



भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

Course Starting: 7 August, 2016

Application Deadline: 15 June, 2016

Big Data Analytics (Batch 1)





Certificate Programme on Big Data Analytics

BATCH 1 (2016-17)

In God We Trust,
All Others Must
Bring Data

W Edwards Deming

The discipline and practice of management is rapidly becoming evidence-based. Consider the area of marketing, which once idolised freewheeling thinkers, elevating them to the status of visionaries for their wild, imaginative inputs. Although these inputs are helpful, functions such as advertising and sales today have started to become software-assisted. Cold, calculative reasoning is required to justify any business decision. In this digital world, there is little utility in seeking validation via sound bytes from the person on the street. Want to know how well a disruptive ad campaign is faring? Gather what thousands of fans say about it on social networks, and apply a host of analytical techniques.

The role of analytics in solving business problems has increased manifold in recent years, causing a spike in the demand for trained professionals. In his book titled, “Competing on Analytics: The New Science of Winning”, Thomas Davenport claims that a significant proportion of high-performance companies employ personnel with high analytical skills. Another study reveals that close to 60% of organisations do not possess the requisite information for decision making. Tackling this issue warrants company-wide

effort to collect data from different functions, and versatile forms of data storage that can support querying in real time. Finally, the company must have professionals on board, who can work with as well as enhance software that helps them make sense of the results.

In an article¹ based on a survey of nearly 3000 executives, MIT Sloan Management Review reported that there is a striking correlation between a firm’s analytical sophistication and its competitive performance. The biggest obstacle to adopting big data analytics is the lack of knowhow among personnel, whose decisions impact business performance. Big data analytics draws from a diverse mix of statistics and operations research, machine learning, deep learning, algorithm design, and systems engineering. This programme is designed to provide an in-depth knowledge of big data techniques, and their applications in improving business processes and decision making.

Big Data Analytics

A triad of terms captures the essence of “big data”: volume, velocity and variety. The volume and pace at which data is created can challenge existing computing infrastructure. For example, every flight of a Boeing 777 can generate up to 1 terabyte (~1000 gigabytes) of data. Making sense of this data is imperative for decision making and troubleshooting. The theory of bounded rationality proposed by Nobel Laureate Herbert Simon is evermore significant today with the increased complexity of business problems; the human mind is constrained in its capacity to evaluate alternatives, given limited time to make conclusions.

Organisations large and small are forced to grapple with problems of big data, which challenge the existing tenets of data science and computing technologies. Techniques in predictive analytics rely

¹ M S Hopkins, S LaValle, F Balboni, N Kruschwitz and R Shockley, “10 Insights: A First Look at The New Intelligence Enterprise Survey on Winning with Data”, MIT Sloan Management Review, Vol. 52, No. 1, 21–31, 2010.

heavily on the validity of statistical concepts such as independent and identically distributed (IID) random variables and the central limit theorem (CLT). When dealing with big data, the validity of these assumptions becomes questionable. Straightforward tasks such as interpreting descriptive statistics have their share of issues. We begin to question the utility of summary measures and diagrams.

Algorithms that work well on “small” datasets crumble when the size of the data extends into the gigabytes. Time series techniques must be revamped to handle streaming data in continuous time. Social media messages have data formats that are unfit to be represented by traditional databases. While these may appear to be difficult problems, there has been a tremendous progress in big data analytics. For example, columnar databases have significantly boosted query speeds. File systems can seamlessly distribute datasets on multiple hard drives, and facilitate analytics on them in real time. Finally, the free and open source nature of several big data platforms promotes rapid adoption.

Program Objectives

This program is designed to equip its participants with an in-depth knowledge of Big Data Analytics (BDA). We will use real case studies to illustrate the applications of key concepts. At the end of the course, the participants will be able to:

- Appreciate the emergence of business analytics and big data as a competitive strategy.
- Analyse datasets by applying techniques from statistics, operations research, machine learning, deep learning, network analysis and data mining.
- Process unstructured data such as social media messages and machine generated clickstream logs.
- Have a working knowledge of languages, platforms and tools that support statistical analysis and visualisation (R/Python), distributed computing (Hadoop/Spark) and network analysis (Gephi).
- Apply the theories, techniques and tools to solve problems from a wide variety of industries such as manufacturing, services, retail, software, banking and finance, sports, pharmaceuticals, and aerospace.

Prior Preparation

Given the ambitious nature of the programme objectives, participants are required to possess a basic understanding of statistics. To help them understand and assimilate the concepts over a comfortable duration, we recommend that participants go through IIMB's MOOCs on edX.org titled Statistics for Business – I and II.² These open courses cover topics in single variable statistics having to do with data analysis, visualisation, probability, Bayes rule, decision making with conditional probability, random variables, binomial, Poisson and normal distributions, and Monte Carlo simulation. The courses supply abundant examples from the industry, and also provide illustrative tutorials to some of the tools we shall be using: spreadsheets and R.

Make no mistake: the programme will be relatively fast paced to help us achieve the goals that we have set. We shall skip covering the above material in the classroom. Instead, the forums on the MOOCs will help address any concern you may have.

Programme Design

The course consists of six modules and a project. All modules are application oriented, without compromising on theoretical aspects. The following sections go over the individual modules and their contents.

Case-based teaching shall be employed wherever possible, across all modules, with cases studies sourced from IIMB, Harvard Business School (HBS), Darden, Ivey, and Kellogg.

The programme shall have a special focus on business analytics as practised in India. A significant proportion of the cases that we use in the programme have been published by IIMB faculty on the Harvard Business Publishing site. A few of them are published by the alumni of other business analytics programmes at IIMB, based on their project work.



Module 1: Foundations of Data Science (5 days)

The process of fact-based decision making requires managers to know how to summarise, analyse, and interpret data, as well as to communicate the results using data visualisation. We shall briefly dwell on univariate analysis, before proceeding to describing and analysing multivariate data. Some of the techniques that we introduce here apply to “small” datasets: they will have to be suitably modified to handle large data volumes.

Contents

- Brief introduction to R. Working with data from different sources: spreadsheets, databases, and the cloud. Data visualisation
- Descriptive statistics. Bayes theorem and applications to decision making.
- Two approaches to inference: estimation and hypothesis testing.
- Machine learning: mapping business problems to methods. Evaluating models with RMSE, precision, recall, sensitivity and specificity. Validating models, problem of overfitting.
- Decision trees, k-nearest neighbour (KNN) method, Naïve Bayes classifier
- Unsupervised methods: cluster analysis, association rule mining
- Presentation of results: knitr

Case Studies:

1. Central Parking Solutions Private Limited (IIMB Case); 2. A Dean's Dilemma: To Admit or Not to Admit (IIMB Case) 3. Analytics in HR – Predicting Job Acceptance (IIMB Case)

Module 2: Big Data Software Tools (2 days)

Advanced R and Python. The objective is to familiarise the participants with the tools we will use throughout the programme.

Module 3: Predictive Analytics (6 days)

Predictive analytics models predict the occurrence of future events such as customer churn, default in loan repayment, etc. based on historical data. In many business problems, we deal with data on several variables, sometime more than the number of observations. Regression models help us understand the relationships among these variables, and how these relationships can be exploited to make decisions. The primary objective of this module is to understand how regression and causal forecasting models can be used to analyse real-life business problems such as prediction, classification and discrete choice problems.

The focus will be on case-based practical problem-solving using predictive analytics techniques to interpret model outputs. The participants will be exposed to software tools such as MS Excel, R, SPSS, and SAS and how to use these software tools to perform regression, logistic regression and forecasting.

Contents

- Regression model building framework: Problem definition, Data pre-processing; Model building; Diagnostics and validation
- Simple linear regression: Coefficient of determination, Significance tests, Residual analysis, Confidence and Prediction intervals
- Multiple linear regression: Coefficient of multiple coefficient of determination, Interpretation of regression coefficients, Categorical variables, heteroscedasticity, Multi-collinearity, outliers,

² <https://www.edx.org/course/statistics-business-i-iimb-qm101-1x>

- Autoregression and Transformation of variables, Regression model building
- Logistic and multinomial regression: Logistic function, Estimation of probability using logistic regression, Deviance, Wald test, Hosmer Lemshow test, Classification table, Gini co-efficient.
- Forecasting: Moving average, Exponential smoothing, Casual models
- Application of predictive analytics in retail, direct marketing, health care, financial services, insurance, supply chain, etc.

Case Studies:

Pricing of Players in the Indian Premier League (IIMB Case), Package Pricing at Mission Hospital (IIMB Case), Colonial Broadcasting Company (HBS Case), Pedigree vs Grit: Predicting Mutual Fund Manager Performance (Kellogg Case), Breaking Barriers – Micro-Mortgage Analytics (IIMB Case), A Game of Two Halves: In-Play Betting in Football (IIMB Case); HR Analytics – Predicting Probability of Renege (IIMB Case), Predicting Demand for Food at Apollo Hospital (IIMB Case)

Module 4: Big Data Eco-System (3 days)

We are ready to scale up the solution techniques with the help of big data platforms such as Hadoop, Spark and Big Query. The second part of this module examines the network approach to data representation and analysis, along with tools to derive important measures of centrality such as PageRank. We address the variety aspect of big data by analysing unstructured social media messages and clickstream data.

Contents

- Big data: changes in approach and analysis
- Cloud computing platforms for big data analysis, using Google Cloud, BigQuery, Amazon AWS and Microsoft Azure
- Hadoop ecosystem: Storing and managing large-scale structured and unstructured data

- Large scale data ingestion using Sqoop, Flume, NoSQL and Kafka
- Apache Spark with interfaces in Python and R
- Gephi tool, R libraries for graph processing
- MLIB (Machine Learning Library): How to use APIs to develop linear models, Decision trees, Random forests, Gradient boosted trees and Collaborative filtering

Module 5: Machine Learning for Prediction (3 days)

This module introduces the participant to machine learning algorithms such as recommender systems, clustering, text analytics, spatio-temporal analysis, association rule mining and Monte Carlo simulation.

Contents

- Nonlinear dimension reduction
- High-dimensional clustering
- Advanced recommender system
- Tree-based methods—Decision tree, Bagging, Random forest, boosting, stochastic gradient boosting
- Support vector machine, Meta-learning, Multi-task SVM

Module 6: Advanced Analytics for Big Data (5 days)

Big-data is defined using the volume of the data, the velocity at which the data is created, and the variety in the data. Sources of big-data include social networks, telecom and mobile services, healthcare and public systems and machine generated data. In this module, we introduce the participant to big data technologies, and explore the challenges.

Contents

- Advanced regression and classification, Zero-inflated regression, Dynamic logistic regression, Dynamic linear model, Multinomial logistic regression
- Bayesian regression; Bayesian optimization
- Artificial neural networks and deep learning
- Advanced modelling techniques: Bag of little bootstrap, variational Bayes, large-scale logistic regression, ADMM, parallel matrix factorisation.
- Graphical models for analytics

Course Evaluation

The participants will be evaluated through take-home assignments and a project work. At the end of each module, the participants will be given a take-home assignment that should be completed and submitted within 4 weeks.

Course Project

Each participant should carry out an individual project for 4 months based on a real-life problem/dataset. IIMB encourages students to publish cases studies based on their course project.

Who Should Attend?

In October 2012, Harvard Business Review claimed that the position of a “Data Scientist” will be the sexiest job of the 21st century. This certificate programme will equip the participants with a large suite of analytical tools, as well as prepare them for corporate roles in analytics bases consulting in marketing, operations, supply chain management, finance, insurance and general management in various industries. The course is suitable for those who are already working in analytics, and wish to enhance their knowledge. We also welcome participants with a strong analytical aptitude, who would like to start their career in analytics.



Why IIMB?

IIMB is one of the first institutes in the world to offer Business Analytics certificate course since 2010. IIMB faculty members have published more than 20 analytics case studies at the Harvard Business Publishing, a record for any Indian institute. Students admitted to this group come from different organizations and different geographical locations across the world making it a great learning experience for students. Students get a chance to work on real world problems as part of the course.

IIM Bangalore has invested heavily in building big data processing capabilities. Participants shall be working with a powerful cluster of 12 nodes on the campus. While we shall explore the cloud route to big data analysis, we shall be able to explain to the participants, aspects of the technologies required to run big data computations.

Eligibility Criteria and Selection Process:

The participants should have a Bachelor degree in engineering/science/commerce or arts with mathematics as one of the subjects during their Bachelor's programme. Preferable work experience is 3 years; in exceptional cases, applicants with less than 3 years are admitted into the programme. It is essential that the applicants have programming knowledge.

Selection Process:

Candidates will be shortlisted for test/interview based on their past academic performance, quality of work experience and fit for analytics job. Interviews will be conducted in July 2016.

Tentative Course schedule:

The course schedule is as follows:

Programme Schedule			
Module		Dates	Venue
1	Foundations of Data Science	7-12 August 2016	IIM Bangalore
2	Big Data Software Tools	8–9 September 2016	IIM Bangalore
3	Predictive Analytics	17-21 October 2016	IIM Bangalore
4	Big Data Eco-System	23–25 November 2016	IIM Bangalore
5	Machine Learning	12-14 January 2017	IIM Bangalore
6	Advanced Analytics for Big Data	17-18 February, 2-4 March 2017	IIM Bangalore

Class timings: 9 am to 5:15 pm (IST)

Project:

Students are expected to do a live project as part of this course. The project report should be submitted by 15 May 2017. The participants have to submit the project proposal by 31 January 2017. The projects will be supervised by an IIMB faculty member.



Program Directors:

Professors U Dinesh Kumar, Shankar Venkatagiri and Pulak Ghosh

For any queries contact: Professor Shankar Venkatagiri (Email: shankar@iimb.ernet.in)



Programme Delivery

The programme will be conducted live in the classroom at IIMB.

Programme Fee

Rs.4,00,000/- + service tax (as applicable) per participant. The fee is payable in three installments as per indicated schedule. The payment schedule is as follows:

Rs.2,25,000/- + service tax	I installment on admission
Rs.1,75,000/- + service tax	II installment on or before 22nd Nov 2016

Award of Certificate

A certificate of completion will be awarded by IIMB to the participants at the end of the programme, upon successful completion of the modules and satisfying all the programme requirements.

Alumni

On successful completion of the programme, the participants are eligible to be admitted to IIM Bangalore Alumni Association on a one-time payment of Rs. 3000/-.

Important Dates

Application Deadline:	15 June 2016
Interview of Shortlisted Candidates:	4th week of June
Announcement of Decision:	5 July 2016
Course Commencement:	August 2016



The Indian Institute of Management Bangalore (IIMB) has been ranked for the eighth successive year as the No. 1 Business School in Central Asia by Eduniversal, a French Consultancy Group. IIMB has obtained the European Quality Improvement System (EQUIS) accreditation awarded by the European Foundation for Management Development (EFMD). Established in 1973, Indian Institute of Management Bangalore today offers a range of post-graduate and doctoral level courses as well as executive education programmes. With a faculty body from amongst the best universities worldwide, Indian Institute of Management Bangalore is fast emerging as a leader in the area of management research, education and consulting.

IIMB's distinctive feature is its strong focus on leadership and entrepreneurial skills that are necessary to succeed in today's dynamic business environment.

IIMB has been ranked among the Top-50 global schools by the Financial Times Executive Education Rankings 2015. Indian Institute of Management Bangalore is the only B-school from India to figure in this elite list.



Registration

The organizations interested in nominating their employees and individuals interested in the programme may apply online.

The Administrative Officer

Executive Education Programmes
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore 560 076
Phone: +91 - 80 - 26993817, 26993475
Fax: +91 - 80 - 2658 4004
Email: openpro@iimb.ernet.in
Web: www.iimb.ernet.in/eep

Participants interested in the programme may contact IIMB at the above-mentioned address for clarifications, if any. Once registration is accepted, cancellation / refund queries and requests will not be entertained.



INDIAN INSTITUTE OF MANAGEMENT BANGALORE
Bannerghatta Road, Bangalore 560 076