✕

-

(http://play.google.com/store/apps/details?id=com.analyticsvidhya.android)
Get FLAT 40% OFF on Introduction to Data Science Course | Use COUPON CODE: **DATA40** | Offer Ends 29th Aug 23:59
hrs IST | **Buy Now** (https://trainings.analyticsvidhya.com/courses?
utm_source=blog&utm_medium=web&utm_campaign=DATADAY33)

≡

(https://www.analyticsvidhya.com/blog/)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?
utm_source=AVTopBanner60off)

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

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

ANALYTICS VIDHYA CONTENT TEAM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/AVCONTENTTEAM/), SEPTEMBER 16, 2016

Introduction

Careful! These question can make you think THRICE!

Machine learning and data science are being looked as the drivers of the next industrial revolution happening in the world today. This also means that there are numerous exciting startups (https://www.analyticsvidhya.com/blog/2016/09/18-data-science-iot-startups-y-combinator-summer-2016/) looking for data scientists. What could be a better start for your aspiring career!

However, still, 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.

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

Interview Questions on Machine Learning

Q1. You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked 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 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-analysis-python/) and pick the components which can explain the maximum variance in the data set.
5. Using online learning algorithms like Vowpal Wabbit (available in Python) is a possible option.
6. Building a linear model using Stochastic Gradient Descent is also helpful.
7. We can also apply our business understanding to estimate which all predictors can impact the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in significant loss of information.

Note: For point 4 & 5, make sure you read about online learning algorithms (https://www.analyticsvidhya.com/blog/2015/01/introduction-online-machine-learning-simplified-2/) & Stochastic Gradient Descent (https://www.quora.com/Whats-the-difference-between-gradient-descent-and-stochastic-gradient-descent). These are advanced methods.

(https://datahack.analyticsvidhya.com/contest/datafest-closing-ceremony/)

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.

Know more: PCA (https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/).

Q3. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

Answer: This question has enough hints for you to start thinking! Since, the data is spread across 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 mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

Q4. You are given a data set on cancer detection. You've build a classification model and achieved 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 following steps:

1. We can use undersampling, oversampling or SMOTE to make the data balanced.
2. We can alter the prediction threshold value by doing probability caliberation (https://www.analyticsvidhya.com/blog/2016/07/platt-scaling-isotonic-regression-minimize-logloss-error/) and finding a optimal threshold using AUC-ROC curve.
3. We can assign weight to classes such that the minority classes gets larger weight.
4. We can also use anomaly detection.

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

Q5. Why is naive Bayes so 'naive' ?

Answer: naive Bayes is so 'naive' because it assumes that all of the features in a data set are equally important and independent. As we know, these assumption are rarely true in real world scenario.

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 probability that the word 'FREE' is used in previous spam message is likelihood. Marginal likelihood is, the probability that the word 'FREE' is used in any message.

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 decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?

Answer: Time series data is known 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 failed to provide robust predictions because it couldn't map the linear relationship as good as a regression model did. Therefore, we learned that, a linear regression model can provide robust prediction given the data set satisfies its <u>linearity assumptions (https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/)</u>.

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 three things:

1. There exist a pattern.
2. You cannot solve it mathematically (even by writing exponential equations).
3. You have data on it.

Always look for these three factors to decide if machine learning is a tool to solve a particular problem.

Q9. You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?

Answer: Low bias occurs when the model's predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results.

In such situations, we can use bagging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling. 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.

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?

Answer: Chances are, you might be tempted to say No, but that would be incorrect. Discarding correlated variables have a substantial effect on PCA because, in presence of correlated variables, the variance explained by a particular component gets inflated.

For example: You have 3 variables in a data set, of which 2 are correlated. If you run PCA on this data set, the first principal component would exhibit twice the variance than it would exhibit with uncorrelated variables. Also, adding correlated variables lets PCA put more importance on those variable, which is misleading.

Q11. After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models 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.

Q12. How is kNN different from kmeans clustering?

Answer: Don't get mislead by 'k' in their names. You should know that the fundamental difference between both these algorithms is, kmeans is unsupervised in nature and kNN is supervised in nature. kmeans is a clustering algorithm. kNN is a classification (or regression) algorithm.

kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.

kNN algorithm tries to classify an unlabeled observation based on its k (can be any number) surrounding neighbors. It is also known as lazy learner because it involves minimal training of model. Hence, it doesn't use training data to make generalization on unseen data set.

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 R² isn't as good as you wanted. For improvement, your remove the intercept term, your model R² becomes 0.8 from 0.3. Is it possible? How?

Answer: Yes, it is possible. We need to understand the significance of intercept term in a regression model. The intercept term shows model prediction without any independent variable i.e. mean prediction. The formula of $R^2 = 1 - \sum(y - y')^2/\sum(y - ymean)^2$ where y' is predicted value.

When intercept term is present, R² value evaluates your model wrt. to the mean model. In absence of intercept term (`ymean`), the model can make no such evaluation, with large denominator, `∑(y - y´)²/∑(y)²` equation's value becomes smaller than actual, resulting in higher R².

Q15. After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he's true? Without losing any information, can you still build a better model?

Answer: To check multicollinearity, 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/)

Q16. When is Ridge regression favorable over Lasso regression?

Answer: You can quote ISLR's authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.

Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.

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.

Know more: Causation and Correlation (https://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/)

Q18. While working on a data set, how do you select important variables? Explain your methods.

Answer: Following are the methods of variable selection you can use:

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

Answer: Yes, we can use ANCOVA (analysis of covariance) technique to capture association between continuous and categorical variables.

Q21. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?

Answer: The fundamental difference is, random forest uses bagging technique to make predictions. GBM uses boosting techniques to make predictions.

In bagging technique, a data set is divided into n samples using randomized sampling. Then, using a single learning algorithm a model is build on all samples. Later, the resultant predictions are combined using voting or averaging. Bagging is done is parallel. In boosting, after the first round of predictions, the algorithm weighs misclassified predictions higher, such that they can be corrected in the succeeding 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 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?

Answer: A classification trees makes decision based on Gini Index and Node Entropy. In simple words, the tree algorithm find the best possible feature which can divide the data set into purest possible children nodes.

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. We can calculate 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 of that split

Entropy is the measure of impurity as given by (for binary class):

$$\text{Entropy} = -p \log_2 p - q \log_2 q$$

Here p and q is probability of success and failure respectively in that node. Entropy is zero when a node is homogeneous. It is maximum when a both the classes are present in a node at 50% − 50%. Lower entropy 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.

Q24. You've got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?

Answer: In such high dimensional data sets, we can't use classical regression techniques, since their assumptions tend to fail. When p > n, we can no longer calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.

To combat this situation, we can use penalized regression methods like lasso, LARS, ridge which can shrink the coefficients to reduce variance. Precisely, ridge regression works best in situations where the least square estimates have higher variance.

Among other methods include subset regression, forward stepwise regression.

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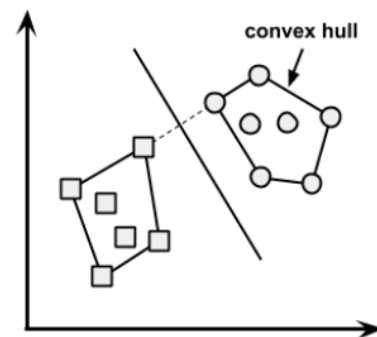
Q26. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn't. How ?

Answer: Don't get baffled at this question. It's a simple question asking the difference between the two.

Using one hot encoding, the dimensionality (a.k.a features) in a data set get increased because it creates a new variable for each level present in categorical variables. For example: let's say we have a variable 'color'. The variable has 3 levels namely Red, Blue and Green. One hot encoding 'color' variable will generate three new variables as `Color.Red`, `Color.Blue` and `Color.Green` containing 0 and 1 value.

In label encoding, the levels of a categorical variables gets encoded as 0 and 1, so no new variable is created. Label encoding is majorly used for binary variables.

Q27. What cross validation technique would you use on time series data set? Is it k-fold or LOOCV?

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, 8 variables have missing values higher than 30%. How will you deal with them?

Answer: We can deal with them in the following ways:

1. Assign a unique category to missing values, who knows the missing values might decipher some trend
2. We can remove them blatantly.
3. Or, we can sensibly check their distribution with the target variable, and if found any pattern we'll keep those missing values and assign them a new category while removing others.

29. 'People who bought this, also bought…' recommendations seen on amazon is a result of which algorithm?

Answer: The basic idea for this kind of recommendation engine comes from collaborative filtering.

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

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

Q31. **You are working on a classification problem. For validation purposes, you've randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?**

Answer: In case of classification problem, we should always use stratified sampling instead of random sampling. A random sampling doesn't takes into consideration the proportion of target classes. On the contrary, stratified sampling helps to maintain the distribution of target variable in the resultant distributed samples also.

Q32. You have been asked to evaluate a regression model based on R², adjusted R² 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 R² as opposed to R² to evaluate model fit because R² increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R² would only increase if an additional variable improves the accuracy of model, otherwise stays same. It is difficult to commit a general threshold value for adjusted R² because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R² and still provide fairly good predictions, as compared to a stock market data where lower adjusted R² implies that model is not good.

Q33. In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance ?

Answer: We don't use manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, euclidean metric can be used in any space to calculate distance. Since, the data points can be present in any dimension, euclidean distance is a more viable option.

Example: Think of a chess board, the movement made by a bishop or a rook is calculated by manhattan distance because of their respective vertical & horizontal 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.

Q35. I know that a linear regression model is generally evaluated using Adjusted R² or F value. How would you evaluate a logistic regression model?

Answer: We can use the following methods:

1. Since logistic regression is used to predict probabilities, we can use AUC-ROC curve along with confusion matrix to determine its performance.
2. Also, the analogous metric of adjusted R² in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.
3. Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

Know more: Logistic Regression (https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/)

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.

In short, there is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

Q37. Do you suggest that treating a categorical variable as continuous variable would result in a better predictive model?

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

Q38. When does regularization becomes necessary in Machine Learning?

Answer: Regularization becomes necessary when the model begins to ovefit / underfit. This technique 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 :

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

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

(https://www.analyticsvidhya.com/wp-content/uploads/2015/07/error-of-a-model.png)

Bias error is useful to quantify how much on an average are the predicted values different from the actual value. A high bias error means we have a under-performing model which keeps on missing important trends. **Variance** on the other side quantifies how are the prediction made on same observation different from each other. A high variance model will over-fit on your training population and perform badly on any observation beyond training.

Q40. OLS is to linear regression. Maximum likelihood is to logistic regression. Explain the statement.

Answer: OLS and Maximum likelihood are the methods used by the respective regression methods to approximate the unknown parameter (coefficient) value. 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 types of questions asked at startups in machine learning. I'm sure these questions would leave you curious enough to do deeper topic research at your end. If you are planning for it, that's a good sign.

Did you like reading this article? Have you appeared in any startup interview recently for data scientist profile? Do share your experience in comments below. I'd love to know your experience.

Looking for a job in analytics? Check out currently hiring jobs (https://www.analyticsvidhya.com/jobs/) in machine learning and data science.

You can also read this article on Analytics Vidhya's Android APP

GET IT ON Google Play

(//play.google.com/store/apps/details?
id=com.analyticsvidhya.android&utm_source=blog_article&utm_campaign=blog&pcampaignid=MKT-Other-global-all-co-prtnr-py-PartBadge-Mar2515-1)

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/)

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/)

●●●

PREVIOUS ARTICLE

AWS / Cloud Engineer – Pune (4+ Years of Experience)

(https://www.analyticsvidhya.com/blog/2016/09/aws-cloud-engineer-pune-4-years-of-experience/)

(https://www.analyticsvidhya.com/blog/author/avcontentteam/)

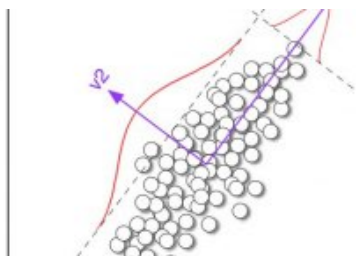**Analytics Vidhya Content Team (Https://Www.Analyticsvidhya.Com/Blog/Author/Avcontentteam/)**

Analytics Vidhya Content team

RELATED ARTICLES

ANALYTICS VIDHYA CONTENT TEAM (HT...   **SUNIL RAY (HTTPS://WWW.ANALYTICSVI...**   **ANKIT GUPTA (HTTPS://WWW.ANALYTIC...**

(https://www.analyticsvidhya.com/
predictions-test-data-principal-
component-analysis/)

**Making Predictions on Test Data
after Principal Component
Analysis in R
(https://www.analyticsvidhya.com/
blog/2016/07/making-predictions-
test-data-principal-component-
analysis/)**

(https://www.analyticsvidhya.com/
graphlab-python/)

**Tutorial – Getting Started with
GraphLab For Machine Learning
in Python
(https://www.analyticsvidhya.com/
blog/2015/12/started-graphlab-
python/)**

(https://www.analyticsvidhya.com/
skilltest/)

**25 Questions to test a Data
Scientist on Support Vector
Machines
(https://www.analyticsvidhya.com/
blog/2017/10/svm-skilltest/)**
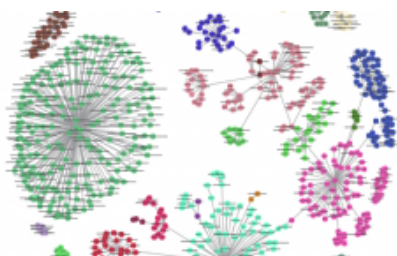
AISHWARYA SINGH (HTTPS://WWW.ANA...

(https://www.analyticsvidhya.com/
nearest-neighbor-introduction-
regression-python/)

**A Practical Introduction to K-
Nearest Neighbors Algorithm for
Regression (with Python code)
(https://www.analyticsvidhya.com/
blog/2018/08/k-nearest-neighbor-
introduction-regression-python/)**

SHIVAM BANSAL (HTTPS://WWW.ANALY...

(https://www.analyticsvidhya.com/
language-processing-made-easy-
using-spacy-%e2%80%8bin-
python/)

**Natural Language Processing
Made Easy – using SpaCy (in
Python)
(https://www.analyticsvidhya.com/
blog/2017/04/natural-language-
processing-made-easy-using-
spacy-%e2%80%8bin-python/)**

PRANAV DAR (HTTPS://WWW.ANALYTIC...

(https://www.analyticsvidhya.com/
7-github-repositories-march-
2018/)

**Top 7 Data Science & Machine
Learning GitHub Repositories in
March 2018
(https://www.analyticsvidhya.com/
blog/2018/04/top-7-github-
repositories-march-2018/)**

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

33 COMMENTS

KAVITHA   **Reply**

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/#comment-116110)

thank you so much manish

ANALYTICS VIDHYA CONTENT TEAM   **Reply**

September 16, 2016 at 9:17 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116122)

Hi Kavitha, I hope these questions help you to prepare for forthcoming interview rounds. All the best.

GIANNI   **Reply**

September 16, 2016 at 7:12 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116111)

Thank you Manish, very helpfull to face on the true reality that a long long journey wait me 🙂

ANALYTICS VIDHYA CONTENT TEAM   **Reply**

September 16, 2016 at 9:15 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116119)

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

ANALYTICS VIDHYA CONTENT TEAM   **Reply**

September 16, 2016 at 9:17 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   **Reply**

September 16, 2016 at 7:46 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 !
I am sure it will be very useful to the budding data scientists whether they face start-ups or established firms.

ANALYTICS VIDHYA CONTENT TEAM   **Reply**

September 16, 2016 at 9:20 am (https://www.analyticsvidhya.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 be asked anywhere. But, with growth in machine learning startups, facing off ML algorithm related question have higher chances, though I have laid emphasis on statistical modeling as well.

SRINIVAS   **Reply**

September 16, 2016 at 9:05 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116117)

Thank you Manish.Helpful for Beginners like me.

ANALYTICS VIDHYA CONTENT TEAM   **Reply**

September 16, 2016 at 9:14 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116118)

Welcome 🙂

CHIBOLE   **Reply**

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.

CHIBOLE   **Reply**

September 16, 2016 at 9:38 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116126)

* stastics = Statistics

ANALYTICS VIDHYA CONTENT TEAM   **Reply**

September 16, 2016 at 9:39 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116127)

Hi Chibole,
True, statistics in an inevitable part of machine learning. One needs to understand statistical concepts in order to master machine learning.

CHIBOLE   **Reply**

September 16, 2016 at 9:55 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-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.

ANALYTICS VIDHYA CONTENT TEAM   **Reply**

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.

CHIBOLE   **Reply**

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

KARTHIKEYAN SANKARAN (HTTP://BIT.LY/31KART8)   **Reply**
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)

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

ANALYTICS VIDHYA CONTENT TEAM   **Reply**
September 19, 2016 at 12:36 pm (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116209)

Most Welcome ! 🙂

NICOLA   **Reply**
September 16, 2016 at 7:22 pm (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116145)

Hi, really an interesting collection of answers. From a merely statistical point of view there are some imprecisions (e.g. Q40), but it is surely useful for job interviews in startups and bigger firms.

ANALYTICS VIDHYA CONTENT TEAM   **Reply**
September 17, 2016 at 4:11 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?

RAJU (HTTP://NONE)   **Reply**
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

RAJU (HTTP://NONE)   **Reply**

September 17, 2016 at 1:53 am (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....

AMIT SRIVASTAVA   **Reply**
September 17, 2016 at 7:07 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116154)

Great article. It will help in understanding which topics to focus on for interview purposes.

ANALYTICS VIDHYA CONTENT TEAM   **Reply**
September 19, 2016 at 12:35 pm (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116208)

Hi Amit,
Thanks for your encouraging words! The purpose of this article is to help beginners understand the tricky side of ML interviews.

CHINMAYA MISHRA   **Reply**
September 17, 2016 at 1:13 pm (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116161)

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   **Reply**
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)

Hi Manish,

Great Job! It is a very good collection of interview questions on machine learning. It will be a great help if you can also publish a similar article on statistics. Thanks in advance

---

ANALYTICS VIDHYA CONTENT TEAM
Reply
September 19, 2016 at 12:34 pm (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.

KARTHI V (HTTPS://GITHUB.COM/VKARTHI46)
Reply
September 20, 2016 at 3:32 pm (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116253)

Hi Manish,

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.

3.High Correlation Filter.
Data columns with very similar trends are also likely to carry very similar information. In this case, only one of them will suffice to feed the machine learning model. Here we calculate the correlation coefficient between numerical columns and between nominal columns as the Pearson's Product Moment Coefficient and the Pearson's chi square value respectively. Pairs of columns with correlation coefficient higher than 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.

4.Random Forests / Ensemble Trees

Decision Tree Ensembles, also referred to as random forests, are useful for feature selection in addition to being effective classifiers. One approach to dimensionality reduction is to generate a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset 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 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 classification performance with the selected machine learning algorithm.

6.Forward Feature Construction.

This is the inverse process to the Backward Feature Elimination. We start with 1 feature only, progressively adding 1 feature at a time, i.e. the feature that produces the highest increase in performance. Both algorithms, Backward Feature Elimination and Forward Feature Construction, are quite time and computationally expensive. They are practically only applicable to a data set with an already relatively low number of input columns.

RAMIT PANDEY

Reply

September 22, 2016 at 7:28 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116334)

Hi Manish ,
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

---

RAHUL JADHAV

Reply

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.

NIKHILS

Reply

September 25, 2016 at 3:44 pm (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 for the share. Please share the pdf format of this blog post if possible. Have also taken note of Karthi's input!

VINAYA

Reply

October 5, 2016 at 9:19 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116815)

ty manish…its an awsm reference…plz upload pdf format also…thanks again

PRASANNA

Reply

October 7, 2016 at 5:32 am (https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/#comment-116907)

Great set of questions Manish. BTW.. I believe the expressions for bias and variance in question 39 is incorrect. I believe the brackets are messed. Following gives the correct expressions.
https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff
(https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff)

SIDD

Reply

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.

TOP ANALYTICS VIDHYA USERS

Rank	Name	Points
1	Rohan Rao (https://datahack.analyticsvidhya.com/user/profile/Rohan_Rao)	8856
2	SRK (https://datahack.analyticsvidhya.com/user/profile/SRK)	8817
3	aayushmnit (https://datahack.analyticsvidhya.com/user/profile/aayushmnit)	7739
4	mark12 (https://datahack.analyticsvidhya.com/user/profile/mark12)	6798
5	sonny (https://datahack.analyticsvidhya.com/user/profile/sonny)	5947

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

 (https://trainings.analyticsvidhya.com/courses/course-

v1:AnalyticsVidhya+AIBL101+AIBL_T1/about?utm_source=AVBannerbelowcommunity)

 (https://trainings.analyticsvidhya.com/courses/course-

v1:AnalyticsVidhya+ITE001+2018_T1/about?utm_source=AVBlogbanner2)

POPULAR POSTS

24 Ultimate Data Science Projects To Boost Your Knowledge and Skills (& can be accessed freely)
(https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-
and-skills/)

A Complete Tutorial to Learn Data Science with Python from Scratch
(https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)

Essentials of Machine Learning Algorithms (with Python and R Codes)
(https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)

7 Types of Regression Techniques you should know!
(https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/)

20 Challenging Job Interview Puzzles which every analyst should solve atleast once
(https://www.analyticsvidhya.com/blog/2016/07/20-challenging-job-interview-puzzles-which-every-analyst-
solve-atleast/)

Understanding Support Vector Machine algorithm from examples (along with code)
(https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)

A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)
(https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/)

A comprehensive beginner's guide to create a Time Series Forecast (with Codes in Python)
(https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/)

RECENT POSTS

The Ultimate Guide to 12 Dimensionality Reduction Techniques (with Python codes) (https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/)

AUGUST 27, 2018

The Ultimate Data Science and Machine Learning Blogathon – More than $2500 up for grabs! (https://www.analyticsvidhya.com/blog/2018/08/data-science-machine-learning-blogathon-lucrative-prizes-bonus/)

AUGUST 24, 2018

A Hands-On Guide to Automated Feature Engineering using Featuretools in Python (https://www.analyticsvidhya.com/blog/2018/08/guide-automated-feature-engineering-featuretools-python/)

AUGUST 22, 2018

A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code) (https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/)

AUGUST 22, 2018

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

Most

ANALYTICS VIDHYA

DATA SCIENTISTS

COMPANIES

JOIN OUR COMMUNITY :

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

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://trainings.analyticsvidhya.com/)

(https://www.facebook.com/Analytics (https://twitter.com/Analytics datahack-summit-2018/?

Hackathon (https://datahack.analyticsvidhya.com/about)

Hiring (https://www.facebook.com/Analytics (https://twitter.com/Analytics

Discussions (https://discuss.analyticsvidhya.com/)

Hackathons (https://datahack.analyticsvidhya.com/)

Followers   Followers

(https://www.facebook.com/Analytics (https://twitter.com/Analytics

Apply Jobs (https://www.analyticsvidhya.com/career-a/)

Advertising (https://www.analyticsvidhya.com/jobs/)

(https://plus.go (https://www.linkedin.com/company/analyticsvidhya.com/contact/)

Contact Us (https://www.analyticsvidhya.com/contact/)

Leaderboard (https://datahack.analyticsvidhya.com/users/)

Reach Us (https://www.analyticsvidhya.com/contact/)

2972   7513

utm_source=BlogStickyBanner1)

(https://plus.go (https://www.linkedin.com

Write for us (https://www.analyticsvidhya.com/about-me/write/)

Followers   Followers

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

(https://plus.go (https://www.linkedin.com

utm_source=StickyBanner2)

Subscribe to emailer   >

Privacy Policy (https://www.analyticsvidhya.com/privacy-policy/)

Don't have an account? Sign up (https

Terms of Use (https://www.analyticsvidhya.com/terms/)

Refund Policy (https://www.analyticsvidhya.com/refund-policy/)