Predicting Case Status for Visa Petitions

Dweepa Honnavalli
01FB16ECS138
dweepa.prasad@gmail.com

Ishita Bhandari
01FB16ES143
ishita.bhandari123@gmail.com

Kavya Varma
01FB16ECS162
kavya.varma98@gmail.com

***Abstract*—The paper aims to predict whether a visa applicant to the United States of America is certified or denied a visa. The work presented determines how visa status outcome of three most applied-to visa-categories (H1-B, L1, and F1) are influenced by attributes of user application metadata. The model uses different approaches to predict the outcome of the different classes of visas. The results of this paper can be utilized by visa aspirants to predict the outcome of their petition and get recommended ways of maximizing success of acceptance. We developed a prediction model which trains on historical data of the same application domain (i.e based on which visa the applicant has applied to) and uses machine learning classifiers to predict the outcome of visa petitions. The allotment of most these visas, like H1-B for example, is said to be a lottery system not depending on any specific criteria. Thus, a foreign national is not guaranteed selection for an H1B visa merely because they fulfill all the qualifying criteria. This work aims to use machine learning on known factors to make predictions under this uncertainty, with accuracy better than that achieved using the prior probability.**

***Index Terms*—visa, status, prediction, classification, regression, model, analysis**

I. INTRODUCTION

The current political and socio-economic status of the world makes entering countries like the United States of America, arduous work. Granting visas to foreign nationals has been restricted. It is hard to predict whether a visa application will be accepted or rejected, especially as different visas are applied for different reasons and have different metadata and criteria for acceptance or rejection. *H1-B:* The H-1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1-B visa, a US employer must offer a job and petition for H-1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college / higher education (Masters, PhD) and work in a fulltime position. H-1B visa class is very industry relevant and many individuals and companies rely heavily on this yearly allotment
L1: The L-1 visa facilitates the temporary transfer of foreign worker in the managerial, executive or specialized knowledge category to the U.S. to continue employment with an office of the same employer, its parent, branch, subsidiary or affiliate. This is also one of the most common visa application type. It is a non-immigrant visa, and is valid for a relatively short amount of time, from three months to five years, based on a reciprocity schedule. With extensions, the maximum stay is seven years.
F1: An F1 visa is a nonimmigrant visa for those wishing to study in the U.S. You must file an F1 visa application if you plan on entering the US to attend a university or college, high school, private elementary school, seminary, conservatory, language training program, or other academic institution.

Performing data analytics and predictive analytics on datasets such as these which include a lot of human intuition and decision making to arrive at the outcome, presents some interesting results. These results can be used to understand the less understood part of decision making and the visa process as a whole. Applicants can get more insight on their application and can try to maximise their success based on the results gleaned.

There are many models out there, which can predict the outcomes of visa status with significant accuracy but currently there is no way for providing recommendations to increase the acceptance rate on the basis of the user data, using our model we try to obtain this.

II. LITERATURE SURVEY

The paper 'Understanding factors that influence L1-visa outcomes in US' by Nihar Dalmia, Meghana Murthy and Nianthrini Vivekanandan attempts to understand the various influencers in the outcome of a L1 visa application. This paper views the problem from the perspective of the employer i.e. they endeavor to present the factors that influence the outcome of an application thus allowing employers to recruit the most appropriate candidate to offer a job through an L1 visa. The dataset used has details about L1 visa applications with over 3 million rows.. The authors combined this dataset with other relevant datasets like religious diversity of the country of origin, rise of GDP of country of origin, education rate of employer state etc, allowing for a more informed model. Their approach starts off with understanding the extent to which a feature affects the outcome, by performing significance tests. They then construct models for Decision Trees, Logistic Regression, Random Forest, ML Perceptron, Gradient Boosting, Naive Bayes. Two models are created for each approach, one with down sampling and another with SMOTE (Synthetic Minority Oversampling Technique). The results indicate that decision trees and logistic regression seem to give better results. The results also state the features that were important or influential for analysis according to the resulting models. Results indicate that down sampling leads to inferior results as compared to SMOTE. The dataset used is a large one with over a hundred features. The authors do not seem to have done any of dimensionality reduction. LDA could be a good technique in the situation. Despite running the statistical significance tests, the features proving to be insignificant have not been eliminated. This could perhaps have bettered the results by eliminating irrelevant features.

The paper 'Predicting Case Status of H-1B Visa Petitions in US' by Induja Sreekanthan, Jahnavi Singhal, Prahal Arora and Rahul Dubey attempts to discover how the outcome of a visa application is influenced by attributes of a users application. The classifier designed in this report could be utilized by both, H-1B aspirants and employers, to gauge the likelihood of visa certification, before and after filing the petition. Features were chosen for prediction by visualising the variations and relations between features. For example, a positive correlation was observed between acceptance rate and year. Similarly, a plot between employer and acceptance rate showed a non uniform plot indicating it is an important feature. They used 4 different classification techniques namely, Gaussian Naive Bayes, Logistic Regression, Random forest Classifier and AdaBoost Classifier. The metrics on which the results of the various models were compared are classification accuracy, false positive rate, false negative rate, balanced error rate, F1 Score, precision and recall. AdaBoost Classifier performed the best in terms of accuracy and F1 score over others with a classification accuracy of 86.139% and Balanced Error Rate of 0.142 on validation data. They observed that the most important feature to considered for their model is the acceptance ratio for the employer and the number of petitions filed by the employer. One inferred drawback of this model is that all the null values were directly dropped. An alternate approach could be to impute the values with a suitable mechanism like substitution with mean or median for discrete variables and KNN for categorical values.

The paper 'Predicting the Outcome of H-1B Visa Applications' by Beliz Gunel and Onur Cezmi Mutlu aims to predict the outcome of H-1B visa applications that are filed by many high-skilled foreign nationals every year by treating it mainly as a classification problem. During the stage of preprocessing, they attempt to weed out applications belonging to certain minority classes either as they are small in number or because they may be incorrect. For example, by creating a feature called applications per employer, they managed to remove companies with misspelled names. New features that were considered positive addition were generated such as company success rate. The final datasets on which the classification algorithms are a combination of the original features along with the newly generated ones. Classification algorithms like Logistic Regression, Naive Bayes, Support Vector Machines and Neural Network with L2 regularization were used to create models. They preferred stochastic gradient descent over batch gradient descent due to the size of the dataset. They also performed different types of regularization such as l1,l2 and ElasticNet (combination of l1 and l2) on the model to increase the test accuracy. Out of the models they tried, Neural Network with l2 regularization outperformed all the other models with 98% training accuracy and 82% test accuracy on the balanced test data. Overall, regularized models performed better by decreasing the variance between training and test accuracy through bounding the size of theta or the weight vector. The authors reported that certain limitations of the model could have been overcome by using one hot representation on categorical variables, trying neural network with L1 regularization (only L2 regularization was attempted on neural networks) and adjusting number of layers in the neural network. Another observed limitation of this approach is that in the preprocessing stage, they have eliminated many of the minority classes. While this removes erroneous records, it also removes the factually correct minority classes. Another approach could have been used to detect these incorrect classes, for example in case of the employee name example taken above, employee names of the dataset could be run against a validated list of employees in the US. The minority classes which lack in number but are factually correct could be upsampled to create a balanced dataset.

The paper 'Predictive Modeling of the Journey from H-1B to Permanent US Work Visa' by Shibbir Khan creates a model that predicts the likelihood of a person with an H-1B visa being given a permanent US work visa. The majority of the paper details the various visual exploratory analysis undertaken in an attempt to understand patterns in the dataset. Plots of acceptance on the basis of job title, country of origin, wage offer, degree of education etc have been made. As the dataset mostly consists of categorical values, probabilistic methods were used, such as Gaussian Naive Bayes, Logistic Regression, Gradient Boosted Classifier, Tree Ensemble Learner Classifier. Then the feature importance graph shows the feature importance of the variables in the model and lists out the attributes most influential to the outcome. All the classifiers indicate accuracies of over 80%. The visualization approach is an excellent one, but the results which show certain features as insignificant could have been incorporated into the predictive model to lessen the confusion created by features deemed irrelevant.

The paper 'H-1B Visa Data Analysis and Prediction by using K-Means Clustering and Decision Tree Algorithms' by Renchi Liu and Jinglin Li attempts to apply machine learning algorithms to the dataset containing application details of H-1B visas to ascertain which will be accepted. Along with predictive analysis, they also propose to answer trend related questions like what are the top companies that have apply to the H-1B for employees, what is the trend in the total number of H-1B application is, what is the top popular job title and worksites for H-1B visa holders, what is the salary mean values of respective job titles etc. The predictive models created were using the techniques decision trees and K-means clustering. The authors reported certain limitations of their approach which could have enhanced their results. The dataset could have been cleaned and processed, naive bayesian and SVM techniques could have been implemented and the living expense of each states of US could have been combined with the salary of the application and provide and predict which cities have more comfortable life and people there could save more compared with other places. This could be more meaningful to relocate for the individuals. Another inferred limitation of this approach is that the models delivered have not been compared with meaningful metrics like confusion matrices, accuracy etc. Without these values, it is hard to determine the success of a model.

III. Problem Statement

To predict whether an L1, H1-B or F1 visa application to the United States is accepted or denied. To Identify and quantify the factors that have an influence in the application decision and help the applicants and employers to identify drawbacks and make informed decisions.

IV. Dataset

Data covers applications in the years 2012 to 2017 and includes information on employer, position, wage offered, job posting history, employee education and past visa history, associated lawyers, and final decision. This data was collected and distributed by the US Department of Labor. The dataset consists of 154 columns containing data on the visa application of applicants.

V. Solution Approach

As stated in the above problem statement, we aim to predict the outcomes of the three most prominent visa categories, namely H1-B, F1 and L1 and futhermore, we aim to build a recommendation system which recommends potential changes or improvements in the application's attributes that will increase the chance of acceptance of the user visa application.

Our research shows that not many studies are held on how to improve a users application to maximize the rate of acceptance of visa application by aspirants. An independent study was done by Beliz Gunel and Onur Cezmi Mutlu of Stanford University [1] where they used many machine learning algorithms to predict the outcome of H1-B visa applications. Another similar project was done by the students of UC Berkley[7] 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. Many other studies are conducted to predict the outcomes of the petitions or something closely related.

We feel that a recommendation system that helps the user to improve their chances to get certified visa petition is very useful for people without prior experience of certification process, trying to apply or visa in the US. To our knowledge, providing users suggestions to improve their application is an uncharted domain and we would like to make an attempt to study the relationship between visa applications attributes and visa status and thus provide suggestions to the user on which attribute of their application they have to work on to guarantee a better chance of visa certification.

VI. Cleaning and Preprocessing

he main and the very first task when it comes to analyzing a huge data is the process of cleaning. We performed cleaning in multiple stages.
-Null values: We checked which of the one hindered and fifty columns contain huge amount of null values. We set a threshold of 75% and dropped all the columns which contained more percentage of null values as they wont contribute enough to the analysis part, and hence will be useless for training of the dataset for prediction.
-Manual: After this step of data cleaning was over, we manually checked which columns are irrelevant to the problem we are foccusing on. For eg: the column named agentcity was irrelevant in detecting the case status ot the outcome of the petition, whereas the columns like the salary might be very important in prediction.
-Based on acceptance ratio: The third and the last method of cleaning the data was graphical analysis of the remaining columns, i.e plot the column entities (encoded if categorical) against the acceptance ratio that is the fraction of applications that received CERTIFIED status. This gave a clear understanding of which columns are actually relevant by showing a non-uniform graph for those particular columns.

The dataset contains 58 classes of visas. The three most popular visa applications are for the H1-B, L1 and F1 and they comprise 84 percent of the entire dataset. Each of these classes of admission have a different thought behind decision making as the purpose of applicants are different. Hence a different classification model is used to predict the outcome of each of the visas separately.

The working dataset has multiple missing values. The columns with more than 75 persent missing values have been completely eliminated. Of the resulting 70 rows, irrelevant attributes like employer address etc. are manually eliminated. This brought our dataset down to 27 columns. We then converted our categorical variables into numerical values with one hot encoding. This dataset has the necessary features along some missing values. We then eliminated the all the rows which have a missing visa type as they cannot be used for any kind of prediction. Imputation of the missing values was our next step. We attempted to implement this using KNN where in we found the nearest neighbours and selected the mode of these as the value to be imputed. The dataset has now been cleaned.

There are 27 columns in the new working dataset which are one-hot encoded. Linear discriminant analysis or PCA (depending on the type of visa category under consideration) is used to reduce the dimensions and complexity of the dataset.

VII. Model

A comparative analysis must be done to check which supervised modelling technique works on which visa type after which the technique will be finalized.

VIII. Recommendation

Along with predicting the outcome of an application, through our project, we also plan to provide a helpful tool which shows the applicant the means by which they can maximise their chances of getting positive results. We plan to do this by clustering the data points based on the outcome i.e. each cluster represents one of the two outcomes. We then find the distance of the test data from the various clusters that have been found. The recommendation aspect comes in as we try to show the best way a test data can move from the denied status to an accepted one by looking at the nearby positive clusters and finding the change in variables that could alter the result.

References

[1] http://cs229.stanford.edu/proj2017/final-reports/5208701.pdf

[2] https://github.com/Jinglin-LI/H1B-Visa-Prediction-by-Machine-Learning-Algorithm

[3] https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf

[4] https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a068.pdf

[5] https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/aml_project_report.pdf

[6] https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2876-200018.pdf

[7] https://www.ischool.berkeley.edu/projects/2016/project-alien-worker