STUDENT WARNING: This course syllabus is from a previous semester archive and serves only as a preparatory reference. Please use this syllabus as a reference only until the professor opens the classroom and you have access to the updated course syllabus. Please do NOT purchase any books or start any work based on this syllabus; this syllabus may NOT be the one that your individual instructor uses for a course that has not yet started. If you need to verify course textbooks, please refer to the online course description through your student portal. This syllabus is proprietary material of APUS.

DATS411

Course Summary

Course : DATS411 **Title : Advanced Data Science**
Length of Course : 8 Weeks
Prerequisites: DATS311 Intermediate Data Science
Credit Hours : 3
Contact: frank.appiah@mycampus.apus.edu

Description

This course completes the three-course sequence in Data Science. This advanced course takes students through the application of more advanced methods in regression and time series models. It includes discussions about causal inference, and a wide-range of time series models. This course emphasizes tools and methods used to capture key patterns and generate insight from data.

Course Scope:

This course is intended for students seriously considering a career in data science. It assumes a fairly high level of mathematics in calculus as well as probability and statistical foundation with the ability to interpret outcomes effectively. It provides students with knowledge and skills for analyzing unstructured, numerical, categorical and time series data. It will also equip students with tools including a variety of computational methods used to conduct analyze unstructured, times series and count data respectively. Students will complete this course with the basic understanding of how to conduct analyses on data that is unstructured, count data and data with time component.

Objectives

At the conclusion of this course, students will be able to:

- Identify and discuss different types of data structures and types of outcome variables including time-series data, panel data, count data, censored data, and duration data.
- Validate key assumptions and run sensitivity analyses for ordinary least squares models.
- Estimate, evaluate and graphically present results from interaction effects.
- Address issues of autocorrelation in time series and panel data.
- Estimate, evaluate and interpret results from discrete choice models such as logit, ordinal logit, and multinomial logit models.
- Estimate count models such as Poisson and negative binomial models.

Course Materials

References:

R for Data Science by Dan Toomey
Natural Language Processing with Python by Steve Bird, Ewan Klein and Edward Loper
Python for Data Analyses by Wes McKinney
https://www.tidytextmining.com/tidytext.html
Python practice: https://app.finxter.com/learn/computer/science/
Python text http://www.datascienceassn.org/sites/default/files/Natural%20Language%20Processing%20with%20Python.pdf
https://towardsdatascience.com/a-guide-to-forecasting-in-r-6b0c9638c261

http://r-statistics.co/Time-Series-Analysis-With-R.html#What%20is%20Autocorrelation%20and%20Partial-Autocorrelation?

http://r-statistics.co/Time-Series-Analysis-With-R.html

Software (Required)

- Install R and RStudio on windows from: https://cran.r-project.org/bin/windows/base/
and https://rstudio.com/products/rstudio/download/ (scroll down to All installers and select windows version)
- Install R and Rstudio on mac from: https://cran.r-project.org/bin/macosx/
and https://rstudio.com/products/rstudio/download/ (scroll down to All installers and select mac version)

Software (Optional)

- Python version 3 on mac and windows available at: https://www.python.org/downloads/
- Other software usage will require approval from the instructor.

Readings:

Week	Topic	Reading Link/Book	Reading Notes

1	Introduction to unstructured data	https://www.tidytextmining.com/tidytext.html	
2	Accessing & Processing Unstructured data Python	http://www.datascienceassn.org/sites/default/files/Natural%20Language%20Processing%20with%20Python.pdf	
3	Analyzing sentiments	https://www.tidytextmining.com/sentiment.html	
4	Term frequency inverse document frequency (TFIDF)	https://www.tidytextmining.com/tfidf.html	
5	Review linear regression check and validate key assumptions	None	.Rmd file in folder
6	Review logistic regression models and validation techniques.	Chapter 4.3 from **intro to statistical learning**	Read .Rmd file on this chapter
7	Poisson regression model and validation	Chapter 3.3 in the **Intro to categorical analyses textbook**	Notes in .Rmd and churn data (for analyses)
8	Introduction to time series data - including visualization, terminologies etc.	http://r-statistics.co/Time-Series-Analysis-With-R.html#What%20is%20Autocorrelation%20and%20Partial-Autocorrelation http://r-statistics.co/Time-Series-Analysis-With-R.html	Notes in .Rmd file

Outline

Week 1 Introduction to unstructured data

At completion of the modules this week students will be able to:

- Get all required software installed (R, Python and anaconda jupyter notebook)
- Install tm package in R
- Use R and python to access unstructured data.
- Recognize and analyze unstructured data.

Module 1: Install Python, R and jupyter notebook using one of the two sources below

Windows:

Install R: https://www.r-project.org/
install R Studio https://rstudio.com/products/rstudio/download/
Install python/Anaconda: https://www.python.org/downloads/
Install jupyter notebook: https://www.anaconda.com/products/individual#windows

Mac: https://www.r-project.org/
install R Studio https://rstudio.com/products/rstudio/download/
Install python/Anaconda: https://www.python.org/downloads/
Install jupyter notebook: https://www.anaconda.com/products/individual#windows

Module 2: Install tm package in R

Module 3: Practice loading and analyzing unstructured data in R

Reading:

https://www.tidytextmining.com/tidytext.html

Lab 1 – Reference lab1.rmd file.

HW: reference hw1.rmd file reference

Week 2: Accessing & Processing Unstructured data Python

At completion of the modules this week students will be able to:

- Understand how to install packages in Python with pip install.
- Use python scripting language to load unstructured data
- Understand the basics of natural language processing such as, tokens, stop words removal, word frequency etc.

Module 4: Run and discuss lab 2

Reference lab2

Module 5: Unit 3.1 Accessing text data from web and disc

Module 6: Unit 3.2 Text processing at lower level

Module 7: Unit 3.4 Regular expressions for detecting word pattern

Module 8: Unit 3.6 Normalizing text

HW run all the script in DATS411_HW_2

Week 3: Analyzing sentiments

At completion of the modules this week students will be able to:

- Replicate all the analyses in the reading list in R
- Visualize text via word cloud and bar graphs
- Compare and contrast the relative visual representations
- Understand sentiments data and manipulate them

Module 9: **Unit 2.1-2.2** The sentiments datasets – dealing with sentiment datasets in R

Module 10: **Unit 2.3** Comparing the three sentiment dictionaries

Module 11: **Unit 2.4-3.6** Most common positive and negative words

Reading: https://www.tidytextmining.com/sentiment.html

HW 3: Complete the codes in DATS411_HW_3.rmd

Week 4: Term frequency inverse document frequency (TFIDF)

At completion of the modules this week students will be able to:

- Understand TFIDF
- Be able to replicate examples
- Understand what a corpus is and be able to analyze it.

Model 12: **3.1-3.2** Term frequency in Jane Austen’s novels & Zipf’s law

Model 13 Unit 3.3 The bind_tf_idf() function

Module 14: **Unit 3.4-3.5** A corpus of physics texts

Reading: https://www.tidytextmining.com/tfidf.html

HW 4 Included as a .rmd file

Week 5: Review linear regression check and validate key assumptions

At completion of the modules this week students will be able to:

- Understand fundamental linear regression assumptions

- Understand how to examine the different types of assumptions
- Understand how to fix the different types of deviations from the assumptions.

Model 16: Review Unit 3.1-3.7 and attached reading on ‘What is OLS’

Model 17: Read attached document on ‘Diagnostics of linear models and how to remedy them’

Reading: .rmd file in folder and DATS 201 textbook (chapter 3)

http://ezproxy.apus.edu/login?url=https://ebookcentral.proquest.com/lib/apus/detail.action?docID=6312402
HW 5: HW5.rmd

Week 6: Review logistic regression models and validation techniques

At completion of the modules this week students will be able to:

- Understand the need for logistic regression as a classifier
- fit a logistic regression
- Perform model fit checks and validation techniques

Model 18: Unit 4.1-4.3 An Overview of Classification ad logistic regression

Model 19: Fitting and validating logistic regression (Notes: Read .rmd file on this chapter)

Reading: Review ch 4.3 of 201 textbook
http://ezproxy.apus.edu/login?url=https://ebookcentral.proquest.com/lib/apus/detail.action?docID=6312402
HW 6: HW6.rmd

Week 7: Poisson regression model and validation

At completion of the modules this week students will be able to:

- Fit and interpret Poisson regression model
- Fit and validate Poisson regression model

Model 20: **Unit 3.3.1-3.3.3** Poisson regression and Overdispersion

Model 21: **Unit 3.3.4** Negative binomial regression

Model 22: **Notes** Fitting and validating Poisson Regression Models

Reading: chapter 3.3 in the Intro to categorical analyses textbook http://ezproxy.apus.edu/login?url=https://ebookcentral.proquest.com/lib/apus/detail.action?docID=290465

Notes in .rmd and include churn data
Hw 7

Week 8: Introduction to time series data -include visualization, terminologies etc.

At completion of the modules this week students will be able to:

- Understand the mathematical representation of time series models
- Fit and interpret time series data
- Perform time series decomposition, autocorrelation and (de) seasonality checks

Module 24: **Notes** What is time series and how are they represented mathematically

Module 25: Notes lags and Decomposition of time series data

Module Notes Autocorrelation and de-seasonalizing times series data

HW 8

Add forecasting to hw

Final Exam- comprehensive

Covers 5 broad questions:

Two short questions on python and R-make practical and discuss unstructured data

One question to tokenize, create word freq, word cloud and tfidf in R

One question to analyze sentiments and fit logistic regression

One question on time series forecasting

Evaluation

A variety of weekly exercises (labs, homework and discussions) will be used to reinforce the material covered in this course. This course will include a final exam, weekly homework, discussions and labs. The grade distribution of the course is shown below. The final exam is comprehensive which means it will cover all the concepts learned in the course.

- Homework 40%
- Laboratories 15%
- Discussions 15%
- Comprehensive Final Exam 30%

Late Assignments

- Students are expected to submit classroom assignments by the posted due date and to complete the course according to the published class schedule. The due date for each assignment is listed under each Assignment.
- Generally speaking, late work may result in a deduction up to 15% of the grade for each day late, not to exceed 5 days.
- As a working adult I know your time is limited and often out of your control. Faculty may be more flexible if they know ahead of time of any potential late assignments.

Policies

Please see the Student Handbook to reference all University policies. Quick links to frequently asked question about policies are listed below.

Drop/Withdrawal Policy

Plagiarism Policy

Extension Process and Policy

Disability Accommodations

Writing Expectations

All written submissions should be submitted in a font and page set-up that is readable and neat. It is recommended that students try to adhere to a consistent format, such as that described below.

- Typewritten in double-spaced format with a readable style and font and submitted inside the electronic classroom (unless classroom access is not possible and other arrangements have been approved by the professor).
- 11 or 12-point font in a style such as Arial, Helvetica or Times New Roman.

Citation and Reference Style

Assignments completed in a narrative essay or composition format must follow a widely accepted citation style, such as APA, Turabian or MLA. Please refer to the APUS Online Library for further examples, or contact the instructor with questions.

Late Assignments

Students are expected to submit classroom assignments by the posted due date and to complete the course according to the published class schedule. As adults, students, and working professionals, I understand you must manage competing demands on your time. Should you need additional time to complete an assignment, please contact me **before the due date** so we can discuss the situation and determine an acceptable resolution. Routine submission of late assignments is unacceptable and may result in points deducted from your final course grade.

Netiquette

Online universities promote the advancement of knowledge through positive and constructive debate – both inside and outside the classroom. Forums on the Internet, however, can occasionally degenerate into needless insults and “flaming.” Such activity and the loss of good manners are not acceptable in a university setting – basic academic rules of good behavior and proper “Netiquette” must persist. Remember that you are in a place for the rewards and excitement of learning which does not include descent to personal attacks or student attempts to stifle the Forum of others.

- **Technology Limitations:** While you should feel free to explore the full-range of creative composition in your formal papers, keep e-mail layouts simple. The Sakai classroom may not fully support MIME or HTML encoded messages, which means that bold face, italics, underlining, and a variety of color-coding or other visual effects will not translate in your e-mail messages.
- **Humor Note:** Despite the best of intentions, jokes and especially satire can easily get lost or taken seriously. If you feel the need for humor, you may wish to add “emoticons” to help alert your readers: ;-), :)

Disclaimer Statement

Course content may vary from the outline to meet the needs of this particular group.

Online library

The Online Library is available to enrolled students and faculty from inside the electronic campus. This is your starting point for access to online books, subscription periodicals, and Web resources that are designed to support your classes and generally not available through search engines on the open

Web. In addition, the Online Library provides access to special learning resources, which the University has contracted to assist with your studies. Questions can be directed to **librarian@apus.edu**.

- **Charles Town Library and Inter Library Loan:** The University maintains a special library with a limited number of supporting volumes, collection of our professors' publication, and services to search and borrow research books and articles from other libraries.
- **Electronic Books:** You can use the online library to uncover and download over 50,000 titles, which have been scanned and made available in electronic format.
- **Electronic Journals:** The University provides access to over 12,000 journals, which are available in electronic form and only through limited subscription services.
- **Tutor.com**: AMU and APU Civilian & Coast Guard students are eligible for 10 free hours of tutoring provided by APUS. Tutor.com connects you with a professional tutor online 24/7 to provide help with assignments, studying, test prep, resume writing, and more. Tutor.com is tutoring the way it was meant to be. You get expert tutoring whenever you need help, and you work one-to-one with your tutor in your online classroom on your specific problem until it is done.

Library Guide (http://apus.campusguides.com/SCIN134)

The AMU/APU Library Guides provide access to collections of trusted sites on the Open Web and licensed resources on the Deep Web. This course guide provides links to a number of sources relevant to this course, including journals, books, and web sites. Also, you can directly contact the librarian assigned to this course for assistance in locating information.