w241_final_report

Emily Fernandes, George Rodriguez, Giulia Olsson, Jason Yang

4/9/2022

Introduction

How marketable is a UC Berkeley Master's Degree in Data Science? Over the last five years, the job market for data scientists has grown notoriously consistently. There is now an overwhelming demand for data scientists and an overall lack of candidates to fill in every role. The job market, as of early 2022, has been called the "candidate's market." That being said, data Science as a university discipline is still fairly new. We wanted to understand if by adding a data science degree to a resume, a candidate had a better chance at being interviewed. In our experiment for this final project, we wanted to answer that question, therefore we tested how competitive a UC Berkeley Master's Degree in Data Science is in getting to the first round of interviews for early career professionals.

Those who have joined MIDS (UC Berkeley's Master's Degree in Information and Data Science program) believe that obtaining a Master's in Data Science from UC Berkeley will give them a leg up in obtaining data science jobs. There are economic opportunities that are associated with a Bachelor's Degree, which leads to the question of whether or not a master's degree also provides greater economic opportunities. We tested this theory by understanding the causal implications of getting a response from the employer (e.g. request for first-round interview or phone screen).

As a guiding point, there have been experiments in the past that investigated our concept. In 2017, LinkedIn named data science the fastest-growing job in the United States, and in 2018, Glassdoor ranked data science roles the best jobs in the country. By 2026, an estimated 11.5 million new data science jobs will be created in the US, according to the US Bureau of Labor Statistics. Given these impressive stats, research, although fairly new, has been done on whether or not having an advanced degree in the role improves one's chances of getting hired during an interview and job search process.

Also in 2017, two researchers at Drake University, Troy Strader and Andrew Bryant, identified the characteristics of universities offering data science programs, which at the time, was a fairly new concept (less than five years old). The researchers concluded that universities with higher access to businesses and governmental organizations were more likely to develop data analytics and data science programs. The researchers make the assumption that students in these programs would be able to later on contribute greatly to work at governmental organizations given their skill sets. The researchers also make an interesting point that growth in data analytics jobs has led to an increase in student interest for analytics degree programs. Universities have responded by implementing programs that teach students skill sets to succeed in these roles. The researchers answer the questions – what are the characteristics of the universities that have data analytics related programs, and what are some of the reasons that motivated colleges and universities to develop their undergraduate data analytics programs – but it does not answer the questions that this appear is looking to solve: Is an early career professional's time better spent on pursuing a MIDS degree to get a data science job? Or is the time better spent through gaining credibility at work?"

In order to inform ourselves before going into our research, we also looked at another study titled, "The effect of post-undergraduate training on passing first-screening for Data Science job applications," which could be seen as an advancement on the first study we looked at. The purpose of the experiment was to investigate the causal effect of a Berkeley MIDS degree on an applicant's ability to obtain a Data Science position by studying the effect of a MIDS degree on passing first-screening for Data Science job applications. The

study did not have enough responses to reach a conclusion. However, the researchers made some interesting points that could inform our study. Most data science roles require a Master's Degree, and the applicant in question was coming out of an academic background with no data science experience; this could have severely impacted results.

Research Question

Does having a UC Berkeley Master's in Data Science on a resume impact success in pursuing a job?

Hypothesis

Individuals with Masters of Information and Data Science will receive more responses from recruiters when applying to jobs.

Experimental Details

Employers look for validation from universities to show competency. Having a data science degree from Berkeley shows 1) dedication to education, 2) the candidate learned the MIDS degree course material 3) the candidate has proven this knowledge through applied projects. We assume that by including a Masters in Data Science, job recruiters will be more interested in Sam as a potential employee.

Our control resume states that Sam is a self motivated learner who has programming and statistics knowledge taught in the MIDS program. Sam even demonstrates that they can apply these skills through the Kaggle competition project. Our hypothesis is that employers will value these skills but are risk averse and would prefer the validation of a degree from Berkeley. We designed the resumes so that Sam has the same skills, knowledge, and applied data science experience. We vary only on the inclusion or exclusion of a MIDS degree.

Treatment

Our control resume will be designed to represent a young professional (age 25) who is interested in becoming a Data Scientist and is currently working as an Analyst. The details of this control resume were inspired by what our team would imagine a first semester MIDS student's resume would look like: above average school record, an interest in data through either work experience or personal projects, and some level of programming experience.

We chose to name this individual Sam Anderson. Sam is a gender neutral name to avoid any unintended bias and Anderson was chosen from the top 20 common American last names. We sent Sam to the University of Washington, a well performing university but without huge name recognition. We gave Sam an undergraduate major outside of STEM fields so that the recruiters would be forced to focus on Sam's work experience and graduate studies. Sam's contact information will reflect that they are applying to a job near them. In order to display Sam's interest in Data Science, we will include a personal project in this resume in which Sam will describe their personal project of working on a Kaggle competition.

To ensure that receiving a response is solely dependent on the inclusion or exclusion of the MIDS degree, all other fields in the resume were identical between the control and treatment groups. The treatment group in this experiment is the resume with the MIDS degree included, completing the program part time while still getting the same years of experience.

Inclusion Criteria

Job postings were collected through job boards LinkedIn and Indeed, with searching for roles for Data Analyst and Data Scientist. When selecting roles to apply for, we limited the roles based in the United States with experience ranging from entry to senior level. Through observations, we noticed that there were conflicting experience levels for these roles, based on the basic requirements of the job description.

Randomization

A Python script was created that ensured an even distribution of the control and treatment groups, based on whether the role was for a data analyst or scientist. Distribution of the random assignments are shown in Figures 1 and 2

Outcome Measure

The outcome measure is a binary response as the following:

- 1 - Contact with an organization regarding an invitation to either a phone screening or an interview, prior to the close of the study
- 0 - No contact with an organization prior to close of the study or contact declining further consideration

Power Calculation

```
# power calculation
# https://med.und.edu/daccota/_files/pdfs/berdc_resource_pdfs/sample_size_r_module.pdf
pwr.t.test(d=0.25, sig.level = 0.05, power =0.8, type = "one.sample", alternative="greater")
```

```
##
##      One-sample t test power calculation
##
##               n = 100.2877
##               d = 0.25
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
```

To calculate how many samples we need to collect, we calculated the power of our experiment. Assuming a medium effect size (d=0.25), significance of 0.05, power of 0.8, and a one-tailed t-test, our sample size was approximately 100.3. We assumed a one-tailed test, because the addition of the MIDS degree in the resume should make applicants more marketable and help them receive responses from the recruiter. We assumed a medium effect size because the control resume had a bachelor's degree that is irrelevant to data science. With a relevant graduate degree in data science, the treatment should have at least a medium effect size in the experiment.

Data

In the experiment, we acquired 84 samples in both the treatment and control groups. We measured the outcome variable as requests for interviews from the job application. When observing the response rates, we see that the Treatment condition with MIDS degree had a higher response rate at 8.33%, though not with statistical significance as we detailed in the modeling section.

Resume_Type	Data_Analyst	Data_Scientist	Total	Response_Rate
Control	39	45	84	4.76%
Treatment	39	45	84	8.33%

Table 1: Summary of experiment data

The data was gathered in the span of 6 weeks from late February 2022 to mid-April 2022. As noted in the power calculation, we needed to acquire at least 100 samples in each treatment and covariate groups to reach statistical significance with medium effect size. As seen in the table above, we did not reach the necessary sample size to reach statistical significance. This was due to the time limitation of the experiment period and the amount of time required to generate each sample. We underestimated the amount of time required to apply for each sample.

We included data analyst and data scientist role covariates. The response rate of data analyst and data scientist roles are likely different, so we decided to block on this variable with proper randomization.

Models

To evaluate our experiment, we built three linear regression models for comparison: 1) Basic Model: to evaluate response by resume type 2) Job Type as Variable: basic model with job type as a covariate 3) Heterogeneous: model to evaluate the heterogeneous effects of resume type and job type

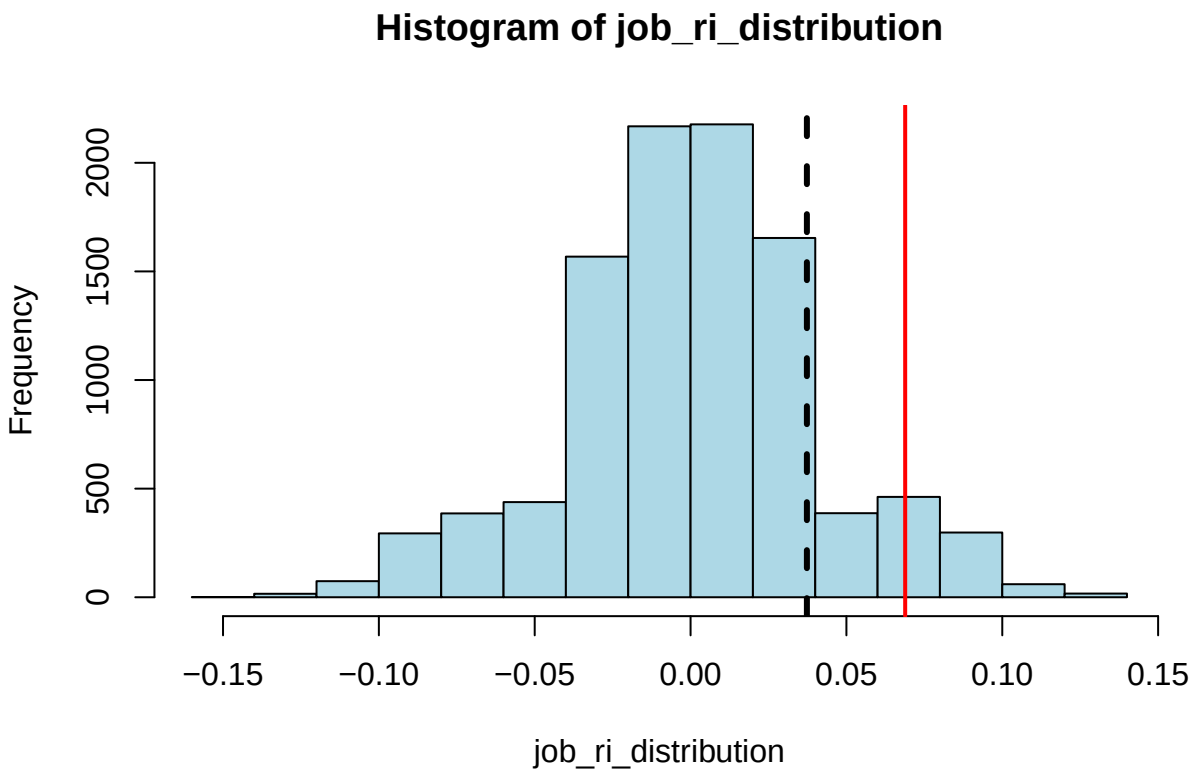
```
d$data_analyst <- ifelse(d$job_type=="A",TRUE,FALSE)
d$data_scientist <- ifelse(d$job_type=="S",TRUE,FALSE)
d$control <- ifelse(d$Resume_type==1,TRUE,FALSE)
d$treatment <- ifelse(d$Resume_type==2,TRUE,FALSE)

d$data_analyst_contol = d$data_analyst & d$control
d$data_analyst_treatment = d$data_analyst & d$treatment
d$data_scientist_contol = d$data_scientist & d$control
d$data_scientist_treatment = d$data_scientist & d$treatment

mod_basic <- lm(recruiter_responded ~ Resume_type,data=d)
mod_job <- lm(recruiter_responded ~ Resume_type + data_analyst,data=d)
mod_sat <- lm(recruiter_responded ~  data_scientist_treatment +
                data_analyst_contol + data_analyst_treatment ,data=d)
mod_sat <- lm(recruiter_responded ~  data_scientist_treatment +
                data_analyst_contol + data_analyst_treatment ,data=d)

robust_se_b <- sqrt(diag(vcovHC(mod_basic, type = "HC1")))
robust_se_j <- sqrt(diag(vcovHC(mod_job, type = "HC1")))
robust_se_s <- sqrt(diag(vcovHC(mod_sat, type = "HC1")))

#stargazer(mod_basic,mod_job, type = "text",
#          se = list(robust_se_b,robust_se_j),
```

```
#          column.labels = c('Basic','Job Type as Variable'))
stargazer(mod_basic,mod_job,mod_sat, type = "text",
          se = list(robust_se_b,robust_se_j,robust_se_s),
          column.labels = c('Basic','Job Type as Variable','Heterogeneous'))
```

```
##
## ================================================================================================
##                                              Dependent variable:
##                          ------------------------------------------------------------------
##                                              recruiter_responded
##                                Basic         Job Type as Variable      Heterogeneous
##                                 (1)                  (2)                    (3)
## ------------------------------------------------------------------------------------------------
## Resume_type                   0.037               0.036
##                              (0.042)             (0.040)
##
## data_analyst                                     0.171***
##                                                  (0.043)
##
## data_scientist_treatment                                                 -0.000
##                                                                         (0.000)
##
## data_analyst_contol                                                      0.132**
##                                                                         (0.056)
##
## data_analyst_treatment                                                   0.211***
##                                                                         (0.067)
##
## Constant                      0.023               -0.054                 0.000
##                              (0.062)             (0.060)                (0.000)
##
## ------------------------------------------------------------------------------------------------
## Observations                    165                 165                    165
## R2                             0.005               0.105                  0.110
## Adjusted R2                   -0.001               0.094                  0.093
## Residual Std. Error      0.270 (df = 163)    0.257 (df = 162)       0.257 (df = 161)
## F Statistic             0.786 (df = 1; 163) 9.474*** (df = 2; 162) 6.636*** (df = 3; 161)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

```
job_ate = mod_basic$coefficients[2]
N=length(d$recruiter_responded)


#Defines Function to calculate ATE from dataframe
calc_ATE <- function(d) {
  #m <- d[ ,list(mean=mean(recruiter_responded)), by=Resume_type]
  mean_t = mean(d$recruiter_responded[d$Resume_type==2])
  mean_c = mean(d$recruiter_responded[d$Resume_type==1])

  ate <- mean_t - mean_c
  return(ate)
}
```

```r
#Defines Function to that randomizes the data n times
#and calculates all those ATEs
run_RI <- function(n, d) {
  d_i = d
  distribution_n <- vector( "numeric" , n )

  steps = seq(1, n, by=1)
  for (i in steps )
  {

  Resume_type_i<- rep(c(1, 2), each = ceiling(N/2))
  Resume_type_i<- sample(Resume_type_i,N)



  d_i$Resume_type = Resume_type_i
  ATE_i = calc_ATE(d_i)
  distribution_n[i]=ATE_i
  }
  return(distribution_n)
}

#Run Random Inference
n = 10000
job_ri_distribution <- run_RI(n, d)
stat_sig <- qnorm(0.95)*sd(job_ri_distribution)
#Plot Histogram of RI(n=10,000) with our actual experiment's ATE
hist(job_ri_distribution, col="lightblue")
abline(v = job_ate, col="black", lwd=3, lty=2)
abline(v = stat_sig, col="red", lwd=2)
```



Conclusion

Future Works