DATA SCIENCE

Interview Questions & Answers

1)Differentiate between Data Science, Machine Learning, and AI.

Data Science vs Machine Learning

Criteria	Data Science	Machine Learning	Artificial Intelligence
Definition	Data Science is not exactly a subset of machine learning but it uses machine learning to analyze and make future predictions.	A subset of AI that focuses on a narrow range of activities.	A wide term that focuses on applications ranging from Robotics to Text Analysis.
Role	It can take on a business role.	It is a purely technical role.	It is a combination of both business and technical aspects.
Scope	Data Science is a broad term for diverse disciplines and is not merely about developing and training models.	Machine learning fits within the data science spectrum.	AI is a sub-field of computer science.
AI	Loosely integrated	Machine learning is a subfield of AI and is tightly integrated.	A sub-field of computer science consisting of various tasks like planning, moving around in the world, recognizing objects and sounds, speaking, translating, performing social or business transactions, creative work.

2) Python or R – Which one would you prefer for text analytics?

The best possible answer for this would be Python because it has a Pandas library that provides easy to use data structures and high-performance data analysis tools.

3) Which technique is used to predict categorical responses

Classification technique is used widely in mining for classifying data sets.

4) What is logistic regression? Or State an example when you have used logistic regression recently.

Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

5) What are Recommender Systems?

A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

6) Why does data cleaning play a vital role in the analysis?

Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because - as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

7) Differentiate between univariate, bivariate, and multivariate analysis.

These are descriptive statistical analysis techniques that can be differentiated based on the number of variables involved at a given point in time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as univariate analysis.

If the analysis attempts to understand the difference between 2 variables at the time as in a scatterplot, then it is referred to as bivariate analysis. For example, analyzing the volume of sales and spending can be considered as an example of bivariate analysis.

Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as multivariate analysis.

8) What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve. The random variables are distributed in the form of a symmetrical bell shaped curve.

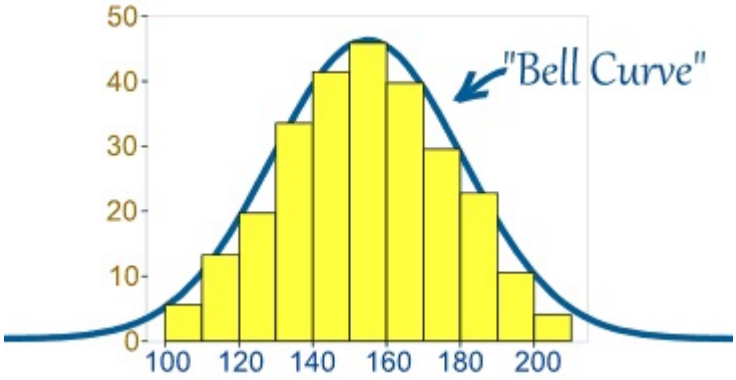


Image Credit: mathisfun.com

9) What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.

10) What are Interpolation and Extrapolation?

Estimating a value from 2 known values from a list of values is Interpolation. Extrapolation is approximating a value by extending a known set of values or facts.

11) What is power analysis?

An experimental design technique for determining the effect of a given sample size.

12) What is Collaborative filtering?

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources, and multiple agents.

13) What is the difference between Cluster and Systematic Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. A cluster sample is a probability sample where each sampling unit is a collection or cluster of elements. Systematic

sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it progresses from the top again. The best example for systematic sampling is the equal probability method.

14) Are expected value and mean value different?

They are not different but the terms are used in different contexts. Mean is generally referred to when talking about a probability distribution of sample population whereas expected value is generally referred to in a random variable context.

For Sampling Data

The mean value is the only value that comes from the sampling data.

Expected Value is the mean of all the means i.e. the value that is built from multiple samples. The expected value is the population mean.

For Distributions

Mean value and Expected value are the same irrespective of the distribution, under the condition that the distribution is in the same population.

15) What does P-value signify about the statistical data?

P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.

- P-Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.

- P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.

- P-value=0.05is the marginal value indicating it is possible to go either way.

16) Do gradient descent methods always converge to the same point?

No, they do not because in some cases it reaches a local minima or a local optima point. You don't reach the global optima point. It depends on the data and starting conditions

17) What is the benefit of shuffling a training dataset when using a batch gradient descent algorithm for optimizing a neural network?

18) A test has a true positive rate of 100% and false-positive rate of 5%. There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?

Let's suppose you are being tested for a disease, if you have the illness the test will end up saying you have the illness. However, if you don't have the illness- 5% of the time the test will end up saying you have the illness, and 95% of the time the test will give an accurate result that you don't have the illness. Thus there is a 5% error in case you do not have the illness.

Out of 1000 people, 1 person who has the disease will get a true positive result.

Out of the remaining 999 people, 5% will also get true positive results.

Close to 50 people will get a true positive result for the disease.

This means that out of 1000 people, 51 people will be tested positive for the disease even though only one person has the illness. There is only a 2% probability of you having the disease even if your reports say that you have the disease.

19) What is the difference between Supervised Learning and Unsupervised Learning?

If an algorithm learns something from the training data so that the knowledge can be applied to the test data, then it is referred to as Supervised Learning. Classification is an example of Supervised Learning. If the algorithm does not learn anything beforehand because there is no response variable or any training data, then it is referred to as unsupervised learning. Clustering is an example of unsupervised learning.

20) What is the goal of A/B Testing?

It is a statistical hypothesis testing for a randomized experiment with two variables A and B. The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. An example of this could be identifying the click-through rate for a banner ad.

21) What is an Eigenvalue and Eigenvector?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing, or stretching. Eigenvalue can be referred to as the strength of the transformation in the direction of the eigenvector or the factor by which the compression occurs.

22) What are various steps involved in an analytics project?

- Understand the business problem
- Explore the data and become familiar with it.
- Prepare the data for modeling by detecting outliers, treating missing values, transforming variables, etc.

- After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step till the best possible outcome is achieved.

- Validate the model using a new data set.

- Start implementing the model and track the result to analyse the performance of the model over the period of time.

23) How can you iterate over a list and also retrieve element indices at the same time?

This can be done using the enumerate function which takes every element in a sequence just like in a list and adds its location just before it.

24) During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored.There are various factors to be considered when answering this question-

- Understand the problem statement, understand the data and then give the answer.Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.
- If it is a categorical variable, the default value is assigned. The missing value is assigned a default value.
- If you have a distribution of data coming, for normal distribution give the mean value.
- Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

25) Explain about the box cox transformation in regression models.

For some reason or the other, the response variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.

26) Can you use machine learning for time series analysis?

Yes, it can be used but it depends on the applications.

27) Write a function that takes in two sorted lists and outputs a sorted list that is their union.

First solution which will come to your mind is to merge two lists and short them afterwards

Python code-

```
def return_union(list_a, list_b):
        return sorted(list_a + list_b)
```

R code-

```
return_union <- function(list_a, list_b)
{
list_c<-list(c(unlist(list_a),unlist(list_b)))
return(list(list_c[[1]][order(list_c[[1]])]))
}
```

Generally, the tricky part of the question is not to use any sorting or ordering function. In that case you will have to write your own logic to answer the question and impress your interviewer.

Python code-

```
def return_union(list_a, list_b):
        len1 = len(list_a)
        len2 = len(list_b)
    final_sorted_list = []
        j = 0
        k = 0

        for i in range(len1+len2):
        if k == len1:
            final_sorted_list.extend(list_b[j:])
        break
        elif j == len2:
            final_sorted_list.extend(list_a[k:])
        break
```

```
        elif list_a[k] < list_b[j]:
        final_sorted_list.append(list_a[k])
        k += 1
        else:
        final_sorted_list.append(list_b[j])
        j += 1
        return final_sorted_list
```

Similar functions can be returned in R as well by following the similar steps.

```
return_union <- function(list_a,list_b)
{
#Initializing length variables
len_a <- length(list_a)
len_b <- length(list_b)
len <- len_a + len_b

#initializing counter variables
j=1
k=1
#Creating an empty list which has length equal to sum of both the lists
list_c <- list(rep(NA,len))
#Here goes our for loop
for(i in 1:len)
  {
        if(j>len_a)
        {
```

```
        list_c[i:len] <- list_b[k:len_b]
        break
        }
        else if(k>len_b)
        {
        list_c[i:len] <- list_a[j:len_a]
        break
        }
        else if(list_a[[j]] <= list_b[[k]])
        {
        list_c[[i]] <- list_a[[j]]
        j <- j+1
        }
        else if(list_a[[j]] > list_b[[k]])
        {
        list_c[[i]] <- list_b[[k]]
        k <- k+1
        }
 }
 return(list(unlist(list_c)))
 }
```

28) What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)?

In Bayesian estimation we have some knowledge about the data/problem (prior) .There may be several values of the parameters which explain data and hence we can look for multiple parameters like 5 gammas and 5 lambdas that do this. As a result of Bayesian Estimate, we get multiple models for making multiple predictions i.e. one for each pair of parameters but with the

same prior. So, if a new example needs to be predicted then computing the weighted sum of these predictions serves the purpose.

Maximum likelihood does not take prior into consideration (ignores the prior) so it is like being a Bayesian while using some kind of a flat prior.

29) What is Machine Learning?

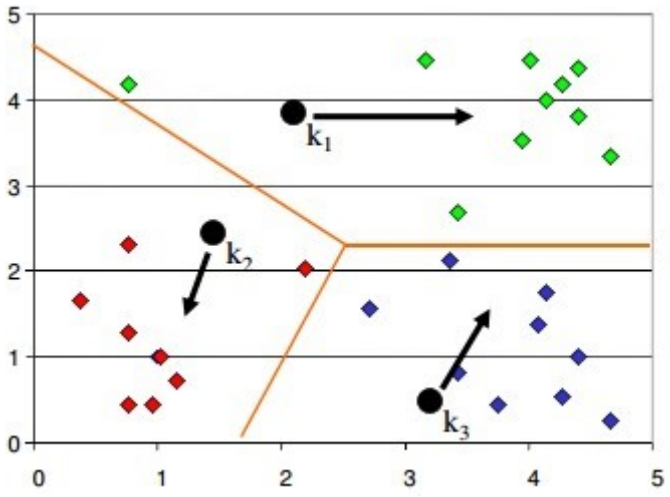
The simplest way to answer this question is – we give the data and equation to the machine. Ask the machine to look at the data and identify the coefficient values in an equation.

For example for the linear regression y=mx+c, we give the data for the variable x, y and the machine learns about the values of m and c from the data.

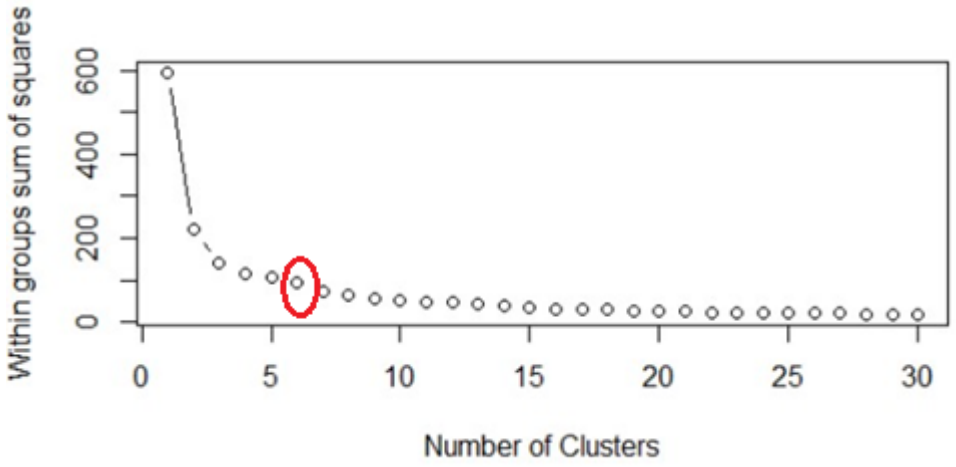
30) How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question will mostly be asked in reference to K-Means clustering where "K" defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.



Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.



Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don't see any decrement in WSS. This point is known as the bending point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendrograms and identify the distinct groups from there.

31) Is it possible to perform logistic regression with Microsoft Excel?

It is possible to perform logistic regression with Microsoft Excel. There are two ways to do it using Excel.

a) One is to use Add-ins provided by many websites which we can use.

b) Second is to use fundamentals of logistic regression and use Excel's computational power to build a logistic regression

But when this question is being asked in an interview, the interviewer is not looking for a name of Add-ins rather a method using the base excel functionalities.

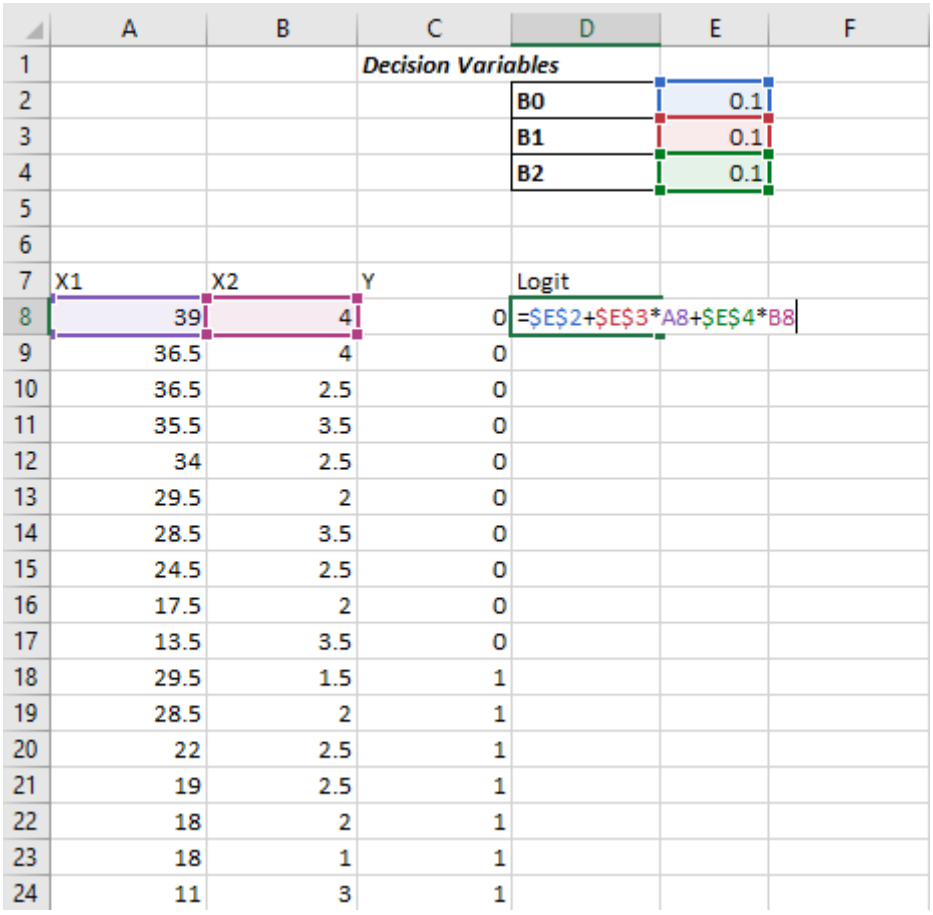
Let's use a sample data to learn about logistic regression using Excel. (Example assumes that you are familiar with basic concepts of logistic regression)

	A	B	C
6			
7	X1	X2	Y
8	39	4	0
9	36.5	4	0
10	36.5	2.5	0
11	35.5	3.5	0
12	34	2.5	0
13	29.5	2	0
14	28.5	3.5	0
15	24.5	2.5	0
16	17.5	2	0
17	13.5	3.5	0
18	29.5	1.5	1
19	28.5	2	1
20	22	2.5	1
21	19	2.5	1
22	18	2	1
23	18	1	1
24	11	3	1
25	11	2.5	1
26	7.5	2	1
27	5	3	1

Data shown above consists of three variables where X1 and X2 are independent variables and Y is a class variable. We have kept only 2 categories for our purpose of binary logistic regression classifier.

Next we have to create a logit function using independent variables, i.e.

Logit = L = $\beta 0 + \beta 1 * X1 + \beta 2 * X2$



	A	B	C	D	E	F
1			*Decision Variables*			
2				**B0**	0.1	
3				**B1**	0.1	
4				**B2**	0.1	
5						
6						
7	X1	X2	Y	Logit		
8	39	4	0	=$E$2+$E$3*A8+$E$4*B8		
9	36.5	4	0			
10	36.5	2.5	0			
11	35.5	3.5	0			
12	34	2.5	0			
13	29.5	2	0			
14	28.5	3.5	0			
15	24.5	2.5	0			
16	17.5	2	0			
17	13.5	3.5	0			
18	29.5	1.5	1			
19	28.5	2	1			
20	22	2.5	1			
21	19	2.5	1			
22	18	2	1			
23	18	1	1			
24	11	3	1			

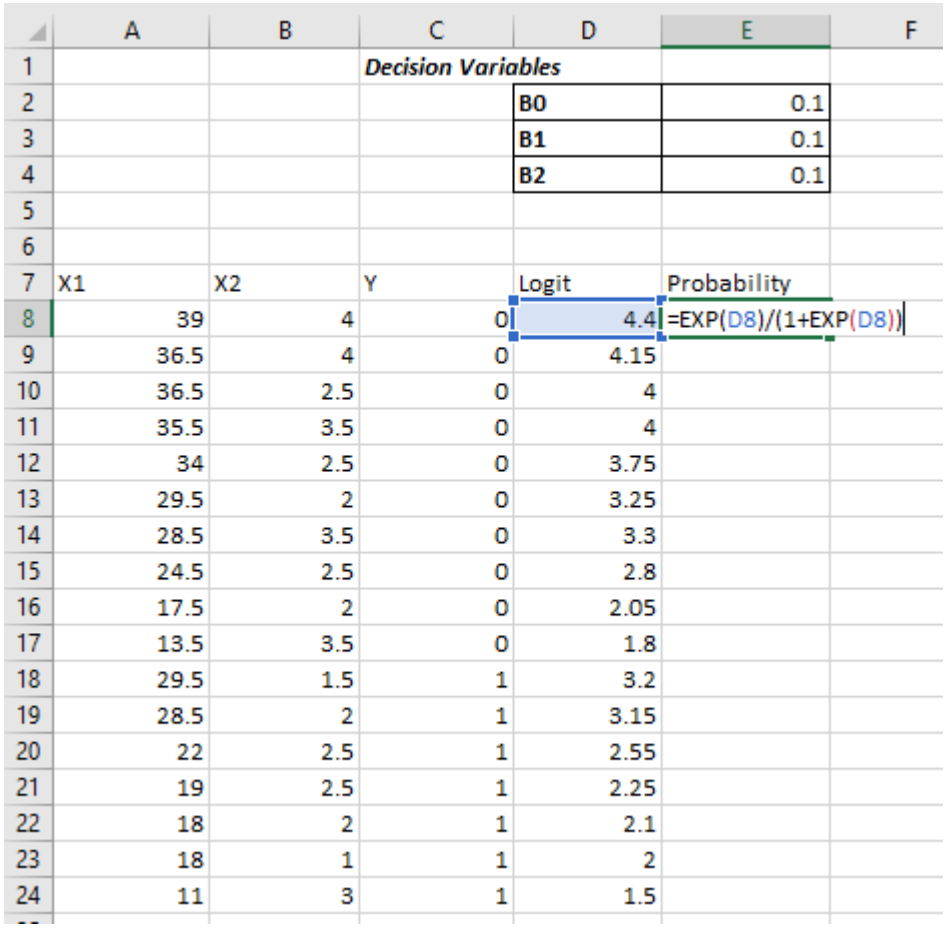
We have kept the initial values of beta 1, beta 2 as 0.1 for now and we will use Excel Solver to optimize the beta values in order to maximize our log likelihood estimate.

Assuming that you are aware of logistic regression basics, we calculate probability values from Logit using following formula:

Probability= e^Logit/(1+ e^Logit)

e is base of natural logarithm i.e. $e = 2.71828163$

Let's put it into an excel formula to calculate probability values for each of the observation.



	A	B	C	D	E	F
1			*Decision Variables*			
2				**B0**	0.1	
3				**B1**	0.1	
4				**B2**	0.1	
5						
6						
7	X1	X2	Y	Logit	Probability	
8	39	4	0	4.4	=EXP(D8)/(1+EXP(D8))	
9	36.5	4	0	4.15		
10	36.5	2.5	0	4		
11	35.5	3.5	0	4		
12	34	2.5	0	3.75		
13	29.5	2	0	3.25		
14	28.5	3.5	0	3.3		
15	24.5	2.5	0	2.8		
16	17.5	2	0	2.05		
17	13.5	3.5	0	1.8		
18	29.5	1.5	1	3.2		
19	28.5	2	1	3.15		
20	22	2.5	1	2.55		
21	19	2.5	1	2.25		
22	18	2	1	2.1		
23	18	1	1	2		
24	11	3	1	1.5		

The conditional probability is the probability of Predicted Y, given a set of independent variables X.

And this *p* can be calculated as-

$$P[(X)]^{Yactual}*[1-P[(X)]^{(1-Yactual)}]$$

Then we have to take natural log of the above function-

$$ln[[\ ]\ P[(X)]^{Yactual}*[1-P(X)^{(1-Yactual)}\ ]]$$

Which turns out to be –

$$Yactual*ln[[\ ]\ P(X)]*(Yactual-\ 1)*ln[1-P(X)]$$

Log likelihood function LL is the sum of above equation for all the observations

	A	B	C	D	E	F	G
1			***Decision Variables***				
2				**B0**	0.1		
3				**B1**	0.1		
4				**B2**	0.1		
5							
6							
7	X1	X2	Y	Logit	Probability	P(Y=y|X)	
8	39	4	0	4.4	0.987871565	=C8*LN(E8)+(1-C8)*LN(1-E8)	
9	36.5	4	0	4.15	0.984480243		
10	36.5	2.5	0	4	0.98201379		
11	35.5	3.5	0	4	0.98201379		
12	34	2.5	0	3.75	0.97702263		
13	29.5	2	0	3.25	0.962673113		
14	28.5	3.5	0	3.3	0.964428811		
15	24.5	2.5	0	2.8	0.942675824		
16	17.5	2	0	2.05	0.885947619		
17	13.5	3.5	0	1.8	0.858148935		
18	29.5	1.5	1	3.2	0.960834277		
19	28.5	2	1	3.15	0.958908722		
20	22	2.5	1	2.55	0.927573515		
21	19	2.5	1	2.25	0.904650535		
22	18	2	1	2.1	0.890903179		
23	18	1	1	2	0.880797078		
24	11	3	1	1.5	0.817574476		

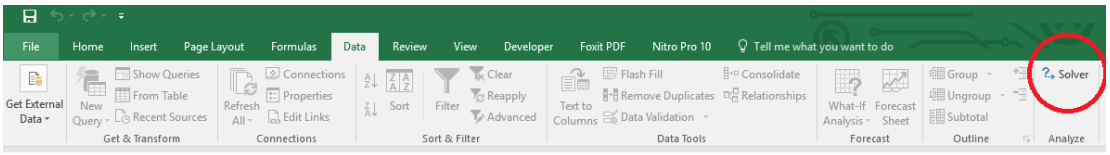
Log likelihood LL will be sum of column G, which we just calculated

	A	B	C	D	E	F	G	H	I
1			*Decision Variables*				*Log Likelihood*		
2				**B0**	0.1			=SUM(F8:F27)	
3				**B1**	0.1				
4				**B2**	0.1				
5									
6									
7	X1	X2	Y	Logit	Probability	P(Y=y|X)			
8	39	4	0	4.4	0.987871565	-4.41220258			
9	36.5	4	0	4.15	0.984480243	-4.16564145			
10	36.5	2.5	0	4	0.98201379	-4.01814993			
11	35.5	3.5	0	4	0.98201379	-4.01814993			
12	34	2.5	0	3.75	0.97702263	-3.77324546			
13	29.5	2	0	3.25	0.962673113	-3.28804137			
14	28.5	3.5	0	3.3	0.964428811	-3.33621926			
15	24.5	2.5	0	2.8	0.942675824	-2.85903283			
16	17.5	2	0	2.05	0.885947619	-2.17109745			
17	13.5	3.5	0	1.8	0.858148935	-1.95297761			
18	29.5	1.5	1	3.2	0.960834277	-0.03995333			
19	28.5	2	1	3.15	0.958908722	-0.04195939			
20	22	2.5	1	2.55	0.927573515	-0.07518323			
21	19	2.5	1	2.25	0.904650535	-0.10020656			
22	18	2	1	2.1	0.890903179	-0.11551952			
23	18	1	1	2	0.880797078	-0.12692801			
24	11	3	1	1.5	0.817574476	-0.20141328			

The objective is to maximize the Log Likelihood i.e. cell H2 in this example. We have to maximize H2 by optimizing B0, B1, and B2.

We'll use Excel's solver add-in to achieve the same.

Excel comes with this Add-in pre-installed and you must see it under Data Tab in Excel as shown below

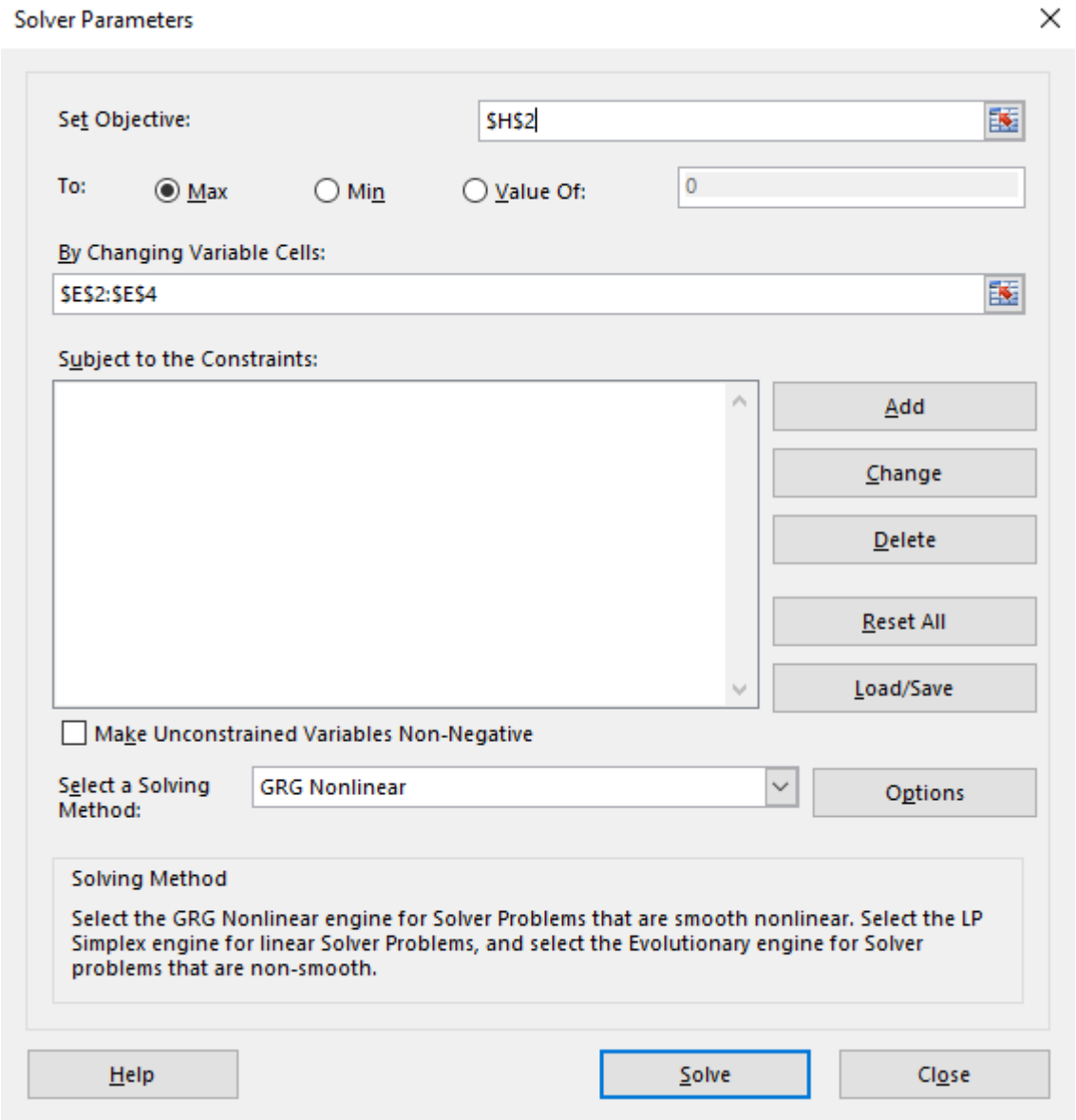


If you don't see it there then make sure if you have loaded it. To load an add-in in Excel,

Go to *File >> Options >> Add-Ins* and see if checkbox in front of required add-in is checked or not? Make sure to check it to load an add-in into Excel.

If you don't see Solver Add-in there, go to the bottom of the screen (Manage Add-Ins) and click on OK. Next you will see a popup window which should have your Solver add-in present. Check the checkbox in-front of the add-in name. If you don't see it there as well click on browse and direct it to the required folder which contains Solver Add-In.

Once you have your Solver loaded, click on Solver icon under Data tab and You will see a new window popped up like –

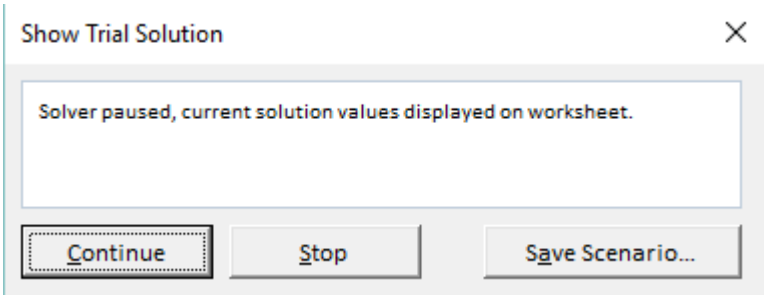


Put *H2* in the set objective, select max and fill cells *E2 to E4* in the next form field.

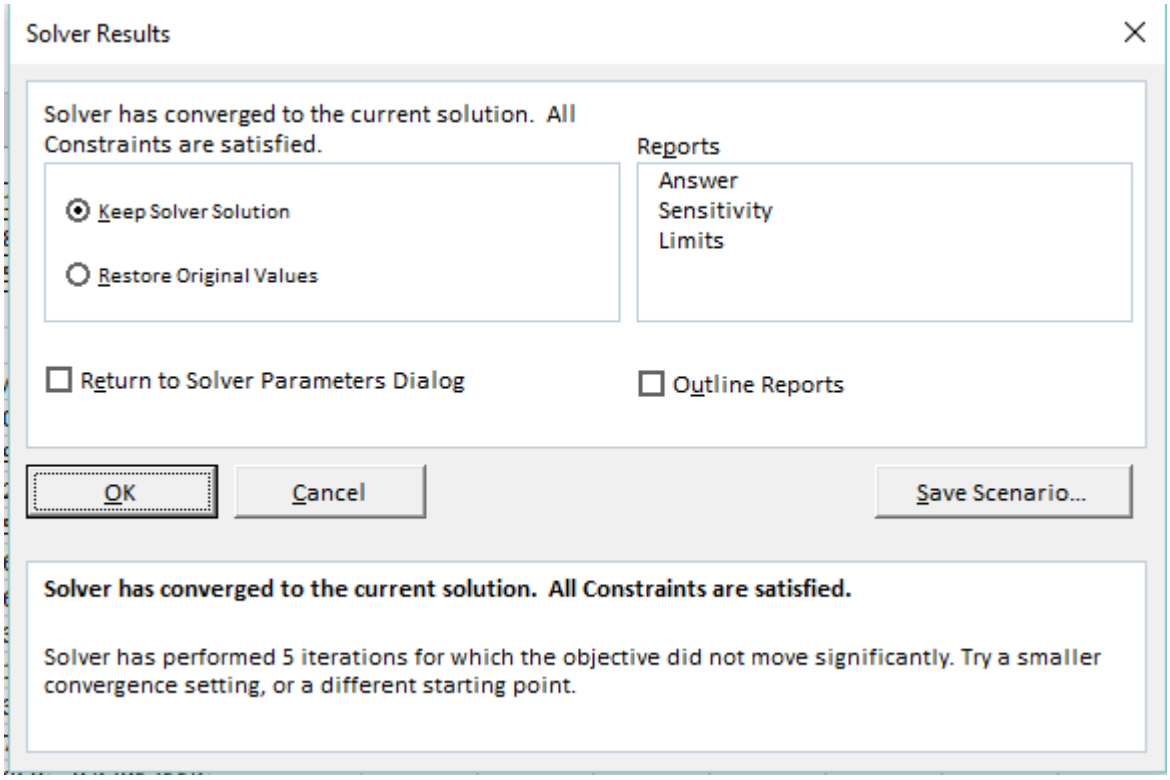
By doing this we have told Solver to Maximize H2 by changing values in cells E2 to E4.

Now click on Solve button at the bottom –

You will see a popup like below -



This shows that Solver has found a local maxima solution but we are in need of Global Maxima Output. Keep clicking on Continue until it shows the below popup



It shows that Solver was able to find and converge the solution. In case it is not able to converge it will throw an error. Select “Keep Solver Solution” and Click on OK to accept the solution provided by Solver.

Now, you can see that the value of Beta coefficients from B0, B1 B2 have changed and our Log Likelihood function has been maximized.

	A	B	C	D	E	F	G	H
1			***Decision Variables***				***Log Likelihood***	
2				**B0**	12.48309171			-6.65456
3				**B1**	-0.23406877			
4				**B2**	-2.93832567			
5								
6								
7	X1	X2	Y	Logit	Probability	P(Y=y|X)		
8	39	4	0	-8.3988928	0.000225066	-0.00022509		
9	36.5	4	0	-7.8137209	0.000403988	-0.00040407		
10	36.5	2.5	0	-3.4062324	0.032101254	-0.0326278		
11	35.5	3.5	0	-6.1104893	0.002214549	-0.00221701		
12	34	2.5	0	-2.8210605	0.056196661	-0.05783746		
13	29.5	2	0	-0.2985882	0.425902646	-0.55495629		
14	28.5	3.5	0	-4.4720079	0.011295312	-0.01135959		
15	24.5	2.5	0	-0.5974072	0.354937108	-0.43840746		
16	17.5	2	0	2.510237	0.924856362	-2.58835382		
17	13.5	3.5	0	-0.9609765	0.276682734	-0.32390733		
18	29.5	1.5	1	1.1705746	0.763248868	-0.27017113		
19	28.5	2	1	-0.0645194	0.483875735	-0.72592715		
20	22	2.5	1	-0.0122353	0.496941214	-0.69928354		
21	19	2.5	1	0.689971	0.665960476	-0.40652496		

Using these values of Betas you can calculate the probability and hence response variable by deciding the probability cut-off.

32) What is the difference between skewed and uniform distribution?

When the observations in a dataset are spread equally across the range of distribution, then it is referred to as uniform distribution. There are no clear perks in a uniform distribution. Distributions that have more observations on one side of the graph than the other are referred to as skewed distribution.Distributions with fewer observations on the left (towards lower values) are said to be skewed left and distributions with fewer observation on the right (towards higher values) are said to be skewed right.

33) You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?

Since the question asked is about post model building exercise, we will assume that you have already tested for null hypothesis, multicollinearity and Standard error of coefficients.

Once you have built the model, you should check for following –

· Global F-test to see the significance of group of independent variables on dependent variable

· R^2

- Adjusted R^2
- RMSE, MAPE

In addition to above mentioned quantitative metrics you should also check for-

- Residual plot
- Assumptions of linear regression

34) What do you understand by Recall and Precision?

Recall measures "Of all the actual true samples how many did we classify as true?"

Precision measures "Of all the samples we classified as true how many are actually true?"

We will explain this with a simple example for better understanding -

Imagine that your wife gave you surprises every year on your anniversary in the last 12 years. One day all of a sudden your wife asks -"Darling, do you remember all anniversary surprises from me?".

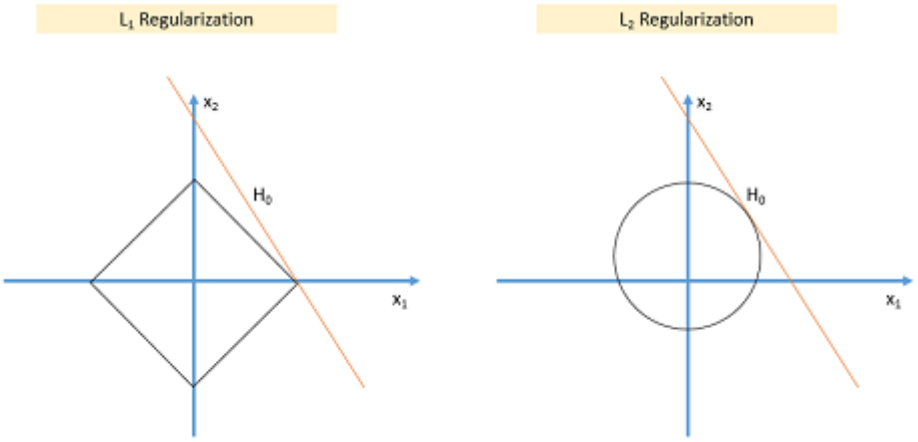
This simple question puts your life into danger.To save your life, you need to Recall all 12 anniversary surprises from your memory. Thus, Recall(R) is the ratio of number of events you can correctly recall to the number of all correct events. If you can recall all the 12 surprises correctly then the recall ratio is 1 (100%) but if you can recall only 10 surprises correctly of the 12 then the recall ratio is 0.83 (83.3%).

However , you might be wrong in some cases. For instance, you answer 15 times, 10 times the surprises you guess are correct and 5 wrong. This implies that your recall ratio is 100% but the precision is 66.67%.

Precision is the ratio of the number of events you can correctly recall to a number of all events you recall (combination of wrong and correct recalls).

35) Why L1 regularization causes parameter sparsity whereas L2 regularization does not?

Regularizations in statistics or in the field of machine learning is used to include some extra information in order to solve a problem in a better way. L1 & L2 regularizations are generally used to add constraints to optimization problems.



In the example shown above H0 is a hypothesis. If you observe, in L1 there is a high likelihood to hit the corners as solutions while in L2, it doesn't. So in L1 variables are penalized more as compared to L2 which results in sparsity.

In other words, errors are squared in L2, so the model sees higher error and tries to minimize that squared error.

36) How can you deal with different types of seasonality in time series modelling?

Seasonality in time series occurs when time series shows a repeated pattern over time. E.g., stationary sales decrease during the holiday season, air conditioner sales increase during the summers etc. are few examples of seasonality in a time series.

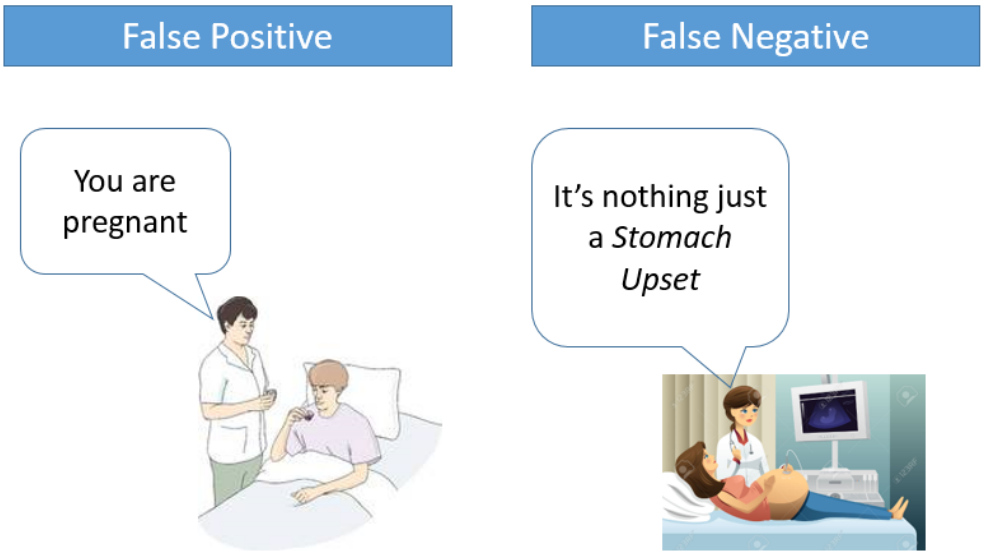
Seasonality makes your time series non-stationary because the average value of the variables at different time periods. Differentiating a time series is generally known as the best method of removing seasonality from a time series. Seasonal differencing can be defined as a numerical difference between a particular value and a value with a periodic lag (i.e. 12, if monthly seasonality is present)

37) Can you cite some examples where a false positive is more important than a false negative?

Before we start, let us understand what are false positives and what are false negatives.

False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.

And, False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.



In the medical field, assume you have to give chemotherapy to patients. Your lab tests patients for certain vital information and based on those results they decide to give radiation therapy to a patient.

Assume a patient comes to that hospital and he is tested positive for cancer (But he doesn't have cancer) based on lab prediction. What will happen to him? (Assuming Sensitivity is 1)

One more example might come from marketing. Let's say an ecommerce company decided to give a $1000 Gift voucher to the customers whom they assume to purchase at least $5000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above 5K.

Now what if they have sent it to false positive cases?

38) Can you cite some examples where a false negative is more important than a false positive?

Assume there is an airport 'A' which has received high security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to shortage of staff they decided to scan passengers being predicted as risk positives by their predictive model.

What will happen if a true threat customer is being flagged as non-threat by the airport model?

Another example can be the judicial system. What if Jury or judge decides to make a criminal go free?

What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

39) Can you cite some examples where both false positive and false negatives are equally important?

In the banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point of time they don't want to acquire bad customers. In this scenario both the false positives and false negatives become very important to measure.

These days we hear many cases of players using steroids during sports competitions. Every player has to go through a steroid test before the game starts. A false positive can ruin the career of a Great sportsman and a false negative can make the game unfair.

40) Can you explain the difference between a Test Set and a Validation Set?

Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid Overfitting of the model being built. On the other hand, a test set is used for testing or evaluating the performance of a trained machine learning model.

In simple terms ,the differences can be summarized as-

- Training Set is to fit the parameters i.e. weights.
- Test Set is to assess the performance of the model i.e. evaluating the predictive power and generalization.
- Validation set is to tune the parameters.

41) What do you understand about the statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, RF etc.). Sensitivity is nothing but "Predicted TRUE events/ Total events". True events here are the events which were true and the model also predicted them as true.

Calculation of sensitivity is pretty straightforward-

Sensitivity = True Positives /Positives in Actual Dependent Variable

Where, True positives are Positive events which are correctly classified as Positives.

42) What is the importance of having a selection bias?

Selection Bias occurs when there is no appropriate randomization achieved while selecting individuals, groups or data to be analyzed. Selection bias implies that the obtained sample does not exactly represent the population that was actually intended to be analyzed. Selection bias consists of Sampling Bias, Data, Attribute, and Time Interval.

43) How do data management procedures like missing data handling make selection bias worse?

Missing value treatment is one of the primary tasks which a data scientist is supposed to do before starting data analysis. There are multiple methods for missing value treatment. If not done

properly, it could potentially result in selection bias. Let see few missing value treatment examples and their impact on selection-

Complete Case Treatment: Complete case treatment is when you remove an entire row in data even if one value is missing. You could achieve a selection bias if your values are not missing at random and they have some pattern. Assume you are conducting a survey and few people didn't specify their gender. Would you remove all those people? Can't it tell a different story?

Available case analysis: Let say you are trying to calculate a correlation matrix for data so you might remove the missing values from variables that are needed for that particular correlation coefficient. In this case, your values will not be fully correct as they are coming from population sets.

Mean Substitution: In this method, missing values are replaced with the mean of other available values. This might make your distribution biased e.g., standard deviation, correlation and regression are mostly dependent on the mean value of variables.

Hence, various data management procedures might include selection bias in your data if not chosen correctly.

44) What are the basic assumptions to be made for linear regression?

Normality of error distribution, statistical independence of errors, linearity and additivity.

45) Can you write the formula to calculate R-square?

R-Square can be calculated using the below formula -

1 - (Residual Sum of Squares/ Total Sum of Squares)

46) What is the advantage of performing dimensionality reduction before fitting an SVM?

Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

47) How will you assess the statistical significance of an insight whether it is a real insight or just by chance?

Statistical importance of an insight can be accessed using Hypothesis Testing.

48) How would you create a taxonomy to identify key customer trends in unstructured data?

The best way to approach this question is to mention that it is good to check with the business owner and understand their objectives before categorizing the data. Having done this, it is always good to follow an iterative approach by pulling new data samples and improving the model accordingly by validating it for accuracy by soliciting feedback from the stakeholders of the

business. This helps ensure that your model is producing actionable results and improving over the time.

49) How will you find the correlation between a categorical variable and a continuous variable ?

You can use the analysis of covariance technique to find the correlation between a categorical variable and a continuous variable.

50) How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values. The most common ways to treat outlier values –

1) To change the value and bring in within a range

2) To just remove the value.

51) How can you assess a good logistic model?

There are various methods to assess the results of logistic regression analysis-

- Using a Classification Matrix to look at the true negatives and false positives.
- Concordance that helps identify the ability of the logistic model to differentiate between the event happening and not happening.
- Lift helps assess the logistic model by comparing it with random selection.

52) What is multicollinearity and how you can overcome it?

SVM and Random Forest are both used in classification problems.

a) If you are sure that your data is outlier free and clean then go for SVM. It is the opposite - if your data might contain outliers then Random forest would be the best choice

b) Generally, SVM consumes more computational power than Random Forest, so if you are constrained with memory go for the Random Forest machine learning algorithm.

c) Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose the Random Forest machine learning algorithm.

d) Random Forest machine learning algorithms are preferred for multiclass problems.

e) SVM is preferred in multi-dimensional problem set - like text classification

but as a good data scientist, you should experiment with both of them and test for accuracy, or rather you can use an ensemble of many Machine Learning techniques.

53) You are given a dataset with 1500 observations and 15 features. How many observations will you select in each decision tree in a random forest?

Each decision tree has a subset of features but includes all the observations from the dataset. In this case, the answer will be 1500 as the tree will include all the observations from the dataset.

54) How will you evaluate the performance of a logistic regression model?

It's very much obvious that you would mention accuracy as the answer to this question but since logistic regression is not the same as linear regression it will mislead. You should mention how you will use the confusion matrix to evaluate the performance and the various statistics related to it like Precision, Specificity, Sensitivity, or Recall. You get bonus points for mentioning Concordance, Discordance, and AUC.

55) What is Selection Bias ?

Selection bias is the term used to describe the situation where an analysis has been conducted among a subset of the data (a sample) with the goal of drawing conclusions about the population, but the resulting conclusions will likely be wrong (biased), because the subgroup differs from the population in some important way. Selection bias is usually introduced as an error with the sampling and having a selection for analysis that is not properly randomized.

56) What is bias-variance trade-off?

Bias: Bias is an error introduced in your model due to oversimplification of the machine learning algorithm. It can lead to underfitting. When you train your model at that time the model makes simplified assumptions to make the target function easier to understand. Low bias machine learning algorithms — Decision Trees, k-NN and SVM High bias machine learning algorithms — Linear Regression, Logistic Regression.

Variance: Variance is an error introduced in your model due to a complex machine learning algorithm, your model learns noise also from the training data set and performs badly on the test data set. It can lead to high sensitivity and overfitting.Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.
Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

- The k-nearest neighbour algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbours that contribute to the prediction and in turn increases the bias of the model.
- The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

57) Differentiate between point estimates and confidence intervals .

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters. A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by 1 — alpha, where alpha is the level of significance.

58) What do you understand by pruning in decision trees ?

Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. So, when we remove sub-nodes of a decision node, this process is called pruning or the opposite process of splitting.

59) What do you understand by Covariance and Correlation ?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.
Correlation: Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are
related.
Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable are reciprocal by a corresponding change in another variable.

60) What is the use of a confusion matrix?

The confusion matrix is a 2X2 table that contains 4 outputs provided by the **binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are

derived from it. *Confusion Matrix.* A data set used for performance evaluation is called a **test data set**. It should contain the correct labels and predicted labels. The predicted labels will be exactly the same if the performance of a binary classifier is perfect.

The most common measures derived from the confusion matrix include -

1. Error Rate = (FP+FN)/(P+N)
2. Accuracy=(TP+TN)/(P+N)
3. Sensitivity(Recall or True positive rate) = TP/P
4. Specificity(True negative rate) = TN/N
5. Precision(Positive predictive value) = TP/(TP+FP)
6. F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/(b2PREC+REC) where b is

commonly 0.5, 1, 2.

61) **What cross-validation technique would you use on a time series data set?**

Instead of using k-fold cross-validation, you should be aware of the fact that a time series is not randomly distributed data — It is inherently ordered by chronological order.

In case of time series data, you should use techniques like forward=chaining — Where you will model on past data then look at forward-facing data.

fold 1: training[1], test[2]
fold 1: training[1 2], test[3]
fold 1: training[1 2 3], test[4]
fold 1: training[1 2 3 4], test[5]

62) Explain the working of a Random Forest Machine Learning algorithm.

The underlying principle of this technique is that several weak learners combined to provide a keen learner. The steps involved are

- Build several decision trees on bootstrapped training samples of data
- On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors
- Rule of thumb: At each split $m=p\sqrt{m=p}$
- Predictions: At the majority rule

63) What is Survivorship bias ?

Survivorship bias is the logical error of focusing on aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.

64) How often should you update a machine learning algorithm ?

- When you want the machine learning model to evolve as data streams through infrastructure
- When the underlying data source is changing
- There is a case of non-stationarity

65) Mention about the types of bias in Sampling .

1. Selection Bias

2. Undercoverage Bias

3. Survivorship Bias

66) What are confounding variables in statistics ?

These are extraneous variables in a statistical model that correlates directly or inversely with both the dependent and the independent variable. The estimate fails to account for the confounding factor.

67) What are the disadvantages of a linear regression model ?

- The assumption of linearity of the errors
- It can't be used for count outcomes or binary outcomes
- There are overfitting problems that it can't solve

68) How will you maintain an already deployed machine learning model ?

The steps to maintain a deployed model are:

1) Monitor

Constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it's doing what it's supposed to do.

2) Evaluate

Evaluation metrics of the current model are calculated to determine if a new algorithm is needed.

3) Compare

The new models are compared to each other to determine which model performs the best.

4) Rebuild

The best performing model is re-built on the current state of data.

69) Explain the law of large numbers.

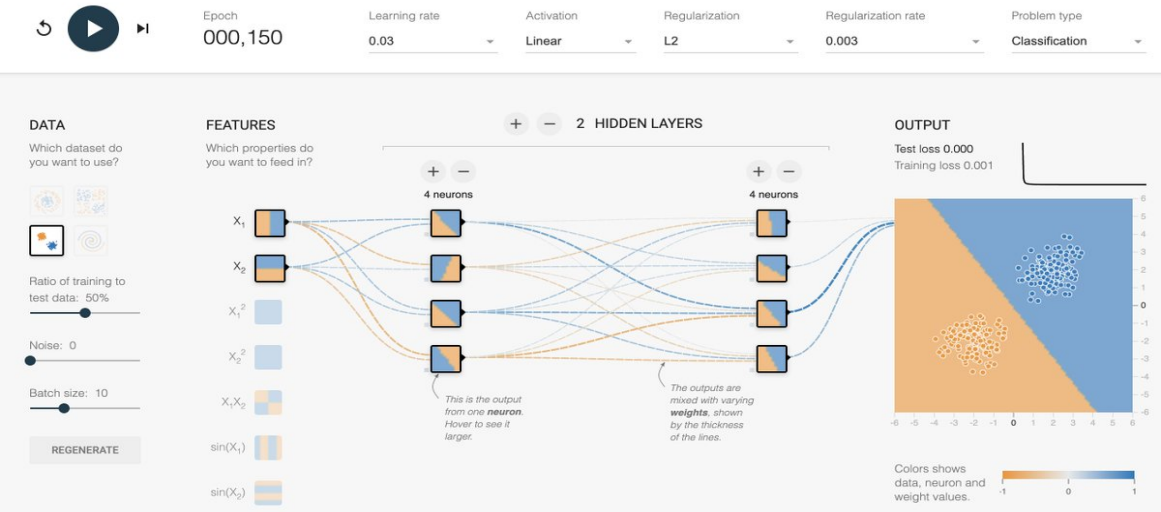
It is a theorem that describes the result of performing the same experiment very frequently. This theorem forms the basis of frequency-style thinking. It states that the sample mean, sample variance and sample standard deviation converge to what they are trying to estimate.

70) Explain the importance of introducing non-linearities in a neural network.

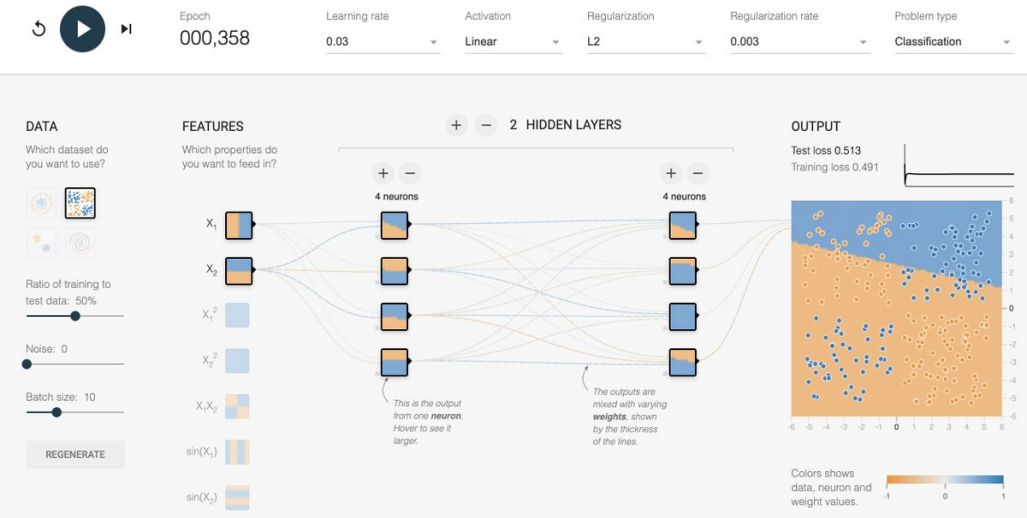
Here is a neural network with 2 hidden layers of 4 neurons. The activation is set to "Linear."

In just a few epochs, the network finds the correct solution.

Notice how the network uses a single dividing line in the output. That's all it can do with linear activations.

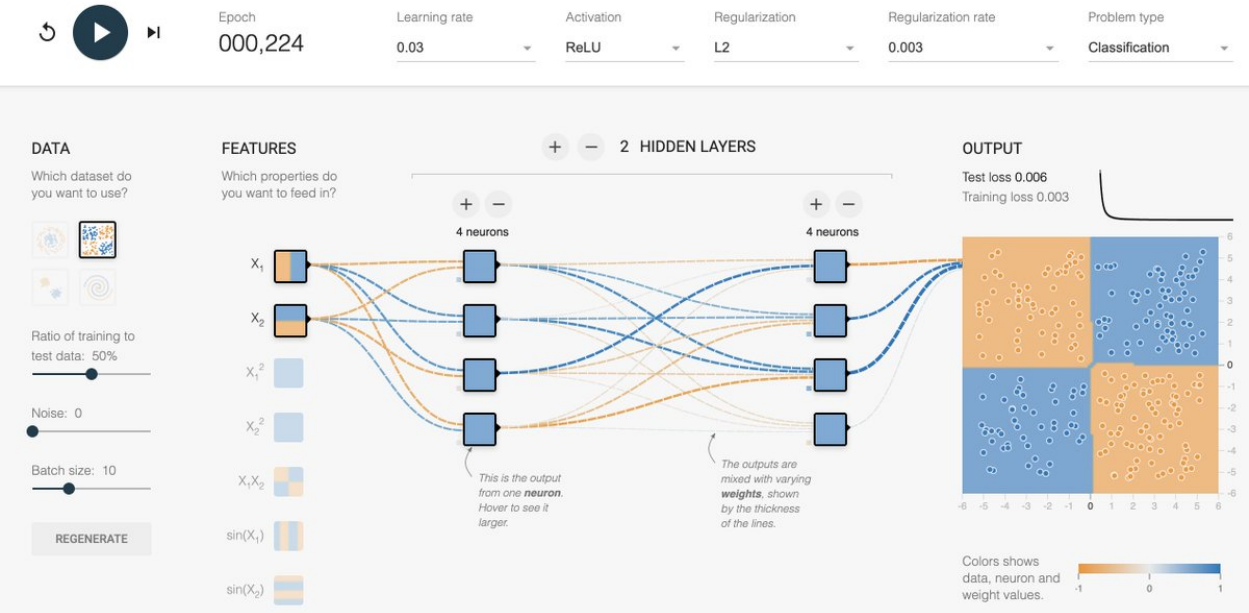


If we try the same network on the more complex problem, it will struggle to classify the data correctly.
We haven't introduced non-linearities in this network, so it won't find the proper solution for this type of problem.



Let's now change the activation to ReLU. This will introduce the nonlinearity we need for the network to become more powerful.

Same exact network structure, with just a different activation, and we are now in business!

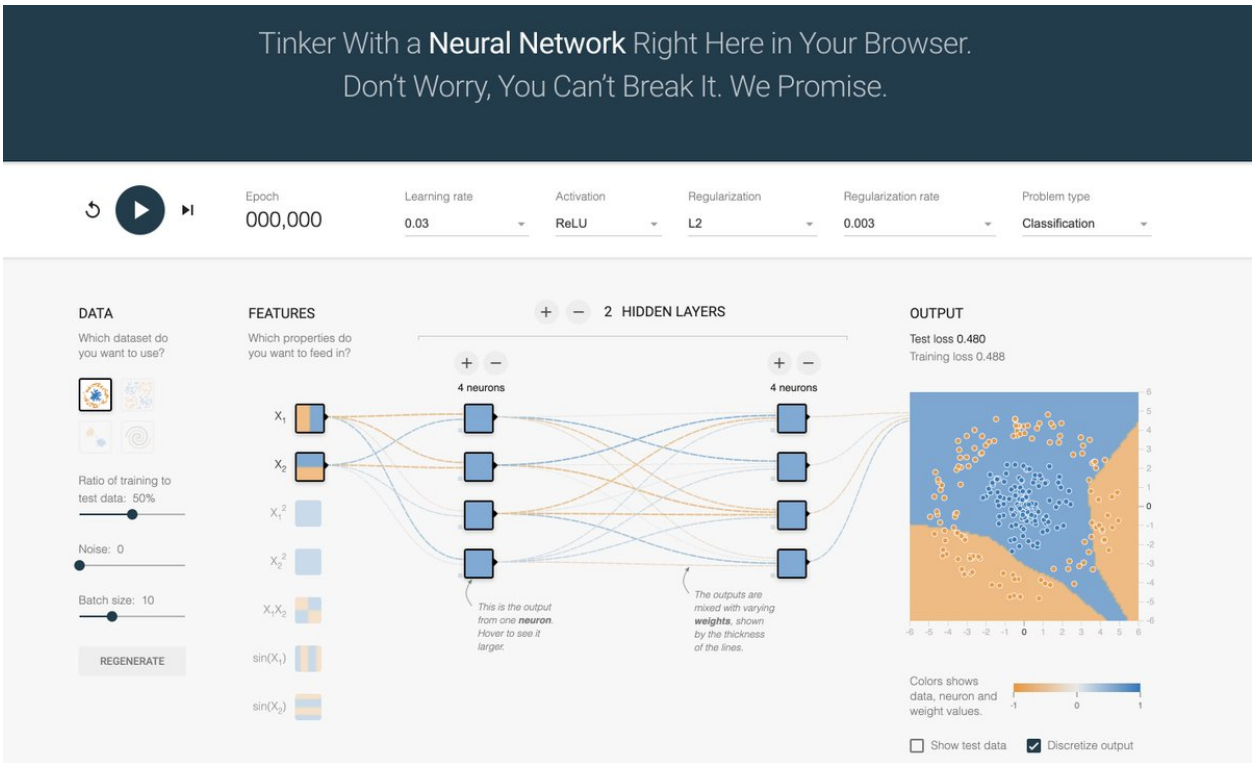


A neural network with linear activation functions and n layers with m hidden units is equivalent to a linear neural network without hidden layers.

In other words, a neural network without non-linear activation functions can only find linear separation boundaries.

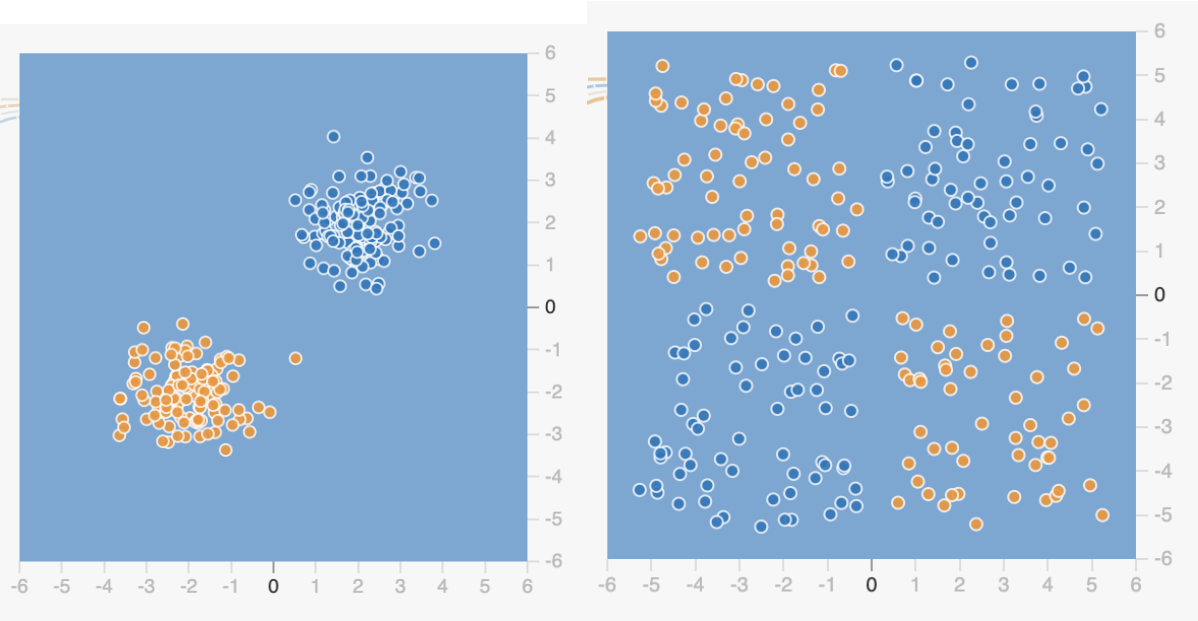
The picture below comes from the TensorFlow Playground.

If you want to see how a neural network works and analyze how it works, take a look at it!



The short answer: So we can solve more interesting problems.

The left image shows a classification problem that can be solved using a single dividing line. The image on the right is much more complex.



Common Data Scientist Interview Questions

1) What is K-means? How can you select K for K-means?

2) How can you make data normal using Box-Cox transformation?

3) Why is it not advisable to use a softmax output activation function in a multi-label classification problem for a one-hot encoded target?

4) Why is vectorization considered a powerful method for optimizing numerical code?

5) What is Gradient Descent?

6) Differentiate between a multi-label classification problem and a multi-class classification problem.

7) What is the difference between gradient descent optimization algorithms Adam and Momentum?

8) What is Regularization and what kind of problems does regularization solve?

9) How will you tackle an exploding gradient problem?

10) How will you tackle a vanishing gradient problem?

11) How do you decide whether your linear regression model fits the data?

12) What is the difference between squared error and absolute error?

13) How are confidence intervals constructed and how will you interpret them?

14) If the training loss of your model is high and almost equal to the validation loss, what does it mean? What should you do?

15) How can you overcome Overfitting?

16) Differentiate between wide and tall data formats?

17) Is Naïve Bayes bad? If yes, under what aspects.

18) How would you develop a model to identify plagiarism?

19) How important it is to introduce non-linearities in a neural network and why?

20) What do you understand by Fuzzy merging ? Which language will you use to handle it?

21) What do you understand by Hypothesis in the content of Machine Learning?

22) How will you find the right K for K-means?

23) Explain evaluation protocols for testing your models? Compare hold-out vs k-fold cross validation vs iterated k-fold cross-validation methods of testing.

24) What do you understand by conjugate-prior with respect to Naïve Bayes?

25) What makes a dataset gold standard?

26) Given that you let the models run long enough, will all gradient descent algorithms lead to the same model when working with Logistic or Linear regression problems?

27) Differentiate between Batch Gradient Descent, Mini-Batch Gradient Descent, and Stochastic Gradient Descent.

28) What are the advantages and disadvantages of using regularization methods like Ridge Regression?

29) What do you understand about long and wide data formats?

30) What do you understand by outliers and inliers? What would you do if you find them in your dataset?

31) Write a program in Python that takes input as the diameter of a coin and weight of the coin and produces output as the money value of the coin.

32) Is it better to have too many false negatives or too many false positives?

33) In experimental design, is it necessary to do randomization? If yes, why?

34) What are the benefits of using a convolutional neural network over a fully connected network when working with image classification problems?

35) What are the benefits of using a recurrent neural network over a fully connected network when working with text data?

36) What do you understand by feature vectors?

37) What are categorical variables?

38) What is the benefit of weight initialization in neural networks?

39) How does the use of dropout work as a regularizer for deep neural networks?

40) How beneficial is dropout regularization in deep learning models? Does it speed up or slow down the training process and why?

41) How will you explain logistic regression to an economist, physician-scientist, and biologist?

42) What is the benefit of batch normalization?

43) Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.

44) What is the curse of dimensionality?

45) Do you need to do feature engineering and feature extraction when applying deep learning models?

46) How will you calculate the accuracy of a model using a confusion matrix?

47) According to the universal approximation theorem, any function can be approximated as closely as required using single collinearity. Then why do people use more?

48) Explain the use of Combinatorics in data science.

49) How will you declare a time series data as stationary ?

50) What makes Tensorflow the most preferred deep learning library?