

Conductor: ticket please

Passenger: please give me one ticket for Hyderabad

Conductor: oh sorry passenger! This bus is not going to Hyderabad. You should go to platform number 1

Passenger: thank you

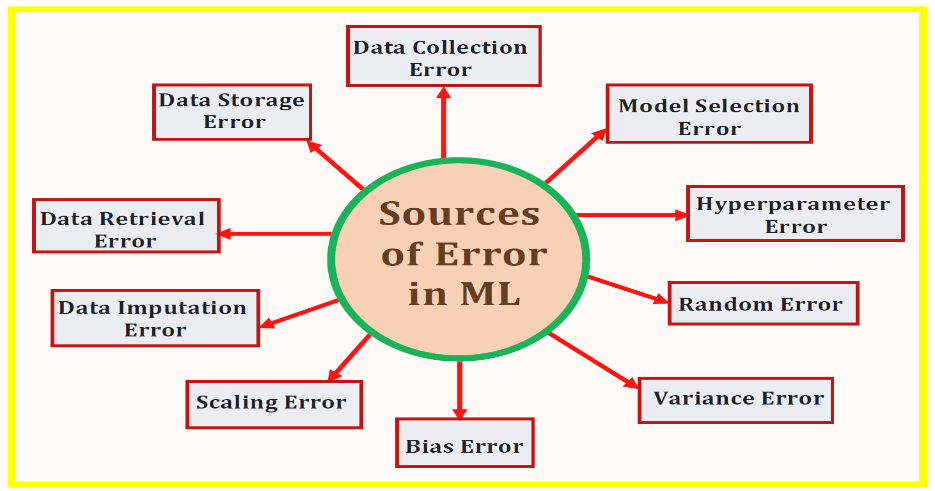
Here passenger is wrongly predicted the platform number

**Error= predicted value - actual value**

What is error?

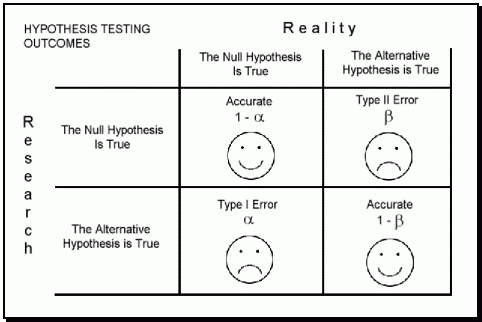
The first step is always understanding the cost of errors. Would you rather error on guessing something is true when it is not, or not guessing it at all.

Sources of error in machine learning:



**Statistical Background:**

In [statistical test theory](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing), the approach of a [statistical error](https://en.wikipedia.org/wiki/Errors_and_residuals) is an integral part of [hypothesis testing](https://en.wikipedia.org/wiki/Hypothesis_Testing). The test goes about choosing about two competing propositions called [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis), denoted by H0 and [alternative hypothesis](https://en.wikipedia.org/wiki/Alternative_hypothesis), denoted by H1 . This is conceptually comparable to the judgement in a court trial. The null hypothesis compare to the position of defendant: just as he is presumed to be innocent until proven guilty, so is the null hypothesis presumed to be true until the data provide convincing evidence against it. The alternative hypothesis corresponds to the position against the defendant.



If the result of the test compare with reality, then a correct decision has been made. However, if the result of the test does not correspond with reality, then an error has appeared. There are two situations in which the decision is wrong. The null hypothesis may be true, whereas we reject H0. On the other hand, the alternative hypothesis H1 may be true, whereas we do not reject H0. Two types of error are acclaimed : Type I error and type II error

True Positive (TP)- A true positive test result is one that detects the condition when the condition is present

**False Positive (FP)-**Also known as a **Type I error**, a false positive test result is one that detects the condition when the condition is absent

**False Negative (FN)-**Also known as a **Type II error**, a false negative test result is one that does not detect the condition when the condition is present.

**True Negative (TN)-** A true negative test result is one that does not detect the condition when the condition is absent

Model performance evaluation metrics:

Regression performance (y-continuous):

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable(y) and one or more independent variables(x)

- R2 and adjusted R2

- Mean squared error(MSE), or root mean squared error(RMSE)

- Mean error, or mean absolute error

- Median error, or median absolute error

Classification performance(y-discrete):

A classification is a division or class in a system which divides things into groups or types.

- Confusion matrix

- Precision

- Recall (aka sensitivity)

- Specificity

- Accuracy

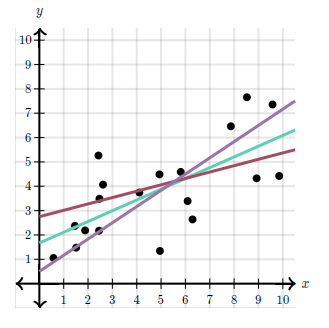
- Lift

- Area under the ROC curve (AUC)

**Residuals:**

We run into a problem in statistics when we're trying to fit a line to data points in a scatter plot. The problem is this: It's hard to say for sure which line fits the data best.

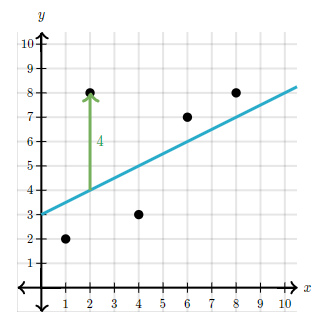
For example, imagine three scientists Andrew, Jeremy, and brooke are working with the same data set. If each scientist draws a different line of fit, how do they decide which line is best?



A residual is a measure of how well a line fits an individual data point.

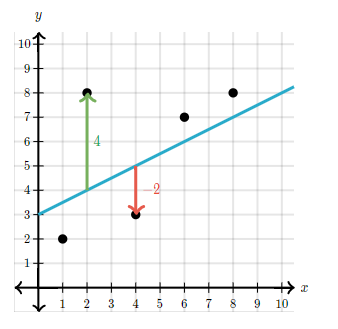
Consider this simple data set with a line of fit drawn through it

and notice how point (2,8) is 4 units above the line:



This vertical distance is known as a **residual**. For data points above the line, the residual is positive, and for data points below the line, the residual is negative

For example, the residual for the point (4,3) is -2



The closer a data point's residual is to 000, the better the fit. In this case, the line fits the point (4,3) is better than it fits the point (2,8)

**Residuals = fitted value – actual value**

Standardized residuals:

The standardized-residualis the [residual](http://www.r-tutor.com/node/97) divided by its [standard deviation](http://www.r-tutor.com/node/43)



What is standard deviation?

The standard-deviationof an observation variable is the square root of its [variance](http://www.r-tutor.com/node/42)

What is variance?

The **variance**is a numerical measure of how the data values is disappear around the [mean](http://www.r-tutor.com/node/35). In particular, the **sample variance**is defined as:



Similarly, the **population variance**is describe in terms of the population mean *μ* and population size *N*:



What is mean?

The **mean**of an observation variable is a numerical measure of the central location of the data values. It is the sum of its data values divided by data count.

Hence, for a data sample of size *n,*its **sample mean**is defined as follows:



Similarly, for a data population of size *N,*the **population mean**is:

