Analytics Vidhya

Machine Learning Interview Guide

50 Most Common Questions

Dear Reader,

Thank you for your interest in the Machine Learning Interview and for purchasing this book.

First, congratulations on picking data science as your future career! There hasn't been a better time to get into this field with the demand for data science professionals far outstripping the supply.

We understand how difficult landing that first data science role can be. The biggest hurdle can often by clearing the interviews! You might have studied up on all the techniques and tools, brushed up on your previous experience, and yet you can't quite crack the interview process.

That's why we combined our collective experience of conducting hundreds of interviews and created the interview Guide.

This book is a sample of the far more comprehensive handbook of data science interview questions. We have picked out some of the most intriguing questions from that book for your perusal here.

The complete handbook is part of the **'Ace Data Science Interviews' course**. It contains over 240 questions based on data science concepts. There is no other resource quite like it anywhere.

Wondering what else to expect in the course? We have put together a comprehensive 7-step framework to help you land your first data science role. We have spoken more about this framework in an article complete with an illustrated infographic detailing each step.

We hope you find this e-book useful and we'll see you in the course!

You can also check out our Ace Data Science Interviews website and podcast for more tips and tricks on how to crack the interview process.

Thanks,
Analytics Vidhya Team

Table of Contents

Puzzles

Question 1

You pull out 2 balls, one after another, from a bag which has 20 blue and 13 red balls in total. If the balls are of similar colour, then the balls are replaced with a blue ball, however, if the balls are of different colours, then a red ball is used to replace them.

Once the balls are taken out of the bag, they are not placed back in the bag, and thus the number of balls keep reducing. Determine the colour of last ball left in the bag.

Answer 1

If we pull out 2 red balls, we need to replace them with a blue ball. On the other hand, if we draw one red and one blue, it is replaced with a red one. This implies that the red ball would always be in odd numbers, whether we remove 2 together, or remove 1 while adding 1.

This also indicates that the last ball to stay in the bag would be a red one.

Question 2

There are 3 mislabeled jars, with apple and oranges in the first and second jar respectively. The third jar contains a mixture of apples and oranges. You can pick as many fruits as required to precisely label each jar. Determine the minimum number of fruits to be picked up in the process of labeling the jars.

Answer 2

A noticeable aspect in this puzzles is the fact that there’s a circular misplacement, which implies if apple is wrongly labelled as Apple, Apple can’t be labelled as Orange,
i.e., it has to be labeled as A+O. We are acquainted with the fact that everything is wrongly placed, which means A+O jar contains either Apple or Orange (but not both).

The candidate picks one fruit from A+O, and let’s assume he gets an apple. He labels the jar as apple, however, jar labelled Apple can’t have A+O. Thus, the third jar left in the process should be labelled A+O.

Question 3

There are 5 pirates in a ship. Pirates have hierarchy C1, C2, C3, C4 and C5. C1 designation is the highest and C5 is the lowest. These pirates have three characteristics: a. Every pirate is so greedy that he can even take lives to make more money. b. Every pirate desperately wants to stay alive. c. They are all very intelligent.

There are total 100 gold coins on the ship. The person with the highest designation on the deck is expected to make the distribution. If the majority on the deck does not agree to the distribution proposed, the highest designation pirate will be thrown out of the ship (or simply killed). Only the person with the highest designation can be killed at any moment. What is the right distribution of the coins proposed by the captain so that he is not killed and does make maximum amount?

Answer 3

The solution of this problem lies in thinking through what will happen if all the pirates were thrown one by one and then thinking in reverse order.

Let us name pirates as A,B,C,D and E in hierarchy (A being highest).

If only D and E are left at end, D will simply give 0 coins to E and still escape because majority cannot be reached. Hence, even if E gets 1 coin he will give his vote to the distributor.

If C, D and E are there on the deck, C will simply give one coin to E to get his vote. And D simply gets nothing. Hence, even if D gets 1 coin he will give his vote to the distributor.

If B,C,D and E are there on the deck, B will simply give one coin to D to get his vote. C & E simply gets nothing.

If A,B,C,D and E are there on the deck, A simply gives 1 coin each to C and E to get their votes.

Hence, in the final solution A gets 98 coins and only C & E get 1 coin each.

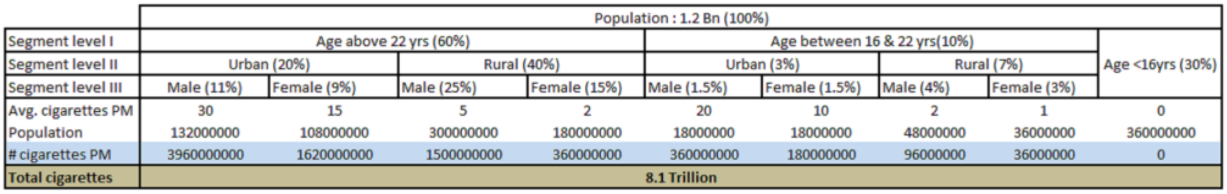
Guesstimates

Question 4

Estimate the number of cigarettes consumed monthly in India.

Answer 4

The population of India, i.e., 1.2 billion. Following is an effective way to segment this population:



	Population : 1.2 Bn (100%)								
Segment level I	Age above 22 yrs (60%)				Age between 16 & 22 yrs(10%)				Age <16yrs (30%)
Segment level II	Urban (20%)		Rural (40%)		Urban (3%)		Rural (7%)		
Segment level III	Male (11%)	Female (9%)	Male (25%)	Female (15%)	Male (1.5%)	Female (1.5%)	Male (4%)	Female (3%)	
Avg. cigarettes PM	30	15	5	2	20	10	2	1	0
Population	132000000	108000000	300000000	180000000	18000000	18000000	48000000	36000000	360000000
# cigarettes PM	3960000000	1620000000	1500000000	360000000	360000000	180000000	96000000	36000000	0
Total cigarettes	**8.1 Trillion**								

Following were the key considerations in building the segmentation and the intermediate guesses:

The rural population consumes far lesser cigarettes than urban because of the purchasing power difference.

Male consume more cigarettes than female in both urban and rural populations.

Children below 16 years consume a negligible number of cigarettes.

Male to Female ratio in Urban is closer to 1 than that of Rural.

Male to Female ratio in younger generations is closer to 1 than that of older. This is because of the increase in awareness level.

Bulk of population start smoking after getting into a job and hence the average number cigarettes are higher in older groups.

Total number of cigarettes from the supply side also come to around 10 Trillion, which gives a good sense check on the final number.

Question 5

How many iPhones are currently being used in China?

Answer 5

Identify the variables to apply to this problem.

Population of China: Approximately 1.4 billion people.

There are several different approaches from this point; one approach is to make assumptions around the number of people that can afford iPhones rather than considering the number of households.

Based on very basic knowledge of China, even though the country is experiencing extraordinary economic growth, you might assume that the majority of the population is still very low-income and cannot afford an iPhone. Thus, you might estimate that 20% of the population could afford an iPhone.

Therefore, the total potential market size is 20% × 1.4 billion = 280 million iPhones.
What percent of this total market size is penetrated? There are many competing products that are cheaper, therefore we estimate that 20% of this segment is currently using an iPhone.

Using these estimates, 20% × 280 million = 56 million iPhones are currently being used in China.

Case study

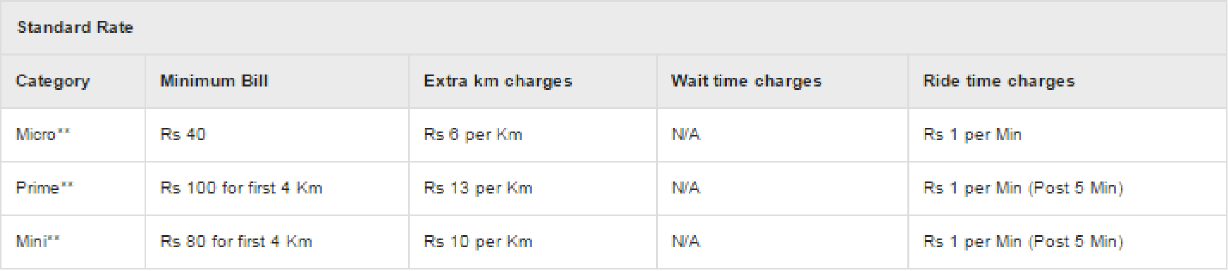
Question 6

There are multiple cab services these days, and you, as a customer has to make an efficient decision, and select the cab based on your requirement. Some important terms that we will use throughout are defined below:

1. **Base Fare**: Initial amount billed to sit in a cab
2. **Excess km fare**: Billed amount on distance after complimentary ride
3. **Time fare**: Billing on the time spent in the cab
4. **Minimum fare**: This is the minimum amount you will be billed
5. **Tolls and excess fee**: This is the excess charge you need to pay to compensate for the long distances outside main city
6. **Taxes**: Taxes is are over and above the bill
7. **Premium multiplier**: In crowded/congested time, you will be bill something like 1.4 – 2.5 X of the actual bill amount.

Here are the details of the three cab services that you must compare.

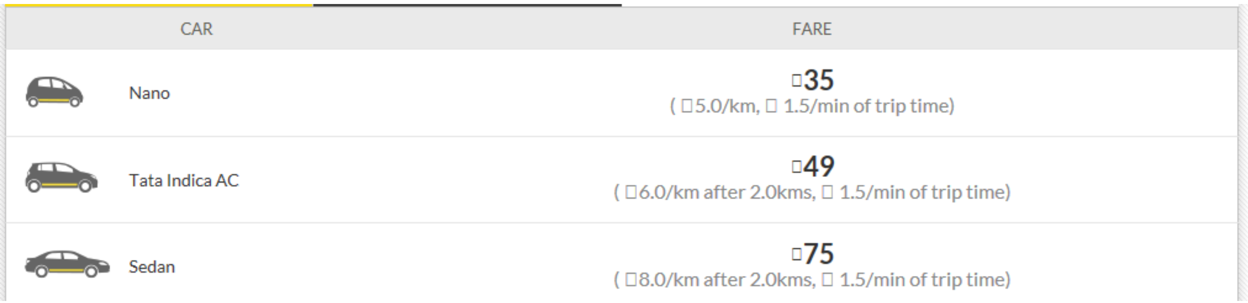
1) The first cab service, let’s call it A, has the following fare for the three types of cabs - micro, mini, prime.



Standard Rate				
Category	Minimum Bill	Extra km charges	Wait time charges	Ride time charges
Micro**	Rs 40	Rs 6 per Km	N/A	Rs 1 per Min
Prime**	Rs 100 for first 4 Km	Rs 13 per Km	N/A	Rs 1 per Min (Post 5 Min)
Mini**	Rs 80 for first 4 Km	Rs 10 per Km	N/A	Rs 1 per Min (Post 5 Min)

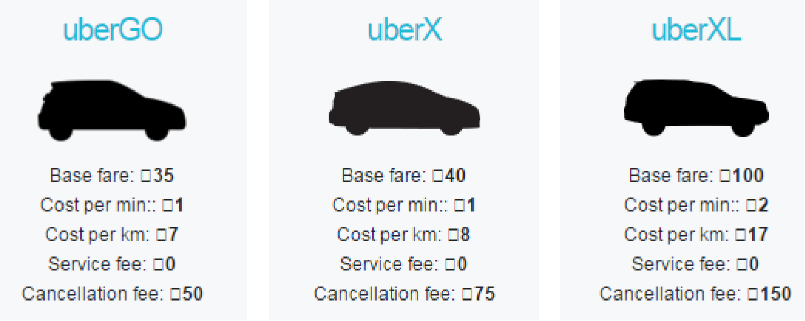
In addition to this, minimum fare of Micro is fixed at Rs. 50.

2) The second service, called B, also has three types of cabs - nano (equivalent to micro), tata indica (equivalent to mini), sedan (equivalent to prime), and here are the prices of the same.



CAR	FARE
Nano	□35 (□5.0/km, □ 1.5/min of trip time)
Tata Indica AC	□49 (□6.0/km after 2.0kms, □ 1.5/min of trip time)
Sedan	□75 (□8.0/km after 2.0kms, □ 1.5/min of trip time)

3) The third company, company C, has the following fare for the cabs offered. They do not have a micro or nano but on the other hand provide an XL.



uberGO

Base fare: □35
Cost per min:: □1
Cost per km: □7
Service fee: □0
Cancellation fee: □50

uberX

Base fare: □40
Cost per min:: □1
Cost per km: □8
Service fee: □0
Cancellation fee: □75

uberXL

Base fare: □100
Cost per min:: □2
Cost per km: □17
Service fee: □0
Cancellation fee: □150

You need to suggest the most optimised cab, for the following situations-

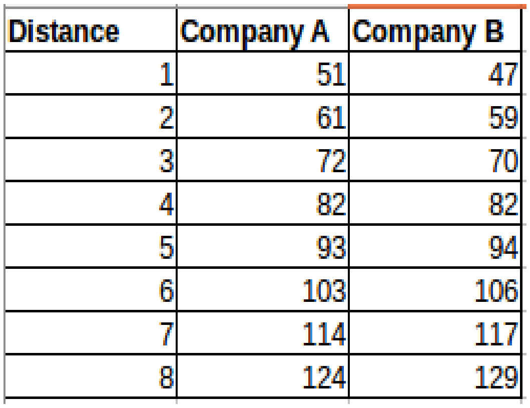
1. Which of the MICRO vehicles will be cheapest if your distance lies between 1 to 8 kms?
2. Which MINI vehicles are the cheapest if your distance is between 1 to 10 kms?

3. If you get a free upgrade from Company A - Micro to Mini, will it be cheaper than Company C, Mini for distance 2-6 kms?
4. Company C is charging a multiplier of 2.1 and Company A is charging a multiplier of 1.4 on their Sedan Vehicles (Company A Prime vs. Company C X). Which one will cost less ?
5. You have already booked UberGo for a multiplier of 1.5 and now you are getting a Company A Mini vehicle without peak charges? The challenge is that if you cancel an Company C, you will incur a cancellation charge penalty. However, if you choose to cancel, you stand a chance to save on peak charges. At what distance will you break even on the cancellation charges on Company C, in case you choose to go ahead with Company A?

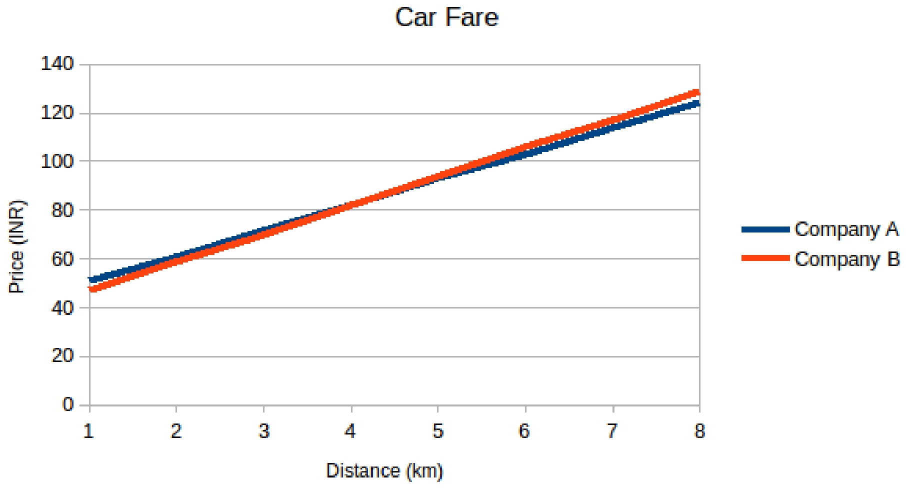
Answer 6

We shall take up the questions one by one.

1. Since only company A and B offer the micro cabs, we will choose a cab from one of the two. Since the distance is provided to be 1-8 km range, let's check out the price for this distance for both the company.

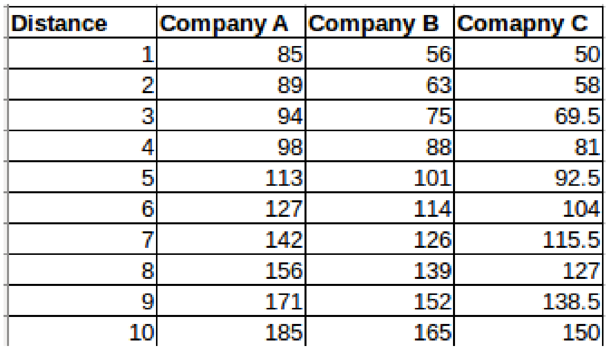


Distance	Company A	Company B
1	51	47
2	61	59
3	72	70
4	82	82
5	93	94
6	103	106
7	114	117
8	124	129

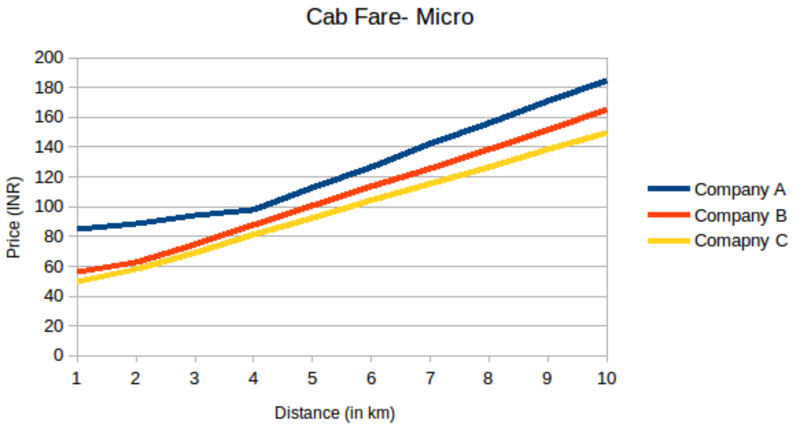


If the distance is less than 4, a cab from company A should be booked while for distance more than 4, cab from company B should be preferred.

2. Since all the companies have an option of a cab type mini, so we must check the cab fare for all the three companies.

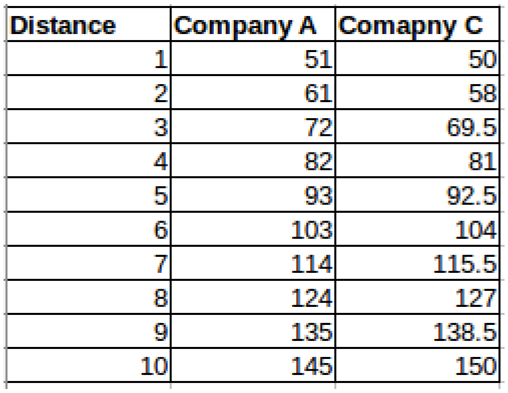


Distance	Company A	Company B	Comapny C
1	85	56	50
2	89	63	58
3	94	75	69.5
4	98	88	81
5	113	101	92.5
6	127	114	104
7	142	126	115.5
8	156	139	127
9	171	152	138.5
10	185	165	150

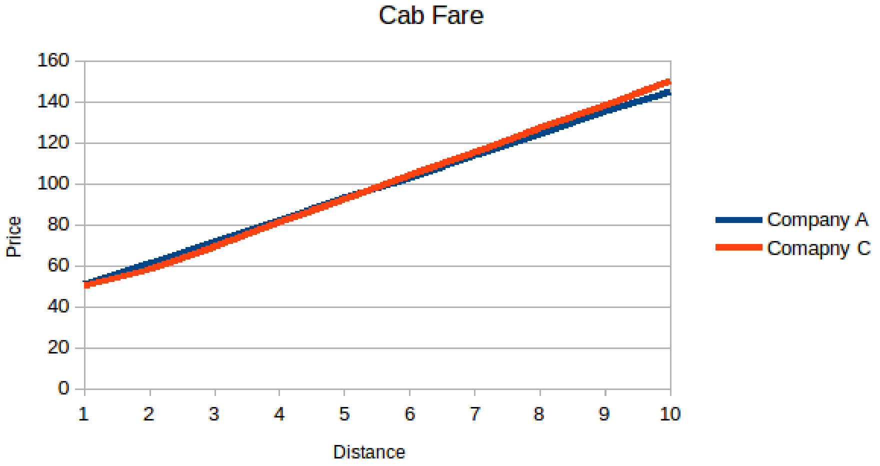


We clearly see that Company C is the cheapest car for the distance range and also seem to be the cheapest option for distances beyond 10 km looking at the trend.

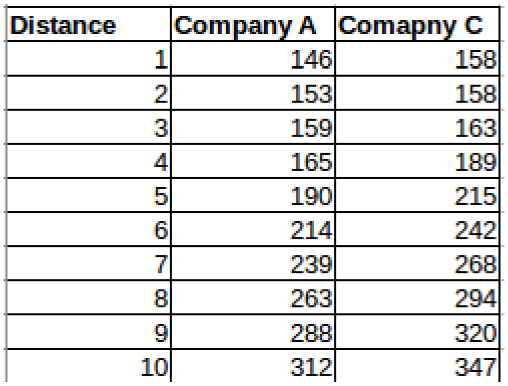
3. We will compare company A micro and company C mini. The trend looks to be very interesting. *Company C mini* starts at a slightly lower rate than *Company A Micro* but Company A takes over after 5 km distance. So, again our answer will be *cannot be determined* in the distance range. However, it is interesting to notice that a *Company A Micro* taxi comes out to be more expensive than *Company C Mini* car for shorter than 5 km distance.



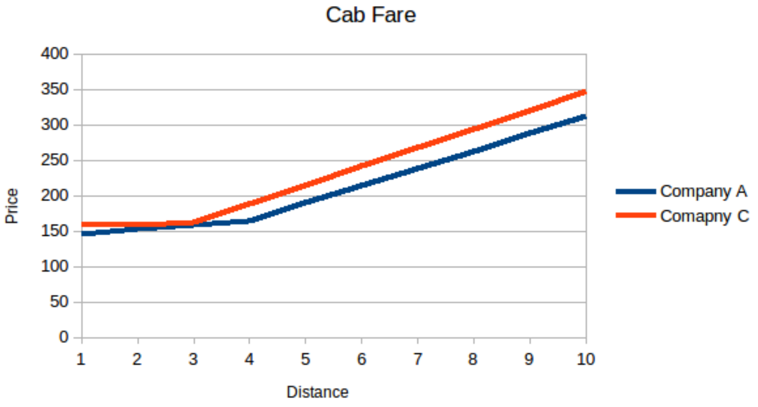
Distance	Company A	Comapny C
1	51	50
2	61	58
3	72	69.5
4	82	81
5	93	92.5
6	103	104
7	114	115.5
8	124	127
9	135	138.5
10	145	150



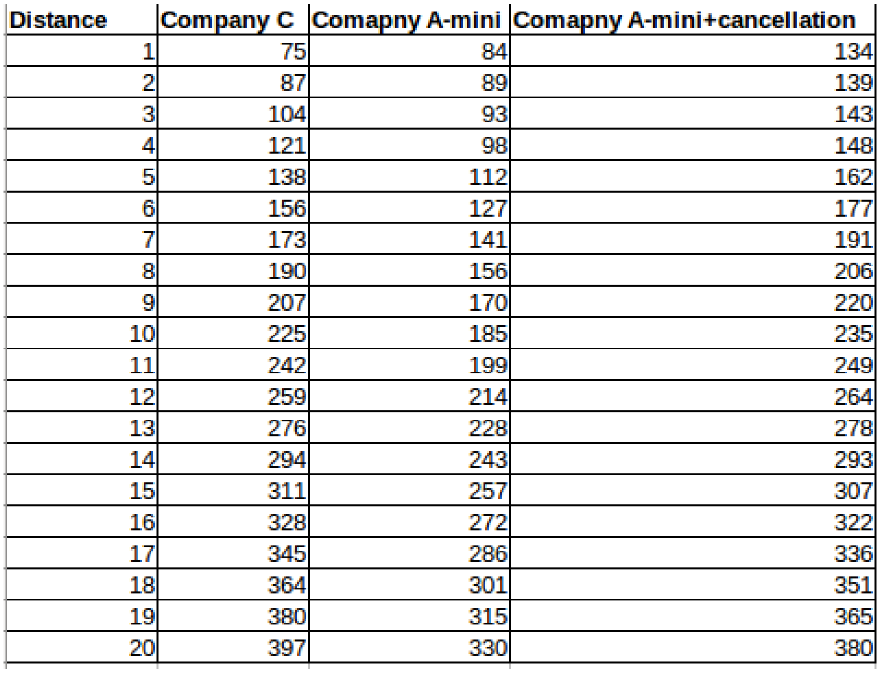
4. Multipliers are often added in peak traffic hours. It becomes very difficult to compare rates in such cases. Here’s one of those scenarios. Company A is generally more expensive compared to Company C, but in such extreme multiplier cases, we see that Company A comes out to be cheaper option throughout.



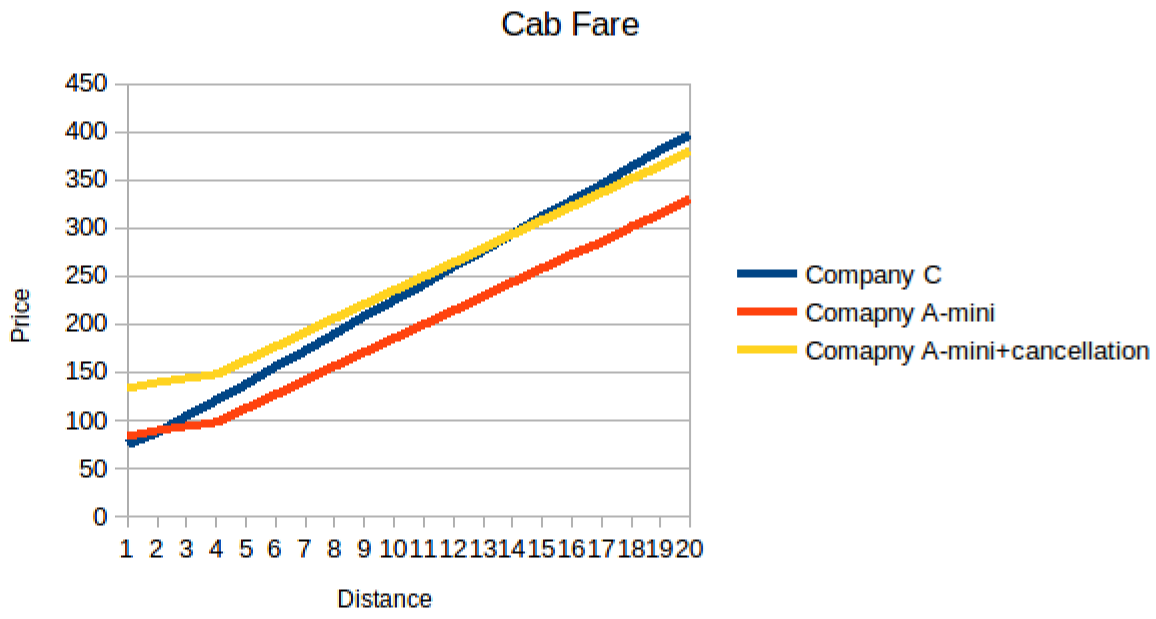
Distance	Company A	Comapny C
1	146	158
2	153	158
3	159	163
4	165	189
5	190	215
6	214	242
7	239	268
8	263	294
9	288	320
10	312	347



5. What you need to keep in mind is the cancellation charges of Company C. Here is the table for Company A vs. Company C :



Distance	Company C	Comapny A-mini	Comapny A-mini+cancellation
1	75	84	134
2	87	89	139
3	104	93	143
4	121	98	148
5	138	112	162
6	156	127	177
7	173	141	191
8	190	156	206
9	207	170	220
10	225	185	235
11	242	199	249
12	259	214	264
13	276	228	278
14	294	243	293
15	311	257	307
16	328	272	322
17	345	286	336
18	364	301	351
19	380	315	365
20	397	330	380



As it can be seen from both table and the graph, the break-even only happens between 13-14 kms. Hence, you should make a switch only if you want to travel more than 13 kms

Question 7

The two most commonly used data structures in python are lists and dictionaries. Can you list down the basic differences between list and dictionary.

Answer 7

Lists store values, in an ordered sequence. Each element is numbered, starting from zero, i.e. the first element in a list is numbered 0, the second 1, the third 2, and so on. We can remove values from the list, and add new values to the end.
Example: list1 = [1,2,3,4,10,100]

Dictionary is an unordered set of ***key: value*** pairs. Unlike lists, instead of using numbers to represent values, we have keys in a dictionary. The values in the dictionary can be removed/ modified and new values can be added.
Example: dict={‘English’: 90, ‘Mathematics’: 95, ‘Physics’:80}

Question 8

Suppose I have a list list1=['a', 'b', 1, {'Name': 'ABC', 'age': 22}], what will be the result of the following command?

```
>>list1[1]
>>list1[3]
>>list1[3]['age']
```

Answer 8

- The list has index starting from 0, so the output of the command list1[1]

will ba ‘b’.
- Since the fourth element in the list is a dictionary, the result would be {'Name': ‘ABC’, 'age': 22}
- List1[3]['age'] will give the output 22

Question 9

In order to add new elements to a list, we can use one of the two commands - list.append() or list.extend()? For a given list: *a=[1,2,3,4]*, what will be the result of the following commands?

```
>>a.append([6,7,8])
>>a.extend([6,7,8])
```

Answer 9

Append is used to add a single value at the end of the list. For the given example, append assumes [6,7,8], to be a list. So the result will be = [1,2,3,4,[6,7,8]]

Extend is used to add multiple elements to the list. In this case, the numbers 6,7 and are considered to be three different elements. The output of second command will be = [1,2,3,4,6,7,8]

Question 10

What will be the output of the following?

➢
```
for i in range(2.0):
   print(i)
```
Answer In python range function can only iterate using integer values. Since 2.0 is a floating point number thus the above code will result in an error.

➢
```
for i in range(2,10):
   print(i)
```
Answer The above code prints numbers from 2 to 9.

➢
```
for i in range(2,10,2):
   print(i)
```

Above code prints all the even numbers from 2 to 8(10 is exclusive), because the third parameter in range is the step-size for iteration.

Question 11

What is the difference between the following code lines?

```
>> df.drop([‘Age’], axis = 1)
>> df.drop([‘Age’], axis = 1, inplace=True)
```

Answer 11

The first case, since the value for parameter *inplace* is not specified, the default *inplace=False* is considered. It prints the dataframe without the column specified in the parameter, ‘Age’ in this case. But this does not make any change to the original dataframe.

When the parameter inplace is set to be true, the column is dropped from the original dataframe.

Question 12

What is the difference between file.tsv and file.csv? How are they used?

Answer 12

- .tsv and .csv are two different file formats referring to tab separated values and comma separated values respectively.
- .tsv are often used in text data as comma can represent a punctuation.
- .tsv and csv files can be loaded into a pandas dataframe using the read_csv function.

Question 13

What are lambda functions, when to use them and when not to use them.?

Answer 13

Lambda functions are anonymous functions and are not bound to any name/identifier. They are used when we can to perform a “small set of actions” and which does not require defining a separate function. They are lightweight and efficient. They are not recommended to use when the set of actions is not “small” or are too complex. Defining a separate function is better in this scenario.

Question 14

Which is more efficient, Iteration or recursion?

Answer 14

Python is not primarily a functional language and therefore there is a little overhead when using recursions, Therefore iterative methods are mostly efficient. But, there are problem where applying iteration can be tedious (Eg: Tower of Hanoi) and therefore can be efficiently solved using the recursion techniques.

Question 15

What do you mean by scope of a variable?

Answer 15

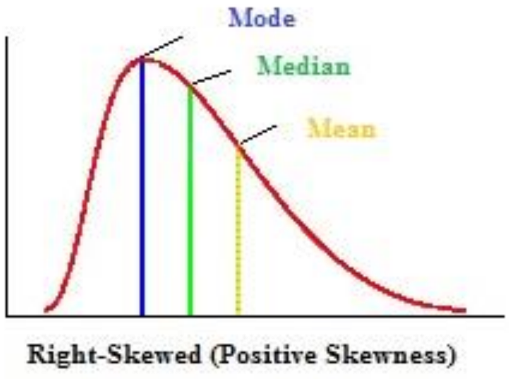
A scope of a variable refers to the environment to which it is restricted to, A variable of higher scope can used used at lower scope but reverse is not true.

Question 16

The mean of a distribution is 25, the median is 23, and the mode is 20. Based on the given information can you determine whether the plot for this distribution, will be positively skewed or negatively skewed?

Answer 16

Since the mode value is less that the mean and median, the data will be positively skewed. Here is a diagrammatic representation for the same.



Question 17

The interquartile range is the difference between the minimum and maximum value of the numbers in a given series. Is the statement correct?

Answer 17

No, the difference between the minimum and maximum value is the range of the series. The interquartile range is the difference between the the **first quartile** (value below which the lower 25% of the data exist) and the **third quartile** (value below which the lower 75% of the data exist).

Question 18

How many values, in the below series, fall within one standard deviation of the mean?
180, 313, 101, 255, 202, 198, 109, 183, 181, 113, 171, 165, 318, 145, 131, 145, 226, 113, 268, 108

Answer 18

On calculation, the mean and the standard deviation of the given data comes out to be 181.25 and 64.65 respectively. Now, one standard deviation above an below the mean would give the range 117 to 245. Total 11 numbers fall in this range.

Question 19

Two unbiased coins are tossed. What is the probability of getting at most one head?

Answer 19

The probability is ¾. Since the total possible outcomes when two coins are tossed will be 4 (HH, HT, TH, TT). We are to find at most 1 head, so the favorable outcomes are (TT, HT, TH). Therefore, Probability = ¾.

Question 20

Normal distribution is symmetric about the origin (0,0)?

Answer 20

No, the normal distribution is symmetric about the mean. But for a standard normal curve the values are scaled between -1 to1. In this case, the value of mean is 0 and standard deviation is 1. So the standard normal curve is symmetric about 0.

Question 21

What happens to the confidence interval when we introduce some outliers to the data?

Answer 21

The confidence interval depends on the standard deviation of the data. On introducing outliers in the data, the standard deviation increase, and so does the confidence interval.

Question 22

What does confidence interval = 95% mean?

Answer 22

On repetitive sampling, 95% of the times, the interval estimates will contain the population mean.

Question 23

What does the p-value for a statistical data signify?

Answer 23

The p value for a statistical test is used to draw conclusions, basically deciding to accept of reject the null hypothesis.The range of p-value is always between 0 and 1. Generally the threshold for p value is set to be 0.05. When the value is below 0.05, the null hypothesis is rejected. If the value is equal to or greater than the threshold, 0.05 in this case, the null hypothesis is not rejected.

Question 24

Consider the following two situations -

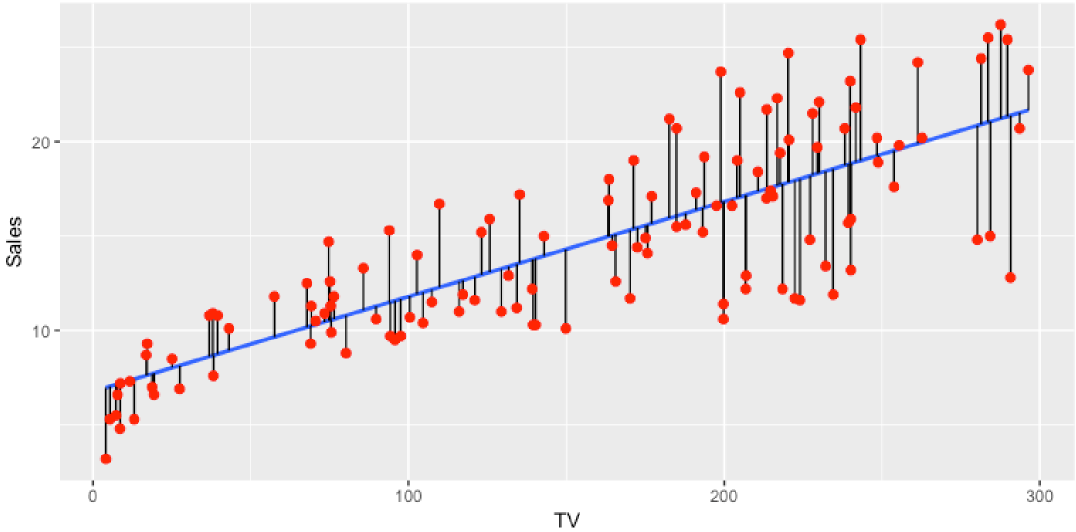
Situation 1: The doctor identifies a male patient as pregnant.

Situation 2: The doctor identifies a female pregnant lady as not pregnant.

Given that the null hypothesis is that the patient is not pregnant, classify the above two situations as Type I and Type II error.

Answer 24

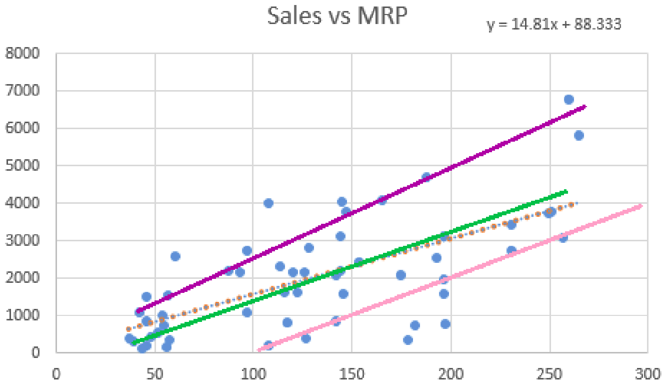
For the first situation, the null hypothesis is incorrectly rejected, which is a type II error. In the second situation, the null hypothesis is falsely accepted hence it is type I error.



Question 25

The linear regression model fits a line in order to model the relationship between independent and dependent variables. So how does the model decide the right line?

Answer 25



- As in the figure above, there could be multiple lines fit on the data points. The right line is decided such that the error value is minimum (as compared to other lines).
- Selecting the best fit line is equivalent to selecting the right set of coefficients for the line. This is an optimization problem.
- To select the best fit line, we use gradient descent optimization technique.

- ➢ Initially the coefficients for independent variables are randomly assigned and the error is calculated. These weights are then continuously updated to reduce the error.

(if you are not familiar with the concept of gradient descent algorithm, you must read this article: Introduction to Gradient Descent Algorithm)

Question 26

If you are given the two variables V1 and V2 such that
1. If V1 increases then V2 also increases
2. If V1 decreases then V2 behavior is unknown
What is the value of pearson’s correlation between V1 and V2?

Answer 26

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V1 as x and V2 as |x|. The correlation coefficient would not be close to 1 in such a case.

Question 27

Suppose we have a classification problem where the target variables have 4 classes - A, B, C, D. How can we use logistic regression for this problem statement?

Answer 27

The logistic regression can be used for multiclass classification by implementing the “one-versus-rest” technique where we take up each class and try to classify whether the data points belong to that particular class or not.

Suppose we have four classes A, B, C and D. We can start with classifying the data points as A vs B/C/D (not A), and then for class B vs not B, so on.

Model->	Mode l1	Model 2	Model 3	Model 4
	(A vs B ,C	(B vs A, C, D)	(C vs A, B, D)	(D vs A, B,

	,D) Calculating probability of outcome A	Calculating probability of outcome B	Calculating probability of outcome C	C) Calculating probability of outcome D

Question 28

Suppose I applied a logistic regression model on data and got training accuracy X and testing accuracy Y. Now I add 2 new features in data. Select option(s) which are correct in such case.

Note: Consider remaining parameters are same.

1. Training accuracy always decreases.
2. Training accuracy always increases or remain same.
3. Testing accuracy always decreases
4. Testing accuracy always increases or remain same

Answer 28

Training accuracy will increase but testing accuracy would increases only when the new variables are found to be significant.

Question 29

Suppose we have 2000 data points and 20 features. What is the maximum number of leaf nodes possible with a decision tree? Also, can you calculate the number of levels to which the tree will grow (depth of tree).

Answer 29

The maximum number of leaf nodes will be equal to the number of datapoints in the dataset (such that each leaf node has only one sample/datapoint). In this case the maximum number of leaf nodes will be 20,000.

Question 30

If we let the tree grow to the maximum length, there are chances of overfitting. How will you make sure that your decision tree model does not overfit?

Answer 30

If we let the tree grow to the maximum length, there are high chances that the model overfits on the training data. So we should reduce the size of the tree, which is called tree pruning. To do so, we can set certain conditions or constraints on the various parameters of decision tree model - such as max depth, min leaf sample size etc.

Consider the example of max depth, if we have a total of 500 data points, then the tree grows till each leaf node has only one sample. Instead, if we set a constraint on the parameter max_depth = 3, the tree would only grow for three levels.

Question 31

A random forest model is an ensemble of multiple decision trees. Consider the following two scenarios -

Case 1: Build a random forest with 10 estimator (i.e. 10 decision trees).
Case 2: Creating 10 decision trees and averaging the predictions from these trees to calculate the final predictions.

Will both the models give the same results? (considering the dataset and other parameters are same in both the cases)

Answer 31

When training a decision tree classifier, splits are done using all of the data points and all of the features whereas in case of random forests every estimator(decision tree) is based only on a random subset of the dataset(thus the name random forest).

Hence, in the case of multiple decision trees trained on the same dataset , all the trees will be same along with their predictions for a given data point whereas because of change in distribution of data for different decision trees in case of random forests, the predictions can be different for the same data point from different estimators.Thus Case 1 and Case 2 will have different results.

Question 32

Two most important ensemble learning methods are bagging and boosting. Can you explain the difference between bagging and boosting?

Answer 32

Bagging (or Bootstrap Aggregating): In this technique, multiple subsets of the complete data are created and a model is trained on each of these subsets. The final prediction is made by combining the values predicted from each of these models trained on different subsets. Various algorithms which use bagging technique are Bagging estimator, Random Forest, Extra trees.

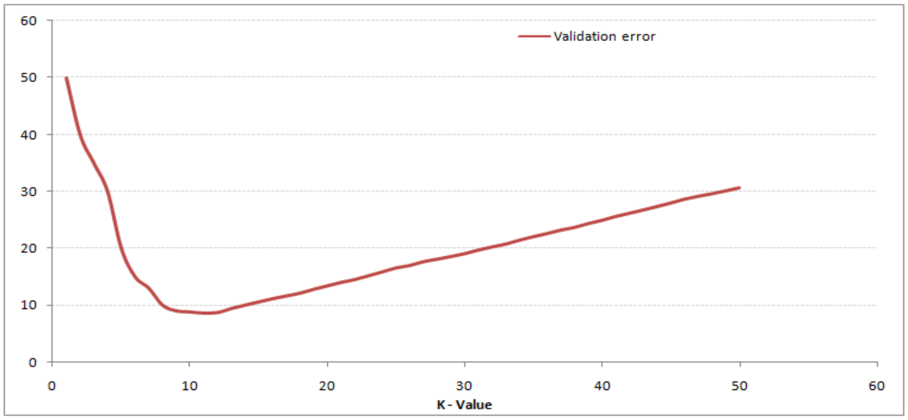
Boosting: It is a sequential process where each subsequent model attempts to correct the errors from the previous model. This is done by giving higher weights to the observations which were incorrectly predicted. Final model (strong learner) is the weighted mean of all the models (weak learners). AdaBoost GBM, XGBoost, etc are some of the algorithms which use boosting technique.

Question 33

What does the K in KNN mean? How do you find the optimal value for K?

Answer 33

K is a hyperparameter for the KNN algorithm which decides the number of training data points that are used for predicting the label for the data point during the testing phase. For finding the optimal value of k , we can try all possible values of K from 3 upto a specific value and monitor the loss for the validation data. Another possible solution is the elbow technique where we plot the total sum of squared error for different values of k and find the point where the total sum of squares starts increasing.



Here k=8 seems like the optimal value for our dataset.

Question 34

While doing K means algorithm, consider the following two cases-
Case1: The cluster centers are updated after every datapoint assignment
Case2: The cluster centers are updated after assigning every data point to some cluster.

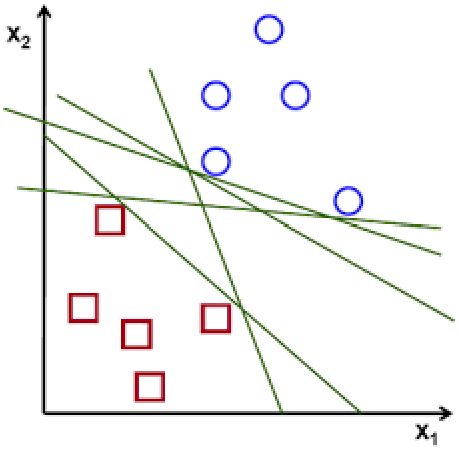
We have applied k means on our dataset to calculate the cluster centers, now one feature of the dataset was scaled by some factor M, do the cluster centers change? What if all the features were scaled?

Answer 34

In case only one feature was scaled , the cluster centers will need to be updated as it will increase/decrease the value of single feature that will lead to the change of cluster centers thus there might be some points whose clusters will get changed.But in case all the features were scaled , the cluster centers will also get scaled by the same factor but in this case the clusters will not change.

Question 35

Suppose we were given a dataset that consists of all numerical features, we applied linear regression but the results were not satisfactory. Further we also used SVM but still the performance did not improve , what seems to be the problem? Find out the best decision boundary from the ones given below



Answer 35

If both linear regression and SVM are not giving satisfactory results on a dataset, this implies that the classes are not linearly separable. In such a situation we can use the kernel trick of SVMs to project the dataset into higher number of dimensions such that the classes become linearly separable.

The best decision boundary is the one that maximizes the margin on both sides.(SVM theory)

Question 36

Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?

Answer 36

The major difference between random forest and gradient boosted trees lies in the fact that the individual estimators in random forests are independent of each other whereas in case of gradient boosted trees the individual estimators are not independent. The estimators in random forests are built parallely whereas in GBTs the estimators are built incrementally. Also the data points for which the current estimator was not able to perform well are given more preference in the next estimators so that the whole model is able to generalize well over the complete dataset.

Question 37

Which of the following are mandatory data pre-processing step for XGB?

1. Impute Missing Values
2. Remove Outliers
3. Convert data to numeric array / sparse matrix
4. Input variable must have normal distribution
5. Select the sample of records for each tree/ estimators

Answer 37

Converting data to numeric array/sparse matrix is a mandatory pre-processing step for XGBOOST as xgboost cannot handle categorical features.

Question 38

The features of our dataset show strong correlation , is Naive Bayes a good choice for such a dataset? How will you modify the dataset such that Naive Bayes shows good results?

Answer 38

Strong correlation between features implies dependence. Thus Naive Bayes is not a good choice for such a dataset because Naive Bayes assumes that the features of the given dataset are independent of each other. In such a case we can apply PCA to identify the independent dimensions from the dataset and then use Naive Bayes.

Question 39

The categorical variables can further be classified as cardinal and ordinal. Can you explain the difference between them and how can we deal with these variables?

Answer 39

The variables which have a certain number of categories fall under the term 'Cardinal variables'. These variables do not have an order between them. For example, the variable 'Gender' can have two categories - Male and Female, and there is no order to these categories. Most machine learning models cannot deal with categorical variables. So we use one hot encoding technique to convert these categorical variables into binary.

While Ordinal variables have categories which contain an order between them, such as 'class of a student' which can be I, II,....IX, X, and so on. We know that IV and V are higher than I and II. So we can say that these categories have an order. Although one hot encoding technique can be applied on these variables, it is preferred to use label encoding (since the order in the variables is important)

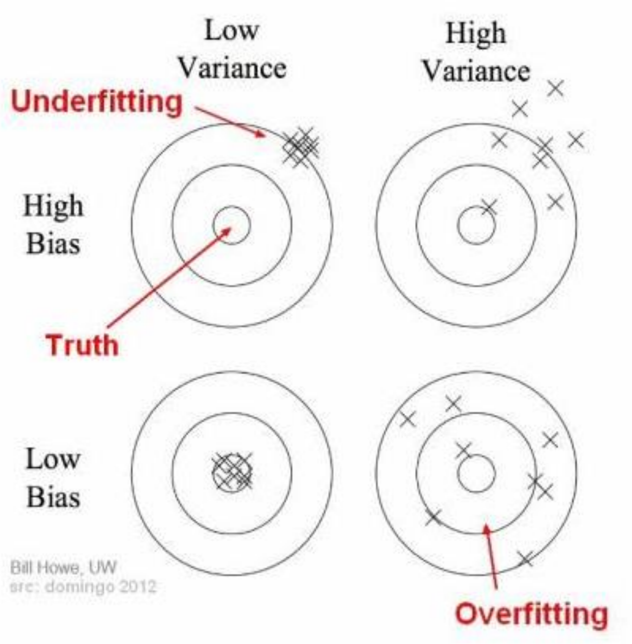
Question 40

Have built two models "Model 1" and "Model 2". "Model 1" has high bias and low variance where as "Model 2" has low bias and high variance.

Which of these two models have higher chance of overfitting or under-fitting?

Answer 40

A model with high bias and low variance tends to overfit on certain feature or data points. The model will not be generalized and performs poorly, thus has a higher chance of underfitting. On the other hand, a model with high variance and low bias will have a higher chance of overfitting. A simple diagram (shown below) can be used to verify the above statements.



Question 41

There are multiple machine learning algorithms, such as linear regression, logistic regression, random forest, xgb etc. How will you select which model to implement?

Answer 41

The model to be used is often decided based on the dataset we need to work on.

- ➢ For a regression problem, we can choose linear regression, random forest regressor etc.

- ➢ Similarly for classification problem, the choice of models would be slightly different.
- ➢ For imbalanced dataset, boosting techniques work really well.
- ➢ In case the dataset has a very high number of categorical variables, we can use CatBoost.
- ➢ For very large datasets, light GBM has proved to work better than other algorithms.

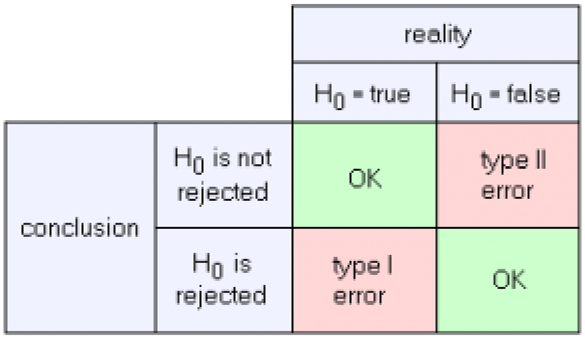
Question 42

What do you understand by Type I vs Type II error ?

Answer 42

Type I error is committed when the null hypothesis is true and we reject it, also known as a ‘False Positive’. Type II error is committed when the null hypothesis is false and we accept it, also known as ‘False Negative’.

In the context of confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).

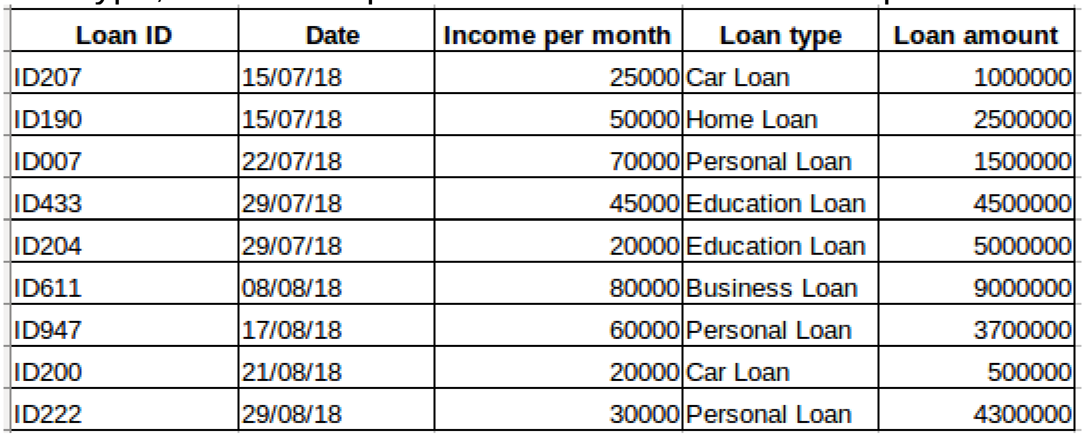


		reality	
		H_0 = true	H_0 = false
conclusion	H_0 is not rejected	OK	type II error
	H_0 is rejected	type I error	OK

Question 43

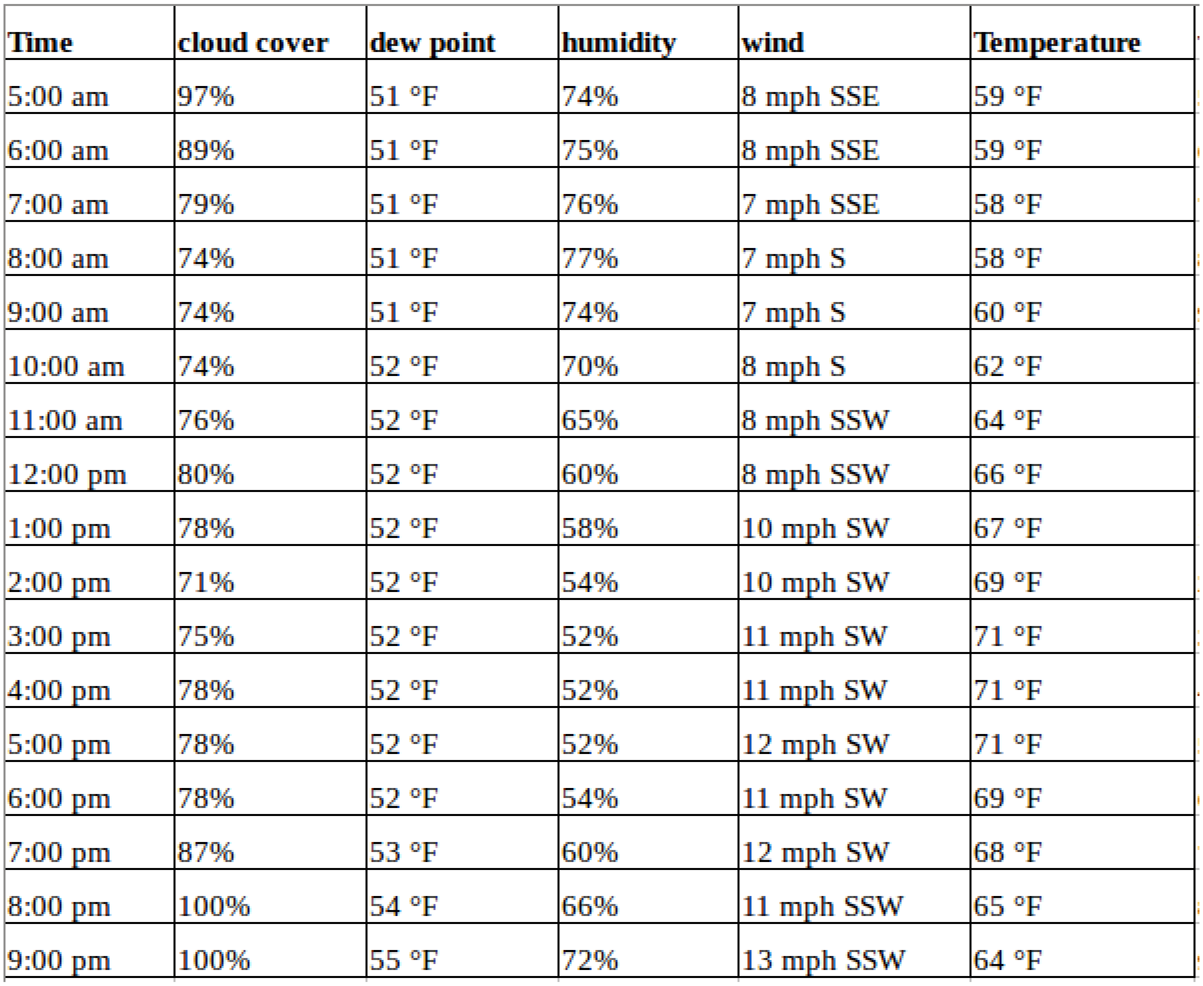
How can you differentiate or classify a problem statement as time series problem? For instance, consider the following two problem statements; which one would you apply a time series model on?

Problem 1- Based on - the income per month, date of issuing loan, and the loan type, we have to predict the loan amount for the person.



Loan ID	Date	Income per month	Loan type	Loan amount
ID207	15/07/18	25000	Car Loan	1000000
ID190	15/07/18	50000	Home Loan	2500000
ID007	22/07/18	70000	Personal Loan	1500000
ID433	29/07/18	45000	Education Loan	4500000
ID204	29/07/18	20000	Education Loan	5000000
ID611	08/08/18	80000	Business Loan	9000000
ID947	17/08/18	60000	Personal Loan	3700000
ID200	21/08/18	20000	Car Loan	500000
ID222	29/08/18	30000	Personal Loan	4300000

Problem 2- The target is to predict the hourly temperature using the for a particular region given the past data for the same.



Time	cloud cover	dew point	humidity	wind	Temperature
5:00 am	97%	51 °F	74%	8 mph SSE	59 °F
6:00 am	89%	51 °F	75%	8 mph SSE	59 °F
7:00 am	79%	51 °F	76%	7 mph SSE	58 °F
8:00 am	74%	51 °F	77%	7 mph S	58 °F
9:00 am	74%	51 °F	74%	7 mph S	60 °F
10:00 am	74%	52 °F	70%	8 mph S	62 °F
11:00 am	76%	52 °F	65%	8 mph SSW	64 °F
12:00 pm	80%	52 °F	60%	8 mph SSW	66 °F
1:00 pm	78%	52 °F	58%	10 mph SW	67 °F
2:00 pm	71%	52 °F	54%	10 mph SW	69 °F
3:00 pm	75%	52 °F	52%	11 mph SW	71 °F
4:00 pm	78%	52 °F	52%	11 mph SW	71 °F
5:00 pm	78%	52 °F	52%	12 mph SW	71 °F
6:00 pm	78%	52 °F	54%	11 mph SW	69 °F
7:00 pm	87%	53 °F	60%	12 mph SW	68 °F
8:00 pm	100%	54 °F	66%	11 mph SSW	65 °F
9:00 pm	100%	55 °F	72%	13 mph SSW	64 °F

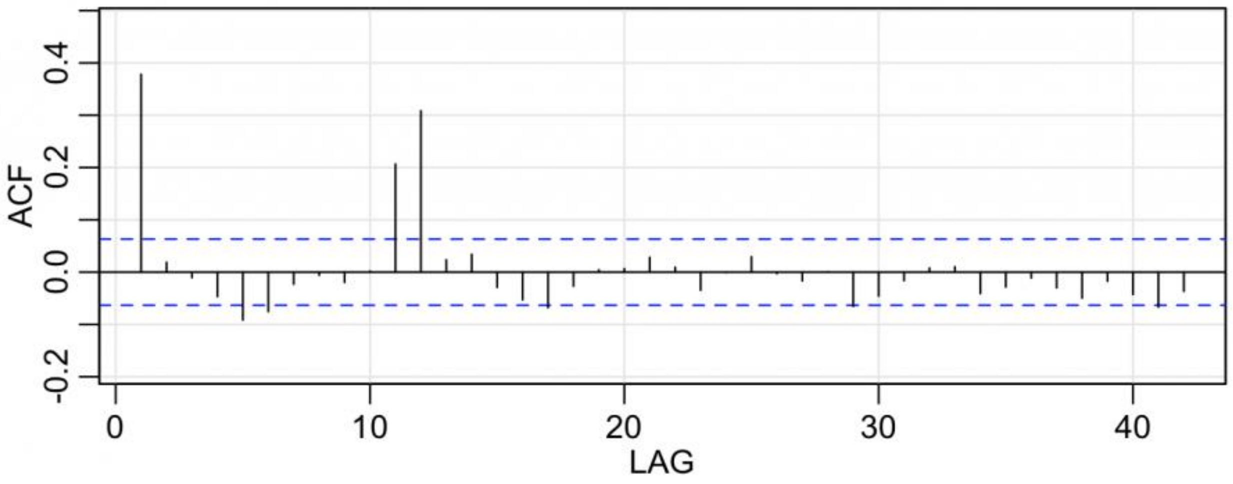
Answer 43

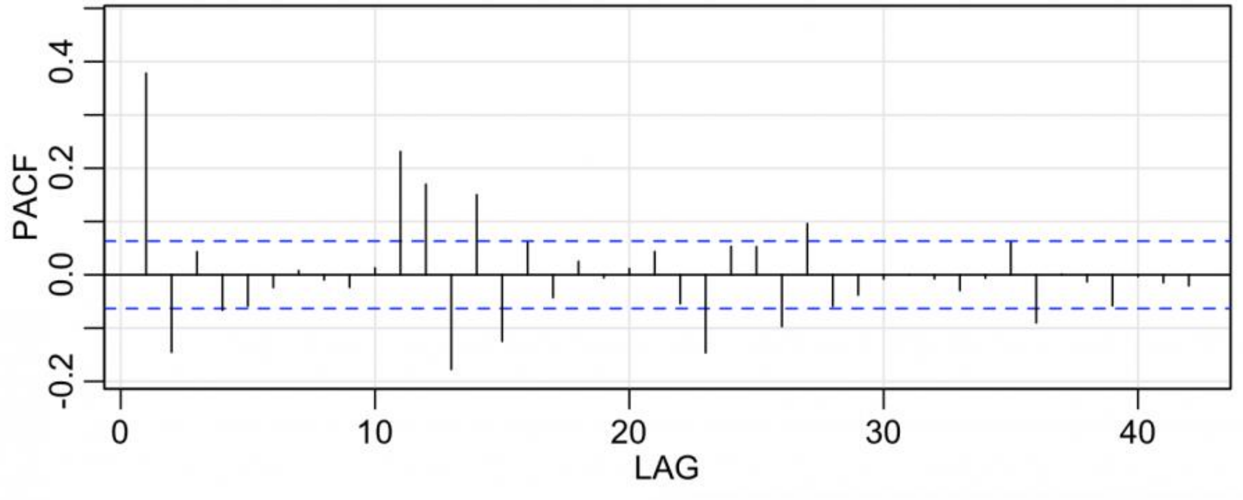
The data points in a time series dataset are equally spaced and the target variable for each row is dependent on the past dependent and independent variables.

From the given two problems, the second problem represents a time series since the next hour's temperature depends on the previous value. While in the first problem, the loan amount for an individual person does not depend on the loan taken by the previous person.

Question 44

Identify the p, q values from the following acf pacf plots-





Answer 44

The value of p and q are taken from as the point where the dotted line is intersected the first time. ACF plot is used to determine the value of q while PACF plot is used to find the parameter p.

In the ACF plot shown above, the first intersection is at the value one, so the value of q must be taken as 1. Similarly, the value of p is also 1.

Question 45

Suppose we have the following architecture for a Multi-Layer Perceptron a. Number of nodes at input layer = 10 b. Nodes at hidden layer = 5 c. Number of output node= 1 What are the total number of connections?

Answer 45

Each node in the input layer has a connection to each node in the hidden layer. So this makes the total number of connections between input and hidden = 50. Then we will have 5 connections between the hidden and output layer. So total number of connections in the above architecture is 55.

Question 46

There are many different activation functions like relu, sigmoid, softmax or tanh. How can one decide which activation function to use? How will you decide the activation functions for input and output layer?

Answer 46

Sigmoid is an activation function that returns a value between 0 and 1, it is primarily used for the output node of the Binary Classification, but can also be used in the hidden layers.

Tanh is an activation function that returns a value between -1 and 1, it is primarily used in hidden layers of Neural Networks.

Relu is an activation function that returns a linear value when the input is positive and returns 0 when the input is 0 or negative. It is primarily used in hidden layers of Neural Networks.

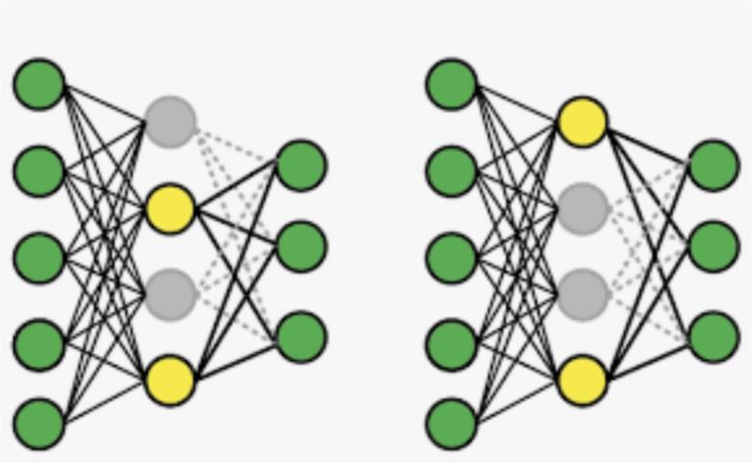
Softmax is an activation function which is primarily used in the last/output layer of multi-class classification problem.

Question 47

Dropout and DropConnect are both regularization techniques for Neural Network. Is there a difference between these two? How is setting dropout =0.3 different from drop connect =0.3?

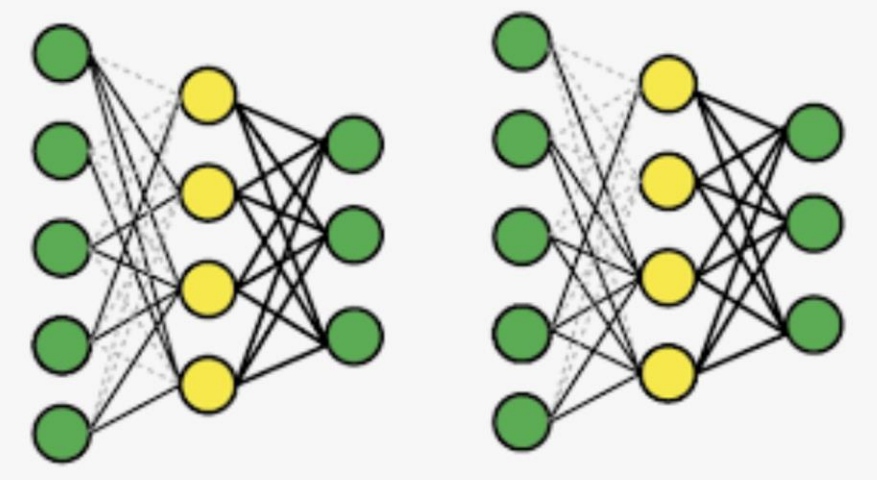
Answer 47

Drop-Out



The function drop-out in a layer assigns a probability p to every node in that layer such that, that node will not be included in the computation during the runtime with respect to probability p. (0.3 in question)

Drop-Connect



The function drop-connection in a layer is a probability p for every node in that layer such that, there is a chance p such that , that node will skip a connection to consecutive layer by the probability p.

Question 48

During feature selection - we remove features thus giving less information to the model for training. Will this affect performance of the model?

Answer 48

Feature selection is selecting the most relevant attributes for the predictive model. Only the redundant features are removed from the model so that it does not adversely affect the accuracy of the model. Having fewer but relevant features reduces the model complexity and training time.

Question 49

There are broadly two common methods to convert categorical variable to number, Label Encoding and One Hot Encoding. How do you decide which method to use when?

Answer 49

Most machine learning models cannot deal with categorical variables and hence we need to perform one hot encoding or label encoding before feeding the data to the model.

The variables which have a certain number of categories fall under the term ‘Categorical variables’. These variables do not have an order between them. For example, the variable ‘Gender’ can have two categories - Male and Female, and there is no order to these categories. In this case, we can use the one hot encoding technique to convert the categorical columns into two columns with binary values.

Question 50

You came to know that your model is suffering from low bias and high variance. Which
algorithm should you use to tackle it? Why?

Answer 50

Low bias occurs when the model's predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results. In such situations, we can use bagging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling.

About Analytics Vidhya

Analytics Vidhya is the World's Leading Data Science Community & Knowledge Portal. The mission is to create the next-gen data science ecosystem! This platform allows people to learn & advance their skills through various training programs, know more about data science from its articles, Q&A forum, and learning paths. Also, we help professionals & amateurs to sharpen their skillsets by providing a platform to participate in Hackathons. Our viewers remain updated with the latest happenings around the world of analytics using our monthly newsletters. Stay in touch with us to be a perfect and informative data practitioner.

Our Other Platforms

Courses: https://courses.analyticsvidhya.com/

Blog: https://www.analyticsvidhya.com/blog/

DataHack: https://datahack.analyticsvidhya.com/contest/all/

Jobs: https://jobsnew.analyticsvidhya.com/jobs/all

Bootcamp: https://www.analyticsvidhya.com/data-science-immersive-bootcamp/

Initiate AI: https://initiateai.analyticsvidhya.com/

Discuss: https://discuss.analyticsvidhya.com/

Analytics Vidhya