Big Data analytics (Predictive Analytics for Big Data)

Lecturer: Maria Chiara Debernardi

Language

English

Course description and objectives

Big Data is the hot new buzzword in ICT circles. The proliferation of digital technologies with digital storage and recording media has created massive amounts of diverse data, which can be strategically exploited by companies in all sectors. Big Data, which can take up terabytes and petabytes of storage space in diverse formats including text, video, sound, images, and more, is actually analyzed with the same methods as more traditional data are, with the only difference being the tools required to support the specific characteristics defining Big Data.
The course gives an overview of the Big Data phenomenon, focusing on how to extract value from Big Data using predictive analytics techniques.

Upon successful completion of this course, students should be able to:

- Understand the Big Data phenomenon, and the differences between Big Data and Traditional Data
- Understand the potential use of Data in a corporate environment
- Understand the use of Predictive Analytics tools to be used for extracting valuable information from raw data

Audience

The course is open to all Bocconi students. In particular it is targeted at:

- All those who want to understand what Big Data really is, and how to exploit it
- All those who want to gain some practical, analytical skills and confidence with the Data Scientist Toolkit

Prerequisites

It is essential that participants have attended and successfully passed the IT course provided for in their study plan or have equivalent skills.
It is also desirable to have a good knowledge of the fundamentals of descriptive and inferential statistics, corresponding to the first Statistics exam of one's study plan.

Duration

12 hours

Teaching mode

It will be possible to join the course in presence and/or in distance, by connecting remotely and following the streaming of the lesson held in the classroom.

Calendar

Lecture	Date	Time	Room	Lesson in person with groups by student ID number
1	Mon 16/11/2020	18.40 - 20.10	Info AS05	Even
2	Wed 18/11/2020	18.40 - 20.10	Info AS05	Even
3	Mon 23/11/2020	18.40 - 20.10	Info AS05	Odd
4	Wed 25/11/2020	18.40 - 20.10	Info AS05	Odd
5	Mon 30/11/2020	18.40 - 20.10	Info AS05	Even
6	Wed 02/12/2020	18.40 - 20.10	Info AS05	Even

Syllabus of the course

Lesson	Topics	Book reference
1	**Introduction** - Big Data: definition and taxonomy - Predictive Analytics: definitions ***Setting up the KNIME environment*** ***KNIME sample flow***	Parr. 1.1, 12.1 + slides

Lesson	Topics	Book reference
2	**Business understanding and Data preparation** - CRISP-DM: how to efficiently create a predictive analytics model - Data preparation: the most important task for a Data Scientist - Data exploration: understanding the dataset ***Introduction to OpenRefine*** ***Exercise***	Parr. 3.3, 4.2, 4.3, 4.4, 2.2, 2.3, 6.1 + slides
3	**Predictive analytics techniques** - Supervised *vs*. unsupervised / black-box *vs*. rule-based models - Predictive analytics applications: when to use which model - Predictive analytics algorithms: characteristics and taxonomy - Sampling: train, test and cross-validation ***Exercise***	Parr. 11.5, 1.3, 1.4, 7.1, 7.2, 7.3 + slides
4	**Classification algorithms** - Data preparation for classification tasks - Setting up a classification model - Model performance & evaluation ***Exercise***	Parr. 9.6, 9.7, 10.2, 11.1, 7.4, 7.5 + slides
5	**From rule-based to black box and back** - From the Linear regression to the Neural Network - How to extract rules from results ***Exercise***	Parr. 9.1, 9.2, 9.4, 11.6 + slides
6	**Model ensembles** - Bagging - Boosting - Random forest - Stacking ***Exercise***	Parr. 11.3, 11.2.3 + slides

Suggested bibliography

Skiena Steven S., *The Data Science Design Manual*, Springer, 2017

More references will be given during the course.

Software used

Those who attend remotely must have already installed the following free applications on their machine before the first lesson starts:

- KNIME version 4.2.1 or higher:
  knime.com (download from knime.com/downloads/download-knime)
- OpenRefine 3.4:
  openrefine.org (download from openrefine.org/download)

Available seats

This activity is limited to **60** participants. Registrations cannot be carried out once this number has been reached or after closing of the registration period.