Data Science Training

R, Predictive Modeling, Machine Learning, Python, Bigdata & Spark

Introduction:

This is a comprehensive course which builds on the knowledge and experience a business analyst and data scientist will have obtained after some years in the role. This course takes the business analysts / predictive modelers to the next level in terms of delivering effective and realistic solutions to machine learning and bigdata problems. This course provides techniques for data cleaning, visualizing the data, predictive modelling, machine learning & bigdata

Course Duration & Features

- Data Analytics training is a **200 HR long Self-paced and instructor-led** online course
- 100% hands-on training
- Contains real word business applications and examples. Project work at the end of each module
- Rich material and handouts for student reference

After Completion of this training:

- Gain exposure to key disciplines and skills needed to fulfil the role of a business analyst /predictive modeler / data scientist
- Build predictive models using linear, logistic regression and decision trees.
- Build machine learning models using Neural nets, SVM and Random forest

Prerequisite: Before attending this course, candidate should

- Have experience using applications, such as SAS/R/word processors/spreadsheets
- **No** statistical background is necessary
- Data Analysis and Reporting background is necessary
- This is a 100% hands on training. Every participant should have access to a computer.

Tools

- **R , Hadoop, Python & Spark**

Course Fee

- Please call us to discuss

Contact us

- info@statinfer.com
- Give a call at: +91 8048511820 or 9886760678
- #647 2nd floor 1st Main, Indira Nagar 1st Stage, 100 feet road,Indranagar, Bangalore,Karnataka, Pin code:-560038
- Landmarks: Opp. Namma Metro Pillar 48, Same Building as Muthoot Finance, Diagonally opposite to magnum honda

Course Contents

Week1-R Programming, Data Handling and Basic Statistics

1. Introduction Analytics Tool(R)
    a. Introduction to Data Analysis
    b. Introduction to R programming
    c. R Environment and Basic Commands
2. Data Handling in R
    a. Importing data
    b. Sampling
    c. Data Exploration
    d. Creating calculated fields
    e. Sorting & removing duplicates
3. Basic Descriptive Statistics
    a. Population and Sample
    b. Measures of Central tendency
    c. Measures of dispersion
4. Reporting and Data Validation
    a. Percentiles & Quartiles
    b. Box plots and outlier detection
    c. Creating Graphs and Reporting

Week2-Project-1 - Data Exploration, Validation and Cleaning Project

1. Project on Data handling
2. Data exploration
3. Data validation
4. Missing values identification
5. Outliers identification
6. Data Cleaning
7. Basic Descriptive statistics

Week3-Regression Analysis & Logistic Regression Model Building

1. Regression Analysis
    a. Correlation
    b. Simple Regression models
    c. R-Square
    d. Multiple regression
    e. Multicollinearity
    f. Individual Variable Impact
2. Logistic Regression
    a. Need of logistic Regression
    b. Logistic regression models
    c. Validation of logistic regression models
    d. Multicollinearity in logistic regression
    e. Individual Impact of variables
    f. Confusion Matrix

Week4-Decision Trees & Model Selection

1. Decision Trees
    a. Segmentation
    b. Entropy
    c. Building Decision Trees
    d. Validation of Trees
    e. Fine tuning and Prediction using Trees
2. Model Selection and Cross validation
    a. How to validate a model?
    b. What is a best model?
    c. Types of data
    d. Types of errors
    e. The problem of over fitting
    f. The problem of under fitting
    g. Bias Variance Tradeoff
    h. Cross validation
    i. Boot strapping

Week5-Project2 -Predictive Modelling Project

1. Objective
2. Model building-1
3. Model building-2
4. Model validation
5. Variable selection
6. Model calibration
7. Out of time validation

Week6-Neural Network, SVM and Random Forest

a. Neural Networks
    a. Neural network Intuition
    b. Neural network and vocabulary
    c. Neural network algorithm
    d. Math behind neural network algorithm
    e. Building the neural networks
    f. Validating the neural network model
    g. Neural network applications
    h. Image recognition using neural networks
b. SVM
    a. Introduction
    b. The decision boundary with largest margin
    c. SVM- The large margin classifier
    d. SVM algorithm
    e. The kernel trick
    f. Building SVM model
    g. Conclusion
c. Random Forest and Boosting
    a. Introduction
    b. The decision boundary with largest margin

c. SVM- The large margin classifier
d. SVM algorithm
e. The kernel trick
f. Building SVM model
g. Conclusion

Week7-Project3- Machine Learning Project

1. Objective
2. ML Model-1
3. ML Model-2
4. ML Model-3
5. Model calibration
6. Out of time validation

Week8-Big Data, Hadoop, Hive

1. Introduction to Big Data and Hadoop
    a. Generic Definition of big data
    b. Bigdata vs conventional data
    c. Philosophy of distributed computing
    d. What is Hadoop
    e. HDFS & Map Reduce
    f. Map Reduce Basics
    g. Map Reduce Program
    h. LAB: Hadoop / Hive Demo
    i. Conclusion
2. Hadoop Local Installation / VM/ Sandbox
    a. Using Hadoop VMware Image
    b. Hadoop VMware Configuration
    c. Working with Hadoop environment
    d. LAB1: Hadoop environment test
3. Hive
    a. What is Hive
    b. How hive works
    c. Hive Architecture
    d. Internal & External Tables
    e. Hive commands
    f. Hive hands on example on bigdata
4. Pig
    a. What is Pig
    b. How Pig works & Pig Architecture
    c. Pig commands
    d. Pig example on bigdata
    e. Writing queries – SPLIT, FILTER, JOIN, GROUP, SAMPLE, ILLUSTRATE

Week9-Project4 – Bigdata

1. Bigdata importing on to Hadoop
2. HDFS
3. Map Reduce Code
4. Hive query

5. Pig Script for reporting

Week10- Sqoop, Flume

1. Sqoop
    a. What is Sqoop
    b. Features of Sqoop
    c. Sqoop Architecture
    d. Importing data from RDBMS using Sqoop
    e. Exporting data from RDBMS using Sqoop
    f. Sqoop commands
    g. Working with Sqoop-Example
2. Flume
    a. What is Flume
    b. Features of Flume
    c. Importing data from non-RDBMS sources
    d. Exporting data from non-RDBMS sources
    e. Flume commands
    f. Working with Flume-Example

Week11-Project5- Bigdata tools & Distributed computing

1. Project Part-1: Bigdata importing using Sqoop
2. Project Part-2: Streaming data importing using Flume

Week12-Python Introduction & Project-6

1. Python Introduction
    a. What is Python & History
    b. Installing Python & Python Environment
    c. Basic commands in Python
    d. Data Types and Operations
    e. Python packages
    f. Loops
    g. My first python program
    h. If-then-else statement
2. Data Handling in Python
    a. Data importing
    b. Working with datasets
    c. Manipulating the datasets
    d. Creating new variables
    e. Exporting the datasets into external files
    f. Data Merging
3. Python Basic Statistics
    a. Taking a random sample from data
    b. Descriptive statistics
    c. Central Tendency
    d. Variance
    e. Quartiles, Percentiles
    f. Box Plots
    g. Graphs
4. Python Data Handling project

a. Project on Data handling
b. Data exploration
c. Data validation
d. Missing values identification
e. Outliers identification
f. Data Cleaning
g. Basic Descriptive statistics

Week13 Python Predictive Modelling & Project-7

1. Regression Analysis
    a. Correlation
    b. Simple Regression models
    c. R-Square
    d. Multiple regression
    e. Multicollinearity
    f. Individual Variable Impact
2. Logistic Regression
    a. Need of logistic Regression
    b. Logistic regression models
    c. Validation of logistic regression models
    d. Multicollinearity in logistic regression
    e. Individual Impact of variables
    f. Confusion Matrix
3. Decision Trees
    a. Segmentation
    b. Entropy
    c. Building Decision Trees
    d. Validation of Trees
    e. Fine tuning and Prediction using Trees
4. Model Selection and Cross validation
    a. How to validate a model?
    b. What is a best model?
    c. Types of data
    d. Types of errors
    e. The problem of over fitting
    f. The problem of under fitting
    g. Bias Variance Tradeoff
    h. Cross validation
    i. Boot strapping

Week14 Python Machine Learning

a. Neural Networks
    a. Neural network Intuition
    b. Neural network and vocabulary
    c. Neural network algorithm
    d. Math behind neural network algorithm
    e. Building the neural networks
    f. Validating the neural network model

   g. Neural network applications
   h. Image recognition using neural networks
b. SVM
   a. Introduction
   b. The decision boundary with largest margin
   c. SVM- The large margin classifier
   d. SVM algorithm
   e. The kernel trick
   f. Building SVM model
   g. Conclusion
c. Random Forest and Boosting
   a. Introduction
   b. The decision boundary with largest margin
   c. SVM- The large margin classifier
   d. SVM algorithm
   e. The kernel trick
   f. Building SVM model
   g. Conclusion

Week15-Project8- Python Machine Learning

1. Objective
2. ML Model-1
3. ML Model-2
4. ML Model-3
5. Model calibration
6. Out of time validation

Week16-Data Science Hackathon / Competition Project-9

- Enroll to data online science completion
- Data exploration
- Model building
- Testing the score and rank
- Variable selection
- Future reengineering
- Checking the score and rank