DATA SCIENCE

Data play with python

ABSTRACT

One of the key technologies for future is data science. With high volume, variety and velocity of data; Industries are heading towards mining the potential benefits underneath the data to maximize innovations on productivity. Python being highly versatile language will enhance analytical capabilities.

Rahul Patnala(MBA-Business Analytics)

Contents

Introduction To Data Science

Jargon Busting
Analytics Problem Solving Framework
Language of Data Analysts
Business and Data Understanding
Overview of analytics tools & their popularity
Data Dictionary & Data Granularity
Data Quality & Cleaning
Data Preparation
Data Visualization
Case Study

Python For Data Science

Overview of Python
Need of Python for data science
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 etc.)
Variable & Value Labels – Date & Time Values
Basic Operations - Mathematical - string – date
Importing Data from various sources (csv, txt, excel, xml etc.)
Database Input (Connecting to database)
Viewing Data objects - sub setting methods
Exporting Data to various formats (Different File systems)
Important python modules (NumPy, SciPy, scikit-learn, Pandas, Matplotlib, etc)
Puzzles/Exercise

Data Wrangling In Python

Cleansing Data with Python

Data Manipulation steps (Sorting, filtering, duplicates, merging, appending, subsetting, derived variables, sampling, Data type conversions, renaming, formatting etc.)

Data manipulation helpers (Operators, Functions, Packages, control structures, Loops, arrays etc.)

Python Built-in Functions (Text, numeric, date, utility functions)

Python User Defined Functions (UDFs)

Formatting data

Puzzles/Exercise

Introduction to Statistics

Basic Statistics - Measures of Central Tendencies and Variance

Building blocks - Probability Distributions - Normal distribution - Central Limit Theorem

Descriptive statistics, Frequency Tables and summarization

Univariate Analysis and Bivariate Analysis

Inferential Statistics -Sampling - Concept of Hypothesis Testing

Statistical Methods - Z/t-tests (One sample, independent, paired), Anova, Correlations and Chi-square

Data Visualization And Statistics 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,

Important Packages for Exploratory Analysis and for statistical methods (NumPy Arrays, Matplotlib, seaborn, Pandas and scipy.stats etc.)

Case Study

Introduction to Machine Learning & Predictive Modelling

Types of Business problems - Mapping of Techniques - Regression vs. classification vs. segmentation vs. Forecasting

Machine Learning Framework

Major Classes of Learning Algorithms -Supervised vs. Unsupervised Learning

Different Phases of Predictive Modelling (Data Pre-processing, Sampling, Model Building, Validation)

Over fitting (Bias-Variance Trade off) & Performance Metrics

Feature engineering & dimension reduction (PCA)

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, and confusion metrics)

Linear Regression (SLR, MLR, Generalised Linear Regression, Regularization Regression)

Supervised Classification (K-NN, Naïve Bayes, Logistic Regression, Support Vector Machines, Decision Trees, Neural Network)

Concept of Distance and related math background

Un-Supervised learning (K-Means Clustering, Hierarchical Clustering)

Time series forecasting, Time Series Components (Trend, Seasonality, Cyclicity and Level) and Decomposition

Basic Techniques of time series - Averages, Smoothening, etc.

Advanced Techniques of time series - AR Models, ARIMA, etc.

Understanding Forecasting Accuracy of time series - MAPE, MAD, MSE, etc.

Concept of Ensembling and Methods of Ensembling

Association Rule Mining

Case Study and project for Applying different algorithms to solve the business problems and bench mark the results