Certificate Program in Big Data Analytics & AI

(May 2018)

By
FORE and University of California, Riverside, Extension

Table of Contents

About Program

Certificate Program in Big Data Analytics and AI covers 4+1 complementary subjects. The subjects are as follows:

Subject	Total hours for the subject
1. Introductory Business Statistics	15
2. Data Mining and Data Analytics	107
3. Business Analytics Capstone (Python Oriented)	20
4. Web Analytics	8
5. Students Exercises/Projects	--

The subject of Data Mining and Data Analytics, in turn, is sub-divided into three distinct modules with a common theme of Analytics. Detailed content under each subject and module follows. We lay special and continued stress throughout the program on performing exercises on the part of students. Details about these are listed subsequently.

Data Mining & AI**	
a. Machine Learning algorithms (using R and Python)	51
b. Hadoop and Kafka Eco System; Data stream processing and analysis	26
c. NoSQL and Graph Databases	10
d. Deep learning, AI & NLP	20

** No of hours are indicative. Execution of projects by students being the focus, actual hours generally exceed than those specified.

Program Objectives

Applications of Big Data transcend disciplines. Use of predictive analytics pervades diverse disciplines as oil and gas, marketing and sales, sports, molecular biology, drug-designing, waste management, finance and the list is very long. Smart cities, for example, are the melting pot where variety of big data technologies mesh with one another to transform a city into a semi-intelligent being. In Marketing and Sales, for example, Big Data is fast emerging as a potent tool to gain deeper insights into Customer behavior and thereby act as a strong driver in spurring innovation. In manufacturing, operations managers are employing advanced analytics on historical process data to identify patterns and

relationships among discrete process steps and inputs, and then optimize the factors that prove to have the greatest effect on yield. Application of computer vision

Broadly the course has three parts: one the analytics part, second the technological part and third the contemporary Deep-Learning and Computer Vision. The analytics part is about learning machine learning algorithms and implementing them, the technological part is about learning to work in hadoop and apache kafka layered-system as also developing skills in NoSQL databases. Computer vision is an interdisciplinary field that deals with how computers can be made for gaining high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do. At the end of this course, given a large dataset from any domain, a participant should:

a. Be able to clean, transform and visualize the dataset to gain deeper insights and make it ready for analysis
b. Be able to select a subset of appropriate machine learning algorithms that could be applied to get the desired predictive results
c. Gain sufficient proficiency in tools necessary to implement algorithms
d. Put to use relevant tools and techniques to get a reasonable predictive accuracy
e. Apply the knowledge of image processing and image analysis to a wide array of disciplines such as health, process control, navigation and others.

Further:

f. Should be able to himself install, setup and configure and experiment with a complete hadoop and Kafka ecosystem
g. Should be able to install, configure and be sufficiently familiar with the variety of NoSQL databases and decide for himself which one to use, when and how

(g) and (g) are important objectives as they instill a sense of confidence in students in handling and experimenting with open-source technologies.

This course is project oriented: All tools, data and platforms including hadoop-ecosystem and Kafka-streaming technologies necessary for learning data-analytics are provided to the participants in advance. There is a heavy emphasis on open-source technologies universally used almost throughout the industry. Each participant, at the beginning of the course, receives two Virtual Machines (VMs) fully equipped with all the software platforms, tools, packages and data to work on. Assembling such VMs independently and by himself is also an important part of our education; students are able to work at ease with open-source technologies that are central to Analytics. We make the whole process very simple and stress-free. Details of the two Virtual Machines are more fully described below.

Complete Program is project based. We have experience with several Industrial projects. An e-book illustrating the projects executed by us can be downloaded from here. Students execute these and other projects while implementing techniques learnt and as part of weekly exercises.

Who should attend

Data being ubiquitous, the program cuts-across job or academic profiles. The techniques taught are generic in nature. These will be valuable to anyone who wishes to interpret data to advance

his/her knowledge and insights of environment. Specifically, the course will be useful to:

Executives

Ambitious Executives (from Private/Public sectors) looking forward to sharpening their skills in making sense of data in order to innovate and add more value to their organization and to society.

Academicians

Lecturers and Professors for extending the horizon of their knowledge through deepening their research skills.

Data Scientists/ Developers

Techniques taught to them will have applications in a broad array of disciplines.

Students/Research Scholars

IInd year students currently enrolled in Engineering / PGDM/ MBA or any graduate or post-graduate program who have had an introductory course in statistics. These students can look forward to better placement opportunities with added skill set.

Eligibility

A graduate in any discipline.

Subject wise details

1.

Introductory Business Statistics

Data Mining is intimately intertwined with Statistics. Knowledge of basic statistics is essential for a successful analyst. Descriptive statistics is invariably used to explore data. Concepts of inferential statistics are used in comparing machine learning models. In this subject, we refresh as also learn statistical fundamentals and essential inferential statistics. Concepts learnt here are reinforced when we use them in subsequent subjects throughout the duration of the program. Besides, as and when needed, we cover additional statistical concepts as they arise under different subjects.

S No	Subject	Session Hours
1	Measures of Central Tendency and Dispersion	1.5
2	Probability Theory (Different Approaches, Rules of Probability, Baye's Theorem)	2
3	Random Variables and Probability Distributions Discrete Probability Distributions - Binomial and Poisson Distribution	2
4	Continuous Probability Distributions – Normal Distribution	1.5
5	Correlation and Regression Analysis: Simple & Multiple Regression	2
6	Concept of Hypotheses Testing, Type I & Type II Errors, Power of The Test, Hypothesis Testing of Mean and Proportion, Two Sample Tests, Tests for Difference in Means and Proportions.	4
7	Chi-Square Goodness-of-Fit Test, Test of Independence	2
Total		15

Reference Book

1. Essentials of Statistics for Business and Economics by David R Anderson, Dennis J Sweeney and Thomas A Williams, Cengage Learning.
2. Statistics for Management by Richard Levin and Sanjay Rastogi; Pearson Publications

2.

Data Mining and Data Analytics

Subject Objectives

Data Mining is about knowledge discovery in databases--structured or unstructured-- searching through large volumes of raw data to find useful information-patterns. Data Scientists and decision makers can use this information for new sources of advantages and differentiation or for developing new business models. Broadly speaking the subject objectives are two-fold:

- Generate familiarity with Big Data, Data Visualization and Data Mining methods: In generating this familiarity there is special emphasis on conceptual understanding of techniques rather than on mathematics. Analytics is a creative process and students are encouraged to be creative.
- Develop skills to set up predictive models across various types of disparate data sets. This is intended to bring home the point that predictive analytics offers a generic set of tools that can be applied on different types of datasets within intersecting set of disciplines.
- Think differently: Expose students through projects as to how novel ways of applying Big Data technologies are changing business models.

Brief about Subject Contents

This subject is divided into four distinct modules. Module 2.1 is about Machine Learning Algorithms. In this module, we use variety of tools besides R and Python. Module 2.2 is about Hadoop and Kafka eco-system: we learn to work on hadoop and its layers; perform data extraction and pipe it into analytics engine. Analyzing streaming data is becoming a major subject in its own right: in this respect we experiment with Apache Kafka and related technologies. Module 2.3 relates to NoSQL and Graph databases. The new millennium and the explosion of web content has marked a new era for database management systems. A whole generation of new databases have emerged, all categorized under the name of NoSQL databases with focus on "task-oriented" database management system; selecting the right tool for the job depending upon its characteristics, nature and requirements. We cover, in depth, some often used NoSQL databases. Module 2.4 pertains to the exploding field of deep learning and AI. In this part we cover deep-learning technologies in depth as also techniques of Natural Language Processing.

Pedagogy

We strongly believe that a course in data analytics can only be practice-based rather than theory based. We also believe that a practice based course requires constant interaction with the teacher during lecture hours in real time. As it is a distance online course, the teaching pedagogy is like this: First the algorithm (or theory part) is conceptually explained without getting into mathematics and then a project is undertaken to implement the techniques. Datasets for implementation are made available in advance and so also a copy of code (or hints on it) that we need to execute. The code is numbered and copiously commented so that long after the lecture has finished, students can go back through the code/comments and refresh their knowledge. During the lecture, we execute this code (or prompt students to fill in the gaps), line-by-line and explain the steps. At his end, the student executes the required code on his laptop. Consequently, results are available at our end as also with the Students immediately. In short, both the teacher and students are working on their respective laptops simultaneously; students solve their problems and ask any questions to clarify. The whole experience is just as if everyone is sitting in a laboratory and working together.

Module details

Module 2.1: Machine Learning Algorithms (using R and Python)*

S No	Subject**	Projects/Datasets for projects***	Session hours*#
1	Developing familiarity with R; Data structures; Summarizing data; Data Exploration and transformation; integrating datasets; data & dates wrangling	1. UCI Repo: Bank Marketing Datasets **2.** Ta Feng Grocery Store dataset	5
2	Data Visualization and story-telling. Developing relationships between various features and plotting distributions	Kaggle: Sherbank Exploratory Analysis: Sherbank Russian Housing Market Kaggle Project: Visualizing Otto data set	4
3	Data Mining: Measures of Proximity; Cluster Analysis: Curse of Dimensionality; K-means clustering and Model based clustering	Kaggle Project: MNIST Handwritten digits dataset Kaggle Project: Clustering countries in World Happiness Report 2016	5
4	Text clustering and Agglomeration clustering;	Clustering Wikipedia Articles	5
5	Evaluation of clusters; Cluster Validation; Clustering tendency	Lecture	1
6	Classification Analysis: Decision tree Induction; Cross-validation, parameter tuning & grid search	Kaggle Project: Otto Group Product Classification	5
7	Techniques of Dimensionality Reduction: PCA and SVD (Singular Value Decomposition)	Kaggle: Statoil/C-CORE Iceberg Classifier Challenge	3
8	Neural Network	Telco churn dataset	5
9	Random Forest and Regression Trees; Determining feature importance with Boruta	Sentiment analysis of IMDB movie database. Kaggle Project: TFI: Predicting annual restaurant sales	4

S No	Subject**	Projects/Datasets for projects***	Session hours*#
10	Gradient Boosting Technique for Machine Learning & grid search of its parameters	Kaggle Project: Predict a biological response of molecules from their chemical properties	2
11	Evaluating Classification: ROC, AUC, Precision, Recall, Specificity, Sensitivity; kappa metric; Overfitting; Bias-variance trade-off; L1 & L2 regularization	Lecture	3
12	Ensemble modeling: A review of variety of techniques; Balancing datasets	Education Analytics: Student Drop-outs from schools	2
13	eXtreme Gradient Boosting (XGBoost)	Kaggle Project: Predicting RedHat Business Value	4
14	LightGBM: Light Gradient Boosting Machine	Kaggle Project: Identify potential purchasers of caravan insurance policies	3
Total			51

* Extra classes, beyond normal schedule, will be held to introduce students to Python. These classes will be scheduled on weekends.
**Teaching sequence may alter somewhat depending upon feedback from students
***Datasets other than those mentioned here may also be introduced during classes to achieve better clarity. Datasets needed for Kaggle projects are to be downloaded from their site even though freely available; this is as per site requirements.
***#** No of session hours are indicative. Execution of projects by students being the focus, actual hours generally exceed than those specified. Students need to be prepared for extension of classes beyond specified hours.

Module 2.2: Hadoop and Kafka Eco System; Processing streaming data and analysis

S No	Subject**	Projects/Datasets for projects***	Session hours*#
1	Introduction to Hadoop and its ecosystem	Lecture & Tutorial	2
	Hadoop file storage formats		
2	Linux and Hadoop shell commands	Tutorial	2
3	Hadoop streaming	Using awk, sed and other utilities to execute map-reduce jobs to manipulate data in Hadoop	2
4	Hive on Tez and hadoop	Book-crossing dataset	3
5	Pig on Tez and hadoop	NYSE dividends and S&P Index data	2
6	Pyspark and SparkSQL: Data storage and Extraction with SQL; Executing ML algorithms (including grid-search)) using MLlib and ML libraries	Analyse Adult Data set, Breast cancer dataset and Forest Cover dataset to make target predictions.	6
7	Recommender Engine using Mahout on hadoop	Recommender Engine with MovieLens Dataset	2
8	Installation of Hadoop ecosystem	Complete installation of Bigtop hadoop system on a Virtual Machine and testing of hive & pig.	2
9	Apache Kafka: Stream data processing	Experimenting with Apache Kafka system; Develop custom message producers and consumers with hands-on exercises; coupling Apache Kafka with SparkR as also with Apache Samza	5
Total			26

**Teaching sequence may alter somewhat depending upon feedback from students

***Datasets other than those mentioned here may also be introduced during classes to achieve better clarity. Datasets needed for Kaggle projects are to be downloaded from their site even though freely available; this is as per site requirements.

***#** No of session hours are indicative. Execution of projects by students being the focus, actual hours generally exceed than those specified. Students need to be prepared for extension of classes beyond specified hours.

Module 2.3: NoSQL and Graph Databases

S No	Subject	Brief topics covered	Session hours*
1	Introduction to NoSQL Databases and CAP theorem; Comparison with RDBMS	Lecture	2
2	Redis in-memory data structure store	Installation of redis. Data structures in redis; Multi key queries; Publication and Subscriptions. Use cases	1
3	MongoDB Document Database	Installation of mongoDB; Use cases; Data storage and CRUD operations; Access controls; Sharding operations; Data import and export;	3
4	Hbase column family database on hadoop	Installation of hbase; Exploratory exercises	1
5	Neo4j Graph Database	Installation of neo4j. Use cases. Graph database concepts; Creating graphs and querying graph databases; importing and modeling a relational database into neo4j	3
Total			10

* No of session hours are indicative. Execution of projects by students being the focus, actual hours generally exceed than those specified.

*Module 2.4: **Deep learning, NLP & AI***

S No	Subject	Brief topics covered	Session hours*
1	Autoencoders and anomaly detection	Recognizing similar Olivetti faces	3

2	Deep Learning with Convolution Neural Network	Building Powerful Image classifiers with very less images	3
3	Using very Deep Convolution networks and Data Augmentation	Experimenting with VGG16 and data augmentation techniques	2
4	Transfer Learning	ResNet50: Working with ResNet50: Kaggle Invasive Species Prediction using multiple images	2
5	Generative-Adversarial Networks (GAN)	GAN working and experiment on self-generated dataset	3
6	Recurrent Neural Networks & LSTM	Sequence classification and Sentiment analysis of tweets on Twitter	3
7	Natural Language Processing & Word2Vec transformation	Analyzing Social Media dataset and classifying emotions	4
Total			20

* No of session hours are indicative. Execution of projects by students being the focus, actual hours generally exceed than those specified.

Virtual Machines for course participants

At the commencement of the course, each participant is given two virtual machines (VMs):

a) Virtual Machine with complete Hadoop-eco system as also NoSQL databases
b) Virtual Machine to experiment in Deep Learning & AI

Both VMs are installable on Windows/Mac/Linux systems with 4GB of RAM. They can be installed on Laptop or desktop. Both Virtual Machines contain all the software tools that the participant will work on. They also contain plenty of data to experiment with along with reading materials. Every software installed on VM is fully licensed. Virtual machines make it easy for participants to learn as also practice weekly exercises at home/workplace.

Applications installed on the Hadoop VM are as follows:

- R and Python: R (with more than 200 packages pre-loaded); RStudio Server; Vowpal Wabbit (both as R package and as a binary).
- Hadoop eco-system: Hadoop; Yarn Resource manager; Hive/ hiveserver2; Pig; Apache SparkR; sparklyR,; PySpark, Mahout; Hbase; Hue; Apache Drill and Apache Phoenix
- Apache Kafka and Apache Samza
- Visual Frameworks: H2o; KNIME; Orange; Gephi (for social network analyses)
- NoSQL Databases: Redis, MongoDB, Hbase and Neo4j

We may mention that besides this virtual machine, we have a separate Hadoop-cluster of ten machines with Cloudera server installed (with around 120GB RAM). This large cluster helps participants to work in groups remotely.

Applications installed on the Deep Learning & AI, VM are as follows:

- Anaconda 5
- OpenCV
- Theano
- Tensorflow

Reference Materials:

Reading material for each Module is placed on e-learning site and also study material is sent by mail.

****

3.

Business Analytics Capstone (Python Oriented)

Introduction:

Python is fast emerging as a preferred tool of data science for many analysts. It is often praised for its easy-to-understand syntax. Like R, python is also open source with several highly developed Integrated Development Environments (IDEs) that make learning python a fun. We use Anaconda that comes both with a set of powerful packages (distribution) as also two well-known IDEs. Our approach here is to use python for data cleaning and transformation, visualization, and model development. Participant will find a great degree of similarity in data manipulation using R and using python. This similarity makes learning python easier. This subject builds upon our knowledge of modeling techniques learnt earlier.

Learning outcomes and course objectives

By the end of the class students will be familiar with the underlying statistics for data science; how to approach different data types, appropriate analysis and visualization methods and presenting analysis results and finally build high fidelity models using various techniques in Python language.

Assignments

There will be hands on homework and project assignments to give students an opportunity to apply what they learn in the class.

S No	Subject***	Projects/Datasets for projects**	Sessions hours*
1	Introduction to python; Using iPython; Basic data types and data structures in python and pandas; Loops and Conditionals in python;	Tutorial	**2**
2	Exploring data with pandas—Quick Start	UCI Repo: Adult Dataset	**2**
3	Numpy: Arrays; Basic arrays operations; Comparison operators and value testing for arrays; Array item selection and manipulation;	Tutorial	**2**
4	Data Visualization in python; Data Visualization using t-distributed stochastic neighbor embedding (t-sne)	Insurance Dataset	**2**
5	k-means clustering with scikit-learn	UCI Repo: Car Evaluation Dataset/ Tamilnadu Electricity Board Hourly Readings Data Set	**2**
6.	Decision trees classifier	UCI Repo: German Credit Dataset	**2**
7	Ensemble Modeling	Kaggle Project: Forest Cover type Prediction	**2**

8	Logistic Regression (along with Dimensionality Reduction, PCA)	MNIST Digits dataset	**2**
9	Support Vector Machines	Marketing campaign dataset in retail	3
9.	Introduction to Keras on Tensorflow	CIFAR10 small image classification	**2**
Total			**20**

* No of session hours are indicative. Execution of projects by students being the focus, actual hours generally exceed than those specified.

**Datasets other than those mentioned here may also be introduced during classes to achieve better clarity. Datasets needed for Kaggle projects are to be downloaded from their site even though freely available; this is as per site requirements.

***Teaching sequence may alter somewhat depending upon feedback from students

Python Resources

1. Online Book--Automate the Boring Stuff with Python: Great intro to python as a programming language with links to worked out examples and links to videos: .https://automatetheboringstuff.com
2. The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, Jerome Friedman; https://web.stanford.edu/~hastie/ElemStatLearn/
3. Python.org has a lot of useful information along with the home pages for numpy and scipy as introduced in the notebooks discussed in the class.(numpy: https://github.com/numpy/numpy and scipy: https://www.scipy.org)

****

4

Web Analytics

Introduction:

Successfully business brands today require a well-balanced blend of art and science. This course introduces students to the science of web analytics, while casting a keen eye toward the artful use of numbers found in the digital space. The goal is to provide marketers with the foundation needed to apply data analytics to real-world challenges they confront daily in their professional lives. Students will learn to identify the web analytics tools right for their specific needs, understand valid and reliable ways to collect, analyze, and visualize data from the web, and utilize data in decision making for their agencies, organizations, or clients.

Objectives:

- Gain an understanding of the motivations behind data collection and analysis methods used by business professionals.
- Learn to evaluate and choose appropriate web analytics tools and techniques
- Understand frameworks and approaches to measuring consumers' digital actions.
- Gain an understanding of a step-by-step approach to planning, collecting, analyzing, and reporting data
- Utilize tools to collect data using today's most important online techniques: performing bulk downloads, tapping APIs, and scraping webpages
- To understand business analytics practices in digital world

Pedagogy:

- Lectures,
- live project,
- hands on sessions

Session Plan:

The course will consist of the following three broad modules

Session No.	Session Theme	Reading/Cases	Session hours
1	Basics of Web analytics	• Best Web Metrics/KPIs for a Small, Medium or Large Sized Business	2.5
2	Analytic techniques and Tools: Google trends, Google Website optimizer, Google Analytics, Google Tag manager	• The Consumer Decision Journey • Best Social Media Metrics: Conversation, Amplification, Applause, Economic Value • Building Brands with Social Media	2.5
3	Data Analysis and Data Visualization	• Nothing is perfect ... especially data: Data quality sucks, let's just get over it • Tidying data for analysis: Tidy data	3
Total			8

Reference Book

Eric Peterson, *Web Analytics Demystified*, 2004 (available for free download from Web Analytics Demystified) and link also available on e-learning site.

****

5

Students Exercises/Projects

Introduction:

The ultimate beneficiary of this program are students. Our experience shows that students learn faster, if they attempt exercises, make mistakes and learn from them. Students are, therefore, expected to undertake exercises and projects.

Exercises serve another purpose. We try to make students learn some of the important topics not possible to cover in the class. We give exercises with sufficient hints to attempt them.

There is also a third advantage. The more students perform exercises, the more a teacher can move faster and also cover advanced concepts.

An illustrative list of projects, topic-wise, is given below. To assist students, for each project we provide sufficient steps/code on our e-learning site. For each project, our steps/codes are quite detailed and students should be able to execute the projects on their own by following the listed steps (or at times by stealing a glance at our code).

Students will be assessed based upon their performance in Exercises and Projects.

S No	Subject	Projects/Datasets for projects
1	Data Visualization and story-telling.	Used cars dataset Analyze Human Development Report
3	K-means clustering Model based clustering	UCI Machine Learning Repository: Whole sale Customers Dataset Wiki Text Clustering
4.	Dimensionality reduction and t-sne visualization	Kaggle: Visualize Gender Voice dataset
5	Decision trees Induction	Kaggle Project: Airbnb, New Users Booking Wiki Text Classification
6	K-Nearest Neighbour	Lower back pain classification
7	Neural Network	Kaggle Project: Predict a biological response of molecules from their chemical properties
8	Naïve Bayes Modeling	UCI Repo: Breast Cancer Data
9	Random Forest	Kaggle Project: Rossmann Drug Store; Forecast sales using store, promotion, and competitor data.

10	Feature plotting	Feature plotting of Credit Card Fraud dataset
11	eXtreme Gradient Boosting (XGBoost)	Kaggle Project: Springleaf; Determine whether to send a direct mail piece to a customer
12	Support Vector Machines	Kaggle Project: Predict Grant Applications results
13	Regression trees	Kaggle Project: How do Housing features add up to its prices
14	Apache Pig Exercises	Drivers event and miles-logged dataset
15	Analyse data on Spark/PySpark	Predict type of Forest Cover
16	mongoDB Exercises	Primer-dataset (json) of restaurants collection
17	Deep-Learning: Autoencoder	Kaggle Project: Find structure in data: Predict backorder risks for the product with historical data
18	Deep Learning	Differentiate images of urban and non-urban areas

****