

EasyVisa Case Study



Contents



Business Overview



Problem Statement



Data Overview



Exploratory Data Analysis



Model Performance



Business Recommendations

Business Overview

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.

Problem Statement

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 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 Overview

case_id: ID of each visa

applicationcontinent: 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

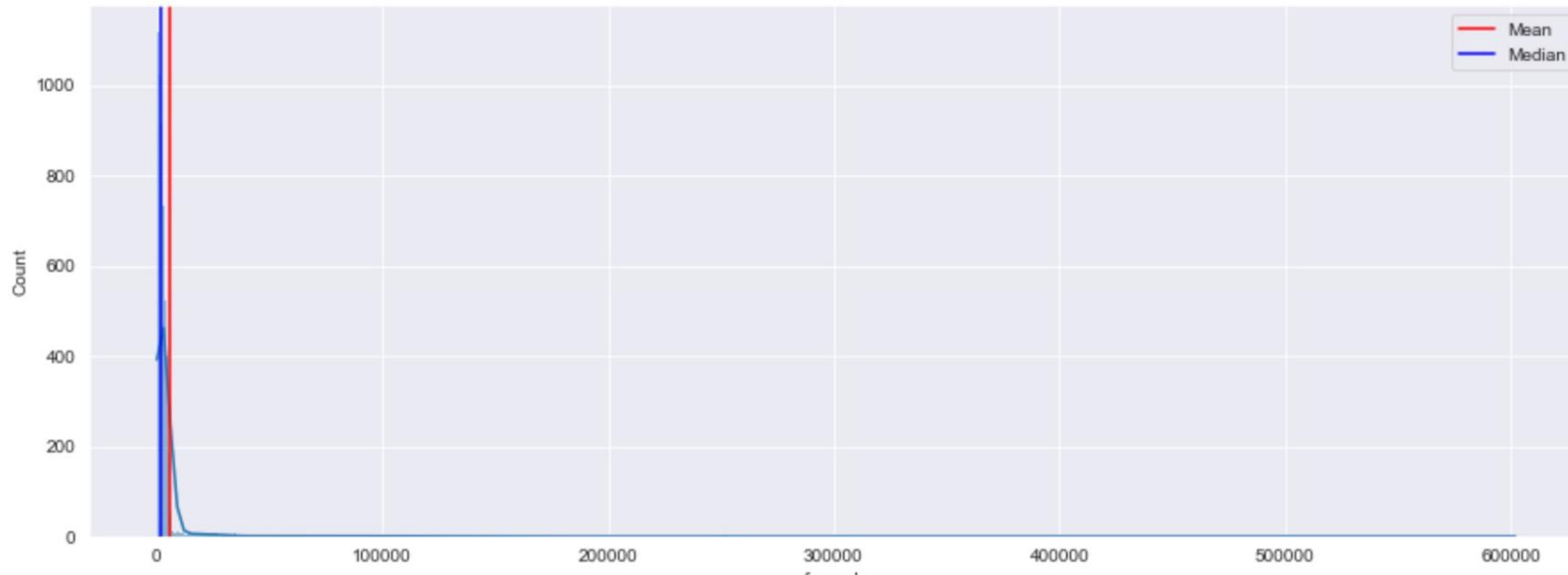
Exploratory Data Analysis

The dataset contains 25480 and 12 features

There are no duplicates in the dataset

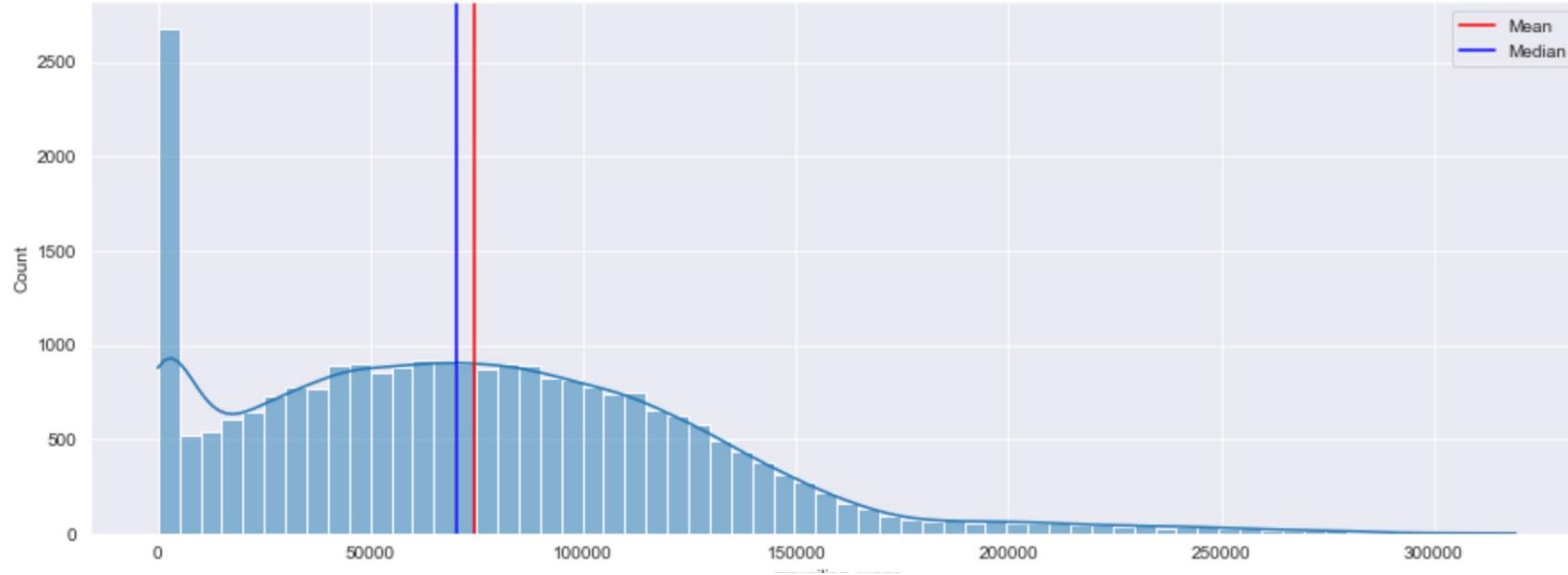
There are no missing values in the dataset

Number of employees and Prevailing wage are numeric values



- Highly skewed data
- The range goes up to 60,000 employees

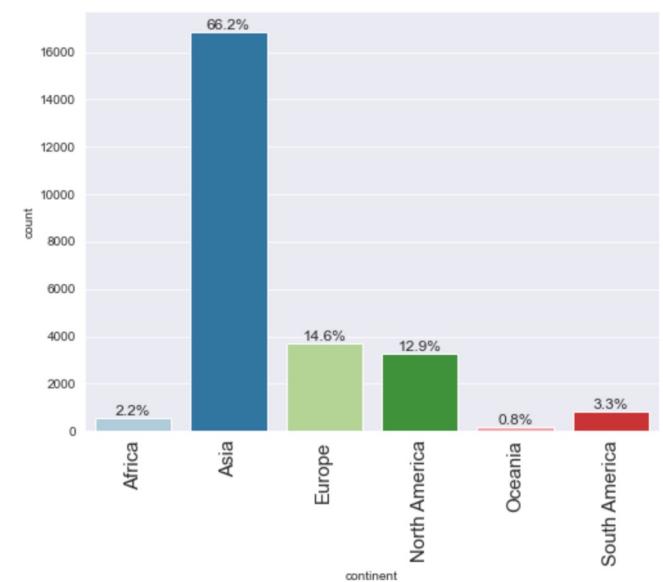
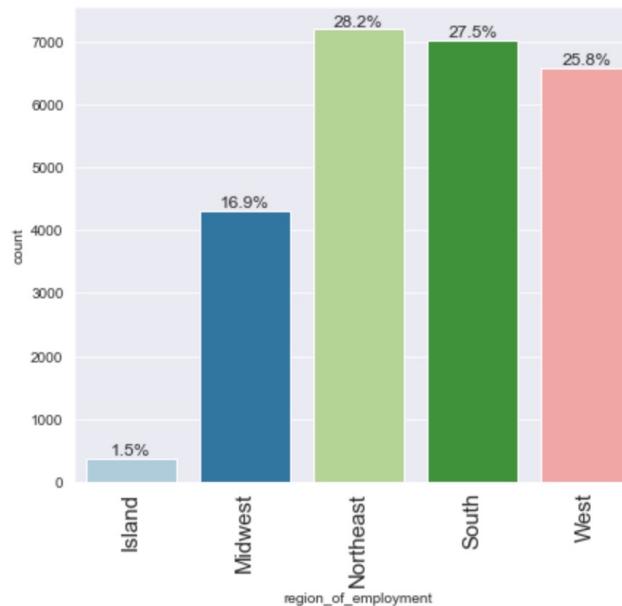
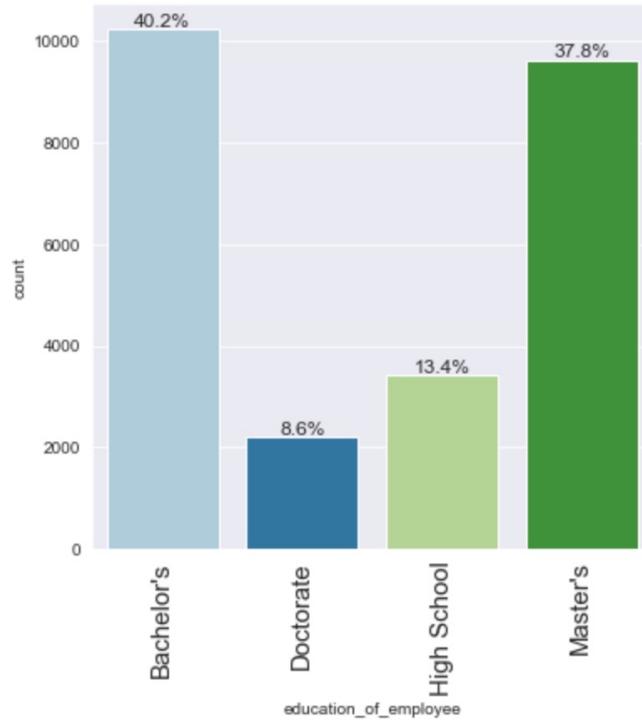
Number of Employees



- The distribution is normal
- The range is between 0 – over 300,000

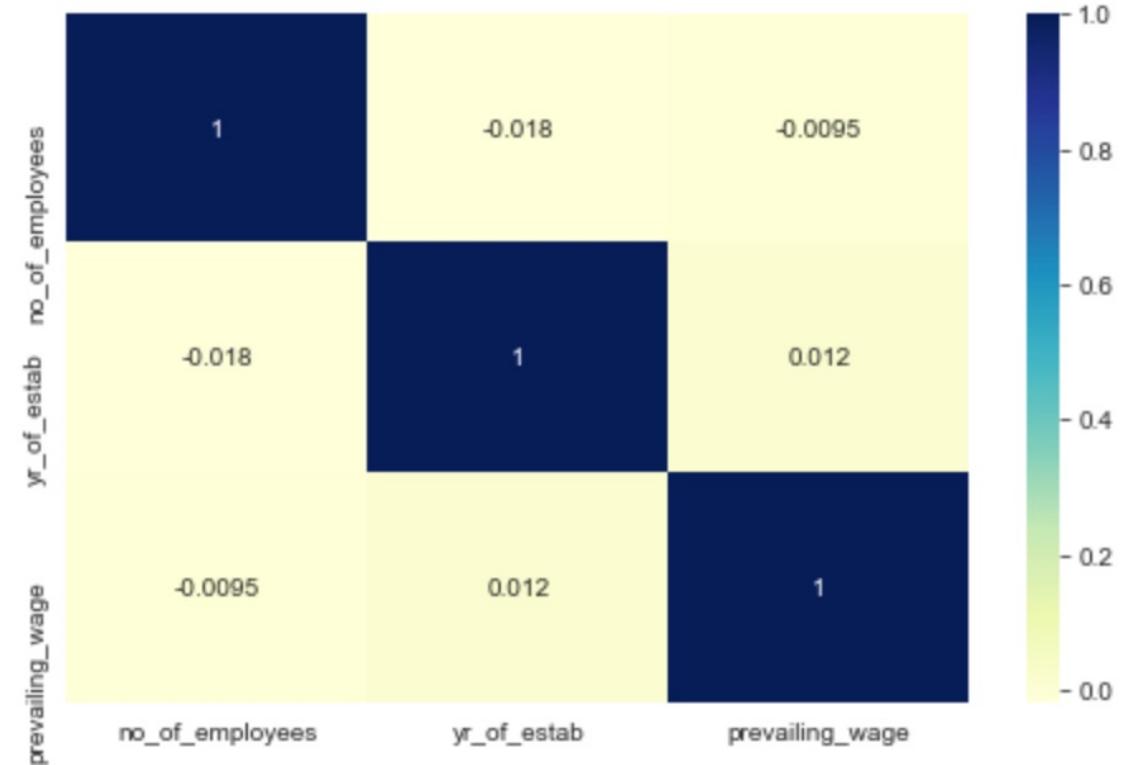
Prevailing Wages

Cases by Continent, Education, Region

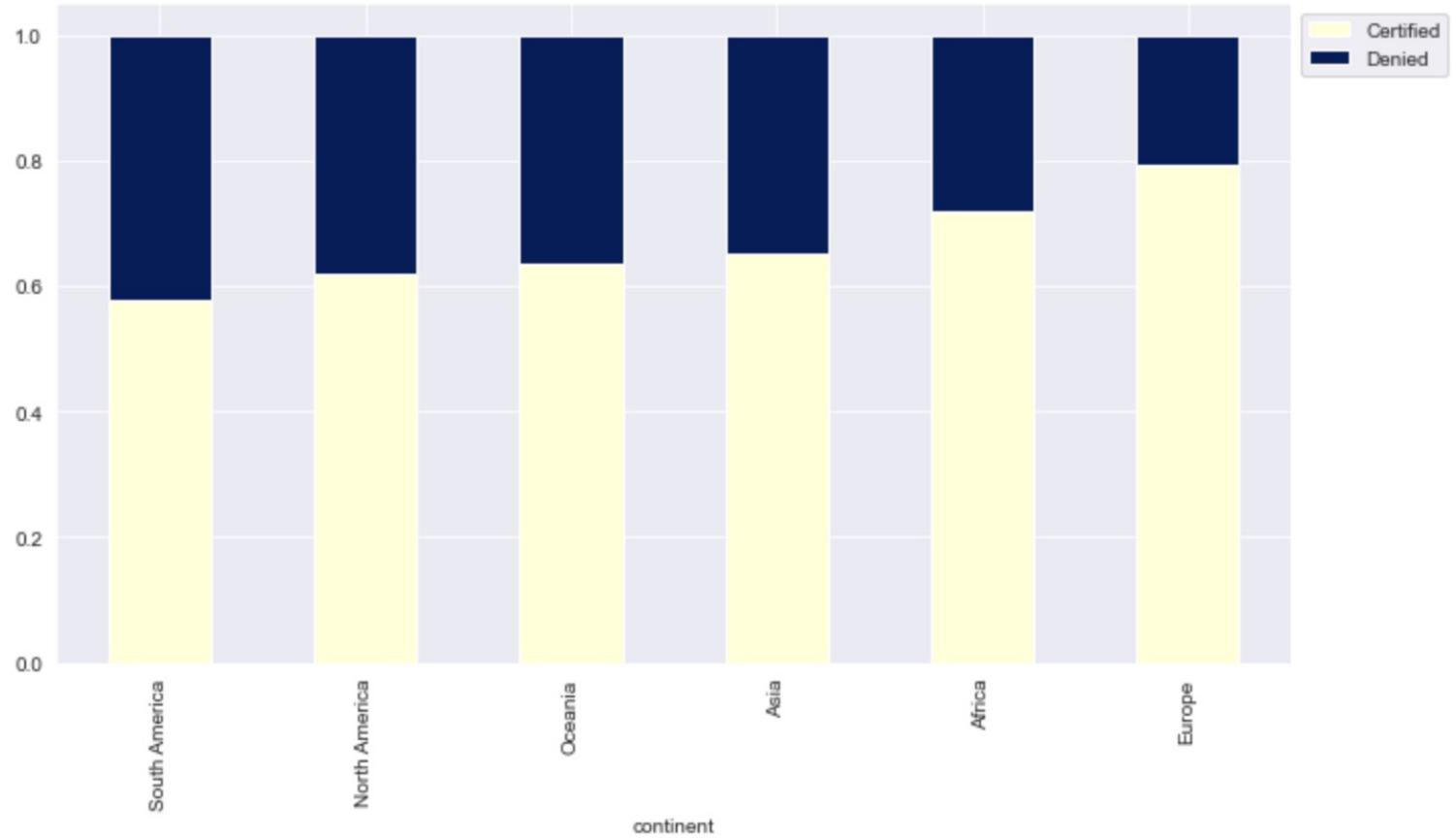


Correlation Analysis

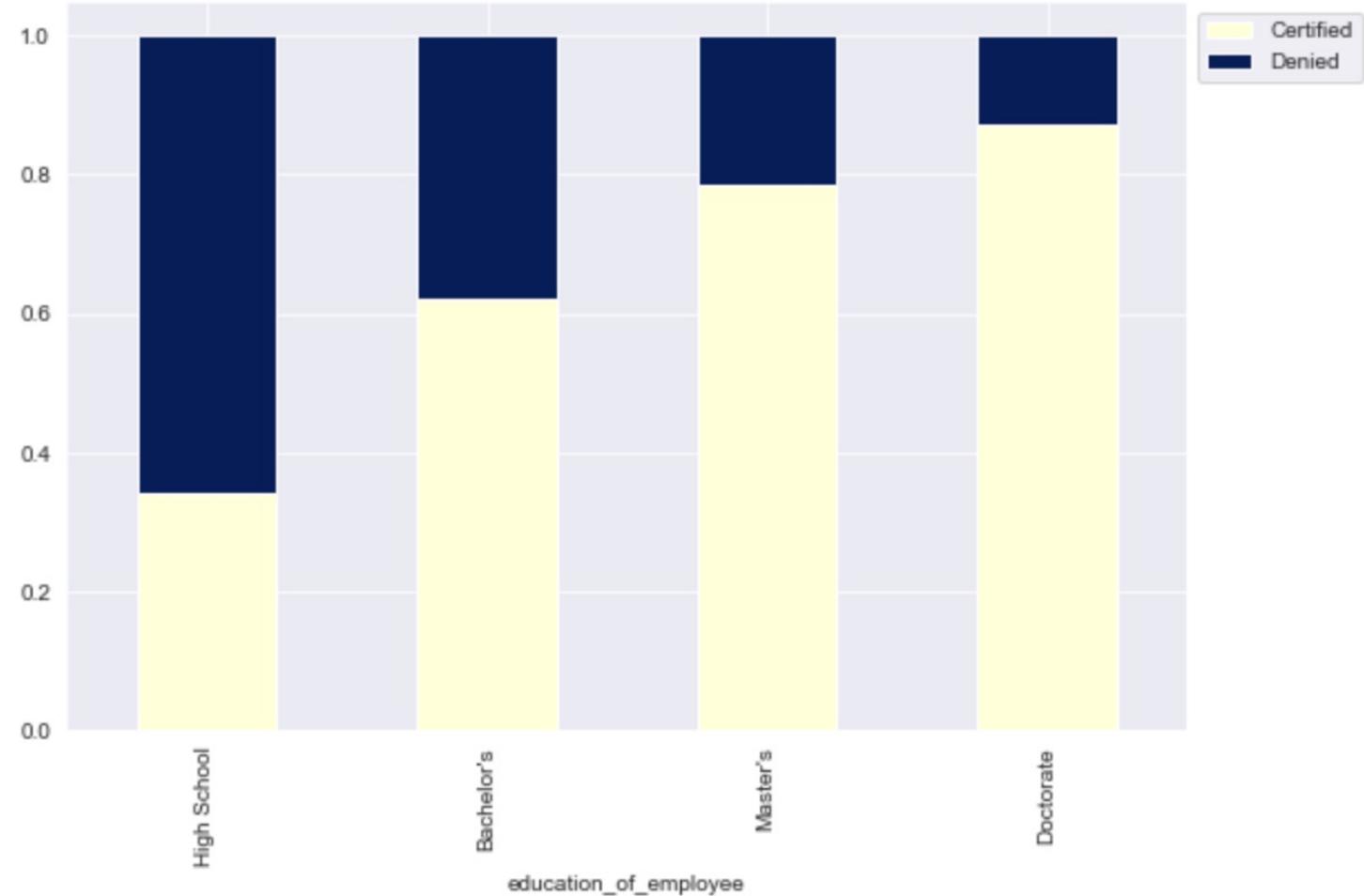
No significant correlation



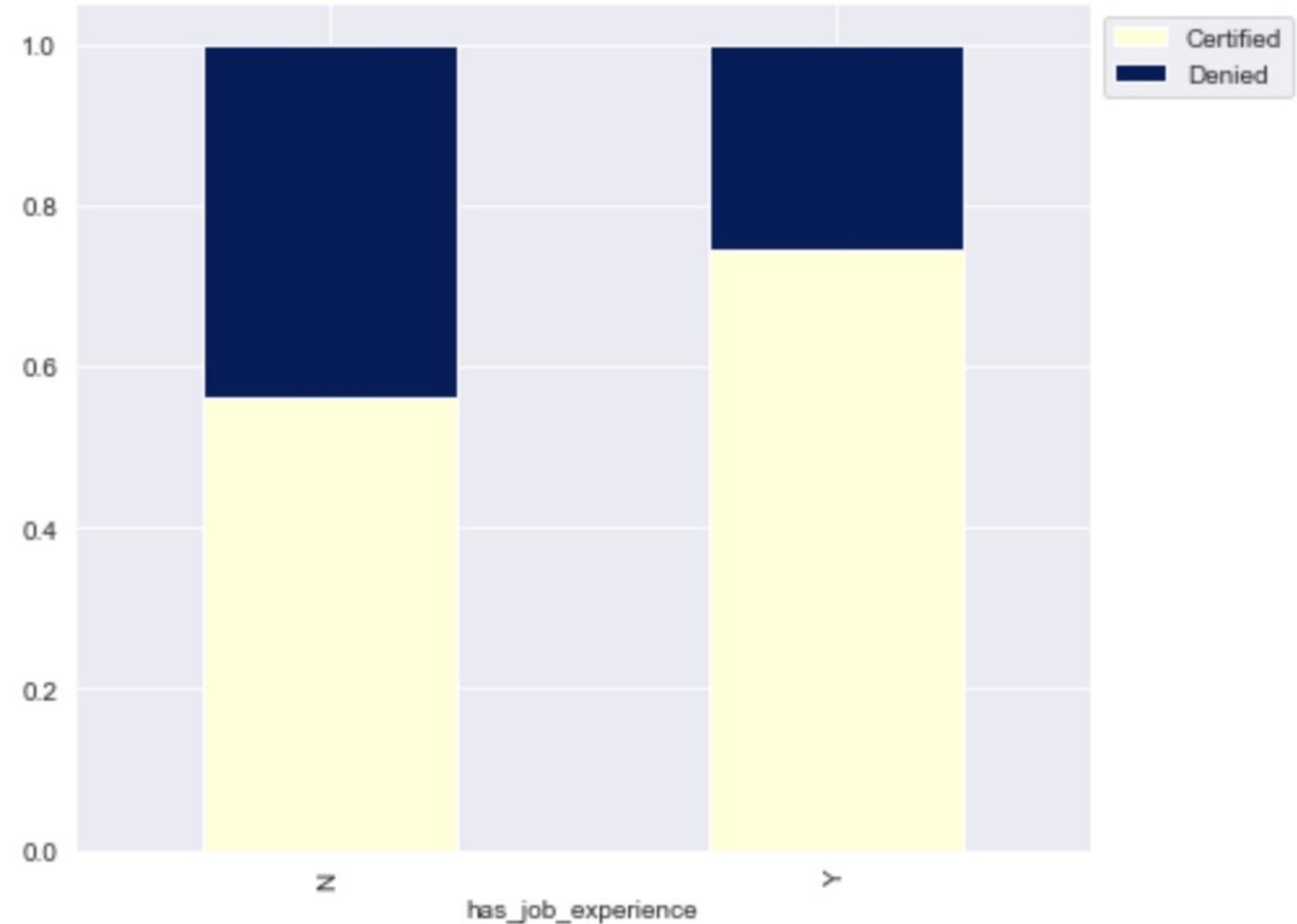
Case Result based on Continent



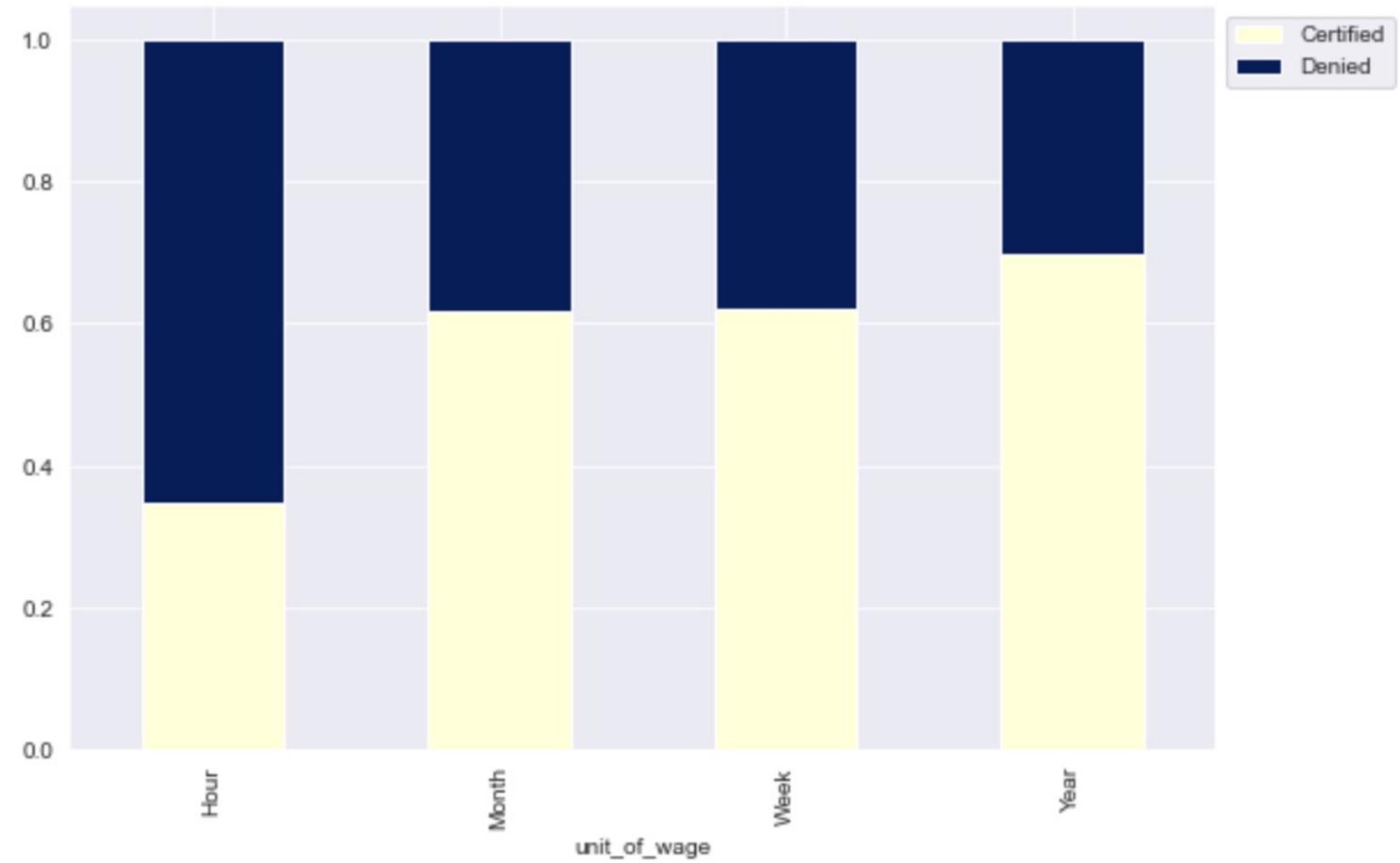
Case Result based on Education



Case Result based on Previous Job Experience



Case Result based on Unit Wage



Data Cleaning



Year of Establishment

Range between 1800-2016

Converted into value variable

Companies before year 2000 and after year 2000



Number of Employees

Some companies have negative values for number of employees

Dropped those observations



Has job experience, requires job training, full time position and case status

Converted Y/N to 1/0 values



Models Developed

- Decision Tree
- Bagging Classifier
- Random Forest Classifier
- Bagging Classifier With Weights
- Random Forest With Weights
- Tuned Decision Tree
- Tuned Bagging Classifier
- Tuned Random Forest
- AdaBoost Classifier
- Gradient Boosting Classifier
- XGBoost Classifier
- Tuned AdaBoost Classifier
- Tuned Gradient Boosting Classifier
- Tunes XGBoost Classifier

Model Performance

Training Data

Training performance comparison:

		Accuracy	Recall	Precision	F1
	Decision Tree	1.000000	1.000000	1.000000	1.000000
	Bagging Classifier	0.982540	0.984454	0.989359	0.986900
	Random Forest Classifier	0.999944	1.000000	0.999916	0.999958
	Weighted Bagging Classifier	0.984505	0.986387	0.990381	0.988380
	Weighted Random Forest Classifier	1.000000	1.000000	1.000000	1.000000
	Decision Tree Estimator	0.711599	0.932605	0.719108	0.812059
	Bagging Estimator	0.999663	0.999832	0.999664	0.999748
	Random Forest Estimator	0.792106	0.906134	0.806567	0.853457
	AdaBoost Classifier	0.737649	0.887899	0.759871	0.818911
	Gradient Boosting Classifier	0.756512	0.877227	0.784003	0.827999
	XGBoost Classifier	0.823041	0.924370	0.830063	0.874682
	AdaBoost Estimator	0.748653	0.879832	0.774580	0.823858
	Gradient Boost Classifier - Init AdaBoost	0.756456	0.876303	0.784414	0.827816
	Gradient Boost Estimator	0.755333	0.874874	0.783961	0.826926
	XGBoost Estimator	0.760330	0.884706	0.784209	0.831431

Model Performance

Testing Data

Testing performance comparison:

		Accuracy	Recall	Precision	F1
	Decision Tree	0.663261	0.754166	0.744965	0.749537
	Bagging Classifier	0.696660	0.777691	0.770441	0.774049
	Random Forest Classifier	0.713687	0.818467	0.768169	0.792521
	Weighted Bagging Classifier	0.703733	0.789453	0.772196	0.780729
	Weighted Random Forest Classifier	0.713294	0.820231	0.766862	0.792649
	Decision Tree Estimator	0.709103	0.929034	0.718248	0.810155
	Bagging Estimator	0.731238	0.872770	0.760376	0.812705
	Random Forest Estimator	0.749050	0.871594	0.779043	0.822724
	AdaBoost Classifier	0.734512	0.877671	0.761395	0.815408
	Gradient Boosting Classifier	0.748134	0.866693	0.780544	0.821366
	XGBoost Classifier	0.735691	0.857675	0.772013	0.812593
	AdaBoost Estimator	0.744990	0.870025	0.775603	0.820105
	Gradient Boost Classifier - Init AdaBoost	0.748527	0.865713	0.781455	0.821429
	Gradient Boost Estimator	0.747479	0.863948	0.781244	0.820518
	XGBoost Estimator	0.749574	0.874338	0.778224	0.823486

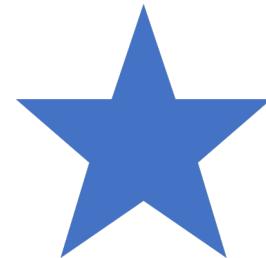
Model Selection



The Decision Tree Estimator
has highest recall score



The Boosting Estimators have
similar scores



Overall, the XGBoost Estimator
has the best performance

Important Features



Education of Employee

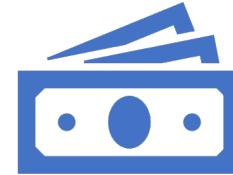
The employees with high school education have very high chances of denial

The employees with master's and doctorate education have high chances of getting certified



Previous Job Experience

Having previous job experience improves chances of getting certified



Prevailing Wage

A fair prevailing wage is important factor for getting certified

Recommendations

- EasyVisa company can use the predictive model to:
 - Identify the cases that can be certified
 - Identify the cases that can be denied
- The factors that affect the case denial are - Education and Previous Job Experience
- The higher the education of the employee, higher are chances of getting certified. The degree of Doctorate and Masters are highest ranking degrees
- The application for the employees with high school education need to be based on the skills they bring to the company
- Prevailing wage is the next important factor. The cases which offer fair wage based on the skill set
- EasyVisa can add the prevailing wage data for specific region and job role to this dataset. This can be used to match with each case application. This help determine case status more accurately
- The company related information like number of employees and year established have little to none impact on the result
- The continent of the employee holds more importance. Employee from Europe has highest chances to get certified.
- The region of the company has no significant impact on the result
- The company can gather more information on the employee's projects for further analysis and better prediction.
- Previous job experience is vital for the case judgement. The company can gather additional data surrounding previous job. For example - number of years, industry, technical skills, etc.