greatlearning
Power Ahead

EasyVisa Business Case - Analysis

SARAVANAN RANGARAJAN (SR)

Contents

- Business Problem Overview
- Solution Approach
- Data Overview
- Education vs Case Status
- Continents vs Case Status
- Work Experience vs Case Status
- Wage Unit vs Case Status.
- Prevailing wage vs Case Status
- Education vs Prevailing Wage
- Bi Variate Analysis
- Model Summary – Initial
- Model Summary – Final
- Model Performance Summary
- Model Insights
- Recommendations
- Questions

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 :
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.

Solution Approach

Per the requirement is to perform Model building and perform Classification, Identify influential features and provide actionable insights.

Following are the steps to perform

1. Understand Dataset
2. Sanity Check
3. Missing values treatment
4. Outliers Detection and Treatment
5. EDA
6. Test & Train Model
7. Build various models like Decision Tree, Bagging Classifier, Random Forest Classifier, Boosting Models such as AdaBoost & Gradient Boost.
8. Hyper Parameter Tuning
9. Comparison of Performance of Models
10. Business Recommendations and Data Insights

Data Overview

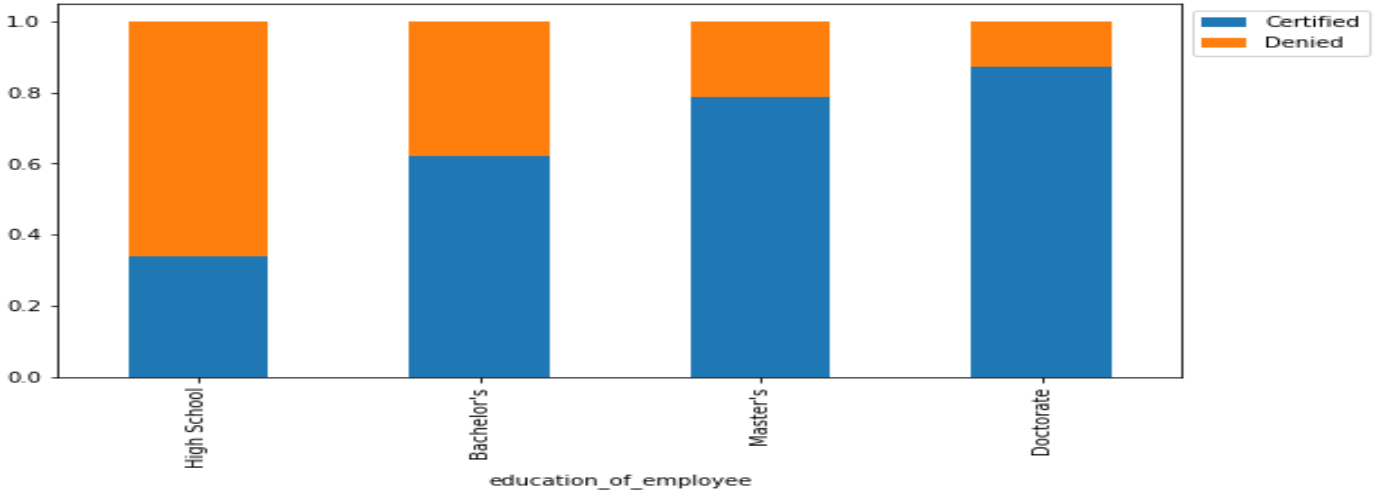
- **Data Dictionary**

- 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

Data Overview – Contd.

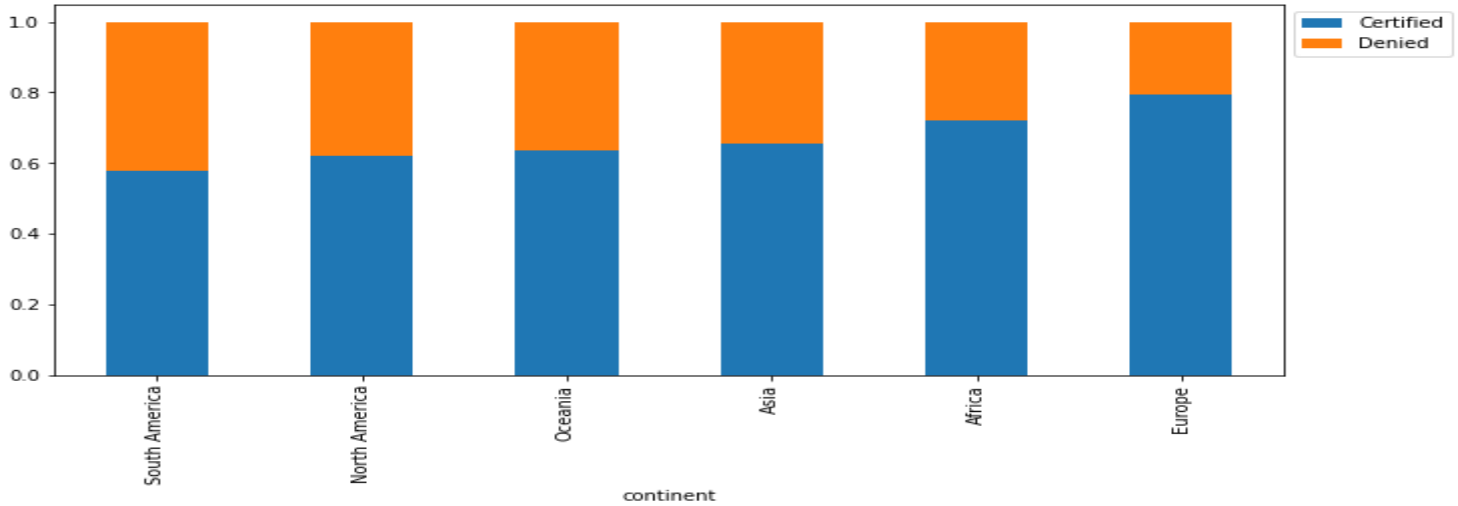
- It has 3 Numerical values & 9 Categorical values
- Null checks have been done to check the data. Null values doesn’t exist .
- Duplicate checks have been done. No Duplicate values exist.
- Check the distinct values in the Categorical columns and check the data spread for general information.
- Case Id columns looks like Unique identifier which will not help in prediction which we can drop.
- No Missing Values but No. of Employees column have negative which has been imputed by Multiplying by 1 for analysis.
- One hot Encoding was used for Continent, Region of Employment.
- Some outlier values in Prevailing wage column that has yearly salary less than 10,000 has been dropped.
- Converted prevailing wage in to One unit of wage for analysis. It has been converted to yearly.
- Data has 24432 rows and 16 columns after Pre-processing

1.Does education play a role in Visa certification?



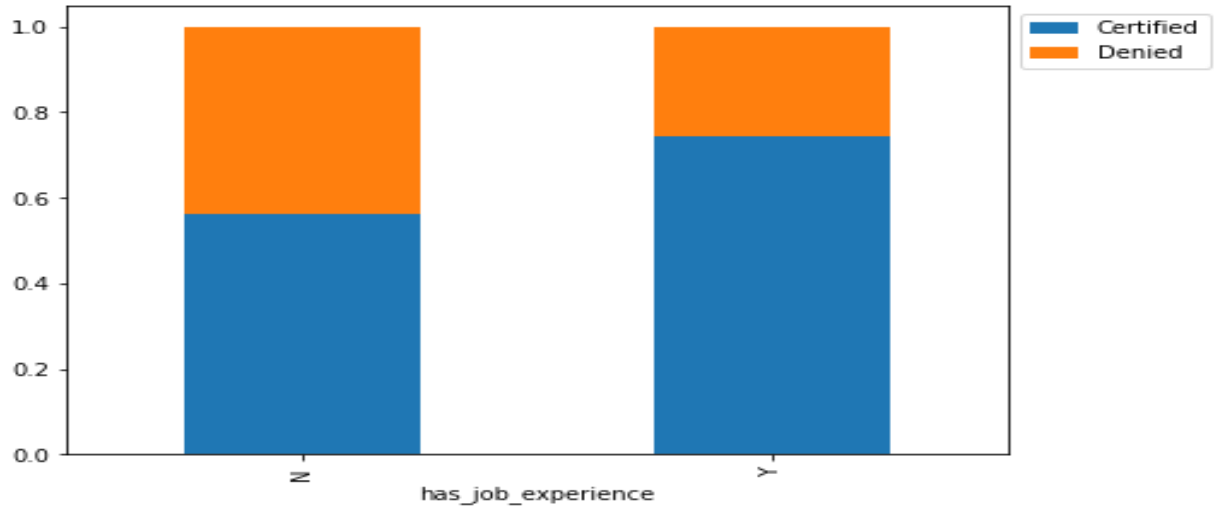
- It is obvious from above analysis, Education does have good influence in Certification status
- Doctorate has highest approval status with around 90% followed by Master's and Bachelor's
- Highschool has lowest.

2. visa status vary across different continents?



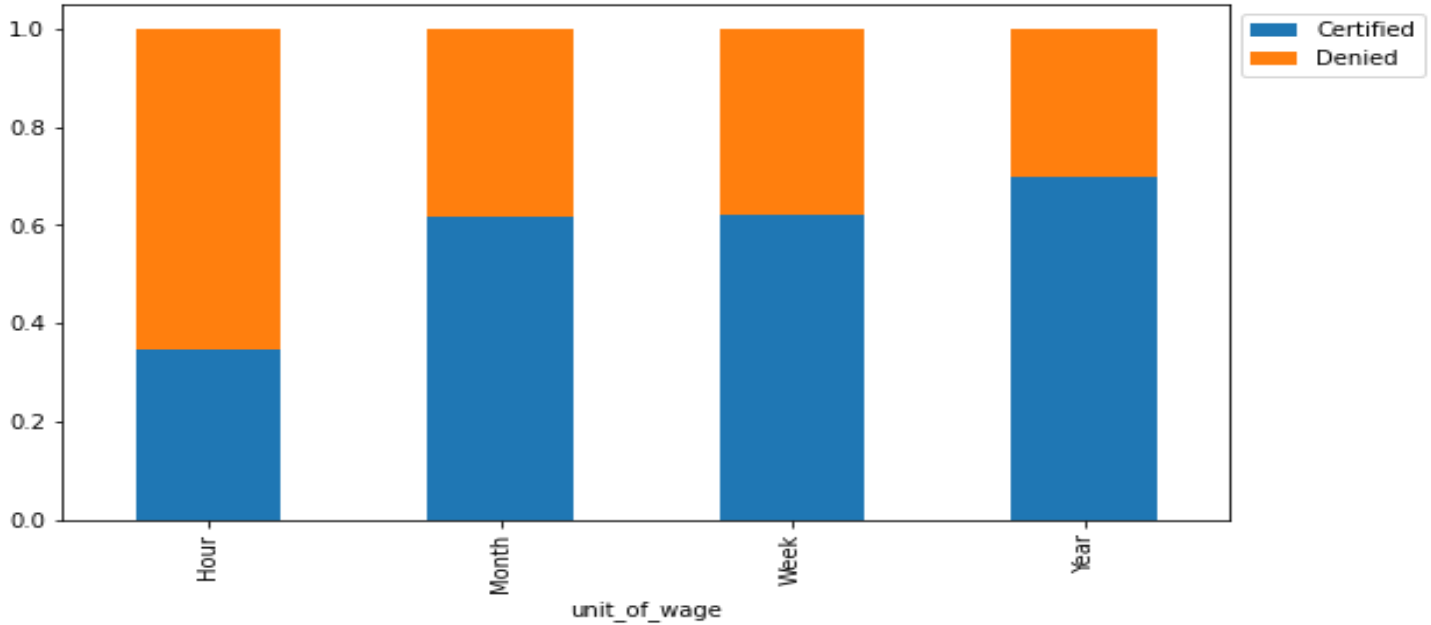
- Above analysis very clearly shows that Visa status varies across continents.
- Europe has highest approval rat followed by Africa and Asia.
- Surprisingly America's comes at the bottom of chart.

3. Does work experience influence visa status?



- Normally work experience would influence visa status specifically for Work visa types.
- Data Analysis clearly shows that having work experience does have edge over not having one.
- Job experience applicants have success rate around 75% and Fresh graduates have success around 55%.

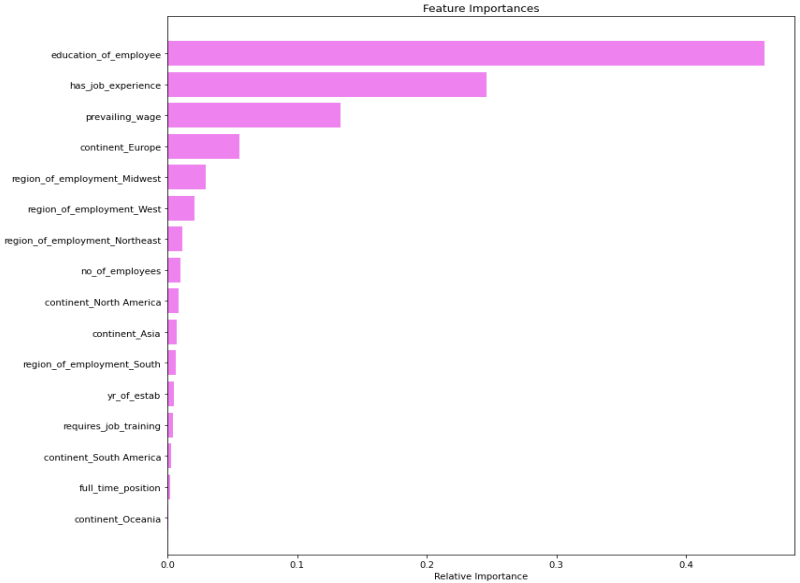
4. Which pay unit is most likely to be certified for a visa?

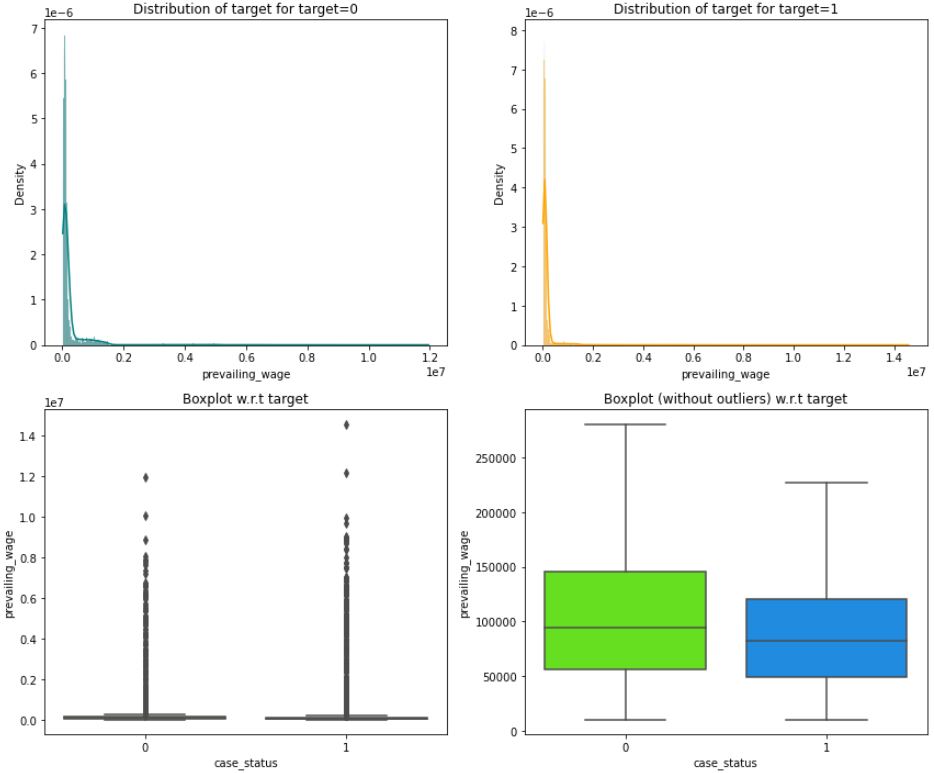


- Except hourly unit of wage, others have almost similar percentage of success in approval rate.
- But Yearly unit of wage has slight edge over others.

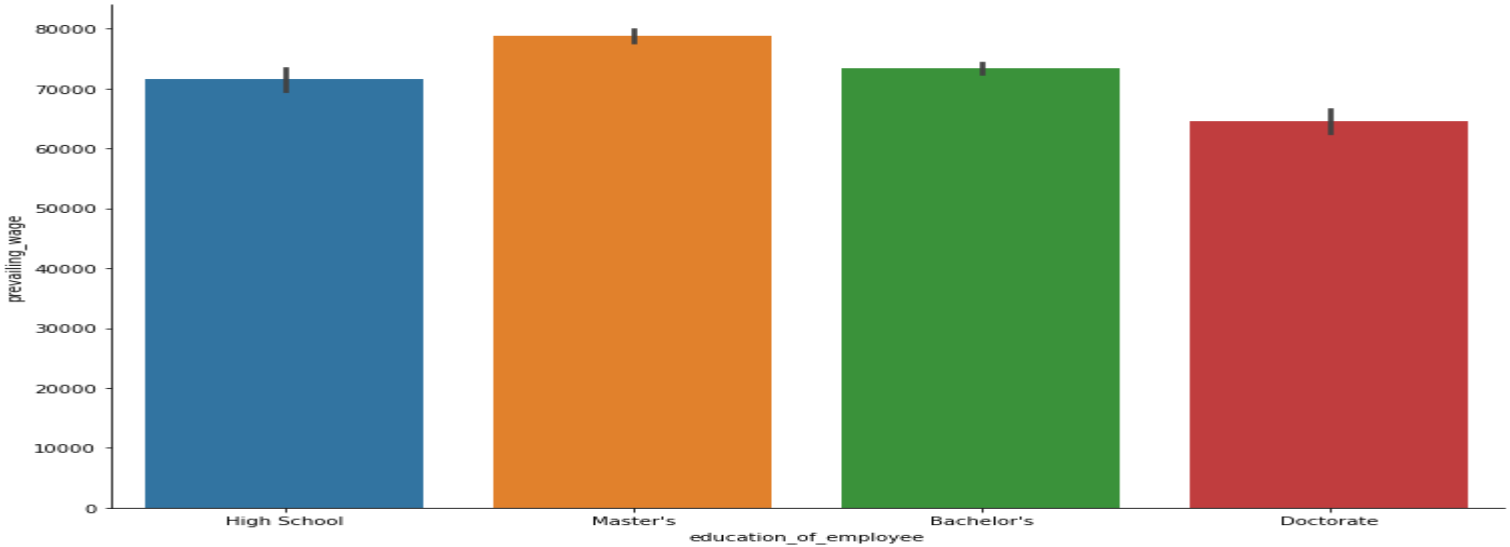
5. How does the visa status change with the prevailing wage?

- Prevailing wage doesn't have too much influence on visa status. Atleast data doesn't show much dependency or deciding factor.
- But for sure it has some influence. After tuning the model with Random Forest classifier, it shows that it does have some influence. It is the 3rd top most feature listed.



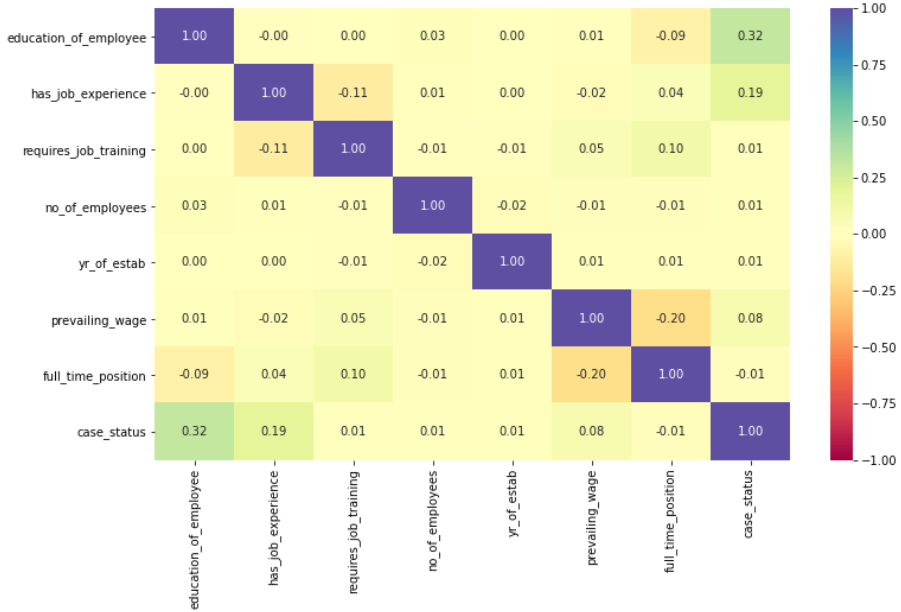


6. Education Influence Prevailing wage.



- Above chart shows analysis of Prevailing wage across Education of Employee.
- Interestingly Master's employees have high prevailing wage compare to Doctorate applicants and even lesser than Bachelors.
- The above is contradiction to approval status where in the Doctorate has most approvals followed by Masters and Bachelors.

Bi variate Analysis (Correlation)



- As expected Education of employee has high correlation with case status indicating better education has higher chance for approval / certified status.
- Job Experience and Prevailing wage also has positive Correlation indicating chances of influence in getting certified / approval.
- Requires job training has negative correlation with Job experience that is obvious.
- Another interesting relationship / positive correlation is between prevailing wage and requires job training.
- Full time position has negative correlation with case status that is kind of surprise.

Model Analysis Summary - Initial

- Initial Data size 25480 rows with 12 Columns
- After preprocessing 16 columns (due to dummy values for Categorical values) and 24432 rows (has to drop low level outliers in prevailing_wage) was the size taken for model analysis.
- Approx 10,000 rows were dropped.
- Model was split in to 70:30 ratio for train and testing.
- The Target variable in this case was “case_status”
- Predications can be
- Predicting an application will get Certified but in reality the Application will be denied.
- Predicting an application will get denied but in reality the application gets certified.
- F1 Score was used as metric (reduce FP and FN).
- Following Classifier models were used for analysis
- Decision Tree Classifier
- Random Forest Classifier
- Bagging Classifier
- AdaBoost Classifier
- GradientBoost Classifier
- Stacking Classifier.
- *XGBoost was not used due to performance issues*
- *Some classifiers with Class Weights used to check influence of Y variable.*

Model Analysis Summary - Final

- Most of the training sets resulted in value near to 1, representing model is overfitting.
- After tuning many of the models showed very negligible improvement. Hyperparameter tuning was done, but not at deep level.
- Due to performance issues, only 50% of Training data was used to train model in GradientBoosting. It may have influence in results.
- Many of the moels have produced nearly same F1 score (greater than or qual to 80%).
- Gradient Boosting model seem to have better number across Accuracy and F1 score. Decision Tree model has better Recall rate and Bagging Weights had better Precision score.
- As per features, Education has most influence followed by Job Experience, Prevailing wage, Europe Continent, Midwest Employment region followed by West and North East.

Models Performance Summary

	Decision Tree	Decision Tree Tuned	Random Forest Estimator	Random Forest Weights	Random Forest Tuned	Bagging	Bagging Weights	Bagging Tuned	Adaboost Classifier	Adaboost Tuned	Gradient Boost Estimator	Gradient Boost Tuned	Stacking Classifier
Accuracy	0.66	0.71	0.72	0.72	0.74	0.70	0.70	0.66	0.73	0.73	0.74	0.74	0.73
Recall	0.74	0.93	0.84	0.84	0.89	0.78	0.76	1.00	0.88	0.88	0.87	0.87	0.85
Precision	0.75	0.72	0.76	0.76	0.76	0.77	0.78	0.66	0.75	0.75	0.77	0.77	0.76
F1	0.74	0.81	0.80	0.80	0.82	0.78	0.77	0.80	0.81	0.81	0.82	0.81	0.81

- The un-tuned gradient boosting model is the best model here. It has the highest F1 score of approx 82% and highest Accuracy score.
- Gradient boosting, Random Forest Tuned Adaboost, and stacking Classifier are the top 4 models. They are all giving a similar performance

Model Insights

* Education of Employee has the most impact on the certification, followed by Job Experience and Prevailing wage.
* People from Europe have higher success rate.
* Experience definitely makes impact on the success rate.
* As per the analysis, Yearly paid employees gets most certification. It may be a factor showing they have long term commitment.
* Prevailing wage has impact on visa status obviously. From Pairplot we can observe higher prevailing wage has good chance of approval status.
* Interestingly applicants from Midwest region have good approval rate.
* Year of Establishment or Employee count have less influence on visa status

Business Recommendations

*Employers looking for Employees trying to get their cases certified should look for Good Education with Job Experience from Europe Continent has good chances of certification.

Questions

- Questions give different Perspective, most of the time result in idea.

greatlearning
Power Ahead

Happy Learning !