Preparing for data science interview

<https://towardsdatascience.com/mastering-the-data-science-interview-15f9c0a558a7>

<https://www.edureka.co/blog/interview-questions/data-science-interview-questions/>

Review for Interviews:

1. Algorithms and data structure (leetcode)
   1. SQL
2. Data Science product questions
3. Data Science concepts
4. Behavioral interviews

Example: Chat app for patients who have questions for doctors

“How would you define user retention for the chat app”

“how would you go about identifying variables that are important for user retention”

“how would you improve user retention”

“A given category of an e-commerce marketplace, for instance jeans, is not doing well. How would you estimate if it is a demand or supply problem?”

“We made a change to our subscription offering adding new features. We expect this to increase subscription retention. How can we test if the change is successful?”

commerce\_user\_actions: date | sessionid | userid | event

Give me the avg number of sessions/user per day for the last 30 days

SELECT COUNT( DISTINCT cua.sessionid)/COUNT(DISTINCT cua.userid) as avg\_sessions\_per\_user

From commerce\_user\_actions cua

Where cua.date >= current\_date - 30

GROUP BY cua.date

time\_spent\_per\_session: date | sessionid | time\_spent\_sec

calculate the number of Daily active user for yesterday

SELECT COUNT(DISTINCT joined\_table.userid)

FROM

{

SELECT cua.date, cua.sessionid, cua.userid, cua.event, tsps.time\_spent\_sec

FROM commerce\_user\_actions cua

JOIN time\_spent\_per\_session tsps ON cua.date = tsps.date AND cua.sessionid = tsps.sessionid

WHERE tsps.time\_spent\_sec > 10

} joined\_table

WHERE joined\_table.date = CURRENT\_DATE - 1

“A Product Manager has this idea about implementing a call to action pop up to encourage users to post things they want to sell on fb marketplace. Why did you think this is a good idea, and how would you evaluate whether this is worth doing?”

how would you calculate Daily Active Users for marketplace?

**My answer was along the line of this:**

* This might be a good idea because then users might end up listing the items they want to sell on facebook, and as a result we get more data on them and we get increased engagement from users.
* I would run an A/B test to see if this is worth implementing. And the things that we want to measure:
  + Number of listing per user
  + Engagement (e.g. comments/likes/posts per user)
* Prior to the experiment I’d conduct an A/A test to ensure the assignment of control and treatment would work well.
* Then we need to calculate how long we need to do this for:
  + We can use the power calculation to do this
  + Assuming that we want a power of 80%, If we can get the expected treatment effect as well as the standard deviation for this new pop up, we can calculate the number of samples we need for the test.
  + The number of samples dictate for how long we should run the test for.

**A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a heads?**

What is your model selection process? How do you go about deciding what model to use. Maybe try a few different models and then comparing the AUC of them. Speed, performance, ability to give feature importance, what kind of features can it deal with. Logical not just intuition and experience.

ANOVA - Analysis of variance

GAM - Generalized Additive Models (rare)

MLE - Maximum Likelihood Estimation (rare)

Gower Distance to cluster dataset

[**https://quizlet.com/armand\_kok/folders/interview-prep/sets?x=1xqt&i=1v89o3**](https://quizlet.com/armand_kok/folders/interview-prep/sets?x=1xqt&i=1v89o3)

**Why are you interviewing right now at the end of the introduction. Keep it under 90 seconds.**

https://www.interviewquery.com/blog-the-google-data-scientist-interview

What do you think is important for a data scientist. -cmmunication, technical, thinking like a data scientist.

5 markov

Why is R^2 important? Why is it not perfect. If you care about making accurate predictions, R^2 might be important in ML. Statical stand point x vs Y. Relationship between two variables, if you’re just looking at relationship, not strength of correlation, it might not be very important. Depends on your usage. Diet vs cholesterol R^2 might not be great. If you include exercise, etc it might improve R^2 but at this point you just want to see correlation.

L1 vs L2 normalization = penalty in training for linear regressions. Used in the context of gradient descent.

Clustering methods. Hierarchical clustering. K means

How do you decide what K to use. Elbow method. Gap statistics (preferred). Intracluster correlation. Domain knowledge might play into it.

Advantages and disadvantages of different clustering methods. Speed, reasonable approximation, intuitive, each data assigned to 1 cluster

Cons maybe some cases you want more than 1 cluster for 1 data point. Sensitive to outliers, assumes spherical clustering. Number of k’s is not clear if you don’t have domain knowledge. GMM Gaussian clustering

Dbscan as a clustering method.

Prepare 3 to 4 data science stories. Situation action results.

A-B testing.

Leet code focus on SQL. Ansi-standard functions for SQL. PostgreSQL

The product manager interview.

confusion matrix.

Interesting story where you applied data science.

What is your biggest nightmare as a data scientist.

All things considered equal story. Failure? Interesting story? fears?

Logistic regression, Linear regression, svm, random forest, decision trees ..one vs the other..pros and cons..their cost function..regularization..drop out..neural network..

3 assumptions for unbiased estimate of average treatment effects in A/B testing and experiments. Excludability (placebo example for exclusion restriction), non-interference assumption, random sampling assumption.

Selection bias question. Sampling bias, time interval, attrition.

LinkedIn has tested a new UI with the goal to increase the number of likes per user. They test it by giving the new UI to a random subset of users

Test wins by 5% on the target metric. What do you expect to happen after the new UI is applied to all users? Will that metric actually go up by ~5%, more, or less? Assume there is no novelty effect here.

Lets say the the positive effect is real, the non-interference assumption is likely violated. This is because when the treatment group likes more, their friends and family in the control group might be affected by them to like more. The new UI now has an indirect effect on the control group. So you might be underestimating the actual impact of the new UI since the control group is over posting as well.

Imbalanced data set - often you model under performs, or you over guess the larger set, look at confusion matrix to see what’s happening. You can under sample the large set but …, you can over sample the small set but … Use informed oversampling and informed under sampling. Use Knn to generate samples that are similar to the your small set, then apply informed under sampling which is….

Product data science questions. Lets say you calculate a power and you decide you need 3 weeks of running the experiment to get the answer. You see the p value became statistically significant at week 1. Your boss tells you to stop running the experiment. What would you tell your boss to stop that from happening? When you’re calculating p-values as you accumulated data, the p value will oscillate up and down. The smaller the data size, the larger the oscillation, as you increase data size the p value will converge. You don’t have a large enough data set yet so your statistically significant p value might just be the low end of a big oscillation and isn’t reliable. stopping an a/b test early is like declaring the team who has the lead at half time the winner in a basketball/football/soccer game

Boosting

Out of bag error

Cons of K means -

Certain types of data distribution doesn’t work well

Hard to choose k

K means is sensitive to outliers - outliers might become it’s own cluster if you initialize near it.

Don’t like assigning observation to exactly 1 cluster - GMM covers this (read up on dbscan)

What do you consider when you’re choosing your model

What is your goal, what are your features and what does your data look like, categorical vs continuous, How many features/data, time constraints, interpretability. Model/Inference time for example k-nn you need to load all the data points into memory first before you make inference.

pros/cons of RF vs. SVM

SVM

<https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>

HOW DO YOU GO ABOUT DESIGNING A MODEL?

A diagram of different differences

Description automatically generated

Data science:

Present a current problem we have and see how they will approach solving the problem

If you hit a road block in a project, how would they approach solving.

General data science problems

In our line of work we build models to classify defects so we can keep track of how our yield is impacted by a certain defect and how this defect affects our reliability down the line. Let’s say we have a defect that we have explored and know some characteristic patterns from this defect, how would you approach building a model to classify this defect?

1. Contextualize the problem. What kind of classification do we need? A yes or no? Or some kind of scaling regressive classification that lets us know how bad the defect is? What is considered a success for this model? How do I measure this success?
2. Collect the data? How do I collect the data? Do I manually do so? Can I speed up this work? maybe build a more rudimentary model and work from there?
3. Explore the data to see what kind of inputs are going to go into your model, are they purely numeric or are there categorical variables? Are there outliers that you should get rid of? Any systematic issues that may confuse the model? For a regression model are there multiple parameters that already correlate to each other? Are there missing values?
4. Split the data into training and test set. How many times do I want to split the training and test set to evaluate my model?
5. Are you overfitting or undercutting the model? How can you tell?
6. What happens if your model is not performing as well as you have hoped? What would you do then?