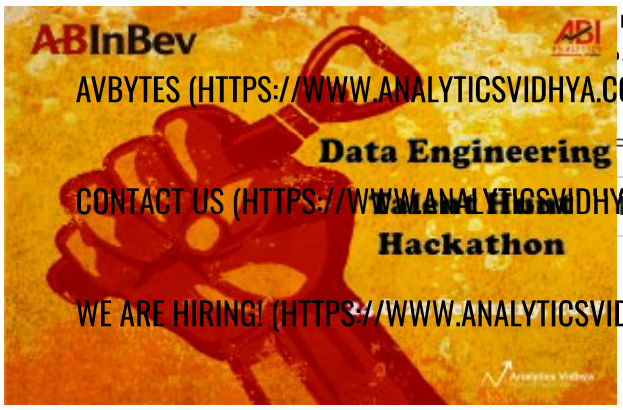
40 Interview Questions asked at Startups in Machine Learning / Data Science

MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/)

TOP ANALYTICS VIDHYA USERS

Rank	Name
1	vopani (https://datahack. /user/profile/Roh
2	SRK (https://datahack. /user/profile/SRk
3	binga (https://datahack. /user/profile/bing



(https://datahack.analyticsvidhya.com/contest
/data-engineering-talent-hunt-hackathon

Carefull Those question can make you think
RICE!

(https://datahack.analyticsvidhya.com/contest
/the-young-optimization-crackerjack-contest
/?utm_source=AVblog_sidebottom)

e are being looked as the
ion happening in the world
e are numerous exciting
ya.com/blog/2016/09/18-
ator-summer-
What could be a better start
for your aspiring career!

However, getting into these roles 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.

	4	Aayushmnit (https://datahack. /user/profile/aay
	5	Mark Landry (https://datahack. /user/profile/mar

More Rankings (http://datahack.analyticsvidhya.com/users)

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

(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebar)

lachine Learning

aving 1000 columns and 1 based on a classification ked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)

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

POPULAR POSTS

- Essentials of Machine Learning Algorithms (with Python and R Codes) (https://www.analyticsvidh /blog/2017 /09/common-machine-learning-algorithms/)
- 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/)
- 7 Types of Regression Techniques you should know! (https://www.analyticsvidh /blog/2015 /08/comprehensive-guide-regression/)
- 11 most read Deep Learning Articles from

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- ...nponents which can explain ... set.

...ke Vowpal Wabbit (available ...

...chastic Gradient Descent is ...

...understanding to estimate ...the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in significant loss of information.

...read about online learning ...yticsvidhya.com/blog/2015 ...arning-simplified-2/) & ...://www.quora.com/Whats- ...escent-and-stochastic- ...ced methods.

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.

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.

icsvidhya.com/blog/2016
nent-analysis-python/)

ta set has missing values
iation from the median,
ain unaffected? Why?

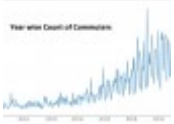
Answer: . This question has enough hints for you to start thinking! Since the data is spread across median, let's assume , in a normal distribution,
d deviation from mean (or
% of the data unaffected.
uld remain unaffected by

cer detection. You've build
d an accuracy of 96%. Why shouldn't you be happy with your model performance? What can you do about it?

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 to be poor, we can undertake the

RECENT POSTS

(https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/)
Natural Language Processing for Beginners: Using TextBlob (https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/)
SHUBHAM JAIN , FE...



following steps:

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

(https://datahack.analytics...
/data-engineering-talent-hunt-hackathon

...threshold value by doing ...//www.analyticsvidhya.com ...mic-regression-minimize- ... optimal threshold using ...es such that the minority ...h.

(https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/)

(https://datahack.analytics...
/the-young-optimization-crackerjack-contest
/?utm_source=AVblog_sidebar)

...cause it assumes that all of ...e equally important and ...ssumption are rarely true in ...

Q6. Explain prior probability, likelihood and marginal likelihood in context of naiveBayes algorithm?

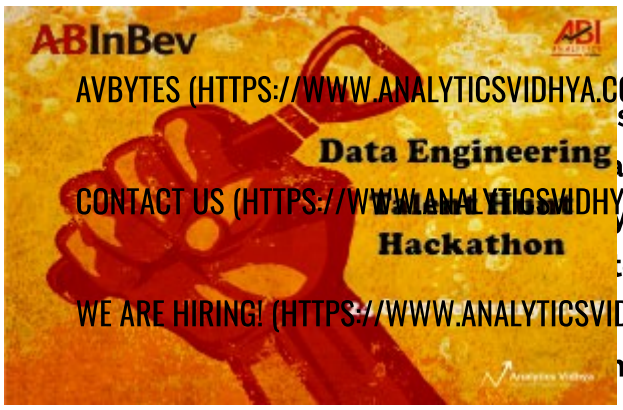
Answer: Prior probability is nothing but, the proportion of dependent (binary) variable in the data set. It is the closest guess you can make about a class, without any further information. For example: In a data set, the dependent variable is binary (1 and 0). The proportion of 1 (spam) is 70% and 0 (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

(https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/)
7 methods to perform Time Series forecasting (with Python codes) (https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/)
GURCHETAN SINGH ,...

(https://www.analyticsvidhya.com/blog/2018/02/introductory-guide-regularized-greedy-forests-rgf-python/)
An Introductory Guide to Regularized Greedy Forests (RGF) with a case study in Python (https://www.analyticsvidhya.com/blog/2018/02/introductory-guide-regularized-greedy-forests-rgf-python/)
ANKIT CHOUDHARY ...

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



(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

Q7. You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the k-NN algorithm since you know it works fairly well for all kinds of data. Later, you tried a time series regression model and got higher accuracy than k-NN. Why?

Answer: Time series data is known to posses linearity. On the other hand, k-NN algorithm is known to work best to

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

Q8. You are assigned a new project 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

(https://www.analyticsvidhya.com/blog/2018/02/demystifying-security-data-science/)
Demystifying Information Security Using Data Science (https://www.analyticsvidhya.com/blog/2018/02/demystifying-security-data-science/)
GUEST BLOG , FEBR...

(http://www.edvancer.in/certified-data-scientist-with-python-course?utm_source=AV&utm_medium=AVads&

three things:

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

s to decide if machine

roblem.

model is suffering from low bias and high variance, which algorithm should you use to tackle it? Why?

e model's predicted values

words, the model becomes

g data distribution. While it

ut not to forget, a flexible

ilities. It means, when this

ata, it gives disappointing

ging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling. Then, these samples are 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.

(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=blog_sidebutton)

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 n presence of correlated a particular component

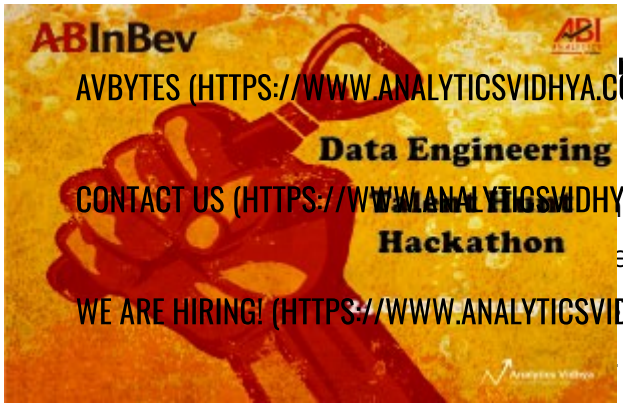
a data set, of which 2 are ...on in PCA, you think the first principal component would exhibit twice the variance than it would ... Also, adding correlated ...portance on those variable,

..., you are now anxious to ...a result, you build 5 GBM ...ithm would do the magic. ...could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. 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.



(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

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 ... in a regression ... del prediction without any ... tion. The formula of $R^2 = 1$ ... predicted value.

... of intercept term (ymean), the model can make no such evaluation, with large denominator, $\Sigma(y - y\acute{})^2/\Sigma(y)^2$ equation's value becomes ... er R^2.

...ur manager has informed ...ring from multicollinearity. ...rue? Without losing any ...tter model?

...we can create a correlation matrix to identify & remove variables having correlation above 75% (deciding a threshold is subjective). 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/2016/07/deeper-regression-analysis-assumptions-plots-solutions/)

(https://www.analyticsvidhya.com/data-engineering-talent-hunt-hackathon/)

able over Lasso

rs Hastie, Tibshirani who

In presence of many variables with small / medium sized effect use ridge regression.

regression (L1) does both shrinkage, whereas Ridge shrinkage and end up e model. In presence of on might be the preferred s best in situations where gher variance. Therefore, it

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

Know more: Ridge and Lasso Regression (https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/)

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.

(https://www.analyticsvidhya.com/blog/2015/06/establish-
/data-engineering-talent-hunt-hackathon)

Q18. How do you select important variables? Explain your methods.

Answer: Following are the methods of variable selection you

1. Remove correlated variables prior to selecting important
2. Use linear regression and select variables based on p values
3. Use Forward Selection, Backward Selection, Stepwise
4. Use Random Forest and plot variable importance
5. Use Lasso Regression.

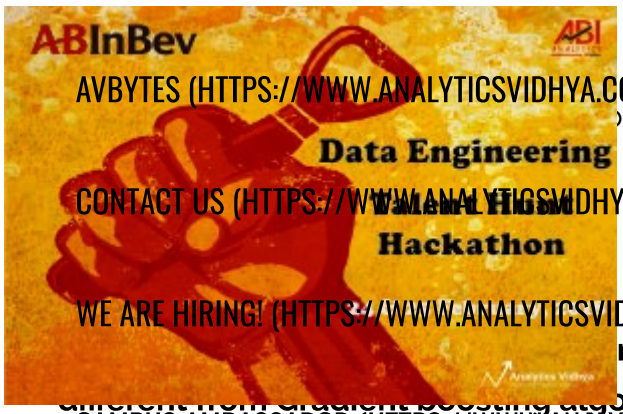
(https://datahack.analyticsvidhya.com/contest
/the-young-optimization-crackerjack-contest
/?utm_source=AVblog_sidebottom)

6. Measure information gain for the available set of features and select top n features accordingly.

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?

(analysis of covariance) between continuous and

m, how is random forest orithm (GBM)?

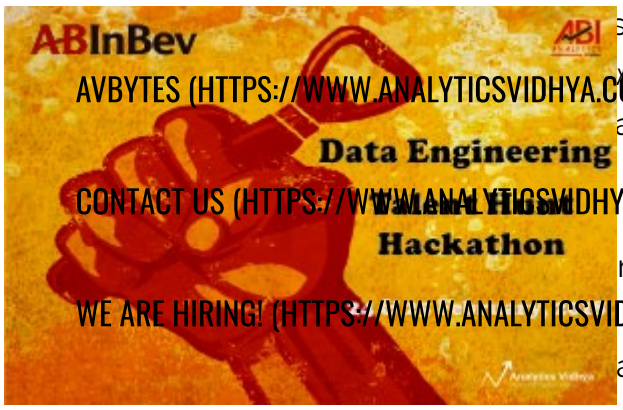
Answer: The fundamental difference is, random forest uses tions GBM uses boosting

ivided into n samples using single learning algorithm a r, the resultant predictions raging. Bagging is done is round of predictions, the dictions higher, such that ding round. This sequential process of giving higher weights to misclassified predictions continue until a stopping criterion is reached.

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



(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

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 an divide the data set into

ms from a population at

alculate Gini as following:

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node

s given by (for binary class):

$$-q \log_2 q$$

ess and failure respectively

node is homogeneous. It is

re present in a node at 50% is desirable.

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?

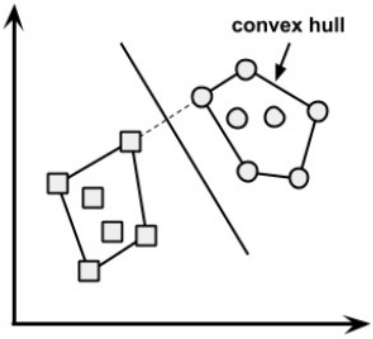
use? Why?

data sets, we can't use their cons, one can't calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.

use penalized regression which can shrink the Precisely, ridge regression least square estimates have

ubset regression, forward

Q25. What is convex hull ? (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 ?

tion. It's a simple question

onality (a.k.a features) in a

eates a new variable for

iables. For example: let's

oding 'color' variable will

categorical variables gets

variable is created. Label

ariables.

(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

ue would you use on time
V?

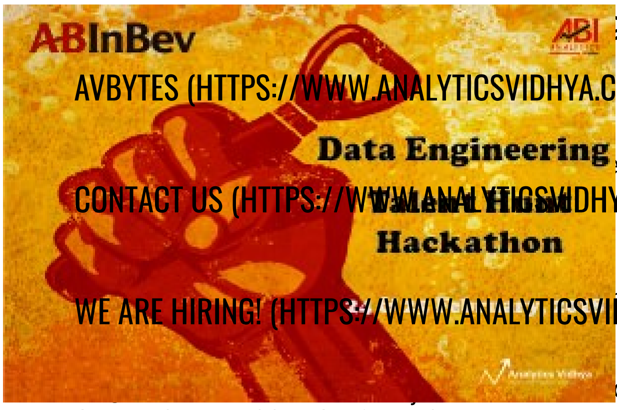
Answer: Neither.

In time series problem, k fold can 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



... following ways:

... ing values, who knows the ... e trend

... distribution with the target variable, and if found any pattern we'll keep those missing values and assign them a new category while removing

... ought...' recommendations ... algorithm?

... of recommendation engine

... nsiders "User Behavior" for recommending items. They exploit behavior of other users and items in terms of transaction history, 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

(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

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 s positive (1) when it is ccurs when we classify a y positive(1).

ssification problem. For n unseen data since your ver, you get shocked after ent wrong?

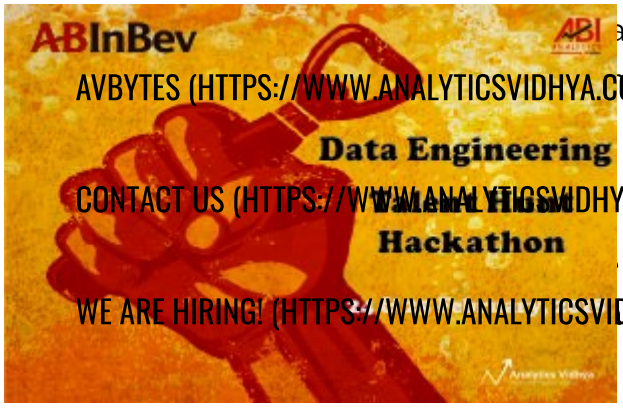
roblem, we should always ndom sampling. A random deration the proportion of tratified sampling helps to variable in the resultant

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 good ... a stock market data ... models category.



(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

... euclidean distance to ... nearest neighbors? Explain.

Answer: ... We don't use manhattan distance because ... or vertically only. It has ... hand, euclidean metric can ... distance. Since, the data ... sion, euclidean distance is a ...

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

... the movement made by a ... anhattan distance because ... tal movements.

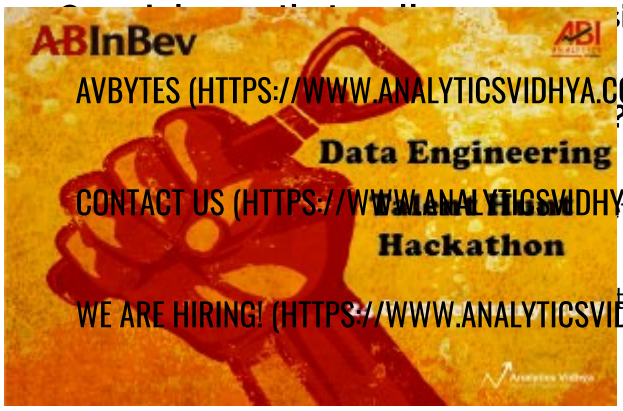
Q34. Explain machine learning to me like a 5 year old.

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.



(https://datahack.analytics vidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analytics vidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

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

(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

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.

ithm for all situations. We rstand which algorithm to

categorical variables as **in a better predictive**

categorical variable can be only when the variable is

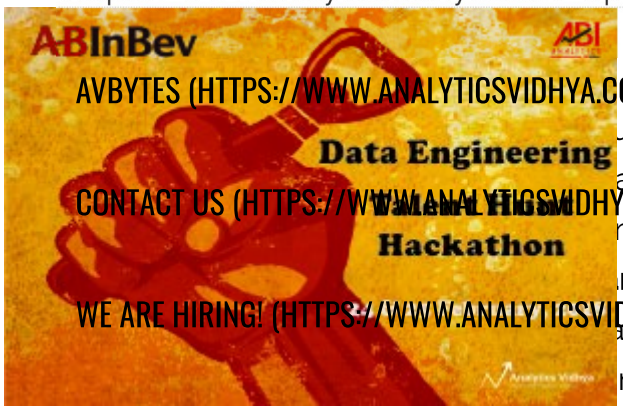
becomes necessary in

ecessary when the model chnique introduces a cost term for bringing in more features 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 :

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\hat{f}(x) - E[\hat{f}(x)]\right]^2 + \sigma_e^2$$



uch on an average are the
actual value. A high bias
ming model which keeps
nce on the other side
ade on same observation
nce model will over-fit on
beyond training.

Maximum likelihood is to
ement.

od are the methods used
thods to approximate the
ue. In simple words,

Ordinary least square(OLS) is a 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.

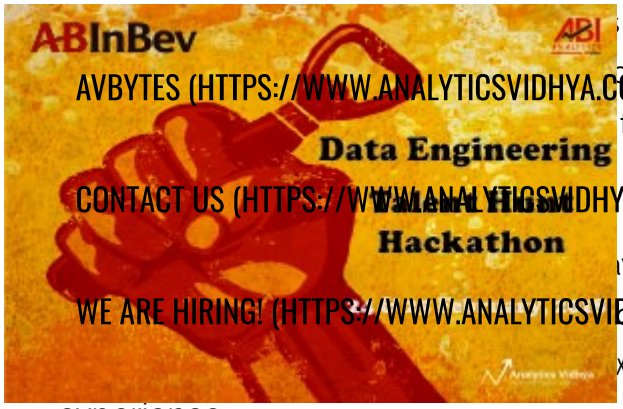
End Notes

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.

These questions are meant to give you a wide exposure on the [illegible] in machine learning. I'm [illegible]ou curious enough to do [illegible]f you are planning for it,

[illegible]ave you appeared in any [illegible]w. I'd love to know your experience.



(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hackathon)

[illegible]? Check out currently [illegible]ww.analyticsvidhya.com [illegible]d data science.

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

TAGS: BAGGING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BAGGING/), BIAS VARIANCE TRADEOFF (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BIAS-VARIANCE-TRADEOFF/), BOOSTING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BOOSTING/), CROSS-VALIDATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CROSS-VALIDATION/), DATA SCIENCE IN STARTUPS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-SCIENCE-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
(https://www.analyticsvidhya.com)

(https://datahack.analyticsvidhya.com/contest
/data-engineering-talent-hunt-hackathon

(https://datahack.analyticsvidhya.com/contest
/the-young-optimization-crackerjack-contest
/?utm_source=AVblog_sidebottom)

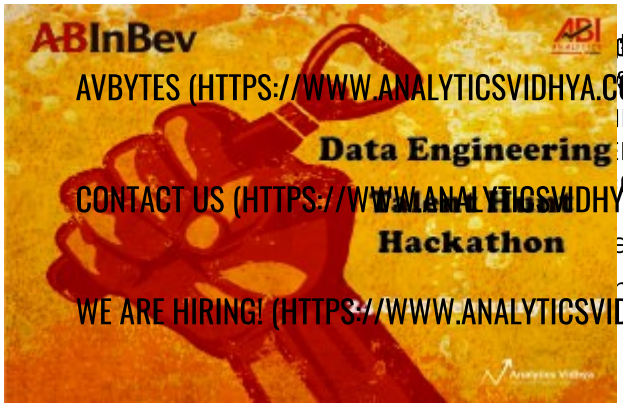
Next Article
Comprehensive Introduction to Apache Spark, RDDs & Dataframes (using PySpark)
(https://www.analyticsvidhya.com/blog/2016/09/comprehensive-introduction-to-apache-spark-rdds-dataframes-using-pyspark/)

a Content Team
nalyticsvidhya.com
/contentteam/)
ntent team

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



(https://datahack.analyticsvidhya.com
/data-engineering-talent-hunt-hackathon

(https://datahack.analyticsvidhya.com
/the-young-optimization-crackerjack-contest
/?utm_source=AVblog_sidebottom)

Analytics Vidhya Content Team says:

ese questions help you to
ning interview rounds. All
full to face on the true reality
me 🙂

Analytics Vidhya Content Team says:

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

Analytics Vidhya Content Team says:

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

(https://datahack.analytics vidhya.com /data-engineering-talent-hunt-hackathon

(https://datahack.analytics vidhya.com/contest /the-young-optimization-crackerjack-contest /?utm_source=AVblog_sidebottom)

Prof Ravi Vadlamani says:
SEPTEMBER 16, 2016 AT 7:40 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:
...DHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116123)

Hi Prof Ravi, You are right. These questions can ...startups, facing off ML ...question have higher ...have laid emphasis on ...g as well.

...-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116117)

Thank you Manish Helpful for Beginners like me.

Analytics Vidhya Content Team says:
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 🙂

chibele says:
SEPTEMBER 16, 2016 AT 9:35 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.

(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

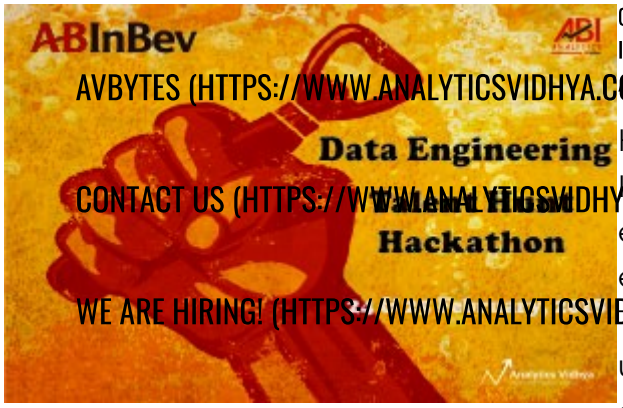
(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

...n inevitable part of machine ...ds to understand statistical ...o master machine learning.

...says:
...HYA.COM/BLOG/2016/09/40-INTERVIEW-...S-IN-MACHINE-LEARNING-DATA-SCIENCE
...16, 2016 AT 9:55 AM
...WW.ANALYTICSVIDHYA.COM/BLOG/2016...ERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116128)

I was wondering, 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://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?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/31KArSB) says:

SEPTEMBER 16, 2016 AT 1:45 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116135)

(https://datahack.analyticsvidhya.com/contest/data-engineering-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

Hi Manish – Interesting & Informative set of questions & answers. Thanks for compiling the same!

tent Team says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116209#RESPOND)
SEPTEMBER 16, 2016 AT 10:22 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116209)

…ction of answers. From a …w there are some … is surely useful for job …gger firms.

ontent Team says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116150#RESPOND)
SEPTEMBER 17, 2016 AT 7:41 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116150)

Hi Nicola,

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

Raja (http://None) says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116148#RESPOND)
SEPTEMBER 17, 2016 AT 1:45 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116148)

I think you got Q3 wrong.

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

Raja (http://None) says:
SEPTEMBER 17, 2016 AT 5:54 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116149)

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116149#RESPOND)

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

Chinmaya Mishra says:
SEPTEMBER 17, 2016 AT 10:03 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116161)

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/?REPLYTOCOM=116161#RESPOND)

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

Sampath says:
SEPTEMBER 18, 2016 AT 5:33 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116185)

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

/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 ... It will be a great help if ... Thanks

... Team says:
... SCIENCE/?REPLYTOCOM=116207#RESPOND)
... PM
...IDHYA.COM/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116207)

Hi Sampath,

...ggestion. I'll surely consider ...g 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 ...ONS-ASKED-AT-STARTUPS-IN-MACHINE- ...6253)

Kudos to you!! Good Collection 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.
ion coefficient between
en nominal columns as the
efficient and the Pearson's
airs of columns with
han a threshold are
reduced to only one. A word of caution: correlation is scale sensitive; therefore column normalization is required for a meaningful correlation comparison.

le Trees

so referred to as random
e selection in addition to
ne approach to
o generate a large and
trees against a target
attribute's usage statistics to
oset of features. Specifically, we can generate a large set (2000) of very shallow trees (2 levels), with each tree being trained on a small fraction (3) of the total number of attributes. 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

(https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

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 [illegible] the selected machine [illegible]

[illegible] the Backward Feature [illegible] only, progressively [illegible] the feature that produces the highest increase in performance. Both algorithms, Backward Feature Elimination and Forward Feature Construction, are quite time and computationally [illegible] only applicable to a data [illegible] low number of input

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

[illegible]-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-[illegible]SPOND)
[illegible]S://WWW.ANALYTICSVIDHYA.COM
[illegible]ONS-ASKED-AT-STARTUPS-IN-MACHINE-[illegible]6334)

After going through these question I feel I am at 10% of knowledge required to pursue 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 5:38 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.

AVBYTES (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/AVBYTES/)
CONTACT US (HTTPS://WWW.ANALYTICSVIDHYA.COM/CONTACT/)
WE ARE HIRING! (HTTPS://WWW.ANALYTICSVIDHYA.COM/ABOUT-ME/CAREER-ANALYTICS-VIDHYA/)
CAMPUS AMBASSADOR (HTTPS://WWW.ANALYTICSVIDHYA.COM/CAMPUS-AMBASSADOR/)
(https://datahack.analyticsvidhya.com
/data-engineering-talent-hunt-hackathon
CORPORATE (HTTPS://WWW.ANALYTICSVIDHYA.COM/CORPORATE/)

(https://datahack.analyticsvidhya.com
/the-young-optimization-crackerjack-contest
/?utm_source=AVblog_sidebottom)

Nikhil S says:
NOVEMBER 7, 2016 AT 4:44 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM
/BLOG/2016/09/40-INTERVIEW-QUESTIONS-ASKED-AT-STARTUPS-IN-MACHINE-LEARNING-DATA-SCIENCE/#COMMENT-116458)

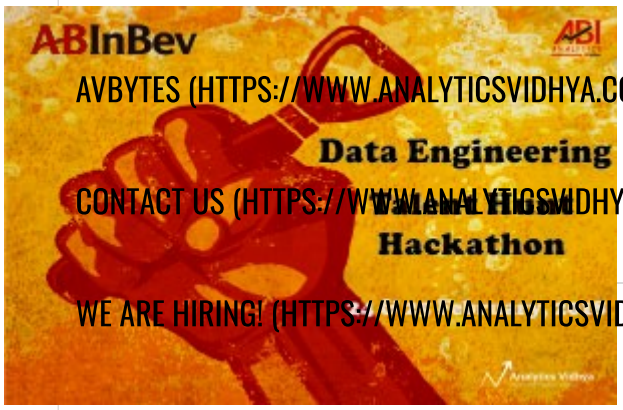
An awesome article for reference. Thanks a ton Manish sir
pdf format of this blog
also, thanks again

Siddharth says:
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://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)

(https://datahack.analyticsvidhya.com/data-engineering-talent-hunt-hackathon)

(https://datahack.analyticsvidhya.com/contest/the-young-optimization-crackerjack-contest/?utm_source=AVblog_sidebottom)