Applied Data Science II - ES 4062

Winter - 2022

Basic Information

Instructor: Kyle Scot Shank ‘14

E-mail: ksshank@coa.edu (note: I will always try to get back to you within 24 hours, but there may be occasions when it takes a bit longer)

Pronouns: he/him/his

Class Meeting: CHE 103 (Center for Human Ecology - General Classroom), Tuesdays 6:00PM - 9:00PM

Office Hours: Thursdays, 8:00PM - 9:00PM via Zoom (link to be provided)

Course Description

Applied Data Science II is a deeper exploration of statistical modeling skills that was briefly covered in Applied Data Science I. Specifically, whereas Applied Data Science I focused on data-cleaning, organization, and visualization, Applied Data Science II will be focused on designing and building quantitative models from data and using those models for predictive purposes. We’ll focus on two main topic areas and ob in this course:

1. **Practical, field-invariant skills for designing complex models from data**
   Think of this as developing the “why” and “when” of using a model
2. **Building various types of predictive models and evaluating their performance**
   Think of this as developing the “which one” and “how” of building models

Leaving this class, students will be able to immediately apply these predictive modeling skills to a broad array of interests and topic areas. For example: students should have the ability to build predictive models for population sizes in ecological settings based

upon field date or design a machine-learning based classification model for whether or not a given student will attend ACM in a given week.

Classes will be taught as a mix of live coding exercises, lectures, and group discussions. Prior familiarity with the R programming language and statistics - such as building linear models and hypothesis testing - are required. Students will need to use either their personal laptop or a COA loaner laptop for class and programming exercises. Evaluation will be through class participation and discussion, weekly homework assignments, and a final project.

Course Evaluation

There will be no quota of A's, B's, etc. You may take this class according to any grading structure you prefer (letter grades, pass/fail, etc.) - please feel free to reach out early so we can discuss the best plan for you.

Evaluation will be through class participation and discussion, weekly problem sets and exercises, as well as a final project. The problem sets will take the form of code-based exercises from our primary text as well as additional data explorations that I will provide. The final project will take the form of an oral presentation of a modeling project based on a dataset of your choice. This can be either done in a group or as an individual and may be of any topic of sufficient interest to the student(s) involved.

In general, the breakdown of course credit will be as follows:

- Weekly Homework Assignments: **60%**
- Final Project: **30%**
- Class Participation: **10%**

Weekly Homework Assignments

Throughout this class we will be working on weekly modeling/coding assignments to help build your practice of data modeling. These assignments **must** be submitted individually via Google Classroom - but you may work together in groups if you like (and are very much encouraged to do so!) so long as you make note of who all worked together on your submission. All assignments for the week (as shown in the Schedule of Assignments below) are due at **5:59PM on Tuesdays, right before class begins**. Note

that all files will need to be submitted as RMarkdown files. We will discuss this requirement and go over the template on the first day of class.

Final Project

There will be a final modeling project (either individual or group) focused on an in-depth analysis of a specific data set. We will discuss the final form of this project during the semester.

Class Participation

I define class participation as a balance between ***presence***, ***attention***, and ***preparation***. Being ***present***, from my perspective, is attending class and being actively engaged in the learning process. ***Attention*** can take a variety of different forms: some students may be more comfortable asking questions in class (either of me or of your peers), others may be more inclined to take notes and digest and absorb information on their own time, etc. All of these forms of attention - as well as whatever other forms attention may take - are welcomed and encouraged in this class. ***Preparation*** is having completed the previous readings and assignments to the best of your ability and being ready to discuss any problems (or insights!) you may have had in doing so. This final piece will be especially important to our class due to our limited number of class meetings.

I've stolen the quote below from Dave Feldman to hopefully give an even better explanation for how I envision class participation as a mechanism of learning:

> *We should all work to create an inviting atmosphere and ensure that there is opportunity for all to contribute. At the same time, there is no need for everybody to contribute equally. It is natural for some people to talk more than others, and I think this is normal and good. Also, I expect that students will engage and contribute at different levels, depending on prior coursework. I see this diversity of backgrounds as a strength and not a weakness; there are roles for everyone to play. Asking good questions is as important as providing answers.*

Late Submission Policy

It is important to turn in work when it is due. That said - the world is pretty crazy right now, so things can (and will) come up. Every student in this course may submit up to **one** of the weekly homework assignments late without penalty. The guidelines for this policy are as follows:

1. The work must be submitted no later than 7 days after it was originally due (i.e. - you'd have to submit the late work *prior* to/in conjunction with submitting the following assignment).
2. After 7 days, no late work will be accepted.

Other subsequent assignments that are submitted late will receive a 5% penalty per-day. This policy does not apply to the final project.

Course Schedule & Flow

Because we will only have ten meetings over the course of the term, we'll be maximizing each meeting period by trying our best to follow this schedule:

- **6:00 - 6:30**: General Announcements / Homework Review
- **6:30 - 7:30**: Overview of topic (based on previously assigned readings)
- **7:30 - 7:45**: Water/Tea/Cookie break
- **7:45 - 9:00**: In-class Lab

We'll adjust as needed, especially given any potential disruptions due to the continued COVID-19 pandemic.

The general "flow" of topics covered will be as follows:

- **Week 1**: Introductions, syllabus and expectations overview, using RMarkdown, overview of predictive modeling
- **Week 2**: Linear Models. Simple and multivariate models, diagnostics, measures of predictive accuracy
- **Week 3**: Linear Models. Model selection and regularization. Shrinkage methods (Ridge, Lasso), dealing with high-dimensional data.
- **Week 4**: Classification Models. Logistic regression.
- **Week 5**: Resampling methods. Cross-validation and bootstrapping.
- **Week 6**: Moving Beyond Linear Models. Additive models and splines.
- **Week 7**: Tree-based methods. Random forests.
- **Week 8**: Deep Learning. Neural networks.
- **Week 9**: Project working today + selected topics.
- **Week 10**: In-class presentations of final projects, course wrap-up

Course Texts

We will be using the following texts. Please note that these texts are available online for free and purchase is only strictly necessary if you'd like to have your own physical copy for reference.

- James, Witten, Hastie, Tibshirani, *An Introduction to Statistical Learning* (ISLR)
    - This will be the **primary** textbook for this course - so make sure you download the 2nd edition from the website (or if you plan to buy it - buy the 2nd edition!)
    - You'll also frequently want to access the data in the text to follow along. These data are available here.
    - If you're a more visual learner / want some supplemental video materials, check out these resources here.
- Max Kuhn and Kjell Johnson - *Applied Predictive Modeling* (APM)
    - This will be mostly a **secondary** textbook for this course - there will be assigned readings but it is not necessary to complete any homework. No need to buy it (a PDF version is in our Google Drive) but it could be fairly useful for future use.
    - We will likely refer to the caret package frequently - so best to bookmark the website for this library as well.

Tentative Schedule of Assignments

Note: in general, the assignments that are due each week will be based upon the readings and in-class lab from the week prior. The only exceptions to this are the assignments due in week 2 (which can be completed based upon previous experience with R + the reading) and week 10 (which is the final assignment). I will also provide supplemental/optional readings through Google Classroom for those interested. I may also provide questions through Google Classroom that can be included as extra credit - so keep your eyes peeled!

- **Week 1**
    - *Readings*:
        - N/A
    - *Assignment*:
        - N/A
- **Week 2**
    - *Readings*:
        - ISLR: 2.1, 2.2, 3.1, 3.2, 3.3
            - If you're feeling a little rusty with R, you should work through 2.3 on your own.
        - APM: 1.1 through 1.4, Chapter 2, Chapter 5

        - The Introduction to RMarkdown
    - *Assignment*:
        - ISLR: 2.4: 1, 2, 8, 10
- **Week 3**
    - *Readings*:
        - ISLR: 6.1 - 6.4
        - APM: Chapter 4, Chapter 6
    - *Assignment*:
        - ISLR: 3.7: Questions 8, 9, 10, 11, 13, 14
- **Week 4**
    - *Readings*:
        - ISLR: 4.1 - 4.6
        - APM: Chapters 11, 12
    - *Assignment*:
        - ISLR: 6.6: Questions 8, 9, 10, 11
- **Week 5**
    - *Readings*:
        - ISLR: 5.1, 5.2
        - APM: Chapters 3, 4
    - *Assignment*:
        - ISLR: 4.8: Questions 8, 13, 14, 15, 16
        - Special Classification Exercise TBD
- **Week 6**
    - *Readings*:
        - ISLR: 7.1 - 7.7
    - *Assignment*:
        - ISLR: 5.4: Questions 5, 6, 8, 9
- **Week 7**
    - *Readings*:
        - ISLR: 8.1, 8.2
        - APM: Chapters 8, 14
    - *Assignment*:
        - ISLR: 7.8: Questions 6, 7, 8, 9, 10, 11, 12
- **Week 8**
    - *Readings*:
        - ISLR: 10.1 - 10.8
        - APM: Chapters 7, 13
    - *Assignment*:
        - ISLR: 8.4: 6, 7, 8, 9, 10, 11, 12
- **Week 9**

    - *Readings*:
        - TBD
    - *Assignment*:
        - ISLR: 8.4: 7, 8, 9, 10, 11, 12
- **Week 10**
    - *Readings*:
        - N/A
    - *Assignment*:
        - **Final Project Submission**

Diversity and Inclusion Statement

It is my intent that students from all backgrounds and perspectives be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength, and benefit. It is my intent to present materials and activities that are respectful of diversity: gender, sexuality, disability, age, religion, socioeconomic status, ethnicity, race, and culture.

Learning about diverse perspectives and identities is an ongoing process. I am always looking to learn more about power and privilege and the harmful effects of racism, sexism, homophobia, classism, and other forms of discrimination and oppression. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups. If something was said or done in class (by anyone, including me) that made you feel uncomfortable, please talk to me about it. You may also reach out to the Provost or Associate Deans for further information or discussion.

COVID-19 & Remote Instruction

The goal for this term is to be *in-person* and *together* as much as possible. With that said - we're still in the midst of a pandemic, so who knows what might happen. If we're going to be switching to remote instruction at any point I will make sure to post in Google Classroom as well as directly email all of you as far ahead of time as I can. If we need to be remote, we'll be using Zoom.

Standard Disclaimers

150 hours of academic engagement

Our accreditation requires that we communicate that students should expect 150 hours of academic engagement for a one-credit COA course This total includes weekly meetings, field trips, office hours, film screenings, readings and other assignments, service-learning, practicum, or other course requirements. You should therefore expect to spend a minimum of 150 academically engaged hours associated with this one-credit course. These 150 hours will be spent roughly as follows: 3 hr/wk in a classroom environment, 4 hr/wk reading, 8 hr/wk on homework.

Plagiarism

By enrolling in an academic institution, a student is subscribing to common standards of academic honesty. Any cheating, plagiarism, falsifying or fabricating of data is a breach of such standards. A student must make it his or her responsibility to not use words or works of others without proper acknowledgment. Plagiarism is unacceptable and evidence of such activity is reported to the academic dean or his/her designee. Two violations of academic integrity are grounds for dismissal from the college. Students should request in-class discussions of such questions when complex issues of ethical scholarship arise.

Library Resources

Thorndike Library offers many resources and services that can assist you in your academic endeavors, including individualized research support and access to resources beyond COA. Study spaces are also available. The library is open 7 days/week. Remote access to the research databases is available 24/7. Contact library@coa.edu or visit the library website for details.