NEW YORK UNIVERSITY
STERN
Leonard N. Stern School of Business

DATA MINING WITH R

MSBA Program
2017

SYLLABUS

Professor	**Luis Torgo,** Faculty of Sciences, University of Porto, Portugal
Office; Hours	
Email	ltorgo@dcc.fc.up.pt **Begin subject: [DMBA] …** **-> note!**
Class Dates	Sunday, Aug 13, 2017 Monday, Aug 14, 2017

1. Course Overview

The goal of this course is to provide hands-on experience on key data mining technologies using one particular tool – the R environment.

R is a fast growing technology that has been witnessing widespread acceptance both in academia and industry. Recent surveys have even put it in the top regarding usage by professional data miners (e.g. Rexer Analytics survey, 2013). There are many factors contributing for this acceptance but clearly these include the price (free), being open source (trustworthy software that can be easily inspected/checked for flaws), the extension of available methods (exponential growth of the set of available methods for different application areas), and the available support from the community (an extremely large community of knowledgeable experts proving top-notch support for free).

In this course we will illustrate the use of R for several key data mining processes. This illustration will be driven by concrete case studies that we will "solve" using R. This course can be regarded as a hands-on complement of the Data Mining for Business Analytics course.

After taking this course you should be able to use R for:

1. *Understanding your data.* Exploratory analysis of data frequently provides key insights to data properties and problems that can have a big impact on posterior mining steps and may help in mapping business problems into data mining tasks. We will provide practical illustrations of methods for summarizing, visualizing and preparing your data for model construction.

2. *Master frequently used modeling techniques.* Data can be modeled in many different ways. The outcomes of these models can provide useful information for decision makers. We will address several concrete modeling tasks with frequently used techniques. We will learn how to obtain and apply these models in R.

3. *Correctly assess the performance of models*. Performance assessment is a key step for taking advantage of the results of data mining models. Being able to carry out this task in a reliable way is of key importance to make sure future deployment of data mining pays off.

4. *Easily report and deliver results of data mining*. The outcome of data mining often needs to be reported, communicated and made available to different kinds of people (e.g. decision makers, key personnel of other departments lacking knowledge of data mining, etc.). We will learn how specific tools available in R can boost our productivity in this type of tasks.

2. Focus and interaction

The course will illustrate the use of R for several key data mining tasks. The main focus will be on how to carry out these tasks in R and not on the principles and theory behind these approaches. This means that this is a practical course that will illustrate several of the techniques you have learned on previous courses. We will cover the full data mining cycle from importing data into R till deployment of the data mining results. The course is driven by concrete case studies and full solutions will be provided to allow you to replicate and re-use the solutions, in the spirit of open source software like R.

You are expected to be prepared for class discussions by reading the material that is provided for the pre-module. The classes will address in detail two concrete data mining case studies. The pre-module assigned readings are essential to fully take advantage of the classes and participate in the discussions and exercises that we will have during the classes.

You are expected to attend the entire class sessions, to arrive prior to the starting time, and to follow basic classroom etiquette, including having all electronic devices turned off and put away for the duration of the class (this is Stern policy, see below) and refraining from chatting or doing other work or reading during class.

If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please detain me after class, send me an email with your questions, or ask on the discussion board. The discussion board is much better than sending me email, as it will allow others to take advantage of the discussions. Also, please try to answer your classmates' questions. In grading your class participation I will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class).

I will check my email at least once a day during the week (M-F). ***Your email will get my priority if you include the special tag [DMBA] in the email subject header***. I use this tag to make sure to process class email first. If you do not include the special tag, I may not read the email for a while. If you forget and send without the tag and then remember, just send it again including the tag.

3. Readings

Book: The textbook for the class will be:

Data Mining with R: learning with case studies, 2nd edition by Luis Torgo (2017), CRC Press.

This book illustrates the use of R for data mining through the use of concrete case studies, some of which we will cove in this course. We will complement the book with discussions of some of the case studies, new material not appearing in the book and some extra illustrations.

- Supplemental books (optional):

  - *Dynamic Documents with R and knitr*
  by Yihui Xie
  CRC Press R book series
    - This book provides an extensive coverage of the knitr infra-structure to produce dynamic documents (presentations and reports) using R, by the author of knitr himself.

  - *Data Science in R*
  by Deborah Nolan and Duncan Temple Lang
  CRC Press R book series
    - This recent book is also a collection of data mining casa studies as the text book of this course. This can be a good complement to the course text book as it includes different case studies

  - *R for Data Science*
  by G. Grolemund and H. Wickham
  O'Reilly
    - This book provides an extensive coverage of some essential steps on any data science project using R. It is available for free at http://r4ds.had.co.nz/

  - *Software for Data Analysis, programming with R*
  by John Chambers
  Springer
    - This book provides an extensive coverage of advanced topics of R programming. It is only of use for those wanting to delve deep into R programming and R software development.

  - *Advanced R*
  by Hadley Wickham
  CRC Press R book series
    - This book provides an extensive coverage of advanced topics of R programming, as the previous book. Again it is only of use for those wanting to delve deep into R programming and R software development. It is available for free at http://adv-r.had.co.nz/

4. Software

This course is about using R for data mining. In this context, you are required to have an up-to-date installation of R in your laptop. R can be freely obtained from http://www.r-project.org . RStudio is another free software tool that provides and integrated development environment to R. I strongly recommend that you use RStudio as your tool for interacting with R. RStudio can be freely downloaded from http://www.rstudio.com .

R comes with an extensive set of tools pre-installed. Still, it can be easily extended through the (free) installation of extra packages. We will use several of these. You can easily install any of these extra packages in RStudio. One that we will use extensively is the R package accompanying the textbook that contains many functions and case studies that we will use throughout the course. The name of the package is "DMwR".

IMPORTANT: ***In order to install the above mentioned software you must have access to a computer on which you can install software. Moreover, due to the practical nature of this course you are expected to have your laptop with you during the classes, as it will be necessary for hands-on exercises.***

5. Requirements

There will be pre-module and post-module assignments.

***Pre-module*:**

I will assume that you have already basic knowledge of R (obtained in previous courses). Still, in case you want to refresh your knowledge I also include some optional slides (01_basicConcepts slide set) in the course documentation that provide short introductions to basic topics of R.
Both in the assigned readings and during the classes we will use several data sets for illustration purposes. You will find the data either in course documentation or through the packages you will install.

Assignment #1: Read the slide sets 02_dataImport, 03_dataPreProcess (the last section on Handling Big Data in R is optional), 04_summarization, 05_visualization, sections 3.1 through 3.3 (pages 43 till 86), and also sections 4.1 through 4.5 (pages 193 till 213) of the text book.

Assignment #2: Read the slide set 06_reporting and complete Assignment #2 (group assignment with your study group) in your course documents.

Assignment #3: Read the slide set 07_predAnalytics up to slide 75 and complete Assignment #3 (group assignment with your study group) in your course documents

Assignment #4: Read the rest of the slide set 07_PredAnalytics and complete Assignment #4 (group assignment with your study group) in your course documents.

***Post-module*:**

Assignment #5: Go to the web site http://shiny.rstudio.com/tutorial/ and take the tutorial on building Web applications using Shiny.

Assignment #6: Complete Assignment #6 (group assignment with your study group) in your course documents

Assignment #7: Submit a proposal to use some of the material learned during the course in your capstone project. After my approval, this will be your term project that you should also submit. This is a group assignment within your capstone group.

***Grading Breakdown*:**

1. Assignments: 55%
2. Term Project: 35%
3. Participation & Class Contribution (includes discussion forums): 10%

Grading of Assignments:
Assignments will be graded according to the following criteria

1. Meeting of requirements: 10%
2. Overall quality of the delivered report (assign. 2,3 and 4) or web app (assign. 6): 10%
3. Technical quality of the solution: 30%
4. Demonstrated knowledge of the topics involved in the assignment: 30%
5. Extensiveness of the delivered assignment: 10%
6. Creativity (thinking out of the box) of the proposed solution: 10%

All assignments will receive a final grade and also some feedback on the scores on each of the above criteria. There will also be written comments/feedback on the assignments.

Policy for late submissions:
The following penalties will apply to those whose work is submitted after the specified submission deadlines.

1. One to five calendar days late: one grade subtracted (i.e. an A grade would become a A-)
2. Six to 15 calendar days late: two grades subtracted (i.e. an A grade would become a B+)
3. Over 15 days (provided that work is submitted before the next module): the highest grade possible will be a "C".
4. Work will not normally be accepted after the next module begins and you will be given a failing grade.
5. If there is more than one assignment for a course, the penalties mentioned above would apply to the entire course grade even if one of the assignments is late.
6. If the Capstone is not approved for presentation and additional work is required, the highest grade possible for the Capstone will be a “C”.
7. If a student misses Capstone presentation, the highest grade possible will be a “C”.

The Academic Director may grant extensions without penalty in certain exceptional circumstances (e.g. severe illness). Excessive pressure at work will not be accepted as an excuse for late submission.

6. Bibliographic profile

Luis Torgo is an Associate Professor of the Department of Computer Science of the Faculty of Sciences of the University of Porto, Portugal. He is a senior researcher of LIAAD / INESC Tec, and a current member of the board of this research lab.

Luis Torgo is also an invited professor of the Stern Business School of the New York University where he has been collaborating in the last 3 years at the Master of Science in Business Analytics.

He has been doing research in the area of Data Mining and Machine Learning since 1990, and has published over 100 papers in several foruns of these areas. Luis Torgo is the author of the widely acclaimed *Data Mining with R* book published by CRC Press in 2010 with a strongly revised second edition comming out in January of 2017. He has been involved in many research projects under different roles and involving different types of organizations.

His current broad research interests revolve around analyzing data from dynamic environments, with a particular focus on time and space-time dependent data sets, in the search for unexpected events. In terms of application domains his research is frequently linked with ecological/biological as well as financial domains.

Luis Torgo main contributions to the state of the art on data mining and machine learning are related with tree-based regression methods and more recently with utility-based forecasting methods.

He has a strong experience of teaching different subjects at different academic levels but also in non-academic settings. He is frequently invited for giving short courses on using R for data mining around the world.

Luis Torgo is the CEO and one of the founding partners of KNOYDA a company devoted to training and consulting within data science.

Class Schedule

Day 1

Time	*Duration*	*Topics*	*Readings*
08:00 08:30	00:30	*Breakfast*	
08:30 09:30	01:00	Introduction; Presentation of the Fraud Detection case study; Exploratory analysis; Data pre-processing (dealing with unknown values)	Slides #08 Ch. 6
09:30 10:30	01:00	Hands on exploratory analysis, data pre-processing and reporting	
10:30 11:00	00:30	*BREAK*	
11:00 11:30	00:30	Defining the data mining task; Evaluation criteria; Experimental methodology to be used	Slides #08 Ch. 6
11:30 12:00	00:30	Unsupervised approaches to the problem (Boxplot rule and LOF)	Slides #08 Ch. 6
12:00 12:30	00:30	Hands on unsupervised approaches	
12:30 13:30	01:00	*LUNCH*	
13:30 14:15	00:45	Classification approaches to the problem; Unbalanced classes and SMOTE; Naive Bayes; AdaBoost	Slides #08 Ch. 6
14:15 14:45	00:30	Hands on classification approaches	
14:45 15:05	00:20	Semi-supervised approaches to the problem; Self training a model	Slides #08 Ch. 6
15:05 15:30	00:25	Hands on semi-supervised approaches	
15:30 16:00	00:30	*BREAK*	
16:00 16:45	00:45	Presentation of the Financial Prediction case study; Defining the prediction task	Slides #09 Ch. 5
16:45 17:30	00:45	Hands on Financial Data	

Day 2

Time	*Duration*	*Topics*	*Readings*
08:00 08:30	00:30	*Breakfast*	
08:30 09:15	00:45	Evaluation criteria; Performance estimation for time series models; MARS models	Slides #09 Ch. 5
09:15 09:45	00:30	Hands on Modeling and Performance Estimation using Monte Carlo	
09:45 10:30	00:45	Deploying model predictions: from predictions into actions; Trading policies and a trading simulator	Slides #09 Ch. 5
10:30 11:00	00:30	*BREAK*	
10:30 11:15	00:45	Hands on Trading with Model Outcomes	
11:15 12:00	00:45	Model selection using Monte Carlo Simulations; A trading workflow	Slides #09 Ch. 5
12:00 12:30	00:30	Hands on Model Selection and Final Evaluation	
12:30 13:30	01:00	*LUNCH*	
13:30 14:30	01:00	Deploying Data Mining results; Web Applications using Shiny	Slides #10 sec. 3.6.2
14:30 15:30	01:00	Hands on Web apps with R	
15:30 16:00	00:30	*BREAK*	
16:00 16:45	00:45	Handling geo-referenced data in R	Slides #11
16:45 17:30	00:45	Hands on spatial data	