Class5	- *Applications* of multinomial logistic regression and seemingly unrelated bivariate probit regression. - Key concepts in machine learning. - Supervised and unsupervised analysis. - Confusion matrix. -The ROC curve and AUC. - AUCs for classifiers and multiclass classifiers. - Cross-validation. - Training, validation, and test. - K-fold cross-validation. - Bias and variance tradeoff.	
Aug 9 Class6	- Mini-project #2 presentations and Midterm	- Mini-project 2, 3-page report. - Power Point presentation of Mini-project 2 results. - Midterm
Aug 13 Class7	- Regularization - Introduction to Ridge, LASSO, LASSO-logit, and Least Angle Regression - *Applications* of model building (shrinkage) algorithms.	

Aug 14 Class8	- Introduction to Python: Jupyter Notebook, Pandas and Scikit-learn. - Decision trees and random forests.	
Aug 16 Class9	- Introduction to clustering and unsupervised classification. - K-means, K-modes, and latent class analysis. - *Applications* for segmentation.	- Mini-project 3, 3-page report. - Power Point presentation of Mini-project 3 results.
Aug 20 Class10	- *Applications*: initial combination of supervised and unsupervised methods.	
Aug 21 Class11	- Applications: Full combination of supervised and unsupervised methods. - Maxi-project *proposals*	- Maxi-project *proposal* presentations
Aug 23		- Maxi-project

Grading

Your grade will be calculated using the following scale:

Letter Grade	Grade Point	Percentages (%)	Description	Learning Outcome
A+	*4.0*	*97-100*	*The student has acquired additional insight, far beyond the standards set forth for the course material.*	*Exceptional*
A	*3.8*	*92-96*	*The student has done an excellent work, developing a complete mastery of the material as intended for the course.*	*Superior*
A-	*3.5*	*87-91*	*The student has acquired a very good mastery of the course material and the necessary ability to use this ability elsewhere.*	*Commendable*
B+	*3.2*	*81-86*	*The student has demonstrated proficient mastery of course material yet partial success on some assessments.*	*Proficient*
B	*3.0*	*71-80*	*The student has demonstrated a foundational level of understanding about the course material and partial success on some of the assessments.*	*Satisfactory*
B-	*2.7*	*65-70*	*The student approaches mastery of course material and, accordingly, needs extra assistance to achieve a*	*Approaching*

			foundational understanding and to apply the main course skills.	
C	*2.0*	*<65%*	*Gaps in mastery of course material. The student shows difficulties understanding at least some of the main concepts and with applying the skills of the course as expected by the program.*	*Developing*
U	*0*	*0*	*Unsatisfactory*	*Ineffective*

Note: Continual matriculation at CGU requires a minimum GPA of 3.0 in all coursework taken at CGU. Students may not have more than two incompletes. Details of the policy are found on the Student Services webpage.

Other content on expectations and logistics:

- *Focus*: This course—taught through the Department of Politics and Government—will make use of some political data. Even though the course is in line with Political Science as a discipline, it is not focused on the "political" but rather on the "science" aspect of it. Accordingly, this is neither a partisan nor an ideological course. What concerns us is the data and the rigorous statistical analysis of the data; nothing less; nothing more. We are here to improve our understanding of how the world works, and not to reaffirm our personal views of the world and opinions.

- *Due dates*: All course assignments should be submitted not later than midnight of the due date. Late submissions will be penalized with a 20% grade reduction, and no assignments will be accepted after 24 hours late. The grade for such late assignments will be zero. Exceptions will be made only under truly exceptional circumstances.

- *Statistical software*: We will use the three most used statistical software in the market for machine learning analysis: Stata, R, and Python. Stata is known for being an intuitive, easy-to-use statistical software. Stata is accessible at CGU, so you will not have to buy it. R and Python are also free and accessible in all CGU computers.
    - We will run a plethora of data processing, analyses, and visualizations. The great majority will be done in Stata and R. However, you will have access to some replications and exercises processed in more than one statistical software. For example, we may see in class an exercise running a Least Angle Regression, yet you may still have access to a Ridge and LASSO regression exercises in R (all three: LARS, LASSO and Ridge regressions are model selection algorithms).
    - It is also impossible to teach three—well, not even one—statistical software in four weeks. This means that we will try to go with the flows, and will use the software of preference among students. Students, therefore, are free to study and run analyses in whatever software they prefer—no software will be given a special preference.
        - You may feel, however, a little overwhelmed with using different software. However, let me tell you a piece of information: That's how it is out there in the labor market. Some organizations prefer Stata (industry on policy and program evaluation and overall data analysis), others R (more prominent in the academia and campaigning), others Python (more prominent among the online industry).

- Although it is not expected at all that you will walk out of this course being an expert in machine learning methods using a specific statistical software, you will indeed (for your advantage) be exposed—at least to some extent—to the three of them: Stata, R, and Python.

- *Assignment submission format*:
  - All assignment instructions will be posted on Canvas, and students are required to submit them through our Canvas platform, as well.
  - Assignments should include your name, e-mail, and the title of the assignment (e.g., Mini-project #2).

- *Working in teams*:
  - *Coauthored assignments*: Many times, students will work in teams. For example, Celia Lacayo and Mark Sawyer will work together. It is Celia's and Mark's responsibility to submit a *single* Word document. It is their responsibility to split the work equally, and to be ethical and assume equal responsibility on the quality of their work. It is expected that each student will help the other with editing, such that the complete document will be written professionally. Both, Celia and Mark, will receive the same grade for that assignment.
  - *Separate assignments*: If the assignment allows you to work in teams but each student is required to submit her/his own Word document, final assignments should be *absolutely independent*. They should not look alike, at all.
  - *Professionalism*: Be professional with your classmates. They are, and will be for years to come, your professional network. Thus, if for example, you will miss class or cannot attend a meeting, you should notify your group members (as well as the professor).
  - *Presentations*: Presentations can be submitted as one Power Point document, but it is expected that each student will present their part independently. Each student will receive independent grades for their presentations.

- *High writing quality*:
  - All written work should be double-spaced, using 12 point fonts, one-inch margins, numbered pages, *and professionally written*. You are Master's and Ph.D. degree students, and that is what is expected from you: Top, graduate-level writing. Throughout your professional and academic lives, you will always be evaluated, and will advance, on the basis of your writing. Graduate professionals and researchers are, in essence, writers.
    - To write is not a component of research. To write *is* to do research. To write is to re-write. Edit. Edit. Edit. All students may not be good writers, but through the re-writing process, all students have the potential to be good editors of their writing. All that said, and given the centrality of writing to the academic experience, your academic performance will also be evaluated on the basis of your writing. Good writing will be rewarded; poor writing will be penalized.

- For some mysterious reason, I've noticed that CGU's Writing Center is one of the most underutilized resources at CGU. I highly recommend that you seek assistance from the Writing Center—even if you are a good writer.
- Presenting well is also a professional necessity. And guess what—the Writing Center recently added a presentation-assistance option!
- The Writing Center is located in a blue-grey house at 141 E. 12th St, and you can also get their assistance online. You do not need a referral to go to the Writing Center.

- *Attendance:*
  - Students are expected to attend all classes. It is my experience that when students miss one class, they will struggle for the rest of the course. Needless to say, this is aggravated if the course is an intensive 4-week course.
  - Students who are unable to attend class must seek permission for an excused absence from the Professor.
    - Unapproved absences or late attendance for three or more classes may result in a lower grade or an "incomplete" for the course.
    - If a student has to miss a class, s/he should arrange to get notes from a fellow student, and it is strongly encouraged to meet with the Teaching Assistant to obtain the missed material. Missed extra-credit quizzes and papers will not be available for re-taking.

- *Scientific and Professional Ethics:*
  - The work you do in this course must be your own. Feel free to build on, react to, criticize, and analyze the ideas of others but, when you do, make it known whose ideas you are working with. You must explicitly acknowledge when your work builds on someone else's ideas, including ideas of classmates, professors, and authors you read. If you ever have questions about drawing the line between others' work and your own, ask the course professor who will give you guidance.
    - The Midterm exam must be completed independently. Any collaboration on answers to exam, unless expressly permitted, may result in an automatic failing grade and possible expulsion from the Program. Additional information on CGU academic honesty is available on the Student Services webpage.

  - Do NOT plagiarize. Please note that plagiarism is determined by the *act*, not the *intent*. Be careful to keep good records and give good citations and references.
    - Both CGU and I take academic integrity very seriously. Cheating is grounds for failure. One form of cheating is plagiarism.
    - Faculty are required by university policy to report all cases of plagiarism—even if they are simply *apparent*—to the office of the Vice President of Academic Affairs. I follow this requirement.

- *The basic rule to avoid plagiarism is **very simple***: give credit where credit is due. That's it. Always give a citation when you use the ideas, words, figures, or data of others. Easy. It is better to use too many citations than to use too few.

- *Feedback and Communication:*
  - The best way to get in touch with me is via e-mail: javier.rodriguez@cgu.edu
    - I will respond to e-mail messages usually within 24 hours.
      - Always be professional in your communication with the Teaching Assistant and the professor.

- *Course Policies:* The CGU institutional policies apply to each course offered at CGU. Students are encouraged to review the student handbook for the program as well as the policy documentation within the bulletin and on the Registrar's pages.

Additional important content

Accommodations for Students with Disabilities: If you would like to request academic accommodations due to temporary or permanent disability, contact Dean of Students and Coordinator for Student Disability Services at DisabilityServices@cgu.edu or 909-607-9448. Appropriate accommodations are considered after you have conferred with the Office of Disability Services (ODS) and presented the required documentation of your disability to the ODS.

Mental Health Resources: Graduate school is a context where mental health struggles can be exacerbated. If you ever find yourself struggling, please do not hesitate to ask for help. If you wish to seek out campus resources, here is some basic information about Monsour. https://www.cuc.claremont.edu/mcaps/

"Monsour Counseling and Psychological Services (MCAPS) is committed to promoting psychological wellness for all students served by the Claremont University Consortium. Our well-trained team of psychologists, psychiatrists, and post-doctoral and intern therapists offer support for a range of psychological issues in a confidential and safe environment."

Phone: 909-621-8202
Fax: 909-621-8482
After hours emergency: 909-607-2000
Tranquada Student Services Center, 1st floor
757 College Way
Claremont, CA 91711

Title IX: If I learn of any potential violation of our gender-based misconduct policy (rape, sexual assault, dating violence, domestic violence, or stalking) by any means, I am required to notify the CGU Title IX Coordinator at Deanof.Students@cgu.edu or (909) 607-9448. Students can request confidentiality from the institution, which I will communicate to the Title IX Coordinator. If

students want to speak with someone confidentially, the following resources are available on and off campus: EmPOWER Center (909) 607-2689, Monsour Counseling and Psychological Services (909) 621-8202, and The Chaplains of the Claremont Colleges (909)621-8685. Speaking with a confidential resource does not preclude students from making a formal report to the Title IX Coordinator if and when they are ready. Confidential resources can walk students through all of their reporting options. They can also provide students with information and assistance in accessing academic, medical, and other support services they may need.