Introduction to BIG Data Science/Data Analytics

What background is required?
What is Data Science?
Why Data Science?
BIG Data Science/Analytics trend
What is Machine Learning?
Data Science Life Cycle

Tools for Data Science/Analytics

Anaconda Distribution package
Open Source: Python/R
Visualization tools: Matplotlib, Seaborn, introduction of Tableau

Data Analytics Problems/Use-cases

From Kaggle competitions
Types of Data: Structured, Unstructured (Image, Text…..)
Predictive Analytics Problems: Classification, Regression, Recommenders
Descriptive Analytics Problems: Clustering, Market Basket Analysis, PCA
Business Verticals: Retail, Real Estate, Banking, Financial, Social, Web, Medical, Scientific, Logistics,

Visualization tools:

Matplotlib,
Seaborn,
Introduction of Tableau

Statistics for Data Scientist

Descriptive Statistics for single variables
Mean, Median, Mode, Quartile, Percentile
Interquartile Range
Standard Deviation
Variance
Descriptive Statistics for two variables
Z-Score
Co-variance/ Co-relation
Chi-squared Analysis / Hypothesis Testing

Calculus for Data Scientist

Limits
Derivatives
Partial Derivatives
Gradients
Significance of Gradients

Probability for Data Scientist

Basic Probability
Conditional Probability
Properties of Random Variables
Expectations
Variance
Entropy and cross-entropy
Covariance and correlation
Estimating probability of Random variable
Understanding standard random processes

Data Distributions

Normal Distribution
Binomial Distribution
Multinomial Distribution
Bernoulli Distribution
Probability, Prior probability, Posterior probability
Bayes Theorem
Naive Bayes
Naive Bayes Algorithm
Normal Distribution

Mastering Python/R Language

How to install python (Anaconda)
How to install sciKit Learn (Anaconda)
How to work with Jupyter Notebook
How to work with Spyder IDE
Strings
Lists
Tuples
Sets

Dictionaries
Control Flows
Functions
Formal/Positional/Keyword arguments
Predefined functions (range, len, enumerates etc…)
Data Frames
Packages required for data Science in R/Python
Lab/Coding

Introduction to NumPy

One-dimensional Array
Two-dimensional Array
Pr-defined functions (arrange, reshape, zeros, ones, empty)
Basic Matrix operations
Scalar addition, subtraction, multiplication, division
Matrix addition, subtraction, multiplication, division and transpose
Slicing
Indexing
Looping
Shape Manipulation
Stacking

Introduction to Pandas

Series
DataFrame
df.GroupBy
df.crosstab
df.apply
df.map

Decision Trees

What are Decision Trees?
Gini, Entropy criterions
Decision trees in Classification
Decision trees in Regression
Ensembles
Random Forest
Boosting (Ada, Gradient, Extreme Gradient)
SVM
Ensembles

Overfitting/Under fitting

Understand what is overfitting and under fitting model
Visualize the overfitting and under fitting model
How do you handle overfitting?

Data Preparation Techniques

Structured Data Preparation
Data Type Conversion
Category to Numeric Conversion
Numeric to Category Conversion
Data Normalization: 0-1, Z-Score
Handling Skew Data: Box-Cox Transformation
Handling Missing Data

Re-sampling Techniques

K-fold
Repeated Hold-out Data
Bootstrap aggregation sampling

Exploratory Data Analysis (EDA)

Statistical Data Analysis
Data Visualization (Matplotlib, Seaboarn)
Exploring Individual Features
Exploring Bi-Feature Relationships
Exploring Multi-feature Relationships
Feature/Dimension Reduction: PCA
Intuition behind PCA
Covariance & Correlation
Relating PCA to Covariance/Correlation
Intuition to math
Applications of PCA: Dimensionality Reduction

Feature Engineering (FE)

Combine Features
Split Features

Data Visualization

Bar Chart
Histogram
Box whisker plot
Line plot
Scatter Plot
Heat Map

Tree Based Algorithms

Gini Index
Entropy
Information Gain
Tree Pruning

Classification (Supervised Learning)

What is Classification?
Finding Patterns/Fixed Patterns
Problems with Fixed Patterns
Machine learning approach over fixed pattern approach
Decision Tree based classification
Ensemble Based Classification
Logistic Regression (SGD Classifier)
Accuracy measurements
Confusion Matrix
ROC Curve
AUC Score
Multi-class Classification
Softmax Regression Classifier
Multi-label Classification
Multi-output Classification

Ensemble models

Random Forest
Bagging
Boosting
Adaptive Boosting
Gradient Boosting
Extreme Gradient Boosting
Heterogeneous Ensemble Models
Stacking / Voting

Regression (Supervised Learning)

What is regression?
Regression example in business verticals
Solution strategies for Regression
Linear Regression
Explanation of statistics
Evaluation metrics
Root Mean Squeare(RMSE)
R-Squre,
Adj R-Squre
Feature selection methods
Linear regression

Multiple/Polynomial Regression (scikit-learn)

Multiple Linear Regressions (SGD Regressor)
Gradient Descent (Calculus way of solving linear equation)
Feature Scaling (Min-Max vs Mean Normalization)
Feature Transformation
Polynomial Regression
Matrix addition, subtraction, multiplication and transpose
Optimization theory for data scientist

Optimisation Theory (Gradient Descent Algorithm)

Modelling ML problems with optimization requirements
Solving unconstrained optimization problems
Solving optimization problems with linear constraints
Gradient descent ideas
Gradient descent
Batch gradient descent
Stochastic gradient descent

Model Evaluation and Error Analysis

Train/Validation/Test split
K-Fold Cross Validation
The Problem of Over-fitting (Bias-Variance tread-off)
Learning Curve
Regularization (Ridge, Lasso and Elastic-Net)
Hyper Parameter Tuning (GridSearchCV)

Recommendation Problem

What is Recommendation System?
Top-N Recommender
Rating Prediction
Content based Recommenders
Limitations of Content based recommenders
Machine Learning Approaches for Recommenders
User-User KNN model, Item-Item KNN model
Factorization or latent factor model
Hybrid Recommenders
Evaluation Metrics for Recommendation Algorithms
Top-N Recommnder: Accuracy, Error Rate
Rating Prediction: RMSE

Clustering (Unsupervised Learning)

Finding pattern and Fixed Pattern Approach
Limitations of Fixed Pattern Approach
Machine Learning Approaches for Clustering
Iterative based K-Means Approaches
Density based DB-SCAN Approach
Evaluation Metrics for Clustering
Cohesion, Coupling Metrics
Correlation Metric

Support Vector Machine (SVM)

SVM Classifier (Soft/Hard – Margin)
Linear SVM
Non-Linear SVM
Kernel SVM
SVM Regression

PCA (Unsupervised Learning)

Dimensionality Reduction
Choosing Number of Dimensions or Principal Components
Incremental PCA
Kernel PCA
When to apply PCA?
Eigen vectors
Eigen values

Model Deployment

Pickle (pkl file)
Model load from pkl file and prediction

Association Rules

A priori Algorithm
Collaborative Filtering (User-Item based)
Collaborative Filtering (User-User based)
Collaborative Filtering (Item-Item based)

Deep Learning:

Introduction to Deep Learning
Tensorflow
Keras
Setting up new environment for Deep Learning
Perceptron model for classification and regression
Perceptron Learning
Limitations of Perceptron model
Multi-layer FF NN model for classification and regression
ML-FF-NN Learning with backpropagation
Applying ML-FF-NN and parameter tuning
Pros and Cons of the Model

Image classification

Image Data Preparation
Converting to gray scale
Pixel Value Normalization
Building Pixel Intensity Matrix
Neural Networks
Fully connected Neural Networks
Feed Forward Neural Networks
Convolution Neural Networks
Filters, Max Pooling
Functional APIs

Text analytics:

Bag of words
Glove Dictionary
Text Data Preparation
Normalizing Text
Stop word Removal
Whitespace Removal
Stemming
Building Document Term Matrix
NLP (Natural Language Processing)