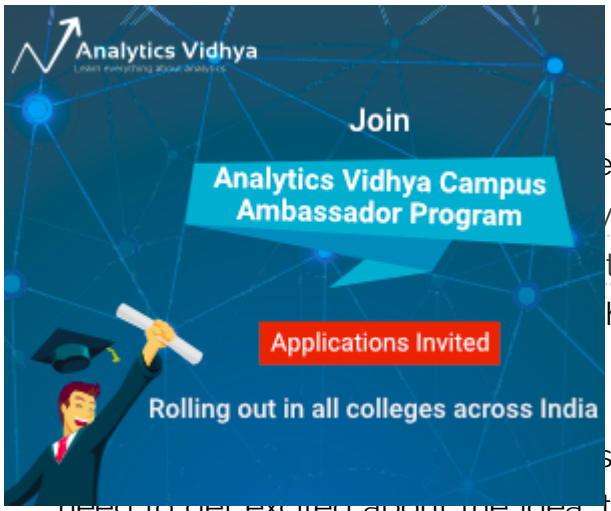
40 Interview Questions asked at Startups in Machine Learning / Data Science

achine-learning-data-science/) g+ (https://plus.google.com/share?url=https:
csvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-
(http://pinterest.com/pin/create/button/?url=https://www.analyticsvidhya.com/blog/2016/09/40-
stions-asked-at-startups-in-machine-learning-data-science/&media=https:
csvidhya.com/wp-content/uploads/2016/09/Interview-questions.png&
%20in%20Machine%20Learning%20

on can make you think
TTANCE!



are being looked as the
on happening in the world
are numerous exciting
ya.com/blog/2016/09/18-
tor-summer-
hat could be a better start

s is not easy. You obviously need to get excited about the idea, team and the vision of the company. You might also find some real difficult techincal questions on your way. The set of questions asked depend on what does the startup do. Do they provide consulting? Do they build ML products ? You should always find this out prior to beginning your interview preparation.

To help you prepare for your next interview, I've prepared a list of 40 plausible & tricky questions which are likely to come across your way in interviews. If you can answer and understand these question, rest assured, you will give a tough fight in your job interview.

POPULAR POSTS

- 25 Open Datasets for Deep Learning Every Data Scientist Must Work With (https://www.analyticsvidh /blog/2018 /03/comprehensive-collection-deep-learning-datasets/)
- A Complete Tutorial to Learn Data Science with Python from Scratch (https://www.analyticsvidh /blog/2016 /01/complete-tutorial-learn-data-science-python-scratch-2/)
- Essentials of Machine Learning Algorithms (with Python and R Codes) (https://www.analyticsvidh /blog/2017 /09/common-machine-learning-algorithms/)

*Note: A key to answer these questions is to have concrete **practical** understanding on ML and related statistical concepts.*

Answer: Processing a high dimensional data on a limited memory machine is a strenuous task, your interviewer would be fully aware of that. Following are the methods you can use to tackle such situation:

1. Since we have lower RAM, we should close all other applications in our machine, including the web browser, so that most of the memory can be put to use.
2. We can randomly sample the data set. This means, we can create a smaller data set, let's say, having 1000 variables and 300000 rows and do the computations.
3. To reduce dimensionality, we can separate the numerical and

7 Types of Regression Techniques you should know! (https://www.analytics vidhya.com/blog/2015/08/comprehensive-guide-regression/)
Understanding Support Vector Machine algorithm from examples (along with code) (https://www.analytics vidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)
6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R) (https://www.analytics vidhya.com/blog/2017/09/naive-bayes-explained/)
A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python) (https://www.analytics vidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/)
A Complete Tutorial on Time Series Modeling in R (https://www.analytics vidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/)

categorical variables and remove the correlated variables. For numerical variables, we'll use correlation. For categorical variables, we'll use chi-square test.

4. Also, we can use PCA (https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component- the components which in the data set.

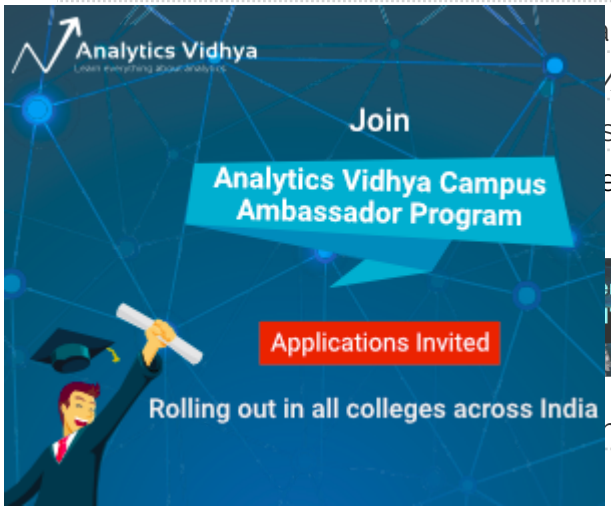
ms like Vowpal Wabbit ption.

chastic Gradient Descent is

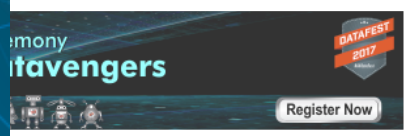
understanding to estimate the response variable. But, to identify useful predictors nformation.

Note: For point 4 & 5 make sure you read about online learning algorithms (https://www.analyticsvidhya.com/blog/2015 arning-simplified-2/) & //www.quora.com/Whats- scent-and-stochastic- ed methods.

(https://www.analyticsvidhya.com/wp-content/uploads/2016/09/banner_5years.png)



(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=blog_sidebottom)

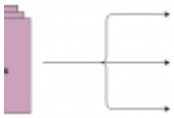


n/contest/datafest-

Q2. Is rotation necessary in PCA? If yes, Why? What will happen if you don't rotate the components?

Answer: Yes, rotation (orthogonal) is necessary because it maximizes the difference between variance captured by the component. This makes the components easier to interpret. Not to forget, that's the motive of doing PCA where, we aim to select fewer components (than features) which can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn't change, it only changes the actual coordinates of the points.

RECENT POSTS



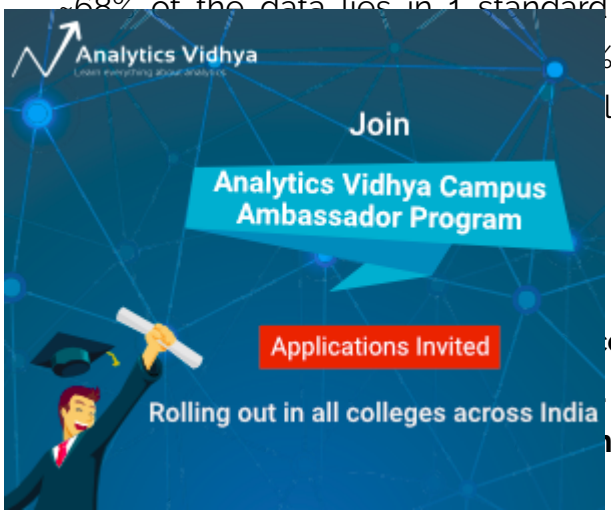
(https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/)

A Comprehensive Guide to Understand and Implement Text Classification in Python (https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/)

SHIVAM BANSAL , A...

(https://www.analyticsvidhya.com/blog/2018/04/avbytes-ai-ml-developments-this-week-230418/)

AVBytes: AI & ML Developments



If we don't rotate the components, the effect of PCA will diminish and we'll have to select more number of components to explain variance in the data set.

ticsvidhya.com/blog/2016
onent-analysis-python/)

ta set has missing values
viation from the median,
main unaffected? Why?

h hints for you to start
cross median, let's assume
it's a normal distribution. We know, in a normal distribution,
~68% of the data lies in 1 standard deviation from mean (or
% of the data unaffected.
ld remain unaffected by

er detection. You've build
an accuracy of 96%. Why
odel performance? What

Answer: If you have worked on enough data sets, you should deduce that cancer detection results in imbalanced data. In an imbalanced data set, accuracy should not be used as a measure of performance because 96% (as given) might only be predicting majority class correctly, but our class of interest is minority class (4%) which is the people who actually got diagnosed with cancer. Hence, in order to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine class wise performance of the classifier. If the minority class performance is found to be poor, we can undertake the

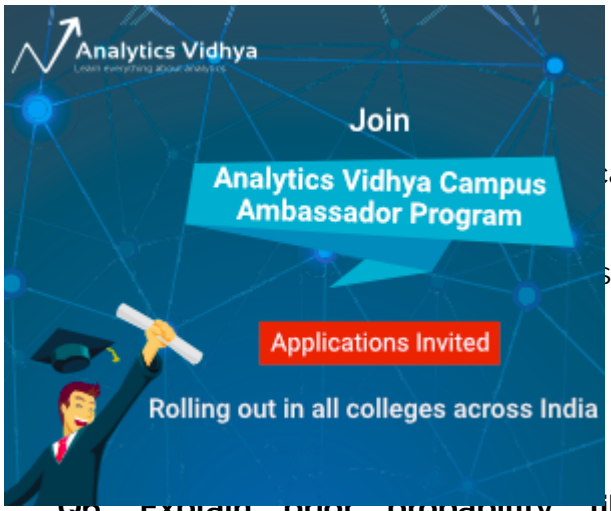
this week – Pandas to end Python 2 Support, Intel's Framework-Neutral Library, Google's Cancer Detection Algo, etc. (https://www.analyticsvidhya.com/blog/2018/04/avbytes-ai-ml-developments-this-week-230418/)
PRANAV DAR , APRI...

(https://www.analyticsvidhya.com/blog/2018/04/5-years-building-analytics-vidhya-journey-learnings/)
5 years of building Analytics Vidhya – our journey and learnings (https://www.analyticsvidhya.com/blog/2018/04/5-years-building-analytics-vidhya-journey-learnings/)
KUNAL JAIN , APRIL ...

An Overview of Regularization

following steps:

1. We can use undersampling, oversampling or SMOTE to make the data balanced.



Q8. Explain prior probability, likelihood and marginal likelihood in context of naiveBayes algorithm?

Answer: Prior probability is nothing but, the proportion of dependent (binary) variable in the data set. It is the closest guess you can make about a class, without any further information. For example: In a data set, the dependent variable is binary (1 and 0). The proportion of 1 (spam) is 70% and 0 (not spam) is 30%. Hence, we can estimate that there are 70% chances that any new email would be classified as spam.

Likelihood is the probability of classifying a given observation as 1 in presence of some other variable. For example: The

probability that the word 'FREE' is used in previous spam messages. Likelihood is, the probability that the word 'FREE' is used in any message.

es data set. You manager acy model. You start with you know it works fairly ter, you tried a time higher accuracy than n? Why?

to posses linearity. On the other hand, a decision tree algorithm is known to work best to detect non – linear interactions. The reason why decision tree ecause it couldn't map the a regression model did. ar regression model can ata set satisfies its linearity ticsvidhya.com/blog/2016 umptions-plots-

which involves helping a food delivery company save more money. The problem is, company's delivery team aren't able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?

Answer: You might have started hopping through the list of ML algorithms in your mind. But, wait! Such questions are asked to test your machine learning fundamentals.

This is not a machine learning problem. This is a route optimization problem. A machine learning problem consist of

three things:

1. There exist a pattern.
2. You cannot solve it mathematically (even by writing

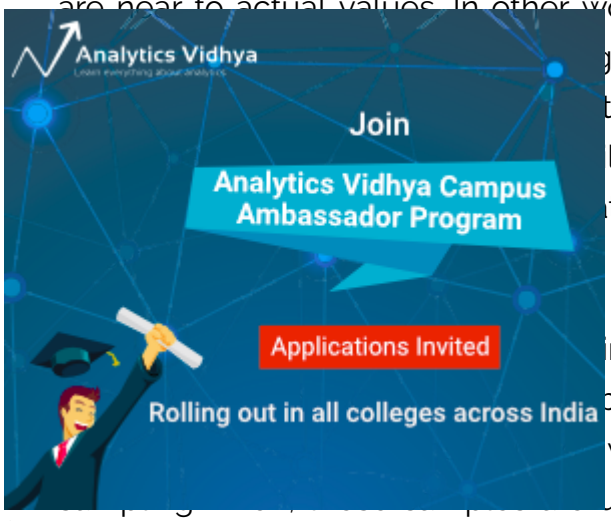
rs to decide if machine

problem.

model is suffering from

algorithm should you use

Answer: Low bias occurs when the model's predicted values are near to actual values. In other words, the model becomes g data distribution. While it t not to forget, a flexible lities. It means, when this ta, it gives disappointing



ing algorithm (like random blem. Bagging algorithms with repeated randomized used to generate a set of models using a single learning algorithm. Later, the model predictions are combined using voting (classification) or averaging (regression).

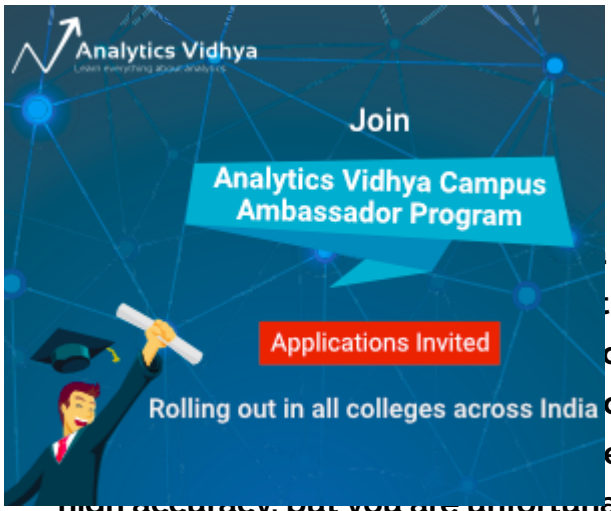
Also, to combat high variance, we can:

1. Use regularization technique, where higher model coefficients get penalized, hence lowering model complexity.
2. Use top n features from variable importance chart. May be, with all the variable in the data set, the algorithm is having difficulty in finding the meaningful signal.

Q10. You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?

empted to say No, but that
related variables have a
in presence of correlated
y a particular component

a data set, of which 2 are
ata set, the first principal
e variance than it would
exhibit with uncorrelated variables. Also, adding correlated variables lets PCA put more importance on those variable,

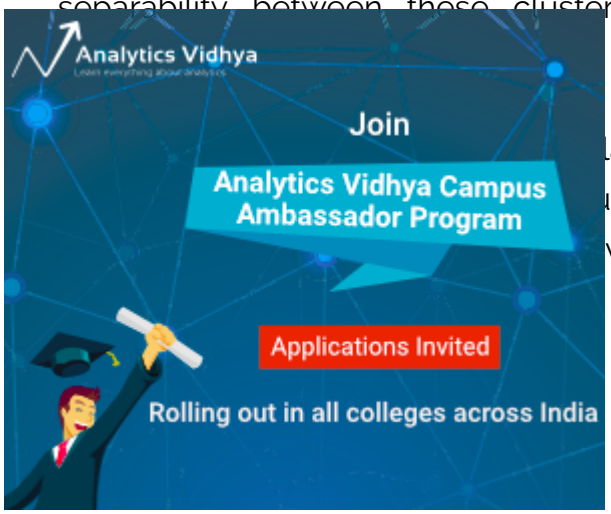


you are now anxious to result, you build 5 GBM hm would do the magic. ould perform better than cided to combine those els are known to return ate. Where did you miss?

Answer: As we know, ensemble learners are based on the idea of combining weak learners to create strong learners. But, these learners provide superior result when the combined models are uncorrelated. Since, we have used 5 GBM models and got no accuracy improvement, suggests that the models are correlated. The problem with correlated models is, all the models provide same information.

For example: If model 1 has classified User1122 as 1, there are high chances model 2 and model 3 would have done the same, even if its actual value is 0. Therefore, ensemble learners are

built on the premise of combining weak uncorrelated models to obtain better predictions.



Q13. How is True Positive Rate and Recall related? Write the equation.

Answer: True Positive Rate = Recall. Yes, they are equal having the formula (TP/TP + FN).

Know more: Evaluation Metrics (https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/)

Q14. You have built a multiple regression model. Your model

R^2 isn't as good as you wanted. For improvement, your remove the intercept term, your model R^2 becomes 0.8 from 0.3. Is it possible? How?

...need to understand the ...m in a regression ...del prediction without any ...ction. The formula of R^2 = 1 ... predicted value.

... of intercept term (ymean), ... equation's value becomes smaller than actual, resulting in higher R^2.

...ur manager has informed ...ing from multicollinearity. ...ue? Without losing any ...ter model?

...e can create a correlation ...s having correlation above ...e). In addition, we can use calculate VIF (variance inflation factor) to check the presence of multicollinearity. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Also, we can use tolerance as an indicator of multicollinearity.

But, removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression. Also, we can add some random noise in correlated variable so that the variables become different from each other. But, adding noise might affect the prediction accuracy, hence this approach should be carefully used.

Know more: Regression (https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-regression-analysis-assumptions-plots-solutions/)

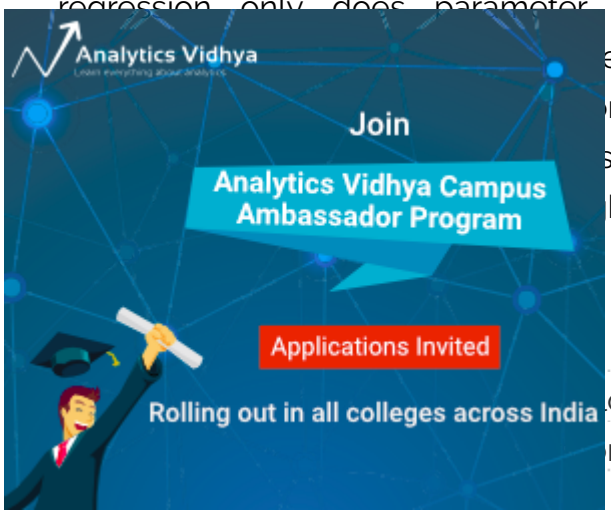
(https://www.analyticsvidhya.com/?utm_source=banner5yearanniversary)

able over Lasso

ors Hastie, Tibshirani who
bles with greater magnitude
resence of many variables
ridge regression.

egression (L1) does both
variable selection and parameter shrinkage, whereas Ridge
regression only does parameter shrinkage and end up
e model. In presence of
n might be the preferred
s best in situations where
her variance. Therefore, it

Lasso Regression
og/2016/01/complete-
n/)



(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

Q17. Rise in global average temperature led to decrease in number of pirates around the world. Does that mean that decrease in number of pirates caused the climate change?

Answer: After reading this question, you should have understood that this is a classic case of "causation and correlation". No, we can't conclude that decrease in number of pirates caused the climate change because there might be other factors (lurking or confounding variables) influencing this phenomenon.

Therefore, there might be a correlation between global average temperature and number of pirates, but based on this information we can't say that pirated died because of rise in global average temperature.

Q19. What is the difference between covariance and correlation?

Answer: Correlation is the standardized form of covariance.

Covariances are difficult to compare. For example: if we calculate the covariances of salary ($) and age (years), we'll get different covariances which can't be compared because of having unequal scales. To combat such situation, we calculate correlation to get a value between -1 and 1, irrespective of their respective scale.

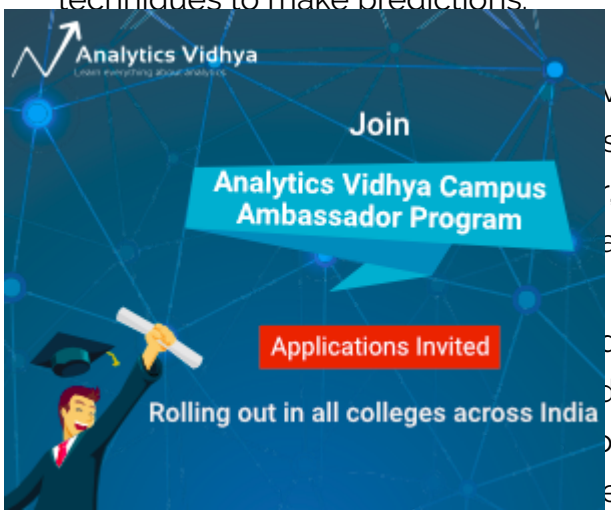
Q20. Is it possible capture the correlation between continuous and categorical variable? If yes, how?

A (analysis of covariance) between continuous and

nm, how is random forest

orithm (GBM)?

e is, random forest uses boosting techniques to make predictions.

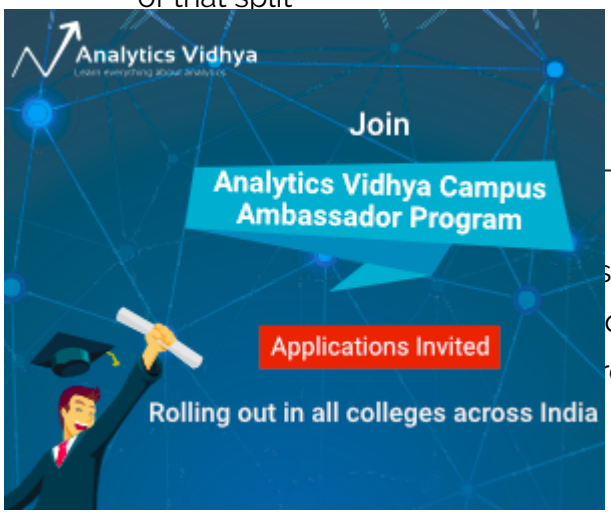


vided into n samples using
single learning algorithm a
, the resultant predictions
aging. Bagging is done is
round of predictions, the
dictions higher, such that
ding round. This sequential
misclassified predictions
eached.

Random forest improves model accuracy by reducing variance (mainly). The trees grown are uncorrelated to maximize the decrease in variance. On the other hand, GBM improves accuracy my reducing both bias and variance in a model.

Know more: Tree based modeling (https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/)

Q22. Running a binary classification tree algorithm is the easy



part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?

s decision based on Gini
words, the tree algorithm
can divide the data set into

ems from a population at
calculate Gini as following:

g formula sum of square of
(p^2+q^2).
Calculate Gini for split using weighted Gini score of each node
of that split

given by (for binary class):

$$- q \log_2 q$$

s and failure respectively in
ode is homogeneous. It is
e present in a node at 50%

Q23. You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?

Answer: The model has overfitted. Training error 0.00 means the classifier has mimiced the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on unseen sample, it couldn't find those patterns and returned prediction with higher error. In random forest, it happens when we use larger number of trees

than necessary. Hence, to avoid these situation, we should tune number of trees using cross validation.

aving p (no. of variable) > bad option to work with?

use? Why?

data sets, we can't use

their cross products to the

calculate a unique least

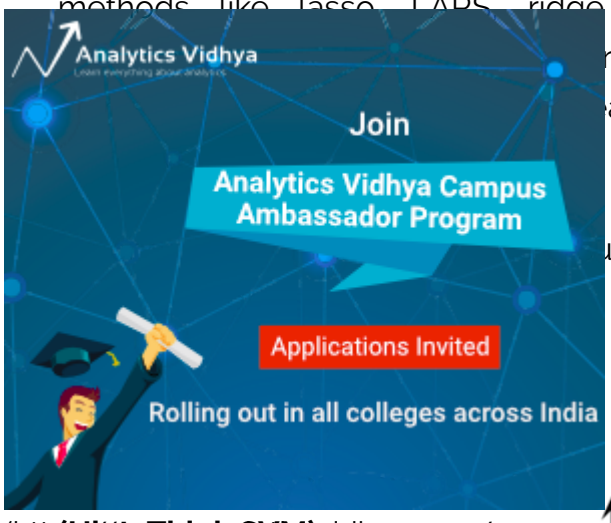
ances become infinite, so

To combat this situation, we can use penalized regression methods like lasso, LARS, ridge which can shrink the

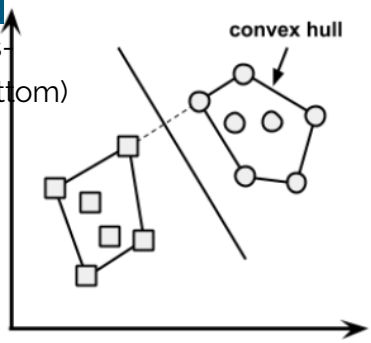
recisely, ridge regression

ast square estimates have

ubset regression, forward



(Hint: Think SVM)



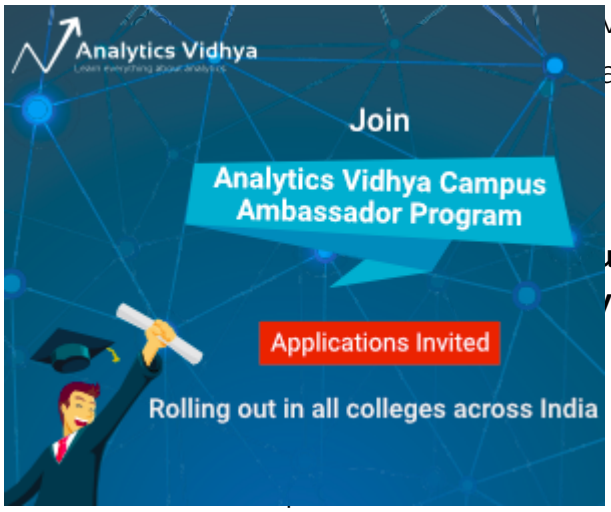
Answer: In case of linearly separable data, convex hull represents the outer boundaries of the two group of data points. Once convex hull is created, we get maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to create greatest separation between two groups.

Q26. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn't. How?

stion. It's a simple question

ionality (a.k.a features) in a reates a new variable for riables. For example: let's oding 'color' variable will or Red, Color_Blue and e.

In label encoding, the levels of a categorical variables gets variable is created. Label ariables.

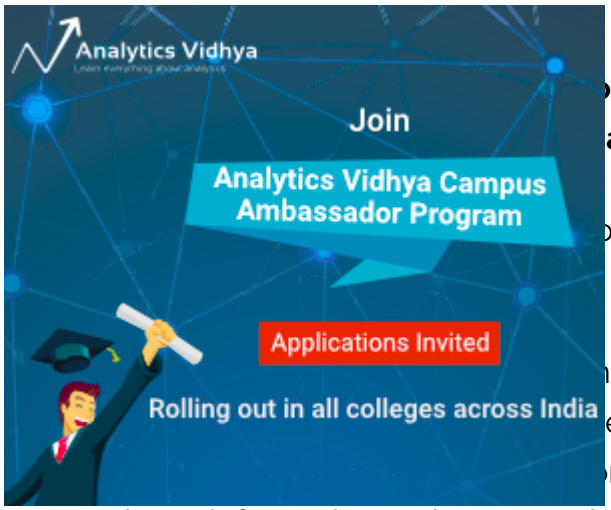


ue would you use on time y?

be troublesome because there might be some pattern in year 4 or 5 which is not in year 3. Resampling the data set will separate these trends, and we might end up validation on past years, which is incorrect. Instead, we can use forward chaining strategy with 5 fold as shown below:

- fold 1 : training [1], test [2]
- fold 2 : training [1 2], test [3]
- fold 3 : training [1 2 3], test [4]
- fold 4 : training [1 2 3 4], test [5]
- fold 5 : training [1 2 3 4 5], test [6]

where 1,2,3,4,5,6 represents "year".



Q28. You are given a data set consisting of variables having more than 30% missing values? Let's say, out of 50 variables, gher than 30%. How will

e following ways:

sing values, who knows the me trend

distribution with the target we'll keep those missing

category while removing

ought…' recommendations algorithm?

f recommendation engine

nsiders "User Behavior" for ehavior of other users and ry, ratings, selection and purchase information. Other users behaviour and preferences over the items are used to recommend items to the new users. In this case, features of the items are not known.

Know more: Recommender System (https://www.analyticsvidhya.com/blog/2015/10/recommendation-engines/)

Q30. What do you understand by Type I vs Type II error ?

Answer: Type I error is committed when the null hypothesis is

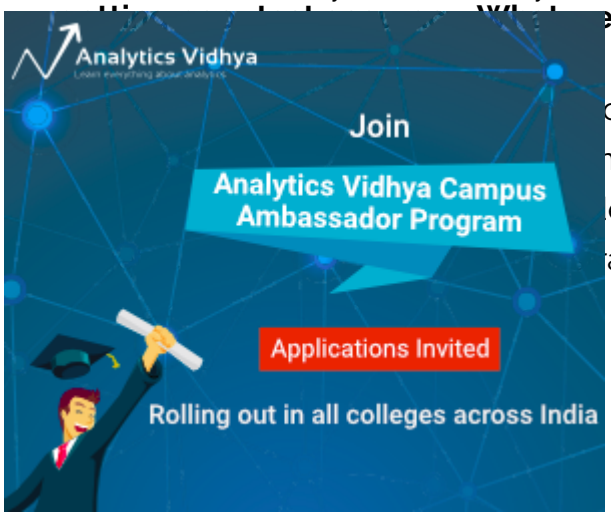
true and we reject it, also known as a 'False Positive'. Type II error is committed when the null hypothesis is false and we accept it, also known as 'False Negative'.

we can say Type I error as positive (1) when it is ccurs when we classify a ly positive(1).

(https://courses.analyticsvidhya.com/pages/login?utm_source=banner5yearanniversary)

ssification problem. For nly sampled the training u are confident that your validation accuracy is high. However, you get shocked after ent wrong?

oblem, we should always ndom sampling. A random eration the proportion of atified sampling helps to variable in the resultant

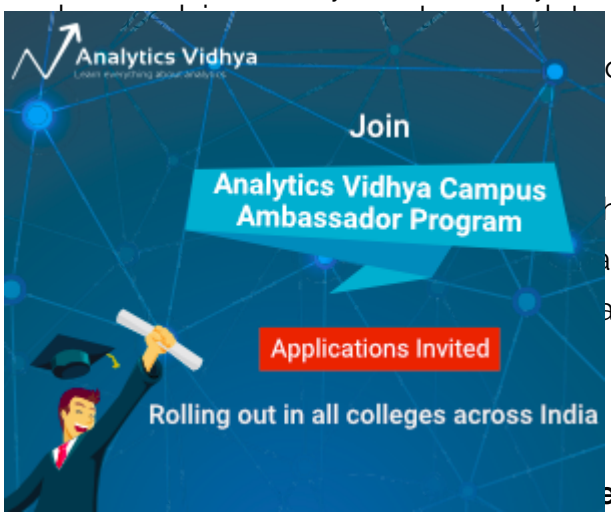


(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

Q32. You have been asked to evaluate a regression model based on R^2, adjusted R^2 and tolerance. What will be your criteria?

Answer: Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of percent of variance in a predictor which cannot be accounted by other predictors. Large values of tolerance is desirable.

We will consider adjusted R^2 as opposed to R^2 to evaluate model fit because R^2 increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R^2 would only increase if an additional variable improves the



accuracy of model, otherwise stays same. It is difficult to comment a general threshold value for adjusted R^2 because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R^2 and still provide fairly o a stock market data odel is not good.

e euclidean distance to

ttan distance because

or vertically only. It has

distance. Since, the data

on, euclidean distance is a

ne movement made by a

nhattan distance because

l movements.

e like a 5 year old.

Answer: It's simple. It's just like how babies learn to walk. Every time they fall down, they learn (unconsciously) & realize that their legs should be straight and not in a bend position. The next time they fall down, they feel pain. They cry. But, they learn 'not to stand like that again'. In order to avoid that pain, they try harder. To succeed, they even seek support from the door or wall or anything near them, which helps them stand firm.

This is how a machine works & develops intuition from its environment.

Note: The interview is only trying to test if have the ability of explain complex concepts in simple terms.

sion model is generally value. How would you l?

ethods:

to predict probabilities, we with confusion matrix to

f adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we m AIC value.

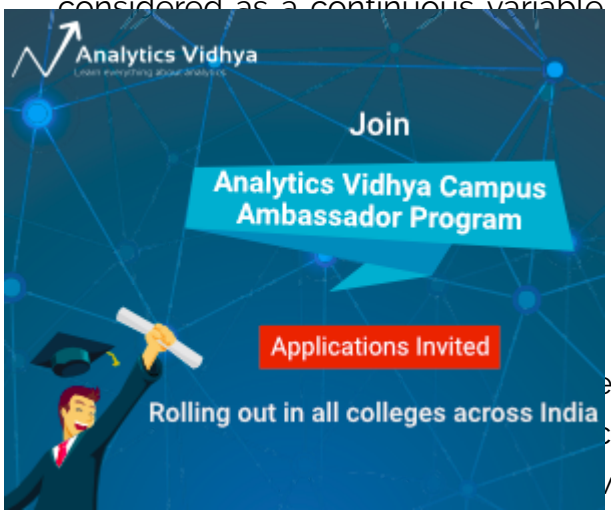
onse predicted by a model lower the value, better the s the response predicted by variables. Lower the value,

og/2015/11/beginners-

Q36. Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?

Answer: You should say, the choice of machine learning algorithm solely depends of the type of data. If you are given a data set which is exhibits linearity, then linear regression would be the best algorithm to use. If you given to work on images, audios, then neural network would help you to build a robust model.

If the data comprises of non linear interactions, then a boosting



or bagging algorithm should be the choice. If the business requirement is to build a model which can be deployed, then we'll use regression or a decision tree model (easy to interpret and explain) instead of black box algorithms like SVM, GBM etc.

rithm for all situations. We erstand which algorithm to

a categorical variable as in a better predictive

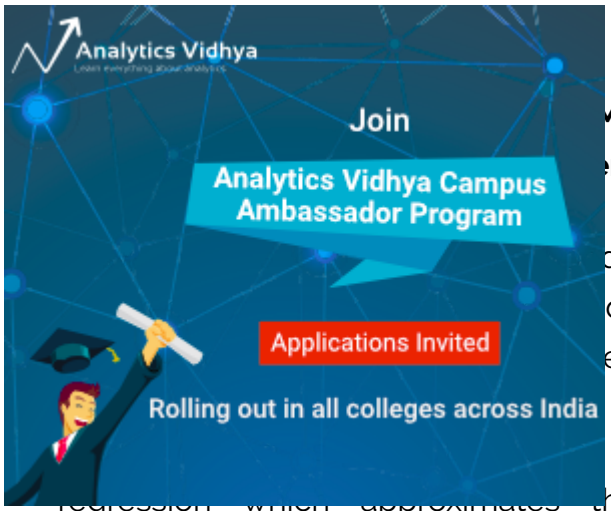
Answer: For better predictions, categorical variable can be considered as a continuous variable only when the variable is

becomes necessary in

ecessary when the model chnique introduces a cost with the objective function. Hence it tries to push the coefficients for many variables to zero and hence reduce cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

Q39. What do you understand by Bias Variance trade off?

Answer: The error emerging from any model can be broken down into three components mathematically. Following are these component :



uch on an average are the actual value. A high bias rming model which keeps ance on the other side ade on same observation badly on any observation

Maximum likelihood is to ement.

d are the methods used by ds to approximate the e. In simple words,

method used in linear regression which approximates the parameters resulting in minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the the values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.

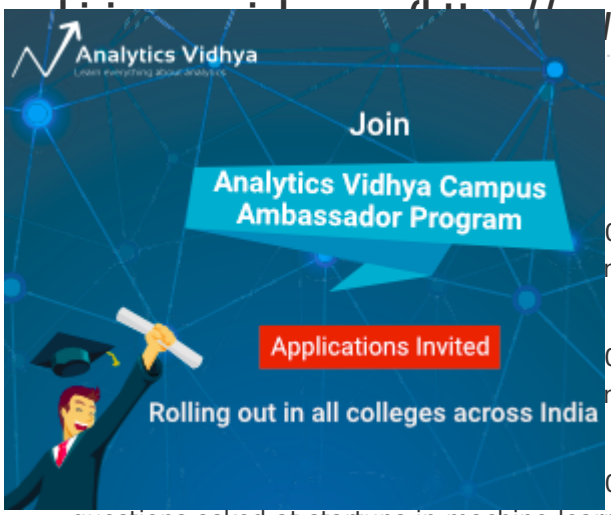
End Notes

You might have been able to answer all the questions, but the real value is in understanding them and generalizing your knowledge on similar questions. If you have struggled at these

questions, no worries, now is the time to learn and not perform.

You should right now focus on learning these topics scrupulously.

(https://id.analyticsvidhya.com/accounts/login/?utm_source=banner5yearanniversary)



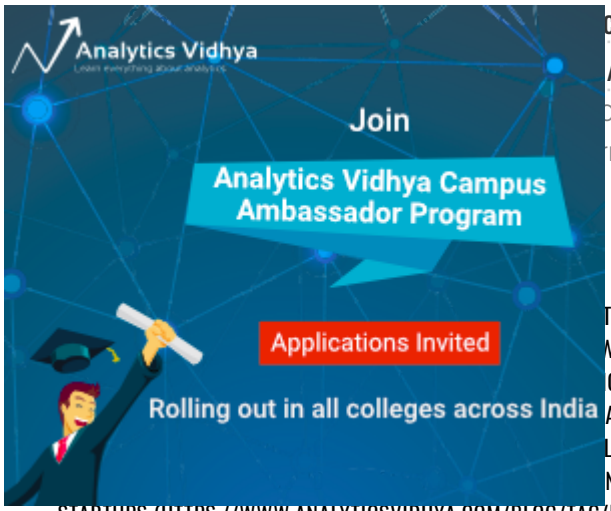
(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

(https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/?share=twitter&nb=1)

(https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/?share=pocket&nb=1)

(https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/?share=reddit&nb=1)

(https://www.analyticsvidhya.com/blog/2016/12/top-35-articles-for-the-analytics-vidhya-community-in-5-years/?utm_source=banner5yearanniversary)



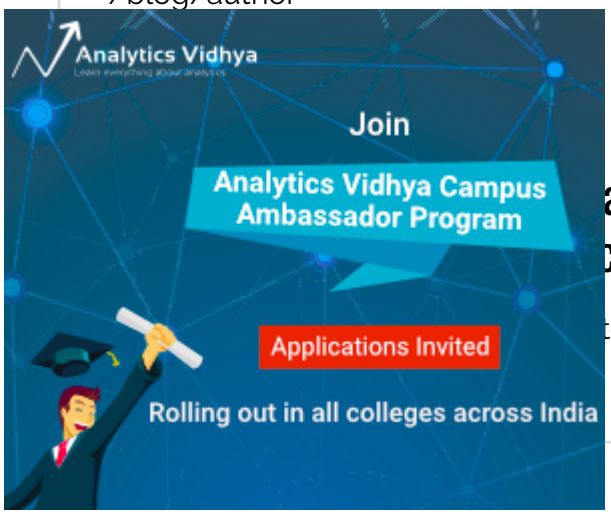
(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog-sidebottom)

/10/interview-data-scientist-gregory-piatetsky-shapiro-president-kdnuggets/)
Interview with Data Scientist- Gregory Piatetsky-Shapiro, President KDnuggets
(https://www.analyticsvidhya.com/blog/2015/10/interview-data-scientist-gregory-piatetsky-shapiro-president-kdnuggets/)
October 13, 2015
In "Business Analytics"

TICSVIDHYA.COM/BLOG/TAG/BAGGING/), WW.ANALYTICSVIDHYA.COM/BLOG/TAG G (HTTPS://WWW.ANALYTICSVIDHYA.COM ATION LIDATION/), DATA SCIENCE IN STARTUPS NCE-IN-STARTUPS/), DATA SCIENTIST IN STARTUPS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-SCIENTIST-IN-STARTUPS/), DATA SCIENTIST INTERVIEW (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-SCIENTIST-INTERVIEW/), INTERVIEW QUESTIONS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/INTERVIEW-QUESTIONS/), MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MACHINE-LEARNING/), MACHINE LEARNING ENGINEER INTERVIEW QUESTION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MACHINE-LEARNING-ENGINEER-INTERVIEW-QUESTION/), MAXIMUM LIKELIHOOD (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MAXIMUM-LIKELIHOOD/), ORDINARY LEAST SQUARE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ORDINARY-LEAST-SQUARE/), REGULARIZATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/REGULARIZATION/)

Previous Article
AWS / Cloud Engineer – Pune (4+ Years of

Next Article
Comprehensive Introduction to Apache Spark, RDDs & Dataframes (using PySpark)



This is article is quiet old now and you might not get a prompt response from the author. We would request you to post this comment on Analytics Vidhya **Discussion portal** (https://discuss.analyticsvidhya.com/) to get your queries resolved.

33 COMMENTS

kavitha says:
SEPTEMBER 16, 2016 AT 7:09 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116110#RESPOND)
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-

Analytics Vidhya Content Team says:

SEPTEMBER 16, 2016 AT 5:47 AM

Good to know you found them helpful! All the best.

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116121#RESPOND)

Analytics Vidhya Content Team says:

SEPTEMBER 16, 2016 AT 5:47 AM

(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116121)

Hi Gianni, I am happy to know that these question would help you in your journey. All the best.

Prof Ravi Vadlamani says:

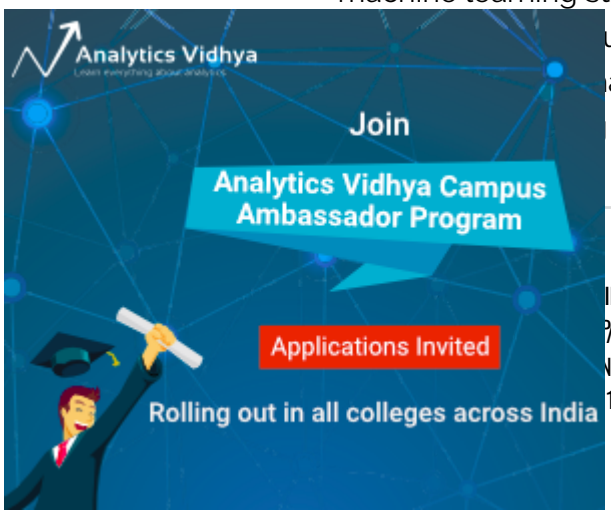
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116113#RESPOND)
SEPTEMBER 16, 2016 AT 9:00 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116113)

Good collection compiled by you Mr Manish ! Kudos ! to the budding data

Content Team says:

SCIENCE/?REPLYTOCOM=116123#RESPOND)
/IDHYA.COM/BLOG/2016/09/40-
ED-AT-STARTUPS-IN-MACHINE-
OMMENT-116123)

right. These questions can
machine learning startups, facing off ML
uestion have higher
ave laid emphasis on
as well.



(https://www.analytics vidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-
POND)
/WWW.ANALYTICSVIDHYA.COM
NS-ASKED-AT-STARTUPS-IN-MACHINE-
117)

Beginners like me.

Analytics Vidhya Content Team says:

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116118#RESPOND)
SEPTEMBER 16, 2016 AT 9:44 AM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116118)

Welcome 🙂

chibole says:

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116124#RESPOND)
SEPTEMBER 16, 2016 AT 9:56 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116124)

It seems Stastics is at the centre of Machine Learning.

Hi Chibole,

True, statistics is an inevitable part of machine
s to understand statistical
o master machine learning.

ndering, do you recommend
for somebody to special in a specific
field of ML? I mean, it is
recommended to choose between
supervised learning and
unsupervised learning algorithms,
and simply say my specialty is this
during an interview. Shouldn't
organizations recruiting specify their
specialty requirements too?

....and thank you for the post.

(https://id.analyticsvidhya.com/account/login/?utm_source=banner5yearanniversary)

(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116129#RESPOND)

Analytics Vidhya Content Team says:

SEPTEMBER 16, 2016 AT 10:40 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116129)

Hi Chibole,

It's always a good thing to establish yourself as an expert in a specific field. This helps the recruiter to understand that you are a detailed oriented person. In machine learning, thinking of building your expertise in supervised learning would be good, but companies want more than that. Considering, the variety of data these days, they want someone who can deal with unlabeled data also. In short, they look for someone who isn't just an expert in operating Sniper Gun, but can use other weapons also if needed.

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116125#RESPOND)

chibole says:

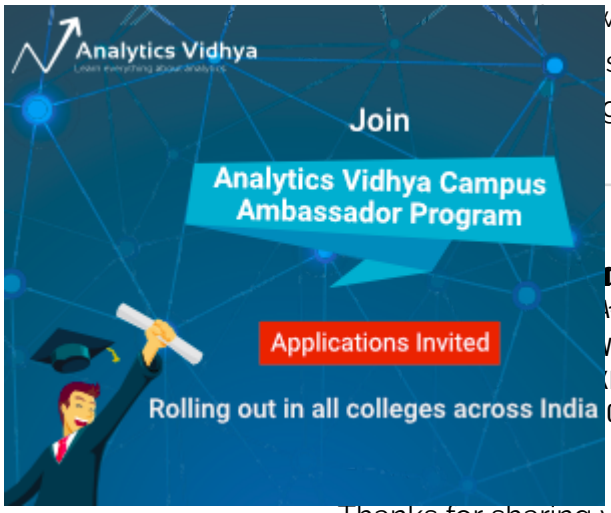
SEPTEMBER 16, 2016 AT 9:36 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116125)

* stastics = Statistics

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116135#RESPOND)

Karthikeyan Sankaran (http://bit.ly/31KA-T9) says:

SEPTEMBER 16, 2016 AT 1:10 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116135)



ntent Team says:

Hi really an interesting collection of answers. From a
v there are some
s surely useful for job
ger firms.

ntent Team says:

Thanks for sharing your thoughts. Tell me more about Q40. What's about it?

Raja (http://None) says:

I think you got Q3 wrong.

It was to calculate from median and not mean.
how can assume mean and median to be same

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=...#RESPOND)

Raju (http://None) says:
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116149)

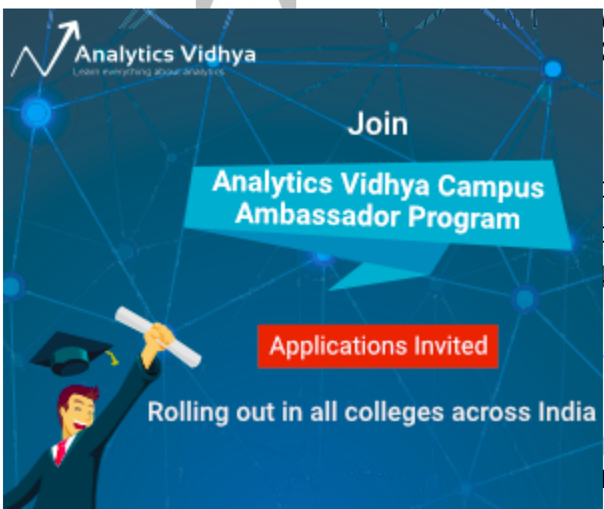
Don't bother....Noted .....you assumed normal distribution....

(https://datahack.analyticsvidhya.com/...
/?utm_source=...

INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-
...ND) (HTTPS://WWW.ANALYTICSVIDHYA.COM
...NS-ASKED-AT-STARTUPS-IN-MACHINE-
...154)

...derstanding which topics to
...es.

Analytics Vidhya Content Team says:
ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116208#RESPOND)
SEPTEMBER 16, 2016 AT 12:38 PM
...VIDHYA.COM/BLOG/2016/09/40-
...KED-AT-STARTUPS-IN-MACHINE-
...COMMENT-116208)



...couraging words! The
...cle is to help beginners
...ky side of ML interviews.

INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-
...POND) (HTTPS://WWW.ANALYTICSVIDHYA.COM
/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-
LEARNING-DATA-SCIENCE/#COMMENT-116161)

(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

Dear Kunal,

Few queries i have regarding AIC

1)why we multiply -2 to the AIC equation
2)where this equation has been built.

Rgds

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116185#RESPOND)

Sampath says:
SEPTEMBER 16, 2016 AT 6:00 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM

/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116185)

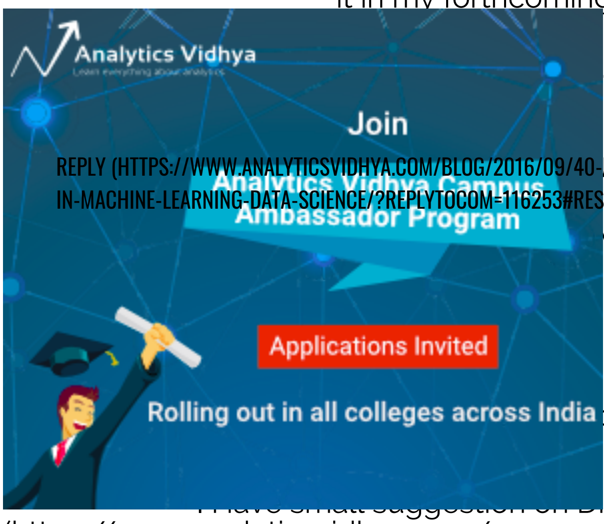
Hi Manish,

Great Job! It is a very good collection of interview ... g. It will be a great help if ... Thanks

Content Team says:
...34 PM
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116207#RESPOND)
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116207)

Hi Sampath,

Thanks for your suggestion. I'll surely consider it in my forthcoming articles.



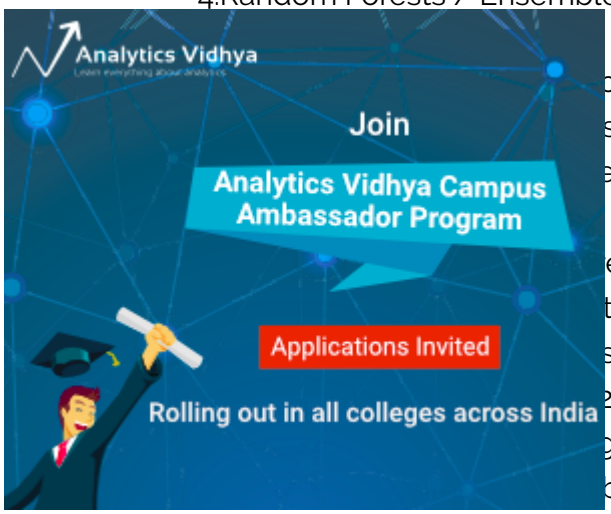
... (vkarthi46) says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116253#RESPOND)
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116253)

... on for beginners

I have small suggestion on Dimensionality Reduction,We can also use the below mentioned techniques to reduce the dimension of the data.

1.Missing Values Ratio
Data columns with too many missing values are unlikely to carry much useful information. Thus data columns with number of missing values greater than a given threshold can be removed. The higher the threshold, the more aggressive the reduction.

2.Low Variance Filter
Similarly to the previous technique, data columns with little changes in the data carry little information. Thus all data columns with variance lower than a given threshold



are removed. A word of caution: variance is range dependent; therefore normalization is required before applying this technique.

2.High Correlation Filter

r trends are also likely to
n this case, only one of
achine learning model.
tion coefficient between
een nominal columns as the
efficient and the Pearson's
pairs of columns with
han a threshold are
f caution: correlation is
mn normalization is
relation comparison.

4.Random Forests / Ensemble Trees

o referred to as random
selection in addition to
e approach to
generate a large and
ees against a target
ttribute's usage statistics to
set of features. Specifically,
2000) of very shallow trees
g trained on a small fraction
outes. If an attribute is often
selected as best split, it is most likely an informative feature to retain. A score calculated on the attribute usage statistics in the random forest tells us – relative to the other attributes – which are the most predictive attributes.

5.Backward Feature Elimination

In this technique, at a given iteration, the selected classification algorithm is trained on n input features. Then we remove one input feature at a time and train the same model on n-1 input features n times. The input feature whose removal has produced the smallest increase in the error rate is removed, leaving us with n-1

input features. The classification is then repeated using n-2 features, and so on. Each iteration k produces a model trained on n-k features and an error rate e(k). Selecting the maximum tolerable error rate, we define the smallest number of features necessary to reach that

h the selected machine

on

the Backward Feature

the feature that produces

mance. Both algorithms

and Forward Feature

nd computationally

e, it is probably quite unfeasible to a data set with an already relatively low number of input

INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-
POND)
W.ANALYTICSVIDHYA.COM
NS-ASKED-AT-STARTUPS-IN-MACHINE-
334)

estion I feel I am at 10% of
e career in Data Science .
Excellent Article to read. Can you Please suggest me any book or training online which gives this much deep information . Waiting for your reply in anticipation . Thanks a million

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116357#RESPOND)

Rahul Jadhav says:

SEPTEMBER 22, 2016 AT 5:36 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116357)

Amazing Collection Manish!
Thanks a lot.

Nikhil says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116458#RESPOND)
NOVEMBER 7, 2016 AT 6:42 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116458)

An awesome article for reference. Thanks a ton Manish sir
e pdf format of this blog
om/blog/category/avbytes/)

(https://id.analyticsvidhya.com/accounts/login/?utm_source=banner5yearanniversary)

(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

Siddharth says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=118048#RESPOND)
NOVEMBER 7, 2016 AT 6:42 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-118048)

Really awesome article thanks. Given the influence young, budding students of machine learning will likely have in the future, your article is of great value.

LEAVE A REPLY

Your email address will not be published.

(https://courses.analyticsvidhya.com/pages/black-friday-sale?utm_source=banner5yearanniversary)

(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

(https://id.analyticsvidhya.com/accounts/login/?utm_source=banner5yearanniversary)

(https://www.analyticsvidhya.com/campus-ambassador/?utm_source=AVblog_sidebottom)

ANALYTICS VIDHYA
About Us (http://www.analyticsvidhya.com/team/)
DATA SCIENTISTS
Blog (https://www.analyticsvidhya.com/blog/)
Hackathon (https://datahack.analyticsvidhya.com)
Discussions (https://discuss.analyticsvidhya.com)
Leaderboard (https://datahack.analyticsvidhya.com/users/)
COMPANIES
Post Jobs (https://www.analyticsvidhya.com/corporate/)
Trainings (https://trainings.analyticsvidhya.com)
Hiring Hackathons (https://datahack.analyticsvidhya.com)
Advertising (https://www.analyticsvidhya.com/contact/)
Reach Us (https://www.analyticsvidhya.com/contact/)
JOIN OUR COMMUNITY :
48782 Followers
15593 Followers
2734 Followers
5065 Followers
Subscribe to emailer >
Policy (https://www.analyticsvidhya.com/privacy-policy/)
Use (https://www.analyticsvidhya.com/terms/)
Policy (https://www.analyticsvidhya.com/refund-policy/)
Don't have an account? Sign