Predicting the outcome of H-1B visa using ANN algorithm

Raghav Khaterpal, Harit Ahuja, Jatin Goel, Karanveer Singh, Rahul Manoj

Abstract*—H-1B visa allows US employers to employ nonimmi- grant workers on a temporary basis. This visa only allows specialty workers to gain employment in the United States which means people who have a bachelor's degree or equivalent work experience are eligible. The duration of H-1B visa is 3 years and it may extend to 6 years. The H-1B visa is the most sought after visa in the world, however it has a low approval rate. In 2019, 200000 people applied for the visa worldwide of which only 85000 people were selected which means an approval rate of only 42%. This fight to obtain the visa is getting more competitive as the US Economy improves. This selection depends upon a number of factors such as employer, wage etc. This paper helps to predict whether an individual can gain the H1B visa or not taking in account all the relevant factors. The proposed system secured a high accuracy of 96% by using ANN algorithm.

***Index Terms*—ANN, One-hot encoding, H-1B visa, ReLU**

I. INTRODUCTION

The H-1B visa allows only a small proportion of individuals to work in the United States. This type of visa are applied by a number of highly skilled foreign nationals. The visa request is applied by the foreign national's company to the US embassy. So, one year of experience counts as one point. Hence, the user has to collect a total of 12 points in order to qualify for the visa. The chances of getting accepted are very slim due to the restrictions imposed by the United States Government. It has become even more difficult to secure the H-1B visa due to the changes in the procedures and the laws. In this paper, the proposed system plays the role of predicting the corresponding petition issued for a H-1B visa. The selection depends upon a plethora of factors such as wage, employer, work experience, proficiency, field of work and many more. Our proposed model takes into account all such factors and using ANN algorithm predicts the outcome of the visa application. The dataset taken consists of 2 million petitions which allows the model to predict more accurately. The dataset was then cleaned by removing the outliers and then applied one hot encoding to convert the data into numeric form for further processing. Finally, the AUC and ROC values were calculated and the F1 score was determined.

Revised Manuscript Received on May 21, 2020.

* Corresponding author

Raghav Khaterpal, Department of Mechanical Engineering, Thapar Insti- tute if Engineering and Technology, Patiala, India

Harit Ahuja, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

Jatin Goel, Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

Karanveer Singh, Department of Mechanical Engineering, Thapar Institute if Engineering and Technology, Patiala, India

Rahul Manoj, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

II. RELATED WORKS

Prateek and Shweta Karun in [1] implemented different classification algorithms that in turn is used to predict the H-1B visa eligibility once all the features are selected. The dataset that was used was H-1B Visa Petitions 2011-2016 from Kaggle. The corresponding dataset contains around three million petitions for the 6 years. Histograms were implemented that were used to differentiate the data into different labels. During feature extraction, only the important features were taken in order to reduce the processing time. Classification algorithms like decision tree, RF, logit were used in the paper. Finally, the F-score was evaluated to get the final accuracy.

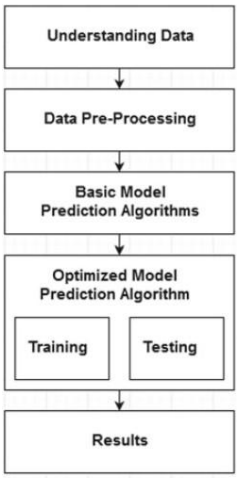

Fig. 1. Prediction Methodology

The paper [2] written by Anay Dombe makes use of machine learning algorithms like neural network to make the corresponding predictions. It takes into account several important factors like the job designation, work location etc. The dataset used in this paper is H-1B Visa Petitions 2011- 2016 which contains approximately 3 million petitions. But for this paper, only one million petitions were extracted in order to improve the efficiency. The dataset was then normalized using the normalization method to bring all the values in the dataset between 0 and 1 using the formula :

$$y = \frac{x - min\{X\}}{max\{X\} - min\{X\}} \quad (1)$$

The final F1 score for logistic regression was 96% and F1 score for neural network came out to be 97%.

Mandeep singh in [3] compared the former machine learning techniques with the new machine learning techniques for a better understanding. The dataset is taken from the Foreign Labor Certification (OLFC) from Kaggle. The dataset contains forty different attributes in the dataset. The next step is the cleaning of the dataset and then the corresponding features are extracted. The training and testing of the of the model is carried out. The results of different machine algorithms like decision tree, c5.0, SVM, Naïve bayes and neural network. The conclusion stated that the cs5.0 had the highest accuracy of 94.71%.

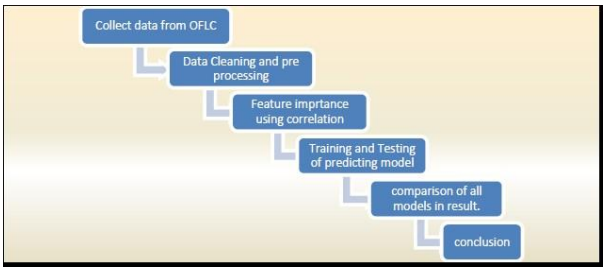

Fig. 2. Methodology

In paper [4], Beliz Gunel predicts the allotment of the H-1B visa applications using a number of different machine learning algorithms that include SVM, Naive Bayes, Logistic regression and Neural network. The dataset used in the paper was the 2011-2016 H-1B dataset which consists of 3 million petitions. Few attributes were extracted from the dataset that were found useful for further processing. SGD Classifier was used that is present in the scikit-library to implement all the machine learning algorithms. For training, 80% of the data was kept aside and 20% of the data was used for testing. In terms of the final result, Neural network was better in performance compared to other algorithms.

	Balanced				Unbalanced			
	Tr. Acc.	Test Acc.	Prc.	Rcl.	Tr. Acc.	Test Acc.	Prc.	Rcl.
Naive Bayes	94%	72%	75%	96.7%	94%	94%	98.6%	95.2%
Log. reg w/cl	98%	74%	72%	99.9%	98%	98%	98.5%	99.2%
Linear SVM w/ElasticNet	98%	78%	75%	99.4%	98%	97%	98.1%	100%
Neural Net.	98%	76%	75%	99.9%	98%	96%	98.2%	100%
Neural Net. w/cl	98%	82%	81%	99.9%	98%	97%	98.5%	100%

Fig. 3. Results

Debrata Swain in [5], implemented a system which calculates the probability of chances of getting a H-1B visa considering few important factors like county of origin, gender, job etc. The dataset used is H-1B dataset 2011-2016 which contains more than 3 million different petitions. Different machine algorithms like K means clustering, Random forest and Logistic Regression were used to train and test the data in order to create a suitable model. Their paper also suggests users to improve their profile in order to get accepted. The final result displays the accuracy of whether the corresponding user will get the approval for H-1B visa or not. The total accuracy of the paper came out to be 86%.

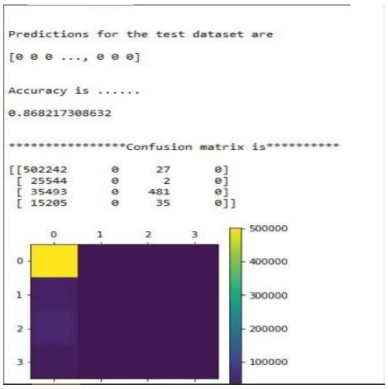

Fig. 4. Accuracy for Model

The paper [6], Dhanasekar Sundararaman implemented a system which predicts whether the H-1B request filed by the corresponding user will be positive or negative. A model was developed which used the machine algorithm called Random Forest which was trained intensively with the cleaned data set. The data set was extracted from the 2016-2017 H-1B data set which contains sufficient entries for the visa for training and testing of the model. The final accuracy that was achieved came out to be 99% when around 100 decision trees were used for the estimation.

III. PROPOSED WORK

A. Overview

In this paper, for predicting the outcome of the approval of H-1B visa, the 2011-2016 H-1B dataset and 2016-17 H-1B dataset is used which contains more than 4 million petitions in total from both the datasets. Histograms were utilized in order to eliminate the outliers. One-hot encoding was used to convert data into appropriate format. Finally, ANN algorithm was used to train the data and predict the final outcome, whether the petition is accepted or not.

B. Dataset

Two datasets were taken, i.e 2011-2016 H-1B dataset and 2016-17 H-1B dataset that were extracted from Kaggle.com. 2 million petitions were cut down from the two datasets. The most relevant factors were taken from all the petitions that included name of the employee, title of the job, position, wage, status of the caste etc.

C. Pre-Processing Data

After the truncation of dataset into relevant data, histograms were plotted to remove the outliers. Outlier is basically a observation point that is distant from other observation point so they might give false accuracy. To detect the outliers, the box plot technique was used. Lastly, the values were normalised so that the values were between the range of 0 and 1.

D. One-hot Encoding

Majority of the factors in the dataset were in text form. The text form needed to be converted to numeric form in order to feed it to the neural network. For example, the factor full time position had two values , i.e Yes or No. So using one-hot encoding, Yes is replaced with the value 1 and No is replaced with the value 0. After all the pre processing data, a total of 27,000 data was resulted. Then 80% of the dataset was kept for training purpose and remaining 20% of the data was kept aside for testing.

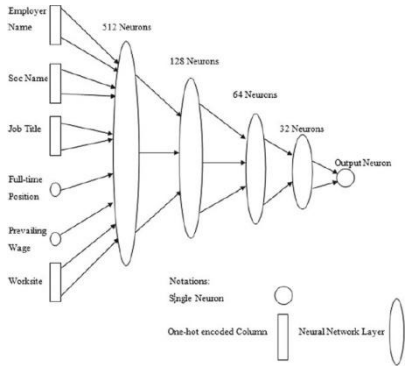

Fig. 6. Architecture of ANN

[!]

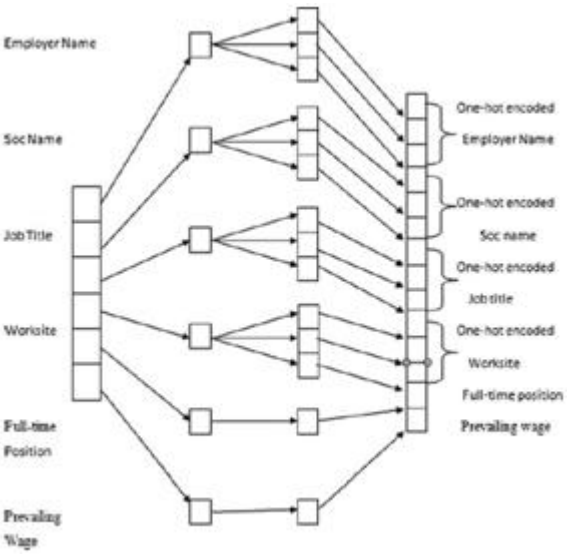

Fig. 5. One-hot encoding pictorial representation

E. Artificial Neural Network (ANN):

The next step in the methodology is to develop a suitable model with the help of ANN algorithm. ANN is relevant for the proposed system as it contains layers that are connected with each other and each layer contains an activation function which converts the input into output. For the next step, the ReLU (rectified linear unit) and MLP (multi layer perceptron) were applied. The output layer is called sigmoid that predicts the final outcome of the visa petition. The final output proba- bility is then approximated to a whole number between 0 and
1. If the final output is 0, then the visa petition is denied and if the value is 1, then the visa petition is approved.

IV. RESULTS

After applying the model was created by using ANN algo- rithm, next step was to apply the testing dataset to calculate the accuracy and determine the final results. Figure [7] shows the training accuracy that is achieved by using the corresponding ANN model. Figure [8] shows the Loss plot for the model that was calculated on 100 epochs. The final accuracy came out to be 98%.

Next, the ROC (Receiver Operating Characteristics) and AUC (Area under Curve) graph was plotted to measure the

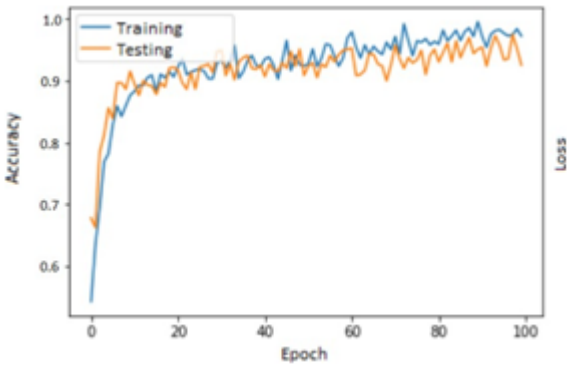

Fig. 7. Training accuracy

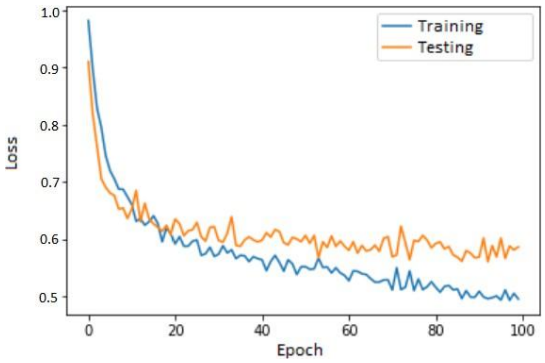

Fig. 8. Loss plot of model

probability of the curve and the degree of separability in Figure [9]. The AUC value came out to be 0.96. Then, finally the confusion matrix and the F1 score was calculated in figure [10]. According to the confusion matrix, the Average precision-recall score was 0.93 and the F1 score was 0.96. From Figure 6, we can also conclude that the False-positive rate is 16.

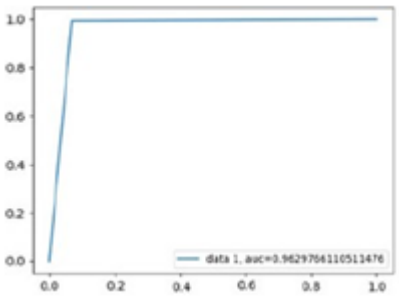

Fig. 9. ROC curve for ANN

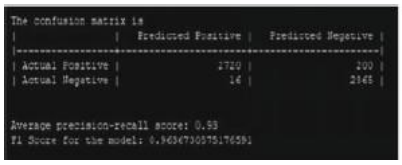

Fig. 10. Confusion Matrix and F1 score

V. CONCLUSION

To conclude, it is indeed possible to predict the outcome of H-1B visa applications based on the attributes of the applicant using machine learning and ANN(artificial Neural Networks). Out of the models we tried, Neural Network outperformed all the other models with 96%. Final accuracy and achieving an overall 0.96 F1 score on the balanced test data. That's likely because artificial neural networks are inherently better at explaining the complexities in the data. This paper further can be treated as a future reference for additional research on H-1B visa. In future, we will be using the recent and updated dataset and develop a more integrated system which will help the corresponding user's profile in order to get the H-1B visa approved.

REFERENCES

1. Karun, S. (2019). Predicting the Outcome of H-1B Visa Eligibility. In Advances in Computer Communication and Computational Sciences (pp. 355-364). Springer, Singapore.
2. Dombe, A., Rewale, R., & Swain, D. (2020). A Deep Learning-Based Approach for Predicting the Outcome of H-1B Visa Application. In Machine Learning and Information Processing (pp. 193-202). Springer, Singapore.
3. Thakur, P., Singh, M., Singh, H., Rana, P. S. (2018). An allotment of H1B work visa in USA using machine learning. International Journal of Engineering Technology, 7(2.27), 93-103.
4. Gunel, B., & Mutlu, O. C. Predicting the Outcome of H-1B Visa Applications.
5. Swain, D., Chakraborty, K., Dombe, A., Ashture, A., & Valakunde, N. (2018, December). Prediction of H1B Visa Using Machine Learning Al- gorithms. In 2018 International Conference on Advanced Computation and Telecommunication (ICACAT) (pp. 1-7). IEEE.
6. Sundararaman, D., Pal, N., & Misraa, A. K. (2017). An analysis of nonimmigrant work visas in the USA using Machine Learning. International Journal of Computer Science and Security (IJCSS), 6.
7. Jing-Lin. H-1B Visa Data Analysis and Prediction by using K-means Clustering and Decision Tree Al- gorithms. [Online] Available: https://github.com/Jinglin- LI/H1B-Visa-Prediction-by- Machine-Learning-Algorithm/blob/master/H1B%20Prediction%20Research%20Report.pdf.
8. Dhanasekar Sundararaman , Nabarun Pal , Aashish Kumar Mis-raa ,(2017),” An analysis of nonimmigrant work visas in the USA using Machine Learning”, International Journal of Computer Science an Security(IJCSS), Vol. 6.
9. Doran, K., Gelber, A. and Isen, A., 2014. The effects of high-skilled immigration policy on firms: Evidence from H-1B visa lotteries (No. w20668). National Bureau of Economic Research.
10. Peri, Giovanni, Shih, Kevin, Sparber, Chad: STEM workers, H-1B visas, and productivity in US cities. J. Labor Econ. 33(S1), S225 S255 (2015).

AUTHOR PROFILE

Raghav Khaterpal is currently pursuing mechanical engineering from Thapar Institute of Engineering and Technology, Patiala. He is passionate about data science and machine learning. He aims to study the various ways in which Data Science can be used to improve existing approaches. He plans to pursue his Masters in Data Science.

Harit Ahuja, is current pursuing his Bacherlor's degree in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai, to be completed in the year 2020. His area of interest includes Artificial Intelligence, Machine Learning, networking and image processing. He is a CCNA ceritified. He represented his college in AUVSI SUAS which was held in Maryland, USA. He takes on any project that involves image processing and deep learning.

Jatin Goel, is currently pursuing his Bachelors Degree in Information Technology and Engineering from SRM Institute of Science and Technology, Chennai, to be completed in the year 2020. His area of interest includes Finance, Cryptocurrency, Data Science and Machine Learning. His self-driven attitude and keen acumen enables him to take on any project that involves Data Science.

Karanveer Singh is currently pursuing mechanical engineering from Thapar Institute of Engineering and Technology, Patiala. He actively takes part in competitive programming contests. He enjoys working Machine Learning projects and is passionate about solving financial problems using it.

Rahul Manoj, is completing his bachelors degree in Computer Science and Engineering from SRM Institute of Science and Technology. He has published various papers in the field of Artificial Intelligence and Stock Market. His interests primary lie in the fields of Artificial Intelligence and Big Data.