3.7 Decision Trees #DecisionTrees

3.7.1 Intro

> 🔑 *A decision tree is a **non-parametric supervised learning algorithm**, which is **utilized** for both **classification** and **regression** tasks. It has a **hierarchical, tree structure**, which **consists** of a **root node, branches, internal nodes** and **leaf nodes**.*

Three types of nodes can be distinguished:

- #DEF **Root Node:** node without any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes;
- #DEF **Decision Node:** internal nodes;
- #DEF **Leaf Node:** these nodes represent all the possible outcomes within the dataset, they tipically assign the fraud labels.

There are various algorithms to implement the decisions needed to build a decision tree. They are:

- #DEF #SplittingDecision **Splitting decision:** Which variable to split at what value (e.g., Transaction amount is > $100, 000 or not)
- #DEF #StoppingDecision **Stopping decision:** When to stop adding nodes to the tree?
- #DEF #AssignementDecision **Assignment decision:** What class (e.g., fraud or no fraud) to assign to a leaf node? Look at the majority class within the leaf node to make the decision (*winner-take-all learning*).

Decision Trees can be divided in:

1. Classification Trees: for categorical variables;
2. Regression Trees: for continuous variables.

3.7.2.1 Impurity #Impurity

To correctly answer a splitting decision, fist the concept of impurity (or chaos) must be defined.

🎯 **Minimal impurity** occurs when all customers are either good or bad.
🎯 **Maximal impurity** occurs when one has the same number of good and bad customers.

> *Decision trees **aim** at **minimizing** the **impurity** in the data.*

Two ways to measure impurity:

- **Entropy**
  --> #DEF #Gain **Gain** = weighted decrease in entropy;
- **Gini**.

3.7.3 Classification Trees #ClassificationTrees

3.7.3.1 Splitting Decision #SplittingDecision

To answer the splitting decision, various candidate splits must be evaluated in terms of their decrease in impurity .

🎯 A higher gain is preferred.

The decision tree algorithm considers different candidate splits for its (root) node. We can have two strategies:

1. **Greedy and recursive:** Pick the one with the highest gain;
2. **Perfectly parallel:** advantage of increased efficiency.

3.7.3.2 Stopping Decision #StoppingDecision

Within a decision tree we must have a stopping decision , because If the tree continues to split, it will have one leaf node per observation. This phenomena is known as **Overfitting** and implies:

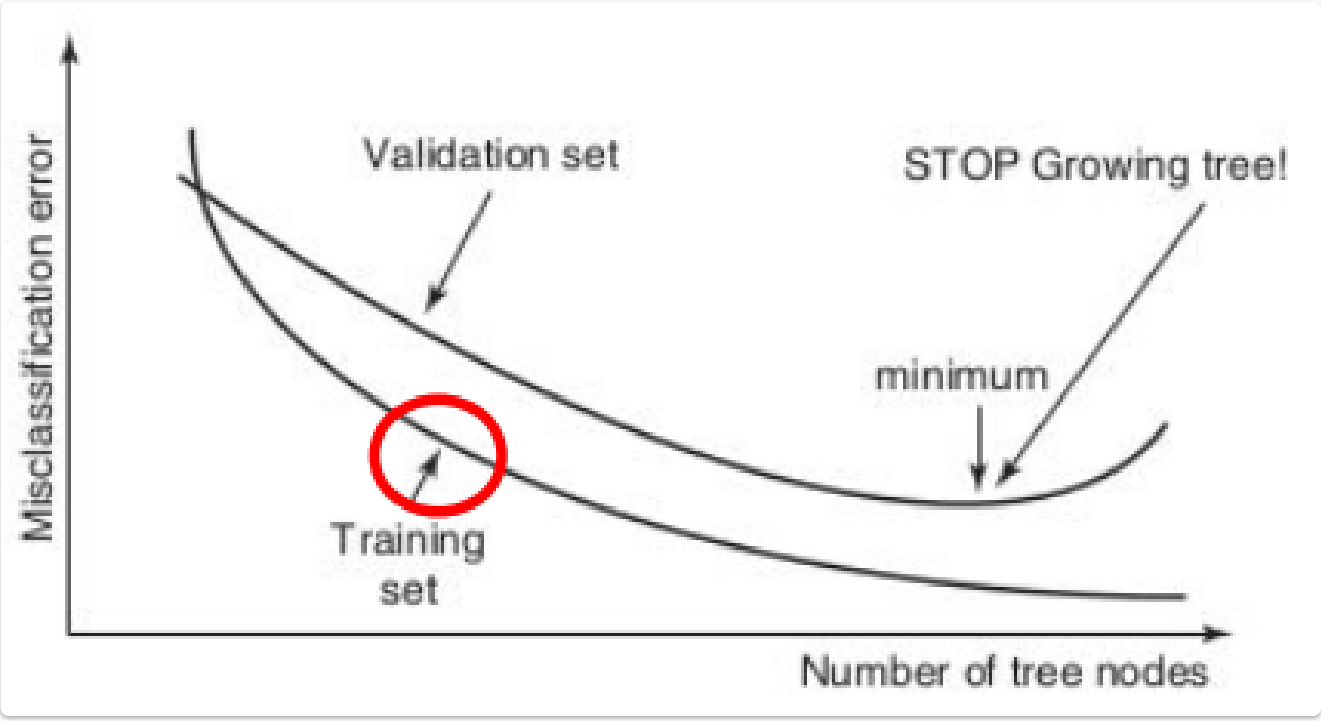
- The tree has become too complex and fails to correctly model or highlight trends in the data;
- It will generalize poorly to new unseen data.

To avoid this, split data into:

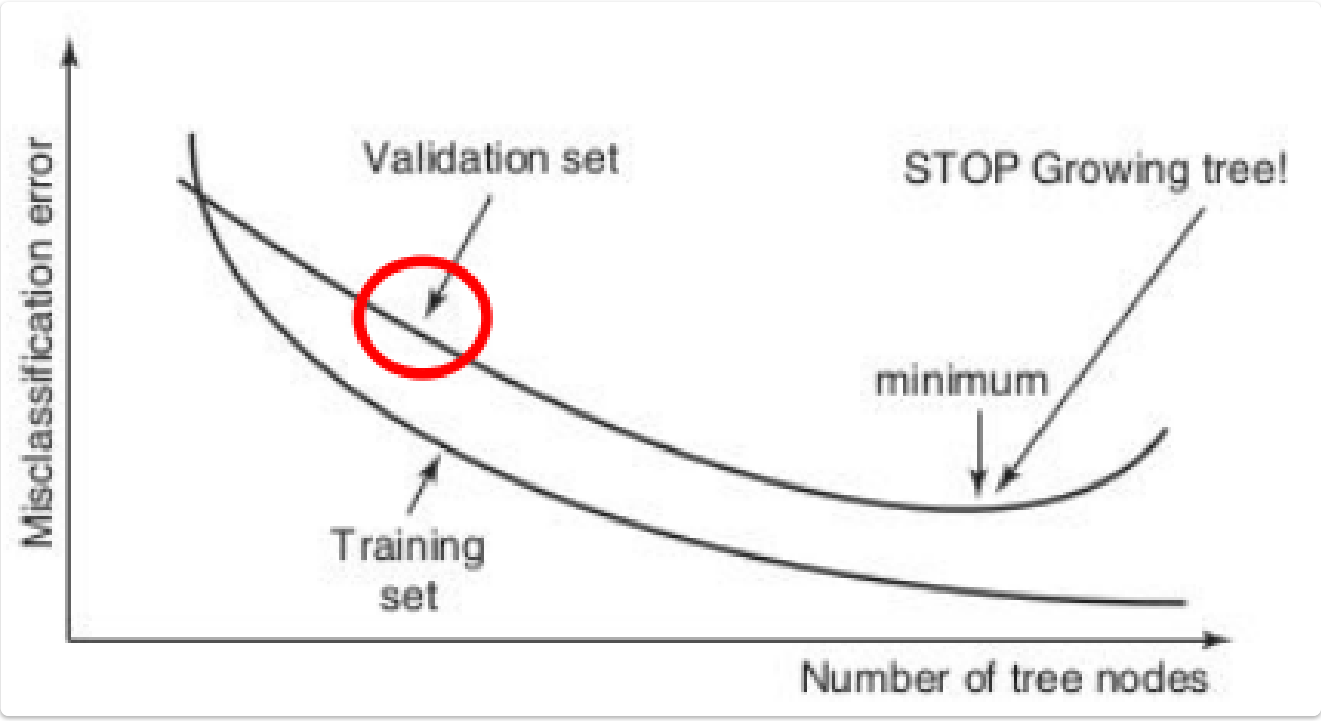
1. **Training set:** make splitting decision;

2. **Validation set:** independent sample to monitor the misclassification error (or any other performance metric) as the tree is grown.

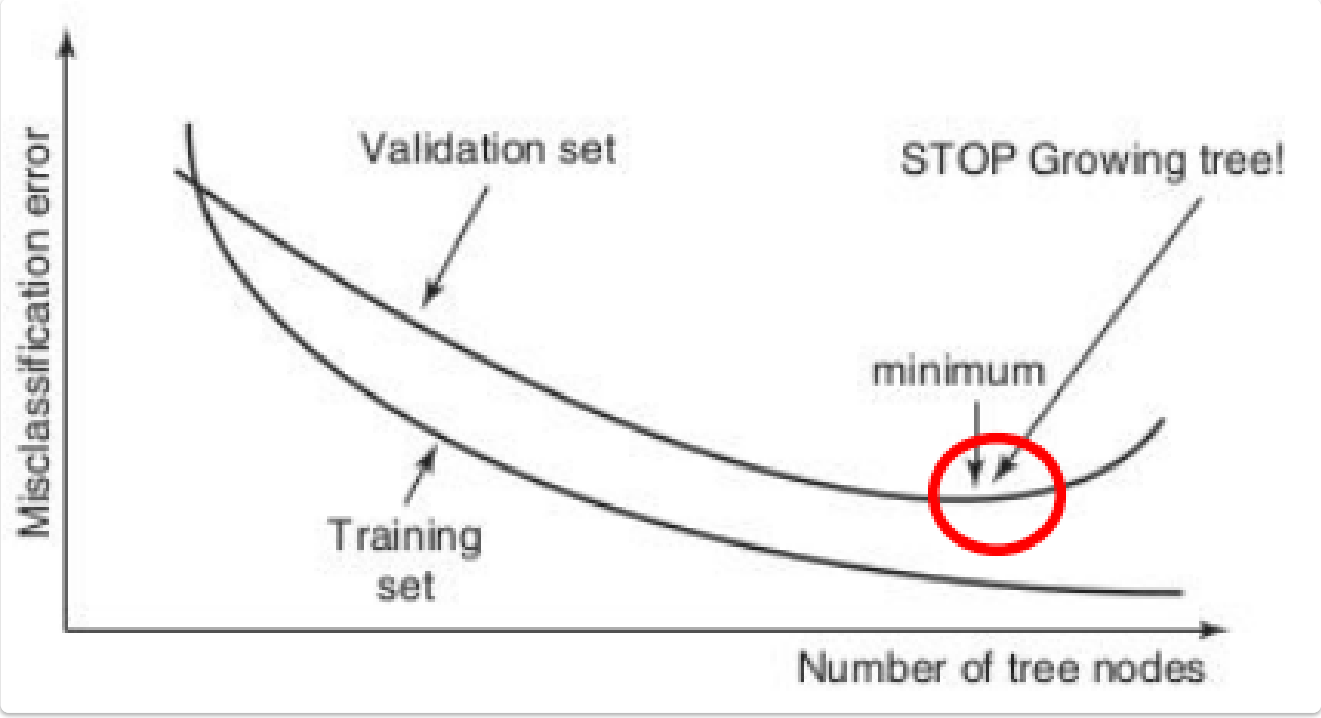
Typical split: 70% of the dataset composes the training set while the remaining 30% makes up the validation set.



The error on the training sample keeps on decreasing as the splits become more and more specific and tailored towards it.



On the validation sample, the error will initially decrease, which indicates that the tree splits generalize well.



At some point the error will increase since the splits become too specific for the training sample as the tree starts to memorize it.

> 🔑 ***Where** the **validation set curve** reaches its **minimum**, the **procedure** should be **stopped**, as **otherwise overfitting** will occur.*

3.7.4 Regression Trees #RegressionTrees

Decision trees can be used to predict continuous targets.

3.7.5.1 Splitting Decision #SplittingDecision

The impurity, for determining the splitting decision, needs to be measured in another way. Two alternatives:

1. **M**ean **S**quared **E**rror (**MSE**);
2. **Variance** (**ANOVA**) **Test** and **F-statistic**;

MSE #MSE

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

where:

- n represents the number of observations in a leaf node;
- Y_i the value of observation i ;
- Y the average of all values in the leaf node;

VARIANCE TEST #VARIANCETEST

$$F = \frac{\frac{SS_{between}}{B-1}}{\frac{SS_{whithin}}{n-B}} \sim F_{n-B,B-1}$$
$$SS_{between} = \sum_{b=1}^{B} n_b \cdot (\bar{Y}_b - \bar{Y})^2$$
$$SS_{whithin} = \sum_{b=1}^{B} \sum_{i=1}^{n_b} (Y_{bi} - \bar{Y}_b)^2$$

where:

- B the number of branches of the split;
- n_b the number of observations in branch b;
- Y_b the average in branch b;
- $Y_b i$ the value of observation i in branch b;
- Y the overall average.

WRAPPING UP

🎯 For the MSE metric: it is **desirable** to have a **low MSE** in a **leaf node** since this indicates that the **node** is more **homogeneous**.

🎯 For the Variance Test metric: **Good splits** favor homogeneity within a node (**low** SS_{within}) and heterogeneity between nodes (**high** $SS_{between}$).

3.7.5.2 Stop Decision #StoppingDecision

The stopping decision is similar to classification trees but a regression based performance measure is used (e.g., mean squared error, mean absolute deviation, R-squared) on the Y-axis.

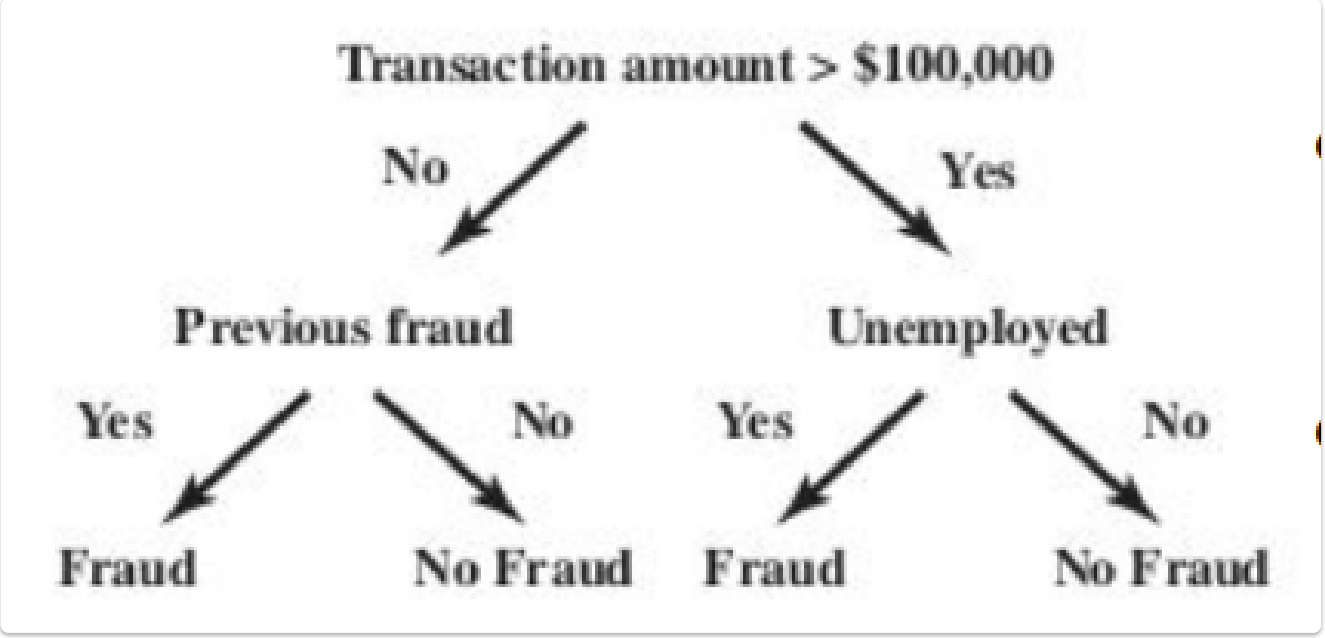
3.7.5.3 Assignment Decision #AssignementDecision

The assignment decision can be made by assigning the mean (or median) to each leaf node.

3.7.5 Decision Trees Properties #DecisionTreesProperties

Every tree can also be represented as a **rule set**. This means that every path from a root note to a leaf node can be viewed as a simple if-then rule.

Example:



Rules:

1. **If** Transaction amount **>** $100, 000 **And** Unemployed **=** No **Then** No Fraud
2. **If** Transaction amount **>** $100, 000 **And** Unemployed **=** Yes **Then** Fraud
3. **If** Transaction amount **≤** $100, 000 **And** Previous fraud **=** Yes **Then** Fraud
4. **If** Transaction amount **≤** $100, 000 **And** Previous fraud **=** No **Then** No Fraud

3.7.6 Decision Trees in Fraud Analytics #DecisionTreesFraudAnalytics

Regarding variables selection:

- Variables that occur at the top of the tree are more predictive;
- Measure the predictive strength of a variable by calculating the **Gain** of a characteristic to gauge its predictive power.

Advantages:
🟢 Decision tree gives a white-box model with a clear explanation: interpretable;
🟢 Operationally efficient;
🟢 Powerful techniques and allow for more complex decision boundaries than a logistic regression;
🟢 Nonparametric, no normality or independence assumptions are needed.

Disadvantages:
🔴 Highly dependent on the sample that was used for tree construction. A small variation in the underlying sample might yield a totally different tree.

> 🔑 *An **analytical fraud model** is usefule when used directly into the business environment.*

Next chapter: Neural Networks