FORE School of Management, New Delhi
Big Data and Data Analytics

Data Mining and Data Analytics
Course details

Learning outcomes

This course is about knowledge discovery in databases; searching through large volumes of raw data to find useful information-patterns that are implicitly embedded in the raw data. Data Scientists and decision makers can use this information for new sources of advantages and differentiation or for developing new business models. In layman's terms, devoid of predictive analytics terminology, some of the tasks that a student will be able to perform after course completion will be:

- Segment customers. (For example, segment customers by their purchase habits)
- Identify groups with similar characteristics. (For example, motor insurance policy holders with similar risk profiles)
- Target marketing (predict if a customer is likely to remain loyal or which customers can turn into loyal customers)
- What is the best product to upsell to a customer after they purchase a product
- What visit volume be expected on the website next week
- What is the estimated customer lifetime (CLV) value of each customer
- Predicting customer churn in a telecommunication company

Course Objectives

Broadly speaking the program's 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 build 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.

Brief Course Contents

This course is divided into three distinct modules. Module 1 is about Data Science and Data Mining, Module 2 is about Projects on Industrial data that students will execute. Module 3 is regarding Hadoop eco-system based analytics. Module 1 and Module 2 progress in parallel while Module 3 follows after completion of Module 1 and 2. Each participant, at the beginning of the course, receives a Virtual Machine fully equipped with all the software tools, packages and data to work on. Modules and Virtual Machine are more fully described below.

Module 1: Data Science and Data Mining

Introduction to Big Data. Data Exploration and data cleaning using RStudio and python.

Data Visualization and story-telling with RStudio, python and Tableau/MS Power BI. Data pre-processing, integration and data transformations. Business Forecasting: Exponential smoothing, Holt-winters smoothing, ARIMA and Hybrid approaches. Data Mining: Measures of Proximity; Cluster Analysis: K-means and Agglomeration clustering; Density based clustering methods; Kohonen self-organizing maps.

Classification Analysis: Decision Trees induction; Neural Network; Support Vector Machines; Naïve Bayes Method; Ensemble Methods (Random Forest, Gradient Boosting Machines, Extreme Gradient Boosting); Auto-encoders and Deep Learning. Model tuning: Cross validations and Grid Search; Evaluating classification results: ROC, AUC, Precision, Specificity and kappa metric.

Module 2: Projects in Machine Learning

This is the projects module. We have experience with a number of Industrial projects. An e-book illustrating the projects executed by us can be downloaded from here. Students execute these projects while implementing techniques learnt and also as part of weekly exercises. Ta Feng Grocery Store: Segmenting customers; Kaggle Project: Airbnb What would be the next destination of a visitor; Kaggle Project: Otto Group, Product Classification Challenge; Kaggle Project: Rossmann Sales Revenue Prediction; Kaggle Project: Predict TFI Restaurant Sales; Kaggle Project: Determine whether to send a direct mail piece to a customer; Kaggle project: Predict handwritten digits; Kaggle Project: Santander Customer Satisfaction; Kaggle Project: Predict Biological Response to drug molecules; Kaggle Project: Housing Prices, Advanced Regression Techniques.

Module 3: Hadoop based Analytics

Hadoop eco-system: Developing familiarity with Hadoop filesystem/Hadoop-cluster; using Hadoop shell and Hue; Data extraction with Hive, Pig, Tez and Impala using SQL; Developing analytical models (such as Recommender Systems and Logistic Regression) with SparkR, H2o, and Mahout over Hadoop. Frequent Itemset Mining and data exploration and developing predictive analytical models on BigML Cloud. NoSQL databases.

Virtual Machine for course participants

At the commencement of the course, each participant is given a virtual machine that is installable on Windows/Mac/Linux systems with 4GB of RAM. It can be installed on Laptop or desktop. The Virtual Machine contains all the software tools that the participant will work on and also data. Every software installed is fully licensed. The virtual machine makes it easy for participants to practice weekly exercises at home/workplace. Applications installed on the VM are as follows:

- R and Python: R (with more than 200 packages pre-loaded); RStudio Server; Anaconda for python (ipython and Spyder); Vowpal Wabbit (both in R and as a binary).
- Hadoop eco-system: Hadoop; Yarn Resource manager; Hive/ hiveserver2; Pig; Apache SparkR; Mahout; Hbase; Hue; Apache Drill
- Visual Frameworks: H2o; KNIME; Orange; Gephi (for social network analyses)
- NoSQL Databases: MongoDB and Hbase

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 comfortably in the Computer lab.

****