Visa Project notes

Permanent Wage level Determination – PW level 9089 has the highest number of nulls – might be useful in terms of predicting if person gets accepted or how fast based on PW wage level. Also good because it classifies the type of work into 4 categories

Unfortunately this is not related to the class of visa awarded – Usually these are awarded EB2 or EB3’s. Level 2 and above could qualify for EB2

*1) Why the null values?*

*2) Not sure if the null values correspond to EB1’s , 2’s and 5’s*

Class of Admission – might not need this column but can keep it for now

*1) Why null values*

Wage level usually determines if person get EB2 or EB3

**Permanent Workers**

Approximately 140,000 immigrant visas are available each fiscal year for aliens (and their spouses and children) who seek to immigrate based on their job skills. If you have the right combination of skills, education, and/or work experience and are otherwise eligible, you may be able to live permanently in the United States. The five employment-based immigrant visa preferences (categories) are listed below.

| **Preferences** | **General Description** | **Labor Certification Required?** |
| --- | --- | --- |
| [**First Preference EB-1**](https://www.uscis.gov/node/41759) | This preference is reserved for persons of extraordinary ability in the sciences, arts, education, business, or athletics; outstanding professors or researchers; and multinational executives and managers. | No |
| [**Second Preference EB-2**](https://www.uscis.gov/node/41726) | This preference is reserved for persons who are members of the professions holding advanced degrees or for persons with exceptional ability in the arts, sciences, or business. | Yes, unless applicant can obtain a national interest waiver (See the [“Labor Certification”](http://www.uscis.gov/tools/glossary/labor-certification) page for more waiver information.) |
| [**Third Preference EB-3**](https://www.uscis.gov/node/42265) | This preference is reserved for professionals, skilled workers, and other workers. (See[Third Preference EB-3](https://www.uscis.gov/node/42265) page for further definition of these job classifications.) | Yes |
| [**Fourth Preference EB-4**](https://www.uscis.gov/node/41502) | This preference is reserved for “special immigrants,” which includes certain religious workers, employees of U.S. foreign service posts, retired employees of international organizations, alien minors who are wards of courts in the United States, and other classes of aliens. | No |
| [**Fifth Preference EB-5**](https://www.uscis.gov/node/48483) | This preference is reserved for business investors who invest $1 million or $500,000 (if the investment is made in a targeted | No |

**Level 1:** Entry/beginning level

* Degree & experience at or below range

**Level 2:** Qualified/competent in basic tasks

* Degree or experience at low end of range

**Level 3:** Experienced/exercise judgment

* Degree or experience at higher end of range AND supervisory requirements or special skills

**Level 4:** Senior/management

* Degree & experience outside the range

Distribution of people who apply for visas (Across US)

Distribution of people who apply for visas (Across US)

Immigration reform

Number of applications processed has been increasing over lst couple of years.

No of apps processed was at an all time low in 2013 and quite low in 2008-2009.

Do we have missing data in 2013 that is accounting for the all time low – No data was processed over all 12 months.

Was it low in 2008-2009 because of the financial crisis? Unemployment rate was at an all time high, so maybe many companies were not sponsoring visas and they did not receive many applications?

http://data.bls.gov/timeseries/LNS14000000

Percentage of applications accepted has climbed over last 8 years from 81 to 93%

Economic growth of country?

Number of applications processed has almost tripled from 2013 to 2015 – Trend in 2016 is continuing to grow –

Do more applications get denied in the later part of the year than the first part as spots fill up? - STATs

Other qns

Very little data was processed in 2009 July – wonder if there was some renovation in the offices then?

Or were the numbers low in general over previous years?

Predicting the time it takes to process a permanent work visa

The goal of this project was to predict the time it takes to process a work visa application in the US.

This is useful for individuals applying for a job to plan their job search process, and for companies to better tailor their hiring process.

I downloaded the data from the US Dept. of Labor Forces (USDLF) [here](https://www.foreignlaborcert.doleta.gov/performancedata.cfm). I then proceeded to:

1. Clean the data
2. Explore and Visualize the data to gain some insights
3. Run some statistical tests to see if certain features had an impact on the processing time
4. Use Linear Regression to predict the number of days to process a visa
5. Use Logistic Regression to predict whether the processing time fell within a certain rage, and compare the performance to Linear Regression

I am currently working on:

1. Implementing other Machine Learning Algorithms like Random Forests
2. Trying to find better features that might improve the predictions
3. Adding more data
4. Developing a web app, so that a potential visa applicant can key in his information and get a predicted wait time

Let’s have a look at the data from 2015.

There were 125 columns, most of which were not relevant to this analysis. Some examples of these less relevant columns were the information of the attorney who filled out the visa application, or the name of the newspaper with the job posting. I kept 13 columns I felt were most relevant, and can include additional ones if needed based on the results of the analysis. The columns are as follows:

CASE\_NUMBER: Unique Identification number for each applicant. (e.g. A-14220-96665 )

‘A’ – Atlanta Processing Center; ‘14’ (last 2 digits of year application was created) 220 (julian date application was created)

DECISION\_DATE: Date decision was reached on status of application

CASE\_STATUS: ‘Certified’ and ‘Certified-Expired’ : Application was accepted. ‘Denied’ : Application denied. ‘Withdrawn’ : Application withdrawn by applicant

CASE\_RECEIVED\_DATE: Date application was received by Processing Center

EMPLOYER\_NAME: Name of employer

PW\_JOB\_TITLE\_9089: Type of job (Falls into categories like Software Developer, Pharmacists, etc)

PW\_LEVEL\_9089: Level of experience of applicant. Level 1: Entry level; Level 2: Qualified; Level 3: Experienced; Level 4: Fully Competent

WAGE\_OFFER\_FROM\_9089: Wage offer

WAGE\_OFFER\_UNIT\_OF\_PAY\_9089: Hourly / Weekly / Bi-weekly/Monthly/ Yearly

JOB\_INFO\_WORK\_CITY: City of Employment

JOB\_INFO\_WORK\_STATE: State of Employment

COUNTRY\_OF\_CITIZENSHIP: Citizenship of Applicant

CLASS\_OF\_ADMISSION: Immigration visa the worker currently holds

I also included a column ‘TIME\_TAKEN’, that has the number of days it took to process the application.

Cleaning the data:

There were a total of 98453 applications processed in 2015.

First, I drop any rows with null values.

That leaves me with 85556 Applications, still a large dataset.

Second, I drop all cases where the application was ‘Withdrawn’ because these were incompletely processed and it doesn't make sense to take them into account.

So all in all I have 81740 Applications.

Out of them, 77253 applications were accepted and 4487 applications were denied.

*Note about null values:* Most of the null values come from incomplete Wage Levels and Class of Admission. It might be that individuals who did not hold an immigrant visa in the US did not fill out the ‘Class of Admission’ section. However, there were many other individuals who filled out ‘Not in USA’, so since this is unclear I dropped those values.

Data Exploration: Countries

The first set of question I was really interested in was:

1. What are the nationalities of the people who are applying for jobs, and how are the applications distributed?
2. Where are the most number of applications coming form?
3. Does it take longer to process applications from countries that have more number of people applying?

More Data Cleaning: Countries

I wanted to visualize this on a world map. To display the number of applications by country on a world map – I needed the 3 digit country codes. I used the ISO 3166 library of country codes.

Some of the countries that applicants said they were from did not exist in the ISO 3166 library (e.g. Yugoslavia, which split after 1991). Moreover, there were some differences in spelling (e.g. Macau / Macao) which had to be corrected. I did this by correcting each country that threw up an error, but there might be more efficient ways to do this.

Data Visualization: Countries

This world map shows the number of applications by country. Most of the applications come from citizens of India, followed by citizens of China. It also looks like there is either no data, or there are no applications from Central Africa. You can hover over the map to get the name of the country and the number of applications.

These are the top 8 countries applicants come from. India takes up a large percentage, followed by China and South Korea, then Canada, Mexico, Philippines, UK and Taiwan.

This map shows the mean time taken to process applications from each country (Hover for standard deviations). It does not look like the number of days needed to process an application (green world map) depends on the number of applications from that country (purple world map).

However, some countries have very large standard deviations, implying that there might be outliers. Let’s see if the median processing time depends on the number of applications. The median times are definitely more consistent than the mean times, however, it still does not look like the processing time depends on the number of applications.

Statistical Analysis: Countries

To confirm this, I calculate the correlation co-efficient. The correlation coefficient is 0.0185 which shows that there is no correlation between time taken and number of applications.

You can see this better when looking at the data on a scatter plot. I split the data up into two scatter plots so it is easier to see that there is no trend. The first plot has all countries with <1000 applications, and the second has all countries with >1000 applications.

Data Exploration: States

The permanent work visa applications are filed only after receiving job offers from employers in the US. The second set of questions I was really interested in was:

1. How are the job offers distributed across the 50 US states?
2. Which states are most of the job offers coming from?
3. Does it take longer to process applications in states where more number of people are receiving offers?

More Data Cleaning: States

I corrected any spelling mismatches in the states, and dropped all territories to make plotting the maps easier.

Data Visualization: States

This US map shows the number of jobs by state. Most of the job offers are from California, followed by Texas. You can hover over the map to get the name of the state and the number of applications for jobs.

These are the top 9 states people get job offers from. California and Texas take up large percentages, followed by New Jersey, New York, Illinois, Massachusetts, Washington, Virginia and Michigan.

This map plots the mean time taken to process applications for each state (Hover for standard deviations). Again, to ensure that outliers are not affecting the data too much, I also plot the median time for each state. The median times look pretty similar to the means, so we will focus on the mean times for each state.

It looks like the time taken to process applications for certain states might be higher because there are more job offers? For example, the number of job offers in Texas is high, and the processing time for those offers is also high. The opposite is true for Wyoming – where there were only 20 job offers, and the mean processing time is low.

Statistical Analysis: Countries

You might be able to see this better when looking at the data on a scatter plot – it looks like there might be a slightly positive trend. To quantify the trend, I calculate the correlation co-efficient. The correlation coefficient is 0.1913, which shows that there is a small correlation between the time taken and number of jobs in each state.

Data Exploration: Level of Experience

From the previous set of analyses, we know that where a person is from, or where he has a job offer, do not have a significant impact on processing time. I wondered if other factors, like the level of experience the person has, might affect the processing time. The question I try to answer here is:

1) Does increased experience mean faster processing time?

Statistical Analysis: Level of Experience

To answer this question, I grouped the data into low and high levels of experience, where low experience is defined as anyone who fell into the Level 1 and 2 categories (under PW\_LEVEL\_9089), and high is anyone who fell into Levels 3 and 4.

The box plot below shows the distributions of processing times for low and high levels of experience. The mean processing time was smaller for higher levels of experience than lower levels of experience. Moreover, even though the maximum time taken was larger for higher levels of experience, the third quartile was still smaller. This shows that at least 75% of the data falls within a lower processing time range when the level of experience is higher.

I then performed a t-test with unequal variance to determine quantitatively if an increased level of experience meant that the application was processed faster.

Null Hypothesis: There is no difference in processing time for applicants with a lower or higher level of experience.

Alternative Hypothesis: There is a shorter processing time for applicants with a higher level of experience, than those with a lower level of experience.

The p value of 2e-23 suggests that we can reject the null hypothesis.

Higher experience does mean a faster processing time and the 2 distributions are significantly different.

Machine Learning: Predicting the time it takes to process the visa.

Let's try to predict how long it will take for a person’s visa application to be processed, using ML algorithms. I will train the model on these features:

1. nationality
2. the state their job is in
3. level of experience
4. visa class they are currently on.

Even though feature 1 did not have a significant impact on the processing time (as explored above), we only looked at the correlation coefficient. I am going to include this feature for now, in case there is something in the data I did not capture in the exploration phase.

Linear Regression

First, I am going to try using linear regression to predict the exact number of days. There are a total of 50 states, 137 countries, 4 levels of experience and 48 visa classes. I use one hot encoding to pre-process my data since these are all categorical features.

The SSE is very high and the R2 value is highly negative showing that the prediction by this model is very poor.

It could be that I am using too many categories in each of my features to train my model (50 states, 137 countries, 4 levels of experience and 48 visa classes)

I am going to reduce the number of categories by grouping all the categories of each feature into 4 categories: countries, states and visa classes where the processing time was < 200 days; 200-300; 300-400 or > 400 days. I picked these 4 categories based on the distribution I observed from the histogram below.

I run linear regression again after reducing the number of features.

Reducing the number of features significantly decreases the sum of squares error and improves the R squared, at least now the R squared is above zero.

However, the R squared value is only 0.04, which means that there is still room for improvement in the linear model. I notice that the model is capturing the mean of the data well, but not the standard deviation.

### When looking at the difference between the predicted and actual values more closely, I noticed that part of the reason the performance of linear regression was so poor was because there is a lot of variance in the data. The number of days needed can vary by a few months, and so an exact prediction will not be very precise and the SSE will be very large.

### Thus, instead of trying to predict the exact number of days, I am going to try to predict whether the visas will be processed within 4 months, within 4-8 months ; 8 - 12 months; 12-16 months; 16-20 months or > 20 months using Logistic Regression. This will still be very useful to a potential applicant, when they check what time range their processing time could fall under, based on features in their application.

I am going to try Logistic Regression next, but here are some other ideas I also had for things to try when using Linear Regression: define better features, regularization, try polynomial regression, add more data, try normalizing the y value (number of days) so it is between 0 and 1.

Logistic Regression

I define the output categories and implement logistic regression:

The Logistic Regression does a much better job than Linear Regression at predicting the processing times - the accuracy of the model is 70.6%. Next, I am going to explore the data a little more to improve the performance of the model.

Data Exploration & Improving Logistic Regression (1)

First, I plot the distribution of the processing times to see if there are any outliers. The histogram below reflects the fact that 96% of the points fall within ~600 days - let's try eliminating these outliers and see if we get a better prediction

The accuracy of the model improves by about 4%! Let's look at the confusion matrix to see how well the model does with true and false positives. I was worried that the model might predict all times as falling within the 4 – 8 month category. It doesn’t do that – it does predict processing times in other categories. However, there are still plenty of false positives in the 4-8 month category, and the model doesn't perform much better than predicting everything as falling within 4-8 months. This is probably because the dataset is quite skewed towards these values. I am going to explore if there are new features that might reduce the number of false positives.

Data Exploration & Improving Logistic Regression (2)

I looked at the data again, and it seemed like if the application was denied, it was usually processed in < 4 months. (Compare Orange (Denied) and Blue (Accepted) Histograms) Adding the case status as another feature could improve the performance of the model

However, this is not something potential applicants will know in advance when they key in details about their application.

#### The model does a little better – accuracy improves by 0.5%. There are more true positives in categories 3,4 and 5. However, there are still plenty of false positives in the 4-8 month category, and so there is definitely room for improvement.

Few things that I can try next:

1. Define new features which are better predictors (maybe wage amount)
2. Regularization
3. Add more data
4. Try a different ML algorithm (Random Forests)
5. Try grid search

The next steps I going to try to pick up more skills are:

1. Build a web app that outputs processing times based on information of applicant
2. Try converting tables to a SQL database
3. Try some web scraping