DATA SCIENCE USING R AND PYTHON

GOOGLE'S SELF-DRIVING CARS AND ROBOTS GET A LOT OF PRESS, BUT THE COMPANY'S REAL FUTURE IS IN MACHINE LEARNING, THE TECHNOLOGY THAT ENABLES COMPUTERS TO GET SMARTER AND MORE PERSONAL.

– ERIC SCHMIDT (GOOGLE CHAIRMAN)

This course is intended to give a holistic understanding on statistical & machine learning and its application using Python and R. The workshop will cover

- An introduction to business analytics
- An introduction to Python for data analysis
- An introduction to R for data analysis
- An introduction to supervised machine learning algorithms
- An introduction to unsupervised machine learning algorithms
- Understanding the core of machine learning – Gradient Descent Algorithm
- Understanding of various sampling strategies and its efficacy in learning process
- An introduction to ensemble methods for handling imbalanced data
- Hands-on using the Python code and R Code on the real life dataset

DATA SCIENCE USING R AND PYTHON

OBJECTIVE

We are living in an era where computing moved from mainframes to personal computers to cloud. And while it happened, we started generating humongous amount of data. However the multi-folds increase in computing power also brought in advancement in application of algorithms which can be used to get insights from huge amount of data being generated. In this course, you will learn to nuances of building supervised and unsupervised machine learning models on real life datasets. We'll introduce you to Anaconda framework, Python and R kernel through Jupyter Notebook for applying some of the statistical and machine learning algorithms which will become handy in solving challenging problems.

At the end of the course you will develop a clear understanding of the need of machine learning algorithms and the context in which to apply these algorithms to solve complex problems from the field of business.

HARDWARE AND SOFTWARE

1. Participants should bring their laptop (preferably Windows 7 or higher/ Mac OS installed).
2. Operating System (any of the following):
   - Mac OS X with XQuartz
   - Windows (Version XP or later) is required.
3. Minimum 8 GB RAM on the system is advisable.

INSTALLATIONS:

- For Windows, go to https://docs.anaconda.com/anaconda/install/windows.html
- For MacOS, go to https://docs.anaconda.com/anaconda/install/mac-os
- For Linux, go to https://docs.anaconda.com/anaconda/install/linux

More about anaconda can be found at https://docs.anaconda.com. Participants are expected to resolve any installation issues of the software prior to the commencement of the session.

PRE-REQUISITE

1. Participants should have basic programming skills and should be in able to understand the scripting language.
2. High speed internet connection will be provided at the training venue.

COURSE DELIVERABLE

1. Python code, R Code and dataset
2. Soft copy of the content being covered (PDF file)

COURSE OUTLINE

Day 1: Understanding Anaconda Framework platform and other useful packages in Python

Session 1 & 2–Introduction to Business Analytics

- What is Business Analytics
- Why is it needed and how industries are adopting it
- Different components of analytics
- Applications of analytics in different domains
- Statistical learning vs. Machine learning
- What is Data Science and skills of a data scientist
- Introduction to cloud machine learning engines
- Different types of machine learning algorithms–Supervised, Unsupervised and Reinforcement learning

Session 3 & 4–Introduction to Anaconda and Python

- Overview of Anaconda framework
- Python – Variables, objects, loops, conditions, function.
- Python Data structures – lists, tuples, dictionaries, sets
- Introduction to Numpy – ndarrays, ndarrays indexing, ndarrays datatypes and operations, statistical sorting and set operation
- Introduction to Pandas – Data ingestion, descriptive statistics, visualization, frequent data operations, merging dataframes, parsing timestamps
- Introduction to visualization – Matplotlib

Session 5–Introduction to R Platform

- Overview of R
- Fundamentals of R – Reading Data Files, Data Manipulation
- R Data structures – lists, dataframes
- Loops and Functions in R
- Useful packages for visualization and modelling
- Run through with few of the twenty three example codes in R

Day 2: Understanding basis statistics and R framework for data analysis

Session 1, 2 & 3–Basics of Statistics

- Probability basics – Benford Law (application in fraud analysis)
- Random Variable – Discrete and Continuous
- Probability density function and Cumulative density function
- Distribution Family – Gaussian Distribution, Standard Normal Distribution
- Population and Sample

- Central limit theorem, Demonstration of Central limit theorem on finance data
- Hypothesis testing – Z test, t test, test for proportion, analysis of variance (ANNOVA)
- Degree of freedom, Covariance and Correlation
- Partial and Semi-partial Correlation

Session 4 & 5–Introduction to useful packages in R

- R Data structures – lists, dataframes
- Loops and Functions in R
- Useful packages for visualization and modelling
- Run through with few of the twenty three example codes in R
- Introduction to R Markdown, Rattle, Shiny

Day 3: Understanding supervised learning algorithms

Session 1, 2 & 3–Lab 1: Linear Regression

- Introduction to simple and multiple linear regression
- Regression diagnostic–R-squared, t-test, F-test, error terms distribution, heteroscedasticity, identifying multi-collinearity and handling, AIC - model selection strategy
- Common task framework for model evaluation – training and test set.
- Case study using regression techniques and hands-on using R code for regression

Session 4 & 5–Lab 2: Logistic Regression

- Introduction to logistic regression
- Logistic regression diagnostic: Wald statistics, Hosmer Lemeshow test, Classification Matrix, Sensitivity, Specificity, ROC Curve, precision, recall, F1-score
- Strategy to find the optimal cut-off
- Bias and variance in the model, Bias vs. variance tradeoff
- Case study using logistic regression techniques and hands-on using Python code for regression.

Day 4: Understanding supervised learning and gradient descent algorithm

Session 1–Lab 2: Logistic Regression

- Introduction to logistic regression
- Logistic regression diagnostic: Wald statistics, Hosmer Lemeshow test, Classification Matrix, Sensitivity, Specificity, ROC Curve, precision, recall, F1-score
- Strategy to find the optimal cut-off
- Bias and variance in the model, Bias vs. variance tradeoff
- Case study using logistic regression techniques and hands-on using Python code for regression.

Session 2&3 –Lab 3: Introduction to Gradient Descent

- Hypothesis formulation for linear regression
- Deriving the cost function for linear regression
- Cost function– Intuition for linear regression with one parameter and two parameters
- Gradient descent algorithm–application in linear regression
- Hypothesis formulation for logistic regression
- Deriving the cost function for logistic regression
- Cost function– Intuition for logistic regression
- Gradient descent algorithm–application in logistic regression

Session 4–Lab 4: Decision Trees

- Decision tree – Classification and regression trees (CART), Gini Index, Entropy
- Decision tree – Chi-square automatic interaction detection (CHAID)
- Case study using decision tree techniques
- Hands-on using Python code

Session 5: Naïve Bayes classifier

- Naïve Bayes classifier on structured data
- Case study using decision tree techniques
- Hands-on using R code for CART and Naïve Bayes Classifier using caret package

Day 5: Understanding unsupervised learning and ensemble methods

Session 1–Lab 6: Clustering and Segmentation

- Supervised and Unsupervised learning
- Clustering–Hierarchical, K means
- Clustering diagnostic–Dendrogram , Calinski and Harabasz index, Silhouette width
- Case study using hierarchical clustering and K–means clustering tree techniques
- Hands-on using Python code for Hierarchical and K–means cluster

Session 2 & 3–Lab 7: Other Machine learning models (Ensemble Methods)

- What is Machine learning
- Different sampling strategies–Bootstrapping, Up–Sample, Down–Sample, Synthetic Sample, Cross–Validation Data
- Introduction to Bagging–Random Forest
- Other Bagging algorithms
- Introduction to Boosting– Adaptive boosting
- Other Boosting algorithms
- Case study of an imbalanced data and application of sampling strategies & ensemble methods
- Hands-on using R code on an imbalanced data

Session 4 & 5–Lab 8: Multivariate Gaussian Model for Anomaly Detection

- What is a fraud/anomaly
- Gaussian Model for anomaly detection
- Multivariate Gaussian Model for anomaly detection

COURSE SCHEDULE

Day 1: Understanding Anaconda Framework platform and other useful packages in Python

This day will be primarily cover introduction to business analytics, introduction to Anaconda and Python

Topic	Session	From	To
Introduction to Business Analytics	1	9 AM	10:15 AM
Introduction to Business Analytics...cont.	2	10:30 AM	11:15 AM
Introduction to Anaconda and Python platform	3	12:00 PM	1:15 PM
Introduction to Anaconda and Python platform...cont.	4	2:15 PM	3:30 PM
Introduction to R	5	3:45 PM	5:00 PM

Day 2: Understanding basis statistics and R framework for data analysis

Day is primarily devoted to concept building on supervised learning and hands-on using Python code for the same

Topic	Session	From	To
Basic of Statistics	1	9 AM	10:15 AM
Basic of Statistics...cont.	2	10:30 AM	11: 45 AM
Basic of Statistics...cont.	3	12:00 PM	1:15 PM
Introduction to R	4	2:15 PM	3:30 PM
Introduction to R...cont.	5	3:45 PM	5:00 PM

Day 3: Understanding supervised learning algorithms

Day will cover concept building on unsupervised learning, sampling strategy and hands-on using Python code for ensemble methods

Topic	Session	From	To
Lab 1: Multiple Linear Regression	1	9 AM	10:15 AM
Lab 1: Multiple Linear Regression...cont.	2	10:30 AM	11: 45 AM
Lab 1: Multiple Linear Regression...cont.	3	12:00 PM	1:15 PM
Lab 2: Logistic Regression	4	2:15 PM	3:30 PM
Lab 2: Logistic Regression...cont.	5	3:45 PM	5:00 PM

Day 4: Understanding unsupervised learning and ensemble methods

Day will cover concept building on unsupervised learning and gradient descent

Topic	Session	From	To
Lab 2: Logistic Regression...cont.	1	9 AM	10:15 AM
Lab 3: Gradient Descent	2	10:30 AM	11: 45 AM
Lab 3: Gradient Descent...cont.	3	12:00 PM	1:15 PM
Lab 4: Decision Tree	4	2:15 PM	3:30 PM
Lab 5: Naïve Bayes	5	3:45 PM	5:00 PM

Day 5: Understanding unsupervised learning and ensemble methods

Day will cover concept building on unsupervised learning and ensemble methods

Topic	Session	From	To
Lab 6: Clustering and Segmentation	1	9 AM	10:15 AM
Lab 7: Ensemble Methods	2	10:30 AM	11: 45 AM
Lab 7: Ensemble Methods...cont.	3	12:00 PM	1:15 PM
Lab 8: Gaussian Model	4	2:15 PM	3:30 PM
Lab 8: Gaussian Model...cont.	5	3:45 PM	5:00 PM

ABOUT INSTRUCTOR

Rahul Kumar is an alumnus of NIT Jaipur and IIM Bangalore. He has more than 12 years of experience spanning across software development, business consulting, analytical modelling and leading process improvement initiatives. He started his career in Information Technology sector and worked in companies like Satyam Computers, Nokia Siemens and Deloitte Consulting before venturing into his own business.

He co-founded a start-up ARIMA Research in June 2014 and was involved in internal operations and consulting engagement for various clients till December 2015.

On the technical front, he currently works as a consultant at Indian Institute of Management Bangalore and has executed several projects in analytics domain. His recent work in the field of analytics includes govt. of India funded research project on fraud analytics and credit scoring model for urban co-operative banks, predicting NPS for a reputed hospital chain of India and using machine learning algorithms to predict earnings manipulations. He has taken several public as well as client specific trainings on "Introduction to Data Science using Python/R" and "Introduction to Machine Learning Concepts using Python/R". He has presented papers in several national and international conferences. Few of the prominent ones are:

- Paper on "Using Machine learning algorithms to predict earnings manipulations" accepted for presentation in IOT and Machine learning conference (Paper 102: IML 2017) 17-18 October 2017.
- Paper on "Predicting Net Promoter Score (NPS) to Improve Patient Experience at Manipal Hospitals" published at Harvard Business Publishing, September 2017.
- Paper on "Behavioral Modeling to Predict Renege" published at Harvard Business Review, January 2016.
- Paper Presentation at CMMI conference organized by CMMI Institute, 10-11 Dec 2014 at Shenzhen, China.
- Paper Presentation at CCSE conference organized by China computer and software enterprises, 6th Dec 2014 at Shenzhen, China.
- Paper Publication and Presentation at 6th International ITSM Conference organized by QAI Global Services in Bangalore, August 2013.
- Paper Presentation at SEPG Europe conference organized by SEI | Carnegie Mellon University, 5-7 June 2012 at Madrid, Spain.

He has also undergone workshop on the usage of statistical models and techniques from ISI Bangalore. His other certifications include DB2 certification from IBM and ISO 9001:2008 lead auditor certification by DNV India.