Course Syllabus

Course Name	Data Mining for Business Analytics
Course Code	ISOM 3360, 3 credits
Course Schedule and Classroom	Lecture (LSK1011): Tu and Thu, from 10:30 am to 11:50 am Lab (LSKG005): Mo, from 9:30 am to 10:20 am.
Instructor	Carlos Fernández-Loría, ISOM Office: LSK5054A Email: imcarlos@ust.hk (please try to use 'Discussions' in Canvas to ask about course material)
Teaching Assistant	Sophie Gu Office: LSK4065 Email: imsophie@ust.hk (please try to use 'Discussions' Canvas to ask about course material)
Office Hours	By appointment (with Sophie or Carlos)
First - Last Class	Sep 2 – Nov 30

Overview

This course will change the way you think about data and its role in business. Firms, governments, and individuals create massive collections of data as part of their everyday activity. Increasingly, businesses seek to exploit such data to improve decision making. In this course, you will study the fundamental concepts of data science, learn to use predictive models to make better decisions, and discuss business cases to illustrate the data mining process and develop a data-analytic thinking towards decision making.

Learning Objectives

This course's goal is for you to learn how to approach business problems data-analytically. You will learn how to:

1. Use data-driven models to improve decision-making in several business settings.
2. Analyze opportunities and limitations of data science projects.
3. Transform business problems into data science tasks.
4. Build data-driven models to solve predictive problems.
5. Competently discuss topics related to data science and machine learning.

What to Expect

The course content will be taught through lectures, lab sessions, class discussions, readings, case studies, tutorials, and take-home assignments. The emphasis will be on understanding and applying fundamental concepts and techniques of data science to predictive business problems. We will discuss the technical details of data science algorithms to better understand some concepts, but the emphasis of the course will not be on explaining algorithms. However, many of the most popular algorithms and techniques in data science are direct applications of the concepts we will study in class.

Be Prepared

Before each class, you are expected to read the assigned readings (which will cover a large part of the course content) and prepare the assigned case studies (which will be discussed in class). This is an intensive and hands-on course with a heavy workload. **Expect to invest at least 10 weekly hours outside of class.** So, please consider this before registering. However, if you are interested in learning about data science and getting hands-on experience, then you will definitely get a lot out of this course. I promise.

Assignments also play an important role in this course because the only way you will truly learn how to do data science is by doing data science. Throughout the course, we will work on several hands-on assignments and case studies using Python—one of the most popular programming languages for doing data science. Most of the Python-related content will be covered in the lab sessions, so please make sure you attend them: they will be **critical** for you to learn the necessary skills to complete the take-home assignments, all of which will involve coding. Prior experience with Python or coding is desirable but not mandatory.

Finally, please try to use the discussion board in Canvas to ask questions related to coding, homework, and class material. This is to track better your questions and allow other students to benefit from the answers. **Answering your classmates' questions before Sophie or me will count as part of the participation grade**. You will not be penalized for being "wrong" when trying to contribute to the discussion board (or the class discussion).

Course Materials

The course book is "*Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.*" I will provide the first two chapters of the book (so that you don't have to buy a copy in case you decide to drop this course). **However, you must get your own copy before the beginning of the third week of class.**

Python and Jupyter Notebook are the official data mining tools of the course. Please download Anaconda (a data science toolkit that includes both Python and Jupyter notebooks) before the first lab session (https://www.anaconda.com/products/individual (Links to an external site.)).

Grading

The grade breakdown is as follows:

1. Short quizzes (15%)
2. Class participation (15%)
3. Group assignments (40%, 10% each assignment)
4. Midterm Exam (10%)
5. Final Exam (20%)

Short quizzes: There will often be a quiz at the beginning of lectures. Their purpose is to make sure you come prepared to class. I will ignore the two quizzes with lowest grades.

Class participation: Public speaking, and being able to voice your opinion, is a skill that is valuable and often necessary in many professional contexts. So, I will cold call students at random to encourage discussion and diversity in class participation. It is OK if you give the "wrong" answer; all contributions are part of the discussion. However, please come prepared to class. Answers such as "I don't know", or "I didn't read" are unacceptable. Answering questions in the discussion board also counts as class participation. A few other things:

- Attending lectures is **mandatory** and will be considered part of class participation.
- Attending lab sessions is not mandatory, but doing so is critical for you to solve the assignments.
- Students who can't physically attend lectures because they can't enter Hong Kong are required to **turn on their cameras**. I will also consider this part of class participation.

Group assignments: Assignments should be submitted in groups of 3 to 4 members. These assignments are meant for you to apply the concepts learned in class, and they are due before the beginning of class sessions. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After a week, late assignments will receive no credit. As part of the assignments, you will be asked to evaluate the other members of your group.

Exams: Each exam will evaluate the material covered up to that point (so the final exam covers all the course material). You must bring your laptop to the exams. The exams are open book and open everything, but you are not allowed to communicate during exams. The final exam will also have a coding component.

Course Outline

Week	Date and session type	Topics (subject to change as class progresses)	Book Readings	Additional Readings & Deliverables
1	Sep 2 Lecture	Introduction to data mining	Ch. 1	- Read syllabus - TelCo case

2	Sep 6 Lab	Introduction to Python & Jupyter NB Introduction to Jupyter notebooks		
	Sep 7 Lecture	The CRISP Data Mining Process	Ch. 2	- TelCo case
	Sep 9 Lecture	Formulation of Predictive Problems		- Target reading
3	Sep 13 Lab	Data processing & visualization (pandas)		
	Sep 14 Lecture	Learning from data (Part 1) Supervised segmentation	Ch. 3	- Dolva case - Deadline for group formation in Canvas - Assignment #1 starts
	Sep 16 Lecture	Learning from data (Part 2) Learning as an optimization problem	Ch. 4	- Pedal Bikeshare case
4	Sep 20 Lab	Model building (decision tree)		
	Sep 21 Lecture	Complexity control and overfitting	Ch. 5	- Pedal Bikeshare case
	Sep 23 Lecture	Regularization and cross-validation	Ch. 5	- Elysium case
5	Sep 27 Lab	Complexity control (CV & Grid search)		
	Sep 28 Lecture	Model evaluation (Part 1) Expected value framework	Ch. 7	- Red cross case

				- Assignment #1 is due - Assignment #2 starts
	Sep 30 Lecture	Model evaluation (Part 2) Alternative evaluation measures	Ch. 8	- Miskatonic case
6	Oct 4 Lab	Model Evaluation (AUC, lift curves)		
	Oct 5 Lecture	Midterm review		
	Oct 7 Lecture	**[no class] Midterm exam** **(Bring your laptops!)**		
7	Oct 11 Lab	End-to-end example		
	Oct 12 Lecture	Mining fine-grained data	Ch. 9	- CartOnline case
	Oct 14 Lecture	**[no class] Public holiday**		
8	Oct 18 Lab	High-dimensional sparse data		
	Oct 19 Lecture	Text mining	Ch. 10	- Trans-American Airlines case **- Assignment #2 is due** - Assignment #3 starts
	Oct 21 Lecture	Invited speaker:		

		Ben Carterette from Spotify (attendance will count as a quiz)		
9	Oct 25 Lab	Text mining workshop		
	Oct 26 Lecture	Unsupervised Learning & Clustering	Ch. 6	- Dureas case
	Oct 28 Lecture	Expected Value Framework (revisited)	Ch. 11	- Telco case (Part 2)
10	Nov 1 Lab	Clustering workshop		
	Nov 2 Lecture	Causal Inference (Part 1) Simpson's Paradox & Experiments		- Causal Inference - McKeen Sea case **- Assignment #3 is due** - Assignment #4 starts
	Nov 4 Lecture	Causal Inference (Part 2) Methods for Observational Data		- Causal Inference - Saint Gerard Bank case
11	Nov 8 Lab	Causal inference workshop		
	Nov 9 Lecture	Popular data science algorithms (Part 1)	Ch. 12	
	Nov 11 Lecture	Popular data science algorithms (Part 2)		
12	Nov 15 Lab	Advanced machine learning algorithms		

		(with an application to image recognition)		
	Nov 16 Lecture	Invited speaker: Foster Provost from NYU & Compass (attendance will count as a quiz)		
	Nov 18 Lecture	eXplainable Artificial Intelligence (XAI)		
13	Nov 22 Lab	Methods to interpret black-box models		
	Nov 23 Lecture	Algorithmic Fairness		- CallFast case **- Assignment #4 is due**
	Nov 25 Lecture	Data science as a competitive advantage	Ch. 13	- Stitch fix reading
14	Nov 29 Lab	To-be-defined		
	Nov 30 Lecture	Wrap-up	Ch. 14	