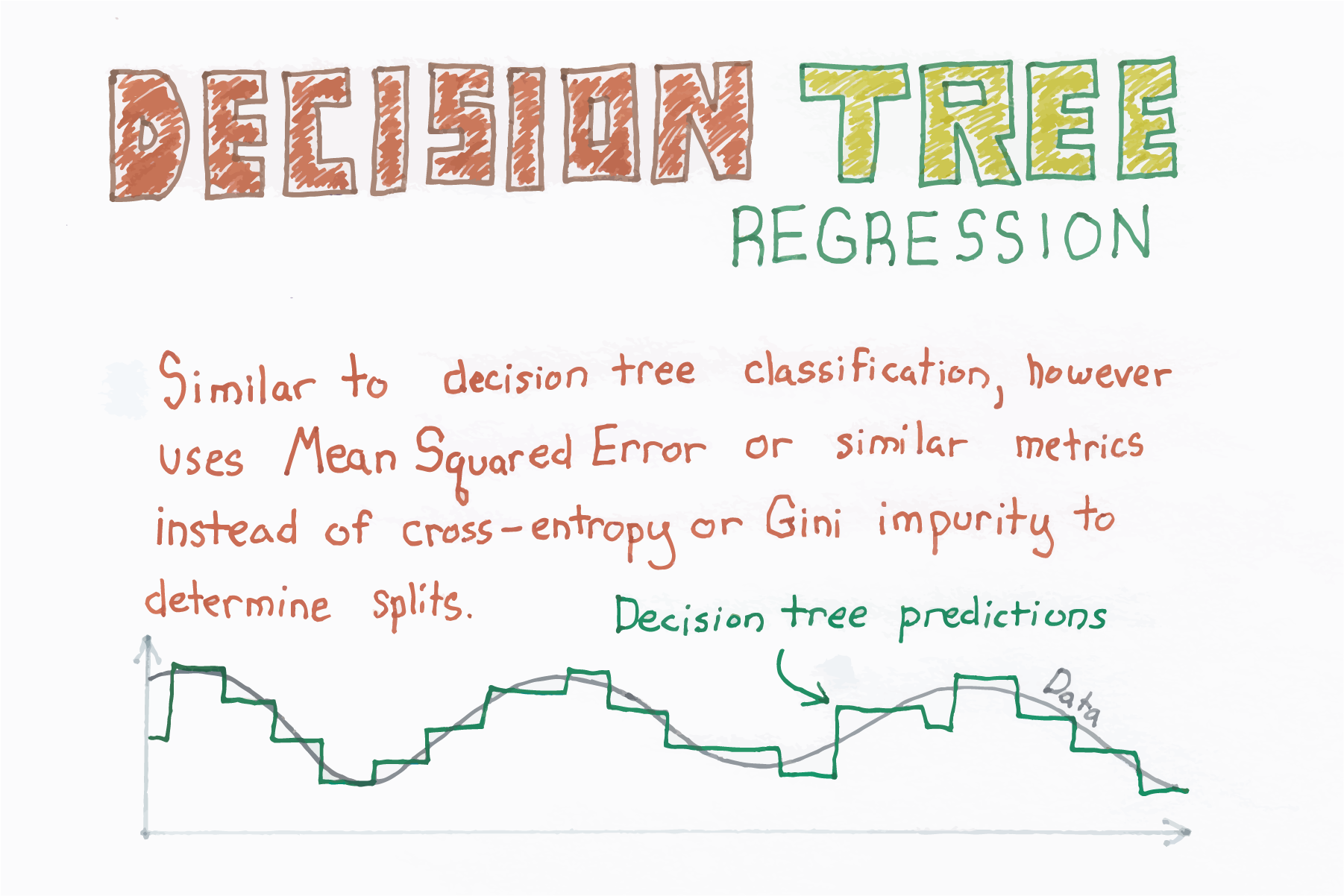
Using Decision Trees for Regression Problems

**Introduction** :

The goal of the blogpost is to equip beginners with basics of Decision Tree Regressor algorithm and quickly help them to build their first model. We will mainly focus on the modeling side of it . The data cleaning and preprocessing parts would be covered in detail in an upcoming post.



Source : https://chrisalbon.com/

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.

The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.

The Mean Squared Error is given by:



Enough of theory , let’s start with implementation.

**Problem Statement :**

To predict the median prices of homes located in boston area given other attributes of the house.

**Data details**

Boston House Prices dataset  
===========================  
  
Notes  
------  
Data Set Characteristics:   
  
 :Number of Instances: 506   
  
 :Number of Attributes: 13 numeric/categorical predictive  
   
 :Median Value (attribute 14) is usually the target  
  
 :Attribute Information (in order):  
 - CRIM per capita crime rate by town  
 - ZN proportion of residential land zoned for lots over 25,000 sq.ft.  
 - INDUS proportion of non-retail business acres per town  
 - CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)  
 - NOX nitric oxides concentration (parts per 10 million)  
 - RM average number of rooms per dwelling  
 - AGE proportion of owner-occupied units built prior to 1940  
 - DIS weighted distances to five Boston employment centres  
 - RAD index of accessibility to radial highways  
 - TAX full-value property-tax rate per $10,000  
 - PTRATIO pupil-teacher ratio by town  
 - B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town  
 - LSTAT % lower status of the population  
 - MEDV Median value of owner-occupied homes in $1000's  
  
 :Missing Attribute Values: None  
  
 :Creator: Harrison, D. and Rubinfeld, D.L.  
  
This is a copy of UCI ML housing dataset.  
<http://archive.ics.uci.edu/ml/datasets/Housing>  
  
This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.  
  
The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic  
prices and the demand for clean air', J. Environ. Economics & Management,  
vol.5, 81-102, 1978. Used in Belsley, Kuh & Welsch, 'Regression diagnostics  
...', Wiley, 1980. N.B. Various transformations are used in the table on  
pages 244-261 of the latter.  
  
The Boston house-price data has been used in many machine learning papers that address regression  
problems.

**Tools used** :

Pandas , Numpy , Matplotlib , scikit-learn

**Python Implementation with code :**

1. **Import necessary libraries**

Import the necessary modules from specific libraries.

|  |
| --- |
|  |

1. **Load the data set**

Use pandas module to read the taxi data from the file system. Check few records of the dataset.

|  |
| --- |
|  |

1. **Select the predictor and target variables**

|  |
| --- |
|  |

1. **Train test split :**

|  |
| --- |
|  |

**5. Training / model fitting :**

Fit the model to selected supervised data

|  |
| --- |
|  |

1. **Model parameters study :**

The coefficient R^2 is defined as (1 - u/v), where u is the residual sum of squares ((y\_true - y\_pred) \*\* 2).sum() and v is the total sum of squares ((y\_true - y\_true.mean()) \*\* 2).sum().

|  |
| --- |
|  |

1. **Check the accuracy and report it with test data :**

Visualize the goodness of the fit with the predictions being visualized by a line

|  |
| --- |
|  |

1. **What are your conclusions?**