30E03000 – Data Science for Business I (6cr)

Syllabus

Version 4.5 (24.08.2023)

Instructors' contact information	Course information
Names: Pekka Malo **Teaching Assistant(s):** Antti Suominen, Iaroslav Kriutchkov **E-mail:** firstname.surname@aalto.fi **Professors' webpages:** https://people.aalto.fi/pekka_malo	**Status of the course:** Advanced Studies in Master's degree program in Information and Service Management (DR2013); Business application course in the Aalto level minor on Analytics and Data Science **Academic Year, Period:** 2023-2024, Period I **Location:** mostly online **Language of Instruction:** English **Course Website:** https://mycourses.aalto.fi/course/view.php?id=40799

1. Overview

"Data Science for Business – What You Need to Know About Data Mining and Data Analytic Thinking"

The objective of the course is to provide an introduction to practical data science from the perspective of business analysts. The course takes a problem-based approach for teaching data-analytic thinking and fundamental concepts/tools that are needed for data-driven decision making in business. As such, the course is not a replacement for algorithm-centered courses that give deeper insights to the data mining techniques.

The course will discuss the following topics:

- **Fundamentals of Predictive Analytics** – Topics covered include basics of predictive modeling, classification, variable selection, hyperparameter tuning, evaluation of models, expected value framework, problem of overfitting and its avoidance.
- **Data Science Tools for Business Analysts** – A technical objective of the course is to introduce Python programming for business analysts. The module does not require prior knowledge in Python but benefits from prior experience in scripting/programming. In addition to fundamentals of data analysis with Python, the module offers an opportunity to experiment with more advanced techniques such as solving of high-dimensional regression problems via regularization techniques. Given the diverse background of course participants, parts of the module are designed to be suitable for self-study to allow progress at different paces and at varying levels of challenge. Other tools (including commercial platforms) may be introduced during the visiting lectures from different companies.

2. Target group and prerequisites

The course is intended for participants with diverse backgrounds:

- Business analysts and developers who will be implementing and evaluating data science solutions;
- Aspiring future data scientists;
- Business people who will be working with data scientists, managing data science-oriented projects, or investing in data-driven ventures.

The course has a strong focus on empirical assignments, which require prior knowledge in statistics and basic skills in programming/scripting (or at least willingness to learn). However, more theoretical or mathematical aspects of data analytics are beyond the scope of the course.

3. Learning outcomes

After completing the course, the students will be able to

- identify the role of data as a business asset;
- understand the principles of predictive modeling;
- recognize how different data science methods can support business decision-making;
- learn basic data analytic techniques for solving business problems;
- understand the promises and limitations of big data;
- gain some experience in using data analytic tools that are widely used in companies.

4. Assessment, assignments, and grading

The course assessment is comprised of the following parts:

- Assignments 30%.
- Course project (**team case**; cannot be done individually) 40%
- Final MyCourses quiz on theory and practice 30%

All assignments must be completed to pass the course. Additional details and instructions about this will be provided during the course.

5. Readings

- Provost, F. and Fawcett, T. (2013) "Data Science for Business: What you need to know about data mining and data-analytic thinking." O'Reilly Media, 1st Edition.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009): "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer Texts in Statistics.
- Shalev-Shwartz, S., & Ben-David, S. (2014). "Understanding machine learning: From theory to algorithms." Cambridge university press.

6. Preliminary schedule

All lectures and tutorials will be pre-recorded and uploaded on Monday each week. Please note that the schedule may be subject to change during the course.

Thursday Q&A session via Zoom: 16:30-18:30

Week	Dates	Topic
1	4.9. – 8.9.2023	**Introduction to Predictive Analytics** Lecture: • Fundamental concepts and definitions ◦ Data mining as a process (CRISP-DM) ◦ Data mining tasks and popular algorithms ◦ Supervised vs. unsupervised learning • Introduction to classification algorithms ◦ Decision trees ◦ Support vector machines Theory to practice (Tutorial 1): • Data loading, preprocessing, one-hot encoding, train/test splitting, model assessment Wednesday: live course kick-off session (room R030/C105 T2, voluntary): formation of project groups.
2	11.9. – 15.9.2023	**Model performance and feature selection** Lecture: • Evaluating performance of classification models ◦ Expected value framework ◦ Visualizing model performance • Quick review on linear regression • Feature selection ◦ Forward and backward feature selection ◦ Information criteria Theory to practice (Tutorial 2): • Feature engineering • Forward and backward feature selection Thursday: Q&A in Zoom (ask anything, e.g. related to lectures, tutorials, assignments, projects)

Week	Dates	Topic
3	**18.9. – 22.9.2023**	**Hyperparameter tuning and cross validation** Lecture: • Hyperparameter optimization • Grid search, random search, successive halving • Bayesian optimization Theory to practice (Tutorial 3): • Using hyperparameter optimization to improve model performance Thursday: Q&A in Zoom (ask anything, e.g. related to lectures, tutorials, assignments, projects)
4	**25.9. – 39.9.2023**	**Ensemble methods** Lecture • Bagging • Boosting • Gradient boosting • Boosted tree models Theory to practice (Tutorial 4): • Applying ensemble methods Thursday: Q&A in Zoom (ask anything, e.g. related to lectures, tutorials, assignments, projects)
5	**2.10. – 6.10.2023**	**Regularized regression models** Lecture • Regularized regression models ◦ Lasso regression ◦ Ridge regression Theory to practice (Tutorial 5): • Applying regularized models with categorical and continuous response variables
6	**9.10. – 13.10.2023** **Mon 16:30–18:30** **Tue 16:30–18:30** **Wed 16:30–18:30**	**Team Case Presentations (live event in room R030/C105 T2!)** Team Case Presentations - session 1, Monday 9.10.2023 Team Case Presentations - session 2, Tuesday 10.10.2023 Team Case Presentations - session 3, Thursday 12.10.2023 *Please note: you are required to attend at least the day when your own team presents (and encouraged to attend the other two sessions as well)*

7. Course workload

The following is a tentative breakdown of the workload of the course. All hours are academic hours.

Theory and tutorials	54h
Class preparation	12h
Assignments	48h
Course project	46h
Total	**160h (6op)**

8. Ethical rules

Aalto University Code of Academic Integrity and Handling Thereof:
https://into.aalto.fi/pages/viewpage.action?pageId=3772443

9. Student selection

The maximum number of seats on this course is 120. Note that Sisu allows more than 120 people to register, so a successful registration does not equal a spot on the course (no "first come first serve"). Instead, the selection is made by Sisu **automatically** based on the priority groups . The priority for the student selection is as follows:

1. Aalto ISM MSc students.
2. Students in Master's Programme in ICT Innovation (EIT digital).
3. Bachelor's students in Business with 150 credits complete.
4. Other Aalto students.