EasyVisa

Eugenie Seholm

Contents

Business Problem - Overview

Problem:

- In FY2016, the Office of Foreign Labor Certification (OFLC) processed a 9% increase in the overall number of processed applications from the previous year
- The process of reviewing every case is a tedious task as the number of applications is increasing every year.

Objective:

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

Business Problem Overview - Financial Implications

False positive: predicting an application for Visa should be certified when it should be denied

- Allows a foreigner to enter the country to fill a position that should be filled by US worker
- Contributes to an increase in unemployment

False negative: predicting an application for Visa should be denied when it should be certified

- Leaves vacancy for employment position that cannot be filled by US workers due to workforce shortages
- Can be detrimental to the company with the unfilled position

Business Problem - Solution Approach

- Data preparation
- Split the data into train and test sets
- Train models using training data
- Try to improve the model performance using hyperparameter tuning
- Test the performance on the test data, maximizing F1-score in order to increase the chances of predicting both classes correctly to minimize false positives and false negatives

Data Overview

- 25,480 rows (observations)
- 12 columns (variables)
- 0 duplicated rows
- 0 missing values

Variable	Description
case_id	ID of each visa application
continent	Information of continent of the complopyee
education_of_employee	Information of education of the employee
has_job_experience	Whether or not employee has job experience
requires_job_training	Whether the employee requires any job training
no_of_emlpoyees	Number of employees in employer's company
yr_of_estab	Year in which the employer's company was established
region_of_employment	Information of 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
unit_of_wage	Unit of prevailing wage (hour, weekly, monthly, yearly)
full_time_position	Whether the position of work is full-time
case_status (target)	Flag indicating if the Visa was certified or denied

Data Manipulations

Changed structure of:

- education_of_employee
- has_job_experience
- requires_job_training
- full_time_position
- case_status

- education_of_employee
  - High school → 1
  - Bachelor's → 2
  - Master's → 3
  - Doctorate → 4
- has_job_experience
  - N → 0, Y → 1
- requires_job_training
  - N → 0, Y → 1
- full_time_position
  - N → 0, Y → 1
- case_status
  - Denied → 0, Certified→ 1

Data Manipulations

- Limit no_of_employees
- Limit prevailing_wage
- Limit unit_of_wage

- no_of_employees
  - Greater than or equal to 0
- prevailing_wage
  - Used Foreign Labor Certification Data Center Online Wage Library to find data reported for 2016
  - Greater than or equal to 7.31
  - Less than or equal to 290690
- unit_of_wage
  - Built models without limiting and with limit to just year
  - Models with limit had prevailing wage below poverty line dropped

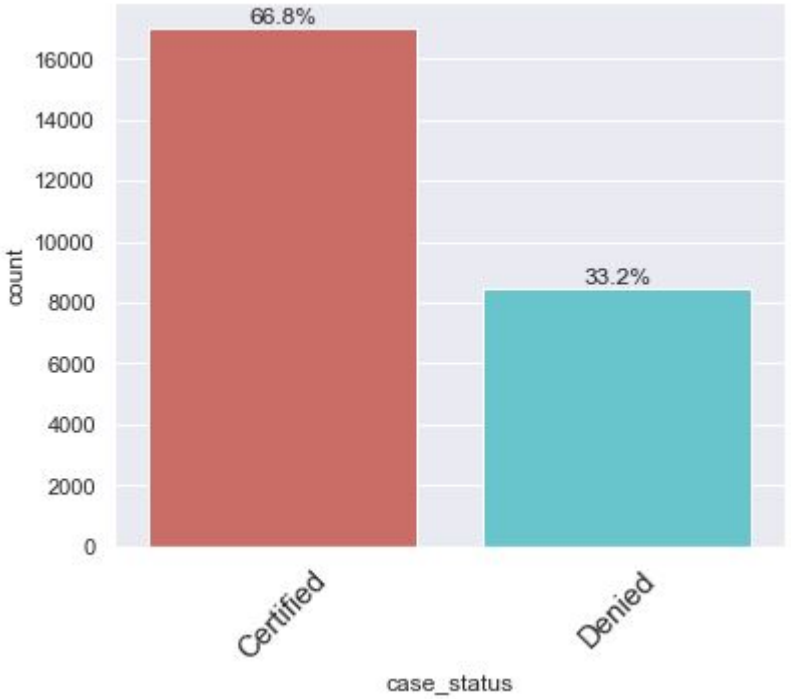
Data Manipulations

Train test split

- 70:30 ratio
- Consistency throughout train and test sets for distribution of classes in target variable

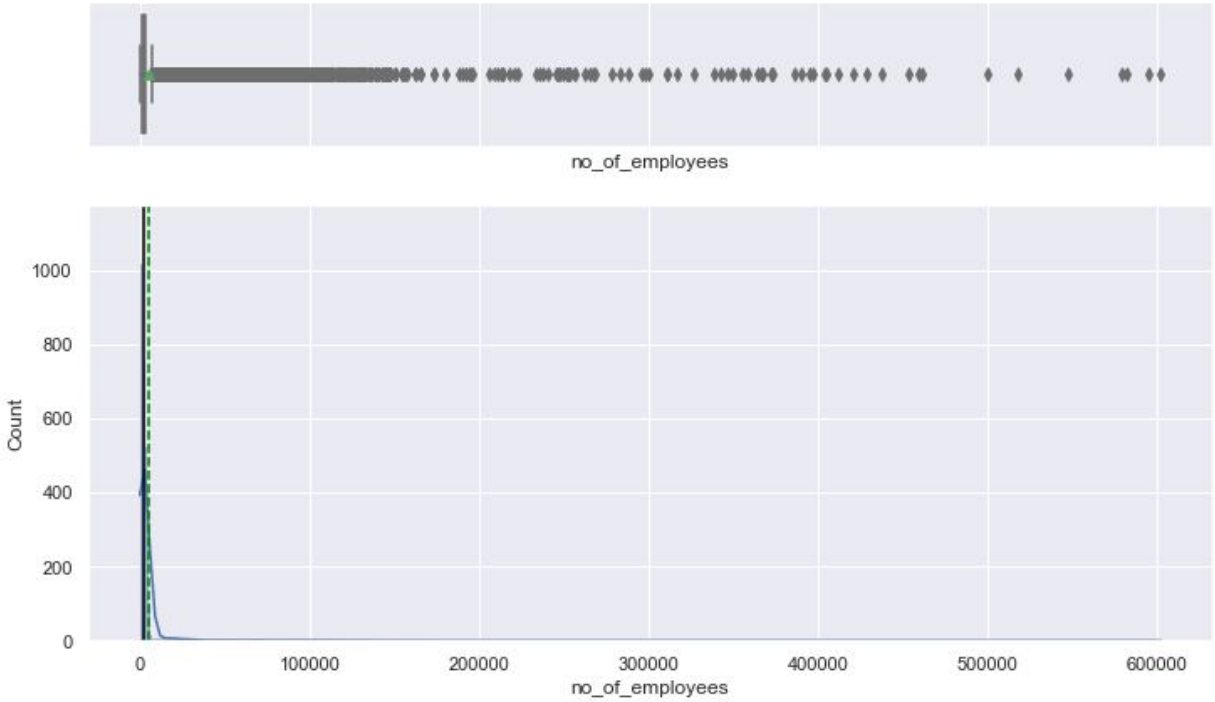
EDA: Case Status (Target)

- Majority (approx two-thirds) certified
- Certified 66.8%
- Denied 33.2%

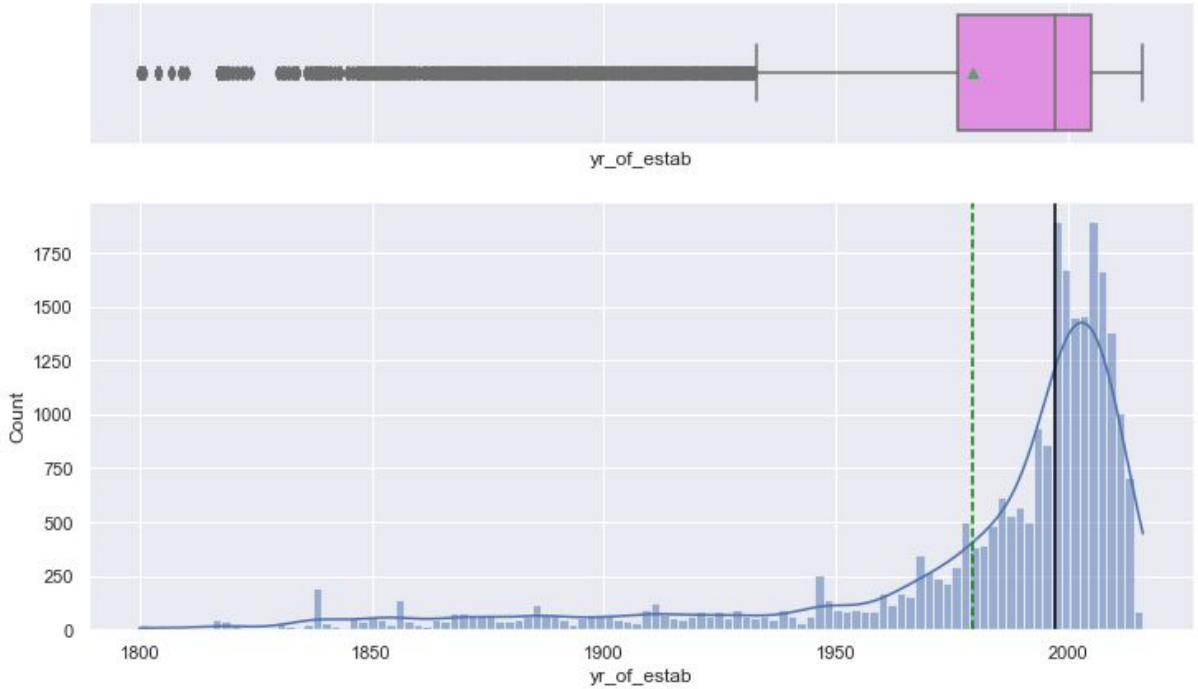


EDA: Number of Employees

- Heavily skewed
- Some employers have negative number of employees prior to data manipulation



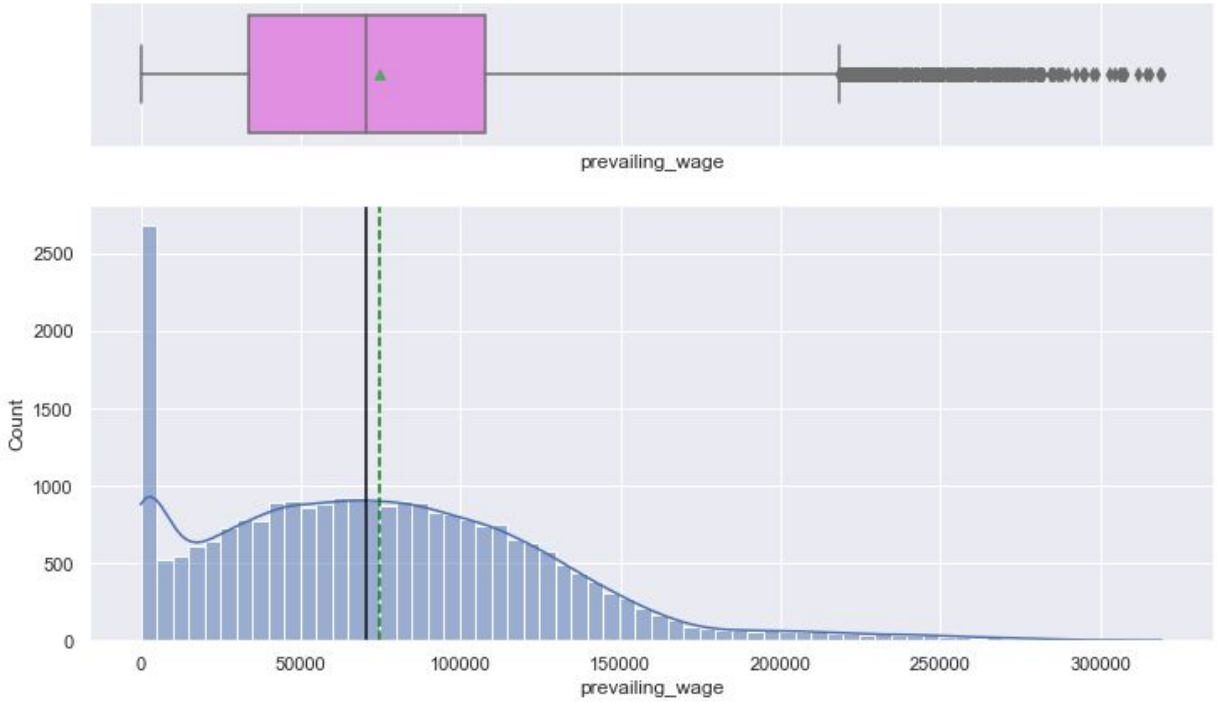
EDA: Year of Establishment



- Heavily skewed to the left
- Large majority of data is from 1930s and on
- 75% data covered by last 40 years

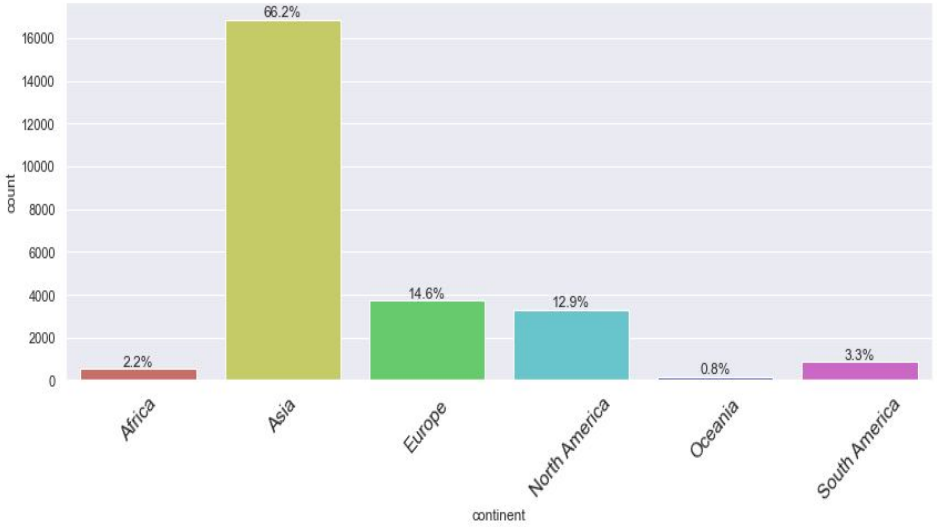
EDA: Prevailing Wage

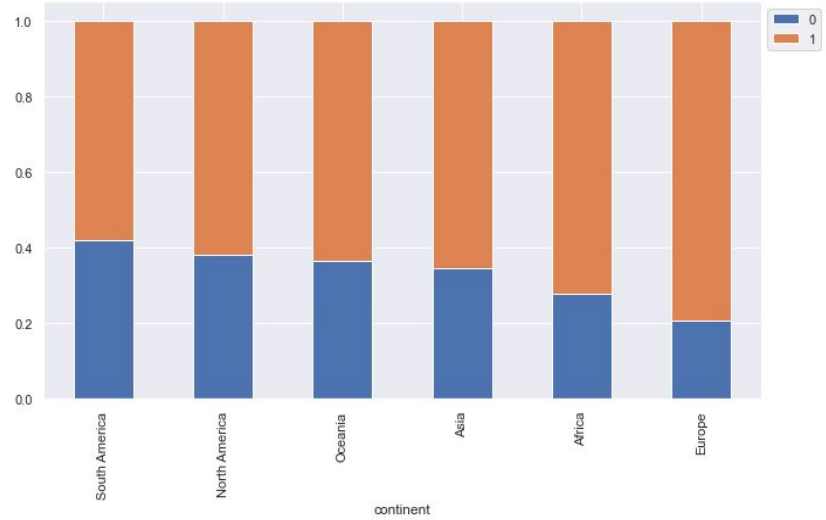
- Large number of jobs with low reported prevailing wage
- Positive skew
- Median wage is approximately what is expected for median salary
- Many salaries below poverty line



EDA: Continent

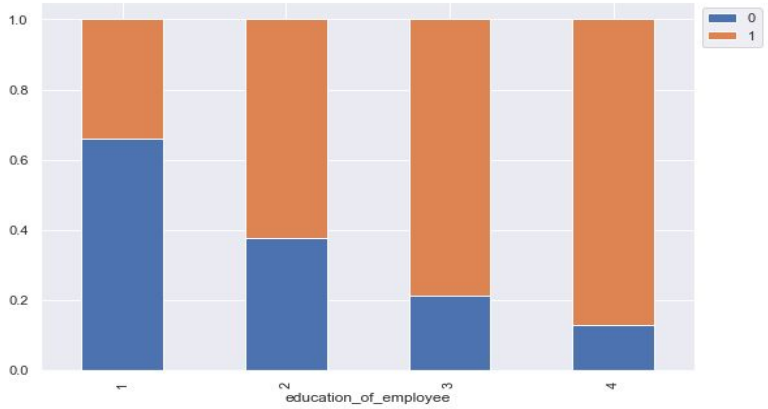
- Most from Asia
- Very few from Oceania
- Asia 66.2%
- Europe 14.6%
- North America 12.9%
- South America 3.3%
- Africa 2.2%
- Oceania 0.8%

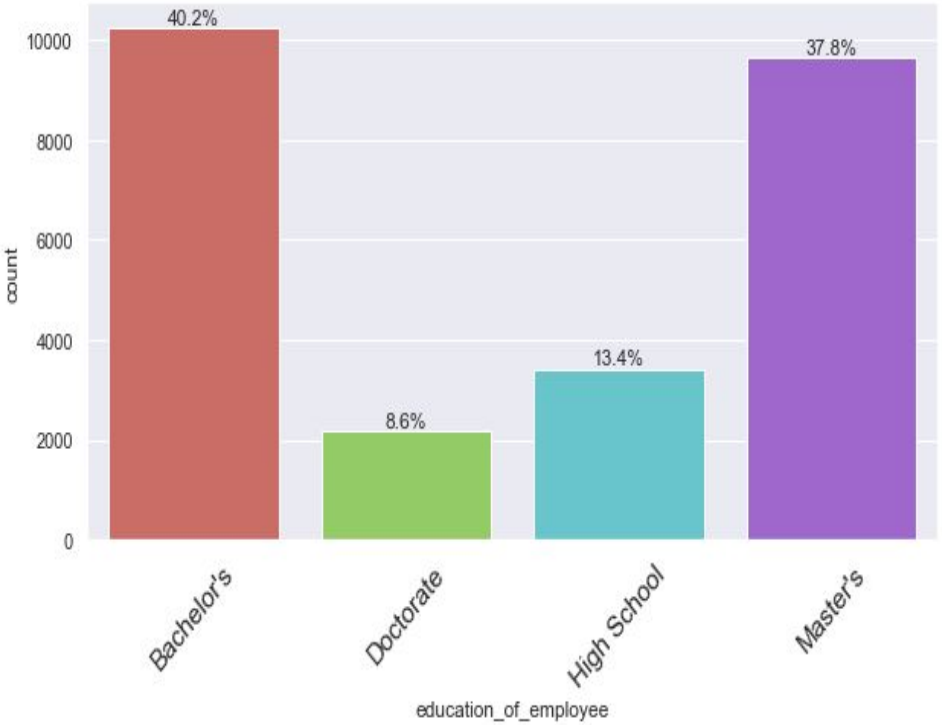




EDA: Education of Employee

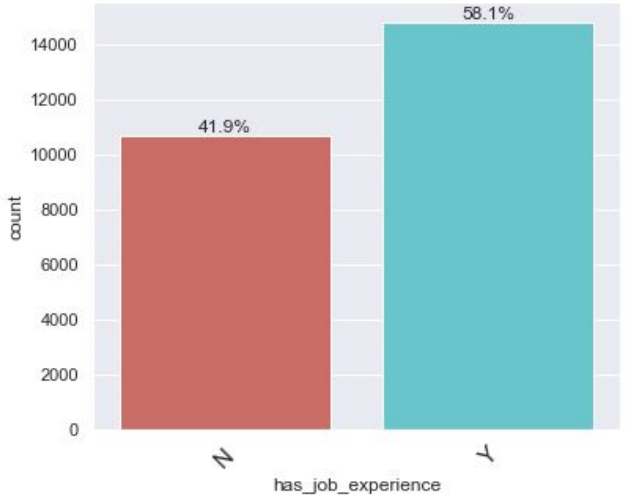
- All applicants have at least high school degree
- 40.2% Bachelor’s
- 37.8% Master’s
- 13.4% High School
- 8.6% Doctorate

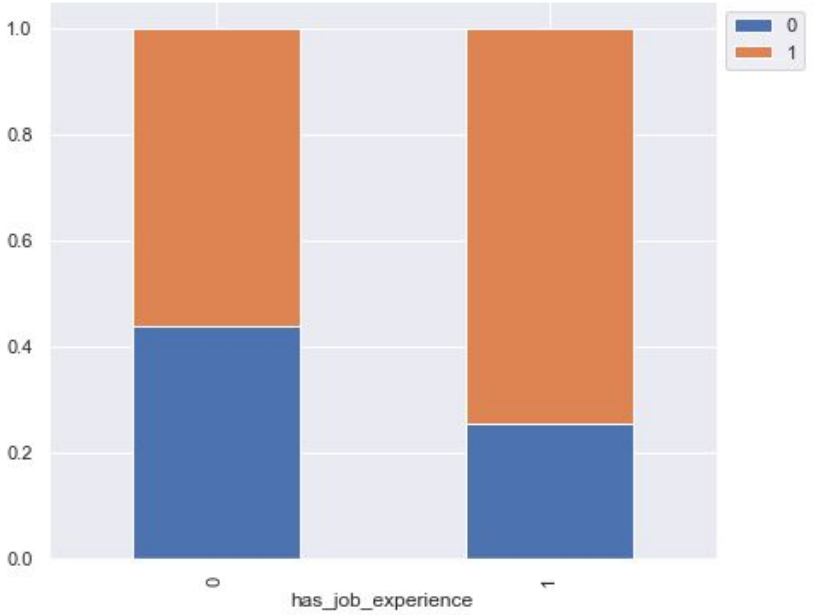




EDA: Has Job Experience

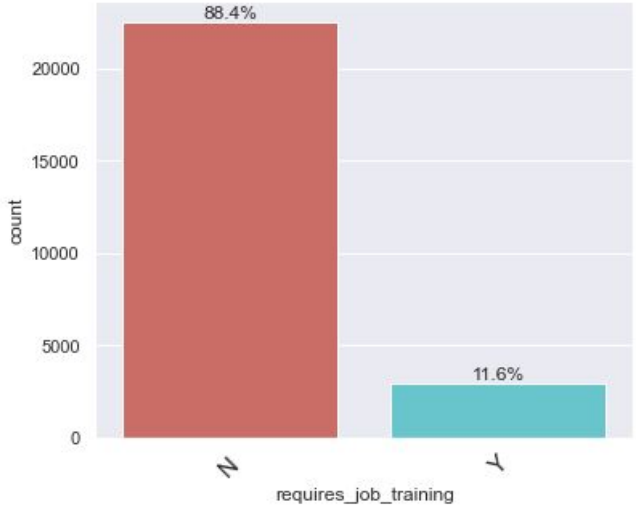
- 58.1% have work experience (majority)
- 41.9% do not have work experience
- Higher proportion of those with work experience are certified

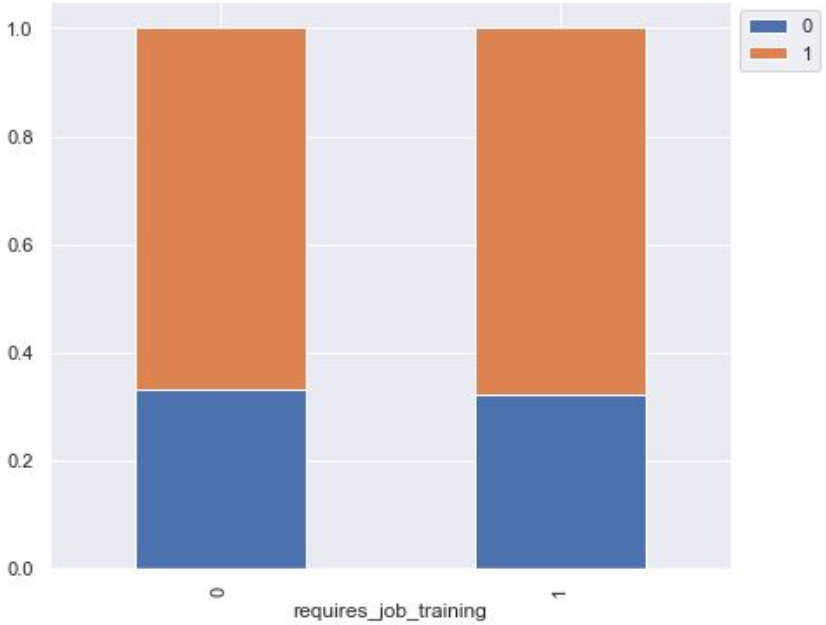




EDA: Requires Job Training

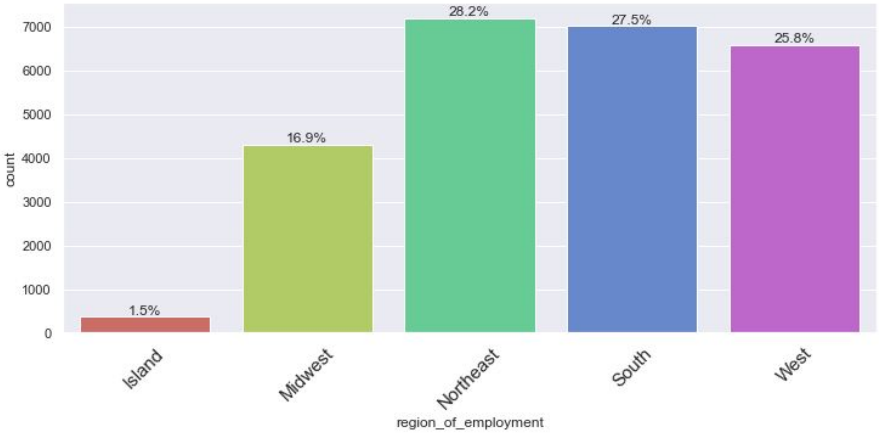
- 88.4% jobs do not require training (majority)
- 11.6% job require training
- Similar proportion approved/denied



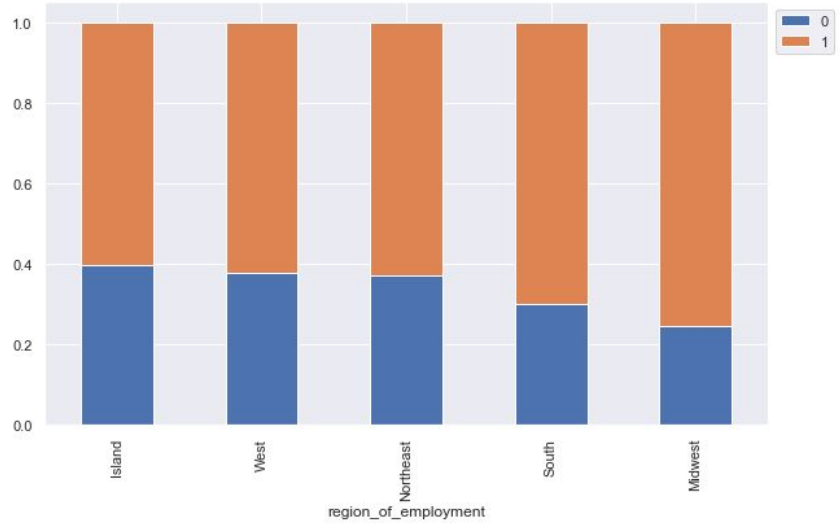


EDA: Region of Employment

- Not many for Midwest and few for Island
- Northeast 28.2%
- South 27.5%
- West 25.8%
- Midwest 16.9%
- Island 1.5%

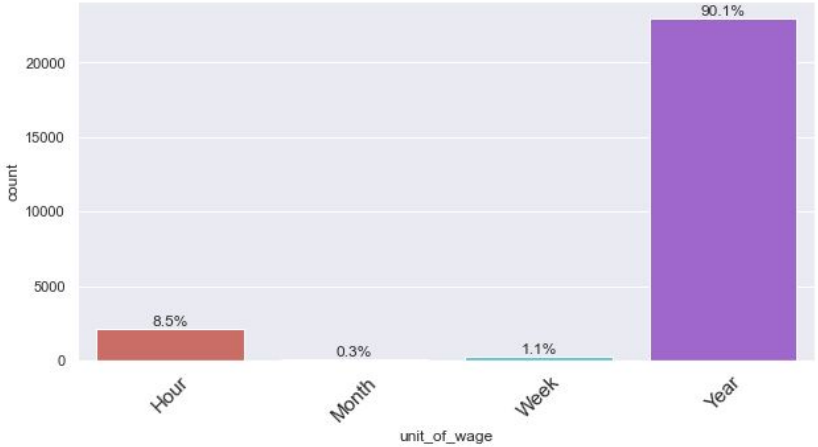


- Island has lowest proportion certified, followed by Midwest, West, Northeast, then South
- In absolute terms, the Island actually had lowest number certified, followed by Midwest, W, NE, S

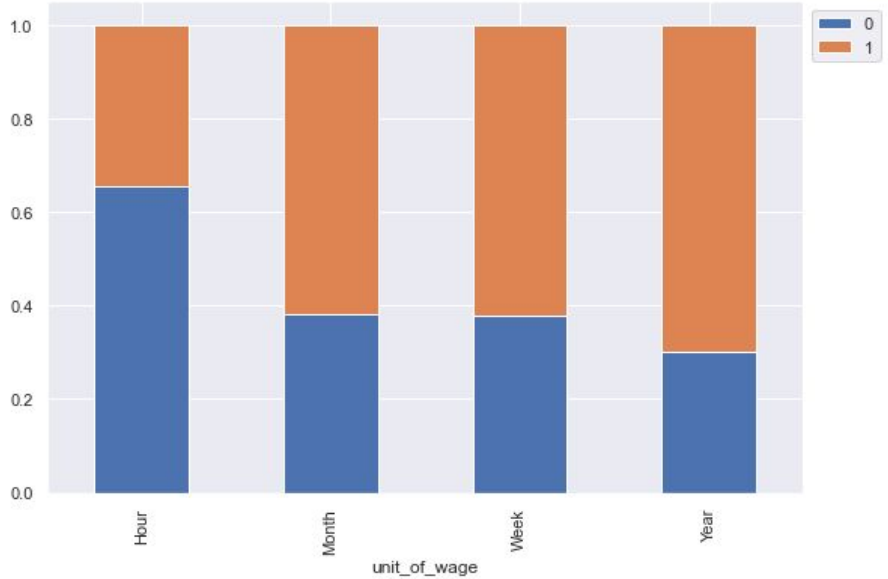


EDA: Unit of Wage

- Mostly year unit of wage, hardly any for month
- 90.1% Year
- 8.5% Hour
- 1.1% Week
- 0.3% Month

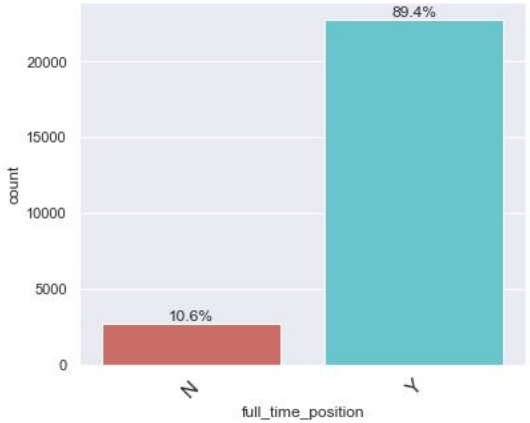


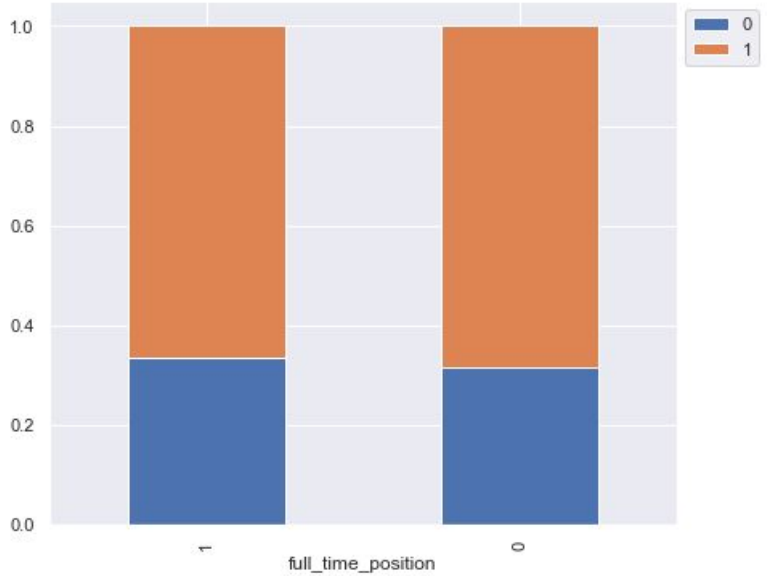
- Year: highest proportion of certified, followed by week, month, then hour
- In absolute terms, year had most total number certified, followed by hour, week, then month



EDA: Full Time Position

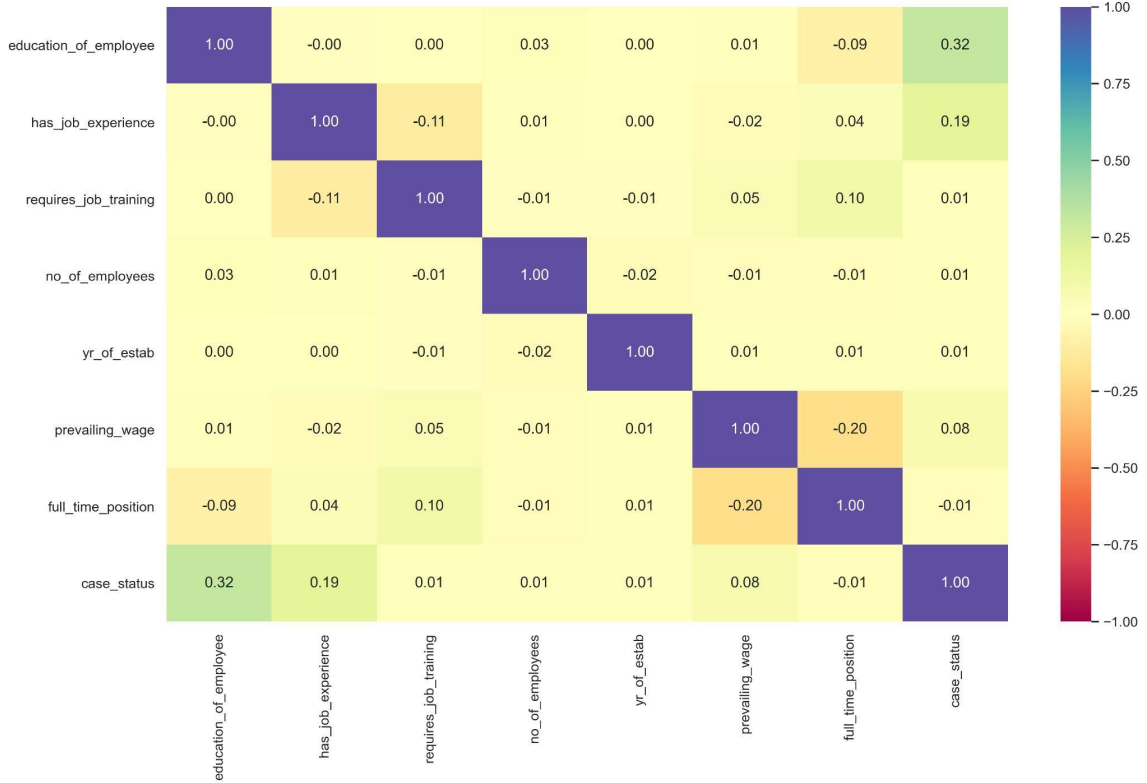
- Majority full-time positions
- Full-time had slightly lower proportion certified, but much higher absolute number
- Full-time 89.4%
- Not full-time 10.6%





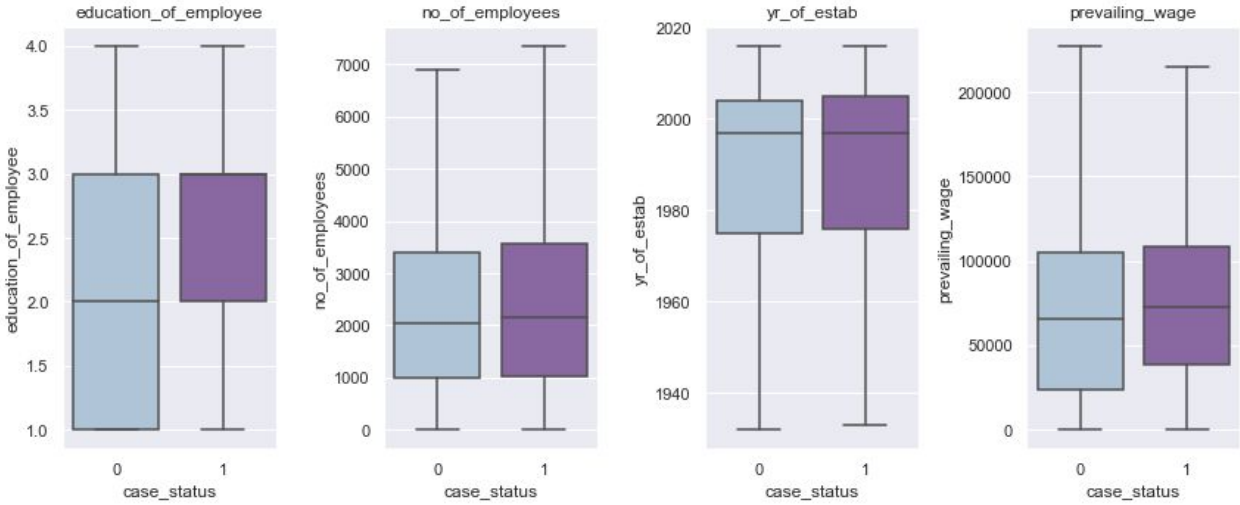
EDA: Correlation Heatmap

- No strongly correlated variables
- Weak positive correlation between case status and education of employee
- Very weak correlation between case status and job experience
- Weak negative correlation between full-time position and prevailing wage



EDA: Case Status vs Education and Numerical Variables

- Proportion of denied applicants decreases with higher degrees
- Number of employees similar between classes
- Year of establishment similar between classes
- Prevailing wage slightly higher for certified, although range is smaller



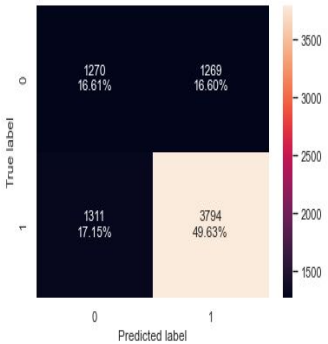
EDA Key Findings and Insights

- Some variables appear to be significant in classifying Visa status, according to EDA:
  - Prevailing wage
  - Education of employee
  - Has job experience
  - Number of employees
  - Year of establishment
- Build models to determine which variables contribute to classifying Visa status

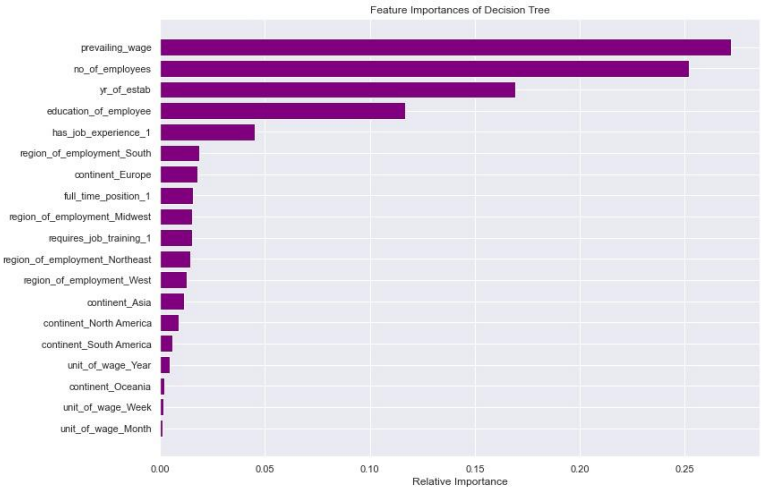
Model Overview and Performance

Decision Tree Classifier

	Accuracy	Recall	Precision	F1
Training performance	1.0	1.0	1.0	1.0
Testing performance	0.66248	0.743193	0.749358	0.746263



- Overfitting
- Training performance perfect
- Test performance lowest F1-score

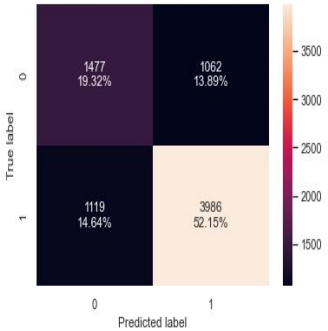


Most important features:

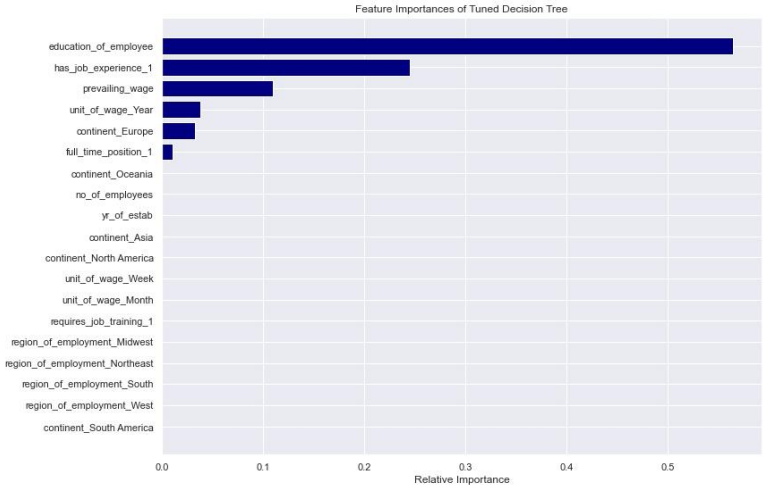
1. Prevailing wage
2. Number of employees
3. Year of establishment

Tuned Decision Tree Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.716192	0.778477	0.792853	0.785599
Testing performance	0.714678	0.780803	0.78962	0.785187



- Not overfitting
- Test performance F1-score better than decision tree and Bagging Classifier

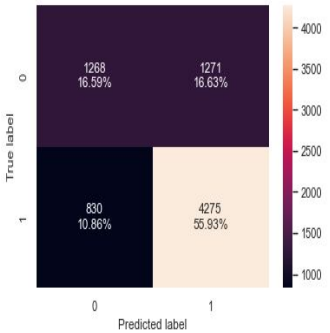


Most important features:

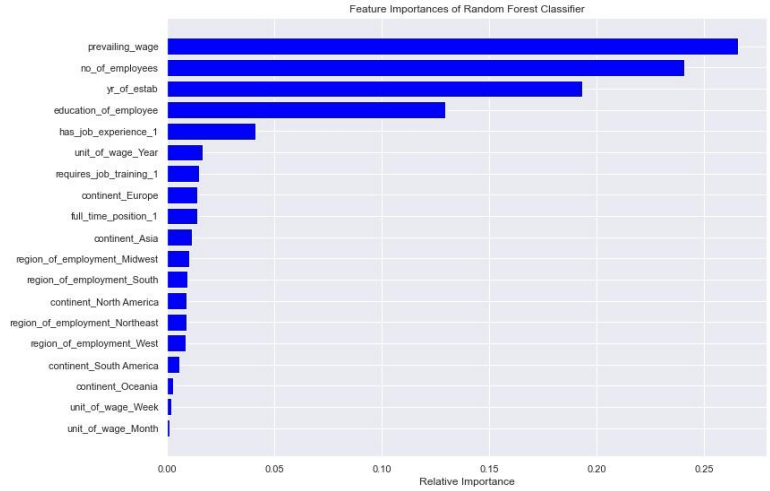
1. Education of Employee
2. Has Job Experience
3. Prevailing wage

Random Forest Classifier

	Accuracy	Recall	Precision	F1
Training performance	1.0	1.0	1.0	1.0
Testing performance	0.725144	0.837414	0.770826	0.802742



- Overfitting
- Perfect training performance
- Test performance has F1-score higher than decision trees and bagging classifier

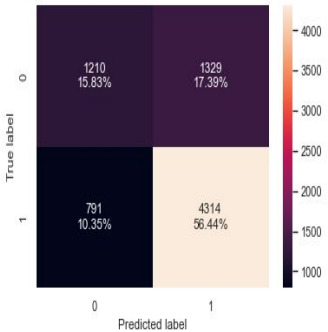


Most important features (similar to DT):

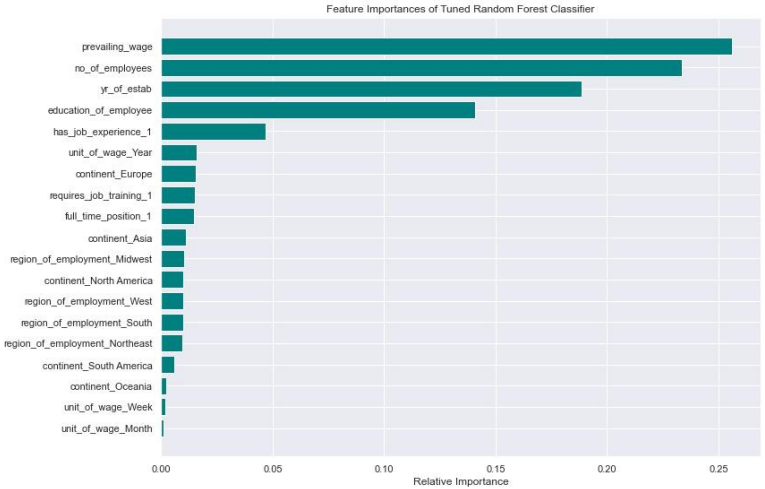
1. Prevailing wage
2. Number of employees
3. Year of establishment

Tuned Random Forest Classifier

	Accuracy	Recall	Precision	F1
Training performance	1.0	1.0	1.0	1.0
Testing performance	0.722658	0.845054	0.764487	0.802754



- Overfitting training data
- Test performance has F1-score very close to random forest (not tuned)



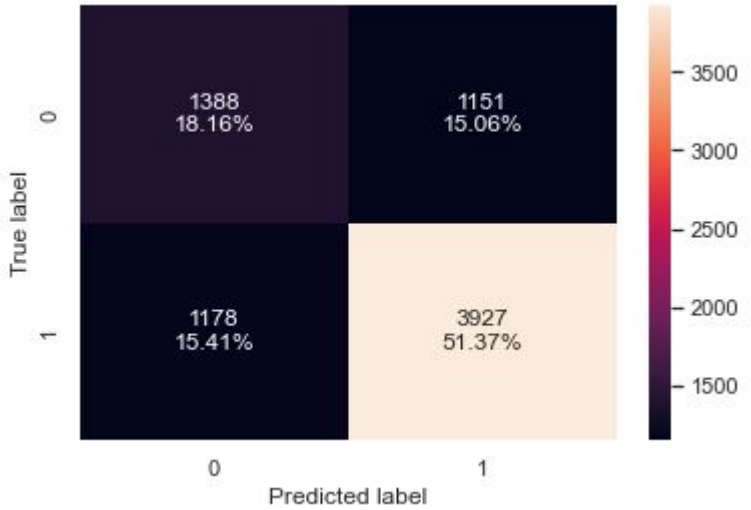
Most important features (similar to DT & RF):

1. Prevailing wage
2. Number of employees
3. Year of establishment

Bagging Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.98318	0.984723	0.990041	0.987375
Testing performance	0.695317	0.769246	0.773336	0.771285

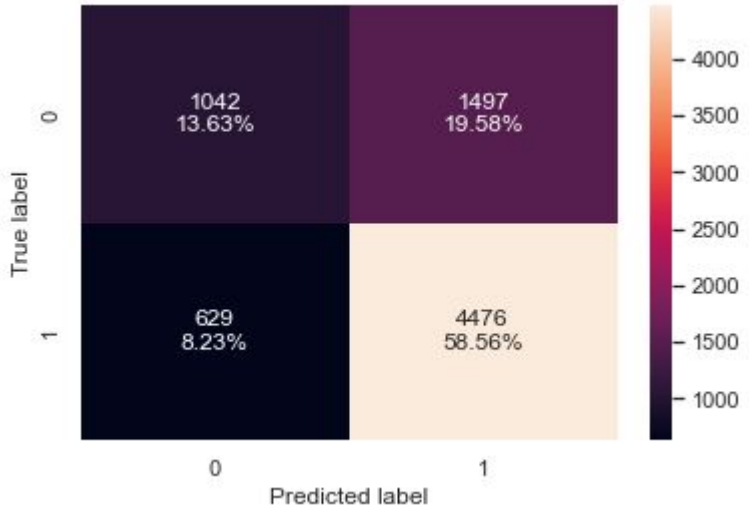
- Overfitting training data
- Test performance F1-score second lowest



Tuned Bagging Classifier

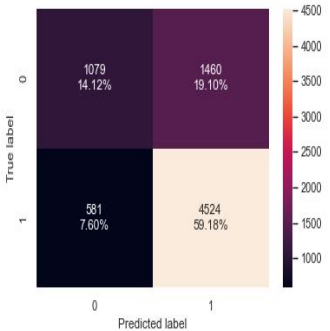
	Accuracy	Recall	Precision	F1
Training performance	0.986824	0.998489	0.982084	0.990219
Testing performance	0.721873	0.876787	0.749372	0.808088

- Overfitting not as severe but still overfitting training data
- Test performance F1-score highest so far

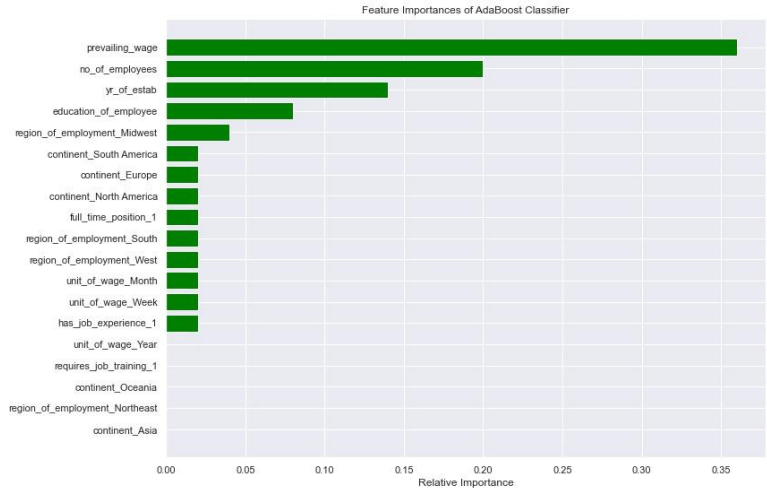


AdaBoost Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.737778	0.889868	0.759058	0.819274
Testing performance	0.732993	0.88619	0.756016	0.815944



- Not overfitting
- Test performance has highest F1-score so far

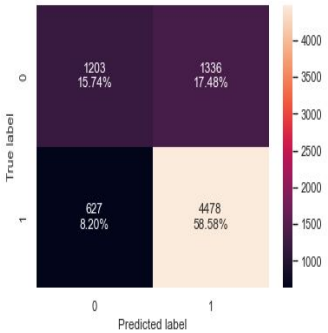


Most important features (similar to DT/RF):

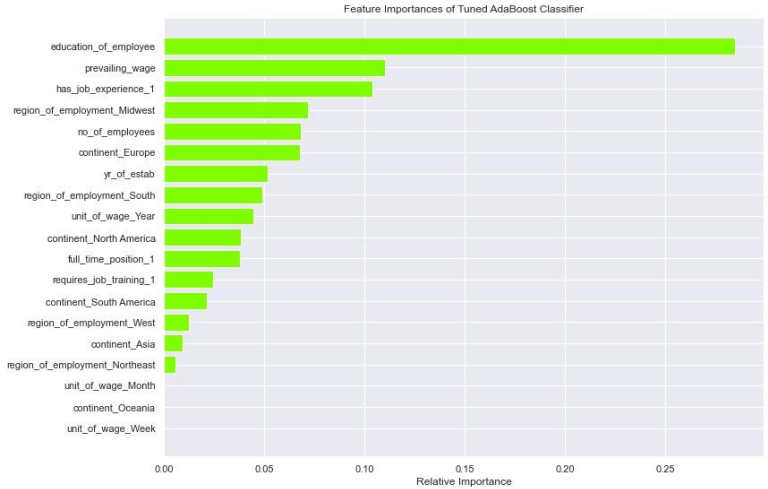
1. Prevailing wage
2. Number of employees
3. Year of establishment

Tuned AdaBoost Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.753476	0.88097	0.778908	0.826801
Testing performance	0.743197	0.877179	0.77021	0.820222



- Not overfitting training data
- Test performance F1-score higher than all others so far

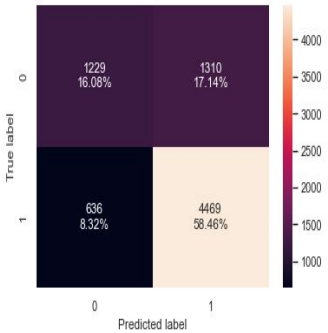


Most important features (similar to tuned DT but different order):

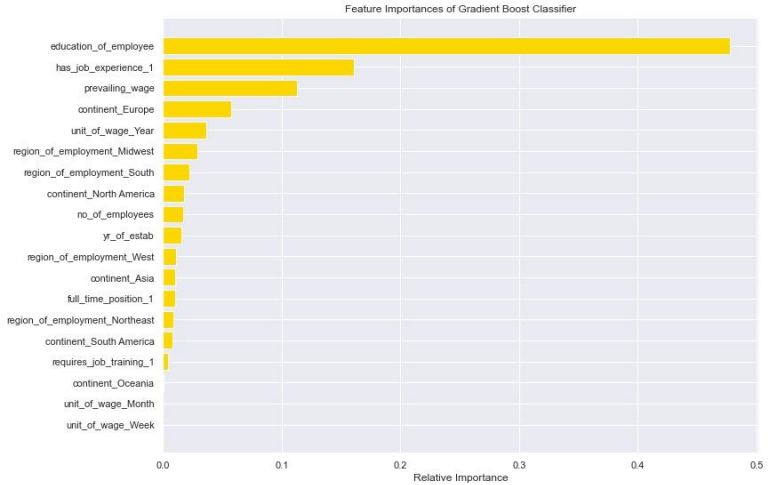
1. Education of Employee
2. Prevailing wage
3. Has Job Experience

Gradient Boosting Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.757681	0.88181	0.782845	0.829386
Testing performance	0.745421	0.875416	0.773317	0.821205



- Not overfitting training data
- Test performance F1-score higher than all others so far

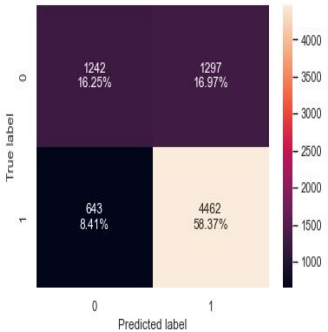


Most important features (similar to tuned DT):

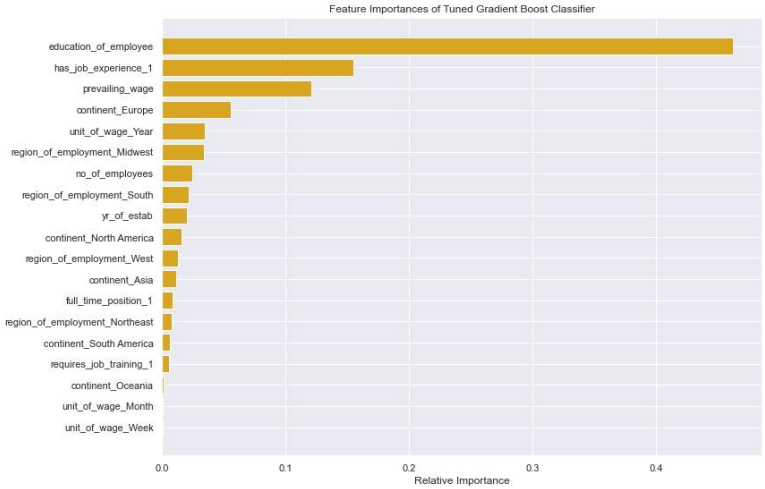
1. Education of Employee
2. Has Job Experience
3. Prevailing wage

Tuned Gradient Boosting Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.757737	0.879292	0.784174	0.829013
Testing performance	0.746206	0.874045	0.774787	0.821429



- Not overfitting training data
- **Test performance F1-score highest of all**

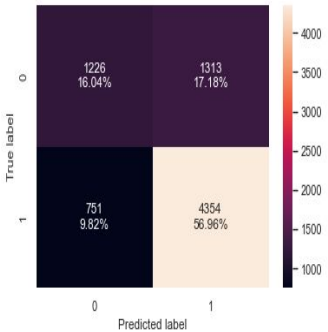


Most important features (similar to tuned DT & GB):

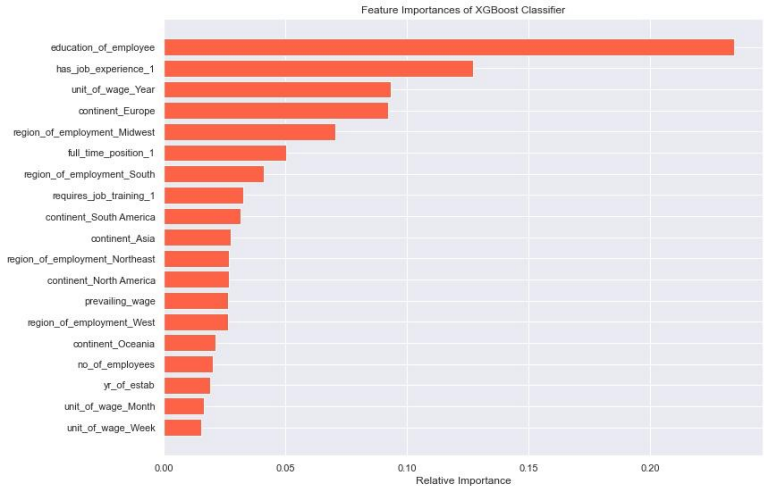
1. Education of Employee
2. Prevailing wage
3. Has Job Experience

XGBoost Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.841612	0.931503	0.846711	0.887086
Testing performance	0.729984	0.852889	0.768308	0.808392



- Overfitting training data
- Test performance F1-score good but not higher than Gradient Boost

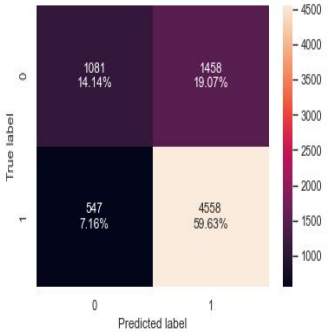


Most important features (similar to tuned DT):

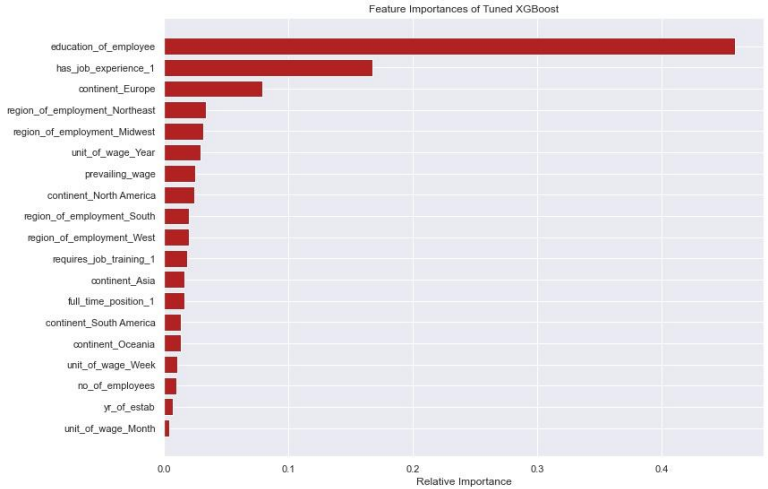
1. Education of Employee
2. Has Job Experience
3. Unit of Wage Year

Tuned XGBoost Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.756952	0.903383	0.771691	0.83236
Testing performance	0.737703	0.89285	0.757646	0.81971



- Not overfitting training data
- Test performance F1-score but but not higher than Gradient Boost



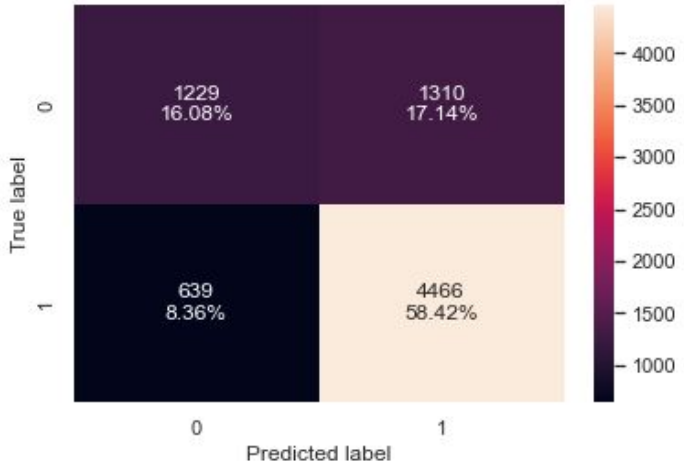
Most important features (similar to tuned DT):

1. Education of Employee
2. Has Job Experience
3. Continent Europe

Stacking Classifier

	Accuracy	Recall	Precision	F1
Training performance	0.755102	0.877025	0.782563	0.827106
Testing performance	0.745029	0.874829	0.773199	0.82088

- Not overfitting training data
- Test performance high F1-score but not better than Tuned Gradient Boost



Stacked 4 estimators

1. Tuned AdaBoost Classifier
2. Tuned Gradient Boosting Classifier
3. Tuned Decision Tree Classifier
4. Tuned XGBoost Classifier

Training Performance Comparison

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adaboost Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	0.716192	1.0	1.0	0.983180	0.986824	0.737778	0.753476	0.757681	0.757737	0.841612	0.756952	0.755102
Recall	1.0	0.778477	1.0	1.0	0.984723	0.998489	0.889868	0.880970	0.881810	0.879292	0.931503	0.903383	0.877025
Precision	1.0	0.792853	1.0	1.0	0.990041	0.982084	0.759058	0.778908	0.782845	0.784174	0.846711	0.771691	0.782563
F1	1.0	0.785599	1.0	1.0	0.987375	0.990219	0.819274	0.826801	0.829386	0.829013	0.887086	0.832360	0.827106

Test Performance Comparison

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	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	0.662480	0.714678	0.725144	0.722658	0.695317	0.721873	0.732993	0.743197	0.745421	0.746206	0.729984	0.737703	0.745029
Recall	0.743193	0.780803	0.837414	0.845054	0.769246	0.876787	0.886190	0.877179	0.875416	0.874045	0.852889	0.892850	0.874829
Precision	0.749358	0.789620	0.770826	0.764487	0.773336	0.749372	0.756016	0.770210	0.773317	0.774787	0.768308	0.757646	0.773199
F1	0.746263	0.785187	0.802742	0.802754	0.771285	0.808088	0.815944	0.820222	0.821205	0.821429	0.808392	0.819710	0.820880

Business Recommendations

Most important features of Gradient Boosting Classifier:

1. Education of Employee
2. Prevailing wage
3. Has Job Experience

- Based on our analysis, we can say Visa certification has the following features in comparison to Visa denial:
  - High employee education level
  - High prevailing wage
  - Has job experience
  - Full time position
  - Continent of Europe
  - Year Unit of Wage
  - Midwest Region of Employment
- OFLC can use model to identify applications to certify, reducing cost & increasing efficiency

Business Recommendations

Subset by or consult expert regarding

- unit of wage
- prevailing wage below poverty line

- Need to consult an expert on prevailing wage as unit of wage as some details were not clear
- It seems unlikely applicants would want a job that pays below the poverty guideline, especially with higher degrees of education.
- Companies often pay immigrants higher wages due to lack of skilled US workers available to fill positions
- However skilled immigrants sometimes fill positions below their education level or well below their earning potential because they may be force to settle for lower-skill positions
- Due to time constraints (models taking over 3 hours to build/fit), models did not complete in time to make good comparisons with subsets

Business Recommendations

Collect More Data

- Factors such as international conflict, climate change, pandemics, economic changes, etc. can influence the picture of the typical Visa applicants
- Collect data from different years and continue to check the models to see if they perform as well
- Adjustments may be necessary