Working Status Analysis of H-1B Visa Workers Using H-1B LCA Disclosure Data (2020–2024)

Yufei Wu

**Authors**

* Yufei Wu (ORCID: 0009-0003-5051-0623)

**Author affiliations**

1. Department of Chemistry, University of Georgia, Athens, GA, USA.

Corresponding author: yufeiwu@uga.edu

# 1. Summary/Abstract

To be added.

# 2. Introduction

## 2.1 General Background Information

H-1B is a nonimmigrant visa category that allows employers to hire foreign workers in specialty occupations, like methematics, engineering, technology, and medical sciences (1). It requires the employee to have a bachelor’s degree or equivalent in the specific specialty. Typically, the maximun duration of an H-1B visa is six years.It is the most common work visa in the US. There are 65,000 available H-1B visas each year, with 20,000 additional visas for candidates with a master’s or doctorate degree from a U.S. institution. If there are more than 65,000 applications, USCIS will run a lottery to decide who can file an H-1B petition. As USCIS is receiving more and more H-1B registration these years, it’s harder for a foreign worker to get an H-1B visa. The Labor Condition Application (LCA) is an application filed by employers to apply for work authorization on behalf of employees as a prerequisite for H-1B. The LCA contains essential details such as job title, wage, and location about the proposed H-1B employment. Therefore, analyzing the LCA data can provide some ideas on the H-1B application status (2).

## 2.2 Questions/Hypotheses to be addressed

This analysis will primarily focus on certified cases. The key research questions to be examined include:

(1)Distribution Analysis:

What is the distribution of certified cases across various features such as SOC titles, locations, wages, and employer names?

(2)Variable Relationships:

How do wages vary across different states, SOC titles, and employers?

(3)Employment Trends and Remote Work Patterns:

How have employment patterns evolved over time?

What trends can be observed in remote work prevalence before and after the pandemic?

(4)Predictive Modeling:

Can a predictive model be developed to estimate wages based on job-related and employer-specific features?

Which factors contribute most significantly to wage determination in certified cases?

# 3. Methods

## 3.1 Data aquisition and cleaning

The dataset is H1B LCA Disclosure Data (2020-2024) from the U.S. Department of Labor. The data is provided by employers and includes final determinations issued by the Department’s Office of Foreign Labor Certification (OFLC) under the Employment and Training Administration (ETA). It includes information such as case status, job title, Standard Occupational Classification (SOC) title, location, and wages. It can be easily downloaded from [Kaggle](https://www.kaggle.com/datasets/zongaobian/h1b-lca-disclosure-data-2020-2024/data).

To ensure data quality, the dataset was filtered to include only certified cases and those specifically related to H-1B visa applications.

Additionally, the dataset was refined to include only the following essential variables:

RECEIVED\_DATE (date the application was received); SOC\_TITLE (occupational title classified under the SOC system); EMPLOYER\_NAME (name of the employer); EMPLOYER\_STATE (employer’s location); WORKSITE\_STATE (location of the worksite); WAGE\_RATE\_OF\_PAY\_FROM (wage offered to the worker); WAGE\_UNIT\_OF\_PAY (unit of pay values); PREVAILING\_WAGE (prevailing wage for the job); PW\_UNIT\_OF\_PAY (unit of pay values)

Observations with missing values in the selected variables were removed. All character columns except state names were standardized to lowercase.

## 3.2 Statistical analysis

### 3.2.1 Exploratory/Descriptive Analysis

The method began with an exploratory analysis to assess the overall structure of the dataset. Univariate techniques were applied to examine the distribution of key variables, such as job titles, employers, and geographic locations, by analyzing application counts over a multi-year period. Additionally, the distribution of prevailing wages was evaluated using visualizations to understand the range and concentration of wage values.

### 3.2.2 Initial Statistical Analysis

For the initial statistical analysis, a half-month subset of the data (from January 1–15, 2024) was selected. The continuous wage variable was then converted into categorical ranges to facilitate further analysis. Bivariate analyses were conducted using Random Forest classification models with individual predictors, including SOC title, employer location, worksite location, and employer name. These single predictor models were compared against a null model to assess their predictive utility.

### 3.2.3 Full Analysis

In the full analysis phase, multiple advanced modeling techniques were employed to predict wages. The methodology involved developing models using Random Forest, XGBoost, and SVM, with a comprehensive cross-validation approach (5-fold with 5 repetitions) to ensure robust evaluation. A comparative analysis of the training and test performance of these models was carried out to identify the approach that best balanced model complexity and generalization capability for wage prediction.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

### 4.1.1 Univariate Analysis: Distribution Analysis

The [Figure 1](#fig-top10_soc_title) displays the 10 job titles with the highest number of applications filed between 2020 and 2024, with “Software Developers, Applications” leading at over 500,000 cases. The [Figure 2](#fig-top10_emp_name) highlights the top 10 employers by application count, with “COGNIZANT TECHNOLOGY SOLUTIONS US CORP” topping the list at approximately 93,000 cases. The [Figure 3](#fig-top10_emp_state) presents the 10 states with the most applications, with California recording over 600,000 cases. Meanwhile, the [Figure 4](#fig-distribution_PREV_WAGE) illustrates the distribution of prevailing wages, revealing that most cases offer wages near $100,000 per year, while fewer cases fall at the lower end (around $15,000) or approach higher values near $800,000 (see Supplement Table 1).

|  |
| --- |
| Figure 1: Top 10 Job Titles. |

|  |
| --- |
| Figure 2: Top 10 Employers. |

|  |
| --- |
| Figure 3: Top 10 Employer Locations. |

|  |
| --- |
| Figure 4: Distribution of Prevailing Wage. |

### 4.1.2 Bivariate Analysis

The [Figure 5](#fig-wage_by_soc_title) shows the 10 job titles with the highest median wage, and they are all medical practitioners. The [Figure 6](#fig-wage_by_employer) highlights the 10 employers with the highest median wage, spanning healthcare, financial services, life sciences, technology, and manufacturing. The [Figure 7](#fig-wage_by_state) indicates that Washington has the highest median wage by state. Plots for the 10 lowest median wages are provided in Supplement Figures 1–3. The [Figure 8](#fig-remote_trend_absolute) illustrates the trend of remote work over the years. The number of remote work cases surged from 2019 to 2020 due to the pandemic and remained relatively stable from 2020 to 2023. Post-pandemic, there was a slight decline from 2023 to 2024. However, as remote work patterns became more established, the decrease was not significant. This analysis uses absolute values instead of proportions because remote work is identified based on discrepancies between employer and worksite states. Some remote work cases may not be captured under this definition, potentially underestimating the actual proportion. Therefore, absolute values provide a more accurate reflection of the remote work trend.

|  |
| --- |
| Figure 5: Top 10 Job Titles by Highest Median Wage. |

|  |
| --- |
| Figure 6: Top 10 Employers by Highest Median Wage. |

|  |
| --- |
| Figure 7: Top 10 States by Highest Median Wage. |

|  |
| --- |
| Figure 8: Remote Work Trends Over Years. |

## 4.2 Initial statistical analysis

The variable wage was selected as the outcome of interest to study the wages of H-1B workers. I first conducted a bivariate analysis to identify potential predictor variables. To perform the analysis, I selected a half-month subset of the data.

The [Figure 9](#fig-distribution_WAGE_Jan2024) illustrates a histogram of wage distribution during January 1–15, 2024. The continuous wage variable was categorized into four ranges: “(0, 75,000)”, “(75,000, 150,000)”, “(150,000, 225,000)”, and “Above 225,000”, converting it into a categorical variable.

A random forest model was applied for classification using a single predictor, with 1,380 observations in the training set and 346 observations in the testing set. [Table 1](#tbl-combined_accuracy1) to [Table 4](#tbl-combined_accuracy4) summarize the model fit for the predictors: SOC title, employer location, worksite location, and employer name. The results indicate that these variables exhibit moderate predictive performance.

Next, a multivariate random forest model was developed. [Table 5](#tbl-combined_accuracy_all) shows that the multivariate model outperformed the bivariate models.

All the models performed better than null model ([Table 6](#tbl-combined_null_accuracy)), with the multivariate model performed best.

|  |
| --- |
| Figure 9: Distribution of Wages (January 1–15, 2024). |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Model Fit for Wage Based on Job Title.   | dataset | .metric | .estimator | .estimate | | --- | --- | --- | --- | | Train | accuracy | multiclass | 0.6985507 | | Test | accuracy | multiclass | 0.6531792 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Model Fit for Wage Based on Employer Locations.   | dataset | .metric | .estimator | .estimate | | --- | --- | --- | --- | | Train | accuracy | multiclass | 0.6195652 | | Test | accuracy | multiclass | 0.6242775 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Model Fit for Wage Based on Worksite Locations.   | dataset | .metric | .estimator | .estimate | | --- | --- | --- | --- | | Train | accuracy | multiclass | 0.6268116 | | Test | accuracy | multiclass | 0.6763006 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Model Fit for Wage Based on Employer.   | dataset | .metric | .estimator | .estimate | | --- | --- | --- | --- | | Train | accuracy | multiclass | 0.8673913 | | Test | accuracy | multiclass | 0.6936416 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5: Model Fit for Wage Based on Job Title, Employer Location, Worksite Location, and Employer.   | dataset | .metric | .estimator | .estimate | | --- | --- | --- | --- | | Train | accuracy | multiclass | 0.8217391 | | Test | accuracy | multiclass | 0.7341040 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 6: Null Model.   | dataset | .metric | .estimator | .estimate | | --- | --- | --- | --- | | Train | accuracy | multiclass | 0.6065217 | | Test | accuracy | multiclass | 0.6156069 | |

## 4.3 Full analysis

The initial statistical analysis shows that variables SOC title, employer location, worksite location, and employer name can serve as predictors for outcome variable wage. Based on that, Random Forest, XGBoost, and SVM were used to fit the model, including cross-validation (5-fold and 5 times repetition).

Tables summarizing the performance of the three models can be found in Supplement Table 2-4. [Figure 10](#fig-model_compare) shows the cmoparison of model performance among Random Forest, XGBoost, and SVM. SVM achieves the highest accuracy on the training data (0.886), while Random Forest (0.820) and XGBoost (0.738) show slightly lower training accuracy compared to SVM. On the test set, random forest outperform among the three models with an accuracy of 0.740. The SVM model’s high training performance compared with its relatively lower test performance (0.717) indicates that it might be capturing noise or very specific patterns in the training data that do not carry over to unseen data. In contrast, Random Forest keeps a better balance between fitting the training data and maintaining generalization on the test data. Therefore, the Random Forest model can be chosen to predict the wage since it appears more robust on the test data.

|  |
| --- |
| Figure 10: Comparison of Model Performance: Random Forest, XGBoost, and SVM. |

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References

1. The H-1B Visa Program and Its Impact on the U.S. Economy. *American Immigration Council*. 2016;

2. Mahajan P, Morales N, Shih K, et al. The impact of immigration on firms and workers: Insights from the h-1b lottery. 2024;