CEU MSc in Business Analytics

Course: Data Science for Business

Term: Winter 2015
Credits: 2 (4 ECTS)
Department of Economics / CEU Business School
Instructor: Szilard Pafka

Course description:

This course will provide a brief overview of Data Science, the field aimed at extracting business value from data. Despite the new name and the recent hype, Data Science is actually not new, it has solid foundations in *statistics* and *computing technology* that go back several decades. A Data Science project usually involves several iterations of the following steps: business understanding, data acquisition, exploratory data analysis, data cleaning, feature engineering, advanced statistical modeling, model validation, technical implementation and deployment and communication of results to decision makers. This course will discuss these steps along with the knowledge and the technology necessary to be able to perform them; some topics will be discussed in very details, while others will have pointers to other courses in the MSc in Business Analytics program for further coverage.

A large part of this current course will be dedicated to advanced statistical modeling / machine learning / predictive analytics. We will discuss methods for supervised learning such as neural networks, decision trees, naive Bayes, k-nearest neighbors, support vector machines, random forests or gradient boosted machines. We will discuss important issues regarding model evaluation and validation (bias and variance, overfitting, training and test sets, cross-validation, data leakage etc.). We will also cover methods for unsupervised learning such as principal component analysis and clustering (k-means, hierarchical).

Other topics, equally important for Data Science will be just briefly discussed here with more details following in other courses. For example, students will get hands-on experience with *exploratory data analysis*, *data manipulation/preparation* and cleaning, *data visualization*, *programming* with data and tools that help *reproducibility* in the Tools for Analytics Lab (the R Track). Data storage, databases, data transformations (data pipelines/ETL) and SQL will be discussed in The Big Data Computing course (also providing a *systems* view). *Data visualization* (a very important component in exploratory data analysis and also in the communication of results to decision makers) will be discussed in further details in the Data Visualization elective course (highly recommended). Some of the more traditional *statistical modeling* topics (such as linear regression) have been already covered in the Data Analysis I and II courses.

Assessment:

20% class participation
40% final exam
40% data analysis project

Course schedule:

1. The Data Science process: business understanding, data acquisition, exploratory data analysis, data cleaning, feature engineering, advanced statistical modeling, model validation, technical implementation and deployment, communication of results to decision makers.

2. Tools for Data Science (R/Python, databases). Exploratory data analysis. Data preparation/munging. Data visualization

3. An example project

4. Supervised learning. Linear models. k-nearest neighbors

5. Decision trees. Random forests. GBM

6. Model evaluation and validation: bias and variance, overfitting, training and test sets, cross-validation, data leakage

7. Naive Bayes. Neural networks. SVM

8. Unsupervised learning. Clustering (k-means, hierarchical). Principal component analysis

9. Q&A. Discuss requirements for data analysis project

10. Other miscellaneous topics