**ML Flow**

* Input file Source (CSV/SQL/TXT..etc)
* Problem statement
* Problem type
  + Supervised (Regression/Classification)
  + Unsupervised (Clustering)
* Target identification
* choose evaluation metric
* EDA
  + Get missing value columns and proportions
  + Plot (libraries: ggplot2)
  + Univariate
    - boxplot (what are q1, q2, q3, q4, IQR, median, outliers?)
    - scatterplot (to know the pattern/trend)
    - histogram (to see the distribution)
  + Bivariate
    - scatter plot (to check correlation)
  + Multivariate
    - correlation matrix (heat map)
  + Variable interaction
    - weight of evidence and information value(r remaining)
    - chi square
    - anova
    - correlation
    - vif (collinearity)
  + Missing value handling (package: mice)
    - why missing? why to handle missing (in ppt description of each slide)
    - by domain
    - mean/median/mode
    - more in the ppt slide and description
  + Outliers:
    - why to handle? what are outliers?
    - how to handle (box plot/mean/median/mode/remove)
* feature engineering
  + date interactions
  + domain features
  + transformed features (log/square/cube/square root..etc)
  + binning
  + one hot encoding/label encoding
* Scaling techniques:
  + min-max
  + (x-mu)/sigma
* Imbalanced:
  + what is imbalanced
  + what problem you get
  + how to handle? (oversampling/undersampling/SMOTE)
* Models:
  + naïve bayes
    - R package and syntax
    - assumptions
    - concepts: prior prob/likelihood/marginal likelihood
  + linear regression (In R: lm)
    - R Syntax
    - durbin watson test: range = (0, 4)
    - assumptions
    - coefficients interpretation
    - output explain
    - advantages
    - disadvantages
    - plot the fit object and check for assumptions
      * residual vs fitted:
        + should not have any pattern
        + if curve present - non linearity in data(then apply polynomial like multiplying with x2,log,exp …..)
        + if funnel/cone: heteroscadasticity: means highbias(underitting)
      * normality q-q plot(applied for residuals,has to foem bell curve means most values are found at 0)
      * scale location plot
  + Decision trees (packages: rpart: recursive partitioning and regression trees)
    - R syntax
    - advantages: interpretability,
    - disadvantages: overfitting, cant extend trend in regression
    - split criterion differences (gini vs entropy)
    - tree types differences (chaid vs c4.5 vs cart vs id3)

|  |  |  |  |
| --- | --- | --- | --- |
| ID3 | CART (R uses RPART) | C4.5 | CHAID |
| Iterative Dichotomizer | Classification and Regression Trees |  | CHisquare Automatic Interaction Detection |
| Uses info gain | Uses Gini | Uses Gain Ratio | Uses Chi-square |
| Cant handle missing | Can handle | Can | Can |
| Only numeric input | Num + Categorical | N + C | N +C |
| Cant handle outliers | Can | Can |  |
| No pruning | Cost based | Error based |  |
|  |  |  |  |

* how tree handles missing values: surrogate splits
* pruning techniques:
  + pre-pruning: advantages/disadvantages
  + post-pruning: types (error based and cost complexity pruning), advantages and disadvantages
* plot and check trees
* Ensemble models:
* rpackages and syntax
* advantages and disadvantages of each
* boosting (adaboost/gradient/xgb/light gbm)
* bagging
* random forest
* differences between each
* variable importance from each models
* Neural networks
* layers: input/hidden/output
* concepts: dropout/activation function (type + advantages + differences)
* parameters: learning rate/number of layers/number of nodes in each layer
* perceptron
* fully connected
* advantages and disadvantages
* PCA
* R syntax for building and prediction
* concept and assumptions
* advantages and disadvantages
* what is eigen vector and eigen value
* default rotation used: varmax
* what does each component represent
* how dimensions are selected to be reduced
* Validation: (KFold/Out of bag)
* why
* how
* package: caret
* Clustering (In R: kmeans/hclust)
* R packages and syntax
* advantages and disadvantages
* distance and similarity measures (formulae/when and where to use what metric)
* how to select the number of clusters !
* how to evaluate the clusters
* Kmeans
* DBSCAN
* Hierarichal
* PAM
* Metrics: Silhoutte score/hopkins test/scree plot - formulae/ranges/interpretation
* Time series
* Additive vs multiplicative
* stationary
* components
* ARIMA
* ACF and PACF
* Holt winters
* Time series to supervised
* Metrics:
* Confusion matrix (document)
* precision/recall/accuracy/roc/auc/fscore - definition and formulae and range of values
* rmse/mse/mape/r-square/adjusted r-square - definition, formulae and range of values
* AIC/BIC: (to compare models)
  + - * goodness of fit, penalizes complex model, BIC penalizes more than AIC so BIC is better!
* Concordance/Discordance
* Handle overfitting:
* depth control
* more data
* parameter tuning
* regularization:
  + lasso (L1) : Least absolute selection and shrinkage operator, Range = [0, infinity)
    - on increasing L1, coef becomes zero
  + ridge (L2): [0, infinity)
    - on increasing L2, coef becomes close to zero
    - stable so better and generalizes better
* Concepts:
* Parameter tuning: document
* Gradient descent (Vanilla/Batch/Stochastic)
* loss function (MSE/Logloss) - formulae
* Bias - Variance trade off
* bias is inversely proportional to variance
* low bias = high variance = complex model = overfitting
* high bias = low variance = base learner = underfitting
* Parametric vs Non parametric models
* Models that can be represented as f(X) = Y they are parametric (logistic, svm, neural networks)
* Otherwise non parametric models (decision tree, bagging, xgb..etc)
* Parametric models have high variance and low bias
* non parametric models have high bias and low variance
* Multi class classification (one vs one/ one vs rest)
* ggplot
  + - layers
    - syntax for all types of plots
    - scatter/bar/stacked bar/line