Visa certification

By: Narges Shahmohammadi(yxs662)

Background

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

Motivation

In 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. I will analyze the dataset with the help of a classification model:

* Facilitate the process of visa approvals.

* Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

The analysis will primarily focus on:

Those with higher education may want to travel abroad for a well-paid job. Does education play a role in Visa certification?

How does the visa status vary across different continents?

Experienced professionals might look abroad for opportunities to improve their lifestyles and career development. Does work experience influence visa status?

In the United States, employees are paid at different intervals. Which pay unit is most likely to be certified for a visa?

The US government has established a prevailing wage to protect local talent and foreign workers. How does the visa status change with the prevailing wage?

Description of the Data

The dataset has 25480 rows and 12 columns.

There are no missing and duplicated values.

9 columns type are object, and 2 columns are integer and just one is float. Some columns need to be dummy like {Continent, education_of_employee or region_of_employment}.

Proposed Analysis

Our analysis of the data will comprise of logistic regression, decision tree and random forest models and visual representations of the data analysis such as histograms to help interpret the information.

Through the use of logistic regression, we will try to explore what variables caused a visa status denied. The use of this type of model can be utilized because the response variable, visa status, is categorical.

The decision tree and random forest are both tree base models. They will help me to predict whether or not the visa status is denied in the future.

Decision tree (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

The Random forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees. The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

Also, by the visualization, I can get a deep insight into the variables. for example, checking the correlation by heatmap, distribution by histogram, outliers by scatterplot, and so on.

Variable Descriptions

The data contains the different attributes of employee and the employer. The detailed data dictionary is given below.

* **case_id:** ID of each visa application

* **continent:** Information of continent the employee

* **education_of_employee:** Information of education of the employee

* **has_job_experience:** Does the employee has any job experience? Y= Yes; N = No

* **requires_job_training**: Does the employee require any job training? Y = Yes; N = No

* **no_of_employees:** Number of employees in the employer's company

* **yr_of_estab:** Year in which the employer's company was established

* **region_of_employment:** Information of foreign worker's intended region of employment in the US.

* **prevailing_wage:** Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.

* **unit_of_wage:** Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.

* **full_time_position**: Is the position of work full-time? Y = Full Time Position; N = Part Time Position

* **case_status:** Flag indicating if the Visa was certified or denied