H-1B Visa Petitions 2011 – 2016

Data Science 6105 Summer 2019

Group 4 – Fantastic Four

1. **Introduction which explains the problem you are trying to address and why it's a big deal?**

A: Dataset we have for this project is H-1B Visa Petition 2011-2016. We analyzed the dataset of H-1B applicants' application information, including the number of applicants, popular positions, popular employers and regions from 2011 to 2016. Meanwhile, we predict the application status of H-1B applicants based on the existing data set, because it can help incoming H-1B applicants modify their applications. The goal of our model is to predict whether the applications are qualified or not based on the application information of applicants.

1. **Details regarding the dataset used, how many rows, features**

A: H-1B Visa Petition 2011 – 2016, total shape is (3002458, 11), Columns include:

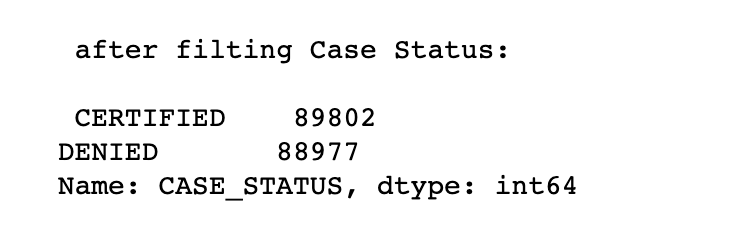
|  |  |
| --- | --- |
| Column | Description |
| CASE\_STATUS | The CASE\_STATUS field denotes the status of the application after LCA processing. Certified applications are filed with USCIS for H-1B approval. CASE\_STATUS: CERTIFIED does not mean the applicant got his/her H-1B visa approved, it just means that he/she is eligible to file an H-1B. |
| EMPLOYER\_NAME | Name of the employer submitting labor condition application. |
| SOC\_NAME | Occupational name associated with the |
| JOB\_TITLE | Title of the job. |
| FULL\_TIME\_POSITIONY | F = Full Time Position; N = Part Time Position. |
| PREVAILING\_WAGE | Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position. |
| YEAR | Year in which the H-1B visa petition was filed. |
| lon | Longitude |
| lat | latitude |

We are going to predict the CASE\_STATUS,

A screenshot of a cell phone

Description automatically generated

Most of them is CERTIFIED, and as we are going to predict CERTIFIED or DENIED, so we drop off the extra Status, as well as some CERTIFIED rows, in order to keep the similar number of rows for CERTIFIED and DENIED.



About other columns,

**Categorical Columns:**

'CASE\_STATUS', 'EMPLOYER\_NAME', 'SOC\_NAME', 'JOB\_TITLE', 'FULL\_TIME\_POSITION', 'WORKSITE'

**Continuous Columns:**

'PREVAILING\_WAGE', 'YEAR', 'lon', 'lat'

1. **Methodology, which algorithm/s did you use?**

A: This prediction is a Classifier Question, and we use Random Forest, Gradient Boosting, and DNN algorithms. Applied to three models and then find out the different performance for that different model.

1. **Results, which metric did you use and why? What was the performance of your model on the test set?**

A: We use Confusion Matrix and ROC Plot for each model, as it is a Binary-Classifier question. At same time, we evaluate model with percison score and f1 score

1. Random Forest Classifier

A screenshot of a cell phone

Description automatically generatedA screenshot of a map

Description automatically generated

1. Gradient Boosting Classifier

A screenshot of a cell phone

Description automatically generatedA screenshot of a map

Description automatically generated

1. Deep Neural Network

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

For DNN, we also find relation between the accuracy and epochs

A screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

After the evaluation, we compare all three models,

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated