# Educational Data Mining with Python and Apache Spark: A Hands-on Tutorial

Lalitha Agnihotri
McGraw-Hill Education
281 Summer Street
Boston, Massachusetts, US
lalitha.agnihotri
@mheducation.com

Shirin Mojarad
McGraw-Hill Education
281 Summer Street
Boston, Massachusetts, US
shirin.mojarad
@mheducation.com

Nicholas Lewkow
McGraw-Hill Education
281 Summer Street
Boston, Massachusetts, US
nicholas.lewkow
@mheducation.com

Alfred Essa
McGraw-Hill Education
281 Summer Street
Boston, Massachusetts, US
alfred.essa
@mheducation.com

## ABSTRACT

Enormous amount of educational data has been accumulated through Massive Open Online Courses (MOOCs), as well as commercial and non-commercial learning platforms. This is in addition to the educational data released by US government since 2012 to facilitate disruption in education by making data freely available. The high volume, variety and velocity of collected data necessitate use of big data tools and storage systems such as distributed databases for storage and Apache Spark for analysis.

This tutorial will introduce researchers and faculty to real-world applications involving data mining and predictive analytics in learning sciences. In addition, the tutorial will introduce statistics required to validate and accurately report results. Topics will cover how big data is being used to transform education. Specifically, we will demonstrate how exploratory data analysis, data mining, predictive analytics, machine learning, and visualization techniques are being applied to educational big data to improve learning and scale insights driven from millions of student's records.

The tutorial will be held over a half day and will be hands on with pre-posted material. Due to the interdisciplinary nature of work, the tutorial appeals to researchers from a wide range of backgrounds including big data, predictive analytics, learning sciences, educational data mining, and in general, those interested in how big data analytics can transform learning. As a prerequisite, attendees are required to have familiarity with at least one programming language.



© 2016 ACM. ISBN 978-1-4503-4190-5
DOI: http://dx.doi.org/10.1145/2883851.2883857

## CCS Concepts

•**Information systems** → **Data mining;** *Data cleaning;* •**Mathematics of computing** → **Exploratory data analysis;** •**Computing methodologies** → **MapReduce algorithms; Modeling and simulation;** *Modeling methodologies; Model verification and validation;*

## Keywords

educational data mining, learning analytics, python, spark, predictive analytics, machine learning, exploratory data analysis, big data, data mining, visualization, simulation, parallel computing

## 1. MOTIVATION

A number of commentators, including Clayton Christensen, have argued that education is prime for disruption [1]. Education has become a part of big data revolution as educational data meets the four main elements of big data including volume, velocity, variety and veracity. Large volume of educational data is collected through online learning platforms such as MOOCs and commercial e-learning products. This data is oftentimes generated real-time as students' activities happen on the e-learning platform and comes in different formats including clickstreams, text and comments, short answers, attendance, performance, etc. The uncertainty in educational data including biases, noise and abnormality necessitates a process in which the data is stored and mined meaningful to the problem being analyzed.

Educational data mining (EDM) is an interdisciplinary field entailing data mining, machine learning and statistics, and their application to education settings to transform teaching and learning. EDM has received much attention from researchers to derive insights from learners'activities and has been adopted by many institutions to improve the services they provide and for increasing student grades and retention.

Since LAK's primary focus is learning analytics, this tutorial will be beneficial for the majority of audience, providing accessible tools to perform predictive analytics and data mining on big educational data sets.

## 2. OBJECTIVES

Given the focus on practical skill-building, the primary objectives of the proposed tutorial are:

- Understand the basics of exploratory data analysis (EDA).
- Learn to use Python programming language.
- Learn how to build and validate predictive models.
- Learn the foundations of parallel computing for working with large datasets and large scale models.

While there are several conference series (e.g., LAK, EDM) focusing on the intersection between educational and computing research, to the best of our knowledge there have not been any tutorials covering all aspects of data analysis including EDA, predictive analytics, reporting results using appropriate statistical measures, and visualizing results. A subset of this tutorial was presented at ECTEL 15 (European Conference on Technology Enhanced Learning) as a part of a half day tutorial. The session was very well received by the audience and the recording is available online at http://educate.gast.it.uc3m.es/wapla/.

Most of the available tools for predictive analytics require enviable knowledge of technical details of underlying algorithms. However, there are also attempts to develop platforms for broader communities of researchers. Weka is a widely used data mining platform in java. Although easy to use, Weka's main packages are not optimized for memory intensive tasks on large data sets. Our goal in this tutorial is to make programming and mining big data in education using Python accessible for a wide range of audience. The Python programming language - with its predictive modeling and visualization libraries such as scikit-learn and Bokeh - provides all the necessary components for mining big data in an efficient and effective manner.

## 3. AGENDA

The preliminary tutorial agenda is as follow:

- Exploratory data analysis (45 minutes).
- Predictive modeling (60 minutes).
- Applying predictive models to educational datasets (60 minutes).
- Introduction to Spark (30 minutes).

## 4. TARGET GROUP

The tutorial appeals to researchers from a wide range of backgrounds including big data, predictive analytics, learning sciences, educational data mining and researchers interested in learning more about how analytics can transform teaching and learning. As a prerequisite, attendees are required to have familiarity with at least one programming language.

According to previous experience from presenting similar tutorial at ECTEL, we will be expecting an estimated number of 20 participants. We will have our colleagues from McGraw-Hill Education to help with the logistics and to help the participants with the installation and use of material.

## 5. FORMAT

Each of the four objectives mentioned above will be covered by a corresponding IPython notebook. Participants can download the data and notebooks before or during the session from our GitHub and stay engaged by running the notebook as we walk through them. At the end of each objective, we will form multiple groups and will assign a mini project to each group to solve. We will discuss the outcome of the exercises before moving to the next objective.

## 6. REFERENCES

[1] Christensen, C. M., Horn, M. B., Caldera, L., and Soares, L. Disrupting college: How disruptive innovation can deliver quality and affordability to postsecondary education. *Innosight Institute* (2011).

## 7. ORGANIZERS' BIOGRAPHY

Lalitha Agnihotri is a Senior Data Scientist at McGraw-Hill Education. Lalitha Agnihotri holds a Ph.D. from Columbia University in Computer Science and has over 15 years of experience in the Data Mining/Modeling area. She has authored over 40 peer reviewed conference and journal papers and has presented at several international conferences. She has applied a wide variety of learning algorithms to huge amounts of data to enable applications related to prediction of outcomes.

Shirin Mojarad is a Data Scientist at McGraw-Hill Education. She has wide experience in framing and conducting complex analyses and experiments using large datasets to find trends in diverse data sources and analyze behavioral patterns using advanced statistical modeling and data mining techniques. Shirin was formerly a senior analytics specialist in the Advanced Analytics team at the Canadian Imperial Bank of Commerce (CIBC) and prior to that, a data mining consultant with a leading software company in predictive analytics. She received her Ph.D. in Electrical Engineering and her M.Sc. in Communications and Signal Processing from Newcastle University U.K., where she specialized in predictive modeling and artificial neural networks.

Nicholas Lewkow is a Data Scientist at McGraw-Hill Education. He holds a PhD in computational astrophysics from the University of Connecticut. Additionally, he has conducted research in astrophysics and high performance computing at Oak Ridge National Laboratory and the Harvard-Smithsonian Center for Astrophysics. Nicholas is currently interested in big data analytics, parallel computing, and machine learning.

Alfred Essa is Vice President R&D and Analytics at McGraw-Hill Education. Previously he was Director of Analytics Research & Strategy at Desire2Learn, where he led product development of the Student Success System the acquisition of Degree Compass and the architecture of the joint analytics solution with IBM. He was Associate Vice Chancellor and Deputy CIO at Minnesota State Colleges & Universities where he led academic online strategy, enterprise infrastructure services, network security academic technologies and web development. Previously he was CIO at MIT's Sloan School of Management, where he won an MIT Excellence Award, was Principal Investigator of the iLearn project, and founded an Open Source project called dotLRN.