B-PALs

Day one we just looked at the data and tried to filter out unnecessary columns and rows. We found that we only got a subset of the data there are companies that jump from having 5 job postings at one time event to having 32 job postings in the next event associated with that job. We also found that for each job hash there were times that the job title was not consistent between job events even with the same hash. There were six hundred duplicated jobhashes out of 236635 job postings, meaning there were two job titles under the same job hash.

On this slide, the font of the text indicates the number of jobs and the color indicate the number of companies in that field. So the larger the text is the more jobs there are. And the darker the text is the more companies there are too. Thus, Indeed should reach out to more companies that has lighter color or smaller font as their category.

The ratio between number of job applies to number of job postings within a state. The darker the color, the more applicants per job.
For example: Number of companies and number of job postings in ID and NV are fairly close. However, the color indicates that there were more job applicants per job posting in NV compare to ID. This tells us that there are more job demand in NV and Indeed.com should consider recruiting more companies in the darker region like NV.

We found that on average over a one week period in 2017 the most applies were on Tuesday and Monday. And the least application were sent out on the weekends. From this information we know that it will be best for companies to make postings late sunday night or very early sunday morning so the company will get the most potential employees to apply. We also looked at data on job postings and applications over all the months of 2017 and found that most job postings are posted in july but most applications are submitted in January and March. Similar to our first conclusions we can tell companies to try to make postings closer to when people are looking for jobs if they want to get more applicants.

On the fourth slide, we use a scatter plot to represent the relation between the number of applies and the number of applications which are reviewed. Each data point represents one job post, and the size of it represents the length of its description, the bigger size indicates the longer length. The darkness of each data point represents the age of the job post, the darker it is, the older the job post is. The shape of the data point indicates whether the job requires any education or not. According to this plot, the relation between the number of applications and the number of applications is a linear relation, and there are no relation between the length of description and the age to the number of applications and reviews.

Insight on Job Openings @indeed.com

By The Stat Wars:
Ruby Ru, Carol Liu, Rebecca Wang, Elaona Lemoto, and Starry Zhou

We all had very different ideas coming into this project, but we collectively thought of using our economic interests to make some insight on this project.

As economists, after looking through the data on indeed.com, we decided to look at job postings over time since an increase in job postings could be an indicator of economic growth. When we pulled up the model for job postings over time, while it did produce an interesting polynomial regression, we just thought that could be explained by outside factors. We then thought we could look at economic indicators like unemployment rate or Producer Price Index to explain some of the variability in the data. We found data on FRED regarding PPI and the trend looked very similar to our model on jobs postings over time. While we did create a model using PPI to predict the number of job postings, we could not use the data because the units for our external data was off and looking back at the data for job postings over time, we saw that there was a jump in the data given. The line is not polynomial at all. There are actually two different lines. Because of this, we had to throw out our model and start again.

After this realization, we tried researching why there was this jump in the data. We searched on the events happening and the unemployment rate during the gap time. As there are no significant events taking place and the unemployment rate remains stable at 4.3%, we did not find any appropriate explanation for the gap. Thus, we chalked this up to a model that we could not use.

Instead, we started looking at the data again and focusing on other variables. An interesting find was made in education. Approximately 0.6% of the jobs on indeed.com, according to this data, require higher education. However, this is counter-intuitive because many jobs (e.g. engineer, computer scientist) that would seem to need at least a higher education level did not have that requirement. Research shows that 35% of job openings in the US require a bachelor's degree. Therefore, we can conclude that the educational requirement on indeed is not a reliable enough resource for data analysis on studies such as "what kinds of jobs require a higher education" because of the lack of the requirements of a more comprehensive data entry.

DataFest 2018 Write Up

Scott Cohn, Chester Moses, Dan Brickell, Dennis Hofmann, Gurudeep Machupalli

3/25/2018

Introduction

In macroeconomic theory, the Phillips Curve represents an inverse relationship between rising wage rates and corresponding rates of inflation. This tradeoff is often visible in the short-run. In the medium to long-run, this relationship is not as evident. In the long-run of the economy, the inflation rate is represented graphically as a vertical line over the natural rate of inflation indicating no effect. Typically, this relationship is used by central banking agencies to dictate how to adjust inflation rates given expectations and unemployment levels. We tested how the variables, including several other economic indicators, affected the job market.

Data

The data was given by career website Indeed.com. Indeed is the biggest job search engine. The provided set had 4.5 million rows and 27 variables. Some of these jobs are aggregated from third-aprties, while others are input directly. Each row represents the information about a job on a particular day. Each column was a piece of information relevant to each job. Some jobs were posted multiple times under the same job identifier value.

Methods

The data was imported and a random sample of 75000 rows was selected. From this sample, it was filtered by industry and saved into a smaller sample. The industry specific dataframe trimmed to contain only dates and number of job applicants. This was binded with log-based monthly economic indicator data from the Federal Reserve Economic Data (FRED). Outliers were pulled in through our log-log model. However, we did test for outliers using the minimum value ellipsoid and robust residiaul tests. Given our large sample size, the few outliers had no significant affect on our analysis. We then regressed these variables on lnApplies.

```
# lm(lnapply ~ UNRATE + INFLATION + PAYEMS + PCEPILFE + IceCreamSales, data = x)
```

We ran multiple regressions using different sample sizes across different industries. We then utilized ANOVA tests to remove insignificant variables. We later got excited by the prospect of novel correlations and tested the change in applies against Industry Sales — specifically against the change of sales of hard ice cream nationally.

Conclusion

Our explained variance, measured by R^2 was less than one percent. The unemployment rate was seen as a significant variable with a p-value less and 0.05. This was likely due to overfitting our model given the degrees of freedom we had. These economic indicators that suggest movement in the labor market did not have any measurable affect on the number of job applications in our dataset. There is no effect in the short run on Indeed job application numbers given the indicators we tested against.

- Who benefits from categorizing job applicant fitness?
    - Job Seeker
        - Identify correlation between applicant features and fitness
        - Identify new jobs one is fit for
    - Hiring Managers
        - Increase efficiency of finding fit applicants
        - Guide expectations for hiring process
    - Indeed
        - Identify trends in user's fitness
- Who benefits from categorizing job applicant deadline?
    - Job Seeker
        - When likely or not likely to be reviewed for a job?
    - Hiring Managers
        - How long to expect before finding right candidate?
    - Indeed
        - Forecast how long job will remain on portal
- What is our method?
    - Assume hiring managers hire on rolling basis
    - Aggregate indeed data by specified parameters to categorize job
    - Find relationship between applicant fitness and job application deadline using:
        - Expected # Reviews by Employer
        - Job Age
        - Cumulative Job Applications

Slide 3

Blue line in graph is total number of expected reviews by employer. Once the cumulative reviewed applications rises above this line, a jobseeker should expect on average not to be considered even if they are fit for the job. The red line represents the cumulative number reviewed assuming 100% review rate. The green line represents the cumulative number reviewed assuming 50% review rate. The places where these lines intersect the blue line are the expected dates by which the job will have been filled. The percent reviewed is identical to the minimum percentile of an applicant in terms of fitness for being reviewed. This model gives us a map between percentile and deadline.

Slide 4

We use our simple model to identify the first relation below and rearrange for the second relation, which gives the minimum applicant percentile as a function of data in indeed's DB. Our third relation comes from plotting our data using calculated percentiles and finding that the relationship is close to linear when one takes exp(1/%). Thus we can find the day a job is filled.

TR = % * CA(DF) | % = TR/CA(DF) | DF = K*e^(1/%) +C

R = Expected total reviews

% = Minimum percentile of applicant

CA = Cumulative applications received

DF = Expected age of job when filled

Team R-Some | DataFest 2018

The purpose of our set of graphics is to help people who are living in the U.S. and looking for a job. Our goal is to help them decide which industry to pursue based on their educational level, where in the country has the best opportunities for this industry, and how to best use Indeed.com in their job-search process.

Our goal is to help the **INDIVIDUAL** use Indeed.com in the most customized, efficient way!

Our test case is a college student named Alex who is looking for a job. Alex, who is on track to complete an undergraduate degree in Mathematics and Psychology, lives in Tallahassee and attends Florida State University. They would like to get a job in an industry with a plenty of opportunities and most closely complements their skill set.

To begin their job-searching process, Alex will consult our first graph. Because they are working towards attaining their college diploma, they will look for the green bars. The largest green bars are their most optimal choices for industries to pursue, as these are the industries with the highest proportion of postings for that given education level. However, the largest green bar represents the industry for Non-Profit/Association, but being a Mathematics and Psychology major, Alex decides that this is not a good industry for them and looks for the next largest green bar. This brings them to the green bar representing the Education industry. Thus, based on this graphic, Alex will search for a job in Education.

Alex now needs to evaluate the best regional options for their chosen industry, Education and thus will consult our second graph. This graph shows the six states with the overall greatest number of jobs available, and for each state their top ten industries with the greatest number of jobs available in that state. Living in Florida, Alex looks to the graph for Florida. However, when Alex consults the graph for Florida, they see that Education is not one

of the top ten industries for jobs in the state. They then look to the other states to see where jobs for Education are most plentiful, which directs them to the graph for Ohio. Compared to the graphs for the other five states, Ohio's bar for Education is the largest. Thus, Alex knows that Florida is not a good spot for their career prospects in education, and they will optimally move from Florida to Ohio, which has good prospects for educational careers.

Finally, Alex needs to decide how to best utilize Indeed.com to initiate their application process. Our third graph shows the job and posting application activity on Indeed.com over the time frame of about a year. From this graph, Alex sees that the number of postings and applications in Education is not volatile and that there is not an optimal time of year to apply for this industry. Therefore, Alex can determine that Indeed.com's steady postings for this industry will allow them to have flexibility in their availability and time for their job search process.

In our analysis process, we aimed to dig for the reasons behind the appearance of unusual points on our scatterplot matrices of five variables which are Clicks, jobagedays, descriptionLengthChars and EmployerJobCount.

At our first trial, we focused on small companies, and hence filtered the full dataset for companies that had employee count from 1-49. We believed that small companies would want to recruit more competent employees. We executed the backward elimination and forward selection to choose the best variables. The result was not satisfying because we didn't extract conclusion. Then we tried the interaction analysis of the variables. Though we did get the interaction coefficient, we were stuck on the calculation and interpretation of the interaction terms. We ran the model again and found out that the adjusted r-square was 0.0014, which indicates that the model is not useful in predicting the clicks.

At our second trial, we randomly extracted the full dataset into a dataset that contains only 1000 entries of data. Within the extracted dataset, we tried finding the relationship between clicks, job age days, description length characters and employer job count through plotting them using R. We found that there were two entries that had extremely high employer job counts. We wanted to know the reason behind this phenomenon. After considering factors such as "employerAccountDateCreated", "admin 1" and "city",we found out that the location for these two entries, Seaford, DE and Quinlan, TX, are relatively small cities with small population. Job seekers are less likely to find a job in a city that is so remote and less populated. This is why we concluded that the reason for a high employer job count for those two entries was most likely because they were re-posted regularly over a short amount of time.

From our trials, we did expect significant relationship between clicks and the length of job description. From a common sense point of view, the more detailed the description is, the more attractive the job is. However, there is no evidence indicating a linear relationship between those two variables. Statistics also indicate that whether a job requires higher education or not is also an important factor that applicants would consider before they click on the job. The linear model we ran might not explain the model efficiently. There are also other variables that need to be taken into account. Hence new models would be constructed and analyzed with other variables and statistical approaches.

Aidan Price, Diana Hansen, Henry Noke, Samuel Baltrus, Kristen Bonsall
Statistically Significant

Improving Job Posting Interaction Rates

This analysis is designed to help employers who use Indeed maximize traffic towards their job postings. To make our analysis relevant to college students, we narrowed the dataset to only intern and entry level positions.

Exploratory Analysis

We first looked at the relationship between clicks and applies, finding that applies is a perfect linear function of clicks. This seems very unlikely, so we recognize the applies data might not be accurate. However, it would make sense if applies and clicks are proportional, so we decided to continue with this relationship. We also only had one month in 2016 and 2018, so we decided to only look at the 2017 data. However, having more full years would make our analysis stronger.

We looked at a correlation graph of the continuous variables, and saw that few variables had a strong relationship. Clicks, applies, and candidate reviews had a positive correlation with each other, and job age days had a negative correlation with most variables. This implies that the longer a job post has been up, the less activities and clicks it will have. We wanted to further analyze this, so we split the data into four quarters and ran a F-test of clicks verse quarters, with the null hypothesis of the Betas of each quarter equal to zero. The F-statistic is 19.92 and the p-value is <2e-16. With an alpha of 0.05, there is at least one Beta that is not equal to zero and there is significant difference in the number of clicks for each quarter. This means that we have shown seasonality in our full dataset.

Statistical Testing

Understanding that the quarters are significantly different, we split the data into different industries and performed the same F-test on five industries. The first industry we looked at was Accounting/Consulting, and we found the p-value was 6.692e-13. We ran a general linear hypotheses test where we adjusted for multiple hypotheses by using the Bonferroni correction, multiplying our obtained p-values by a factor of six. Testing the Accounting/Consulting data, the fourth quarter was significantly different from the rest. This result implies that it is optimal for Accounting/Consulting companies to post jobs in the fourth quarter to get the most activity on Indeed.

Testing a total of five industries, we found that, based on the industry, certain quarters had significant difference in their mean clicks. This supports our argument of seasonality because the different mean number of clicks in certain quarters show a non-standard trend of clicks, highlighting certain industries experience highs and lows in terms of students looking for employment.

Maximize Clicks

After finding that seasonality impacts click numbers, we looked at other aspects to see if there is a way to boost those numbers more. A job will get its most clicks on the first few weeks of posting, and will then soon level out afterwards, remaining constant. A company posting a job in the season of their industry will help them capture those maximum number of clicks in the first few weeks, as well as avoid being buried by low traffic.

Trends in New York City Health Industry

Team name: MoData MoMoney

From Mount Holyoke College

For this project, we looked at three data sets: Indeed, Bureau of Labor Statistics, and Google Trends. Some of the variables that we used were name of the city, education requirement, date that a job was posted, duration of a job posting and job title.

From the Indeed dataset, we noticed that most of the jobs posted on Indeed came from the healthcare industry. The majority of the jobs were posted from New York City. We decided to take a deep dive into job postings related to the healthcare industry in New York City. Second, we looked the external dataset from Bureau of Labor Statistics for job listings in New York City.. We noticed that the job posting pertaining to the healthcare industry were not limited to clinicians such as doctors, nurses or pharmacists. There is a high demand for practical jobs as well. For example, drivers are needed to transport medical equipment to hospitals. Here, we also explore the trend of posted job titles in healthcare industry and the trend is related to education requirement and salary trend.

Therefore, we connected the Indeed dataset with the second dataset. We found that most of the people applied to these jobs directly rather than through a staffing company. We also found that for most of the jobs, some level of education was required since majority of the jobs were for technical or managerial positions.

Additionally, we looked at “jobs nyc” search term in Google Trends. We found out the search trend of healthcare job on Google in 2017 and we related the tend to the supply side of the jobs on Indeed. In this way, we found how the trend of application is influenced by supply and demand of the job in healthcare industry.

We Don't Work in a Vacuum

datacHAMPs1 - Kirsten Lydic, Brooke Fitzgerald, Cindy Fang, Kyoko Sano, Hunter Johns

Datafest 2018

Employment is a very personal process. In fact, there are many external factors that influence whether or not an employer gets the applicants they are looking for.

Figure 1: Lifetime of a job

Firstly, we present generalized additive model for both the number of applications and the number of candidates reviewed for each individual job, across days within the listing's age.

There is a drop in both numbers of applicants and candidates reviewed after approximately two weeks, which we consider a reflection of the "effective lifetime" of a job on the website. In short, if a job listing doesn't get the number of applications that the employer is looking for, they might be out of luck.

This also indicates that competitiveness to apply and review may be highest within the first few weeks of a job's listing on Indeed. In our exploration of this idea, we considered other factors relevant to competitiveness.

Figure 2: Competitiveness within Industries

This bar graph represents competitiveness as a measure of applications per individual job for each of the industries categorized in the data. For simplicity, we are only presenting our sample of the fuller plot; we show here the five industries of highest competitiveness, and the five industries of lowest competitiveness.

Figure 3: Job Availability per State

Here we show a map of job availability by state in the US as represented by the job listings per capita, using state populations. The northeast and upper midwest have higher job availability; this was also reflected in the rightside map of applications per state.

The applications follow job availability; from this we can derive that location is another factor of influence on whether jobs listed will receive substantial numbers of applicants.

Figure 4: Postings and Unemployment

Finally, we show a three-part graph of unemployment, job availability, and number of applications over time and see that they are very related.

Preventative Medicine: Valid or Not?

Three Musketeers: Audrey Cheng, Margaret Chien, Fengling Hu, Bonnie Lin, Lucy Yuan

A large population of U.S. citizens do not have access to affordable health insurance and care. The Affordable Care Act (ACA), one of the most controversial acts of recent times, was signed into law in 2010 to address this public health crisis. A major aspect of Obamacare is that it rations federal funding for states to expand Medicaid and provide health insurance coverage to more of the low-income population. As of today, 32 states have chosen to adopt the expansion and 18 have not.

The pros and cons of ACA have been the subject of heated discourse. Specifically, CEOs of health insurance companies assert that it is not economically feasible; they argue more coverage will lead to more individuals seeking care, which will create too much demand for healthcare that companies and hospitals cannot supply. However, ACA architects argue it will decrease healthcare spending in the long run. They argue for a preventative model of care - by allowing people to get affordable regular checkups, emergency room visits will decrease and chronic diseases will be detected earlier, when lifestyle changes have larger positive effects on prognosis.

We sought to study this discourse by investigating demand for employees in the healthcare sector across all 50 states during 2017, seven years after ACA was enacted. We hypothesized that, depending on whether ACA increased/decreased demand for care, healthcare sectors in the 32 states which expanded Medicaid would post more/less job offerings compared to those in the other 18 states.

To approach this question, we collapsed the dataset by unique job posting; each row corresponded to one such posting. Then, we joined the dataset with Medicaid expansion data (whether or not the state expanded Medicaid) and total population for each state. This allowed us to find density of job postings per capita per state.

Analyzing this data for the healthcare sector, we found a significant difference in job posting densities between states that did and did not expand Medicaid ($p < 0.05$). Importantly, expanded states seemed to have higher job posting densities. That is to say, healthcare employers in expanded states were looking to hire more new employees than healthcare employers in non-expanded states. Additionally, investigating this relationship in a sector less related to ACA expansion, the technology sector, we found no significant relationship. This more strongly suggests the correlation we see in the healthcare sector is actually present.

Next, we examined this issue from a supply perspective to see if 1) there is enough supply to keep up with heightened demand and 2) there were any significant limiters of adequate supply. As a proxy for supply, we used total clicks on each posting per capita per state. Looking at the ratio of job posting density to click density, we saw no significant difference between expanded and non-expanded states. This suggests supply keeps up with the heightened demand resulting from Medicaid expansion.

Even so, some difficulties may arise in filling these job postings. Looking at education requirements for healthcare jobs, we see that, when compared to three of the next largest industries in the Indeed dataset (technology, retail, food), the proportion of jobs requiring supplemental degrees and licenses was highest in the healthcare industry. However, we note that this proportion is still relatively low. Thus, the healthcare industry could, but is unlikely to, have difficulty reaching demand necessitated by Medicaid expansion. Costs of hiring these new employees may still be prohibitive.

Our analysis provides important insight about the implementation of ACA. Healthcare is a universal human right and inarguably furthers the human condition. However, discussions of economic costs and supply are unavoidable. We suggest that, though supply of employees seems to be adequate, ACA architects may be incorrect, with respect to costs, in their hypothesis of cost reduction through preventative medicine; expanded states seemed to demand more, not less, healthcare employees. But this conclusion must be taken with a grain of salt. Change does not happen overnight, and in the context of long-term, nationwide healthcare reform, the seven years between ACA implementation and this dataset is certainly a very short time.

Alicia Bochnak
Michaela Digan
Sarah Manlove
Natalie Slabczynski
Chinh Do

Relationship Between Industries, Traffic on Indeed, and Layoffs

We started by looking through the data using summary statistics in R. After we saw that there was not much need to clean up the data, we moved on to studying the data with Tableau. We explored the data in Tableau by using statistical models such as line graphs, bar plots, and geographical maps to try to find some abnormalities in the data. We originally wanted to compare the relationship between clicks and job applications. After conferring with the Google Document, we realized that this was not a viable option and continued to explore the data. We were looking at the average number of job applications for each respective industry per month. During this, we observed an anomaly within the automotive industry. The average number of applications in July skyrocketed compared to other months and we did not see this anomaly within other industries such as education, chemical engineering, or civil engineering. Overall, automotive applications averaged to 6.317 applications, but the average of automotive applications in July was 19.50. We began to try to pinpoint what happened in July to cause this spike in job applications. After some research, we came across a *New York Times* article which was published on July 4, 2017 and Ford June 2017 Sales Report about the plummet in sales in the automotive industry. To see if there was data to back up these claims, we began to look through government agencies to find some data about the job market in the automotive industry. After some probing, we found the Bureau of Labor Statistics website which displayed a line graph regarding the number of employees in the automotive industry (https://data.bls.gov/timeseries/CES3133600101?data_tool=XGtable). Interestingly enough, there was a dramatic decrease in employment in July of 2017. The line graph figure on the Bureau of Labor Statistics was an inverse of the graph that we produced in Tableau. This was fascinating to our group, so we decided to pursue this on a deeper level and think about the implications of discovering this phenomenon.

By comparing these graphs, we were able to notice inverse trends and correlations between layoffs and increased applications to careers where the layoffs occurred. For Indeed, this is an opportunity to not only increase revenue thanks the the pay-per-click business model, the company can also do a public service. Predictive technologies and machine learning can determine future layoffs, and the company can plan accordingly for the incoming web traffic. A push for sponsored posts for jobs in the ailing industries can help users who fall victim to the layoffs. Users can get more economic opportunities, and Indeed's business clients can find more

potential employees. Hopefully with this data, Indeed can adjust its business practices to be more efficient for users and also be more profitable as a whole.