DBA3803/DSC3216: PREDICTIVE ANALYTICS IN BUSINESS

NUS Business School   Department of Analytics & Operations (DAO)

23 May, 2022

Administrative Information

- Instructor: Dr. Long ZHAO
  - Office: BIZ1 8-61
  - Office Hours: By Appointment
  - Email: longzhao@nus.edu.sg
- Prerequisite: DAO2702
  - I will review linear algebra (matrix multiplication) and normal distribution.
- Evaluation:
  - Individual Assignments: 20%
  - Group projects: 50%
    - Default: two projects with detailed instructions. 25% + 25%
      - I want to reduce the workload of this course.
    - You could do an additional project.
      - Choose the topic as you like.
      - If you decide to do three projects, they will be 12.5% + 12.5% + 25% because the additional one requires both report and e-presentation.
  - Two Quizzes: 12% + 18%
- Coding: R or Python
  - No prerequisite on the skill level. Individual assignments will be challenging but manageable for a rookie coder.
  - Python has PyCaret package which is extremely easy to implement (non-deep -learning) machine learning methods. Thus, I recommend Python if you have no preference.
  - However, because I only use R, if you choose Python, I might not be able to help you effectively.
  - The coding examples for projects are written in R.
  - Python does not handle categorical variables well. This means that you need to write more codes to preprocess data.

Course Outline and Schedule

This course aims to develop an understanding of forecasting methods from data science for analyzing complex issues and solving business problems. We will make productive use of analytics tools available in R and Python. Because these packages are mature and convenient, we will instead focus on the thinking behind the methodology, which also applies to more advanced tools. Moreover, we will also learn the limitations of forecasting methods and common illusions in predictive analytics. Although the class focuses on simplified models, it aims to bridge the classroom knowledge and business applications, such as portfolio construction.

Week	Date	Topic	Remark
1	Aug.	Introduction to Data Science	-
2	Aug.	Loss Function + Linear Regression 1	Assignment 1 Due
3	Aug.	Linear Regression 2 & Bias-Variance Tradeoff	Assignment 2 Due
4	Sep.	Regularization: Lasso, Ridge, PCR	Assignment 3 Due
5	Sep.	Project 1	Assignment 4 Due
6	Sep.	Trees & Quiz 1 Review	-
-	Sep.	Recess Week	Project 1 Due
7	Sep.	Enhancement of Trees & Classification Trees	In-class Quiz 1
8	Oct.	Gradient Boosting & AdaBoost	Assignment 5 Due
9	Oct.	Logistic Regression & Project 2	Assignment 6 Due
10	Oct.	Summary & Overfitting in Validation	Project 2 Due
11	Oct.	Smart Choice of Objective & Quiz 2 Review	Assignment 7 Due
12	Nov.	Unsupervised Learning & SVM	In-class Quiz 2
13	Nov.	No Class	
14	Nov.	3rd Project e-Presentation	3rd Project Due

Resources

- Datacamp.com may be the best way to start learning data science coding.
  - Here is the invitation link. **Python** is the **default** choice.
  - If you want to join **R** track, please put your name here. I will manually assign you. Sorry for the inconvenience.
  - You should be able to use DataCamp for free with your **NUS email** (domain: @u.nus.edu).
  - If you cannot join it, please let me know ASAP. All our individual assignments are from DataCamp.

Optional Books & Videos:

- [ISLA] An Introduction to Statistical Learning with Applications in R
  - Here are the videos, Youtube: StatsLearning.
- [DSB] Data Science for Business
- [LFD] Learning From Data
  - Here are the videos, Youtube: Caltech's Machine Learning Course.
  - The videos are fantastic, but they are too high-level for our course.

Individual Assignments

All of the individual assignments come from DataCamp. The majority of the time, there will be two parallel tracks for R and Python. That is to say, if you use Python, you could ignore the corresponding R courses. Meanwhile, when it is only available in one language, you have no choice but to finish it.

Each assignment contains **1-2** mandatory mini-courses which take a rookie about 3-6 hours while a guru 1-3 hours. There are also **1-2** optional mini-courses for the rookies to catch up in coding. The grade of individual assignments (in this course) is defined as

$$\text{grade} = 70\% \times \text{completion} + 30\% \times \text{XP Percentage}.$$

Here the XP percentage is calculated as your **new** XP divided by the XP target which is **95%** of the total XP. This 5% gap buffers for evaluation mistakes, DataCamp gliches, and too technical problems.

For example, if you complete all assignments and achieve 80% as the XP percentage, then your grade will be $70\% \times 100\% + 30\% \times 80\% = 94\%$.

For students who have completed some DataCamp courses before, send me an email and I will waive the corresponding assignment for you.

Two (or Three) Group Projects

Because you will mainly work with strangers in a company, to enhance your abilities in such cooperation, **I will 'dictate' the groups for the first two projects**. The size of a group is 3 or 4 (preferred). *Each group will use the same programming language.* I have an algorithm to generate groups such that one will work with different people across projects. I will first introduce each project during class and then provide a brief discussion. I expect all of them are challenging for the following reasons.

- One's default score is 70%. Playing safe will not be enough.
- 6-page limit (with reasonable font). One will face penalty if his/her report goes beyond 6 pages (excluding cover). Thus, one needs to think about how to deliver the message concisely.
- All have unique characteristics that make it special. That is to say, you need to have a customized approach.
- It is always hard to write code from scratch.

Presentation(Pre-recorded Video): the 3rd project (optional) requires both report and e-presentation. Each group will have 15-min e-presentation. Please video record the presentation and upload to LumiNUS. Explain the problem, managerial decisions you are making, data source, approaches, evaluation and insights. The content of the presentation should highlight the key findings.

Two Quizzes

Quizzes will focus on conceptual arguments. They will be available on LumiNUS during the class time.

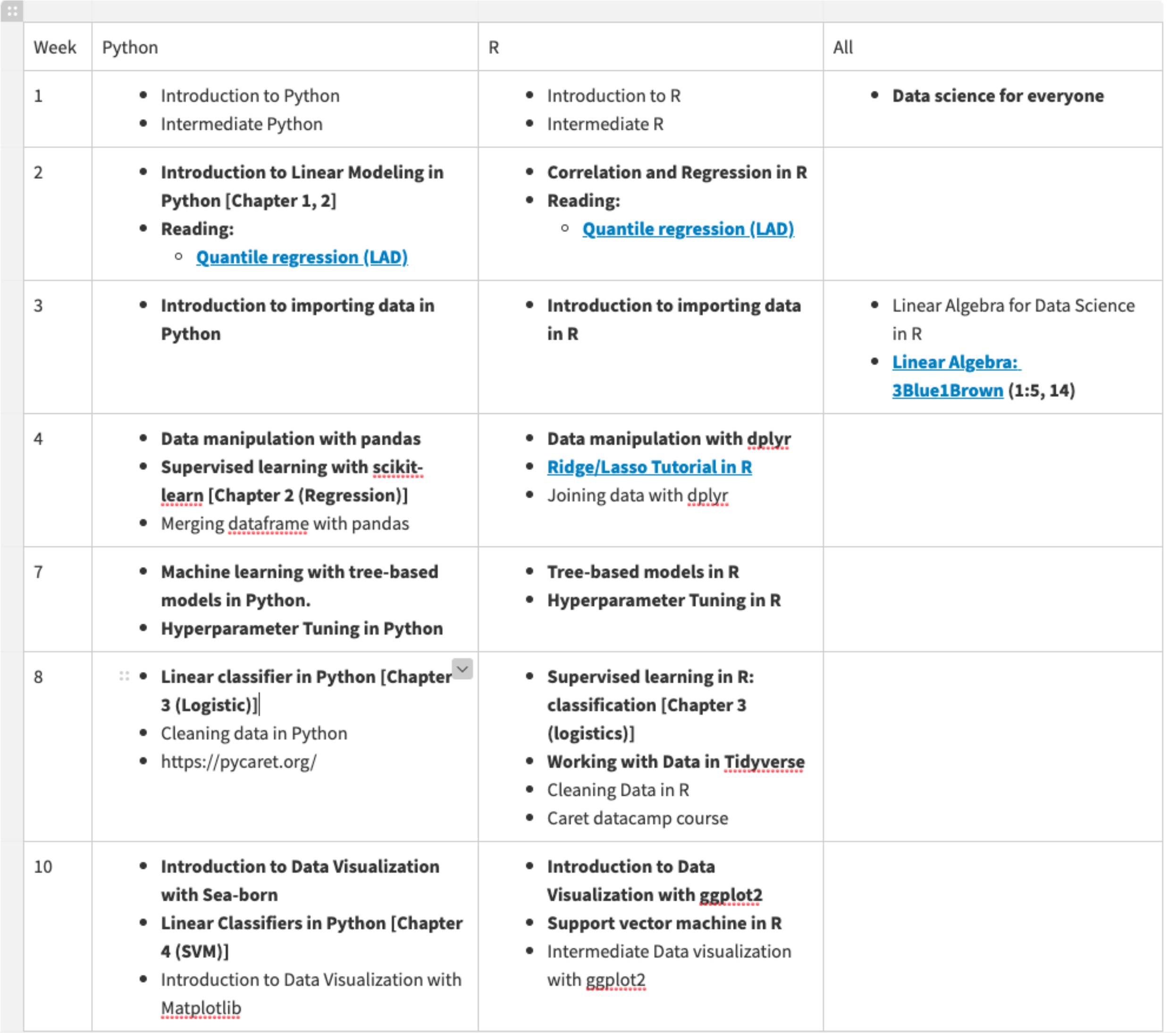
FAQ

1. I have completed some assigned DataCamp courses before but I am not satisfied with my XPs. What should I do?
   - You could obtain a waiver from Long for the corresponding course. Namely, your existing low XPs do not matter to your grade.
2. If I choose Python track, could I also access R courses?
   - You could access any DataCamp course and project for free and learn XP by doing it. However, your completion score is only determined by the assigned ones.
3. My DataCamp code is identical to the solution, but DataCamp considers it wrong. What should I do?
   - It also happens to Long. Thus, the XP target is 95% of the total XP to compensate the existence of such thing.
4. I am also interested in other DataCamp courses, could I use hints freely for other courses?
   - Yes, you could use hints without hurting your grades of individual assignments. However, there is a risk that the courses you do will be assigned as homework in the future. Thus, try not use hints too aggressively.

5. Is my DataCamp permanent?
   - Yes, it is permanent but the premium only lasts 6 months. The starting date is Aug. 2nd, 2021.

DataCamp Tentative Plan

- You could search the bolded courses on the next page in DataCamp to complete the individual assignment in advance.You might want to focus on data manipulation and visualization because I want to introduce the machine learning methods myself. My way is more consistent across the whole course.



Week	Python	R	All
1	• Introduction to Python • Intermediate Python	• Introduction to R • Intermediate R	• **Data science for everyone**
2	• **Introduction to Linear Modeling in Python [Chapter 1, 2]** • **Reading:** ◦ **Quantile regression (LAD)**	• **Correlation and Regression in R** • **Reading:** ◦ **Quantile regression (LAD)**	
3	• **Introduction to importing data in Python**	• **Introduction to importing data in R**	• Linear Algebra for Data Science in R • **Linear Algebra: 3Blue1Brown (1:5, 14)**
4	• **Data manipulation with pandas** • **Supervised learning with scikit-learn [Chapter 2 (Regression)]** • Merging dataframe with pandas	• **Data manipulation with dplyr** • **Ridge/Lasso Tutorial in R** • Joining data with dplyr	
7	• **Machine learning with tree-based models in Python.** • **Hyperparameter Tuning in Python**	• **Tree-based models in R** • **Hyperparameter Tuning in R**	
8	• **Linear classifier in Python [Chapter 3 (Logistic)]** • Cleaning data in Python • https://pycaret.org/	• **Supervised learning in R: classification [Chapter 3 (logistics)]** • **Working with Data in Tidyverse** • Cleaning Data in R • Caret datacamp course	
10	• **Introduction to Data Visualization with Sea-born** • **Linear Classifiers in Python [Chapter 4 (SVM)]** • Introduction to Data Visualization with Matplotlib	• **Introduction to Data Visualization with ggplot2** • **Support vector machine in R** • Intermediate Data visualization with ggplot2	

Figure 1: Tentative Plan

- Here are the links.
  - Solving optimization problems from Khan Academy
  - Python Quantile regression (LAD)
  - R Quantile regression
  - Linear Algrebra: 3Blue1Brown (1:5, 14)
  - Ridge/Lasso Tutorial in R