DATA SCIENCE PROJECT CASE STUDIES | MATH 6397 | CLASS NUMBER: 18396 | COURSE ID: 031328

YEAR COURSE OFFERED: 2019

SEMESTER COURSE OFFERED: Spring Session

Instructor (I): Luis Arregoces | email @uh.edu| (Working on getting UH email addresses)

Instructor (II): Sid Bhattacharya | email @uh.edu|

Instructor (III): Brandon Ellet? | email @uh.edu |

DEPARTMENT: MATH

COURSE DESCRIPTION:

Apply multiple techniques for exploratory data analysis, visualize and understand the data using Inferential Statics, Descriptive Statistics, and probability Distributions.

Visualization is an important part of the typical Data Science workflow, in the class we will leverage plotting libraries like matplotlib, and seaborn in Python and ggplot2 in R. If you are familiar with other exploratory analysis tools such as Tableau, Excel and D3.js, you can use them in the problem definition and understanding section.

Leverage the results from the exploratory analysis to guide the analytic approach that will be used to solve the problem. For example: (1) create predictions on key variables, (2) search for patterns in Time Series, (3) using data augmentation (based on data provided) to explain the behavior of key variables.

LEARNING OJECTIVES:

Interpret Time Series data

Present information to the class concisely

Develop a solution that leverages Machine Learning, or Statistical methods to solve the problem or enhance the understanding of a specific problem in a reasonable amount of time

Share the results with the group

EVALUATION:

(1/4) Data exploration results and presentation
(1/4) Implementation of the proposed algorithm
(1/4) evaluation methods KPI's
(1/4) Final presentation to the group

TIMELINE PART I (Lead Instructor – Luis Arregoces)

Week 1: Discussion of data sources, Data visualization, select the approach to solve the problem from (3) choices
Week 2: Code the exploratory analysis in R or Python. (Time Series)
Week 3: Begin the coding for the selected modeling approach

Week 4: Review of proposed methodology (Confusion Matrix / KPI evaluation / goodness of fit etc)
Week 5: Selection of the MVP (minimum viable product)
Week 6: Final Presentation

TIMELINE PART II (Lead Instructor – Sid Bhattacharya)

Week 1: Discussion of data sources, Data visualization, select the approach to solve the problem from (3) choices, Code the exploratory analysis in R or Python. (Time Series)
Week 2: Begin the coding for the selected modeling approach
Week 3: Review of proposed methodology (Confusion Matrix / KPI evaluation / goodness of fit etc)
Week 4: Selection of the MVP (minimum viable product)
Week 5: Final Presentation

TIMELINE PART III (Lead Instructor – HyperGiant Brandon Ellet?)
Week 1:
Week 2:
Week 3:
Week 4:
Week 5:

PROGRAMMING LANGUAGE:

The main programing languages are R and primarily Python, the use of Java or C++ is allowed, as long as the functions or procedures are "wrapped" in a Python script and the execution of the steps occurs in only one package.

The use of editors such as RStudio for R, Spyder for Python or Notebooks like Jupyter is highly recommended.

1. https://www.rstudio.com/
2. https://www.anaconda.com/

TEXTBOOK:

There is no required textbook for the class, below are a few links to resources on the web

1. https://pandas.pydata.org/
2. https://scikit-learn.org/stable/
3. https://seaborn.pydata.org/
4. https://www.statsmodels.org/stable/index.html
5. https://keras.io/

CASE STUDIES:

Time permitting, we will cover a variety of case studies related to the energy industry. The plan is to use 20-30 minutes each class meeting to go over these case studies.

Current state of Analytics in Oil and Gas (23 Slides - Overview)

Data driven consulting for energy (36 Slides – Methodology)

Digital Twin Analytics (11 Slides – Solution)

Natural Language Processing (51 Slides – Overview / Case Study)

LNG Analytics (11 Slides – Case Study)

PowerGeneration for Oil and Gas (10 Slides – Case Study)

Data Architecture for Big Data Analytics (30 Slides – Overview)

University of Houston
Department of Mathematics
Master of Science in Statistics and Data Science (MSDS)

DATA SCIENCE PROJECT CASE STUDIES
MATH 6397 | CLASS NUMBER: 18396 | COURSE ID: 031328

2019

Mo	Tu	We	Th	Fr	Sa	Su	
December 31	January 1	January 2	January 3	January 4	January 5	January 6	
January 7	January 8	January 9	January 10	January 11	January 12	January 13	
January 14 First Day of Classes	January 15	January 16 Lecture 7-9pm	January 17	January 18	January 19	January 20	**Case Study I** Accenture - Oil and Gas
January 21	January 22	January 23 Lecture 7-9pm	January 24	January 25	January 26	January 27	
January 28	January 29	January 30 Lecture 7-9pm	January 31	February 1	February 2	February 3	
February 4	February 5	February 6 Lecture 7-9pm	February 7	February 8	February 9	February 10	
February 11	February 12	February 13 Lecture 7-9pm	February 14	February 15	February 16	February 17	
February 18	February 19	February 20 Lecture 7-9pm	February 21	February 22	February 23	February 24	
February 25	February 26	February 27 Lecture 7-9pm	February 28	March 1	March 2	March 3	**Case Study II** Accenture - Electricity
March 4	March 5	March 6 Lecture 7-9pm	March 7	March 8	March 9	March 10	
March 11	March 12	March 13	March 14	March 15	March 16	March 17	
<----------Spring Break ---------->							
March 18	March 19	March 20 Lecture 7-9pm	March 21	March 22	March 23	March 24	
March 25	March 26	March 27 Lecture 7-9pm	March 28	March 29	March 30	March 31	
April 1	April 2	April 3 Lecture 7-9pm	April 4	April 5	April 6	April 7	
April 8	April 9	April 10 Lecture 7-9pm	April 11	April 12	April 13	April 14	**Case Study III** Hyper Giant? Hopefully some type of marketing data. (Data besides oil/gas industry)
April 15	April 16	April 17 Lecture 7-9pm	April 18	April 19	April 20	April 21	
April 22	April 23	April 24 Lecture 7-9pm	April 25	April 26	April 27	April 28	
April 29	April 30	May 1 Lecture 7-9pm	May 2	May 3	May 4	May 5	
May 6	May 7	May 8 Last Day of Class Lecture 7-9pm?	May 9	May 10	May 11	May 12	
May 13	May 14	May 15	May 16	May 17	May 18	May 19	

Week 1	Discussion of data sources, Data visualization, select the project from (3) choices
Week 2	Code the exploratory analysis in R or Python. (Time Series)
Week 3	Begin the coding for the selected modeling approach
Week 4	Review of proposed methodology (Confusion Matrix / KPI evaluation / goodness of fit etc)
Week 5	Selection of the MVP (minimum viable product)
Week 6	Case Study I: Final Presentation
Week 7	Discussion of data sources, Data visualization, select the project from (3) choices
Week 8	SME Presentation on Refinery Operating standards
Week 9	**Spring Break**
Week 10	Coding for the SME recommeded solution
Week 11	Selction of MVP (minimum viable product)
Week 12	Case Study II: Final Presentation
Week 13	Begin Case III (HyperGiant)
Week 14	
Week 15	
Week 16	
Week 17	Case Study III: Final Presentation