Data Science Scenario Based Interview Questions

How to Approach Data Science Case Study Questions

1. Clarify: The first step is used to gather more information. More often than not, these case studies are designed to be confusing and vague!
2. Make Assumptions: The next step is where the thought process really starts to be outlined. With all the data provided, it's important to start investigating and discarding possible hypotheses.
3. Hypothesize and Propose a Solution: Now that a hypothesis is formed, gathering context is the next step towards fleshing out an answer. This is where the problem should be reframed given the new information gathered in the last two steps.
4. Provide Data Points and Analysis: Finally, providing data points and analysis involves choosing and prioritizing a main metric.
5. Consider Potential Pitfalls: Every case question tends to have multiple solutions. Therefore, you should absolutely consider and communicate any potential trade-offs of your chosen method.
6. effective communication: All the analysis in the world isn't going to help if interviewees cannot verbally work through and highlight their thought process within the case study

Note: The given answers are just sample answers. These are just hint for you to frame your own unique and wonderful answers in interviews.

1.How Do You Design an Email Spam Filter?

```
Building a spam filter involves the following process:
●   The email spam filter will be fed with thousands of emails
●   Each of these emails already has a label: 'spam' or 'not spam.'
●   The supervised machine learning algorithm will then determine which type of emails are
being marked as spam based on spam words like the lottery, free offer, no money, full
refund, etc.
●   The next time an email is about to hit your inbox, the spam filter will use
statistical analysis and algorithms like Decision Trees and SVM to determine how likely
the email is spam
●   If the likelihood is high, it will label it as spam, and the email won't hit your
inbox
●   Based on the accuracy of each model, we will use the algorithm with the highest
accuracy after testing all the models
```

2. At Netflix, we offer a subscription where customers can enroll for a 30 day free trial. After 30 days, customers will be automatically charged based on the package selected. We want to measure the success of acquiring new users through the free trial. How can we measure acquisition success and what metrics can we use to measure the success of the free trial?

```
One way we can frame the concept specifically to this problem is to think about
controllable inputs, external drivers, and then the observable output. Start with the
major goals of Netflix:

Acquiring new users to their subscription plan.
Decreasing churn and increasing retention.
How does that affect how Netflix might acquire new users?
```

Goal of this strategy is customer acquisition.
User Journey: The prospective customer enrolls for the 30 day free trial after providing his/her payment details. There can be 3 outcomes:
1. Once the free-trial period is over, the payment will be deducted from his account – the prospective customer becomes a customer
2. The prospective customer uses the free trail for 30days and then on the last day opts out.
3. The prospective customer opts out somewhere in the middle of the trial period.
The metrics which need to be measured to determine acquisition success are as follows:
1. # of prospective customers that converted to customers after 30 days.
2. # of customers who stayed with Netflix for atleast or more than 1 month after trial period ended.
3. Customer churn rate – number of customer who cancelled as soon as trial period is over + number of customers who used and opted out within a month after trial period.
----------The metrics which need to be measured to determine success of free trial:
4. Customer Acquisition Cost/Life Time Value ratio: for a successful business like Netflix it should be atleast 1:3
5. % of new customers who have been onboarded through this campaign daily, weekly, monthly – this determines the efficacy of the campaign

3. How would you measure the success of Facebook Groups?

Start by considering the key function of Facebook Groups. You could say that Groups are a way for users to connect with other users through a shared interest or real-life relationship.

Hint: With this in mind, what how could we use the goals of Facebook Groups to measure success?

Relevancy of content associated with a group can also be a metric to check group quality.

Posts having words similar to the word cloud associated with the group can be also another metric

Relevancy - Number of Associated Posts / Number of Posts.

Sometimes a group can become more interesting if the posts are in line to what the group is expected to be. This attracts more users and thus makes it more active

Other Metrics can be -

number of groups created per month
Churn on groups per day
Number of posts in the group per day
Number of reactions to posts/number of posts per day
Number of comments per day/number of posts
Number of groups closed
Avg Number posts reported per group
positive Sentiment of coments on posts

4.Suppose you're working as a data scientist at Facebook. How would you measure the success of private stories on Instagram, where only certain chosen friends can see the story?

While I think the approach to look at retention and recalculate the active comments per user by only focusing on the power/active users sounds good, it’s actually incomplete and will lead to wrong conclusions.

This is because when you restrict the denominator to mean active/retained users, you must make the groups in the numerator match what’s in the denominator.

We suggest:

numerator: all comments

denominator: all active users

This gives you all comments per active user, which doesn’t make any sense. You can’t just associate comments that active users didn’t make with active users. This will give you a false proxy for engagement since those active users didn’t make any of those comments. In short, if you are going to segment the denominator by a particular group, you must also segment the numerator with the same group for an apples-to-apples comparison between two variables that results in an appropriate proxy for an engagement metric.

Thus the appropriate new metric has to be:

numerator: all active comments from active users

denominator: all active users

This will now give you the correct percentage of active comments per active user, and will allow us to better chart what’s going on with the data. Taking Jay’s example, let’s say the active users are 7500, and we see that there are 50k comments. We then need to take the percentage of the 50k posts written by active users. If the proportion is growing, from say 50% in Feb, and 60% in March, we’ll get 25k comments and 36k comments from active users for both months respectively.

What does this scenario tell us as a whole? It means the following:

While total users are growing and retention is slowing down, active users represent the majority of activity in the comments section. This means that engagement from new users are going down, which might lead to issues later on.

So in short, if we’re going to change the denominator by a particular group, we must do the same thing for our numerator. This will then allow us to examine how much existing users represent the total comments, if they’re dominating, etc., giving us a much fuller picture.

While the number of new user is increasing, the company has to pay attention to number of retained users. Sometimes the number of retained users are very small even though a lot of new users, that means users dont use our service after first time;Sometimes new users growing very slow, but we have a pretty stable number of retained users, that means we are doing good but the market is probably saturated, we have to figure out how to bring more new customers.

Those are some examples of key metrics new companies should pay attention to.

5. We're working on a new feature for LinkedIn chat, and we want to implement a green dot to show an “active user”. Given engineering constraints, we can't AB test it before release.

How would you analyze the effectiveness of this new feature?

When you approach case study questions, remember to always clarify any vague terms. In this case, "effectiveness" is very vague. To help you define that term, you would want to first consider what the goal is of adding a green dot to LinkedIn chat.

What would be the expected business impact with the green dot implementation - is it to boost the user chat on LinkedIn? If so: Assume the chat frequency was evenly distributed throughout the day before the 'Green Dot' is implemented. Then Randomly sample the user chat activity and frequency distributions while users are online and offline - define null hypothesis that there's no statistically significant difference between the chat frequencites of users being online and being offline. Test the means of chat frequencies from online to offline to see if the null hypothesis can be rejected or not. We'd conclude whether the green dot implementation has made a difference on user chat activities.

6. Let's say that you're a data scientist on the engagement team. A product manager comes up to you and says that the weekly active users metric is up 5% but email notification open rates are down 2%. What would you investigate to diagnose what's happening?

What assumptions can you make about the relationship between weekly active users and email open rates? With a case question like this, you'd want to first answer that.

Hint: Open rate can decrease when its numerator decreases (fewer people open emails) or its denominator increases (more emails are sent). Taking these two factors into account, what are some hypotheses we can make about our decrease in open rate compared to our increase in weekly active users?

WAU = new users + resurrected users + churned users Is there any increase in particular category? If yes then check % of users by various channels, w/w ? Is there % increase through emails or decrease through emails or same?

If email channel % users decreased, if absolutes also decreased then it shows other channels causing an increase and we can look into marketing efforts, Check demographics of users sent emails, compare platforms and type of notifications? Are the same notifications sent to same users or sent to distinct users? Since email notification rate = no of emails opened / no of emails sent

If % of users through email notification remains same, and only rate goes down, means too many users sent emails, check the type of users and segment them to see if there is any change compared to last week

If % of users through email notification increase, and only rate goes down,it again means too many users sent emails, check the type of users and segment them to see if there is any change compared to last week. Are users getting distracted by too many emails? Check email sent per distinct users and see if they unsubscribed?

Apart from this, we need to check product performance metrics, email open to page load time or page response time or CTA issues? Tracking issues? Bugs? Check for false positives?

If email open rate is within 12 hours or standard 24 hours measured? Check for external factoras, Is it holiday season where people are not checking emails? Is there any hoax news for data privacy related?

7. Describe how you would build a model to predict Uber ETAs after a rider requests a ride.

Common case study problems like this are designed to explain how you build a model. Many times this can be scoped down to specific parts of the model building process. For example taking the example above, we could break it up to:
- How would you evaluate the predictions of an Uber ETA model?
- What features would you use to predict the Uber ETA for ride requests?

Our recommended framework breaks down a modeling and machine learning case study to individual steps in order to tackle each one thoroughly. In each full modeling case study, you'll want to go over:

Data processing
Feature Selection
Model Selection
Cross Validation
Evaluation Metrics
Testing and Roll Out

8. You work at a bank that wants to build a model to detect fraud on the platform.The bank wants to implement a text messaging service that will text customers when the model detects a fraudulent transaction in order for the customer to approve or deny the transaction with a text response. How would we build this model?

Let’s start out by understanding what kind of model would need to be built. We know that since we’re working with fraud, there has to be a case where there either is a fraudulent transaction or there isn't.

Hint: This problem is then a binary classification problem. Now given the problem scenario, what considerations do we have to think about when first building this model? What would the bank fraud data look like?

Training Data (chargebacks due to fraud & historical fraud txns)

If imbalanced data, then use the resampling technique to boost the fraud example
Model perspective (Use a simple Logistic Regression / RandomForest based so that realtime inference is fast). We could also explore Ensemble of multiple models (if runtime perspective if that is fast enough)

A simple classification model emitting probability score (a threshold could be chosen) to take the decision

Features

POS / web/ phone
Time (temporal features)
Realtime features (location)
txn_amount
txn_currency
Card in person
Was the card stolen or on hold
Is there a overdraft protection
Is txn_amount > 500 USD
# of times card used in the previous month
average txn amount for the last 90 days
is_international card / txn ?
Does it involve multi-currency ?
Is this from a different location than the correct Address ?
Cost consideration (False Positive vs False Negative)

Measure based on the cost consideration (As this could be highly class imbalanced data, we could consider F1-score along with accuracy and AUC)

9. We want to build a model to predict booking prices on Airbnb. Between linear regression and random forest regression, which model would perform better and why?

Hint: What are the main differences between linear regression and random forest?
Let's see how each model is applicable to Airbnb's bookings. One thing we need to do in the interview is to understand more context around the problem of predicting bookings. To do so, we need to understand which features are present in our dataset.

A linear regression is a linear model. Random forest is a tree-based model that grows trees in parallel. These are two completely different models so there are a lot of differences:
Linear Regression has many assumptions 1) normal distribution of error terms 2) independence of the predictors 3) mean residuals = 0 and constant variance 4) no multi-collinearity or autocorrelation. Random Forest, on the other hand, does not have these assumption requirements.
Linear Regression cannot handle cardinality and can be affected by extreme outliers. Random Forest handles missing values and cardinality very well and is not influenced by extreme outliers
A linear regression will work better if the underlying distribution is linear and has many continuous predictors. Random Forest will tend to be better with categorical predictors.
Both will give some semblance of a “feature importance.” However, linear regression feature importance is much more interpretable
In general, random forest will outperform linear regression. But as mentioned before, it’ll depend on the distribution of the dataset and the different available predictors. There is no cookie cutter, “ this model always performs better.”

Random forest out perform linear regression if the data is more complex (non linear relationship).

10. Suppose we have a binary classification model that classifies whether or not an applicant should be qualified to get a loan. Because we are a financial company, we have to provide each rejected applicant with a reason why. Given that we don't have access to the feature weights, how would we give each rejected applicant a reason why they got rejected?

Hint: How would the problem change if we had 10, 1000, or 10K applicants that had gone through the loan qualification program?

Given we do not have access to the feature weights, we are unable to tell each applicant which were the highest contributing factors to their application rejection. However, if we have enough results, we can start to build a sample distribution of application outcomes, and then map them to the particular characteristics of each rejection.

For example, if a rejected applicant had a recurring outstanding credit card balance of 10% of their monthly take-home income: if we know that the percentile of this data point falls within the middle of the distribution of rejected applicants, we can be fairly certain it is at least correlated with their rejection outcome. With this methodology, we can outline a few standard factors that may have led to the decision.

or
if it is a regression classifier, we can perform feature importance by changing each feature, while keeping other features the same value, and seeing how the scores fluctuates. If the prediction jumps from yes to no, vice versa, that indicates the positive or negative effect of changing that feature. We can use the score change to identify important features.

if it is a tree model, we can get access to the structure of the tree, some metrics will also be available to look at the feature importance. (Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node.)

We can also look at the percentile of each feature among all others, combined with the direction of that feature, we can infer how much that feature has positive/negative impact on the results. Again, this assume linear relationship between features and results.

11. Say you're running an e-commerce website. You want to get rid of duplicate products that may be listed under different sellers, names, etc... in a very large database.

For example: iPhone X and Apple iPhone 10

How do you go about doing this?

To formulate this problem, We need to identify all the keyword which depicts the same product. For example "iphone" in this case. Other keys that we need to consider is different numeric representations as X and 10. Identifying the keys is the most important part. than we can group entire dataset based on these keys. and drop duplicates. Other parameters that we can consider is seller origin or region/country of operation. As for different countries, there might be different offering which doesn't qualify as a duplicate record.

12. You work as a data scientist for a ride-sharing company. An executive asks how you would evaluate whether a 50% rider discount promotion is a good or bad idea. How would you implement it? What metrics would you track?

Hint: Be sure you ask for clarification on a question like this. What does good or bad mean to a ride-sharing business? What metrics will help you measure the pros and cons of the promotion?

13. 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?

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:

We can use undersampling, oversampling or SMOTE to make the data balanced.
We can alter the prediction threshold value by doing probability caliberation and finding a optimal threshold using AUC-ROC curve.
We can assign weight to classes such that the minority classes gets larger weight.
We can also use anomaly detection.

14. 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?

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 linearity assumptions.

15. 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?

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.

16. Say Chase Bank is looking into creating a new partner card (think Starbucks Chase credit card or Whole Foods Chase credit card). You have access to all of its customer spending data. How would you determine what the next partner card should be?

Hint: Chase creates partnerships with different merchants to increase acquisitions of new customers and retain existing customers. How can you find metrics to measure that? Will you need to look outside the current dataset?

Spending habits of customers play a big part here. So if a certain merchant customers are more transactors(they pay their bills completely) or they tend to roll over their balance from one month to the other. The FICO score of each customer population, age, region distribution. What is the brand loyalty of the new partner, whole foods has high loyalty compared to a regional grocery store. What is the partner looking into in terms of rewards system, are we splitting the cost of the rewards (negotiation plays a big role). How many total customers they have that are eligible to get a card (% of customers under 18).

17. Let's say that you're working at Netflix. The company executives are working to renew a deal with another TV network that grants Netflix exclusive licensing to stream their hit TV series (think something like Friends or The Office). One of the executives wants to know how to approach this deal. We know that the TV show has been on Netflix for a year already. How would you approach valuing the benefit of keeping this show on Netflix?

Here's how to approach a question like this: Start by trying to understand the reasons why Netflix would want to renew the show. Netflix mainly has three goals for what their content should help achieve:

Acquisition: To increase the number of subscribers.
Retention: To increase the number of active subscribers and retain them as paying members.

```
Revenue: To increase overall revenue.
With this in mind, how would you go about calculating the loss of subscribers caused by
not renewing the show?

You can ask these questions too

- Has the viewership stayed consistent or waned? Social media backlash?
- Is there a new season coming up?
- How many people who viewed this on Netflix renewed their subscriptions?
- What percent of viewers who saw this completed the entire series?
- Has there been a steady rise in the number of subscriptions/year?
```

18. 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?

```
 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:
We can use undersampling, oversampling or SMOTE to make the data balanced.
We can alter the prediction threshold value by doing probability caliberation and finding
a optimal threshold using AUC-ROC curve.
We can assign weight to classes such that the minority classes gets larger weight.
We can also use anomaly detection.
```

19. Assume that you are working at DataFlair and you have been assigned the task of developing a machine learning algorithm that predicts the number of views an article attracts. During the process of analysis, you include important features such as author name, number of articles written by the author in the past etc. What would be the ideal evaluation metric that you would use in this scenario?

```
Number of views that an article attracts on the website is a continuous target variable
which is a part of the regression problem. Therefore, we will make use of mean squared
error as our primary evaluation metric.
```

20. Suppose that you have to work with the data present on social media. After you have retrieved the data have to develop a model that suggests the hashtags to the user. How will you carry this out?

```
We can carry out Topic Modeling to extract significant words present in the corpus. To
capture the top n-gram words and their combinations. And, for learning repeating contexts
in the sentence we train a word2vec model.
```

21. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should you use?

```
One-way ANOVA

Explain more on this approach
```

Let me address three questions that inevitably come up.

1. Where is the answer key?

If you are asking for an answer key, you've missed the point about case interviews. Case interviews are open-ended by design. If these questions have answer keys, then they'd be useless to assess critical thinking. The hiring manager is listening for how you approach problems, and structure the analysis.

2. How do I know I have a good answer?

Try your answer out on a few friends, or better, a hiring manager. Then, try again. If you're doing it right, these different attempts should move along different paths, because the interview questions are designed to be open-ended.

3. There is not enough information to come to a conclusion.

That is exactly what every real-world data problem is like. You never have enough data, or all the right data, which means you need to make sensible assumptions, and keep moving along. A good case interview is a dialogue – you gather more information by asking your interviewer questions.

Happy Learning:)