ML based solution for predicting US work visa application status outcomes

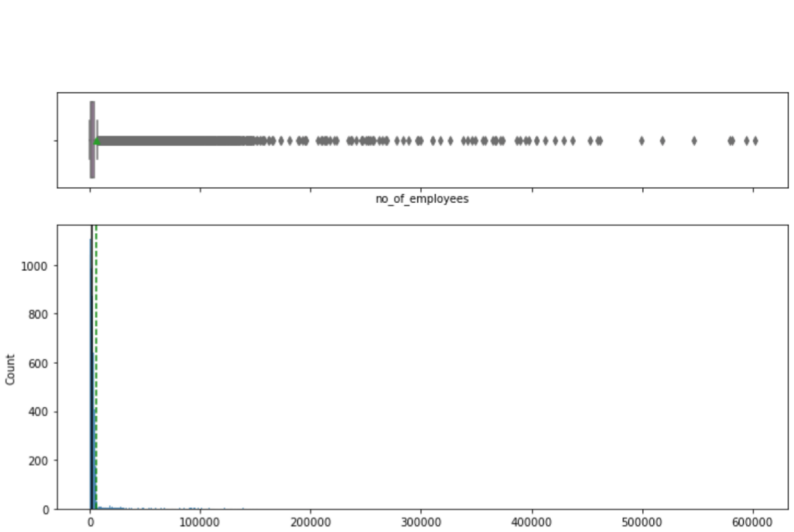
Need: Number of US work visa applicants are increasing 10 folds compared to previous years and process of reviewing the same is becoming more tedious. This calls for a machine learning solution that can help in shortlisting candidates having higher chances of VISA approval by studying the following attributes -

- ***case_id:*** ID of each visa application
- ***continent:*** Information of continent of 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

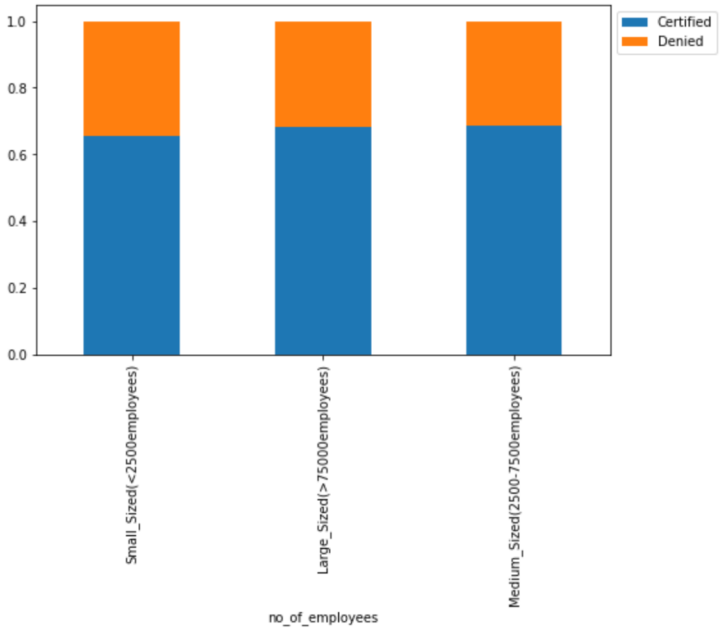
Predictor variable --> case_status: Flag indicating if the Visa was certified or denied

Exploratory Data Analysis – Univariate & Bivariate

(1) Number of employees

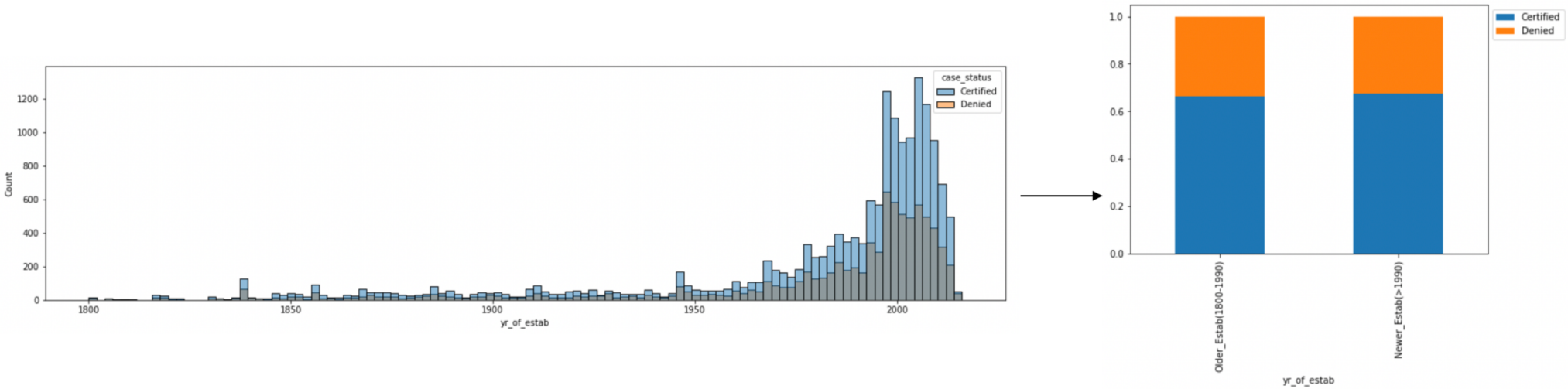


Binning into categories →



- The distribution of number of employees is skewed right with several outliers. However, greater than twice the number of cases (i.e., 65%) are certified than denied both for employers having lesser as well as more number of employees
- While a decision tree ML model is robust to outliers, binning into practical bins (3+) will decrease model building time & help in visualizing the trend
- 58% are small sized companies (less than 2500 employees), 36% are medium sized and 6% are large sized companies (more than 7500 employees)

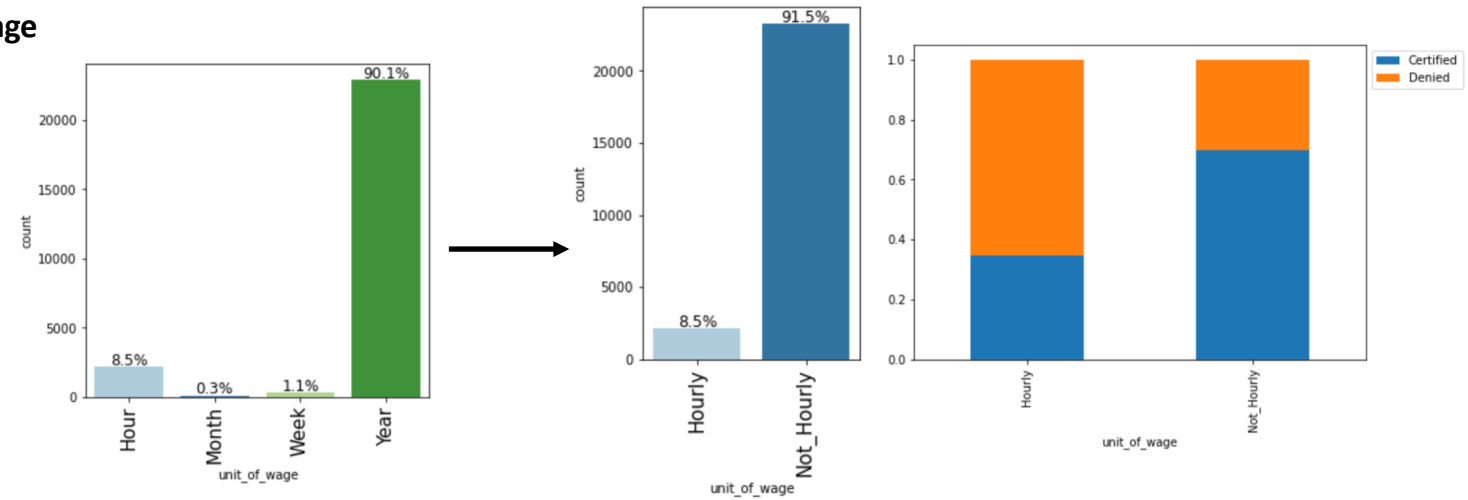
(2) Year of establishment of employers



- Distribution for year of establishment is skewed left. ~65% cases are certified for employer's independent of year of establishment
- Binning likewise, we observe around 61% of employers were established after 1990 and 39% of employers before 1990

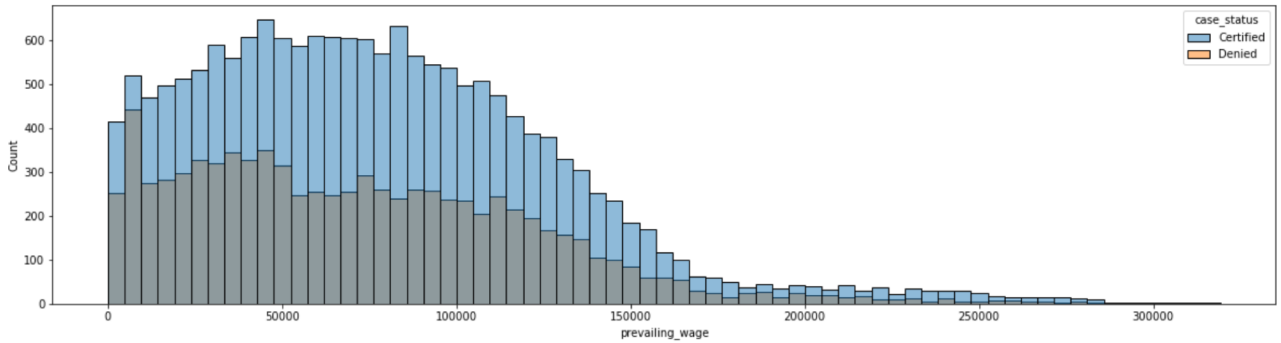
(3) Prevailing wage & unit of wage

- Unit of wage is assumed to be not hourly, when employee receives fixed salary irrespective of number of hours worked (i.e.,Week, Month, or Year) & hourly otherwise (i.e., Hour)
- Almost 92% of all entries are with unit_of_wage as not hourly & only 8% entries as hourly
- 70% of cases are certified when the unit_of_wage is not hourly, and only 35% cases are certified when the unit_of_wage is hourly



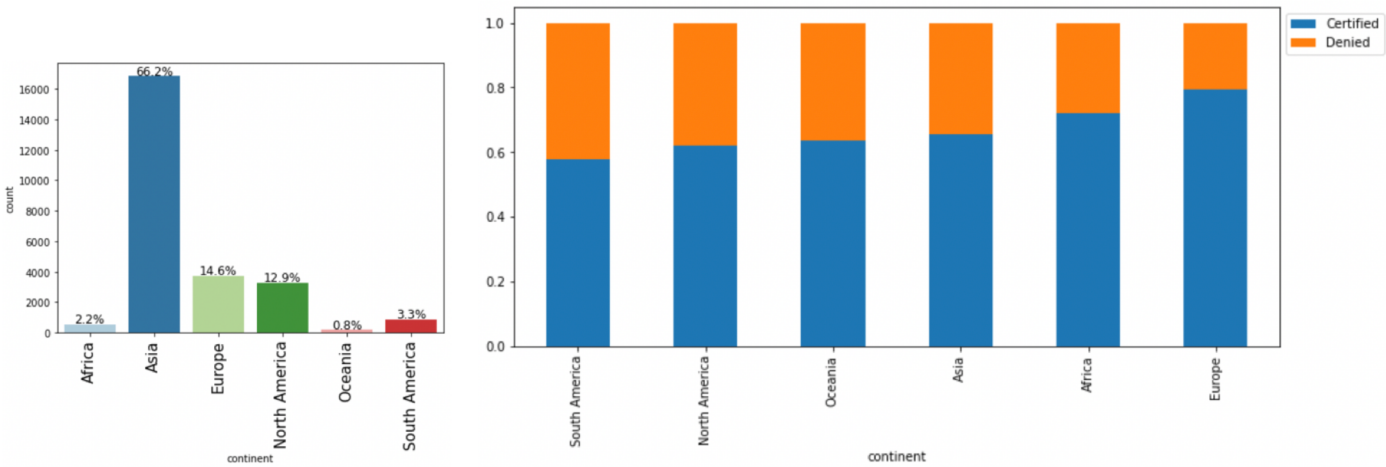
Count & % certification for visa statuses by unit of wage

- Prevailing_wage was cleaned up to only contain annual wages
- Values on the lower end (<US$14K) are concerning as well there are outliers on the higher end (>US$200K). These are not treated further as decision tree ML model is robust to outliers
- General trend that ~ twice the cases are certified more than denied, dropping slightly & increasing slightly on lower & upper end of wages respectively



(4) Continent of employee

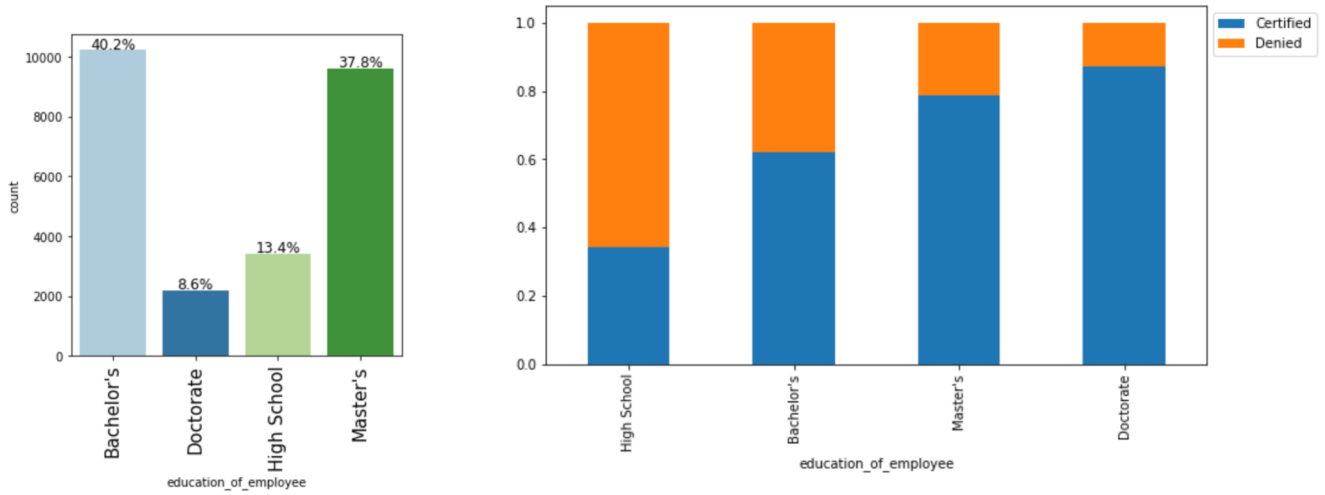
- Majority of employees (>50%) are from Asia
- Cases getting certified is highest for Europe (80%), then Africa (72%), then Asia (65%), & least for S.America & N.America (around 60%)



Count & % certification for visa statuses by continent of employees

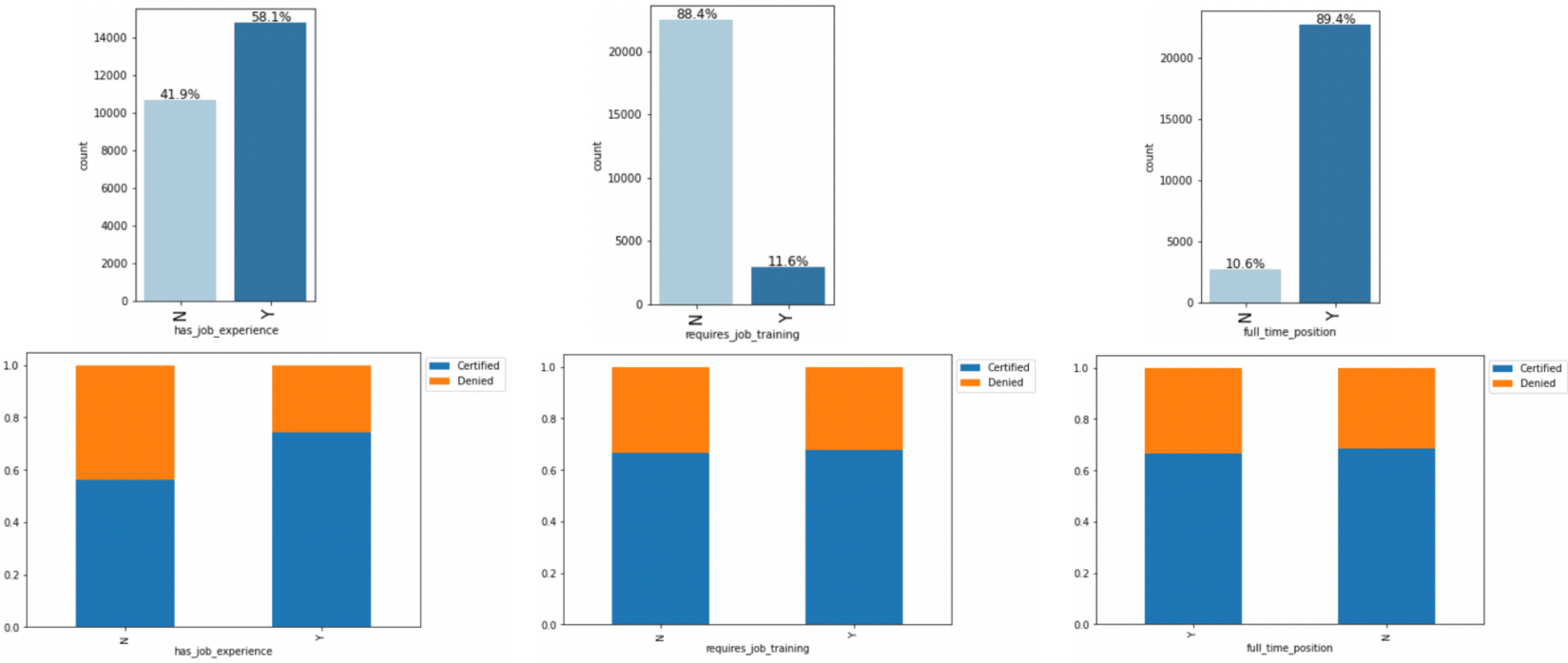
(5) Education of employee

- Majority of employees have either a bachelor's (40%) or a master's (38%) and minority of applicants have either a doctorate (8%) or only a high school diploma (13%)
- Cases getting certified is highest for doctorate degree (>86%),followed by master degree (>76%), then bachelor's (~62%) & high school (<35%)



Count & % certification for visa statuses by education of employees

(6) Prior work experience of employees, if employees require job training, if employment is a full time opportunity

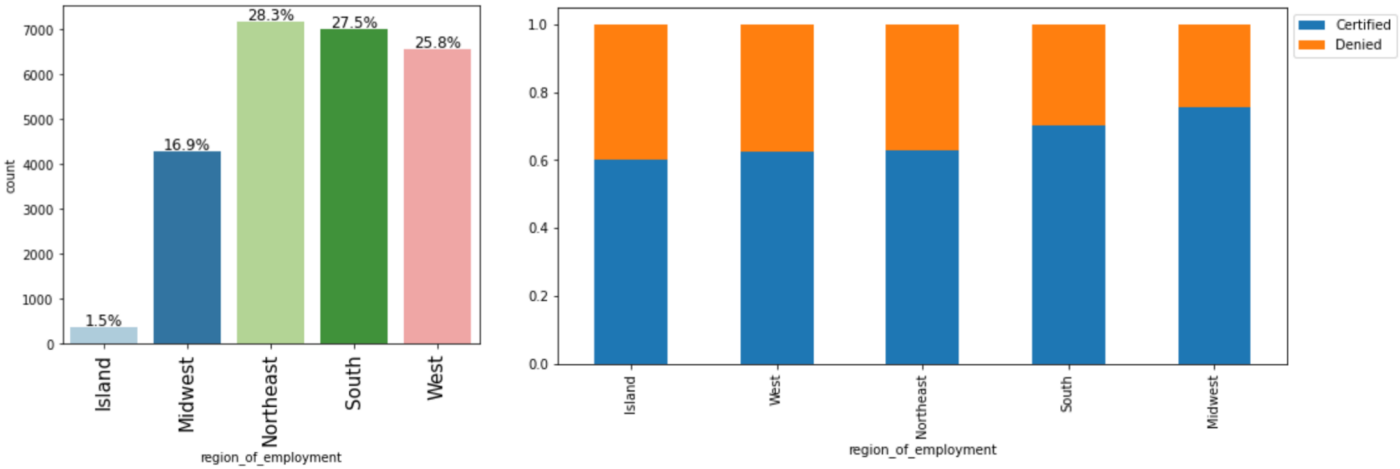


Count & % certification for visa statuses for different attributes

- 58% of all applicants have prior job experience and 42% do not. Cases certified is high for applicants with prior job experience (75%) & low for applicants without prior job experience (~56%)
- Majority do not require the employee to receive any additional job training & are are full time rather than part time opportunities. These attribute were not found to have an impact on the case statuses with equal number of cases getting certified independent of the attributes

(7) Region of employment in the US

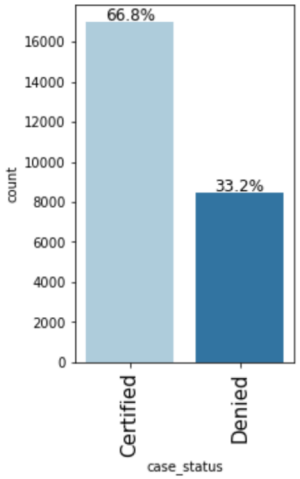
- Majority of the applications are to Northeast (28.3%), then South (27.5%), then West (25.8%), Midwest (16.9%) and least to Island (1.5%) regions
- Cases certified follows Midwest (75%), then South (70%), then Northeast, West, & Island (60%)
- Region of employment being Midwest is an important attribute contributing positively to a case being certified



Count & % certification for visa statuses by region of employment

(7) Visa case status

Approximately, 67% of all cases are approved and 33% of all cases are denied



Correlation score between attributes is not studied. Unlike linear regression, classification based on decision tress & related ensemble methods are not affected by multicollinearity.

Model Evaluation Criteria

The model can make wrong predictions as:

- *Certifying a case when the required criteria are not met*
    - This could have an adverse impact leading to job loss for citizens and locals in the US as well, the foreign workers will be hired at less than ideal pay for market rate & their calibre

- *Denying a case when the required criteria are met*
    - If cases are denied where criteria are met, it could also have an adverse impact leading to human resource shortages across occupations in the US, slowing down the economy

We would want F1-Score to be maximized, the greater the F1-Score higher the chances of predicting both the classes correctly

The dataset is split into 70:30 training : testing groups, and the following models are built & optimized:

- Decision Tree & Hyperparameter Tuned Decision Tree
- Random Forest & Hyperparameter Tuned Random Forest
- Bagging Classifier & Hyperparameter Tuned Bagging Classifier

Boosting models –

- AdaBoost Classifier & Hyperparameter Tuned AdaBoost Classifier
- Gradient Boosting Classifier & Hyperparameter Tuned Gradient Boosting Classifier
- XGBoost Classifier & Hyperparameter Tuned XGBoost Classifier
- Stacking classifier (Using Hyperparameter Tuned Decision Trees, Hyperparameter Tuned AdaBoost & Gradient Boosting classifier as base estimators and Hyperparameter Tuned XGBoost Classifier as the final estimator)

The confusion matrix is plotted for testing dataset & metrics – Accuracy, Recall, Precision and F1_score are studied and compared across models

Model Comparison

Testing performance comparison:

	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.661559	0.709103	0.676621	0.724951	0.688016	0.728225	0.735560	0.745514	0.748527	0.746431	0.737525	0.746955	0.745907
Recall	0.743384	0.929034	0.760047	0.761419	0.757106	0.877475	0.877671	0.861596	0.865517	0.865321	0.858851	0.887473	0.866301
Precision	0.748372	0.718248	0.756931	0.814768	0.771628	0.755316	0.762432	0.780362	0.781554	0.779446	0.773345	0.769244	0.778404
F1	0.745869	0.810155	0.758486	0.787191	0.764298	0.811826	0.816003	0.818970	0.821395	0.820141	0.813858	0.824140	0.820004

- Decision tree, Random forest (default & tuned), Bagging classifier (default & tuned) & XGBoost (default) were found to overfit the training dataset. This is concluded as the measured metrics are higher for training dataset than testing dataset for these ML models
- Decision tree tuned, AdaBoost (default & tuned), Gradient boost (default & tuned) and XGBoost (tuned) were found to give generalized (similar) performance on the training & testing datasets. Of these, *the XGBoost_tuned has the highest F1 score* (although all models have more or less similar performance)

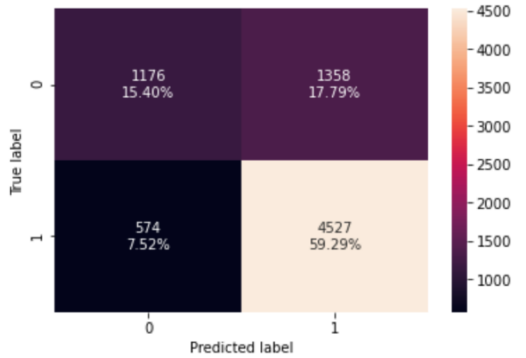
- The XGBoost_tuned ML model is able and is able to explain over 80% of information (accuracy of 75% on test dataset & F1 score of 82% on test dataset).
    - The precision & recall are likewise both high (77% & 88% respectively)
    - Confusion matrix (test data) is able to identify a higher % of cases getting certified (>88%), but only a smaller % of cases getting denied correctly (~54%). This limitation has to be borne in mind, and perhaps a re-evaluation of cases getting denied can be carried out in case there is a a prevailing human resource shortage in the US (the model still considerably saves time)

Training performance:

	Accuracy	Recall	Precision	F1
0	0.753818	0.898739	0.770811	0.829874

Testing performance:

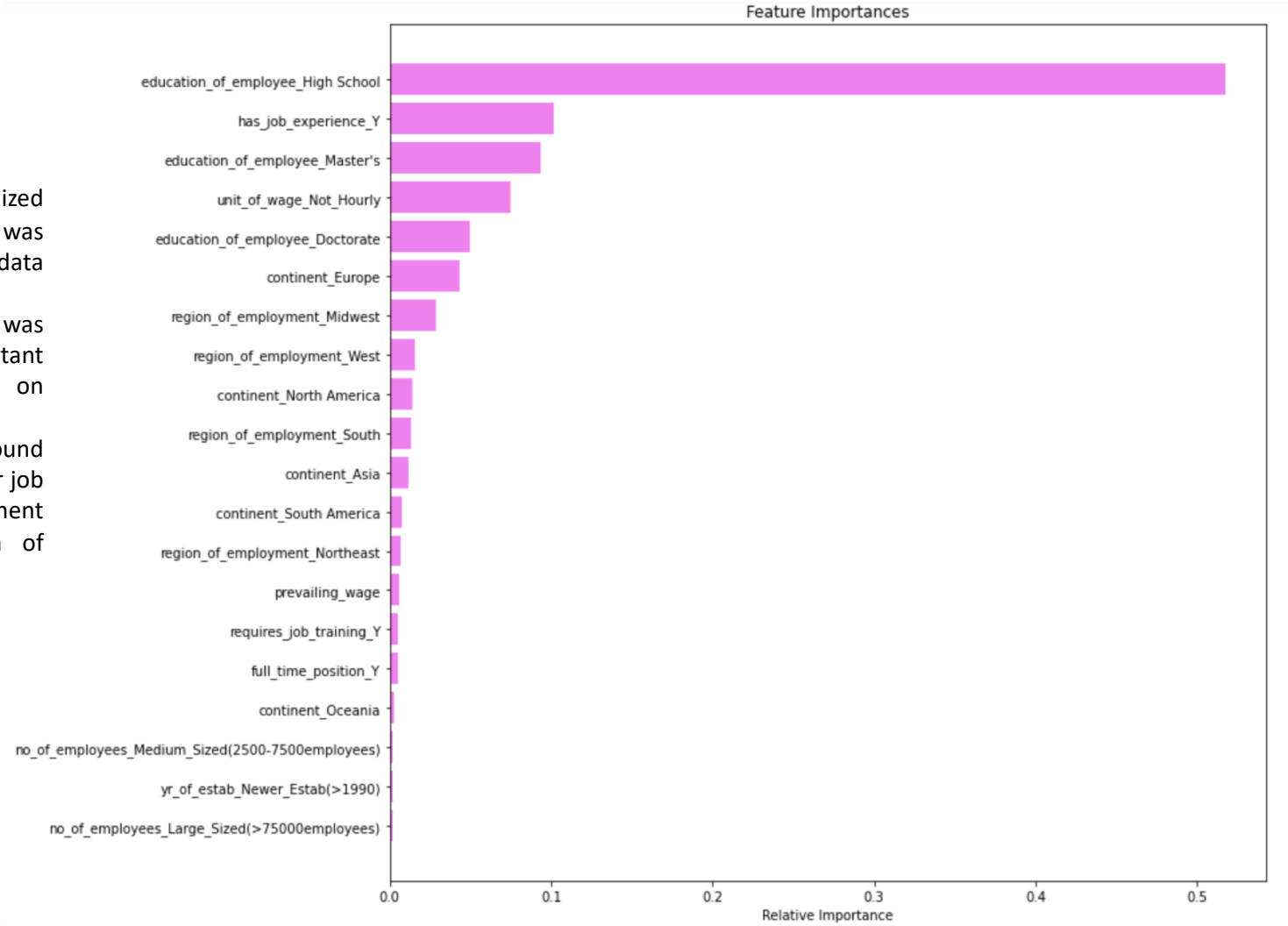
	Accuracy	Recall	Precision	F1
0	0.746955	0.887473	0.769244	0.82414



Confusion matrix (testing) & metric score for training & testing using XGBoost (tuned) model

XGBoost (tuned) Feature Importance

- The findings from the optimized model is similar to what was observed post exploratory data analysis
- Education of the employee was found to be the most important attribute having an influence on visa certifications
- Other important attributes found were - if an employee has prior job experience, unit of wage, continent of the employee, & region of employment in the US



Insights and Recommendations

- Based on the EDA and the XGBoost tuned model, the following features were identified as important for visas getting certified than denied
  - ***(1) Education of employee***
    - an employee with only a high school certification has over 65% chance of visa getting denied in comparison to an employee with a doctorate degree with over a 85% chance of visa getting certified
  - ***(2) Unit of wage***
    - an employee with an hourly pay likewise has over 65% chance of visa getting denied in comparison to an employee with a non-hourly pay (weekly, monthly or yearly) with over 70% chance of visa getting certified
  - ***(3) Continent the employee is from & Prior work experience of employee***
    - an employee from Europe has over 80% chance of visa getting certified
    - an employee with prior work experience has over 75% chance of visa getting certified
  - ***(4) Region of the US the employment opportunity is in***
    - over 70% cases getting certified if the region is Midwest or South

- Attributes like if the job opportunity is full time/ part time ; if an employee requires further job training ; the annual prevailing wage of the occupation in the US ; year of establishment of the employer or the number of employees in the organization are not important attributes & do not have much bearing on a case getting certified vs denied

- ***The findings can help build a suitable profile of candidates to facilitate the process of visa approvals.*** Our model is able to capture over 80% of the information while making predictions. % certifications correctly identified is high as per test confusion matrix (>88%) while % denied is on the lower end (~54%), which indicates requiring a re-evaluation of cases getting denied. This approach is still expected to save over 60% processing time