Statistics Fundamentals & R

Introduction to Data Science

- What is Data Science?
- Analytics Landscape
- Life Cycle of Data Science Projects
- Data Science Tools & Technologies

R for Data Science

- R Installation, R Studio, Understanding Data Structures in R – Lists, Matrices, Vectors
- Intro to R Programming
- R Base Software
- Understanding CRAN
- RStudio the IDE
- Basic Building Blocks in R
- Understanding Vectors in R
- Basic Operations Operators and Types
- Handling Missing Values in R
- Subsetting Vectors in R
- Matrices and Data Frames in R
- Logical Statements in R
- Lapply, Sapply, Vapply and Tapply Functions

Statistical Learning

- Measures of Central Tendency in Data
- Measures of Dispersion
- Understanding Skewness in Data
- Probability Theory
- Bayes Theorem
- Probability Distributions
- Hypothesis Testing

Analysis of Variance and Covariance

- One-Way Analysis of Variance
- Assumption of ANOVA
- Statistics Associated with One-Way Analysis of Variance
- Interpreting the ANOVA Results
- Two-Way Analysis of Variance
- Interpreting the ANOVA Results
- Analysis ofCovariance

Data Visualization using R

- Grammar of Graphics
- Bar Charts
- Histograms
- Pie Charts
- Scatter Plots
- Line Plots and Regression
- Word Clouds
- Box Plots
- GGPLOT2

Data Science with R

Exploratory Data Analysis with R

- Merge, Rollup, Transpose and Append
- Missing Analysis and Treatment
- Outlier Analysis and Treatment
- Summarizing and Visualizing the Important Characteristics of Data
- Univariate, Bivariate Analysis
- Crosstabs, Correlation

Linear Regression

- What is Regression Analysis
- Covariance and Correlation
- Multivariate Analysis
- Assumptions of Linearity Hypothesis Testing
- Limitations of Regression
- Implementing Simple & Multiple Linear Regression
- Making Sense of Result Parameters
- Model Validation
- Handling Other Issues/Assumptions in Linear Regression
- Handling Outliers, Categorical Variables, Autocorrelation, Multicollinearity, Heteroskedasticity Prediction and Confidence Intervals

Logistic Regression

- Implementing Logistic Regression
- Making Sense of Result Parameters: Wald Test, Likelihood Ratio Test Statistic, Chi-Square Test Goodness of Fit Measures
- Model Validation: Cross Validation, ROC Curve, Confusion Matrix

Decision Trees

- Introduction to Predictive Modelling with Decision Trees
- Entropy & Information Gain
- Standard Deviation Reduction (SDR)
- Overfitting Problem
- Cross Validation for Overfitting Problem
- Running as a Solution for Overfitting

Linear Discriminant Analysis

- Multi-class classification

Data Science with Python

Pandas

- Introduction To Pandas
- IO Tools
- Basics Of Numpy
- Numpy Functions
- Pandas – Series and Dataframes

Basics of Python for Data Science

- Python Basics
- Data Structures in Python
- Control & Loop Statements in Python
- Functions & Classes in Python
- Working with Data
- Spyder IDE
- Jupyter Notebook
- Single and Multiline Comments

Exceptions and Files

- Exception Handling
- Raising Exceptions
- Assertions
- Working with Files

Data Frame Manipulation

- Data Acquisition (Import & Export)
- Indexing
- Selection and Filtering Sorting & Summarizing
- Descriptive Statistics
- Combining and Merging Data Frames

- Removing Duplicates
- Discretization and Binning
- String Manipulation

Exploration of Data Analysis

- Data Visualization & EDA

Dimensionality Reduction

- Principal Component Analysis (PCA)
- Scree Plot
- One-Eigenvalue Criterion
- Factor Analysis

Supervised Learning

- Linear Regression
- Linear Regression with Stochastic Gradient Descent, Batch GD
- Optimizing Learning Rate
- Momentum

Unsupervised Learnings

- K-Means Clustering

Linear Regression

- Implementing Simple & Multiple Linear Regression with Python
- Making Sense of Result Parameters
- Model Validation
- Handling Outliers, Categorical Variables, Auto-Correlation, Multicollinearity,
- Heteroskedasticity
- Prediction and Confidence Intervals
- Use Cases

Logistic Regression

- Logistic Regression with Stochastic Gradient Descent, Batch GD
- Optimizing Learning Rate
- Momentum
- Implementing Logistic Regression with Python
- Wald Test, Likelihood Ratio Test Statistic, Chi-Square Test
- Goodness of Fit Measures
- Model Validation: Cross Validation, Roc Curve, Confusion Matrix
- Use Cases

Decision Tree

- Fundamental Concepts of Ensemble
- Hyper-Parameters
- Implementing Decision Trees using Python
- Homogeneity
- Entropy
- Information Gain
- Gini Index
- Standard Deviation Reduction
- Vizualizing & Prunning a Tree

Random Forest

- Implementing Random Forests using Python
- Random Forest Algorithm
- Important Hyper-Parameters of Random Forest for Tuning the Model
- Variable Importance
- Out of Bag Errors

K Nearest Neighbour

- Understanding KNN
- Voronoi Tessellation
- Choosing K
- Distance Metrics – Euclidean, Manhattan, Chebyshev

Support Vector Machines

- What is SVM?
- When to use SVM?
- Understanding Hyperplane
- What is Support Vector?
- Understanding Lagrangian Multiplier, Karush Kuhn Tucker Conditions
- SVM Kernels – Radial Basis Function, Gaussian Kernel, Linear Kernel
- Optimizing the C Parameter
- Regularization

Time Series Forecasting

- Understand Time Series Data
- Visualizing Time Series Components
- Exponential Smoothing
- Holt's Model
- Holt-Winters Model
- ARIMA
- ARCH & GARCH
- ACF/PACF Functions

Data Visualization

- Basics of Data Visualization
- Line Plots
- Bar Charts
- Pie Charts
- Histograms
- Scatter Plots
- Parallel Coordinates

Machine Learning

Introduction to Machine Learning

- Machine Learning Modelling Flow
- How to treat Data in ML
- Parametric & Non-parametric ML Algorithm
- Types of Machine Learning
- Performance Measures
- Bias-Variance Trade-Off
- Overfitting & Underfitting
- Optimization Techniques
- Scikit-Learn Library

ML Algorithms – Supervised Learning

- Linear Regression with Stochastic Gradient Descent
- Logistic Regression with Stochastic Gradient Descent
- K-Nearest Neighbour
- Eager Methods Vs. Lazy Methods
- Nearest Neighbor Classification
- Building Kd-Trees
- Support Vector Machine
- Perceptron Algorithm

ML Algorithms – Unsupervised Learning

- What is Clustering?
- K-Means Algorithm
- Types of Clustering
- Evaluating K-Means Clusters

Ensemble Algorithms

- Ensemble Techniques
- Bootstrap Aggregation
- Random Forest
- Boosting

Neural Networks

- Understanding Neural Networks
- The Biological Inspiration
- Perceptron Learning & Binary Classification
- Backpropagation Learning
- Object Recognition

Hands-on Project Work

- Project #1: Real Estate Price Prediction using Linear Regression
- Project #2: Bankruptcy Prediction using Logistic Regression
- Project #3: Identifying Good and Bad Customers for Granting Credit Using Decision Trees
- Project #4: Forecasting and Predicting the Sales of Furniture of the Superstore

SAS Programming

Introduction to SAS and SAS Programs

- What is SAS?
- Key Features
- Submitting a SAS Program
- SAS Program Syntax Examining SAS Datasets Accessing SAS Libraries
- Sorting and Grouping Reporting Data
- Using SAS Formats

Reading and Manipulating Data

- Reading SAS Datasets
- Reading Excel Data
- Reading Raw Files
- Reading Database Data
- Creating Summary Reports
- Combining Datasets

Data Transformation

- Writing Observations
- Writing to Multiple Datasets
- Accumulating Total for a Group of Data
- Data Transformations

Macros

- Introduction to Macro Variables
- Automatic Macro Variables
- User Defined Macro Variables
- Macro Variable Reference

- Defining and Calling Macros
- Macro Parameters
- Global and Local Symbol Table
- Creating Macro Variables in the Data Step

SQL

- Introduction to SQL
- How Does RDBMS Work?
- SQL Procedures
- Specifying Columns
- Specifying Rows
- Presenting Data
- Summarizing Data
- Writing Join Queries using SQL
- Working with Subqueries
- Indexes and Views
- Set Operators
- Creating Tables and Views Using Proc SQL

Data Visualization with Tableau

Tableau Basics

- Introduction to Visualization
- Working with Tableau
- Visualization in Depth
- Data Organisation
- Advanced Visualization
- Mapping
- Enterprise Dashboards Data Presentation

Best Practices for Dashboarding and Reporting and Case Study

- Have a Methodology
- Know Your Audience
- Define Resulting Actions
- Classify Your Dashboard
- Profile Your Data
- Use Visual Features Properly
- Design Iteratively

Case Studies

1. Linear Regression – Boston Dataset – Using Sklearn Linear Model & Gradient Descent Model
2. Logistic Regression – Iris Dataset – Sklearn Logistic Model & Stochastic Average Gradient Descent
3. Decision Tree & Random Forest – Bank Marketing Dataset – Decision Tree Classifier, Random Forest Classifier, Adaboost Classifier & Bagging Classifier
4. KNN – Breast Cancer Dataset – KNN Classifier & How to Choose the K Value
5. SVM – Default of Credit Card Clients Dataset – SVM Classifier using Different Kernels (Linear, Polynomial, Radial Basis Function)
6. K-Means Clustering – Cars Dataset
7. Neural Network – Predict Close Value of Stock – Dow Jones Industrial Average (DIJA) Dataset

Hands on Projects

PROJECT 1

Property Price Prediction using Linear Regression in R

PROJECT 2

Bank Credit Card Default Prediction using Logistic Regression in R

Bankruptcy Prediction using Logistic Regression **:** *Use financial ratios to predict if a company is going to be bankrupt*

PROJECT 3

Predict Wine Quality with Decision Tree (Regression Trees *Classification Trees)

PROJECT 4

Multi-Class Classification with Linear Discriminant Analysis

PROJECT 5

Forecasting and Predicting the Furniture Sales using ARIMA

PROJECT 6

Reduce Data Dimensionality for a House Attribute Dataset using PCA

PROJECT 7

Use K-means Clustering to Group Teen Students into Segments for Targeted Marketing Campaigns

PROJECT 8

Real Estate Price Prediction using Linear Regression

PROJECT 9

Identifying Good and Bad Customers for Granting Credit using Decision Trees

PROJECT 10

Breast Cancer Prediction – KNN Classifier & How to Choose the K Value

PROJECT 11

Bank Marketing Analytics –Decision Tree & Random Forest Classifier

PROJECT 12

Default Prediction of Credit Card Clients – SVM Classifier using Different Kernels

PROJECT 13

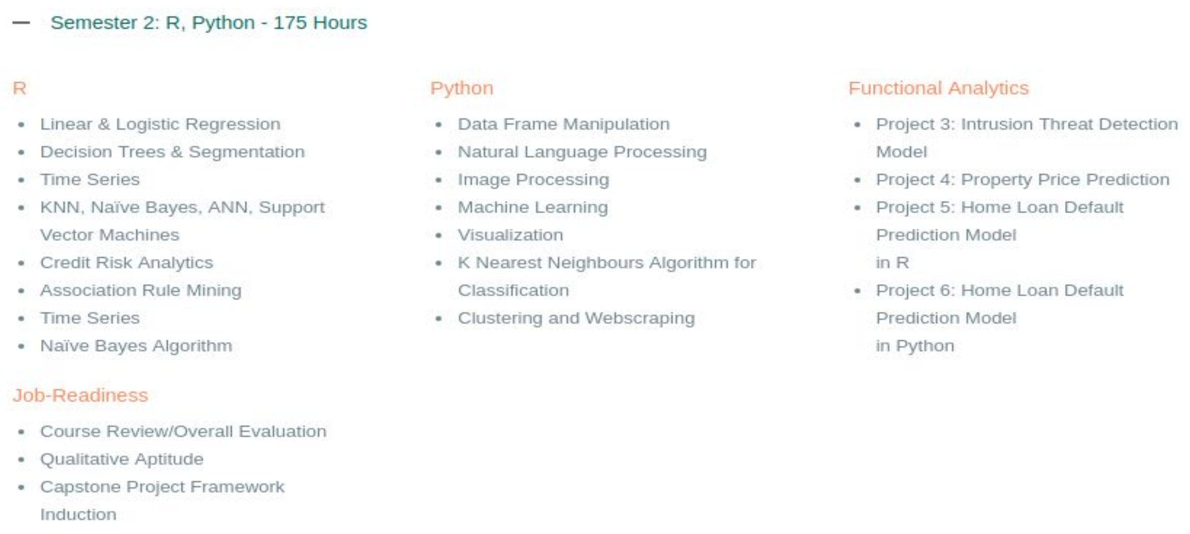
Store Data Analytics in SAS

PROJECT 14

Building Tableau Dashboard

PROJECT 15

Forecasting the Sale of Furniture of a Superstore - Time Series Forecasting



— Semester 2: R, Python - 175 Hours

R

- Linear & Logistic Regression
- Decision Trees & Segmentation
- Time Series
- KNN, Naïve Bayes, ANN, Support Vector Machines
- Credit Risk Analytics
- Association Rule Mining
- Time Series
- Naïve Bayes Algorithm

Job-Readiness

- Course Review/Overall Evaluation
- Qualitative Aptitude
- Capstone Project Framework Induction

Python

- Data Frame Manipulation
- Natural Language Processing
- Image Processing
- Machine Learning
- Visualization
- K Nearest Neighbours Algorithm for Classification
- Clustering and Webscraping

Functional Analytics

- Project 3: Intrusion Threat Detection Model
- Project 4: Property Price Prediction
- Project 5: Home Loan Default Prediction Model in R
- Project 6: Home Loan Default Prediction Model in Python