Topics - Data Analytics / Data Science

Saturday, October 9, 2021   7:51 PM

Data Science Project Pipeline
_____

- [ ] EDA
  - [ ] Description of data
  - [ ] Correlation, covariance, VIF
  - [ ] Univariate, bivariate analysis
- [ ] Missing value treatment
- [ ] Outlier Treatment
- [ ] Feature Engineering
  - [ ] Backpropogation
  - [ ] Feature scaling
  - [ ] Normalization
  - [ ] Transformation
  - [ ] standardization
  - [ ] Data Visualization
    - [ ] Plots
    - [ ] Heatmap
    - [ ] Bar charts, line charts
    - [ ] Qq plots
  - [ ] Sampling
- [ ] Feature Selection
- [ ] Model creation
- [ ] Algorithm selection
- [ ] Model evaluation
  - [ ] Precision and recall
  - [ ] Confusion matrix
  - [ ] AUC ROC curse
  - [ ] Type 1 and Type 2 error
  - [ ] F1 score
  - [ ] Gini impurity & KS statistic
  - [ ] Different types of errors
  - [ ] R squared, adjusted R squared
  - [ ] Information Gain
  - [ ] Entropy, AIC
- [ ] Model validation
  - [ ] kfold
- [ ] Hyperparameter Tuning
  - [ ] Grid search - random, stratified
- [ ] Model Deployment

_____

Statistics
_____

- [ ] Descriptive Statistics
  - [ ] Mean, median, mode
  - [ ] Standard Deviation, Variance
  - [ ] Range
  - [ ] Boxplots
  - [ ] histogram
- [ ] Inferential Statistics
  - [ ] Sample and population
  - [ ] Hypothesis Testing
    - [ ] Confidence interval
    - [ ] P value
    - [ ] Significance level
  - [ ] Central limit theoram
  - [ ] Normal Distribution
  - [ ] Binomial, poission, gaussian etc.
  - [ ] A/B testing
  - [ ] Skewness and Kurtosis
  - [ ] probability
  - [ ] Statistical tests
    - [ ] Z-test
    - [ ] T-test
    - [ ] Chi-square test
    - [ ] Annova
    - [ ] One & two tailed tests

Supervised ML
_____

- [ ] Algorithms
  - [ ] Regression
    - [ ] Linear
    - [ ] Ridge and Lasso
    - [ ] Polynomial regression
    - [ ] Multi-linear
    - [ ] Non-linear
    - [ ] OLS
    - [ ] Time series forecasting
  - [ ] Classification
    - [ ] Logistic
    - [ ] kNN
    - [ ] Decision trees
    - [ ] Naïve bayes
    - [ ] Random Forest
    - [ ] XGBoost
    - [ ] Light GBM
    - [ ] GBM
    - [ ] SVM
    - [ ] ANN
    - [ ] Catboost, adaboost
- [ ] Concepts
  - [ ] Ensemble techniques
  - [ ] Bootstraping, stacking
  - [ ] Bagging and Boosting
  - [ ] Regularization
  - [ ] Gradient descent
  - [ ] Loss function
  - [ ] Overfitting underfitting
  - [ ] Multicollinearity
  - [ ] Assumptions of linear regression
  - [ ] Vanishing gradient
  - [ ] Learning rate
  - [ ] Exploding gradient
  - [ ] Batch normalization
  - [ ] Bias-Variance Tradeoff
  - [ ] ARIMA Forecasting
    - [ ] Moving average
    - [ ] Smoothing exponential

Unsupervised ML
_____

- [ ] Clustering
  - [ ] Kmeans
  - [ ] DB Scan
  - [ ] heirarchial
- [ ] Appriori - association rule mining
- [ ] Recommendation systems
  - [ ] Content filtering
  - [ ] Collaborative filtering
  - [ ] Hybrid filtering
- [ ] Dimensionality reduction
  - [ ] PCA
    - [ ] Orthogonal vectors
    - [ ] Eigen vector & eigenvalue
  - [ ] t-SNE
  - [ ] ICA
  - [ ] SVD
  - [ ] Autoencoding
- [ ] Concepts
  - [ ] Elbow curve
  - [ ] Silhouette curve
  - [ ] Gaussian distribution
  - [ ]

Business case studies/ problems
_____

- [ ] Customer segmentation
- [ ] Customer retention
- [ ] Propensity modeling
- [ ] Sentiment analysis / Emotion mining
- [ ] Fraud detection
- [ ] Anomaly detection
- [ ] Upsell cross-sell
- [ ] Time series Forecasting
- [ ] Market basket analysis
- [ ] Recommendar systems
- [ ] LTV - lifetime value prediction
- [ ] Price sensitivity
- [ ] Root cause analysis
- [ ] Stock prediction
- [ ] Weather forecasting
- [ ] Credit risk modeling