**Introduction to data science**

**Course Prerequisites**

* Background in R
* Probability and Statistics

**Course Description**

Today, businesses, consumers, and societies leave behind massive amounts of data as a by-product of their activities. Leading-edge companies in every industry are using analytics to replace intuition and guesswork in their decision-making. As a result, managers are collecting and analyzing enormous data sets to discover new patterns and insights and running controlled experiments to test hypotheses.

This course prepares students to understand Data Science and become leaders in these areas in business organizations. This course teaches the scientific process of transforming data into insights for making better business decisions. It covers the methodologies, issues, and challenges related to analyzing business data. It will illustrate the processes of data science by allowing students to apply Data Science algorithms and methodologies to business problems. The use of examples places Data Science techniques in context and teaches students how to avoid the common pitfalls, emphasizing the importance of applying proper Data Science techniques.

**Learning Objectives**

After taking this course, students should be able to:

* approach business problems data-analytically. Students should be able to think carefully and systematically about whether and how data science can improve business performance.
* Develop Data Science ideas, analyze data using data science software, and generate business insights.

**Topics Covered**

* Linear Regression
* Regression Diagnostics
* Variable Selection
* Indicator Variables and Interaction Terms
* Nonlinear Transformation Models
* Logistic Regression
* Resampling Methods
* Treatment Effect
* Social Network Analytics
* Text Analytics
* Classification and Regression Trees
* Association Rules (Market Basket Analysis)
* Analytics in Business (Executive/Expert Interviews)

**Textbooks**

* Required: (ISLR) *Introduction to Statistical Learning*. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. ISBN-13: 978-1461471370, ISBN-10: 1461471370. Downloadable at<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf> or available for purchase at Amazon.
* Required: (Galit) Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., &Lichtendahl, K. C. (2018). *Data mining for business analytics concepts, techniques, and applications in R*. Hoboken, NJ, USA: Wiley. ISBN-13: 978-1118879368, ISBN-10: 1118879368. This book is available for purchase from Amazon.

**Software Requirements**

We will be learning Data Science with the help of open-source and free software applications that are provided for educational use. Please follow instructions provided in their respective websites and install the following software in your personal laptop:

1. R: <https://www.r-project.org/>
2. RStudio: <https://www.rstudio.com/>
3. Gephi: <https://gephi.org/>

There are plenty of resources on how to learn R:

* *R for Datascience*, <http://r4ds.had.co.nz/>
* <https://www.datacamp.com/courses/free-introduction-to-r>
* <https://www.rstudio.com/online-learning/>

**Hardware Requirements**

Note that tablets, Chromebooks, and old laptops don’t work well for this class.Make sure that you have admin rights on your laptop since occasionally you will need to install R, RStudio, packages in R,and other software like Gephi.

**Grading**

Grades will be assigned on the following basis:

Homework Assignment (4; worth 5% each) 20%

Midterm Exam – Part 1 25%

Midterm Exam – Part 2 15%

Final Exam, Part 1 40%

**Course Schedule**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** |  | **Topic** | **Readings** |  |
| 1 |  | Linear Regression | * Galit, Ch. 6
* ISLR, Sections 3.1, 3.2
 |  |
| 2 |  | Regression Diagnostics | * ISLR, Section 3.3.3
 |  |
| 3 |  | Variable Selection | * ISLR, Sections 6.1, 6.2
 |  |
| 4 |  | Indicator Variables and Interaction Terms | * ISLR, Section 3.3
 |  |
| 5 |  | Nonlinear Transformation Models | * Interpreting Nonlinear Models
 |  |
| 6 |  | Logistic Regression | * Galit, Ch. 10
* ISLR, Section 4.3
 |  |
| 7 |  | Resampling Methods | * ISLR, Ch. 5
 |  |
| 8 |  | Treatment Effect | * Program Evaluation and the Diff-in-Diff Estimator
 |  |
|  |  |  |  |  |
| 10 |  | Social Network Analytics | * Galit, Ch. 19
 |  |
| 11 |  | Text Analytics |  |  |
| 12 |  | Classification & Regression Trees | * Galit, Ch. 9
* ISLR, Ch. 8
 |  |
| 13 |  | Association Rules (Market Basket Analysis) |  |  |
|  |  |  |
| 14 |  | Data Scince in Business (Executive/Expert Interviews) |  |  |
|  |  |  |  |  |