DATA SCIENCE

Kelly Technologies

1. Introduction to Data Science
   a. What is data science?
      - ✓ How is data science different from Bi and Reporting?
   b. Who are data scientists?
      - ✓ What skillsets are required?
   c. What do they do?
      - ✓ What kind of projects they work on?

2. Business statistics
   a. Data types
      - ✓ Continuous variables
      - ✓ Ordinal Variables
      - ✓ Categorical variables
      - ✓ Time Series
      - ✓ Miscellaneous
   b. Descriptive statistics
   c. Sampling
      - ✓ Need for Sampling?
      - ✓ Different types of Sampling
      - ✓ Simple random sampling
      - ✓ Systematic sampling
      - ✓ Stratified Sampling
   d. Data distributions
      - ✓ Normal Distribution – Characteristics of a normal distribution
      - ✓ Binomial Distribution
   e. Inferential statistics
   f. Hypothesis testing
      - ✓ Type I error
      - ✓ Type II error
      - ✓ Null and alternate hypothesis
      - ✓ Reject or acceptance criterion

3. Introduction to R
   a. A Primer to R programming
   b. What is R? similarities to OOP and SQL
   c. Types of objects in R – lists, matrices, arrays, data.frames etc.
   d. Creating new variables or updating existing variables
   e. IF statements and conditional loops - For, while etc.
   f. String manipulations
   g. Sub setting data from matrices and data.frames
   h. Casting and melting data to long and wide format.

i. Merging datasets

4. Exploratory data analysis and visualization
   a. Getting data into R – reading from files
   b. Cleaning and preparing the data – converting data types (Character to numeric etc.)
   c. Handling missing values – Imputation or replacing with place holder values
   d. Visualization in R using ggplot2(plots and charts) – Histograms, bar charts, box plot, scatterplots
   e. Adding more dimensions to the plots
   f. Visualization using Tableau(Introduction)
   g. Correlation – Positive , negative and no correlation
   h. What is a spurious correlation
   i. Correlation vs. causation

5. Introduction to Python:
   a. Understanding the reason of Python's popularity
   b. Basics of Python: Operations, loops, functions, dictionaries
   c. Advanced operations with text: Finding, Sequencing and basic analytics
   d. Ground-up for Deep-Learning

6. Predictive analytics
   a. Different types of predictive analytics – prediction, forecasting, optimization, segmentation etc.
   b. Supervised learning
      ✓ Prediction (Linear)
      1. Simple Linear Regression
      2. Assumptions
      3. Model development and interpretation
      4. Sum of least squares
      5. Model validation – tests to validate assumptions
      6. Multiple linear regression
      7. Disadvantages of linear models

      Classification
      1. Logistic Regression
      1. Need for logistic regression
      2. Logit link function
      3. Maximum likelihood estimation
      4. Model development and interpretation
      5. Confusion Matrix – error measurement
      6. ROC curve

7. Measuring sensitivity and specificity
8. Advantages and disadvantages of logistic regression models

2. Decision trees
   1. C5.0
   2. Classification and Regression trees(CART)
      a. Process of tree building
      b. Entropy and Gini Index
      c. Problem of over fitting
      d. Pruning a tree back
      e. Trees for Prediction (Linear) – example
      f. Tress for classification models – example
      g. Advantages of tree based models?

3. KNN – K nearest neighbors
   1. Advantages and disadvantages of KNN

c. Re-Sampling and Ensembles Methods
   1. Bagging
   2. Random Forests
   3. Boosting – Gradient boosting machines

b. Advanced methods
   1. Support Vector machines
   2. Neural networks
   3. Introduction to deep learning
   4. Introduction to online learning

d. Un-Supervised learning

   Cluster analysis
   1. Hierarchical clustering
   2. K-Means clustering
   3. Distance measures
   4. Applications of cluster analysis – Customer Segmentation

e. Time series analysis - Forecasting
   1. Simple moving averages
   2. Exponential smoothing
   3. Time series decomposition
   4. ARIMA
   Collaborative filtering

  5. User based Filtering
  6. Item based Filtering

7. Model validation and deployment
 a. Error measurement
   1. RMSE – Root Mean squared error
   2. Misclassification rate
   3. Area under the curve (AUC)

8. Practical use cases and best practices
 a. Business problem to an analytical problem
  ✓ Problem definition and analytical method selection
 b. Guidelines in model development

9. Introduction to big-data and other tools (Python and R-Server)
 a. Big data and analytics?
  ✓ Leverage Big data platforms for Data Science
 b. Introduction to evolving tools e.g Spark
  ✓ Machine learning with Spark

10. Introduction to Azure cloud and Big-Data computing over cloud
 a. Creation of R-Server clusters
 b. Computation of Big-Data ML algorithms over the Azure cloud

11. Introduction to Deep Learning
 a. What is DL and how does it score better over traditional MLs?
 b. Convolutional and Perceptron models
 c. Comparison between DL and ML performances over the MNIST dataset

12. Analytical Visualisation with Tableau
 a. Why is it important for Data-Analyst
 b. Tableau workbook walkthrough
 c. Instruction of creation of your own workbooks
 d. Demo of few more workbooks

13. Offerings from Kelly.
 a. Mock interviews questions and case studies walkthrough over Azure Cortana gallery
 b. Guidance to prepare resumes
 c. Information on companies and industry trends on data science