USING NEURAL NETWORK TO PREDICT H-1B VISA SALARY

PROJECT

CONTENTS

1. PROBLEM DESCRIPTION AND CURRENT STATE OF DOMAIN

For most of international students, the expected salary is always one of the largest concerns. In our common sense, the job location, job title, education background, major, previous work experience etc. all related to the salary level of the international students. As an international student facing the same problem, I want to build a model to predict the H1B candidates' salary based on the current data set I have. I really hope my analysis result can help international to predict the future salary which in turn give the candidates some guidance for choosing job location, function, education etc.

2. DATASET DESCRIPTION: ORIGIN, DATA POINTS, VARIABLES

The original dataset comes is from the U.S. Department of Labor which contains H-1B candidates records from 2015 to 2017.

The original dataset includes five categories and 128 variables and more than 200,000 records. The five categories are:

a. Case information
b. Job information
c. Agency information
d. Candidates information
e. Others

The primary variables are listed below:

Seq	Variable Name	Data Format	Data Type	Description
1	**WAGE_OFFER**	Ratio	Dependent	the wage offer
2	**CASE_STATU**	Nominal	Independent	Status associated with the last significant event or decision. Valid values, include “Certified,” “Certified-Expired,” “Denied,” and “Withdrawn
3	**EMPLOYER_NAME**	Nominal	Independent	Name of employer requesting permanent labor certification
4	**EMPLOYER_STATE**	Nominal	Independent	Contact information of the employer requesting permanent labor certification
5	**EMPLOYER_NUM_EMPLOYEES**	Ratio	Independent	Total Number of employees employed by employer
6	**PW_LEVEL_9089**	Nominal	Independent	Level of the prevailing wage determination. Valid values include

				"Level I," "Level II," "Level III," and "Level IV"
7	**JOB_INFO_JOB_TITLE**	Nominal	Independent	Common name or payroll title of the job being offered
8	**JOB_INFO_TRAINING**	Binary	Independent	Identifies whether or not training is required for the job
9	**JOB_INFO_FOREIGN_LANG_REQ**	Binary	Independent	Indicates if knowledge of a foreign language is required to perform the job duties
10	**FW_INFO_TRAINING_COMP**	Nominal	Independent	Indicates whether the foreign worker completed the training required for the requested job opportunity
11	**FW_INFO_REQ_EXPERIENCE**	Nominal	Independent	Indicates whether the foreign worker has the experience as required for the requested job opportunity
12	**FW_INFO_ALT_EDU_EXPERIENCE**	Nominal	Independent	Indicates whether the foreign worker possesses the alternate combination of education and experience
13	**FW_INFO_REL_OCCUP_EXP**	Nominal	Independent	Indicates whether the foreign worker has the experience as required for the requested job opportunity
14	**FOREIGN_WORKER_INFO_EDUCATION**	Nominal	Independent	Highest Education achieved by the foreign worker
15	**FOREIGN_WORKER_INFO_INST**	Nominal	Independent	Name of the institution where the relevant education achieved by the foreign worker
16	**NAICS_US_CODE**	Nominal	Independent	Industry code associated with the employer requesting permanent labor certification, as classified by the North American Industrial Classification System (NAICS)
17	**CLASS_OF_ADMISSION**	Nominal	Independent	Indicates the class of immigration visa the foreign worker held at the time the permanent labor certification application was 18submitted for processing (if applicable)
18	**COUNTRY_OF_CITIZENSHIP**	Nominal	Independent	Country of citizenship of the foreign worker being sponsored by the employer for permanent employment in the United States.
19	**FOREIGN_WORKER_INFO_STATE**	Nominal	Independent	State of the foreign worker
20	**PW_UNIT_OF_PAY_9089:**	Categorical	Independent	Unit of Pay. Valid values include "Hourly (hr)", "Weekly (wk)," "Bi-Weekly (bi)," "Monthly (mth)," and "Yearly (yr)"

The basic statistic related the independent variable – Wages:

WAGE_OFFER	
Mean	101577.3382
Standard Error	529.5677537
Median	97889
Mode	105000
Standard Deviation	44040.17384
Skewness	1.911396312
Range	699991.95
Minimum	8.05
Maximum	700000
Count	6916

3. DATA PREPROCESSING ACTIVITIES AND RESULTS

Data preprocessing included the following actions:

Feature Selection

- **Step 1:** Screen the raw data and eliminate unrelated variables. The original data set has 128 attributes categorized into five categories: Case information, Job information, Agency information, Candidates' information, Others. Eliminate the three categories, H1B case information, agency information, and others, that are not related to predict the salary by the common sense. Within the left two categories, Job information and Candidates' information, select 19 variables (shown in the above table) as the independent variables.

- **Step2:** Regression. From a statistical point of view run the regression to eliminate the variables that not significantly related to the dependent variable. after the coefficiency analysis, the total number of variables deducted to eight variables(Appendix A).

Data Transformation

- **Step 1:** Eliminate null-value rows, convert categorical data into upper case, convert numerical data in to numerical format.
- **Step 2:** Filter the data for US based employer, CERTIFIED H1B cases, Yearly paid salary, and for H-1B class of candidates.
- **Step 3:** Categorize:

- **Step 3.1 :** Categorized Education into 6 categories: Primary, Bachelor, Master, Doctorate, None, and others
- **Step 3.2:** Categorized EMPLOYER_STATE and FOREIGN_WORKER _INFO_STATE into 4 regions: West, Midwest ,Northeast, South

- **Step4:** Separated salary into 3 categorical bins: Less than 50,000, 50,000 – 100,000, greater than 10,000 **(Appendix B**).

4. INTENDED ALGORITHMS AND RATIONALE

Neural Network Analysis

The Neural Network has been employed to predict the H-1B salary levels. the best rank of variable importance to determine whether H-1B visa status would be certified following an offer of employment. The variables available for inclusion in the Neural Network: salary bins/ranges, job training requirement, employee training status, employee work experience, alternate combination of education and experience, education level, and employer region.

The Neural Network algorithm was chosen because it can detect complex nonlinear relationships between dependent and independent variables. The resulting Neural Network can provide higher accuracy result that can provide a greater probability of a positive outcome for an H-1B visa seeker. Furthermore, we want to use the result from another methodology to compare with the result from decision tree that we accomplished in step 1.

5. EXECUTION

Independent variables:

The independent variables used in the neural network are described as follows:

- **Job training requirement:** Identifies whether or not training is required for the job
- **Job foreign language requirement:** Indicates if knowledge of a foreign language is required to perform the job duties
- **Employee training status:** whether the foreign worker completed the training required for the requested job opportunity
- **Employee work experience:** Indicates whether the foreign worker has the experience as required for the requested job opportunity
- **Alternate combination of education and experience:** Indicates whether the foreign worker possesses the alternate combination of education and experience

- **Education level:** We converted the education level category into six categories: Primary, Bachelor, Master, Doctorate, None, and others
- **Employer region :** We converted the employer states category into four categories: West, Midwest ,Northeast, South

Neural Network:

By default, the Neural Network was built by using the Resilient Backpropogation algorithm (RPROP+). RPROP is a fast algorithm and doesn't require as much tuning as classic backpropogation. The one hidden layers with five hidden nodes have been designed in the algorithm to achieve better accuracy. Salary was separated into three salary levels: 0- 50,000, 50,000 – 100,000, 100,000 and more. The result shows that the model has 62.88% overall accuracy. The evaluation of the prediction is shown below. Overall we have a good accuracy output since we have three outputs.

```
p
      0   1   2
  0   0  44  10
  1   0 512 238
  2   2 308 508

Overall Statistics

               Accuracy : 0.6288533
                 95% CI : (0.6048178, 0.6524214)
    No Information Rate : 0.5326757
    P-Value [Acc > NIR] : 0.000000000000003213501

                  Kappa : 0.2843251
 Mcnemar's Test P-Value : 0.000000000001351182497

Statistics by Class:

                        Class: 0  Class: 1  Class: 2
Sensitivity          0.000000000 0.5925926 0.6719577
Specificity          0.966666667 0.6860158 0.6420323
Pos Pred Value       0.000000000 0.6826667 0.6210269
Neg Pred Value       0.998724490 0.5963303 0.6915423
Prevalence           0.001233046 0.5326757 0.4660912
Detection Rate       0.000000000 0.3156597 0.3131936
Detection Prevalence 0.033292232 0.4623921 0.5043157
Balanced Accuracy    0.483333333 0.6393042 0.6569950
```

Predictive Results:

The analysis described above provided the most useful results, using information derived from descriptive analysis to segment the salary dependent variable into bins. The Model demonstrated an accuracy of 63%. We anticipate the need for feature selection and data clean to incorporate more independent variables in order to increase the accuracy of the described model. The complete visualization of the Neural Network model can be seen in Appendix C

6. CONCLUSION

1. **The company size is not critical to predict the salary:** At beginning I chose EMPLOYER_NUM_EMPLOYEES as one independent variable since in my common sense, the bigger the company is, the higher the salary should be. However, when I run the regression to select feature, the result was out of my expectation. The EMPLOYER_NUM_EMPLOYEES Is not statistical significant factor to predict the salary. In other words, the result indicates in the job selection, the company size is not a factor to determine the H1-B candidates' salary.

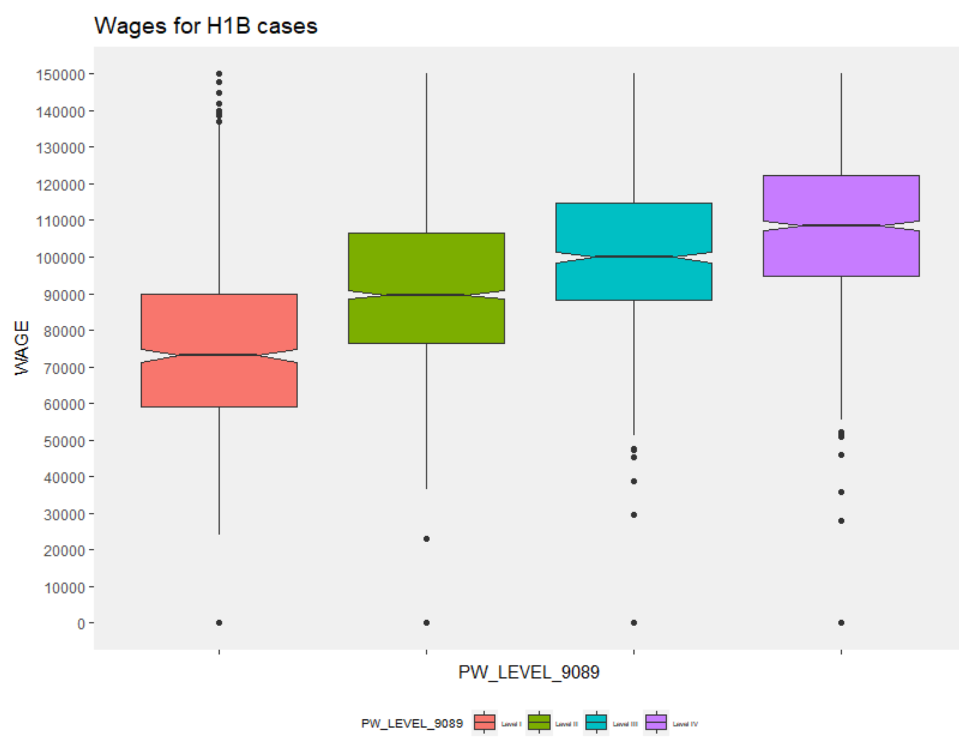
2. **63% overall accuracy:** from the trained algrosm, we can achive 63% of overall accuracy which is much better than the 33% of random guess. I satisfied with this prediction accuracy. The next step I can try increase the hidden layer nodes or include other variables to increase the accuracy. The model can be used to predict the possible salary range by given the parameters such like working region, education level, work experience etc. which can be really helpful in the next step of our research.

APPENDIX A. MULTIPLE REGRESSION

```
Coefficients:
                                      Estimate            Std. Error  t value              Pr(>|t|)    
(Intercept)                 2.0780850139299876  0.0333713325054282 62.27156 < 0.000000000000000222 ***
PW_LEVEL_9089Level I       -0.2213302304004628  0.0307174125309698 -7.20537 0.00000000000065847243 ***
PW_LEVEL_9089Level II      -0.0572395308437731  0.0276266383784168 -2.07190             0.03832251 *  
PW_LEVEL_9089Level III      0.1115093977118036  0.0297497647605655  3.74824             0.00017994 ***
PW_LEVEL_9089Level IV       0.3882551623742176  0.0283169911919109 13.71103 < 0.000000000000000222 ***
JOB_INFO_TRAININGY          0.6636867505843139  0.1872816997712802  3.54379             0.00039776 ***
EMPLOYER_NUM_EMPLOYEES      0.0000000009065173  0.0000000007246064  1.25105             0.21097126    
JOB_INFO_FOREIGN_LANG_REQY -0.4099864074279990  0.0498361368936518 -8.22669 0.00000000000000023902 ***
FW_INFO_TRAINING_COMPN     -0.1146322805151382  0.1684360609681865 -0.68057             0.49617375    
FW_INFO_TRAINING_COMPY     -0.2671677449334030  0.1798669720006960 -1.48536             0.13750611    
FW_INFO_REQ_EXPERIENCEN     0.0268664107264971  0.0171255017730462  1.56880             0.11675422    
FW_INFO_REQ_EXPERIENCEY    -0.0854203873640944  0.0163538431122409 -5.22326 0.00000018239959113361 ***
FW_INFO_ALT_EDU_EXPERIENCEN 0.0502965850472740  0.0201401503464514  2.49733             0.01254284 *  
FW_INFO_ALT_EDU_EXPERIENCEY 0.0430648226112376  0.0194657992272839  2.21233             0.02698541 *  
FW_INFO_REL_OCCUP_EXPN      0.1559746864580324  0.0465070585129974  3.35379             0.00080264 ***
FW_INFO_REL_OCCUP_EXPY      0.2200291942905896  0.0171607208530296 12.82168 < 0.000000000000000222 ***
EducationDoctorate          0.1464650505430622  0.0276295375519760  5.30103 0.00000011973217382694 ***
EducationMaster             0.0443833276301711  0.0149572208784539  2.96735             0.00301696 ** 
EducationNone              -0.0447350229502726  0.0751405019111910 -0.59535             0.55163339    
EducationOther              0.5017696708482156  0.0371662135079881 13.50069 < 0.000000000000000222 ***
Educationprimary           -0.1588851236721142  0.0998397043906679 -1.59140             0.11157767    
EMPLOYER_REGIONNortheast    0.0954529928988859  0.0207213631054519  4.60650 0.00000418917009856314 ***
EMPLOYER_REGIONSouth       -0.0460944651404245  0.0186102907547840 -2.47683             0.01328604 *  
EMPLOYER_REGIONWest         0.2809774052167702  0.0181448087142005 15.48528 < 0.000000000000000222 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4748921 on 5396 degrees of freedom
Multiple R-squared:  0.2909295, Adjusted R-squared:  0.2879071 
F-statistic: 96.25923 on 23 and 5396 DF,  p-value: < 0.000000000000000222044
```

APPENDIX B BOX PLOT OF THE H1B Salary



APPENDIX C VISUALIZATION OF NEURAL NETWORK