

PYTHON FOUNDATION: VISUALIZATION USING 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
 - Project plan for Analytics project & key milestones based on effort estimates
 - Build Resource plan for analytics project
- Why Python for data science?

PYTHON FOUNDATION: ESSENTIALS (CORE)

- Overview of Python- Starting with Python
- Why Python for data science?
- Anaconda vs. python
- Introduction to installation of Python
- Introduction to Python Editors & IDE's(Jupyter,/Ipython)
- Understand Jupyter notebook & Customize Settings

- Concept of Packages - 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
- Control flow & conditional statements
- Debugging & Code profiling
- Python Built-in Functions (Text, numeric, date, utility functions)
- User defined functions – Lambda functions
- Concept of apply functions
- Python – Objects – OOPs concepts
- How to create class and modules?
- How to call classes and modules?
- Concept of pipelines in Python

PYTHON FOUNDATION: OPERATIONS WITH NUMPY (NUMERICAL PYTHON)

- What is NumPy?
- Overview of functions & methods in NumPy
- Data structures in NumPy
- Creating arrays and initializing
- Reading arrays from files
- Special initializing functions
- Slicing and indexing

- Reshaping arrays
- NumPy Maths
- Combining arrays
- Basic algebraic operations using NumPy arrays
- Solving linear equations
- Matrix inversions
- Calculating Eigen vectors

PYTHON FOUNDATION: OVERVIEW OF PANDAS

- What is pandas, its functions & methods
- Pandas Data Structures (Series & Data Frames)
- Creating Data Structures (Data import – reading into pandas)

PYTHON FOUNDATION: ACCESSING/IMPORTING AND EXPORTING DATA USING PYTHON MODULES

- Importing Data from various sources (Csv, txt, excel,etc.)
- Database Input (Connecting to database)
- Viewing Data objects - sub setting, methods
- Exporting Data to various formats
- Understanding of data
- Important python modules: Pandas, NumPy

PYTHON FOUNDATION: CLEANSING DATA WITH PYTHON

- Understand the data
- Sub Setting Data or Filtering Data or Slicing Data o Using [] brackets o Using indexing or referring with column names/rows o Using functions
- Dropping rows & columns
- Mutation of table (Adding/deleting columns)

- Binning data (Binning numerical variables in to categorical variables using cut() and qcut() functions)
- Renaming columns or rows
- Sorting
 - o By data/values, index
 - o By one column or multiple columns
 - o Ascending or Descending
- Type conversions • Setting index
- Handling duplicates
- Handling missing values – detect, filter, replace
- Handling outliers
- Creating dummies from categorical data (using get_dummies())
- Applying functions to all the variables in a data frame (broadcasting)
- Data manipulation tools(Operators, Functions, Packages, control structures, Loops, arrays etc.)
- Important Python modules for data manipulation (Pandas, NumPy, re, math, string, datetime etc.)

PYTHON FOUNDATION: VISUALIZATION USING PYTHON

- Exploratory data analysis
- Descriptive statistics, Frequency Tables and summarization
- Uni-variate Analysis (Distribution of data & Graphical Analysis)
- Bi - Variate Analysis(Cross Tabs, Distributions & Relationships, Graphical Analysis)
- Creating different Graphs using multiple python packages- Bar/pie/line chart/histogram/stack chart/boxplot/scatter/ density etc)
- Important Packages for Visualization (graphical analysis) – Pandas, Matplotlib, Seaborn, Bokeh etc.)

PYTHON FOUNDATION: BASIC STATISTICS & IMPLEMENTATION OF STATS METHODS IN PYTHON

- Basic Statistics - Measures of Central Tendencies and Variance
- What is probability distribution?
- Important distributions (discrete & continuous distributions)

- Deep dive of normal distributions and properties
- Concept of sampling & types of sampling
- Concept of standard error and central limit theorem
- Inferential Statistics - Concept of Hypothesis Testing
- Statistical Methods - Z/t-tests (One sample, independent, paired), ANOVA, Correlation and Chi- square
- Important modules for statistical methods: NumPy, SciPy, Pandas

PYTHON MACHINE LEARNING: INTRODUCTION TO MACHINE LEARNING

- Introduction to Predictive Modelling
- Types of Business problems - Mapping of Techniques
- Relevance in industry and need of the hour
- Difference between jargons i.e. data science, data analysis, data analytics, data mining
- What is Machine Learning?
- What is the goal of Machine Learning?
- Applications of ML (Marketing, Risk, Operations, etc.)
- Key components of ML
- Overall process of executing the ML project (Data Pre-processing, Sampling, Model Building, Validation)
- Common mistakes done in ML project and how to overcome
- Different terms to know for ML

PYTHON MACHINE LEARNING: LEARNING ALGORITHMS

- Major Classes of Learning Algorithms -Supervised vs. Unsupervised Learning vs. Semi supervised vs. Reinforcement Learning
- Important Consideration like fitment of techniques
- Concept of Over fitting and Under fitting (Bias-Variance Trade off) & Performance Metrics
- Concept of optimization - Gradient descent algorithm

- Concept of feature engineering
- Regularization (LASSO, LARS, Elastic net and Ridge regression)
- Types of Cross validation (Train & Test, K-Fold validation etc.)
- Cost & optimization functions

PYTHON MACHINE LEARNING: SUPERVISED LEARNING - REGRESSION PROBLEMS USING LINEAR REGRESSION

- Introduction - Applications
- Assumptions of Linear Regression
- Building Linear Regression Model
- Important steps in Model building
- Need of Data preparation
- Data Audit Report and Its importance
- Consolidation/Aggregation - Outlier treatment - Flat Liners - Missing values- Dummy creation - Variable Reduction
- Variable Reduction Techniques - Factor & PCA Analysis
- Understanding standard metrics (Variable significance, R-square/Adjusted R-square, Global hypothesis ,etc.)
- 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

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: CLASSIFICATION PROBLEMS USING LOGISTIC REGRESSION

- Introduction - Applications
- Linear

Regression Vs. Logistic Regression Vs. Generalized Linear Models

- Building Logistic Regression Model

- Important steps in model building
- Understanding standard model metrics (Concordance, Variable significance, Gini, KS, Misclassification, 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, etc.)
- Interpretation of Results - Business Validation
- Implementation on new data and Tracking the model

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: CLASSIFICATION & REGRESSION PROBLEMS USING DECISION TREES

- Over view of Decision Trees
- Types of decision trees (Regression Trees, Classification trees, Oblique Decision Trees)
- Types of decision tree algorithms (CART vs. CHAID vs. C50 etc.)
- Concept of objective segmentation
- How to use decision trees to solve regression, classification & segmentation problems)
- Rule Based Knowledge: Logic of Rules, Evaluating Rules, Rule Induction and Association Rules
- Construction of Decision Trees through Simplified Examples; Choosing the "Best" attribute at each Non-Leaf node;
- Splitting criteria: Entropy; Information Gain, Gini Index, Chi Square; ANOVA)
- Generalizing Decision Trees; Information Content and Gain Ratio; Dealing with Numerical Variables; other Measures of Randomness
- Pruning decision tree
- Cost as a consideration
- Fine Tuning model using tuning parameters
- Model validation
- Over fitting - Best Practices to avoid
- Implementation of Solution

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: CLASSIFICATION & REGRESSION PROBLEMS USING ENSEMBLE LEARNING

- What is concept of Ensemble learning (Stacking, Mixture of Experts)?
- Types of ensemble models (homogeneous, heterogeneous)
- Logic, Practical Applications • Ensemble learning techniques
- Bagging
- Random Forest
- Boosting
- AdaBoost
- Gradient Boost
- XGBoost
- Fine tuning the model using tuning parameters

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: CLASSIFICATION & REGRESSION PROBLEMS USING KNN

- What is concept of Instance based learning?
- What is KNN?
- KNN method for regression & classification
- KNN method for missing imputation
- Computation of Distance Matrix
- The Optimum K value
- Model Building, validation & Evaluation of Model
- Advantages & Disadvantages of KNN Models
- Applications of KNN in collaborative filtering, digit recognition
- KNN in collaborative filtering, digit recognition

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: CLASSIFICATION & REGRESSION PROBLEMS USING BAYESIAN TECHNIQUES

- Fundamentals of Probability; Conditional and Marginal Probability; Bayes Theorem and Its Applications
- Probabilities - The Prior and Posterior Probabilities
- Bayesian Belief nets, MAP, Naïve Rule and Naïve Bayes
- Naïve Bayes for classification - Data Processing - Discretization of Features
- Applications of Naïve Bayes in Text Mining, Spam Engines and Classifications
- Model Building, Validation and Evaluation of model
- Pros/Cons of Naïve Bayes Models

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: REGRESSION & CLASSIFICATION PROBLEMS USING SUPPORT VECTOR MACHINES

- What is Support vector machines?
- Understanding SVM
- Concepts of Linearly separable vs. non separable data
- Mathematical Intuition (Kernel Methods Revisited, Quadratic Optimization and Soft Constraints)
- Train/Test/Tune the Model using SVM
- Applications and Interpretation

PYTHON MACHINE LEARNING: UNSUPERVISED LEARNING: SEGMENTATION PROBLEMS USING CLUSTER ANALYSIS

- Introduction to Segmentation
- Types of Segmentation (Subjective Vs. Objective, Heuristic Vs. Statistical)
- Heuristic Segmentation Techniques (Value Based, RFM Segmentation and Life Stage Segmentation)
- Concept of Distance and related math background
- Segmentation Techniques

PYTHON MACHINE LEARNING: UNSUPERVISED LEARNING: SEGMENTATION PROBLEMS USING CLUSTER ANALYSIS

- K-Means/K-Medians Clustering
- Density Based clustering (DBSCAN)
- Identifying number of segments (Pseudo F-value, Silhouette score, elbow method etc.)
- Cluster evaluation and profiling
- Identifying the characteristics of segmentation
- Interpretation of results - Implementation on new data
- Overview of other unsupervised learning techniques (Factor analysis, Hidden Markov models, Gaussian mixture models etc.)

PYTHON MACHINE LEARNING: FORECASTING OVERVIEW AND BASICS OF TIME SERIES

- What is forecasting?
- Applications of forecasting
- Time Series Components (Trend, Seasonality, Cyclicity and Level) and Decomposition
- Types of Seasonality (Hourly, daily, weekly, monthly, quarterly etc.)
- Classification of Techniques (Pattern based - Pattern less)
- Important terminology: lag, lead, Stationary, stationary tests, auto correlation & white noise, ACF & PACF plots, auto regression, differencing
- Classification of Time Series Techniques (Univariate & Multivariate)

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: FORECASTING PROBLEMS USING TIME SERIES ANALYSIS Stationary Time Series Methods

- Moving Averages
- Weighted moving averages
- Exponential Smoothing
- Comparison between MA & ES Trend Based Time Series

- Linear Regression
- Double exponential smoothing (Holt's Method)
- Comparison between Regression & DS Seasonal Time Series
- Decomposition - CMA Method Advanced Techniques • Box Jenkins Methodology
- AR, MA, ARMA Models
- ARIMA/SARIMA
- ARIMAX, SARIMAX

PYTHON MACHINE LEARNING: EVALUATION OF FORECASTING

- Understanding Forecasting Accuracy
- Goodness Metrics: MSE, MAPE, RMSE, MAD

PYTHON MACHINE LEARNING: SUPERVISED LEARNING: REGRESSION & CLASSIFICATION PROBLEMS USING NEURAL NETWORKS

- Motivation for Neural Networks and Its Applications
- Understand Neural Networks
- Structure of Networks
- Perceptron and Single Layer Neural Network, and Hand Calculations
- Learning In a Multi Layered Neural Net: Back Propagation and Conjugant Gradient Techniques
- The ANN Model • Types of Activate functions
- Train/Test/Tune the ANN Model

PYTHON TEXT MINING NLP/NLG: INTRODUCTION TO TEXT MINING

- Unstructured vs. Semi-structured Data
- Text Mining - characteristics, trends
- Domain presentation - discussion of various areas and their applications
- Programming languages designated for working on Text Mining analysis

- Data Scientist - a profession comprising mainly of working with Text Mining
- Social Media – Characteristics of Social Media
- Applications of Social Media Analytics
- Examples & Actionable Insights using Social Media Analytics

PYTHON TEXT MINING NLP/NLG: TEXT PROCESSING USING BASE PYTHON & PANDAS, REGULAR EXPRESSIONS

- Text processing using string functions & methods
- Understanding regular expressions
- Identifying patterns in the text using regular expressions

PYTHON TEXT MINING NLP/NLG: TEXT PROCESSING WITH SPECIALIZED MODULES LIKE NLTK, SKLEARN ETC

- Getting Started with NLTK
- Introduction to NLP & NLTK
- Introduction to NLTK Modules (corpus, tokenize, Stem, collocations, tag, classify, cluster, tbl, chunk, Parse, ccg, sem, inference, metrics, app, chat, toolbox etc)

PYTHON TEXT MINING NLP/NLG: INITIAL DATA PROCESSING AND SIMPLE STATISTICAL TOOLS

- Fundamentals of information retrieval
- Reading data from file folder/from text file, from the Internet & Web scrapping, Data Parsing
- Cleaning and normalization of data
- Sentence Tokenize and Word Tokenize, Removing insignificant words,(“stop words”), Removing special symbols, removing bullet points and digits, changing letters to lowercase, stemming /lemmatisation /chunking
- Creating Term-Document matrix
- Finding associations
- Removing rare terms (Sparse terms)

- Measurement of similarity between documents and terms
- Visualization of term significance in the form of word clouds

- Tagging text with parts of speech

- Word Sense Disambiguation

PYTHON TEXT MINING NLP/NLG: ADVANCED DATA PROCESSING AND VISUALIZATION

- Sentiment analysis

- vocabulary approach, based on Bayesian probability methods

- Name entity recognition (NER)

- Methods of data visualization

- word length counts plot

- word frequency plots

- word clouds

- correlation plots

- letter frequency plot

- Heat map

- Grouping texts using different methods

PYTHON TEXT MINING NLP/NLG: ADVANCED DATA PROCESSING AND VISUALISATION

- Data-centric methods

- K-means

- Classification Models (spam detection, topic modelling)

- K Nearest Neighbours

- SVM (Linear Support Vector Classifier)

- Naive Bayes

- Decision tree

- Semantic similarity between texts
- Language Models and n-grams -- Statistical Models of Unseen Data (Smoothing)

PYTHON TEXT MINING NLP/NLG: FINAL PROJECTS

- Sentiment Analysis (Classification, weighted score etc)
- Word cloud analysis (Examples)
- Segmentation using K-Means/Hierarchical Clustering (Grouping the similar words)
- Classification (Spam/Not spam)
- Topic Modeling (LDA, LSA, Louvain etc)
- Text Summarization