HOUSE OF INNOVATION

7313

DATA SCIENCE ANALYTICS

M.Sc. course 21/22

EMELIE FRÖBERG, Assistant Professor
Stockholm School of Economics
Department of Entrepreneurship, Innovation and Technology
PhD Business Administration 2010-2016

Date: Oct, 2021

HOUSE OF INNOVATION

Course name

Data Science Analytics

Course Director

Emelie Fröberg

SSE Factulty

Emelie Fröberg

Panel Participants

- Patrik Tran (Validio, Stockholm AI, Talk About AI-podcast)
- Salla Franzén (IKEA, previously SEB)
- Thomas Falk (Sveriges Riksbank, previously Blocket)
- Errol Koolmeister (TBC) (The AI Framework, previously H&M)
- Rebecka Cedering Ångström (TBC) (Ericsson)
- Magnus Petersson (TBC) (Spotify)

General Course Description

Intended Learning Objectives

To demonstrate an ability to choose, apply and evaluate statistical and machine learning methods for predictive modeling that harness value using data in a business domain.

Format

This course is hands-on and has a "learning by doing" format, where self-learning through trial and error is encouraged. Participants should be comfortable using R. The main format is lab sessions, but a mixture of pre-recordings, live panel discussions with practitioners and seminars are also used for variation.

Labs to facilitate learning

The course builds on “learning-by-doing” and as a reflection of that, there are many labs in the course. For every lab session, there is a designated lab available. The labs are designed to help you complete the hand-ins and the take-home exam. You are advised to work in your groups before the lab sessions, such that you can come to the lab sessions with questions. During the session, the teacher (Emelie Fröberg) will be assisting students individually. If you have completed the designated lab of the session, you are advised to work on the next lab or the hand-in. You can ask any question you want during the labs. In previous (pre-pandemic) years, most students participated in the lab sessions even if they had no specific questions to ask. *In case that the teacher cannot physically attend, the lab sessions will be transformed to online sessions where you book a slot*.

Formative feedback in that learning process

The course uses the group-hand ins as an opportunity to provide formative feedback during your learning journey. Based on previous years' feedback, these are now also (a) graded and (b) with rubrics for clarity of what is expected. The take-home exam will be like the hand-ins and graded similarly.

Structure

The course follows the six phases of the cross-industry standard process data mining (CRISP-DM), based on a fictious case using a pseudo-anonymized real-world dataset. This is a programming course where business and economic domain knowledge is applied for data science analytics. The course covers: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. For the hand-ins and access to the modules, #1 and #2 as well as #4 and #5 are grouped. Details are provided below.

BUSINESS AND DATA UNDERSTANDING

Will unlock Oct 29. In this module, you will:

* Get introduced to the course, through pre-recordings and in a half-group seminar (Oct 29) and by reading the introductory chapter of ISLR.

* Learn to formulate a business use case and translate it to data terminology, by browsing recommended articles and through a panel discussion with practitioners (Salla Franzén, Thomas Falk, and Patrik Tran, on Nov 3).

DATA PREPARATION

Will unlock Nov 3. In this module, you will:

* Learn to access and transform data in a MySQL-database on AWS, by watching pre-recorded videos (Nov 10), doing two labs (Nov 12, 15) and use the MySQL Reference Manual.

* Import data into R and handle missing data, through reading materials (specifically van Buuren) and one lab (Nov 17).

* Each group can book one virtual time slot for 20 min, either on Nov 19 or on Nov 22, to get help for the deadline.

MODELING AND EVALUATION

Will unlock Nov 15. In this module, you will:

* Learn to fit and evaluate models based on supervised machine learning. You will closely follow the ISLR, where each lab is designed to focus on one chapter. You will practice doing the textbook examples working with the real-world dataset. We work exclusively with in-class labs during this lab.

DEPLOYMENT

Will unlock Nov 29. In this module, you will:

* Learn to communicate results with data visualization using Ggplot, through one lecture and one lab.

* Reflect upon and problematize deployment, through a panel discussion with practitioners (Errol Koolmeister, Rebecka Cedering Ångström, Magnus Petterson, **to be confirmed**, Dec 10). This part is not covered in the hand-in, but will be in the take-home exam.

Canvas

The modules in Canvas are used to enforce the format and structure. There are four modules: (a) Business and Data Understanding, (b) Data Preparation, (c) Modeling and Evaluation, and (d) Deployment. These are unlocked in the same order. The labs are available through the modules, named with their dates. If you have questions, the labs are constructed as Discussion-topics and thus you can reply below the instructions. If you have a general question, you can access the general topic through the home page (syllabus). The group assignments, take-home exam and self-reflections are handed in through Canvas. The formative feedback will be available through Canvas.

Prerequisites

R (e.g., course 5304, 4317, 7316, or self-study course, such as R Programming on Coursera, R Basics on Udemy)

Introductory level algebra and basic statistics.

Literature

MAIN

Gareth, J., Witten, D., Hastie, T. & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. (8th Ed.). Springer (available both in print and free online:https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf)

BUSINESS AND DATA UNDERSTANDING

Choose one or several articles you find interesting, browse through the content:

- Athey, S. & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11. 685-725. DOI: 10.1146/annurev-economics-080217-053433
- Bertomeu, J. (2020) Machine learning improves accounting: discussion, implementation and research opportunities. *Review of Accounting Studies*, 25(3). 1135–1155. DOI: 10.1007/s11142-020-09554-9
- Cao, S., Jiang, W., Wang, J. L., & Yang, B. (2021). From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses (No. 28800; *NBER Working Papers*).

- Davenport, T. (2020). Beyond Unicorns: Educating, Classifying, and Certifying Business Data Scientists. *Harvard Data Science Review*, 2(2). DOI: 10.1162/99608f92.55546b4a
- George, G., Osinga, E. C., Dovev, L. & Scott, B. A. (2016). Big Data and Data Science Methods for Management Research. *Academy of Management Journal*, 59(5). 1493–1507. DOI: 10.5465/amj.2016.4005
- Gu, S., Kelly, B. & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5). 2223–2273. DOI: 10.1093/rfs/hhaa009
- Karolyi, G. A. & Van Nieuwerburgh, S. (2020). New Methods for the Cross-Section of Returns. *The Review of Financial Studies*, 33(5). 1879–1890. DOI: 10.1093/rfs/hhaa019

DATA PREPARATION

ISLR does not cover data preparation (even though this is what data science is mostly about), please refer to:

- Anscombe, Francis J. (1973). Graphs in statistical analysis. The American Statistician, 27, 17–21. doi: 10.2307/2682899.
- MySQL 8.0 Reference Manual: Retrieving Data; Data Types; Functions
- Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press: https://stefvanbuuren.name/fimd/
- Wickham, Hadley. (2014). Tidy Data. Journal of Statistical Software, 59(10). 1-23. DOI: 10.18637/jss.v059.i10

MODELING AND EVALUATION

ISLR has full coverage, except for XGBoost:

- XGBoost Documentation: https://xgboost.readthedocs.io/en/latest/

DEPLOYMENT

Choose one or several articles you find interesting, browse through the content:

- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. DOI: 10.1037/xge0000033
- Rai, A. Explainable AI: from black box to glass box. *J. of the Acad. Mark. Sci.* 48, 137–141 (2020). DOI: 10.1007/s11747-019-00710-5
- Vinuesa, R., Azizpour, H., Leite, I. et al. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*. 11(233). 1–10. DOI: 10.1038/s41467-019-14108-y

Examination

Hand-ins in groups of 4 throughout the course (40 points).

Take-Home Exam (60 points). An individual take-home exam will be made available on Dec 16 and is due on Dec 22. Please note that it is available throughout the entire exam period, so that you can also take other exams. The recommendation is to block one working day to complete the exam.

* Individuals can book one virtual time slot for 15 min, on Dec 13, limited to 11 people per time slot.

Reflections of Learning (P/F). Reflections on learning in the course is mandatory (pass/fail) to get a final grade and is due on Jan 12 (half a page PDF).

Attendance

There is no mandatory attendance required. You are strongly encourage to participate in the panel discussions on Nov 3 and Dec 10, either in room R or on Zoom.

Granted educational support

Have you been granted educational support because of a documented disability? If so, please contact the course director as soon as possible so that adequate accommodations can be made. Contact specialneeds@hhs.se if you have questions about being granted educational support or need a copy of your educational support certificate. Read more at Special Needs on the Portal.

Schedule

Course Summary:

Date	Details	Due
Fri Oct 29, 2021	Hybrid (pre-recording + half-class) (Seminar group 1)	1:15pm to 2pm
	Hybrid (pre-recording + half-class) (Seminar group 2)	2:15pm to 3pm
Wed Nov 3, 2021	Hybrid panel discussion	1:15pm to 3pm
	Online (pre-recording)	1:15pm to 3pm

Date	Details	Due
Wed Nov 10, 2021	**Assignment:** Group hand-in (Nov 10)	due by 4pm
Fri Nov 12, 2021	In-class lab	1:15pm to 3pm
Mon Nov 15, 2021	In-class lab	1:15pm to 3pm
Wed Nov 17, 2021	In-class lab	1:15pm to 3pm
Fri Nov 19, 2021	Online lab: Book slot	1:15pm to Nov 22 at 3pm
	Online lab: Book slot	1:15pm to 3pm
Mon Nov 22, 2021	Online lab: Book slot	1:15pm to 3pm
	Assignment: Group hand-in (Nov 22)	due by 4pm
Wed Nov 24, 2021	In-class lab	1:15pm to 3pm
Fri Nov 26, 2021	In-class lab	1:15pm to 3pm

Date	Details	Due
Mon Nov 29, 2021	In-class lab	1:15pm to 3pm
Wed Dec 1, 2021	In-class lab	1:15pm to 3pm
Fri Dec 3, 2021	In-class lab	1:15pm to 3pm
Mon Dec 6, 2021	In-class lecture	1:15pm to 3pm
	Assignment: Group hand-in (Dec 6)	due by 4pm
Wed Dec 8, 2021	In-class lab	1:15pm to 3pm
Fri Dec 10, 2021	Hybrid panel discussion	1:15pm to 3pm
Mon Dec 13, 2021	Q&A: Book slot	1:15pm to 3pm
	Q&A: Book slot	1:15pm to 3pm
	Assignment: Group hand-in (Dec 13)	due by 4pm
Wed Dec 22, 2021	**Assignment:** Take-home exam (Dec 22)	due by 11:59pm

HOUSE OF INNOVATION

Date	Details	Due
Wed Jan 12, 2022	**Assignment:** Self-reflections (Jan 12)	due by 11:59pm