How to Build a Data Science Portfolio

Michael Galarnyk Follow
Jul 8, 2018 · 18 min read

A portfolio is one way to show people you are that data science unicorn.

How do you get a job in data science? Knowing enough statistics, machine learning, programming, etc to be able to get a job is difficult. One thing I have found lately is quite a few people **may have the required skills to get a job, but no portfolio**. While a resume matters, having a portfolio of public evidence of your data science skills can do wonders for your job prospects. Even if you have a referral, **the ability to show potential employers what you can do instead of just telling them you can do something is important**. This post will include links to where various data science professionals (data science managers, data scientists, social media icons, or some combination thereof) and others talk about what to have in a portfolio and how to get noticed. With that, let's get started!

Robinson Chief Data Scientist at DataCamp when he was interviewed by Marissa Gemma on Mode Analytics blog. He was asked about landing his first job in industry and said,

> *The most effective strategy for me was doing public work. I blogged and did a lot of open source development late in my PhD, and these helped give public evidence of my data science skills. But the way I landed my first industry job was a particularly noteworthy example of the public work. During my PhD I was an active answerer on the programming site Stack Overflow, and an engineer at the company came across one of my answers (one explaining the intuition behind the beta distribution). He was so impressed with the answer that he got in touch with me [through Twitter], and a few interviews later I was hired.*

You may think of this as a freak occurrence, but you will often find that the more active you are, the greater chance you have of something like this occuring. From David's blog post,

> *The more public work you do, the higher the chance of a freak accident like that: of someone noticing your work and pointing you towards a job opportunity, or of someone who's interviewing you having heard of work you've done.*

People often forget that software engineers and data scientists also Google their issues. If these same people have their problems solved by reading your public work, they might think better of you and reach out to you.

Portfolio to get around an Experience Requirement

Even for an entry level role, most companies want to have people with at least a little bit of real life experience. You may have seen memes like the one below.

The question is how do you get experience if you need experience to get your first job? If there is an answer, the answer is **projects**. Projects are perhaps the best substitutes for work experience or as Will Stanton said,

> *If you don't have any experience as a data scientist, then you absolutely* have to *do independent projects.*

In fact, when Jacqueline Nolis interviews candidates, she wants to hear about a description of a recent problem/project that you have faced.

> *I want to hear about a project they've worked on recently. I ask them about how the project started, how they determined it was worth time and effort, their process, and their results. I also ask them about what they learned from the project. I gain a lot from answers to this question: if they can tell a narrative, how the problem related to the bigger picture, and how they tackled the hard work of doing something.*

If you don't have some data science related work experience, the best option here is to talk about a data science project that you have worked on.

Types of Projects to Include in a Portfolio

Data science is such a broad field that it is hard to know what kind of projects hiring managers want to see. William Chen, a Data Science Manager at Quora, shared his thoughts on the subject at Kaggle's CareerCon 2018 (video).

> *I love projects where people show that they are interested in data in a way that goes beyond homework assignments. Any sort of class final project where you explore an interesting dataset and find interesting results… Put effort into the writeup… I really like seeing really good writeups where people find interesting and novel things…have some visualizations and share their work.*

advice on what kind of datasets you should use. He also echos one of William's points about working with interesting data.

> *I find that the best portfolio projects are less about doing fancy modeling and more about working with interesting data. A lot of people do things with financial information or Twitter data; those can work, but the data isn't inherently that interesting, so you're working uphill.*

One of his other points in the article is that webscraping is a great way to get interesting data. If you are interested in learning how to build your own dataset by webscraping in Python, you can see my post here. If you are coming from academia, it is important to note that your thesis can count as a project (a very large project). You can hear William Chen talk about it here.

Traffic Cruising Data Science for Social Good Project (https://github.com/uwescience/TrafficCruising-DSSG2017). This is an example of a project I personally find interesting, but there are so many interesting projects out there. Credit (Orysya Stus, Brett Bejcek, Michael Vlah, Anamol Pundle)

Types of Projects NOT to Include in a Portfolio

One thing I have found very common (to the point of it appearing multiple times in this blog post) in a lot of portfolio/resume advice is not to have common projects in your portfolio.

> *personal projects.*
>
> *When in doubt, here are some projects that hurt you more than they help you:*
>
> ** Survival classification on the Titanic dataset.*
>
> ** Hand-written digit classification on the MNIST dataset.*
>
> ** Flower species classification using the iris dataset.*

The image below shows partial examples of classification of Titanic (A), MNIST (B), and iris (C) datasets. There aren't a lot of ways to use these datasets to distinguish yourself from other applicants. Make sure to list novel projects.

Titanic (A), MNIST (B), and iris (C) classification

Portfolios are Iterative

Favio Vazquez has an excellent article where he talked about how he got his job as a data scientist. Of course, one of his tips is to have a portfolio.

> *Have a portfolio. If you are looking for a serious paid job in data science do some projects with real data. If you can post them on GitHub. Apart from Kaggle competitions, find something that you love or a problem you want to solve and use your knowledge to do it.*

One of the other interesting findings was that you always have to keep on improving as you go through the job hunt.

> *be honest). I studied a lot, programmed everyday, read a lot of articles and posts. They helped a lot.*

As you learn more and improve yourself, your portfolio should also be updated. This same sentiment is echoed in many other advice articles. As Jason Goodman said,

> *The project isn't done when you post it publicly. Don't be afraid to keep adding on to or editing your projects after they're published!*

This advice is especially true when you are looking for a job. There are many stories of successful people like Kelly Peng, Data Scientist at Airbnb, who really persevered and kept on working and improving. In one of her blog posts, she went over how many places she applied for and interviewed with.

> *Applications: 475*
>
> *Phone interviews: 50*
>
> *Finished data science take-home challenges: 9*
>
> *Onsite interviews: 8*
>
> *Offers: 2*
>
> *Time spent: 6 months*

She clearly applied to a lot of jobs and kept on persisting. In her article, she even mentions how you need to keep on learning from your interviewing experiences.

> *Take note of all the interview questions you got asked, especially those questions you failed to answer. You can fail again, but don't fail at the same spot. You should always be learning and improving.*

*If you aren't getting interviews yet, apply for more jobs and keep on finding ways to learn and improve.

Incorporating Portfolio into 1 Page Resume

One of the ways someone finds your portfolio is often through your resume so it is worth a mention. A data science resume is a place to focus on your technical skills. Your resume is a chance to succinctly represent your qualifications and fit for that particular role. Recruiters and hiring managers skim resumes very quickly, and you only have a short time to make an impression. Improving your resume can increase your chance of getting an interview. You have to make sure every single line and every single section of your resume counts.

William Chen, a Data Science Manager from Quora has 9 Tips for making your data science resume. **Notice in the brief summary of his points below, that projects and portfolio are points 6, 7, 8, and arguably 9.**

1. **Length**: Keep it simple and one page max. This gives you the most impact for a quick skim. Recommends a simple one column resume as it is easy to skim.

Sample Resume used in Video (latex: https://github.com/sb2nov/resume)

2. Objective: Don't include one. They don't help you distinguish yourself from other people. They take away space from the more important things (skills, projects, experience etc). Cover letters are extremely optional unless you really personalize it.

Objectives don't help you distinguish yourself from other people. A lot of them say very similar things.

Examples of relevant coursework displayed on various resumes.

4. Skills: Don't give numerical ratings for your skills. If you want to rate yourself on your skills, use words like proficient or familiar or things like that. You can even exclude assessments altogether.

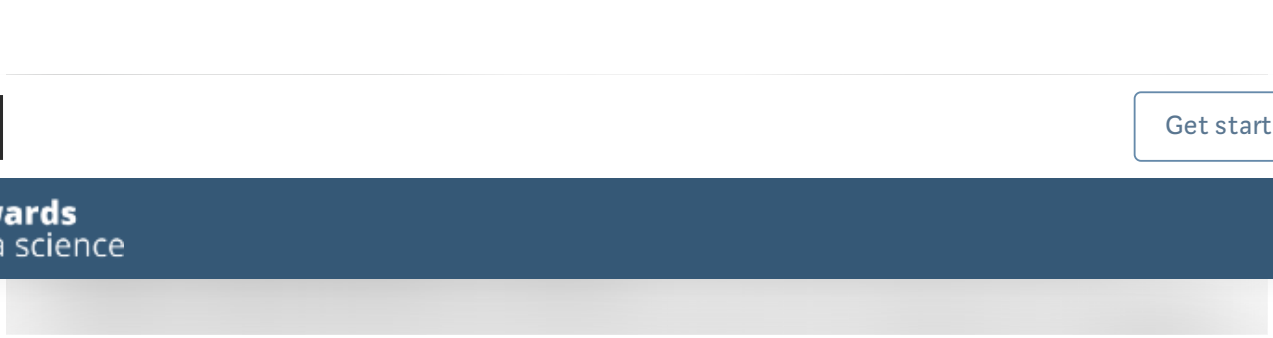
Don't give numerical ratings for your skills

5. Skills: Do list technical skills that the job description mentions. The order you list your skills in can suggest what you are best at.

Examples of how you can list your skills on your resume

6. Projects: Don't list common projects or homework. They aren't that helpful in distinguishing you from other applicants. List projects that are novel.

7. Projects: Show results and include links. If you participated in Kaggle competition, put percentile rank as it helps the person reading your resume understand where you are in the competition. In projects sections, there is always room for links to writeups and papers as they let the hiring manager or recruiter dig in deeper (bias to real world messy problems where you learn something new).



Good example project sections

Notice that in one of the projects sections above, a person has an additional link to a blog that lets the recruiter or hiring manager find out more. This is one way to link to various parts of your portfolio from your resume.

8. Portfolio: Fill our your online presence. The most basic is a LinkedIn profile. It is kind of like an extended resume. Github and Kaggle profiles can help show off your work. Fill out each profile and include links to other sites. Fill out descriptions for your GitHub respositories. Include links to your knowledge sharing profiles/blog (medium, quora). Data science specificially is about knowledge sharing and communicating what the data means to other people. You don't have to do all of them, but pick a few and do it (More on this later).

9. Experience: Tailor your experience towards the job. Experience is the core of your resume, but if you don't have work experience what do you do? Focus your resume on independent projects, like capstone projects, independent research, thesis work, or Kaggle competitions. These are substitutes for work experience if you don't have work experience to put on your resume. Avoid putting irrelevant experience on your resume.

Importance of Social Media

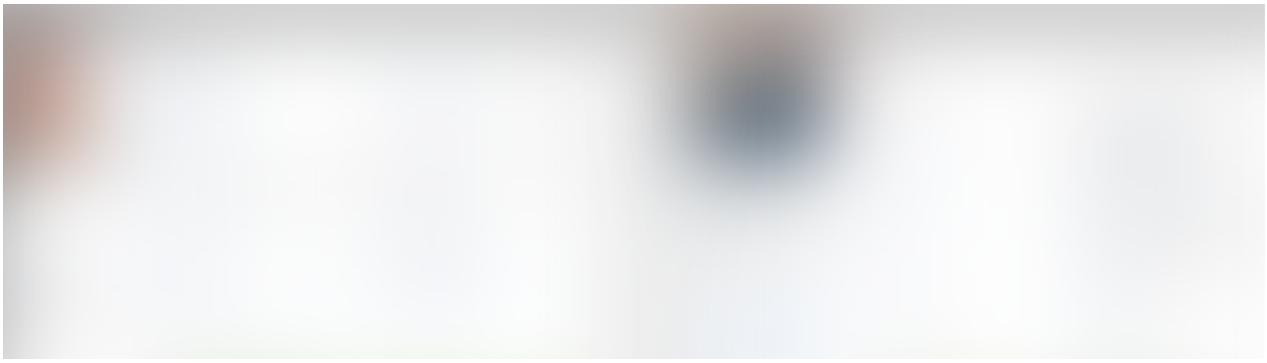
This is very similar to the Importance of a Portfolio section, just divided into subsections. Having a Github page, a Kaggle profile, a Stack Overflow, etc can provide support for your resume. Having online profiles filled out can be a good signal for hiring managers.

As David Robinson phrases it,

> *Generally, when I'm evaluating a candidate, I'm excited to see what they've shared publicly, even if it's not polished or finished. And sharing* anything *is almost always better than sharing nothing.*
>
> The reason why data scientists like seeing public work is as Will Stanton said,
>
> *Data scientists use these tools to share their own work and find answers to questions. If you use these tools, then you are signaling to data scientists that you are one of* them*, even if you haven't ever worked as a data scientist.*

A lot of Data science is about communication and presenting data so it is good to have these online profiles. Besides from the fact that these platforms help provide valuable experience, they can also help you get noticed and lead people to your resume. People can and do find your resume online through various sources (LinkedIn, GitHub, Twitter, Kaggle, Medium, Stack Overflow, Tableau Public, Quora, Youtube, etc). You will even find that different types of social media feed into eachother.

Github



A Github profile is a powerful signal that you are a competent data scientist. In the projects section of a resume, people often leave links to their GitHub where the code is stored for their projects. You can also have writeups and markdown there. GitHub lets people see what you have built and how you have built it. At some companies, hiring managers look at an applicants GitHub. It is another way to show employers you aren't a false positive. If you take the time to develop your GitHub profile, you can be better evaluated than others.

It is worth mentioning that you need to have some sort of README.md with a description of your project as a lot of **data science is about communicating results**. Make sure the README.md file clearly describes what your project is, what it does, and how to run your code.

Kaggle

Participating in Kaggle competitions, creating a kernel, and contributing to discussions are ways to show some competency as a data scientist. It is important to emphasize that Kaggle is not like an industry project as Colleen Farrelly, mentions in this quora question. Kaggle competitions take care of coming up with a task, acquire data for you, and clean it into some usable form. What it does is give you practice analyzing data and coming up with a model. Note that there is a good reason why Kaggle Grandmasters continue to participate in Kaggle competitions. Reshama Shaikh has a post To Kaggle Or Not where she talked about the value of Kaggle competitions. From her post,

> *It is true, doing one Kaggle competition does not qualify someone to be a data scientist. Neither does taking one class or attending one conference tutorial or analyzing one dataset or reading one book in data science. Working on competition(s) adds to your experience and augments your portfolio. It is a complement to your other projects, not the sole litmus test of one's data science skillset.*

I completely agree with Reshama's view on this. In particular, the point about how taking a class in something doesn't make you an expert in something nor does it give you a job. I literally have made a course called Python for Data Visualization and I go

Linkedin

Unlike a resume, which is confined by length, a LinkedIn profile allows you to describe your projects and work experience in more depth. Udacity has a guide on making a good LinkedIn profile. An important part of LinkedIn is their search tool and for you to show up, you must have *relevant keywords* in your profile. Recruiters often search for people on LinkedIn. LinkedIn allows you to see which companies have searched for you and who has viewed your profile.

Checking where your searchers work and how many times people have viewed your profile.

Besides companies finding you and sending you messages on your availability, LinkedIn also has many features like Ask for a Referral. Jason Goodman in his article Advice on Applying to Data Science Jobs uses LinkedIn to **indirectly** ask for referrals.

> *I never, never, never applied to any companies without an introduction to someone who worked at the company…once I was interested in a company, I would use LinkedIn to find a first- or second- degree connection at the company. I would write to that connection, asking to talk to them about their experience at the company and, if possible, whether they'd be able to connect me to someone on the Data Science team. Whenever I could, I did in-person meetings (coffee or lunch) instead of phone calls. As an aside, Trey Causey recently wrote a great post on how to ask for just these kinds of meetings. I would never ask for a job directly, but they would usually ask for my resume and offer to submit me as an internal referral, or put me in touch with a hiring manager. If they didn't seem comfortable doing so…I'd just thank them for their time and move on.*

Aman Dalmia learned something similar by Interviewing at Multiple AI Companies and Startups.

> *Networking is* NOT *messaging people to place a referral for you. When I was starting off, I did this mistake way too often until I stumbled upon* this *excellent article by* Mark Meloon, *where he talks about the importance of building a* real *connection with people by offering our help first.*

One other point he had is that LinkedIn is great for getting your content/portfolio out.

> *Another important step in networking is to get your content out. For example, if you're good at something, blog about it and share that blog on Facebook and LinkedIn. Not only does this help others,* it helps you as well.

Medium and/or Other Blogging Platforms

Having some form of blog can be highly beneficial. A lot of data science is about communication and presenting data. Blogging is a way of practicing this and showing you can do this. Writing about a project or a data science topic allows you to share with the community as well as encourages you to write out your work process and thoughts. This is a useful skill when interviewing.

As David Robinson said,

A blog is your chance to practice the relevant skills.

- ***Data cleaning****: One of the benefits of working with a variety of datasets is that you learn to take data "as it comes", whether it's in the form of a supplementary file from a journal article or a movie script*
- ***Statistics****: Working with unfamiliar data lets you put statistical methods into practice, and writing posts that communicate and teach concepts helps build your own understanding*

- ***Visualization**: Having an audience for your graphs encourages you to start polishing them and building your personal style*
- ***Communication**: You gain experience writing and get practice structuring a data-driven argument. This is probably the most relevant skill that blogging develops since it's hard to practice elsewhere, and it's an essential part of any data science career*

By writing a blog, you can practice communicate findings to others. It also is another form of advertising yourself. Blogs about Using Scrapy to Build your Own Dataset, and ironically Python Environment Management with Conda have taught me a lot and have gotten me a lot of opportunities I would normally not have gotten. Recently, my boxplot blog brought me the opportunity to create my own Python for Data Visualization course. One of the major benefits I have found is that throughout the process of people critiquing my projects and suggesting improvements (though the comments section of the blog) makes it so interviewers aren't the first ones pointing out these same flaws. The more obvious benefit is that by making a blog you tend to read a lot more data science/machine learning blog posts and hence learn more.

As for what platform to blog on, I recommend using Medium. Manali Shinde in her blog post How to Construct a Data Science Portfolio from Scratch had a really good point on why she choose Medium for her blog.

> *I thought of creating my own website on a platform such as WordPress or Squarespace. While those platforms are amazing to host your own portfolio, I wanted a place where I would get some visibility, and a pretty good tagging system to reach greater audiences. Luckily Medium, as we know, has those options (and it's also free).*

If you don't know what to write about, I suggest you look at David Robinson's advice.

https://twitter.com/drob

Twitter

Being active on Twitter is a great way to identify and interact with people in your field. You can also promote your blog on Twitter so that your portfolio can be that much more visible. There are so many opportunities to interact with people on twitter. One of them as Reshama Shaikh said in her famous blog post "How Do I Get My First Data Science Job?" was,

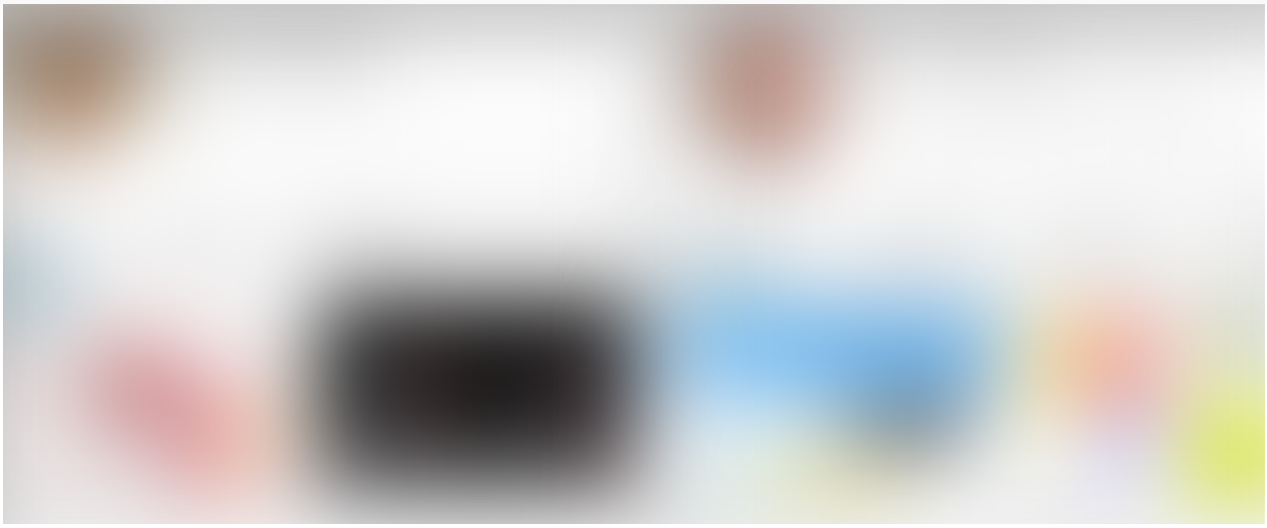
> *David Robinson generously offers to retweet your first data science post. With 20K+ followers, that's an offer that can't be refused.*

Twitter can be used for other things than self promotion. Data Science Renee has a post "How to use Twitter to Learn Data Science (or Anything)" that is quite insightful about taking Twitter to learn skills. One other takeaway from her article was how much her Twitter presence helped her network and get opportunities.

> *I have been asked to be interviewed on podcasts and blogs (some of those should be coming up soon), offered contract work, and offered free admission to a conference I unfortunately couldn't go to, but was excited to be considered for. "Famous" people in the industry are now coming to me to work with them in some way.*

Tableau Public

Not every data science job uses Tableau or other BI tools. However, if you are applying to jobs where these tools are used, it is important to note that there are websites where you can put dashboards for public consumption. For example, if you say you are learning or know Tableau, put a couple dashboards on Tableau Public.



Conclusion

Remember a portfolio is a process. Keep on improving.

Having a strong resume has long been the primary tool for job seekers to relay their skills to potential employers. These days, there is more than one way to showoff your skills and get a job. A portfolio of public evidence is a way to get opportunities that you normally wouldn't get. It is important to emphasize that a portfolio is an iterative process. As your knowledge grows, your portfolio should be updated over time. Never stop learning or growing. Even this blog post will be updated with feedback and with increasing knowledge. If you want interview advice/guides, time to check out Brandon Rohrer's advice on how to survive a data science interview,

Data Science | Interview | Portfolio | Python | Towards Data Science

16.4K claps

WRITTEN BY

Michael Galarnyk

Data Scientist at Scripps Research Institute

Follow

Towards Data Science

A Medium publication sharing concepts, ideas, and codes.

Follow

See responses (70)