STAT 571 Final Project: Exploration of H1B Candidates for Compensation, Career, and Approval

Yifan Jiang, Yang Yi, Kexin Zhu

Contents

Executive Summary

Today, with rising demand for high-tech professionals, more and more companies are willing to sponsor H1B working visa for foreign employees to obtain legal status of residence. Given the fact that the United States has boasted a large number of top high-tech and internet companies throughout the world, international students are dreaming of working for these brilliant organizations and to the largest extent train their expertise. As such, it is reasonable for prospective individuals to care about the expected salary they would receive and whether they can be accepted by the H1B program. Moreover, the application process for the H1B program is troublesome and full of uncertainty. Thus, a clearer understanding of related information and what to expect may largely help international students to make future plans in advance. As potential candidates who will be experiencing this process, we decided to look deeply into the H1B disclosure data from the United States Department of Labor (See more on: https://www.foreignlaborcert.doleta.gov/performancedata.cfm), which provides information of each H1B applicants in terms of levels of income, job titles, application status, worksite state, etc., to find some insights about the factors of receiving high salaries, including the corresponding job titles and employers. By analyzing these data, we figure out that most international students and foreign professionals have a very decent annual compensation, with the average of 90 thousands dollars, which is much higher than the average compensation level across the United States. We also find out that workers who have been promoted to principals, senior levels, and managers should have a relatively high salary, while jobs such as accountant, assistant and associate are more likely to be correlated with lower levels of compensation. We predict that the most important factor for the approval of H1B, based on our analysis, is the annual salary.

H1B Data Anaylsis

Introduction

As an immigrant country, the United States has long been welcoming talents from all over the world to obtain self-development or realize their dreams, and accordingly, the H1B program has been carried out to retain skilled labor in high-tech industries. The H1B program allows companies in the United States to temporarily employ foreign workers in occupations that require the theoretical and practical application of a body of highly specialized knowledge and a bachelor's degree or higher in the specific specialty. Simply put, H1B Visa is a temporary working visa officially given to aliens with specialized capabilities to work and live in the United States. H1B is extremely important for foreign employees because this is equivalent to their immigrant status within the States. Their working permission, in other words, is not fully controlled by the employers solely but also partially by the government.

The dataset we will be using is this public disclosure file from the Department's Office of Foreign Labor Certification, Employment and Training Administration, by which this file contains administrative data from employer petitions for prevailing wage determinations (ETA Form 9141) processed (The dataset could be fully retrieved from: https://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2019/H-1B_Disclosure_Data_FY2019.xlsx). This file records the information with date of either the initial prevailing wage determination or redetermination was issued on or after October 1, 2018, and on or before December 31, 2018. All data were extracted from the Office of Foreign Labor Certification's iCERT Visa Portal System, an electronic filing and application processing system of employer requests for prevailing wage determinations (Information retrieved from the record layout).

In this project, Our goal is to analyze the important factors and identify the keywords of job title which employers are able to get sponsored, receive higher salaries, and get H1B Visa approved. With respect to the methodology, we use regressions featuring LASSO selection, classification methods including random forest and neural network, and text mining tools in our study. The overall result of our experiment shows that we can to a large extent predict the salary, precisely capture popular job titles to get sponsored, and whether the of H1B Visa will be approved.

Since all these data are from previous cases, we could thus build models to predict future conditions using previous bases. Our analysis is mainly composed of three parts: first, we do some basic cleaning for the dataset. Specifically, we exclude some irrelevant or unnecessary features from the dataset and regularize some data from the original data, and choose the possible risk factors to predict. Second, we try building predicting models with respect to compensations for foreign employees; finally, we explore the keywords for occupations and employers that most welcome aliens.

Data Cleaning and Explanatory Analysis

Our preliminary data (`H-1B_Disclosure_Data_FY2019.csv`) is composed of 97572 observations with 52 variables, based on which we conducted a cleaning process and engineered some variables to ease the coding burden.

Specifically, we excluded some unnecessary columns which are not necessary to fill to ensure consistency. Also, we excluded all the information related to the attorney as we would not look into the applications per se through a legal perspective. Besides, we removed the specific dates of cases and addresses for employers but only keeping roughly geographical information to prepare for further analysis across states.

After that, in order to better obtain the insights from a complete dataset, we decided to exclude all the NA records with the following manipulations.

To simplify, we only considered the cases of which the salaries are recorded as annual. Therefore, we finally obtained a dataset with 56443 observations and 18 variables. The following is a quick summary for variables we would be dealing with from the filtered dataset.

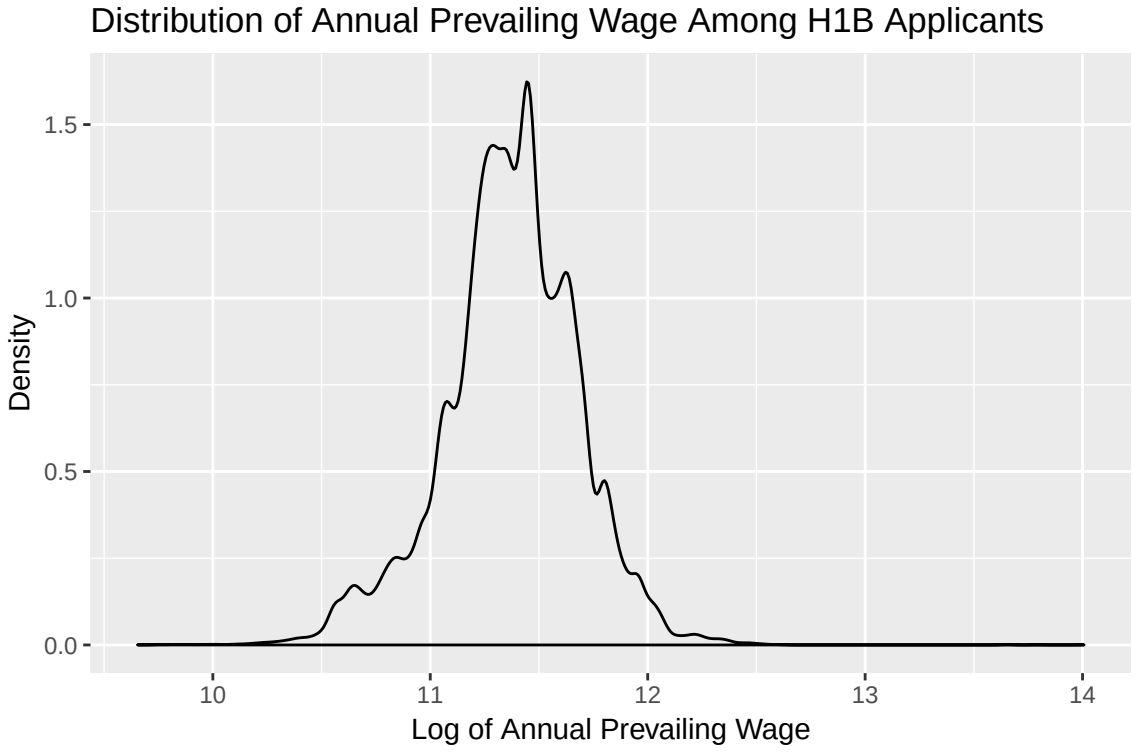
Table 1 Detailed Description of Variables in the Cleaned Dataset

Variable	Description
CASE_STATUS	Status associated with the last significant event or decision.
VISA_CLASS	Type of visa classification supported by the employer request for a prevailing wage determination.
AGENT_REPRESENTING_EMPLOYER	The employer by which the applicant is employed.
JOB_TITLE	Title of the employer's job opportunity.
SOC_NAME	Occupational name associated with an occupational code.
TOTAL_WORKERS	The total number of workers is responsible for the H1B Visa.
NEW_EMPLOYMENT	Label indicates whether the applicant is newly employed.
CONTINUED_EMPLOYMENT	Label indicates whether the applicant is continued employed.
CHANGE_PREVIOUS_EMPLOYMENT	Label indicates whether the applicant changed job before.
NEW_CONCURRENT_EMP	Label indicates whether the applicant has several jobs at the same time.
CHANGE_EMPLOYER	Label indicates whether the applicant changed emplyer before.
AMENDED_PETITION	Label indicates whether the applicant amended petition for H1B application.
FULL_TIME_POSITION	Label indicates whether the applicant have a full time job or not.
PREVAILING_WAGE	Number for the applicants's wage per year.
PW_WAGE_LEVEL	Label indicates the applicant current position level.
H1B_DEPENDENT	Label indicates whether applicant has dependent spouse or children.
WILLFUL_VIOLATOR	The agency finds that the employer has committed either a willful failure or a misrepresentation of a material fact .

Variable	Description
WORKSITE_STATE	The location (states level) where the applicant is hired.

Before heading to our main part of analysis, we first plotted the salary distribution to get a picture of how annual salary distributes among foreign employees. Hence, we had a preliminary impression that: most applicants are receiving a wage within the range of 70,000 and 105,000, and the mean annual wage of the sample employees is around 90,000 dollars. The distribution of salaries is close to the normal distribution which is slightly right-skewed.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   73226   87506   90814  105893 1211118
```



Mapping the data to a heatmap

A heatmap is useful to provide easily-read visualized data with geographical information. In this section, we generated a heat map to display summary statistics at the states level. Specifically, we mapped the mean salaries with corresponding states. Accordingly, we created a variable named `med.income`. Additionally, we extracted the mean of `med.income`, mean income rate by state among other statistics. Here, `n` equals the number of observations within each state (See Appendix 0).

Here, we could observe that, in general, applicants who work in states including California, Washington, Alaska, and New York may have higher levels of income compared with that employed in other states, which is somewhat against our intuition as we may expect that the states along the east coast of the United States may obtain higher income. Take one step further, the pattern showing higher income levels along the west coast may reversely demonstrate that professionals working for high-tech companies are expected to be paid more.

Compensation Prediction

In this section, we would be looking for a model that could predict the compensation of employees who are eligible for the H1B program. To prepare for the corresponding prediction, we excluded the job title and created a subset named `h1b.data.1` to solely look into the levels of income for which we would be using text mining tool to explore the corresponding job titles.

Method 1: Linear regression

We started from the a simple linear regression model (Summary See Appendix 1.1).

To further examine whether all categorical variables (`factor variables in r`) are nexessary in the regression model, we ran the Anova Type II tests. The results are shown as follows:

```
## Anova Table (Type II tests)
##
## Response: PREVAILING_WAGE
##                                 Sum Sq    Df    F value    Pr(>F)
## CASE_STATUS                  5.2175e+10     3    41.5433 < 2.2e-16 ***
## VISA_CLASS                   2.7959e+09     3     2.2262 0.0828961 .
## AGENT_REPRESENTING_EMPLOYER 3.5864e+11     1   856.6786 < 2.2e-16 ***
## TOTAL_WORKERS                1.8075e+08     1     0.4318 0.5111307
## NEW_EMPLOYMENT               1.4352e+08     1     0.3428 0.5582026
## CONTINUED_EMPLOYMENT         8.1353e+08     1     1.9433 0.1633171
## CHANGE_PREVIOUS_EMPLOYMENT  8.7111e+08     1     2.0808 0.1491650
## NEW_CONCURRENT_EMP           3.4667e+06     1     0.0083 0.9274933
## CHANGE_EMPLOYER              8.4732e+07     1     0.2024 0.6527933
## AMENDED_PETITION             2.8766e+08     1     0.6871 0.4071419
## FULL_TIME_POSITION           1.2406e+07     1     0.0296 0.8633243
## PW_WAGE_LEVEL                1.3575e+13     3 10808.5043 < 2.2e-16 ***
## H1B_DEPENDENT                5.7561e+09     1    13.7495 0.0002091 ***
## WILLFUL_VIOLATOR             2.0299e+06     1     0.0048 0.9444850
## WORKSITE_STATE               6.0644e+12    55   263.3808 < 2.2e-16 ***
## Residuals                    2.3597e+13 56367
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the output above, we concluded that among all categotical variables, only `CASE_STATUS`, `AGENT_REPRESENTING_EMPLOYER`, `PW_WAGE_LEVEL`, `H1B_DEPENDENT`, and `WORKSITE_STATE` are significant at 0.001 level. Therefore, we selected these variables and built another linear regresson (Summary See Appendix 1.2). A comparison was conducted between the regression model with full variables (`fit.wage.1.0`) and the five-variable model (`fit.wage.1.1`) using Chi-square test.

```
## Analysis of Variance Table
##
## Model 1: PREVAILING_WAGE ~ CASE_STATUS + VISA_CLASS + AGENT_REPRESENTING_EMPLOYER +
##     TOTAL_WORKERS + NEW_EMPLOYMENT + CONTINUED_EMPLOYMENT + CHANGE_PREVIOUS_EMPLOYMENT +
##     NEW_CONCURRENT_EMP + CHANGE_EMPLOYER + AMENDED_PETITION +
##     FULL_TIME_POSITION + PW_WAGE_LEVEL + H1B_DEPENDENT + WILLFUL_VIOLATOR +
##     WORKSITE_STATE
## Model 2: PREVAILING_WAGE ~ CASE_STATUS + AGENT_REPRESENTING_EMPLOYER +
##     PW_WAGE_LEVEL + H1B_DEPENDENT + WORKSITE_STATE
##   Res.Df        RSS  Df   Sum of Sq     F   Pr(>F)
## 1  56367 2.3597e+13
## 2  56379 2.3635e+13 -12 -3.7381e+10 7.441 6.96e-14 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

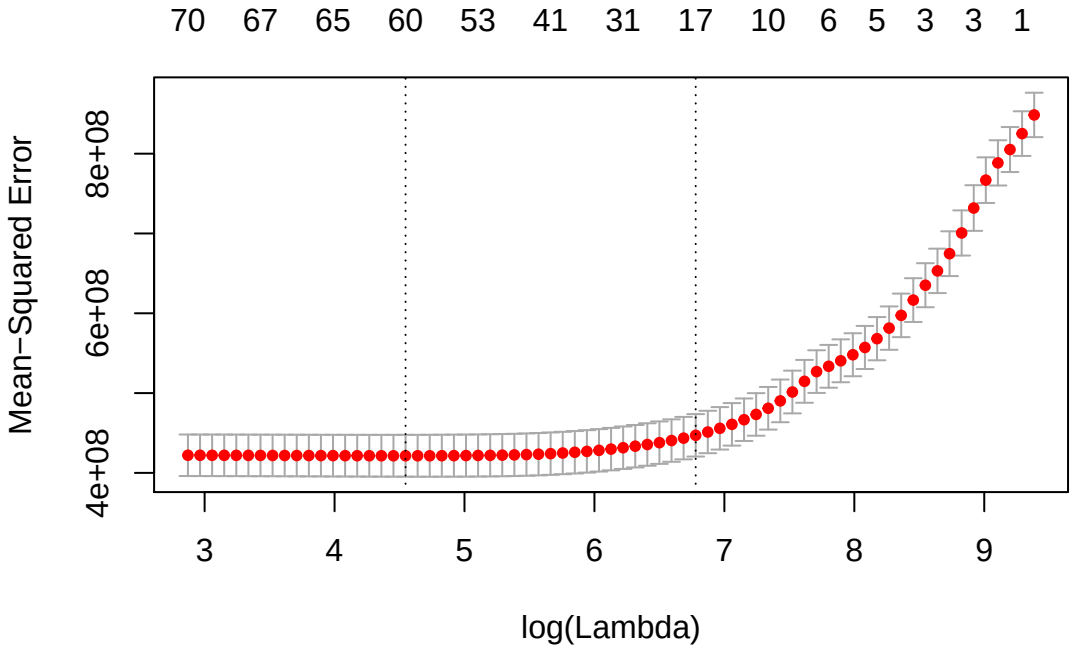
Clearly, we observed that the second model with five selected variables (`fit.wage.1.1`) is far better than the full model as the former one manifests a larger F statistics and smaller p-value (nearly 0) which is also significant at 0.001 level. We thus did figure out that our model with fewer variables performed better.

Method 2: LASSO

Another way we chose to build the prediction model is to first use LASSO regression analysis to select the best subset of variables, based on which we fitted the linear model. The merit of LASSO is that it performs both variable selection and regularization to enhance the accuracy of prediction and interpretability of the model it produces.

We looked for the lambda with the smallest mean square error (`MSE`) while best keep the simplicity by running the cross validation. The visualized graph drew a picture of the number of variables used for a specific lambda, as well as the corresponding `MSE`. From the following graph, we were able to observe the two recommended points displayed as two vertical lines. The left line indicates the minimum lambda (lambda.min) outputting the best model with largest AUC, while the right line represents the lambda (lambda.1se) that outputs fewer variables but remains error within one standard error of the best model.

Figure 2 log(Lambda) with Mean Square Error for Compensatin LASSO



We realized that the difference of `MSE` between lambda.min and lambda.1se is tiny, but that the numbers of variables change a lot. To avoid underfitting we decided to use `lambda.min` to include a few more variables to be our candidates. Then, we built a linear regression based on the variables chosen and proceed Type II tests to see if the selected variables were significant. The results of Type II tests are shown as follows:

```
Anova(fit.wage.2.0)
```

```
## Anova Table (Type II tests)
##
## Response: PREVAILING_WAGE
##                               Sum Sq    Df    F value    Pr(>F)
## CASE_STATUS                4.8006e+10     3    38.1932 < 2.2e-16 ***
```

```
## AGENT_REPRESENTING_EMPLOYER 4.3085e+11     1  1028.3535 < 2.2e-16 ***
## NEW_EMPLOYMENT               3.3115e+09     1     7.9039 0.0049346 **
## CHANGE_PREVIOUS_EMPLOYMENT  5.2875e+09     1    12.6200 0.0003819 ***
## CHANGE_EMPLOYER              1.7471e+10     1    41.6989 1.073e-10 ***
## AMENDED_PETITION             2.3049e+09     1     5.5014 0.0190050 *
## PW_WAGE_LEVEL                1.3595e+13     3 10816.4456 < 2.2e-16 ***
## WORKSITE_STATE               6.2673e+12    55   271.9767 < 2.2e-16 ***
## Residuals                    2.3620e+13 56376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above output, all variables prove to be significant at 0.01 level. Therefore, we obtained a LASSO regression model with eight variables: `CASE_STATUS`, `AGENT_REPRESENTING_EMPLOYER`, `NEW_EMPLOYMENT`, `CHANGE_PREVIOUS_EMPLOYMENT`, `CHANGE_EMPLOYER`, `AMENDED_PETITION`, `PW_WAGE_LEVEL`, and `WORKSITE_STATE`.

To test if the models were performing well, we randomly picked a sample from the dataset and see the prediction results.

Table 2 Sample Prediction Results with Two Models

Actural Value	Prediction with Linear Regression	Prediction with LASSO Regression
70200	85283.91	85098.21

Overall, we observed a quite satisfying performance of the model, within the fault tolerance range of 20% (approximately), rather than a high accuracy prediction though.

Job Title Text Mining

Here, we will be doing text mining especially for job titles to find out high-pay jobs for international students and foreign professionals, so that we could hopefully help H1B candidates to consider their future in advance, especially their choice of whether to experience the tiring H1B application and indispensable negotiation with employers as well.

For technical convenience, we will be using the word term frequency table to extract texts into words frequencies. To achieve this goal, first we formed a bag of words where all the words appeared in the documents say N; then for each document (row) we recorded the frequency (count) of each word in the bag which give us N values (notice: most of the entries are 0); and finally we output the document term matrix (`dtm`) as an input for future transformations.

To align with our research goal, we only looked into the column describing the job titles of H1B candidates in our dataset. Accordingly, we have 59990 variables in the original dataset, and the type is defined as character. The following are the first five elements.

```
## [1] "PRINCIPAL SOFTWARE ENGINEER" "SOFTWARE ENGINEER"
## [3] "ASSISTANT PROFESSOR"         "ARCHITECTURAL GRADUATE"
## [5] "SOFTWARE ENGINEER"
```

Data preparation

Here we display detailed data preparation for our text mining precedure.

First, to obtain best results, we considered all the word as candidate keywords. We converted all the keywords into low case and removed punctuation, numbers and stopwords to match the word package in the R software.

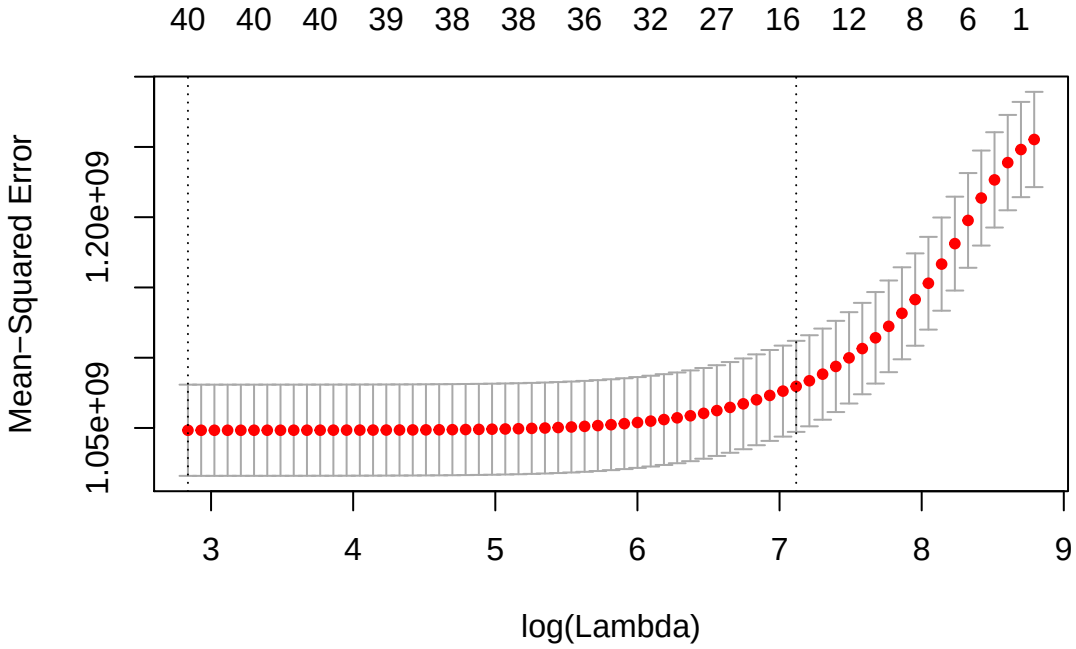
Then we set a sparsity threshold to remove keywords occur below threshold, and we also combined all the origin feature with text mining to output a new dataset for future analysis.

Model Building

We used LASSO regression for text mining analysis. The procedure is similar to what we have done previously for salary prediction, and our results indicated a bag of selected words that are the most important to the dependent variable.

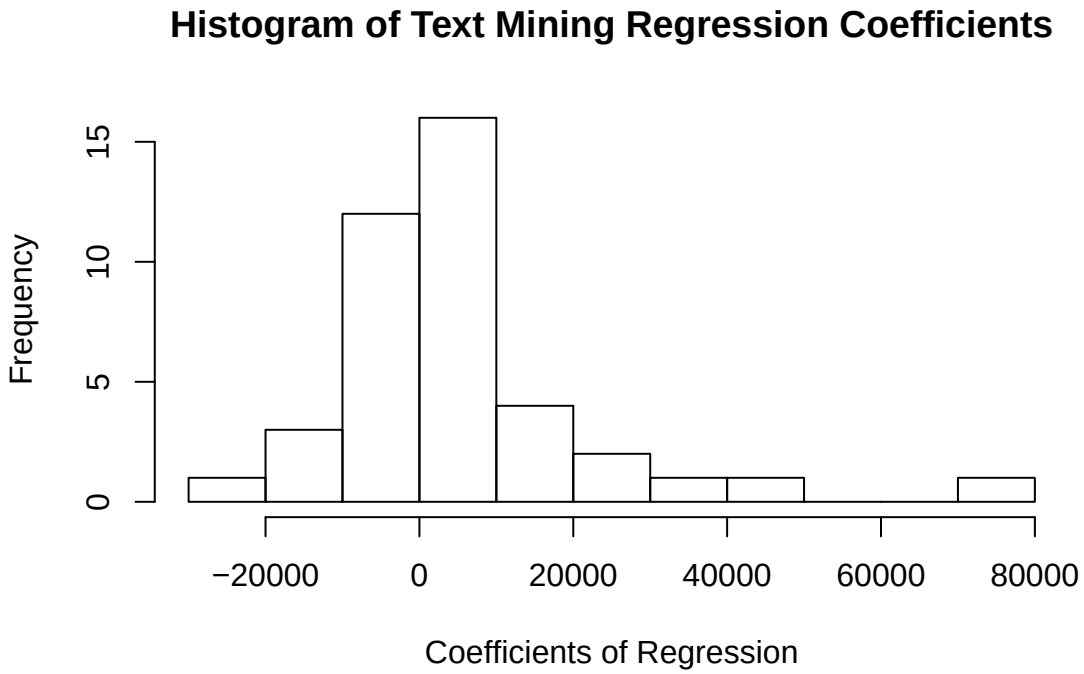
The folowing is the LASSO plot with mean square error for text mining.

Figure 3 log(Lambda) with Mean Square Error for Text Mining LASSO



We will be using `lasso.min` to choose variables because `lasso.min` output the smallest mean square error. Accordingly, we built a linear regression based on the variables chosen by LASSO. Thus, we obtained the regression model `result.lm` as the result of text mining procedure.

After that, we pulled out all the positive coefficients and the corresponding words. We ranked the coefficients in decreasing order and reported the leading 2 words and the coefficients, where we observed a right-skewed frequency distribution of coefficients.

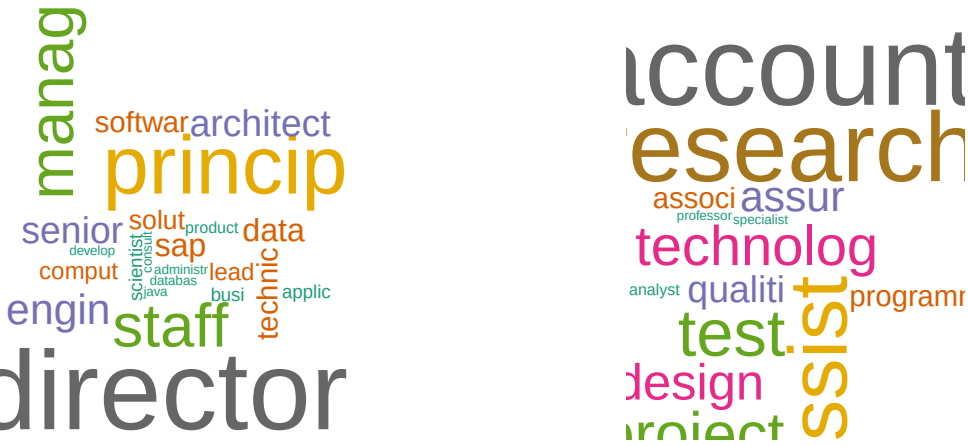



```
## (Intercept)    director
##    74294.87    43347.43
```

Finally, we output word clouds to visualize the correlation between positive words and negative words. Specifically, we ordered the `result.lm` positive coefficients (aka positive words) and output a word cloud, where the size of words displayed indicates the strength of positive correlation between that word and the chance of being a good rating. The logic is similar for negative words.

The following images are the word clouds respectively for positive words and negative words, namely, the high-pay job titles and the low-pay job titles.

Figure 4 Word Cloud for High-Pay and Low-Pay Job Titles



Based on the word clouds shown above, we can infer that:

1. Higher level of positions leads to higher compensation, which is very intuitive. For example, we can observe that the top largest words of the high-pay word clouds are `director`, `principal`, and `manager`, which are basically primary managing or leader positions. Comparatively, we can see entry-level positions like `specialist`, `analyst`, and `assistant` included in the low-pay word cloud.

2. Specialty does matter the compensation for foreign employees. We found that high-pay jobs, to a large

extent, are related to data science and computer science, including `software engineer`, `computer scientist` and `technician`, where professionals in `JAVA` and `database` seems to be in high demand relatively. Besides, `architect` and `consultant` are also positions with a competitive salary. It is noteworthy that even in the low-pay word cloud, we still observed keywords like `system`, `technology`, and `programmer`. This result indicates that technological talents are in great demand where foreigners are more likely to find a job regardless of the level of income.

3. Researchers and postdoctoral fellows account for another large proportion of H1B candidates. Most of the keywords occur in the low-pay group, where we found `professor`, `research`, and `postdoctor` based on the word cloud. This condition suggests that The United States is still proactively absorbing scholars and providing an academic environment featuring openness and freedom, even though the pay of these positions is not necessarily competitive.

To sum up, foreigners specialized in technology such as computer science and data science are welcomed and could obtain competitive income if being promoted to higher levels of positions. Scholars and researchers are also supported compared with other titles. Business talents may not be a large group in terms of H1B candidates (we saw very few words describing business related positions, such as `business` and `accountant`).

H1B Approval Prediction

As we know that H1B petitions need to be approved by the government to finally confirm eligibility of candidates, we decided to look into the approval data and expected to find some insights through analysis. We did a few transformations for our dataset to match the requirements for regression and decision tree analysis in R software.

The response in the original dataset is the variable called `CASE_STATUS` which includes four levels of case status `CERTIFIED`, `CERTIFIED-WITHDRAWN`, `DENIED`, and `WITHDRAWN` representing whether the H1B was approved or not. For those of cases which were withdrawn, we didn't know the whether the employee had been terminated, or the the employee had left the company. So we got rid of these cases and thus our final dataset consists of 54771 cases. We assigned `DENIED` as 0 and `CERTIFIED` or `CERTIFIED-WITHDRAWN` as 1 in our case.

We split the dataset into training data and testing data for predictive model building as well as performance test. A dominant advantage of splitting data is that we could use the training model to generate predictions for testing data without sample overlapping because it makes little sense if we validated our model with the same sample used to build the model.

We first drew a boxplot to see case status are related with annual salary, and we found applications with extremely high or low salary are more likely to be denied. The boxplot is shown as Appendix 3.0.

Direct Logistic Regression

In this section, we first used the training data to fit the logistic regression model by including all the predictors. From the summary, we found that there are many categorical variables with some kind of levels. `Anova()` is a good way to drop categorical predictors which have small p-values. Following the results of `Anova()`, we used the backward selection method to keep only variables whose coefficients are significantly different from 0 at .05 level. We kicked out the variable with the largest p-value first, and then re-fitted the model to see if there were other variables that should be excluded. The final model we got contains 11 variables which were `VISA_CLASS`, `AGENT_REPRESENTING_EMPLOYER`, `TOTAL_WORKERS`, `NEW_EMPLOYMENT`, `CONTINUED_EMPLOYMENT`, `CHANGE_PREVIOUS_EMPLOYMENT`, `CHANGE_EMPLOYER`, `AMENDED_PETITION`, `FULL_TIME_POSITION`, `PREVAILING_WAGE` and `H1B_DEPENDENT`.

From Appendix 3.5, we could figure out the model does not satisfy the linearity.

Elastic Net Logistic Regression

Apart from the direct logistic regression, we also tried to select a model with a few important variables through elastic net. It is better to filter a subset of variables that are most important to the dependent variable, otherwise, the workload will be too much with very little efficiency. We used LASSO regression to realize this step, by setting $\alpha = 1$ and apply 5 fold Cross Validation (CV) to minimize the deviance.

The plot of log(Lambda) vs. Mean-Squared Error is shown in Appendix 3.1). The first vertical line is the `lambda.min`, or the λ which gives the smallest cvm, while the second vertical line is `lambda.1se`, or largest λ whose cvm is within the cvsd bar for the `lambda.min` value. In this section, we chose `lambda.min` to select a set of variables to build our model because we wanted the lowest mean square error.

Based on the variables chosen by `lambda.min`, we refitted logistic regression model accordingly. To make a comparison, we also took a look at the important variables chosen by `lambda.1se`. However,there are no variables selected by `lambda.1se`.

TO Write

Single Decision Tree

Next, we fitted a single decision tree, which is a popular model in data mining. The idea of decision tree is to partition the space into J boxes $R_1, R_2, \ldots, R_J$. The target is to minimize the RSS

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

where $\hat{y}_{R_j}$ is the mean of the training samples in the region R_j. Since decision tree is not a linear model, it can be more flexible and can take interactions among variables. Using a single tree, we can show the interpretation by visualizing the split of the tree.

We tuned the deep of the tree by changing the parameter `mindev`. If the value is too small, there might be only one split. However, we would get an overfitting tree if we set too small value to the `mindev`. After trying for multiple times, we chose 0.0006 for the `mindev`, which is neither too small nor deep. The plot of the single decision tree can be found in Appendix 3.2.

The final tree has 168 terminal nodes (leaves) and contains 11 variables which are `H1B_DEPENDENT`,`VISA_CLASS`, `AGENT_REPRESENTING_EMPLOYER`, `PREVAILING_WAGE`, `CONTINUED_EMPLOYMENT`, `PW_WAGE_LEVEL`, `AMENDED_PETITION`, `TOTAL_WORKERS`, `CHANGE_EMPLOYER`, `CHANGE_PREVIOUS_EMPLOYMENT`, and `NEW_EMPLOYMENT`.

Ramdon Forest

One problem of single decision tree described above is its high variance, which, fortunately, could be reduced by bagging. However, bagging produces highly correlated trees which turns out will not help reduce variance substantially. Thus, we introduced random forest in this section to tackle this problem by forcing to split only a subset of predictors.

The algorithm of the random forest is as follows:

For b=1 to B, we took a bootstrap sample of size n. Then we built a tree using the bootstrap sample recursively until the n_min is met. Here in classification tree, we randomly selected m variables (mtry = $\sqrt{p}$). For each variable, we found the best split point such that the misclassification errors could be minimized by **majority vote**. When we found the best variable and split point, splitted the node into two, and the end node would output the majority vote either 0 or 1.

To implement the random forest algorithm, we used the randomForest package in r, and fit the model using training data. Tuning parameter `m` is thus introduced. If m is too small, we might miss important variables;

if `m` is too large, there would be more correlations between trees, which is not goood either. We used default setting of `mtry`, which is $\sqrt{p}$. The choices for Bootstrap size B (ntree) affect the performance, and usually, the error decreases as the size of ntree increases. Therefore, we used a large number `ntree = 100`. All other parameters we used in this study are default settings as well.

Evaluation

Model Evaluation

Misclassification Error

After fitting all the models using training data, we will be evaluating the performace of each model by predict on the testing data, and calculating the misclassification error.

For direct logistic regression and LASSO with logistic regression, we predicted the probability on the testing data, and chose the threshold of 1/2. All probabilities that are greater than 1/2 are classified as 1, while those less than 1/2 are classified as 0. For random forest, the prediction is estimated by the majority vote of the aggregated trees, while the probability is estimated by the sample proportion of 1's among aggregated trees.

To illustrate the classification results, we show the confusion matrics of all models (See Appendix 3.3). Confusion matrix is a table to describe the performance of a classification model on a set of test data for which the true values are known.

We then obtained the misclassification error of each model by calculating the mean values of missclassifications

$$MCE = \frac{1}{n}\sum_{i=1}^{n}\{\hat{y}_i \neq y_i\}$$

The misclassification errors of all the model tried in the section 3 is shown below. Logistic regression models didn't display a good predicting perforemance because our dataset is unbalanced. However, Decision trees seem to work quite well in predicting H1B status.

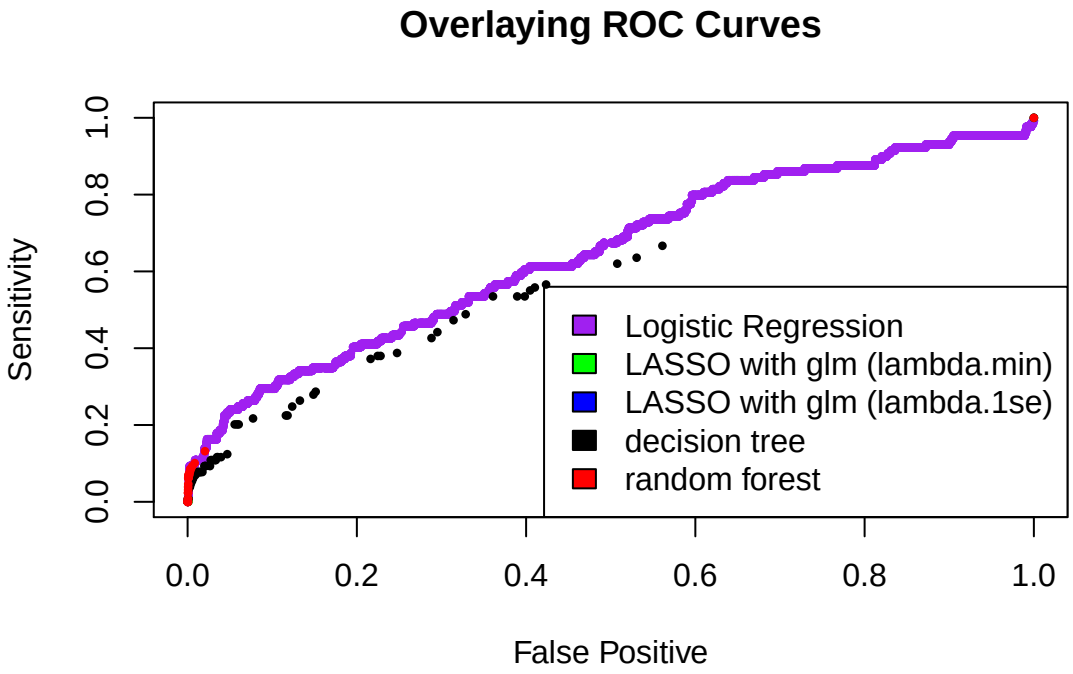
Table 3 Summary of Misclassification Error for Models

Model Name	Misclassification Error
Single Decision Tree	0.009
Random Forest	0.008

Warning: We do not include misclassification error for logistic and lasso, because the model are not appropriate for h1b data. Logistic model require balanced dataset while our h1b dataset is extremly unbalanced.

ROC curve and AUC

For each model or analyzing process, there will be a pair of sensitivity and specificity given a threshold or a classifier. We then plotted ROC curve which includes all the pairs of **False Positive** and **True Positive** by changing the thresholds. We overlayed ROC curves of all the models, shown as below, to measure their performance. As we can see, all models have the similar trend with direct logistic regression while LASSO performs a little better.



AUC is "the area under the curve", which is also used to measure the performance of the classifier as a whole. The classifier with larger AUC means the classifier is better. We summarized AUCs of all the model in the following table.

Table 4 Summary of AUCs of Models

Model Name	AUC
Direct Logistic Regression	0.65
LASSO with glm (lambda.min)	0.51
Single Decision Tree	0.59
Random Forest	0.55

Bayes Rule with Unequal Loss

In reality, the loss of mislabeling may be different, and we should consider the weighted loss into our classification to be best implemented in real-world scenario. Thus, we tried to use Bayes Rule to treat this problem. Specifically, we first made the corresponding equation for the preparation of conducting Bayes Rule classification. Then, we used the testing data to see how well the classifier performed. In our case, the two type of the mistakes cost almost the same, so we set the loss ratio to 1:1.

The loss function is: $a_{1,0} = L(Y = 1, \hat{Y} = 0)$: the loss (cost) of making an "1" to a "0".
$a_{0,1} = L(Y = 0, \hat{Y} = 1)$: the loss of making a "0" to an "1".
$a_{0,0} = a_{1,1} = 0$

We calculated weighted misclassification error with $\frac{a_{0,1}}{a_{1,0}} = 2$ and used Bayes rule:

$$P(Y = 1|x) > \frac{\frac{a_{0,1}}{a_{1,0}}}{1 + \frac{a_{0,1}}{a_{1,0}}}$$

The weighted misclassification errors of all the model tried is shown below.

Table 5 Summary of Weighted Misclassificaton Error of Models

Model Name	Weighted Misclassification Error
Single Decision Tree	0.01
Random Forest	0.009

From the summary table, we could see that random forest is absolutely the best model for the H1B approval prediction. Logistic regression and LASSO could not take variables interactions. For H1B disclosure data with all categorical variables, random forest model ought to be the most accurate model for workers to predict their approval status of H1B application.

Variable Evaluation

To explore which identify the important risk factors that a loan will be fully paid or not, we listed the predictors selected by each model summarized as follows:

Table 6 Summary of Predictors Selected by Models

Predictors	Logistic Regression	LASSO with glm (min)	Single Tree	Random Forest
VISA_CLASS	v	v	v	v
AGENT_REPRESENTING_EMPLOYER	v	–	v	v
TOTAL_WORKERS	v	–	v	v
NEW_EMPLOYMENT	v	–	v	v
CONTINUED_EMPLOYMENT	v	–	v	v
CHANGE_PREVIOUS_EMPLOYMENT	v	–	v	v
NEW_CONCURRENT_EMP	–	–	–	v
CHANGE_EMPLOYER	v	–	v	v
AMENDED_PETITION	v	–	v	v
FULL_TIME_POSITION	v	–	–	v
PREVAILING_WAGE	v	–	v	v
PW_WAGE_LEVEL	–	–	v	v
H1B_DEPENDENT	v	–	v	v
WILLFUL_VIOLATOR	–	–	–	v

Since random forest is an ensemble method, it includes almost all the variables. Take a step further, we introduced a metric to measure the importance of variables. To get the variable importance of each variable, we sum over the decrease of gini index (RSS for regression trees) of the splits using the given variable in each tree, and then average across the bagged trees the sum of gini index reduction(See the graph in Appendix 3.4).

From the above analysis, we concluded that predictors `PREVAILING_WAGE` and `VISA_CLASS` are two important factors with respect to random forest. While predictors `WILLFUL_VIOLATOR` and `NEW_CONCURRENT_EMP` are less important.

Conclusion and Discussion

Compensation Prediction and Text Mining

From LASSO and text mining model, we figured out international employees are targeting high-tech company jobs. Most international students and foreign professionals have a very decent annual compensation, with the

average of 90 thousands dollars, which is much higher than the average compensation level across the United States. We also conlcuded that higher salaries are tied with higher levels of postiions, more economically flourishing states, and the specialties included in the job titles. Furthermore, we observed that workers who have been promoted to principals, senior levels, and managers should have a relatively high salary, while jobs such as accountant, assistant and associate are more likely to be correlated with lower levels of compensation.

H1B Approval Prediction

Overall, it is very unlikely a H1B petition would get denied after being selected as part of H1B project, except for the condition where an employee is doing a job which is highly replaceable or his/her annual salary is below the H1B standard records for this specific job title. The most important factor for the approval of H1B, based on our analysis, is the annual salary. In other words, if the employee were receiving a relatively high salary, doing a job which is closely related to his/her degree major (undergrad or graduate degrees), then he or she would be less likely to be denied through the final consideration of H1B projects.

Discussions

As an end note, although we successfully obtained some useful insights with respect to H1B applications, our conclusion is still prone to critiques.

The major problem is that H1B candidates are largely selected by the lottery system in oppose to a manual filter (Retrieved from https://www.uscis.gov/archive/archive-news/uscis-reaches-fy-2017-h-1b-cap), which contains too many uncontrollable factors and uncertainty. One of the most direct evidence shown in our analysis is the astonishingly bad performance of regression models in predicting the approval status, of which the misclassification errors are around 0.99, indicating a large degree of stochasticity within. However, we still decided to explore the H1B dataset because we are part of the eligible candidates with somewhat worries about career paths after graduation. We would like to see if there is any easily-captured trend so that international students could carefully think about their future beforehand.

Another critical issue is that we didn't consider different pool of H1B candidates, where people from Singapore or Chile are eligible for a separate H-1B1 visa application ratified by the Singapore-United States Free Trade Agreement (Retrieved from https://www.govtrack.us/congress/bills/108/hr2739) and the Chile-United States Free Trade Agreement (Retrieved from https://www.nytimes.com/2003/06/07/business/chile-and-us-sign-accord-on-free-trade.html) in 2003. Thus, there should be different analysis for these two countries since they are not on the same page of H1B candidates from other countries such as China and India. In future exploration, researchers should also conduct analysis supported by legal documents to best capture the subtle regularity for foreign working visa applications and other similar certifications.

More importantly, the explanatory power of our analysis is limited. We noticed that our dataset is restricted to a narrow period of time, which couldn't manifest the changes throughout the past decades where some more profound insights might be inferred or uncovered. To be more specific, the analysis towards H1B could be more compelling if being put in a bigger picture, say, an exploration of the preferences or trends hidden in the demand of the job markets within the United States. A slight touch upon the data retrieved from one year is faint in explaining or even predicting the macro employment patterns for foreign talents. A ten-year based investigation is probably needed for more valuable and rigorous conclusions.

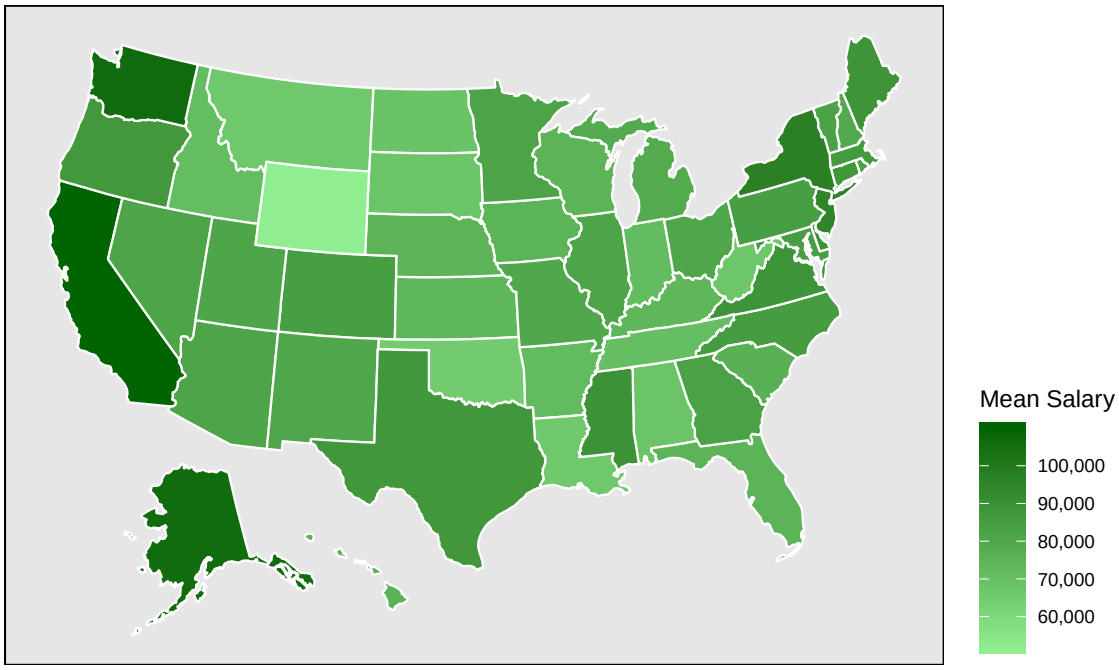
In addition, we found some vital limitations of the dataset. For example, we noticed that the education levels of H1B candidates are not included, whereas candidates with undergraduate degrees and those with graduate degrees are handled respectively according to the H-1B Visa Reform Act and recent announcements from United States Immigration and Citizenship Services (Retrieved from https://timesofindia.indiatimes.com/world/us/a-us-masters-degree-will-increase-your-chances-of-an-h1-b-visa/articleshow/66902664.cms), whereas we were failed to find education-related records in the dataset. This missing data might account for the weakness of our models which we kept being skeptical of. It would be very helpful if future researched could dig deeply into the correlations between education, salaries, employers submitting the petitions, and job titles for foreign talents.

Acknowledgement

We would like to thank Prof Linda Zhao for this impressive class and guidance along the way, every TA who has once helped us out, and every fellows who played a role in the success of our final project. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the University of Pennsylvania or the Wharton School.

Appendices

Appendix 0 Explanatory Analysis: Mean Salary across States



Appendix 1 Compensation Prediction: Linear Regression

Appendix 1.1 Simple Linear Regression Summary: Full model

```
summary(fit.wage.1.0)
```

```
##
## Call:
## lm(formula = PREVAILING_WAGE ~ ., data = h1b.data.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -116379   -9490     787    8066 1085619
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     59277.46   21567.05   2.749 0.005988 **
## CASE_STATUSCERTIFIED-WITHDRAWN  -2301.90     246.89  -9.324  < 2e-16 ***
## CASE_STATUSDENIED                5033.96     961.74   5.234 1.66e-07 ***
## CASE_STATUSWITHDRAWN            -1625.49     511.79  -3.176 0.001494 **
## VISA_CLASSH-1B                   5936.45    2573.98   2.306 0.021096 *
## VISA_CLASSH-1B1 Chile           -3884.90    9509.10  -0.409 0.682875
## VISA_CLASSH-1B1 Singapore        2883.47    6965.47   0.414 0.678900
## AGENT_REPRESENTING_EMPLOYERY     5872.95     200.65  29.269  < 2e-16 ***
## TOTAL_WORKERS                     379.02     576.83   0.657 0.511131
## NEW_EMPLOYMENT                   -337.68     576.72  -0.586 0.558203
## CONTINUED_EMPLOYMENT             -810.45     581.38  -1.394 0.163317
```

```
## CHANGE_PREVIOUS_EMPLOYMENT       -852.19     590.77  -1.443 0.149165
## NEW_CONCURRENT_EMP                 57.34     630.11   0.091 0.927493
## CHANGE_EMPLOYER                   262.54     583.57   0.450 0.652793
## AMENDED_PETITION                 -479.72     578.71  -0.829 0.407142
## FULL_TIME_POSITIONY              3527.97   20494.20   0.172 0.863324
## PW_WAGE_LEVELLevel II           21623.82     277.52  77.917  < 2e-16 ***
## PW_WAGE_LEVELLevel III          39861.79     307.43 129.661  < 2e-16 ***
## PW_WAGE_LEVELLevel IV           56141.22     347.63 161.496  < 2e-16 ***
## H1B_DEPENDENTY                   -737.43     198.87  -3.708 0.000209 ***
## WILLFUL_VIOLATORY                -318.90    4579.61  -0.070 0.944485
## WORKSITE_STATEAL               -24108.53    6372.12  -3.783 0.000155 ***
## WORKSITE_STATEAR               -26278.13    6252.01  -4.203 2.64e-05 ***
## WORKSITE_STATEAZ               -17607.20    6198.42  -2.841 0.004505 **
## WORKSITE_STATECA                 6420.23    6172.90   1.040 0.298313
## WORKSITE_STATECO               -12860.36    6220.62  -2.067 0.038703 *
## WORKSITE_STATECT                -9760.48    6213.86  -1.571 0.116244
## WORKSITE_STATEDC                -7478.83    6266.34  -1.193 0.232682
## WORKSITE_STATEDE               -12253.52    6311.08  -1.942 0.052192 .
## WORKSITE_STATEFL               -23050.05    6187.76  -3.725 0.000195 ***
## WORKSITE_STATEGA               -15170.66    6186.56  -2.452 0.014202 *
## WORKSITE_STATEGU               -21541.24    7099.99  -3.034 0.002415 **
## WORKSITE_STATEHI               -25818.83    7246.51  -3.563 0.000367 ***
## WORKSITE_STATEIA               -23199.30    6275.23  -3.697 0.000218 ***
## WORKSITE_STATEID               -25887.39    6636.83  -3.901 9.61e-05 ***
## WORKSITE_STATEIL               -16306.72    6181.78  -2.638 0.008345 **
## WORKSITE_STATEIN               -25611.25    6227.48  -4.113 3.92e-05 ***
## WORKSITE_STATEKS               -23808.54    6324.70  -3.764 0.000167 ***
## WORKSITE_STATEKY               -25782.41    6318.91  -4.080 4.51e-05 ***
## WORKSITE_STATELA               -28546.12    6470.79  -4.412 1.03e-05 ***
## WORKSITE_STATEMA               -10631.56    6183.59  -1.719 0.085562 .
## WORKSITE_STATEMD                -9405.00    6204.00  -1.516 0.129536
## WORKSITE_STATEME               -10444.43    6827.10  -1.530 0.126060
## WORKSITE_STATEMI               -20844.61    6191.11  -3.367 0.000761 ***
## WORKSITE_STATEMN               -18252.79    6203.69  -2.942 0.003260 **
## WORKSITE_STATEMO               -14516.89    6212.92  -2.337 0.019465 *
## WORKSITE_STATEMP               -33516.18   21372.33  -1.568 0.116839
## WORKSITE_STATEMS                -7900.30    6624.11  -1.193 0.233008
## WORKSITE_STATEMT               -23679.78    8122.79  -2.915 0.003556 **
## WORKSITE_STATENC               -14411.47    6187.84  -2.329 0.019863 *
## WORKSITE_STATEND               -19164.79    6840.83  -2.802 0.005088 **
## WORKSITE_STATENE               -22427.34    6329.94  -3.543 0.000396 ***
## WORKSITE_STATENH               -17296.40    6418.13  -2.695 0.007042 **
## WORKSITE_STATENJ                -4319.18    6180.65  -0.699 0.484665
## WORKSITE_STATENM                -9324.66    6552.69  -1.423 0.154734
## WORKSITE_STATENV               -19273.02    6460.59  -2.983 0.002854 **
## WORKSITE_STATENY                -3763.96    6178.01  -0.609 0.542361
## WORKSITE_STATEOH               -19012.38    6192.73  -3.070 0.002141 **
## WORKSITE_STATEOK               -29266.05    6402.82  -4.571 4.87e-06 ***
## WORKSITE_STATEOR               -13723.06    6234.80  -2.201 0.027737 *
## WORKSITE_STATEPA               -14947.95    6185.00  -2.417 0.015660 *
## WORKSITE_STATEPR               -50170.27    8940.58  -5.612 2.01e-08 ***
## WORKSITE_STATEPW               -21257.95   21371.16  -0.995 0.319885
## WORKSITE_STATERI               -18416.13    6324.32  -2.912 0.003593 **
## WORKSITE_STATESC               -20961.14    6297.57  -3.328 0.000874 ***
```

```
## WORKSITE_STATESD               -22318.13    7617.77  -2.930 0.003394 **
## WORKSITE_STATETN               -25689.33    6225.14  -4.127 3.69e-05 ***
## WORKSITE_STATETX               -10117.09    6175.78  -1.638 0.101388
## WORKSITE_STATEUT               -22678.49    6294.75  -3.603 0.000315 ***
## WORKSITE_STATEVA                -7613.51    6190.36  -1.230 0.218740
## WORKSITE_STATEVI               -49574.86   13327.71  -3.720 0.000200 ***
## WORKSITE_STATEVT               -17780.20    7152.01  -2.486 0.012920 *
## WORKSITE_STATEWA                 9486.77    6179.19   1.535 0.124722
## WORKSITE_STATEWI               -23611.71    6225.01  -3.793 0.000149 ***
## WORKSITE_STATEWV               -31111.84    6768.76  -4.596 4.31e-06 ***
## WORKSITE_STATEWY               -23815.06    8124.23  -2.931 0.003376 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20460 on 56367 degrees of freedom
## Multiple R-squared:  0.5076, Adjusted R-squared:  0.507
## F-statistic: 774.8 on 75 and 56367 DF,  p-value: < 2.2e-16
```

Appendix 1.2 Simple Linear Regression Summary: Five-variable model

```
summary(fit.wage.1.1)
```

```
##
## Call:
## lm(formula = PREVAILING_WAGE ~ CASE_STATUS + AGENT_REPRESENTING_EMPLOYER +
##     PW_WAGE_LEVEL + H1B_DEPENDENT + WORKSITE_STATE, data = h1b.data.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -116067   -9552     831    7848 1085381
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     68671.5     6180.3  11.111  < 2e-16 ***
## CASE_STATUSCERTIFIED-WITHDRAWN  -2219.9      245.1  -9.058  < 2e-16 ***
## CASE_STATUSDENIED                4763.7      955.6   4.985 6.22e-07 ***
## CASE_STATUSWITHDRAWN            -1607.0      511.8  -3.140 0.001690 **
## AGENT_REPRESENTING_EMPLOYERY     5954.2      199.1  29.908  < 2e-16 ***
## PW_WAGE_LEVELLevel II           21570.4      277.2  77.829  < 2e-16 ***
## PW_WAGE_LEVELLevel III          39828.8      307.0 129.721  < 2e-16 ***
## PW_WAGE_LEVELLevel IV           56113.1      346.7 161.839  < 2e-16 ***
## H1B_DEPENDENTY                   -814.7      198.4  -4.107 4.01e-05 ***
## WORKSITE_STATEAL               -24113.7     6376.5  -3.782 0.000156 ***
## WORKSITE_STATEAR               -26193.4     6256.2  -4.187 2.83e-05 ***
## WORKSITE_STATEAZ               -17549.1     6202.6  -2.829 0.004667 **
## WORKSITE_STATECA                 6519.1     6177.0   1.055 0.291257
## WORKSITE_STATECO               -12844.3     6224.8  -2.063 0.039078 *
## WORKSITE_STATECT                -9853.8     6218.1  -1.585 0.113040
## WORKSITE_STATEDC                -7511.0     6270.5  -1.198 0.230989
## WORKSITE_STATEDE               -12323.4     6315.3  -1.951 0.051019 .
## WORKSITE_STATEFL               -23095.8     6191.8  -3.730 0.000192 ***
## WORKSITE_STATEGA               -15182.8     6190.7  -2.453 0.014189 *
## WORKSITE_STATEGU               -21904.7     7103.3  -3.084 0.002045 **
## WORKSITE_STATEHI               -25811.4     7251.4  -3.559 0.000372 ***
## WORKSITE_STATEIA               -23212.1     6279.5  -3.697 0.000219 ***
```

```
## WORKSITE_STATEID              -25842.9     6641.3  -3.891 9.98e-05 ***
## WORKSITE_STATEIL              -16332.1     6185.9  -2.640 0.008288 **
## WORKSITE_STATEIN              -25706.5     6231.7  -4.125 3.71e-05 ***
## WORKSITE_STATEKS              -23801.6     6329.0  -3.761 0.000170 ***
## WORKSITE_STATEKY              -25829.0     6323.2  -4.085 4.42e-05 ***
## WORKSITE_STATELA              -28552.7     6475.1  -4.410 1.04e-05 ***
## WORKSITE_STATEMA              -10609.8     6187.8  -1.715 0.086418 .
## WORKSITE_STATEMD               -9402.9     6208.2  -1.515 0.129882
## WORKSITE_STATEME              -10444.3     6831.7  -1.529 0.126320
## WORKSITE_STATEMI              -20806.5     6195.3  -3.358 0.000784 ***
## WORKSITE_STATEMN              -18258.2     6207.8  -2.941 0.003271 **
## WORKSITE_STATEMO              -14497.0     6217.1  -2.332 0.019716 *
## WORKSITE_STATEMP              -33485.7    21386.9  -1.566 0.117422
## WORKSITE_STATEMS               -7956.9     6628.6  -1.200 0.229995
## WORKSITE_STATEMT              -23608.5     8128.2  -2.905 0.003680 **
## WORKSITE_STATENC              -14390.9     6192.0  -2.324 0.020124 *
## WORKSITE_STATEND              -19170.1     6845.5  -2.800 0.005106 **
## WORKSITE_STATENE              -22358.4     6334.2  -3.530 0.000416 ***
## WORKSITE_STATENH              -17299.1     6422.5  -2.694 0.007072 **
## WORKSITE_STATENJ               -4329.0     6184.8  -0.700 0.483971
## WORKSITE_STATENM               -9155.7     6557.1  -1.396 0.162634
## WORKSITE_STATENV              -19291.9     6464.7  -2.984 0.002844 **
## WORKSITE_STATENY               -3772.5     6182.2  -0.610 0.541715
## WORKSITE_STATEOH              -18988.6     6196.9  -3.064 0.002183 **
## WORKSITE_STATEOK              -29262.3     6407.2  -4.567 4.95e-06 ***
## WORKSITE_STATEOR              -13964.4     6238.7  -2.238 0.025202 *
## WORKSITE_STATEPA              -14967.7     6189.2  -2.418 0.015594 *
## WORKSITE_STATEPR              -50344.9     8946.6  -5.627 1.84e-08 ***
## WORKSITE_STATEPW              -21316.1    21385.7  -0.997 0.318892
## WORKSITE_STATERI              -18508.9     6328.6  -2.925 0.003450 **
## WORKSITE_STATESC              -20952.4     6301.9  -3.325 0.000885 ***
## WORKSITE_STATESD              -22720.3     7621.4  -2.981 0.002873 **
## WORKSITE_STATETN              -25672.8     6229.3  -4.121 3.77e-05 ***
## WORKSITE_STATETX              -10058.1     6179.9  -1.628 0.103625
## WORKSITE_STATEUT              -22533.4     6299.0  -3.577 0.000347 ***
## WORKSITE_STATEVA               -7640.3     6194.5  -1.233 0.217434
## WORKSITE_STATEVI              -50080.3    13336.6  -3.755 0.000173 ***
## WORKSITE_STATEVT              -17893.3     7156.9  -2.500 0.012417 *
## WORKSITE_STATEWA                9493.4     6183.4   1.535 0.124715
## WORKSITE_STATEWI              -23676.7     6229.2  -3.801 0.000144 ***
## WORKSITE_STATEWV              -31122.8     6773.4  -4.595 4.34e-06 ***
## WORKSITE_STATEWY              -23923.0     8129.7  -2.943 0.003256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20470 on 56379 degrees of freedom
## Multiple R-squared:  0.5068, Adjusted R-squared:  0.5063
## F-statistic: 919.7 on 63 and 56379 DF,  p-value: < 2.2e-16
```

Appendix 2 Compensation Prediction: LASSO

Appendix 2.1 Coefficients for the subset selected by LASSO analysis

```
coef.min <- coef(fit.cv, s="lambda.min")  #s=c("lambda.1se","lambda.min") or lambda value
coef.min <- coef.min[which(coef.min !=0),]   # get the non=zero coefficients
coef.min  # the set of predictors chosen
```

```
##                   (Intercept) CASE_STATUSCERTIFIED-WITHDRAWN
##                   51618.51702                   -2177.35942
##              CASE_STATUSDENIED          CASE_STATUSWITHDRAWN
##                    3950.88508                   -1070.40828
##                 VISA_CLASSH-1B   AGENT_REPRESENTING_EMPLOYERY
##                    3193.96504                    5884.24398
##                NEW_EMPLOYMENT          CONTINUED_EMPLOYMENT
##                     26.57887                    -299.31868
##               CHANGE_EMPLOYER              AMENDED_PETITION
##                    403.40645                     -63.16238
##          PW_WAGE_LEVELLevel II         PW_WAGE_LEVELLevel III
##                   20582.74674                   38797.56935
##          PW_WAGE_LEVELLevel IV                 H1B_DEPENDENTY
##                   55012.83593                    -554.45604
##                WORKSITE_STATEAL               WORKSITE_STATEAR
##                    -7873.97879                   -10450.51269
##                WORKSITE_STATEAZ               WORKSITE_STATECA
##                    -2261.20937                    20987.49959
##                WORKSITE_STATECO               WORKSITE_STATECT
##                     946.70330                     4055.54178
##                WORKSITE_STATEDC               WORKSITE_STATEDE
##                    5973.41695                     1023.31941
##                WORKSITE_STATEFL               WORKSITE_STATEGU
##                    -7919.79427                    -3390.93871
##                WORKSITE_STATEHI               WORKSITE_STATEIA
##                    -6884.13284                    -7303.94419
##                WORKSITE_STATEID               WORKSITE_STATEIL
##                    -8525.05059                    -1253.09131
##                WORKSITE_STATEIN               WORKSITE_STATEKS
##                   -10085.94363                    -7644.07726
##                WORKSITE_STATEKY               WORKSITE_STATELA
##                    -9696.61922                   -11824.31351
##                WORKSITE_STATEMA               WORKSITE_STATEMD
##                     3538.99636                     4492.27557
##                WORKSITE_STATEME               WORKSITE_STATEMI
##                     1054.97932                    -5596.50692
##                WORKSITE_STATEMN               WORKSITE_STATEMS
##                    -2848.96668                     4124.76551
##                WORKSITE_STATEMT               WORKSITE_STATEND
##                    -3375.91801                    -1564.91141
##                WORKSITE_STATENE               WORKSITE_STATENH
##                    -6219.73164                     -745.42939
##                WORKSITE_STATENJ               WORKSITE_STATENM
##                    10015.32849                     2774.48003
##                WORKSITE_STATENV               WORKSITE_STATENY
##                    -2412.94518                    10611.50216
##                WORKSITE_STATEOH               WORKSITE_STATEOK
##                    -3742.62027                   -12739.31475
##                WORKSITE_STATEPR               WORKSITE_STATERI
##                   -28404.24670                    -2230.37702
```

```
##               WORKSITE_STATESC               WORKSITE_STATESD
##                    -4995.26743                    -3166.59996
##               WORKSITE_STATETN               WORKSITE_STATETX
##                   -10121.09098                     4262.26217
##               WORKSITE_STATEUT               WORKSITE_STATEVA
##                    -6497.38612                     6510.98226
##               WORKSITE_STATEVI               WORKSITE_STATEWA
##                   -21833.64763                    23812.71337
##               WORKSITE_STATEWI               WORKSITE_STATEWV
##                    -8062.48743                   -13415.50348
##               WORKSITE_STATEWY
##                    -3918.79037
```

```
rownames(as.matrix(coef.min)) # shows only names, not estimates
```

```
##  [1] "(Intercept)"                    "CASE_STATUSCERTIFIED-WITHDRAWN"
##  [3] "CASE_STATUSDENIED"              "CASE_STATUSWITHDRAWN"
##  [5] "VISA_CLASSH-1B"                 "AGENT_REPRESENTING_EMPLOYERY"
##  [7] "NEW_EMPLOYMENT"                 "CONTINUED_EMPLOYMENT"
##  [9] "CHANGE_EMPLOYER"                "AMENDED_PETITION"
## [11] "PW_WAGE_LEVELLevel II"          "PW_WAGE_LEVELLevel III"
## [13] "PW_WAGE_LEVELLevel IV"          "H1B_DEPENDENTY"
## [15] "WORKSITE_STATEAL"               "WORKSITE_STATEAR"
## [17] "WORKSITE_STATEAZ"               "WORKSITE_STATECA"
## [19] "WORKSITE_STATECO"               "WORKSITE_STATECT"
## [21] "WORKSITE_STATEDC"               "WORKSITE_STATEDE"
## [23] "WORKSITE_STATEFL"               "WORKSITE_STATEGU"
## [25] "WORKSITE_STATEHI"               "WORKSITE_STATEIA"
## [27] "WORKSITE_STATEID"               "WORKSITE_STATEIL"
## [29] "WORKSITE_STATEIN"               "WORKSITE_STATEKS"
## [31] "WORKSITE_STATEKY"               "WORKSITE_STATELA"
## [33] "WORKSITE_STATEMA"               "WORKSITE_STATEMD"
## [35] "WORKSITE_STATEME"               "WORKSITE_STATEMI"
## [37] "WORKSITE_STATEMN"               "WORKSITE_STATEMS"
## [39] "WORKSITE_STATEMT"               "WORKSITE_STATEND"
## [41] "WORKSITE_STATENE"               "WORKSITE_STATENH"
## [43] "WORKSITE_STATENJ"               "WORKSITE_STATENM"
## [45] "WORKSITE_STATENV"               "WORKSITE_STATENY"
## [47] "WORKSITE_STATEOH"               "WORKSITE_STATEOK"
## [49] "WORKSITE_STATEPR"               "WORKSITE_STATERI"
## [51] "WORKSITE_STATESC"               "WORKSITE_STATESD"
## [53] "WORKSITE_STATETN"               "WORKSITE_STATETX"
## [55] "WORKSITE_STATEUT"               "WORKSITE_STATEVA"
## [57] "WORKSITE_STATEVI"               "WORKSITE_STATEWA"
## [59] "WORKSITE_STATEWI"               "WORKSITE_STATEWV"
## [61] "WORKSITE_STATEWY"
```

Appendix 2.2 LASSO regression

```
summary(fit.wage.2.0)
```

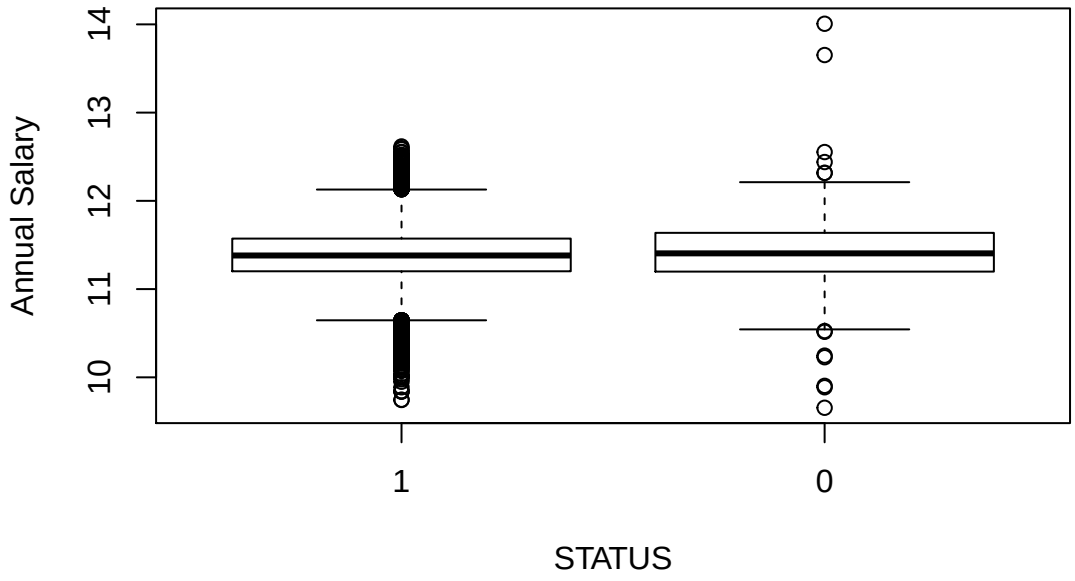
```
##
## Call:
## lm(formula = PREVAILING_WAGE ~ CASE_STATUS + AGENT_REPRESENTING_EMPLOYER +
##     NEW_EMPLOYMENT + CHANGE_PREVIOUS_EMPLOYMENT + CHANGE_EMPLOYER +
##     AMENDED_PETITION + PW_WAGE_LEVEL + WORKSITE_STATE, data = h1b.data.1)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -115812   -9506     716    8065 1085610
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    68524.59    6178.48  11.091  < 2e-16 ***
## CASE_STATUSCERTIFIED-WITHDRAWN -2179.36     246.24  -8.851  < 2e-16 ***
## CASE_STATUSDENIED               4967.08     954.32   5.205 1.95e-07 ***
## CASE_STATUSWITHDRAWN           -1500.41     511.52  -2.933 0.003356 **
## AGENT_REPRESENTING_EMPLOYERY    6109.21     190.51  32.068  < 2e-16 ***
## NEW_EMPLOYMENT                    43.03      15.31   2.811 0.004935 **
## CHANGE_PREVIOUS_EMPLOYMENT      -434.43     122.29  -3.552 0.000382 ***
## CHANGE_EMPLOYER                  548.31      84.91   6.457 1.07e-10 ***
## AMENDED_PETITION                -154.71      65.96  -2.345 0.019005 *
## PW_WAGE_LEVELLevel II          21376.75     270.15  79.128  < 2e-16 ***
## PW_WAGE_LEVELLevel III         39685.84     305.05 130.095  < 2e-16 ***
## PW_WAGE_LEVELLevel IV          56056.12     347.38 161.367  < 2e-16 ***
## WORKSITE_STATEAL              -24264.78    6374.59  -3.806 0.000141 ***
## WORKSITE_STATEAR              -26554.63    6254.17  -4.246 2.18e-05 ***
## WORKSITE_STATEAZ              -17924.35    6200.47  -2.891 0.003844 **
## WORKSITE_STATECA                6221.32    6175.24   1.007 0.313717
## WORKSITE_STATECO              -13091.99    6222.92  -2.104 0.035397 *
## WORKSITE_STATECT              -10131.64    6215.91  -1.630 0.103117
## WORKSITE_STATEDC               -7777.54    6268.48  -1.241 0.214708
## WORKSITE_STATEDE              -12595.98    6313.09  -1.995 0.046024 *
## WORKSITE_STATEFL              -23363.07    6189.87  -3.774 0.000161 ***
## WORKSITE_STATEGA              -15469.44    6188.69  -2.500 0.012435 *
## WORKSITE_STATEGU              -21886.98    7101.26  -3.082 0.002056 **
## WORKSITE_STATEHI              -25973.07    7249.30  -3.583 0.000340 ***
## WORKSITE_STATEIA              -23442.76    6277.42  -3.734 0.000188 ***
## WORKSITE_STATEID              -25961.75    6639.45  -3.910 9.23e-05 ***
## WORKSITE_STATEIL              -16631.97    6183.87  -2.690 0.007156 **
## WORKSITE_STATEIN              -25909.90    6229.72  -4.159 3.20e-05 ***
## WORKSITE_STATEKS              -24016.41    6327.08  -3.796 0.000147 ***
## WORKSITE_STATEKY              -26050.21    6321.04  -4.121 3.77e-05 ***
## WORKSITE_STATELA              -28668.52    6473.34  -4.429 9.50e-06 ***
## WORKSITE_STATEMA              -10844.57    6185.85  -1.753 0.079586 .
## WORKSITE_STATEMD               -9595.06    6206.32  -1.546 0.122107
## WORKSITE_STATEME              -10662.01    6829.65  -1.561 0.118498
## WORKSITE_STATEMI              -21057.97    6193.35  -3.400 0.000674 ***
## WORKSITE_STATEMN              -18597.66    6205.67  -2.997 0.002729 **
## WORKSITE_STATEMO              -14794.89    6215.01  -2.381 0.017292 *
## WORKSITE_STATEMP              -33536.84   21380.86  -1.569 0.116760
## WORKSITE_STATEMS               -7968.95    6626.74  -1.203 0.229158
## WORKSITE_STATEMT              -23865.24    8126.00  -2.937 0.003316 **
## WORKSITE_STATENC              -14671.09    6190.02  -2.370 0.017786 *
## WORKSITE_STATEND              -19227.77    6843.55  -2.810 0.004962 **
## WORKSITE_STATENE              -22620.55    6332.24  -3.572 0.000354 ***
## WORKSITE_STATENH              -17603.54    6420.32  -2.742 0.006111 **
## WORKSITE_STATENJ               -4617.43    6182.74  -0.747 0.455172
## WORKSITE_STATENM               -9311.09    6555.29  -1.420 0.155499
## WORKSITE_STATENV              -19499.02    6462.82  -3.017 0.002553 **
```

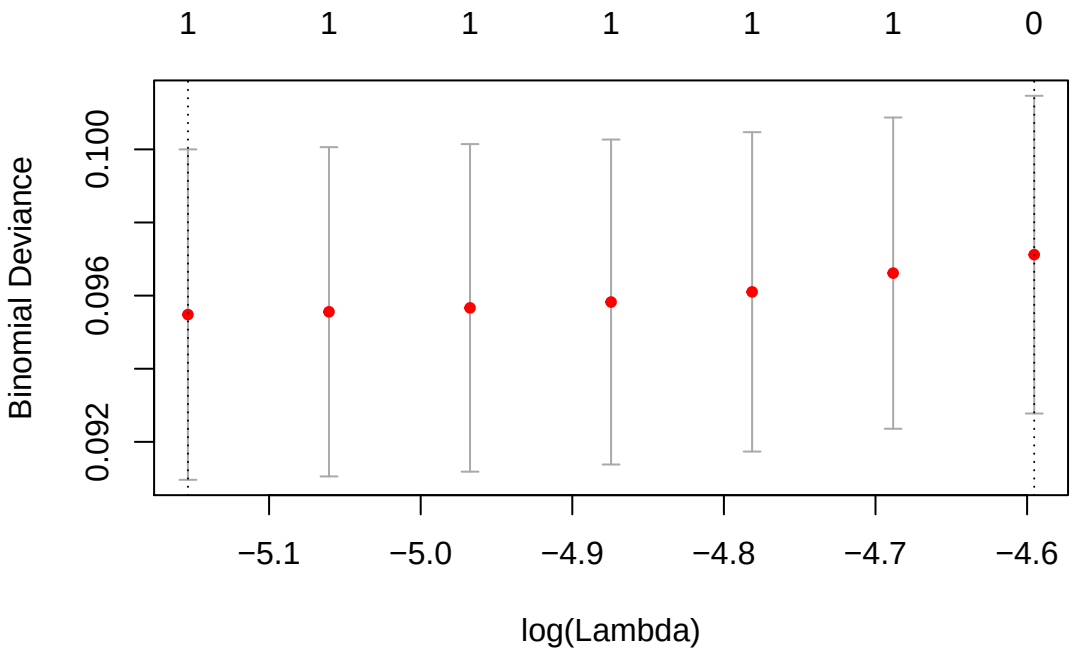
```
## WORKSITE_STATENY               -3941.96    6180.38  -0.638 0.523595
## WORKSITE_STATEOH              -19342.04    6194.65  -3.122 0.001795 **
## WORKSITE_STATEOK              -29402.32    6405.25  -4.590 4.43e-06 ***
## WORKSITE_STATEOR              -14298.97    6236.69  -2.293 0.021868 *
## WORKSITE_STATEPA              -15189.59    6187.23  -2.455 0.014092 *
## WORKSITE_STATEPR              -50373.96    8944.08  -5.632 1.79e-08 ***
## WORKSITE_STATEPW              -20975.85   21379.68  -0.981 0.326542
## WORKSITE_STATERI              -18740.79    6326.39  -2.962 0.003055 **
## WORKSITE_STATESC              -21271.89    6299.68  -3.377 0.000734 ***
## WORKSITE_STATESD              -22662.17    7619.55  -2.974 0.002939 **
## WORKSITE_STATETN              -25889.02    6227.46  -4.157 3.23e-05 ***
## WORKSITE_STATETX              -10383.09    6177.95  -1.681 0.092833 .
## WORKSITE_STATEUT              -22766.48    6297.18  -3.615 0.000300 ***
## WORKSITE_STATEVA               -7921.69    6192.47  -1.279 0.200816
## WORKSITE_STATEVI              -49891.22   13332.78  -3.742 0.000183 ***
## WORKSITE_STATEVT              -18019.99    7154.59  -2.519 0.011783 *
## WORKSITE_STATEWA                9500.98    6181.67   1.537 0.124309
## WORKSITE_STATEWI              -23929.52    6227.16  -3.843 0.000122 ***
## WORKSITE_STATEWV              -31214.80    6771.44  -4.610 4.04e-06 ***
## WORKSITE_STATEWY              -23945.12    8127.46  -2.946 0.003218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20470 on 56376 degrees of freedom
## Multiple R-squared:  0.5071, Adjusted R-squared:  0.5066
## F-statistic: 878.9 on 66 and 56376 DF,  p-value: < 2.2e-16
```

Appendix 3 H1B APPROVAL PREDICTION

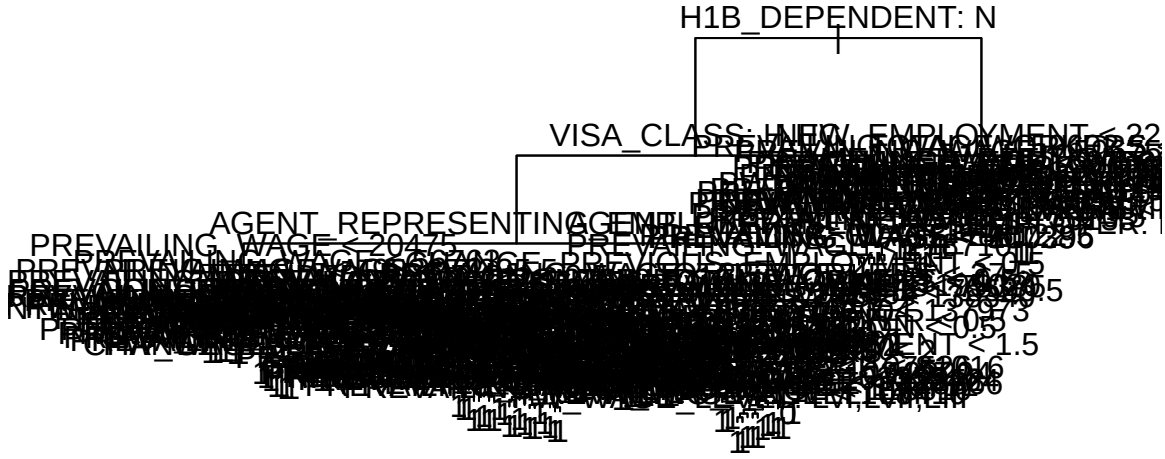
Appendix 3.0 Boxplot of Annual Salary with Certified and Denied Status



Appendix 3.1 Plot of log(Lambda) vs. Mean-Squared Error



Appendix 3.2 Plot of single tree to predict H1B approval



Appendix 3.3 Confusion Matrix to predict H1B approval

```
##
## fit.glm.predict     1     0
##               0 14641   129
##               1     1     0
```

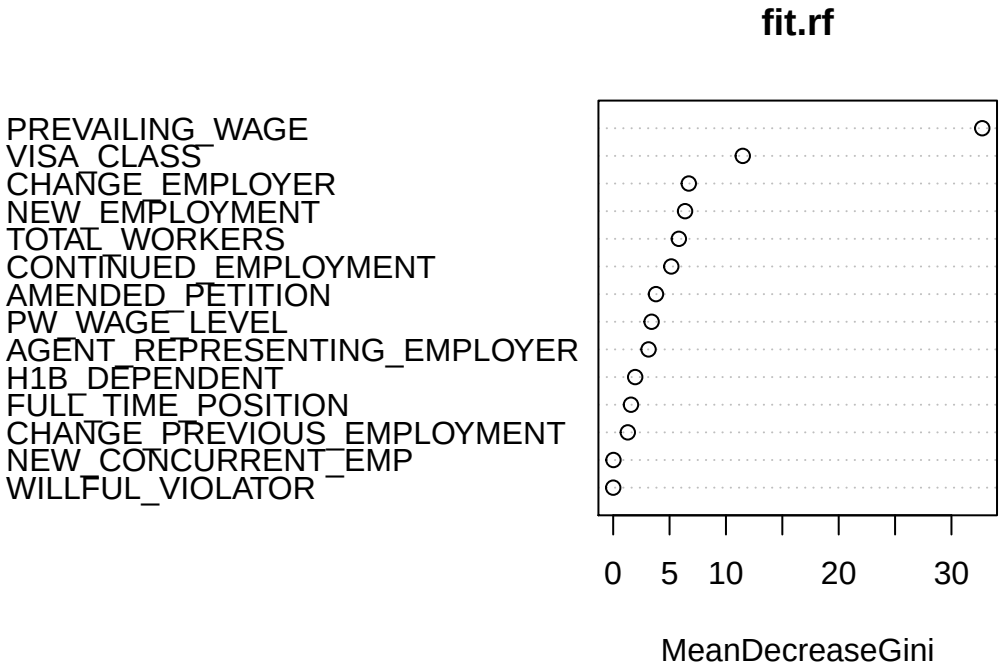
```
##
## fit.lasso.min.predict     1     0
##                     0 14642   129
```

```
##
## tree.predict.label     1     0
##                    1 14637   128
##                    0     5     1
```

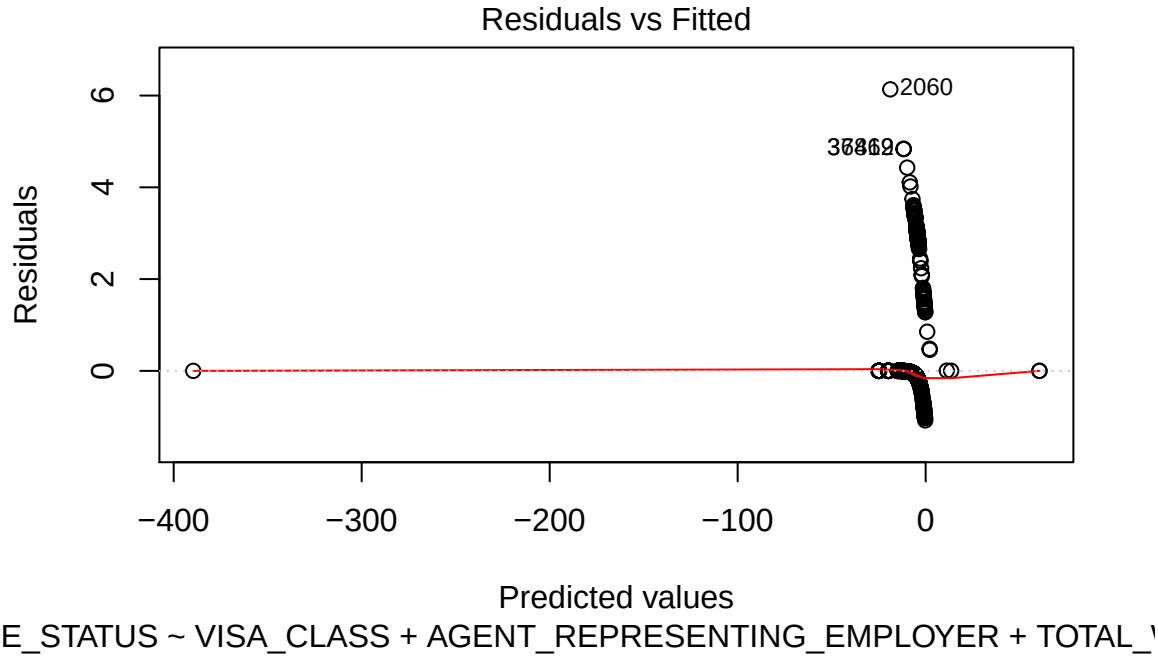
```
##
## rf.predict.label     1     0
##                  1 14641   129
```

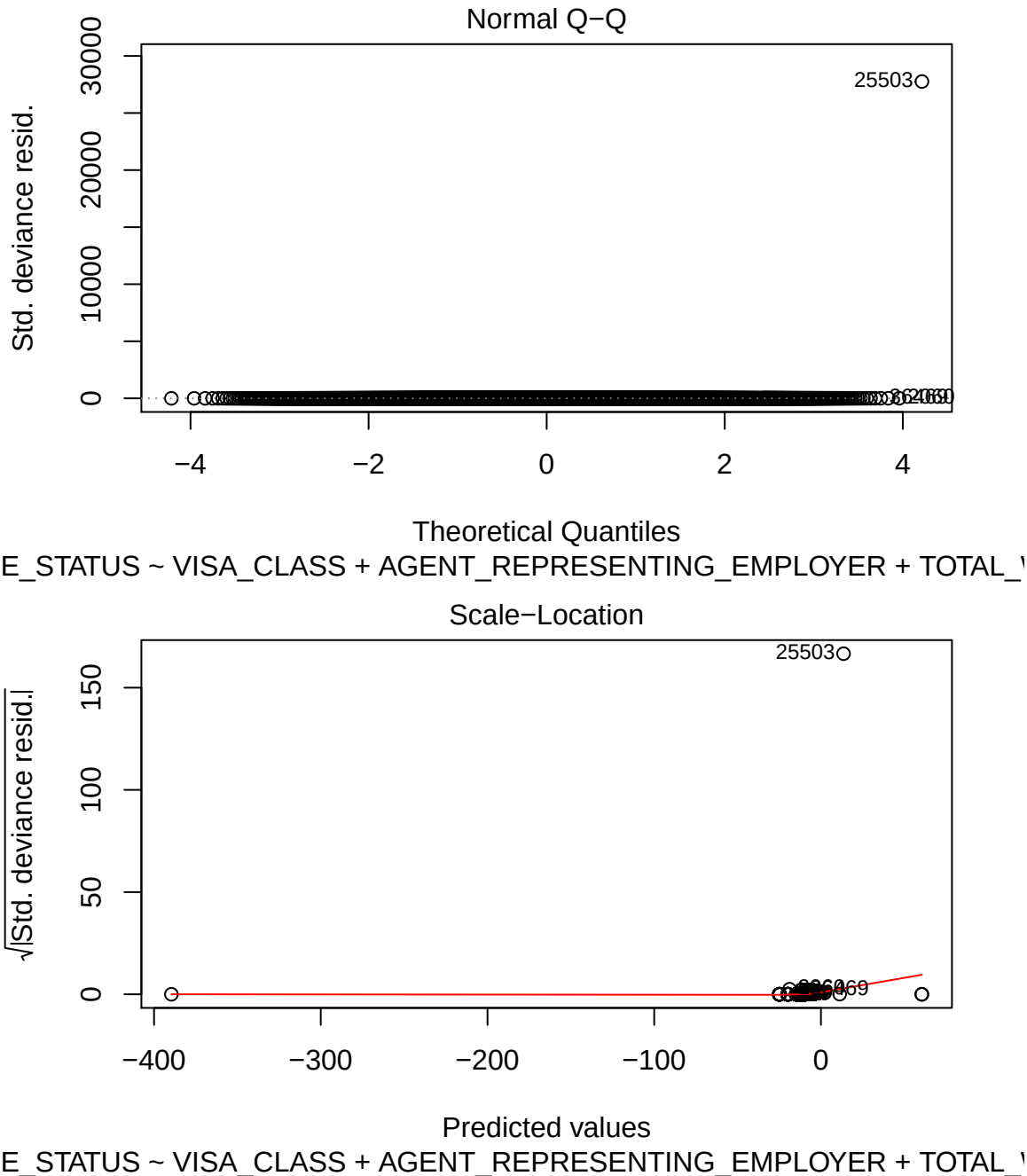
```
##                   0      1      0
```

Appendix 3.4 Variable Importance Measure in Random Forests to predict H1B approval



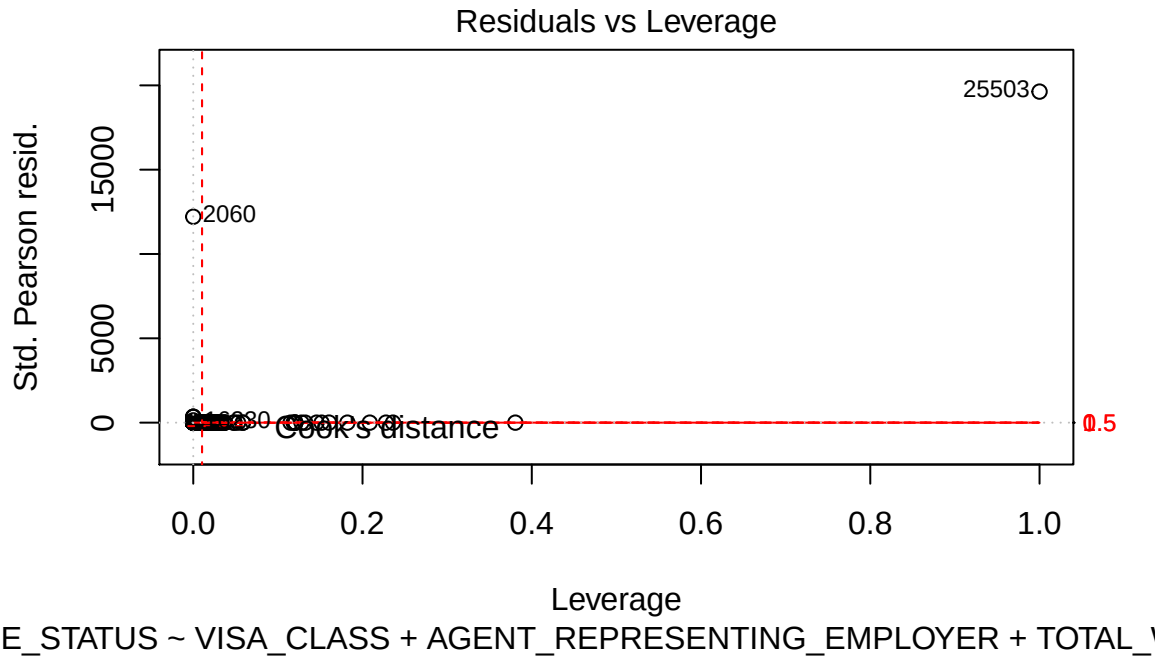
Appendix 3.5 Model diagonises for H1B Logistic






```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



E_STATUS ~ VISA_CLASS + AGENT_REPRESENTING_EMPLOYER + TOTAL_