IST 707 Final report

H1B Application Result Determinants Analysis

Minyang Wang

Introduction

H-1B is a visa category in the United States of America under the Immigration and Nationality Act (INA), which allows U.S. employers to employ foreign workers. The first step employer must take to hire a foreign worker is to file the Labor Condition Application and employers can hire non-immigrant workers only after their LCA petition is approved. The approved LCA petition is then submitted as part of the Petition for a Non-immigrant Worker application for work authorizations for H-1B visa status. For an employee who is seeking to an employer, undoubtfully he would like to apply to the company or organization with 100% LCA approval rate so that he can further apply for the H1B visa.

Therefore, the goal of my project is to analyze the determinants of an approved or denied application and to predict the case status of an application submitted by the employer to hire non-immigrant workers under the H-1B visa program, so that based on the factors which may affect the result of an application, the employee can choose the employers who have the highest possibility of receiving approved LCA.

Method

Two data mining tasks are mainly applied in this project. In terms of finding the determinants of an approved or denied application, I used the association rule model to see what features are generally associated with an approved application or denied application by setting the right-hand side as approved or denied application status. For application status prediction, I combined the discovered features and some other attributes to predict the application status.

Besides the two mentioned tasks, descriptive statistics analysis is also used to describe the features of employers such as their location distribution, industry distribution, title of the job and prevailing wage etc. I also filtered all the instances by the employer's location, employer's industry and by the job tittle to see which state has the most number of H1B applications and the highest approved rate, and to see given an employer's industry which state has the highest approved rate so that for graduates who want to work in that industry, he can send more applications to companies located in that state.

Data

The data is from the Office of Foreign Labor Certification, U.S. Department of Labor Employment and Training Administration[1]. This public disclosure file contains administrative data from employers' Labor Condition Applications and the certification determinations processed by the Department's Office of Foreign Labor Certification, Employment and Training Administration where the date of the determination was issued on or after October 1, 2016, and on or before June 30, 2017.

The original data contains 27 columns and 528135 rows, each row represents an individual LCA case. I removed some irrelevant columns such as employer's zip code or telephone number. I also converted some columns to the form of data that may be considered as the determinant of the application case result. For example, there are two columns indicating the start date and end date of the employee's contract, I created a new column which measures the length of the employee's contract. For some continuous numerical variables such as the employee's annual salary, the employee's contract length, in order to be better analyzed by the classification model and association rule model, I converted them to discrete variables. For instance, I divided the annual salary into 7 categories: 0 to $20,000, $20,000 to $50,000, $50,000 to $80,000, $80,000 to $100,000, $100,000 to $150,000, $150,000 to $200,000, $200,000 to $500,000, and beyond $500,000.

After the data cleaning preprocessing steps, there are 12 columns left:

Table 1: columns label and descriptions

Attribute Label	Description
CASE_STATUS	Denied or certified
EMPLOYER_STATE	State where the employer locates
AGENT_REPRESENTING_EMPLOYER	Whether the employer is represented by an Agent or Attorney
SOC_NAME	Occupational name associated with the job being requested
NAICS_CODE	Industry code associated with the employer
PREVAILING_WAGE	Prevailing Wage for the job being requested
PW_WAGE_LEVEL	Variables include "I", "II", "III", "IV" or "N/A."
H-1B_DEPENDENT	Whether the employer is H1-B dependent or not
WILLFUL_VIOLATOR	Whether the employer has been previously found to be a willful violator or not
WORKSITE_STATE	State of the foreign worker's intended area of employment
CONTRACT_LENGTH	The length of contract that the foreign worker has signed with the employer

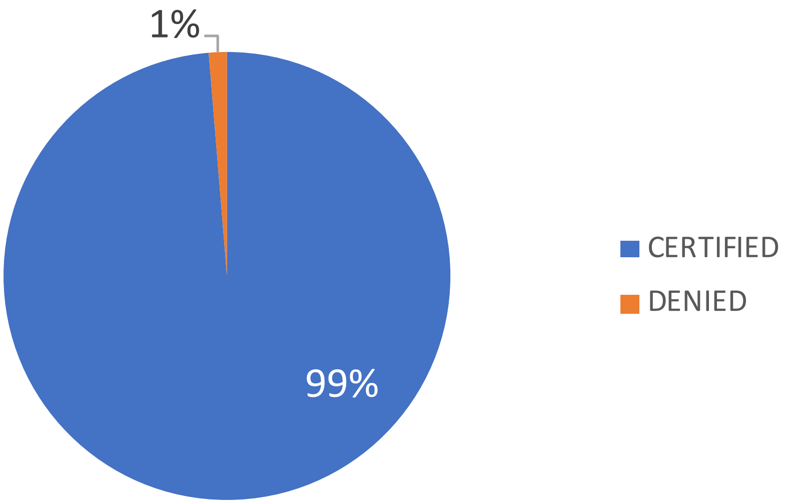
[1] https://www.foreignlaborcert.doleta.gov/performancedata.cfm

DECISION_DURATION	Duration of time that the employee received the case status since it submitted the application

Result

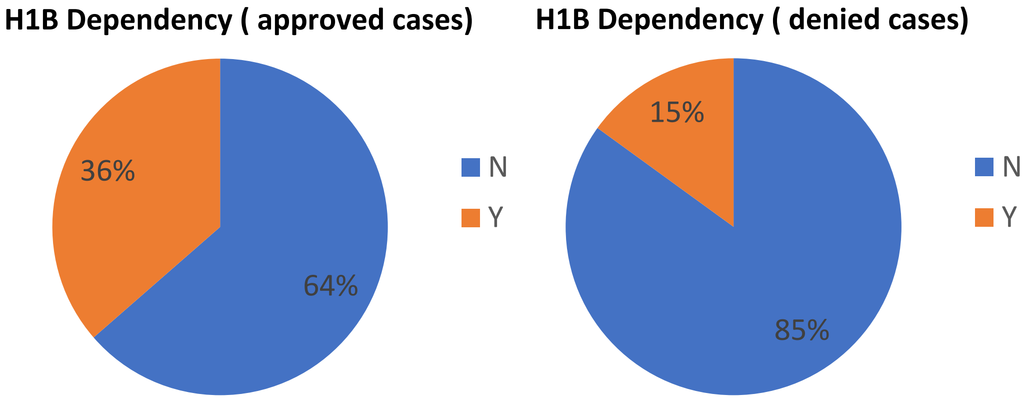
1. Descriptive analysis

Graph 1: Distribution of cases status for all applications



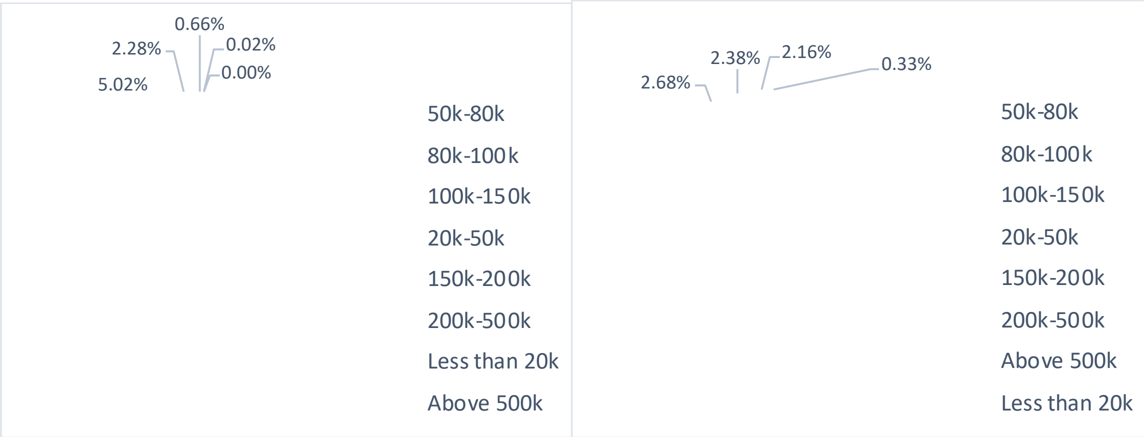
Among all the application cases, around 98.7% cases are approved whereas only 1.3% are denied. Since the data is strongly biased that the absolute majority of applications are approved, the case result distribution baseline ratio should be adjusted to around 98.7%, meaning that for a situation if the portion of approved case is below 98.7%, we can say that in this situation the application approval rate is relatively low.

Graph 2,3: Distribution of employer's H1B dependency for approved and denied cases



According to the U.S. Department of Labor Wage and Hour Division[2], an employer is considered H1B dependent if it has: 25 or fewer full-time equivalent employees and at least eight H-1B nonimmigrant workers; or 6 - 50 full-time equivalent employees and at least 13 H-1B nonimmigrant workers; or 51 or more full-time equivalent employees of whom15 percent or more are H-1B nonimmigrant workers. So a H1B dependent employer has higher demand of foreigner employees, and we can assume that whether the employer is H1B dependent can affect the application results. The data indicates that among all the approved cases, around 36% employers are H1B dependent, whereas the ratio drops to only 15% for the denied cases. So whether the employer is H1B dependent may be a consideration for the government to decide whether or not to approve the Labor Condition Application.

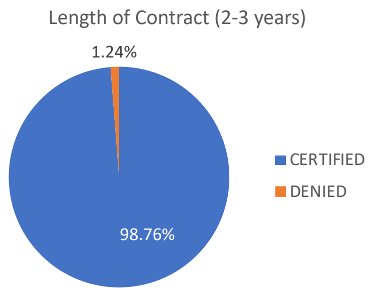
Graph 4,5: Distribution of employee's annual income level for approved and denied cases

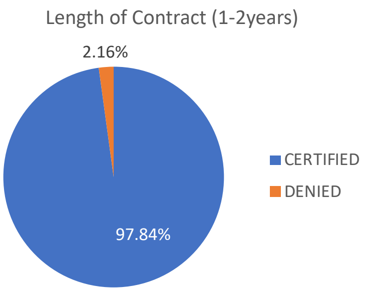


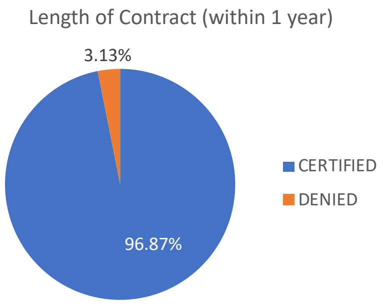
I also noticed that the annual income of the employee can also affect the result of LCA. Among the denied cases, the portion of low income ($20k-$50k) is 7.5% higher than the approved cases. And the portion of extremely high income ($200k-$500k and above $500k) in denied cases is higher than that in approved cases, probably it's because that annual income above $200k is considered not realistic for a H1B applicant. Thus, employees with extreme high or low annual income may have higher probability of being denied.

Graph 6,7: Distribution of cases status for different contract lengths

[2] https://www.dol.gov/whd/regs/compliance/FactSheet62/whdfs62C.pdf



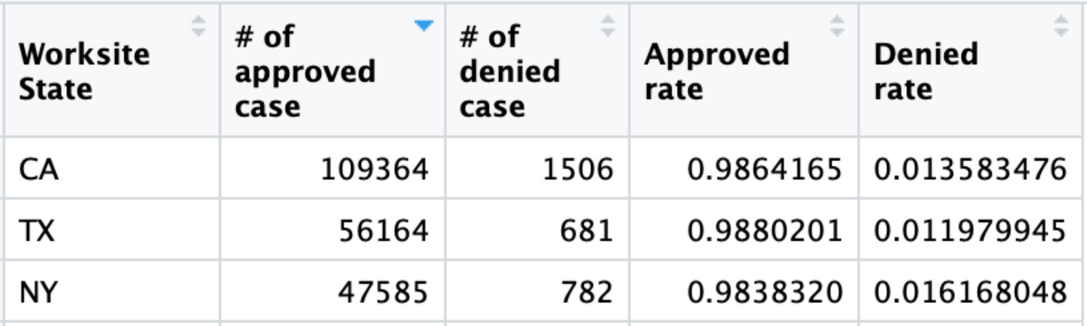




The length of the employee's labor contract may also make a difference to the application result. As can be seen in the pie charts above, for employees with shorter contract length, the portion of denied cases is higher. Since the ratio baseline is 98.7%, applicants with 1-2 years contract or shorter than 1-year contract have higher probability of being denied than the average.

For employers located in different states, the application approved rate may also be different. CA, TX and NY are the top 3 states with both the greatest number of application cases and the greatest number of approved cases.

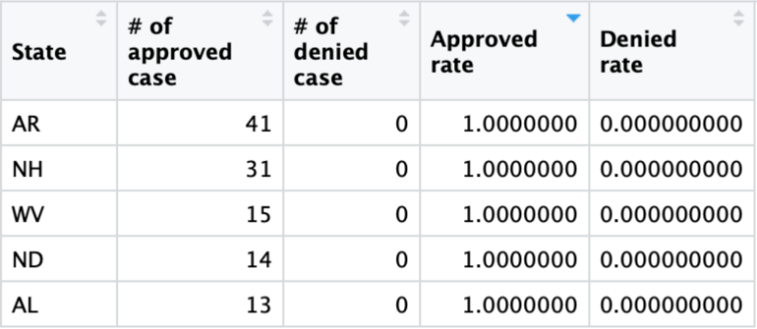
Table 2: Application result statistic by state



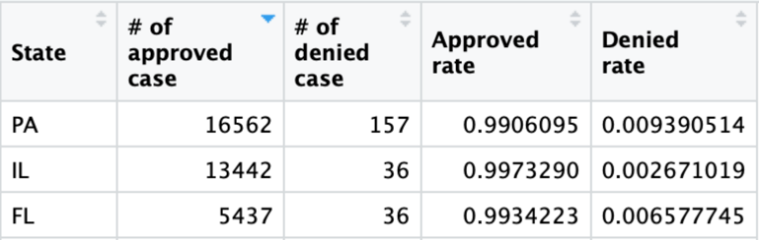
Worksite State	# of approved case	# of denied case	Approved rate	Denied rate
CA	109364	1506	0.9864165	0.013583476
TX	56164	681	0.9880201	0.011979945
NY	47585	782	0.9838320	0.016168048

Given a specific industry, the approved rate in different states may also be different. For applicants who work for employers located in management, scientific, and technical consulting services industry, Arkansas, New Hampshire, West Virginia, North Dakota and Alabama are the states have 100% approving rate; Pennsylvania, Illinois and Florida also have higher than 99% case approving rate. While on the other hand, MT, RI, ME, MS, OR and SC are the states where the applications have relative very high probability of being denied.

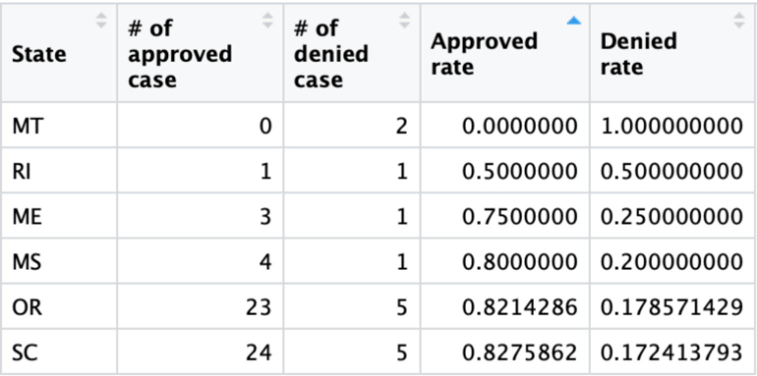
Table 3,4,5: Statistic of application filed by employers in management, scientific, and technical consulting services industry results, by state



State	# of approved case	# of denied case	Approved rate	Denied rate
AR	41	0	1.0000000	0.000000000
NH	31	0	1.0000000	0.000000000
WV	15	0	1.0000000	0.000000000
ND	14	0	1.0000000	0.000000000
AL	13	0	1.0000000	0.000000000



State	# of approved case	# of denied case	Approved rate	Denied rate
PA	16562	157	0.9906095	0.009390514
IL	13442	36	0.9973290	0.002671019
FL	5437	36	0.9934223	0.006577745



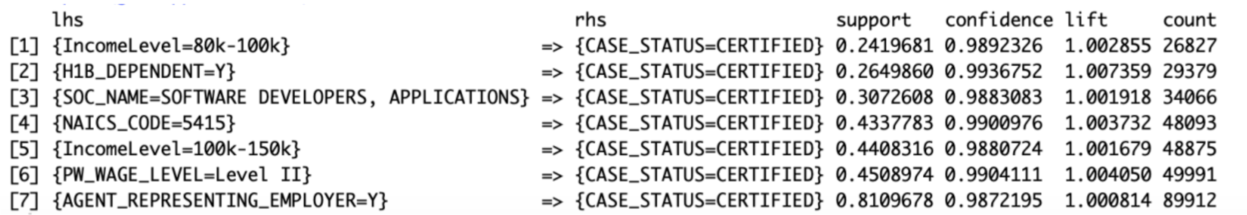
State	# of approved case	# of denied case	Approved rate	Denied rate
MT	0	2	0.0000000	1.000000000
RI	1	1	0.5000000	0.500000000
ME	3	1	0.7500000	0.250000000
MS	4	1	0.8000000	0.200000000
OR	23	5	0.8214286	0.178571429
SC	24	5	0.8275862	0.172413793

2. Association rules task

I did 2 association rules tasks trying to figure out what are the features of approved application cases and denied application cases.

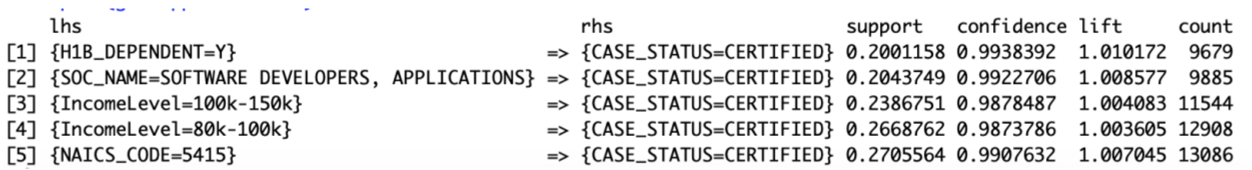
Since employers in CA and NY filed the most applications, first of all I did association rules task for applications that employers locate in CA and NY. I set the right hand side of the rules as approved case status and denied case status to figure out what are the features of approved application cases and denied application cases in CA and NY.

Graph 8: association rules of applications that employers in CA, with approved case status as rhs



```
    lhs                                          rhs                        support   confidence lift     count
[1] {IncomeLevel=80k-100k}                    => {CASE_STATUS=CERTIFIED} 0.2419681 0.9892326 1.002855 26827
[2] {H1B_DEPENDENT=Y}                         => {CASE_STATUS=CERTIFIED} 0.2649860 0.9936752 1.007359 29379
[3] {SOC_NAME=SOFTWARE DEVELOPERS, APPLICATIONS} => {CASE_STATUS=CERTIFIED} 0.3072608 0.9883083 1.001918 34066
[4] {NAICS_CODE=5415}                         => {CASE_STATUS=CERTIFIED} 0.4337783 0.9900976 1.003732 48093
[5] {IncomeLevel=100k-150k}                   => {CASE_STATUS=CERTIFIED} 0.4408316 0.9880724 1.001679 48875
[6] {PW_WAGE_LEVEL=Level II}                  => {CASE_STATUS=CERTIFIED} 0.4508974 0.9904111 1.004050 49991
[7] {AGENT_REPRESENTING_EMPLOYER=Y}           => {CASE_STATUS=CERTIFIED} 0.8109678 0.9872195 1.000814 89912
```

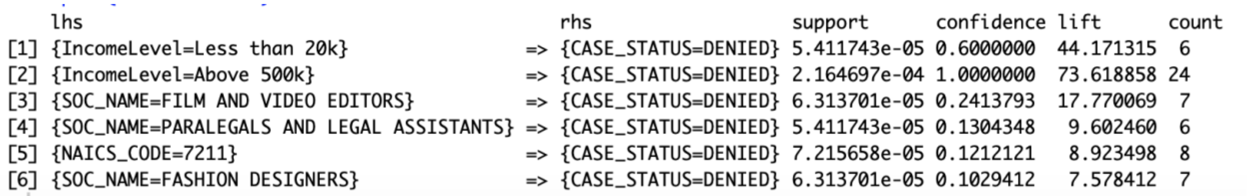
Graph 9: Association rules of applications that employers in NY, with approved case status as rhs



```
    lhs                                              rhs                       support   confidence lift     count
[1] {H1B_DEPENDENT=Y}                             => {CASE_STATUS=CERTIFIED} 0.2001158 0.9938392  1.010172  9679
[2] {SOC_NAME=SOFTWARE DEVELOPERS, APPLICATIONS} => {CASE_STATUS=CERTIFIED} 0.2043749 0.9922706  1.008577  9885
[3] {IncomeLevel=100k-150k}                       => {CASE_STATUS=CERTIFIED} 0.2386751 0.9878487  1.004083 11544
[4] {IncomeLevel=80k-100k}                        => {CASE_STATUS=CERTIFIED} 0.2668762 0.9873786  1.003605 12908
[5] {NAICS_CODE=5415}                             => {CASE_STATUS=CERTIFIED} 0.2705564 0.9907632  1.007045 13086
```

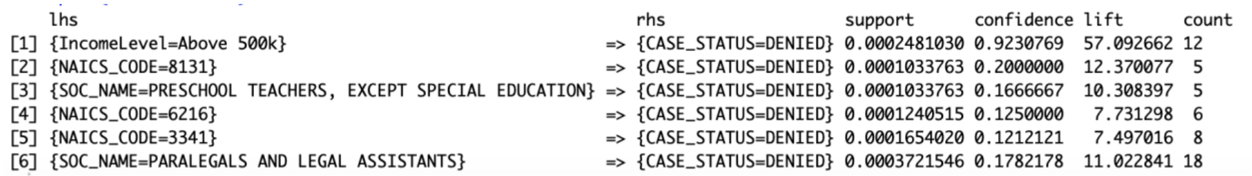
The rules with approved case status as rhs indicate that in CA and NY, the annual income, whether the employer is H1B dependent, the employer's industry and the job tittle are important factors that may contribute to an approved application. Employees with annual salary between $80k and $150k have higher than average probability (98.7%) of having their application approved; applications filed by a H1B dependent employer in CA and NY also have higher than average probability to be approved; employers located in industry with NAICS code as 5415 (Computer Systems Design and Related Services) or employees whose job tittle is software developer may also have almost 100% chance of having the application approved.

Graph 9: Association rules of applications that employers in CA, with denied case status as rhs



```
    lhs                                              rhs                    support      confidence lift      count
[1] {IncomeLevel=Less than 20k}                   => {CASE_STATUS=DENIED} 5.411743e-05 0.6000000  44.171315  6
[2] {IncomeLevel=Above 500k}                      => {CASE_STATUS=DENIED} 2.164697e-04 1.0000000  73.618858 24
[3] {SOC_NAME=FILM AND VIDEO EDITORS}             => {CASE_STATUS=DENIED} 6.313701e-05 0.2413793  17.770069  7
[4] {SOC_NAME=PARALEGALS AND LEGAL ASSISTANTS} => {CASE_STATUS=DENIED} 5.411743e-05 0.1304348   9.602460  6
[5] {NAICS_CODE=7211}                             => {CASE_STATUS=DENIED} 7.215658e-05 0.1212121   8.923498  8
[6] {SOC_NAME=FASHION DESIGNERS}                  => {CASE_STATUS=DENIED} 6.313701e-05 0.1029412   7.578412  7
```

Graph 10: Association rules of applications that employers in NY, with denied case status as rhs



```
    lhs                                                         rhs                    support      confidence lift      count
[1] {IncomeLevel=Above 500k}                                 => {CASE_STATUS=DENIED} 0.0002481030 0.9230769 57.092662 12
[2] {NAICS_CODE=8131}                                        => {CASE_STATUS=DENIED} 0.0001033763 0.2000000 12.370077  5
[3] {SOC_NAME=PRESCHOOL TEACHERS, EXCEPT SPECIAL EDUCATION} => {CASE_STATUS=DENIED} 0.0001033763 0.1666667 10.308397  5
[4] {NAICS_CODE=6216}                                        => {CASE_STATUS=DENIED} 0.0001240515 0.1250000  7.731298  6
[5] {NAICS_CODE=3341}                                        => {CASE_STATUS=DENIED} 0.0001654020 0.1212121  7.497016  8
[6] {SOC_NAME=PARALEGALS AND LEGAL ASSISTANTS}               => {CASE_STATUS=DENIED} 0.0003721546 0.1782178 11.022841 18
```

Table 6: NAICS code and description[3]

NAICS code	Description
7211	Traveler Accommodation
8131	Religious Organizations
6216	Home Health Care Services
3341	Computer and Peripheral Equipment Manufacturing

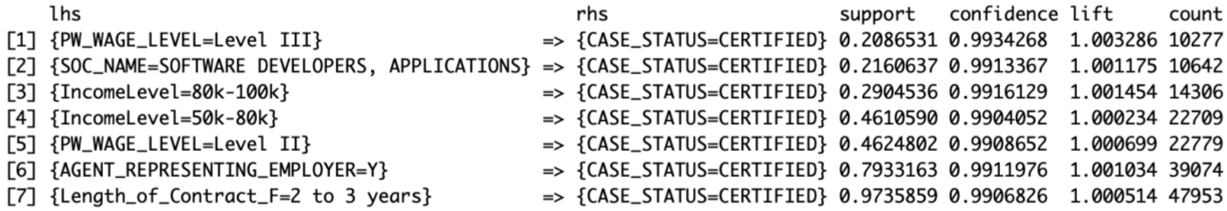
On the other hand, employees with extremely high or low income in CA and NY have almost 100% chance of getting their application denied, which is correspondent with our result in the descriptive analysis. Applications filed by employers in some specific industries such as traveler accommodation, religious organization, home health care services or computer and

[3] https://www.census.gov/programs-surveys/economic-census/guidance/understanding-naics.html

peripheral equipment manufacturing have more than 10% chance of being denied, which is much higher than the baseline 1.3%. Also, for some specific jobs such as film and video editors, paralegals and legal assistants, or fashion designers of companies in CA, and preschool teachers and legal assistants in NY have more than 10% chance of having their applications denied.

Besides targeting application cases with employers located in CA and NY, I also applied association rules machine learning method on application cases with employers located in management, scientific, and technical consulting services industry, to see in this particular industry what kind of employees and employers are the most favored by the government.

Graph 11: Association rules of applications that employers locate in management, scientific, and technical consulting services industry, with approved case status as rhs



```
    lhs                                           rhs                       support   confidence lift     count
[1] {PW_WAGE_LEVEL=Level III}                  => {CASE_STATUS=CERTIFIED} 0.2086531 0.9934268  1.003286 10277
[2] {SOC_NAME=SOFTWARE DEVELOPERS, APPLICATIONS} => {CASE_STATUS=CERTIFIED} 0.2160637 0.9913367  1.001175 10642
[3] {IncomeLevel=80k-100k}                     => {CASE_STATUS=CERTIFIED} 0.2904536 0.9916129  1.001454 14306
[4] {IncomeLevel=50k-80k}                      => {CASE_STATUS=CERTIFIED} 0.4610590 0.9904052  1.000234 22709
[5] {PW_WAGE_LEVEL=Level II}                   => {CASE_STATUS=CERTIFIED} 0.4624802 0.9908652  1.000699 22779
[6] {AGENT_REPRESENTING_EMPLOYER=Y}            => {CASE_STATUS=CERTIFIED} 0.7933163 0.9911976  1.001034 39074
[7] {Length_of_Contract_F=2 to 3 years}        => {CASE_STATUS=CERTIFIED} 0.9735859 0.9906826  1.000514 47953
```

I only set the right hand side as approved case status and found that similar to the previous Arules task, employees with $80k to $150k income level and software developers have the absolute high chance to get their application approved. Besides, employees in the second or the third prevailing wage level, or whose contract length is 2-3 years also have almost 100% probability of being approved.

3. Classification

According to the results of descriptive analysis and association rules, I selected 7 attributes to predict the application status, they are: whether the employer is represented by an agent or attorney; the employee's occupational name; the industry code associated with the employer requesting permanent labor condition; the prevailing wage level; whether the employer is H-1B Dependent; the length of employee's labor contract and the employee's annual salary.

Since the baseline ratio of approved cases verses denied cases is 98.7%, it's hard to evaluate the accuracy of the classification model given such high baseline ratio. Also, under the situation that the majority of application cases are approved, we care more about how well the model can predict the potential denied case so that we can avoid having the potential denied cases.

So I focused only on the application cases filed by employers who locate in industries with approved rate lower than 95%. I tried different classification models including Naïve Bayes, SVMs, Knn, Decision Tree and Random Forest to predict the application case status, and evaluate the model's accuracy by 5-fold cross validation and hold-out test.

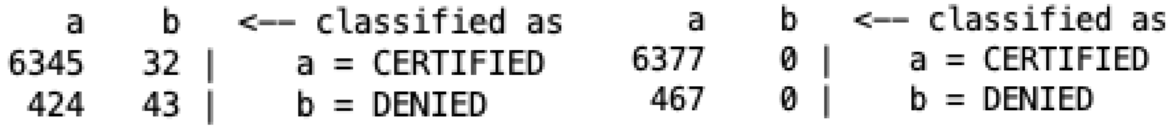
Table 7: classification model accuracies of different models

	5-fold CV	Hold-out test (1:2)
Baseline	93.2%	91.5%
Naïve Bayes	92.3%	92.7%
SVMs(C=10)	93.1%	93.5%
Knn(K=25)	93.5%	93.5%
Decision tree	93.2%	91.5%
Random forest	92.2%	92.5%

The baseline ratio for the cross-validation evaluation is 93.2% and for the hold out test is 91.5%. comparing the accuracy of different models, the Knn model performed the best in both cross-validation and hold-out test that its prediction accuracy is higher than the baseline. Other models such as SVMs and Naïve Bayes did well in the hold-out test but their accuracies in the cross-validation test are below the baseline.

Although the Knn model has the highest prediction accuracy, the accuracy value is only a little above the baseline. According to the confusion matrices, 2 main errors that the models have which caused the relative low accuracy comparing to the baseline:

Graph 12: confusion matrices of classification models



```
   a    b   <-- classified as
6345   32 |    a = CERTIFIED
 424   43 |    b = DENIED
```

```
   a    b   <-- classified as
6377    0 |    a = CERTIFIED
 467    0 |    b = DENIED
```

1. The first is that the models can’t correctly classify the denied application cases that the accuracy of predicting the denied cases is less than 10%, as shown in one of the confusion matrices on the left above.
2. The second is that some models failed to classify the denied cases that they classify all the applications to be approved, as shown in the confusion matrix on the right above.

Conclusion

As for conclusion, the results of descriptive analysis and association rules support our hypothesis that the employee’s income level, the contract length of the employee, the employee’s job tittle, the employer’s H1B dependency and its industry can affect the LCA status. Software developers whose income locates in the range of $80k-$150k, and work in computer systems design or related services industry are the most welcomed foreign labors in the U.S. that have the most chance to get their LCA approved.

Employers in California, Texas and New York filed the most applications and received the greatest number of approved applications in the year 2017, but it doesn’t mean that everyone should consider moving to these 3 states to find a job. Foreigners who want to work in traveler accommodation, religious organization, home health care services or computer and peripheral equipment manufacturing industry are not recommended to apply their jobs in the 3 states. For people who want to work in management, scientific, and technical consulting services industry are highly recommended to apply jobs in Pennsylvania, Illinois and Florida because these 3 states have the most application cases and the approved rate is higher than the average baseline.