Decision Trees

Created Date: 2022-08-15

> Metadata 🗃️
>
> - Title: Decision Trees for Regression (Regression Trees)
> - Author: Andrew Jones
> - Reference: Data Science Infinity

> Links & Tags 📎
>
> - Index: Course Note Index
> - Atomic Tag: #datascience
> - Subatomic Tags: #machinelearning #decisiontrees

High-Level Overview

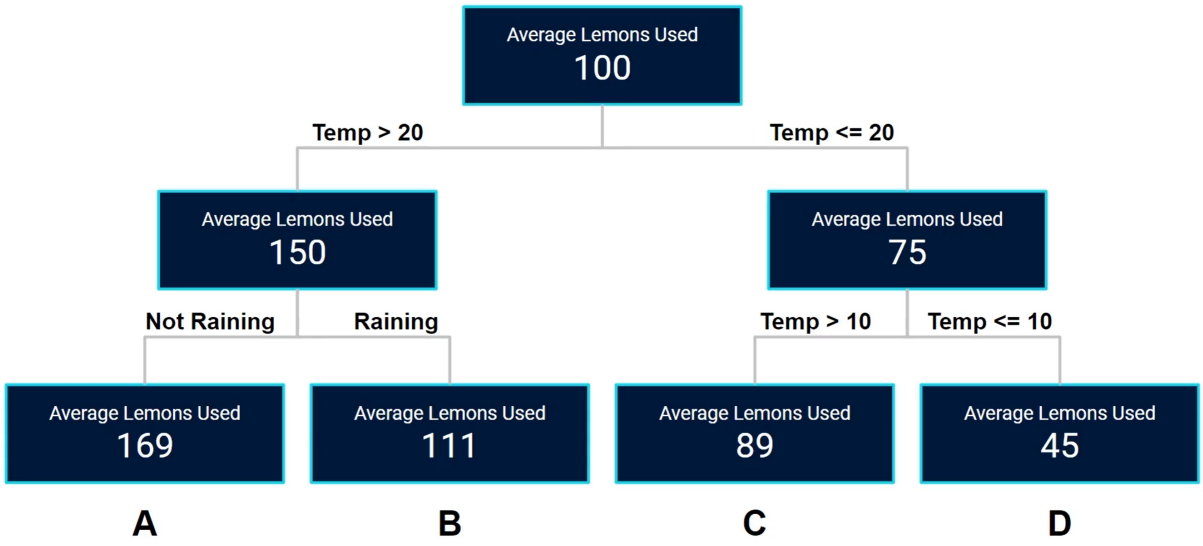
Jupyter Notebook: Basic Decision Tree Template

> A Decision Tree is a model that splits the data into distinct buckets using the input variables, with split decisions being based on how well each potential split explains differences in the output variable.
>
> Also known as C.A.R.T. (Classification and Regression Trees)

- Decision tree starts with all data
  - Model will test all possible ways it can split input variables and select the split with the biggest difference in output
  - Algorithm will split into new buckets and repeat the process

    - Buckets are called nodes, with the last nodes referred to as leaf nodes
  - Due to this splitting, decision tree algorithms do not benefit from the removal of outliers and feature selection in the data preparation process
- Decision trees are able to split on numerical and categorical variables
  - May eliminate the need to encode categorical variables
- Parameters can be specified to tell the model to stop splitting the data
- Predictions are less granular when compared to a linear regression
- Useful for dealing with data that does not follow a linear trend

Decision Tree for Regression



Advanced Theory

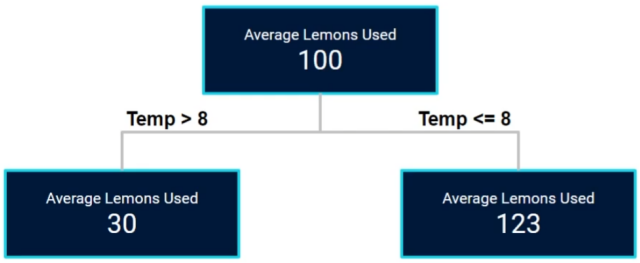
Jupyter Notebook: Advanced Decision Tree Template

Splitting Criteria

- The most commonly used metric to determine splitting criteria is the Mean Squared Error (MSE)
- The algorithm will calculate the MSE at each possible split point and split at the lowest MSE

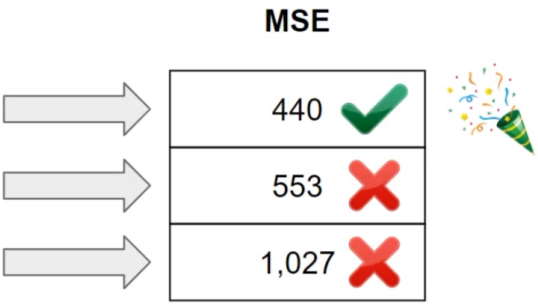
Mean Squared Error (MSE)

Daily Peak Temperature	Lemons Used (Actual)	Lemons Used (Predicted)	Error	Squared Error
8	30	30	0	0
16	92	123	-31	961
24	127	123	4	16
32	151	123	28	784
				Σ = 1761



MSE	440

Daily Peak Temperature	Lemons Used (Actual)
8	30
16	92
24	127
32	151



$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- y_i Actual Output Value
- $\hat{y}_i$ Predicted Output Value
- $(y_i - \hat{y}_i)$ Prediction Error (squared to remove negative values)
- n Number of Data Points
- $\frac{1}{n}$ Same as dividing the rest of the formula by n which results in the mean

> *MSE is an error score, telling us how far out the predicted values were from the actual values. A larger MSE is worse than a smaller MSE.*

Stopping Criteria

- The algorithm will continue to split data until one of the following criteria is met;
  - There is only one data point in each leaf node
  - It cannot find a split that will reduce the MSE
  - It's told to stop (reviewed below and in the template)
- We want to make sure the algorithm is stopped at a point that avoids overfitting
- There are several ways to instruct the algorithm to stop splitting (situation depending);
  - Specify a maximum depth (splits)
    - The algorithm may not split the specified amount if it cannot find a way to reduce the MSE, but it will not proceed pass the maximum specified splits
  - Specify a minimum number of data points that must exist in a node before its allowed to split
  - Specify a minimum number of data points that must exists in both the left and right nodes before its allowed to split
    - Ensures balanced splitting

Evaluating Model Performance

- The most commonly used metric to determine model performance is R Squared
- The R^2 measure (Coefficient of Determination) assesses the accuracy, or goodness of fit, of the decision tree
  - 0 = No Relationship
  - 1 = Perfect Relationship
  - Explains how much of the variation in the output variable (y) is explained by the input variables (x)
  - An $R^2 = 0.70$ is telling us that our input variable(s) are explaining 70% of the variation in the output variable (y)

$R^2 = \frac{SSR[Mean] - SSR[Model]}{SSE[Mean]}$

- SSR Sum of Squared Residuals
- SSR Model Sum of Squared Residuals Using the Predicted Values
- SSR Mean Sum of Squared Residuals Using the Mean for Predicted Values
    - Squared residuals are calculated as the difference between the actual output and predicted (or mean) output, squared

$R^2 = 1 - \frac{RSS}{TSS}$

- RSS Sum of Squared Residuals
- TSS Total Sum of Squares

Where;

$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$

- y_i Sample Value
- $f(x_i)$ Predicted Value
- n Number of Observations

$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

- y_i Sample Value
- $\bar{y}$ Mean Sample Value
- n Number of Observations