Data Scientists Jobs and Careers in Data Science +6

What are some actual projects that data scientists have worked on? What tools and analytical techniques were used, and what mistakes were made?

This question previously had question details. You can find them in the question comments.

Answer Request ▾ Follow 470 Comments 3 Downvote

8 Answers

William Chen, Data Science Manager at Quora
Updated Jan 14, 2014 · Upvoted by Lili Jiang, Data Scientist at Quora and Joe Blitzstein, Professor in the Harvard Statistics Department

It looks like you're looking for the complete end-to-end of a data science project, complete with code, reasoning, thought processes, and dead ends.

It's hard to find industry projects that can be shared to that degree (mine certainly can't). But perhaps you can find some student projects.

I wanted to share the project done by team **Buffalo Capital Management** (Me, Sebastian Chiu, Salena Cui, Carl Gao) for the 2013 Harvard Data Science course taught by Joe Blitzstein and Hanspeter Pfister.

Our project was to **predict directional movement of stock prices**. While not a standard data science problem, it fits in with the **data science process** quite nicely and will give you some insight in how a data science project is completed from end to end.

Related Questions

What is the most interesting project(s) you've worked on as a data scientist?

What are some of your best projects you worked on as a data scientist?

What are some examples of projects that Data Scientists at Disney have worked on?

What is the most common project to work as a data scientist in any company?

What is the actual work of a data scientist/ analyst?

What are some mathematical models or statistical techniques that are useful for scientists working with big data?

What is the workflow or process of a data scientist? What tools do they use?

What algorithms do data scientists actually use at work? What are good resources to learn the same?

As a data scientist, how I am suppose to work on a project without any data?

What can a data scientist create in 1 hour, 1 day, 1 week, or 1 month? What problems can they tackle? What tools would they use?

More Related Questions

Question Stats

469 Public Followers

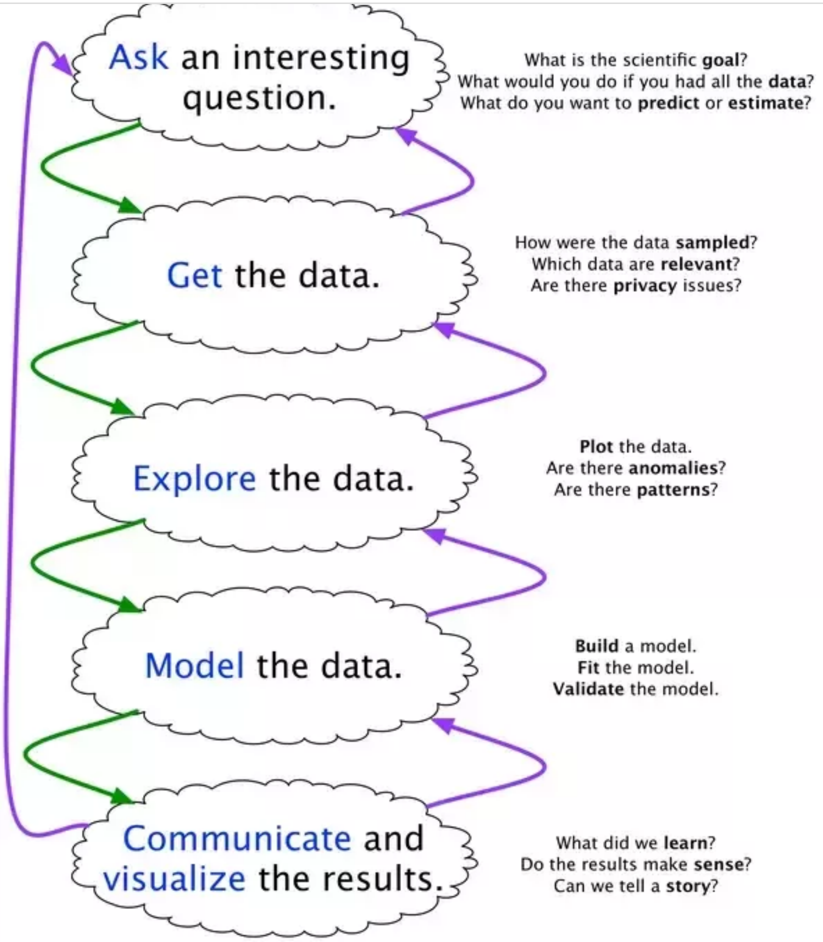
63,453 Views

Last Asked Aug 7

3 Merged Questions

Edits

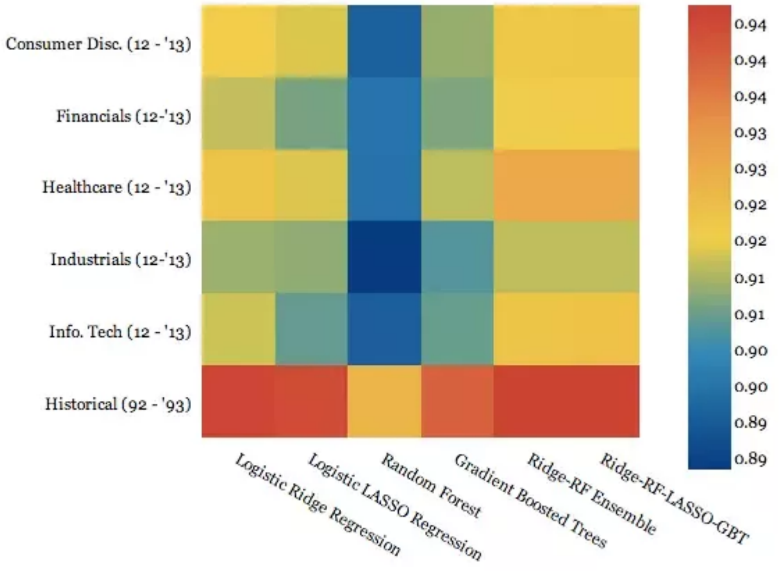
Home 1 Answer Notifications Search Quora Ask Question


Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Ultimately, we found that directional movement of stock prices for blue chip stocks was **highly predictable given today's opening price and open/close prices for the previous 9 days**. Directional movement was predictable with over 90% AUC. We were able to win a predictive modeling competition (2013 Boston Data Week) with our Ridge-RF Blended model.

Upvote 406 Downvote



Our **video** briefly summarizes our motivations, results, and main takeaways.

Our **IPython process notebook** completely outlines our reasoning and thought processes behind each step.

Our **GitHub** contains all of our code.

And lastly our **website** ties everything together.

For more projects like these from our classmates, simply **search on YouTube for CS109** !

Hope that's what you were looking for. Good luck in your data science ambitions!

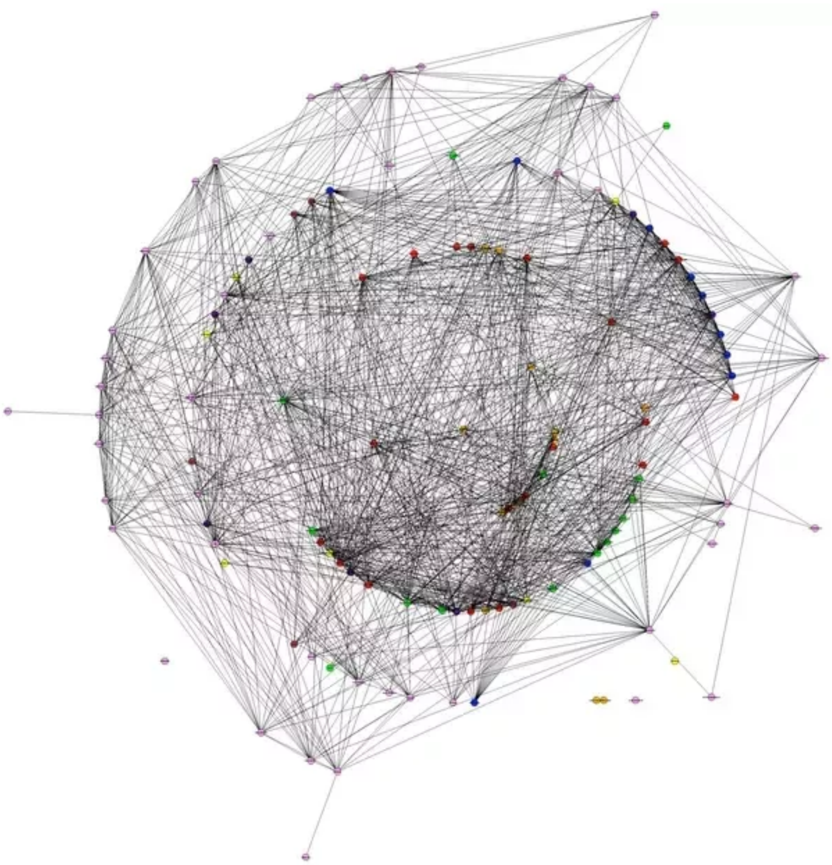
28.8k Views · 406 Upvotes · Answer requested by Neal Wu and Rohan Patil

Ted Conbeer
If this is true... why are you doing anything other than trading blue chip stocks right n...

1 more comments from Nishant Gupta

Bradley Voytek, Former Data Scientist, Uber Inc.
Answered Jan 12, 2014 · Upvoted by Alan Yiu, Data Scientist @ Quora and Don van der Drift, Quora Data Scientist

While I've done a lot of little projects for Uber (company), some of which I make public over on their Uber Blog (#uberdata) , for the purposes of this question I think my brainSCANr project probably fits best. Note that Uber, brainSCANr, and my actual experimental neuroscience research really inform one another in amazingly wonderful ways.



- brainSCANr website
- Automated cognome construction and semi-automated hypothesis generation paper

The Problem

The opening lines of the paper:

> The scientific method begins with a hypothesis about our reality that can be tested via experimental observation. Hypothesis formation is iterative, building off prior scientific knowledge. Before one can form a hypothesis, one must have a thorough understanding of previous research to ensure that the path of inquiry is founded upon a stable base of established facts. But how can a researcher perform a thorough, unbiased literature review when over one million scientific articles are

> calculating associations between concepts in the peer-reviewed literature, we can algorithmically synthesize scientific information and use that knowledge to help formulate plausible low-level hypotheses.

Inception Stage

In May 2010 I was invited to speak at Berkeley's Cognitive Science Student Association (CSSA) Conference. At that conference I sat on a Q&A panel with a hell of a group of scientists, including my friend and colleague George Lakoff and the (then) Chair of Stanford's Psychology department, James McClelland, who helped pioneer Parallel Distributed Processing.

On that panel I A'd many Qs, one of which was a fairly high-level question about the challenge of integrating the wealth of data hidden in the neuroscientific literature. It was a variant on the classic line that neuroscience is "data rich but theory poor". This is a problem I've been struggling with for a long time and I'd had a few ideas.

In my response I said that one of our problems as a field was that we had so many different people with different backgrounds speaking different jargons who aren't effectively communicating. I followed with an off-hand comment that "The Literature" was actually pretty smart when taken as a system, but that us individual puny brains just weren't bright enough to integrate it all.

I went on to claim that, if there was some way to automatically integrate information from the peer-review literature, we could probably glean a lot of new insights. James McClelland *really* seemed to disagree with me, but the idea kept kicking around my brain for a while.

Creation

One night, several months later (while watching Battlestar Galactica (2003–2009 series) with my wife Jessica Bolger Voytek), I turned to her and explained my idea. She asked me how I was planning on coding it up and, after I explained it, she challenged me by saying that she could definitely code that faster than I could.

Fast-forward a couple of hours to around 2am and she had her results. I did not.

Bah.

The idea I discussed with her was *very* simple (and probably simplistic) and was based on the assumption that the more frequently two neuroscientific terms appear in the title or abstracts of peer-reviewed papers together, the more likely those terms are to be associated with one another.

For example, if "learning" and all of its synonyms appears in 100 papers with "memory" and all of *its* synonyms while both of those terms appear in a total of 1000 papers without one another, then the probability of those two terms being associated is 100/1000, or 0.1.

We calculated such probabilities for every pair of terms using a dictionary that we *manually curated*. It contained 124 brain... (more)

Upvote 247 Downvote

Steve Carnagua
Now that I've taken the time to read what you two actually did, I'm amazed at how eas

Thia Kai Xin, Data scientist at Lazada, Co-Founder of DataScience SG.
Answered Mar 2, 2016

Great question. There are lots of details that data scientists cannot share in a public forum, but I can share some of my thoughts on a recent project.

1. All data science projects start with a business or research question. The problem statement of this project is to predict the probability of cancelations and returns for products sold through the eCommerce platform.
2. With the business problem in mind, we started examining the data available, the current solution we have, accuracy we are targeting and timeline for the project.
3. Initially, I saw this as a pretty standard, straightforward question. I took historical data, train several models (boosted trees, randomForest, linear models) and chose the one with the highest cross-validation accuracy (boosted trees). To further improve the accuracy, I created new features that brought in the historical return rates of the product and buyer.
4. The model was 4x faster and 2x more accurate than the current solution we had. So I submitted the model and lived happily ever after...not really. If only life was so easy.
5. What they did not teach you in school and Kaggle is that model deployment is difficult.

- Through the model building process, I had to switch databases multiple times due to internal migration. The data, while similar, is not identical across the databases because of the schemas, update frequency and a host of various issues. We fixed that by creating a virtual or staging table to hold the columns needed by the model. We also painstakingly checked the data sources to ensure consistency.
- Next, the model accuracy fluctuates across the different countries we operate. Some countries have longer delivery time that results in more cancellation. Some have a high number of cash on delivery orders that are prone to cancellation. In essence, each country was not only geographically separated, but they were also different regarding external factors like purchase behavior, credit card penetration rate, local promotions, etc. Thus, we had to extend the model to include a decay factor for long shipping items, bring in business rules to adjust the prediction specifically for cash on delivery and train separate models for each country.
- Finally, the most difficult part is handling unexpected downtimes. Databases can go down. Connection can fail. The server can run out of resources. Queries can timeout. As a wise man said, the best way to deal with errors is to log everything and that's what we did. We also added auto restart feature on the script and backup of predictions to flat files so we will never lose any data.

All in all, the job of a data scientist is not just in making fancy models.

If you like to hear more war stories, many companies have detailed technical blogs by data scientists, here are some of the best I know:

- Facebook Research Blog

- The Unofficial Google Data Science Blog
- research | Twitter Blogs
- Yahoo Research
- Yahoo Engineering

2.7k Views · 25 Upvotes

Upvote 25 Downvote

Add a comment... Recommended All

Top Stories from Your Feed