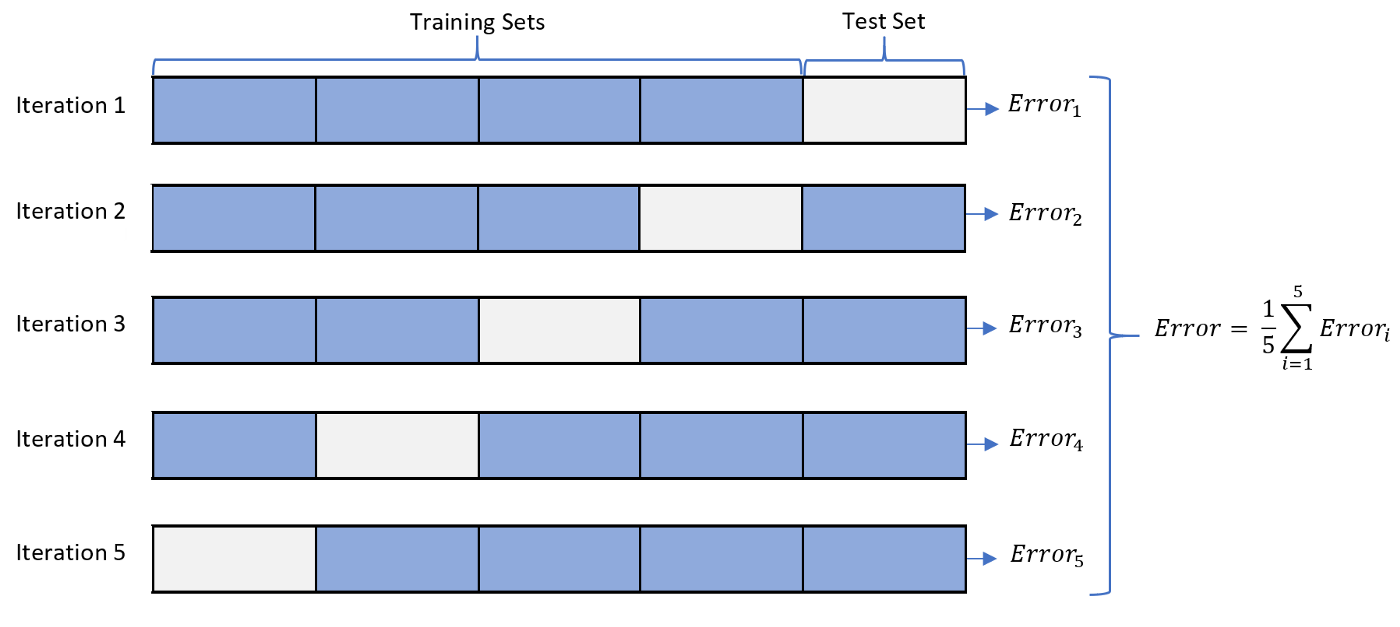


**Cross Validation:**

Like traditional Train/Test split, it is used to assess the performance of your model by randomizing the data, breaking the data up into K groups, training on K-1 groups, and testing on 1 group. You continue until each group as been tested on and end up with K number of accuracies.

This gives each point a chance to be trained on and tested on. Also reduces chances of getting a false high accuracy just by the way the data was sampled. Ensures unbiased performance

With time series, cross validation works similarly, except with a sliding window. As you continue up the X-axis, the testing dataset will always be the tail (Sliding) and the training will always be the head (Incremental). Note that you cannot randomize the data since it is time dependent

Chart

Description automatically generated

**Linear Regression:**

Statistical regression method to understand the relationship between X features and a continuous target

Must meet the following assumptions:

1. Linearity: Between all X and y

2. Homoscedasticity: The variance of residual is the same across the domain

3. Normality: All features

4. Multicollinearity: Between X features

Pros: Explainable, Interpretable, Fast ML model to train

Cons: Must meet all assumptions, sensitive to outliers, can easily underfit

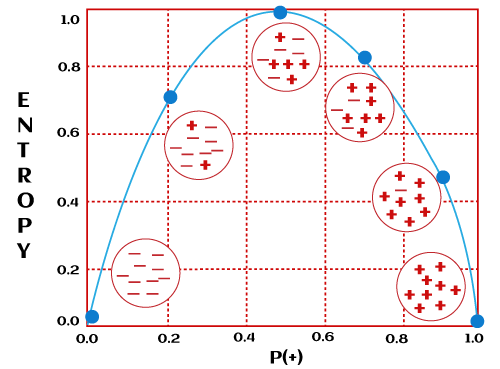
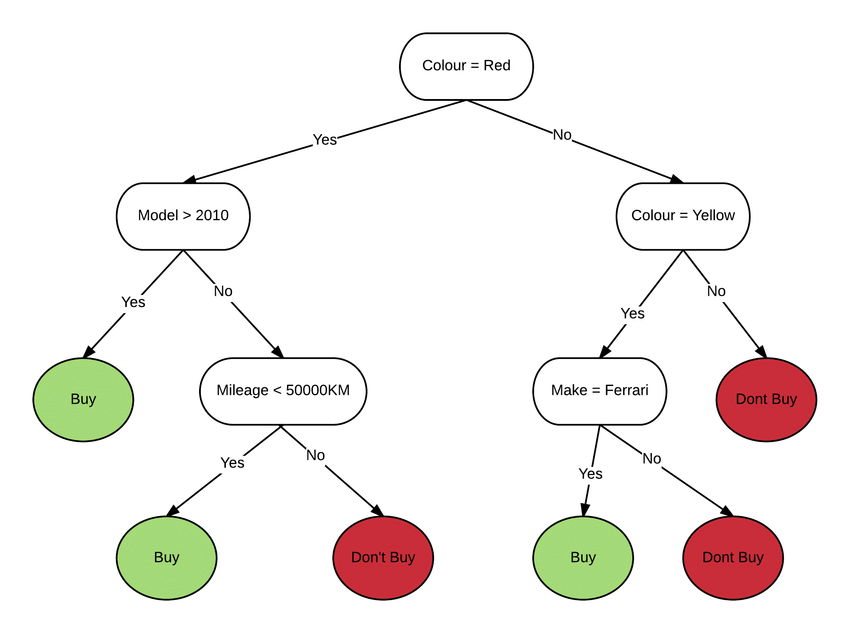
**Logistic Regression:**

A probabilistic classification model that models the relationship between X features and a categorical output

Pros: Fast to train, interpretable, can use regularization to reduce overfitting

Cons: Can overfit when number of features > number of records, linear boundaries

**Decision Trees:**

* A tree-like supervised machine learning model used to predict both continuous and categorical targets
* Like a flowchart
* Top-down algorithm, so the top of the tree has the more important features, and get less important as you get down to the root notes
* At each level, there is a decision node, which uses Entropy to decide what feature to split on
* Entropy is the measure of disorder range [0,1], where 1 would imply the highest disorder (uncertainty). High disorder means that the sample is not pure or not homogenous
* Information gain (IG) measures how much “information” a feature gives us about the class. The information gain is based on the decrease in entropy after a dataset is split on an attribute. Ranges from [0,1], where 1 represents the most lift or gain from splitting a node
* Pruning is the process of cutting back the tree
* Pros: Interpretable (Explainable), can handle NULL values
* Cons: Sensitive to outliers, Easy to overfit

**Random Forest:**

* A tree-based model, random forest takes decision trees to the next level
* Train a large number of decision trees (Usually 100s), in parallel, and each independent of each other
* Each tree is sampled randomly with replacement (Bootstrap Sampling)
* The output of each tree with be aggregated (ensembled to form one single prediction
* Bagging
* Pros: Less likely to overfit
* Cons: Training can be complex, not interpretable