EM-623 Final Project

Student: Ekta Solanki

Instructor: Dr. Carlo Lipizzi

1. Business understanding

Many foreign nationals want to work in the United States. Although U.S. jobs usually start off as temporary, U.S. employment can often lead to a green card (U.S. permanent residency). Every year, lots of people come to USA to study, to work, or just as a tourist. Some decides to return some just falls in to love with this beautiful country and decides to stay. But this application is not as easy as it sounds. It often takes years to obtain a permanent visa in US.

US government provides different kinds of non-immigrant visas to foreigners like H1B, L1, O1, E1, TN etc. These non-immigrant visa holders can apply for permanent visa through some process which requires them to qualify for it. Almost all the students coming to US are interested in applying for permanent visa after completion of their education and few years of work experience. This data mining project can help future students coming for studies or current employer of US companies or companies that are affiliated with US companies, to help them realize their chances of getting a future permanent visa.

This data mining project is about the same. The main focus of this data-mining is on what kind of profile gets certified for permanent visa from the people who are not the citizen of USA.

2. Data Understanding

File type: us-perm-visas.csv

Dataset size: 70MB

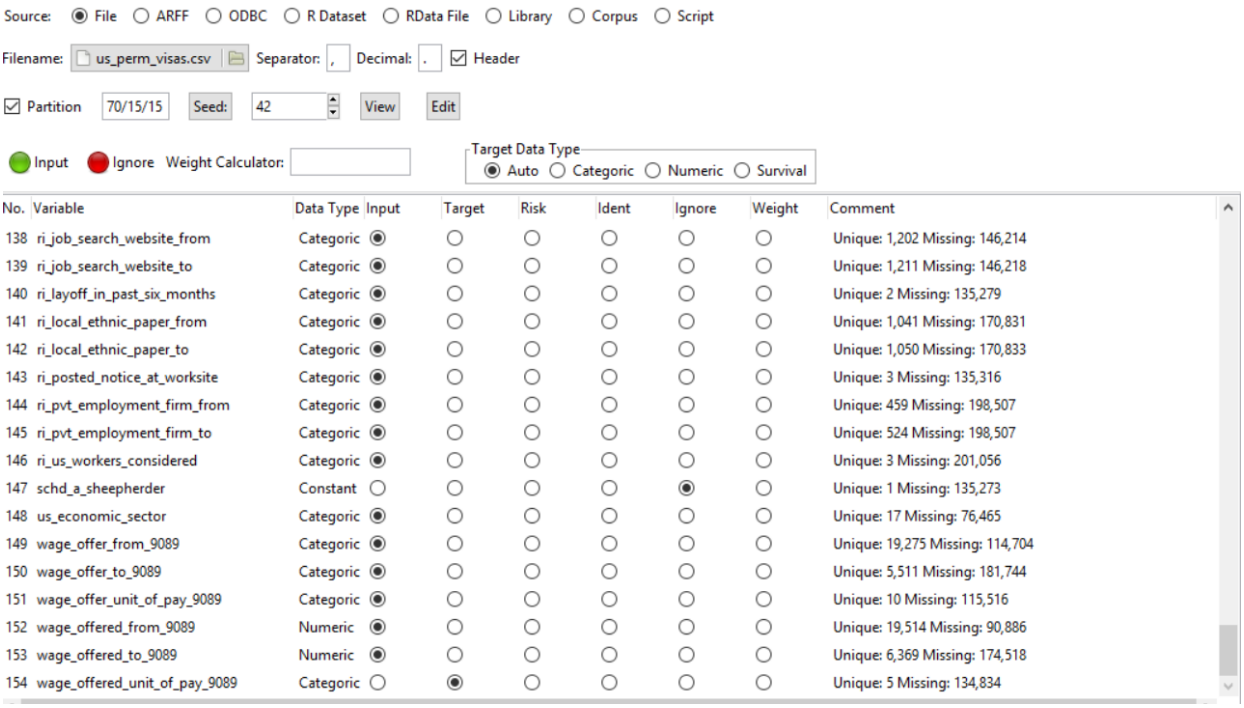
Dataset Shape: 154*205533

Dataset Source: https://www.kaggle.com/jboysen/us-perm-visas

Origin of dataset: https://www.dol.gov/ (**US Department of Labor)**

A permanent labor certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. In most instances, before the U.S. employer can submit an immigration petition to the Department of Homeland Security's U.S. Citizenship and Immigration Services (USCIS), the employer must obtain a certified labor certification application from the DOL's Employment and Training Administration (ETA). The DOL must certify to the USCIS that there are not sufficient U.S. workers able, willing, qualified and available to accept the job opportunity in the area of intended employment and that employment of the foreign worker(H1-B) will not adversely affect the wages and working conditions of similarly employed U.S. workers. In addition to this many no- working visa holders also submits their application for permanent residency to US DOL for example, F1, J1, B1-B3 etc.

To get the basic idea of like number of variables and variable types we run the csv file in to rattle,



No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
138	ri_job_search_website_from	Categoric	◉	○	○	○	○	○	Unique: 1,202 Missing: 146,214
139	ri_job_search_website_to	Categoric	◉	○	○	○	○	○	Unique: 1,211 Missing: 146,218
140	ri_layoff_in_past_six_months	Categoric	◉	○	○	○	○	○	Unique: 2 Missing: 135,279
141	ri_local_ethnic_paper_from	Categoric	◉	○	○	○	○	○	Unique: 1,041 Missing: 170,831
142	ri_local_ethnic_paper_to	Categoric	◉	○	○	○	○	○	Unique: 1,050 Missing: 170,833
143	ri_posted_notice_at_worksite	Categoric	◉	○	○	○	○	○	Unique: 3 Missing: 135,316
144	ri_pvt_employment_firm_from	Categoric	◉	○	○	○	○	○	Unique: 459 Missing: 198,507
145	ri_pvt_employment_firm_to	Categoric	◉	○	○	○	○	○	Unique: 524 Missing: 198,507
146	ri_us_workers_considered	Categoric	◉	○	○	○	○	○	Unique: 3 Missing: 201,056
147	schd_a_sheepherder	Constant	○	○	○	○	◉	○	Unique: 1 Missing: 135,273
148	us_economic_sector	Categoric	◉	○	○	○	○	○	Unique: 17 Missing: 76,465
149	wage_offer_from_9089	Categoric	◉	○	○	○	○	○	Unique: 19,275 Missing: 114,704
150	wage_offer_to_9089	Categoric	◉	○	○	○	○	○	Unique: 5,511 Missing: 181,744
151	wage_offer_unit_of_pay_9089	Categoric	◉	○	○	○	○	○	Unique: 10 Missing: 115,516
152	wage_offered_from_9089	Numeric	◉	○	○	○	○	○	Unique: 19,514 Missing: 90,886
153	wage_offered_to_9089	Numeric	◉	○	○	○	○	○	Unique: 6,369 Missing: 174,518
154	wage_offered_unit_of_pay_9089	Categoric	○	◉	○	○	○	○	Unique: 5 Missing: 134,834

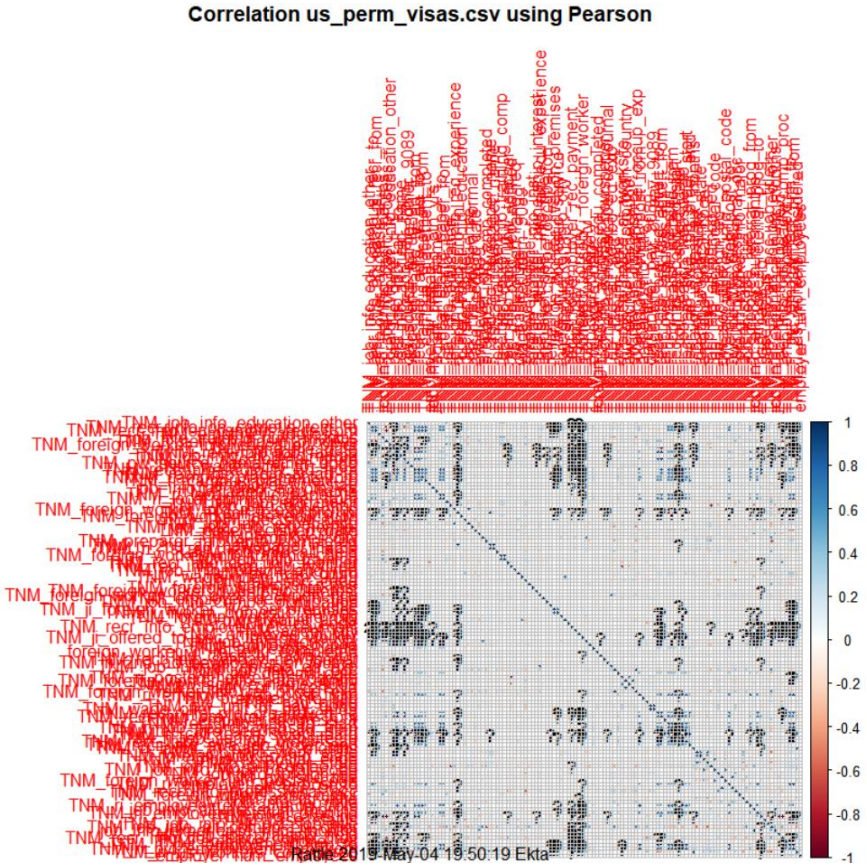
So, the dataset consists of 154 columns and 205,533 rows. The columns contain details like, employer name, employer city, foreign worker education, foreign worker citizenship, wage offered by employer, Class of admission (visa status before application for permanent visa), Industry of foreign worker, NAICS(North American Industry Classification System) codes, Prevailing wages according to ETA form-9089, different types of codes regarding to the application, and case status.

Here we will select “Case Status” as our target variable to see what kind of profile gets certified for permanent work or residency visa.

Data in the different columns are numeric as well as categoric.

Tools used: We will use Excel, R-studio (Rattle), Canopy (Python) for data cleaning, data analysis and data-mining.

Correlation mattrix of this data after converting the non-numeric data in to catagotic data, the matrix looks pretty complicated and hard to decide which variable to keep and which varible to ignore. It will be easier for us to logically decide which variables are more related for our analysis and keeping them only.



3. Data Preparation:

The file “us-perm-visas.csv” contains 154 variables and 2,05,533 entries.

Cleaning is done manually using Excel.

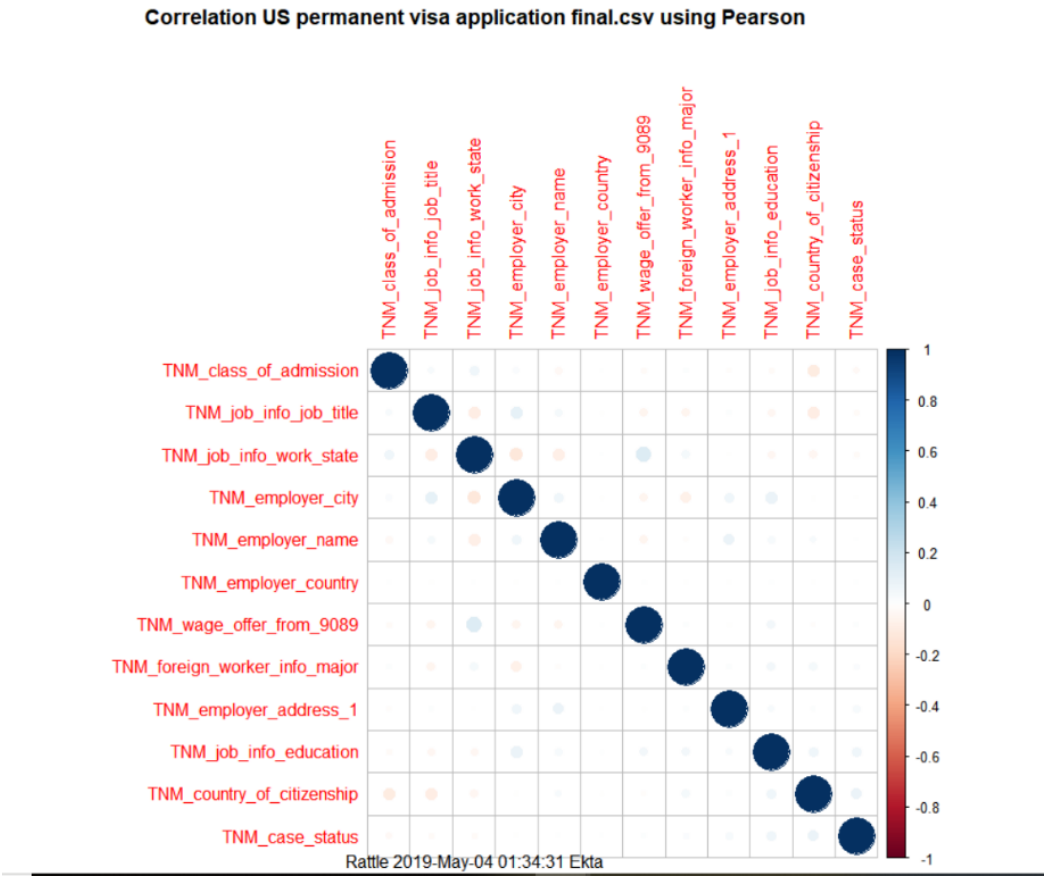
Some of them contains 70% of missing data some of them has no data. For better results we delete these columns.

Some two columns have similar input but missing values in either of the columns. So, I combined these two columns in to one and deleted the unnecessary column.

Out of 205533 entries, first 135270 entries have more than 80% missing data(50% is our threshold here). So, these rows have been deleted.

There is lack of coherence in “wage_offer_from_9089” variable, as it contains hourly and yearly wage provided by employer. So all the hourly wage entries has been deleted.

We keep the threshold of 0.2, means if there is correlation between variables <0.2 then we can keep them. After cleaning the data-set the correlation matrix looks something like this

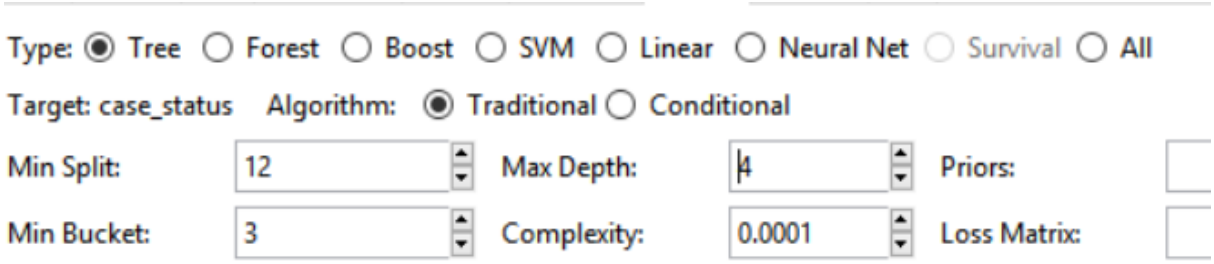


Case_status is our **target variable** which has bare minimum correlation with all the other variable that we selected. So now we can start the modeling process.

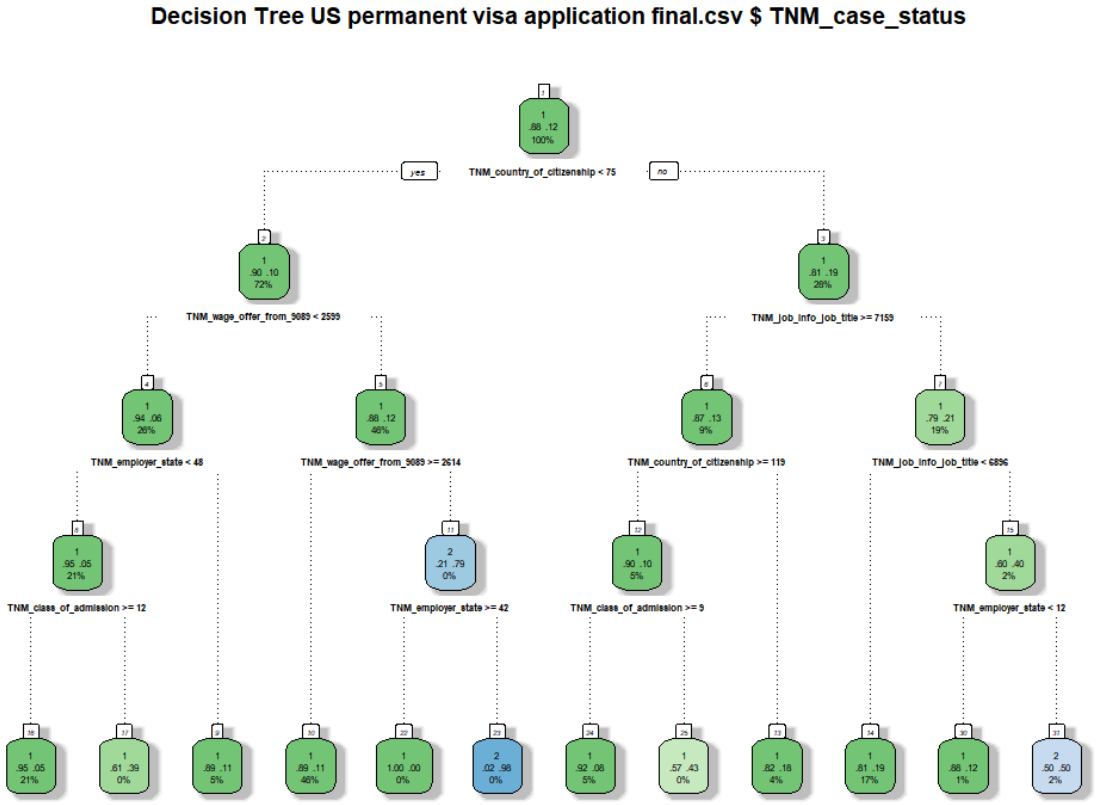
4. Modelling:

We are dealing with supervised leaning as we know the target variable.

After iterating multiple modeling parameters to make an optimal decision tree, we finalize below values for modeling parameters.



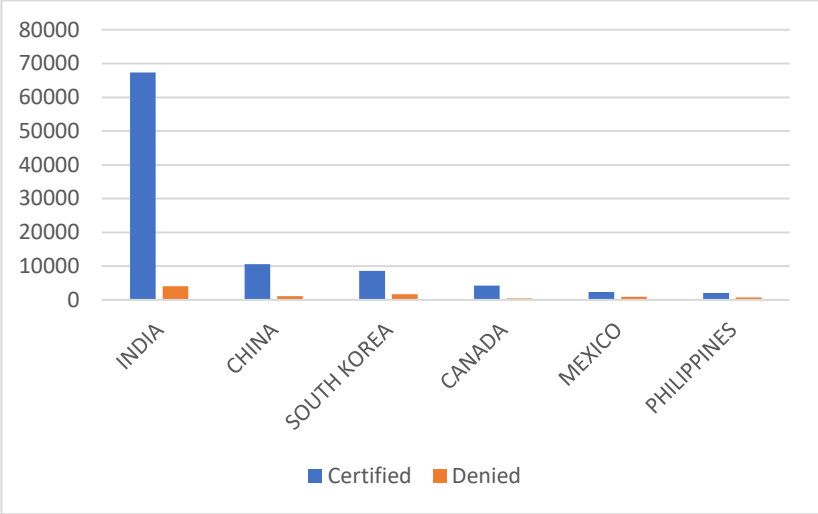
Which gives us the decision tree as below,



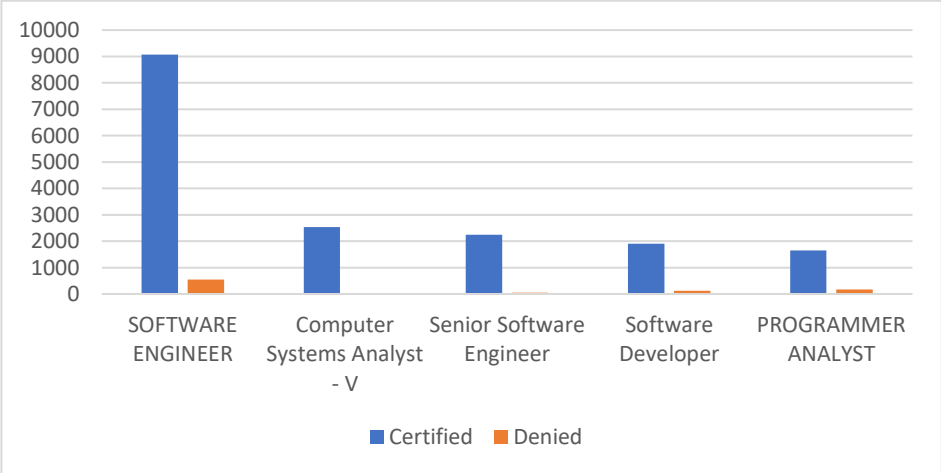
This model considers variables like County of citizenship of applicant, Prevailing Wage offered by employer, Class of admission (visa status before application for permanent visa), Job title of employee, and employer state.

This could be explained that getting approval for permanent visa in us in vastly dependent on your Current citizenship, Wage offered by your employer(skills) and Employer state.

We can verify it by doing a simple calculation in Excel.



On the left side is the graph of application by different **citizenship** holders and their acceptance and denial graph. We can see that top 5 countries to apply for permanent visa to US are India, China, South Korea, Canada and Mexico. India's acceptance rate is much higher than any other country during 2012-2017.



Similarly, it can be computed in Excel to verify that **Job title** is also a very important variable in getting acceptance for permanent visa. As we can see top 5 job title are all computer related.

5. Evaluation:

We devide dataset in to 75/25 for training vs testing data to do the evaluation of our model.

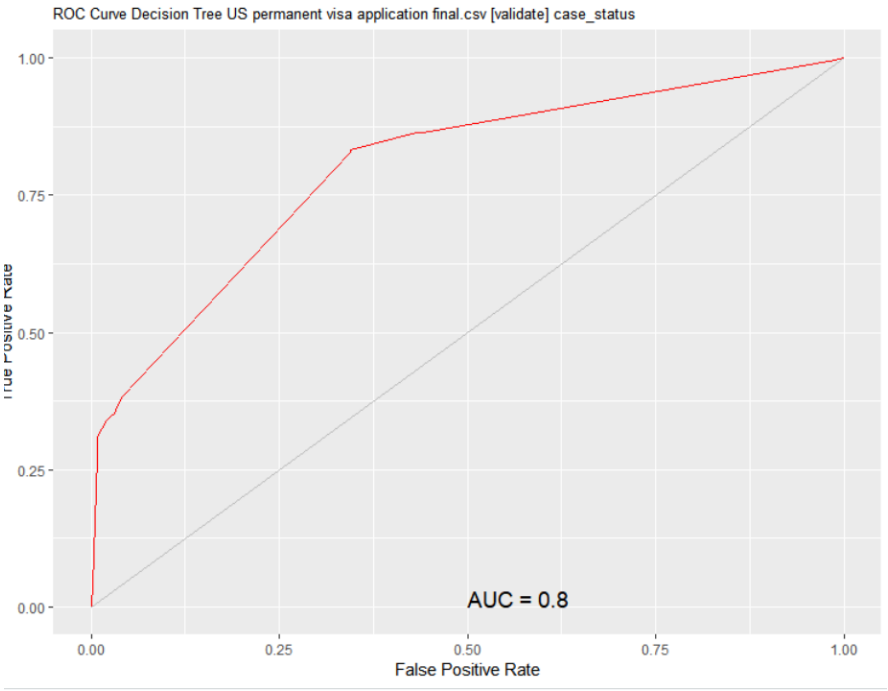
After running the error matrix in “evaluation” tab in rattle we get below results.

```
Error matrix for the Decision Tree model on US perm vi

      Predicted
Actual    1  2 Error
     1 5931 54   0.9
     2  726 73  90.9

Error matrix for the Decision Tree model on US perm vi

      Predicted
Actual    1   2 Error
     1 87.4 0.8   0.9
     2 10.7 1.1  90.9

Overall error: 11.5%, Averaged class error: 45.9%
```

This model shows inefficiency of 90% for false negative means it does not predict well the denial as case status with given variables.

But the overall accuracy of model is high as the overall error is 11.5% only So this model can be considered.

Area Under the Receiver operating characteristic curve is 0.8 which means the model when tested on rest of the testing dataset, the model we builder gives 80% of accuracy. Which is a good number.



6. Deployment:

This model can help you **predict** future possibility of getting approval or denial of US permanent visa based on your citizenship, your employer, your education and wage offered by your employer.

We can improve this model by adding recent data and doing an in-depth and extended analysis beyond this exam.

For the people who are planning to come to US or are currently in college can focus on which industry they want to go and which skill can help them get hired by a US employer to improve their probability to get approval of US permanent visa.