Machine Learning with Python

01 LINEAR REGRESSION

- What are different types of Regression Analysis?
- How to fit Simple linear regression?
- What are key assumptions underlying Linear Regression?
- How to validate if model assumes each assumption?
    - Residual Plots
    - Normal QQ Plots
    - Auto-correlation
    - Constant Variance
    - Outliers - Leverage
- What is Multiple Linear Regression?
- What is Interaction between variables?
- What is Polynomial Regression?
- What are different components of regression summary & what they mean?
- How to balance in Model Accuracy and Interpret ability?
    - Subset Selection (variables)
    - Bias Variance Trade-off
    - Under-fitting vs over-fitting
- How to choose the Optimal Model (Indirect Estimate) from train metrics?
- How to compare between different models via Indirect Estimates?
- How to choose the Optimal Model (Direct Estimate) from test metrics?
    - Leave On Out – Cross Validation
    - K-fold Cross Validation
- What is Shrinkage / Regularization? Ridge - L2, Lasso - L1
- What is Dimension Reduction

02 LOGISTIC REGRESSION

- Intuition behind modifying linear regression to fit binary outcome
- Extending OLS to Generalized Linear Models
- Assumptions in GLM
- What is link functions?
- Logit function explained
- Probability and Odds
- Ordinary least squares to Maximum likelihood estimator
- Specialized Variants of Regression

03 DECISION TREES

- Intuition of Decision Tree - Basics of if-else rule and Information value (IV)
- How to measure Information Value –
    - Variance in Continuous target
    - Impurity in categorical target

- Entropy and Gini Index (Basics and mathematics behind it)
- Information Gain as driving force for decision making
- How Decision tree propagates
- Predictor space Segmentation
- Components of Decision Tree - Explained
- Over fitting issue & remedies
    - Pre-Pruning (Tree Depth, Max. no of leaves & Samples in leaf)
    - Post-Pruning
- Ensemble Trees - Why they are so valuable?
    - Bagging (Bootstrap Aggregation)
- Feature importance (Calculation in Single tree vs Ensemble tree)

04 BAGGING, RANDOM FOREST & BOOSTED TREES

- Ensemble Trees - Why they are so valuable?
    - Bagging (Bootstrap Aggregation)
- Feature importance (Calculation in Single tree vs Ensemble tree)
- Basics of Bagging, Random Forest and Ada-Boost
- Tuning parameters in Ensemble Trees

06 K-MEANS

- Intuition of neighborhood concept - Euclidean Distance
- Minimizing “within sum of squares” – Inertia – Heterogeneity
- K-means ++ to overcome local minima issue with K-means
- Elbow method to find best no of clusters
- Tuning parameters in K-means
- Cost/Time complexity in K-means

07 HIERARCHICAL CLUSTERING

- What is Hierarchical Tree?
- Clustering in the form of Hierarchical Tree.
- Variants of Hierarchical Tree?
    - Agglomerative
    - Divisive
- What is dendrogram?
- Process of Agglomerative Algorithm and mathematics behind it (distance matrix)

08 DBSCAN CLUSTERING

- Concept of leaf, branch, core points, border points and noise
- Tuning parameters in DBSCAN (ε, Minimum Points)
- Process of DBSCAN Algorithm

09 TIME SERIES ANALYSIS

- Rollup data to specific time step (hourly, daily, weekly, monthly)
- Parse dates
- Check missing time periods
- Clean data if required
- Check Stationary (Plotting Rolling Statistics or Dickey-Fuller Test)
- Log transformation
- Aggregation (MA,WMA, EWMA)
- Smoothing
- Polynomial Fitting
- Eliminating Trend and Seasonality (Differencing / Decomposition)
- Find auto-regressive terms (p) - PACF
- Find no. of differences (d) -ACF
- Find no. of MA (q)
- Build MA Model
- Build ARMA Model
- Build ARIMA Model
- Build S-ARIMAX Model
- Residual Analysis
- Add back trend & seasonality
- Evaluate Model performance
- Forecasts for future time period
- Plot forecasts with confidence intervals

10 TEXT MINING

- Connecting API and fetching data
- Scraping Data
- Data parsing (Txt, JSON, PDF)
- Sentence & Word Tokenization
- Lower case words
- Remove HTML Characters, Punctuation, Stop-words, Digits, Low frequency words & Word Length < 3
- Split attached words
- Standardizing words (Repeating letters)
- Stemming or Lemmatization
- Word Frequency & Word Cloud
- Word Embedding
- POS tagging
- Generate
- Vocabulary
- Create DTM (tf)
- Create tf-idf
- Cosine Similarity
- Model Learning & Model Deployment