Who Is Likely to Succeed: Predictive Modeling of the Journey from H-1B to Permanent US Work Visa

Shibbir Dripto Khan
Clark University, Graduate School of Management, Worcester, MA, USA

April 8 - 11 | Denver, CO
#SASGF
Predictive Modeling of the Journey from H-1B to Permanent US Work Visa

Shibbir Khan
Clark University, Graduate School of Management, Worcester, MA, USA

ABSTRACT

The purpose of this study is to help US employers and legislators predict which employees are most likely to succeed in the US job market and therefore making sure that the time and money spent is for the most eligible ones. A permanent labor certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. Here we build several predictive models to signal among the applicants who has been awarded with the opportunity of H-1B visa which is a temporary work permit, are likely to sustain and succeed eventually. Then we shall do a probability measure of each H1B visa holders which we would later use as a valuation factor. This gives an idea as to which areas the U.S. government should emphasize to encourage domestic students to develop more local workforce in those job domains for the future.

The model will allow legislators and employers in US job market to pass legislation and conduct hiring process based on and targeted towards those workers who are possibly going to make the most out of the process.

DATA & METHODS

The H1B data set and the Permanent work authorization dataset is collected from United States Department of Labor - Office of Foreign Labor Certification (OFLC). The Office of Foreign Labor Certification (OFLC) generates disclosure data that is useful information about the immigration programs including the H1-B visa. We removed all the data points which had empty values for case status, in our models, we only included the cases ‘CERTIFIED’ and ‘DENIED’. We decided to ignore ‘CERTIFIED- WITHDRAWN’ and ‘WITHDRAWN’ since those were decisions taken by the applicant and/or employer. The data set includes 40 columns and covers a total of 3 million records spanning from 2011-2016.

EXPLORATORY ANALYSIS

Exploratory Analysis: We analyzed the distribution of various features, their relation with the number of applications, and with the acceptance rate. These charts show significant trend in terms of these features.

Further exploratory analysis were done in terms of case status based on level of graduate and undergraduate education, year of approval, pay scale and jobs that hire he most. These charts show significant trend in terms of these features. Not so surprisingly, Various Information Technology (IT) roles dominate the H-1B visa applications.

We have also used PROC summary in SAS Enterprise guide to get some basic statistical measures. Such as Mean, Variance and confidence intervals. We found the descending order of employers based on the number of applications submitted. We also found with PROC summary most frequent occupation code in the application pattern. We also used some data preparation tools such as string to number, Row filters and column filters. Missing value node was used to decide measure and imputation. Since logistic regression will not work with missing values, we derived Non Null distribution through R.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration. Other brand and product names are trademarks of their respective companies.
Predictive Modeling of the Journey from H-1B to Permanent US Work Visa
Shibbir Khan
Clark University, Graduate School of Management, Worcester, MA, USA

EXPLORATORY ANALYSIS

The following map shows top cities with Visa Applications for Case Status as Certified. Large number of applications is indicated by a larger circle, we included Cary North Carolina. We used R’s free third-party basemaps which were added using the addProviderTiles() function, this is implemented using leaflet-providers minimap plugin. minimap shows map tiles; markers, polygons, and other layers on the main map. We used Esri USGS WorldStreet minimap here.

RESULTS CONTINUED

For the predictive modeling we initially had a couple of different procedures in mind including Clustering and nearest-neighbor methods, but usually those are ideally suited for use with numeric data. However, our dataset mostly consists of categorical values, so using a probabilistic method, such as the Naive Bayes Classifier (NBC) is likely to provide more accurate results.

We have used four different classifiers for the predictive task and evaluated the results of each:

A. Gaussian Naive Bayes
B. Logistic Regression
C. Gradient Boosted Classifier
D. Tree Ensemble Learner Classifier

RESULTS CONTINUED

Partitioning: before diving into the predictive modeling we have partitioned the datasets. We have performed sampling of 70% and 30% for training and testing and we have chosen stratified sampling based on the strata in case status column.

A. Gaussian Naive Bayes
Naive Bayes assumes all features are conditionally independent given labels. It calculates p(x|y = 1), p(x|y = 0) and p(y) by taking their maximum likelihood estimates in the joint likelihood of the data. While making a prediction, it calculates p(y = 1|x) both for p(y=1) and p(y=0) using the Bayes rule and compares the two. For the Naive Bayes learner we have created 4.2 PMML compatible model and we have selected case status as the classification column as it is our target column.

B. Logistic Regression
For this we have again picked target column as ‘case status’ and reference category as ‘certified’. We have selected solver as ‘stochastic average gradient’ In logistic regression, probability of the response taking a particular value is modeled based on combination of values taken by the predictors. Estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values. The advantage of Logistic Regression model is that it gives the confidence of prediction as a probability. The disadvantage is that it assumes that the classes are linearly separable in feature space.

C. Gradient Boosted Classifier
For this we picked as missing value handling method : ‘XGBoost’, which we have later reproduced for H1B dataset through R Studio. we used a static random seed for this classifier, we changed prediction column name as Prediction (case status) GB.
RESULTS CONTINUED

In this work, Gaussian Naive Bayes, Logistic Regression, Tree Ensemble Learner Classifier and Gradient Boosted Classifier were considered for determining the status of PERM visa applications from H1B visa applications. Gradient Boosted Classifier performed the best in terms of accuracy as 88.16% were correctly classified. We achieved a 86.069% classification accuracy on validation data for Logistic Regression. Although we had high expectation for Naïve Bayes Classifier but we ended up having an accuracy of 82.34%, this can be due to the fact that it assumes that all our features are independent and we might have at least some form of correlation between the columns. We also inferred from our relative variable importance plot that: Particular well known Employer, prevailing wage, prevailing SOC occupation codes (type of occupation) play an important role in determining the case status of the visa application.

REFERENCES


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.