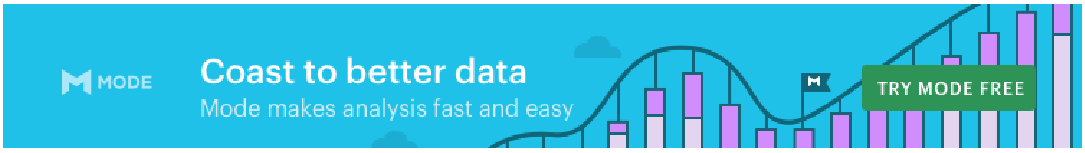
KDnuggets

Subscribe to KDnuggets News | | Contact

search KDnuggets Search

- SOFTWARE
- NEWS
- Top stories
- Opinions
- Tutorials
- JOBS
- Companies
- Courses
- Datasets
- EDUCATION
- Certificates
- Meetings
- Webinars



Coast to better data! Mode makes analysis fast and easy. Try Mode Free

KDnuggets Home » News » 2017 » Feb » Tutorials, Overviews » 17 More Must-Know Data Science Interview Questions and Answers (17:n07)

17 More Must-Know Data Science Interview Questions and Answers

Feb 2017 Gold KDnuggets Blog

Previous post
Next post

Like 210 Share 210 Share 582 Tweet G+1 14 Share 302

Tags: Anomaly Detection, Bias, Classification, Data Science, Donald Trump, Interview questions, Outliers, Overfitting, Variance

17 new must-know Data Science Interview questions and answers include lessons from failure to predict 2016 US Presidential election and Super Bowl LI comeback, understanding bias and variance, why fewer predictors might be better, and how to make a model more robust to outliers.

NEW YORK UNIVERSITY
NYU STERN
LEONARD N. STERN SCHOOL OF BUSINESS
FIND WISDOM IN DATA
APPLY NOW>>
BUSINESS ANALYTICS FOR EXECUTIVES
NYU MS in Business Analytics for Professionals - apply now

Pages: 1 2

By Gregory Piatetsky, KDnuggets.

Q4. Why might it be preferable to include fewer predictors over many?

Anmol Rajpurohit answers:

Here are a few reasons why it might be a better idea to have fewer predictor variables rather than having many of them:

Redundancy/Irrelevance:

If you are dealing with many predictor variables, then the chances are high that there are hidden relationships between some of them, leading to redundancy. Unless you identify and handle this redundancy (by selecting only the non-redundant predictor variables) in the early phase of data analysis, it can be a huge drag on your succeeding steps.

It is also likely that not all predictor variables are having a considerable impact on the dependent variable(s). You should make sure that the set of predictor variables you select to work on does not have any irrelevant ones – even if you know that data model will take care of them by giving them lower significance.

Note: Redundancy and Irrelevance are two different notions –a relevant feature can be redundant due to the presence of other relevant feature(s).

Overfitting:

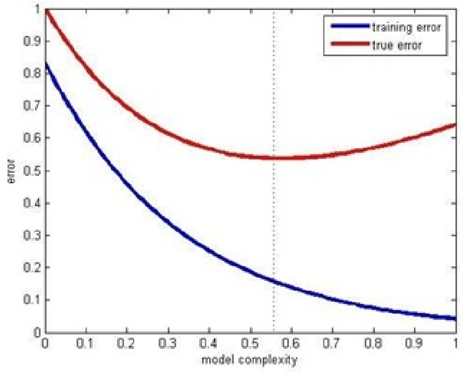
Even when you have a large number of predictor variables with no relationships between any of them, it would still be preferred to work with fewer predictors. The data models with large number of predictors (also referred to as complex models) often suffer from the problem of overfitting, in which case the data model performs great on training data, but performs poorly on test data.

Productivity:

Let's say you have a project where there are a large number of predictors and all of them are relevant (i.e. have measurable impact on the dependent variable). So, you would obviously want to work with all of them in order to have a data model with very high success rate. While this approach may sound very enticing, practical considerations (such of amount of data available, storage and compute resources, time taken for completion, etc.) make it nearly impossible.

Thus, even when you have a large number of relevant predictor variables, it is a good idea to work with fewer predictors (shortlisted through feature selection or developed through feature extraction). This is essentially similar to the Pareto principle, which states that for many events, roughly 80% of the effects come from 20% of the causes.

Focusing on those 20% most significant predictor variables will be of great help in building data models with considerable success rate in a reasonable time, without needing non-practical amount of data or other resources.



Training error & test error vs model complexity (Source: Posted on Quora by Sergul Aydore)

Understandability:

Models with fewer predictors are way easier to understand and explain. As the data science steps will be performed by humans and the results will be presented (and hopefully, used) by humans, it is important to consider the comprehensive ability of human brain. This is basically a trade-off – you are letting go of some potential benefits to your data model's success rate, while simultaneously making your data model easier to understand and optimize.

This factor is particularly important if at the end of your project you need to present your results to someone, who is interested in not just high success rate, but also in understanding what is happening "under the hood".

Q5. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

Prasad Pore answers:

Binary classification involves classifying the data into two groups, e.g. whether or not a customer buys a particular product or not (Yes/No), based on independent variables such as gender, age, location etc.

As the target variable is not continuous, binary classification model predicts the probability of a target variable to be Yes/No. To evaluate such a model, a metric called the confusion matrix is used, also called the classification or co-incidence matrix. With the help of a confusion matrix, we can calculate important performance measures:

1. True Positive Rate (TPR) or Hit Rate or Recall or Sensitivity = TP / (TP + FN)
2. False Positive Rate(FPR) or False Alarm Rate = 1 - Specificity = 1 - (TN / (TN + FP))
3. Accuracy = (TP + TN) / (TP + TN + FP + FN)
4. Error Rate = 1 – accuracy or (FP + FN) / (TP + TN + FP + FN)
5. Precision = TP / (TP + FP)
6. F-measure: 2 / ((1 / Precision) + (1 / Recall))
7. ROC (Receiver Operating Characteristics) = plot of FPR vs TPR
8. AUC (Area Under the Curve)
9. Kappa statistics

You can find more details about these measures here: The Best Metric to Measure Accuracy of Classification Models.

All of these measures should be used with domain skills and balanced, as, for example, if you only get a higher TPR in predicting patients who don't have cancer, it will not help at all in diagnosing cancer.

In the same example of cancer diagnosis data, if only 2% or less of the patients have cancer, then this would be a case of class imbalance, as the percentage of cancer patients is very small compared to rest of the population. There are main 2 approaches to handle this issue:

1. **Use of a cost function**: In this approach, a cost associated with misclassifying data is evaluated with the help of a cost matrix (similar to the confusion matrix, but more concerned with False Positives and False Negatives). The main aim is to reduce the cost of misclassifying. The cost of a False Negative is always more than the cost of a False Positive. e.g. wrongly predicting a cancer patient to be cancer-free is more dangerous than wrongly predicting a cancer-free patient to have cancer.

Total Cost = Cost of FN * Count of FN + Cost of FP * Count of FP

2. **Use of different sampling methods**: In this approach, you can use over-sampling, under-sampling, or hybrid sampling. In over-sampling, minority class observations are replicated to balance the data. Replication of observations leading to overfitting, causing good accuracy in training but less accuracy in unseen data. In under-sampling, the majority class observations are removed causing loss of information. It is helpful in reducing processing time and storage, but only useful if you have a large data set.

Find more about class imbalance here.

If there are multiple classes in the target variable, then a confusion matrix of dimensions equal to the number of classes is formed, and all performance measures can be calculated for each of the classes. This is called a multiclass confusion matrix. e.g. there are 3 classes X, Y, Z in the response variable, so recall for each class will be calculated as below:

Recall_X = TP_X/(TP_X+FN_X)

Recall_Y = TP_Y/(TP_Y+FN_Y)

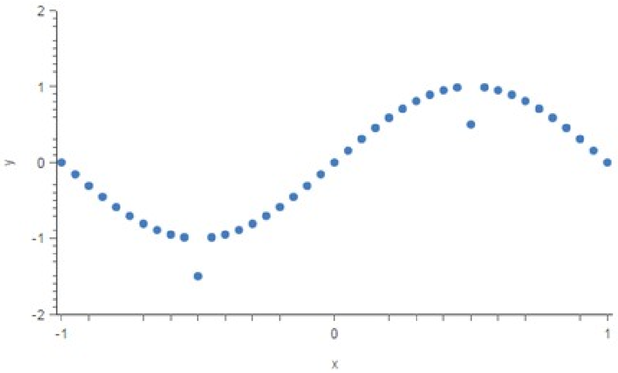
Recall_Z = TP_Z/(TP_Z+FN_Z)

Q6. What are some ways I can make my model more robust to outliers?

Thuy Pham answers:

There are several ways to make a model more robust to outliers, from different points of view (data preparation or model building). **An outlier** in the question and answer is assumed being unwanted, unexpected, or a must-be-wrong value to the human's knowledge so far (e.g. no one is 200 years old) rather than a rare event which is possible but rare.

Outliers are usually defined in relation to the distribution. Thus outliers could be removed in the pre-processing step (before any learning step), by using standard deviations (for normality) or interquartile ranges (for not normal/unknown) as threshold levels.



Outliers. Image source

Moreover, **data transformation** (e.g. log transformation) may help if data have a noticeable tail. When outliers related to the sensitivity of the collecting instrument which may not precisely record small values, **Winsorization** may be useful. This type of transformation (named after Charles P. Winsor (1895–1951)) has the same effect as clipping signals (i.e. replaces extreme data values with less extreme values). Another option to reduce the influence of outliers is using **mean absolute difference** rather mean squared error.

For model building, some models are resistant to outliers (e.g. tree-based approaches) or non-parametric tests. Similar to the median effect, tree models divide each node into two in each split. Thus, at each split, all data points in a bucket could be equally treated regardless of extreme values they may have. The study [Pham 2016] proposed a detection model that incorporates interquartile information of data to predict outliers of the data.

References:

[Pham 2016] T. T. Pham, C. Thamrin, P. D. Robinson, and P. H. W. Leong. Respiratory artefact removal in forced oscillation measurements: A machine learning approach. IEEE Transactions on Biomedical Engineering, 2016.

This Quora answer contains further information.

Here is part 2 and part 3 with more answers.

9 Comments **KDnuggets** 1 **Login**

♡ **Recommend** **Share** **Sort by Best**

Join the discussion…

Philip Leitch • 6 days ago

I disagree with the outlier answer. It is true for frequentist data analysis, but not for Baysean approaches.

Increasingly Bayesian approaches are being used in machine learning, and Bayesian approaches consider all data as significant, and therefore identifying "outliers" and removing them removes information from the analysis. The only true "outliers" are those that are not actually part of the sample group. For example, a machine that occasionally produces an incorrect reading is producing outliers, The incorrect reading would be a separate distribution (of reading errors), and therefore can be identified and removed.

Otherwise outliers are perfectly valid stochastic events and expected values from almost any distribution. In fact, lack of extreme values is an indication of a non-random process/sampling. I don't always stick to this rule - especially when performing frequentist analysis as outlined above - but removing outliers to fit the solution is generally a flawed approach.

For example, if you were analysing your expected return from lottery ticket, any random sampling would be highly unlikely to include the 1st place winner, yet if NO sampling can ever include a 1st place winner, either something is wrong with sample (rigged lottery) or outlier elimination has screened out the winner because the 1st place winner would accurately be identified as an outlier (but an expected and meaningful outlier). 1st place lottery winners are outliers, and are expected, and add information to the expected return from purchasing a lottery ticket.

^ | v • Reply • Share ›

Anna Olecka ➦ Philip Leitch • 2 days ago

I don't belive there is a contradiction here, as long as we all agree with Thuy Pham's assumption that the only outliers we want to "treat" are the unwanted data points, not possible in the real distribution. What we don't want, is remove mechanically the data points statisticians call "outliers", e.i. points distant from the distribution mean which Thuy Pham calls "rare events". These actually often form super predictive features. Case in point: in scoring businesses for a likelyhood of bankruptcy non payment of government taxes would be a very rare event, a "statistical outlier" in the distribution of all US businesses. Yet as a feature in a bankruptcy prediction model it will show 95% correlation with an upcoming bankruptcy, regardless of the modeling framework chosen

^ | v • Reply • Share ›

T Pham ➦ Philip Leitch • 3 days ago

Thanks Philip Leitch for your comment and sharing your knowledge. I would improve the clarify of the question and answer from your comment.

Frankly I m not want to dive in a debatable talk between frequentist and Baysian. It happens for long time and I tend to agree with the post of declaring it's over by Rafa Irizarry, Roger Peng, and Jeff Leek ("Imposing Bayesian or frequentists philosophy on us would be a disaster.")

However, I think the question and the answer really need a common definition of what is an "outlier"? Outliers I have been encountered mostly are unwanted, unexpected, a must-be-wrong value to the human's knowledge so far (e.g. no one is 200 years old) rather than a rare event which is possible but rare. They are different. The purpose of question is how a machine can ignore such kind of "not-resonable" values in data when machines are learning. Thus the example I gave is a case of "unwanted" data when the condition of recording changed. The researchers want to focus on only the data of a specific condition thus they need to remove that "outlier" of the data before doing further analysis steps.

^ | v • Reply • Share ›

Timo Mechsner • a month ago

I believe there is a mistake in your description of the confusion matrix: What you described as the False Positive Rate is actually the True Negative Rate, namely of all real negative data, how many did the model get. The False Positive Rate is the counterpart of the True Positive Rate, so 1 - TPR.
However, thank you for the nice article, I really like those "Questions you must know" things! :-)

^ | v • Reply • Share ›

Gregory Piatetsky Mod ➦ Timo Mechsner • a month ago

typo corrected: False Positive Rate(FPR) = 1 - Specificity = 1- (TN / (TN + FP))

^ | v • Reply • Share ›

Dimitrios Fotiadis ➦ Timo Mechsner • a month ago

My good friend Timo. (TN)/(TN+FP) = TNR the proportion of negatives correctly identified as negatives.

Please double-check .
and 1-TNR= FPR

^ | ⌄ • Reply • Share ›

Timo Mechsner ➦ Dimitrios Fotiadis • a month ago

This time it was only a typo: I meant the TNR and corrected it, thank you for checking back! ;-)
Now we got it. :-)

^ | ⌄ • Reply • Share ›

Dimitrios Fotiadis ➦ Timo Mechsner • a month ago

Timo,
FPR is actually not TNR, but rather its all negatives that were predicted to be positives. It is actually 1-TNR and not 1-TPR as you indicate. False positive means that you predict positive when in reality you have negative.Also FPR is not the counterpart of TPR. If by counterpart you imply complement , its wrong. TPR +FPR is not 1. Rather TPR+FNR=1 and FPR+TNR=1.

(TPR , FNR) are complementary and (FPR,TNR) are complementary.

^ | ⌄ • Reply • Share ›

Timo Mechsner ➦ Dimitrios Fotiadis • a month ago

You are right, Dimitrios, thank you for correction.
However TN / (TN + FP) is the TNR, not FPR as the article states in 5.2 :-)

^ | ⌄ • Reply • Share ›

✉ **Subscribe** **Add Disqus to your site Add Disqus Add** 🔒 **Privacy**

Pages: 1 2

◂ Previous post
Next post ▸

Top Stories Past 30 Days

Most Popular

1. The 10 Algorithms Machine Learning Engineers Need to Know
2. 7 More Steps to Mastering Machine Learning With Python
3. An Overview of Python Deep Learning Frameworks
4. Gartner 2017 Magic Quadrant for Data Science Platforms: gainers and losers
5. Every Intro to Data Science Course on the Internet, Ranked
6. 17 More Must-Know Data Science Interview Questions and Answers
7. 50 Companies Leading The AI Revolution, Detailed

Most Shared

1. 50 Companies Leading The AI Revolution, Detailed
2. An Overview of Python Deep Learning Frameworks
3. 7 More Steps to Mastering Machine Learning With Python
4. The Data Science Project Playbook
5. What Is Data Science, and What Does a Data Scientist Do?
6. What makes a good data visualization – a Data Scientist perspective
7. Getting Up Close and Personal with Algorithms

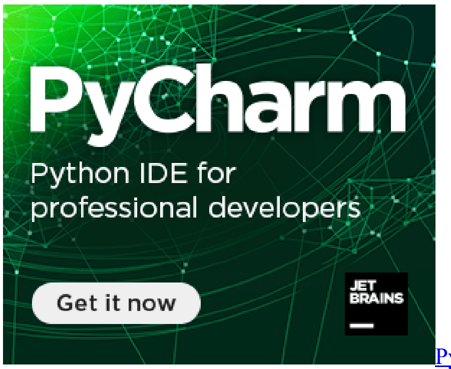
Latest News

- Live.xyz: Senior Data Scientist
- Webinar: Improve Your CLASSIFICATION with CAR...

- Top Stories, Mar 20-26: What Is Data Science,...
- From Big Data Platforms to Platform-less Mach...
- Turner: Sr. Operations Research Analyst / Dat...
- What is Structural Equation Modeling?

PENN STATE | ONLINE
Programs include Data Analytics and Business Analytics
▸ GET STARTED NOW – REQUEST A BROCHURE

Penn State Online: MS in Data Analytics, Business Analytics

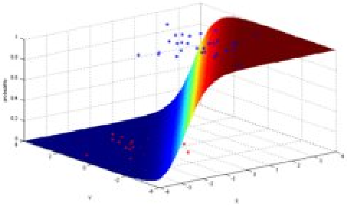


PyCharm: Python IDE for Developers
Get it now

**Top Stories
Last Week**

Most Popular

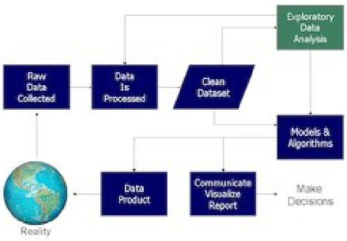
1. ▲ **The 10 Algorithms Machine Learning Engineers Need to Know**



2. NEW **50 Companies Leading The AI Revolution, Detailed**
3. ▲ **17 More Must-Know Data Science Interview Questions and Answers**
4. NEW **Getting Started with Deep Learning**
5. NEW **The Most Underutilized Function in SQL**
6. NEW **Analytics 101: Comparing KPIs**
7. NEW **How to think like a data scientist to become one**

Most Shared

1. **What Is Data Science, and What Does a Data Scientist Do?**



2. **Getting Up Close and Personal with Algorithms**
3. **Getting Started with Deep Learning**

4. **Key Takeaways from Strata + Hadoop World 2017 San Jose**
5. **How to think like a data scientist to become one**
6. **The Most Underutilized Function in SQL**
7. **Data Scientists Might Have It Made For 2017**

More Recent Stories

- What is Structural Equation Modeling?
- Last Chance: Big Data San Francisco, April 19 & 20
- Analytics and Machine Learning training in Q2
- Turner: Advisor Analytics Architect
- Apple: Software Engineer – Local Search
- PeachIH: Data Scientist, Machine Learning Engineer
- Getting Started with Deep Learning
- IBM Chief Data Officer Strategy Summit, March 29-30, San Franc...
- Key Takeaways from Strata + Hadoop World 2017 San Jose, Day 1
- Unsupervised Investments: A Comprehensive Guide to AI Investors
- Make Analytics Pay with Live Immersive Training
- How to think like a data scientist to become one
- What Is Data Science, and What Does a Data Scientist Do?
- Apache Big Data: top projects, people and technologies –...
- Top tweets, Mar 15-21: Reverse-engineering a $500M AI compa...
- Eindhoven University of Technology: Full Professor Database Te...
- Webinar: Predictive Analytics: Failure to Launch – Apr 13
- Deep Learning Summit & Deep Learning in Healthcare Summit...
- What Top Firms Ask: 100+ Data Science Interview Questions
- Why A/B Testers Have The Best Jobs In Tech

KDnuggets Home » News » 2017 » Feb » Tutorials, Overviews » 17 More Must-Know Data Science Interview Questions and Answers (17:n07)

© 2017 KDnuggets. About KDnuggets

Subscribe to KDnuggets News

X