
Data Science with Python Course Content

Course Description:

Data Science is one of the hottest fields of the 21st century. Data Science is a trending technology that gives useful information and insights by analyzing structured and unstructured data using scientific methods, processes, algorithms, and systems. The data science concept blends statistics, data analysis, machine learning, and related methods in order to understand and analyze the data. Data science with Python programming language has much scope in the IT industry and has a huge demand across the globe with honchos like Amazon, Google, Microsoft paying great salaries and perks to Data scientists, Data analytics. Learning Data Science with Python language will give you an extra edge in your career in the software industry.

Hachion's Data Science with Python online training is prepared by the trained masters with all basic and advanced concepts of python programming language. This course provides you structured syllabus from scratch including basics of Python, data analysis, data scraping, data visualization, machine learning algorithms, etc. If you enjoy mathematics and statistics and have a piece of good practical business knowledge and the ability to present ideas in meaningful ways will definitely have a great impact to build a strong career in the Data Science field. The complete course will enhance your practical knowledge and programming skills by solving the assignments included within the Python Data Science tutorial. This course also provides hands-on experience through real-time projects.

Course Content:

Introduction to Data Science with Python

- What is Analytics & Data Science?
- Common terms in Analytics
- Analytics vs. Data warehousing, OLAP, MIS Reporting
- Relevance in industry and need of the hour
- Types of problems and business objectives in various industries
- How leading companies are harnessing the power of analytics?
- Critical success drivers
- Overview of analytics tools & their popularity
- Analytics Methodology & problem-solving framework
- List of steps in Analytics projects
- Identify the most appropriate solution design for the given problem statement
- The project plan for Analytics project & key milestones based on effort estimates
- Build a Resource plan for an analytics project
- Why Python for Data science?

Python: Essentials (Core)

- Overview of Python - Starting with Python

- Introduction to installation of Python
- Introduction to Python Editors & IDE's (Canopy, pycharm, Jupyter, Rodeo, Ipython etc...)
- Understand Jupyter notebook & Customize Settings
- Concept of Packages/Libraries - Important packages (NumPy, SciPy, scikit-learn, Pandas, Matplotlib, etc)
- Installing & loading Packages & Name Spaces
- Data Types & Data objects/structures (strings, Tuples, Lists, Dictionaries)
- List and Dictionary Comprehensions
- Variable & Value Labels – Date & Time Values
- Basic Operations - Mathematical - string - date
- Reading and writing data
- Simple plotting
- Control flow & conditional statements
- Debugging & Code profiling
- How to create class and modules and how to call them?

Scientific Distributions Used in Python for Data Science

- Numpy, scify, pandas, scikitlearn, statmodels, nltk etc

Accessing/Importing and Exporting Data Using Python Modules

- Importing Data from various sources (CSV, txt, excel, access, etc)
- Database Input (Connecting to the database)
- Viewing Data objects - subsetting, methods
- Exporting Data to various formats
- Important python modules: Pandas, beautiful soup

Data Manipulation – Cleansing – Munging Using Python Modules

- Cleansing Data with Python
- Data Manipulation steps (Sorting, filtering, duplicates, merging, appending, subsetting, derived variables, sampling, Data type conversions, renaming, formatting etc)
- Data manipulation tools (Operators, Functions, Packages, control structures, Loops, arrays, etc)
- Python Built-in Functions (Text, numeric, date, utility functions)
- Python User Defined Functions
- Stripping out extraneous information
- Normalizing data
- Formatting data
- Important Python modules for data manipulation (Pandas, Numpy, re, math, string, DateTime etc)

Data Analysis – Visualization Using Python

- Introduction exploratory data analysis
- Descriptive statistics, Frequency Tables and summarization
- Univariate Analysis (Distribution of data & Graphical Analysis)
- Bivariate Analysis(Cross Tabs, Distributions & Relationships, Graphical Analysis)

- Creating Graphs- Bar/pie/line chart/histogram/ boxplot/ scatter/ density etc)
- Important Packages for Exploratory Analysis(NumPy Arrays, Matplotlib, seaborn, Pandas and scipy.stats etc)

Introduction to Statistics

- Basic Statistics - Measures of Central Tendencies and Variance
- Building blocks - Probability Distributions - Normal distribution - Central Limit Theorem
- Inferential Statistics -Sampling - Concept of Hypothesis Testing
- Statistical Methods - Z/t-tests(One sample, independent, paired), Anova, Correlations and Chisquare
- Important modules for statistical methods: Numpy, Scipy, Pandas

Introduction to Predictive Modeling

- Concept of model in analytics and how it is used?
- Common terminology used in analytics & modeling process
- Popular modeling algorithms
- Types of Business problems - Mapping of Techniques
- Different Phases of Predictive Modeling

Data Exploration for Modeling

- Need for structured exploratory data
- EDA framework for exploring the data and identifying any problems with the data (Data Audit Report)
- Identify missing data
- Identify outliers data
- Visualize the data trends and patterns

Data Preparation

- Need of Data preparation
- Consolidation/Aggregation - Outlier treatment - Flat Liners - Missing values- Dummy creation -
- Variable Reduction
- Variable Reduction Techniques - Factor & PCA Analysis

Segmentation: Solving Segmentation Problems

- Introduction to Segmentation
- Types of Segmentation (Subjective Vs Objective, Heuristic Vs. Statistical)
- Heuristic Segmentation Techniques (Value Based, RFM Segmentation and Life Stage Segmentation)
- Behavioral Segmentation Techniques (K-Means Cluster Analysis)
- Cluster evaluation and profiling - Identify cluster characteristics
- Interpretation of results - Implementation on new data

Linear Regression: Solving Regression Problems

- Introduction - Applications
- Assumptions of Linear Regression
- Building Linear Regression Model

- Understanding standard metrics (Variable significance, R-square/Adjusted R-square, Global hypothesis ,etc)
- Assess the overall effectiveness of the model
- Validation of Models (Re running Vs. Scoring)
- Standard Business Outputs (Decile Analysis, Error distribution (histogram), Model equation, drivers etc.)
- Interpretation of Results - Business Validation - Implementation on new data

Logistic Regression: Solving Classification Problems

- Introduction - Applications
- Linear Regression vs. Logistic Regression Vs. Generalized Linear Models
- Building Logistic Regression Model (Binary Logistic Model)
- Understanding standard model metrics (Concordance, Variable significance, Hosmer Lemeshov Test, Gini, KS, Misclassification, ROC Curve etc)
- Validation of Logistic Regression Models (Re running Vs. Scoring)
- Standard Business Outputs (Decile Analysis, ROC Curve, Probability Cut-offs, Lift charts, Model equation, Drivers or variable importance, etc)
- Interpretation of Results - Business Validation - Implementation on new data

Time Series Forecasting: Solving Forecasting Problems

- Introduction - Applications
- Time Series Components(Trend, Seasonality, Cyclicity and Level) and Decomposition
- Classification of Techniques(Pattern based - Pattern less)
- Basic Techniques - Averages, Smoothing, etc
- Advanced Techniques - AR Models, ARIMA, etc
- Understanding Forecasting Accuracy - MAPE, MAD, MSE, etc

Machine Learning -Predictive Modeling – Basics

- Introduction to Machine Learning & Predictive Modeling
- Types of Business problems - Mapping of Techniques - Regression vs. classification vs. segmentation vs. Forecasting
- Major Classes of Learning Algorithms -Supervised vs. Unsupervised Learning
- Different Phases of Predictive Modeling (Data Pre-processing, Sampling, Model Building, Validation)
- Over fitting (Bias-Variance Trade off) & Performance Metrics
- Feature engineering & dimension reduction
- Concept of optimization & cost function
- Overview of gradient descent algorithm
- Overview of Cross validation (Bootstrapping, K-Fold validation etc)
- Model performance metrics (R-square, Adjusted R-square, RMSE, MAPE, AUC, ROC curve, recall, precision, sensitivity, specificity, confusion metrics)

Unsupervised Learning: Segmentation

- What is the segmentation & Role of ML in Segmentation?
- Concept of Distance and related math background

- K-Means Clustering
- Expectation Maximization
- Hierarchical Clustering
- Spectral Clustering (DBSCAN)
- Principle component analysis (PCA)

Supervised Learning: Decision Trees

- Decision Trees - Introduction - Applications
- Types of Decision Tree Algorithms
- Construction of Decision Trees through Simplified Examples; Choosing the "Best" attribute at each Non-Leaf node; Entropy; Information Gain, Gini Index, Chi Square, Regression Trees
- Generalizing Decision Trees; Information Content and Gain Ratio; Dealing with Numerical Variables; other Measures of Randomness
- Pruning a Decision Tree; Cost as a consideration; Unwrapping Trees as Rules
- Decision Trees - Validation
- Over fitting - Best Practices to avoid

Supervised Learning: Ensemble Learning

- Concept of Ensembling
- Manual Ensembling Vs. Automated Ensembling
- Methods of Ensembling (Stacking, Mixture of Experts)
- Bagging (Logic, Practical Applications)
- Random forest (Logic, Practical Applications)
- Boosting (Logic, Practical Applications)
- Ada Boost
- Gradient Boosting Machines (GBM)
- XGBoost

Supervised Learning: Artificial Neural Networks (ANN)

- Motivation for Neural Networks and Its Applications
- Perceptron and Single Layer Neural Network, and Hand Calculations
- Learning In a Multi Layered Neural Net: Back Propagation and Conjugant Gradient Techniques
- Neural Networks for Regression
- Neural Networks for Classification
- Interpretation of Outputs and Fine tune the models with hyper parameters
- Validating ANN models

Supervised Learning: Support Vector Machines

- Motivation for Support Vector Machine & Applications
- Support Vector Regression
- Support vector classifier (Linear & Non-Linear)
- Mathematical Intuition (Kernel Methods Revisited, Quadratic Optimization and Soft Constraints)

- Interpretation of Outputs and Fine tune the models with hyper parameters
- Validating SVM models
- Supervised Learning: KNN

Supervised Learning: KNN

- What is KNN & Applications?
- KNN for missing treatment
- KNN For solving regression problems
- KNN for solving classification problems
- Validating KNN model
- Model fine tuning with hyper parameters

Supervised Learning: Naive Bayes

- Concept of Conditional Probability
- Bayes Theorem and Its Applications
- Naïve Bayes for classification
- Applications of Naïve Bayes in Classifications

Text Mining & Analytics

- Taming big text, Unstructured vs. Semi-structured Data; Fundamentals of information retrieval, Properties of words; Creating Term-Document (TxD);Matrices; Similarity measures, Low-level processes (Sentence Splitting; Tokenization; Part-of-Speech Tagging; Stemming; Chunking)
- Finding patterns in text: text mining, text as a graph
- Natural Language processing (NLP)
- Text Analytics – Sentiment Analysis using Python
- Text Analytics – Word cloud analysis using Python
- Text Analytics - Segmentation using K-Means/Hierarchical Clustering
- Text Analytics - Classification (Spam/Not spam)
- Applications of Social Media Analytics
- Metrics (Measures Actions) in social media analytics
- Examples & Actionable Insights using Social Media Analytics
- Important python modules for Machine Learning (SciKit Learn, stats models, scipy, nltk etc)
- Fine tuning the models using Hyper parameters, grid search, piping etc.

R Programming

- While loop
- If loop
- For loop
- Arithmetic operations

Statistics

- Correlation

- Linear Regression
- Non-Linear Regression
- Predictive time series forecasting
- K means clustering
- P-value
- Find outlier
- Neural Network
- Error Measure

Leading Topics

- Overture of R Shiny
- What is Hadoop
- Integration of Hadoop in R
- Data Mining using R
- Clinical research preface in R
- API in R (Twitter and Facebook)
- Word Cloud in R