Work Visa in the U.S.A, a real lottery?

Ugo LABBE

1 Problem Understanding

1.1 Dataset

The chosen dataset is about H-1B, H-1B1, E-3 Visa Petitions in 2022. These 3 visas are work visa, where companies fill a petition in order to make a foreigner come to the U.S.A to work for them.

It was found on Kaggle[1] but the original data was delivered by the Office of Foreign Labor Certification[2].

1.2 Motivation

As a previous resident of the U.S.A I was interested in working there for a bit, however work visas are really hard to find, even considered as impossible to get. Indeed there is a cap of work visas delivered every year but after that cap, companies wishing to get more foreign employees can enter the 'H1B Lottery'. There is 65000 work visas delivered per fiscal year, as well as 20000 for masters degree owner. Researcher, teacher and some other specific type of employment are found to have no limit.[3]

My goal is then to see if there is a clear difference between certified and denied Visa as well as to see if we can predict if a visa will be approved or not.

The full dataset contain 626084 observations and 12 variables, with informations relatives to the Visa, its status, the company that filled the application, the job title and information about it (wage, full time position, location).

In most of the analysis provided, I've decided to exclude Withdrawn visas. (576221 Certified, 3096 Denied).

2 Data Understanding

The full dataset contain 626084 observations and 12 variables, with informations relatives to the Visa, its status, the company that filled the application, the job title and information about it (wage, full time position, location), the quarter when the petition was filled and the case status.

In most of the analysis provided, I've decided to exclude Withdrawn visas. (576221 Certified, 3096 Denied).

[1] https://www.kaggle.com/datasets/jishnukoliyadan/lca-programs-h1b-h1b1-e3-visa-petitions?resource=download

[2] https://flag.dol.gov/programs/lca

[3] https://www.uscis.gov/working-in-the-united-states/temporary-workers/h-1b-specialty-occupations-and-fashion-models/h-1b-cap-season

Later we will focus on Visas only concerning jobs in the the field of data, in order to narrow down the analysis.

3 Data Preparation

Multiple modification have been done on the dataset in order to help the analysis.

In a first time, the Prevailing was normalized to an annual wage, as the unit of pay was varying from Hour, Bi-weekly, Weekly , Monthly and Annually.

Then all categorical variables have been put to lowercase in order to limit the number of different modalities (e.g: Amazon and amazon).

In order to better analyze the location the location was split into city and state, going from the form : New York, New Jersey to 2 different columns, one for the state one for the city.

To later conduct an analysis on data focused job and create predictive machine learning models I also created, using all informations available a new dataset including:

- The quarter when the petition was filled
- The information on if the job was full time or not
- The normalized Prevailing wage (between 0 and 1, using a scaler)
- The ratio of denied visa of the specific job
- The ratio of denied visa of the specific company
- the ratio of denied visa of the specific state
- The Case Status, the label to predict

Each ratio have been computed by :

$$\frac{Denied}{(Denied + Certified)}$$

4 Modeling

4.1 Descriptive Statistics

When digging into this kind of huge dataset, it's always interesting to observe the most reoccurring modalities.

As we have information about the jobs, the companies as well as the location, we can be interested into knowing which is filling the most petitions in 2022. This rank is summarized in the following table.

Rank	**Company**	*petition*	**Position**	*petition*	**State**	*petition*
1	Amazon	16304	Software Engineer	36515	California	132749
2	Cognizant	13906	Software developer	25415	Texas	70671
3	Google	11945	Sr. software engineer	9754	New Jersey	61757
4	Microsoft	11547	Manager	5531	Washington	43639
5	Tata	11172	Assistant Professor	5320	New York	43275

With this table it appear clearly that IT related jobs are the most petitioned, and that the big tech companies, part of the GAFAM[4] (or not) are the one petitioning the most.

Looking at other available information:

- 98.67% of petition were for full time jobs
- The median prevailing wage for Denied Visa is $89,824.5while the prevailing wage for Certified Visa is $98,842, a clear different of almost $10,000.
- On the denied visa, 5,75% were part time jobs, while on the certified visa, only 1,28% were.

4.2 Data visualisation

To visualize the data. the location was used to illustrate some informations about the visas, the frequency of certified visa per state1 as well as the ratio of denied visa per state2.

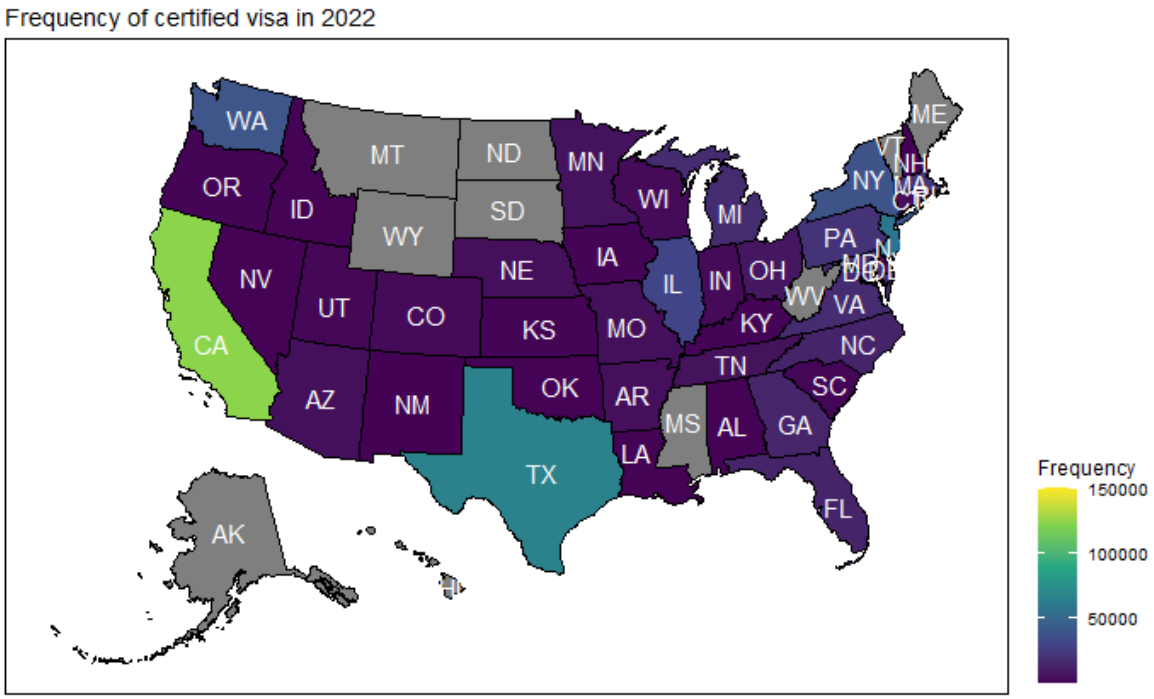

Figure 1: Frequency of Certified Visa per state

On Figure 1 We can clearly see that California is way ahead of everything else, but it isn't surprising as it contains most tech companies, which seems to be the one petitioning the most to bring foreign workers. Other populated states appears to also have quite a lot of petition, such as Texas and New Jersey. The USA North Mid West seems to have close to none petition but it isn't really surprising as they are the least populated states[5] (in grey):

- Wyoming (WY): Least populated state
- Montana (MT): 8th least populated
- North Dakota (ND): 4th least populated
- South Dakota (SD): 5th least populated

[4] https://en.wikipedia.org/wiki/Big_Tech
[5] https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population

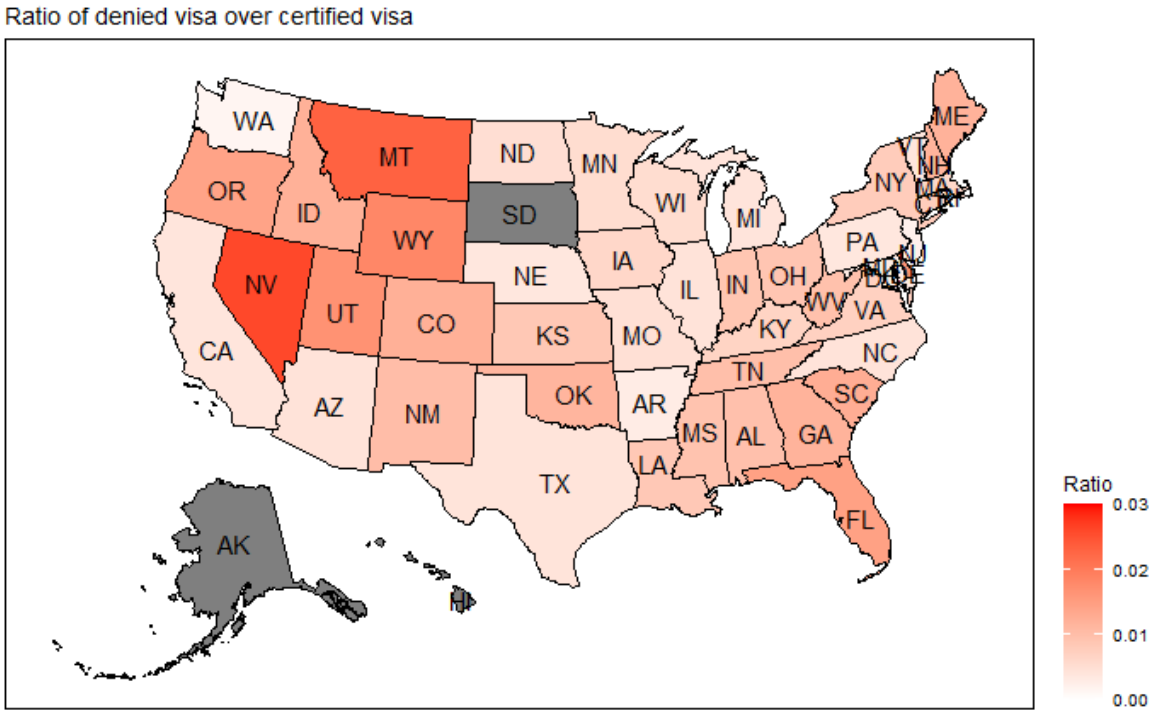

Figure 2: Ratio of Denied Visa per state

The Figure 2 is a bit less expected, as we can see that states like Nevada and Montana have the highest ratio of denied Visa. On the other way wen can see that Washington have a really low denial rate.

4.3 Predictive Models and Machine Learning techniques

By narrowing down the dataset to only include jobs related to data (found by using the job title variables), I wanted to see if it was possible to predict the outcome of a visa application with information about the company, the state, the job (salary, full time position or not and ratio of denial).

This new dataset contains 34606 observation, of which only 102 are denied visas, a really unbalanced dataset then.

In order to see if a visual data separation was possible, a Principal Component analysis[6] (PCA) was conducted3.

[6] `https://en.wikipedia.org/wiki/Hyperparameter_optimization`

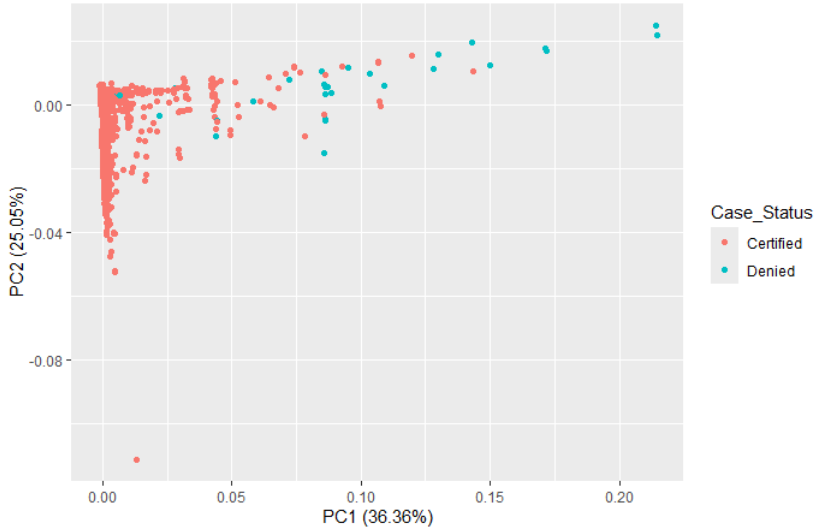

Figure 3: PCA of the "data job" dataset

The PCA is explaining 61.41% of the total variance and we can see some sort of separation between Certified and Denied Visa

After doing some data preparation, a train set and test set were created, with respectively 80% and 20% of the total dataset. The train set was then re scaled to have balance between the classes.

2 Different classifiers (Support Vector Machine[7] and K-nearest neighbor[8]) were trained on the data, using special hyper parameters chosen using a grid search method[9] and cross validation technique[10].

4.4 Evaluation of the classifiers

Each classifer were trained on the train data (80% of the dataset, rescaled) and tested on the test data (20% of the dataset), with 6838 Certified visas and 13 Denied.

After getting the result of each classifier, a confusion matrix can be computed[11]. From this matrix we can then compute different metrics in order to assess the quality of the results. As our data are highly unbalanced, I chose to evaluate the 2 models using the F1-measure:

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

I also chose to compute the accuracy as a second measure:

$$\frac{Correct\_Predictions}{All\_predictions}$$

[7] https://en.wikipedia.org/wiki/Support_vector_machine
[8] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[9] https://en.wikipedia.org/wiki/Hyperparameter_optimization
[10] https://en.wikipedia.org/wiki/Cross-validation_(statistics)
[11] https://en.wikipedia.org/wiki/Confusion_matrix

4.4.1 Support Vector Machine (SVM)

Pred/Ref	**Certified**	**Denied**
Certified	6496	0
Denied	342	13

We can see that this algorithm is struggling more to identify Denied Visa because it actually predicted 342 certified visa into the wrong class, resulting into worse scores. 95% for the F1-measure and an accuracy of 97.43%. However it didn't misclassified any Denied visa.

4.4.2 K-nearest neighbor

The K-nearest neighbor actually predicted the classes correctly resulting in an accuracy of 100% and a F1-measure of 100%

5 Conclusion

From the different anlysis we can draw some conclusions, first, on the kind of job and companies filling the most petitions: Big tech companies, such as Amazon or Google. Mostly for full time jobs with a quite high wage (around $90,000 annualy). Most of them being in highly populated area, such as California. Highest denial rates were found surprisingly in states like Nevada, Montana or even Florida.

As on if the work visas are actually distributed by a lottery system, this is not what we observed on our predictive models as they were able to classify quite well between a certified visa and a denied visa using information about the filling company, the job title and job information as well as the state where the petition was filled.

Some more information on the dataset would have been nice to explore, such as information of the applicant (nationality, age, sex) but these informations are private hence not usable.