EasyVisa

Travis Bruns

Table of Contents

- Core business idea
- Problem to tackle
- Financial implications
- How to use ML model to solve the problem

Business Problem Overview

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.

Objective

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.

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 your firm EasyVisa for data-driven solutions. You as a data scientist 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 Overview

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

- case_id: ID of each visa application
- continent: Information of continent the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee has any job experience? Y= Yes; N = No
- requires_job_training: Does the employee require any job training? Y = Yes; N = No
- no_of_employees: Number of employees in the employer's company
- yr_of_estab: Year in which the employer's company was established
- region_of_employment: Information of foreign worker's intended region of employment in the US.
- 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.
- unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
- full_time_position: Is the position of work full-time? Y = Full Time Position; N = Part Time Position
- case_status: Flag indicating if the Visa was certified or denied

Raw Data Summary

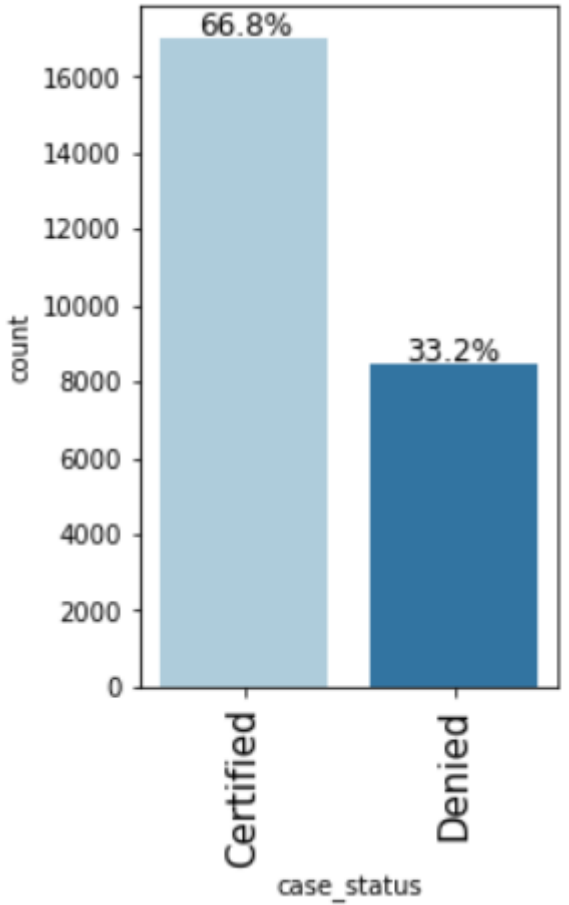
- The Source Data includes 25,480 rows and 13 columns.
- The Source Data did not duplicates.
- The Data types are 1 float, 2 integers, and 9 objects.
  - Float: prevailing_wage
  - Integer: no_of_employees, yr_of_estab
  - Object: case_id, continent, education_of_employee, has_job_experience, requires_job_training, region_of_employment, unit_of_wage, full_time_position, case_status
- The Data does not include any missing values.
- The Dependent Variable is Case Status.

Data Summary

- no_of_employees: Average number of employees is 5,667. The minimum value is negative and needs to be further investigated.
- yr_of_estab: The average year of establishment is 1979.
- prevailing_wage: The average prevailing wage is 52,816 ranging from $2.13 to $319,210.
- case_id is a unique value assigned to each value.
- continent contains six options with Asia being the most common.
- education_of_employee contains four values with Bachelor's being the most common.
- has_job_experience contains two values and the majority of applicants have job experience.
- requires_job_training contains two values and the majority of applicants do not need training.
- region_of_employment contains five values with the most common being te Northeast.
- unit_of_wage contains four values and the most common is year/annual.
- full_time_position contains two values and the majority of the applicants are for full time roles.
- case_status contains two values with the majority of applicatns being certified.

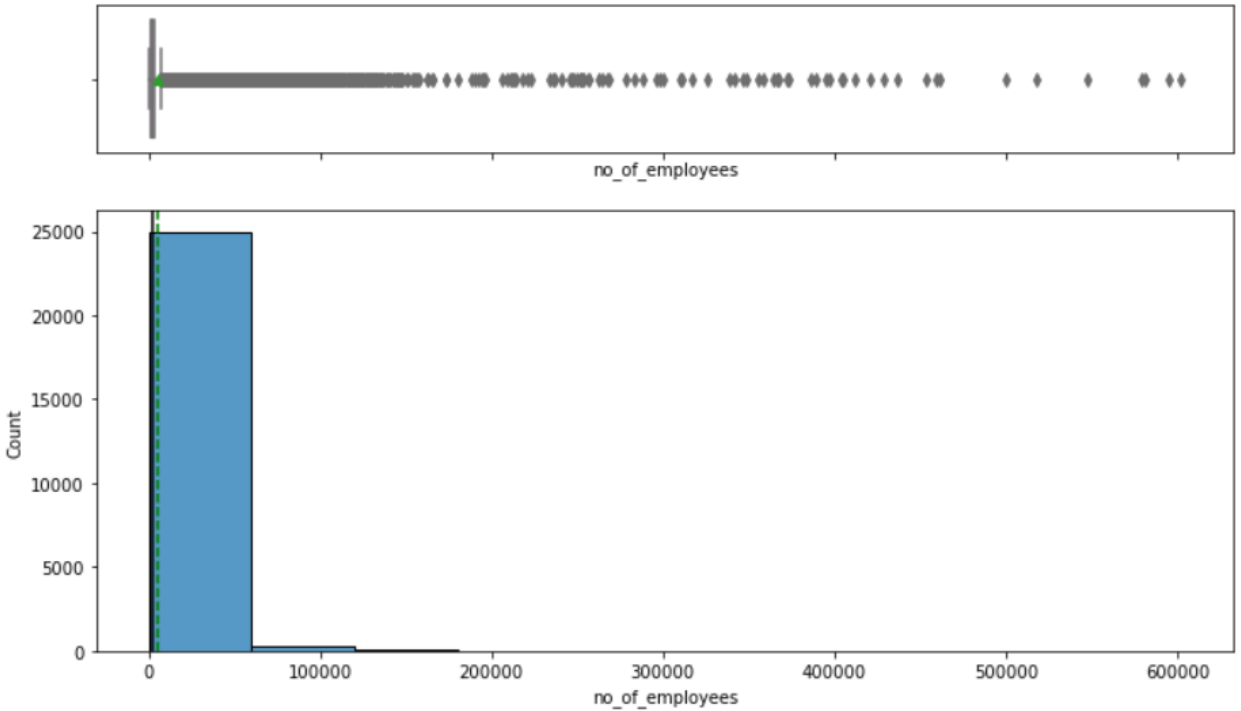
Categorical Data Summary

- Asia is the most common homeland for Visa applicants.
- The majority of applicants – 22,060/25,480 or 86.6% - are educated at the Bachelor's level or higher.
- The most common area of employment for Visa Applicants is the Northeast and is closely followed by the South.
- Denials occur 8,462/25,480 or 33.2% of the time.
- Most Visa applicants – 22,962/25,480 or 90% - are paid in an annual salary format.
- Most Visa applicants do not require job training – 22,525/25,480 or 88.4%.
- The majority of Visa Applicants have job experience – 14,802/25,480 or 58.1%.



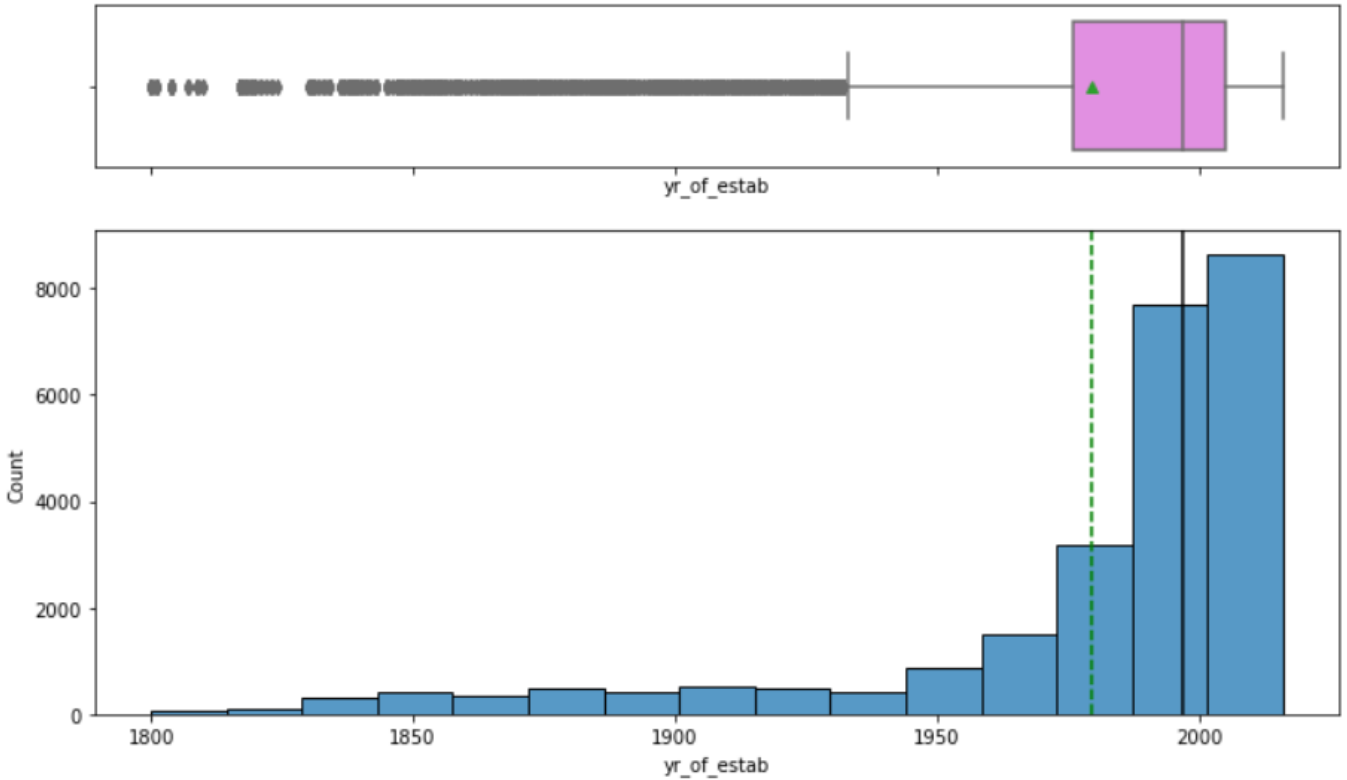
Univariate Data Analysis – Number of Employees

- The data is right skewed with almost all companies sponsoring Visa applicants having less than 10,000 employees.



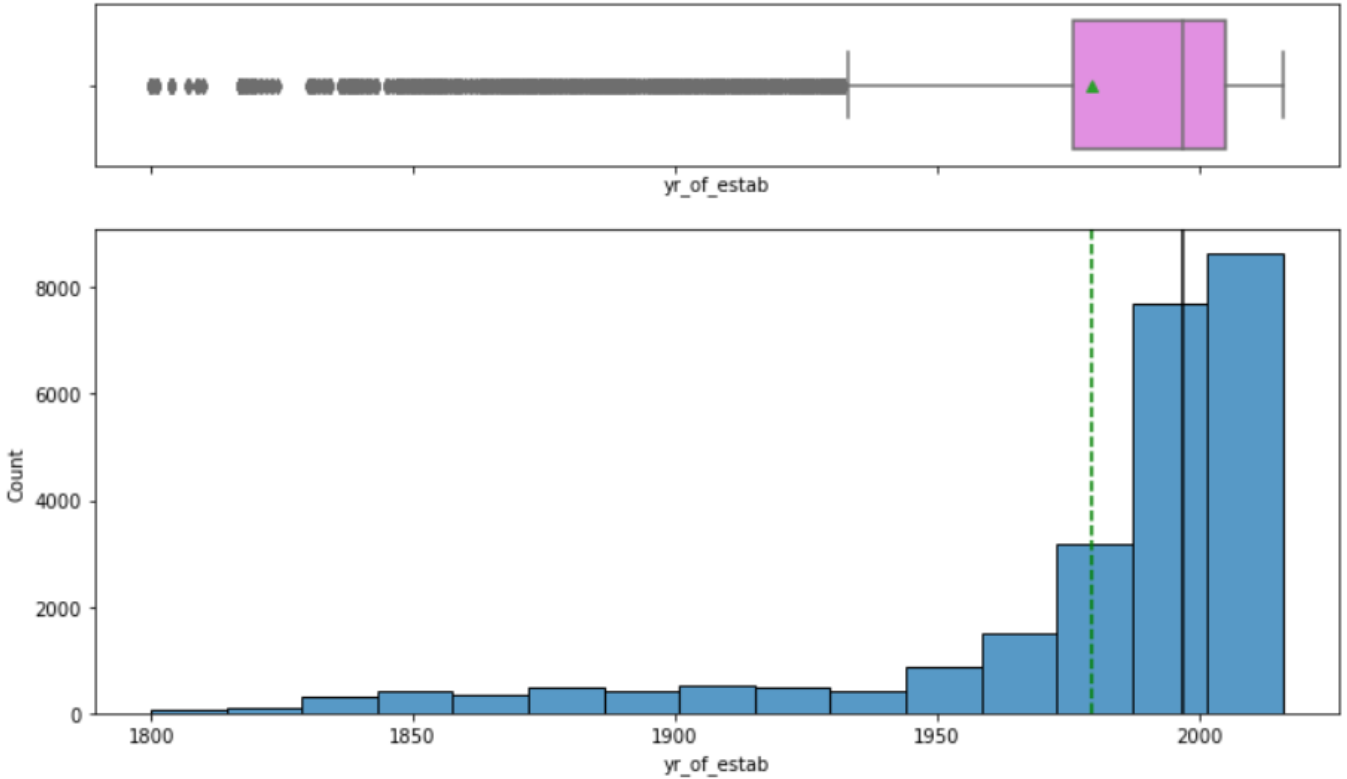
Univariate Data Analysis – Company Established Year

- The data is left skewed with most sponsors of Visa applicants being established after 1975.



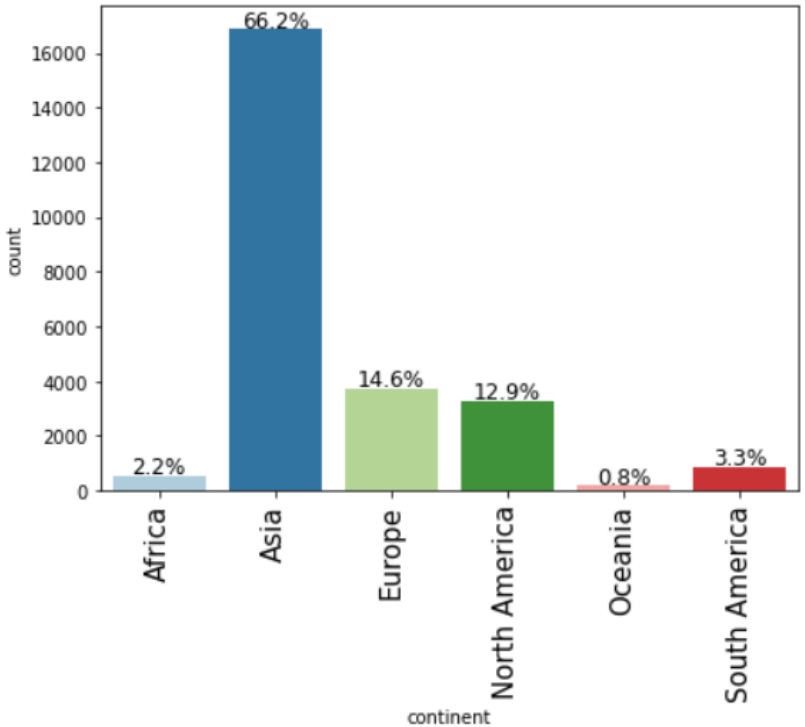
Univariate Data Analysis – Prevailing Wages

- The data is right skewed with most Visa applicants being offered wages greater than $55k.



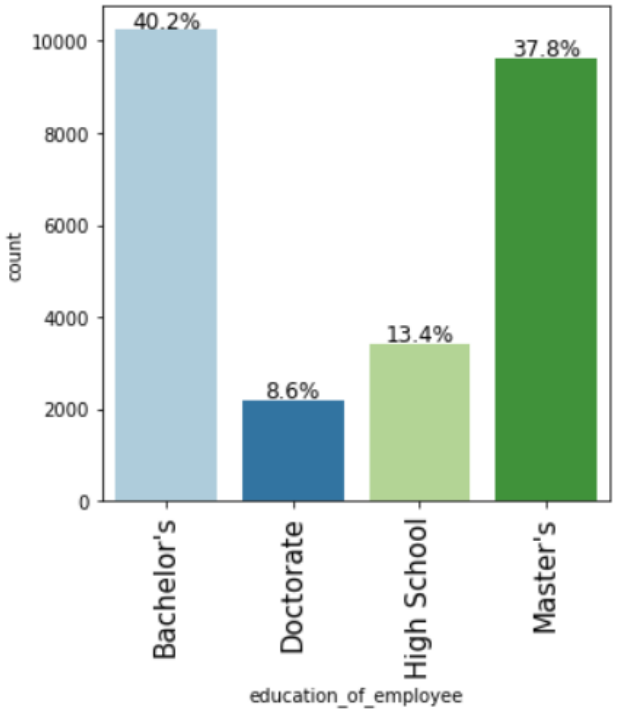
Univariate Data Analysis – Continent of Origin

- Asia is the continent most likely to have citizens apply for Visa certification.



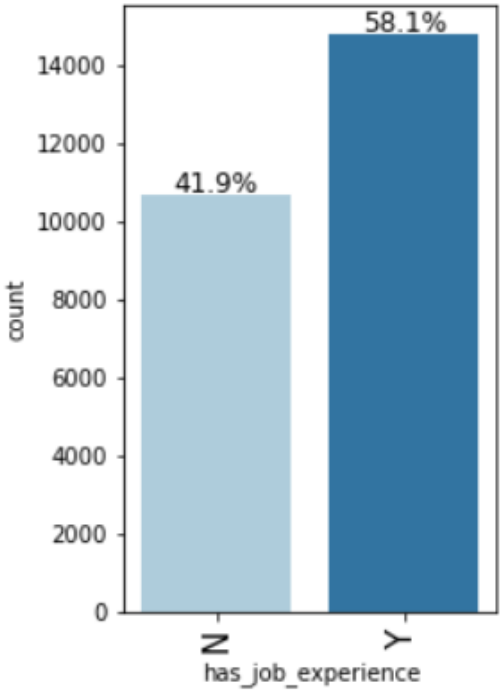
Univariate Data Analysis – Education

- Most Visa Applicants graduated with a Bachelor's or Master's degree.



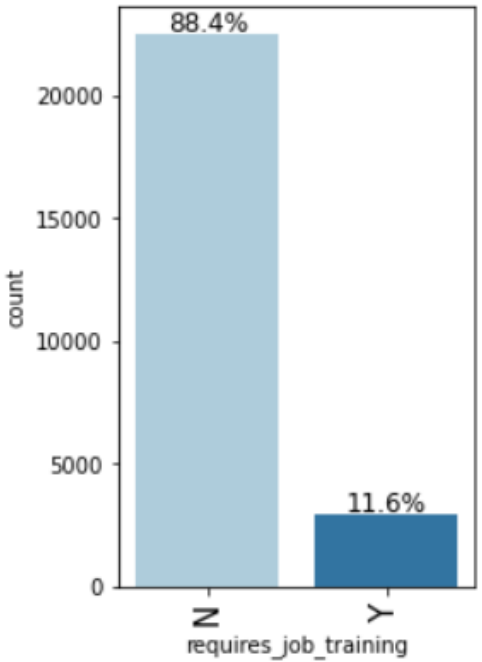
Univariate Data Analysis – Job Experience

- Most Visa Applicants have job experience.



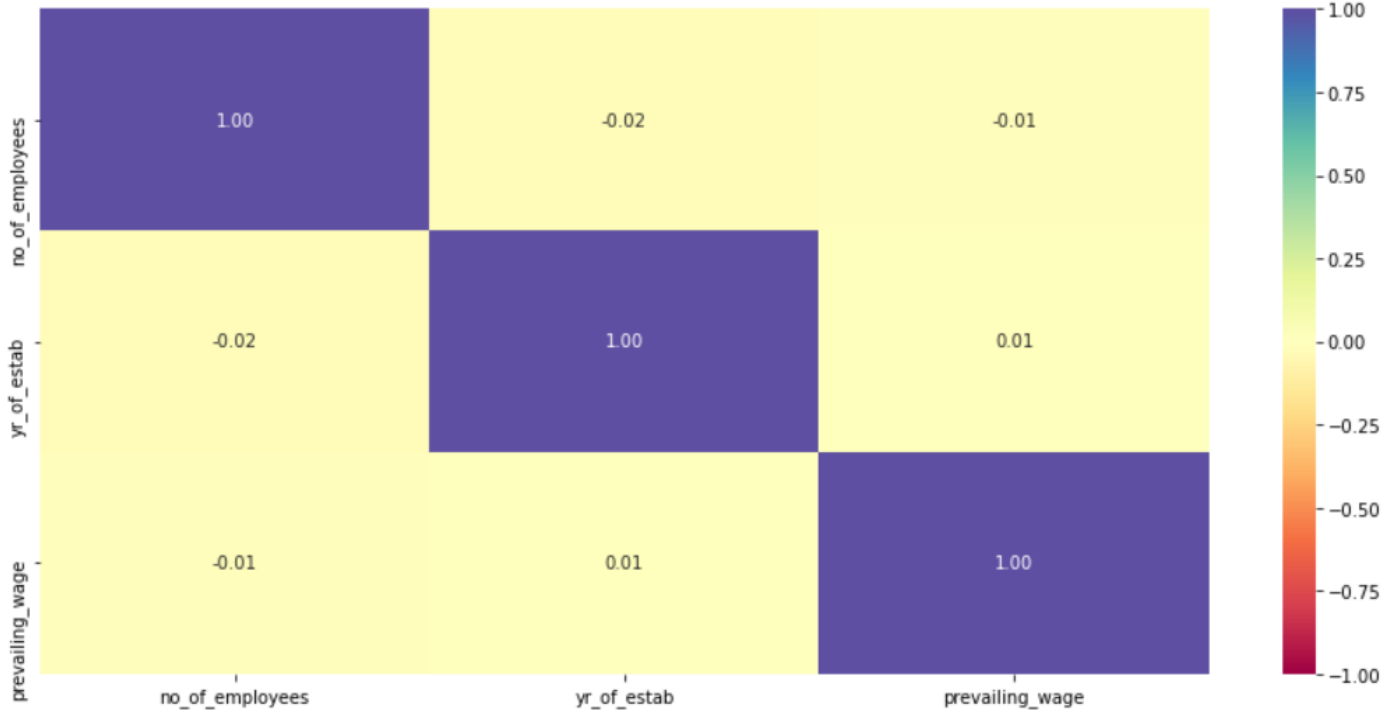
Univariate Data Analysis – Job Experience

- Most Visa applicants do not require job training.



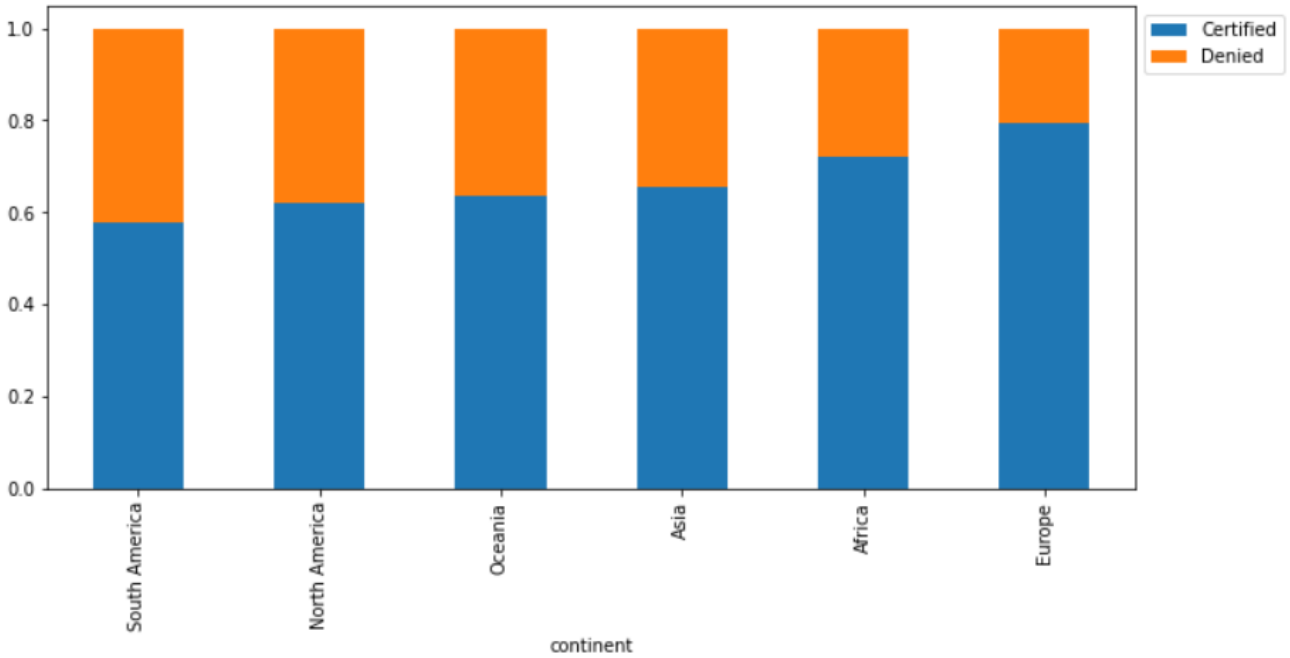
Bivariate Data Analysis – Integer Data

- Number of employees, years of establishment, and prevailing wages do not appear to be correlated.



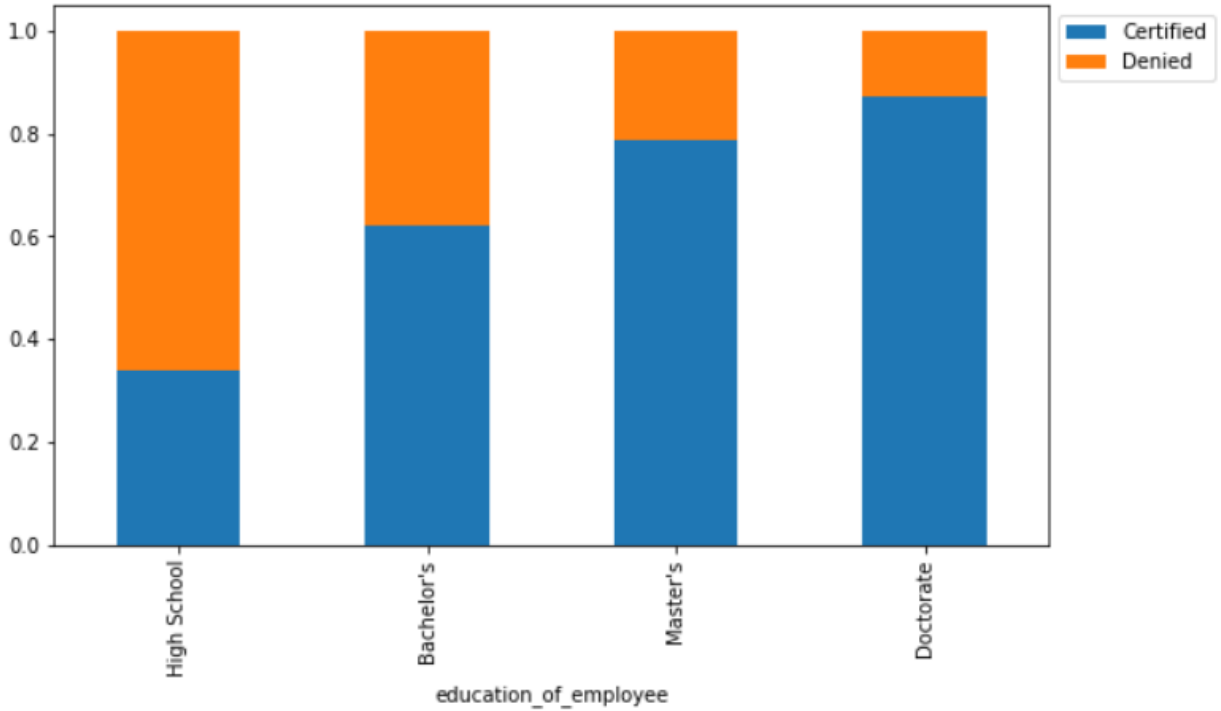
Bivariate Data Analysis – Certified per Continent

- South American Visa applicants have the lowest certification rate and Europe has the highest.



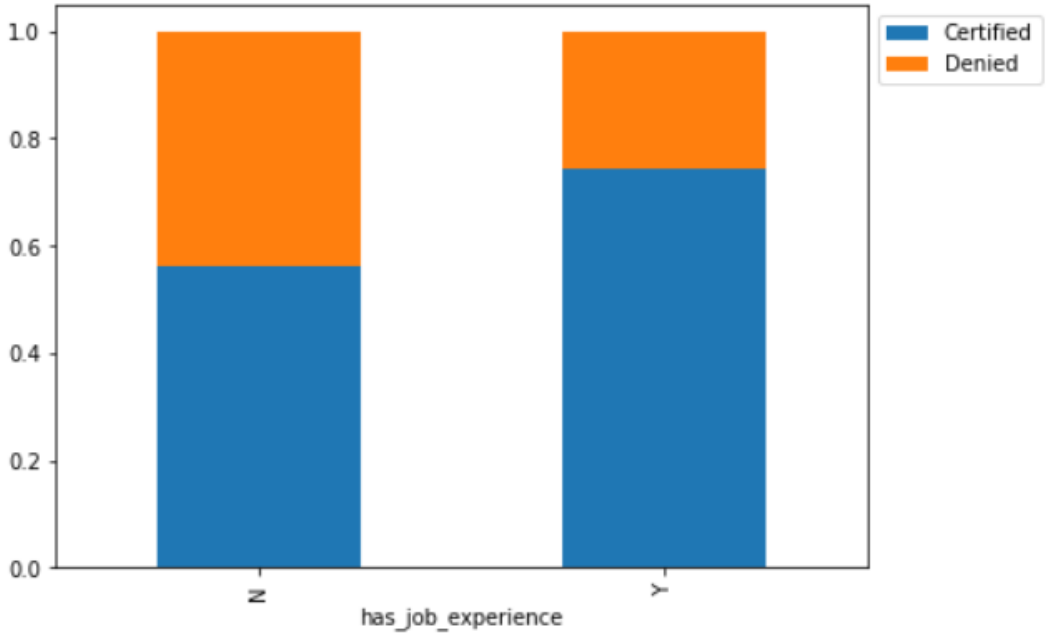
Bivariate Data Analysis – Certified per Education Level

- The level of education appears to affect the ability to of an applicant to become certified.



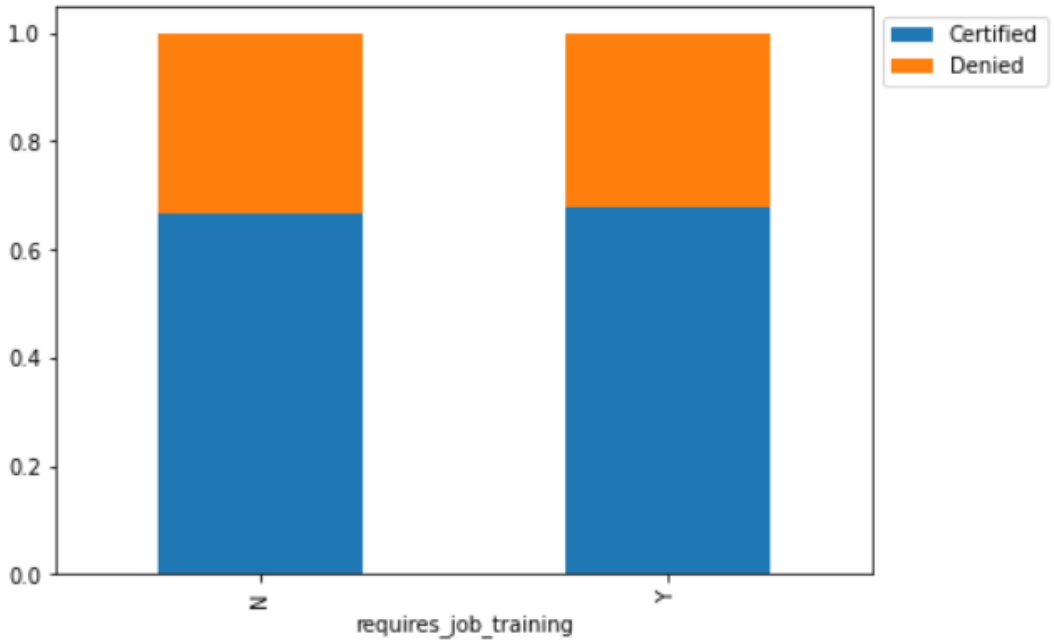
Bivariate Data Analysis – Certified per Job Experience

- Job experience appears to be positively correlated to becoming a certified Visa applicant.



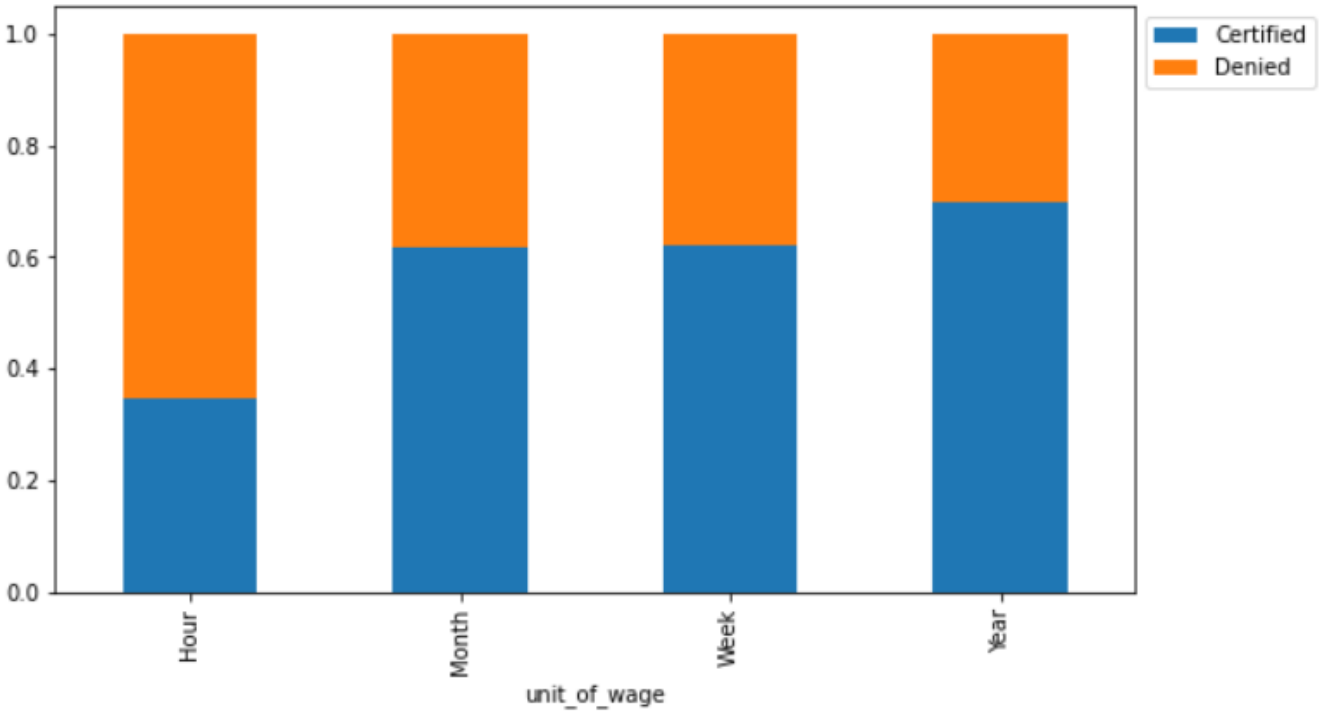
Bivariate Data Analysis – Certified per Requiring Training

- Requiring job training does not appear to affect a Visa applicant's success.



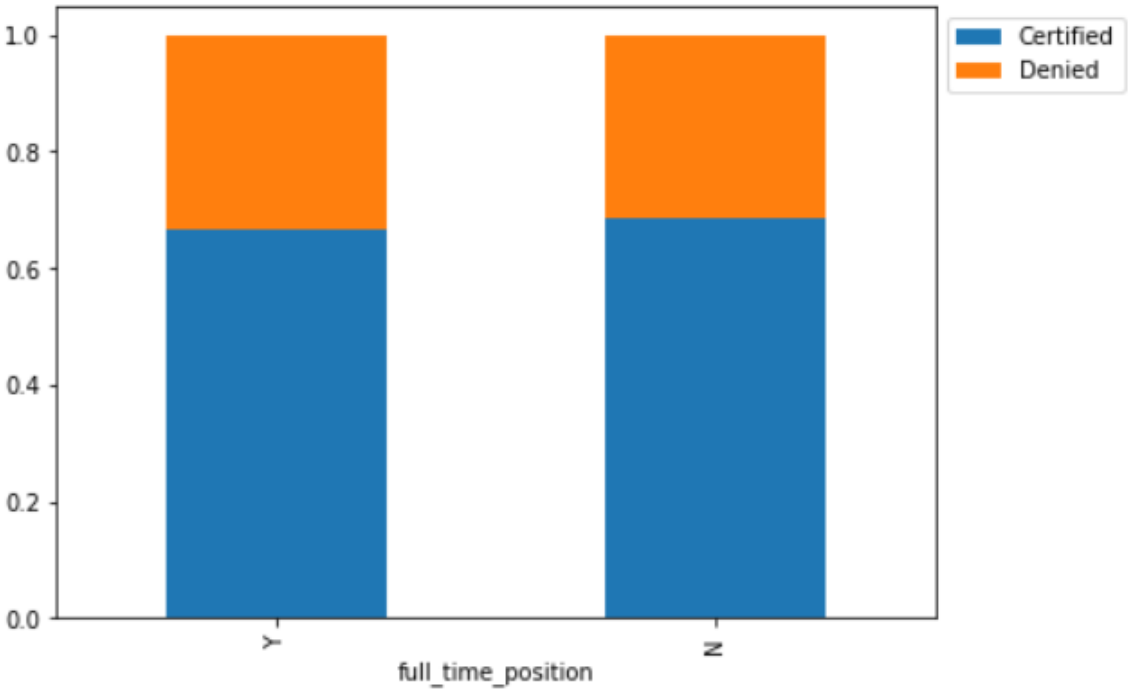
Bivariate Data Analysis – Certified per Wage Units

- Hourly employees appear to be the least likely to have a Visa application approved.



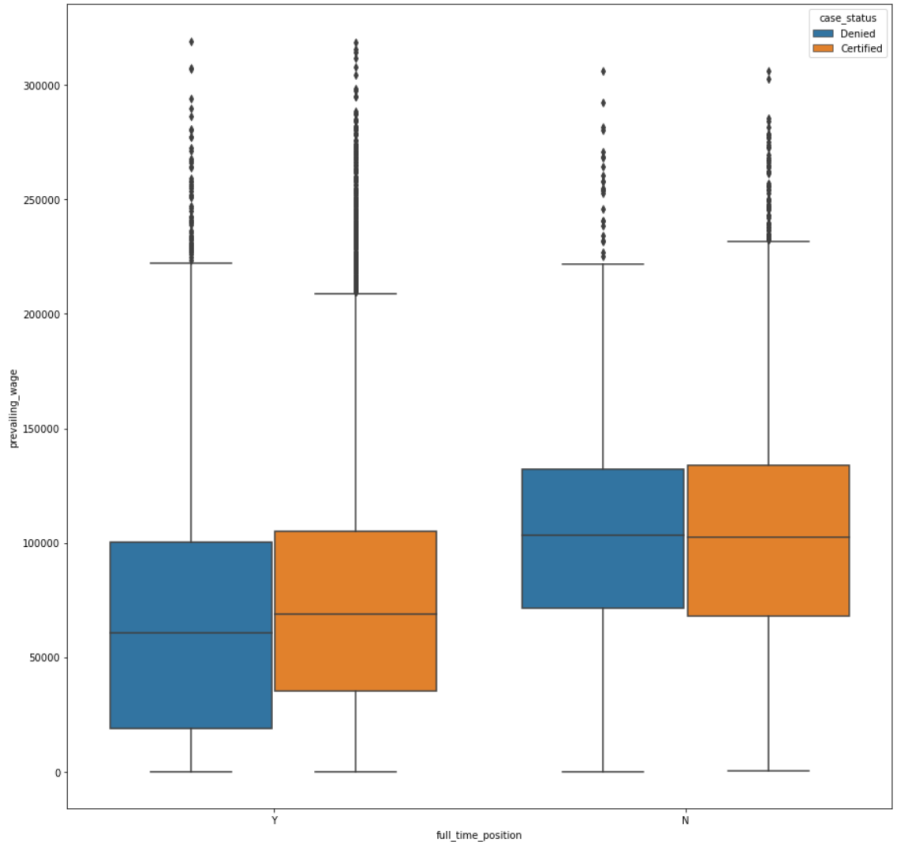
Bivariate Data Analysis – Certified per Job Type

- A job being full time or part time does not appear to correlate to certification or denial.



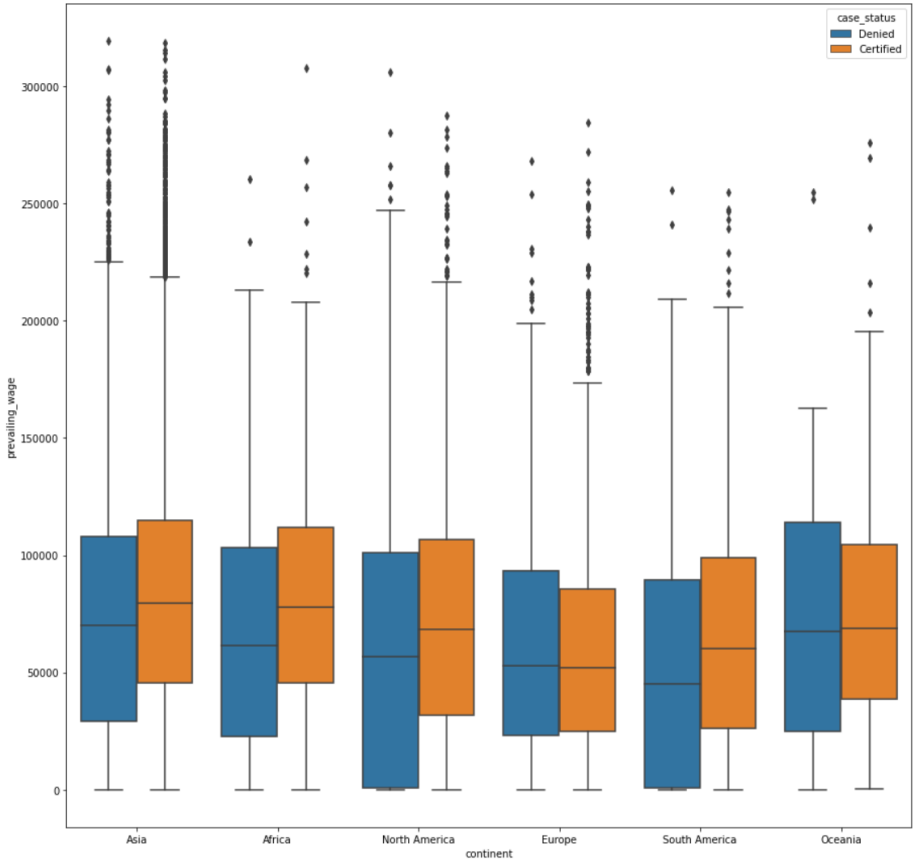
Multi-Variate Data Analysis – Role types and Wages

- Full time roles pay more on average than part time roles.
- Wages do not appear to affect the chances of an applicant being certified or denied.



Multi-Variate Data Analysis – Origin and Wages

- Applicants with low prevailing wages from North America and South America appear more likely to be denied.
- Certification appears to be more likely for applicants from Asia, Africa, North America, and South America with higher prevailing wages.



Exploratory Data Analysis Observations

- Visa Certification is greater for those with higher levels of education. A doctorate is the most likely level of education to be approved for Visa, followed by a Master's.
- Applicants from Europe, followed by Africa, Asia, Oceana, North America, and South America, respectfully. No continent has less than a 50% denial rate.
- Work experience improves the chances of a Visa applicant being certified. Individuals with work experienced are certified at a rate of ~75% compared to ~56% for those without work experience.
- Being paid an annual rate leads to the most likelihood of Visa certification. An hourly rate leads to the least likely rate of Visa certification.
- A higher prevailing wage might lead to a greater chance of certification, but it appears that the values visualized would not support this when diving deeper with statistical analysis.

Ensemble Model

Building the Model

Model evaluation criterion

Model can make wrong predictions as:

- Predicting a Visa applicant will be certified but should be denied.
- Predicting a Visa applicant will be denied but should be certified.

Which case is more important?

- Predicting a Visa applicant as certified but should be denied.

Which metric to optimize?

- Precision should be maximized. The greater the Precision, the higher the chances of minimizing false positives. The reason we want to minimize false positive model predictions is that a certified applicant who should be denied can slip through the cracks, but a denied applicant who should be certified has the opportunity to appeal the decision and be certified through a manual/secondary process.

Ensemble Model Overview

Details:

- Three approaches were used: decision tree, random forest, and bagging.
- Each approach was tuned to generalize the training populations to their respective test populations.

Techniques:

- **Ensemble methods** use this same idea of combining several predictive models (supervised ML) to get higher quality predictions than each of the models could provide on its own**.**
- **Bagging**, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset.
- **Decision tree** is one of the predictive modelling approaches used in statistics, data mining and machine learning.
- A **random forest** is a machine learning technique that's used to solve regression and classification problems.

Ensemble Model Performance Metrics

Metrics of Each Model:

Training Metrics:

Generalized

Generalized

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned
Accuracy	1.0	0.638316	1.0	0.676441	0.965239	0.965351
Recall	1.0	0.704542	1.0	0.712308	0.906635	0.904947
Precision	1.0	0.470250	1.0	0.509172	0.987677	0.989843
F1	1.0	0.564033	1.0	0.593849	0.945423	0.945493

Test Metrics:

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned
Accuracy	0.698849	0.641418	0.739927	0.684066	0.740581	0.746860
Recall	0.492320	0.714061	0.420638	0.727846	0.471839	0.461599
Precision	0.552364	0.473615	0.673817	0.517357	0.651087	0.673563
F1	0.520616	0.569499	0.517944	0.604811	0.547157	0.547792

Ensemble Model Performance Summary

Observations

- Several techniques were unable to generalize models to both the training and test data.
- The Decision Tree Estimator and Random Forest Tuned generalize the data best between their training and test data.
- Random Forest Tuned reduces false negatives and false positives as evidenced by its highest F1 score.
- Bagging Estimator Tuned predicts the least amount of False Negatives as evidenced by its high Precision score amongst test data.
- A combination of Random Forest Tuned and Bagging Estimator tuned should be used together to automate Visa certifications and denials due to the Random Forest Tuned does the best at generalizing the data and reducing total errors and Bagging Estimator Tuned should be used as it is the best model to reduce False Negatives.

Boosting Model Performance Summary

Building the Model

Model evaluation criterion

Model can make wrong predictions as:

- Predicting a Visa applicant will be certified but should be denied.
- Predicting a Visa applicant will be denied but should be certified.

Which case is more important?

- Predicting a Visa applicant as certified but should be denied.

Metrics:

- Accuracy, Precision and Recall but the metric of interest here is precision.
- Precision - It gives the ratio of False positives to Actual positives, so high Precision implies low false positives, i.e. low chances of predicting a denial as a certified.

Boosting Model Overview

Details:

- Two approaches were used: AdaBoost and Gradient Boosting
- Each approach was tuned to generalize the training populations to their respective test populations.

Techniques:

- **AdaBoost** algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.
- **Gradient boosting** is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

Boosting Model Performance Metrics

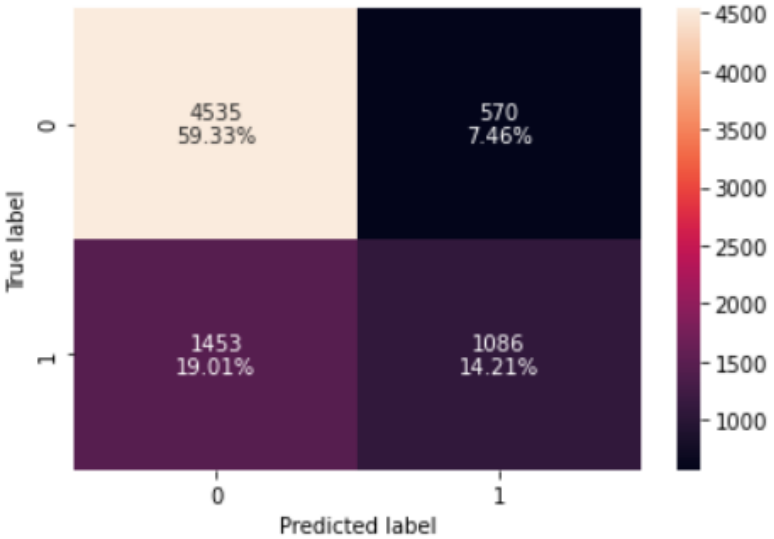
Metrics of Each Model:

	Adaboosting Training	AdaBoosting Test	Gradient Boosting Training	Gradient Boosting Test
Accuracy	0.7385624	0.7353479	0.7512334	0.7509157
Recall	0.4399797	0.4277274	0.5005909	0.5029539
Precision	0.6594129	0.6557971	0.6671917	0.6654507

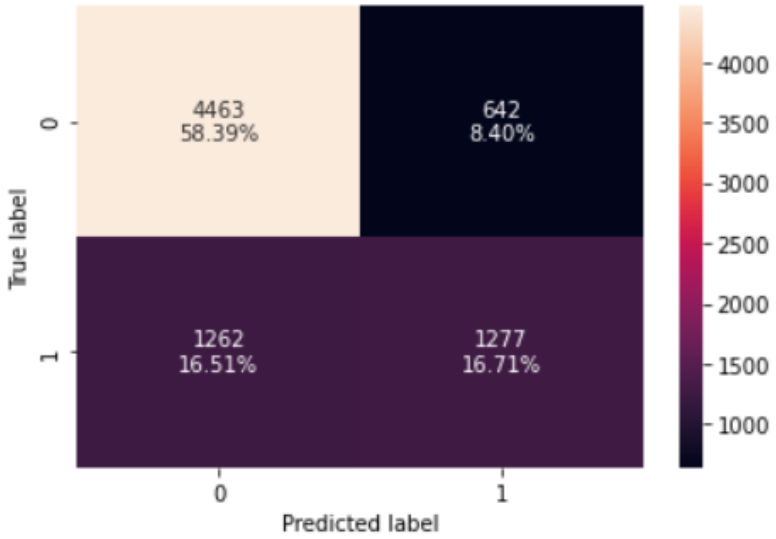
- Each model is generalized.
- Gradient boosting is a slightly better model than Adaboosting.

Boosting Model Confusion Matrices

AdaBoosting:



Gradient Boosting:



Boosting Model Performance Summary

Observations

- Adaboosting leads to lower Accuracy, Recall, and Precision than Gradient Boosting.
- Gradient Boosting is the superior model for this data set.
- The data can be further processed to improve the processing speeds of the machine learning models.
    - The Case ID should have been removed from the ML models since these datapoints are unique to all Visa applicants.

Business Insights

- The data collection can be improved to identify additional variables to assess Visa certifications.
- Additional data points to collect or standardize:
  - Seasonal job: Yes or No
  - All prevailing wages to be annualized
  - Criminal Records/No Fly List Review
  - Key skillsets

Business Recommendations

- EasyVisa should recommend a three part model for automated certifications and denials leveraging the random forest tuned, bagging estimator tuned, and gradient boosting.
- Any Visa applicant that is certified by all three models should be fast tracked to certification.
- Any Visa applicant that is denied by all three models are automatically denied and the applicant can appeal the denial if the sponsoring company desires a reassessment.
- Any Visa applicant who is not certified by all three models should be reviewed manually to determine if the applicant should be certified or denied.
- An assurance process should be developed to conduct manual random sampling of the automated certified and denial populations to ensure the models is accurate and assess when the models should be retrained.

greatlearning
Power Ahead

Happy Learning !