Learning Path to become a Data Scientist

The job market for a data science professional has never been higher. I see a lot of people are either starting their careers in analytics or making a switch from other professionals like software engineers. And for sure the intake of students for Masters in Business Analytics has exponentially increased in no time. In spite of a huge growth in the supply of potential candidates, the demand for a data scientist has been a steady rise and will be in the near future. Technology companies such as Google, Apple have always fallen short of employees, and start-ups that are evolving in the bay area are hungry for professionals. Financial companies are also adapting to leverage data science to take business decision and go data-driven. It is a fair and reasonable thing for students who are interested in data science to take this path, where the opportunity knocks on your door.

From my recent experience in finding and landing a job as a Data Scientist, I could share a few tips that might help improve the candidacy.

1. **Technical Skills**: This goes without saying that Data Scientist is expected to be very strong in technical skills and concepts.
    a. Probability - Random variables, Expected values, different probability distributions, conditional probability, Bayes' Theorem and problem solving.
    b. Statistics - Descriptive statistics, Inferential Statistics, A/B testing, Power Analysis, Different kinds of hypothesis testing, Multi Variant Analysis, ANOVA, Chi-square.
    c. Advanced Statistics - Linear regressions, Logistic regression, Maximum Likelihood estimation, Polynomial regression, Assumptions of linear regressions and their violations, Time series forecasting, Performance evaluation metrics.
    d. Machine Learning – Bagging and Boosting concepts, Decision Trees, Random Forest, GBM (Xgboost), SVM, ensemble techniques, Clustering techniques (K-means, Hierarchal clustering), Regularization methods (dimension reductions methods)
    e. Natural Language Processing (NLP) - Text processing methods, Text classification methods, Sentiment Analysis using Naïve Bayes algorithms and SVM, document clustering (similarity), Topic Modeling (LDA). There are a lot of courses on coursera on NLP (Stanford University, University of Michigan etc.)

    These are just a few that I mentioned here. There are a lot more and I highly recommend to take up a Machine learning course and check out the syllabus.

2. **Tools and Languages**: This is very important to understand how to implement algorithms or functions in a language of your choice. https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/
    a. R - Open source data science statistical language which is very widely used across every industry. A lot of statistical analysis could be carried out using R. There are numerous APIs and libraries that come in handy to solve business problems.
    b. Python - Another open source programming language that has gained huge popularity in data science in the recent years. The best thing about python is the in-built packages and their documentation that makes beginner to learn and implement very soon. Anyone with a little bit of programming language can adapt and pick up this language quite comfortably.
    c. SQL - No matter what you do, you can't get rid of SQL. The core querying language used almost daily by a data scientist. You need to master this!
    d. SAS - SAS has been in the industry even before the whole buzz about data science started. Industries such as banking and finance where the data should be highly

secured and confidential, SAS played a major role. With that being said, as the tech world is moving towards open source, to be collaborative in building projects, SAS has been dominated by R and Python lately. I see a lot of job descriptions stressing on R and python more than SAS. And the main drawback about SAS is that it is licensed software. So, spending huge amount of money of a tool when you can get R or python free is not the right decision for some of the companies.

e. Business Intelligence Tools: These tools are used mostly for visualization and reporting the data and also while performing a ETL projects.
    i. Tableau – highly recommend to learn this tools completely. The functionalities, the story boarding and integrating with other tools make this tool stand out.
    ii. QlikView – Another competitive visualization tool.
    iii. d3 - visual appearance is really great and it is open source. However, you need to learn java script which could be challenging for non-computer background candidates
f. Big Data: The next big thing! As the data is increasing astronomically, it's a challenge to store and process it to gain insights from it. In that aspects, we have Hadoop (Map Reduce algorithm, Pig, Hive), Apache Spark (Pyspark, sparkR, scala), H2O, AWS as starters. There are a lot of new tools and techniques are in the market to explore. Spending enough time on Big Data is must.

3. **Business knowledge**: Only after giving multiple interviews from across various domains, I understood the importance of having business knowledge. It's not a mandatory requirement but having to know what the company does and how it works will have an undue advantage in talking business with the hiring manager. For example, if you are looking at Healthcare Insurance companies, I would recommend learning the terminology and concepts that are mainstream, such as deductibles, claims, co-insurance etc. In this manner, you would understand the business model of the company and that will help you think how you, a data scientist, could be of any help to them. There are numerous articles and white papers specific to domains, detailing about how analytics is being used to solve bigger problems in the industry. Reading about them should put you in situation that separates from the rest of the crowd. This also shows your interests in the domain and the company you are applying. And from my previous interactions with data scientists, the most important tipping point is when you explain to them a concept that coincides with their daily work. For example, I interviewed for a data scientist position and the interviewer, who is also a data scientist, has worked on text classification problem and I happened to have a project on that. This made him curious to know what and how I did: the approach, the packages used and the business impact. We hit it off really well immediately. So, when you are interviewing with someone, study about them on LinkedIn thoroughly, and pace your interview in that manner. Your answers to same question should vary with the people interviewing. An MBA grad (Manager) would expect your answers in terms of deep business impact, while a data scientist would expect the technical part of the project too. So, structure your answers accordingly.
4. **Case interviews**: This has been my daily bread and butter while searching for the job. Companies are more interested in your analytical thinking while solving some hypothetical problems than pure technical skills. Lot of job seekers who dint know about this had trouble answering consultant like problems, even if they applied to data science roles or business intelligence roles. Even though data scientist typically won't be solving such problems, however, having this knowledge is THE main key for cracking the interview. There are a lot of resources to study, books such as Case in Point by Marc Cosentino, Case Interview sessions by

Victor Chen, frameworks etc. These materials would cover most of the case related scenarios that would help you in case interviews. Do finish these book and practice randomly with a friend (more like a discussion) so that you wouldn't forget. Try solving guestimate problems as well. There is no right answer for these questions but definitely your approach matters!

5. **Miscellaneous**: Always stay positive with the whole process of interview and having patience is very important. The process might take 1-2 months from the start of first round, hence holding tight and going one step at a time should help you get through.
    a. I see a lot of data challenges as part of the interview process to test your programming skills, asking you to give insights on the data or build a model or recommendations. In that case, ask the recruiters clearly about their expectation from this analysis, which language to use and in which format to send back the solutions. Sometimes more than the analysis, your presentation might have more weightage, so being more cautions is needed.
    b. Resume is the first way of communicating with the companies, so having this perfect is as important as getting a job. Get it proof read by a data scientist as well because they would know how to structure projects and business impacts.
    c. If you are into coding, have a GitHub profile and save all your codes and approaches for the projects you did. This clearly tells how interested, organised you are and all the more evidence that you know the programming languages. This part helped me a lot!!
    d. Try participating in Kaggle and AnalyticsVidhya as these are very important sources of knowledge. Reading their blogs will help you understand every aspect of data science.
    e. There are many sources such as coursera – Machine learning by Andrew Ng, ISLR & Elements of statistical learning by Trevor Hastie, Youtube – Khan Academy, Sentdex, DataCamp, Brandon Fotlz, MarinStatsLectures and many more.
    f. Whenever you don't clear an interview, just send a casual mail to them asking what went wrong. This is totally okay. Analyse the feedback and try not doing the same mistake again.
    g. I don't want to go so philosophical but do not lose hope. You might have some setbacks but don't let it ruin your ongoing calls. Think how you can rectify and better structure your answers. And always show the enthusiasm while in the phone or video or face to face. That shows you are very energetic. And have very good questions in the end to ask your interviewer to know more about the role and impress them further as well.

There is a vast ocean of concepts to learn in data science. There are definitely more things that you can learn from the things mentioned above. This document could be starter for many more to come. You can reach out to me any time if you want to talk and need any help what so ever. I would learn as well and get to know someone too. That's a win-win situation. Wish you all the very best!

Aditya Padala

LinkedIn Gmail GitHub