EasyVisa

Business Case

Background

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

Objective

The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. OFLC has hired the firm EasyVisa for data-driven solutions. You as a data scientist at EasyVisa have to analyze the data provided and, with the help of a classification model:

- Facilitate the process of visa approvals.
- Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

Data Information

The data contains information about the business problem

Variable	Description	Type of Variable
case_id	ID of each visa application	Object
continent	Information of continent the employee	Object
education_of_employee	Information of education of the employee	Object
has_job_experience	Does the employee have any job experience? Y= Yes; N = No	Object
requires_job_training	Does the employee require any job training? Y = Yes; N = No	Object
no_of_employees	Number of employees in the employer's company	Int64
yr_of_estab	Year in which the employer's company was established	Int64
region_of_employment	Information of foreign worker's intended region of employment in the US	Object
prevailing_wage	Average wage paid to similarly employed workers in a specific occupation in the area of intended employment (The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment)	Float64
unit_of_wage	Unit of prevailing wage (Values include Hourly, Weekly, Monthly, and Yearly)	Object
full_time_position	Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position	Object
case_status	Flag indicating if the Visa was certified or denied	Object

Observations	Variables
25,480	12

Manipulations to Raw Data:

1.Object variables were converted to Category
2.Removal of negative employee figures

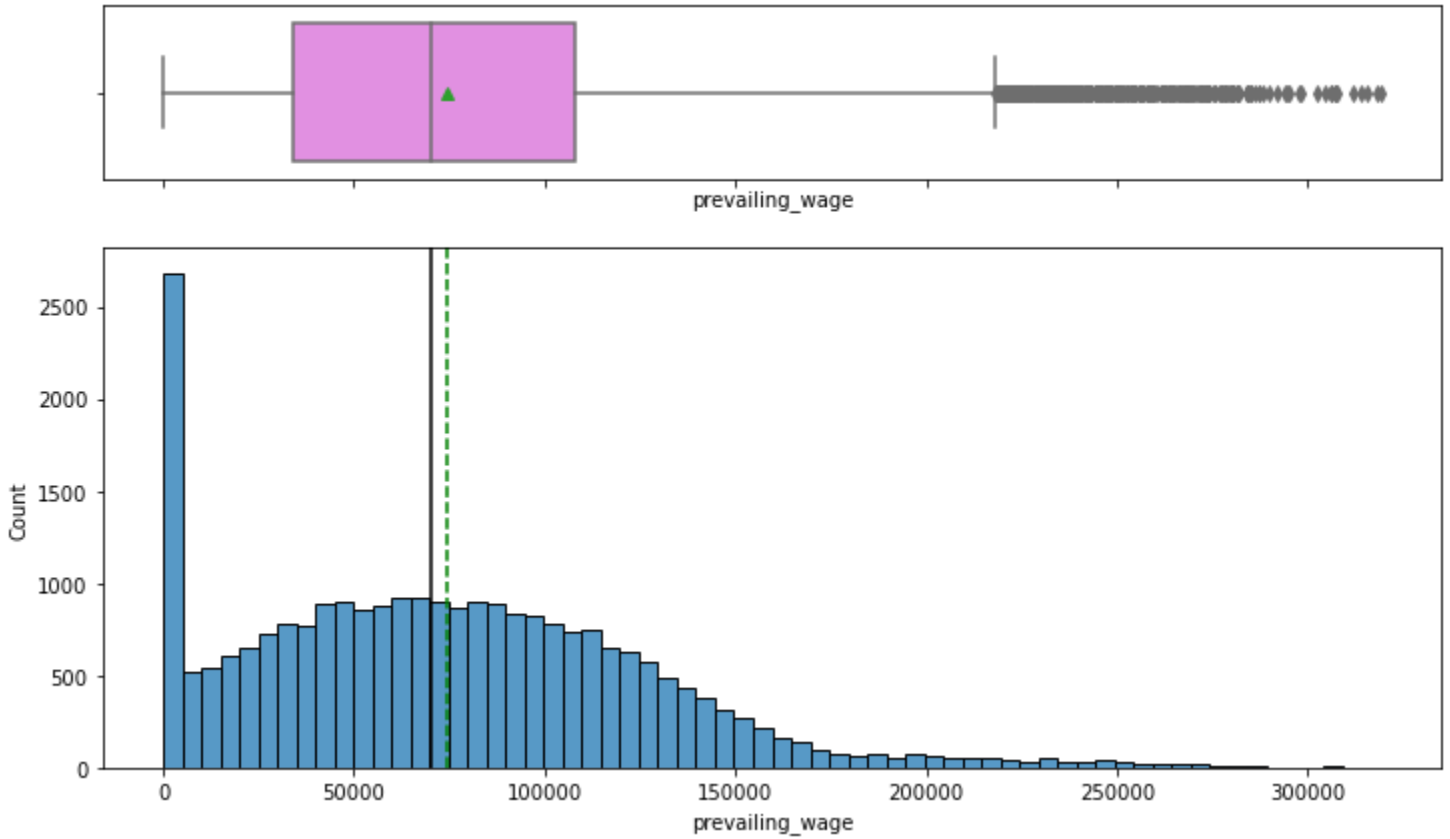
Exploratory Data Analysis – Prevailing Wage

This data contains the prevailing wage

Observations:

1. The prevailing wage is heavily left skewed
2. The largest number of people earn quite small compared to the rest of the workers indicating a significant wage disparity

Prevailing Wage



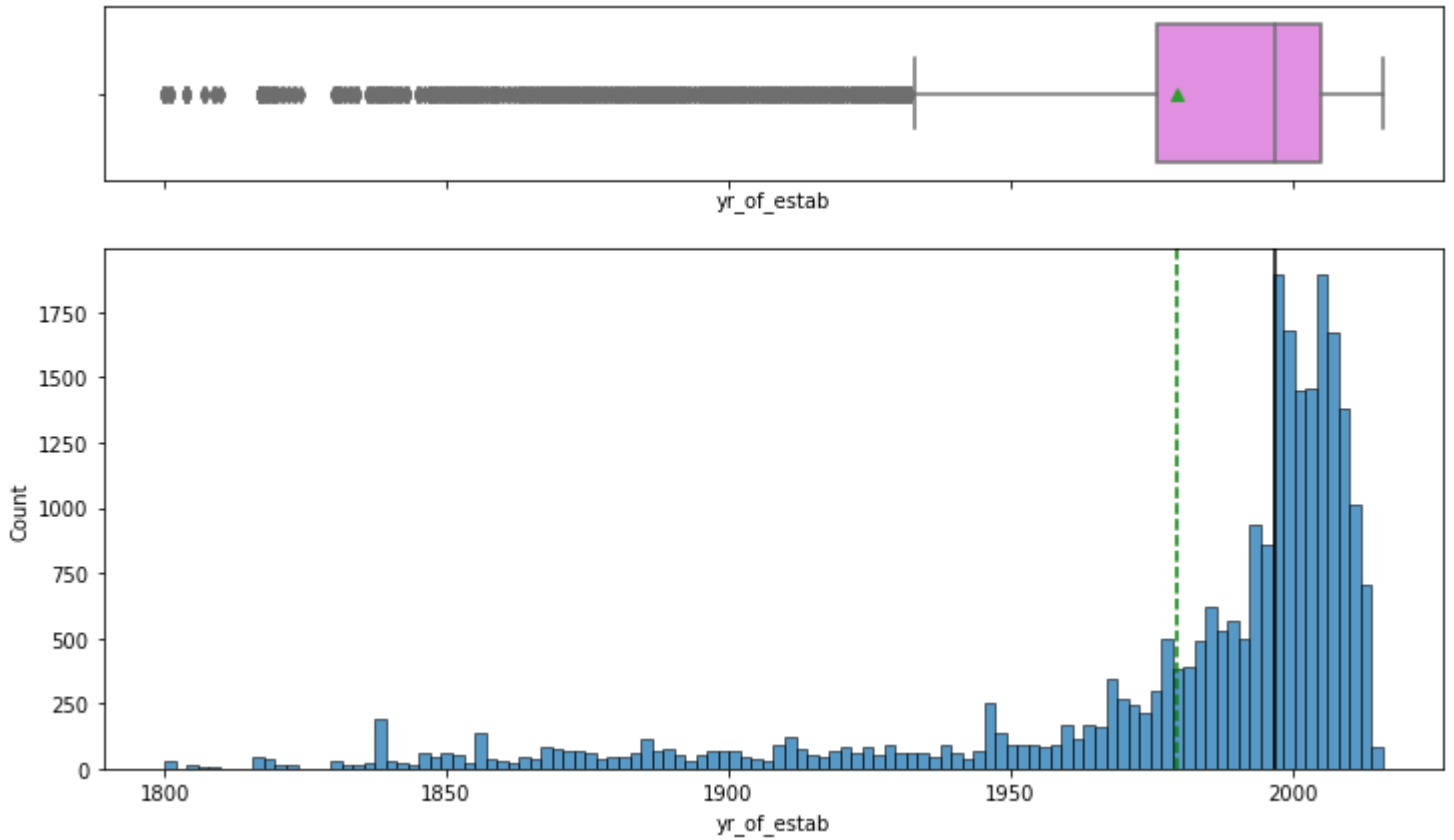
Exploratory Data Analysis – Year of Establishment

This data contains the year of establishment

Observations:

1. The year of establishment of the companies is heavily right skewed which makes sense as most companies would have been formed after a certain time.
2. Most of the companies have their formation date around the 2000's

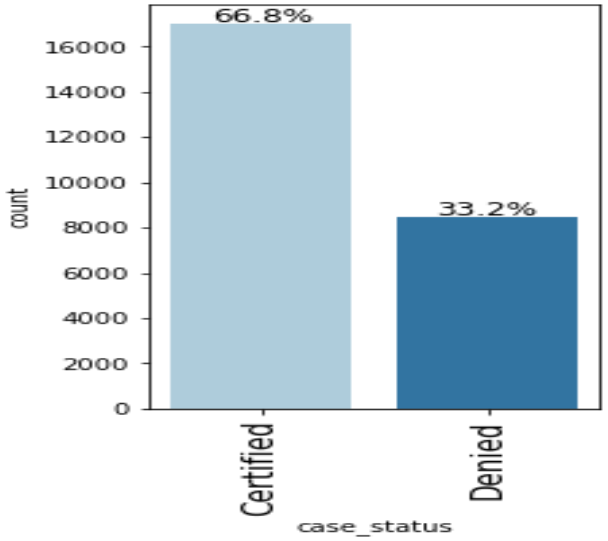
Year of Establishment



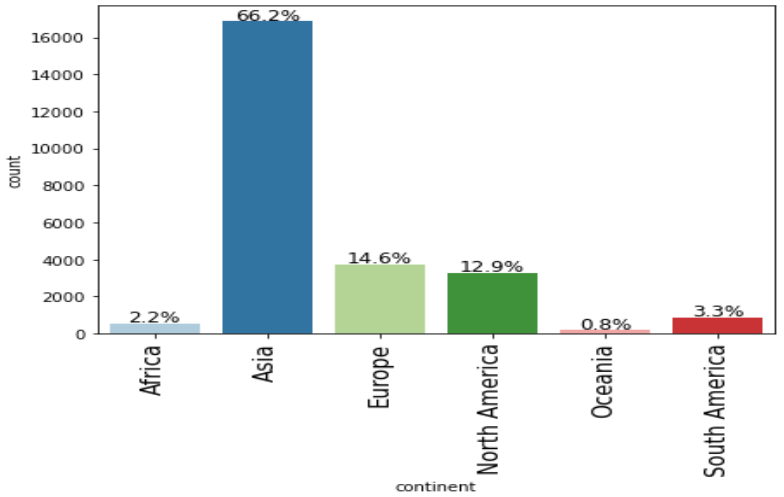
Exploratory Data Analysis – General Information

General Information about the Data

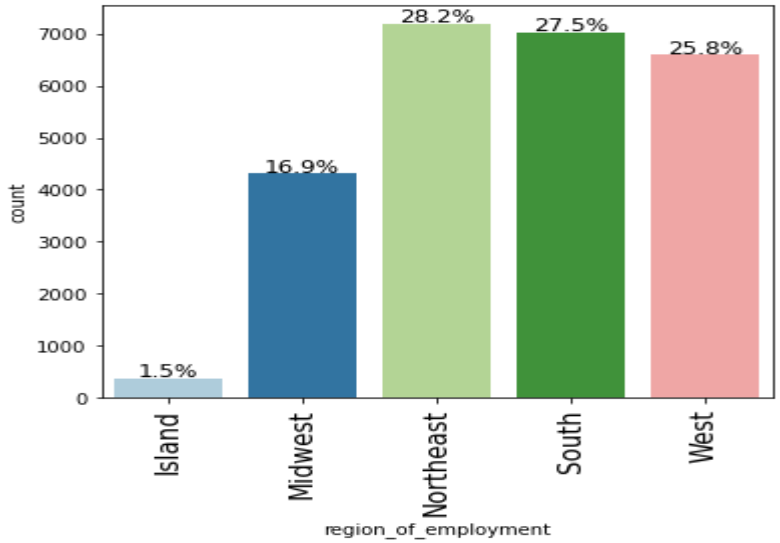
Percentage split of Visa Status



Percentage split of Continent



Percentage split of Employee Region



Observations:

1. The number of certified applicants is 66.8% indicating that most of the applicants get their visa certified

Observations:

1. Asia has the highest number (~16,900) of visa applicants by a very large margin
2. Europe and North America combined have approximately 7,000 applicants, which is slightly more than a quarter
3. The other 3 continents contribute very little

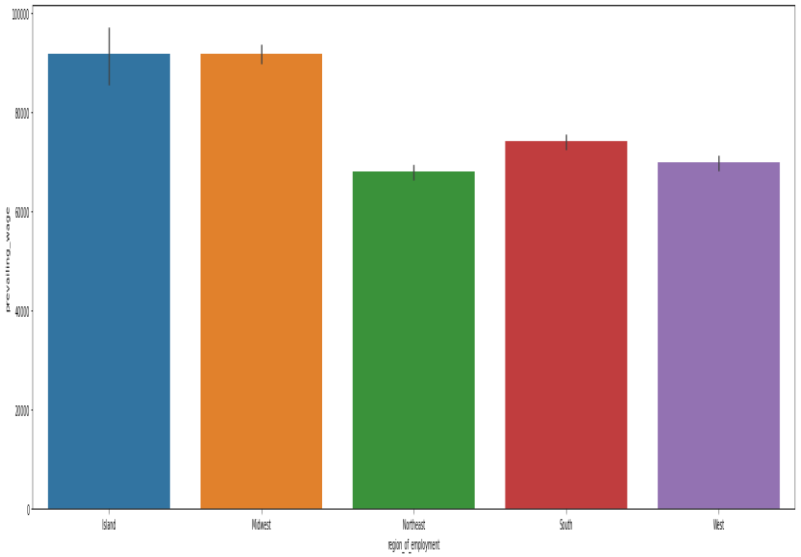
Observations:

1. The Northeast has the highest number of applicants, with the South and West coming a close second and third respectively. Together they make up over 80% of applications

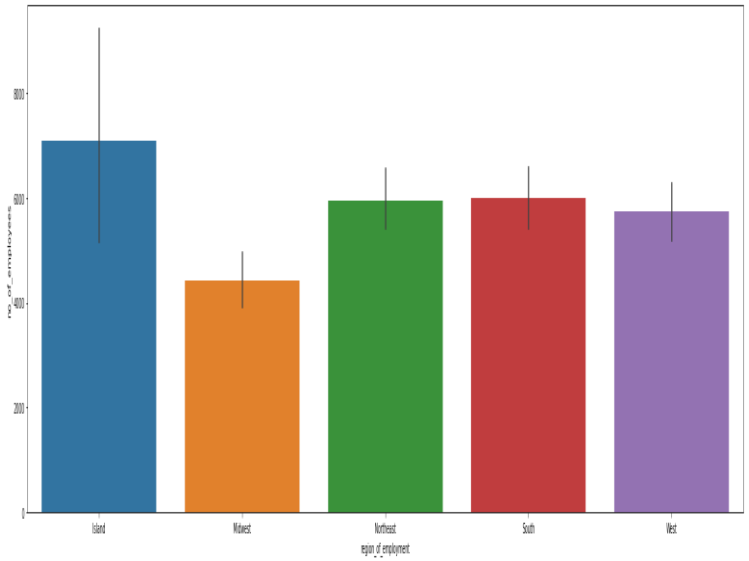
Exploratory Data Analysis – General Information

General Information about the Data

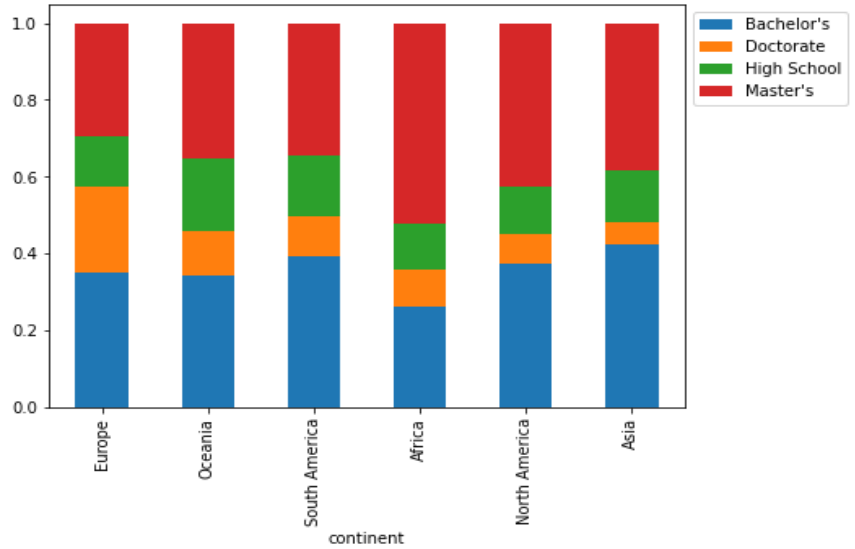
Region wrt Prevailing Wage



Region wrt No. of Employees



Continent wrt Education Status



Observations:

1. The highest prevailing wages come from the Island and Midwest regions. However, as shown in the previous observation, they have the lowest number of applicants indicating that they are probably known for not taking in as many applicants as the other regions

Observations:

1. The Island region has the highest number of employees. This could be another reason why their applicants are so low.
2. Northwest and West seem to have an equal number of employees
3. However, Midwest and South, though having lower numbers, the difference is not too significant.

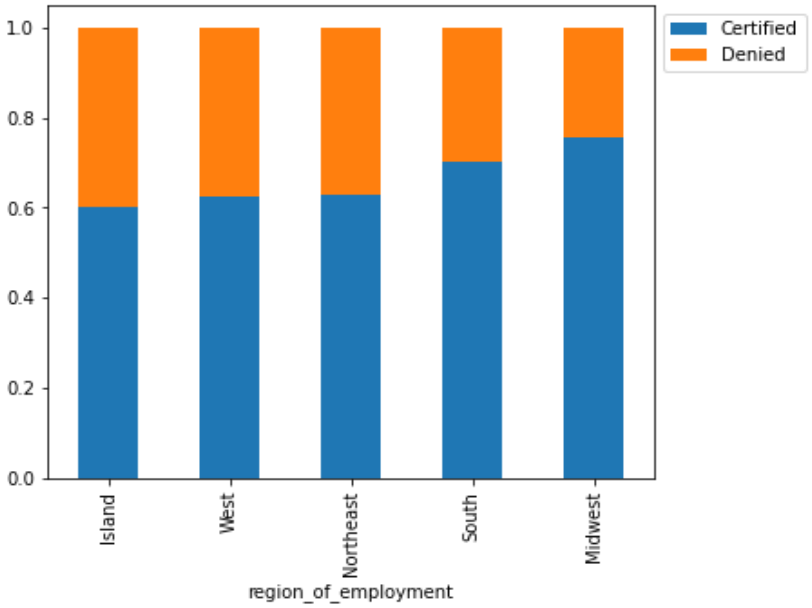
Observations:

1. Europe has the highest percentage of Doctorates while Asia has the highest percentage of Bachelor degrees.
2. Africa has the highest percentage of Masters
3. The degree split is very varied amongst the various continents

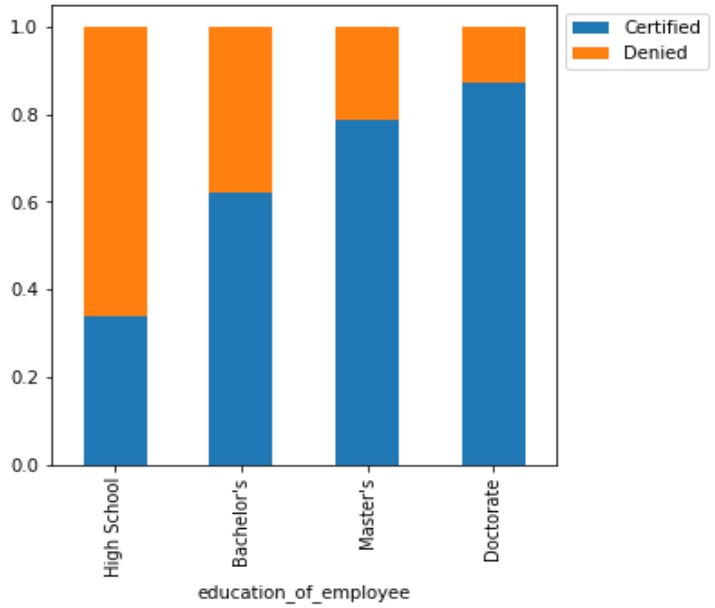
Exploratory Data Analysis – Visa Status

Information abut Visa Status

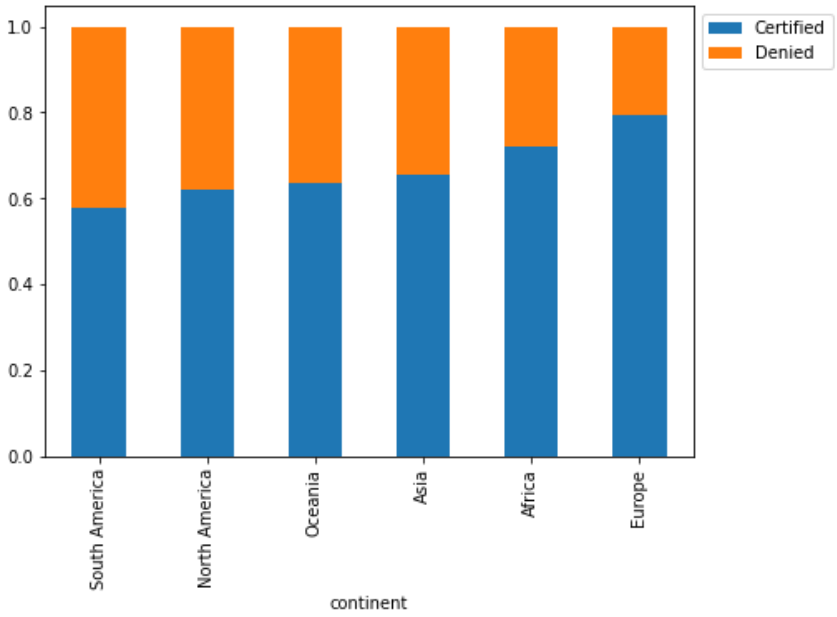
wrt employment region



wrt education level



wrt to employee continent



Observations:

1. Midwest gives the highest number of certified applicants while Island has the lowest number.
2. However, the disparity is not too large with none of the regions reach 80%

Observations:

1. Candidates with a doctorate seem to have a much higher rate of being certified with about 90% of them being certified.
2. Candidates with only a High School Diploma do not get certified with only about 30% getting certified

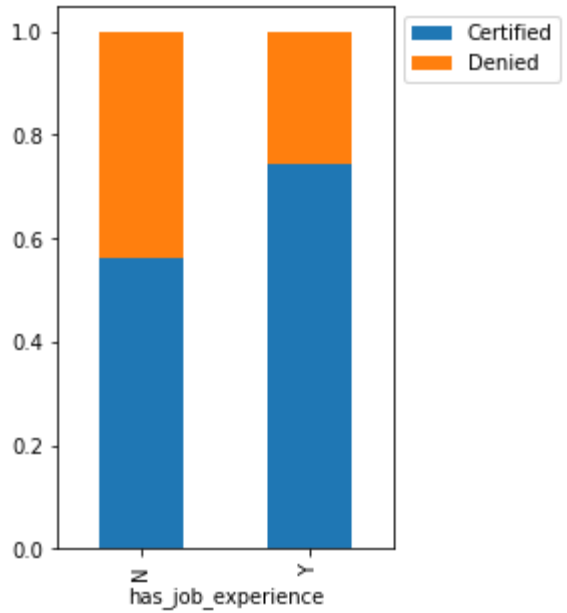
Observations:

1. South America has the lowest rate of visa certification while Europe has the highest.
2. Europe is well above the average indicating that more European workers get their visa certified

Exploratory Data Analysis – Visa Status

Information abut Visa Status

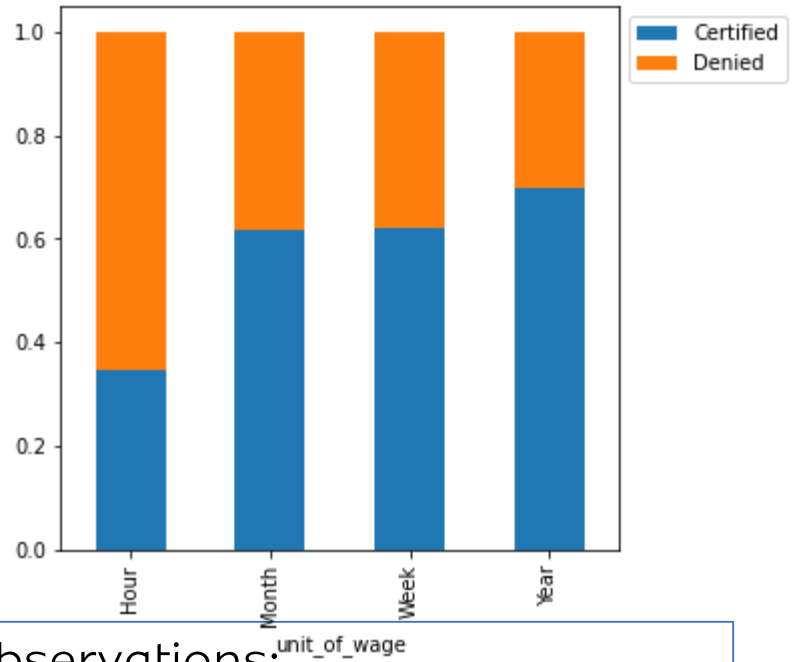
wrt job experience



Observations:

As expected, applicants with job experience get their visa's certified with almost 80% of them being successful

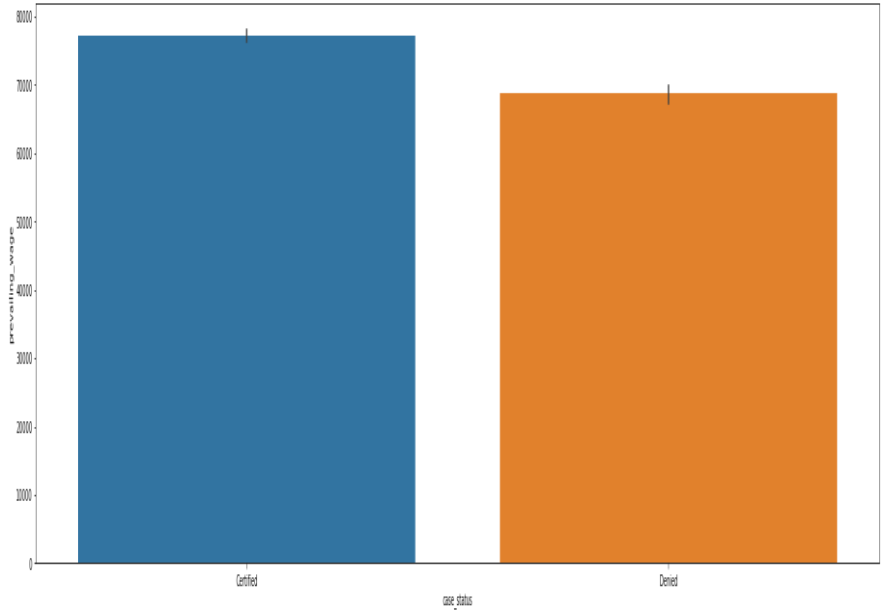
wrt payment schedule



Observations:

Applicants that get yearly wages have a higher chance of getting their visas certified while those that work for hourly wages have the lowest chances

wrt to prevailing wage



Observations:

The higher wages seem to have more visas certified than those with lower prevailing wages

Prediction Model – Bagging Classifier

Two methods were used under the Bagging Classifier models (Bagging Classifier and Random Forest Classifier). Also, hyperparameter tuning was done on the models in order to get the best prediction possible. Tuned Random Forest was chosen as the best of the Bagging models.
Below is a summary of the results:

Confusion Matrix (Random Forest with class_weights)

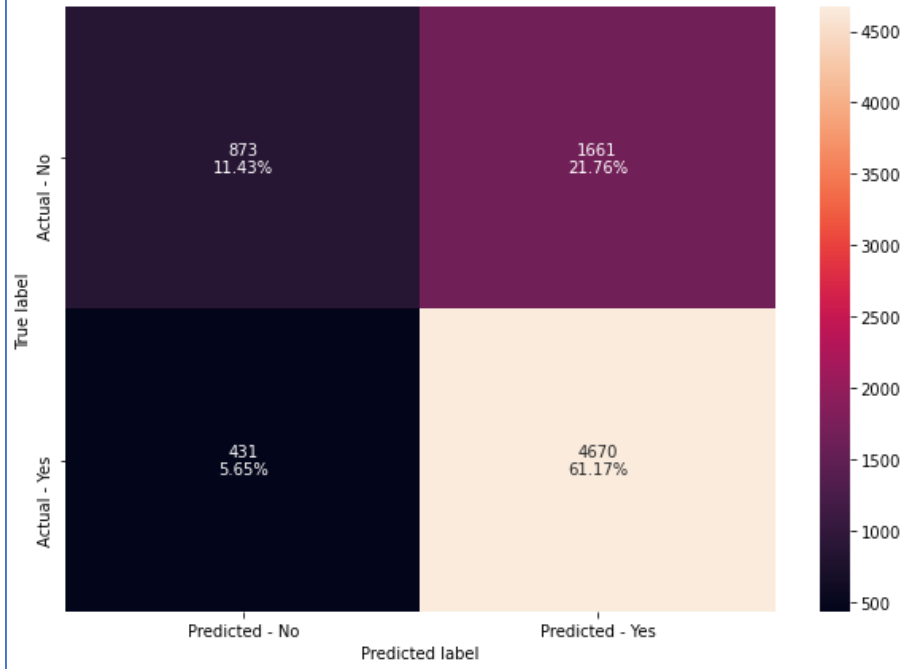
Model Comparisons						
Model Name	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Bagging classifier with default parameters	0.98	0.70	0.99	0.78	0.99	0.77
Tuned Bagging Classifier	0.67	0.67	1	1	0.67	0.67
Bagging classifier with base_estimator=LR	0.67	0.67	1	1	0.67	0.67
Random Forest with deafult parameters	1	0.72	1	0.83	1	0.77
Tuned Random Forest Classifier	0.76	0.75	0.88	0.87	0.79	0.78
Random Forest with class_weights	0.72	0.70	0.98	0.97	0.71	0.70

The most important features as identified by the model are:
a. High school education (Education of Employee)
b. Prevailing wage
c. Number of Employees
d. Job Experience

Prediction Model – Boosting Classifier

Two methods were used under the Bagging Classifier models (Adaboost Classifier and Gradient Classifier). Also, hyperparameter tuning was done on the models in order to get the best prediction possible. Tuned Gradient Boosting was chosen as the best of the Boosting models.
Below is a summary of the results:

Confusion Matrix (Tuned Gradient Booster)



Model Comparisons						
Model Name	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
AdaBoost with default paramters	0.74	0.73	0.89	0.88	0.76	0.76
AdaBoost Tuned	0.69	0.69	0.97	0.97	0.69	0.69
Gradient Boosting with default parameters	0.76	0.75	0.88	0.87	0.78	0.78
Gradient Boosting with init=AdaBoost	0.76	0.75	0.88	0.87	0.78	0.78
Gradient Boosting Tuned	0.73	0.73	0.92	0.92	0.74	0.74

The most important features as identified by the model are:
a. High school education (Education of Employee)
b. Prevailing wage
c. Number of Employees
d. Job Experience

Conclusion

Due to the high number of visa applications for the limited number of jobs, a profile of the 'ideal' visa candidate is needed in order to facilitate a smoother, more effective and correct visa approval process. based on the data analysed above, the most important factors in approving Visas, have been determined. They include, but are not limited to

- Minimum of a high school education
- The prevailing wage of the company/industry
- Job experience of the candidate
- Location of candidate
- Region of employment
- Type of wage candidate is willing to work for

Based on all these factors, the ability of the model to properly classify and predict which candidate is going to be certified will enable their time and resources more effectively and efficiently.

The classification model chosen is *Tuned Gradient Boosting Model* because it has both a very high recall score with a sufficiently high accuracy

Recommendation

Based on the analysis, these are the following recommendations that can help the business determine what criteria should be used for Visa approval and help facilitate the process for Visa approval:

- As Asia has the highest number of applications, a dedicated Asia section should be created. This is done to handle the high number of cases from that specific continent
- The other regions can be grouped together for the time being, until they become bigger and contribute more
- The regions with the largest number of applicants should also be investigated in order understand why they are coming from there and why there is a disconnect between the number of applications and certified status
- Special focus should be paid on low contributing continents in order to drum up business by introducing special rates and promotions.
- The company can partner with schools in order to grow the number of candidates with degrees so as to increase certified statuses
- Companies should be split into two concerning *job experience* and *type of position* and clearly indicated on the site
- The company needs to identify a minimum requirement/criteria list from each region/company/industry so that applicants are aware of what they are applying for. This will enable immediate weeding out

- The ideal profile for a certified applicant should consist of the following characteristics:
  - Minimum of Bachelors degree
  - Applying to regions that is known for having a lot of certified applicants
  - Relevant job experience
  - Prevailing wage of the industry

*This profile should be circulated to all aspirational candidates so they can be forewarned before applying
*This profile is subject to adjustment by the company