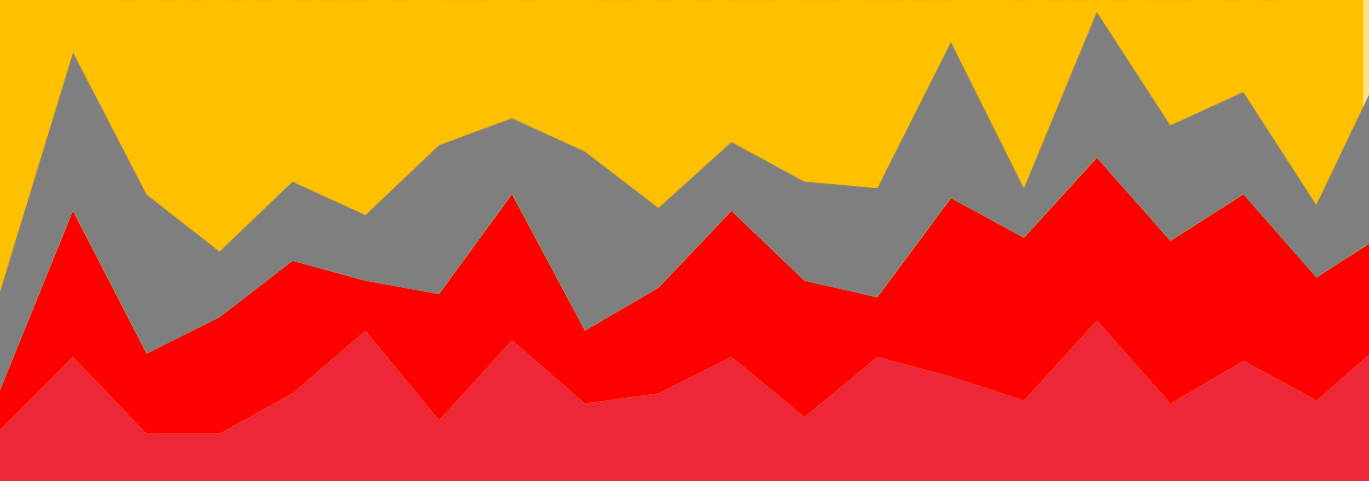
Interview Preparation Document

15

DATA ANALYTICS QUESTIONS EVERY ANALYST SHOULD KNOW



SUNIL KAPPAL | DATA DOJO

DATA ANALYTICS INTERVIEW QUESTIONS

What is R2? What are some other metrics that could be better than R2 and why?

- goodness of fit measure. variance explained by the regression / total variance
- the more predictors you add the higher R^2 becomes.
  - hence use adjusted R^2 which adjusts for the degrees of freedom
  - or train error metrics

What is the curse of dimensionality?

- High dimensionality makes clustering hard, because having lots of dimensions means that everything is "far away" from each other.
- For example, to cover a fraction of the volume of the data we need to capture a very wide range for each variable as the number of variables increases
- All samples are close to the edge of the sample. And this is a bad news because prediction is much more difficult near the edges of the training sample.
- The sampling density decreases exponentially as p increases and hence the data becomes much more sparse without significantly more data.
- We should conduct PCA to reduce dimensionality

Let's say you're given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?

- PCA

DATA DOJO

DATA ANALYTICS INTERVIEW QUESTIONS

Is more data always better?

- **Statistically,**
  - It depends on the quality of your data, for example, if your data is biased, just getting more data won't help.
  - It depends on your model. If your model suffers from high bias, getting more data won't improve your test results beyond a point. You'd need to add more features, etc.

- **Practically,**
  - Also there's a tradeoff between having more data and the additional storage, computational power, memory it requires. Hence, always think about the cost of having more data.

What are advantages of plotting your data before performing analysis?

- Data sets have errors. You won't find them all but you might find some. That 212 year old man. That 9 foot tall woman.

- Variables can have skewness, outliers etc. Then the arithmetic mean might not be useful. Which means the standard deviation isn't useful.

- Variables can be multimodal! If a variable is multimodal then anything based on its mean or median is going to be suspect.

DATA DOJO

How can you make sure that you don't analyze something that ends up meaningless?

- Proper exploratory data analysis.
- In every data analysis task, there's the exploratory phase where you're just graphing things, testing things on small sets of the data, summarizing simple statistics, and getting rough ideas of what hypotheses you might want to pursue further.
- Then there's the exploitatory phase, where you look deeply into a set of hypotheses.

The exploratory phase will generate lots of possible hypotheses, and the exploit

What is the role of trial and error in data analysis? What is the the role of making a hypothesis before diving in?

Data analysis is a repetition of setting up a new hypothesis and trying to refute the null hypothesis.

The scientific method is eminently inductive: we elaborate a hypothesis, test it and refute it or not. As a result, we come up with new hypotheses which are in turn tested and so on. This is an iterative process, as science always is.

How can you determine which features are the most important in your model?

Run the features though a Gradient Boosting Machine or Random Forest to generate plots of relative importance and information gain for each feature in the ensembles.

Look at the variables added in forward variable selection

DATA DOJO

How do you deal with some of your predictors being missing?

- Remove rows with missing values - This works well if 1) the values are missing randomly
- If you don't lose too much of the dataset after doing so.
- Build another predictive model to predict the missing values - This could be a whole project in itself, so simple techniques are usually used here.
- Use a model that can incorporate missing data - Like a random forest, or any tree-based method.

You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?

- Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.
- Leave the model as is, despite multicollinearity. The presence of multicollinearity doesn't affect the efficiency of extrapolating the fitted model to new data provided that the predictor variables follow the same pattern of multicollinearity in the new data as in the data on which the regression model is based.
- principal component regression

DATA DOJO

What is the main idea behind ensemble learning? If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse?

- The assumption is that a group of weak learners can be combined to form a strong learner.
- Hence the combined model is expected to perform better than an individual model.
- Assumptions:
  - average out biases
  - reduce variance
- Bagging works because some underlying learning algorithms are unstable: slightly different inputs leads to very different outputs. If you can take advantage of this instability by running multiple instances, it can be shown that the reduced instability leads to lower error. If you want to understand why, the original bagging paper(http://www.springerlink.com/cont...) has a section called "why bagging works"
- Boosting works because of the focus on better defining the "decision edge". By reweighting examples near the margin (the positive and negative examples) you get a reduced error (see http://citeseerx.ist.psu.edu/vie...)
- Use the outputs of your models as inputs to a meta-model.
- For example, if you're doing binary classification, you can use all the probability outputs of your individual models as inputs to a final logistic regression (or any model, really) that can combine the probability estimates.
- One very important point is to make sure that the output of your models are out-of-sample predictions. This means that the predicted value for any row in your dataframe should NOT depend on the actual value for that row.

DATA DOJO

DATA ANALYTICS INTERVIEW QUESTIONS

You have 5000 people that rank 10 sushis in terms of saltiness. How would you aggregate this data to estimate the true saltiness rank in each sushi?

- Some people would take the mean rank of each sushi. If I wanted something simple, I would use the median, since ranks are (strictly speaking) ordinal and not interval, so adding them is a bit risque (but people do it all the time and you probably won't be far wrong).

How would you come up with an algorithm to detect plagiarism in online content?

- Reduce the text to a more compact form (e.g. fingerprinting, bag of words) then compare those with other texts by calculating the similarity

You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters to include?

- KMeans
- choose a small value of k that still has a low SSE (elbow method)
- https://bl.ocks.org/rpgove/0060ff3b656618e9136b

DATA DOJO

Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach this problem?

- collaborative filtering

A certain metric is violating your expectations by going down or up more than you expect. How would you try to identify the cause of the change?

- breakdown the KPI's into what consists them and find where the change is
- then further breakdown that basic KPI by channel, user cluster, etc. and relate them with any campaigns, changes in user behaviors in that segment

You're a restaurant and are approached by Groupon to run a deal. What data would you ask from them in order to determine whether or not to do the deal?

- for similar restaurants (they should define similarity), average increase in revenue gain per coupon, average increase in customers per coupon, number of meals sold

DATA DOJO

Contact: Sunil Kappal | skappal7@gmail.com | datageek7@gmail.com |
Analytics Consultation | Trainings | Lean Six Sigma Implementation |
Project Management




DATA DOJO