

**DSEG6111 : Predictive Modeling**

**Instructor:**

Fotios Kokkotos, Ph.D.

Email: [kokkotosf@merrimack.edu](mailto:kokkotosf@merrimack.edu)

**Course Term: Fall 2023**

**Live Meetings: Every Wednesday 5:30 pm EST (subject to change with prior notice)**

**Course Description:**

This course offers students an introduction to the basic concepts, practices and applications of developing forward-looking statistical models. The class will begin with a full background and explanation of prescriptive statistics and predictive modeling, followed by a conceptual overview of multivariate statistics and a detailed discussion of commonly used types of multivariate predictive models. The second part of the course focuses on hands-on applications of model fitting and evaluation using programming applications in R or Python. Topic wise, students will develop an understanding of numerous core statistical notions, including variable types, statistical learning, linear regression, classification, non-linear modeling and tree based methods.

**Course Learning Objectives:**

Upon completion of this course, students should have:

* The ability to assess the validity and reliability of analytic outcomes
* The ability to deploy and ‘productize’ final modeling solutions
* The ability to manipulate and process available data

**Required Course Materials:**

James G., Witten D., Hastie T., & Tibshirani R. (2021). An Introduction to Statistical Learning with Applications in R, Second Edition. This is an open source textbook and can be downloaded from the [www.StatLearning.com](http://www.statlearning.com) website, or a printed copy can be ordered from [Amazon](https://www.openintro.org/redirect.php?go=amazon_os3&referrer=/stat/textbook.php).

Supplemental Learning Resources:

SciKits Learn Machine Learning in Python

Data Science Central: [www.datasciencecentral.com](http://www.datasciencecentral.com) (an excellent portal for data scientists)

**Required Software:**

R & RStudio or Python

**Important Deadlines for the Term:**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Check with the Registrar’s Office |  |  |
|  |  |  |

## Grading & Assignments:

|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **B** | **C** | **F** |
| **Excellent** | **Good** | **Fair/Poor** | **Poor** |
| 93+ A | 87-89 B+ | 77-79 C+ | <70 |
| 90-92 A- | 83-86 B | 73-76 C |  |
|  | 80-82 B- | 70-72 C- |  |

Student requirements and their corresponding grades are summarized as follows:

|  |  |
| --- | --- |
| **Assignment** | **Weight** |
| Exercises | 40% |
| Final Project | 60% |
|  |  |
|  |  |
|  |  |

The following is a brief explanation of each requirement.

**Weekly & Final Project Assignments:**

**(1) Chapter Exercises (40%)**

Submit R Markdown files for all assignments where the use of R and RStudio is required. Make sure to include the R code for all assignments which require computations. Where programming commands are required, make sure that you test them to ensure that they are executable prior to submitting the lab assignment. For each exercise, answer all parts with well composed sentences including numbers or graphical images depending on the nature of the exercise. Begin each exercise with the appropriate number and label the subparts of each exercise with letters following the conventions used in the text. When you have finished the exercises, submit a Word document with your answers.

**(2) Final Project (60%)**

The final project assignment requires students to select a data science topic of interest, propose a research question which may be answered using descriptive and/or inferential statistics, locate a suitable data file related to the data science topic and research question, analyze the data file using appropriate statistical methods and computational procedures and prepare a report summarizing your findings and documenting the procedures used. Students are strongly encouraged to begin plans for this project early in the course and discuss topics of interest with the instructor and employers by or before Week 5 of the course. A draft outline of the final project is also due on Week 5.

**Grading-Related Policies for this Course:**

Labs, exercises and final project are to be submitted electronically, on blackboard, by the **11:59 pm of the due date**. There will be **NO** late work accepted. Final grades will be based on a percentage of possible points earned. But the instructor reserves the right to give a curve to adjust the average course grade.

## Class Policies:

**Credit Justification Statement (required policy)**

Throughout the term, students will spend approximately 180 hours engaged with the material of the course.

Live class sessions – 2 hours per week 28 hours

Reading textbook – 2 hours per week 26 hours

Weekly homework assignments 48 hours

Weekly lab assignments 38 hours

Final Project 40 hours

**Academic Integrity (required policy)**

All written work for this class must be your original work.  Presenting material from other sources, either print or electronic, as one’s own work constitutes plagiarism. Please review Merrimack College’s Academic Integrity Code: <http://catalog.merrimack.edu/content.php?catoid=9&navoid=202#academic-integrity>

Please consult the library’s web site for a complete discussion on academic integrity [http://libguides.merrimack.edu/content.php?pid=120821 for a complete discussion of academic integrity.](http://libguides.merrimack.edu/content.php?pid=120821)

**Academic Accommodations from the Accessibility Services Office (required policy)**

Regardless of whether the course or the student is on-campus or remote, Merrimack College provides reasonable accommodations for students with documented disabilities through the Accessibility Services Office. Students who have, or think they may have, a disability are invited to contact the Accessibility Services Office via the online request form found on the Accessibility Services website: www.merrimack.edu/aso, email accessibilityservices@merrimack.edu or by visiting us on the third floor of McQuade Library (subject to change if the college is remote).

Students are encouraged to contact the office as soon as possible via the website or via email at accessibilityservices@merrimack.edu to ensure adequate time to meet and create a plan. Students already registered with Accessibility Services are encouraged semesterly to request for their letters to be emailed and students are responsible to then email the letter to their instructors personally. The Accessibility Services Testing Center remains available to students whether in-person or remote. While it is understood that some students will not use all accommodations in all courses, accommodations cannot be made retroactively.

**Live Meeting Attendance**

Participation in weekly live meetings is highly encouraged, but it is not mandatory. All sessions will be recorded and recordings will be posted; however, when not attending students forego the ability to directly ask questions about topics being discussed, or other parts of the course.

**Requests for Extensions**

The general policy is that, outside of properly verified serious medical emergencies\* (as defined below), extensions are not given, which applies to the Final Exam, Position Papers and Analytic Plan. Missing an assignment without an acceptable reason (to be clear, that means a serious medical emergency, as defined below) will result in 0 points for the exam or a project. The intent here is not to penalize anyone – quite to the contrary, it is to create a level playing field so that no one has a unique and an unfair advantage. The exam and project due dates are published (see the Live Meetings, Topics & Assignments section below) and will not change, barring a natural or other emergency – please consider those dates when planning any non-class related activities.

\**Serious medical emergency is defined as an injury or illness that is acute and poses an immediate risk to a person’s life or long term health. To be “properly verified”, the said serious medical condition must be attested to by hospitalization and related medical treatment documentation*.

**Weekly Topics, Live Meetings & Assignments**

|  |  |  |
| --- | --- | --- |
| **Date** | **Topic** | **Assigned Readings** |
| **Week 1** | **HW Chapter 2** | |
| Week 1 | * Introduction and Organization of Class * Types of variables and their distinguishing characteristics * Prescriptive Statistics v. Predictive Modeling * Understanding response and explanatory variables * What is statistical learning? * Assessing model accuracy * Use of R and RStudio for basic data tasks * HW Chapter 2: Problems 8, 9, 10 | Chapter 1**,** Chapter 2 **,**  Chapter 2 Lab |
| **Week 2** | **HW Chapter 3** | |
| Week 2 | * Assessing model accuracy * Simple and Multiple Linear Regression * HW Chapter 3: Problems 9, 10, 14 | Chapter 2, Chapter 3  Chapter 3 Lab |
| **Week 3** | **HW Chapter 4** | |
| Week 3 | * Logistic Regression * Linear Discriminant Analysis * Quadratic Discriminant Analysis * K-Nearest Neighbors * HW Chapter 4: Problems 13, 16 | Chapter 4  Chapter 4 Lab |
| **Week 4** | **HW Chapter 5** | |
| Week 4 | * Cross-Validation * The Bootstrap * HW Chapter 5: Problems 6, 9 | Chapter 5  Chapter 5 Lab |
| **Week 5** | **HW Chapter 6 & Draft Outline of Final Project** | |
| Week 5 | * Linear Model Selection * Subset Selection * Shrinkage Methods * Dimension Reduction Methods * HW Chapter 6: Problems 9, 11 | Chapter 6  Chapter 6 Lab |
| **Week 6** | **HW Chapter 8** | |
| Week 6 | * Tree Based Methods * Bagging, Random Forests, Boosting * Continue working on the final project | Chapter 8  Chapter 8 Lab |
| **Week 7** | **HW Chapter 8 , cont.** | |
| Week 7 | * Bagging, Random Forests, Boosting * HW Chapter 8: Problems 8, 9, 10 | Chapter 8  Chapter 8 Lab |
| **Week 8** | **Final Project** | |
| Week 8 |  |  |