Data-Driven Predictive Modeling (in R)

Course numbers:	455 / 555
Instructor:	Pradeep Pendem
Contact:	pradeepp@uoregon.edu
Class hours:	Tuesday/Thursday 8:00 am to 9:50 am, Chiles 225
Office hours:	Tuesday/Thursday 1:00 pm to 2:00 pm, Lillis 432

Overview of the Course:

This course introduces to basics of programming in R and the fundamentals of predictive modeling. The audience for this course is senior undergraduate and MBA students. Predictive modeling is a sub-field of business analytics. Utilizing historical data and applying various machine learning techniques enables us to develop models that can be used in Regression, Classification, and identifying patterns. A few examples include predicting – a new patient's length of stay in a hospital outpatient department, used cars sale price, whether a customer accepts a loan, whether a customer commits a credit card fraud, etc.

In this course, the main concepts of various techniques include *k*-Nearest Neighbors (*k*-NN), Linear Regression, Logistic Regression, Regression-based Forecasting, Classification Trees, Regression Trees, and Cluster Analysis are discussed. Further, we also discuss the theory and implementation of error or accuracy measures, cross-validation, and model selection.

In this course, theoretical concepts of predictive modeling will be supplemented by applying them to real datasets. For this purpose, R software will be used. R is one of the most popular programming languages for statistics/data science; therefore, our objective is to teach programming fundamentals in this language.

Students who complete this course will:

- Have fair exposure to introductory programming in R
- Learn the process of managing, cleaning, summarizing, and visualizing real-world datasets in R
- Be able to establish the thinking process of defining a real-world data-driven analytics project
- Learn theoretical fundamentals of different predictive modeling techniques
- Learn the implementation and inference of the methods in R
- Learn how to evaluate various predictive models and select the model

Textbook:

The recommended book for the course:

- **An Introduction to Statistical Learning with Applications in R** by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

Electronic version available for download at: https://statlearning.com/

I will provide my presentation slides throughout the term; the textbook serves as a great compliment to the lecture notes and offers more details.

Groups:

Students must form groups of **three** for the homework, final project presentation, and report submission. Students need to create groups on their canvas course page.

Assessment Weights:

- Homework: 20%
- Midterm 1 : 20%
- Midterm 2 : 20%
- Final Project: 40% (Presentation – 10% + Final Report – 30%)

Assessment Components:

- **Homework's**
  - Four homework's
  - All homework's are group submissions. Groups must be the same for all homework assignments
  - You may not seek/receive help from individuals outside your group
  - Homework's are due by 11:59 PM on canvas on their due dates (specified in the schedule below)
  - Only one member of the group submits (in pdf) on Canvas. Please be sure to list all group members' names in the submitted file itself
  - Late submission will result in a zero score for the group

- **Mid-term Exam**
  - Two midterms
  - Midterms are open-book exams in the form of multiple-choice on canvas. The quiz will test your conceptual knowledge, identify the appropriateness of different techniques for the various business scenarios, identify the strengths and shortcomings of these techniques, and interpret the analysis results.

- **Final Project**

  The final term project will pose a relevant business question, gather, and clean data, perform data analysis, and report your results in a detailed write-up.

  - Final projects are group submissions (same as homework groups).
  - You need to specify a business problem and find a relevant dataset. Business context could be in any area, including but not limited to healthcare, operations, marketing, finance, and social media. If you do not have any business problems in mind, please meet me early. I can help and guide you through the thought process
  - You will be asked to make a 10-15-minute presentation of your project and results during one of the classes in Week 10
  - The final report will be a formal report, including the introduction, problem description, data exploration, analysis results, conclusions, and recommendations. The analysis should include output from all the exhaustive models appropriate for your problem and scientifically recommend the best predictive model. The report should be 8-10 pages (tables and graphs). In addition to the report, you should submit the code (penalty of 15 out of 30 points on not submitting the code)
  - The final project report will be due by exam day and end time. Late submissions are not accepted under any circumstances and result in a score of 0.

Requirement for OBA 555 Students: The final project for master's students (enrolled in OBA 555) will go beyond the similar 455-level project regarding data analysis expectations. Those higher expectations may also include, depending on the project, more demanding data preparation and cleaning processes and more in-depth and comprehensive statistical analyses.

Misc. Course Policies

We will adhere to the following policies, the motivations for which should be self-explanatory.

- A missed test will result in a score of zero for that test, so you should be sure to check whether you have a conflict with the announced test dates and times.

Academic Honesty

The University Student Conduct Code (available at http://dos.uoregon.edu/conduct) defines academic misconduct. Students are prohibited from committing or attempting to commit any act that constitutes academic misconduct. By way of example, students should not give or receive (or attempt to give or receive) unauthorized help on assignments or examinations without express permission from the instructor. Students should properly acknowledge and document all sources of information (e.g. quotations, paraphrases, ideas) and use only the sources and resources authorized by the instructor. If there is any question about whether an act constitutes academic misconduct, it is the students' obligation to clarify the question with the instructor before committing or attempting to commit the act. Additional information about a common form of academic misconduct, plagiarism, is available at:
http://researchguides.uoregon.edu/citing-plagiarism/whycite

Business students are also expected to adhere to the principles and values expressed in the Lundquist Code of Professional Business Conduct. We as a community aspire to be open, respectful, and honest, possessing no tolerance for inappropriate behavior such as cheating, plagiarism, or disorderly conduct.

University of Oregon ADA Policy

The University of Oregon is committed to making available to all its students the opportunity for an excellent and rewarding education. The Americans with Disabilities Act of 1990 and Section 504 of the Rehabilitation Act of 1973 provide federal guidelines which help the University ensure that students with documented disabilities have equal access to this opportunity. If you have a documented disability and anticipate needing accommodations in this course, please make arrangements to meet with me soon. Please request that the Counselor for Students with Disabilities send a letter verifying your disability.

Schedule:

Week	Date	Topic & Content	Due
1	Mar 28 (Tue)	**Introduction to Predictive Modeling** Instructor-Student introduction, Importance of Predictive Modeling, Course Logistics	
	Mar 31 (Thu)	**Introduction to R & RStudio** R & R-Studio Installation, Brief introduction to programming	
2	Apr 05 (Tue)	**Data Management** Basic Data management	Group formation 8 pm
	Apr 07 (Thu)	**Data Management & Graphics** Advanced-Data Management & Graphics	
3	Apr 12 (Tue)	***k*-Nearest Neighbors (*k*-NN) as Classification** Theory, Application in R, Inference	Homework 1 11:59 pm
	Apr 14 (Thu)	***k*-Nearest Neighbors (*k*-NN) as Regression** Theory, Application in R, Inference	
4	Apr 19 (Tue)	**Midterm 1**	
	Apr 21 (Thu)	**Linear Regression** Theory, Application in R, Inference	
5	Apr 26 (Tue)	**Model Evaluation & Accuracy Measures - Classification** Data Partition, Accuracy/Error measures	Homework 2 11:59 pm
	Apr 28 (Thu)	**Model Evaluation & Accuracy Measures - Regression** Data Partition, Accuracy/Error measures	
6	May 03 (Tue)	**Cross-Validation** LOOCV & K-fold Cross-Validation	
	May 05 (Thu)	**Logistic Regression** Theory, Application in R, Inference	
7	May 10 (Tue)	**Midterm 2 Review**	Homework 3 11:59 pm
	May 12 (Thu)	**Midterm 2**	
8	May 17 (Tue)	**Regression-based Forecasting** Theory, Application in R, Inference	
	May 19 (Thu)	**Regression Trees** Theory, Application in R, Inference	
9	May 24 (Tue)	**Classification Trees** Theory, Application in R, Inference	
	May 26 (Thu)	**Cluster Analysis** Theory, Application in R, Inference	Homework 4 11:59 pm
10	May 31 (Tue)	**Project Presentations**	
	Jun 02 (Thu)	**Project Presentations**	Last Class
11	Jun 08 (Wed)	**Final report**	Final Report 8 am
12	Jun 14 (Tue)	**Grades Due**	**Instructor** **12 pm**