Retrospective on “On Moving From Statistics to Machine Learning”

Posted on July 4, 2020 by W.D.

My old piece is getting traction thanks to a share on Hacker News, where some of the most insufferable tech guys in California try to dissect in the comments whether I have deep-seated psychological issues.

> This sort of fashionable disparagement of a group of people to signal that you’re not part of the “bad group of tech bro’s” is so trashy. Why are these random people you easily hate? Who are they? Why take glee in shared hatred?

Also, I was mentioned in this blog post at Win Vector LLC, which offers a fair and very good critique of what I wrote, although I should note that I’ve never uttered the post’s titular phrase.

I’m happy people are sharing my post, but I won’t know whether I’ve truly made it until I see myself on n-gate.

It’s been about a year since I’ve written that piece. Where to begin?

A Bone to Pick

My first and biggest regret is the title of the piece. I read somewhere that pieces with longer titles get more engagement on average; appending “the final stage of grief” to the title was my attempt to cash in on that in a campy pseudo-poetic way.

The stage of grief is me coming to terms with moving on from mere “statistics” to wondrous “machine learning.” There was some grief involved in making concessions to the apparition of tech guys because if you haven’t

noticed, a large part of my brand is being skeptical of tech, tech industry, and tech industry culture. But I'm not sure I ever experienced any grief or alienation in having a "data scientist" job title. My work has not changed much other than that my code shifted from being pure Pandas to incorporating some engineery cloud stuff like Kubeflow, I write more SQL, and the data sets I work with are a couple orders of magnitude larger.

For context, a lot of my 2019 writing was coming from a place of resentment at struggling to get a data science job despite thinking I was qualified. My writing was me trying to make sense of what I was getting wrong on my resume and in interviews, digging into topics I wasn't familiar with, and coming up with ways to market myself so I could make more money and escape the trap where I continually get sucked into jobs that require no more than knowing how to perform a ~~VLOOKUP~~ XLOOKUP in Excel. There's nothing wrong with those sorts of jobs; knowing how to write a little Python code doesn't make you better than someone who works primarily in Excel. I had just felt like I knew enough Python at that point to take part in the tech industry gold rush.

The fact that I now have a "data scientist" job and am doing just fine is *prima facie* evidence that my resentment at failing to break into these jobs was mostly warranted. There's no way to expound on this without coming off as haughty, so forgive me for not going into much detail, but basically: I have a "data scientist" job at a medium-sized company, and I seem to do well enough at it to not get fired. But who knows, maybe I'll get fired in a few weeks for poor performance, and anyone who feels insulted or annoyed by me can revel in the schadenfreude.

> The author starts with
>
> *The data science world may reject me and my lack of both experience and a credential above a bachelors degree*
>
> More likely the data science world will reject him because he is so confident a field he has so little experience or knowledge of.

Data science jobs are often just cleaning data, joining data together, summing or averaging stuff in a GROUP BY, and doing some light coding tasks. If you are good at this stuff, you will probably do pretty well at the median data science job. When data science people claim in unison that their jobs are more complicated than this, a nontrivial chunk of them are either lying to gatekeep outsiders and protect their inflated salaries, or they are confessing that they're genuinely not very smart and these banal tasks are pushing the limits of their abilities. And sure, maybe a few of them genuinely have ornate jobs where knowing how to barrage your data with pre-built fancy-sounding algorithms is a real core competency and not hollow posturing.

I'm also sure a lot of people who will think I'm "telling on myself" by describing the median data science role as banal work, but...

Who is Data Science, Exactly?

My second regret is that I really under-appreciated how vague of a term "data science" is. Five data scientists can have five different jobs with the only similarity between them being "I write code relating to data." Even at organizations large enough that you'd expect each employee to wear one hat and not five hats, a data scientist might flip between various tasks regardless: writing API wrappers to parse JSONs, doing some SQL to help with accounting/book-keeping stuff, straight up software engineering, and yes, sometimes you work on machine learning and statistical models.

The stereotypical techies on Hacker News in the comments seem to believe that because sometimes you have to write code bordering on software engineering as part of a data science job, this is evidence that mere statisticians are not ascended enough for the hard knock life of being a data scientist. As I've written before, this is a very commonly employed double standard in the tech industry. It is a double standard because none of these techies believe this standard applies in any other scenario besides writing

code. For example, many tasks data scientists involve doing things that overlap with or are straight up actuarial science, behavioral science, finance, economics, social psychology, sociology, and more. Tech people do not think their lack of qualifications in these fields precludes them from being paid exuberantly to do half-assed jobs juggling these subject matters.

All About the Benjamins

Despite that data science jobs are very heterogeneous, I do *not* regret describing data science as statistics for tech bros, and I do not regret cynically encouraging stats people to call what they do "machine learning" and "data science" to score a higher salary.

A lot of tech employees do mental gymnastics to justify why they make boatloads of money. (Similarly, a lot of tech *companies* do mental gymnastics to justify why they should be funded without any sensible plan of becoming profitable, or why they should even be considered "tech" companies in the first place.) Case in point, there are a handful of people in the Hacker News comments opining about why I shouldn't make as much money as them.

Rich people of all stripes, tech industry or otherwise, will tell you they're rich because they're smart, worked hard, and made all the right choices. The reason you're not rich is supposedly because you don't meet those three criteria.

This is wrong. 100% of people alive today make however much money they do mainly because of luck and circumstance. There's occasionally some effort required, don't get me wrong, but it's peanuts compared to luck. 500 years ago, you would most likely have been either born a peasant or born into royalty, with zero chance of upward mobility if you're in the former category. The fact that you weren't born 500 years ago is itself a stroke of luck.

We don't need to operate in centennial units to appreciate the luck involved in being an American tech worker at this particular time: 30-40 years ago, the vast majority of computer coding jobs did not exist and the few that did were not well funded by a bloated VC industry. 10 years from now, there's no telling whether the sudden shift toward remote work induced by coronavirus will increase the salience of off-shoring, consequentially destroying most lavishly paid tech jobs and driving down salaries.

The economy, so claims the neoclassical model, determines wages based on the marginal product (i.e. contributions to corporate revenue) of people's labor, not based on meritoriousness, one's intelligence, one's grit, or any ethical imperatives. That said, obviously this is not a perfect economic model, otherwise I would not be talking about how you can trick VCs into paying you an extra $20k/year by uttering some magic words in interviews.

If Your Boss Lets You Fuck Around, You'll Never Work a Day in Your Life

In the old piece, I wrote: *"How often are we doing fancy neural network stuff to impress others and feel smart rather than to elucidate important phenomena, which sometimes requires being boring and using OLS and some old-school stuff?"*

Which leads me to my last regret. I regret suggesting the use of OLS here, which is often way too advanced and not boring enough. Very often you just want to cross-tabulate data. If you are a data scientist, be sure to add color to it so people don't question why you get paid the big bucks. I leave it up to readers' discretion whether they'll keep the docstring when they add this function to their `utils.py`:

```
1
2
3
```

```
import pandas as pd
import seaborn as sns
def heat_map(
        df: pd.DataFrame,
        color: str = 'lime',
        fmt: str = '{:.0f}'
) -> 'pd.io.formats.style.Styler':
    cmap = sns.light_palette(color, as_cmap=True)
    return df.style.format(fmt).background_gradient(cmap=cmap, axis=None)
df = pd.DataFrame({
    'x': range(1, 4),
    'y': range(4, 7),
    'z': range(7, 10)
}, index=['a', 'b', 'c'])
heat_map(df)
```

28	
29	
30	

The importance of cross-tabulation is something I learned very early on in my career, and was the side-effect of getting verbally beat down by the PhD economist I was working for. I do not recall the specifics of this story other than that we were doing work-product stuff for a law firm's client, and the client was on a tight budget. I was not following her orders directly because I was trying to be fancy and because I was arrogant. She was upset and explained to me that doing fancy fun stuff isn't what we're being paid to do, and the stuff that's most easily defended in court and explained to jurors, lawyers, judges, and clients are simple things like group averages.

You see, I was a very arrogant and adventurous kid who liked to run regressions and download MATLAB code for fancy economics models and stick data in it. Whenever I got data, I wanted to mess around with it and do as much as I could with it. I'm still arrogant, but my adventurous side has been beaten out of me. I've come to terms with the facts that (A.) running other people's fancy models and pre-built algorithms doesn't make me smart or cool; and (B.) fancy models aren't useful in a lot of contexts.

I think data science tends to select for those types of people: younger starry-eyed adventurous and arrogant types. I'm not sure that many data science people actually get the adventurousness beaten out of them, though, and that's not their faults. Scrum is the way you make your engineers actually get stuff done and it's an effective system once you get over all the jargon like "Kanban board" and "scrum poker." But scrum doesn't work well in a lot of data science contexts, so reasonably, data science teams at many companies tend to operate somewhat orthogonally to an agile sprint. Also, management seems to think of data science as fucky magic stuff because hey, oftentimes it is, but that's used as an excuse if they never provide things of tangible value.

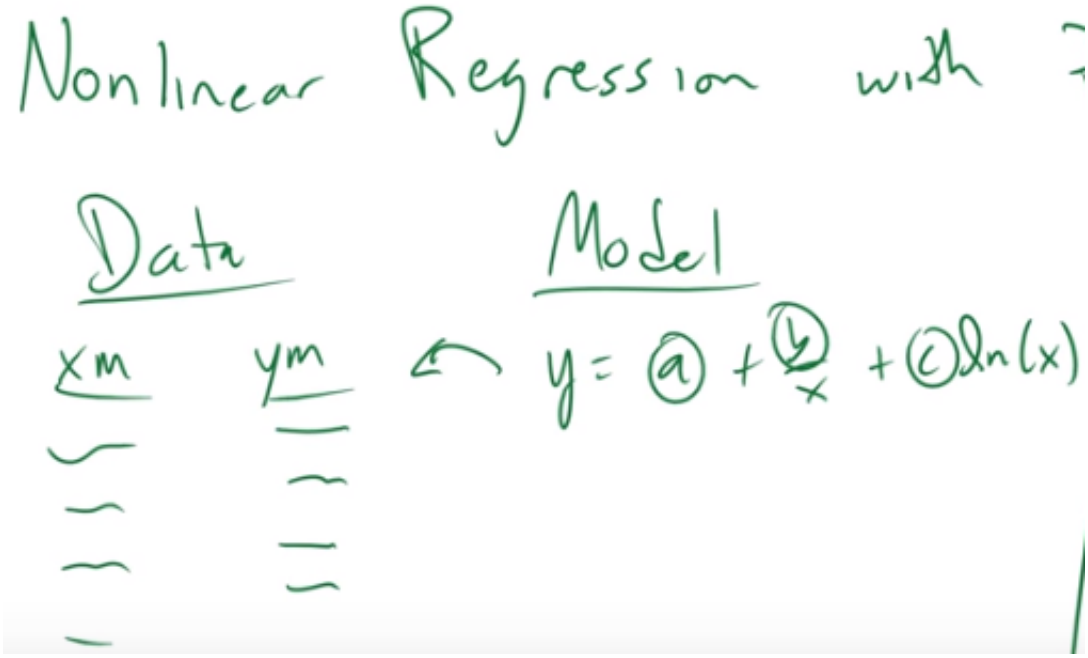
The end result is a system where data scientists can get away with exclusively doing stuff designed to beguile peers, managers, and investors rather than doing anything useful (in the fiduciary, capitalist sense of the term "useful"). I am not slyly referencing my current workplace; I am referencing stories from other data scientists as well as the previous two places I worked, where data scientists were employed and seemingly nobody knew what the data scientists were doing. Long story short for my current company: our CEO is a data science skeptic (maybe he reads my blog). So our company's data science culture reflects a concerted effort to provide tangible value and convince his skeptical side, probably more so than the typical data scientist's workplace.

It's hard to blame the data scientists for these arrangements because it favors them. It's genuinely really awesome to mess around and not need to provide capitalistic value, and it would be neat if more jobs were like that. But it also explains a large chunk of why so much data science is silly voodoo: they didn't have a 65 year-old PhD economist with decades of expert testimony experience to shame them about not doing things that are boring but effective.

I think that's the *main* reason why so much data science is silly voodoo, but there's more to it than that.

A lot of data science is silly voodoo because the incentive at many places is to produce silly voodoo. If you are being hired to dazzle managers, then being boring risks you getting fired.

A lot of data science is silly voodoo because people don't seem to know that simple models can accommodate a lot of use cases. A general rule of thumb, not just in data science but in life, is people who don't know what they're talking about tend to think things need to be more complicated than they actually have to be. For example, here is an example from a Youtube tutorial on nonlinear least squares (NLS) where the instructor doesn't know that you can trivially perform a linear regression on this equation:



Lastly, a lot of data science is silly voodoo because people are lazy. This is not a data science thing, it's just a *thing*. A fundamental facet of the human condition is trying to do less work. Much like I can't blame people for being adventurous instead of productive, I also I can't blame people for being lazy; it's wasteful to squander the privilege of being able to slack off.

A lot of machine learning models are OK substitutes for putting in effort because they're powerful. Neural networks are juggernauts. Just stick some layers into a Keras Sequential class, stick in your unclean data, and watch Tensorflow do its magic.

If you're presenting to management, you can simply spout all the classification algorithms you know in 60 seconds and add a few violin plots and correlograms without explaining the reason you included them; this is also a fine substitute for effort. Sidebar: I get why people do this to management and I will never fault anyone for doing this to management. But it is insulting when data scientists do this shit to an audience of other data scientists. Nobody who does this stuff on the regular is impressed at you speedrunning your dataset through the Scikit-Learn API.