International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 08 | Aug 2018 www.irjet.net p-ISSN: 2395-0072

Predicting filed H1-B Visa Petitions' Status

Darshit A. Pandya

***Abstract*— The H1-B visa in United States allows employers to employ foreign workers in speciality occupations. The report addresses the approach to predict the case status of the filed H1-B Visa petitions using various data such as employer name, job category, job title, location of job, filing year, and prevailing wage. The trained models help to relate the decision with the attributes of the application. As the H1-B visa category is one of the highly coveted ones, this approach can be used by both the individual and the employer in between applying for the visa and getting the final decision to be informed of the outcome before it occurs.

***Keywords*— H1B, status, classification, evaluation, analysis**

1. INTRODUCTION

The US H1-B visa is a non-immigrant visa that allows US companies to employ graduate level workers in specialty occupations that require theoretical or technical expertise in specialized fields such as IT, finance, accounting, architecture, engineering, mathematics, science, medicine, etc. This is one of the highly used visa categories, and companies that usually require foreign talent rely on it to a great extent. The growth of IT, Research & Development and various sectors affecting US economy has forced US established firms to hire foreign talent and hence the rate of H1-B visa petition filing has increase substantially.

This report shows an efficient approach to solve the problem of foreseeing the decision before filing or after filing, and before receiving the decision on the filed petition. The report initially highlights the analysis of the available data by showing the relationships between different attributes through plotting. In the next sections, the following have been described in brief: literature survey, data cleaning, feature selection and creation, application of different classification models on the available data and its evaluation.

2. PROBLEM FORMULATION

The report tries to show the dependency of the decision on the attributes of the application. The attributes of the application here serve as an input and the output is the predicted decision. The data used here has the below meta-data:

- **Name of the employer**: Name of employer submitting labor condition application.
- **Category of the job or SOC Name**: Occupational name associated with the requested job under temporary labor condition. Standard Occupational Classification system defines the codes and the names associated with them.
- **Job title**: the requested job title in the petition.
- **Employment Type**: Full time employment (Y) or a part-time employment (N).
- **Year of Filing**: Year when petition is filed (in between 2011-2016).
- **Prevailing wage**: Prevailing Wage for the job being requested for temporary labor condition.
- **Location of work**: State information of the foreign worker's intended area of employment.

Output can be any of the two values: 1. Certified, 2) Denied. The data used for the problem has been imported from Kaggle[5] and the raw form of it held 3 million records in form of filed petitions. The distribution is in accordance to case status:

TABLE 1. DISTRIBUTION OF STATUS LABELS

CERTIFIED	2615623
CERTIFIED-WITHDRAWN	202659
DENIED	94346
WITHDRAWN	89799
PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED	15
REJECTED	2
INVALIDATED	1

3. SYSTEM DESIGN

A. Pre-Data Analysis

Considering the original data of the petitions before pre-processing, we have crunched the numbers to find out the trends and the jobs for which highest foreign labor is required.

1. *Number of Applications per year* - We divided the data as per the filing year and counted the number

of petitions filed to discover the rise in number of applications per year.

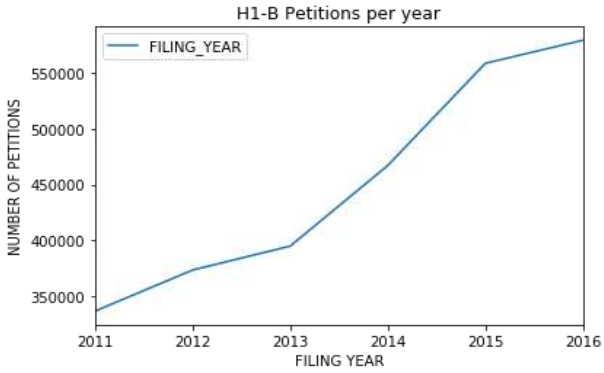

Fig 1. Number of Applications v/s Filing year

2. *Acceptance rate per year* – There are two kind of case status result: 1) Certified, 2) Denied. To find the acceptance, we calculated the ratio of certified petitions by the number of applications for that year.

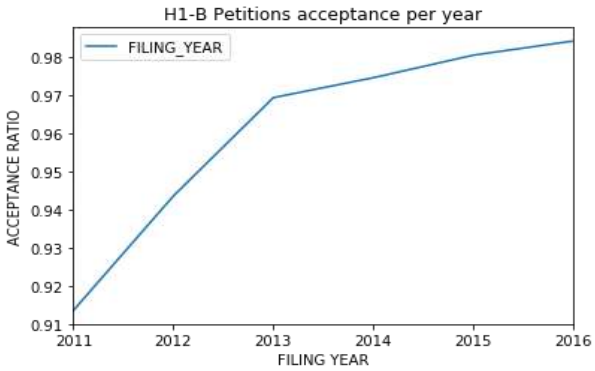

Fig 2. Number of Certified Samples v/s Filing year

3. *Top 15 employers by number of applications* – As per the number of petitions filed by each of the employers, the top 15 are:

TABLE 2. TOP 15 EMPLOYERS AND THEIR APPLICATION COUNT

INFOSYS LIMITED	130241
TATA CONSULTANCY SERVICES LIMITED	64358
WIPRO LIMITED	43679
DELOITTE CONSULTING LLP	36667
ACCENTURE LLP	32983
IBM INDIA PRIVATE LIMITED	28166
MICROSOFT CORPORATION	22373
HCL AMERICA, INC.	22330
ERNST & YOUNG U.S. LLP	18217
LARSEN & TOUBRO INFOTECH LIMITED	16724
CAPGEMINI AMERICA INC	16032
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION	15448
GOOGLE INC.	12545
IGATE TECHNOLOGIES INC.	12196

4. *Top 15 Job categories in demand* – We analysed the top job category for which highest number of H1-B visa petitions filed till date. It highlights the jobs in demand in States.

TABLE 3. TOP 15 JOB CATEGORIES AND ITS APPLICATION COUNT

COMPUTER PROGRAMMERS	372124
COMPUTER OCCUPATIONS, ALL OTHER	360575
COMPUTER SYSTEMS ANALYSTS	164659
SOFTWARE DEVELOPERS, SYSTEMS SOFTWARE	469300
MANAGEMENT ANALYSTS	75806
ACCOUNTANTS AND AUDITORS	62096
FINANCIAL ANALYSTS	49780
MECHANICAL ENGINEERS	46730
NETWORK AND COMPUTER SYSTEMS ADMINISTRATORS	39844
DATABASE ADMINISTRATORS	36219
MARKET RESEARCH ANALYSTS AND MARKETING SPECIALISTS	35303
ELECTRONICS ENGINEERS, EXCEPT COMPUTER	34433
PHYSICIANS AND SURGEONS, ALL OTHER	31782
OPERATIONS RESEARCH ANALYSTS	30641

B. Data Pre-processing and Cleaning

1. *Data Cleaning and Filtering:* The raw data considered for the problem statement needs some pre-processing and cleaning depending on its attributes' value. Firstly, all the records with NULL

or "N/A" value in either of the attributes were pruned as they can't be handled with any random values. The problem discusses about the decision prediction as either "CERTIFIED" or "DENIED", so all the records containing value as "CERTIFIED" or "DENIED" in their case status were considered for further steps.

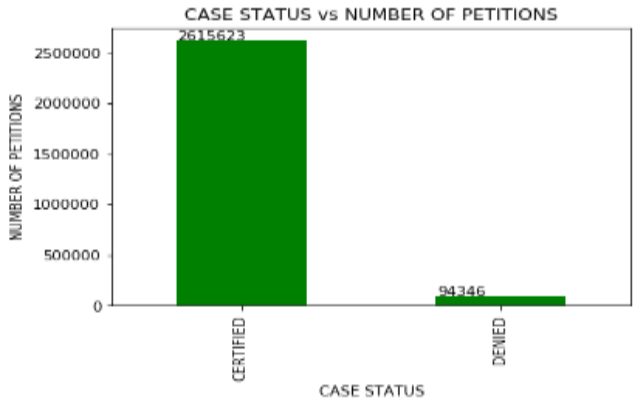

Fig 3. Case Status v/s number of petitions

2. *Data Down Sampling:* As visible, the data is highly imbalanced as the samples with certified status are far more than with the ones with denied status. Hence, data balancing becomes the most important step. To match the number of samples, we down sampled the certified samples to the count of denied samples.

After down sampling, the number of total records are reduced to odd 2,50,000 rows. The distribution turns out to be as below:

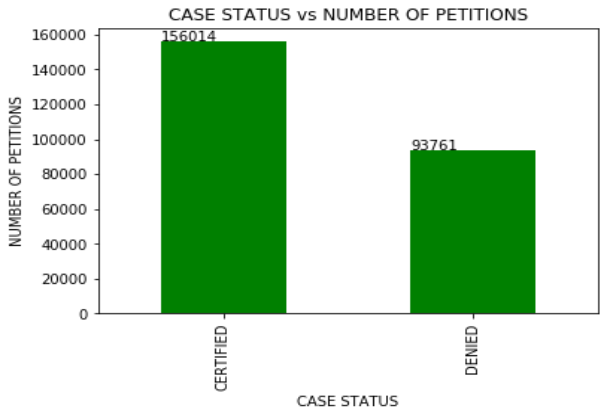

Fig 4. (Down sampled) Case Status v/s number of petitions

3. *Data Consistency through Referencing:* Considering the down sampled data, in SOC names, the data has values like “Computer Systems Analysts” and “Computer Systems Analysts” and many more, which are actually one and the same. Hence, we took the reference of the job categories as an external file input and the job categories in data were mapped and refined using the ratio threshold of string match as 0.94. The threshold was derived after calculating the mean of the string match ratio of various random almost similar strings.

C. Data Conversion and Features Extraction

1. *Categorical to Numerical Values:* For applying the classification models on the pre-processed data, we need to convert the categorical variables to numerical values. There are three standard methods:

- Encode to Ordinal Values
- Feature Hashing
- One Hot encoding

As the values might get interpreted as levels, we avoided encoding to ordinal values. In the data, there are chances that after applying hash function on the values of the attributes, they might be converted into the same values. Hence, we decided to use the one-hot encoding approach to convert the categorical values to numerical.

a. *Data Analysis*: Looking at the unique values of the attributes in the below table, it is difficult to convert them to numerical values using one-hot encoding. Hence, we decided to partition the values into categories according to their values.

TABLE 4. UNIQUE VALUE COUNT OF ATTRIBUTES

Case Status	2
Unique Employment Type	2
Unique Filing Year	6
Unique Worksite State	53
Unique SOCs	925
Unique Job Titles	53272
Unique Employers	80566

b. Feature Creation:

- Case Status: As it has only 2 unique values, it will be easy to apply one-hot encoding
- Wage Category: It has huge number of unique values, we decided it to divide into 5 categories. The distribution is shown in the below table

TABLE 5. WAGES SPLIT INTO CATEGORIES

Low Range	High Range	Category
0	50000	VERY LOW
50000	70000	LOW
70000	90000	MEDIUM
90000	150000	HIGH
150000	200000	VERY HIGH

- Employer name, SOC Name and Job Title: The number of unique employers, job categories and job titles in the down sampled data are huge, hence we analysed the data and created a feature named EMPLOYER_ACCEPTANCE, SOC_ACCEPTANCE and JOB_ACCEPTANCE respectively. To calculate the ratio values, we found the ratio of the certified samples vs the total samples of each kind. Further, depending upon the values of the ratio obtained, they were divided into 5 categories as shown in below table.

TABLE 6. RATIO SPLIT INTO CATEGORIES

Low Range	High Range	Category
0	0.2	VERY LOW ACCEPTANCE
0.2	0.4	LOW ACCEPTANCE
0.4	0.6	MEDIUM ACCEPTANCE
0.6	0.8	HIGH ACCEPTANCE
0.8	1	VERY HIGH ACCEPTANCE

- Employment Type: This attribute has only two values. Y for Full-time employment and N for Part-Time employment.

- Worksite: Originally, the worksite attribute contains the data in city name, state name format. We extracted only the states out of the column values resulting into 53 unique state values which is manageable with one-hot encoding.

- Filing year: Data collection is for the petitions filed between 2011-2016(inclusive), hence 6 unique values.

c. One-Hot Encoding: Case status, employer acceptance category, SOC acceptance category, job title acceptance category, filing year, wage category, employment type, worksite are passed as the column names in the one-hot encoding function and after application, in total 73 columns are created.

2. Feature Selection: Relative effect of the attributes' values on the final decision making, a feature elimination model called Recursive Feature Elimination(RFE) was employed. In RFE, weights are assigned to the features and then least important features are eliminated recursively from the current set of features until the desired number of features are eventually reached.

Considering the output of RFE, attributes employer acceptance level, job acceptance level, wage category, worksite state and filing year. Top attributes selected are: {'EMPLOYER', 'FILING', 'FULL', 'JOB', 'SOC', 'WAGE', 'WORKSITE'}

Hence, employer acceptance category, filing year, job title acceptance category, wage category, work state, job type, soc acceptance category is considered further.

3. EXPERIMENTAL EVALUATION

A. Data Splitting

In classification, we divide the data into two sets viz. 1. Training Data 2. Test Data. We considered 20 percent of the total records for test data and the rest of 80 percent for training the classifiers.

B. Models Applied

We trzined total 7 different classifiers for predication task and evaluated each based on the evaluation metrics.

- **Decision Trees**: In this method of predictive modelling, the data is split into smaller chunks. It uses branching method to create a branch for every possible decision based on the input values of the attributes.

- **k-Nearest Neighbour(k-NN)**: In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). We did parameter tuning to find the best value of k which turned out to be ___

- **Support Vector Machine(SVM)**: SVMs maximize the margin around the separating hyperplane. The decision function is fully specified by a usually a small subset of training samples, the support vectors. Support

vectors are the data points that lie nearest to the decision surface.

- **Logistic Regression:** Logistic Regression(LR) is a predictive analysis model which is a regression analysis when the dependent variable is binary. It models the probability of the dependent variable using the combination of the values of the independent variables.

- **Random Forest**: Random Forest is an ensemble learning method for classification which can be visualised a bunch of decision trees. Unlike decision trees, it solves the main disadvantage of decision tree of overfitting their training data. We considered the number of estimators as 75 and the random state value as 42 to get the best optimum result.

- **Neural Network**: An Artificial Neural network are very powerful brain-inspired computational models. Neural networks, have remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. As the model in our case converges easily, only 5 hidden layers were used with 20 neurons in each of the layers. Additionally, the number of iterations after some point doesn't change the metrics and hence the optimum value of 1000 iterations was used.

- **Gaussian Naive Bayes**: Gaussian Naive bayes is a simple probabilistic classifier model that converges quickly which assumes that there exists features distribution which are independent of each other.

C. Model Evaluation Metrics

All the classifier models were evaluated using three types of metrices.

- **Precision Score**: Precision Score is calculated as the ratio of correctly predicted positive observations to the total predicted positive observations.

  Precision = TP/ TP + FP

- **Recall Score**: Recall is calculated as the ratio of correctly predicted positive observations to the all observations in actual class – yes.

  Recall = TP/TP+FN

- **F1 Score**: F1-Score is calculated as a weighted average of both precision and recall score.

  F1 Score=2*(Recall*Precision)/ (Recall + Precision)

Here,

FP = False Positive: Predicted Yes when Actual is No

FN = False Negative: Predicted No when Actual is Yes

TP = True Positive: Predicted Yes when Actual Yes

TN = True Negative: Predicted No when Actual No

D. Model Evaluation Results

After applying the described models on the data with certain parameter values, the metrics score for each of the model are as below

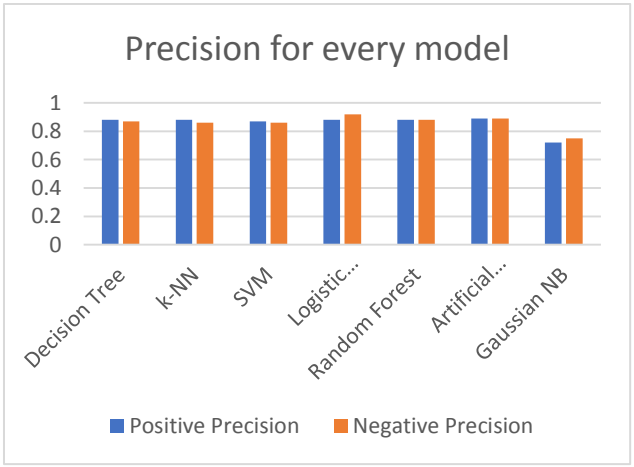

Fig. 5. Classification Model v/s achieved precision score

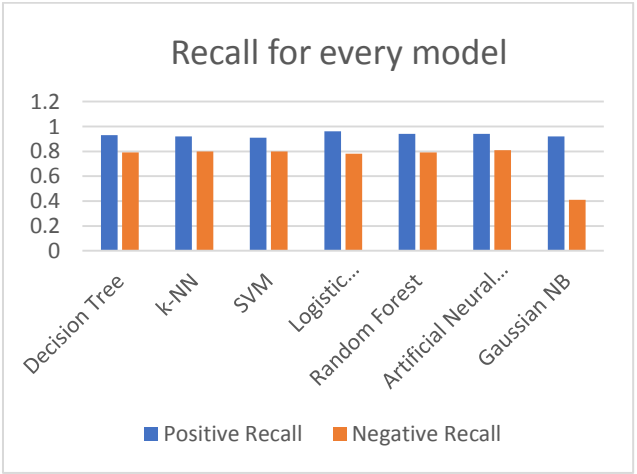

Fig. 6. Classification Model v/s achieved recall score

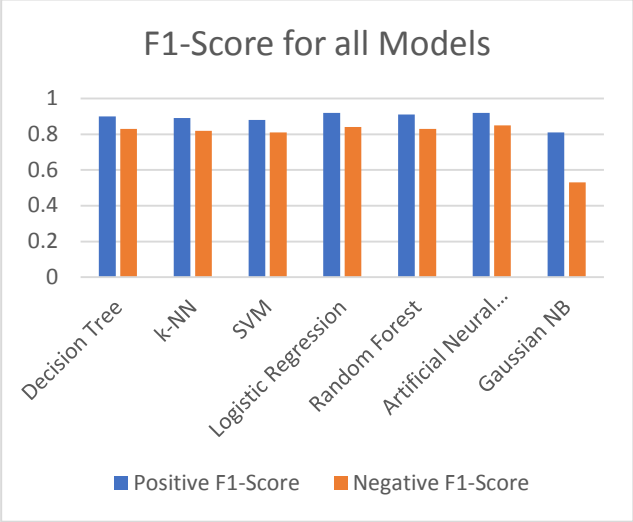

Fig. 7. Classification Model v/s achieved F1- score

4. LITERATURE SURVEY

- There has not been noticeable work done in the field of predicting the case status of H1-B visa petitions. A similar project has been done at UC, San Diego for predicting the decision of the file H1-B visa petition. A project done by the students of UC Berkley [5] tried

  to predict the waiting time to get a work visa for a

  given job title and for a given employer. They used

  K-Nearest Neighbors as the primary model to predict

  'Quickest Certification Rate' across both occupations

  and companies.

- Compared to all approaches presented till date, we have achieved the highest accuracy of prediction task through application of Artificial Neural Network and Logistic Regression along with Random Forest.

CONCLUSION

Analysing the above charts, it is evitable that for the particular problem statement, Logistic Regression and Artificial Neural Network classifier gives the best positive F1-Score of **92**% and negative F1-Score of **84**% and **85**% respectively. Data balancing played a major role in achieving the high accuracy for both the classes. We also infer that job title, employer name/employer acceptance ratio, wages, worksite and filing year play an important role in inferring the value of the case status.

REFERENCES

[1] H-1B ICERT LCA. PDF. Washington, DC: U.S. Department of Labor, 2016.https://www.foreignlaborcert.doleta.gov/docs/Performance_Data/Disclosure/FY15-FY16/H-1B_FY16_Record_Layout.pdf

[2] "OFLC Performance Data." OFLC Performance Data. April 12, 2018. Accessed April 28, 2018. https://www.foreignlaborcert.doleta.gov/performancedata.cfm.

[3] "2018 SOC System." Standard Occupational Classification. 2018. Accessed April 28, 2018. https://www.bls.gov/soc/2018/home.htm.

[4] "North American Industry Classification System (NAICS) Main Page." North American Industry Classification System. May 15, 2012. Accessed April 28, 2018. https://www.census.gov/eos/www/naics/.

[5] R. Zaman, Arif Uz. "H1-B VISA Applications." H1-B VISA Applications. June 14, 2017. Accessed April 28, 2018. https://www.kaggle.com/mazaman/h1-b-visa-applications/data.

[6] Dan, Lucas, Samuel Kabue, and Sarah Neff. Project Alien Worker. PDF. Berkeley: UC Berkeley School of Information, April 2016.

https://www.ischool.berkeley.edu/sites/default/files/projects/project_alien_worker_-_final.pdf

[7] Wilson, Jill H. "Immigration Facts: Temporary Foreign Workers." Brookings. August 02, 2016. Accessed April 28, 2018.

[8] https://www.brookings.edu/research/immigration-facts-temporary-foreign-workers/.

[9] Costa, Daniel, and Jennifer Rosenbaum. "Temporary Foreign Workers by the Numbers: New Estimates by Visa Classification." Economic Policy Institute. March 7, 2017. Accessed April 28, 2018. https://www.epi.org/publication/temporary-foreign-workers-by-the-numbers-new-estimates-by-visa-classification/.

[10] Rapoza, Kenneth. "No. 1: Infosys - Pg.12." Forbes. January 25, 2015. Accessed April 28, 2018. https://www.forbes.com/pictures/eglg45heklg/no-1-infosys-15/#4b6485493f2c.

[11] Nowrasteh, Alex. H-1B Visas: A Case for Open Immigration of Highly Skilled Foreign Workers. PDF. Washington, DC: Competitive Enterprise Institute, October 2010.

[12] http://www.cei.org/sites/default/files/Alex Nowrasteh - H1-B Visas.pdf

[13] Hadar, Yonatan. "Using Categorical Data in Machine Learning with Python." YellowBlog. September 19, 2017. Accessed April 28, 2018. https://blog.myyellowroad.com/using-categorical-data-in-machine-learning-with-python-from-dummy-variables-to-deep-category-66041f734512?gi=cfa270f9aed

WORK DIVISION

The below table shows the work division amongst the team members:

TABLE 7. TASK DIVISION

Name	Tasks
Darshit Pandya	Data collection, Preprocessing and Feature creation, Training Classifiers
Julya Matsakyan	Feature elimination and Data Splitting
Mansi Patel	Model Evaluation

LEARNING EXPERIENCES

As a team, we learnt that

- Data balancing is an important step when the data is imbalanced else the results will be inclined towards the class with larger number of records
- Categorising feature values helps to draw meaning out of the categorical variables
- Artificial Neural Network's performance reduces if the number of hidden layers are comparatively high and the data is not complex. Because, it converges easily and then the rest of layers adds unnecessary overhead.