DATA SCIENCE AND ADVANCED TABLEAU

1) Advanced Excel

a) Introduction to MS Excel, Cell Ref, Basic Functions and Usage
b) Sorting, Filtering, Advance Filtering, Subtotal
c) Pivot Tables and Slicers
d) Goal Seek and Solver
e) Different Charts Graphs – Which one to use and when
f) Vlookup, Hlookup, Match, Index
g) Conditional Formatting
h) Worksheet & Workbook Reference, Error Handling
i) Logical Operators & Functions – IF and Nested IF
j) Data Validation
k) Text Functions
l) Form Controls
m) Dashboard
n) 6 Case Studies from App Cab Aggregators, Insurance, Sports, Sales, Marketing, Web Analytics Industry

2) SQL Queries & Relational Database Management

a) Relational Database Fundamentals
b) Steps to Design Efficient Relational Database Models
c) Case Studies on Designing Database Models
d) Case Study Implementation on Handling Data
e) Importing / Exporting Large Amount of Data into a database
f) SQL Statements - DDL, DML, DCL, DQL
g) Writing Transactional SQL Queries, Merging, joining, sorting, indexing, co-related queries, etc.
h) Hands-on Exercises on Manipulating Data Using SQL Queries
i) Creating Database Models Using SQL Statements
j) Individual Projects on Handling SQL Statements
k) 6 Case Studies from App Cab Aggregators, Ecommerce, Sports Industry

3) Tableau

a) Introduction to Data Visualization
   i) What is Dashboard
   ii) Why do we need Dashboard
b) Introduction of Data Visualization using Tableau
   i) Use of Tableau
   ii) Navigation in Tableau
   iii) Exporting Data
   iv) Connecting Sheets
c) Tableau Basics
   i) Working with Dimension and Measures
   ii) Making Basic Charts like Line, Bar etc.

   iii) Adding Colours
   iv) Working in marks card
d) Working with Sorting and Filters
e) Creating Dual Axis and Combo Charts
   i) Working with Tables
   ii) Creating Data Tables
f) Table Calculations
g) Calculated Field
h) Logical Calculations
   i) If/Then
   ii) IIF
   iii) Case/When
i) Date Calculations
   i) Date
   ii) DateAdd
   iii) DateDiff
   iv) DateParse
   v) Today()
   vi) Now()
j) Parameters
   i) Pre-defined Lists for Faster Filtering
   ii) Top N Filter
   iii) Reference Line Parameter
   iv) Swapping Dimensions or Measures in a View
k) Using Actions to Create Interactive Dashboards
   i) Filter Actions
   ii) Highlight Actions
l) Advanced Charts
   i) Heat maps, Tree Maps, Waterfall Charts etc.
   ii) Working with latitude and Longitude
   iii) Symbol and filled maps
m) Working with data
   i) Joining multiple tables
   ii) Blending of Data
n) Sets
   i) In/Out Sets
   ii) Combines Sets
o) Drilling Up/Down using Hierarchies
p) Grouping
q) Bins/Histograms
r) Analytics
   i) Referencing lines
   ii) Clustering
   iii) Trend Line

s) Building dashboards
   i) Layout and Formatting
   ii) Interactivity with Actions
   iii) Best Practices
t) Story Telling with Data
   i) Working with story
   ii) Highlighting important insights
u) Data Interpreter
   i) Data Preparation
   ii) Data Cleaning
   iii) Pivoting
v) 4 Case Studies on Retail, Airline, Bank datasets

4) Business Statistics

a) Types of data, Graphical representation
   i) Introduction of data
   ii) Types of data
   iii) Data Presentation
   iv) Charts & Diagrams
   v) Assignment on Type of Data and Type of Charts
b) Correlation, Data Modeling & Index Numbers
   i) Correlation
   ii) Data Modeling
   iii) Index Number
c) Measures of Central Tendency & Dispersion
   i) Measures of Central Tendency
   ii) Measures of Central Dispersion
   iii) Measures of Central Dispersion (Variance)
   iv) Normal Distribution
   v) Assignment of Central Tendency and Dispersion
d) Forecasting & Time Series Analysis
   i) Forecasting
   ii) Components of time Series
   iii) Measurement of Secular Trend
   iv) Forecasting Software
e) Probability, Bayesian Theory
   i) Probability
   ii) Computing joint & marginal probabilities
   iii) Bayes' Theorem
f) Probability Distribution and Mathematical Expectation
   i) Random Variables
   ii) Probability Distribution (Discrete)

iii) Probability Distribution (Continuous)
iv) Finding Normal Probabilities

g) Sampling and Sampling Distribution
i) Sample, Types of sample
ii) Sampling Distribution
iii) Example of Sampling
iv) Assignment on Probablity Distribution, Binomial & Poisson, Normal Distribution

h) Theory of Estimation and Testing of Hypothesis
i) Theory of Estimation, Estimation Process,Statistical Inference
ii) Test of Hypothesis, Decision Errors, OneLevel of Significance
iii) Two-tail test, Testing of hypothesis
iv) Degrees of freedom

i) 9. Analysis of Variance
i) Anova
ii) Hypothesis - One way Anova
iii) Two way Anova
iv) Assignment on Hypothesis Testing

j) Regression Models
i) Regression, Linear Regression, Multiple Linear Regression
ii) Coefficient of Determination, R-square, Adjusted R-sqare
iii) Example using Excel
iv) Assignment on Corelationa & Simple Regression

5) Predictive Modeling with R

a) Introduction to R
i) General introduction to R and R Packages
ii) Installing R in Windows
iii) Installing R packages through R using syntax
iv) Basic syntaxes in R

b) Data Handling in R
i) Creating Dataframe
ii) Variables in R
iii) Creating columns with conditions AND, OR
iv) Different numeric functions in R like exp, log, sqrt, sum, prod etc. Sorting in R. Ranking and concatenating strings in R.
v) Exercises on Import / Export of Data
vi) Exercises on Data Handling in R

c) Overview of Analytics and Statistics
i) Types of data variables
ii) What is Population
iii) Mean, Median, or Mode – Their applications
iv) Basic Statistics Exercises

d) String and character functions in R
   i) Substring, string split
   ii) Change name of column and checking mode of variable
   iii) Dividing variable into different buckets
   iv) Creating user defined functions in R
   v) Loops in R
   vi) SQL in R using sqldf
   vii) Scatter plot, Box plot, Histogram, pie chart in R T Test in R
   viii) Exercise: Data Summarization using Financial Retail Datasets
e) Overview of Analytics and Statistics
   i) Standard deviation interpretation
   ii) Population vs Sample
   iii) Univariate & Bivariate Analysis
   iv) Normal distribution
   v) What is Confidence Interval
   vi) Hypothesis Testing
   vii) In-Case Study: Academic Performance Case Study
   viii) Self-Case Study: Health Care Case Study
f) Linear regression in R
   i) Regression
   ii) Residual Analysis
   iii) Multiple Regression
   iv) Model Building
   v) In-class Case Study: Predict Academic Performance of School Students
   vi) Self Case Study: Predict Customer Value for an Insurance Firm
g) Logistic Regression in R
   i) Model theory, Model Fit Statistics
   ii) Reject Reference, Binning, Classing
   iii) Dummy Creation, Dummy Correlation
   iv) Model Development (Multicolinearity, WOE, IV, HLT, Gini KS, Rank Ordering, Clustering Check)
   v) Model Validation (Rerun, Scoring)
   vi) Final Dashboard
   vii) In-class Case Study: Predict Customer Churn for a Telecom firm
   viii) Self Case Study: Predict Propensity to Buy Financial Product among Existing Bank
h) Time Series theory discussion overview
   i) ARIMA, Stationarity & Non stationarity check concepts
   ii) forecasting
   iii) components of Time Series
   iv) Measurement of Circular Trend
   v) Time Series codes overview
   vi) Exponential smoothening theory discussion
   vii) Case Study - Random walk in Time Series

viii) Case Study - Forecasting sales for retail

i) Clustering Concepts and Case Study
   i) K-means Clustering
   ii) Types of Clustering
   iii) Centroids
   iv) Case Study - Airline customer segmenation

j) Feature Engineering & Dimension Reduction and Case Study
   i) Factor Analysis
   ii) PCA
   iii) Methods of Variable Reduction
   iv) Dimensionality Reduction

k) Decision Trees
   i) Pre-reading on basics of segmentation and decision trees
   ii) Intro to Objective Segmentation
   iii) CHAID and CART concept, example, and exercise
   iv) Implement Decision Trees
   v) Advantages and disadvantages of Decision Trees over Prediction
   vi) Multiple Decision Trees
   vii) Case Study – Predict earning of an individual

6) Python for Data Science

a) Python Essentials
   i) Overview of Python- Starting with Python
   ii) Introduction to installation of Python
   iii) Introduction to Python Editors & IDE's(Canopy, pycharm, Jupyter, Rodeo, Ipython etc…)
   iv) Understand Jupyter notebook & Customize Settings
   v) Concept of Packages/Libraries - Important packages(NumPy, SciPy, scikit-learn, Pandas, Matplotlib, etc)
   vi) Installing & loading Packages & Name Spaces
   vii) Data Types & Data objects/structures (strings, Tuples, Lists, Dictionaries)
   viii) List and Dictionary Comprehensions
   ix) Variable & Value Labels – Date & Time Values
   x) Basic Operations - Mathematical - string - date
   xi) Reading and writing data
   xii) Simple plotting
   xiii) Control flow & conditional statements
   xiv) Debugging & Code profiling
   xv) How to create class and modules and how to call them?

b) Scientific Distribution
   i) Numpy, scify, pandas, scikitlearn, statmodels, nltk etc

c) Accessing / Importing and Exporting Data using Python modules
   i) Importing Data from various sources (Csv, txt, excel, access etc)

ii) Database Input (Connecting to database)
iii) Viewing Data objects - subsetting, methods
iv) Exporting Data to various formats
v) Important python modules: Pandas, beautifulsoup

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

e) Visualization using Python
i) Introduction exploratory data analysis
ii) Descriptive statistics, Frequency Tables and summarization
iii) Univariate Analysis (Distribution of data & Graphical Analysis)
iv) Bivariate Analysis(Cross Tabs, Distributions & Relationships, Graphical Analysis)
v) Creating Graphs- Bar/pie/line chart/histogram/ boxplot/ scatter/ density etc)
vi) Important Packages for Exploratory Analysis(NumPy Arrays, Matplotlib, seaborn, Pandas and scipy.stats etc)

f) Introduction to Predictive Modeling
i) Concept of model in analytics and how it is used?
ii) Common terminology used in analytics & modeling process
iii) Popular modeling algorithms
iv) Types of Business problems - Mapping of Techniques
v) Different Phases of Predictive Modeling

g) Modeling on Linear Regression
i) Introduction - Applications
ii) Assumptions of Linear Regression
iii) Building Linear Regression Model
iv) Understanding standard metrics (Variable significance, R-square/Adjusted R-square, Global hypothesis ,etc)
v) Assess the overall effectiveness of the model
vi) Validation of Models (Re running Vs. Scoring)
vii) Standard Business Outputs (Decile Analysis, Error distribution (histogram), Model equation, drivers etc.)

viii) Interpretation of Results - Business Validation - Implementation on new data

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

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

7) Machine Learning

a) Supervised learning : Decision Tree
   i) Decision Trees - Introduction - Applications
   ii) Types of Decision Tree Algorithms
   iii) 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
   iv) Generalizing Decision Trees; Information Content and Gain Ratio; Dealing with Numerical Variables; other Measures of Randomness
   v) Pruning a Decision Tree; Cost as a consideration; Unwrapping Trees as Rules
   vi) Decision Trees - Validation
   vii) Overfitting - Best Practices to avoid
   viii) Case Study on Decision Tree

b) Supervised Learning : Ensemble Learning
   i) Concept of Ensembling
   ii) Manual Ensembling Vs. Automated Ensembling
   iii) Methods of Ensembling (Stacking, Mixture of Experts)
   iv) Bagging (Logic, Practical Applications)

v) Random forest (Logic, Practical Applications)
vi) Boosting (Logic, Practical Applications)
vii) Ada Boost
viii) Gradient Boosting Machines (GBM)
ix) XGBoost
x) Case Study on Random Forest

c) Text Mining and Analytics
i) 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)
ii) Finding patterns in text: text mining, text as a graph
iii) Natural Language processing (NLP)
iv) Text Analytics – Sentiment Analysis using Python
v) Text Analytics – Word cloud analysis using Python
vi) Text Analytics - Segmentation using K-Means/Hierarchical Clustering
vii) Text Analytics - Classification (Spam/Not spam)
viii) Applications of Social Media Analytics
ix) Metrics(Measures Actions) in social media analytics
x) Examples & Actionable Insights using Social Media Analytics
xi) Important python modules for Machine Learning (SciKit Learn, stats models, scipy, nltk etc)
xii) Fine tuning the models using Hyper parameters, grid search, piping etc.
xiii) Case Study on Text Analytics