95-791 – Data Mining (Section C1)

Fall 2021

Instructor

Prof Gongora-Svartzman (Prof. GS)
Email: ggongora@cmu.edu
Office Hours: TDB - Look on CANVAS for update

Course Description

Data mining is the science of discovering structure and making predictions in large, complex data sets. Nowadays, almost every organization collects data, which they hope to use to support improved decision making. Learning from data can enable us to better: detect fraud, make accurate medical diagnoses, monitor the reliability of a system, perform market segmentation, improve the success of marketing campaigns, and much, much more.

This course serves as an introduction to Data Mining for students in Business and Data Analytics. Students will learn about many commonly used methods for predictive and descriptive analytics tasks. They will also learn to assess the methods' predictive and practical utility.

Learning Objectives

By the end of this class students will learn:

1. Be able to produce, comprehend and run Python code for commonly used data mining methods.
2. Understand the advantages and disadvantages of multiple data mining methods. This involves:
    1. Generalizability
    2. Bias-variance trade-off
    3. Interpretability-flexibility tradeoff
3. Be able to compare the utility of different methods through lab exercises, homeworks, and a final project.
4. Understand the concepts behind feature engineering, and be able to place them into practice through different types of data.
5. Be able to choose an appropriate model/s for a dataset and evaluate the performance and reliability of such model/s.
6. Be able to apply methods to real-world data.

Learning Resources

Textbooks

There is one required textbook in this class. It is available for free at the link below. If you find the textbook to be useful, please show your appreciation by purchasing a copy for personal use.

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
  An Introduction to Statistical Learning: with Applications in R - Second Edition
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar - Introduction to Data Mining.
- Andreas C. Müller, Sarah Guido. Introduction to Machine Learning with Python.
- Jake VanderPlas. Python Data Science Handbook, Essential Tools for Working with Data.

Additionally students will be presented with slides, online tutorials and recent papers. Please keep an eye on CANVAS for these resources.

Additional Textbooks Suggestions

In addition to the required text, the following references are highly recommended. Students may find it useful to own a personal copy of one or two of the texts below.

- *Wes McKinney,* Python for Data Analysis
- *Jacqueline Kazil, Katherine Jarmul,* Data Wrangling with Python
- Witten and Frank, Data Mining: Practical Machine Learning Tools and Techniques
- Hastie, Tibshirani, Friedman, Elements of Statistical Learning
- Provost and Fawcett, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking
- Kuhn and Johnson, Applied Predictive Modeling

Helpful resources

There are many resources online that may help you with various parts of the class.

Here are some resources to help you refresh or expand your Python skills and knowledge:

- Pandas Documentation
- Jupyter Notebooks Documentation
- Anaconda
- Markdown in Jupiter Notebooks
- Seaborn
- Matplotlib
- Plotly - Python

TENTATIVE COURSE SCHEDULE

Date (Week)	Topics
WEEK 1 Aug. 30th - Sept 5th	1. Introduction to Data Mining 2. Introduction to Regression **Lab 1** * Jupiter Notebooks - Python * Regression models in Python
WEEK 2 Sept 6th - Sept 12th	**September 6th - Labor Day - No classes** 1. Nonlinear Regression 2. Model Validation
WEEK 3 Sept 13th - Sept 19th	1. Model Selection 2. Classification **Lab 2** * Model validation and model selection in Python * Nonlinear models
WEEK 4 Sept 20th - Sept 26th	1. Classification II 2. Tree-based Methods **Midterm Exam**
WEEK 5 Sept 27th - Oct 3rd	1. Advanced Methods I 2. Advanced Methods II **Lab 3** * Classification * Trees and advanced methods
WEEK 6 Oct 4th - Oct 10th	1. Association Methods 2. Unsupervised Learning I **Lab 4**
WEEK 7 Oct 11th - Oct 17th	1. Unsupervised Learning II 2. Review Session **Final Project (due)**
Friday, October 17th	**Final Project (due)**

****Note: This schedule is subject to changes due to unforeseen events (e.g. snowstorms, outages, etc).*

Assessments

The final course grade will be calculated using the following categories:

Assignment	Grade Percent
Homeworks	18%
Mid-mini Exam	25%
Final Project	30%
Participation and Discussions	9%
Labs	18%
Total Grade	**100%**

*All lectures, labs, and assignments are in **Pittsburgh, PA timezone.***

The final course grade will be calculated using the following scheme:

Grade	Letter
97 - 100 %	A+
93 - 96.99 %	A
90 - 92.99 %	A-
87 - 89.99 %	B+
83 - 86.99 %	B
80 - 82.99 %	B-
77 - 79.99 %	C+
73 - 76.99 %	C
70 - 72.99 %	C-
0 - 69.99 %	R

All work is individual unless otherwise stated. Cheating, copying or any other unethical behavior will not be tolerated, and will lead to an automatic R in the course.

Attendance: This course is REO, it runs synchronously and attendance is mandatory for lectures and labs.

Participation: In-class exercises and Canvas discussions will be considered as participation. This will be clearly played out on CANVAS and Piazza. In-class assignments may be done on a periodic basis, and will not be announced in class beforehand; since regular attendance is the norm, this should not be an issue. A major purpose of these in-class assignments is for both students and faculty to be certain that key concepts are understood and can be applied to basic problems. There will be no make-up for missed in-class assignments but you can be excused with prior permission

Assignments: All assignments (homeworks, mid-mini exam, labs, discussions and final project) should be submitted through CANVAS in their corresponding assignment slot. Instructions will be given for this on CANVAS. Assignments are to be completed by individuals without the assistance of classmates or other students. In this class, students **may NOT look at the code of other students**, **show their code to other students, or get/give material assistance to/from another student or outside the individual.** The only exception to this rule is with the final project, where you will work in groups. Heinz College considers academic integrity to be of great importance, we actively scan for cheating policy violations and will take swift and appropriate measures against those who fail to abide by these standards. You will deeply regret cheating in this class if you are caught, I assure you.

Final Project: There will be one main project that will be developed and evaluated during the duration of the mini.

Exams: There will be no final exam, just a final project that will be developed during the duration of the mini. There will be a mid-mini exam.

Programming projects will be graded based on their correctness, completeness, and quality. We expect very high-quality work and attention to detail at all times in this course - it is up to you to see that this is so. Programs that are missing, substantially incomplete, do not load, do not run, or more than 24 hours late will be assessed penalties of 100%.

Late / Make-Up Work Policy

Due dates for every assignment are provided on the course syllabus and course schedule (and posted in Canvas). Unless otherwise stated, assignments are due on those days. There will be no make-up work. You may still submit an assignment up to 24 hrs later, but for a 50% deduction off your grade in the assignment (this only applies to homeworks, labs and discussions). Mid-mini exam and final project assignments will receive a 0 if handed in late.

I recognize that sometimes “life happens”, for this reason if you have any issues during the course, that interfere with you making deadlines (e.g. health, life-related, etc.) please reach out to me through email (ggongora@cmu.edu). Honesty, communication and transparency are highly welcomed and encouraged in this course.

Course Format and Structure

Instructor's Online Hours

Look on CANVAS for an update (Home page).

I will be available via email and will respond as soon as I am available (generally within 24-48) hours. When emailing me, please place in the subject line the course number/section and the topic of the email (i.e. 95-791 – Assignment 2 Question). This will help me tremendously in locating your emails when I scan the hundreds of emails that seem to make it into my box each day. *For quicker responses, I encourage you to communicate on Piazza so that you can get help from the instructor, the TA, or even your fellow classmates.*

Course Format

This is a Remote Only course and is held synchronously, meaning that students are required to attend (virtually) all lectures and labs. Attendance to labs is mandatory and the participation grade given for the lab cannot be substituted by just handing in the lab exercises and not attending the zoom session. All lectures, labs, and office hours will be held through zoom and the links will be provided through Canvas. No student may record any classroom activity without express written consent from me.

All zoom links - for both lectures and labs - are provided on Canvas (Go to the Zoom tab on the left-hand side). Please make sure that your Internet connection and equipment are set up to use Zoom and able to share audio and video during class meetings. (See this page from Computing Resources for information on the technology you are likely to need). Let me know if there is a gap in your technology set-up as soon as possible, and we can see about finding solutions.

During our class meetings, please keep your mic muted unless you are sharing with the class or your breakout group. If you have a question or want to answer a question, please use the chat or the "raise hand" feature (available when the participant list is pulled up). I [or a TA or a rotating student who serves as the "voice of the chat"] will be monitoring these channels in order to call on students to contribute.

Sharing video

In this course, being able to see one another helps to facilitate a better learning environment and promote more engaging discussions. Therefore, our default will be to expect students to have their cameras on during lectures and discussions. However, I also completely understand

there may be reasons students would not want to have their cameras on. If you have any concerns about sharing your video, please email me as soon as possible and we can discuss possible adjustments. Note: You may use a background image in your video if you wish; just check in advance that this works with your device(s) and internet bandwidth. Our synchronous meetings will periodically involve breakout room discussions, and those will work better if everyone in your small group has their camera turned on.

Every student is allowed to miss two lectures without penalty for any reason. However, if you get Covid-19 or have some other special case and you notify me with the appropriate information as quickly as possible, I am certainly willing to excuse those absences as well.

Communication and Questions

Assignments and class information will be posted on CANVAS.

Email: ggongora@cmu.edu

The Piazza forum should be used for general course-related questions that may be of interest to others in the class. For other types of questions (e.g., to report illness, request various permissions) please contact Prof. GS via email (ggongora@cmu.edu)

Students are strongly encouraged to ask questions during class. The material can be tricky at times and we expect questions to be asked during lectures. All Assignments and class information will be posted through Canvas. I will be available via email and will respond as soon as I am available (generally within 24-48 hours). When emailing me, please place in the subject line the course number/section and the topic of the email. This will help me tremendously in locating your emails quicker when I scan the hundreds of emails that seem to make it into my box each day.

Please be advised that sending an email to your instructors does not create responsibility or obligation to respond to it. Sending us an email does not shift any responsibility from you to us. You are still responsible for on-time, high-quality completion of assignments and projects. Please reserve your email for matters that actually require our attention. In any case, do not expect a response to non-emergency emails in under 48hrs. Emails will not be used to send grade reports either.

Even though we are using Canvas please try to avoid sending me or the TAs any Canvas Inbox messages, we prefer direct emails.

Piazza

The Piazza forum should be used for general course-related questions that may interest others in the class. The quicker you begin asking questions on Piazza (rather than via emails), the quicker you'll benefit from the collective knowledge of your classmates and instructors. We encourage you to ask questions when you're struggling to understand a concept - you can even do so anonymously. Rather than emailing questions to the teaching staff, we encourage you to create a discussion topic/line and post your questions.

We will be monitoring Piazza every day and no question should go more than 24 hours without being answered (in most cases, much sooner). Do not send questions via email to any TAs without first checking Piazza discussions to see if an answer has already been posted. In case the answer has already been posted, the TA will refer you back to Piazza. If you email new questions that are not of a personal nature (like grades, standing in class) your question might be posted on Piazza so that the answer can be useful for everyone. All students are invited to refer to the Piazza etiquette section at the end of this syllabus.

Every student in the course will be enrolled in the Piazza. If you have not been enrolled please do not hesitate to contact us (Prof GS - ggongora@cmu.edu).

If you have any problems or feedback for the developers of Piazza, please send an email to mailto:team@piazza.com

Academic Honesty and Integrity

You are encouraged to discuss homework and lab problems with your fellow students. However, the work you submit **must be your own**. You must acknowledge in your submission any help received on your assignments. That is, you must include a comment in your homework submission that clearly states the name of the student, book, or online reference from which you received assistance.

Submissions that fail to properly acknowledge help from other students or non-class sources will receive no credit. Copied work will receive no credit. Any and all violations will be reported to Heinz College administration.

All students are expected to comply with the CMU policy on academic integrity. This policy can be found online at http://www.cmu.edu/academic-integrity/.

Every line of text and line of code that you submit must be written by you personally. You may not refer to another student's code, or a "common set of code" while writing your own code. You may, of course, copy/modify lines of code that you saw in lecture or lab.

While we encourage you to be helpful to your classmates, you must understand that the work you turn in for evaluation or credit must be your own. You are welcome to talk with other students about general course content, requirements, and technology issues. You are not welcome to offer, or to ask for, substantial, material assistance to, or from, other students in completing specific aspects of graded assignments for individual credit. If there is any doubt in your mind about a particular situation, ask yourself this question: "How would I feel if I observed another student or students engaging in this particular behavior?"

Here are some examples of academic integrity violations related to code:

- Copying line-by-line (or substantial amounts) of another student's code (classmate or peer).
- Copying line-by-line of online (or substantial amounts) code (e.g. Stackoverflow) or code found in one of our textbooks.
- Posting class assignments online to get help. You can only post questions on our Piazza.
- Posting class assignments in a public repository, therefore allowing other classmates to replicate your code.

Note from Prof. GS: "viewing someone's code, sharing code with another, or giving specific instructions, verbally or in writing, for work in any of the phases is definitely out of bounds. Do not post your code to a public repository where it is available to all or a private repository where it is viewable by other students." If there is any doubt if it is allowed please see Prof. GS to get clarity on the matter. Better safe than sorry!

Any student who turns in work for credit that is identical, or similar beyond coincidence, to that of another student may face appropriate disciplinary action at either the department, college, or university level. Your reputation among your peers and among the CMU faculty is one of your most valuable assets. Do not risk damage to your reputation or academic career by engaging in behavior that could be interpreted as dishonest or unethical.

To be clear, any academic integrity violation will result in zero in that assignment and a failing grade in the course. You will be - without a doubt - reported and face appropriate disciplinary action at either the department, college, or university level.

Online Etiquette Guidelines

Your instructor and fellow students wish to foster a safe online learning environment. All opinions and experiences, no matter how different or controversial they may be perceived, must be respected in the tolerant spirit of academic discourse. You are encouraged to comment, question, or critique an idea but you are not to attack an individual. Our differences, some of which are outlined in the University's inclusion statement below, will add richness to this learning experience. Please consider that sarcasm and humor can be misconstrued in online

interactions and generate unintended disruptions. Working as a community of learners, we can build a polite and respectful course ambiance.

Please read the etiquette rules for this course:

- Always be respectful of your instructors, teaching assistants, and fellow students.
- Do not dominate any discussion. Allow other students to join in the discussion.
- Do not use offensive language. Present ideas appropriately. We have zero-tolerance for inappropriate language, as well as racist, sexist, or discriminatory language in any context during the course.
- Additionally, any assignment, documentation, code, or files turned in containing language, comments, or references that are inappropriate will incur an automatic 30% penalty without exemption.
- Be sure to turn off your cellphones, notifications, and other distracting factors during lecture time.
- Be cautious in using the Internet language. For example, do not capitalize all letters since this suggests shouting.
- Avoid using vernacular and/or slang language. This could lead to misinterpretation.
- Keep an “open-mind” and be willing to express even your minority opinion.
- Think and edit before you push the “Send” button.
- Do not hesitate to ask for feedback.
- We intend to start each class on time, so please be respectful and be on time.
- Do not leave the class early unless you have informed your instructor in advance.
- Always double check your microphone is muted unless you wish to speak up, participate, or you're being called out by the instructor or teaching assistants. Otherwise, your background noise can become distractive or disruptive to your fellow students.
- We are trying to give you the same experience as in-person, therefore we ask for you to put on your video and appear during lectures, labs, and office hours.

Student Wellness

Disability Resources

If you have a disability and need special accommodations in this class, please review the steps listed by the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

If you already have an accommodations letter from the Disability Resources Office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate.

Diversity Statement

We must treat every individual with respect. We are diverse in many ways, and this diversity is fundamental to building and maintaining an equitable and inclusive campus community. Diversity can refer to multiple ways that we identify ourselves, including but not limited to race, color, national origin, language, sex, disability, age, sexual orientation, gender identity, religion, creed, ancestry, belief, veteran status, or genetic information. Each of these diverse identities, along with many others not mentioned here, shape the perspectives our students, faculty, and staff bring to our campus. We, at CMU, will work to promote diversity, equity and inclusion not only because diversity fuels excellence and innovation, but because we want to pursue justice. We acknowledge our imperfections while we also fully commit to the work, inside and outside of our classrooms, of building and sustaining a campus community that increasingly embraces these core values.

Resources for Diversity and Inclusion:

* Center for Diversity and Inclusion
* Intercultural Communication Center
* Office of Title IX Initiatives

Wellness Statement

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep, and taking some time to relax. This will help you achieve your goals and cope with stress.

All of us benefit from support during times of struggle. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is almost always helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty, or family member you trust for help getting connected to the support that can help.

If you have questions about this or your coursework, please let me know. Thank you, and have a great semester!