NYU | STERN

# DATA SCIENCE FOR BUSINESS (TECHNICAL)

TECH-GB.2336
Thursdays 6pm

**TENTATIVE SYLLABUS – Spring 2023**

| | |
|---|---|
| Instructor | **Foster Provost,** Information, Operations & Management Sciences Department |
| Office; Hours | FP: KMC 8-86; Office hours: **TBD** |
| Email | fprovost@stern.nyu.edu<br>**Begin subject: [DS Class] …** **← note!** |
| Classroom | TBD |
| Class time | Thurs 6pm-9pm |
| First : Last Class | Feb 9; May 4 |
| Final Quiz | Timing announced in class |
| Course Assistant | TBD |

**NOTE 1: It is required that you have programmed in Python prior to taking this class. That does not mean that you have to be a Python expert or an expert programmer. An introduction to Python class and some experience with programming should suffice. But we will not be teaching either Python or programming in this class, and you likely will not have time to learn on the fly. You will have the time to learn the data analysis and machine learning libraries that you will use through the class (mainly Pandas and sklearn).**

## 1. Course Overview

This course is about the "data science" way of thinking about data and its role in business, and in particular about the fundamentals of machine learning.

Businesses, governments, and individuals create massive collections of data as a by-product of their activity. Increasingly, decision-makers and systems rely on intelligent technology to analyze data systematically in order to improve decision-making. In many cases automating analytical and decision-making processes is necessary because of the volume of data and the speed with which new data are generated.

We will examine how data analytics and machine learning can be used to improve decision-making. We will study the fundamental principles and techniques of data science, machine learning, and data-driven artificial intelligence, and we will examine real-world examples and cases to place data science techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much art as science.

After taking this course you should:

1. *Approach business problems data-analytically.* Think carefully & systematically about whether & how data can improve business outcomes, to make better-informed decisions for product management, marketing, sales, operations, consulting, etc.

2. *Be able to interact competently about data science, machine learning, and data-driven artificial intelligence.* Know the fundamental principles of data science that underlie processes, algorithms & systems. Improve your ability to envision new opportunities, work on data science projects & interact with everyone involved.

3. *Have had hands-on experience mining data.* Be prepared to follow up on ideas or opportunities that present themselves, e.g., by performing pilot studies.

## 2. Focus and interaction

The course will explain the fundamental principles, uses, and some technical details of data science and machine learning. We do this through lectures, discussions, in-class case studies, hands-on work in Python, and real-world examples The emphasis primarily is on understanding and applying the fundamental concepts of data science and business applications of data mining and machine learning. We will discuss the mechanics of how the methods work to a limited degree—only as is necessary to understand and illustrate the fundamental concepts and business applications. This is not an algorithms course. However, many machine learning algorithms and techniques are the embodiment of one or more of the fundamental principles of data science.

We will expect you to be prepared for class discussions by having satisfied yourself that you understand what we have done in the prior classes. The assigned readings will cover the fundamental material. The class meetings will be a combination of lectures/discussions on the fundamental material, discussions of business applications of the ideas and techniques, case discussions, student exercises, and demos.

You are expected to attend every class session, to arrive at or before the starting time, to remain for the entire class, and to follow basic classroom etiquette. In general, we will follow Stern default policies unless we state otherwise. We will assume that you have read them and agree to abide by them:

**http://www.stern.nyu.edu/portal-partners/academic-affairs-advising/policies-procedures/default-policies-stern-courses**

The Brightspace site for this course will contain reading materials, assignments, and late-breaking news. You should check the sites daily, and we will assume that you keep up to date on all announcements and class discussion.

If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please ask after class, come to office hours to see one of us (Foster, Pearl), or ask in the discussion forum. **All questions related to homework, class material, and technical details must be posted in the discussion forum**. This is to track better your questions and allow other students to benefit from the answers. Also, please try to answer your classmates' questions. In grading your class contribution, we will include your contributions to the discussion forum. You will not be penalized for being wrong in trying to contribute on the discussion forum (or in class).

Worth repetition: It is your responsibility to check NYU Classes and your email at least once a day during the week (M-F), and you will be expected to be aware of any announcements within 24 hours of the time the message was sent.

Use email for personal issues or questions specific to your term project. **We will check for class-related email at least once a day.** ***Your email should include the special tag [DS Class] in the email subject header***. We use this tag to make sure to process class-related email first. If you do not include the special tag, we may not read the email

for a while (maybe a long while). If you forget and send without the tag and then remember later, ***just send it again*** including the tag.

## 3. Readings and Lecture Notes

**Book:** The textbook for the class will be the following.

*Data Science for Business: What you need to know about data mining and data analytic thinking. Provost & Fawcett (O'Reilly, 2013) – Foster's new update (as of 2020).*

**This updated version will be available in the bookstore as a coursepack**. As I understand it, you will get a digital version bundled for free.

This book covers the fundamental material that will provide the basis for you to think and communicate about data science and business analytics. We will complement the book with discussions of applications, cases, and demonstrations, and possibly some additional readings or notes for material that is not covered in the book.

You are expected to ask questions about any material that is unclear after our class discussion and after reading the assigned book chapters. Requiring readings from the book frees up class time for more discussion of applications, cases, etc.—so many of your questions may be answered in the book. If not, please let us know! If the class discussion and the book are not adequate to explain a topic we skip, you should ask about it in the discussion forum. We will be happy to follow up.

We may hand out or post some additional required readings as we go along. *Note that some of these readings may be accessible for free only from an NYU computer. If you can't access a link from home, please try it from school.*

For those interested in going further, come and talk to me or Pearl about supplemental books and other material, which give alternative perspectives on and additional details about the topics we cover. These are completely optional; you will not be required to know anything in these supplemental readings that are not also in the primary/required materials or lectures. There are many books that we can recommend, for example if you want a reference to a more mathematical treatment of the topics or a more hands-on guide. Please don't hesitate to come and talk to us about what supplemental material might be best for you, if you want to go further.

For example, one particularly useful book for those interested in going deeper on the "hands-on" component of the class: (OPTIONAL)

*Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd Edition
by Sebastian Raschka & Vahid Mirjalili
https://www.amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow/dp/1789955750/

## 4. Requirements and Grading

The grade breakdown is as follows:

1. Homeworks: 30%
2. Term Project: 30%
3. Class Contribution/Professionalism/Attendance: 10%
4. Final Quiz: 30%

At NYU Stern we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. In my experience, students generally become engaged with this course and do excellent or very good work, receiving As and Bs, and only one or two perform only adequately or below and receive C's or lower. Note that the actual distribution for this course and your own grade will depend upon how well each of you actually perform this particular semester.

**Homework Assignments**

The homework assignments are listed (by due date) in the class schedule below. Each homework comprises questions to be answered and/or hands-on tasks. Except as explicitly noted otherwise (see next paragraph), you are expected to complete your assignments on your own—without interacting with others on the completion of your assignment. You are free of course to discuss the concepts with your classmates, and to discuss similar problems to the ones in the homeworks.

For the hands-on parts of the assignments, if they are challenging for you I encourage you to work with your group members and other classmates to understand how to get Python to do what you need to do generally (not on the specific homework problems), and then to complete your assignment on your own. So, for example, you could have a classmate help you do something similar, such that then you would be able to complete the assignment.

With the support of me, Pearl, and your classmates, we operate under a "**diligent attempt but limited frustration**" policy: (1) If you get stuck on something, spend some time Googling to try to find the answer. If you seem to be moving forward, keep going. That search and discovery will pay off, both in terms of the direct learning about how to do what you need to do, and also in terms of your learning *how to find* such things out. BUT, (2) limit frustration—start your assignments early enough that if you run into a wall, you can just stop searching and ask about it. Let's say, if you feel like you have not moved forward after 15 minutes of being stuck, just stop and ask your classmates or in the discussion forum. If you don't get a resolution, escalate it to Pearl or me.

Completed assignments must be handed in as specified in class, by the due dates and times we provide. Assignments will be graded and returned promptly. Answers to homework questions should be well thought out and communicated precisely and professionally, avoiding sloppy language, poor diagrams, and irrelevant discussion.

The hands-on tasks in the homeworks will be based on data that we will provide. You will mine the data to get hands-on experience in formulating problems and using the various techniques discussed in class.

**IMPORTANT**: ***You must have access to a computer on which you can access the class Colab notebooks. If you do not have such a computer, please see me immediately so we can make alternative arrangements.*** *You should bring your computer to class.* If you need additional help with Colab or Python, please see Pearl.

Generally Pearl (our course assistant and co-instructor; see first page) should be the first point of contact for questions about and issues with the homeworks, hands-on work, and Python/Pandas/sklearn. Pearl will have the responsibility to make sure that all questions are answered in a timely fashion.

Foster will be the main contact for the term project deliverables (but Pearl may be able to provide better help than Foster for advice on coding-related questions).

**Late Assignments**
Assignments are to be submitted by the due date. 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 one week, late assignments will receive no credit. Please turn in assignments early if there is any uncertainty about your ability to turn it in on time.

**Term Project**
Term project deliverables, including the final report, will be prepared by student teams. We will give you the instructions on how to form your teams. Teams are encouraged to interact with Foster and Pearl electronically or face-to-face in developing their projects. You will submit various deliverables through the course. We will discuss the project requirements in class.

**Final Quiz**
The final quiz will be completed during the days following the final class. The subject matter covered and the exact dates will be discussed in class.

**Participation/Professionalism/Contribution/Attendance**
Please see Section 2.

**Regrading**
If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to me describing the error, within one week after the date on which that assignment was returned. Include documentation (e.g., pages in the book, a copy of class notes, etc.). I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go "in your favor" and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

***FOR STUDENTS WITH DISABILITIES*:** If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

***Please read the policies for Stern courses***

**http://www.stern.nyu.edu/portal-partners/academic-affairs-advising/policies-procedures/default-policies-stern-courses**

***Please keep in mind the Stern Code of Conduct***

http://www.stern.nyu.edu/sites/default/files/assets/documents/con_039512.pdf

# *Class Schedule (Example – to be revised)*

| Week | Date | Topics<br>*(subject to change as class progresses)* | Readings | Deliverables |
|---|---|---|---|---|
| 1 | | Introduction to the Course<br>Introduction to Data Analytics<br>Doing Data Science<br>Tools for Business Analytics – Conceptual and Practical<br><br>**Notebooks:**<br>**Dealing with Data w/ Colab, Python & the Command Line** | Ch. 1-2 | **Pre-class Assignment due by 11:59pm Sep 19** |
| 2 | | Intro to Predictive Modeling<br>Supervised Segmentation<br>Conceptual Tools for Creating a Predictive Analytics Solution to a Business Challenge<br><br>**Notebook:**<br>**Predictive Modeling/Supervised Segmentation** | Ch. 3-4 | **HW#1 due by 11:59pm Oct 1** |
| 3 | | Intro to Predictive Modeling, cont.<br><br>Problem formulation, cont. + Fitting the data. More Modeling<br><br>**Notebook: Fitting mathematical models - Take 1** | Ch. 5 | **Team choices and initial project ideas due by 11:59pm Oct 8** |
| 4 | | Model performance analytics & the science of predictive modeling<br>Overfitting the data. Holdout testing, cross-validation +learning curves, domain knowledge validation<br><br>**Notebook: Fitting mathematical models - Take 2** | Ch. 7 | **HW#2 due by 11:59pm Oct 15** |
| 5 | | Model performance analytics II<br>Ranking, true positives, false positives, profit, lift<br>ROC analysis, expected value framework<br><br>**Notebook**: **Model performance analytics** | Ch. 8<br><br>Optional:<br>Stein paper | **Project Proposal due by 11:59pm Oct 22** |
| 6 | | Mining fine-grained data, Prediction via evidence combination, Bayesian reasoning, text classification, "Naïve" Bayes<br><br>**Case Study: Modeling consumer behavior for targeted marketing**<br>**(banking and/or online advertising)** | Ch. 9 & 10 | **HW#3 due by 11:59pm Oct 29** |

| Module | Date | Topics | Readings | Deliverables |
|---|---|---|---|---|
| 7 | | Similarity, Distance, Nearest Neighbors, Clustering | Ch. 6 | **Project update due by 11:59pm Nov 5** |
| 8 | | Toward Analytical Engineering | Ch. 11 | **HW#4 due by TBD** |
| 9 | | ML Strategy – Choosing which ML projects to invest in | Will be posted on Brightspace | |
| 10 | | TBD (Maybe Data Analytics for Causal Inference) | Will be posted on Brightspace | |
| 11 | | **Data Analytics, Management, Strategy & Tactics** | Ch. 12 & 13 | |
| 12 | | Catch Up/Wrap Up | | **Project report Due by 11:59pm TBD** |
| **<u>Final Quiz</u>**:<br>Open book<br>Timing announced in class | | | | |