MACHINE LEARNING FOR PREDICTIVE MODELING OF H1B VISA APPROVAL : A NEXT GENERATION APPROACH TO IMMIGRATION SERVICES

Submitted by,

LIYA C C

ATHILA M S

1. INTRODUCTION

The main goal is to predict the outcome of H-1B visa applications that are filed by many professional foreign nationals every year. Here, we framed the problem as a classification problem and applied it in order to output a predicted case status of the application. The input to our algorithm is the attributes of the applicant. H-1B is a type of non-immigrant visa in the United States that allows foreign nationals to work in occupations that require specialized knowledge and a bachelor's degree or higher in the specific specialty. This visa requires the applicant to have a job offer from an employer in the US before they can file an application to the US immigration service (USCIS). We believe that this prediction algorithm could be a useful resource both for the future H-1B visa applicants and the employers who are considering sponsoring them.

In order to predict the case status of the applicants, we will be feeding the model with the dataset which contains the required fields by which the machine can classify the case status as certified or denied.

1.1 OVERVIEW

Visa is the guide of authorization on a travel permit that gives a permit to the holder to move in, leave or stay in the country for a predetermined timeframe. There are distinctive kinds of foreigner visas, the required structures, and the means in the worker visa process contingent upon the nation one needs to move. Moving to America is a vital and complex decision.. It is intended to carry outside experts with professional educations and specific aptitudes to fill occupations when qualified Americans can't be found. Be that as it may, as of late, worldwide outsourcing organizations have ruled the program, winning a huge number of visas and pressing out numerous American organizations, including littler new companies. The development in the portrayal of the outside conceived among the US workforce was brought down drastically.

1.2 PURPOSE

The main objective is to predict the outcome of H-1B visa applications that are filed by many professional foreign nationals every year. Here, we framed the problem as a classification problem and applied it in order to output a predicted case status of the application.

2 LITERATURE SURVEY

2.1 EXISTING PROBLEM

The dataset that we are studying is available on Kaggle under the name 'H-1B Visa Petitions 2011-2016 dataset' which is processed dataset from the original data available on Office of Foreign Labor Certification (OFLC) website after performing various data transformations on the data. From data analysis performed on this data allow us to finding top Occupations, States, Employers and Industries that contribute to highest number of H1B visa application. Our research says that some independent analysis performed on this data provided some insightful facts and also to predict some details. Astudy by Andrew Shikair explored a way to predict wages of the visa recipients by performing text analysis of application attributes and their study concludes that Occupational Classification and Job Title were the two most important fields to predict the applicants wage as accurately as possible.

2.2 PROPOSED SOLUTION

H-1B visa applications that are filed by many professional foreign nationals every year. Here, we framed the problem as a classification problem and applied it in order to output a predicted case status of the application. The input to our algorithm is the attributes of the applicant. H-1B is a type of non-immigrant visa in the United States that allows foreign nationals to work in occupations that require specialized knowledge and a bachelor's degree or higher in the specific specialty. There are distinctive kinds of foreigner visas, the required structures, and the means in the worker visa process contingent upon the nation one needs to move. Moving to America is a vital and complex decision.. It is intended to carry outside experts with professional educations and specific aptitudes to fill occupations when qualified Americans can't be found.

3 THEORITICAL ANALYSIS

In present work, we prognosis the consequences of all the seven machine learning characterization models on the testing dataset. All the seven approaches are running on the parameters as appeared in Table 5. The accuracy which is computed by utilizing Eq. (5) and Other statistical measures are also present in table 6.These measures are computed by using R function mmetric which computes the classification error metrics. In this work we use metrics like Accuracy, Precision, Total Positive Rate (TPR), Total Negative Rate (TNR), Classification error (CE) and F1 score (F1). Time and accuracy are used for comparison between these seven models. Table 7 shows different results of accuracy and time on 50-50%, 60-40%, 70-30% and 80-20% partition of train and test datasets. In cross-validation, models are accomplished n number of times and accuracy is recorded if accuracy is very fluctuating then that model is over fitted/under fitted/one-sided.

3.1 HARDWARE / SOFTWARE DESIGNING

The hardware required for the development of this project is:

Processor : Intel CoreTM i5-9300H
Processor speed : 2.4GHz
RAM Size : 8 GB DDR
System Type : X64-based processor

SOFTWARE DESIGNING:

The software required for the development of this project is:
Desktop GUI : Anaconda Navigator

Operating system : Windows 10
Front end : HTML, CSS,JAVASCRIPT
Programming : PYTHON
Cloud Computing Service : IBM Cloud Services

4 EXPERIMENTAL INVESTIGATION

IMPORTING AND READING THE DATASET

Importing the Libraries
First step is usually importing the libraries that will be needed in the program.
Pandas: It is a python library mainly used for data manipulation.

NumPy: This python library is used for numerical analysis.

Matplotlib and Seaborn: Both are the data visualization library used for plotting graph which will help us for understanding the data.

csr_matrix() :A dense matrix stored in a NumPy array can be converted into a sparse matrix using the CSR representation by calling the csr_matrix() function.

Train_test_split: used for splitting data arrays into training data and for testing data.

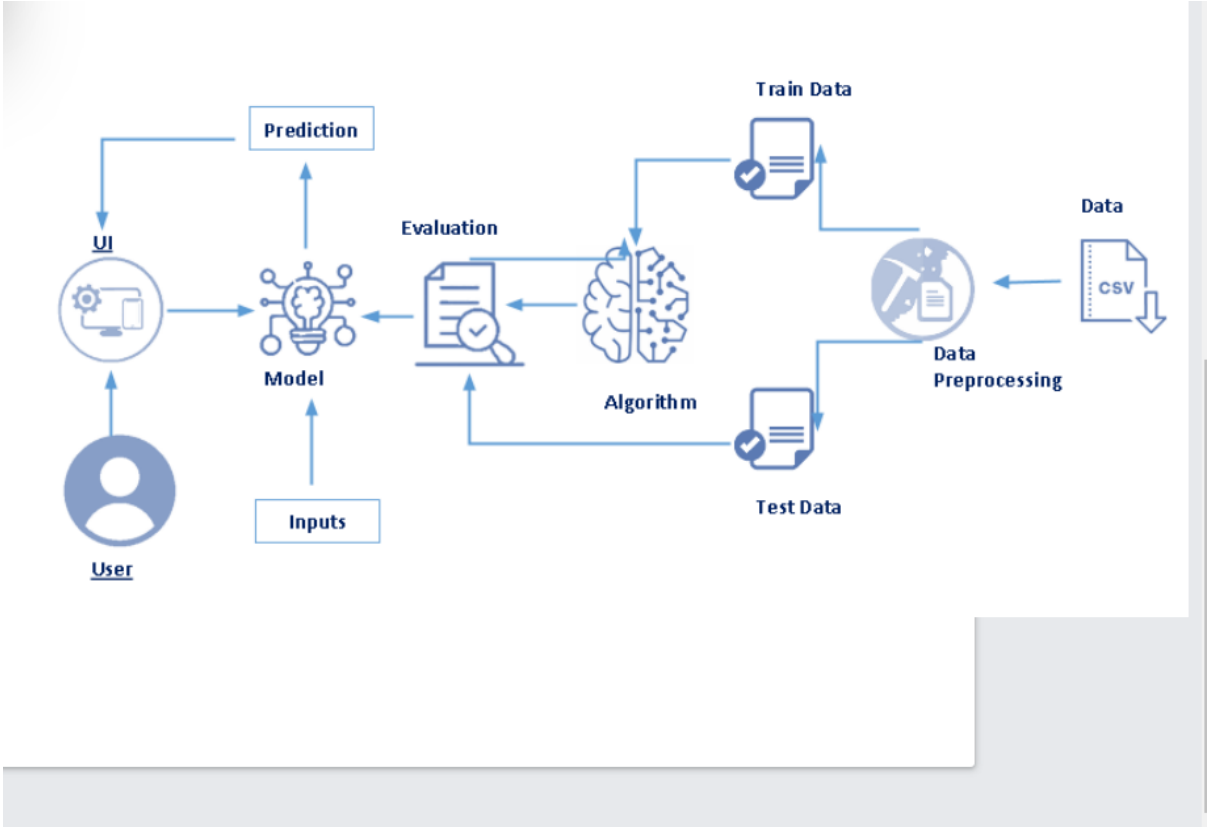
Pickle: to serialize your machine learning algorithms and save the serialized format to a file.

Reading the Dataset

For this project, we make use of three different datasets (Books_Ratings, Books, Users). We will be selecting the important features from these datasets that will help us in recommending the best results.
The next step is to read the dataset into a data structure that's compatible with pandas. Let's load a .csv data file into pandas. There is a function for it, called read_csv().We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program).If the dataset in same directory of your program, you can

directly read it, without any path. After the next Steps we made following bellow:
1.Data visualization
2.Collabrative and filtering
3.Creating the Model
4.Test and save the model
5.Buil Python Code
6.Build HTML Code
7.Run the Application
We are the following above sections we did and investigate it.

5 .FLOWCHART



Project Flow:
• User interacts with the UI (User Interface) to upload the input features.
• Uploaded features/input is analysed by the model which is integrated.

Once a model analyses the uploaded inputs, the prediction is showcased on the UI.
1. Data Collection.
• Collect the dataset or Create the dataset
2. Data Pre- processing.
• Import the Libraries.
• Importing the dataset.

- Exploratory Data Analysis
- Data Visualization

3. Collaborating Filtering

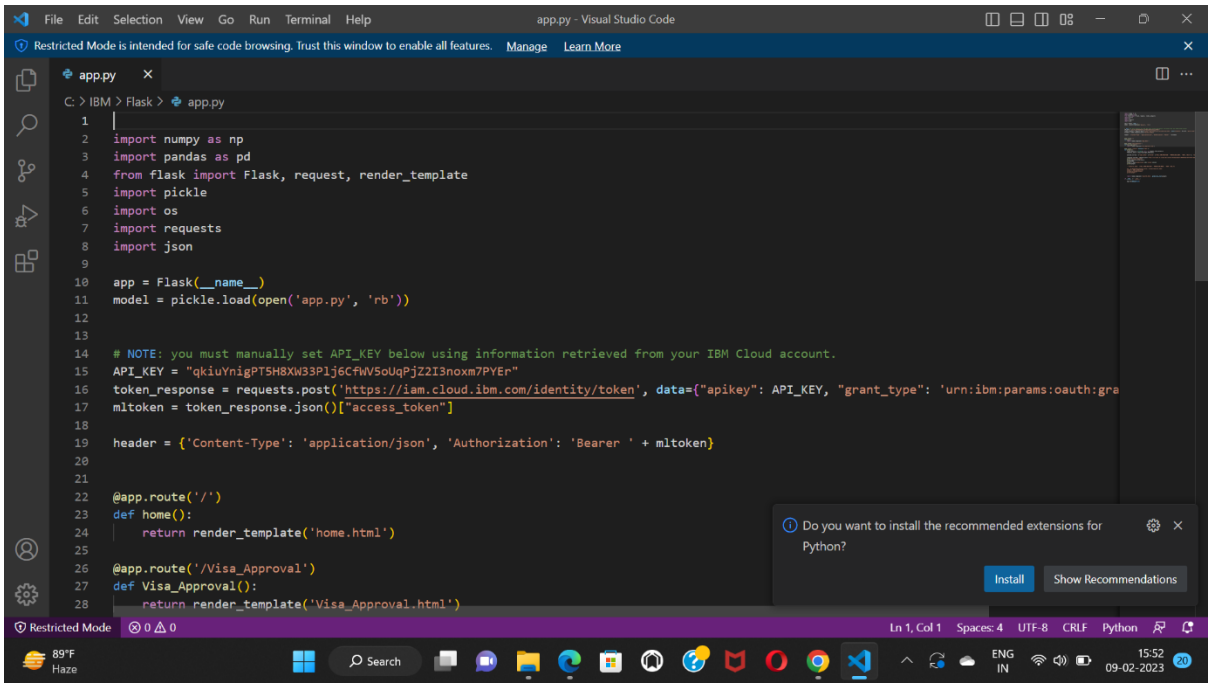
- Merging datasets
- Creating the Model

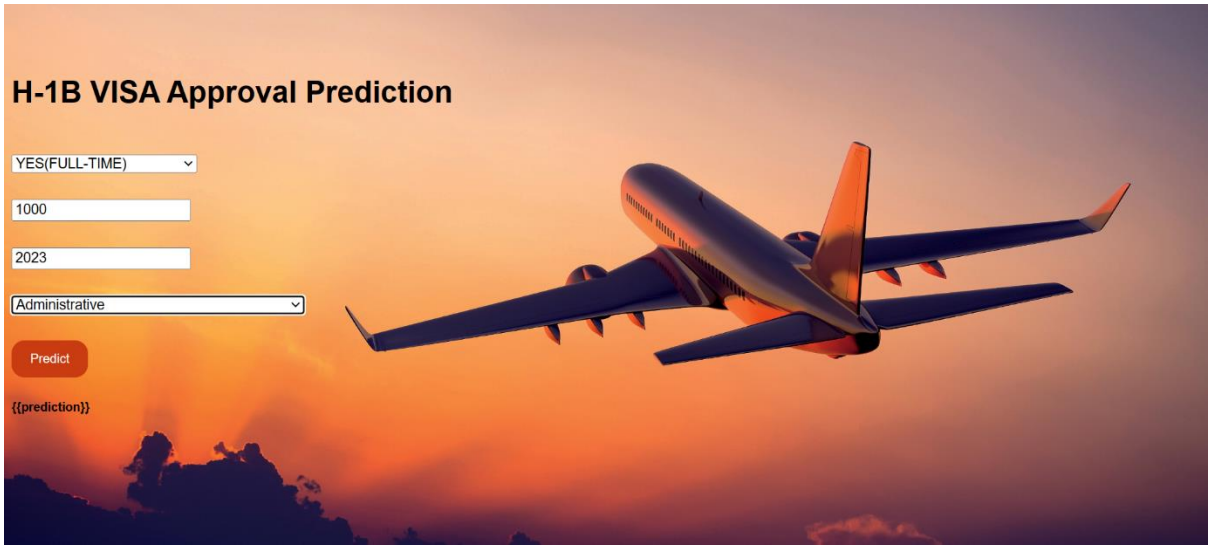
- Predicting the results
- Saving our model and dataset

4. Application Building

- Create an HTML file
- Build a Python Code

6.RESULT





7 ADVANTAGES AND DISADVANTAGES

ADVANTAGES

- H-1B requirements are not as stringent as those for an O-1A.
- Dual intent visa, which means it's totally fine to be anywhere in the green card process.
- Under the current random lottery system, the minimum salary for an H-1B recipient is low, which could be a plus if other compensation, such as equity or stock options, is high.
- Eligible for premium processing.
- Dependent spouse of H-1B holder eligible for a work permit once green card petition approved.

DISADVANTAGES

- H-1B startup founder must be supervised by someone who oversees the founder's work and can fire the founder.
- A founder's equity stake in the startup must be less than 50%.
- Premium processing is often temporarily halted during the H-1B season.
- Allows for a temporary stay, not permanent residence.

8. APPLICATIONS

When applying for an H1B visa, the applicant is sponsored by their U.S. employer that has hired them. The employer will pay for the applicant's visa fees and will submit the required documents on behalf of the applicant to bring them to the U.S. so that they can work for their company.

9. CONCLUSION

H1B visa category is one of the most applied categories among other visas categories. It is designed to overcome the shortage of skilled workers in America but it affects the hiring of American workers and no. of foreign workers increased day by day. So in current work we investigate the machine learning arrangement models with 20 properties to foresee the real H1B visa solicitors with no contribution from any external sources. Based on classification error metrics proposed model give better accuracy of 95.45 as compared single models. The projected model increases the accuracy rate as validated by 10 fold cross validations. There are chances to train data on some other classification models which may give better results. The work can be stretched out to more properties with a better relationship and other computational strategies to upgrade the execution of machine learning techniques. Some pros and cons are that we select important features by using Pearson correlation only, in future other researchers can also go for other feature selection methods which may predict better results by training new models under new conditions. The dataset and source code utilized as a part of the examination are accessible at https://www.kaggle.com/nsharan/h-1bvisa.

10 FUTURE SCOPE

Previously operating as a lottery-based system, the H-1B visa program awards temporary non-immigrant working visas to highly educated foreign professionals, many of whom are most likely to work in STEM-related positions. It is intended to carry outside experts with professional educations and specific aptitudes to fill occupations when qualified Americans can't be found. Be that as it may, as of late, worldwide outsourcing organizations have ruled the program, winning a huge number of visas and pressing out numerous American organizations, including littler new companies. The development in the portrayal of the outside conceived among the US workforce was brought down drastically.

11. BIBILOGRAPHY

[1] Dhanasekar Sundararaman , Nabarun Pal , Aashish Kumar Misraa ,(2017),” An analysis of nonimmigrant work visas in the USA using Machine Learning” ”, International Journal of Computer Science and Security(IJCSS), Vol. 6.
[2] https://www.foreignlaborcert.doleta.gov/performancedata.cfm.
[3] UNITED STATES DEPARTMENT OF LABOR. (2009, Janu-ary 15). OFLC Performance Data. (www.dol.gov) Retrieved September 09, 2017, from UNITED STATES DEPARTMENT OF LABOR Employment & Training Administration.
[4] Trim Bach, S., (2016), Giving the Market a Microphone: Solutions to the Ongoing Displacement of US Workers through the H1B Visa Program. Nw. J. Int'l L. & Bus., 37, p.275.
[5] Doran, K., Gelber, A. and Isen, A., 2014. The effects of highskilled immigration policy on firms: Evidence from H-1B visa lotteries (No. w20668). National Bureau of Economic Research. https://doi.org/10.3386/w20668

APPENDIX

SOURCE CODE OF FLASK

```
import numpy as np
import pandas as pd
from flask import Flask, request, render_template
import pickle
import os
import requests
import json

app = Flask(__name__)
model = pickle.load(open('app.py', 'rb'))

# NOTE: you must manually set API_KEY below using information retrieved from your
IBM Cloud account.
API_KEY = "qkiuYnigPT5H8XW33Plj6CfWV5oUqPjZ2I3noxm7PYEr"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey":
API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}

@app.route('/')
def home():
    return render_template('home.html')

@app.route('/Visa_Approval')
def Visa_Approval():
    return render_template('Visa_Approval.html')

@app.route('/predict',methods=['POST'])
def predict():
    input_features = [float(x) for x in request.form.values()]
```

```
    features_value = [np.array(input_features)]

    payload_scoring = {"input_data": [{"field": [['FULL_TIME_POSITION',
'PREVAILING_WAGE', 'YEAR','SOC_N']], "values": [input_features]}]}

    response_scoring = requests.post('https://us-
south.ml.cloud.ibm.com/ml/v4/deployments/9078763e-b479-4774-9abe-
a28bdab485e9/predictions?version=2021-10-26', json=payload_scoring,
headers={'Authorization': 'Bearer ' + mltoken})
    print("Scoring response")
    pred=response_scoring.json()
    print(pred)
    output = pred['predictions'][0]['values'][0][0]
    print(output)

    '''features_name = ['FULL_TIME_POSITION', 'PREVAILING_WAGE',
'YEAR','SOC_N']

    df = pd.DataFrame(features_value, columns=features_name)
    output = model.predict(df)
    #output=np.argmax(output)
    print(output)'''

    return render_template('resultVA.html', prediction_text=output)

if __name__ == '__main__':

    app.run(debug=False)
```