Recession-Proofing Data Scientists

Melton Scholar: Isaac Sheets[a*], Faculty Advisor: Yuanyang Liu[bc]

[a]Business Analytics and Statistics, University of Tennessee, Knoxville

[b]Business Analytics and Statistics, University of Tennessee, Knoxville

Abstract

In 2020 Q2, job markets across the world were hit hard. This created an interesting question: who was still hiring and why? A particularly interesting trend was noted when evaluating the data science job market.

In 2020 Q2, the top 5, 10, and 20 data science employers significantly increased their percentage of data science job postings compared to the rest of the data science job market. Understanding this spike is vital to identifying traits that can recession-proof data scientists in the future. It can enable data science educators, analytics managers, and even individual data science practitioners to take appropriate steps to ensure that they are ready for the next economic curve ball thrown their way.

Two hypothesis were proposed to explain this spike:

1.) Companies emphasizing the prevalence of social skills in their data scientists were able to ensure that their data scientists remained marketable and/or economically viable during the recession; and 2.) The analytic maturity (for which proportion of total 2019 jobs that were computer-related was used as a proxy) of the top 20 data science employers helped recession-proof the demand for their analysts by enabling these departments to be able to provide valuable prescriptive analytic solutions (as opposed to descriptive analytics) to problems faced by their respective companies during the recession.

In order to test the first hypothesis, a ratio was calculated for each firm with 5 or more data science job postings in 2019. The ratio represented the number of social skills terms listed in all of the texts of every data science job posting for said company in 2019, divided by the number of data science job postings for said company in 2019 (number of social skill terms/number of job postings). There was a statistically significant and positive correlation between this variable and the number of data science job postings in 2020 Q2.

In order to test the second hypothesis, for each company, the proportion of all 2019 data science job postings that were computer-related was calculated. There was also a statistically significant and positive correlation between this variable and the number of data science job postings in 2020 Q2.

1 Introduction

Are there certain occupations that are recession-proof? History has certainly proven that there are, but, with each recession, the list of occupations that are impervious to said recession's effects fluctuates slightly.

Since 2008, one of the most rapidly evolving industries in the world has been data science. This industry has never been fully tested to see how it would hold up in a global recession, until Quarter 2 of 2020.

Frankly, given the nature of the COVID recession, there couldn't have been a better type of recession to test the durability of the data science industry. Data science can be performed remotely at any place, at any time, and anywhere where the practitioner has a stable, secure internet connection.

Even with all those factor's in its favor, data scientist employment rates still suffered in 2020 Quarter 2. However, not all data science employers were hit equally as hard. For the top 20 data science employers, their percentage of the total number of data science job postings increased by approximately 7%. This statistic then begs the question: Why did the top 20 data science employment rates differ so greatly from the rest of the industry?

Do these employers emphasize certain skill sets that are making their data scientists recession proof? Is it possible that the maturity of these companies' analytics departments help stabilize demand for their data scientists? Could it be that these companies have more resources to withstand the recession, and they are using the recession as an opportunity to acquire top data science talent for cheap?

Understanding the answer to these questions is vital for all individuals and entities with ties to the data science industry. If the impervious nature of data scientists at these companies can be attributed to their possession of specific skill sets not found among data scientists across the rest of the industry, then school administrators need to adjust course curriculum to ensure that their graduates posses these skills. If these data science practitioners are more recession-proof because of increased analytical maturity at their company, then it might behoove other data scientists across the industry to speed up the analytical maturation process at their respective companies. If companies are just using the recession as an opportunity to stock up on top talent for cheap, then economists have to ask the question: How will this impact the data scientist job market for years to come?

1.1 Interesting Data Science Job Market Trend

As can be seen in Figure 1, the percentage of data scientist job postings that were posted by the top 20 largest data science employers has been steadily decreasing for quite a while. However, in Quarter 2 of 2020, that percentage spiked substantially. What were the driving factors of that spike? The answer to this question can have significant impacts for the data science industry going forward.

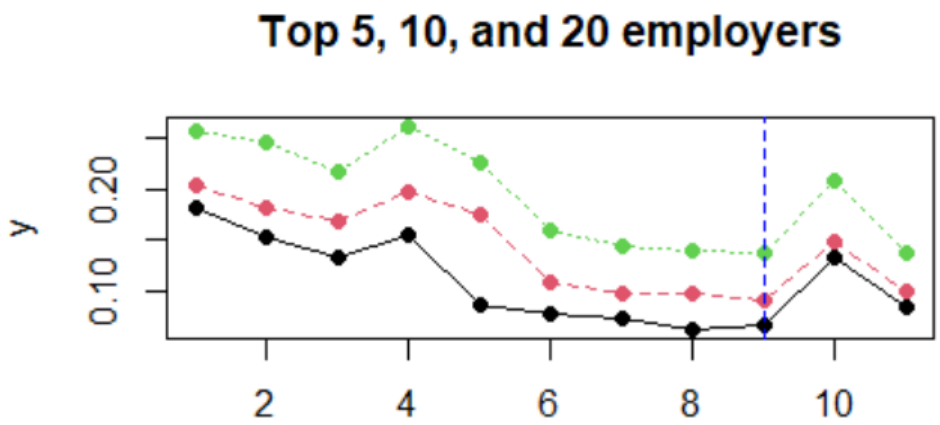

Figure 1: This chart shows the percentage of all job postings (y-axis) that were posted by the top 5 (black line), 10 (red line), and 20 (green line) data science employers, from Q1 2018 through Q3 2020 (x-axis). The blue line is at Q1 2020.

1.2 Question

The purpose of this paper is to identify what caused the aforementioned spike in Figure 1. More specifically, we will be quantitatively testing the following 2 hypothesis:

- Top data science companies require their employees to posses strong social skills. These social skills help data scientists communicate their importance to c-suite executives within the company and, in some cases, the company's customers.
- The analytics department of top data science employers is mature enough to be able to provide critical insights necessary to ensure that company executives manage the company's crisis in the most effective manner.

1.3 Results

The two main findings are that:

- There is a positive, statistically significant correlation between proportion of social skill terms and data science job posting rates for companies.
- There is a positive, statistically significant correlation between percentage of job postings in 2019 that were for computer-related positions and data science job postings in Q2 2020.

2 Theoretical Background

2.1 Previous papers on social skill

We used the following two papers to construct our "social skill" variables.

- Ham et al. (2020): This paper calculated the percentage of specific skills for audit job postings as compared to the total number of audit job postings for each accounting firm. It then compared the skill percentages for each firm to the overall quality of the audits for each firm.[1]

- Deming (2017): Researchers compared the performance of jobs requiring math skills and jobs requiring social skills over the years of 1980-2012. It concluded that jobs that emphasized social skills were growing in importance; while those that didn't require social skills were declining in importance. It determined the skills associated with each job by utilizing the ONET database which paired specific skills with specific occupations.[2]

Our results add to people's understanding in terms of the importance of social skill in labor market for data scientists.

3 Data

3.1 Burning Glass Data

In our analysis, we pulled job postings from Burning Glass, a comprehensive job posting database. We analyzed 800 different companies with $\geq$ 10 data scientist postings in 2019. See Ham et al. (2020), Forsythe et al. (2020) for more information on BGT data.

3.2 Variables

- Y_f: Our predictor variable was the natural log of number of 2020 Q2 data science job postings for a firm f.

- pSocial$_f$: This variable stood for the proportion of social skill terms. To calculate this variable we took the sum of the number of times a social skill was used in all data science job postings in 2019 for each company, and we divided it by the total number of data science job postings for each company.

- pComputer$_f$: This variable stood for the proportion of computer-related job postings. To calculate this variable, for each company we took the number of computer occupation job postings (SOC number starting with 15-11XX). Then, we divided that number by the total number of job postings for said company in 2019. We used this variable as a proxy for analytical maturity.

[1] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3727495
[2] https://academic.oup.com/qje/article/132/4/1593/3861633

- Controls: There were three main controls used when developing our models. These controls were calculated individually for each company.
  - Total job postings.
  - Industry classification.
  - Proportion of all 2019 job postings that were data science postings.

4 Empirical Analysis

4.1 Linear Regression

Regression model:

$$Y_f = \beta_0 + \beta_1 \cdot \text{pSocial}_f + \beta_2 \cdot \text{pComputer}_f + \beta_3 \cdot \text{Controls}_f \tag{1}$$

Results (Table 1):[3]

Table 1: Regression models. This table shows .....

	Model 1	Model 2	Model 3	Model 4
pSocial	0.04*		0.04*	0.02
	(0.02)		(0.02)	(0.02)
pComputer		0.01*	0.01*	0.01***
		(0.00)	(0.00)	(0.00)
pDS				0.04***
				(0.01)
Total				0.48***
				(0.03)
Industry	Yes	Yes	Yes	Yes
Adj. R^2	0.03	0.03	0.04	0.28
Num. obs.	801	801	801	801

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

When evaluated individually, both pSocial and pComputer appear to have a positive, statistically significant correlation to the number of job postings in Q2 2020. However, when they are both evaluated in the same model along with the control variables, we see that the pSocial term becomes no longer statistically significant. Setting aside the near miss of statistical significance for the pSocial term, the regression output for the model described above seems to suggest:

- pSocial: For every 0.01 increase for a firm, we would expect that firm to increase the number of data science job postings by 2 percentage points.
- pComputer: For every one percentage point increase for a firm, we would expect that firm to increase the number of data science job postings by one percentage point.

[3]Save R regression output to latex table: (1) I use "texreg" https://cran.r-project.org/web/packages/texreg/index.html, (2) people also use "stargazer" https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

4.2 Poisson Regression

Given the fact that the job posting data consists of non-negative integers (i.e. you can't have a negative number of job postings), a Poisson regression model was fit to the data in an effort to find a better fitting model.

- Regression equation:

$$log(Y_f) = \alpha + \beta_1 \cdot \text{pSocial}_f + \beta_2 \cdot \text{pComputer}_f + \beta_3 \cdot \text{Controls}_f \tag{2}$$

- Results table:

	Model 1	Model 2	Model 3	Model 4
Prop_SST	0.03*		0.03*	0.01
	(0.01)		(0.01)	(0.01)
pComputer1		0.00*	0.00*	0.01***
		(0.00)	(0.00)	(0.00)
pDS				0.02**
				(0.01)
Total1				0.35***
				(0.03)
Industry	Yes	Yes	Yes	Yes
Deviance	840.95	840.86	835.80	664.90
Num. obs.	800	800	800	800

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Poisson Regression Results

The results in Table 2 also show that both pSocial and pComputer appear to have a positive, statistically significant correlation to the number of job postings in Q2 2020, when these two terms are individually evaluated. However, when they are both evaluated in the same model along with the control variables, we again see that the pSocial term becomes no longer statistically significant. Setting aside another near miss of statistical significance for the pSocial term, the regression output for the model described above seems to suggest:

- pSocial: For every 0.01 increase for a firm, we would expect that firm to increase the number of data science job postings by one percentage point.
- pComputer: For every one percentage point increase for a firm, we would expect that firm to increase the number of data science job postings by one percentage point.

4.3 Matching

The goal with Matching was to find a pair of two firms that are similar in control variables and different in PropSST. Then, calculate their difference in Y. This was done through two different methods. The first method was a simple KNN method, and the second was the synthetic control method.

4.3.1 K-nearest neighbor matching

When performing KNN analysis, the following steps were implemented:

1. Given a industry, find a *treated* firm with the highest pSocial.
2. For the same industry, the *control* firms are those with pSocial smaller than the industry median.
3. For a *treated* firm, find the most similar control firm in all the control variable values.
    - X: (1) pComputer1 (2) pDS (3) Total
    - Use Euclidean distance function to compare two firms' X vector similarity.
    - Find K-nearest neighbor, with K=1.
    - Record the *treated* and *control* Y.
4. A total of 16×2 observations were generated (16 industries each have 1 treated firm and one control firm).
5. A paired t-test was conducted on the number of Q2 2020 job postings for the treated and control firms in each of the 16 different industries.
6. This process was repeated for the pComputer variable as well.

    The p-vaules for both the pSocial ad the pComputer variables were well above 0.05 (0.44 for pSocial and 0.51 for pComputer), and no statistically significant results were generated. However, it was noted that, for a given treated high pSocial/pComputer firm, event its nearest neighbor control firm was often quite different in their X vector (see 3 control variables above). Therefore, Synthetic Control firms were constructed from all control firms (firms with below-median pSocail/pComputer values for their respective industries).

4.3.2 Synthetic control matching

For each treated firm, a linear combination of control firms was created to form a *synthetic control* firm. The trick was to find what weight to ascribe to each firm in the control set in order to create the optimal match to the treated firm. In order to do this, the following steps were completed:

- Each firm is represented by a vector of covariates: $\mathbf{x} = (x1, x2, x3)$.
- Firm A: the treated firm. $\mathbf{x_A} = (x[A, 1], x[A, 2], x[A, 3])$, one row in the data.
- Firm B,; Firm C, Firm D, three control firms (three rows in the data).
- The goal was to find weight for Firm B, C, D, and so forth such that the weighted combination of the control firms is closest to the treated Firm A's covariates.
  - Weights: w_B, w_C, w_D
  - Synthetic control unit:

$$\mathbf{x}_{synth} = w_B \cdot \mathbf{x}_B + w_C \cdot \mathbf{x}_C + w_D \cdot \mathbf{x}_D ... \tag{3}$$

  - $\mathbf{x}_B$, $\mathbf{x}_C$, $\mathbf{x}_D$ represent the control variable vectors for each of the control firms. This means that, given weights w's, $\mathbf{x}_{synth}$ can be calculated.
  - Given a $\mathbf{x}_{synth}$, we could then calculate its similarity to $\mathbf{x_A}$ (the treated firm). The goal was for $\mathbf{x}_{synth}$ to be as similar to $\mathbf{x_A}$ as possible.
- In order to find the *optimal* weights $\mathbf{w} = (w_B, w_C, w_D)$, we had to tackle the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \text{Distance}(\mathbf{x}_A, \mathbf{x}_{synth}) \\ \text{s.t.} \quad & \mathbf{w} \geq 0 \end{aligned} \tag{4}$$

  - $\mathbf{x}_A$: Data
  - $\mathbf{x}_{synth}$: A vector with same dimension of $\mathbf{x}_A$. From Equation 3. Given a vector $\mathbf{w}$, its value can be calculated.
  - Distance function: Euclidean distance.
- In order to solve this optimization problem, we completed the following steps:
  - We created a function called Distance which contained 3 inputs: (weights, treated.firm.vector, control.firms.vectors). Given a treated firm vector, control firms vectors, and a weight vector, the function would calculate the Euclidean distance between the treated firm and the synthetic control firm (weighted sum of the control firms).

The goal of this function was to (1) figure out the number of weight values that needed to be in the weight vector, which is the same as the number of the control firms, and (2) calculate the distance between the treated firm and the synthetic control firm, where the function was only given (1) a treated firm and (2) any size control firm set.

  - Next, we applied the optim() function in R to the Distance() function we wrote, which determined the optimal weights needed to generate the smallest possible distance between the treated firm and the synthetic control firm.

- A total of 16×2 observations were generated (16 industries each have 1 treated firm and one control firm).
- A paired t-test was conducted on the number of Q2 2020 job postings for the treated and control firms in each of the 16 different industries.
- This process was completed for both pSocial and pComputer.

The results of these tests were mixed. For pSocial, there was a near miss in statistical significance with a p-value at 0.0558. However, pComputer was not even close to statistically significant with a p-value at 0.3368.

5 Conclusion

When independently evaluating a firm's emphasis on Social Skills in their data scientists and a firm's analytical maturity, it is clear that there is a positive, statistically significant correlation between these terms and the number of data science job postings for Q2 2020. However, when evaluating these terms together, along with appropriate control variables, the waters become a bit more murky. That being said, the data does certainly suggest that data science practitioners who are very socially adept and who are working in an firm with a high degree of analytical maturity, will likely fare much better during the next recession than their counterparts, who lack social skills and who work in fledgling analytics departments.

References

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 132(4):1593–1640.

Forsythe, E., Kahn, L. B., Lange, F., and Wiczer, D. (2020). Labor demand in the time of covid-19: Evidence from vacancy postings and ui claims. *Journal of public economics*, 189:104238.

Ham, C. C., Hann, R. N., Rabier, M., and Wang, W. (2020). Auditor skill demands and audit quality: Evidence from job postings. *Available at SSRN 3727495*.