Fundamentals of Data Science

I&C SCI X427.05
2.5 Units

Instructor Information

Name: Yu Zhang
Email: Use Canvas “Inbox”
Website: http://ce.uci.edu
Yu Zhang has over 5 years of experience in python programming and uses the big query SQL, python colab, and Tableau in previous and current work. She has over 5 years of experience in academia in statistical analysis and 3 years of experience in industry and government working on complex data and machine learning projects. She is technically sound with full experience in data validation and machine learning modeling. In addition, she has over 6 year’s teaching experience as a lecturer, teaching assistant, and mentor working in UC Santa Barbara, UC Davis, UC Irvine. She is passionate about using data analytics for real-world problem solving and working with diverse students.

Course Description

The goal of this course is to demystify data science and to familiarize students with key data scientist skills, techniques, and concepts. Starting with foundational concepts like analytics taxonomy, the Cross-Industry Standard Process for Data Mining, and data diagnostics, the course will then move on to compare data science with classical statistical techniques. An overview of the most common techniques used in data science, including data analysis, statistical modeling, data engineering, relational databases, SQL and NoSQL, manipulation of data at scale (big data), algorithms for data mining, data quality, remediation and consistency operations will be covered.

Prerequisites

- I&C SCI_X426.64 – Introduction to Python Programming
- I&C SCI X425.99 – Practical Math and Statistics

Course Sequencing

This course is the first required course in the Data Science Certificate Program and should be taken after the program prerequisites have been met.

Student Learning Outcomes

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

- Analyze a customer success story and explain how data science was used to solve a problem.
- Discuss the ethical and privacy considerations in problem solving with data science.
- Describe the data science process and identify risks to consider during a project.
- Develop data science models and explain how to evaluate a model's performance from both technical and business perspectives.
- Utilize technical and business techniques to deliver business insight, competitive intelligence, and consumer sentiment.

Course Material

There are no required materials for this course. Below is an **optional** but recommended textbook.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol: O'Reilly.

Course Outline

Module 1	**Key Topics**	● Data Science as a Field
		● Skills of Successful Data Scientists
		● Applications of Data Science Across the Lines of Business
		  ○ Marketing
		  ○ HR
		  ○ Finance
		  ○ Supply-Chain
		  ○ Healthcare
		● Data: Small and Big Data
		● Skills of Successful Data Scientists
		● Ethical Considerations

	Student Learning Outcomes	By the end of this module, you will be able to: ● Explain what types of business problems can be better understood through data science ● Discuss the ethical and privacy considerations in problem-solving with data science ● Analyze a customer success story and explain how data science was used to solve a problem ● Identify what types of questions or problems cannot be answered with data science ● Define what fields data science can apply to
	Assignments	● Assignment 1 ● Discussion 1

		• Quiz 1

Module 2	**Key Topics**	• CRISP-DM
		• Descriptive, Predictive and Prescriptive Analytics
		• Data Science Problem Formulations
		  ○ Predictive Maintenance
		  ○ Algorithmic Pricing
		  ○ Customer Segmentation
		  ○ Employee Retention
		• Data Science Toolkit
		  ○ Python and R
		  ○ KNIME, Orange, and Rapidminer
		  ○ SAS and IBM/SPSS
		  ○ DataRobot and Dataiku
		  ○ Alteryx

	Student Learning Outcomes	By the end of this module, you will be able to: • Discuss CRISP-DM and the benefits of following this standard data science process • Explain the skills needed to convert business problems into data science solutions • Describe the data science process and identify risks to consider during a project • Identify what type of analytics are most commonly used in business applications
	Assignments	• Assignment 2 • Discussion 2 • Quiz 2

Module 3	**Key Topics**	• Predictive Modeling
		• Supervised vs. Unsupervised Modeling
		• Data Dimensionality
		• Sampling
		• Classification Analysis
		  ○ Decision Trees
		  ○ Random Forest Model
		  ○ Visualizing Classification

	Student Learning Outcomes	By the end of this module, you will be able to: • Discuss the application of supervised classification and identify how it would apply to your work • Build and explain a decision tree classifier and explain the most significant variables • Recognize the applications of predictive analytics

		● Identify the difference between classification type and regression type prediction problems
	Assignments	● Assignment 3 ● Discussion 3 ● Quiz 3

Module 4	**Key Topics**	● Fitting and Overfitting Models
		● Model Generalization
		● Train vs. Test
		● Train and Score
		● Regression Analysis
		  ○ Regression Analysis in Python
		  ○ Visualizing Regression
		● Linear vs. Logistic Regression
		● Modeling to Address Business Objective

	Student Learning Outcomes	By the end of this module, you will be able to: ● Discuss the types of problems that would be better solved with logistic regression instead of decision trees ● Develop a linear regression model and provide the model fit measure ● Identify how a model can be generalized in order to apply to new data ● Recognize the qualities of a regression model
	Assignments	● Assignment 4 ● Discussion 4 ● Quiz 4

Module 5	**Key Topics**	● Cluster Analysis and Segmentation
		  ○ Cluster Analysis in Python
		  ○ Visualizing Clusters
		● Collaborative Filtering
		● Association Rules Mining and Market Basket Analysis

	Student Learning Outcomes	By the end of this module, you will be able to: ● Discuss the differences between supervised and unsupervised learning ● Develop a k-means multivariate clustering model for three clusters and analyze the accuracy of the algorithm ● Recognize how an algorithm determines the number of clusters into which data should be partitioned ● Identify how a market basket analysis functions
	Assignments	● Assignment 5 ● Discussion 5 ● Quiz 5

<table>
<tr><td rowspan="3">Module 6</td><td>Key Topics</td><td>

Classification Type Models

Evaluating Model performance

Confusion matrix
Unbalanced Classes
Costs and Benefits
Rare Event Detection
Model Performance Evaluation

Visualizing Model Performance

Model Performance Visualization in Python
Lift Curves
ROC Curves

Regressions Type Prediction Models

Correlation Versus Regression
Simple Versus Multiple Regression
Visualization

</td></tr>
<tr><td>Student Learning Outcomes</td><td>
By the end of this module, you will be able to:

Differentiate between classification and regression type models
Explain how to evaluate a model’s performance from both technical and business perspectives
Evaluate classification results with multiple performance metrics and build an ROC curve
Recognize how data scientists evaluate classification model performance using a confusion matrix
Identify a calculation for applying a cost-benefit analysis to model evaluation

</td></tr>
<tr><td>Assignments</td><td>

Assignment 6
Discussion 6
Quiz 6

</td></tr>
<tr><td>Module 7</td><td>Key Topics</td><td>

Introduction to Text Analytics
Sentiment Analysis
Text Summarization
Text Classification
Social Media Analytics
Topic Modeling
Natural Language Processing Functions

Corpus
Stemming
Lemmatization
Stop Words
Tokenization
N-grams
Part-of-Speech Tagging

</td></tr>
</table>

		○ Morphology ○ Term-by-Document Matrix (TDM) ○ Singular Value Decomposition (SVD)
	Student Learning Outcomes	By the end of this module, you will be able to: ● Discuss how to leverage social media analytics and natural language processing to improve customer engagement ● Develop a supervised classification model to predict consumer sentiment ● Differentiate between natural language processing and social media analytics ● Identify strategies to get the most accurate sentiment analysis results
	Assignments	● Assignment 7 ● Discussion 7 ● Quiz 7
Module 8	**Key Topics**	Past, Present and Future of Data Science
	Student Learning Outcomes	By the end of this module, you will be able to: ● Discuss the potential for new methods to impact the future of data science ● Develop a regression model numerical dependent variable (DV) and a classification model using the nominal dependent variable ● Calculate the accuracy measures and comment on the most important input variables ● Identify commonly used variable importance methods employed in machine learning and predictive modeling
	Assignments	● Final Assessment ● Discussion 8 ● Quiz 8

Evaluation and Grading

This course will be graded using the following weighted percentages for each of the assignments in the course. Feedback and grades are typically posted within one week of assignments due dates.

Assignments	**% of Grade**
Discussions	20%
Assignments	40%
Final Assessment	30%
Quizzes	10%

Total	**100%**

Grading Scale

This course uses the following grading scale.

Letter Grade	Percentage
A	93% - 100%
A-	90% - 92%
B+	87% - 89%
B	83% - 86%
B-	80% - 82%
C+	77% - 79%
C	73% - 76%
C-	70% - 72%
D+	67% - 69%
D	63% - 66%
D-	60% - 62%
F	59% or less

Technical Requirements

Below are the basic technical requirements for all UC Irvine, Division of Continuing Education courses.

Hardware

To participate in this course, you need a computer or device with reliable internet access. The device should be able to play videos on the screen and audio through headphones or speakers.

Software

For this course, you must have Microsoft Office, Google Docs, OpenOffice, or another compatible word processing software. If additional software is required, your instructor will provide detailed access information.

Skill Requirements

You may be required to use a web camera, copy and paste functions, attach documents, or use upload features in the LMS. In addition, you may need to upload and view documents in PDF format. You can download Adobe Acrobat Reader to open PDF files by going to this site: https://get.adobe.com/reader/. (Please note that Apple OS X includes the Preview application, which allows you to open PDF files.)

Communication Expectations

The majority of course communication takes place in general course forums. The written language has many advantages: more opportunity for reasoned thought, the ability to go in-depth, and more time to think through an issue before posting a comment. However, written communication also has certain disadvantages, such as a lack of the face-to-face signaling that occurs through body language, intonation, pausing, facial expressions, and gestures. As a result, please be aware of the possibility of miscommunication and compose your comments in a positive, respectful, and constructive manner.

UC Irvine Policies

Code of Conduct

All participants in the course are bound by the University of California Code of Conduct, found at https://ce.uci.edu/resources/conduct/.

Academic Honesty Policy

The University is an institution of learning, research, and scholarship predicated on the existence of an environment of honesty and integrity. As members of the academic community, faculty, students, and administrative officials share responsibility for maintaining this environment. It is essential that all members of the academic community subscribe to the ideal of academic honesty and integrity and accept individual responsibility for their work. Academic dishonesty is unacceptable and will not be tolerated at the University of California, Irvine. Cheating, forgery, dishonest conduct, plagiarism, and collusion in dishonest activities erode the University's educational, research, and social roles.

Students who knowingly or intentionally conduct or help another student engage in dishonest conduct, acts of cheating, or plagiarism will be subject to disciplinary action at the discretion of UCI Division of Continuing Education.

Disability Services

If you need support or assistance because of a disability, you may be eligible for accommodations or services through the Disability Service Center at UC Irvine. Please contact the DSC directly at (949) 824-7494 or TDD (949) 824-6272. You can also visit the DSC's website: http://www.disability.uci.edu/. The DSC will work with your instructor to make any necessary accommodations. Please note that it is your responsibility to initiate this process with the DSC.

Privacy

UC Irvine is fully committed to maintaining the integrity of your personal student information, including academic records, in accordance with the Federal Family Education Rights and Privacy Act of 1974 (FERPA). For complete

information on privacy and FERPA, please visit the UCI Privacy and Student Records website: http://www.reg.uci.edu/privacy/.

Accessibility

The University of California and UC Irvine are committed to improving accessibility for all students. For complete information about UC Irvine's policy affecting all information technology systems and software, please visit: https://www.oit.uci.edu/accessibility/accessibility-policies/.

For accessibility information pertaining to the software used in UCI courses, please see below:

Software	Link
Canvas (LMS)	https://www.canvaslms.com/accessibility
Zoom	https://zoom.us/docs/doc/vpat/Zoom%20Product%20Web%20Pages%20VPAT.pdf
VidGrid	https://www.vidgrid.com/accessibility/

UCI-DCE Student Services

The UCI Division of Continuing Education Student Services office is responsible for maintaining enrollment, academic, and financial records while providing academic support services to DCE students, instructors, and staff. Services include:

- Student records and DCE My Account portal access
- Adding, dropping, and waiting lists for all DCE courses and programs
- Cashiering, refunds, and accounts receivable
- Grades, transcripts, candidacy, and certificates
- International F-1 student advising, student academic success, and wellness support
- Important notifications regarding DCE course-related changes such as schedule, location, instructor, and course materials

Descriptions of many of these services can be found at https://ce.uci.edu/resources/. Additionally, UCI-DCE Student Services is available by email (dce-services@uci.edu) or by calling (949) 824-5414 (option 1), Monday-Friday, 8:30am-4:30pm.