DECISION TREE

DATA SCIENCE ENGINEERING METHODS & TOOLS

-SHARDUL CHAVAN

1

DATA PRE-PROCESSING

- Check for the missing values or null values.
- Encoding the Categorical Variables using
- OneHotencoding & Label Encoding

Home_Expn	Balance	Sex_F	Sex_M	Res_status_owner	Res_status_rent	...	Job_status_military
145	0	0	1	1	0	...	0
140	0	0	1	0	1	...	0
0	2200	1	0	1	0	...	0
0	0	0	1	0	1	...	0
228	0	0	1	1	0	...	0
...	...	...	...	...	...	...	...
232	200	0	1	1	0	...	0
280	0	1	0	0	1	...	0
422	200	0	1	1	0	...	0
80	300	0	1	0	1	...	0
140	0	0	1	1	0	...	0

```
[ ]  loan_data.isna().sum()

Sex                0
Age                0
Time_at_address    0
Res_status         0
Telephone          0
Occupation         0
Job_status         0
Time_employed      0
Time_bank          0
Liab_ref           0
Acc_ref            0
Home_Expn          0
Balance            0
Decision           0
dtype: int64
```

TRIED SCALING THE DATA & THE ACCURACY WAS REDUCED. DOES DECISION TREE REALLY NEEDS LESS DATA PRE-PROCESSING?

2

ANSWER

Does data need to be scaled for decision trees?

Decision trees and ensemble methods **do not require feature scaling to be performed** as they are not sensitive to the the variance in the data. Jun 21, 2020

After Scaling V.S. Before Scaling

```
[ ]  from sklearn.metrics import accuracy_score
     print(accuracy_score(Y_test, y_pred))

0.6481481481481481
```

```
from sklearn.metrics import accuracy_score
print(accuracy_score(Y_test, y_pred))

0.6759259259259259
```

3

GINI VS ENTROPY

- **Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.**
- **Gini Index is calculated between 0 and 0.5**
- **While, the calculation of the Gini Index is faster.**

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

Gini Index Formula

- ***Entropy is the measurement of the impurity or randomness in the data points.***
- **Entropy is calculated between 0 and 1**
- **Computationally, entropy is more complex since it makes use of logarithms**

$$\text{Entropy} = -\sum_{i=1}^{n} p_i * \text{Log}_2(p_i)$$

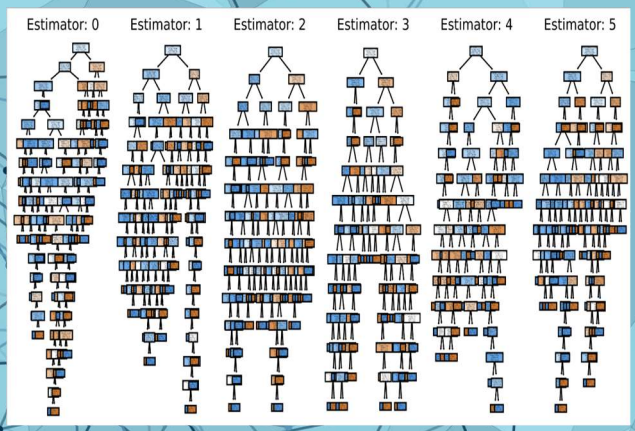
Entropy Formula

Here "p" denotes the probability that it is a function of entropy.

4

RANDOM FOREST

- Random Forest is a supervised Machine learning algorithm used for classification & regression.
- Creates a set of decision trees from a randomly selected subset of the training set
- It collects the votes from different decision trees to decide the final prediction.



5