PDAT 613G: Data Mining

Scott Thatcher, Spring 2022, Online Delivery

Contents

1 General Course Information

Instructor: Dr. Scott Thatcher

Office: Violette 2134

Phone: 660-785-4552 (A voice message will reach me relatively quickly.)

e-Mail: thatcher@truman.edu

Website: All course information is located on Blackboard

General Office Hours: See Blackboard for up-to-date office hours and Zoom meeting IDs.

Weekly Class Meeting: Mondays at 7:00 p.m.

Academic Success Coordinator: Ashleigh Harding, aharding@truman.edu, (660) 785-7403

1.1 Welcome and Introduction

Welcome to PDAT 613! In this course, we'll explore techniques related to Data Mining and Machine Learning.

We'll see several types of predictive modeling algorithms that go beyond basic linear regression. We'll talk about analysis of numeric, categorical and text variables, and we'll look at ways to cluster and display unorganized data. While this course is not designed to go deeply into the mathematics that underlies these methods, we will emphasize some basic concepts that underlie their successful application.

The module on visualization will discuss principles of good data visualization, especially in the multivariate context, and we'll end the course with some readings on topics related to the ethics of Data Science. What can we do, and what should we do?

I hope that this course broadens your appreciation of what's out there, and whets your appetite for continuing with the more rigorous material in PDAT 615!

1.2 Catalog Description

An exploration of techniques used to find patterns in very large data sets, with an emphasis on the statistical structure of the approaches and practical uses of key tools.

1.3 Prerequisites

Successful completion of PDAT 610G: Introduction to Data Science. Recommended: Completion or concurrent enrollment in PDAT 611G: Big Data Management.

1.4 Texts and Resources

There is no required textbook purchase for this course. Students may find the following resources helpful, and some will be used for required readings:

- Roger Peng, *Exploratory Data Analysis* (free-$35). https://leanpub.com/exdata
- Brian Caffo, *Regression Models* (free-$25). https://leanpub.com/regmods
- Garrett Grolemund and Hadley Wickham, *R for Data Science*. https://r4ds.had.co.nz/
- Rafael A. Irizarry, *Introduction to Data Science*. https://rafalab.github.io/dsbook/
- The *Tidymodels* package documentation and tutorials. https://www.tidymodels.org/
- Julia Silge and David Robinson, *Text Mining with R*. https://www.tidytextmining.com/
- The Truman Library website. https://library.truman.edu

1.5 Course Outcomes

On successful completion of this class, students will:

- describe the various skills, techniques and workflow that comprise the general field of Data Mining;
- explain, on a non-mathematical level, the central ideals behind several methods in supervised and unsupervised machine learning, including regression, decision trees, random forests, support vector machines, k-means clustering, hierarchical clustering and principal components analysis;
- use R to apply these methods to real data sets, demonstrating proficiency with commands from several appropriate R packages;
- explain the results of their analysis in clear, non-technical, language;
- apply methods such as train/test sets and cross-validation to achieve honest estimates of a model's predictive ability;
- describe some of the basic principles of good data visualization;

- comment on whether data visualization examples illustrate these principles of good design;
- create displays of multivariate data; and
- discuss ethical issues, such as the ownership of data and data products, the dangers of p-hacking, and algorithmic bias.

1.6 Credit Hours

The minimum investment of time by the average Truman student necessary to achieve the learning goals in this course is not less than 45 hours of student effort per credit hour. This average time per week for an average student may have weekly and per student variations. In this class, time investment will be split between "learning activities" (online lectures), programming assignments, discussion prompts, and reading from assorted reference material. Additionally, students may participate in a weekly live class meeting and office hours.

2 Technology and Skill Requirements

2.1 Minimum Technology Requirements

In order to participate fully and effectively in an online course, students should have a reliable broadband connection (cable modem, DSL, satellite). Students should have a relatively new operating system (Linux, Windows 8 or 10; Mac OSX, etc.) and employ a compatible browser such as Firefox, Chrome or Safari. Courses use Blackboard Learn. For a list of compatible systems and browser types, visit Blackboard (blackboard.com).

This course does use audio and video. Videos are close-captioned. To benefit from the audio you will need a computer equipped with speakers.

While it's not required, you may also want to use our Data Science virtual Linux server at https://fire.truman.edu:6900/vnc.html. If your internet connection goes in and out, sustaining the connection to these virtual desktops might be difficult.

In order to complete assignments, you will need to use R and RStudio and be able to export R Markdown files as PDF. General familiarity with other standard office software is also assumed.

2.2 Minimum Technical Skills

While this isn't an exhaustive list, to be successful in this class, you should at least be able to

- access the Internet and be able to comfortably navigate websites using a web browser,
- maintain contact with me and the class through e-mail, Zoom video conferencing, and Blackboard forums,
- apply your general coding skills and specific knowledge of R to write code and complete assignments, and
- access outside resources that might provide data sets or coding documentation, etc.

Many other skills are likely implied by the ones listed, but at this point in your program, I hope you have a good sense for what's required!

2.3 Technical Expectations for Completing Assignments

Assignments will be submitted through Blackboard, usually by submitting PDF files generated from R Markdown code and by submitting the R Markdown code itself. Discussion assignments will completed using Blackboard forums and optional Zoom meetings.

2.4 Proctoring

There will be no proctoring necessary for assignments in this class.

3 Course Modules

A sample schedule is given below. Note that the course used to have only three large modules, and for historical reasons, the modules are still labeled 1, 2A, 2B, etc.

Module/Week	Topics
Week 1/Mod 1	Introduction to data mining. What is data mining? What is a data mining workflow?
Week 2/Mod 2A	Supervised learning algorithms: regression and classification. Methods include linear regression, logistic regression trees, random forests, and SVM.
Week 3/Mod 2B	Unsupervised learning algorithms: *k*-means and hierarchical.
Week 4/Mod 2C	More on predictive modeling. Topics include cross-validation and the *Tidymodels* packages.
Week 5/Mod 3A	Natural Language processing. Sentiment scores, document-term matrices, and topic analysis.
Week 6/Mod 3B	Introduction to multivariate visualization. Critique and creation of data visualizations.
Week 7/Mod 3C	Ethics in data mining/machine learning. Reading and discussion of several case studies.
Week 8	Catch-up and Capstone topic generation, if you haven't already submitted a capstone topic.

4 Assignments and Grading

4.1 Course Components

Work for this course will include the following components:

Assignment	Percent
Ungraded Quizzes	0%
Weekly Assignments	90%
Discussions	10%

4.2 Grading

While I reserve the right to be more lenient in grading, the following score percentages correspond to the following minimum grades:

Score	Grade
$\geq$ 90%	A
80%–89%	B
70%–79%	C

An overall course grade lower than a "C" is considered a failing grade, and no more than one course grade of "C" may be counted toward the certificate. Students may retake a course to raise a grade not meeting minimum program requirements.

4.3 Quizzes

Short quiz questions on readings and lectures are interspersed within the lecture material. Their purpose is to help you focus on and review material that has been presented. Quiz scores will not be included in the final grade.

4.4 Weekly Assignments

Most weekly assignments will consist of practical application programming assignments will allow students to practice the skills of the course. Programs will be graded on adherence to specifications, program quality, correctness of results, and clear communication of those results to the reader.

Weekly assignments will be due on Wednesdays.

4.5 Discussion

The discussion assignment score will make up 10% of your total grade. I will not impose a strict word limit, but I will be looking for (1) evidence of substantial engagement with the material, and (2) evidence of regular interaction with me and with the rest of the class.

What does "substantial engagement" mean to me? Here are some guidelines:

- A good response to many discussion questions will draw the connection between abstract concepts and specific examples, specifically explaining how that example relates to the characteristics of that concept. Compare these two examples:
  - [Not so good] "I'd use regression to model expected sales because it's linear."
  - [Better] "Linear regression does well when the relationship between response variable and predictor variables can be modeled with a linear equation. Looking at the scatter plots of sales vs. advertising, and sales vs. consumer confidence, both appear to have a linear pattern. Therefore, regression seems appropriate in this situation."
- If a question asks you for a personal viewpoint, try to do some thinking, and share specifics (as you feel comfortable).

- Support opinions with evidence, or relate them back to concepts discussed in class.
- Try to avoid short "me too" responses to others' posts. Instead, ask yourself if you can paraphrase their important points and then add to them, or (respectfully) disagree with them.

4.6 Assignment Redo's

The assignments you turn in through Blackboard should represent your best work. I understand that sometimes something will go wrong with a particular assignment, and I do want to give you some flexibility to try again if necessary. On the other hand, I want to avoid a pattern of multiple submissions that incrementally inch toward a final product.

In an effort to find that balance, my rule for this class is that **students will be allowed a total of two redo's of graded assignments, which should be turned in by a week after the original due date.** If you've submitted multiple redo's of an assignment before I've graded it, I'll grade the most recent version.

All that being said, remember that the best workflow is to

- start assignments with plenty of time,
- talk with me *before* you turn your assignments if you have questions or would like general advice as to whether you're on the right track,
- turn in a good assignment, and
- redo only if necessary.

5 Course Policies and Expectations

5.1 Attendance

Students are strongly encouraged to join the weekly live Zoom meeting in order to better interact with me and other class members. Even if you don't feel "caught up," it is better to check in regularly. Past students have reported the sense of camaraderie to be a plus. These sessions are not, however, strictly required.

5.2 My Expectations of Students

My expectations of students include:

- Putting in time for this class commensurate with its accelerated pace: about 18 hours per week would not be an unreasonable ask for a three-credit, eight-week course. Of course individual students' experiences will vary.
- You are strongly encouraged to take advantage of posted virtual office hours. If you cannot meet during the posted hours, additional appointments can be arranged. I will create a Zoom meeting space for virtual office hours and post this information in Announcements on Blackboard. Individuals needing to consult privately may do so by phone or arrange a time to conference separately.
- Ask questions, and stay in contact with me, especially if you run into roadblocks or "real life" emergencies! I'll help as much as I can, but there's not much I can do if you disappear completely!

- Engage in both the coding and discussion aspects of the class. Discussion is difficult in online classes, but I've found that it can be good with a population of students who are looking to get a positive experience out of it.
- Engage in discussions with me and other students politely and respectfully.
- Know important class dates and keep track of assignments.

5.3 What Students Should Expect of Me as Their Instructor

You can expect me, as the instructor, to

- Respond to e-mail and other communication within 24–48 hours.
- Regularly engage with discussions posted on Blackboard, although I may leave space for students to comment first.
- Grade assignments submitted on time within 2-3 days. If that schedule slips, I'll let you know and give an updated timeline via Blackboard announcement.
- Post regular announcements about the class, especially if there are changes or updates to assignments or schedules.
- In case of unforeseen events that force me to step away from the class for more than 48 hours, notify you as soon as possible, and provide for alternate points of contact if at all possible.

5.4 Response Time and Feedback

As stated above, I will strive to respond to communication within 24–48 hours and grade most assignments within 3 days. Communication that I receive over a weekend or holiday may be answered on the next "business" day, but I'll often answer over the weekend.

5.5 Student Interaction

You've enrolled in a program that features a human on the other end of the internet connection! I strongly encourage you to keep in contact with me, and strive to keep up with the material. The discussion forums are a good place to ask and answer questions, and it's often helpful to be asking questions when other students are also looking at the same material.

We will have weekly live Zoom meetings, which are not required, but they're a quick and easy way to make sure you feel caught up and connected to what's going on in class. I'll also have regular Zoom office hours posted on Blackboard and by appointment.

5.6 Netiquette and Civil Dialog

Etiquette for the net is important in an online or blended online course because of the significant amount of communication taking place in an environment that is not face-to-face (though much of what we discuss here has its applicability in face-to-face settings as well).

A few good rules to keep in mind include:

- Avoid communication strategies that could easily lead to misunderstanding. Avoid using slang and jargon. Avoid jokes and sarcasm. Don't use ALL CAPS or lots of exclamation points!!! These can seem like yelling or incorrectly convey the intensity of emotion.
- Write to communicate with clarity. Avoid using emojis, texting abbreviations, and technical terms that do not apply to the course. Include a clear subject line in email and discussion board communications. Write in a manner that conveys professionalism, as if you were writing a formal letter or a research paper. Make responses in discussion boards substantive by avoiding short responses that do not add to the conversation (e.g. "I agree!" "Good point.")
- Be considerate of your legal and ethical obligations. Be sure to avoid posting content that would constitute plagiarism or violate copyright law. This would include unattributed quotations, posting copies of print articles, sharing photos you do not have permission to share, etc. Do not engage in behavior that would be considered discriminatory or harassing. There may be concrete repercussions for behavior that would violate the Academic Honesty or Discrimination policies articulated elsewhere in this syllabus.

5.7 Academic Honesty

The General Catalog states:

> Students are expected to do their own academic work. Any student involved in cheating on a paper, an examination or in any other form of academic dishonesty is subject to disciplinary action, including suspension or expulsion from the class, the student's academic program, or the University.
>
> Serious cases of academic dishonesty are reported by the faculty member to his or her Department Chair and to his or her Dean, who may take additional disciplinary action against the dishonest student, including suspension or expulsion from classes in the School. The Dean reports the dishonesty to the Vice President for Academic Affairs, who may also report it to the Vice President for Student Affairs. The Dean may also report the dishonesty to the School in which the dishonest student is enrolled as a major; the Dean of this School may suspend or expel the student from the academic program in the major. The Dean of Students may also suspend or expel the student from the University as outlined in the Student Conduct Code for incidents of academic dishonesty.

More information can be found in the General Catalog and in the Student Conduct Code Section 8.050.1

In this class, we'll be writing a lot of code. I expect the code you write to be your own work—after all, the purpose of this class is to help you practice your skills, and you won't get that practice if you simply copy code from someone else. I assume that developing your knowledge and skills is an important reason you're in this program!

Having said that, here is a non-exhaustive list of standards concerning common situations and questions that have come up in the past:

- **Study Groups:** I encourage you to exchange ideas through our Blackboard discussion forums or though study groups you might form with other class members. However, after you discuss an assignment, you should disengage from the group and make sure to write up your work on your own. This is the only way to make sure that you understand what's going on!

- **Credit Where Credit is Due:** If you do work with a study group, give credit to the members of the group in your assignment write-up.
- **Outside Sources:** Especially when writing code, you don't live in a vacuum. It's natural to look online (through Google or Stack Exchange, for example) for solutions to coding problems. When you do this, or access other outside sources (living or non-living), you should also give credit. A link to the site where you found the code snippet you adapted, or an acknowledgment of the human who helped you, etc. Not only is this ethical, but it's also extremely valuable when you need to see that solution again, and perhaps the search result no longer appears at the top of a Google search. This practice has saved me time on multiple occasions
- **Fully Worked-Out Examples:** It's generally **not** OK, in the context of this class, to Google something like "regression analysis of the mtcars data set" in order to find fully-fleshed-out code, graphs, explanation, etc., for a homework problem that I've asked you to do. Not only are you cheating yourself of the experience of learning on your own, but much of the stuff that people post online is of questionable value, at best. There are a lot of bloggers who *think* they understand data science, but fewer who actually do!

6 Important University Policies and Procedures

6.1 Substantive Interaction

Truman policy and federal regulations require that students demonstrate that they are academically engaged in the courses they take. You must meet this requirement within the first calendar week of the semester, beginning at 12:00 a.m. on Monday, January 10 and ending 11:59 p.m. Saturday January 15. Failure to do so, or to provide an explanation of an extenuating circumstance by that date and time will result in your removal from the course. Under certain circumstances, removal could impact your scholarship eligibility or financial aid.

This policy is not intended to be punitive, but rather intended to protect students who never intended to remain enrolled in the course from surprise tuition charges.

For the purposes of this class, establishing academic engagement requires, at a minimum, posting to the online discussion from Module 1 or posting a introductory post in the "Bulldog Cafe" section of the Discussion forum.

6.2 Important Dates

- **Start Date:** Wednesday, January 6.
- **End Date:** Tuesday, March 1.
- **Last Drop Date:** Wednesday, February 16.
- **Last Date to Withdrawal from All Classes:** Friday, April 29.

See the Registrar's Office (registrar.truman.edu) for more detailed calendars and information on fees/refunds associated with dropping/withdrawing at various points in the semester.

6.3 Emergency Procedures

In each classroom on campus, there is a poster of emergency procedures explaining best practices in the event of an active shooter/hostile intruder, fire, severe weather, bomb threat, power outage, and medical emergency. This poster is also available as a PDF (police.truman.edu).

Students should be aware of the classroom environment and note the exits for the room and building. For more detailed information about emergency procedures, please consult the Emergency Guide for Academic Buildings (police.truman.edu).

A six-minute video (police.truman.edu) provides some basic information on how to react in the event there is an active shooter in your location.

Truman students, faculty, and staff can sign up for the TruAlert emergency text messaging service via TruView. TruAlert sends a text message to all enrolled cell phones in the event of an emergency at the University. To register, sign in to TruView and click on the "Truman" tab. Click on the registration link in the lower right of the page under the "Update and View My Personal Information" channel on the "Emergency Text Messaging" or "Update Emergency Text Messaging Information" link. During a campus emergency, information will also be posted on the TruAlert website (trualert.truman.edu).

6.4 Discrimination and Title IX

Truman State University, in compliance with applicable laws and recognizing its deeper commitment to equity, diversity and inclusion, which enhances accessibility and promotes excellence in all aspects of the Truman Experience, does not discriminate on the basis of age, color, disability, national origin, race, religion, retaliation, sex (including pregnancy), sexual orientation, or protected veteran status in its programs and activities, including employment, admissions, and educational programs and activities. Faculty and staff are considered "mandated reporters" and therefore are required to report potential violations of the University's Anti-Discrimination Policies to the Institutional Compliance Officer.

Title IX prohibits sex harassment, sexual assault, intimate partner violence, stalking and retaliation. Truman State University encourages individuals who believe they may have been impacted by sexual or gender-based discrimination to consult with the Title IX Coordinator who is available to speak in depth about the resources and options. Faculty and staff are considered "mandated reporters" and therefore are required to report potential incidents of sexual misconduct that they become aware of to the Title IX Coordinator.

For more information on discrimination or Title IX, or to file a complaint contact:

Ryan Nely, Institutional Compliance Officer, Title IX and Section 504 Coordinator
Office of Institutional Compliance
Violette Hall, Room 1308
100 E. Normal Ave
Kirksville, MO 63501
Phone: (660) 785-4354
titleix@truman.edu

The institution's complaint procedure can be viewed at
http://titleix.truman.edu/files/2015/08/University-Complaint-Reporting-Resolution-Procedure.pdf
and the complaint form is accessible at
http://titleix.truman.edu/make-a-report/

6.5 FERPA

Education records are protected by the Family Education Right to Privacy Act (FERPA) (registrar.truman.edu). As a result, course grades, assignments, advising records, etc. cannot be released to third parties without your permission. There are, however, several exceptions about which you should be aware. For example, education records can be disclosed to employees or offices at Truman who have an "educational need to know." These employees and offices may include your academic advisor, the Institutional Compliance Officer, the Registrar's Office, or Student Affairs depending on the type of information

7 Important Contacts

Various offices that provide services to online students are identified at the One Stop Services page (online.truman.edu). Should you need to consult with administrators that oversee this department and course, here is the contact information for those individuals:

- **PDAT Program Director:** Hyun-Joo Kim
  hjkim@truman.edu
  660-785-4693
  Violette Hall 2234.
- **Statistics Department Chair:** Scott Alberts
  salberts@truman.edu
  660-785-7649
  Violette Hall 2132.
- **Dean, School of Science and Mathematics:** Timothy Walston
  tdwalston@truman.edu
  660-785-4248
  Magruder Hall 2004.

Hopefully your experience with this class is positive. When and if you feel a complaint about this or another course is required, however, the procedure for lodging a complaint can be found on the University's Report a Complaint page (truman.edu).

Students taking an online course from outside of the state of Missouri should follow the complaint procedure offered here (truman.edu). Students are always asked to address their complaint to the professor of the course first when possible, then take their concerns to the Department Chair if the matter cannot be resolved with the faculty member.

8 Student Support

The University provides a range of both academic and student support services to ensure your success. These offices can advise you on learning strategies, point you toward valuable services, and help you troubleshoot technical problems as they arise.

8.1 Center for Academic Excellence

The Center for Academic Excellence (truman.edu) provides advising services for students in their first year for most departments, as well as tutoring services. The Center is located in Kirk Building 112 and it may be

reached at 660-785-7403.

8.2 Counseling Services

Counseling Services (truman.edu) are available on campus at McKinney Center. Appointments may be scheduled by calling (660) 785-4014. An after-hours crisis line is also available at 660-665-5621.

8.3 IT Help Desk

The IT Service Center (truman.edu) has combined the IT Call Center, Help Desk and Telephone Services into a one-stop location to serve you. You will find the following services and more when you stop by Pickler Library 109 or call 660-785-4544. You may submit a customer support ticket at this web address. (truman.edu)

8.4 Disability Services

To obtain disability-related academic accommodations, students with documented disabilities must contact the course instructor and the Office of Student Access and Disability Services (OSA) as soon as possible. Truman complies with ADA requirements. For additional information, refer to the Office of Student Access and Disability Services website (disabilityservices.truman.edu/) or contact by phone at (660) 785-4478 or by email at studentaccess@truman.edu.

This online course is designed for maximum accessibility. If you encounter difficulty accessing materials required for this course, please do not hesitate to inform me and I will work with Truman's support staff to promptly address the issue.

8.5 Writing Center

I encourage you to use the University's Writing Center (truman.edu) for your writing projects. It is not a proofreading service. The writing consultants will read your work and give you feedback, letting you know what is working well (and why) and what might not be working so well (and why). They can help you understand and better your writing craft. They can also do brainstorming if you're having a hard time getting started. And they have an online scheduler, so making an appointment is easy.

8.6 Student Survey of Instruction

You will be asked to complete a survey regarding my instruction in this course at the end of the term. The survey is anonymous and I will not see the results until after grades have been completed. It is very important that I receive this feedback as it helps me to continuously improve this class. It also helps the University make decisions about our overall curriculum. Please be sure to participate in this survey opportunity.