VISA Approval Prediction

1.INTRODUCTION

1.1 Overview

In our project, we aim to predict the outcome of H-1B visa applications that are filed by many high-skilled foreign nationals every year. We framed the problem as a classification problem and applied 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). USCIS grants 85,000 H-1B visa's every year, even though the number of applicants far exceed that number. The selection process is claimed to be based on a lottery, hence how the attributes of the applicants affect the final outcome is unclear. 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 to

sponsor them.

1.2 Purpose

Our goal for this project is to predict the case status of an application submitted by the employer to hire non-immigrant workers under the H-1B visa program. Employer can hire non-immigrant workers only after their LCA petition is approved. The approved LCA petition is then submitted as part of the Petition for a Non-immigrant Worker application for work authorizations for H-1B visa status.

2.LITERARY SURVEY

2.1 Existing Problem

The dataset that we are studying is available on Kaggle under the name 'H1B Disclosure Dataset'
which is processed dataset from the original 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.

VISA Approval Prediction

A project done by the students of UC Berkley aims to predict the waiting time to get a work visa for a given job title and for a given employer. They used KNN as the primary model to predict ‘Quickest Certification Rate’ across both occupations and companies.

2.2 Proposed Solution

The dataset we used for this problem is downloaded from Kaggle. It contained 10 features as shown in the Figure 1.

FULL_TIME_POSITION	PREVAILING_WAGE	PW_SOURCE_YEAR	PW_SOURCE_OTHER	WORKSITE_STATE	CASE_STATI
Y	59197.0	2015.0	OFLC ONLINE DATA CENTER	IL	CERTIFIEDW
Y	49800.0	2015.0	WILLIS TOWERS WATSON SURVEY	IL	CERTIFIEDW

Figure 1: Two datapoints from the unprocessed dataset

FULL_TIME_POSITION:Positions are given in Full_time_position="Y", and Part_time_position="N". We converted them to ”Full Time Position = 1; Part Time Position = 0” format.

PREVAILING_WAGE: Prevailing wage is the average wage paid to employees with similar qualifications in

VISA Approval Prediction

the intended area of employment. We are using this feature as it is.

CASE_SUBMITTED_YEAR: The year when the application was submitted.

CASE_SUBMITTED_MONTH: The month when the application was submitted.

CASE_SUBMITTED_DAY: The day was the application got submitted.

PW_SOURCE_YEAR: This is the year when the average wage paid to employees.

DECISION_DAY: The day when application got approved.

DECISION_YEAR: The year when application got approved.

DECISION_MONTH: The month when application got approved.

CASE_STATUS: This feature give us a complete

VISA Approval Prediction

prediction about either the application has been approved or denied.

3.THEORITICAL ANALYSIS

3.1 BLOCK DIAGRAM

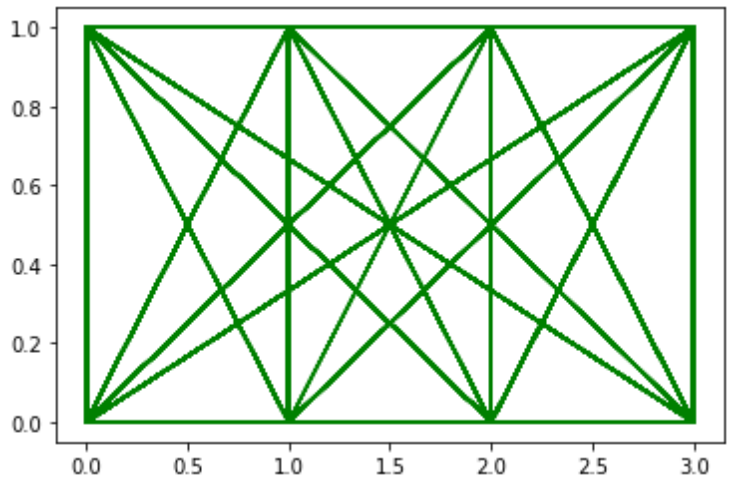


Figure 4: Matplot between "CASE_STATUS" and "FULL_TIME_POSITION"

3.2HARDWARE/SOFTWARE DESIGNING

During the data preprocessing, we used Label Encoding only on 2 features, "FULL_TIME_POSITION",

and"CASE_STATUS". And then simply splitted the data into train and test set in 80:20 ratio.

4. EXPERIMENTAL ANALYSIS

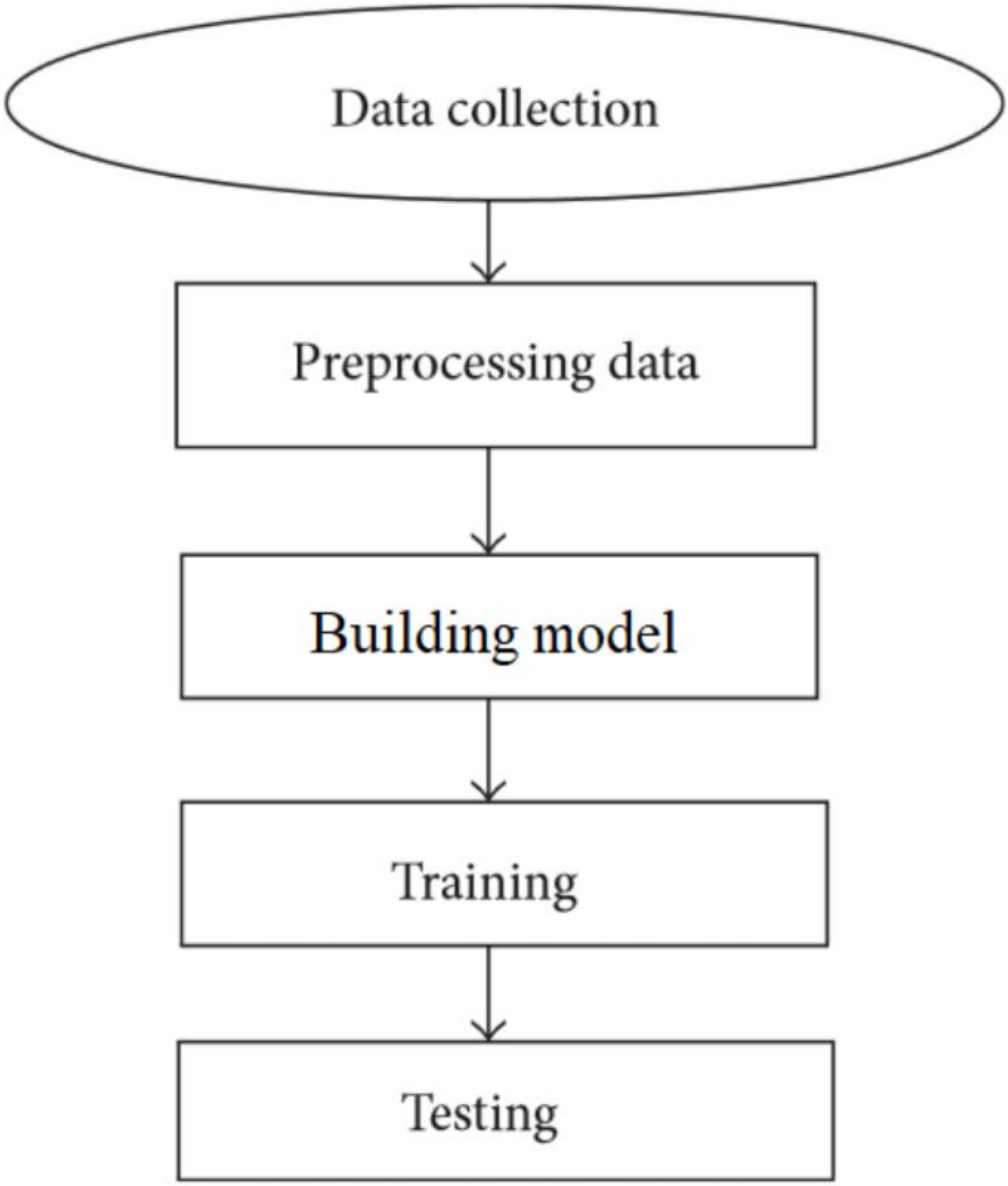
As our dataset is large and its also a classification problem, we thought of using Navie Bayes technique. Either Navie Bayes or Support Vector Machine technique can be used, but here we implemented Navie Bayes technique.

Navie Bayes:
Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.It is a simple and interpretable model which assumes all features are conditionally independent given labels and are in gaussian distribution.
It calculates P(x/y=0), P(x/y=1) and P(y) by taking their maximum likelihood estimates in the joint likelihood of the data. While making a prediction, it considers both P(y=1) and P(y=0) on the Bayes rule and compares the two.

P(A|B) = P(B|A) * P(A) / P(B)

VISA Approval Prediction

5. FLOWCHART



VISA Approval Prediction

6. RESULT

After spliting the data in ratio 80:20, we applied Navie Bayes on the data to predict the outcome. While using the navie bayes algorithm we used the default version of it without changing any hyper-parameter tuning.

CLASSIFIER	ACCURACY
Navie Bayes	0.8420763630511138

VISA Approval Prediction

7.ADVANTAGES AND DISADVANTAGES

As it a classification problem , we used the the navie bayes classifier but there are many pros and cons using this classifier in machine learning.

Advantages:
1. When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models.
2. Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.
3. Naive Bayes is also easy to implement.

Disadvantages:
1. Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.
2. If categorical variable has a category in test data set,

which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency.

8. APPLICATIONS

1. Real-time Prediction: As Naive Bayes is super fast, it can be used for making predictions in real time.
2. Multi-class Prediction: This algorithm can predict the posterior probability of multiple classes of the target variable.
3. Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers are mostly used in text classification (due to their better results in multi-class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)
4. Recommendation System: Naive Bayes Classifier along with algorithms like Collaborative Filtering

VISA Approval Prediction

makes a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

9. CONCLUSION

In this work navie bayes was considered for determining the status of H1B visa application. We have achieved a best of 84% accuracy by the navie bayes classifier. This leads to conclusion that how much important a feature selection and transformation is. Thus one can infer from results that the chance of being certified increases with the amount of wage and how successful your sponsor was in the previous H1B applications.

10. FUTURE WORK

If we had more time and computational resources, there are several directions we could take to improve

VISA Approval Prediction

our prediction algorithm.
We would have tried Support Vector Machine also for the prediction. We would have also tried logistic regression and Neural Network.In addition, we could convert more features into onehot encoding to achieve better accuracy.

BIBLIOGRAPHY

1. H1B Disclosure Dataset - Predicting the Case Status[Online] Available- https://www.kaggle.com/trivedicharmi/h1b-disclosure-dataset

2. Data visualization using Python[Online]; Available- https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7

3. Data preprocessing[Online]; Avialable- https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf

4. Model Building[Online]; Available- https://thesmartbridge.com/documents/spsaimldocs/Machin

VISA Approval Prediction

elearning.pdf

5.W3SCHOOL[Online]; Available-
https://www.w3schools.com/bootstrap/bootstrap_forms_inputs.asp

6. Flask App[Online]; Available-
https://thesmartbridge.com/documents/spsaimldocs/FlaskML.pdf

7. https://acadgild.com/blog/naive-bayesian-model

APPENDIX

finapp.html

```
from flask import Flask,render_template,request
import pickle
import numpy as np
model=pickle.load(open(r"C:\Users\sai srija behara\Desktop\team12\naive.pkl",'rb'))
app=Flask(__name__)
@app.route('/')
```

VISA Approval Prediction

```
def home():
    return render_template("fin.html")
@app.route('/login',methods=['POST'])
def login():
    file=request.form['ap']
    if(file=="Y"):
        s1=1
    if(file=="N"):
        s1=0
    file1=request.form['ag']
    file2=request.form['bp']
    file3=request.form['subday']
    file4=request.form['submonth']
    file5=request.form['subyear']
    file6=request.form['decday']
    file7=request.form['decyear']
    file8=request.form['decmonth']


total=[[s1,int(file1),int(file2),int(file3),int(file4),int(file5),int
(file6),int(file7),int(file8)]]

    y_pred=[[3]]

    y_pred=model.predict(np.array(total))
```

VISA Approval Prediction

```
    if (y_pred==[[0]]):
        return render_template("fin.html",showcase=" your
visa is certified {} "+str(y_pred))

    if(y_pred==[[1]]):
        return render_template("fin.html",showcase="your
visa is certifiedwithdrawn{}"+str(y_pred))
    if(y_pred==[[2]]):
        return render_template("fin.html",showcase="your
visa is denied{}"+str(y_pred))
    else:
        return render_template("fin.html",showcase="your
visa is withdrawn{}"+str(y_pred))

if(__name__)=='__main__':
    app.run(debug=False)
```