DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

DAY 12

Q1. Where is the confusion matrix used? Which module would you use to show it?

Answer:

In machine learning, confusion matrix is one of the easiest ways to summarize the performance of your algorithm.

At times, it is difficult to judge the accuracy of a model by just looking at the accuracy because of problems like unequal distribution. So, a better way to check how good your model is, is to use a confusion matrix.

First, let's look at some key terms.

Classification accuracy – This is the ratio of the number of correct predictions to the number of predictions made

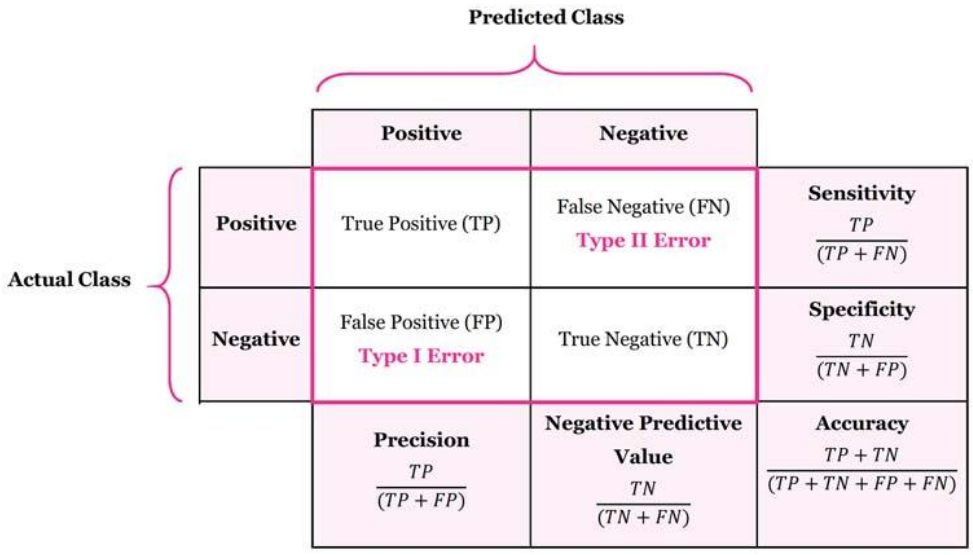
True positives – Correct predictions of true events

False positives – Incorrect predictions of true events

True negatives – Correct predictions of false events

False negatives – Incorrect predictions of false events.

The confusion matrix is now simply a matrix containing true positives, false positives, true negatives, false negatives.

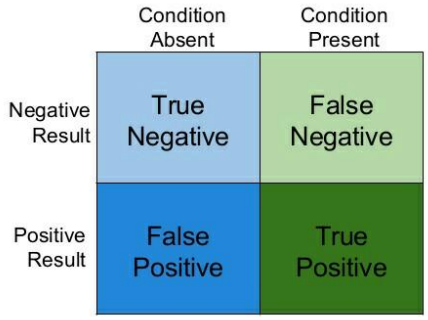


Q2: What is Accuracy?

Answer:

It is the most intuitive performance measure and it simply a ratio of correctly predicted to the total observations. We can say as, if we have high accuracy, then our model is best. Yes, we could say that accuracy is a great measure but only when you have symmetric datasets where false positives and false negatives are almost same.

Accuracy = True Positive + True Negative / (True Positive +False Positive + False Negative + True Negative)



Q3: What is Precision?

Answer:

It is also called as the positive predictive value. Number of correct positives in your model that predicts compared to the total number of positives it predicts.

Precision = True Positives / (True Positives + False Positives)

Precision = True Positives / Total predicted positive

It is the number of positive elements predicted properly divided by the total number of positive elements predicted.

We can say Precision is a measure of exactness, quality, or accuracy. High precision

Means that more or all of the positive results you predicted are correct.

Q4: What is Recall?

Answer:

Recall we can also called as sensitivity or true positive rate.

It is several positives that our model predicts compared to the actual number of positives in our data.

Recall = True Positives / (True Positives + False Positives)

Recall = True Positives / Total Actual Positive

Recall is a measure of completeness. High recall which means that our model classified most or all of the possible positive elements as positive.

Q5: What is F1 Score?

Answer:

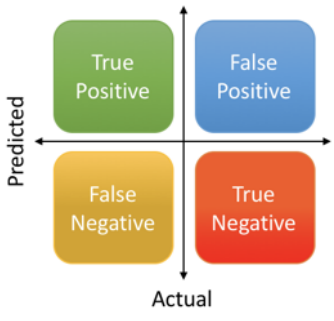
We use Precision and recall together because they complement each other in how they describe the effectiveness of a model. The F1 score that combines these two as the weighted harmonic mean of precision and recall.

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



Q6: What is Bias and Variance trade-off?

Answer:

Bias

Bias means it's how far are the predict values from the actual values. If the average predicted values are far off from the actual values, then we called as this one have high bias.

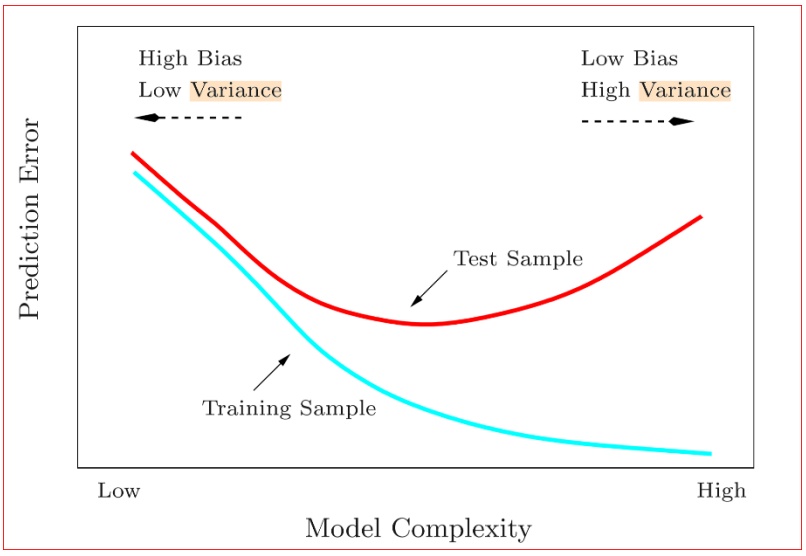
When our model has a high bias, then it means that our model is too simple and does not capture the complexity of data, thus underfitting the data.

Variance

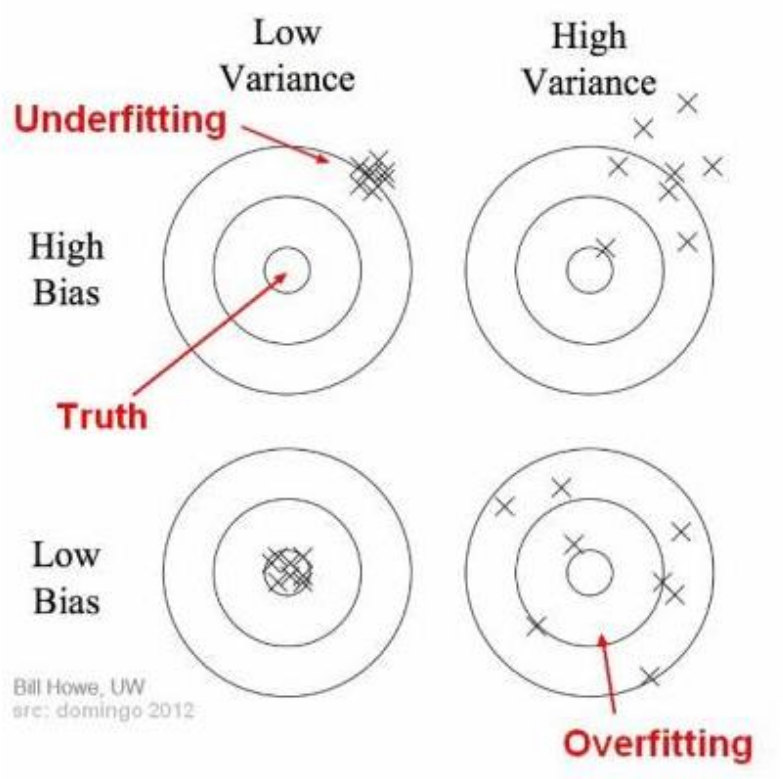
It occurs when our model performs good on the trained dataset but does not do well on a dataset that it is not trained on, like a test dataset or validation dataset. It tells us that actual value is how much scattered from the predicted value.

Because of High variance it cause overfitting that implies that the algorithm models random noise present in the training data.

When model have high variance, then model becomes very flexible and tune itself to the data points of the training set.



Bias-variance: It decomposition essentially decomposes the learning error from any algorithm by adding bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if we make the model more complex and add more variables, We'll lose bias but gain some variance —to get the optimally reduced amount of error, you'll have to tradeoff bias and variance. We don't want either high bias or high variance in your model.



Bias and variance using bulls-eye diagram

Q7. What is data wrangling? Mention three points to consider in the process.

Answer:

Data wrangling is a process by which we convert and map data. This changes data from its raw form to a format that is a lot more valuable.

Data wrangling is the first step for machine learning and deep learning. The end goal is to provide data that is actionable and to provide it as fast as possible.

There are three major things to focus on while talking about data wrangling –

1. Acquiring data

The first and probably the most important step in data science is the acquiring, sorting and cleaning of data. This is an extremely tedious process and requires the most amount of time.

One needs to:

- Check if the data is valid and up-to-date.
- Check if the data acquired is relevant for the problem at hand.

Sources for data collection Data is publicly available on various websites like kaggle.com, **data.gov** ,**World Bank**, **Five Thirty Eight Datasets**, AWS Datasets, Google Datasets.

2. Data cleaning

Data cleaning is an essential component of data wrangling and requires a lot of patience. To make the job easier it is first essential to format the data make the data readable for humans at first.

The essentials involved are:

- Format the data to make it more readable
- Find outliers (data points that do not match the rest of the dataset) in data
- Find missing values and remove them from the data set (without this, any model being trained becomes incomplete and useless)

3. Data Computation

At times, your machine not have enough resources to run your algorithm e.g. you might not have a GPU. In these cases, you can use publicly available APIs to run your algorithm. These are standard end points found on the web which allow you to use computing power over the web and process data without having to rely on your own system. An example would be the Google Colab Platform.

Q8. Why is normalization required before applying any machine learning model? What module can you use to perform normalization?

Answer:

Normalization is a process that is required when an algorithm uses something like distance measures. Examples would be clustering data, finding cosine similarities, creating recommender systems.

Normalization is not always required and is done to prevent variables that are on higher scale from affecting outcomes that are on lower levels. For example, consider a dataset of employees' income. This data won't be on the same scale if you try to cluster it. Hence, we would have to normalize the data to prevent incorrect clustering.

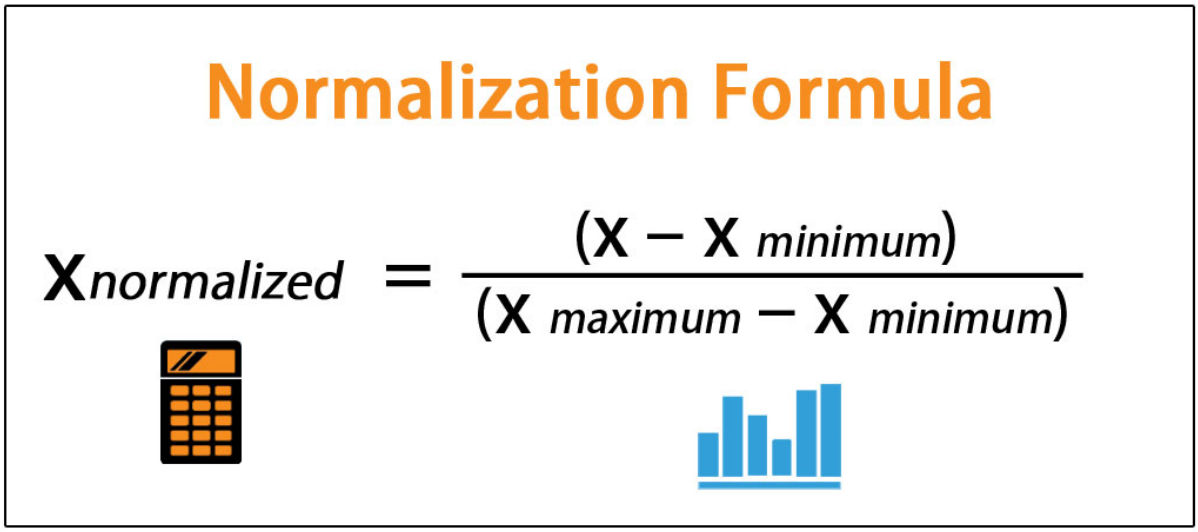
A key point to note is that normalization does not distort the differences in the range of values.

A problem we might face if we don't normalize data is that gradients would take a very long time to descend and reach the global maxima/ minima.

For numerical data, normalization is generally done between the range of 0 to 1.

The general formula is:

Xnew = (x-xmin)/(xmax-xmin)



Q9. What is the difference between feature selection and feature extraction?

Feature selection and feature extraction are two major ways of fixing the curse of dimensionality

1. Feature selection:

Feature selection is used to filter a subset of input variables on which the attention should focus. Every other variable is ignored. This is something which we, as humans, tend to do subconsciously.

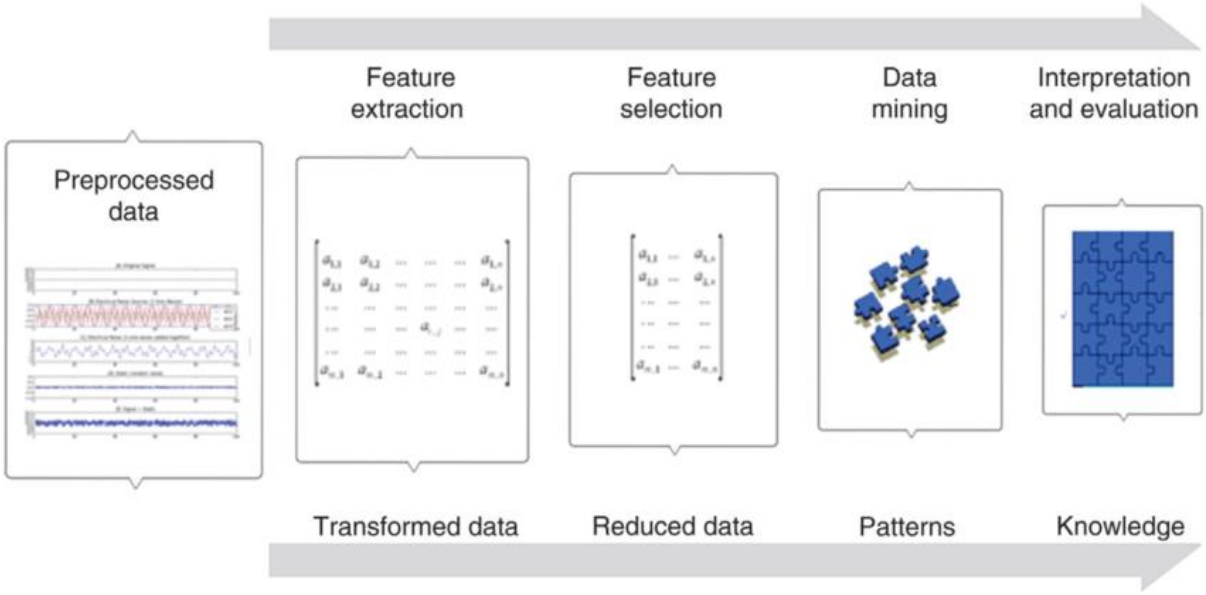
Many domains have tens of thousands of variables out of which most are irrelevant and redundant. Feature selection limits the training data and reduces the amount of computational resources used. It can significantly improve a learning algorithms performance.

In summary, we can say that the goal of feature selection is to find out an optimal feature subset. This might not be entirely accurate, however, methods of understanding the importance of features also exist. Some modules in python such as Xgboost help achieve the same.

2. Feature extraction

Feature extraction involves transformation of features so that we can extract features to improve the process of feature selection. For example, in an unsupervised learning problem, the extraction of bigrams from a text, or the extraction of contours from an image are examples of feature extraction.

The general workflow involves applying feature extraction on given data to extract features and then apply feature selection with respect to the target variable to select a subset of data. In effect, this helps improve the accuracy of a model.



Q10. Why is polarity and subjectivity an issue?

Polarity and subjectivity are terms which are generally used in sentiment analysis.

Polarity is the variation of emotions in a sentence. Since sentiment analysis is widely dependent on emotions and their intensity, polarity turns out to be an extremely important factor.

In most cases, opinions and sentiment analysis are evaluations. They fall under the categories of emotional and rational evaluations.

Rational evaluations, as the name suggests, are based on facts and rationality while emotional evaluations are based on non-tangible responses, which are not always easy to detect.

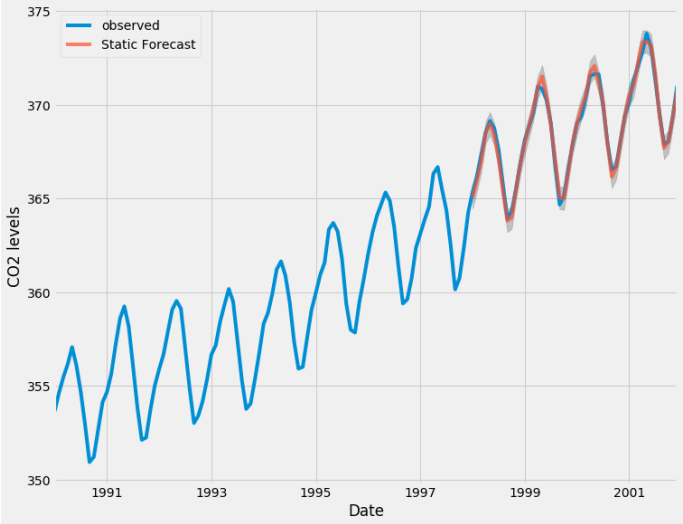
Subjectivity in sentiment analysis, is a matter of personal feelings and beliefs which may or may not be based on any fact. When there is a lot of subjectivity in a text, it must be explained and analysed in context. On the contrary, if there was a lot of polarity in the text, it could be expressed as a positive, negative or neutral emotion.

Q11. When would you use ARIMA?

Answer:

ARIMA is a widely used statistical method which stands for Auto Regressive Integrated Moving Average. It is generally used for analyzing time series data and time series forecasting. Let's take a quick look at the terms involved.

Auto Regression is a model that uses the relationship between the observation and some numbers of lagging observations.



Integrated means use of differences in raw observations which help make the time series stationary.

Moving Averages is a model that uses the relationship and dependency between the observation and residual error from the models being applied to the lagging observations.

Note that each of these components are used as parameters. After the construction of the model, a linear regression model is constructed.

Data is prepared by:

- Finding out the differences
- Removing trends and structures that will negatively affect the model
- Finally, making the model stationary.