Government of Russian Federation

Federal State Autonomous Educational Institution of High Education

«National Research University Higher School of Economics»

National Research University
High School of Economics
Faculty of Computer Science

Syllabus for the course
«Data Science for Business»

MAGoLEGO course
from University pool

Author:
Leonid E. Zhukov, Professor, Ph.D lzhukov@hse.ru
Co-author:
Ilya A. Makarov, Senior Lecturer iamakarov@hse.ru

Approved by:

Head of Data Analysis and Artificial Intelligence School, Sergey O. Kuznetsov, Sept. 2019

Recommended by:

Methodical center DOOP HSE, Feb. 2019

Moscow, 2019

1. Course Description

a. **Title:** "Data Science for Business"

Author: Leonid E. Zhukov, National Research University Higher School of Economics, Faculty of Computer Science, Department of Data Analysis and Artificial Intelligence, professor.

Co-author: Ilya A. Makarov, National Research University Higher School of Economics, Faculty of Computer Science, Department of Data Analysis and Artificial Intelligence, senior lecturer, deputy head.

b. **Prerequisites:** The course assumes no prior knowledge of statistics and is based on basic notions of mathematics from high-school. However, all students are advised to be prepared for studying a mathematical discipline, even if they come from a non-mathematical background.

The following knowledge and competence are needed to study the discipline:

- A good command of the English language, both orally and written.
- A basic knowledge of mathematics
- A basic programming experience

c. **Course Type**: elective, MAGoLEGO course.

Place of the discipline in the MagoLego course from University pool. The course «Data Science for Business» is a course taught in MagoLego course from University pool. It is recommended for non-specialists who wish to get fundamental knowledge in analysis of social networks

Scope of Use. The present program establishes minimum demands of students' knowledge and skills, and determines content of the course.

The present syllabus is aimed at department teaching the course, their teaching assistants, and students of MagoLego course from University pool.
This syllabus meets the standards required by:

- Educational standards of National Research University Higher School of Economics;
- Educational program of Federal Master's Degree Program for 2019 https://www.hse.ru/org/hse/elective_courses/MG_KK;
- University curriculum of the Master's program in «Data Science» (010402) for 2019.

d. **Abstract** (Summary of the course).

Data Science discipline formed recently in response to increasing use of data in business. It utilizes data mining, machine learning and statistical methods, but focuses on business applications, solving real world problems and delivering impact on business. This course is using a case-based approach to teaching, i.e. the students will be learning methods and techniques while solving business case problems. Each case will contain a description of a business problem and available data. The goal would be to convert a business problem into analytical and solve it using data with the help of variety of data mining and machine learning methods. The methods will be introduced as needed for each case solution. The course will be hands-on, during the lectures students will learn the approach and implement and solve the case in their

home assignments.

2. Learning Objectives

The learning objective of the course «Data Science for Business» is to provide students with essential knowledge of data mining methods and algorithms and experience in converting business problems into analytical and solving them.

3. Learning outcomes

After completing the study of the discipline «Data Science for Business» the student should:

- Know basic notation and terminology used in data science
- Be able to visualize, summarize and analyze datasets
- Understand basic principles behind analysis algorithm
- Develop practical skills in Python or R programming
- Being able to formulate and solve analytical problems for given business problem

After completing the study of the discipline «Data Science for Business» the student should have the following competences:

Competence	Code	Code (UC)	Descriptors (indicators of achievement of the result)	Educative forms and methods aimed at generation and development of the competence
The ability to reflect developed methods of activity.	SC-1	SC-M1	The student is able to reflect developed network methods in social sciences	Lectures and tutorials.
The ability to propose a model to invent and test methods and tools of professional activity	SC-2	SC-M2	The student is able to visualize and summarize data, develop mathematical models	Examples covered during the lectures and tutorials. Assignments.
Capability of development of new research methods, change of scientific and industrial profile of self-activities	SC-3	SC-M3	Students obtain necessary knowledge in network science, sufficient to develop new methods in other disciplines.	Assignments, additional material/reading provided.
The ability to describe problems and situations of professional activity in terms	PC-5	IC-M5.3_5.4_5.6_2.4.1	The student is able to describe problems in terms of network science	Lectures and tutorials.

Competence	Code	Code (UC)	Descriptors (indicators of achievement of the result)	Educative forms and methods aimed at generation and development of the competence
of humanitarian, economic and social sciences to solve problems which occur across sciences, in allied professional fields.				

4. Course Plan

Two pairs consist of 2 academic hour for lecture followed by 2 academic hour for computer exercises/labs after lecture. Additional office hours for lectures' content are provided.

№	Topic	Total hours	Contact hours		Self-study
			Lectures	Seminars	
1.	Introduction to Data Science for Business	10	2	2	6
2.	Dealing with data	10	2	2	6
3.	Data mining, machine learning, statistics	11	2	2	7
4.	Case study 1. Customer segmentation	11	2	2	7
5.	Case study 2. Customer churn modeling	11	2	2	7
6.	Case study 3. Pricing	11	2	2	7
7.	Case study 4. Production optimization	11	2	2	7
8.	Case study 5. Sales territory design	11	2	2	7
9.	Dealing with big and fast data	11	2	2	7
10.	Impacting the business	11	2	2	7
Total:		**108**	**20**	**20**	**68**

Course Description

The following list describes the topics that will be covered in the course in correspondence with lecture order.

Topic 1. Introduction to Data Science for Business
Introduction to a new discipline Data Science. Its place in academic world and industry. Examples of real world problems

Topic 2. Dealing with data
Skills needed to work with data. Data cleaning and preparation. Basic data analysis

Topic 3. Data mining, machine learning, statistics.
Major classes of algorithms, applicability, solution quality metrics

Topic 4. Case study 1 – Customer segmentation in marketing

The goal of the case is to group customers into clusters based on some customer similarity metrics

Algorithms: clustering – k-means, agglomerative, dimensionality reduction - PCA

Topic 5. Case study 2 – Customer churn modeling

The goal of the case is to predict which customers are going to leave the service within a given time.

Algorithms: Supervised learning – logistic regression, decision trees, random forest.

Topic 6. Case study 3 - Pricing

The goal of the case is to determine the optimal pricing for goods and services

Algorithms: supervised learning – regression (linear and non-linear models)

Topic 7. Case study 4 – Production optimization

The goal of the case is to predict an output of the production line and find optimal parameter setting

Algorithms: supervised learning – regression, non-linear optimization

Topic 8. Case study 5 – Sales territory design

The goal of the case is to select locations of the sales offices to maximize the coverage under constrained resources

Algorithms: clustering and geo-analytics approaches

Topic 9. Dealing with big and fast data

Handling data in real world – big data and data streams.

Topic 10. Impacting the business

How to create a visible impact on business with analytics

5. Reading List and Materials

We do not follow a particular textbook in this subject, but the student may find the following references useful:

5.1 Recommended Reading

1. Foster Provost , Tom Fawcett. “Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking”. O’Reilly Media, 2013.
2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. “An Introduction to Statistical Learning: with Applications in R”. Springer , 2017

5.2 Supplementary Reading

1. Christopher Bishop. “Pattern Recognition and Machine Learning”. Springer, 2006.
2. Maksim Tsvetovat and Alexander Kouznetsov. “Data Science for Business for Startups”. O’Reilly Media, 2011.

5.3 R and Python programming

1. Robert Knell. “Introductory R: A Beginner's Guide to Data Visualisation, Statistical Analysis and Programming in R”, 2013
2. Wes McKinney. “Python for Data Analysis: Data Wrangling with Pandas, NumPy

and IPython.", O'Reily 2017
3. Robert Kabacoff. "R in action. Data Analysis and graphics with R", Manning Publications, 2011

5.4 Popular Reading

1. Nate Silver. "The Signal and the Noise: Why So Many Predictions Fail--but Some Don't", Penguin books, 2015.
2. Eric Siegel. "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die", Wiley 2016.

5.5 Course webpage

Students are provided with links to the lecture notes, problem sheets and their solutions, assignments and their solutions, and additional readings.

6. Grading System

Type of grading	**Type of work**		
	Homework	5	Solving homework task and examples
Final			As average cumulative grade

The assessment consists of one homework, handed out to the students during the semester. The homework problems are based on each lecture topics.

Final assessment is the final exam. Students have to demonstrate knowledge of probability and statistics theory.

7. Guidelines for Knowledge Assessment

The grade formula:

Final course mark *is obtained from the following formula*:
Final=(HW1+HW2+HW3+HW4+HW5)/5.

The grades are rounded in favor of examiner/lecturer with respect to regularity of class and home works. All grades, having a fractional part greater than 0.5, are rounded up.

Table of Grade Accordance

Ten-point Grading Scale	**Five-point Grading Scale**	
1 - very bad 2 – bad 3 – no pass	Unsatisfactory - 2	**FAIL**
4 – pass 5 – highly pass	Satisfactory – 3	

6 – good 7 – very good	Good – 4	**PASS**
8 – almost excellent 9 – excellent 10 – perfect	Excellent – 5	

8. Methods of Instruction

Course lecturer is advised to use interactive learning methods, which allow participation of the majority of students, such as slide presentations, combined with writing materials on board, and usage of interdisciplinary papers to present connections between probability theory and statistics. The course is intended to be adaptive, but it is normal to differentiate tasks in a group if necessary, and direct fast learners to solve more complicated tasks.

Term Educational Technology

The following educational technologies are used in the study process:

- discussion and analysis of the results during the computer exercises;
- regular assignments to test the progress of the student;
- consultation time on Monday mornings with lecturer and after lecture;
- teleconference lectures
- office hours and classes with tutor and teaching assistants
- tutorship

Recommendations for students

The course is interactive. Lectures are combined with exercises. Students are invited to ask questions and actively participate in group discussions.

The lecturer is ready to answer your questions online by official e-mails that you can find in the "contacts" section**.** This course is taught in English, and students can ask teaching assistants to help them with the language.

In addition to introductory classes on R language you may find useful course from Coursera: https://www.coursera.org/course/rprog

9. Special Equipment and Software Support

The course requires a laptop and projector.
R statistical modeling environment, RStudio IDE, igraph library
R: http://www.r-project.org
RStudio: http://www.rstudio.com
Python (Anaconda distribution) : https://www.anaconda.com/distribution/

Lecture materials, course structure and the syllabus are prepared by Leonid Zhukov.

Also the syllabus part concerning classes' structure was made in collaboration with Ilya Makarov.