ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

A Predictive Model for H1-B Visa Petition Approval

Madhana Sohan Kumar[1], N Naresh[2]

P.G. Scholar, Department of CS & SE, College of Engineering (A), Andhra University, Visakhapatnam, India[1]

Research Scholar, Department of CS & SE, College of Engineering (A), Andhra University, Visakhapatnam, India[2]

ABSTRACT: In this paper, we aim to predict the outcome of H1-B visa petitions filed for many highly skilled foreign nationals every year in US by their employers for speciality jobs. Over 2 Million visa petitions are filed by the employers each year and only 65000 petitions are approved. So, the goal is to explore the petitions filed and their outcomes for past six years i.e., from 2011 to 2016 and to find a pattern to predict the outcome by using a predictive model developed using Machine Learning techniques.

KEYWORDS: Classification model, Feature Selection, Parameter tuning, Resampling.

I. INTRODUCTION

H1-B Visa is one type of non-immigrant temporary visa granted by USCIS (United States Citizenship and Immigration Service) for the foreign nationals. These petitions are filed by the employers for their employees. This visa is also filed by international students after they get admissions into Universities. Since the number of applicants is very large than the number of selections and as the selection process is claimed to be as lottery there is no insight of how the attributes have influence over the outcome. So, we believe that a predictive model generated using all the past data can be a useful resource to predict the outcome for the applicants and the sponsors.

II. LITERATURE SURVEY

The dataset that we are studying is available on Kaggle under the name 'H-1B Visa Petitions 2011-2016 dataset' which is processed dataset from the original data available on Office of Foreign Labor Certification (OFLC) website after performing various data transformations on the data. From data analysis performed on this data allow us to finding top Occupations, States, Employers and Industries that contribute to highest number of H1B visa application.
Our research says that some independent analysis performed on this data provided some insightful facts and also to predict some details. Astudy by Andrew Shikair explored a way to predict wages of the visa recipients by performing text analysis of application attributes and their study concludes that Occupational Classification and Job Title were the two most important fields to predict the applicants wage as accurately as possible. A project done by the students of UC Berkley aims to predict the waiting time to get a work visa for a given job title and for a given employer. They used K-Nearest Neighbours as the primary model to predict 'Quickest Certification Rate' across both occupations and companies. Although some studies provide some insightful facts about our data, my study is to predict the final outcome of the applicants by determining the influence of the attributes over the outcome.

III. DATASET

This dataset is downloaded from the Kaggle which has 9 features and 1 feature containing the class label. The total number of records available for us is more than 3 million points. The features provide the following information about our samples.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

- EMPLOYER_NAME: Name of employer submitting application.
- SOC_NAME: Occupational name associated with the SOC CODE which is an occupational code associated with the job being requested for temporary labour condition, as classified by the Standard Occupational Classification (SOC) System.
- JOB_TITLE: Title of the job
- FULL_TIME_POSTION: There are 2 categories for this feature: Y= Full time position and N = Part Time Position
- PREWAILING WAGE: the average wage paid to employees with similar qualifications in the intended area of employment.
- FILING_YEAR: The year of filing the petition
- WORKSITE: City and state of the applicant's job.
- LONGITUDE & LATITUDE: Exact geographical location of the worksite.

The 1 label in the dataset is divided into 7 classes: (1) CERTIFIED (2) CERTIFIED-WITHDRAWN (3) DENIED (4) WITHDRAWN (5) PENDING QUALITY AND COMPLIANCE REVIEW – UNASSIGNED(6) REJECTED (7) INVALIDATED

The distribution of labels is shown below (Table 1):

Label	Number of applications
CERTIFIED	2600241
CERTIFIED-WITHDRAWN	201479
DENIED	93761
WITHDRAWN	89110
PENDING QUALITY AND COMPLIANCE REVIEW – UNASSIGNED	15
REJECTED	2
INVALIDATED	1

Table 1 Distribution of Labels of the dataset

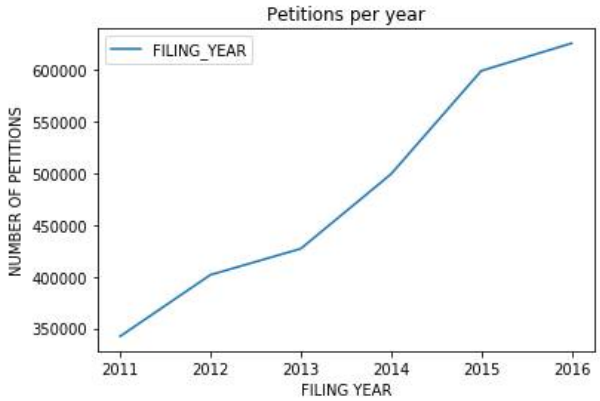

Fig. 1. Petitions per Year

Fig.1 here shows the number of H1-B Visa petitions filed each year. The petitions filed per year has highly increased from 2011 to 2016 has approximately doubled its number.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

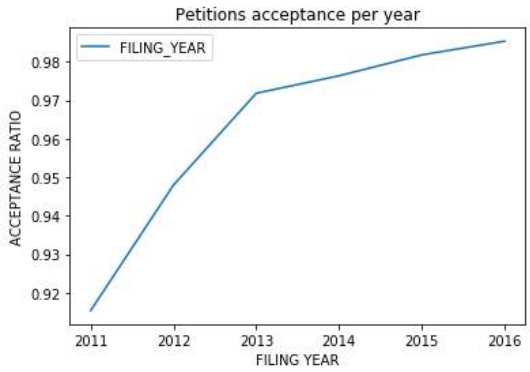

Fig. 2. Petitions acceptance per year

Fig. 2 represents the total percent of the filed Petitions are accepted each year.

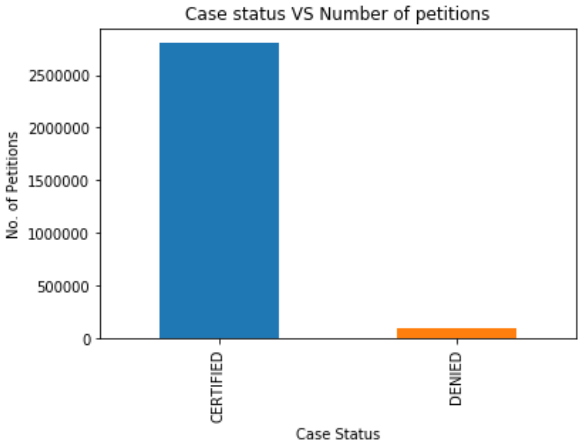

Fig. 3 Case Status in an imbalanced dataset

As can be seen from the Fig. 3, the dataset is highly imbalanced, we apply some pre-processing techniques to create a dataset which has more relevant data to generate a model leaving all the noise in the data. We performed exploratory data analysis to get some facts which the data provides us and basing on them we considered the relevance of relationship among the features and accordingly discarded few features and created some to remove the redundant information.

We observed that features 'LONGITUDE' and 'LATITUDE' have missing values nearly 100000 points hence we remove both the features entirely. Also, we transformed few features into new features. The features of our final dataset after transformation is:

1. Case Status: All the cases 'WITHDRAWN' are excluded completely as these are the decisions made by either the employer or the applicant.
2. Employer Acceptance: Total number of certified petitions / Total number of petitions filed by the Employer.
3. Wages Category:
4. SOC_Acceptance:Total number of certified petitions / Total number of petitions filed in the name of a SOC
5. Job_Acceptance: Total number of certified petitions / Total number of petitions filed by the Employer
6. Filing Year: The year of filing.
7. Worksite State: worksite of the applicant's job
8. Employment Type: Full-time employee or Part-time employee.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

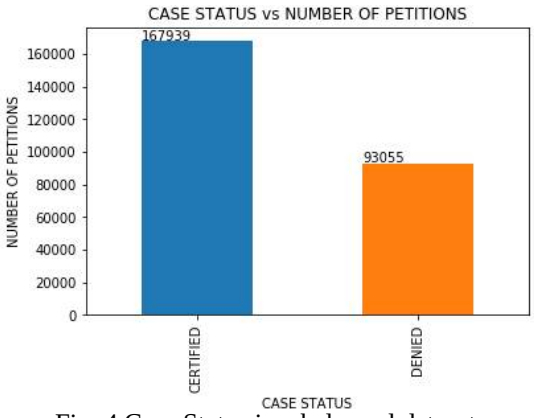

Fig. 4 Case Status in a balanced dataset

Fig. 4 shows that the data pre-processing result is good i.e., the ratio of two classes, CERTIFIED and DENIED, has increased from 98:2 to 65:35 in imbalanced and balanced dataset.

IV. METHODS

After all the pre-processing steps we performed on our dataset to get the final transformed dataset we split the data into train and test data in 80:20 ratio. One-Hot encoding is applied on the final dataset. For the prediction task we use these 4 classifiers. They are Gaussian Naïve Bayes Classifier, Random-Forest classifier and XG-Boost.

1. ***Naïve Bayes:*** Naive Bayes is a simple and interpretable model which assumes all features are conditionally independent given labels and are in gaussian distribution. It calculates $P(\frac{x}{y} = 1)$, $P(\frac{x}{y} = 0)$ and $P(y)$ by taking their maximum likelihood estimates in the joint likelihood of the data. While making a prediction, it considers both $P(y = 1)$ and $P(y = 0)$ on the Bayes rule and compares the two.
2. ***Random-Forest:*** It is an ensemble technique which uses bagging technique. It uses number of meta-classifiers on various sub samples of the dataset and then averages the prediction to improve the final predictive outcome. This classifier can also control overfitting by proper parameter tuning.
3. ***XG-Boost:*** XGBoost or Extreme Gradient Boost algorithm is an ensemble method. It uses 'Bagging and Boosting' techniques. In Bagging technique, trees are grown to their maximum extent and Boosting techniques uses trees with fewer splits. On aggregation of the two models, the final model gives us the outcome with less MSE (Mean Squared Error).

V. EVALUATION AND RESULTS

After the 80:20 split of data into train and test data. We applied all the 3 classifiers on the data to predict the outcome. While using Naïve-Bayes algorithm we used the default version of it without changing any hyper-parameter tuning, but while using random forest and XG-Boost algorithms we have done some parameter tuning to get better results. In Random Forest algorithm, the n-estimators is 400, max_depth to 15 and random state value is 50. we used the default method 'Gini' index for the calculation of splits. While tuning XG-Boost, we used binary: logistic function. We here used n-estimators 100.

CLASSIFIER	PRECISION SCORE	RECALL	F1-SCORE	ACCURACY
Naïve-Bayes	0.9050445103857567	0.6256072546691137	0.7398187156900294	0.8438284258319125
Random Forest	0.8776365780165073	0.723199827269783	0.7929687499999	0.865974443954865
XG-Boost	0.9001343183344527	0.7234697182338335	0.8021905673928658	0.8733692216325983

Table 2 Evaluation Results

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

We consistently have high recalls over di□erent models, which shows that our algorithm doesn't have many false negatives. However, our precision metric is relatively lower across di□erent methods, which demonstrates low false positives rates.

VI. CONCLUSION

In this work, Gaussian Naive Bayes, Random Forest Classifier and XGBoost Classifier were considered for determining the status of H1-B visa applications. XGBoost Classifier performed the best in terms of accuracy, precision and F1 score over others. We achieved a best of 87.3369% classification accuracy. But to the surprise Naïve Bayes classifier has done quite a good performance than expected with 84.382% accuracy and 90.50% precision score. This leads to conclusion that how much important is feature selection and feature transformation is. Our results showed that the most predictive features are EMPLOYER SUCCESS RATE and PREVAILING WAGE. One can infer from these results that the chance of being certified increases with the amount of wage and how successful your sponsor was in the previous H1B applications.

REFERENCES

1) Prof. S. Sarkar IIT Kharagpur, Introduction to machine learning NPTEL :"https://www.youtube.com/playlist?list=PLYihddLF-CgYuWNL55Wg8ALkm6u8U7gps"
2) A. Ng, "CS229 Lecture Notes" [Online]. Available:" http://cs229.stanford.edu/notes/. [Accessed: 14-Dec-2017]."
3) An End-to-End guide to understand the Math behind XGBoost:"https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/"
4) Tianqi Chen. XGBoost: "https://github.com/tqchen/xgboost"
5) AurélienGéron,"Hands-OnMachine LearningwithScikit-LearnandTensorFlow", Wiley, 2017.
6) Dataset: "https://www.kaggle.com/nsharan/h-1b-visa"
7) "PredictingCaseStatusof H-1B Visa Petition"[Online]Available:"https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf."
8) H-1B Visa Petitions 2011-2016 — Kaggle. [Online]. Available: https://www.kaggle. com/nsharan/h-1b-visa/data. [Accessed: 20-Oct-2017].
9) "H-1B Visa Data Analysis and Prediction by using K-means Clustering and Decision Tree Algorithms." [Online]. Available: https://github.com/Jinglin-LI/H1B-VisaPrediction-by-Machine-Learning-Algorithm/blob/master/H1B\%20Prediction\%20Research\%20Report.pdf
10) Shubham Panat, "A Study on H1B – VISA Petitions 2011-2016",South Central SAS Users Group Educational Forum 2016, [Online]: "http://www.scsug.org/wp-content/uploads/2017/10/sp04.pdf"
11) "H-1B Visa Data Analysis and Prediction by using K-means Clustering and Decision Tree Algorithms." [Online]. Available: https://github.com/Jinglin-LI/H1B-VisaPrediction-by-Machine-Learning-Algorithm/blob/master/H1B\%20Prediction\%20Research\ %20Report.pdf
12) Stefan Van Der Walt, S. Chris Colbert, Gaël Varoquaux, "The NumPy array: a structure for efficient numerical computation"Computing in Science and Engineering, Institute of Electrical and Electronics Engineers, 2011, 13 (2), pp.22-30.