**ML NOTES (Regression)**

**INTRODUCTION**

**Machine Learning:** Making predictions and classifications.

**Training Data:** original data

**Testing data:** Data used to test the predictions made by the model. Used to evaluate the ML model.

**\***Don’t be fooled by how well a ML method fits the training data.

**\***No method is 100% accurate, we just pick the method that does the best job predicting any outcome.

**Types:** Regression model, Decision trees, Bayesian etc etc.

A diagram of a graph

Description automatically generated

**Important website for stocks data:** data.nasdaq.com (module name = nasdaqdatalink)

**Important modules:** ‘scikit-learn’, ‘nasdaqdatalink’, ‘pandas’

**Features and Labels:**

Label is a dependent variable, output variable or the target variable that is predicted by the trained model.

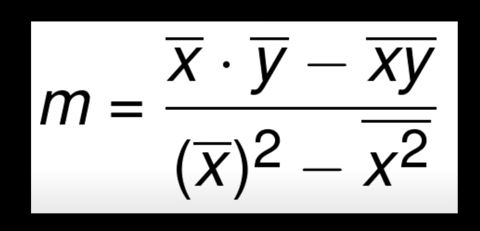
Feature is an independent variable, input variable that is used by the model to make predictions or classifications.

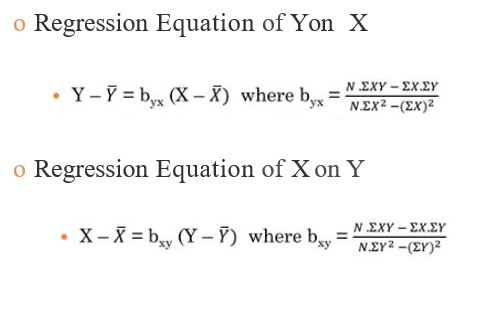
**Scaling:** Ensuring that all the data falls within a specified range. (not always necessary). Module used is ‘proprocessing’

**Cross validation:** process of shuffling the data to make the testing data sample more reliable and varied. Module used is ‘model\_selection’

**Pickling:** Saving the classifier after clf.fit(x\_train, y\_train) step instead of training the model again and again. We might have to retrain the model once or twice a month but we don’t have to retrain it every time we use it.

**Regression Theory (How It Works)**

y = mx + b

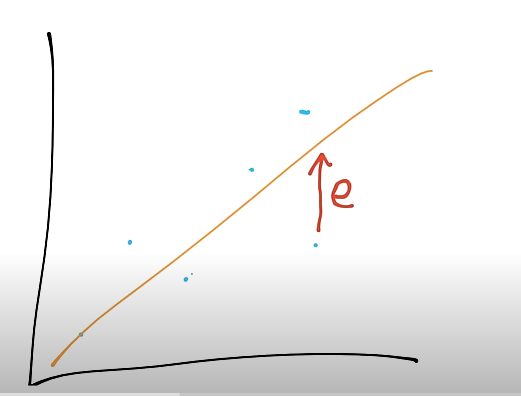
OR (here m = b)  


A close-up of a mathematical equation

Description automatically generated

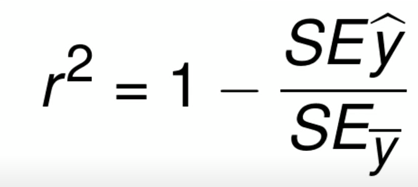
**How to determine the accuracy of the best fit line? (i.e How well does it fit the data?)**

We determine the accuracy through R-square or the coefficient of determination (squared error).



Error is the distance between any point and its distance from the best fit line as shown above. We square the error so we don’t get any negative values as we want the error to only give the magnitude not the direction.

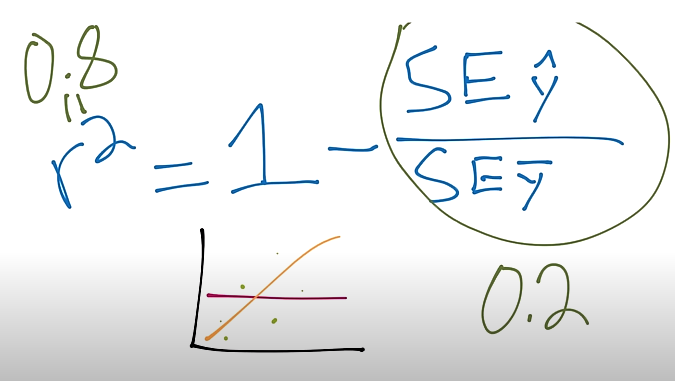
**Formula**



r^2 = r-squared

SEy\_cap = squared error of best fit line from the point

Sey\_bar = squared error of mean of y from the point



Inorder for r^2 to be 0.8, SEy\_cap/SEy\_bar would have to be 0.2

SEy\_cap should be significantly lower than SEy\_bar ideally so SEy\_cap/SEy\_bar is very small and r^2 is as close as possible to 1.

* The higher the r^2 value the better the accuracy of the best fit line