IEOR 135 (Data-X)

Employment Dynamics

Team 10: Mariam Germanyan, Aleksandra Ma, Taline Mardirossian, Riya Prahlad, Ayesha Yusuf

Mentor: James Hodson

Project Overview

Our main goal is to provide information for any party (employers, students, etc.) to be able to use our website interface on AI4good.org to see how and if graduating from UC Berkeley has an impact on future success in terms of employment. Our project is split into three main parts: 1) Nearest Neighbor matching to compare a Berkeley-graduate Bay Area worker in tech with the most similar non-Berkeley-graduate Bay Area worker in tech, 2) Calculating the number of promotions that UC Berkeley alumni receive in their careers compared to their closest matches at other universities, 3) Linear Regression Model to predict the time between graduation and first tech job based on workers' self-reported skills weights and their highest degree level. Our intended user interface is a Jupyter interactive widget that allows users to see various visualizations by choosing different features on the widget, as well as time series plots that show different visualizations/graphs which show trends in the timeline of when employees were hired in Bay Area's list of Top Tech Companies to work for.

Dataset Used

The dataset that we are using comes from our mentor from AI for Good Foundation, who extracted relevant employment information (for employees within the Bay Area) from websites that collect resumes data from its users (i.e. LinkedIn and ResumeBuilder). All the resume profiles are anonymous, with all the identification information removed. The whole dataset consists of 3.7 million unique resumes with 23 million rows and 33 columns altogether, where the column "ID" that we use to identify each unique resume profile can appear multiple times in the dataset representing different records of employment and education.

Techniques Used

Nearest Neighbor Algorithm

Number of Employment Records from Top Companies by School and Gender:

We gave everyone in the dataset *non_tech_berk* a score based on their similarities with each person in the dataset *tech_berk*. The scores are based on 1) *Skill Sets* and *Skillset Weights*, 2) *Degree Type,* 3) *Elite University,* 4) *Major*, with heavier weights on the first two factors. In this way, each Berkeley-graduate Bay Area worker in the tech industry will get matched to a non-Berkeley-graduate Bay Area worker in the tech industry with the greatest similarity score metric we defined, using samples of each dataset.

For each university shown in the graph, we took a sample from the worker profiles of that specific university alumni and matched it with the worker profiles of people who did not graduate from that university, and then plotted the number of employment records from top companies (See appendix for list), as shown in Figure 2a. By the same logic, we also plotted the number of employment records from top tech companies of male Berkeley graduates and female Berkeley graduates who are working in the tech industry in the Bay Area.

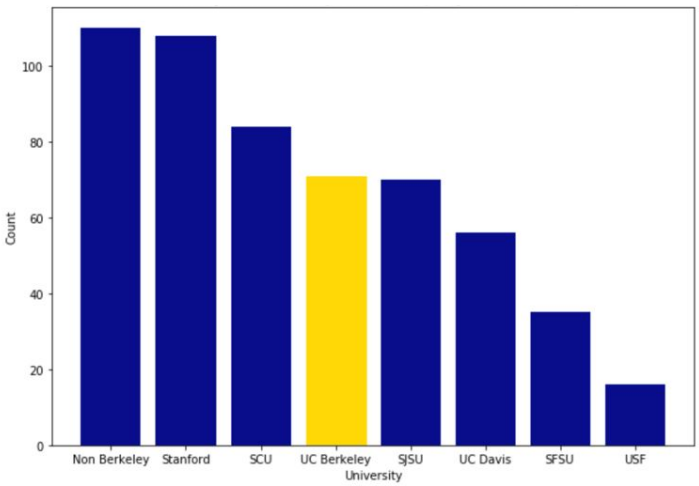

Figure 2a. Count of Bay Area University Alumni in Top Bay Area Tech Companies

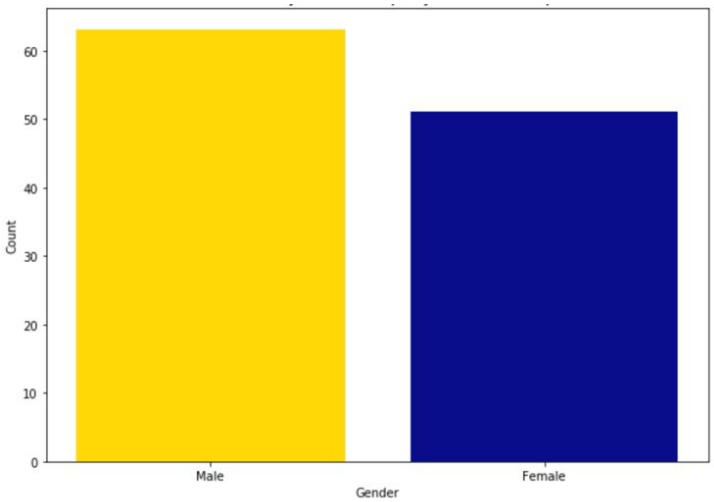

Figure 2b. Count of Berkeley Alumni in Top Bay Area Tech Companies Based on Gender

Number of Promotions Bay Area University Alumni Received (Unmatched vs. Matched):

We generated datasets for Bay Area tech industry employment records of other university alumni like we did with *tech_berk* to see the difference among the total amount of promotions of each school's alumni. In order to get the number of promotions for each unique ID, we counted the total number of companies one has worked for, counted the number of unique companies within the total number of companies, and subtracted the number of unique companies from the total number of companies. In the end, we summed up the total number of promotions we got from samples of different schools and plotted them based on unmatched IDs and matched IDs, as shown in Figure 2c & Figure 2d respectively.

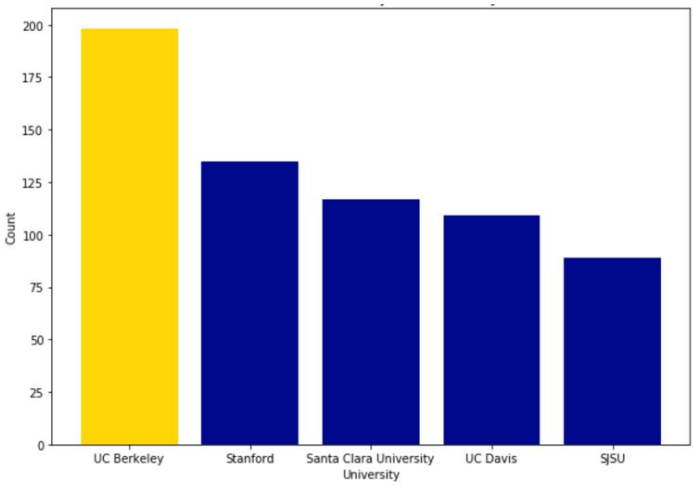

Figure 2c. Counts of Promotions of Bay Area University Alumni (Non-matched)

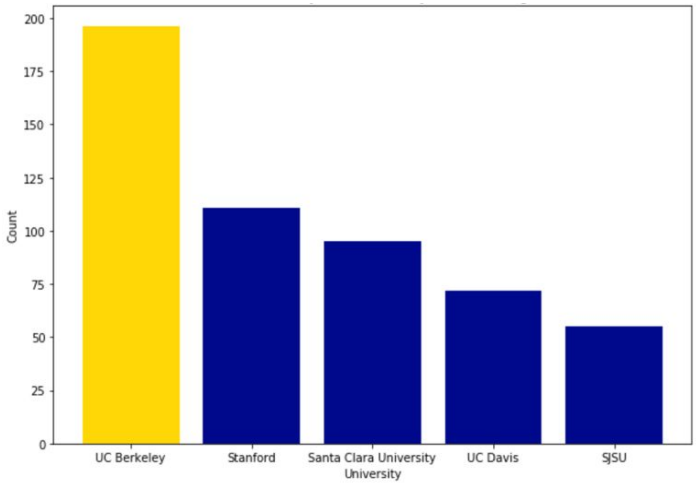

Figure 2d. Counts of Promotions of Bay Area University Alumni (Matched)

Linear Regression Model

We split the initial data into train, validation, and test subsets. This split was done on ID's, so as to retain all the rows of education/employment information for each unique person. After that, we carried out many manipulations on the data, including filtering out invalid start or end dates and produced a "duration" column which represented the time in days between the end date of their last earned education, and the start date of their first job post-graduation in the tech field. Once we extracted this date, we fit the linear regression model on the training set, and applied it to the validation to see our accuracy. Our MAE (mean absolute error) was around the same as the average duration, so we concluded that the features we utilized had fit the model relatively well.

Learning Path

1. Real-world datasets:

   This project gives us a hands-on opportunity to work and experiment with real-world data. After some exploration of different variables, we realized that real-world data can be very noisy, which makes it essential to understand what each variable means and the types of values each column contains. We had the most trouble cleaning the data, as data was extracted from online resumes, so the accuracy of the variables was difficult to quantify. Also, determining what we wanted to include and exclude from our subtables in order to get the most accurate representation of the population was quite challenging.

   It also took us a long time to run our functions since the dataset that we were provided is gigantic, consisting of 23 million rows. In order to optimize our time, we experimented with randomly selected samples from our dataset first and then applied our functions on the whole dataset.

2. Logic of Implementation:

   Our approaches to determine whether attending UC Berkeley leads to better outcomes in the long run were altered as we were exploring our dataset. Initially, we wanted to create a model to predict first promotions in order to see which job offer would be more optimal using the “Length at Job” column, which we ultimately couldn’t use due to it being incorrect in certain instances. Making the most of the few numeric columns we were given, we then decided to create a regression model using highest education level and skill weights as the main feature. We also decided to make the assumption that if someone had the same company listed multiple times in their employment history, they moved up in role and that would count as a promotion. This would help us avoid hours of manually going through every job title listed to determine some sort of hierarchy. Looking back, we could potentially have mapped a few generic keywords in job titles such as “Junior” and “Senior” to some sort of numeric values with corresponding magnitude to get a more accurate measure of promotions and resolve the potential edge case our method would not consider - if someone went directly to “CEO” of their own start-up, for example.

   Some difficulties that we faced along the way included candidates whose birth year was the default value of 2001; essentially, this meant that these resumes did not hold any educational records.

Conclusions

1. Based on our graph in *Figure 2a*, it seems that it isn't necessarily easier or harder for UC Berkeley grads to get into top companies.
2. Berkeley graduates get their first tech jobs 6% faster than other graduates.
3. Berkeley graduates get more promotions in total compared to other graduates.

Appendix

1. List of top tech companies used in analysis:
   Google, Nvidia, Intuit, Oracle, Cisco, PayPal, Salesforce, LinkedIn, Qualcomm, Amazon, Microsoft, IBM, Airbnb, Uber, Lyft, Facebook, Tesla, Genentech, VMWare, Hewlett-Packard (HP), Apple, Intel, and any other company titles that included these companies.