Introduction to Data science with R

40 hours

Course Overview:

Data science involves various disciplines: statistics, computer science, IT, management science, and more. With a lot of data pouring-in from various sources, data science is becoming the "sexiest" profession of the 21st century.

In simple terms, data science can be described as putting on a "detective's hat", and providing answers to business (or research) questions, helping managers reach the right decisions, based on a quantitative analysis. If the data scientist is a detective, than R is his magnifying glass - an important and necessary tool which helps us analyze, make sense and provide insights out of raw data. R is also open source, fun and easy to learn!

If you are a programmer, a BI analyst, an IT professional, a statistician, social scientist, or otherwise a person dealing with data, and want to:

- Learn the basics of data science
- Understand how to harness the power of R for analysis and business decision making
- Get the state-of-the-art tools and know-how to get you going with R

Then this course if for you.

Our course is a 5 days course, packed with 40 hours of frontal classes and exercises in R, which will get you started in no time. The course is aimed at business use cases and is full of example of business problems which are solved with data science modelling in R.

Who should attend:

- Web developers who want to implement data analysis features in their webpage
- Professionals and Data Analysts working in Business Intelligence and aspire to become Data Scientists
- Researchers who perform data analysis including graphs

Required Skills:

- Basic mathematics and good analytical skills
- Basic understanding of statistics, and data structures
- Some programming knowledge (the more the merrier)

Course Contents:

Introduction to data science - what is it all about?

- Skills of a data scientist
- The average day of a data scientist
  - How does a "data science project" looks like?
- A teaser - examples of business problems solved with data science
  - Recommender systems (what was the "netflix challenge")
  - Time series forecasting
  - Classification problems
  - Clustering your customers into groups
  - Predicting the probability a customer is going to buy

  - Shiny dashboards
- Some theory (1): How does a dataset looks like?
- Some theory (2): What is a data model?

“Diving in” - technical introduction

- The IDE (R and RStudio)
- Base R versus packages
- The data.frame and the tibble
- Other data types (factors, strings, matrices, numerics)
- Control structures (“loops and if-else” statements)

The tidyverse.

- Reading the data into R: from excel, csv, SPSS, SAS, json, etc. Packages readr, haven, readxl, jsonlight
- Manipulating the data - “crunching and munging” (packages dplyr, tidyr)
- Functional programming (“looping without loops”) with package purrr and dplyr::do.

Plot your data with ggplot2

- The ggplot philosophy (“the grammar of graphics”)
- Exploring a distribution and detecting outliers (ecdf, histogram, boxplot)
- Exploring the relationship of two or more characteristics (bar plot, line plot, facets)
- Making an interactive chart with ggplot2+plotly.

Modelling and evaluating models

- Regression
  - Intro and theory
  - Linear regression (lm)
  - Non linear regression (glm, lasso, ridge)
  - Variable selection and dimension reduction
- Classification
  - Trees (rpart)
  - Random forests (randomForest)
  - k-NN - nearest neighbors (knn, fnn)
  - Support vector machines
- Clustering (“unsupervised learning”)
  - Kmeans clustering (kmeans)

Next steps - what should you do to sharpen up your skills?

- All about cheatsheets
- Data science competitions
- Additional sources (e.g., R-Bloggers)

(Optional topics) If time permits and depending on the participant’s background, we might also touch upon part of the following:

- Building interactive data apps using shiny
- Scraping (pulling data from websites), and web APIs (package httr)
- Creating your own R server in the cloud
- Creating a web api with R (package plumber)