STEM Labor Market Analysis

Zhexu Li, A14532514

Department of Electrical and Computer Engineering, University of California, San Diego

Abstract

In this project I explore a STEM jobs dataset collected from 2017 to 2021. I analyze the distribution of STEM job opportunities and income, and how they relate to different levels of education, years of experience, and locations. I also preprocess data and select features based on the data analysis. In the end I use several machine learning models to predict the income given features selected and analyzed during the data analysis process, and compare their performance.

I. Introduction

STEM (Science, Technology, Engineering and Mathematics) is the engine of the U.S economy, in fact, according to [1], 69% of the U.S GDP are supported by STEM. Every year, hundreds of thousands of college students in the U.S choose to study in different STEM related fields, planning their careers for a hopefully promising future. While it has become common sense that STEM workers were some of the highest paid people in the US, with the drastic increase of STEM graduates in recent years and the workforce shortage in certain non-STEM fields, it's hard to say if STEM jobs still bring top income. Also, different STEM positions require drastically different educational and experience backgrounds, while income and job availability differs significantly in different states and cities. So for people who have already decided to pursue a career in STEM fields, it's important for them to understand how education and experience level affects their future career, and what places provide the best opportunities. In this project, I will investigate these issues by analyzing the distributions of job opportunities, income, education, location, and the relations between them. After the analysis I will use several machine learning models to predict the income based on one's background and compare their performance. This study could provide valuable insights about the STEM job market, and help current or prospective STEM students to make better career decisions.

II. Dataset

The dataset I used for this project is the Data Science and STEM Salaries Dataset uploaded by Jack Ogozaly on Kaggle. It contains 62462 records of STEM jobs and their corresponding information, stored in 29 columns (some columns are one hot encoded so there are 20 features in total). The dataset was scraped by the uploader from a famous salary posting / job discussion website levels.fyi, and it's a bit dirty in its raw form. There are 32092 records with missing values in columns related to the analysis. Because most missing values (education levels, locations) are not possible to impute without affecting the integrity of the dataset, I believe it's better to drop those rows. Also, I noticed there are titles like "Human Resources", which doesn't seem to be a STEM job, and after some research I chose to include those records in the dataset because apparently there are companies who consider them as STEM positions and require STEM degrees. After the data cleaning process there are 30370 records remaining, the records were posted from July 2017 to August 2021 and include 15 different titles in 1632 companies around the world. All records were verified by proof documents when uploaded to the levels.fyi website.

III. Data Analysis, Visualizations

In this section I will perform data analysis to extract valuable insights from the data to help us understand how education, location and other factors affect one's STEM career. These insights might also be helpful for constructing model later. There will be a lot of visualizations and I will try my best to creates the finest plots.

We first analyze the distribution of annual total income in our dataset. Annual total income is calculated by base salary + stock grants + bonus

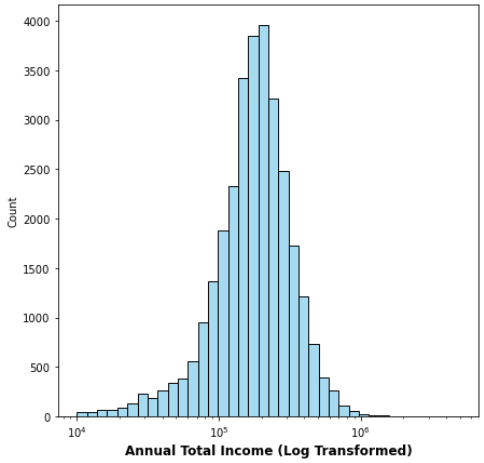

Fig. 1 Log Transformed Distribution of Annual Total Income

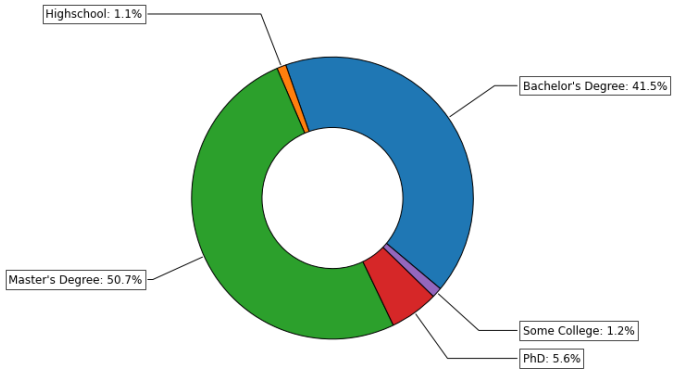

Fig. 2 Distribution of Education Levels of STEM positions

and the unit is in dollars. In figure 1 we can see the distribution of annual total income in our dataset is roughly normal. Notice I applied log transformation when plotting this histogram, because the original distribution was heavily skewed right. The average annual earning in our dataset is 206452 dollars, the median annual earning is 180000 dollars, and the standard deviation is 133967. And according to a recent study in [2], the top 10% annual U.S household income in 2021 is around 201000 dollars, so it seems STEM jobs indeed still bring high income. But we should be aware that records in our dataset were provided by people who are willing to share information about their income, so it's likely to be biased.

Knowing the extremely high income of these STEM people, a natural question to ask is what educational level one needs in order to be qualified for one of those jobs? So now we analyze the distribution of education levels. There are 5 different education levels in the dataset: Highschool, Some College, Bachelor's Degree, Master's Degree and PhD. Based on figure 2, over 97% of STEM positions recorded in the dataset were occupied by people with a Bachelor's Degree or higher, which suggests that people really need to earn at least a Bachelor's Degree in order to get a STEM job. Interestingly, the proportion of Master's Degree holders (50.7%) is higher than the proportion of Bachelor's Degree holders (41.5%) by 9% in our dataset, which further suggests that the competition for these positions is getting so fierce nowadays, and a Master's Degree might get you some advantages when applying for STEM positions.

Knowing the insane competition for these positions, we then analyze how education level affects one's total annual earning. In figure 3 we can see, while the average annual income generally increases with higher degrees, it seems Highschool graduates and Some College degree holders have higher average annual income than Bachelor's Degree holders. I

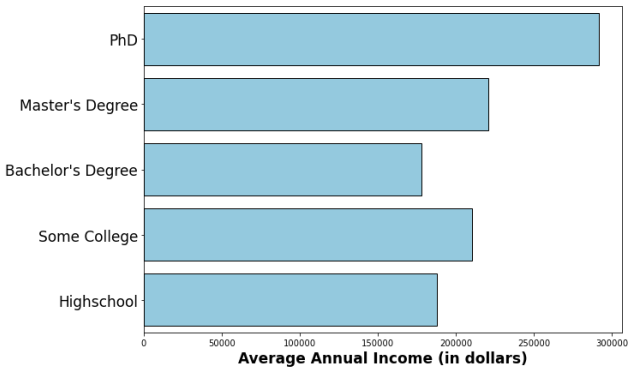

Fig. 3 Average Annual Earning vs. Education Level

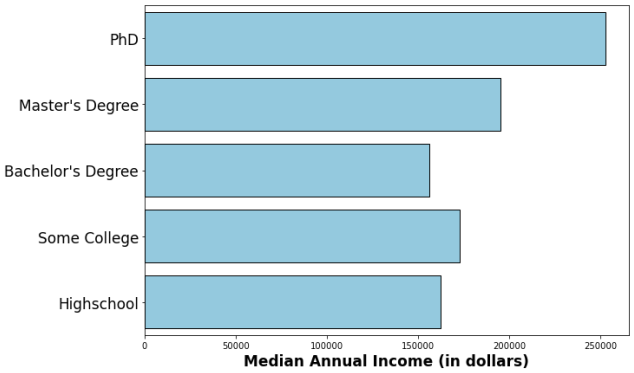

Fig. 4 Median Annual Earning vs. Education Level

initially thought it might be caused by outliers, but in figure 4 we observe the same trend with medians. But as we mentioned above, finding a STEM job without a Bachelor's degree or higher is extremely difficult. And considering the massive increase in earning from Bachelor's to Master's and from Master's to PhD, it seems higher education level indeed leads to higher income. So basically, if you want to get a high paid job in the STEM industry, you may want to earn a higher degree.

Other than education, experience is also an important asset in the STEM job market. In figure 5 we can see the distribution of years of experience in our dataset is heavily skewed right, with 31% positions occupied by people with < 2 years of experience, and 78 % people has < 10 years of experience. To better analyze and visualize this feature, we put years of experience into 5 groups: 0 - 2 years, 2 - 5 years, 5 - 10 years, 10 - 20 years, and 20+ years, and generate a box plot for these groups. Based on figure 6, the annual total income gradually increases with the increase in years of experience, which is expected. To better understand the trend, we fit the dataset to a linear regression model with x = years of experience and y = annual total income. The resulting model has a coefficient of 9290.7, which indicates you earn 9290.7 dollars in addition for every additional year of experience you have. The intercept is 139450, which means the typical total annual income (in our dataset) for

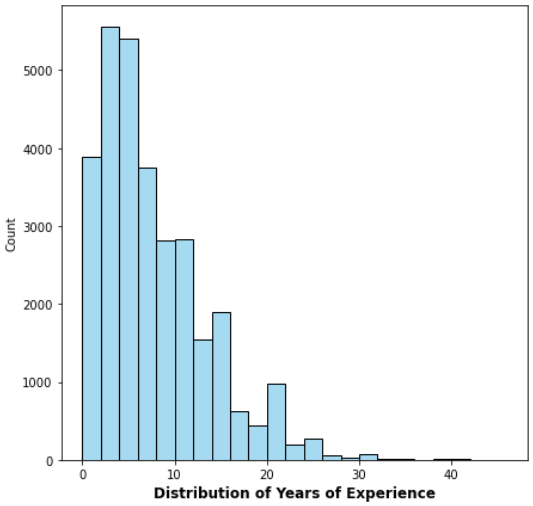

Fig. 5 Distribution of Years of Experience

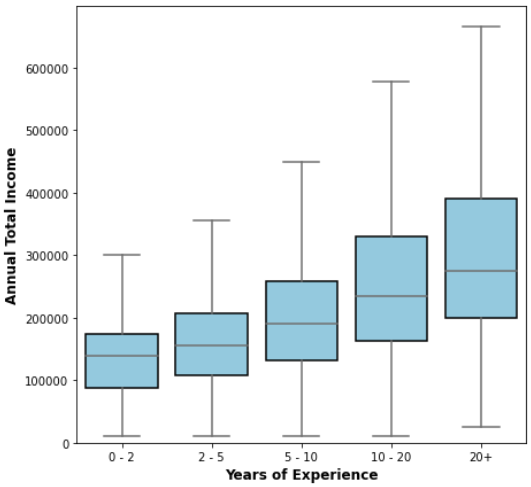

Fig. 6 Annual Total Income vs. Years of Experience

someone with 0 years of experience is 139450 dollars. Even though we know STEM workers are usually well paid, it's still surprising to see someone with 0 years of experience get paid annually over 130k dollars. And the extra earning you get from every additional year of experience is considerably large.

Lastly we analyze the distribution of job opportunities and incomes based on location. We first plot the number of STEM positions in different states, normalized by its population in 2020 collected in [3] (otherwise it will become a choropleth of population). In figure 7 we can see most STEM job opportunities are scattered in California, Washington State and New York. In fact, 71.8% STEM jobs in our dataset are located in these 3 states. Notice that we don't have any data for North Dakota, Hawaii and Alaska which suggests opportunities there might be quite limited. We then plot the average annual income in different states, and based on figure 8 we can see California has the highest average annual income (267700 dollars) for STEM employees, followed by Washington State (245000 dollars) and New York (225000 dollars). So if you are looking for a STEM career, California, Washington State and New York are your best choices, which are actually common sense but now supported by our analysis.

Besides states, we can also visualize which cities are popular locations for these high income STEM jobs in our dataset. A word clouds is an aesthetically pleasing choice for this situation. In figure 9 we observe many familiar places in the plot, like San Jose and Menlo Park. And it seems Seattle, New York City, and San Francisco are undoubtedly the three best cities in terms of STEM opportunities. And San Diego also appears to be a popular choice, which is quite exciting for us. But still, we should understand most cities in the plot are population centers, so more opportunities in these cities also means more intense competitions.

IV. Preprocess, Predictive Modeling

Having a solid understanding of our dataset through data analysis, we now start building our prediction models. We will predict annual total income based on level of education, years of experience, and

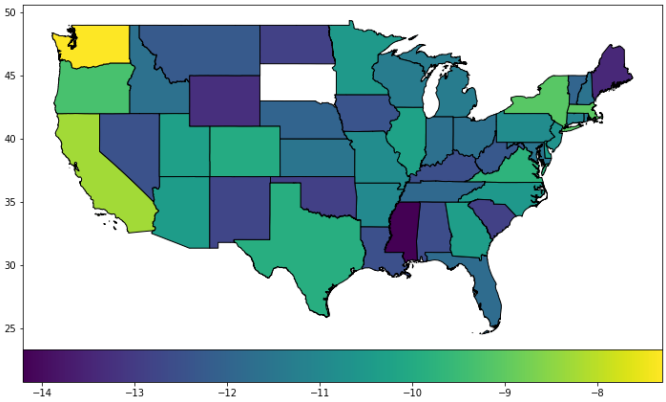

Fig. 7 Number of STEM Jobs by States, Log Transformed

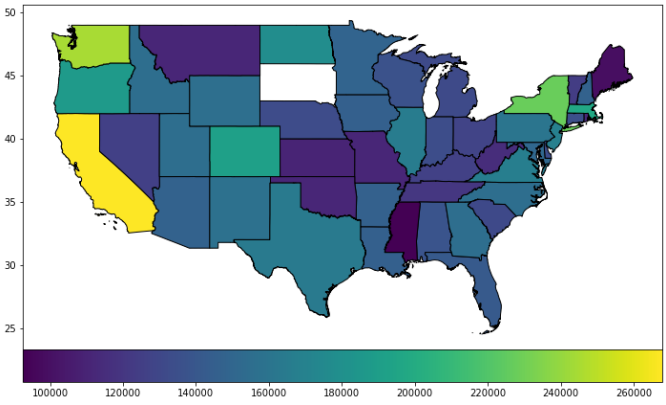

Fig. 8 Average Annual Income by states

Fig. 9 Word Clouds of Top Locations

location. As we mentioned above, we will use only the features we analyzed for our model. We first preprocess the dataset based on our analysis above. We apply ordinal encoding on education level, since level of education is an ordinal variable. We then apply one hot encoding on location, because it's a nominal feature. Years of experience doesn't need any preprocessing because it's already a numerical variable. Because income varies drastically for different job titles, we believe it's better to make predictions within a given job title. We choose Software Engineer because it's the most popular title in our dataset, and we have 15027 records in the cleaned and processed dataset. We use 90% for training and 10% for testing.

We will use three models: Decision Tree Regressor, Random Forest Regressor and Multi-Layer Perceptron Regressor. Decision Tree Regressor is a simple tree model, Random Forest Regressor is a more complex ensemble learning algorithm, and MLP Regressor is a neural network model which is the most complex model in our list. All these models are popular choices for predicting numerical variables using a combination of categorical and numerical features. The prediction accuracy is evaluated by MAE (Mean Absolute Error) because according to our analysis above there are quite a lot of outliers which makes RMSE and MSE inapplicable.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

We then fit models using training data and make predictions. The resulting MSE for every model is displayed in figure 10. Based on our results, the Random Forest Regressor has the best performance, and interestingly, the most complex neural network model MLP Regressor has a higher MAE than the simplest Decision Tree Regressor. This might be because the MLP Regression minimizes MSE using

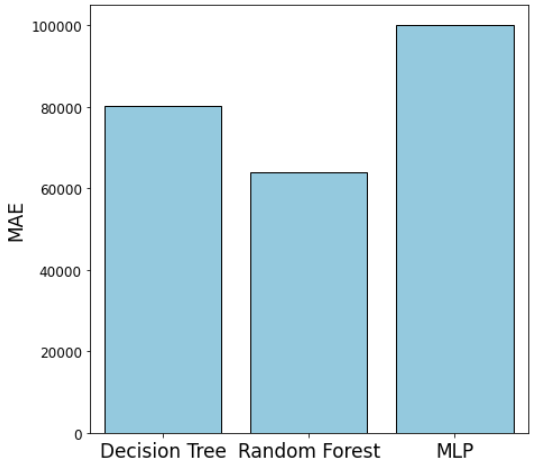

Fig. 10 MAE for Different Models

functions specifically designed for MSE instead of MAE, while Decision Tree Regressor and the Random Forest Regressor were trained by minimizing MAE (set criterion = absolute error). While Decision Tree Regressor is not the best model here, it has an advantage that it can be simply interpreted, which might provide some valuable insights about the features we used. We extract the feature importance from the model, which returns 0.1105 for education level, 0.43 for years of experience, and the remaining are for one hot encoded columns of locations. Surprisingly, based on our Decision Tree model, for annual total income, years of experience is more important than level of education and location. It suggests in the STEM industry, experience is vital for high income.

V. Project Conclusions

In conclusion, we analyzed how level of education, years of experience, and locations affects one's career in the STEM industry. Based on our analysis, a Bachelor's Degree is a minimum requirement for STEM positions nowadays, and Master's Degree is becoming more and more popular. A higher level of education also leads to higher income. For years of experience, while most people in the STEM industry have <10 years of experience, the income is positively correlated with income, and the extra

income you gain from every additional year of experience is 9290.7 (dollars). And for location, based on our analysis, California, Washington State, and New York State are the top three states with the highest number of STEM opportunities and average income for STEM employees. And within these states, Seattle, San Francisco, and New York City are the top 3 popular choices for STEM opportunities. In our predictive modeling process, we constructed three different models for predicting annual total income based on level of education, years of experience, and location. We discovered that Random Forest Regressor is the best model that works with our dataset.

VI. Limitations, Future Work

As we mentioned before, because it's a data exploration project, the scope of this project is limited to this one dataset. And since the records were provided by people who are willing to share information about their income and have access to the levels.fyi website, it's likely to suffer from sample bias. And we have observed class imbalance in our dataset especially for titles and education levels, which limits the generalizability of our models. In order to draw better conclusions and improve model generalizability, more representative datasets are needed. Also, the scope of our analysis is limited to the U.S STEM job market, but if we have accurate and representative dataset from other countries in other domains, we can extend our analysis to those places and possibly draw more interesting insights.

VII. References

[1] STEM Supports 67% of U.S Jobs, *Eos*, 2020/1/28, https://eos.org/agu-news/stem-supports-67-of-u-s-jobs.

[2] Average, Median, Top 1%, and all United States Household Income Percentiles, *DQYDJ*, 2021/5, https://dqydj.com/average-median-top-household-income-percentiles/ FSFSFS F4F4F4

[3] Population by States in 2020, *USDA*, 2021/10/8, https://data.ers.usda.gov/reports.aspx?ID=17827