UNT
TOULOUSE GRADUATE SCHOOL
Advanced Data Analytics

Spring 2021 - SYLLABUS – 8 Week Course
ADTA 5410 - Applications & Deployment of Advanced Analytics

Course Information

- ADTA 5410 – Applications and Deployment of Advanced Analytics
- Spring 2021 – 8W2: March 8 – May 1
- Class Meeting Time/Location – We will meet virtually via Zoom on Thursdays 7:00 pm - 9:30 pm. The synchronous Zoom meetings will be recorded and are optional.

Professor Contact & Communication

- Dr. John Garcia, DBA, MS, CAP, CPA
- Office hours: Monday, Wednesday, Friday 4 pm – 6 pm (virtual & by appointment).
  - I am **happy** to meet with students for academic advising, **help** with assignments, or just to **chat**.
  - Using this link, you can schedule time with me.
- John.Garcia2@unt.edu / (469) 296-8426 (call or text)
- You can also find me on Microsoft Teams if you'd liked to chat.
- You can also send me anonymous feedback through this survey form.

About the Professor

Welcome to ADTA 5410! I am Dr. John Garcia, the professor for this course. Before joining the UNT faculty in January 2020, I was a Finance & Analytics executive at Toyota, where I worked for 15 years in various finance and analytics roles. Before Toyota, I worked at Ernst & Young for six years. I hope to bring my 21 years of industry experience to the classroom and provide you not only the theoretical background but also the practical implications that you will see in business. Like most of my students, I have not followed the traditional academic path, which I think enables me to see both the academic and practitioner viewpoints, thereby helping me add practical, real-world meaning to the textbook material. The variety of career and academic experiences that we all bring will also provide the foundation for exciting course discussions.

I have a Doctorate degree in Finance & Analytics from Creighton University, a Masters in Predictive Analytics from Northwestern University, a Masters in Accountancy from the University of Notre Dame, and I received my Bachelor's degree in business from Cal Poly in San Luis Obispo, CA.

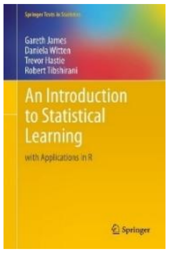
I am excited to have you in this course, and I look forward to learning more about you and your career goals. Together we will explore a variety of advanced analytics tools, learn about how and when to use them, interpret the outputs of the analysis, and describe the results in ways that will help us or others take appropriate actions to achieve the desired outcomes or goals. I believe data science is also part art as it is impossible to consider data divorced from people, and understanding how to influence people is more art than science. Thus, throughout the course, we will also discuss the art of data science. I look forward to our learning journey!

Course Pre-requisites, Co-requisites, and/or Other Restrictions

ADTA 5130 Data Analytics 1, ADTA 5230 Data Analytics 2, ADTA 5250 Large Data Visualization, ADTA 5240 Harvesting, Storing and Retrieving Data, ADTA5340 Discovery and Learning with Big Data

Required Materials

One open-source textbook is required for this course. Students will also need to have access to R and R studio.



Gareth James, D, Witten, T. Hastie, T and R. Tibshirani (2017). *An Introduction to Statistical Learning, with Applications in R. Open Textbook*: Download the eBook

Suggested Textbooks

- L Kuhn, Max, and Kjell Johnson (2018). *Applied predictive modeling*. Vol. 26. New York: Springer. (Available for free from the UNT library as a PDF or EPUB)
- Hastie, T., Tibshirani, R., and Friedman, L (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Download here.(This is a more advanced version of ISLR)
- Wickham,H. & Grolemund, G. (2019). *R for Data Science*. Download the PDF
- Chang, W. (2020). R Graphics Cookbook, 2nd Edition
- Lantz, Brett. (2019). *Machine Learning with R, 3rd* Ed. *[As an R resource]*
- van Buuren, S (2012). Flexible Imputation of Missing Data, Boca Raton: CRC Press
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018 *Learning from imbalanced data sets*. Berlin: Springer. (Available for free from the UNT library as a PDF or EPUB)

Guidelines for Success

While unforeseen events do happen that can make college life and achievement difficult, generally speaking success is a choice. In order to help yourself and others succeed, avoid distractions during class (like cell phones, iPads, social media, or games), and utilize the resources at your side. You have many valuable resources: textbooks, college services, each other, and myself. Don't hesitate to ask for help and always communicate. Be sure to read your assigned readings, be punctual, and save all your assignments (and back them up!) Follow these guidelines and you'll be well on your way.

Course Description

The course focuses on the application of machine learning methods explored in earlier courses, which use data and statistical techniques to predict outcomes. Students will learn through a hands-on approach to build and tune models using R to predict categorical and continuous outcomes, test those models, interpret and present the results. The focus will be on the application of machine learning models implemented in R while balancing the trade-off between prediction power and model interpretability. The course covers how to formulate a model for a given decision problem, perform analysis with the model to generate insights, and effectively communicate those insights.

Course Objectives

By the end of the course, students should be able to:

1. Apply machine learning methods using R to build predictive models and discover patterns in data to develop analytic solutions to practical business problems and enable more informed business decision-making.
2. Students will learn how to fit and evaluate a variety of predictive models, including classification and regression trees, support vector machines, logistic and linear regression models, regularized regression, generalized linear models, GAMs, discriminant analysis, tree ensemble models, Naive Bayes, k-nearest neighbors, neural networks, ensemble learning, and clustering methods.
3. Develop analytic solutions to practical business problems using the R statistical programming language, transforming data into knowledge.
4. Students will learn strategies in data wrangling, feature engineering, missing value imputation, and data pre-processing techniques such as synthetic resampling to improve predictive models.
5. Describe the data analytics project lifecycle and key elements of each phase.
6. Effectively communicate analysis results and insights verbally and in writing, presenting descriptive statistics and models in a business context and employing appropriate data visualizations.

Course Topics

1. Execution of analytics projects using the CRISP-DM framework and the R programming language.
2. Classification models: K-nearest neighbors, naive Bayes, logistic regression, neural networks, classification trees, bagging, random forests, discriminant analysis, boosting, support vector machines, and ensemble learning methods.
3. Regression models: K-nearest neighbors, linear regression, partial least squares regression, regression trees, tree ensembles, generalized additive models, non-linear regression, support vector machines, penalized regression (ridge, lasso, elasticnet), generalized linear models (Poisson, negative binomial), and ensemble learning methods.
4. Linear model selection and regularization and the challenges of high-dimensional data.
5. Model fitting, hyperparameter tuning, and model evaluation.
6. Data pre-processing techniques including resampling techniques and addressing imbalanced datasets using synthetic sampling methods, and advanced missing value imputation methods.
7. Feature selection, measuring variable importance, and visualizing data.
8. Unsupervised learning techniques including principal component analysis and clustering methods.
9. Presenting the story of the data; including the effective structuring of analytical presentations, framing of insights, and supporting data.
10. Model stacking, automated machine learning, and model interpretability.

Course Requirements

Your final grade will be determined based on the assignments noted in the table below. The total number of points received will be divided by the total possible number of points.

Assignments	**Points Possible**	**Percentage of Final Grade**
Quizzes Six quizzes based on the course materials (20 points each)	120 points	12%
Eight R Practice Assignments: 7 DataCamp R Practice Modules and 1 Coursera R project (10 points each)	80 points	8%
R Programming Assignments Eight programming assignments (R labs - 25 points each) (the lowest score will be dropped)	200 points	20%
Three Group Projects (80 points each)	240 points	24%
Final Group Project	360 points	36%
Total Points Possible	**1000 points**	**100%**

Grading

Course grades will be assigned as follows (100% — 89.5%, A; 89.5% — 79.5%, B; 79.5% — 69.5%, C; 69.5% — 60%, D; <60%, F).

Course Assignment, Examination, and or Project Policies

Quizzes (12%)

- There will be an open book quiz for each of the 10 chapters.
- They are worth 20 points each and must be completed by the due date as indicated on Canvas.
- You will have one attempt and 45 minutes to complete the quiz.
- The quizzes will be multiple-choice questions designed to reinforce the textbook content.

DataCamp R Practice Assignments (8%)

- To help gain hands-on experience in applying the statistical learning techniques using R, this course will include eight R DataCamp assignments and one Coursera R project, each worth 10 points.
  - Additional DataCamp modules will be included as optional or extra credit assignments.
- To earn full marks, you only need to finish the DataCamp module, your score on the DataCamp activities will not be factored in your grade (e.g., a completed DataCamp by the assigned deadline will be awarded 10 points).
  - Note that in DataCamp, you may get the answers to the exercises, but try as many of the exercises so you get more practice in R, and if you request the answer, review the code to understand the solution.
  - If you do not finish the assigned DataCamp module by the assigned date, you will earn a 0.

R Programming Assignments (R Labs) (20%)

- To help students gain hands-on experience in applying statistical learning techniques using R, this course will include eight R lab tutorials. *The lowest R lab score will be dropped from your final grade.*
- After completing the R lab tutorials, you may start the R programming assignment at any time during the Thursday-Sunday window that the assignment will be open.
- You must turn in your R code via Canvas. Failure to turn in your code will result in a 20% reduction in your score.

Group Projects (24%)

- There will be three projects, which will include a report and presentation component. Each project is worth 80 points (60 points for the report and 20 points for the presentation).
- Instructions for each of the projects are provided in Canvas.
- The projects will be completed in groups of 2-3 (same group as your final project group)

Final Group Project (36%) *(Final Report 275 / Presentation 85 pts)*

- The final project will be completed in groups of 2-3 (assigned by the professor).
- See Canvas for a detailed description of the project.
- The final project and presentation are due at the end of the course.
- Part of the final project is a brief 10-12 minute presentation.
- You are to use APA style for citations and reference list. The minimum requirement for the paper will be ten pages of content, double-spaced, 1-inch margins, using Arial or Times Roman 12-point font.
- It is recommended that the submitted research paper also include a separate cover page that includes your name and the title of your paper.
- A rubric for the project is available in Canvas.
- The presentation is due on the date specified in the course calendar. Late papers will not be accepted.
- The paper will be submitted for grading via software that checks for plagiarism.

Extra Credit Assignments

There are several extra credit opportunities available (see the course calendar and Canvas for details). If you choose not to complete an extra credit assignment by the due date, you may still submit it after the due date for an opportunity to earn up to 70% of the original potential points.

Late Assignment Policy

All work for this course is due no later than 11:59 pm on the designated due. Any assignment submitted after that time will receive a highest possible score of 60% through 48 hours past the deadline. Additional points may be deducted when the assignment is graded based on the quality of the work submitted. No points will be awarded for assignments turned in 48 hours or more past the due date**.** Please don't lose valuable points this semester by turning in work late. Late work is subject to the penalty described above unless previously approved by the instructor**.**

Attendance

Students are encouraged to login regularly to the online class site and attend class virtually. Students are also required to participate in all class activities such as discussion boards, chats. The weekly synchronous sessions will be optional, but it is highly recommended that you regularly attend.

COVID-19 Impact on Attendance

While attendance is expected as outlined above, it is important for all of us to be mindful of the health and safety of everyone in our community, especially given concerns about COVID-19. Please contact me if you are unable to attend class because you are ill, or unable to attend class due to a related issue regarding COVID-19.

If you are experiencing cough, shortness of breath or difficulty breathing, fever, or any of the other possible symptoms of COVID-19 (https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html) please seek medical attention from the Student Health and Wellness Center (940-565-2333 or askSHWC@unt.edu) or your health care provider. UNT also asks that you contact the UNT COVID Hotline at 844-366-5892 or COVID@unt.edu for guidance on actions to take due to symptoms, pending or positive test results, or potential exposure. While attendance is an important part of succeeding in this class, your own health, and those of others in the community, are more important.

Turnitin Notice

Turnitin is used as a tool to assist students in their scholarly writing to address plagiarism issues. All works submitted for credit must be original works created by the scholar uniquely for the class. It is considered inappropriate and unethical, particularly at an advanced undergraduate/graduate level, to make duplicate submissions of a single work for credit in multiple classes, unless specifically requested by the instructor. It is also considered inappropriate and unethical to work together on individual assignments or share work that is to be created on an individual level. Work submitted at the senior/graduate level is expected to demonstrate higher-order thinking skills and be of significantly higher quality than work produced at the lower undergraduate levels. It is recommended that students use the Turnitin resource to ensure their work is free of copyright issues prior to final submission of their projects.

Grades of Incomplete

Grades of Incomplete will only be given per university policy as outlined by the Office of the Registrar.

Copyright Notice

Some or all of the materials on this course web site may be protected by copyright. Federal copyright law prohibits the reproduction, distribution, public performance, or public display of copyrighted materials without the express and written permission of the copyright owner, unless fair use or another exemption under copyright law applies. Additional copyright information may be located at: http://policy.unt.edu/policy/08-001.

Additional UNT Policies

Please review and familiarize yourself with the additional UNT policies outlined in the Canvas site.

Course Calendar

Below is a tentative schedule. Should any change become necessary, it will be announced in class and in the announcements sent via the UNT email. It is your responsibility to check for changes in the schedule.

Week	***Topic / Reading***	***Assignments***
Week 1 ***Mar 11*** (Module 1)	***1A) Introduction to Machine Learning and R*** *Course overview, Syllabus review, and the CRISP-DM process* *Read the following chapters in An Introduction to Statistical Learning, with Applications in R (ISLR)* Required: 1. ISLR - C1. Introduction Suggested: 1. Applied Predictive Modeling (APM) – Appendix B: An Introduction to R 2. The Elements of Statistical Learning (ESL) – Chap 1 ***1B) Statistical Learning and Exploratory Data Analysis in R*** Required: 1. ISLR - C2. Statistical Learning Suggested: 1. ESL – C2. Overview of Supervised Learning	*Install R and R Studio*[1] *Complete Quiz 1* *Complete DataCamps:* *#1– Intro to R* *#2– EDA in R* *Complete R Lab 1* *Extra Credit (EC -5 pts): DataCamp: Intermediate R*
Week 2 ***Mar 18*** (Module 2)	***2A) Multiple Linear Regression*** Required: *ISLR - C3. Linear Regression* Suggested: 1. MLR – C6**.** Forecasting Numeric Data – Regression Methods (p. 167-199) 2. APM – C3: Data Pre-Processing 3. ESL – C3. Linear Methods for Regression ***2B) Advanced Regression Methods & KNN Regression*** Suggested: 1. APM – C6: Linear Regression and Its Cousins 2. APM – C7.4: K-Nearest Neighbors 3. DataCamp: Reporting with R Markdown	*Complete Quiz 2* *Complete R Lab 2* *Complete Project #1* *Complete DataCamp #3- Supervised Learning in R: Regression (Sections 1-4)*

[1] R: https://cran.r-project.org /RStudio: https://www.rstudio.com

Week	***Topic / Reading***	***Assignments***
Week 3 ***Mar 25*** (Module 3)	***3A) Resampling Methods & Model Tuning*** *Required:* • *ISLR – C5. Resampling Methods* • *APM – C4: Over-Fitting and Model Tuning* Suggested: 1. ESL – C7: Model Assessment and Selection 2. APM – C5: Measuring Performance in Regression Models ***3B) Linear Model Selection and Regularization*** Required: *ISLR – C6.-C6.3 Linear Model Selection and Regularization* Suggested: *ESL – C3.4 Shrinkage Methods*	*Complete Quiz 3* *Complete R labs 3-4* *Complete DataCamp #4– Hyperparamter Tuning* *Extra Credit: Home price Prediction*
Week 4 ***Apr 1*** (Modules 4)	***4A) Dimension Reduction Methods & Non-Linear Models*** Required*: ISLR – C6.3.-C6.7 Linear Model Selection and Regularization* Suggested: APM 6.3-6.5 – Partial Least Squares and Penalized Models ***4B) Non-Linear Models (Polynomial regression, step functions & splines, Local Regression, GAMs and GLMs (Poisson/Negative Binomial)*** Required: • *ISLR – C7. Moving Beyond Linearity* • *APM – C7.2 Multivariate Adaptive Regression Splines* Suggested: ESL – C9.1: Generalized Additive Models	*Complete Quiz 4* *Complete DataCamp #5- Modeing in R w/ GAMS* *Complete R labs 5-6* *Complete Project #2* *Extra Credit: The Ethics of Analytics*

Week 5 ***Apr 8*** (Module 5)	***5A) Classification Models (Part 1 – KNN, Naïve Bayes, Discriminant Analysis)*** Required: ISLR – C4. Classification Suggested: • ESL – C4: Linear Methods for Classification • APM C11. Measuring Performance in Classification Models • APM: C12.3-12.5, C13.5-13.6 ***5B) Classification Models (Part 2 – Neural Networks) and Learning from Imbalanced Datasets*** Required: • APM – 7.1 & 13.2 Neural Networks • APM - C16: Remedies for Severe Class Imbalance • Chapter 2 from: Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Berlin: Springer. Suggested: ESL – C11: Neural Networks	*Complete Quiz 5* *Complete R Lab 7* *Complete DataCamp #6- Classification* *Complete Coursera: Predicting Credit Card Fraud with R* *Extra Credit: KNN, NB, NN Classification*
Week 6 ***Apr 15*** (Modules 6)	***Module 6: Tree-Based Methods, Bagging, Random Forests, & Boosting, and Automated Machine Learning*** *Required:* • ISLR – C8. Tree-Based Methods Suggested: • APM C14. Classification Trees and Rule-Based Model • ESL - C9.2: Tree-Based Methods • ESL – C15: Random Forests	*Complete Quiz 6* *Complete R Lab 8* *Complete DataCamp #7– Tree Based Models*
Week 7 ***Apr 22*** (Module 7)	***7A) Support Vector Machines and Kernel Methods*** *Required:* • ISLR C9. – Support Vector Machines and Kernel Methods Suggested: ESL – C12: Support Vector Machines and Flexible Discriminants APM – 13.4: Support Vector Machines Practice: DataCamp: Support Vector Machines in R ***7B) Unsupervised Learning – K-Means & Hierarchical Clustering*** Required: ISLR – C10. Unsupervised Learning Suggested: • DataCamp: Cluster Analysis in R • ESL – C14: Unsupervised Learning	*Complete Project #3*

Week 8 ***Apr 29*** (Module 8)	***Presentations & Final Project*** *Project Presentations and Paper*	*Final Project Paper Due Fri. Apr. 30th at 11:59pm CST* *Final Project Presentation due Tues. Apr 29 at 6:00pm CST*

SCHOLARLY EXPECTATIONS

UNT Code of Student Conduct

You are encouraged to become familiar with the University's Code of Student Conduct and the Policy of Academic Integrity (Links to an external site.) found on the Dean of Students website. The Dean of Students Office (opens in a new window) (Links to an external site.) enforces the Code. The Code explains what conduct is prohibited, the process the DOS uses to review reports of alleged misconduct by students, and the sanctions that can be assigned. When students may have violated the Code they meet with a representative from the Dean of Students Office to discuss the alleged misconduct in an educational process. The University's expectations for student conduct apply to all instructional forums, including University and electronic classroom, labs, discussion groups, field trips, etc.

Of particular interest are the following terms:

- **Cheating** – intentionally using or attempting to use unauthorized materials, information, or study aids in any academic exercise. The term academic exercise includes all forms of work submitted for credit or hours.
- **Plagiarism** – the deliberate adoption or reproduction of ideas, words, or statements of another person as one's own without acknowledgement.
- **Fabrication** – intentional and unauthorized falsification or invention of any information or citation in an academic exercise.
- **Facilitating academic dishonesty** – intentionally or knowingly helping or attempting to help another to violate a provision of the institutional code of academic integrity.

The policies contained on the course website apply to this course. In addition, you are expected to adhere to the ADTA Academic Integrity Policy outlined below. If you have questions regarding any of the information presented regarding academic integrity, please feel free to contact me.

Academic Integrity

All works submitted for credit must be original works created by the scholar uniquely for the class. It is considered inappropriate and unethical, particularly at the graduate level, to make duplicate submissions of a single work for credit in multiple classes, unless specifically requested by the instructor. Work submitted at the graduate level is expected to demonstrate higher-order thinking skills and be of significantly higher quality than work produced at the undergraduate level.

ADTA Academic Integrity Policy

Occurrence	**Minor Assignments (e.g., Discussions, Homework, and Journals)**	**Major Assignments (e.g., Exams and Projects worth more than 10% of your grade)**
1st Warning	1. First written warning 2. Min. 20% deduction	1. Written warning 2. Min. 15% deduction
2nd Warning	1. Second written warning 2. Min. 50% deduction 3. Inform academic advisor during Dept. Meeting	1. Second written warning 2. Min. 50% 3. Inform academic advisor during Dept. Meeting
3rd Warning	1. Written Letter 2. Min. 0 grade for that assignment	1. Written Letter 2. Min. 0 grade for that assignment