Easy Visa

Ensembled Techniques

02/03/2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- The Office (OFLC) should be more precise with the education of employee, unit of wage, job experience and continent of employee as these are the most important factors in determining the work visa eligibility.

- Most of the ensembled methods have a generalized results without overfitting the training data. The XGBoost Classifier is giving a more generalized performance as compared to the other models. There fore the company can use it to identify which employee is eligible for the work visa. This would help to reduce the time taken to evaluate each applicants and increase the efficiency of the process.

- The model built can be used to predict the work visa eligibility of the applicant and can correctly identify 82% of the eligible applicants.

Executive Summary

- As more than 60% of the high school graduates visa request are denied. The office should brief all the employers to comply with the parameters to accept high school graduates.
- As more than 70% of the applicants with job experience are approved their request to work in USA. The employers should focus on experienced employees.
- More than 70% of the employees with yearly unit of wage are certified.
- As the probability of approving visa for hourly unit of wage is very minimum and more than 70% of the applicants with yearly unit of wage are certified, The Office should advise the employers to avoid hiring foreign employees with hourly unit of wage.

Business Problem Overview and Solution Approach

- **Business problem overview**
- OFLC(Office of Foreign Labor Certification) processes job certification applications for employers seeking to bring foreign workers into the united states. 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 based on the drivers that significantly influence the case status.

- **Solution Approach/ methodology**

- Overview of the data
- Review the patterns and trends of the data using EDA
- Data Pre processing
- Model building – Decision Tree, Bagging, Boosting and Stacking
- Model performance Evaluation and improvement – Decision Tree, Bagging, Boosting and Stacking

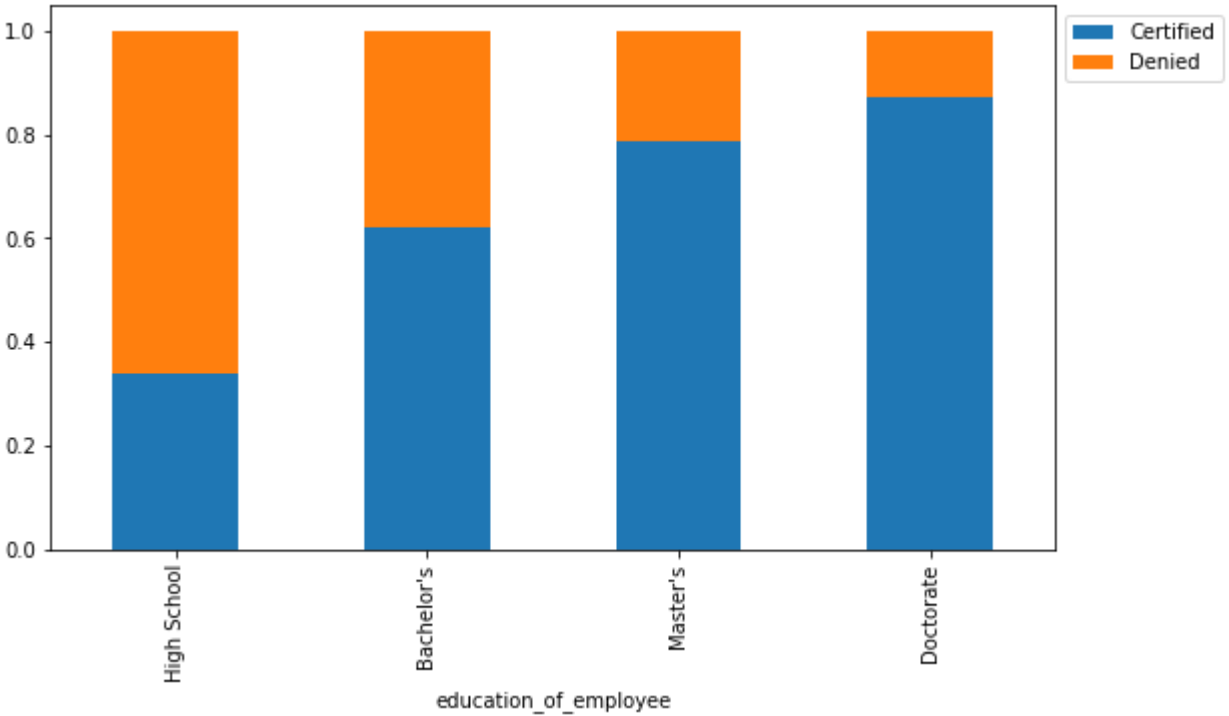
EDA Results

- The data set has 25480 observations and 12 Attributes.
- There are 3 numeric (int and float) and 9 object type variables in our data.
- 78% of the employees has a bachelors and masters of degree holders.
- Most of the applicants has a job experience and doesn't require job training. 88.4% of the employees doesn't require job training in US. they can join the workforce immediately.
- There is a huge difference between mean(5667) and median(2109) of number of employees due to outliers.
- Most of the employers are applied for a full time employment of the foreign employees.
- Asia is the biggest market for the employers, 66% of the employees are from Asian Continent.

Link to Appendix slide on data background check

EDA Results - Continue

- More than 80% of the Masters and doctorate holders are certified.
- More than 60% of the high school graduates are denied.



EDA Results - Continue

- 90% of the applicants unit of wage is yearly.
- 80% of the visa requests from Europe continents are approved.
- 40% of South American applicants are rejected their visa requests.
- More than 70% of the applicants with job experience are approved their request to work in USA. But more than 40% of the applicants without job experience are rejected.
- More than 90% of the employees who have job experience are not requires job training.
- The prevailing wage is higher on midwest and island regions.
- More than 70% of the employees with yearly unit of wage are certified.
- The probability of approving visa for hourly unit of wage is very minimum.

Data Preprocessing

- There is no duplicate value in the data.
- There is no missing value in the data.
- There are outliers in our data. However, we will not treat them as they are proper values.
- Before we proceed to build the model. We encode the categorical features.
- We Split the data into train and test in the ratio 70:30 to be able to evaluate the model that we build on the train data.

Model Building, Hyperparameter tuning and Performance Summary

Overview of ML model and its parameters

- We are building 7 ensembled models. Decision Tree, Bagging Classifier, Random Forest Clasifier, Adaboost Classifier, Gradient Boosting Classifier, XGBoost Classifier and Stacking classifier.
- First, We are building these models with default parameters and then use hyperparameter tuning to optimize the model performance by using GridSerachCV.
- We will calculate all Four metrics - Accuracy, Precision, Recall and F1 Score but the metric of interest here is F1 Score as both the cases are important.
- The greater the F1-Score higher the chances of predicting both the classes correctly.
- Except Decision tree, Baggin classifier, Tuned bagging classifier and Random forest classifiers the rest models are performing well on both training and test data without the problem of overfitting.

Overview of ML model and its parameters

- The Stacking Classifier is giving the highest f1-score on the test data but is slightly overfit the training data.
- The XGBoost Classifier has given the second-highest test f1-score and is giving a more generalized performance as compared to the stacking classifier. There fore it can be selected as the final model.
- Education of employee(Doctorate, High school) are the most important features in identifying the visa eligibility of employees followed by yearly unit of wage and job experience.

Model Building - Bagging

- The decision tree with the default parameters is overfitting the training data with F1 score of 1.0 on training and 0,75 on test data.
- The Bagging Classifier is overfitting the training data with F1 score of 0.99 on training and 0.77 on test data.
- The Random Forest Classifier is overfitting the training data with F1 score of 1.0 on training and 0.8 on test data.
- Let us use hyperparameter tuning to optimize the model performance.

Model Improvement - Bagging

- After improving the decision tree by hyperparameter tuning the parameters (class_weight='balanced', max_depth=5, max_leaf_nodes=2, min_impurity_decrease=0.0001, min_samples_leaf=3, random_state=1). Tuned decision tree is giving a generalized result as the F1 scores on both train and test data are coming to 0.81.
- Even if we improve the bagging classifier by hyperparameter tuning, there is no change on generalizing the results. The F1 score for both training and test data are increased.
- After improving the random forest by hyperparameter tuning the parameters (max_depth=10, max_features='sqrt', min_samples_split=7, n_estimators=20, oob_score=True, random_state=1) . The overfitting is reduced significantly and the model performance is improved.
- Tuned Random forest is giving a slightly higher test f1-score than tuned decision tree and tuned bagging models.

Model Building – Boosting and Stacking

- The Adaboost Classifier with the default parameters is giving a generalized result with F1 score of 0.819 on training and 0.816 on test data.
- The Gradient Boosting Classifier with the default parameters is giving a generalized result with F1 score of 0.83 on training and 0.82 on test data.
- The XGBoost Classifier with the default parameters is giving a generalized result as compared to the rest of models with F1 score of 0.828 on training and 0.821 on test data.
- We are building the stacking model with Adaboost, Gradient boosting, and Random forest, then use tuned XGBoost to get the final prediction.
- The Stacking Classifier with the default parameters is giving the highest F1-score on the test data, but is slightly overfit the training data.

Model Improvement - Boosting

- The F1 score for both training and test data of tuned Adaboost classifier is decreased.
- After hyperparameter tuning of the gradient boosting, the model is slightly overfit the training data.
- The tuned XGBoost classifier model is slightly overfit the training data.

Model Performance Summary

- Training Performance comparison

	DT	Tuned DT	BC	Tuned BC	RFC	Tuned RFC	ABC	Tuned ABC	GBC	Tuned GBC	XGBC	Tuned XGBC	SC
Accuracy	1.0	0.713	0.985	0.996	1.000	0.769	0.738	0.719	0.759	0.764	0.756	0.757	0.768
Recall	1.0	0.932	0.986	1.000	1.000	0.919	0.887	0.781	0.884	0.883	0.884	0.883	0.893
Precision	1.0	0.720	0.992	0.994	1.000	0.777	0.761	0.795	0.783	0.789	0.781	0.781	0.788
F1	1.0	0.812	0.989	0.997	1.000	0.842	0.819	0.788	0.830	0.833	0.8288	0.8289	0.837

- Testing Performance Comparison

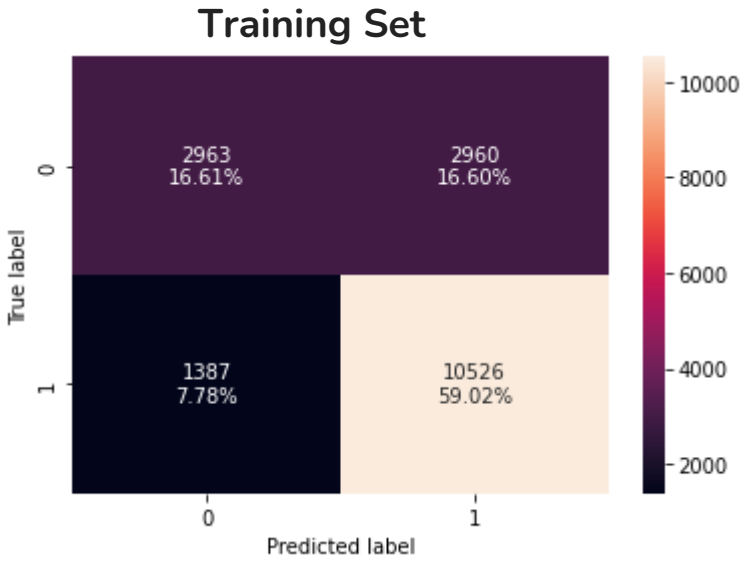
	DT	Tuned DT	BC	Tuned BC	RFC	Tuned RFC	ABC	Tuned ABC	GBC	Tuned GBC	XGBC	Tuned XGBC	SC
Accuracy	0.665	0.707	0.692	0.724	0.721	0.738	0.734	0.717	0.745	0.743	0.745	0.744	0.745
Recall	0.743	0.931	0.764	0.895	0.832	0.899	0.885	0.781	0.876	0.871	0.877	0.876	0.880
Precision	0.752	0.715	0.772	0.744	0.769	0.755	0.758	0.791	0.772	0.773	0.772	0.772	0.771
F1	0.747	0.809	0.768	0.813	0.799	0.821	0.816	0.786	0.821	0.819	0.8211	0.8207	0.821

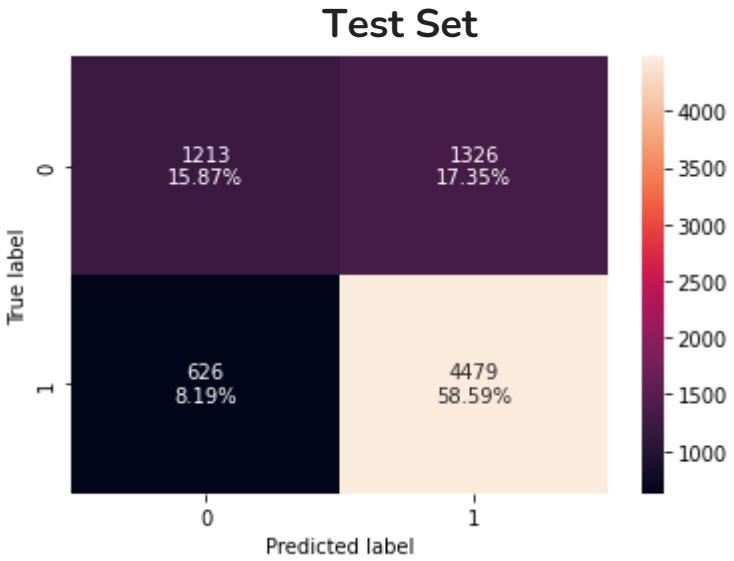
APPENDIX

Data Background and Contents

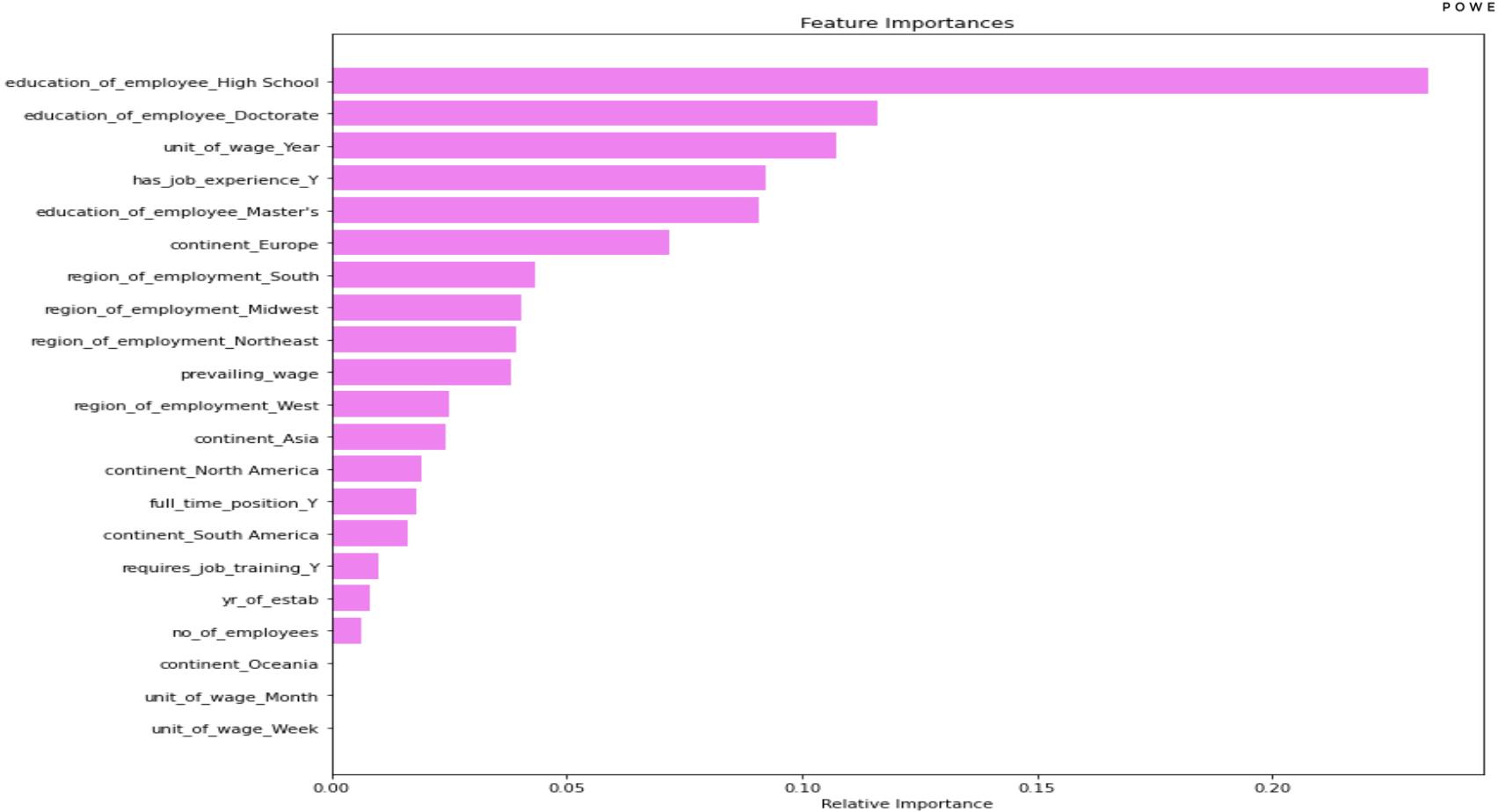
- Confusion matrix : Slide 20
- Feature importance : Slide 21
- EDA graphs and charts : Slide 22 - 40

Confusion Matrix for XGBoost Classifier model

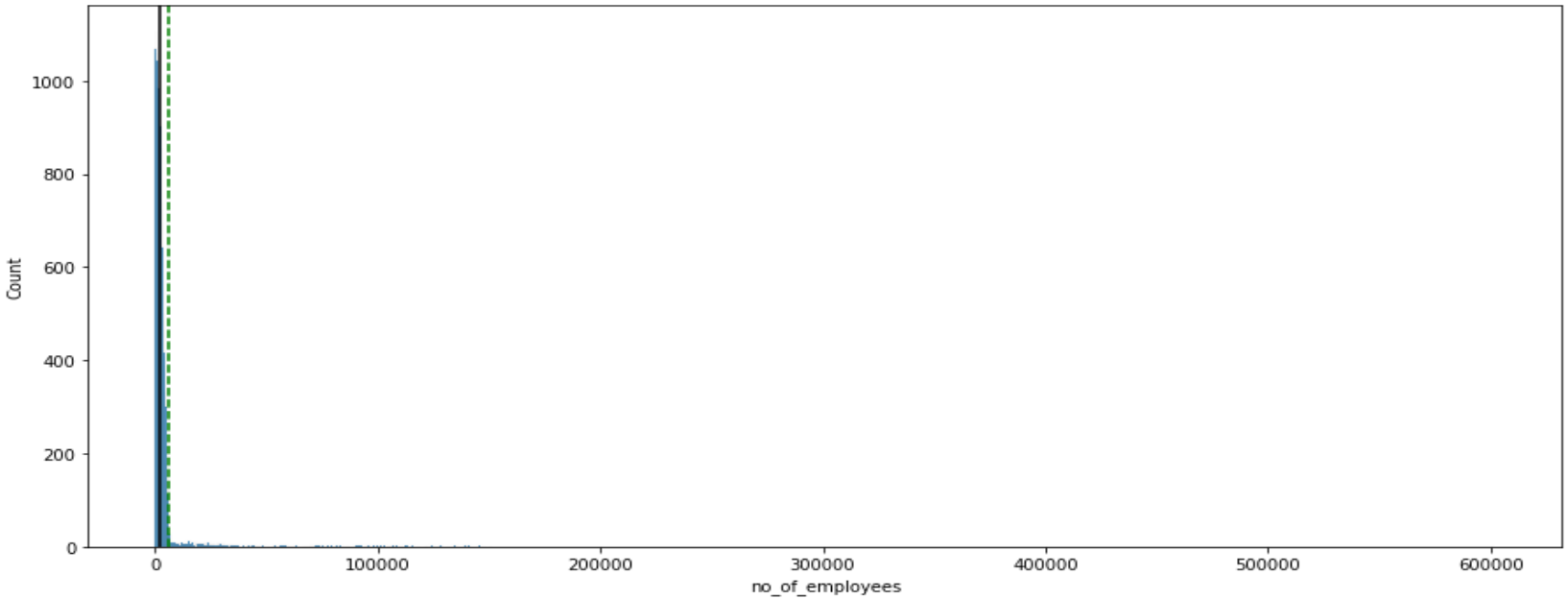




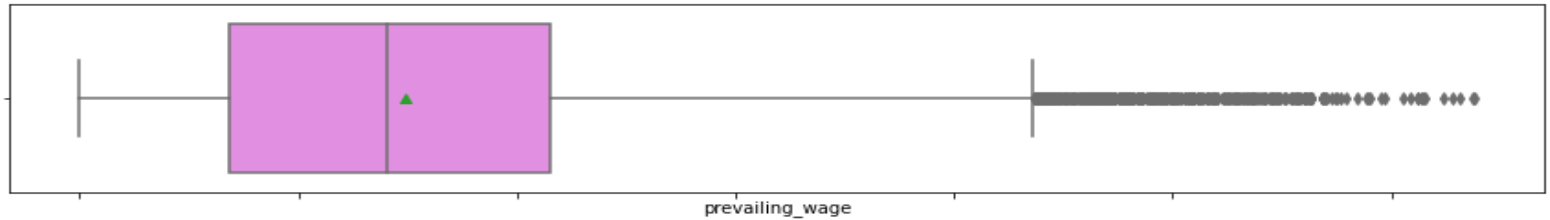
Feature importance of the final model

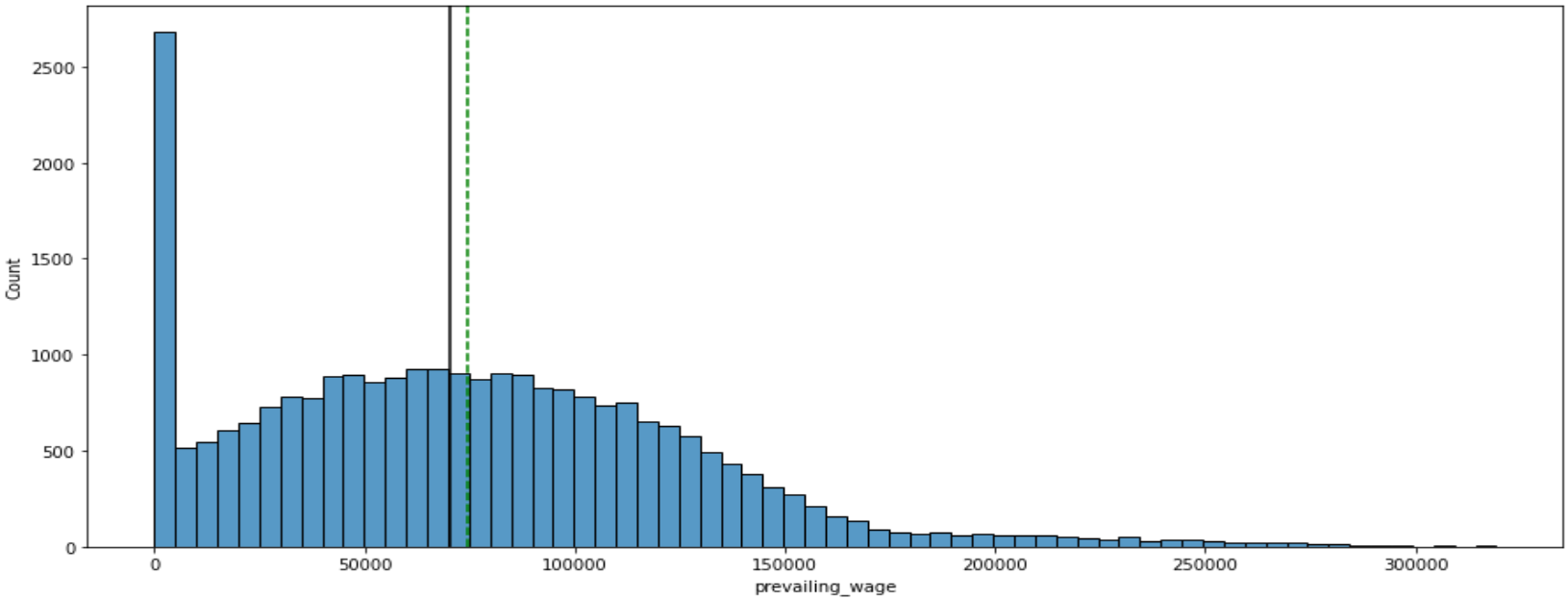


Observation on number of employees

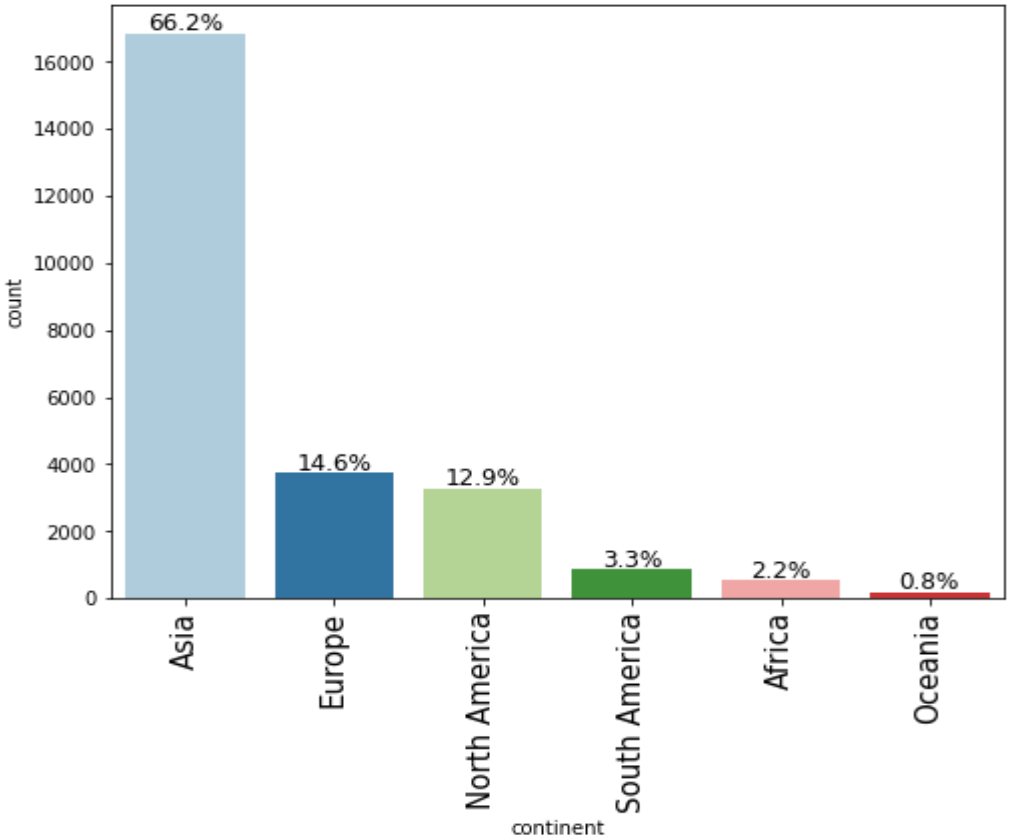


Observation on prevailing wage

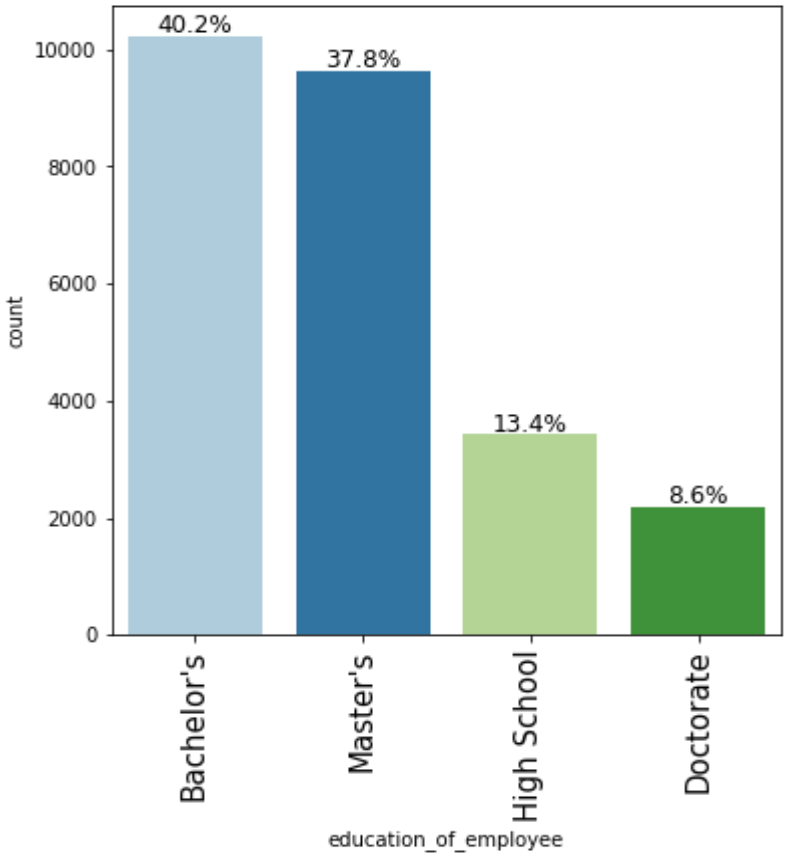




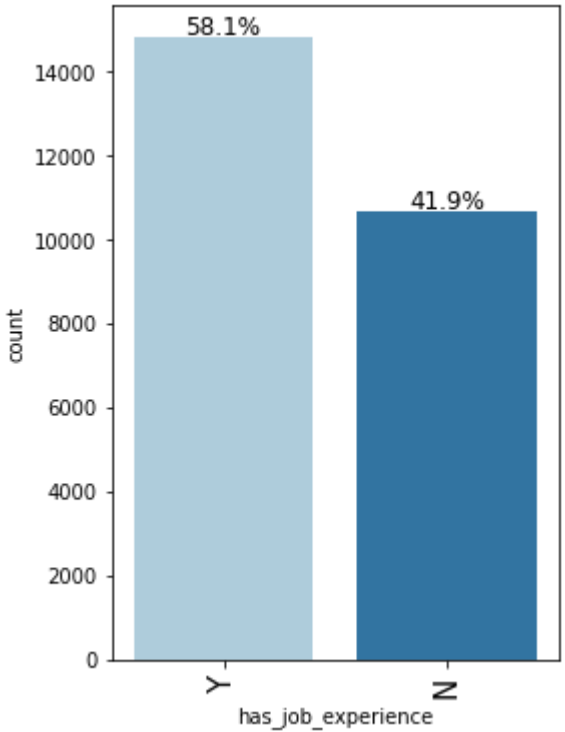
Observation on continent



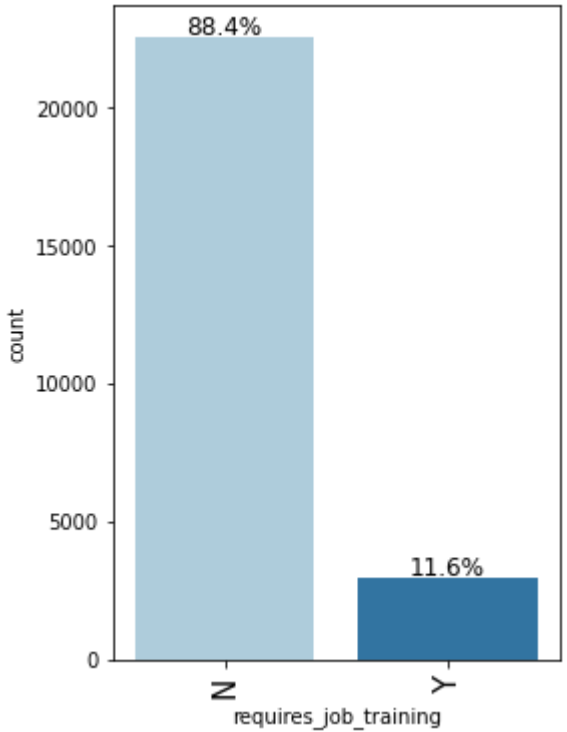
Observation on education of employee



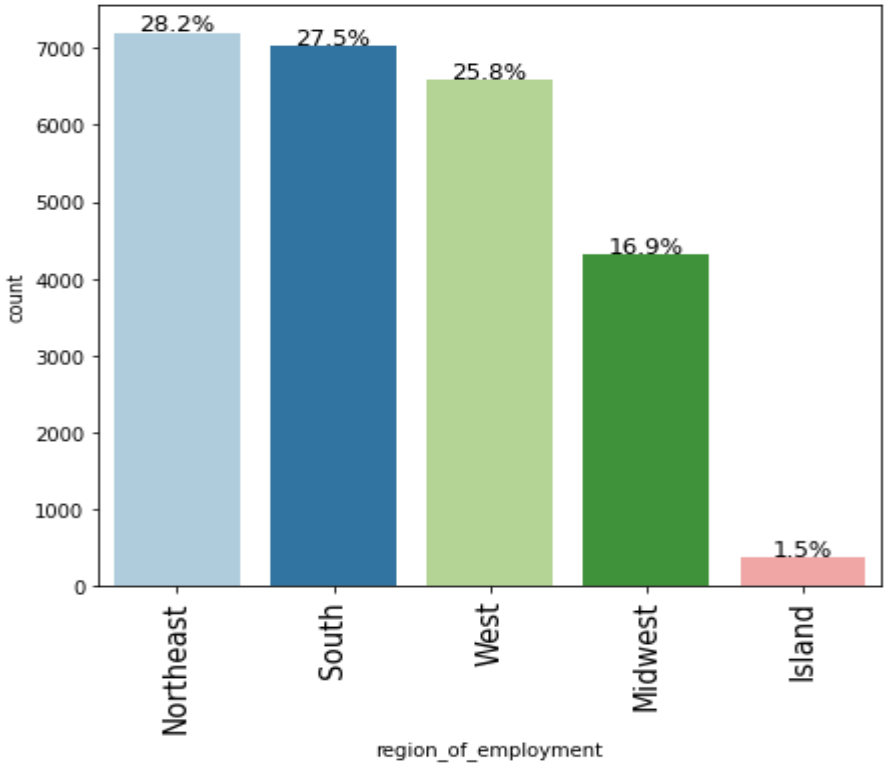
Observation on job experience



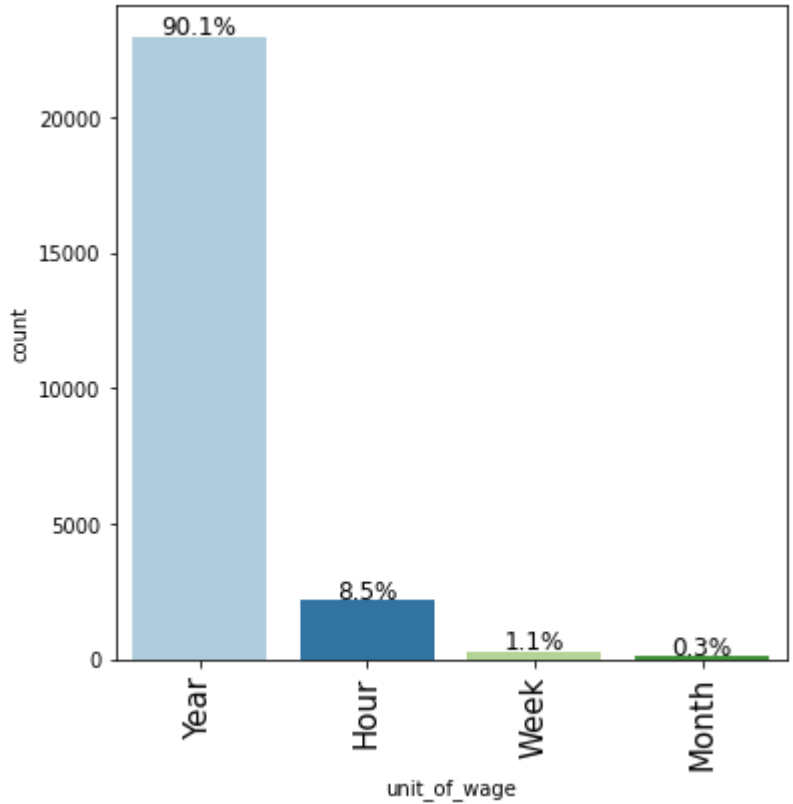
Observation on job training



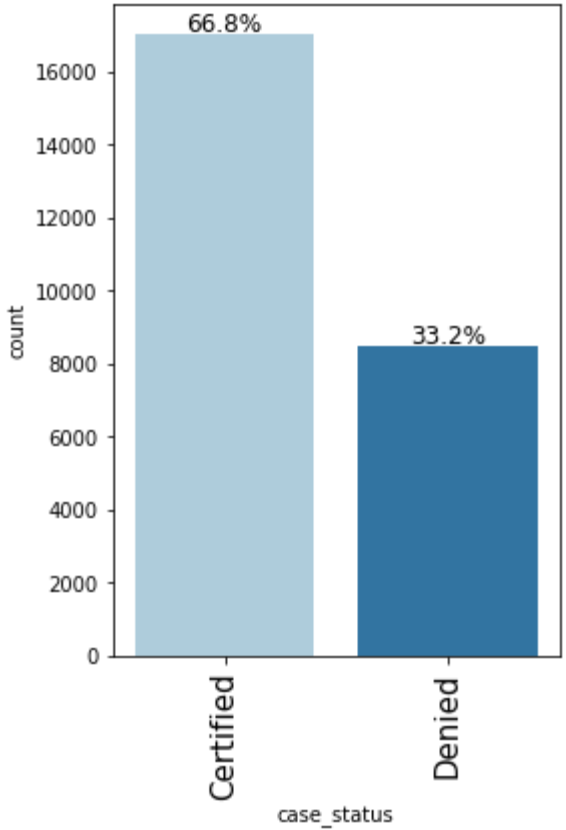
Observation on region of employment



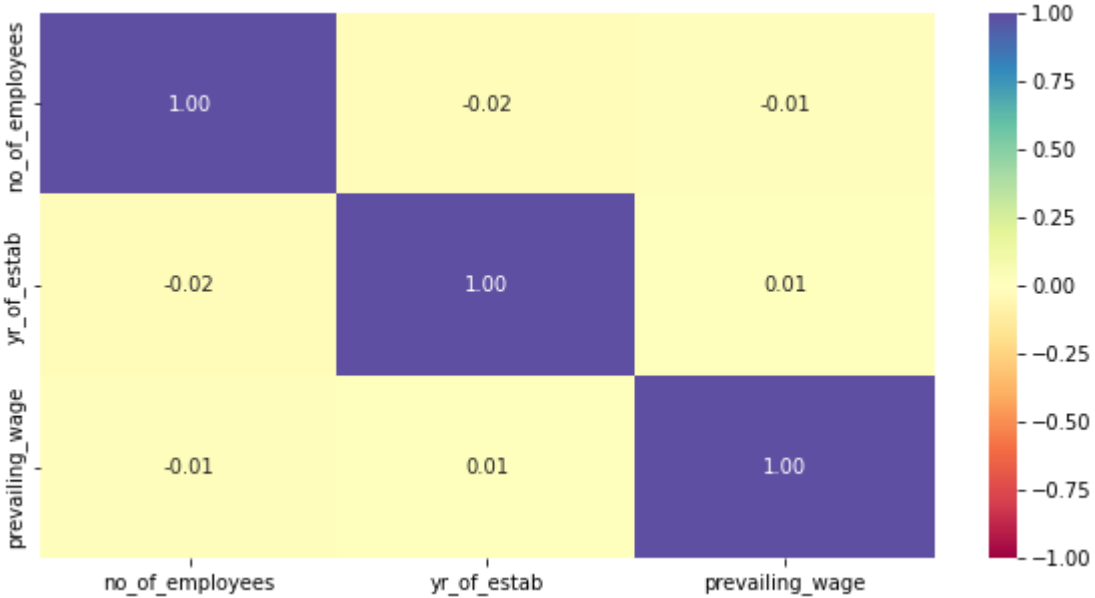
Observation on unit of wage



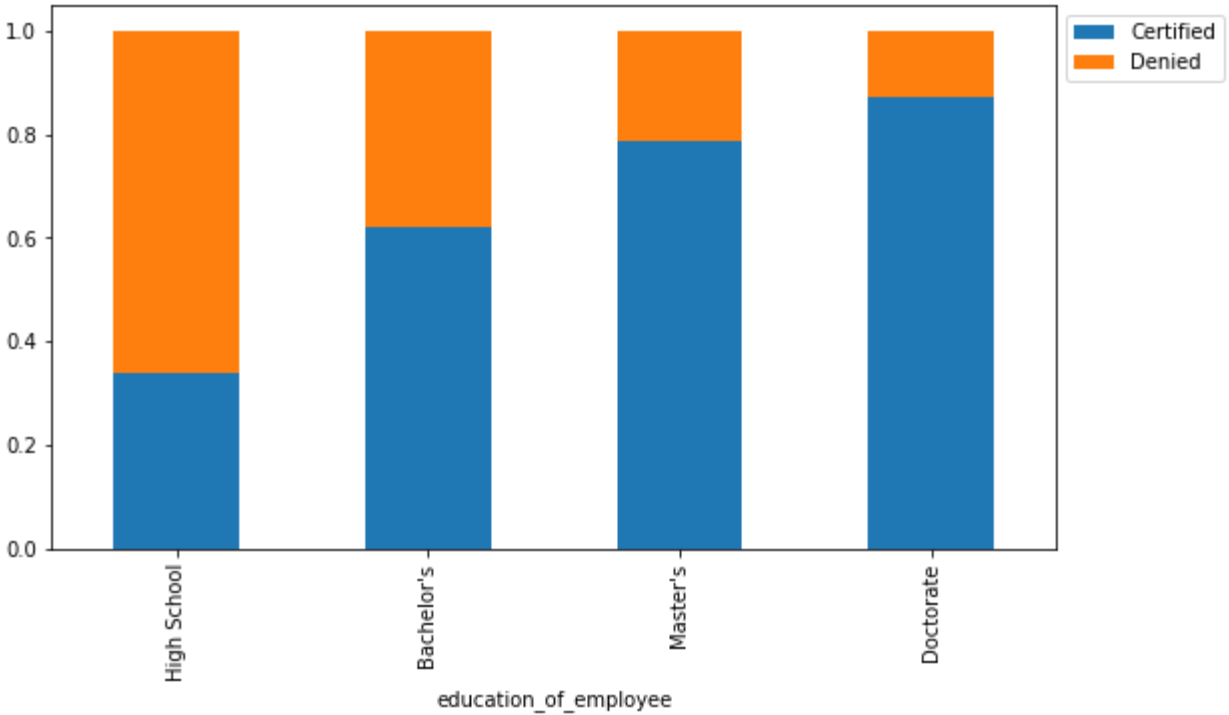
Observation on case status



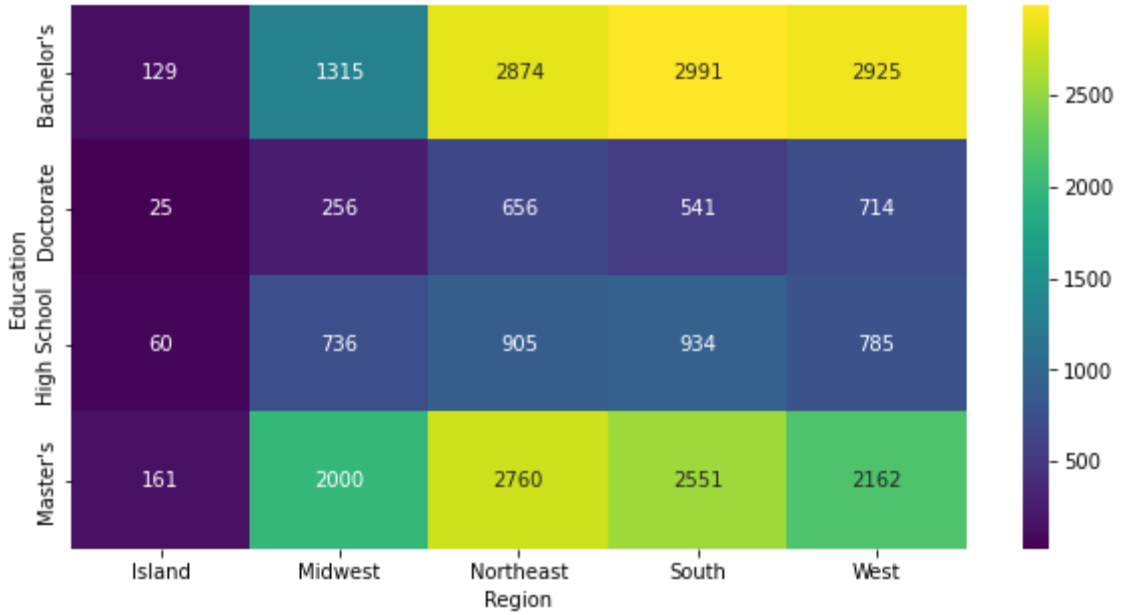
Correlation matrix



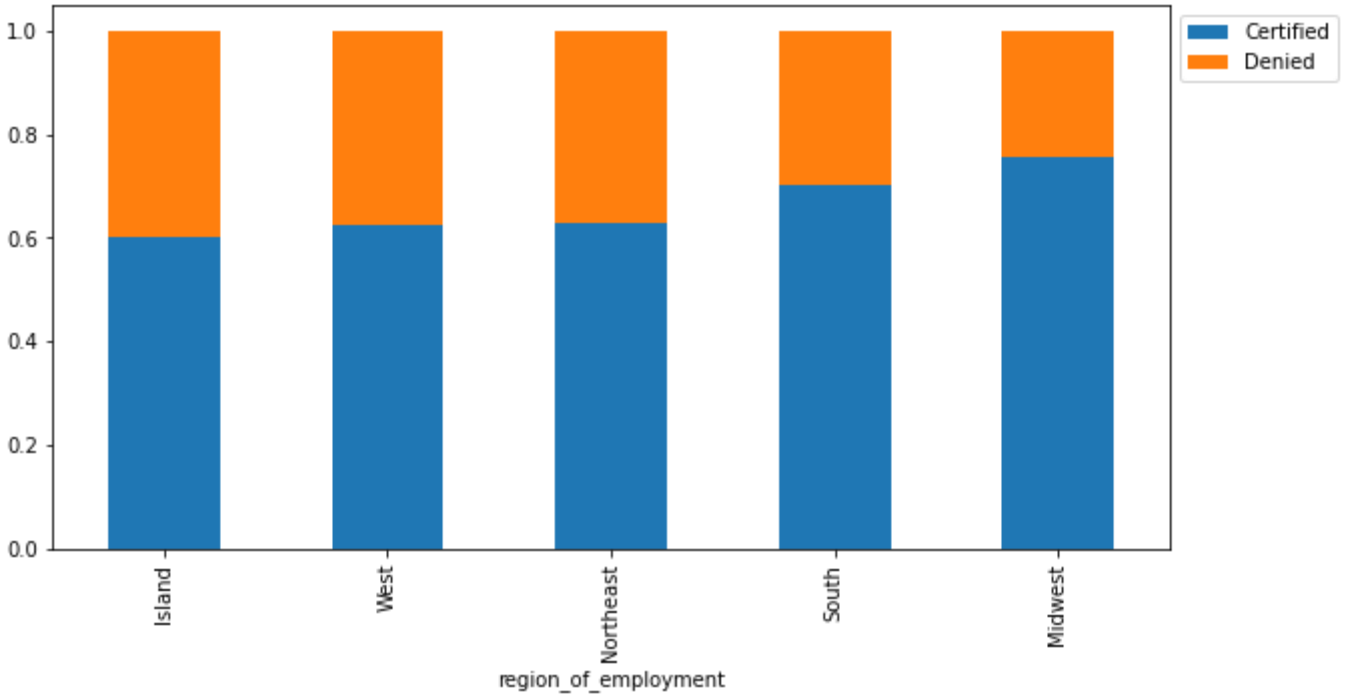
Education of employee Vs case status



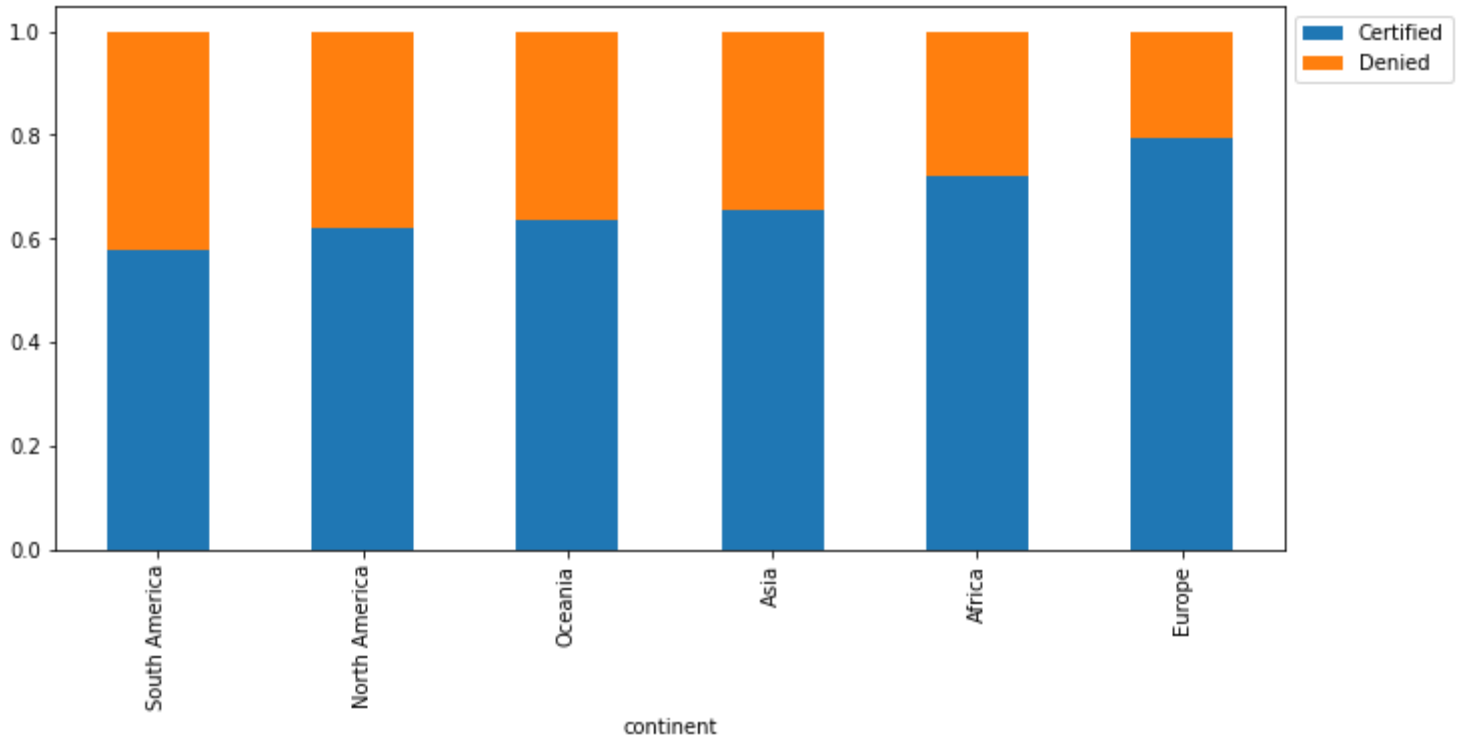
Education of employee Vs region of employment



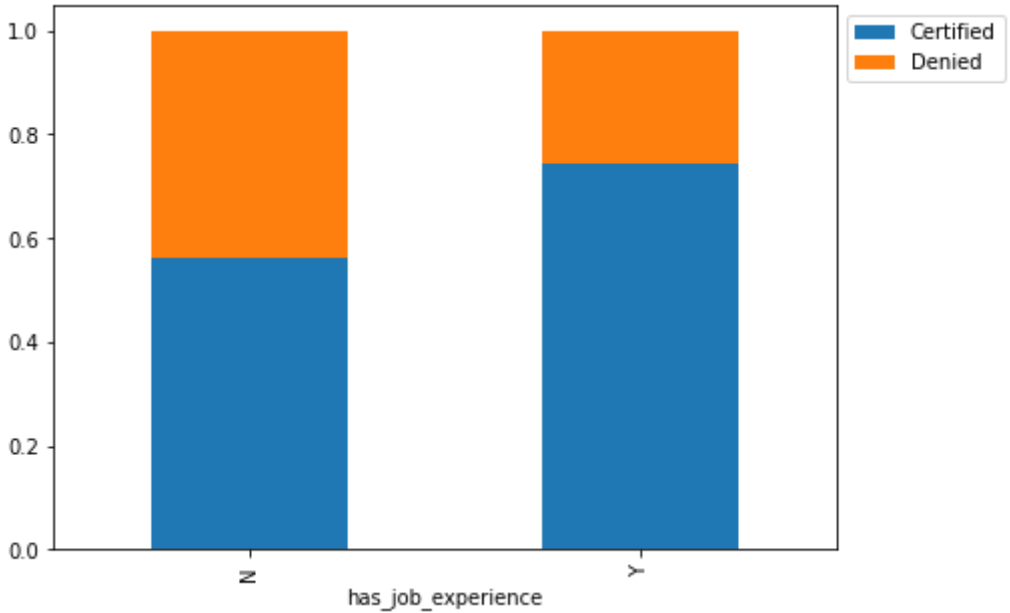
Region of employment Vs case status



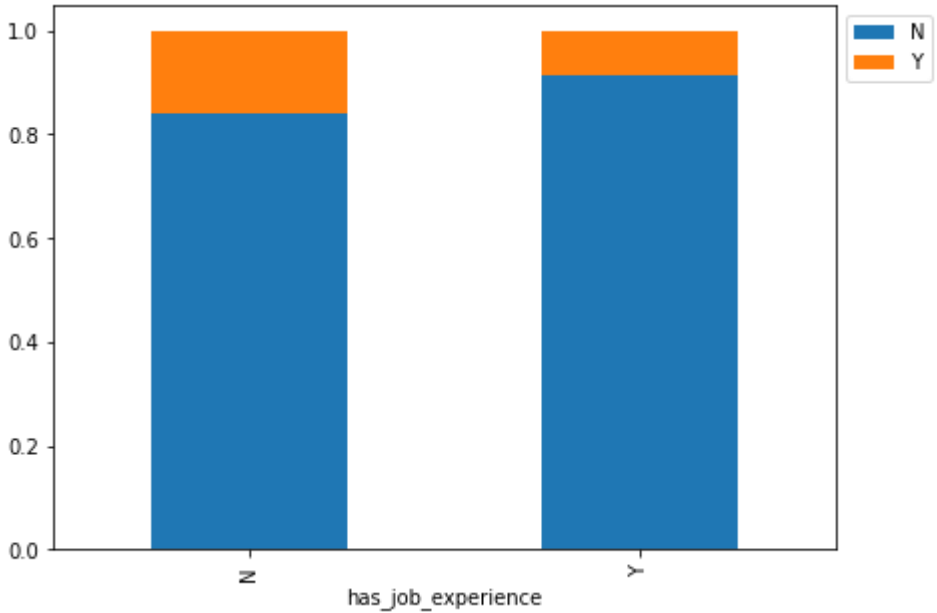
Continent Vs case status



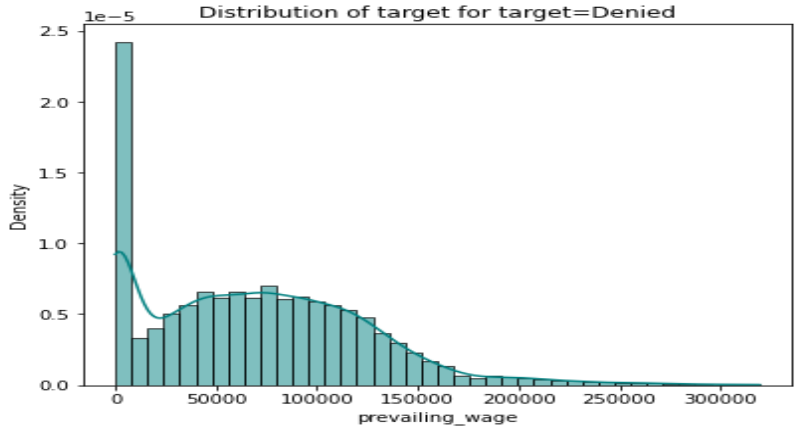
Job experience Vs case status

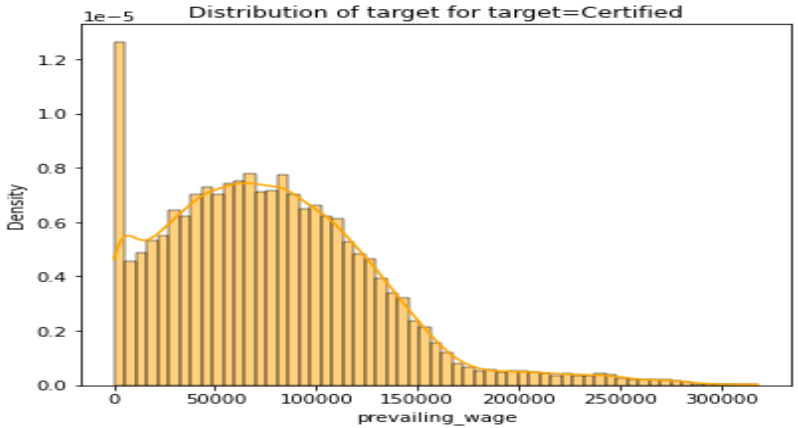


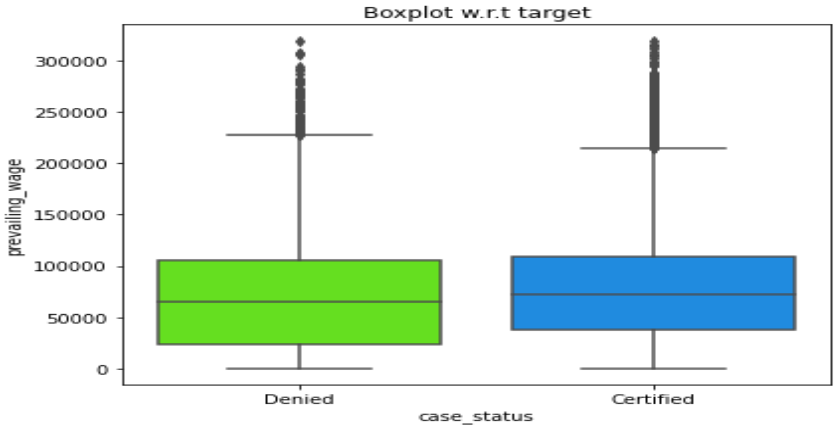
Job Experience Vs Job training

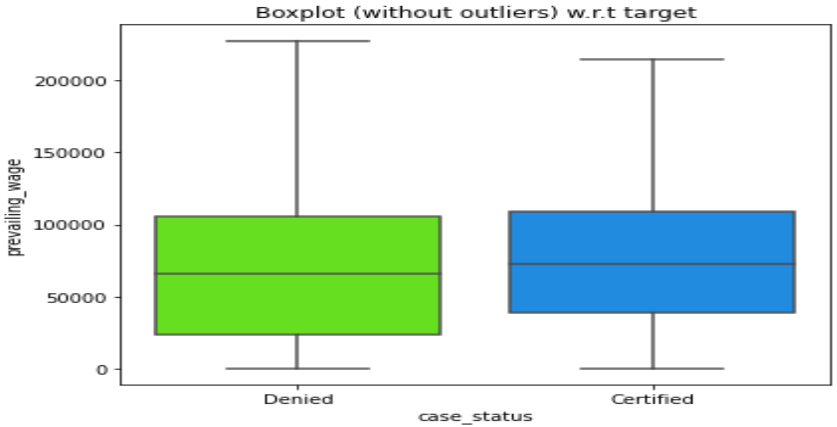


Prevailing wage Vs case status

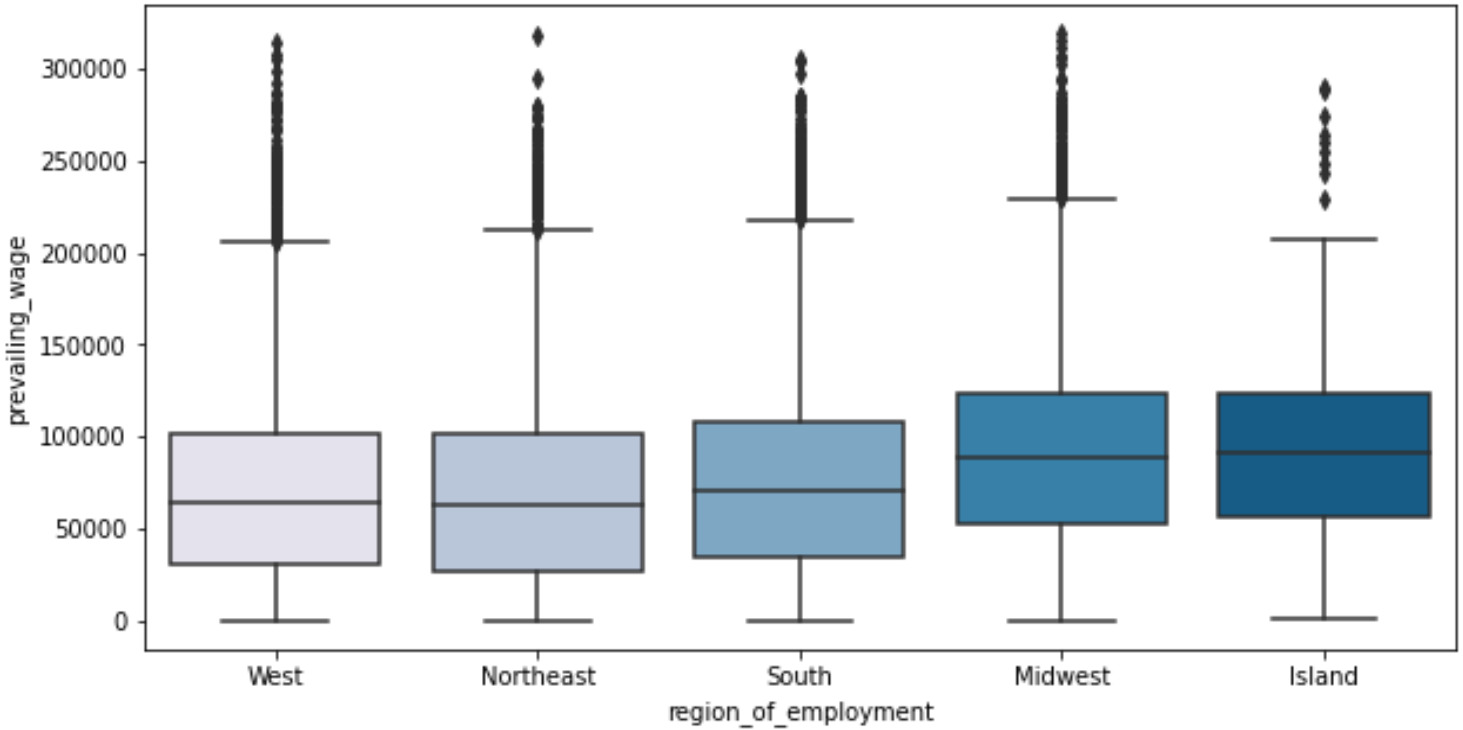




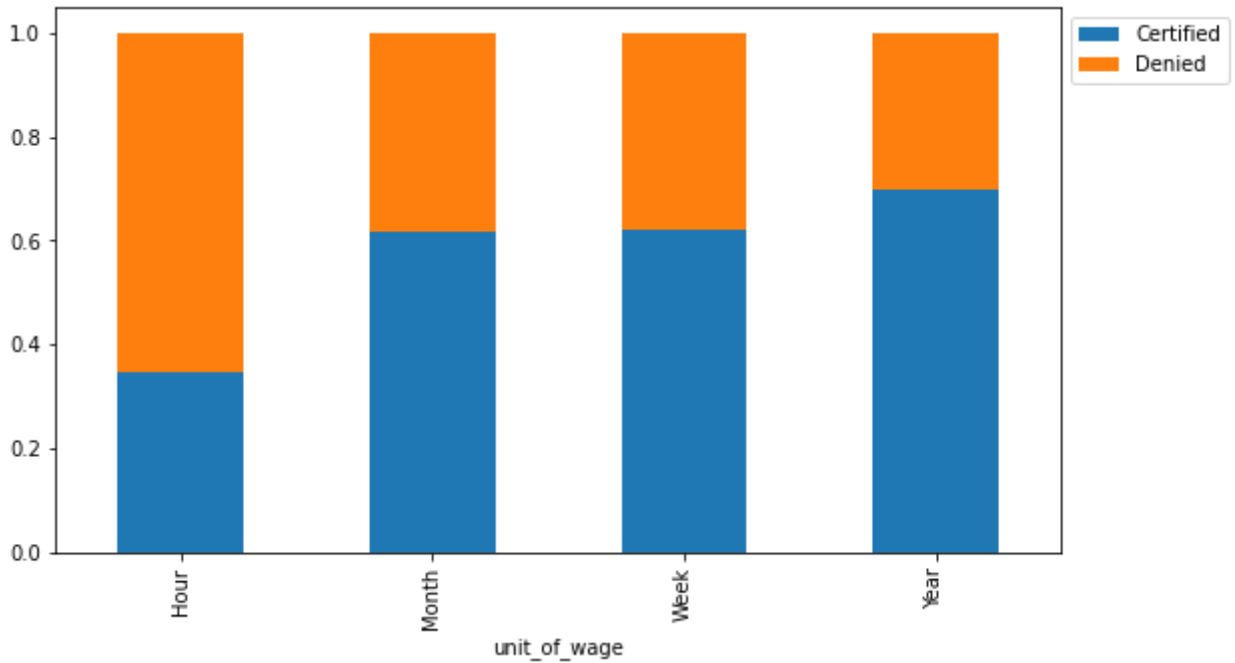




Region of employment Vs prevailing wage



Unit of wage Vs case status



Great Learning

Happy Learning !