As a data scientist, I often get the question,"What do you actually do?"

Notebook:	Data Science		
Created:	2/23/2017 10:21 PM	**Updated:**	2/23/2017 10:34 PM
Author:	sumendar@gmail.com		
Tags:	project tips		
URL:	https://www.springboard.com/blog/data-science-process/?utm_content=buffer390ce&utm_...		

As a data scientist, I often get the question,"What do you actually do?"

Data scientists can appear to be wizards who pull out their crystal balls (MacBook Pros), chant a bunch of mumbo-jumbo (machine learning, random forests, deep networks, Bayesian posteriors) and produce amazingly detailed predictions of what the future will hold. However, as much as we'd like to believe it was, data science is not magic. The power of data science comes from a deep understanding of statistics and algorithms, programming and hacking, and communication skills. More importantly, data science is about applying these three skill sets in a disciplined and systematic manner.

Over the last few years, I've not only worked as an individual data scientist in several companies, but also led a team of data scientists as chief data scientist at Pindrop Security, a hot Andreessen-Horowitz funded cybersecurity startup. My team worked on several cutting-edge projects using a wide variety of tools and techniques. Over time, I realized that despite the variation in the details of different projects, the steps that data scientists use to work through a complex business problem remain more or less the same.

After Pindrop, I joined Springboard as the director of data science education. In this capacity, my role is to design and maintain our data science courses for students, such as our **data science career track bootcamp**. Designing these courses compelled me to reflect on the systematic process that data scientists use at work, and to make sure that I incorporated those steps in each of our data science courses. In this article, I explain this data science process through an example case study. By the end of the article, I hope that you will have a high-level understanding of the day-to-day job of a data scientist, and see why this

role is in such high demand.

The Data Science Process

Congratulations! You've just been hired for your first job as a data scientist at Hotshot, a startup in San Francisco that is the toast of Silicon Valley. It's your first day at work. You're excited to go and crunch some data and wow everyone around you with the insights you discover. But where do you start?

Over the (deliciously catered) lunch, you run into the VP of Sales, introduce yourself and ask her, *"What kinds of data challenges do you think I should be working on?"*

The VP of Sales thinks carefully. You're on the edge of your seat, waiting for her answer, the answer that will tell you exactly how you're going to have this massive impact on the company of your dreams.

And she says, *"Can you help us optimize our sales funnel and improve our conversion rates?"*

The first thought that comes to your mind is: *What? Is that a data science problem? You didn't even mention the word 'data'. What do I need to analyze? What does this mean?*

Fortunately, your data scientist mentors have warned you already: this initial ambiguity is a regular situation that data scientists encounter frequently. All you have to do is systematically apply the data science process to figure out exactly what you need to do.

The data science process: a quick outline

When a non-technical supervisor asks you to solve a data problem, the description of your task can be quite ambiguous at first. It is up to you, as the data scientist, to translate the task into a concrete problem, figure out how to solve it and present the solution back to all of your stakeholders. We call the steps involved in this workflow the “Data Science Process.” This process involves several important steps:

- **Frame the problem:** Who is your client? What exactly is the client asking you to solve? How can you translate their ambiguous request into a concrete, well-defined problem?
- **Collect the raw data needed to solve the problem:** Is this data already available? If so, what parts of the data are useful? If not, what more data do you need? What kind of resources (time, money, infrastructure) would it take to collect this data in a usable form?
- **Process the data (data wrangling):** Real, raw data is rarely usable out of the box. There are errors in data collection, corrupt records, missing values and many other challenges you will have to manage. You will first need to clean the data to convert it to a form that you can further analyze.
- **Explore the data:** Once you have cleaned the data, you have to understand the information contained within at a high level. What kinds of obvious trends or correlations do you see in the data? What are the high-level characteristics and are any of them more significant than others?
- **Perform in-depth analysis (machine learning, statistical models, algorithms):** This step is usually the meat of your project,where you apply all the cutting-edge machinery of data analysis to unearth high-value insights and predictions.
- **Communicate results of the analysis:** All the analysis and technical results that you come up with are of little value unless you can explain to your stakeholders what they mean, in a way that’s comprehensible and compelling. Data storytelling is a critical and underrated skill that you will build and use here.

So how can you help the VP of Sales at hotshot.io? In the next few sections, we will walk you through each step in the data science process, showing you how it plays out in practice. Stay tuned!

Step 1 of 6: Frame the problem (a.k.a. “ask the right questions”)

The VP of Sales at hotshot.io, where you just started as a data scientist, has asked you to help optimize the sales funnel and improve conversion rates. Where do you start?

You start by asking a lot of questions.

- Who are the customers, and how do you identify them?
- What does the sales process look like right now?
- What kind of information do you collect about potential customers?
- What are the different tiers of service right now?

Your goal is to get into your client’s (the VP in this case) head and understand their view of the problem as well as you can. This knowledge will be invaluable later when you analyze your data and present the insights you find within.

Once you have a reasonable grasp of the domain, you should ask more pointed questions to understand exactly what your client wants you to solve. For example, you ask the VP of Sales, *“What does optimizing the funnel look like for you? What part of the funnel is not optimized right now?”*

She responds, *“I feel like my sales team is spending a lot of time chasing down customers who won’t buy the product. I’d rather they spent their time with customers who are likely to convert. I also want to figure out if there are customer segments who are not converting well and figure out why that is.”*

Bingo! You can now see the data science in the problem. Here are some ways you can frame the VP’s request into data science questions:

1. What are some important customer segments?
2. How do conversion rates differ across these segments? Do some segments perform significantly better or worse than others?
3. How can we predict if a prospective customer is going to buy the product?
4. Can we identify customers who might be on the fence?
5. What is the return on investment (ROI) for different kinds of customers?

Spend a few minutes and think about any other questions you'd ask.

Now that you have a few concrete questions, you go back to the VP Sales and show her your questions. She agrees that these are all important questions, but adds: *"I'm particularly interested in having a sense of how likely a customer is to convert. The other questions are pretty interesting too!"* You make a mental note to prioritize questions 3 and 4 in your story.

The next step for you is to figure out what data you have available to answer these questions. Stay tuned, we'll talk about that next time!

Step 2 of 6: Collect the right data

You've decided on your very first data science project for hotshot.io: predicting the likelihood that a prospective customer will buy the product.

Now's the time to start thinking about data. What data do you have available to you?

You find out that most of the customer data generated by the sales department is stored in the company's CRM software, and managed by the Sales Operations team. The backend for the CRM tool is a SQL databasewith several tables. However, the tool also provides a very convenient web-based API that returns data in the popular JSON format.

What data from the CRM database do you need? How should you extract it? What format should you store the data in to perform your analysis?

You decide to roll up your sleeves and dive into the SQL database. You find that the system stores detailed identity, contact and demographic information about customers, in addition to details of the sales process for each of them. You decide that since the dataset is not too large, you'll extract it to CSV files for further analysis.

As an ethical data scientist concerned with both security and privacy, you are careful not to extract any personally identifiable information from the database. All the information in the CSV file is anonymized, and cannot be traced back to any specific customer.

In most data science industry projects, you will be using data that already exists and is being collected. Occasionally, you'll be leading efforts to collect new data, but that can be a lot of engineering work and it can take a while to bear fruit.

Well, now you have your data. Are you ready to start diving into it and cranking out insights? Not yet. The data you have collected is still 'raw data'—which is very likely to contain mistakes, missing and corrupt values. Before you draw any conclusions from the data, you need to subject it to some data wrangling, which is the subject of our next section.

Step 3 of 6: How to process (or "wrangle") your data

As a brand-new data scientist at hotshot.io, you're helping the VP of Sales by predicting which prospective customers are likely to buy the product. To do so, you've extracted data from the company's CRM into CSV files.

But, despite all your work, you're not ready to use the data yet. First, you need to make sure the data is clean! Data cleaning and wrangling often takes up the bulk of time in a data scientist's day-to-day work, and it's a step that requires patience and focus.

First, you need to look through the data that you've extracted, and make sure you understand what every column means. One of the columns is called 'FIRST_CONTACT_TS', representing the date and time the customer was first contacted by hotshot.io. You automatically ask the following questions:

- Are there missing values i.e. are there customers without a first contact date? If not, why not? Is that a good or a bad thing?
- What's the time zone represented by these values? Do all the entries represent the same time zone?
- What is the date range? Is the date range valid? For example, if hotshot.io has been around since 2011, are there dates before 2011? Do they mean anything special or are they mistakes? It might be worth verifying the answer with a member of the sales team.

Once you have uncovered missing or corrupt values in your data, what do you do with them? You may throw away those records completely, or you may decide to use reasonable default values (based on feedback from your client). There are many options available here, and as a data scientist, your job is to decide which of them makes sense for your specific problem.

You'll have to repeat these steps for every field in your CSV file: you can begin to see why data cleaning is time-consuming. Still, this is a worthy investment of your time, and you patiently ensure that you get the data as clean as possible.

This is also a time when you make sure that you have all of the critical pieces of data you need. In order to predict which future customers will convert, you need to know which customers have converted in the past. Conveniently enough, you find a column called 'CONVERTED' in your data, with a simple 'Yes/No' value.

Finally, after a lot of data wrangling, you're done cleaning your dataset, and you're ready to start drawing some insights from the data. Time for some exploratory data analysis!

Step 4 of 6: Explore your data

You've extracted data and spent a lot of time cleaning it up.

And now, you're finally ready to dive into the data! You're eager to find out what information the data contains, and which parts of the data are significant in answering your questions. This step is called exploratory data analysis.

What are some things you'd like to explore? You could spend days and weeks of your time aimlessly plotting away. But you don't have that much time. Your client, the VP of Sales, would love to present some of your results at the board meeting next week. The pressure is on!

You look at the original question: predict which future prospects are likely to convert. What if you split the data into two segments based on whether the customer converted or not and examine differences between the two groups? Of course!

Right away, you start noticing some interesting patterns. When you plot the age distributions of customers on a histogram for the two categories, you notice that there are a large number of customers in their early 30s who seem to buy the product and far fewer customers in their 20s. This is surprising, since the product targets people in their 20s. Hmm, interesting …

Furthermore, many of the customers who convert were targeted via email marketing campaigns as opposed to social media. The social media campaigns make little difference. It's also clear that customers in their 20s are being targeted mostly via social media. You verify these

assertions visually through plots, as well as by using some statistical tests from your knowledge of inferential statistics.

The next day, you walk up to the VP of Sales at her desk and show her your preliminary findings. She's intrigued and can't wait to see more! We'll show you how to present your results to her in our next section.

Step 5 of 6: Analyze Your Data In Depth

In the previous section, we explored a dataset to find a set of factors that could solve your original problem: predicting which customers at hotshot.io will buy the product. Now you have enough information to create a model to answer that question.

In order to create a predictive model, you must use techniques from machine learning. A machine learning model takes a set of data points, where each data point is expressed as a *feature vector.*

How do you come up with these feature vectors? In our EDA phase, we identified several factors that could be significant in predicting customer conversion, in particular, age and marketing method (email vs. social media). Notice an important difference between the two factors we've talked about: age is a numeric value whereas marketing method is a categorical value. As a data scientist, you know how to treat these values differently and how to correctly convert them to features.

Besides features, you also need labels. Labels tell the model which data points correspond to each category you want to predict. For this, you simply use the CONVERTED field in your data as a boolean label (converted or not converted). 1 indicates that the customer converted, and 0 indicates that they did not.

Now that you have features and labels, you decide to use a simple machine learning classifier algorithm called logistic regression. A classifier is an instance of a broad category of machine learning techniques called '*supervised learning,*'where the algorithm learns a model from labeled examples. Contrary to supervised learning, *unsupervised learning* techniques extract information from data without any labels supplied.

You choose logistic regression because it's a technique that's simple, fast and it gives you not only a binary prediction about whether a customer will convert or not, but also a probability of conversion. You apply the method to your data, tune the parameters, and soon, you're jumping up and down at your computer.

The VP of Sales is passing by, notices your excitement and asks, *"So, do you have something for me?"* And you burst out, *"Yes, the predictive model I created with logistic regression has a TPR of 95% and an FPR of 0.5%!"*

She looks at you as if you've sprouted a couple of extra heads and are talking to her in Martian.

You realize you haven't finished the job. You need to do the last critical step, which is making sure that you communicate your results to your client in a way that is compelling and comprehensible for them.

Step 6 of 6: Visualize and Communicate Your Findings

You now have an amazing machine learning model that can predict, with high accuracy, how likely a prospective customer is to buy Hotshot's product. But how do you convey its awesomeness to your client, the VP of Sales? How do you present your results to her in a form that she can

use?

Communication is one of the most underrated skills a data scientist can have. While some of your colleagues (engineers, for example) can get away with being siloed in their technical bubbles, data scientists must be able to communicate with other teams and effectively translate their work for maximum impact. This set of skills is often called *'data storytelling.'*

So what kind of story can you tell based on the work you've done so far? Your story will include important conclusions that you can draw based on your exploratory analysis phase and the predictive model you've built. Crucially, you want the story to answer the questions that are most important to your client!

First and foremost, you take the data on the current prospects that the sales team is pursuing, run it through your model, and rank them in a spreadsheet in the order of most to least likely to convert. You provide the spreadsheet to your VP of Sales.

Next, you decide to highlight a couple of your most relevant results:

- **Age:** We're selling a lot more to prospects in their early 30s, rather than those in their mid-20s. This is unexpected since our product is targets people in their mid-20s!
- **Marketing methods:** We use social media marketing to target people in their 20s, but email campaigns to people in their 30s. This appears to be a significant factor behind the difference in conversion rates.

The following week, you meet with her and walk her through your conclusions. She's ecstatic about the results you've given her! But then she asks you, *"How can we best use these findings?"*

Technically, your job as a data scientist is about analyzing the data and showing what's happening. But as part of your role as the interpreter of data, you'll be often called upon to make recommendations about how others should use your results.

In response to the VP's question, you think for a moment and say, *"Well, first, I'd recommend using the spreadsheet with prospect predictions for the next week or two to focus on the most likely targets and see how well that performs. That'll make your sales team more productive right away, and tell me if the predictive model needs more fine-tuning.*

Second, we should also look into what's happening with our marketing and figure out whether we should be targeting the mid-20s crowd with email campaigns, or making our social media campaigns more effective."

The VP of Sales nods enthusiastically in agreement and immediately sets you up in a meeting with the VP of Marketing so you can demonstrate your results to him. Moreover, she asks you to send a couple of slides summarizing your results and recommendations so she can present them at the board meeting.

Boom! You've had an amazing impact on your first project!

You've successfully finished your first data science project at work, and you finally understand what your mentors have always said: data science is not just about the techniques, the algorithms or the math. It's not just about the programming and implementation. It's a true multi-disciplinary field, one that requires the practitioner to translate between technology and business concerns. This is what makes the career path of data science so challenging, and so valuable.

If you enjoyed reading this and are curious about a career in data science, check out some of our awesome programs and resources:

- **Data science career track**
- Data science interview guide

Also, I've written answers to several related questions on Quora that might be helpful.

AUTHOR

Raj Bandyopadhyay