Determining Predictors of H-1B Salary and Approval

Milestone Report

Wenhao Yu
University of Notre Dame
South Bend, Indiana
wyu1@nd.edu

Luke Duane
University of Notre Dame
South Bend, Indiana
lduane@nd.edu

Will Badart
University of Notre Dame
South Bend, Indiana
wbadart@nd.edu

ABSTRACT

The paper presents the initial findings of the H-1B visa program analysis project for CSE-40647/60647.

ACM Reference format:
Wenhao Yu, Luke Duane, and Will Badart. 2018. Determining Predictors of H-1B Salary and Approval. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 4 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The H-1B visa program, enacted by the Immigration and Nationality Act of 1965, opens the door for immigrants in specialized professions to migrate to the United States for an extendable term of six years. Last year, in 2017, almost 350,000 foreign workers applied for the program, and just under 200,000 were approved.

To decide which of the many applicants are awarded one of the limited number of approvals, The US Citizenship and Immigration Services (USCIS) conducts an annual lottery. The H-1B lottery is a laborious and complex process for both large companies bringing in thousands of migrant employees and small ones onboarding only a couple. A tool which highlights the important features that support H-1B approvals could be a vital strategic asset for these companies. Lots of data exists in this domain, but to integrate it and perform meaningful analysis is beyond the capabilities of companies without established data science practices. We plan to produce a model that shows what features are most valuable in regards to H-1B workersâĂŹ salaries and approval.

2 RELATED WORK

In April of 2017, Glassdoor published an article analyzing the salaries of H-1B immigrants and comparing them to those of domestic workers in similar roles and fields. While the report does not attempt to model H-1B workersâĂŹ salaries based on other features, it offers a comprehensive statistical analysis of their pay. [1]

[1]Glassdoor Comparison on H-1B Visa Salaries vs US Workers

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 PROBLEM DEFINITION

How can we predict the approval status of a given H-1B via application? What tangential analyses provide tangible business value for companies sponsoring H-1B visas? How would the salary range change based on a given occupation?

4 PROPOSED METHODOLOGY

The sheer volume of data available to train our model necessitated that we perform a number of initial analyses before constructing the model. For these initial analyses, we chose to calculate a number of descriptive statistics over our primary data set[2] as well as a couple visualizations to quickly understand the distributions of key features. We have already identified a few outliers in the primary dataset (in particular, in the PREVAILING_WAGE feature) and cleaned our data before producing our initial findings.

After the data cleaning and description phase, we began to train our predictive models. Our baseline model are Naive Bayes model and Decision Tree model, which attempt to predict the status of an H-1B application and the salary range.

We used 5-fold Cross Validation to make sure every data has been used as training and testing. So we used the overall accuracy estimate as the average of the accuracies obtained from each iteration.

Additionally, we will use our findings from the decision tree construction to create a random forest to predict approval status, which we expect to have the best performance. To supplement these findings, we will train a regression to model.

If time permits, we're curious to implement a neural network as a classifier, and pit it against our best performing model of those described above.

While executing this project, we identified another interesting data science task: what meaningful groupings of data points can we discover or create to reduce the computation load of processing three million or more individual data points on H-1B applications? This question maps naturally to the task of clustering. We determined heuristically that JOB_TITLE would be the most logically sound attribute on which to perform the clustering. See the following section for the experiments performed in service of this task.

5 DATA AND EXPERIMENTS

5.1 Datasets

(1) One of the largest freely available datasets on H-1B applications comes from kaggle.com. It contains over 3 million

[2]See 5.1 Datasets

records and tracks 10 different features per application[3]. This data covers applications roughly between 2012 and 2016.
(2) Another key dataset comes from the Foreign Labor Certification Data Center. Its data is organized by year, spanning from 2001 to 2007[4].
(3) OFLC's annual reports also provides a lot of program information and data. Although it is not raw data, it disclosures cumulative quarterly and annual releases of program to assist with external research and program evaluation[5].

5.2 Data Summary

This subsection presents a preliminary description of dataset (1), the Kaggle dataset described in section 5.1 Datasets.

Figure 1 shows the salary distribution.There are 5 Outliers in the original data. The average of salary is 72,221, the median of salary is 66,602, and the standard deviation is 24,704.

Table 1 shows the frequency of each value of the CASE_STATUS feature, the column which labels whether an application was approved. From the dataset's documentation:

> The CASE_STATUS field denotes the status of the application after LCA processing. Certified applications are filed with USCIS for H-1B approval. CASE_STATUS: CERTIFIED does not mean the applicant got his/her H-1B visa approved, it just means that he/she is eligible to file an H-1B.

As demonstrated in the following table, our dataset is characterized by pretty heavy class imbalance. This lead to special considerations in some of our experiments, such as performing stratified sampling in the partitioning of testing and trainging subsets.

Table 1: Approval Status Classes

Class Name	Frequency
CERTIFIED	914,251
NON-CERTIFIED	134,325

Table 1 shows that most H-1B applications are certified (note: this does not mean they are accepted). We also chose to examine the change in volume in H-1B applications over time.

Table 2: Salary Classes

Class Name	Frequency	Range
Very High	90,004	[104042,E99)
High	182,226	[79331,104042)
Middle	361,845	[59155,79331)
Low	181,648	[28963,59155)
Very Low	98,528	[12584,28963)

Table 2 shows the frequency of each value of the WAGE feature. We have five categories and the columns show the frequency and salary range of each class.

[3]See kaggle.com/asavla/h1-visa/data
[4]See flcdatacenter.com/CaseH1B.aspx
[5]Please follow https://www.foreignlaborcert.doleta.gov/pdf/OFLC_Annual_Report_FY2016.pdf

Figure 1: Salary Distribution

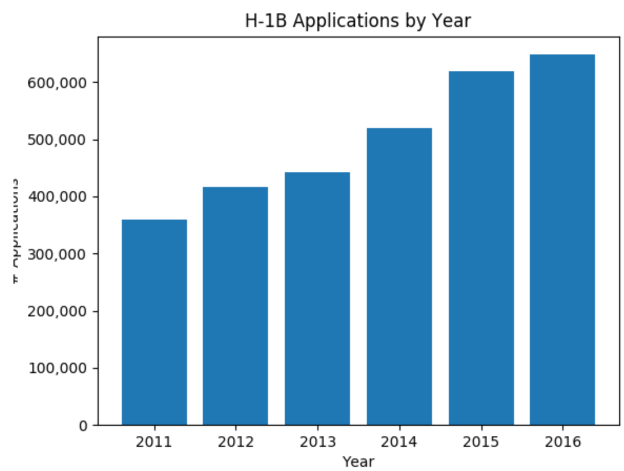

Figure 2: H-1B Application Volume by Year

5.3 Experimental Settings

About the approval status classification, we use four features: EMPLOYER, JOB TITLE, LOCATION, SALARY with the label CASE_STATUS, which has two categories: CERTIFIED and NON-CERTIFIED.

About the salary level classification, we use three features: EMPLOYER, JOB TITLE, LOCATION and the label is SALARY, which has five categories: VERY HIGH, HIGH, MIDDLE, LOW,VERY LOW.

The clustering task was a bit of a different beast. There were 287,551 different job titles listed in the primary dataset, so mapping them into the real space to compute and visualize their groupings was a challenge. The method we decided on first vectorized each job title in a modified one-hot encoding, turning the set of job titles into a sparse, high-dimensional matrix.

In order to visually evaluate the quality of each clustering, we reduced the dimensionality of our data matrix down to 2 through

SVD. We also found that performing the *SVD* transformation significantly improved the runtime performace of the *SpectralClustering* method.

This encoding is compatible with the clustering methods presented by `scikit-learn`. Figure 3 are the results of clustering with K ranging from 2 to 8.

Each cluster can be characterized by the job title terms that appear most frequently within it. In general, we found at least one cluster dominated by C-suite officers, another by directors, another by managers, and sometimes, one by engineering. This provides meaningful groupings of the records according to job position information.

5.4 Evaluation Results

Table 3: Naive Bayes Confusion Matrix for Approval

	Predicted Approved	Predicted Denied
Approved	888435	25814
Denied	98423	35901

Accuracy: 0.91409770402 Specificity:0.26727167

Table 4: Decision Tree Confusion Matrix for Approval

	Predicted Approved	Predicted Denied
Approved	905932	8317
Denied	81756	52568

Accuracy: 0.881516343609 Specificity:0.3913522

Table 5: Naive Bayes Salary Prediction Accuracy

Class	Correct	Wrong
Very High	58282	31722
High	93270	88956
Middle	276983	84862
Low	96456	85198
Very Low	71492	27063
Total Accuracy	65.2	

Table 6: Decision Tree Salary Prediction Accuracy

Class	Correct	Wrong
Very High	67667	22337
High	137068	45158
Middle	317643	44202
Low	151558	30090
Very Low	95306	3222
Total Accuracy	84.1	

6 CONCLUSIONS

REFERENCE

[1] The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011-3

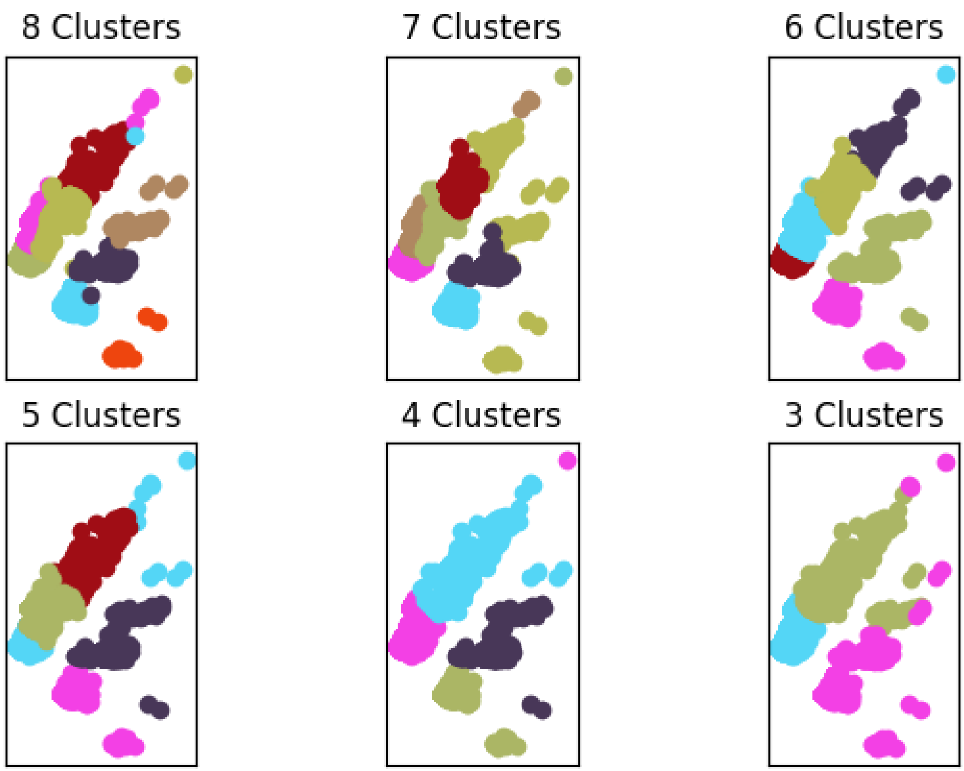

Figure 3: Clustering on JOB_TITLE for first 20,000 records (3,433 unique titles)