H-1B Visa Dataset Analysis

BCIS 5110
Programming Languages for Business Analytics

Project Report

Presented By
Nimmala Vijaysimha Reddy
Group 13

Table of Contents

1. Executive Summary

The project's objective is to analyse the H1B visa dataset, extract useful insights and present a review of the various factors affecting the H1B visa approval rates which help various stakeholders such as employers and job seekers.

This project reviews several studies on H1B visas that have explored different aspects of the U.S. visa program. A crucial component of American immigration law is the H1B visa program, which offers a method for highly educated professionals to enter the nation and boost the economy, with worries regarding its influence on job possibilities, the program has come under increased attention and criticism in recent years. Therefore, this research focused on the understanding the approval time length and predicting the Case Status. This study aims to fill that gap by examining the approval time of H1B visas based on sector and job type. The study will use descriptive statistics and exploratory data analysis to gain knowledge on this topic. Additionally, machine learning models Logistic Regression, Random Forest, and Naïve Bayes will be employed to predict the approval status of H1B visas. The dataset includes variables indicating whether the consent was certified, denied, withdrawn, or certified and withdrawn, and other variables will be used as predictors.

Based on our analysis, we were able to list the top 20 list of job for a H1B visa and the time length for the approval. From the analysis, we were able to understand if the H1B visa approval length is related to the employer location, employer name, job title, or industry. This will assist non-immigrants to know what locations, employers and occupation needed to have higher chances to get approved for H1B visa.

2. Project Background

H1B visa program in the United States, which allows foreign nationals to work in certain occupations that require specialized expertise and a bachelor's degree or higher in the relevant field. The H1B visa was established in the 1960s, and it is the most common type of non-immigrant visa in the U.S.

To apply for an H1B visa, an applicant must have an offer of employment from a U.S. company for a position that requires expertise, a bachelor's degree proof in the field, and evidence from the employer firm that the U.S. lacks such an expert applicant for that role. These are the three requirements that an applicant must meet, according to the USCIS, to apply for an H1B visa (Glennon, 2020). If the application is successful, the employer must submit a petition on the applicant's behalf. The visa application will be entered into a lottery if the applicant's petition is granted, and candidates are randomly selected to receive visas, subject to an annual cap.

Indian citizens account for 74% of all visa holders in the United States, and the demand for H1B visas is primarily emanating from India. This is due to the significant salary disparity between employees in the United States and those in India, with U.S. workers earning far more than their Indian counterparts for the same job.

Research explains that policies related to immigration have a significant impact on the performance of firms and that research has focused on the effects of skilled immigration to the United States on businesses in the U.S.

Overall, the H1B visa program provides significant benefits for skilled workers and companies in the United States. However, it also highlights the potential risks to the availability of highly trained labour in India due to the program's lure of higher salaries in the U.S., so it concludes that research into the effects of skilled immigration and emigration on businesses in both the U.S. and India is necessary to address any research gap.

3. Key questions

1. What are factors that have the most significant impact on the H1B visa approval rates?
2. Do the prediction models give the factors influencing the approval status of the H1B visa?

   The research question will look at the models that best fit the dataset, predict the approval status (case status), compare the models, and evaluate their performance. It looks at how different factors explain the variability of the response (case status), influencing the approval status.
3. Can we predict the likelihood of an H1B visa application being approved based on the full-time or part-time position of the applicant?

4. Is there relationship between length of H1B approval and job type and industry they work in?

5. What is the time taken to process a petition, based on the difference between the CASE_SUBMITTED and DECISION_DATE columns?

6. Total no of H1B applications per state and is there any relationship between H1B visa approval and Employer State?

7. Is there any relationship between visa approval and job industry they work in?

8. Top Employers sponsoring H1-B Visas and its Case Status?

4. Data Source

The H1B visa data collection, which is accessible on Kaggle, https://www.kaggle.com/datasets/jonamjar/h1b-data-set-2017?select=H-1B_Disclosure_Data_FY17.csv, is an extensive dataset that includes details on H1B visa applications.

The data source is a publicly accessible portal that explains the contents of the dataset and makes it easy to understand. The H1B visa program dataset was chosen because it contains crucial information that can help bridge the knowledge gap about its impact on the US economy. The dataset has been underutilized, which provides an excellent opportunity for research on approval times based on sector and job type.

5. Data description

This dataset includes 53 columns, 624650 H1B visa applications.

Below are the field names and descriptions:

FIELD NAME	DESCRIPTION
CASE_NUMBER	Unique identifier assigned to each application submitted for processing to the Chicago National Processing Center
CASE_STATUS	Status associated with the last significant event or decision. Valid values include “Certified,” “Certified Withdrawn,” “Denied,” and “Withdrawn”.
CASE_SUBMITTED	Date and time the application was submitted.
DECISION_DATE	Date on which the last significant event or decision was recorded by the Chicago National Processing Center.
VISA_CLASS	Indicates the type of temporary application submitted for processing.

	R = H-1B; A = E-3 Australian; C = H-1B1 Chile; S = H-1B1 Singapore. Also referred to as “Program” in prior years.
EMPLOYMENT_START_DATE	Beginning date of employment.
EMPLOYMENT_END_DATE	Ending date of employment.
EMPLOYER_NAME	Name of employer submitting labor condition application.
EMPLOYER_BUSINESS_DBA	Trade Name or dba name of employer submitting labor condition application, if applicable.
EMPLOYER_ADDRESS	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_CITY	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_STATE	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_POSTAL_CODE	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_COUNTRY	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_PROVINCE	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_PHONE	Contact information of the Employer requesting temporary labor certification.
EMPLOYER_PHONE_EXT	Contact information of the Employer requesting temporary labor certification.
AGENT_REPRESENTING_EMPLOYER	Y = Employer is represented by an Agent or Attorney; N = Employer is not represented by an Agent or Attorney

AGENT_ATTORNEY_NAME	Name of Agent or Attorney filing an H-1B application on behalf of the employer.
AGENT_ATTORNEY_CITY	City information for the Agent or Attorney filing an H-1B application on behalf of the employer.
AGENT_ATTORNEY_STATE	State information for the Agent or Attorney filing an H-1B application on behalf of the employer.
JOB_TITLE	Title of the job.
SOC_CODE	Occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.
SOC_NAME	Occupational name associated with the SOC_CODE.
NAICS_CODE	Industry code associated with the employer requesting permanent labor condition, as classified by the North American Industrial Classification System (NAICS).
TOTAL_WORKERS	Total number of foreign workers requested by the Employer(s)
NEW_EMPLOYMENT	Indicates requested worker(s) will begin employment for new employer, as defined by USCIS I-29.
CONTINUED_EMPLOYMENT	Indicates requested worker(s) will be continuing employment with same employer, as defined by USCIS I-29.
CHANGE_PREVIOUS_EMPLOYMENT	Indicates requested worker(s) will be continuing employment with same employer without material change to job duties, as defined by USCIS I-29.
NEW_CONCURRENT_EMPLOYMENT	Indicates requested worker(s) will begin employment with additional employer, as defined by USCIS I-29.
CHANGE_EMPLOYER	Indicates requested worker(s) will begin employment for new employer, using the same classification currently held, as defined by USCIS I-29.

AMENDED_PETITION	Indicates requested worker(s) will be continuing employment with same employer with material change to job duties, as defined by USCIS I-29.
FULL_TIME_POSITION	Y = Full Time Position; N = Part Time Position
PREVAILING_WAGE	Prevailing Wage for the job being requested for temporary labor condition.
PW_UNIT_OF_PAY	Unit of Pay. Valid values include “Daily (DAI),” “Hourly (HR),” “Bi-weekly (BI),” “Weekly (WK),” “Monthly (MTH),” and “Yearly (YR)”
PW_WAGE_LEVEL	Variables include "I", "II", "III", "IV" or "N/A."
PW_SOURCE	Variables include "OES", "CBA", "DBA", "SCA" or "Other".
PW_SOURCE_YEAR	Year the Prevailing Wage Source was Issued.
PW_SOURCE_OTHER	If "Other Wage Source", provide the source of wage.
WAGE_RATE_OF_PAY_FROM	Employer’s proposed wage rate.
WAGE_RATE_OF_PAY_TO	Maximum proposed wage rate.
WAGE_UNIT_OF_PAY	Unit of pay. Valid values include “Hour", "Week", "Bi-Weekly", "Month", or "Year".
H-1B_DEPENDENT	Y = Employer is H-1B Dependent; N = Employer is not H-1B Dependent.
WILLFUL_VIOLATOR	Y = Employer has been previously found to be a Willful Violator; N = Employer has not been considered a Willful Violator.
SUPPORT_H1B	Y = Employer will use the temporary labor condition application only to support H-1B petitions or extensions of status of exempt H-1B worker(s); N = Employer will not use the temporary labor condition application to support H-1B petitions or extensions of status

	for exempt H-1B worker(s); N/A = not applicable
LABOR_CON_AGREE	Y = Employer agrees to the responses to the Labor Condition Statements as in the subsection; N = Employer does not agree to the responses to the Labor Conditions Statements in the subsection.
PUBLIC_DISCLOSURE_LOCATION	Variables include "Place of Business" or "Place of Employment."
WORKSITE_CITY	City information of the foreign worker's intended area of employment
WORKSITE_COUNTY	County information of the foreign worker's intended area of employment.
WORKSITE_STATE	State information of the foreign worker's intended area of employment.
WORKSITE_POSTAL_CODE	Zip Code information of the foreign worker's intended area of employment.
ORIGINAL_CERT_DATE	Original Certification Date for a Certified Withdrawn application.

6. Exploratory data analysis

Before analyzing the data, several steps such as data cleaning and preprocessing will be carried out to ensure its accuracy. The analysis will involve descriptive statistics, data exploration, and statistical modeling, which will provide valuable insights to achieve the project objective.

The cleaning process involves dropping some rows, such as those with null values and some irrelevant variables. The dependent variable is “case status,” which shows the status of the H1B visa program application. The rest of the variables are covariates. The remaining data records will be used entirely, although the data is extensive and can be computationally time-consuming.

It will also include dropping some variables irrelevant to the analysis since some contain around 30% or less of the total records. When the data is cleaned and usable, it is taken through descriptive statics such as correlation and numerical and graphical summaries.

The dataset will undergo exploratory data analysis to uncover hidden trends and gain valuable insights. This step is crucial as it involves powerful visualizations of variables highly correlated with the response, which can explain the results adequately. These variables will be identified from the correlation chart. Through this process, the data can be effectively explored and analyzed.

After exploration, the dataset is preprocessed for statistical modeling. This involves balancing the data, scaling it, and dividing it into training and testing sets. Three machine-learning algorithms, namely Logistic Regression, Random Forest and Naïve Bayes are then applied to the preprocessed data. The models are trained using the training set and evaluated using the testing set. Performance metrics such as accuracy, precision, recall, f1-score, and confusion matrix are calculated to assess the model's performance. The two algorithms are compared, and the better performing model is chosen to make predictions.

The data is quite extensive, with 624650 observations explained in 53 variables It is of little use in its original State due to missing values and other columns containing unnecessary information and therefore must undergo a thorough cleaning. Several columns were drooped, as well as rows with missing values. The variables that were used for the analysis are: 'CASE_SUBMITTED,' 'DECISION_DATE,' 'CASE_STATUS,' 'EMPLOYMENT_START_DATE,' 'EMPLOYMENT_END_DATE,' 'EMPLOYER_NAME,' 'EMPLOYER_STATE,' 'JOB_TITLE,' 'SOC_NAME,' 'FULL_TIME_POSITION,' 'PREVAILING_WAGE,' 'PW_UNIT_OF_PAY,' and 'WORKSITE_STATE.' These selected variables were enough to give the information needed during data analysis.

6.1 descriptive statistics

```
In [368]: 1 h1b.describe()
```

Out[368]:

	CASE_SUBMITTED	DECISION_DATE	CASE_STATUS	EMPLOYMENT_START_DATE	EMPLOYMENT_END_DATE	EMPLOYER_NAME	EMPLOYER_STATE	
count	517259.00000	517259.000000	517259.000000	517259.000000	517259.000000	517259.000000	517259.000000	517
mean	1062.98593	149.369780	0.125867	1266.450797	1278.619454	31245.366548	27.312919	40
std	127.18870	61.986462	0.331700	166.982262	229.783129	17893.344574	15.890917	22
min	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1051.00000	121.000000	0.000000	1199.000000	1217.000000	15784.000000	11.000000	17
50%	1097.00000	160.000000	0.000000	1323.000000	1362.000000	29285.000000	30.000000	44
75%	1111.00000	175.000000	0.000000	1375.000000	1425.000000	47998.500000	41.000000	60
max	1210.00000	349.000000	1.000000	1497.000000	1547.000000	61782.000000	55.000000	78

Table (6.1) provides a summary of some data in the dataset, including the minimum and maximum values, as well as the mean and standard deviation. It gives a brief overview of the dataset's descriptive statistics.

6.2 Data information

```
#check data information
h1b.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 624650 entries, 0 to 624649
Data columns (total 53 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   Unnamed: 0                    624650 non-null  int64
 1   CASE_NUMBER                   624650 non-null  object
 2   CASE_STATUS                   624650 non-null  object
 3   CASE_SUBMITTED                624650 non-null  object
 4   DECISION_DATE                 624650 non-null  object
 5   VISA_CLASS                    624650 non-null  object
 6   EMPLOYMENT_START_DATE         624621 non-null  object
 7   EMPLOYMENT_END_DATE           624620 non-null  object
 8   EMPLOYER_NAME                 624594 non-null  object
 9   EMPLOYER_BUSINESS_DBA         43270 non-null   object
 10  EMPLOYER_ADDRESS              624643 non-null  object
 11  EMPLOYER_CITY                 624635 non-null  object
 12  EMPLOYER_STATE                624632 non-null  object
 13  EMPLOYER_POSTAL_CODE          624632 non-null  object
 14  EMPLOYER_COUNTRY              528143 non-null  object
 15  EMPLOYER_PROVINCE             6116 non-null    object
 16  EMPLOYER_PHONE                528142 non-null  object
 17  EMPLOYER_PHONE_EXT            27872 non-null   object
 18  AGENT_REPRESENTING_EMPLOYER   528144 non-null  object
 19  AGENT_ATTORNEY_NAME           624650 non-null  object
 20  AGENT_ATTORNEY_CITY           351344 non-null  object
 21  AGENT_ATTORNEY_STATE          336009 non-null  object
 22  JOB_TITLE                     624645 non-null  object
 23  SOC_CODE                      624648 non-null  object
 24  SOC_NAME                      624647 non-null  object
 25  NAICS_CODE                    624643 non-null  object
 26  TOTAL_WORKERS                 624650 non-null  int64
 27  NEW_EMPLOYMENT                624650 non-null  int64
 28  CONTINUED_EMPLOYMENT          624650 non-null  int64
 29  CHANGE_PREVIOUS_EMPLOYMENT    624650 non-null  int64
 30  NEW_CONCURRENT_EMPLOYMENT     624650 non-null  int64
 31  CHANGE_EMPLOYER               624650 non-null  int64
 32  AMENDED_PETITION              624650 non-null  int64
 33  FULL_TIME_POSITION            624645 non-null  object
 34  PREVAILING_WAGE               624649 non-null  float64
```

Table (6.2) provides concise summary of the Data Frame, including the data types of each column, the number of non-null values.

6.3 unique values

```
In [374]:
1 #lets check number of unique values
2 h1b.nunique()
```

Out[374]:

Unnamed: 0	624650
CASE_NUMBER	624650
CASE_STATUS	4
CASE_SUBMITTED	1323
DECISION_DATE	358
VISA_CLASS	4
EMPLOYMENT_START_DATE	1598
EMPLOYMENT_END_DATE	1645
EMPLOYER_NAME	71170
EMPLOYER_BUSINESS_DBA	9441
EMPLOYER_ADDRESS	64378
EMPLOYER_CITY	5123
EMPLOYER_STATE	57
EMPLOYER_POSTAL_CODE	10962
EMPLOYER_COUNTRY	5
EMPLOYER_PROVINCE	313
EMPLOYER_PHONE	81620
EMPLOYER_PHONE_EXT	1461
AGENT_REPRESENTING_EMPLOYER	2
AGENT_ATTORNEY_NAME	6621
AGENT_ATTORNEY_CITY	1081
AGENT_ATTORNEY_STATE	55
JOB_TITLE	93012
SOC_CODE	736
SOC_NAME	867
NAICS_CODE	2810
TOTAL_WORKERS	55
NEW_EMPLOYMENT	50
CONTINUED_EMPLOYMENT	25
CHANGE_PREVIOUS_EMPLOYMENT	19
NEW_CONCURRENT_EMPLOYMENT	10
CHANGE_EMPLOYER	23
AMENDED_PETITION	20
FULL_TIME_POSITION	2
PREVAILING_WAGE	25132

Table (6.3) shows the number of unique values present in each variable of the dataset that is being shown. It also suggests that some variables have a single unique value that is consistent across all rows, while others have exceptional values that differ for each row in the dataset.

6.4 Visa Class & Employer selection

Now we will explore the data variables relevant to the research project, which provided valuable insights and help answer the research questions. The analysis focused on several aspects, including the duration of visa approvals, the time taken for visa approvals, and the relationship between visa approval and job type, industry, employer location, and the number of approvals for different job types. Interactive visualizations were utilized to make it easy to understand and obtain accurate information from the data.

Table (6.4) shows only data related to H1B visas and employers in the United States were considered for the analysis.

In [375]:
```
#DATA CLEANING
#dealing with H1-B visa and Employers from USA only
h1b = h1b[h1b.VISA_CLASS == 'H-1B']
h1b = h1b[h1b.EMPLOYER_COUNTRY == 'UNITED STATES OF AMERICA']
h1b.head()
```

Out[375]:

	Unnamed: 0	CASE_NUMBER	CASE_STATUS	CASE_SUBMITTED	DECISION_DATE	VISA_CLASS	EMPLOYMENT_START_DATE	EMPLOYMENT_END_DATE	EMP
0	0	I-200-16055-173457	CERTIFIED-WITHDRAWN	2016-02-24	2016-10-01	H-1B	2016-08-10	2019-08-10	PF
1	1	I-200-16064-557834	CERTIFIED-WITHDRAWN	2016-03-04	2016-10-01	H-1B	2016-08-16	2019-08-16	[
2	2	I-200-16063-996093	CERTIFIED-WITHDRAWN	2016-03-10	2016-10-01	H-1B	2016-09-09	2019-09-09	TE
3	3	I-200-16272-196340	WITHDRAWN	2016-09-28	2016-10-01	H-1B	2017-01-26	2020-01-25	IN
4	4	I-200-15053-636744	CERTIFIED-WITHDRAWN	2015-02-22	2016-10-02	H-1B	2015-03-01	2018-03-01	C

5 rows × 53 columns

6.5 Data Cleaning

In this section we perform the data cleaning, like checking for the missing values, dropping them, and selecting only relevant variables, After selecting relevant columns, the columns with null value rows are also dropped.

Refer tables (6.5) (6.5.1) below for the same.

In [323]:
```
#checking for missing values
h1b.isnull().sum()[h1b.isnull().sum() > 0]
```

Out[323]:
```
EMPLOYMENT_START_DATE              15
EMPLOYMENT_END_DATE                15
EMPLOYER_NAME                      37
EMPLOYER_BUSINESS_DBA          475714
EMPLOYER_ADDRESS                    2
EMPLOYER_CITY                      10
EMPLOYER_POSTAL_CODE               11
EMPLOYER_PROVINCE              511357
EMPLOYER_PHONE                      1
EMPLOYER_PHONE_EXT             489885
AGENT_REPRESENTING_EMPLOYER         3
AGENT_ATTORNEY_CITY            173864
AGENT_ATTORNEY_STATE           188644
JOB_TITLE                           3
SOC_NAME                            1
NAICS_CODE                          2
FULL_TIME_POSITION                  4
PW_UNIT_OF_PAY                     25
PW_WAGE_LEVEL                   26520
PW_SOURCE                          24
PW_SOURCE_YEAR                     23
PW_SOURCE_OTHER                  4798
WAGE_RATE_OF_PAY_TO                 1
WAGE_UNIT_OF_PAY                    6
H1B_DEPENDENT                       3
WILLFUL_VIOLATOR                    3
SUPPORT_H1B                    316352
LABOR_CON_AGREE                308226
PUBLIC_DISCLOSURE_LOCATION     517351
WORKSITE_CITY                       9
WORKSITE_COUNTY                   973
WORKSITE_STATE                      7
WORKSITE_POSTAL_CODE               15
ORIGINAL_CERT_DATE             474644
```

In [324]:
```
#select relevant columns
select = ['CASE_SUBMITTED', 'DECISION_DATE', 'CASE_STATUS', 'EMPLOYMENT_START_DATE', 'EMPLOYMENT_END_DATE', 'EMPLOYER_NAME',
          'PREVAILING_WAGE','PW_UNIT_OF_PAY','WORKSITE_STATE']
```

In [325]:
```
#subset the columns
h1b = h1b[select]
```

In [326]:
```
#relavant columns with null values
h1b.isnull().sum()[h1b.isnull().sum() > 0]
```

Out[326]:
```
EMPLOYMENT_START_DATE    15
EMPLOYMENT_END_DATE      15
EMPLOYER_NAME            37
JOB_TITLE                 3
SOC_NAME                  1
FULL_TIME_POSITION        4
PW_UNIT_OF_PAY           25
WORKSITE_STATE            7
dtype: int64
```

In [327]:
```
#drop missing values
h1b = h1b[h1b['EMPLOYMENT_START_DATE'].notnull()]
h1b = h1b[h1b['EMPLOYMENT_END_DATE'].notnull()]
h1b = h1b[h1b['JOB_TITLE'].notnull()]
h1b = h1b[h1b['SOC_NAME'].notnull()]
h1b = h1b[h1b['FULL_TIME_POSITION'].notnull()]
h1b = h1b[h1b['PW_UNIT_OF_PAY'].notnull()]
h1b = h1b[h1b['WORKSITE_STATE'].notnull()]
h1b = h1b[h1b['EMPLOYER_NAME'].notnull()]
```

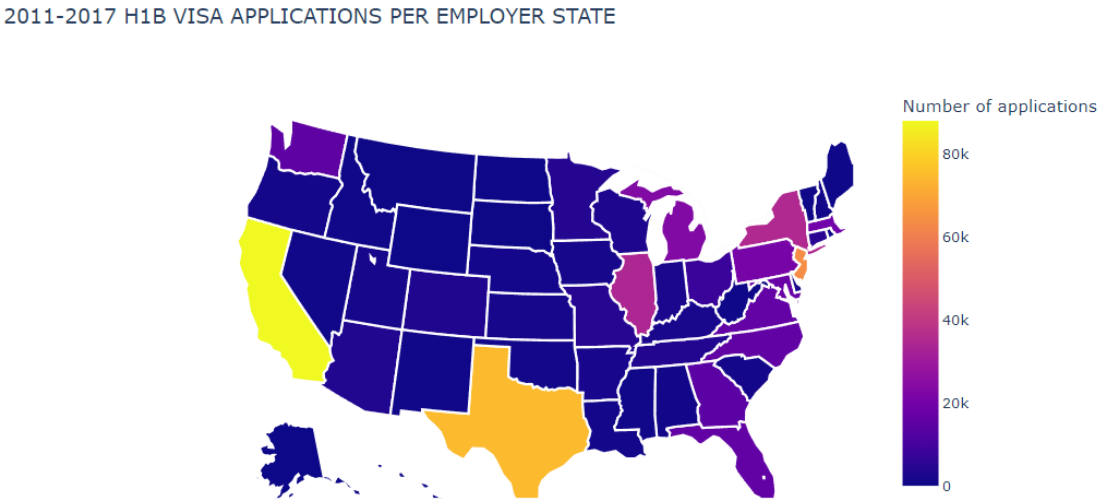
In [328]:
```
#check if missing values were dropped
h1b.isnull().sum()[h1b.isnull().sum() > 0]
```

Out[328]:
```
Series([], dtype: int64)
```

6.6 Applications per state

```
#DATA ANALYSIS
#number of applications per state
h1bst = h1b.groupby('EMPLOYER_STATE',as_index=False).count()[['EMPLOYER_STATE','Count']].sort_values('Count',ascending=False
```

Figure 6.6.1



The above visualization is a distribution map that displays the number of H1B visa applications submitted in various states of the United States of America. The states with the highest number of visa applications are colored yellow, while the states with the lowest number of visa applications are colored blue. The state of California has the highest number of visa applications with 87,914 applicants, followed by Texas with 74,481 visa applications, and New Jersey with the third highest number of applications. This suggests that a significant number of employers are in these states. On the other hand, states such as Alaska, New Mexico, Maine, North Dakota, and others have fewer than 1,000 visa applications, indicating that only a few employers are in these areas. Illinois and New York state have a considerable number of visa applications.

6.7 Leading H1B approved states

The above figure (6.6.1) and table below (6.7) provide insight into the relationship between H1B visa approval and the state where the employer is located, the figure shows that California, Texas, and New Jersey have the highest number of visa applications, which is consistent with the data in the table below. The table only includes certified cases, and it shows that California has the highest number of certified issues with 76,158 out of 87,914 possible applications. Texas follows with 66,855 approved visas out of 74,481 applications. Based on this information, it can be concluded that certain states have a higher likelihood of H1B visa approval than others, which answers Question 6 in third section.

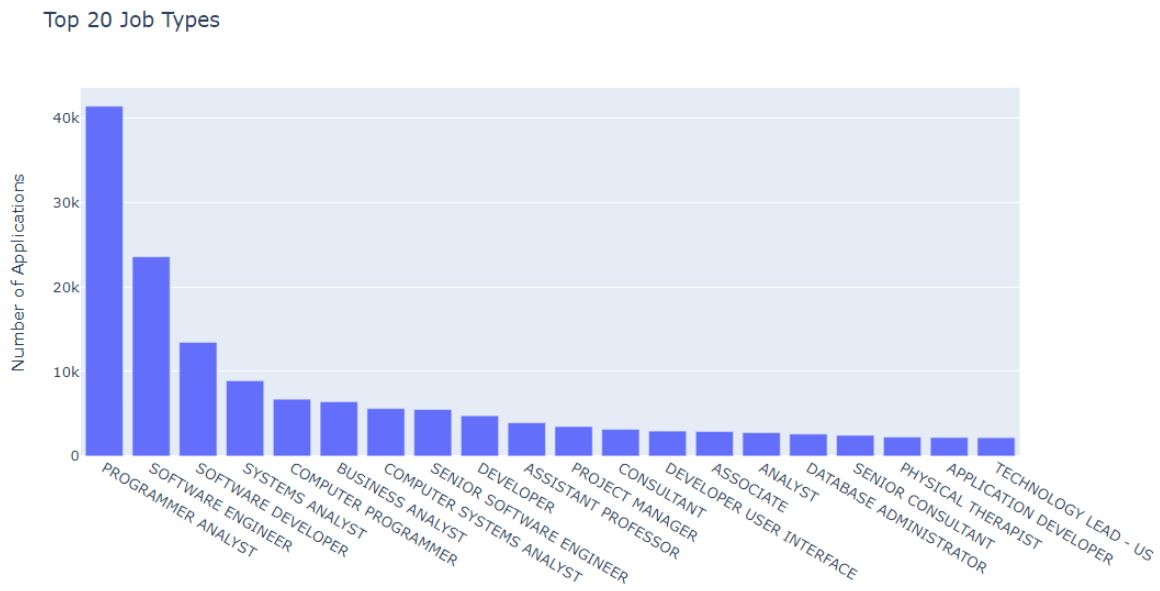
```
#Employer States leading with H1B Visa Approved Applications
cert_loc = h1bcert.groupby('EMPLOYER_STATE',as_index=False).count()[['EMPLOYER_STATE','CASE_STATUS']]
cert_loc.sort_values('CASE_STATUS',ascending= False).head(20)
```

	EMPLOYER_STATE	CASE_STATUS
4	CA	76158
46	TX	66855
33	NJ	56431
15	IL	30807
36	NY	30059
23	MI	20145
40	PA	18542
20	MA	17061
21	MD	16256
9	FL	13894
48	VA	13784
51	WA	13511
29	NC	12910
10	GA	12159
37	OH	6763
6	CT	4056
25	MO	3716
24	MN	3365
45	TN	3263
3	AZ	2866

6.8 Analyzing Job types

```
#Analyzing the job types
#getting the top 20
h1bjob = h1b.groupby('JOB_TITLE',as_index=False).count()[['JOB_TITLE','Count']].sort_values('Count',ascending=False)[0:20]
```

Figure 6.8.1



The above figure (6.8.1) presents the top job types for H1B visa applications, with some job types having more applications than others. The Programmer Analyst job had the highest number of applicants for the visa with 41,369, followed by the Software Engineer job type with 23,562 applications, and Software Developer with 13,443 applications.

The below table (6.8.2) displays the number of certified H1B visas by job industry. It indicates that Software Developers and Applications had the highest number of certified applications at 87,674, followed by Computer Systems Analysts at 67,915, and Computer Programmers at 52,296 certified cases, ranking third. The computing and technology industry dominates the top six positions, which is likely due to technological advancements and the increasing demand for technologically skilled workers globally. This answers Question 7 in third section.

Table 6.8.2

```
#Industries leading with H1B Visa Approved Applications
h1bcert = h1b[h1b.CASE_STATUS == 'CERTIFIED']
cert_ind = h1bcert.groupby('SOC_NAME',as_index=False).count()[['SOC_NAME','CASE_STATUS']]
cert_ind.sort_values('CASE_STATUS',ascending= False).head(20)
```

	SOC_NAME	CASE_STATUS
656	SOFTWARE DEVELOPERS, APPLICATIONS	87674
186	COMPUTER SYSTEMS ANALYSTS	67915
172	COMPUTER PROGRAMMERS	52296
164	COMPUTER OCCUPATIONS, ALL OTHER	38566
185	COMPUTER SYSTEMS ANALYST	13891
662	SOFTWARE DEVELOPERS, SYSTEMS SOFTWARE	13424
438	MANAGEMENT ANALYSTS	9744
10	ACCOUNTANTS AND AUDITORS	9071
507	NETWORK AND COMPUTER SYSTEMS ADMINISTRATORS	8071
466	MECHANICAL ENGINEERS	7507
526	OPERATIONS RESEARCH ANALYSTS	7144
312	FINANCIAL ANALYSTS	6907
216	DATABASE ADMINISTRATORS	5653
449	MARKET RESEARCH ANALYSTS AND MARKETING SPECIAL...	5542
264	ELECTRONICS ENGINEERS, EXCEPT COMPUTER	4957
255	ELECTRICAL ENGINEERS	4706
150	COMPUTER AND INFORMATION SYSTEMS MANAGERS	4706
553	PHYSICIANS AND SURGEONS, ALL OTHER	4348
736	WEB DEVELOPERS	3901
695	STATISTICIANS	3636

6.9 Time Taken based on Job title & Industry.

Table (6.9.1) shows the details of time taken for each case for the approval or rejection.

Table 6.9.1

```
#creating a column for the time taken to approve or reject the H1B visas in days
h1b['Time_Taken'] = (h1b['DECISION_DATE'] - h1b['CASE_SUBMITTED']).dt.days
#creating a column for the length of approval
h1b['Approval_Len'] = (h1b['EMPLOYMENT_END_DATE'] - h1b['EMPLOYMENT_START_DATE']).dt.days
h1b.head()
```

RE_LOCATION	WORKSITE_CITY	WORKSITE_COUNTY	WORKSITE_STATE	WORKSITE_POSTAL_CODE	ORIGINAL_CERT_DATE	Count	Time_Taken	Approval_Len
NaN	RIVERWOODS	LAKE	IL	60015	2016-03-01	1	220	1095
NaN	RIVERWOODS	LAKE	IL	60015	2016-03-08	1	211	1095
NaN	WASHINGTON	NaN	DC	20007	2016-03-16	1	205	1095
NaN	JERSEY CITY	HUDSON	NJ	07302	NaN	1	3	1094
NaN	NEW YORK	NEW YORK	NY	10036	2015-02-26	1	588	1096

Table 6.9.2

```
#Time taken to approve the H1B visas based on Job Type
cert_jt = h1b.groupby('JOB_TITLE',as_index=False)["Time_Taken"].mean().reset_index()[['JOB_TITLE','Time_Taken']]
#20 most job types that took longer for approvals
cert_jt.sort_values('Time_Taken' ,ascending= False).head(20)
```

	JOB_TITLE	Time_Taken
18273	DIRECTOR OF PRODUCT DEVELOPMENT & QUALITY CONTROL	1753.0
11792	CLIENT RELATIONS SPECIALIST	1705.0
45345	PROGRMMER / ANALYST	1266.0
77384	VP; CONS PROD STRATEGY ANALYST IV	1253.0
23370	FINANCE AND OPERATIONS ASSOCIATE (TAX)	1229.0
6308	ASSOCIATE LANGUAGE STRATEGIST	1221.0
24895	GLOBAL CLIENT ANALYST	1213.0
6599	ASSOCIATE PROFESSIONAL SERVICES ANALYST	1205.0
25318	GRADUATE FIRE ENGINEER	1204.0
2062	ANALYST, RESPONSE ANALYTICS	1204.0
49564	RF SYSTEMS ENGINEER IPHONE	1204.0
14359	COORDINATOR, COMMUNITY MINISTRY SERVICES	1202.0
41760	POWER TEST ENGINEER	1202.0
71061	STRUCTURES DESIGN AEROSPACE ENGINEER	1200.0
24002	FOOD SCIENCE RESEARCH ASSOCIATE	1200.0
47967	REFRIGERATION ENGINEER I	1200.0
40045	PEDIATRIC RHEUMATOLOGY FELLOW	1198.0
56130	SENIOR MANAGER, PROJECT MANAGEMENT	1196.0
4054	ASSISTANT DIRECTOR OF ANNUAL GIVING	1196.0
72085	SYSTEMS ADMINISTRATION ENGINEER II	1194.0

The above table 6.9.2 shows the time taken to approve H1B visa-based job title The DIRECTOR OF PRODUCT DEVELOPMENT & QUALITY CONTROL (1753 days) and CLIENT RELATIONS SPECIALIST (1705 days) took longer for their applications to be approved and the rest of the job titles took almost same time for the approval.

Based on the job industry, Table 6.9.3 shows FIRE-PREVENTION AND PROTECTION ENGINEERS (1204 days), MARKET RESEARCH analysts (1131 days), and BIOCHEMICAL ENGINEERS (1101) lead in the time taken to approve the visas as well, due to many applications. The time is calculated in days. These analyses explain the influence of job type and industry on approval rates of H1B visas.

Table 6.9.3

In [400]:
```
#Time taken to approve the H1B visas based on industry type
cert_ind_t = h1b.groupby('SOC_NAME',as_index=False).mean().reset_index()[['SOC_NAME','Time_Taken']]
#industries in which visas took long for approval
cert_ind_t.sort_values('Time_Taken' ,ascending= False).head(20)
```

Out[400]:

	SOC_NAME	Time_Taken
336	FIRE-PREVENTION AND PROTECTION ENGINEERS	1204.000000
475	MARKET RESEARCH ANALYST	1131.000000
75	BIOCHEMICAL ENGINEERS	1101.000000
477	MARKET RESEARCH ANALYSTS & MARKETING SPECIALISTS	1013.000000
591	PHYSICIST	982.000000
416	INDUSTRIAL PRODUCTION MANAGER	978.000000
413	INDUSTRIAL ENGINEERING TECHNOLOGISTS	976.000000
374	GEOPHYSICAL DATA TECHNICIANS	972.000000
401	HOSPITALISTS	965.000000
194	COMPUTER SUPPORT SPECIALISTS	950.500000
514	MEDICAL SCIENTISTS EXCEPT EPIDEMIOLOGISTS	885.000000
76	BIOCHEMIST	874.000000
724	SOFTWARE QUALITY ASSURANCE ENGINEES AND TESTERS	869.000000
772	TRANSPORTATION PLANNERS	691.000000
513	MEDICAL SCIENTIST, EXCEPT EPIDEMIOLOGISTS	666.000000
442	LABOR RELATIONS SPECIALISTS	659.181818
185	COMPUTER PROGRAMMERS R & D	597.800000
532	MOLECULAR AND CELLULAR BIOLOGIST	547.000000
655	SALES REPRESENTATIVE	541.000000
184	COMPUTER PROGRAMMERS NON R & D	523.000000

Thus, the above statistics show that job titles have a weak influence on the approval of an H1B visa since the approval length is almost the same for different job titles, but the industry they work in have the influence of on the time taken for visa approval. This answers Question 4 & 5 in the third section.

6.10 Top Employers

The table below shows the top 30 employers sponsoring the visas based on the visa case status. The case status has four outcomes: withdrawn, denied, certified-withdrawn, and certified. INFOSYS LIMITED leads in the number of Visas with 17,029, and all of them are accredited. This implies that the approval rate for H1B visas for company employees is very high. Tata Consultancy Services Limited follows with 10717 certified keys. CAPGEMINI AMERICA INC. follows third with 7474 certified applications. In the top 20 employers, GOOGLE INC. has the most significant number of H1B visas that were certified but withdrawn. This statistic answers Question 8 in section three employer influence on the approval of an H1B visa.

	EMPLOYER_NAME	CASE_STATUS	Count
33453	INFOSYS LIMITED	CERTIFIED	17029
65876	TATA CONSULTANCY SERVICES LIMITED	CERTIFIED	10717
11766	CAPGEMINI AMERICA INC	CERTIFIED	7474
66080	TECH MAHINDRA (AMERICAS),INC.	CERTIFIED	6876
32156	IBM INDIA PRIVATE LIMITED	CERTIFIED	6517
1034	ACCENTURE LLP	CERTIFIED	5512
18678	DELOITTE CONSULTING LLP	CERTIFIED	5255
23001	ERNST & YOUNG U.S. LLP	CERTIFIED	4994
30012	HCL AMERICA, INC.	CERTIFIED	3680
65333	SYNTEL CONSULTING INC.	CERTIFIED	3669
43911	MICROSOFT CORPORATION	CERTIFIED	3426
75452	WIPRO LIMITED	CERTIFIED	3332
28186	GOOGLE INC.	CERTIFIED	2969
3685	AMAZON CORPORATE LLC	CERTIFIED	2932
14953	COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION	CERTIFIED	1929
38617	L&T TECHNOLOGY SERVICES LIMITED	CERTIFIED	1394
32148	IBM CORPORATION	CERTIFIED	1361
53239	PRICEWATERHOUSECOOPERS ADVISORY SERVICES LLC	CERTIFIED	1341
38989	LARSEN & TOUBRO INFOTECH LIMITED	CERTIFIED	1314
18671	DELOITTE & TOUCHE LLP	CERTIFIED	1286
28187	GOOGLE INC.	CERTIFIED-WITHDRAWN	1280
34151	INTEL CORPORATION	CERTIFIED	1269
36648	JPMORGAN CHASE & CO.	CERTIFIED	1262
23961	FACEBOOK, INC.	CERTIFIED	1165
5173	APPLE INC.	CERTIFIED	1161
53246	PRICEWATERHOUSECOOPERS LLP	CERTIFIED	1110

7. Models and analysis

The analysis method used in this project is a predictive model to predict the CASE_STATUS variable, representing the approval status of the H1B visa applications.

7.1 Logistic Regression

Logistic Regression is a machine learning statistical algorithm that can carry out the classification of categorical data. It was selected because it efficiently

classifies categorical data in the minimum time possible. When used with a high usability dataset, the algorithm also gives high prediction accuracy.

This model was fitted in the dataset using the selected variables, whereby CASE_STATUS was the response, and the rest were the covariates. The dataset was partitioned into training and testing sets, with the training set getting 70% of the observations. The train set was used to train the model so that it can be able to predict unseen data. The testing data was used to evaluate the model's performance.

The CASE_STATUS has four types of values 'CERTIFIED', 'DENIED', 'CERTIFIED WITHDRAWN', 'WITHDRAWN.'

As the no of records with the DENIED status are fewer, to bring the best out of model performance I have considered 'CERTIFIED WITHDRAWN' & 'WITHDRAWN under DENIED category.

In [355]:
```
#data split into train and test sets
X= h1b[['DECISION_DATE', 'CASE_SUBMITTED', 'EMPLOYMENT_START_DATE', 'EMPLOYMENT_END_DATE', 'EMPLOYER_NAME', 'EMPLOYER_STATE'
        'JOB_TITLE', 'SOC_NAME', 'FULL_TIME_POSITION', 'PREVAILING_WAGE', 'WORKSITE_STATE']]
y= h1b['CASE_STATUS']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

In [356]:
```
#fitting a logistic regression model
logit = LogisticRegression(solver='liblinear', C=10.0, random_state=0)
logit.fit(X_train, y_train)
```

Out[356]:
```
LogisticRegression(C=10.0, random_state=0, solver='liblinear')
```

In [357]:
```
#predicting
pred= logit.predict(X_test)
```

In [358]:
```
#classification report
print(classification_report(y_test, pred))
```

```
              precision    recall  f1-score   support

           0       0.90      1.00      0.95    135668
           1       1.00      0.24      0.39     19510

    accuracy                           0.90    155178
   macro avg       0.95      0.62      0.67    155178
weighted avg       0.91      0.90      0.88    155178
```

In [359]:
```
#coefficient and intercept
print(logit.coef_)
```

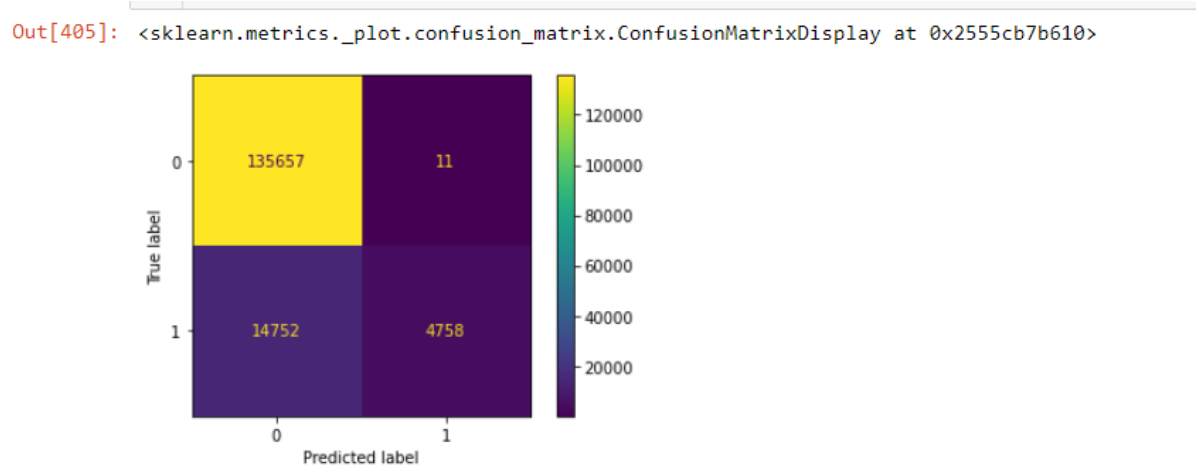
```
[[ 1.52428073e-02 -9.64598759e-03  4.09769890e-03 -1.16897976e-03
   1.38980808e-05  6.57973443e-03  7.48697638e-06  6.30878254e-04
   3.55221024e-04  5.52450343e-05  6.77053431e-03]]
```

In [404]:
```
#coefficient and intercept
print(logit.coef_)
```

```
[[ 1.52428073e-02 -9.64598759e-03  4.09769890e-03 -1.16897976e-03
   1.38980808e-05  6.57973443e-03  7.48697638e-06  6.30878254e-04
   3.55221024e-04  5.52450343e-05  6.77053431e-03]]
```

The Logistic Regression model performed well in predicting the case status of the H1B visa application approval. The prediction accuracy was 0.90. This means that the model can predict the case status with an accuracy of 90%. The model also produced a precision of 0.90 and f1-score of 0.95. This performance by the model shows that the predictors used can significantly explain the response variance (case status). Thus, the factors most influence visa approvals are 'DECISION_DATE', 'CASE_SUBMITTED', 'EMPLOYMENT_START_DATE', 'EMPLOYMENT_END_DATE', 'EMPLOYER_NAME', 'EMPLOYER_STATE', 'JOB_TITLE', 'SOC_NAME', 'FULL_TIME_POSITION', 'PREVAILING_WAGE', 'WORKSITE_STATE'. This helps answer research Questions 1 & 2 in section three.

From the coefficients for FULL_TIME_POSITION a unit increase in FULL_TIME_POSITION is associated with an increase of 0.00036 in the log-odds of the CASE_STATUS being certified, holding all other variables constant, this answers Question 3 in section three.



The confusion matrix shows accurate labels and predicted labels. The response had two titles CERTIFIED (0) and DENIED (1).

7.2 Random Forest

Random Forest is a machine learning technique that involves building an ensemble of decision trees and using them to make a prediction. It is capable of both classification and regression tasks. In this study, it was used as a classifier, and the 'CASE_STATUS' variable was used as the response variable, while the other variables were used as predictors. The Random Forest algorithm is known for its ease of use and flexibility, and it typically achieves high accuracy when trained on an appropriate dataset. The model was trained on a training set and evaluated on a test set by predicting the response variable.

```
In [407]:
#fitting a random forest model
rf = RandomForestClassifier()
rf.fit(X_train, y_train)
```

```
Out[407]: RandomForestClassifier()
```

```
In [408]:
#prediction of the response
pred1= rf.predict(X_test)
```

```
In [409]:
#classification report
print(classification_report(y_test, pred1))
```

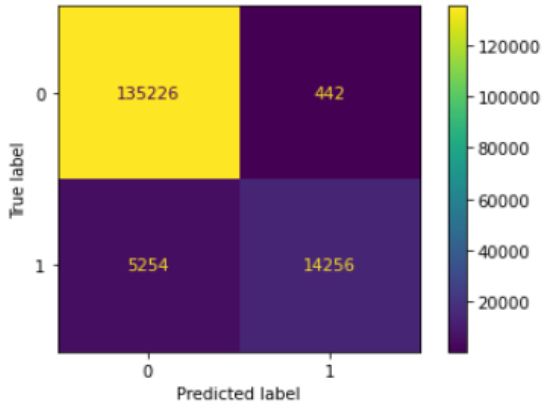
```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98    135668
           1       0.97      0.73      0.83     19510

    accuracy                           0.96    155178
   macro avg       0.97      0.86      0.91    155178
weighted avg       0.96      0.96      0.96    155178
```

The random forest algorithm produced excellent results by scoring an accuracy of 0.96, a precision of 0.96 and f1-score of 0.98. The model performs better than the Logistic Regression model. The accuracy of 0.96 means that the covariates can explain 96% response variability i.e., Case Status

```
Out[410]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x2555cbbb520>
```



The confusion matrix shows accurate labels and predicted labels. The response had two titles CERTIFIED (0) and DENIED (1).

7.3 Naïve Bayes

Naive Bayes is a type of classification algorithm that makes use of the probabilities of certain conditions to determine the probability of an event occurring. In the context of classification, this algorithm can be used to predict the class of a given data point by calculating the probabilities of its features or attributes belonging to each class.

```
Out[413]: GaussianNB()
```

```
In [418]:
1 #prediction of the response
2 pred2= nb.predict(X_test)
```

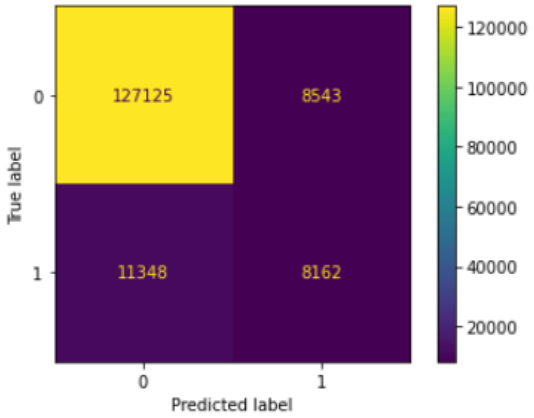
```
In [419]:
1 #classification report
2 print(classification_report(y_test, pred2))
```

```
              precision    recall  f1-score   support

           0       0.92      0.94      0.93    135668
           1       0.49      0.42      0.45     19510

    accuracy                           0.87    155178
   macro avg       0.70      0.68      0.69    155178
weighted avg       0.86      0.87      0.87    155178
```

Out[420]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x256449744f0>



8. Findings and Model Evaluation

From the classification report and the confusion matrix of all the three models the logistic regression model shows that the classification was done with accuracy of 90% and a misclassification error of 0.0951. The Naïve Bayes model performed the worst from in terms of predicting the outcomes. The accuracy from the Naïve Bayes model is 87%, and the misclassification error is 0.1281. The random forest model was the best, with an accuracy of 96%. And misclassification error of 0.03623

9. Conclusion

This research project aimed to analyze the H1B Visa program dataset from 2011 to 2017 to gain useful insights and enhance understanding of the program. The dataset was subjected to data cleaning procedures to eliminate missing values and outliers. performed descriptive statistics and exploratory data analysis to gain insightful information about the dataset, including the correlation between approval length and factors such as job type, job sector, and employer location. The research successfully answered the research questions and achieved its overall goal.

Three machine learning models, namely Logistic Regression, Random Forest, and Naïve Bayes, were utilized to predict the approval status or CASE_STATUS of the H1B visa program dataset. These models displayed good performance with the Random Forest algorithm producing the best results. The prediction accuracy on the test set was 0.90 for logistic regression and 0.96 for the random forest. The misclassification error was found to be 0.095 for the logistic regression model and 0.036 for the random forest model. The high accuracy and low misclassification error of the Random Forest model make it the most suitable algorithm for this dataset. The data analysis presented valuable insights into the H1B visa program and its effects on the US economy.

Examining all variables in the dataset can provide valuable insights into the H1B visa program. Due to the dataset's size, analyzing each variable can be computationally demanding. The random forest classifier took longer than logistic regression and Naïve Bayes, but it provided the best results.

10. python codes

Please refer to the attached files for the code used for the project.

PDF file and ipynb file both are attached.

Project_H1B - Jupyter Notebook.p

Project_H1B.ipynb