Silicon Valley Technology Companies Data Science Machine Learning

What are some ways to prepare for a "data challenge" with a Silicon Valley tech company? How do I make sure I have sufficient fluency with data munging and ML in Python?

4 Answers

I-Kang Ding, Data Science Manager at Capital One
Answered Apr 6, 2016

Originally Answered: How can I successfully tackle a take-home assignment for a data scientist or data analyst interview?

This type of challenge is more specific to data scientists, and will test your skills on data manipulation, cleaning, and predictive modeling. Most often, the employer would give you a small dataset which you can easily deal with in modern laptops, and ask you to analyze the data and answer some specific questions that they asked. The questions often contain both descriptive and predictive components. For example, one of the dataset that I dealt with was the bike share ridership info, which contains information for tens of thousands bike share sessions across a 1-month span. For each entry, there are the date/time, starting/ending station ids, and bike ids. The descriptive questions may ask for "average amount of time a bike spends at a station in seconds", whereas the predictive question will ask for "predict the number of bicycles arriving at a specific station during a specific timespan."

Here are what I learned through performing several rounds of data analysis challenges from interviewing with different companies a while back:

- Stick with the modeling method you are comfortable with, and be prepared to justify why you choose a particular set of modeling methods. The justification shows a deeper layer of understanding than just the scikit-learn's API.
- At the end of the analysis, your goal was not only to have good grasp on the modeling, but also provide a written document to explain, in laymen's terms, what you have found. This is because, as a data scientist, you are often expected to communicate the results with other non-data scientists, and employers want to test your communication ability.
- You must be able to analyze data with code (eg. R, python, and maybe MATLAB) – point-and-click programs such as Excel won't cut it! I did all my analysis in R and wrote reports in RMarkdown; others have used Python / iPython notebook.

19.6k views · View 25 Upvoters

Timothy Schaaf, R/Python/Startup/Random Data Guy
Answered Dec 23, 2015 · Author has **92** answers and **145.4k** answer views

Originally Answered: What are some ways to prepare for a "data challenge" with a Silicon Valley tech company?

The length and difficulty of code challenges for data roles varies, quite extra-ordinarily. I've seen quick and simple (a few problems, less than an hour) to something resembling a marathon (6 or more hours).

The goal of the code challenge is to assess your technical capabilities and analytical thinking, and the best way to prepare for them, in general, is to be technically competent: you need to be able to do the work to do the challenge. This might mean a CS degree, or statistics, or other coursework. Outside of coursework and related fundamental skills, you could improve by participating in Kaggle Competitions, or contributing to relevant data open-source projects.

The fact that it is take home helps in that generally there's no transparency on time required to complete, though there is usually an expectation that it will be completed within a certain period of time (a few days, a week, etc.). But if you find it takes you a long

12.1k views · View 6 Upvoters

Related Questions

I have an interview with Facebook in 2 weeks for Data science role. I am a self taught data scientist and worked on few kaggle projects, how s...

What are some examples of take home data science challenges?

How can I prepare for a case analysis interview question for a Facebook data scientist position?

Giulio Palombo, Ex-Data Scientist @ Airbnb,ex-Caltech, wrote a book about DS takehome challenges
Answered Mar 22, 2016

Originally Answered: What are some ways to prepare for a "data challenge" with a Silicon Valley tech company?

Note: After Airbnb, I started teaching at DSE to help students become data scientists. I have seen hundreds of real data challenges. I also helped several companies in the Silicon Valley prepare their own data challenge and wrote the collection of data science takehome challenges book to help candidates prepare.

Like Timothy said, data challenges can vary a lot, ranging from a simple 1 hr exercise to a several day one. For instance, according to Glassdoor, several candidates reported that it took them 1 week to finish Uber takehome test, while Airbnb one is 3 hours. Quite a difference!

Data challenges are a very recent trend though. Likely, soon, things will get more standardized.

The data challenge is often the hardest part of a data science job interview, especially for candidates with no work experience. Indeed, many candidates on Glassdoor report lack of experience as the main reason for failing the challenge. More in details, lack of experience actually means:

- Candidates with no work experience have often a misguided idea of what a data scientist actually does. They focus too much on over optimizing the model and theoretical knowledge without spending time on things like data cleaning or coming up with product recommendations.
- If the question is open ended, they don't have enough experience to know which direction/approach taking at first to make sure that, within the given time frame, they get some interesting results no matter what.
- They are not familiar enough with R/Python. Therefore, simple data processing tasks take them a long time and they can't finish the challenge.

In my opinion, a good approach to solve a takehome data test is:

- Save some scripts in advance that you will almost surely have to use. Examples of these are: A function that plots all the variables against the label (this will already give you some insights to talk about!). A function that bins continuous variables into classes. A function to extract info from dates. A function building the ROC curve and optimizing the cutoff point. A function to cross-validate. A function that returns partial dependence plots for the top random forest variables. A function that builds a decision tree and automatically extracts the top 3/4 splits.
- Spend the first 20 minutes just checking the data and make sure they make sense. If there are wrong data, clean them. A great model on wrong data is not that useful!
- Look at the plots of each variable against the output, think about the product and set the direction of your work: where the information is likely to be, how this can be actionable, and what story you want to tell.
- Perform some basic feature engineering, such as extract day of the week or week of the yr from dates or create useful dummy variables.

optimizing them. Check that the model predicts well and, then, extract insights from the model via coefficients/splits/partial dependence plots. Highlight the findings that are in line with the given story you are trying to tell. One clear story beats multiple not-so-clear stories.

- Give clear examples of the interesting user segments you discovered and how your findings could improve the company product.

In order to prepare for a data science takehome test, you could:

- Nothing beats experience. I wrote a book "A collection of data science take-home challenges " that gives candidates the opportunity of practicing in advance on takehome challenges just like the ones you'll get in the job interview.
- Think extensively about the product of the company where you are interviewing. What are the pain points in your experience? Chances are you are not the only one having experienced those issues and you can likely find that in the data too. It is so much easier if you already have an idea of what to look for.
- Make sure you are familiar with joining and subsetting data in the language you use. Likely, they will send you more than one data set and you'll have to join/subset them according to certain criteria.
- Try to solve the easiest projects on Archived Problems - Project Euler . This will make you become fast in the data processing part.

Finally, the most important thing is that takehome challenges are often based on actual data scientist work at that company. Use the challenge as an opportunity to figure out whether you'd like working there.

Theresa Watson
Answered May 23, 2019

Below are my answer for the question: What are some ways to prepare for a data challenge with a Silicon Valley tech company How do I make sure I have sufficient fluency with data munging and ML in Python?

TOP 25 TIPS TO BECOME A PRO DATA SCIENTIST2!

Hi friends, I have worked in a head huntiing company since 2014, main field in data science, AI, deep learning.... Let me share amazing tips to become a pro d,ata scientiist as below. I hope that you love it. (ref from kdnuggets).

1. Leverage external datta sources: tweets about your company or your competitors, or datta from your vendors (for instance, customizable newsletter eBlast statistics available via vendor dashboards, or via submitting a ticket)

2. Nuclear physicists, mechanical engineers, and bioinformatics experts can make great datta scientists.

3. State your problem correctly, and use sound metrics to measure yield (over baseline) provided by datta science initiatives.

4. Use the right KPIs (key metrics) and the right datta from the beginning, in any project. Changes due to bad foundations are very costly. This requires careful analysis of your daata to create useful daatabases.

5. Ref this resourrce: 74 secrets to become a pro data scientiist

6. With big daata, strong signals (extremes) will usually be noise. Here's a solution.

7. Big dat,a has less value than useful dat,a.

8. Use big dat,a from third-party vendors, for competitive intelligence.

9. You can build cheap, great, scalable, robust tools pretty fast, without using old-fashioned statistical science. Think about model-free techniques.

11. Correlation is not causation. This article might help you with this issue. Read also this blog and this book.

12. You don't have to store all your dat,a permanently. Use smart compression techniques, and keep statistical summaries only, for old dat,a.

13. Don't forget to adjust your metrics when your da,ta changes, to keep consistency for trending purposes.

14. A lot can be done without da,tabases, especially for big da,ta.

15. Always include EDA and DOE (exploratory analysis/design of experiment) early on in any da,ta science projects. Always create a da,ta dictionary. And follow the traditional life cycle of any da,ta science project.

16. Da,ta can be used for many purposes:

– quality assurance

– to find actionable patterns (stock trading, fraud detection)

– for resale to your business clients

– to optimize decisions and processes (operations research)

– for investigation and discovery (IRS, litigation, fraud detection, root cause analysis)

– machine-to-machine communication (automated bidding systems, automated driving)

– predictions (sales forecasts, growth, and financial predictions, weather)

17. Don't dump Excel. Embrace light analytics. Da,ta + models + gut feelings + intuition is the perfect mix. Don't remove any of these ingredients in your decision process.

18. Leverage the power of compound metrics: KPIs derived from da,tabase fields, that have a far better predictive power than the original d,atabase metrics. For instance, your da,tabase might include a single keyword field but does not discriminate between the user query and search category (sometimes because d,ata comes from various sources and is blended together). Detect the issue, and create a new metric called keyword type – or d,ata source. Another example is IP address category, a fundamental metric that should be created and added to all digital analytics projects.

19. When do you need true real-time processing? When fraud detection is critical, or when processing sensitive transactional d,ata (credit card fraud detection, 911 calls). Other than that, delayed analytics (with a latency of a few seconds to 24 hours) is good enough.

20. Make sure your sensitive d,ata is well protected. Make sure your algorithms cannot be tampered by criminal hackers or business hackers (spying on your business and stealing everything they can, legally or illegally, and jeopardizing your algorithms – which translates in severe revenue loss). An example of business hacking can be found in section 3 in this article.

21. Blend multiple models together to detect many types of patterns. Average these models. Here's a simple example of model blending.

22. Ask the right questions before purchasing software.

23. Run Monte-Carlo simulations before choosing between two scenarios.

24. Use multiple sources for the same d,ata: your internal source, and d,ata from one or two vendors. Understand the discrepancies between these various sources, to have a better idea about what the real numbers should be. Sometimes big discrepancies occur when a metric definition is changed by one of the vendors or changed internally, or data has changed (some fields no longer tracked). A classic example is web traffic data: use internal log files, Google Analytics and another vendor (say Accenture) to track this data.

25. Fast delivery is better than extreme accuracy. All data sets are dirty anyway. Find the perfect compromise between perfection and fast return.

Related Questions

I have an interview with Facebook in 2 weeks for Data science role. I am a self taught data scientist and worked on few kaggle projects, how s...

What are some examples of take home data

I have an interview with Facebook in 2 weeks for Data science role. I am a self taught data scientist and worked on few kaggle projects, how s...

What are some examples of take home data science challenges?

How can I prepare for a case analysis interview question for a Facebook data scientist position?

What is the best site to find A/B tests case studies?

How should I prepare for whiteboard coding interviews for a data scientist position (is there a website like LeetCode)?

How do I prepare for a data scientist interview?

What kind of A/B testing questions should I expect in a data scientist interview and how should I prepare for such questions?

What are common CS questions asked during data scientist interviews?

How can I prepare for product sense in data scientist interview? I failed a dozen DS interviews and all feedback showed I am weak on product s...

How do I prepare for a data analyst/scientist position at Facebook?

What's Facebook Data scientist ETL interview question like?

How do I answer open-ended Data Science interview questions?

How do I prepare for an analytics job interview at Facebook?

What is the best site for preparing data science interview?

What topics are relevant for a Facebook Data Scientist phone interview?

question for a Facebook data scientist position?

What is the best site to find A/B tests case studies?

How should I prepare for whiteboard coding interviews for a data scientist position (is there a website like LeetCode)?

How do I prepare for a data scientist interview?