Welcome to H1B and PERM prediction page!

What is this web page predicting?

If someone would like to apply for h1b or green card, his application would be firstly censored by labor apartment to make sure that he is qualified for his future jobs, which is what we are handling. His application would be sent to USCIS only after certified by labor apartment, and the final censorship is confidential. In another word, we are exactly predicting is one important prerequisite of h1b and green card, a simplified version of h1b/perm prediction.

How should I start predicting?

You don't need to log in or download anything. If all you want is a quick result, just put in your information about your job and then click the "Start/Reset prediction" button, then you can get results that if you would be certified and the probability as well.

Should I upload any personal information?

Absolutely not. All needed information for predicting is relative to your job.

I was confused about some of the options on the prediction page. Can you help to explain?

H1B part

- **Select your OES Wage level** identifies your OES job level. If you don't have an OES certification yet, choose "UNKNOWN".
- **Is there a dependent for your employer** identifies if your employer is H-1B Dependent.
- **Are you a Willful violator** identifies if your employer has been previously found to be a willful violator.

PERM part

- **Is it a refile?** identifies if your application was previously filed.
- **Select OES skill level** identifies for your OES job level. If you don't have an OES certification yet, choose "UNKNOWN".
- **Does foreign worker has ownership interest?** identifies if the foreign worker has ownership interest or familial relationship with the Employer.
- **Select minimum education acceptable** identifies the minimum U.S. diploma or degree required by the employer for the position. Valid values include “None”, “High School”, “Associate’s”, “Bachelor’s”, “Master’s”, “Doctorate", and “Other”.
- **Is training required?** identifies whether or not training is required for the job.
- **Is the alternative field acceptable?** indicates if an alternate field of study will be acceptable for education requirement.
- **Are job requirements normal?** indicates if the job opportunity’s requirements are normal for the occupation being offered.
- **Is foreign language required?** indicates if knowledge of a foreign language is required to perform the job duties.
- **Is it for a professional occupation?** identifies whether or not the application is for a professional occupation, other than a college or university professor. N = Application is not for a professional occupation other than a college or university professor.
- **Do you have any former admissions** indicates the current visa status of the foreign worker. If you do not have any current visa status, choose "UNKNOWN". For more information, please view ETA-9089, Page 7 Section J, Item 8.
- **Is the training done?** indicates whether the foreign worker completed the training required for the requested job opportunity. Y = The foreign worker completed the training. N = The foreign worker did not complete the training. UNKNOWN = Not applicable.

What if my prediction result is 'Denied'?

Our predictions are based on history data and machine learning and would not be 100% correct. Getting a 'denied' result does **NOT** mean that you would be denied in the real certifications. The real certification is much more complex, and you would be asked for much more personal information once you apply. The final result would be strongly relevant to the current policies, which is not included in our model. We also provide some useful information that you could try to improve your certification probability.

What else could I get from this web page?

If you know some data science, you can also enjoy exploring data from the charts we prepared for you or upload your own h1b/perm data to train your unique model for predicting. See details in the following sections.

Upload your own data to build a model

In the update h1b/perm dataset pages, you are allowed to upload your own datasets to build unique models.

What dataset should I upload?

First you have to view the offcial page of US department of labor, click 'disclosure data' to find history datasets. You can download any datasets for H1B part or PERM part after year 2015 **(Don't change the filenames)**. Then, you can go through the datasets and do a selection on **rows**, like, selecting data form March or April, Selecting data only from software engineers, etc. Do not change the column names or any other data in the datasets.

How to build a model and apply it for prediction?

After getting your unique dataset, you can then upload all your selected data and click 'start processing'. After completing upload, you can go 'start training'. Finally, you could choose to use 'user-defined' model in the 'model to use' dropdown on homepage for the corresponding part.

Explorary Data Analysis

We have prepared some interesting charts and plots for you to know more about labor certifications on H1B and PERM.

Looking for more?

We have done our work on kaggle with more EDAs and models for both H1B and PERM part. Enjoy!

How we build our H1B model

How we did feature selection

There are hundreds of features for appliers to submit for each certification. However, most of them are either irrelevant to the result(large p-value) or have high correlation(like your prevailing wage and wage level). So first we group the features with high correlations and then do feature engineering separately.

Our principles:

1.Select features with small P-value.

2.Select features with good interpretability.

3.Cluster the features and see how could we describe a cluster comprehensively with less features.

4.If more than one features are describing the same aspect of a cluster, try to use the one with less correlation with other clusters.(e.g. wage level or annual income).

5.Try to use the features that are stably shown in the reports of each year if possible.

6.Try to use the features that are **NOT** personal information.(This one is hard to apply so we can only choose features that are not so 'personal')

How we build our model

First we have to decide the data we use. Surprisingly we found that the more history data we use, the worse results we get. This is because that the policies are changing fast, as well as some features are missing in history datasets. So we have to mark them as ‘unknown’ which negatively affects the model.

Then we choose to use only 2020 dataset and do model selection. Set certified = positive and denied = negative, we concern on negative predictive value (NPV), and found that Logistic Regression gives better results than others.If you want to know more about it, you can go kaggle notebook and see how we build our model with more detailed information.

How we build our PERM model

How we did feature selection

There are hundreds of features for appliers to submit for each certification. However, most of them are either irrelevant to the result(large p-value) or have high correlation(like your prevailing wage and wage level). So first we group the features with high correlations and then do feature engineering separately.

Our principles:

1.Select features with small P-value.

2.Select features with good interpretability.

3.Cluster the features and see how could we describe a cluster comprehensively with less features.

4.If more than one features are describing the same aspect of a cluster, try to use the one with less correlation with other clusters.(e.g. wage level or annual income).

5.Try to use the features that are stably shown in the reports of each year if possible.

6.Try to use the features that are **NOT** personal information.(This one is hard to apply so we can only choose features that are not so 'personal')

How we build our model

First we have to decide the data we use. Surprisingly we found that the more history data we use, the worse results we get. This is because that the policies are changing fast.

Then we choose to use only 2020 dataset and do model selection. Set certified = positive and denied = negative, we concern on negative predictive value (NPV), and found that random forest gives better results than others. If you want to know more about it, you can go kaggle notebook and see how we build our model with more detailed information.

Contact us

The source code can be found in Github here.
If you have any more questions, please contact Yichun Sun or Liang Xue .