A Predictive Machine Learning Model for H1-B Visa Petition Approval

Abhinav Sarkar | Pushpendra Mishra|Tsering Namgail | Rishabh Aggrawal

Sharda University , Greater Noida

Abstract:

In this project, we are going to predict the acceptance of H1-B visa application with the help of previous data. An H-1B is a type of visa that allows an individual , visa certified , to work in USA with a specialty occupation . The rules and regulations defined for H1B visa is that the individual should have a specialty occupation for example he should have deep understanding of the knowledge he should be working in like examples for such fields are biotechnology, computing , software development , medicine , health , chemistry, journalism etc. and it requires a bachelor's degree or it's equivalent as minimum . For the FY 2021 lottery season, 274,237 petitions were received. Since 65,000 are designated for the visa cap, that's a 23 percent chance of selection. However, you need to subtract the 6,800 petitions allocated for Chile and Singapore, so the odds were a bit lower than 23% for that year. Hence it is very needed to analyse previous data for prediction of visa being certified or not and for further decisions about the work , job and settlement in USA. H1B visa is more popular as it gives many more benefits than other visas. Hence it receives almost three times more petitions than allotted. Hence it is less likely to get H1B visa as the process is of randomly selecting the visa petitions to get certified. That's why our model is of more significance as it gives the prediction as well as work environment details and details about opportunities and some valuable information where you are going to work in.

Keywords:

Classification Model , Logistic Regression , Decision Tree , Encoding , Pre-processing etc.

Introduction:

The H-1B visa is a kind of work permit which allows American companies to hire foreign workers instead of American worker if they can not find suitable employee in America. It also allows foreigners to work in America and this visa is easy to be issued and does not involve complicated rules and a foreigner can work in America for a period of at least six years (3 years is the limit with extension of three years) without worry of visa getting expired. The foreigner must know the required knowledge of the field he is going to work in. When applying for an H-1B visa, the applicant is sponsored by the American company that has hired them. The employer company pays the visa fees and submits the required paperwork on behalf of the applicant who is hired by them. Over 2 million visa petitions are filed by the employers each year and only 65000 petitions are approved. So, our goal is to explore the previous data from 2011 to 2016 and by analysing this data extract the information which can help the applicant the most in getting H1B visa and all the other information which can help them in various ways as they are getting hired for rightful pay for their work or not, they can also predict what will be the work environment of their employer company.

Literature Survey:

The dataset that we are studying is available on Kaggle under the name ’H-1B Visa Petitions 2011-2016 dataset’. This dataset is already processed dataset from the original data available on Office of Foreign Labor Certification (OFLC) website. Various data transformations are already performed on this dataset. There are over 3 million rows and 11 columns in the dataset. By Analysing the data, we can get top 10 employers and average salary given by these employers. We can find the average salary offered to all approved certificates. And by this the applicant can find which company he should join and the average salary he can get according to his qualifications. There is various other information that can be analysed from this dataset.

Step 1 for getting H1B visa is to find an H1B sponsor. American company should be sponsoring for you to enter the U.S.

Step 2 involves the submitting of a Labor Conditions Approval (LCA). Once US employer is willing to sponsor you, He will begin the application process by submitting an LCA to the Department of Labor. By this the employer attests to the government that the employee will receive a wage that is similar or greater to other workers in same position in that geographical area.

Step 3 involves the submitting of the form I-129. Once the LCA has been approved, the employer will then file the Petition for a Non-immigrant Worker, by submitting the Form I-129. For this employer will also need various documents of yours such as your resume, proof of your education etc.

Step 4 involves the completion of application at a U.S. embassy or consulate. Once the petition is approved, the last step is for the applicant to process their visa at their home country's U.S. embassy office or consulate.

Dataset:

This dataset contains five years' worth of H-1B petition data. This dataset has 3002458 rows which is over 3 million and 11 columns. The columns are case status, employer name, worksite coordinates, job title, prevailing wage, occupation code, and year filed.

For more information on individual columns, refer to the column metadata. A detailed description of the underlying raw dataset is available in an official data dictionary.

The Office of Foreign Labor Certification (OFLC) collects program data, including data about H1-B visas. The disclosure data is updated annually . This data is available online.

The raw data available is messy and not immediately suitable for analysis. A set of data transformations were performed making the data more accessible for quick exploration.

- EMPLOYER_NAME: Name of employer who submits the application for applicant.

- SOC_NAME: Occupational name associated with the SOC CODE which is an occupational code associated with the job being requested for temporary labour condition.

- JOB_TITLE: Title of the job.

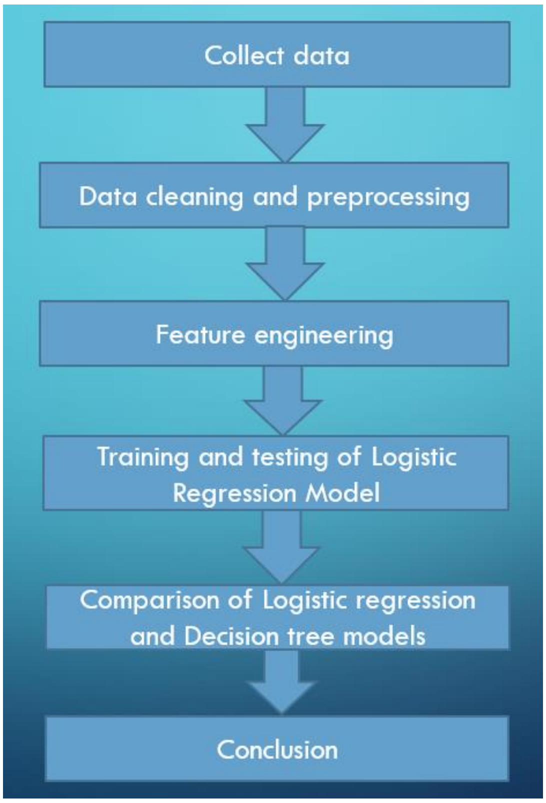
- FULL_TIME_POSTION: This shows if the employee is going to be in a full-time position(Y) or in a part time position(N)
- PREWAILING WAGE: the average wage paid to employees having similar position in the geographical area in which the applicant is going to work in.
- FILING_YEAR: The year in which the petition is filed.
- WORKSITE: City and state in which the applicant is going to work in USA.
- LONGITUDE & LATITUDE: Exact geographical location of the worksite.

Proposed Solution:

Our model is based on Logistic Regression.

Logistic Regression is a Machine Learning method that is used to solve classification issues. It is a predictive analytic technique which relies on the idea of probability. It works on the categorical variables such as True-False, normal-abnormal, yes-no, success-failure, 0-1 etc. This technique predicts the likelihood of a categorical dependent variable. For example, it predicts the probability of yes or no. It also can give the classification of more than 2 valued categorical dependent variables but these variables have to be turned into binary valued categorical variables. For example, of such variables pick one vs. rest.

The goal of Logistic Regression is to discover a link between characteristics and the likelihood of a specific outcome.


Collect data
Data cleaning and preprocessing
Feature engineering
Training and testing of Logistic Regression Model
Comparison of Logistic regression and Decision tree models
Conclusion

We are also trying to use Decision tree as a predictive model and we will be comparing it with our logistic regression model afterwards .

Decision Tree is a technique which has various applications in several different areas. This technique can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions. Prediction is based on leaf nodes. This technique starts with root node and after that leaf(decision) nodes are formed by splitting the dataset. After that the decision nodes are formed from the previous nodes by recursively splitting. The recursive splitting happens till all formed decision nodes give no value to tree.

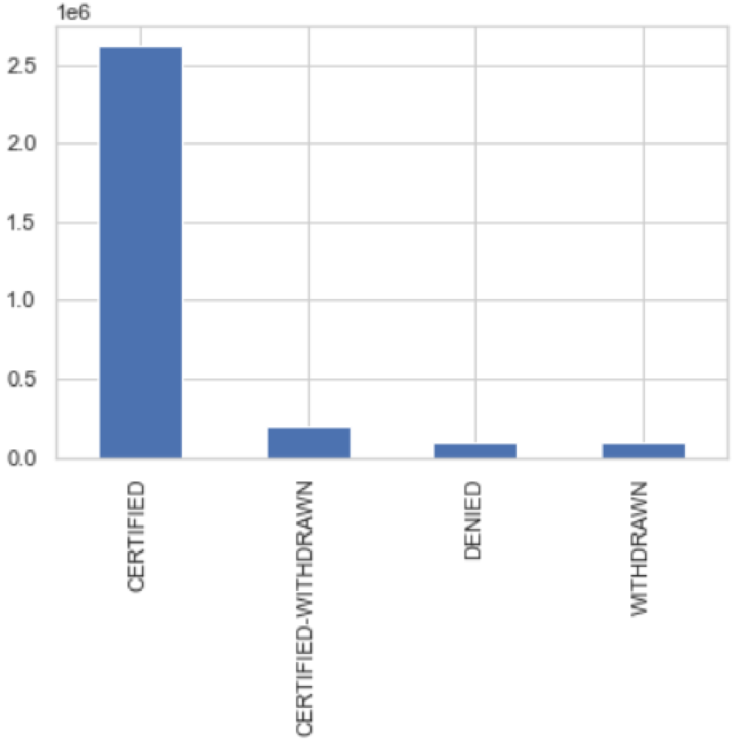
Feature Engineering, Pre-processing and Analysis of the dataset:

As we want to use this dataset to find if the petition is going to be certified or not. We have to make our dataset clean , error free and machine understandable. Hence we will remove the unneeded columns , outliers , and rows with empty cells. Here is the analysis of the data and conclusions based on it are given below.

Univariate Analysis:

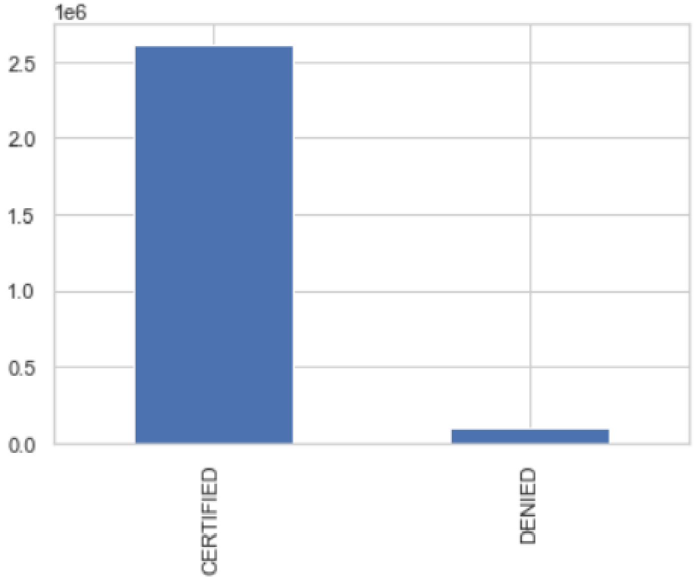
CASE_STATUS:

Here is a graph showing the four categories of applicants getting certified, certified yet withdrawn, Denied and withdrawn. (In graph 1e6 means *10^6)



Conclusion: in four years over 2.5 million applicants were certified and less than 0.5 million people were denied. This shows a very large disparity between two categories that we want to predict that is certified and withdrawn so it will be very difficult in pre-processing the data. We will be removing all other categories as other categories are not needed in this prediction model.

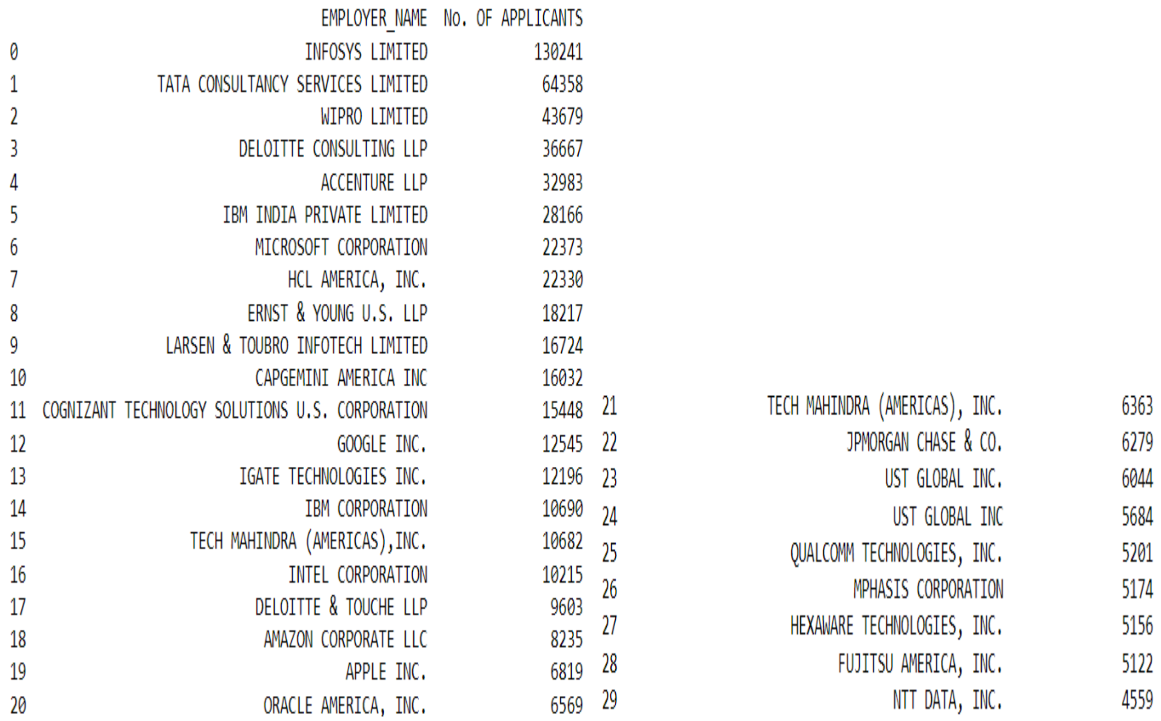
After removing all other categories graph looks like this



Still there is a large disparity in number of certified and denied applicants we will try to remove it in further process.

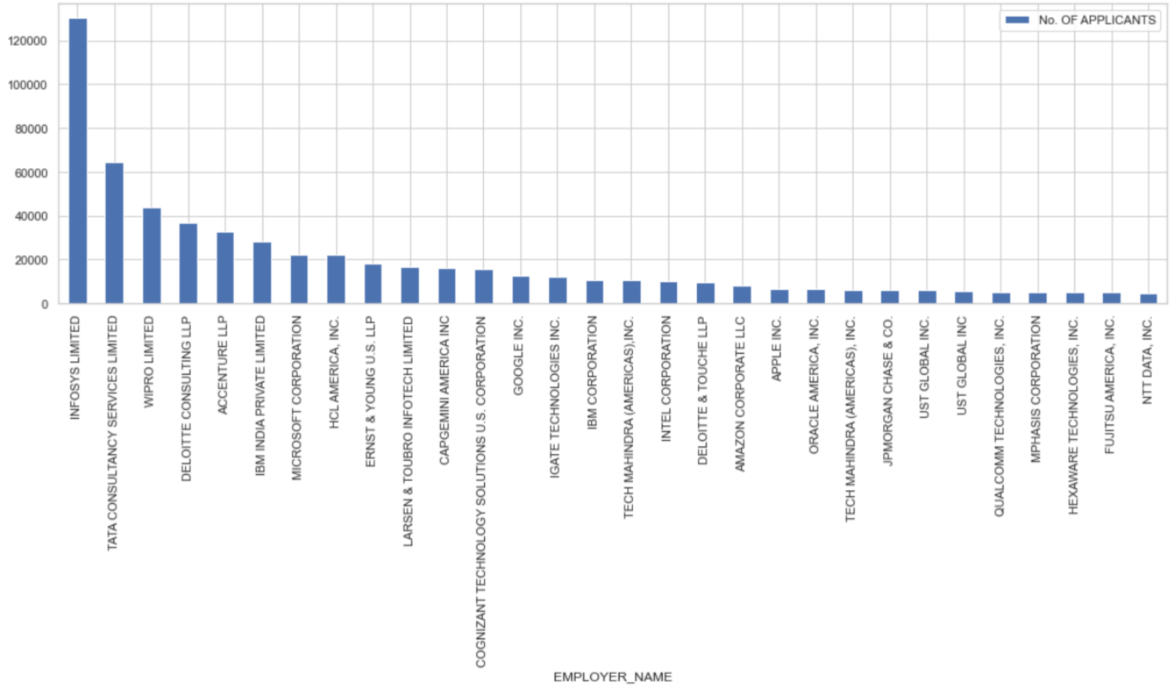
EMPLOYER_NAME:

Here is the list of top 30 employers with number of applicants.



```
                                  EMPLOYER_NAME  No. OF APPLICANTS
0                              INFOSYS LIMITED             130241
1             TATA CONSULTANCY SERVICES LIMITED              64358
2                                WIPRO LIMITED              43679
3                       DELOITTE CONSULTING LLP              36667
4                                ACCENTURE LLP              32983
5                    IBM INDIA PRIVATE LIMITED              28166
6                        MICROSOFT CORPORATION              22373
7                           HCL AMERICA, INC.              22330
8                          ERNST & YOUNG U.S. LLP              18217
9                 LARSEN & TOUBRO INFOTECH LIMITED              16724
10                         CAPGEMINI AMERICA INC              16032
11  COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION              15448
12                                   GOOGLE INC.              12545
13                       IGATE TECHNOLOGIES INC.              12196
14                               IBM CORPORATION              10690
15                   TECH MAHINDRA (AMERICAS),INC.              10682
16                             INTEL CORPORATION              10215
17                          DELOITTE & TOUCHE LLP               9603
18                          AMAZON CORPORATE LLC               8235
19                                    APPLE INC.               6819
20                          ORACLE AMERICA, INC.               6569
21                  TECH MAHINDRA (AMERICAS), INC.               6363
22                           JPMORGAN CHASE & CO.               6279
23                              UST GLOBAL INC.               6044
24                               UST GLOBAL INC               5684
25                   QUALCOMM TECHNOLOGIES, INC.               5201
26                          MPHASIS CORPORATION               5174
27                    HEXAWARE TECHNOLOGIES, INC.               5156
28                        FUJITSU AMERICA, INC.               5122
29                               NTT DATA, INC.               4559
```

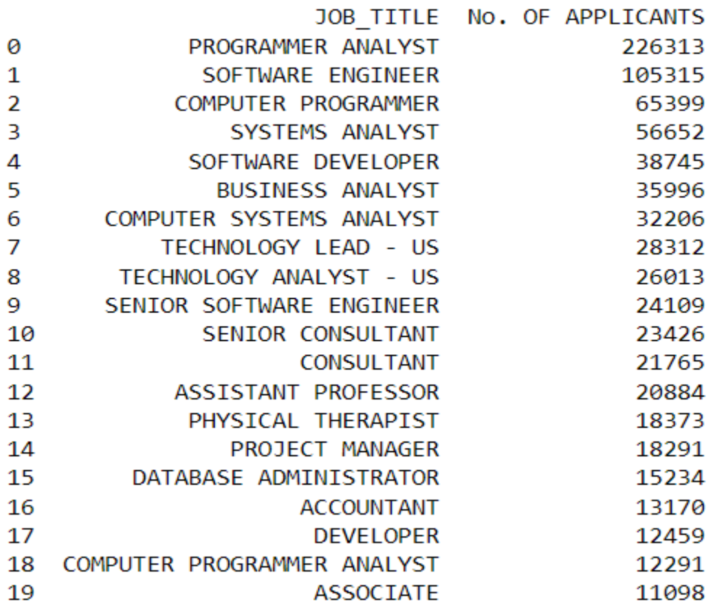
Here is the visual representation of above list which makes further things clear.



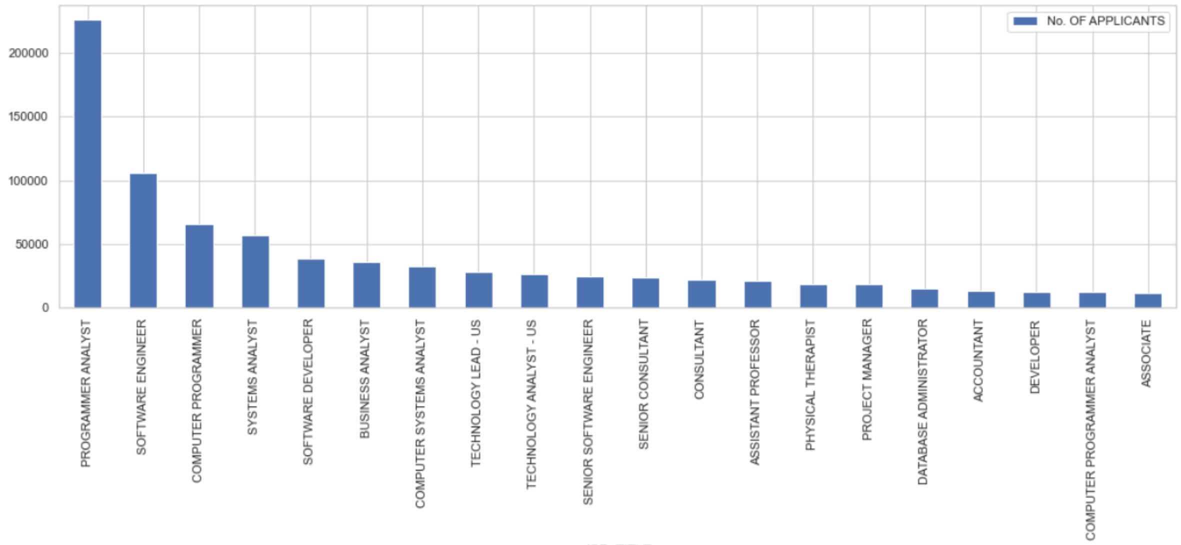
Conclusion: The graph and the list shows that Infosys Limited was the employer that petitioned highest number of applicants.

JOB_TITLE:

Here is a list of top 20 jobs with respect to number of applicants.



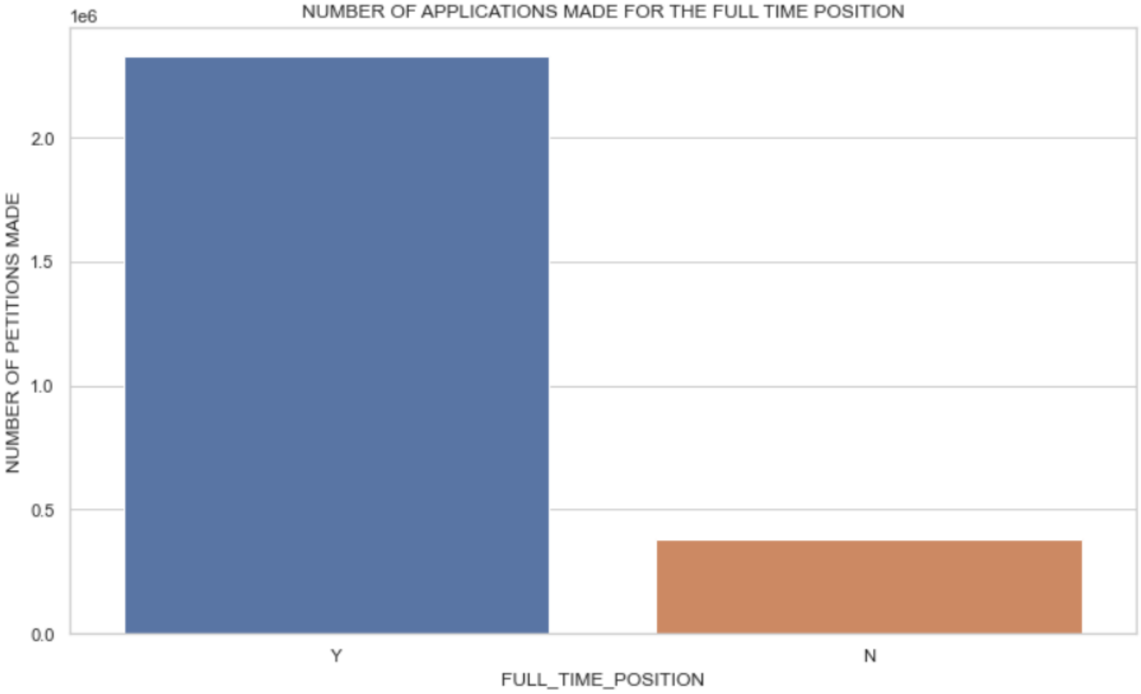
```
                       JOB_TITLE  No. OF APPLICANTS
0               PROGRAMMER ANALYST             226313
1                SOFTWARE ENGINEER             105315
2              COMPUTER PROGRAMMER              65399
3                  SYSTEMS ANALYST              56652
4               SOFTWARE DEVELOPER              38745
5                 BUSINESS ANALYST              35996
6         COMPUTER SYSTEMS ANALYST              32206
7              TECHNOLOGY LEAD - US              28312
8           TECHNOLOGY ANALYST - US              26013
9          SENIOR SOFTWARE ENGINEER              24109
10               SENIOR CONSULTANT              23426
11                      CONSULTANT              21765
12             ASSISTANT PROFESSOR              20884
13              PHYSICAL THERAPIST              18373
14                 PROJECT MANAGER              18291
15          DATABASE ADMINISTRATOR              15234
16                      ACCOUNTANT              13170
17                       DEVELOPER              12459
18   COMPUTER PROGRAMMER ANALYST              12291
19                       ASSOCIATE              11098
```



Conclusion: Above graph represents that Programmer Analyst is the highest preferred job for Employer companies.

FULL_TIME_POSITION:

Here is the graph which shows the number of petitions filed for Y and N category :



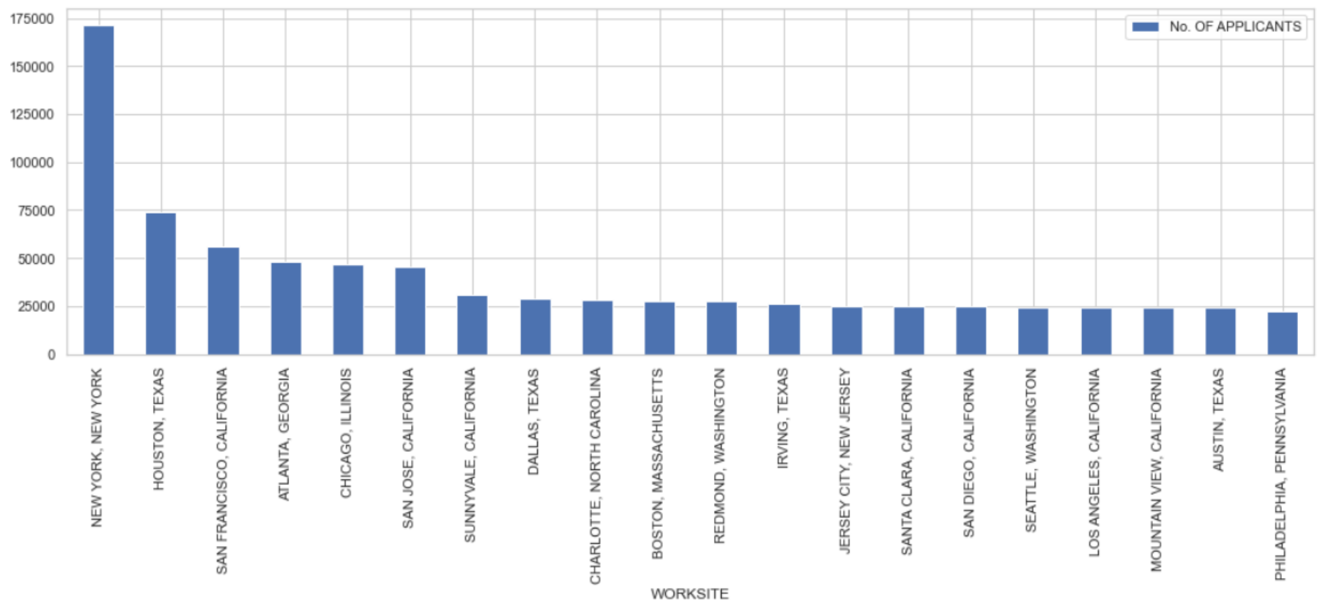
PREVAILING_WAGE:

Average salary offered from the employers to employee is 148152.30

And the 60000.0 was offered in highest frequency.

WORKSITE:

Here is the graph shows the No. of applicants for the top 20 worksites in USA.



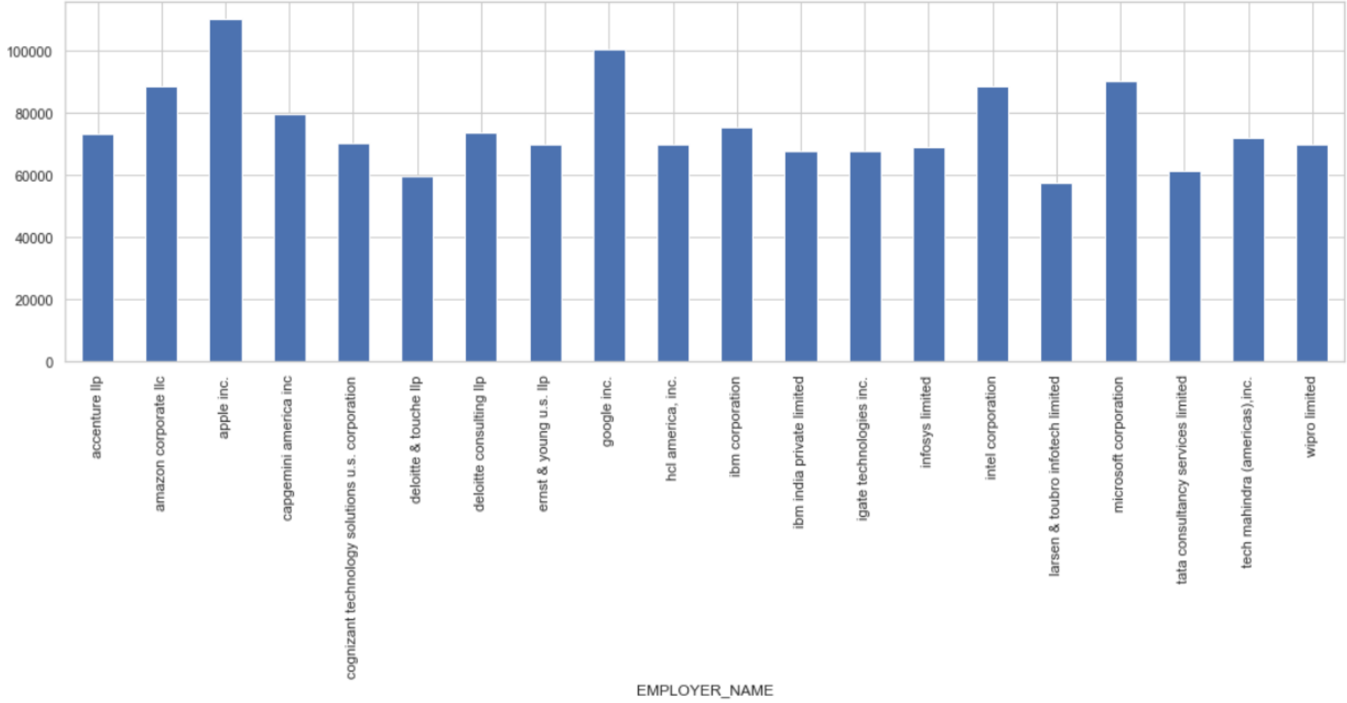
Now we will remove the unwanted columns and fill the null values with mode of the columns as mode is robust and it is not affected by outliers.

YEAR:

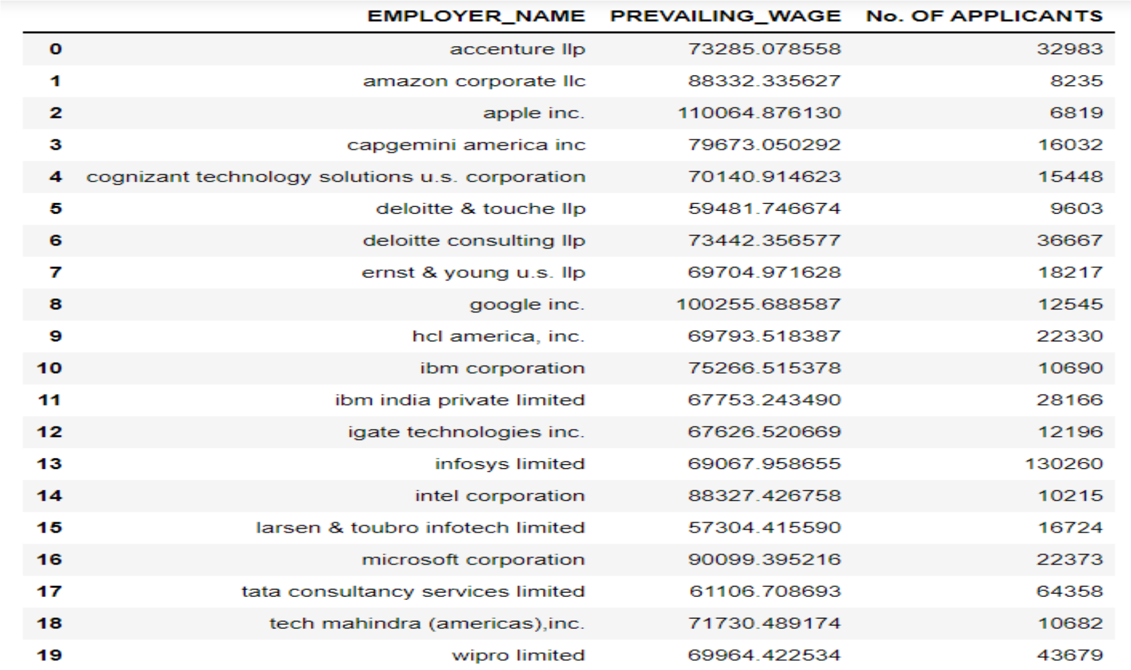
No. of petitions are increasing year by year. we don’t need figures or graphs for that as this trend will go on for decades.

Bivariate Analysis:

Here is the top employers vs. their average prevailing wage graph:

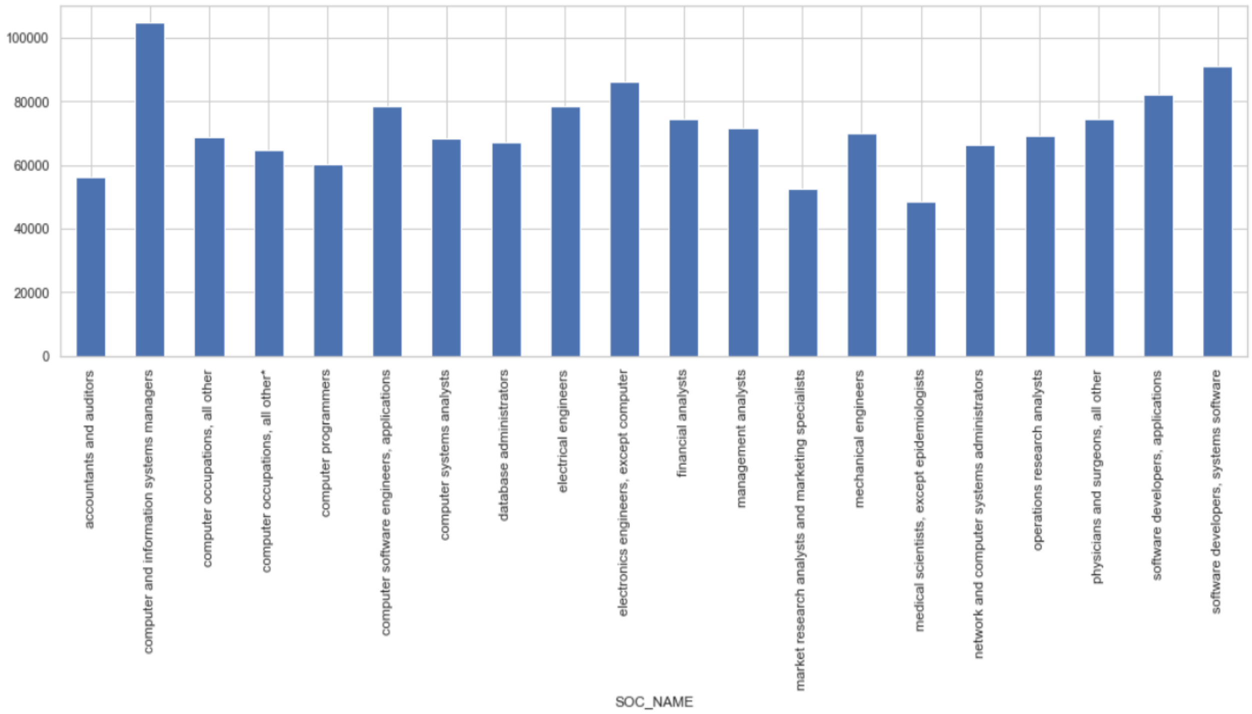


conclusion: This graph shows the top 20 employers and average salary given by them in which Apple Inc. gives the highest average salary to its hired employees. Google is the third and Amazon is the second at giving the highest average salaries to its employees. Here is the list of top 20 companies who hires the most with the average salary given by them.

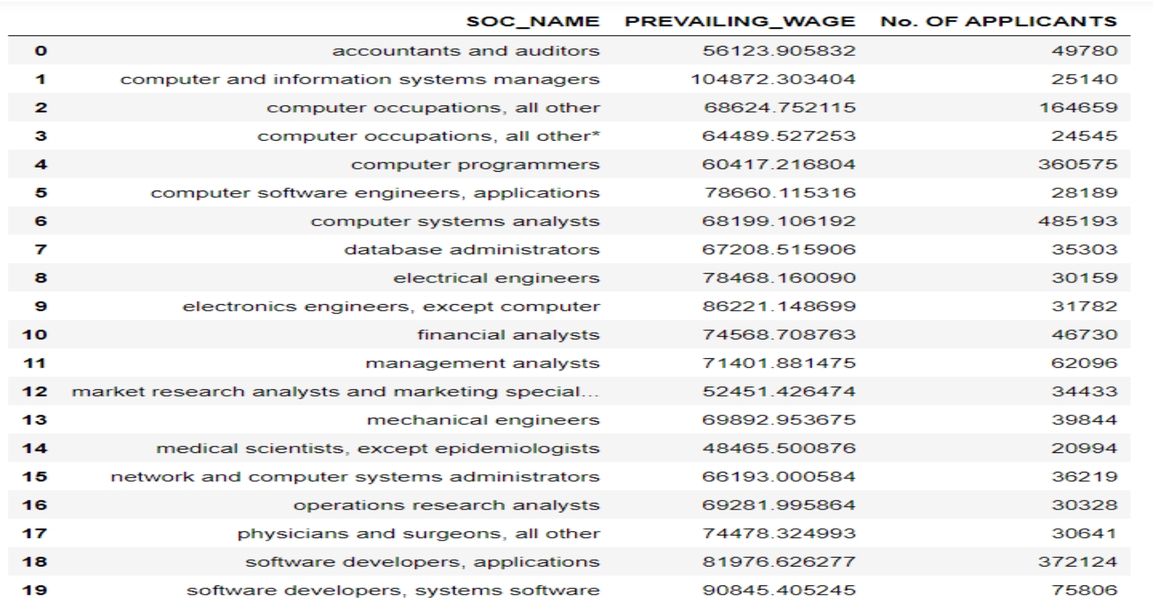


	EMPLOYER_NAME	PREVAILING_WAGE	No. OF APPLICANTS
0	accenture llp	73285.078558	32983
1	amazon corporate llc	88332.335627	8235
2	apple inc.	110064.876130	6819
3	capgemini america inc	79673.050292	16032
4	cognizant technology solutions u.s. corporation	70140.914623	15448
5	deloitte & touche llp	59481.746674	9603
6	deloitte consulting llp	73442.356577	36667
7	ernst & young u.s. llp	69704.971628	18217
8	google inc.	100255.688587	12545
9	hcl america, inc.	69793.518387	22330
10	ibm corporation	75266.515378	10690
11	ibm india private limited	67753.243490	28166
12	igate technologies inc.	67626.520669	12196
13	infosys limited	69067.958655	130260
14	intel corporation	88327.426758	10215
15	larsen & toubro infotech limited	57304.415590	16724
16	microsoft corporation	90099.395216	22373
17	tata consultancy services limited	61106.708693	64358
18	tech mahindra (americas),inc.	71730.489174	10682
19	wipro limited	69964.422534	43679

Here is another graph showing the top 20 SOC_NAME for which visa petitions were petitioned with average prevailing wage offered.

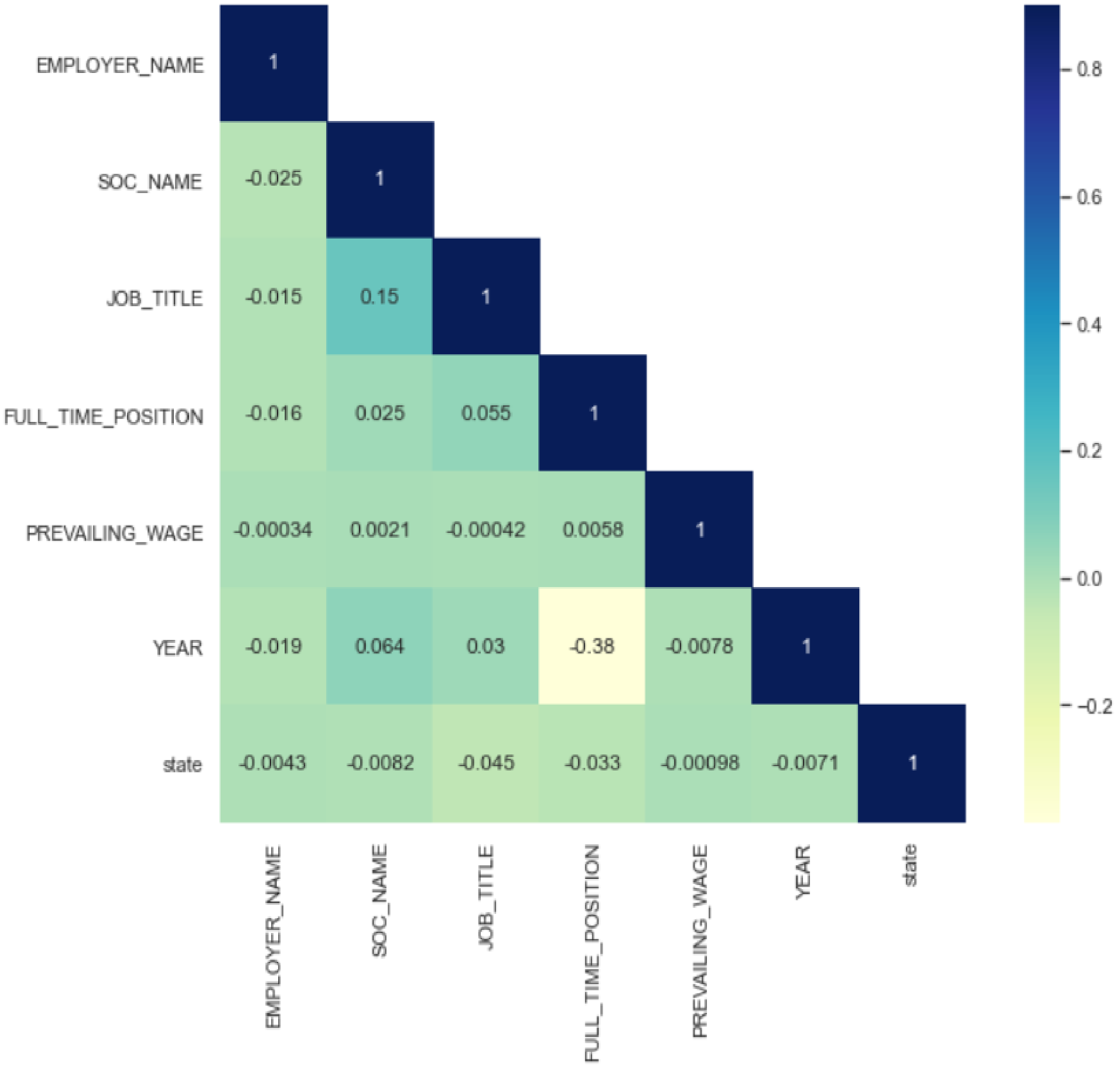


Conclusion: highest average salary is offered to computer and information systems managers SOC_NAME. this means this job is in high demand with highest average salary in USA. On second place software developers and system software and on third place electronic engineers. Hence it shows that IT professionals are in high demand in US with highest average salary offered. Here is the list top 20 SOC_NAME in demand with average prevailing salary.



	SOC_NAME	PREVAILING_WAGE	No. OF APPLICANTS
0	accountants and auditors	56123.905832	49780
1	computer and information systems managers	104872.303404	25140
2	computer occupations, all other	68624.752115	164659
3	computer occupations, all other*	64489.527253	24545
4	computer programmers	60417.216804	360575
5	computer software engineers, applications	78660.115316	28189
6	computer systems analysts	68199.106192	485193
7	database administrators	67208.515906	35303
8	electrical engineers	78468.160090	30159
9	electronics engineers, except computer	86221.148699	31782
10	financial analysts	74568.708763	46730
11	management analysts	71401.881475	62096
12	market research analysts and marketing special...	52451.426474	34433
13	mechanical engineers	69892.953675	39844
14	medical scientists, except epidemiologists	48465.500876	20994
15	network and computer systems administrators	66193.000584	36219
16	operations research analysts	69281.995864	30328
17	physicians and surgeons, all other	74478.324993	30641
18	software developers, applications	81976.626277	372124
19	software developers, systems software	90845.405245	75806

Correlation of labels(columns) is given by the graph given below—



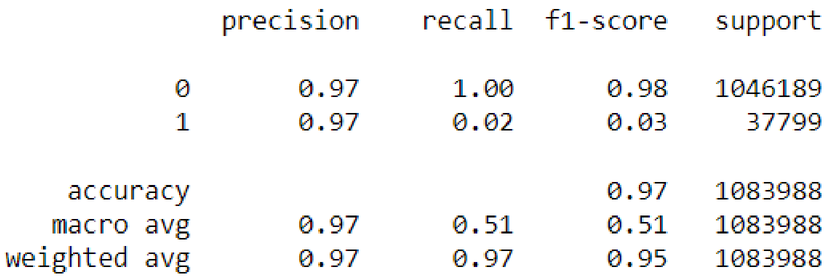
The graph is showing the correlation among the columns . Its shows how the value in one column affects the value in another column if changed and how much it can affect.

Correlation is the first priority in finding out that how much predictive modelling will be successful on the given data it gives us the primary features which are most reliable in predicting the outcome.

We have used label Encoder and one Hot encoder for changing the dataset into machine understandable dataset . as Machine only understands in the values of 0 and 1 we have to change the data accordingly after the pre-processing and one hot encoding, we split the dataset train and test dataset we used these split datasets in training and testing of our model.

Result:

After training our model it gave the accuracy of 96.57% . here is the confusion matrix of our model:



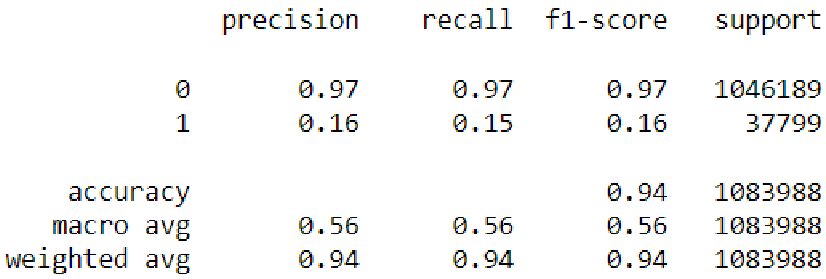
```
              precision    recall  f1-score   support

           0       0.97      1.00      0.98   1046189
           1       0.97      0.02      0.03     37799

    accuracy                           0.97   1083988
   macro avg       0.97      0.51      0.51   1083988
weighted avg       0.97      0.97      0.95   1083988
```

We also trained a Decision Tree model for comparison purpose so that we can find out which machine learning model is better for such problems.

The accuracy for decision tree came out to be 94.2% . here is the confusion matrix for decision tree model:



```
              precision    recall  f1-score   support

           0       0.97      0.97      0.97   1046189
           1       0.16      0.15      0.16     37799

    accuracy                           0.94   1083988
   macro avg       0.56      0.56      0.56   1083988
weighted avg       0.94      0.94      0.94   1083988
```

As the accuracy of the logistic regression model is better hence it is more suited for such problems than decision tree.

Conclusion:

In this paper , we have proposed a logistic regression model by which we can predict the h1b-visa petition can be certified or denied. We have also analysed the data thoroughly as this data gives a deep information on the jobs which are in high demand in USA, the company which hires most number of employees and average wages given by these companies and also the average wages accordance to the soc_name or job title. By analysing this data a person can understand if the job is suitable for him or not, the company is giving him proper wages or not and he should apply for h1b visa or not that’s why we tried to explore the data as much as possible. In the end we also tried to compare logistic regression model with decision tree model which gives us understanding on both machine learning techniques. In the end logistic regression is more suitable for this type of problem than decision tree.

References:

1. https://www.kaggle.com/code/dpandya18/h1b-visa-status-prediction/data
2. https://www.datacamp.com/community/tutorials/predicting-H-1B-visa-status-python
3. http://journalstd.com/gallery/49-july2021asd.pdf
4. http://www.ijirset.com/upload/2018/october/37_A%20Predictive.pdf
5. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
6. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/
7. sp04.pdf (scsug.org)