Overfitting

Overfitting in machine learning is a phenomenon that occurs when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the model learns the training data too well, including its inherent inaccuracies and random fluctuations, leading to the following issues:

Overfitting is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data. When data scientists use machine learning models for making predictions, they first train the model on a known data set. Then, based on this information, the model tries to predict outcomes for new data sets. An overfit model can give inaccurate predictions and cannot perform well for all types of new data.

Why does overfitting occur?
You only get accurate predictions if the machine learning model generalizes to all types of data within its domain. Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead. Overfitting happens due to several reasons, such as:

- The training data size is too small and does not contain enough data samples to accurately represent all possible input data values.
- The training data contains large amounts of irrelevant information, called noisy data.
- The model trains for too long on a single sample set of data.
- The model complexity is high, so it learns the noise within the training data.

<u>Overfitting examples:</u>

Consider a use case where a machine learning model has to analyze photos and identify the ones that contain dogs in them. If the machine learning model was trained on a data set that contained majority photos showing dogs outside in parks , it may may learn to use grass as a feature for classification, and may not recognize a dog inside a room.
Another overfitting example is a machine learning algorithm that predicts a university student's academic performance and graduation outcome by analyzing several factors like family income, past academic performance, and academic qualifications of parents. However, the test data only includes candidates from a specific gender or ethnic group. In this case, overfitting causes the algorithm's prediction accuracy to drop for candidates with gender or ethnicity outside of the test dataset.

How can you prevent overfitting?
You can prevent overfitting by diversifying and scaling your training data set or using some other data science strategies, like those given below.

Early stopping
Early stopping pauses the training phase before the machine learning model learns the noise in the data. However, getting the timing right is important; else the model will still not give accurate results.

Pruning
You might identify several features or parameters that impact the final prediction when you build a model. Feature selection—or pruning—identifies the most important features within the training set and eliminates irrelevant ones. For example, to predict if an image is an animal or human, you can look at various input parameters like face shape, ear position, body structure, etc. You may prioritize face shape and ignore the shape of the eyes.

Regularization
Regularization is a collection of training/optimization techniques that seek to reduce overfitting. These methods try to eliminate those factors that do not impact the prediction outcomes by grading features based on importance. For example, mathematical calculations apply a penalty value to features with minimal impact. Consider a statistical model attempting to predict the housing prices of a city in 20 years. Regularization would give a lower penalty value to features like population growth and average annual income but a higher penalty value to the average annual temperature of the city.

Ensembling
Ensembling combines predictions from several separate machine learning algorithms. Some models are called weak learners because their results are often inaccurate. Ensemble methods combine all the weak learners to get more accurate results. They use multiple models to analyze sample data and pick the most accurate outcomes. The two main ensemble methods are bagging and boosting. Boosting trains different machine learning models one after another to get the final result, while bagging trains them in parallel.

Data augmentation
Data augmentation is a machine learning technique that changes the sample data slightly every time the model processes it. You can do this by changing the input data in small ways. When done in moderation, data augmentation makes the training sets appear unique to the model and prevents the model from learning their characteristics. For example, applying transformations such as translation, flipping, and rotation to input images.

Validation

Validation in machine learning (ML) is a critical process used to evaluate the performance of a model. It involves testing the model on a set of data that is separate from the data used for training. This process is essential for several reasons:

Assessing Model Performance: Validation helps in determining how well the model performs. This includes evaluating its accuracy, precision, recall, F1 score, and other relevant metrics depending on the task (classification, regression, etc.).

Overfitting Prevention: One of the primary goals of validation is to check for overfitting. Overfitting occurs when a model learns the training data too well, including its noise and outliers, and performs poorly on new, unseen data. Validation helps in identifying if the model is generalizing well to new data.

Model Selection and Tuning: By validating different models or different configurations of the same model (hyperparameters), you can compare their performance and select the best one. This is essential for fine-tuning a model to achieve the best results.

Feature Evaluation: Validation can also be used to assess the importance and impact of different features (input variables) on the model's performance.

Model Validation Techniques

There are a number of different model validation techniques, choosing the right one will depend upon your data and what you're trying to achieve with your machine learning model. These are the most common model validation techniques.

Train and Test Split or Holdout

The most basic type of validation technique is a train and test split. The point of a validation technique is to see how your machine learning model reacts to data it's never seen before. All validation methods are based on the train and test split, but will have slight variations.

With this basic validation method, you split your data into two groups: training data and testing data. You hold back your testing data and do not expose your machine learning model to it, until it's time to test the model. Most people use a 70/30 split for their data, with 70% of the data used to train the model.

Resubstitution

The resubstitution validation method is where you use all of your data as training data. Then, you compare the error rate of the machine learning model's output to the actual value from the training data set. This is an easy to do method and it can help you quickly find the gaps in your data.

K-Fold Cross-Validation
A k-fold cross-validation is similar to the test split validation, except that you will split your data into more than two groups. In this validation method, "K" is used as a placeholder for the number of groups you'll split your data into.

For example, you can split your data into 10 groups. One group is left out of the training data. Then you validate your machine learning model using the group that was left out of the training data. Then, you cross validate. Each of the 9 groups used as training data are then also used to test the machine learning model. Each test and score can give you new information about what's working and what's not in your machine learning model.

Random Subsampling
Random subsampling functions in the same way to validate your model as does the train and test validation model. The key difference is that you'll take a random subsample of your data, which will then form your test set. All of your other data that wasn't selected in that random subsample is the training data.

Bootstrapping
Bootstrapping is a form of machine learning model validation technique that uses sampling with replacement. This type of validation is most useful for estimating the quantity of a population.

When using the bootstrapping validation method, you will take a small sample out of your whole data set. From that small sample, you'll find the average or another meaningful statistic. You'll replace the data and include the new statistic that you calculated and then run your model again.

Nested Cross-Validation
Most types of validation techniques are looking to evaluate the error estimation. The nested cross-validation technique is used to evaluate the hyperparameters of your machine learning model. Testing your hyperparameters with this method prevents overfitting.

To use this model you nest two k-fold cross-validation loops inside one another. The inner loop is for hyperparameter tuning while the outer loop is for error testing and estimating accuracy.

Regression

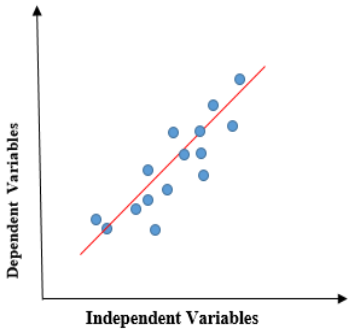
In Regression, we plot a graph between the variables which best fit the given data points. The machine learning model can deliver predictions regarding the data. In naïve words, "Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum." It is used principally for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

Types of Regression models

1. Linear Regression
2. Polynomial Regression
3. Logistics Regression

Linear Regression

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. *If there is a single input variable (x), such linear regression is called* ***simple linear regression****. And if there is more than one input variable, such linear regression is called* ***multiple linear regression*****.** The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is

likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.
To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

y= Dependent Variable.

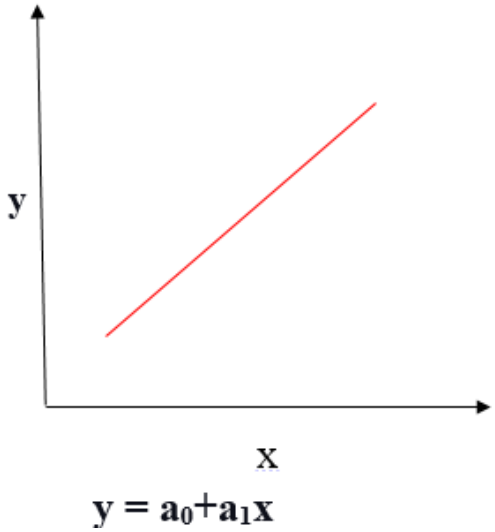
x= Independent Variable.

a0= intercept of the line.

a1 = Linear regression coefficient.

As mentioned above, Linear regression estimates the relationship between a dependent variable and an independent variable. Let's understand this with an easy example:
Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.
A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

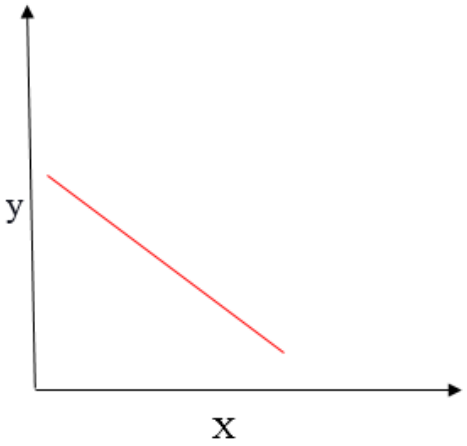
Positive Linear Relationship
If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



$$y = a_0+a_1x$$

Negative Linear Relationship

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



$$y = -a_0 + a_1 x$$

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic Regression is another statistical analysis method borrowed by Machine Learning. It is used when our dependent variable is dichotomous or binary. It just means a variable that has only 2 outputs, for example, A person will survive this accident or not, The student will pass this exam or not. The outcome can either be yes or no (2 outputs). This regression technique is similar to linear regression and can be used to predict the Probabilities for classification problems.

Types of Logistic Regression

Here are the three main types of logistic regression:

1. Binary logistic regression

Binary logistic regression is used to predict the probability of a binary outcome, such as yes or no, true or false, or 0 or 1. For example, it could be used to predict whether a customer will churn or not, whether a patient has a disease or not, or whether a loan will be repaid or not**.**

2.Multinomial logistic regression

Multinomial logistic regression is used to predict the probability of one of three or more possible outcomes, such as the type of product a customer will buy, the rating a customer will give a product, or the political party a person will vote for.

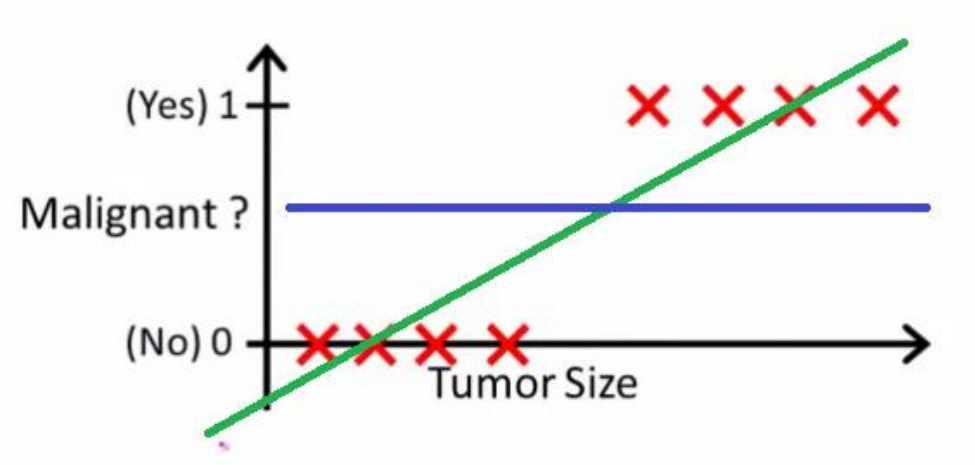
3.Ordinal logistic regression

Ordinal Logistic regression is used to predict the probability of an outcome that falls into a predetermined order, such as the level of customer satisfaction, the severity of a disease, or the stage of cancer.

Why do we use Logistic Regression rather than Linear Regression?

After reading the definition of logistic regression we now know that it is only used when our dependent variable is binary and in linear regression this dependent variable is continuous.
The second problem is that if we add an outlier in our dataset, the best fit line in linear regression shifts to fit that point.

Now, if we use linear regression to find the best fit line which aims at minimizing the distance between the predicted value and actual value, the line will be like this:



The blue line represents the old threshold and the yellow line represents the new threshold which is maybe 0.2 here. To keep our predictions right we had to lower our threshold value. Hence we can say that linear regression is prone to outliers. Now here if h(x) is greater than 0.2 then only this regression will give correct outputs.

Another problem with linear regression is that the predicted values may be out of range. We know that probability can be between 0 and 1, but if we use linear regression this probability may exceed 1 or go below 0.

To overcome these problems we use Logistic Regression, which converts this straight best fit line in linear regression to an S-curve using the sigmoid function, which will always give values between 0 and 1.

Regularization

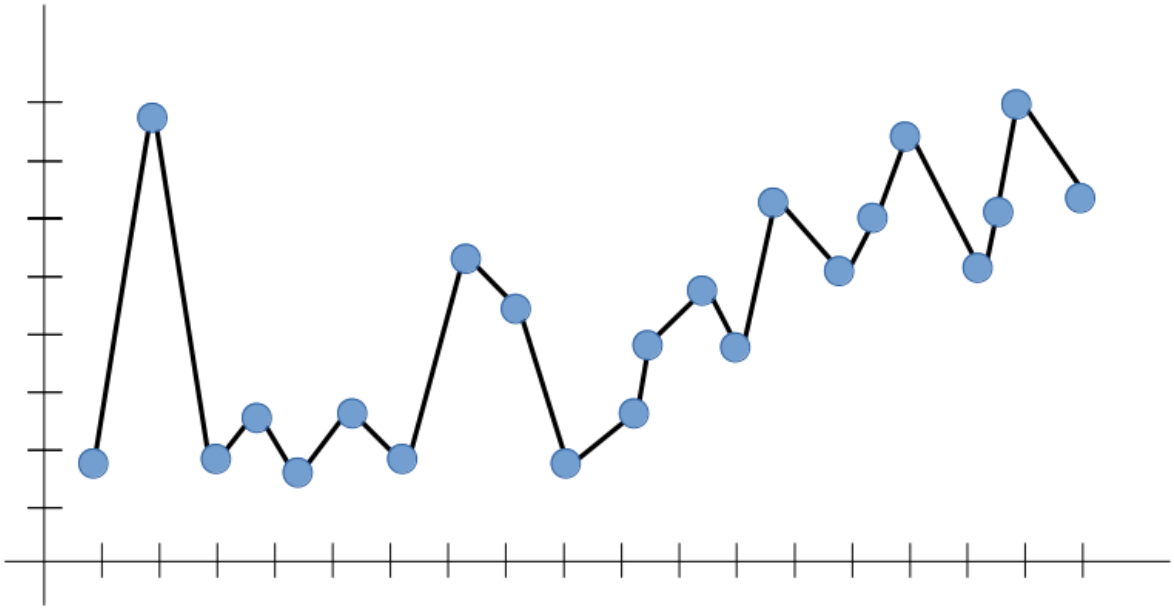
When training a machine learning model, the model can be easily overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit the model to our test set. Regularization techniques help reduce the possibility of overfitting and help us obtain an optimal model.

Understanding Overfitting and Underfitting

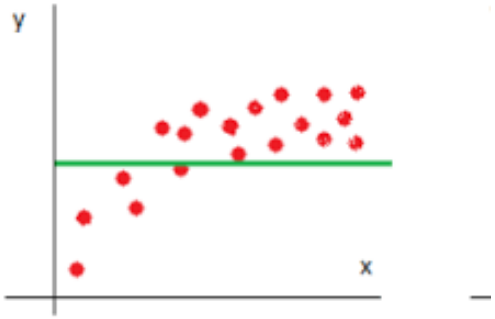
To train our machine learning model, we provide it with data to learn from. The process of plotting a series of data points and drawing a line of best fit to understand the relationship between variables is called Data Fitting. Our model is best suited when it can find all the necessary patterns in our data and avoid random data points, and unnecessary patterns called noise.

If we allow our machine learning model to look at the data too many times, it will find many patterns in our data, including some that are unnecessary. It will learn well on the test dataset and fits very well. It will learn important patterns, but it will also learn from the noise in our data and will not be able to make predictions on other data sets.

A scenario where a machine learning model tries to learn from the details along with the noise in the data and tries to fit each data point to a curve is called Overfitting.
In the figure below, we can see that the model is fit for every point in our data. If new data is provided, the model curves may not match the patterns in the new data, and the model may not predict very well.



Conversely, in the scenario where the model has not been allowed to look at our data enough times, the model will not be able to find patterns in our test data set. It won't fit our test data set properly and won't work on new data either.
A scenario where a machine learning model can neither learn the relationship between variables in the test data nor predict or classify a new data point is called Underfitting.
The image below shows an underequipped model. We can see that it doesn't fit the data given correctly. He did not find patterns in the data and ignored much of the data set. It cannot work with both known and unknown data.



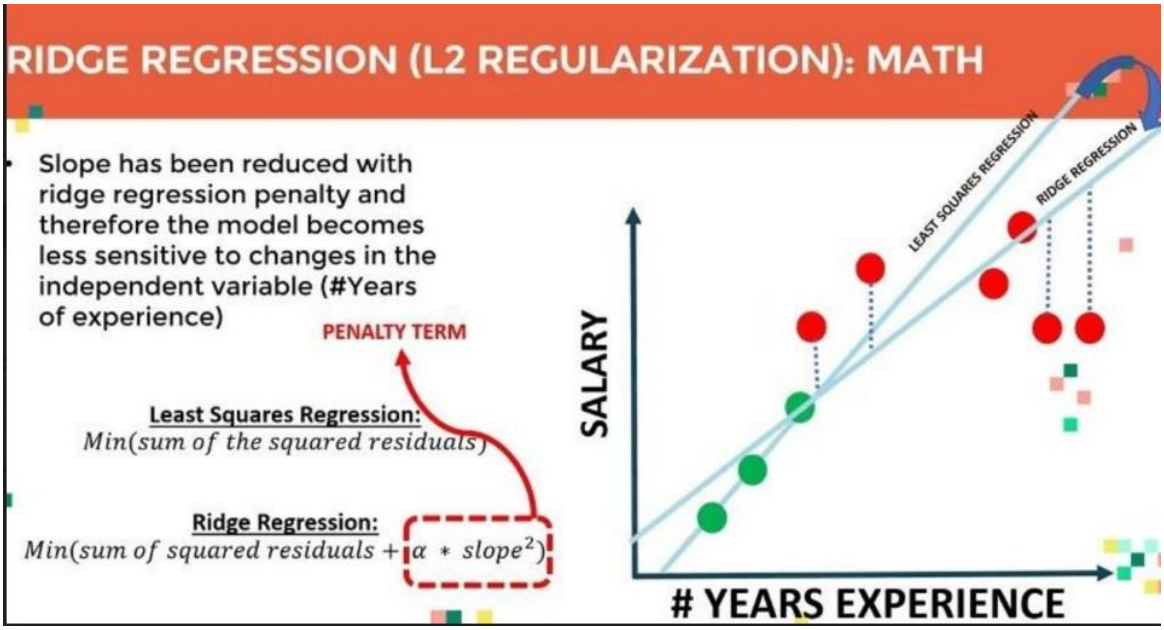
Underfitting (High Bias)

Regularization Techniques

1. Ridge Regularization

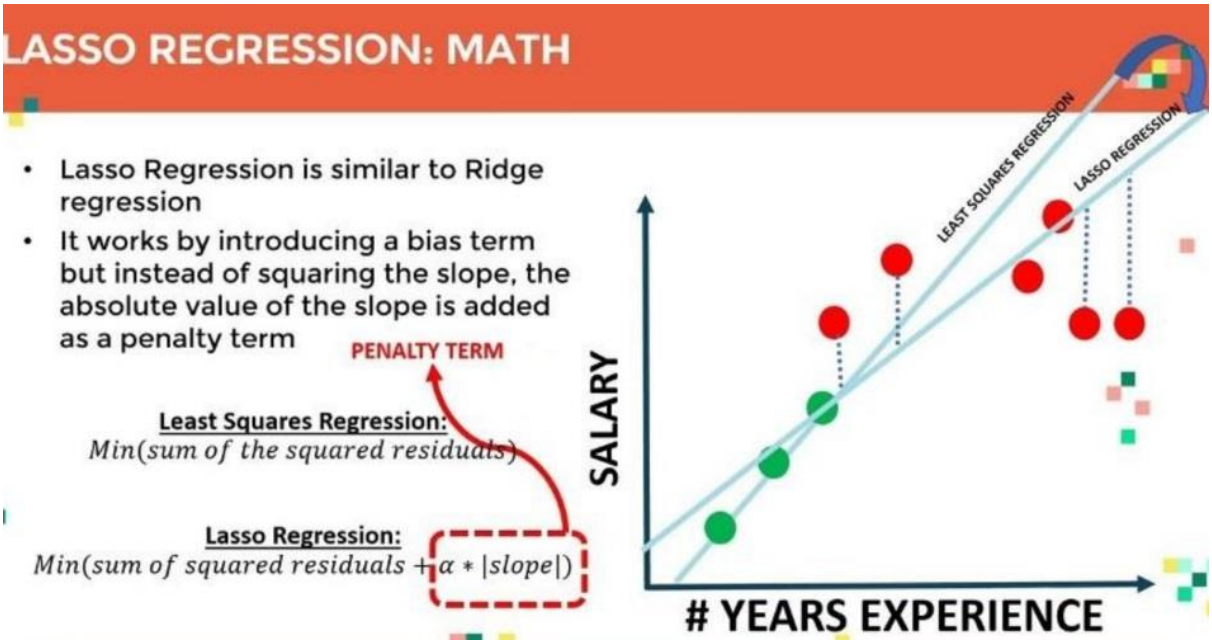
   Also known as Ridge Regression, it adjusts models with overfitting or underfitting by adding a penalty equivalent to the sum of the squares of the magnitudes of the coefficients.

   This means that the mathematical function representing our machine learning model is minimized and the coefficients are calculated. The size of the coefficients is multiplied and added. Ridge Regression performs regularization by reducing the coefficients present. The function shown below shows the cost function of the ridge regression.



2. Lasso Regularization

Modifies overfitted or under-fitted models by adding a penalty equivalent to the sum of the absolute values of the coefficients.
Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the actual values of the coefficients. This means that the sum of the coefficients can also be 0 because there are negative coefficients.
Consider the cost function for the lasso regression.



Resampling Method

Resampling Method is a statical method that is used to generate new data points in the dataset by randomly picking data points from the existing dataset. It helps in creating new synthetic datasets for training machine learning models and to estimate the properties of a dataset when the dataset is unknown, difficult to estimate, or when the sample size of the dataset is small.

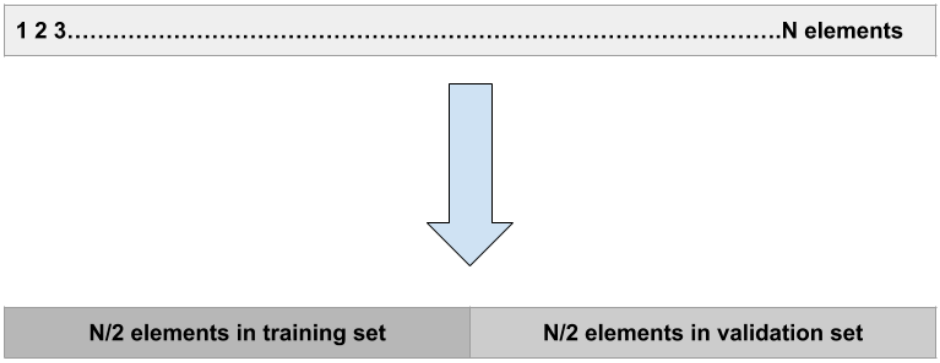
Two common methods of Resampling are

1. Cross Validation
2. Bootstrapping

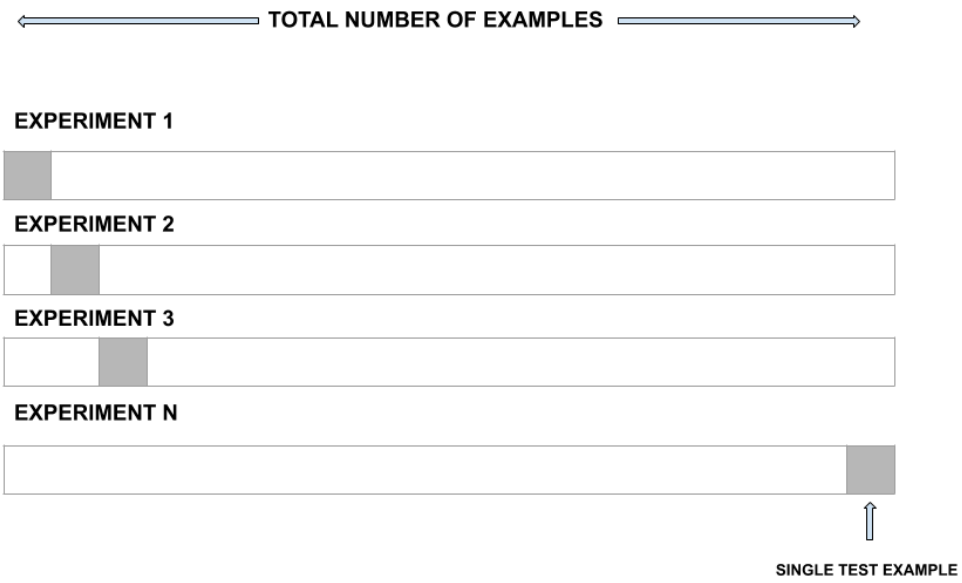
1. Cross Validation

Cross-Validation is used to estimate the test error associated with a model to evaluate its performance.

Validation set approach:
This is the most basic approach. It simply involves randomly dividing the dataset into two parts: first a training set and second a validation set or hold-out set. The model is fit on the training set and the fitted model is used to make predictions on the validation set.

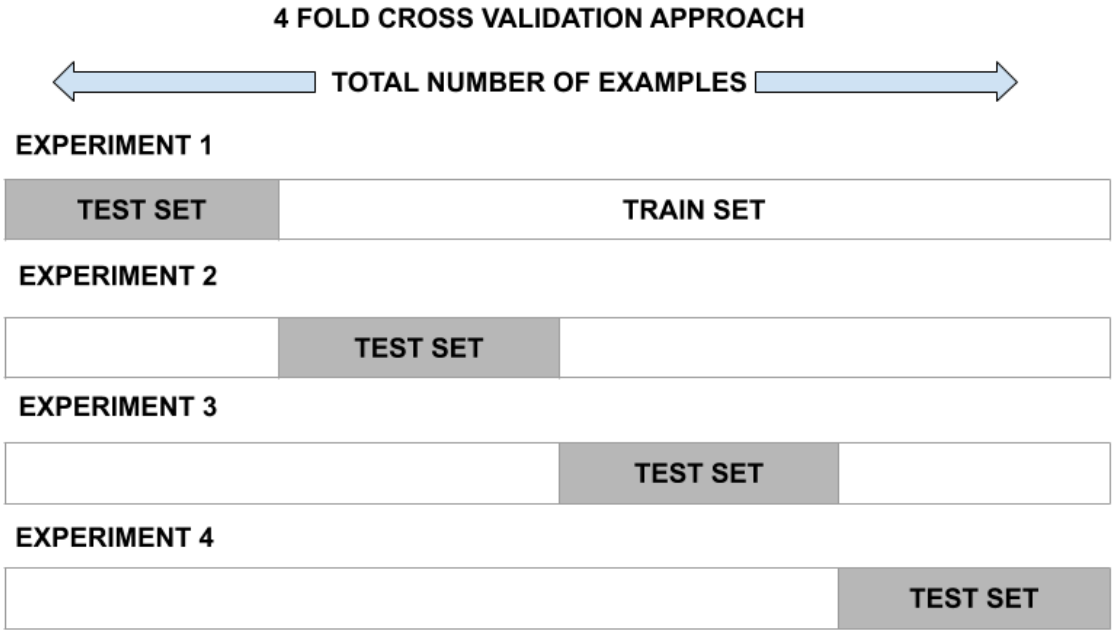


Leave-one-out-cross-validation:
LOOCV is a better option than the validation set approach. Instead of splitting the entire dataset into two halves, only one observation is used for validation and the rest is used to fit the model.



k-fold cross-validation

This approach involves randomly dividing the set of observations into k folds of nearly equal size. The first fold is treated as a validation set and the model is fit on the remaining folds. The procedure is then repeated k times, where a different group each time is treated as the validation set.



2. Bootstrapping

Bootstrapping is a resampling technique that helps in estimating the uncertainty of a statistical model. It includes sampling the original dataset with replacement and generating multiple new datasets of the same size as the original. Each of these new datasets is then used to calculate the desired statistic, such as the mean or standard deviation.

This process is repeated multiple times, and the resulting values are used to construct a probability distribution for the desired statistic.

This technique is often used in machine learning to estimate the accuracy of a model, validate its performance, and identify areas that need improvement.

For example, we can use bootstrap sampling to calculate the population means. Also, the result would be as follows.