Machine Learning for Business, Economics and Finance

Bachelor course (ECTS: 6)

13.00 – 15.30 (CET Ljubljana)

VÉKÁS Péter

Corvinus University of Budapest, Hungary

Course objectives and learning outcomes:

The aim of the course is to offer a brief, interactive and lively introduction to the immensely popular field of Data Science to interested students using the R programming language. Prior exposure to R is not required. During the interactive classes, students will receive an introduction to contemporary Data Visualization, Multivariate Statistical and Machine Learning techniques as they are applied in the 21st century by professionals in the fields of business, economics and finance. Besides numerous real-life applications, there will be an emphasis on methodological questions and finding the most appropriate technique to tackle a given problem. Students who attend the course will be capable of performing complex analyses of data from diverse sources using various popular techniques. Students will apply what they have learned in the course by analyzing their own dataset in pairs and presenting their results in front of the audience in order to master essential communication skills.

Prerequisites for attending the course:

Statistics

Course syllabus/Daily topics:

PROGRAMME DAY	ACTIVITY/TOPIC/SESSION
Monday, 5 July	**LSS Welcome session (no lectures)**
Tuesday, 6 July	**Concepts and Basics of Data Analysis in R** Types of data (cross-sectional, time series, spatial, network, structured and unstructured data). Scalars, vectors, matrices and data frames. Exporting and importing data. Missing data and imputation. Measurement scales of variables. Descriptive statistics. Basic R programming constructs (selection, loops, data transformations). R Studio and the Markdown language. **Application:** Most recent country-level socio-economic indicators of the world.

ONLINE Ljubljana Summer School

5 – 23 July 2021

Wednesday, 7 July	**Data Visualization: a Picture is Worth a Thousand Words** The Grammar of Graphics. Univariate plots. Pairwise plots. Measures and tests of correlation. Correlation vs. causality. Plots of multiple variables. Time series plots. Miscellaneous plots (animated, spatial, network and interactive plots). **Application:** Visual inspection of a database of bank customers.
Thursday, 8 July	**Introduction to Supervised Machine Learning** Supervised and unsupervised learning. Predictive modeling. Classification and regression. Linear regression and the binary logit model. Nonlinear terms. Overfitting, LASSO and ridge regularization. Measures of model fit. Validation methods. Hyperparameters and tuning. **Applications:** Prediction of real estate prices. Classification of bank customers (credit scoring).
Friday, 9 July	**Decision Trees and Random Forests** Breiman's method of Classification and Regression Trees (CART). Growing and pruning the tree. Bagging and boosting ensembles. Breiman's random forest. Measuring variable importance. Hyperparameter tuning. Differences and similarities between statistics and machine learning. **Applications:** Survivors of the Titanic. Predicting salaries of employees.
Monday, 12 July	**Cluster Analysis** Introduction to unsupervised learning. Distance metrics. Standardization of data. Hierarchical clustering methods. Dendrogram. *k*-means clustering. Methods to determine the number of clusters. Interpretation. **Applications:** The European East-West divide. Forbes Top 100 global corporations.
Tuesday, 13 July	**Principal Components Analysis**

	Dimension reduction. The correlation matrix and its eigen-decomposition. Communalities and total variance explained. Geometric and practical interpretation of principal components. Biplots of principal components and clusters. **Applications:** Student skills based on grades. Yield curve analysis.
Wednesday, 14 July	**Artificial Neural Networks** Artificial intelligence. Neurons, weights, bias and activation functions. Layers and network architecture. Gradient descent and backpropagation learning. **Applications:** Credit scoring. Prediction of individual salaries.
Thursday, 15 July	**Deep Learning** Multilayer perceptron. Advanced learning methods. Deep learning using *keras* in R. Ethical dilemmas of artificial intelligence. **Application:** Credit card fraud detection.
Monday, 19 July	**Text Analytics** Unstructured data. Natural language processing. Bag-of-words model. Frequencies, word clouds and associations. *n*-grams. Sentiment analysis. Topic modelling. **Applications**: News headlines and daily stock returns.
Tuesday, 20 July	**Summary and Practice Problems** Using a medley of miscellaneous methods on a dataset of bank customers. Recap on problematic points, reflecting on the individual needs of students. Brief demonstration of further methods of interest.
Wednesday, 21 July	No lectures (preparation for final examination)
Thursday, 22 July	Final examination / Project presentations
Friday, 23 July	**Meeting hours with students & LSS Farewell session**

Online teaching methods and tools/software used:

1. Online teaching via Zoom (or Microsoft Teams, if preferred).
2. Theoretical background using lecture slides, and practice problems using computer software and the R programming language.
3. 10-minute *Kahoot!* interactive online quiz (www.kahoot.com) at the beginning of every teaching day. On the first day, to assess the background of students. On further days, to recap on new concepts learned on the previous day.

4. Emphasis on interactivity: joint discussion of key concepts with the active participation of students, encouraged by several open-ended questions instead of frontal teaching.
5. Flexible selection of subtopics of interest and datasets to be analyzed by participants as a 'parliament', to improve the appeal of the course material and make students more involved.
6. Animated demonstrations of key concepts (e.g., growing a random forest or finding the final cluster means in *k*-means clustering).
7. Creating and analyzing a dataset compiled from the data of participants using a quick online survey, to make the analysis more interesting and personal and engage participants.
8. Rewarding individual ideas as well as fastest or smartest solutions to specific problems with bonus points, to increase motivation.
9. Emphasis on fun and engaging learning environment, using games, cartoons, interesting examples, real-life data, online articles on current topics, etc.

Course materials/List of readings:

(open source e-books, accessible for free at https://www.ossblog.org/grasp-r-programming-open-source-books/):

- Venables, W.N., Smith, D.M. and the R Core Team (2018). *An Introduction To R: Notes on R: A Programming Environment for Data Analysis and Graphics*. R Core Team.
- Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly, CA, United States.

Online examination methods and evaluation criteria (weighted categories):

Students will form teams (pairs of two). Every team will create their own dataset on a topic of their choice, and will perform data visualization and exploratory analysis individually. Additionally, they will choose and perform three more advanced techniques from the course. Pairs of students will briefly present the findings of their analyses in about 20 minutes (+ 5 minutes for questions and answers) via Zoom (or Microsoft Teams, if preferred), with all other students listening and commenting afterwards. Submitted datasets and R scripts as well as scores from *Kahoot!* quizzes will also be considered in the final individual evaluations of students.

The detailed breakdown of the final score:

- Presentation: 90%
  - *Dataset:* *10%*
  - *Visualization:* *15%*
  - *Analysis:* *45%*
  - *Interpretation:* *20%*
- *Kahoot!* quizzes: 10%

Grading scale:

DEFINITION	%	LOCAL SCALE	ECTS SCALE	Grade (USA)
exceptional knowledge without or with negligible faults	92-100	10	A	A+, A, A-
very good knowledge with some minor faults	85-91	9	B	B+, B
good knowledge with certain faults	77-84	8	C	B
solid knowledge but with several faults	68-76	7	D	C+, C, C-
knowledge only meets minimal criteria	60-67	6	E	D+, D
knowledge does not meet minimal criteria	<60	5	F	

Short course leader(s) biography:

Péter Vékás, Ph.D. is an assistant professor of the Institute of Mathematical and Statistical Modelling of Corvinus University of Budapest, Hungary. He is an economist with a keen interest in applied data science and modelling real-world phenomena using the R programming language. He has given university courses and conference talks in numerous countries of Europe, North and South America, Africa, Asia and Australia. Prior to his current position, he was a lecturer of the University of Groningen, the Netherlands, and a life insurance actuary in the private sector. He has published a book, multiple book chapters and several papers on applied data science, insurance and pensions, and has worked on several projects in the private sector as a freelancer.