Exploring Factors of Influence on H-1B Data

Yulin Lei
yulin.lei@duke.edu

Qi Wang
qi.wang3@duke.edu

Hanqiu Xia
hanqiu.xia@duke.edu

Xin Xu
xin.xu3@duke.edu

Youyuan Zhang
youyuan.zhang@duke.edu

Abstract

This paper intends to investigate the factors that would potentially affect the success rate of obtaining H-1B visas for foreign workers. Classical studies on this matter merely focus on individual characteristics, like race and gender, whereas we conduct a more sophisticated analysis by including both individual and company factors. With application status (certified and denied) as the target variable, we apply Frequentist Logistic regression, Frequentist Lasso regression and Bayesian Probit regression for our analysis. We also employ the Principle Component Analysis and the test on Out-of-Sample data to ensure the efficiency of our model. The results from the three above models are consistent, which indicate that job-wise factors, such as salary per year and job types, as well as application-wise factors, such as the length of application process and month of submission, can significantly influence the final outcome of the H-1B visa decision.

1 Introduction and Motivation

1.1 Introduction

H-1B is a program that allows employers to employ qualified foreign workers in the United State based on a nonimmigrant purpose. Current laws limit the number of qualifying foreign workers to 65,000 every year with an additional 20,000 quota under the H-1B advanced degree exemption [1]. As the difficulty of obtaining H-1B visas increases over the past several years, foreign students who plan to work in the United States are interested in how they could increase their chances of getting H-1B visas.

1.2 Motivation

As foreign students ourselves seeking H-1B visas in the future, we are interested in analyzing the available data and finding the possible influential factors on the certified rate of H-1B visas. The existing studies on H-1B visas are mostly outdated and revolve around individual characteristics that affect the chances of being certified. Our current dataset extends beyond individual characteristics and provides sufficient information for the employers characteristics.

By rudimentary exploratory data observation, we can make several assupmtions. One possible significant factor is the companys location. It would allow us to determine which state has the highest chance of being certified. Another area of interest is what types of jobs provide the best chance. Other interesting questions include the month that the case was submitted, the length of the decision period, and wage. Through a thorough analysis on the data, we hope to find significant factors on the certified rate and provide insights to both policy makers and applicants to enhance their chances of getting H-1B successfully.

2 Methods

Since the dependent variable is whether an applicant can obtain the H-1B visa or not, we can conduct binomial logistic regression analysis. We will also produce overall descriptive statistics for all variables.

After conducting Principle Component Analysis to detect the most important variables, we will consider conducting Frequentist Logistic regression, Frequentist LASSO regression and Bayesian Probit regression. We expect to be able to determine the most important variables that are crucial to obtaining the H-1B visa. Our regression model may also include interactions between groups of different characteristics. Since adding interaction terms may cause overfitting, we extract part of the sample as Out-of-Sample and calculate the In-Sample and Out-of-Sample misclassification rate for cross validation.

3 Data and Application

3.1 Data

With the OFLC (Office of Foreign Labor Certification) Performance Data we get from the United States Department of Labor, we obtain data about the information of people who applied for H-1B visa, including application status, submitted date and decision date, job title, location, etc.

3.1.1 Data Cleaning

To make sure that we can process the data efficiently, we clean the data and transform them into forms that are convenient for analysis. The detailed methods are as follows:

First, only keep the subjects whose visa classes are H-1B, and their H-1B statuses are certified or denied. Second, select all variables which would potentially affect the results of H-1B as follows with their corresponding variable names in analysis: wages (*wage*), length of application process(*days*), company location(*location*), company categories(*academic*), job types(*company*), full time or not(*fulltime*) and month of submission(*months_sub*). Third, because of the heterogeneity of the wages unit, we standardize the unit of wage into salary per year. In order to make wage distributed normally, we take log on the wage and also remove the potential outliers (Figure 4). Next, divide the location of companies into four parts: Northeast, West, Midwest and South based on their states location, and remove those not in the 50 states. Then, use Document Term Matrix to get the frequency of words of company and job names. Classify job types into Computer Science, finance and others, and split company categories into academic and non-academic (Table 4 and 5). Finally, apply complete case analysis on missing data.

Because we have prior information about the proportion of obtaining H-1B being about 35%, while the data we get contain 90% certified rate, we use stratified sampling for our analysis[2]. Since we have 10,603 denied data, we sample 1/2 of reject sample size from the certified data to maintain the real certified proportion. Besides, at the beginning of our analysis, we sample 10% for later Out-of-Sample test and use the left 90% to build up our model.

3.1.2 Descriptive Statistics

To get an overall view of our data, we present descriptive statistics. To show the difference in obtaining H-1B throughout the US, we observe by drawing a heated map (Figure 1). The complete interactive map we create can be found here: `http://rpubs.com/Darien/130817`. By calculating the number of days between decision date and submit date, we can observe the length of the applications from Figure 2. By extracting the month information of submit date, we can find the months with the highest probability of getting H-1B from Figure 3. Through interaction plots in Figure 5, we see that interactions could be significant in our model.

3.2 Implementation

3.2.1 Principal Component Analysis

Firstly, we conduct Principle Component Analysis on all of our interested independent variables. Our purpose is to reduce dimension and find out the most important factors that will affect approval of H-1B visa. Since most of the variables are not quantitative, we use `FAMD` function in {`FactoMineR`} package to conduct Factor Analysis for Mixed Data.

As can be seen from Figure 6, the first two components are able to explain only 6.87% and 5.81% of the variance respectively. The angles between the arrows represent the correlations between two variables. Since all of the arrows are in the top right quadrant and the angles are quite small, it is hard to extract an effect principle component. Due to the poor interpretation percentage of the first two principles and the hardness of composing the principle components, PCA is not an ideal way of dimension reduction for this dataset.

3.2.2 Frequentist Approach

We perform a frequentist logistic regression analysis and modeling selection using AIC. We first create a logistic regression model with all the independent variables of interest, and then use a backward selection method to find the best fitted model for our data.

Next, we perform a lasso regression using the `glmnet` package. The selected models generated from the two methods are the same: *status* $\sim$ *wage* + *days* + *months_sub* + *company* + *academic* + *location* + *fulltime*. All variables of interest have significant influence on the rate of success in obtaining H-1B visa. Table 2 shows the coefficients of the two methods, and we can see that they are quite similar. The misclassification rates for both approach are at around 34%.

We also conduct a same set of analyses incorporating interaction terms in our model. Even though the misclassification rates indicate a better fit with model selected from forward selection based on AIC, it is a much larger model and a problem of overfitting could occur. The full model is as follows *status* $\sim$ *company* + *days* + *wage* + *months_sub* + *fulltime* + *location* + *academic* + *days:wage* + *days:months_sub* + *company:location* + *wage:location* + *company:wage* + *wage:academic* + *company:academic* + *company:fulltime*.

3.2.3 Bayesian Approach

Since our response *status* is a categorical variable, we consider using probit regression here. To do Gibbs sampling, we use the latent variable construction where $y_i^* = x_i\beta + \epsilon_i$ and $y_i = I(y_i^* > 0)$, for $\epsilon_i \sim N(0,1)$. The likelihood will be: $L(y_i \mid x_i, \beta, y_i^*) = 1(y_i^* > 0)y_i + 1(y_i^* < 0)(1-y_i)$, $L(y_i^* \mid x_i, \beta) = N(x_i'\beta, 1)$.Prior is $\beta \sim N(b_0, B_0)$, here we choose $b_0 = 0, B_0 = kI_p$. Then, the full conditional posteriors will be:

$$\pi(y_i^* \mid x_i, y_i, \beta) \propto N(x_i'\beta, 1)[1(y_i^* > 0)y_i + 1(y_i^* < 0)(1-y_i)]$$

$$\pi(\beta \mid x_i, y_i^*) \sim N((\sum_{i=1}^{n} x_i x_i' + \frac{1}{k}I)^{-1}(\sum_{i=1}^{n} y_i^* x_i), (\sum_{i=1}^{n} x_i x_i' + \frac{1}{k}I)^{-1})$$

From Figure 7, we could see that the coefficients converge. From the result of Gibbs sampling without interaction terms, the coefficients are shown in Table 2. The misclassification rate is 0.3391.

	Frequentist Logistic	Frequentist LASSO	Bayesian
Without Interaction	0.3395	0.3415	0.3391
With Interaction	0.2589	0.2606	0.2629

Table 1: Misclassification Rates for Our Data (In Sample)

To demonstrate how we can interpret the coefficients, we provide a few examples here. By referring to Table 6 of Exponentiated Coefficients, we can see that wage has little effect on the probability of getting H-1B visa. When wage increases by 1%, the odds of getting H-1B visa is multiplied by 1.41 according to the Frequentist Logistic model. CS related jobs would largely increase the probability of getting H-1B. For a categorical variable, for instance, being in a CS-related company increases the odds of obtaining the visa in the CS industry versus other industries by 2.3.

	Frequentist	LASSO	Bayesian
(Intercept)	-5.6813	-5.4444	-3.1743
wage	0.3430	0.3372	0.1870
days	-0.2502	-0.2418	-0.1392
months_subApril	-0.0422	-0.1455	-0.0359
months_subAugust	0.2548	0.1144	0.1430
months_subDecember	0.1148	0.0000	0.0596
months_subFebruary	0.3401	0.2098	0.1934
months_subJuly	0.0243	-0.0808	0.0035
months_subJune	0.2306	0.0940	0.1319
months_subMarch	0.1884	0.0615	0.0892
months_subMay	0.1775	0.0366	0.0978
months_subNovember	0.1848	0.0414	0.1039
months_subOctober	0.5049	0.3408	0.2726
months_subSeptember	0.3420	0.1948	0.2102
companycs	0.8310	0.8282	0.5096
companyfinance	0.0606	0.0091	0.0367
academicacademic	-0.2218	-0.2040	-0.1217
locationnortheast	0.1526	0.1372	0.0878
locationsouth	0.1495	0.1374	0.0838
locationwest	-0.0786	-0.0805	-0.0503
fulltimeTRUE	0.5016	0.4594	0.2770

Table 2: Coefficients of All Three Methods

4 Discussion

4.1 Testing of Results

To test our results and avoid overfitting, we pull out 10% data at the beginning to do Out-of-Sample test. Using all the three methods with/without interaction, we get the misclassification as follows:

	Frequentist Logistic	Frequentist LASSO	Bayesian
Without Interaction	0.4029	0.3884	0.4422
With Interaction	0.2282	0.2313	0.2237

Table 3: Misclassification Rates for Test Data (Out of Sample)

From Table 3, we could see that although the model with interaction terms seems to be too complex, the Out-of-Sample misclassifications are close to In-Sample misclassifications and are much smaller than Out-of-Sample misclassifications without interaction. So, including interaction in our model will not lead to overfitting.

4.2 Conclusion

All three models prove to be equally effective to construct a regression model for our data, as they all produce similarly small misclassification rates. The results of our models show that the certified rate of the H-1B visas is affected by the factors in our analysis. First of all, time is a key factor. The more days needed to make the decision, the less likely the H-1B visa is certified. Also, submitting application in the months of February, March or May can significantly increase the chance of receiving H-1B visa. Location is another factor that seems to have an influence on the decision. Applying in the northeast area can boost your odds of getting the visa comparing to other areas. There are also job-wise factors we should not ignore. Full-time positions have a huge advantage over other position types. Computer Science and finance related occupations would boost the probability of obtaining H-1B visa, while academic related jobs are disadvantagous when comparing with other jobs. Our results could provide some insights to students looking for jobs in the US. Policy makers could also consider changing the current policy to balance out the job market for foreigners.

References

[1] "Overview of the H-1B, H-1B1 and E-3 Temporary Programs." *United States Department of Labor.* Web. 28 Oct 2015.

[2] "What are the chances of getting selected in the lottery for H1B visas in 2016." *Quora.* Web. 23 Nov 2015.

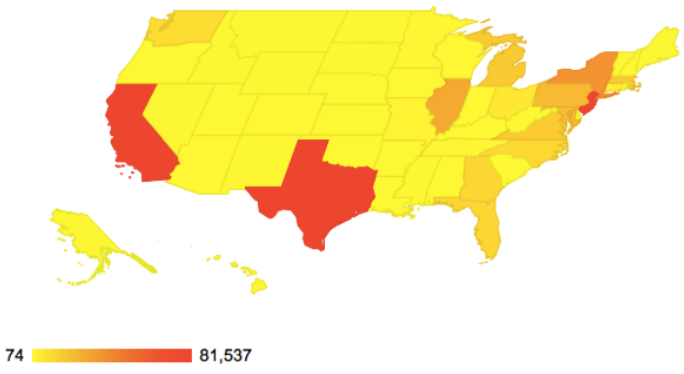
5 Appendix

	Job
computer	222567
analysts	123982
systems	117798
software	105980
developers	94047
applications	75561
programmers	70294
occupations	38525
engineers	35533
managers	16569
except	16564
administrators	15470
teachers	13605
research	13427
specialists	12523
financial	11494
scientists	11279
management	11064
accountants	9085
auditors	8978
network	8756
marketing	8659
postsecondary	7952
operations	7810
mechanical	7455
database	7064
electronics	6561
market	6532
medical	6269
information	6198

Table 4: Word count for Job

	Company
inc	233676
limited	58090
llc	57912
services	36082
corporation	34056
technologies	28099
solutions	25658
infosys	24307
university	22323
llp	22193
systems	17942
consulting	17602
tata	14155
consultancy	14144
group	14074
america	13204
technology	12845
software	10298
global	9780
deloitte	9321
usa	8576
company	8453
international	8346
wipro	8304
infotech	7600
tech	7331
ibm	7053
medical	6421
health	5599
india	5557

Table 5: Word count for Company



State	H1B
AK	116
AL	1189
AR	1557
AZ	3196
CA	81537
CO	3118
CT	5250
DC	1925
DE	1901
FL	12941
GA	13053
GU	210
HI	348
IA	1689

Figure 1: Heatmap

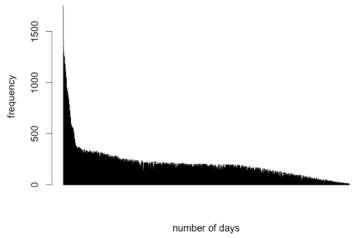


Figure 2: Descriptive: Application Length

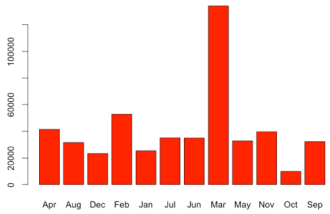


Figure 3: Descriptive: Months

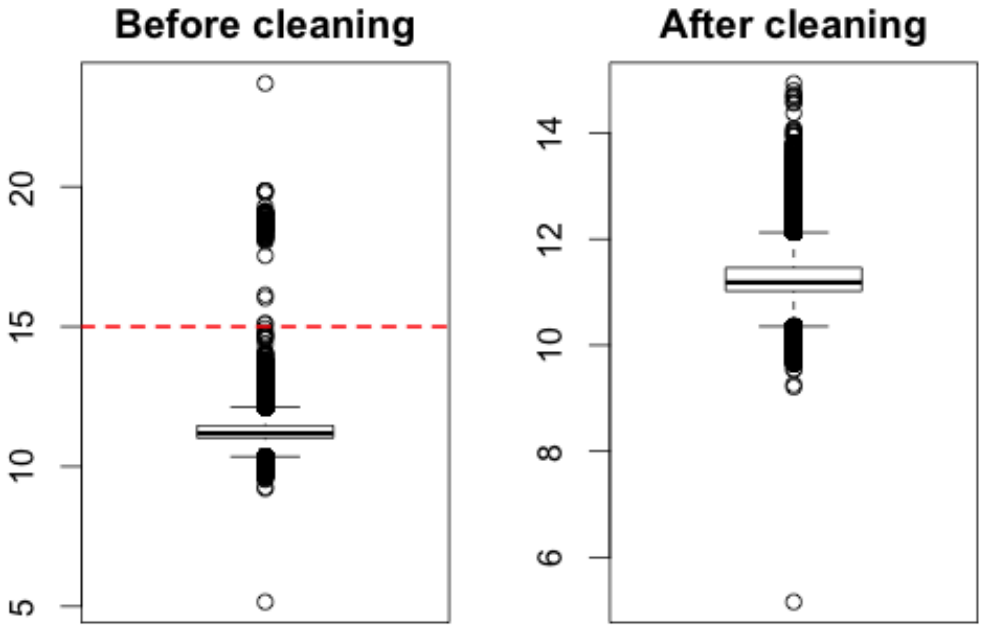


Figure 4: Wage Box Plot

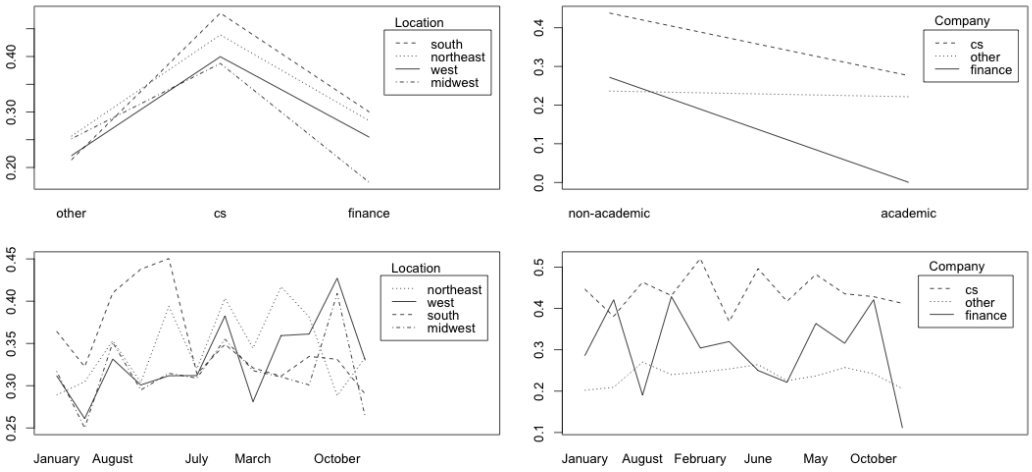


Figure 5: Interaction Plot

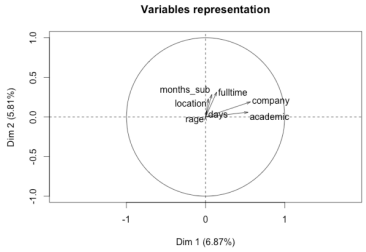


Figure 6: Principle Component Analysis

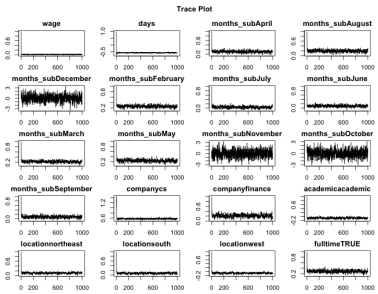


Figure 7: Trace Plot

	Frequentist	LASSO	Bayesian
(Intercept)	0.0034	0.0043	0.0418
wage	1.4091	1.4010	1.2057
days	0.7786	0.7852	0.8701
months_subApril	0.9587	0.8646	0.9648
months_subAugust	1.2902	1.1212	1.1537
months_subDecember	1.1216	1.0000	1.0614
months_subFebruary	1.4051	1.2334	1.2133
months_subJuly	1.0245	0.9223	1.0035
months_subJune	1.2594	1.0986	1.1410
months_subMarch	1.2074	1.0634	1.0933
months_subMay	1.1942	1.0372	1.1028
months_subNovember	1.2030	1.0423	1.1095
months_subOctober	1.6568	1.4061	1.3134
months_subSeptember	1.4077	1.2150	1.2339
companycs	2.2955	2.2891	1.6646
companyfinance	1.0624	1.0091	1.0373
academicacademic	0.8011	0.8154	0.8854
locationnortheast	1.1649	1.1471	1.0918
locationsouth	1.1613	1.1473	1.0874
locationwest	0.9244	0.9227	0.9509
fulltimeTRUE	1.6513	1.5831	1.3192

Table 6: Exponentiated Coefficients of All Methods