Devil in the Data: Write-Up

When we were first exploring the data from Indeed, we noticed that the most common employer industry was the Healthcare/Medical industry. Doing research on the different industries in the US, we found that, overall, healthcare is one of the largest industry in the US. On top of that, healthcare is also one of the fastest growing, with eleven out of the twenty fastest growing occupations in 2016 belonging to the healthcare industry, according to the Bureau of Labor Statistics. While its growth is significant, there is also a large potential for change. It is an uncertain job market and potential for changes in the near future due to federal policy shifts. Because of these factors, we believed a closer look at the healthcare industry was necessary. While the common face of the healthcare industry is primarily made up of doctors and nurses, the data from Indeed showed that the jobs that are the most commonly posted about are those that are somewhat behind the scenes. For instance, management and administration are the two most commonly posted categories.

This led us to deeper questions of which categories, and which jobs within those categories in the healthcare industry are most prevalent on Indeed. We explored the demand for each of the top five categories in the healthcare field in each state. Population data from the U.S. census was then used to figure out the demand for each category in terms of the number of jobs posted in that category on Indeed per capita for each state. Of the many ways to calculate demand, we calculated it as the number of postings for each job over the population of each state.

We found that the five most common categories within the Healthcare/Medical field posted on Indeed were (in order) management, administration, medical nurse, sales, and service. For these five categories, we created bar graphs that show the top 5 states with the highest demand in each category, as well as a map of the demand for healthcare regardless of category for each state to look at similarities and differences between these five categories.

Vermont shows up as a state with high demand in every category, with Maine and New Hampshire showing up in most categories. We noticed an association between these three states and the general states that are most popular on Indeed when adjusted for population. Therefore, it may be more useful to look at the states that are not in the top five most popular on Indeed overall, or when filtering for the healthcare field, but do show up in the top five of the subsets: Nebraska, Kansas, and West Virginia.

Additional analysis could be done with a different metric for demand, possibly with looking at the number of healthcare jobs divided by the number of jobs for each state. This would do a better job of controlling for Indeed's popularity in each state, but a worse measurement of need relative to total population, and therefore total jobs in a state. The overlap between these metrics would indicate states that have an unusual demand for healthcare jobs specifically in the categories management, administration, medical nurse, sales, and service. Indeed could potentially use this for marketing purposes and for an understanding of who and where a large amount of the employers using Indeed are.

Competitiveness: A New Job Search Filter

Team A-super-NOVA, Smith College
Audrey Bertin, Riley Boeth, Emma Livingston, Clara Rosenberg, Kara VanAllen

As a job seeker or an employer looking to hire, it is of vital importance to understand the interaction between supply and demand in the job market. We set out to better understand how competitiveness depends on industry or region. We sought to answer the following question: in which regions or industries is the demand for jobs unusually high or low?

This is a primary issue for both job seekers and employers. Job seekers are concerned with where the job market is less competitive due to a lower demand for employment. The answers can allow one to best determine where to go or what industry-specific skills to develop in order to have the best chance at getting a job. Employers are more interested in areas in which the job market is more competitive, particularly when a company is starting out, opening a new branch, or moving to a new location. These areas are those in which there are more applicants in an industry, and thus more options for new employees.

In order to quantify the competitiveness of an industry or region, we developed a metric that proxied competitiveness. This metric was calculated by dividing average applications per job by the average length that jobs from that industry or region remain posted on Indeed. Higher values in the index correspond to increased competition for jobs. Since these factors might not equally contribute to an industry's true competitiveness, this index can be weighted as needed by industry professionals to better reflect the needs of the job market.

To explore our data and solve for the competitiveness index, we wrote custom functions that allow prospective employees and employers to find the least (or most) competitive region in a specific industry, and the least (or most) competitive industry in a specific region. The functions also allow individuals to specify based on their educational attainment, as companies often have requirements listed. For these functions, we used R packages such as ggplot2 and dplyr to visualize and wrangle the data. In order to compare the competitiveness measure among states, we created an interactive map of the United States using the Shiny package, in which the color scale corresponds to the competitiveness level in the state. Thus, a darker colored state for the technology industry indicates the state is one job seekers should avoid due to smaller chances of employment, and one that employers should consider moving to for more abundant employment opportunities.

We believe that it would be beneficial for Indeed to allow users to sort by the competitiveness of the job market, as this would allow job seekers to put more effort towards jobs that they are more likely to get. For example, a recent high school graduate in New England might be interested to find that the region's least competitive industries are Veterinary, Agriculture, and Analyst. If the grad is curious about the best place to go in, for example, Information Technology, they would be interested in finding that the least competitive region for that industry is the East South Central US (Alabama, Kentucky, Mississippi, and Tennessee).

These insights are extremely beneficial for a job seeker or a company deciding where to go next. Additionally, the inclusion of a competitiveness index would likely bring traffic to Indeed's website, due to the fact that no other major job search sites include one.

Last Stand

- *Internal Usage of Data - What Could We do With the Data Provided*

This is the first way to use the data. Consider it Level 1.

We are trying to make sense of the data internally. We are concentrating on the variables that have been provided to us and trying to extract information independent to the outside world. This is the **interpretation** of our database.

This is a primitive use of data.

- *External Usage of Data - What is the Potential of the Dataset for the Company*

This is the second way to use the data. Consider it Level 2.

We are trying to put our data in context. We try to make sense of the data geographically. The data set is very limited and due to constraints we restricted our analysis. However, we would be interested in looking at where our job-seekers are. We already have where the job postings are in the United States, now we just have to connect it to the people that use the website to look for jobs.

This information would be useful because we would be directing our efforts towards making the website more attractive for the people that are looking for jobs.

This is a better use of data.

- *MetaData - Turning our Attention to the Information of the Information*

This is the third way to use data.

We are concentrated on the aggregate information we could get from our user base.

Indeed.com has to make money. By accumulating information from their users in exchange to the service that they are providing (connecting employers and job-seekers), they can sell this information and become a more profitable company.

Team Tie Dad: Anya Conti, Brittany Pine, Corwin Burdick, Edward McCormick, Jeffrey Spahl

What, Where, and When: An Examination of Opportunities Posted on Indeed

Three major questions on the minds of job-seekers are the following: one, what types of jobs are out there; two, where are they located; and three, when do those positions have openings? We sought to answer these questions through our analysis of the Indeed data set. We opted to restrict our data set to 2017 only to have an equal number of samples for each month, so we eliminated the months of December 2016 and January 2018.

One component of our research consisted of finding the distinct job postings by month for each of the 29 industries in the data set. The total number of jobs posted peaked in the months of March, October, and August, but patterns varied among the individual industries. Postings slumped universally in November and December, so opportunities are most limited around the new year. Furthermore, the two industries with the most job openings were healthcare/medicine and staffing firms. People with a focus in healthcare fields appear to have the most opportunities on Indeed.com.

To express similar data on the state level, we geospatially plotted top industry per US state per month. Similar to the above findings, the top two industries across the states for each month were health care/medicine and staffing firms. The proportion of states with one of these as the top two industries deceased in November and December, which is consistent with our findings in the previous paragraph. No state maintains the same most popular industry from month to month for the entire course of the year.

We then constructed a regression model to better explain job posting trends. We found the number of postings in each industry per month and per state. Then we modeled the natural log of job postings (our dependent variable) as a function of the natural log of state population[1], along with quarter of the year (to account for seasonal changes), each of the eight regions[1] of the United States, and each job industry category. For large populations, we expect a correspondingly high number of jobs, since people tend to aggregate where they can find work. We expect the Northeast to have the highest number of job postings given their population. For the model, we had an adjusted R^2 of .955. Then we did a residual plot, and tested for homoskedasticity using a Breusch-Pagan test, and as a result, we weighted the standard errors to correct the model. Then we performed F-tests on each group of binary variables (quarter, region, and industry). Each test resulted in a statistically significant output at the 5% level, and so we left each of those variables in the model.

All variables discussed below are statistically significantly different from 0 at the 5% level of significance. The model showed that a 1% increase in population corresponds to a 0.869% increase in job postings, holding all other variables constant. The following binary variable coefficients shift the intercept holding other variables constant. The Staffing Industry and Health Care/Medical industry binary variables had the highest coefficient estimates compared to other industries. Each quarter was similar, except for the fourth quarter having a far lower coefficient. The Mid-Atlantic and New England had the highest coefficient compared to other regions. Our conclusions matched our expectations about the model.

Prospective Job Seekers should know that most Indeed.com posting are in cities and the New England and Mid-Atlantic regions. The most prevalent fields for postings are Healthcare/Medical, and the fewest postings occur around the end of the year.

[1] Data from US Census Bureau

Team: 404 Found
Weijia Bao, Xinyi Cheng,Jessica Feng, Keting Yang

Slide 1: Insights for employees

How we processed Data:

1. Filter the data with the condition of “Job age day” equals to 1, to avoid overcounting.
2. Extract top 5 Categories (Industry) and admin1(states)
3. Process the correlation map of industry, states, and education level.

Conclusion:

- Jobs on Indeed.com normally does not require high school degree. Top 5 states have significant more more job supply A universal finding over the five states are most people working in Healthcare/medical.
- A student interested in healthcare could go any of the five states for higher job opportunities. CA has largest supply of jobs, especially for IT. But not Food Industry.

○ Food: FL, TX ○ Health: NY, CA ○ Others: CA, FL○ Retail: NY, PA ○ IT: CA. TX

Slide 2 +3: Insights for employers

How we processed Data:

1. Filter the data with the condition of “Job age day” equals to 1, to avoid overcounting.
2. Exclude all samples with the average rating equals to 0(non-rated company)
3. Extract the clicks, average rating and description length columns
4. Normalized the data to the same scale by subtracting the mean and dividing the max.
5. Feed our data to a neural network with one hidden layer of three nodes.

Conclusion:

The formula we found for predicting the click number with average rating and description length is clicks = 0.09-0.005*average rating-0.04*description length. This indicates very weak linear correlation between the number of clicks and the average rating and description length.

With the scatter plot, we can see companies with 30-40 ratings tend to have the highest number of clicks, but so do the companies with extreme rating as 10 and 50. The reason that there’s an decreasing trend for the rating between 40-50 might be the people who is rating the companies tend to be kind of “extreme” when they “like” the company and give 50 in general.

Smaller firms tend to post more jobs and also get more clicks. As the size of the company increase, they tend to post less jobs, but this trend cease when company size is larger than 1000+. We graphed total count clicks by company scale and job post, and the two graphs are similar. Thus we plot another graph of the ratio between them, which is clicks/(job*day). Averagely there are 75 clicks per job post per day, and this data is same for companies ranging from 0-49 employees to 1000+ employees. Therefore we conclude that the total number of clicks does not depend on size of the company.

Slide 4: Insights for Indeed

1. The employment distribution for each industry from Indeed is consistent with the one from U.S. Bureau of Labor. This implies that the posted jobs on Indeed.com are pretty diverse and cover all the popular industries.
2. We suggest Indeed.com to specify the industries on posted jobs from staffing firms. This will allow Indeed.com to have a better understanding of how the major industries of employment are distributed.
3. Indeed.com can provide our findings on rating and description length to employers when they post a job on Indeed.com.
4. Indeed.com can use that currently there are 75 clicks/job*day to promote sponsorship.

Team Pandas

5C DataFest 2018

March 25, 2018

Hi, we're Jon, Steven, Max, and Wyatt, and we'd like to introduce you to our friend Susan. Susan is a typical high school senior from California. All her life, she has been giving different answers to the question that all adults ask: "what do you want to be when you grow up?" But now that it's time to really decide, she wants to make an informed decision. That's why she asked us to analyse the job market for her. She gave us some criteria and asked us which industry she should get into.

Susan knows she wants to go to college (her parents will kill her if she doesn't) but she really has no idea what she's interested in studying. She wants to choose a career path that will ensure she can always find a job after college, where she gets paid a high salary. She would also like to know where she is likely to find a job once she decides on an industry. Susan definitely wants to stay in America, and she'd prefer not to stray too far from home, in the California Bay area (though she will follow the money).

To answer her questions, we looked through the Indeed database as well as data from the Bureau of Labor Statistics. We merged the two datasets together by lumping the high-resolution job categories from Indeed into the categories that the Bureau uses. For example, Indeed differentiates between different types of engineers, but we combined all engineering data.

From this, we boiled the jobs down to 22 categories, and we plotted those categories against the number of postings in that category on Indeed, the expected 10 year growth of that industry, and the median wages. The first two variables give Susan an idea of how likely she is to find a job and how likely it is that she won't get laid off or that she'll be able for find more jobs down the road. Again, one of her criterium was to have a high salary, so we made sure to plot that as well. We then plotted all of these variables on a 3D scatter plot so we could visualize which job fits all of these criteria the best. We did a little math in the background to normalize all the dimensions and find the optimal point, which is the one circled on the graph.

And what do you know? The Computer Science & Mathematics category fits our criteria the best! So let's take a look at where she might live and work.

This heat map shows where all the Computer Science & Mathematics jobs posted to Indeed are located. Darker regions indicate more jobs. From the previous graphs, we know the median wage for the Computer Science & Math category is about $82k, but as we can see, the areas where the most jobs are

are also where the wages are highest. Places like Wyoming bring the average wage down, but they also have a low cost of living. Computer Science jobs are common all across the country, but it is an industry where you're more likely to find work in large cities, such as LA or NY (as well as the Silicon Valley). So it looks like Susan can stay close to home after all! Though, if she changes her mind about that while she's at shcool, she can always choose to live somewhere else.

In summary, based on Susan's criteria of a high salary and good job prospects, she should study Computer Science or Mathematics. She can expect to make $82k or more. There are currently a lot of jobs available in this area, and she can expect there to be an increased amount of jobs by the time she graduates. Moreover, now Susan has all the information she needs to make an informed decisoin about her life.

Eros Erdős

Michelle Heeney, Abigail Boulin, Kirs Imsong, Roy Jackman, and Jack Kenney

March 25, 2018

Abstract

Trying to help employees find communities, and helping employers communicate effectively.

1 Objective

For our Data Fest project, we decided to apply machine learning methods to produce usable data for both the employees and the employers. Given the distribution of job offers across the United States, we built a model that can predict the quality of a company job posting based on a number of different parameters, so the employers can tailor their posts for specific locations and employee types. For the employees, we have made a heat map based on industry, normalized using city population so you can find a job in a community of similar, or maybe not so similar, industries.

2 Methodology

1. We made several linear regression models, the best of which explained 65% of the variation in the data. Thinking that we could do better than this, we applied a machine learning model known as Multilayer Perception Classification.

2. Applying the machine learning model of neural network classifiers using the fields employeeJobCount, employeeCount, descriptionLengthChars, licenseRequiredJob, noEducationRequirementsJob, highschoolEducationRequirementsJob, higherEducationRequirementsJob, supervisingJob, jobAgeDays, admin1, and city. We predicted the overall company rating using these fields with 75% confidence.

3. We wanted to represent interesting metrics geographically since jobs are so closely tied to the job markets of the cities in which they are posted. This is why we chose to use Tableau which has powerful pre-built mapping tools. We were able to plot industry size by city but we were running into a problem that larger cities were overly represented, which makes sense. To better show different city industries, we decided to normalize our data by city population. We went ahead and pulled in a US Population data from data.un.org and performed a left-outer-join with our original dataset from Indeed.com. We were now able to show the best places to post jobs, based off of industry, and normalized to better represent smaller cities. Heatmap Image

3 Citations

1. Indeed Jobs Data Set

2. UN Population Data

The Cutting Edge DataFest Write-Up

Main Goal:

The main goal of this analysis was to understand the discrepancy between observed search trends and job posting data from Indeed. Specifically, we found that many job postings originated in the Northern regions of the country, per capita, compared to the vast amount of searches that originated in the South.

(Failed) Attempts at Analysis:

- We originally had hoped to find differences and similarities between the different job postings using the K-Means clustering algorithm
    - We found this to not produce specific results due to the fact that many of the positions titles were specific to their industry
- We also used an LSTM Network to create a prediction forecast of total employment numbers in the US
    - This was not used as it was not relevant to our final presentation

Data Used:

We used the given data from indeed.com to get the city and state of the listings, the date they were posted, and the industry the posted job was it. Externally, we used data from the US Bureau of Labor Statistics, the 2010 US Census, and Google Trends.

Data Analysis:

- Determined Indeed's job listing throughout regions of the US
- Normalized to average the population throughout each state to show job listings per capita
- States with the most searches for "Indeed" is observed to have lowest job listings per capita
- Employment data from Bureau of Labour Statistics was used to make predictions about the industries the job searchers were coming from
- Trends for those industries were extrapolated to find potential increases in user base for Indeed

Strategies:

Industries that are falling could result in the unemployment of millions. We could sponsor companies and organizations that provide industry switching education programs. These programs will efficiently allow individuals to experience new opportunities quickly new positions towards employment.

Five College DataFest 2018
Team: We R Zixian Li, Zhu Shen, Jingyi Liu, Kelly Pien, Siyu Lily Qian

COMPETITIVENESS OF INDEED JOBS

We analyze job market circumstances for general job seekers based on Indeed's job postings data. While we recognize individuals have their own preferences, our analysis will focus on education level, region, specific industries, company size and rating, and job descriptions.

As a supplement to Indeed's data, we created two additional variables to help us analyze the job market. One is the general competitive ratio, which is total available positions divided by total applicants, to measure the probability of getting a job. We also used the popularity ratio, which is total applicants divided by the age of the job, as an indicator of the competitiveness of a job.

We categorized jobseekers into three types based on when they intended to enter the job market: after dropping out of high school, after high school graduation, and after pursuing higher education. Jobs with no education requirements are more numerous and are essentially as competitive as those requiring a high school diploma. 37,631 Indeed jobs require no education. Each of those jobs, on average, received 5.45 applications per day. The most jobs for high school dropouts and graduates are offered with the health care/medical industry, staffing firms, the food/beverage industry, the retail/consumer goods industry, and other industries. This shows that high school dropouts and graduates have a high chance of working in the same industry.

While college graduates (or those who pursue higher education) saw fewer Indeed jobs (9,700) requiring their higher education, they faced less competition: those jobs averaged only 4.08 applications per day. For people who pursue higher education, we made a map to show the general competitiveness of job market in each state. When choosing an academic institution, those students could look for colleges in states with less competition in the local job market. The states with the lowest job application ratios are Delaware, Alaska, North Dakota, and Kansas.

For college students who have preferences for particular job industries, we built decision trees to investigate the factors influencing the competitiveness of a job. Random forest model was also developed to detect the variable importance in determining the competitiveness of jobs in different industries. Because the data included so many different industries, we only focused on two specific industries in this project: Financial Services and Manufacturing. For these two industries, the description length of the job, popularity ratio, company size and age of a job are all significant factors deciding the competitiveness. For a job posted on Indeed, if the description of the job is relatively long, the job is recently posted, and the company size is large, the competitiveness of the job is predicted to be low. The error rate of the decision tree model on the test set is 30%. However, there is one difference between these two industries. Compared with manufacturing, jobs in financial services often require a license and the company rating level becomes a more significant predictor. Generally, between industries, the significance of the factors in determining the competitiveness of jobs may differ based on the structure of the industry and required skills, etc.

In conclusion, education level, region, different industries, length of job descriptions, company size and company rating can affect the competitiveness of a job. In the future, we want to build a shiny app to show the competitiveness of different jobs in various industries and combine external resources to conduct more rigorous research.

Datafest 2018 RudeGirlz Write-up

Resources:https://www.linkedin.com/pulse/new-survey-reveals-85-all-jobs-filled-via-networking-lou-adler/
https://www.bls.gov/news.release/pdf/jolts.pdf

Background:
An article on Linkedin from 2016 found that from doing a sampling survey of people in staff/management roles, networking was people's primary means for finding jobs, even trumping direct applying for a job.

Research Question: Is it true that the majority of job openings are not posted on online job search/databases like *Indeed*? If so, what can companies like *Indeed* do to expand their influence/clientele and in turn increase their profit?

By comparing data of the number of hires in 2017 to the total number of job postings on Indeed, we can find the discrepancy in the numbers of actual job openings and the ones that are posted online in *Indeed*.

Variables:

- Count of job posts (per industry, per months of September 2017~January 2018)
- Industry variables: Financial activities, Professional and business services, Education and health services, Manufacturing

External data: The data is from the Bureau of Labor Statistics in the U.S. Department of Labor. The data is presented in a report called Job Openings and Labor Turnover on January 2017, December 2017, and January 2018. This report contains comprehensive labor data compiled over the year of 2017 broken down by industries.

Results: Jobs posted on online job search databases like Indeed are only a small portion of the total jobs available. *Time series Plot*: to show gap between govt. data and Indeed data. 85-90% of jobs are not advertised, meaning only 10-15% of jobs are posted online.

Limitations:

- One job posting does not necessarily equate to one hire
  - Our Assumption: 1 job posting = 10 hire
  - This is the best we can do from what we have, and we believe this shouldn't be a huge problem because it will probably not affect the direction of our result
- Time series for only fall 2017-jan 2018 (time crunch)
- Our external data from government website is the compilation of all jobs in the U.S. including jobs that *Indeed* wouldn't post on their server (ex. Part-time to work in a local cafe, manual labor).

Implication: As there are many more jobs out there that are available but not posted on online databases like *Indeed, Indeed* has much potential to grow as long as it understands the status quo and expands its services through ways such as but not limited to:

Recommendation:

- Establishing special partnerships with companies to gain access to jobs that are not posted online but usually obtained through other paths such as networking