Decision Tree notes

- A simpler model built from a large dataset will perform better than a complex model built with a small dataset.

Overfitting

- Model is built to exactly match (perfectly predict) the dataset being used for training. Model reflects the noise in the dataset more than it does the relationship between independent and dependent variables.
- Would generate different responses for similar datasets.

Underfitting

- Model lacking sufficient training that it doesn't understand anything about the dataset and makes random, incoherent, predictions.
- Would generate different responses for similar datasets

Overfitting solutions

- Cross Validation
- Regularisation
- Ensemble Learning

Cross validation

Splitting the data into **k** subsets.

- Model selection technique
- A good model is one which performs well on an unseen dataset i.e Something other than the training set

Workflow:

- Split the data into multiple subsets
- Keep some data aside to validate the performance of the model - Validation

Example(4-fold cross validation)

1. Split the data into 4 subsets
2. Take one of the subsets as the testing dataset. The other 3 subsets are the training datasets
3. Train your model and validate against the testing subset
4. repeat steps one to three using a different subset as the testing set each time

Regularisation

Penalises models that are too complex

E(New Model) = E(Model) + αReg(Complexity)

A form of regression that shrinks the coefficient estimates towards zero. For a refresher on regression just go down below

(brief) Regression refresher

Technique used to describe a relationship between a dependent variable and one or more independent variables

Before going into regression I'll first define the correlation coefficient r

Correlation Coefficient (r): ranging from -1 to 1 it measures the correlation of data from strong correlation (1, -1) to weak correlation (0)

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

where $x = x_i - \bar{x}$ and $y = y_i - \bar{y}$

Linear Regression: relation of an dependent variable to a single independent variable

$$\hat{y} = b_0 + b_1 x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable and $\hat{y}$ is the predicted value of the dependent variable

to find the the constant b_0 and the regression coefficient b_1 you use the following formulae

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},\ b_0 = \bar{y} - b_1 \bar{x}$$

I should mention that this is called the least squares method

- Minimizes the squared difference between values

I should also mention that your x and y values will be coming from your training data

There are many forms of regression but this should give an idea of how regression functions