DECISSION TREE – Most commonly used in industries

1. One of the methods which is client favorite- (graphical representation)
2. Simplest model to understand and implement

IMP variables in Decision tree -> Root node, Sub-node, Depth of the tree

Entropy – (IG) – Information Gain – Degree of measuring randomness of data

Gini index –

**A] Information Gain – how much information are you gathering**

**Information gain (IG)** measures how much “information” a feature gives us about the class.

**Entropy** is the **measures of impurity, disorder or uncertainty** in a bunch of examples.

1] More the variation in data more will be the Entropy



**B] GINI INDEX**

A picture containing text, clock, watch

Description automatically generated

2] Information Gain:-

Is how much amount of information we are getting after splitting th’e node.

3]Overfitting [ONLY CONTROLING DEPTH OF THE TREE]

Pruning method to handle ovefitting and depth of the tree there are two types of pruning methods.

A] Pre-pruning 🡪 Top up approach

B] Post-pruning 🡪 Bottom up approach 🡪 is Approchable methods

Above methods are controlling on depth of the tree

4]Hyper-parameter tuning (90% of the time)

Parameter that can be controlled

- Depth of the tree

- Split 🡪 one can control the split by 🡪 min\_sample\_split and min\_sample\_leaf there are total 16 ways to control it.

- IG/ Entropy

- min\_sample\_split

- min\_sample\_leaf

- max\_Leaf

- max\_feature

5]Search 🡪 the depth of the tree -> we need to define upper and lower limit to hyperparameter

-Depth of the tree

- Random Search

- Grid Search

6] How to effectively split the data

Train - Validation – Test 🡪 not effectivey for small data set

70% - 15% - 15%

Model classification

Train- Validation

Final model will be tested on the test dataset

We will be applying K-foldon the top of the Grid search

Text

Description automatically generated

Text, letter

Description automatically generated

7] K-fold Validation

K – stands for no of folds you want to split your train data

A screenshot of a computer

Description automatically generated with medium confidence

DECISION TREE – MODEL BUILDING – FLOW

1. Load the data set / csv/ using pandas / then read the file
2. Data preprocessing 🡪 Missing values/Outliers detection/Feature Engineering/Normal distribution/ data prechecks)
3. EDA – Exploratory data analysis 🡪 Visualizations/Tables/graphs/normal distribution/correlation/skewness/courtliness
4. Based on EDA – Will build a Report
5. Model Building

5.1 – Train-test split

5.2 – If required do the scaling on the train dataset

5.3 – Import the required library and then fit he model

5.4 – Hyper parameter tunning

a- K-fold approach

b- Grid-search/Random search/ Stratified search

5.5 – Built the model with the best parameters

5.6 – check the metrics – e.g., R2, RMSE, confusion matrix, ROC, Specificity,

Different kind of methods use different metrics to validate the model

Test dataset

5.7 – Validating the model on the test dataset.

5.8 – Save the model

Decision Tree: - Suffers from overfitting. Missing values/outliers

RANDOM FOREST

Ensemble of trees(distinct) – Bagging technique

1. Results are averaged
2. It is free from overfitting –
3. It can most extend missing values
4. Tree will be designed in parallel

Limitation: -

1. Time consuming
2. Computationally expensive
3. No graphically representation

In Random Forest we are doing more and more tree

Step1. Choosing the train data at random (30% to 70%)

Step2. Choosing the independent variables at random (30% to 50%)

\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\

Different classification methods

K-NN-Nearest Neighbors method

Navies-Bayes method (Bayes Theorem)

SVM

Gradient Descent

XG - Boost

\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\