EasyVisa Project

Contents

* Business Problem Overview and Solution Approach
* Data Overview
* Data Processing Initial Steps
* EDA

Univariate Analysis

Bivariate Analysis

* Data Processing Other Steps

Outlier Detection and Treatment

* Bagging and Boosting Models
* Model Performance Evaluation
* Business Insights and Recommendations

Business Problem Overview and Solution Approach

Context:

- 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:

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 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.

Data Overview

The data contains the different attributes of the employee and the employer. The detailed data dictionary is given below.

- 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 Processing Initial Steps.

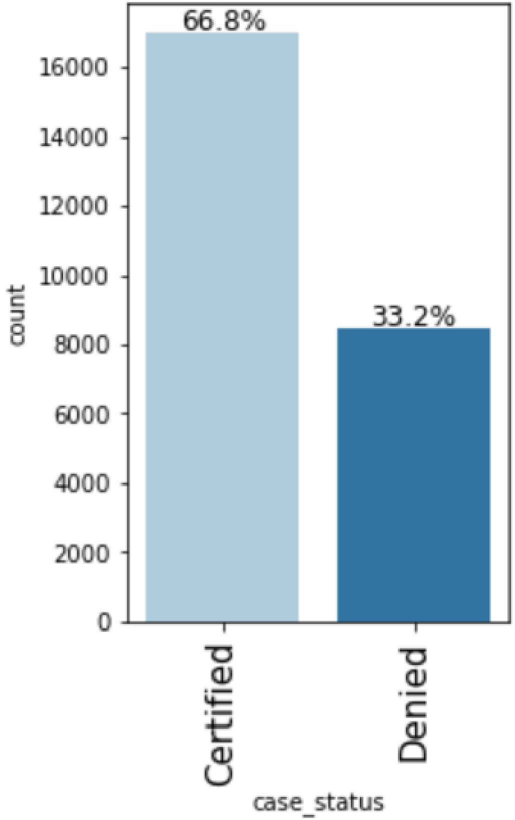
- Dataset has 25480 rows and 12 columns
- Most of the data-types are object, this means we need to convert these into suitable data-type before we feed our data into the model.
- Two columns no_of_employees and yr_of_estab are of int64 data-types and one of the columns prevailing_wage has data-type float64.
- There are no null values in the dataset
- There are no duplicate or missing values in the dataset.

EDA

- **UNIVARIATE ANALYSIS OF case_status**

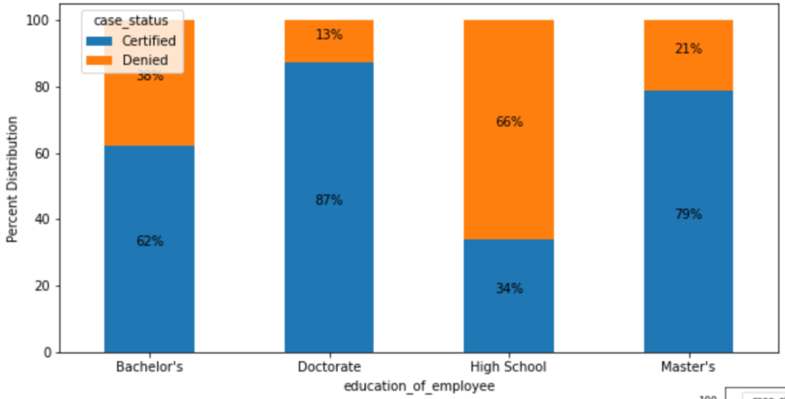
Observations

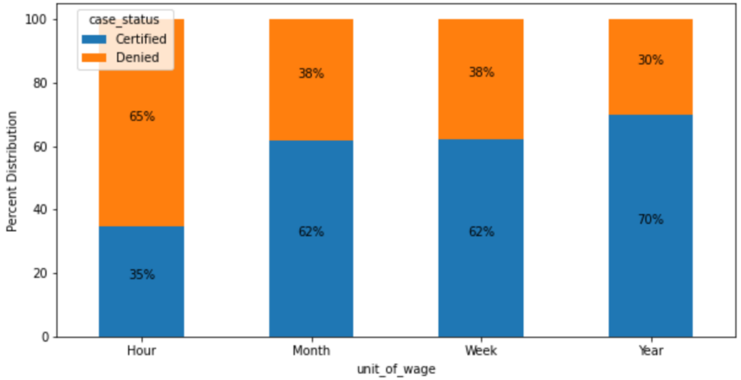
- 66.8% of employees are certified and given visa.
- 33.2% of employees are Denied visa.
- We need to find the variables that have strong relation with case_status and that would result in Visa Certifications instead of Denial.

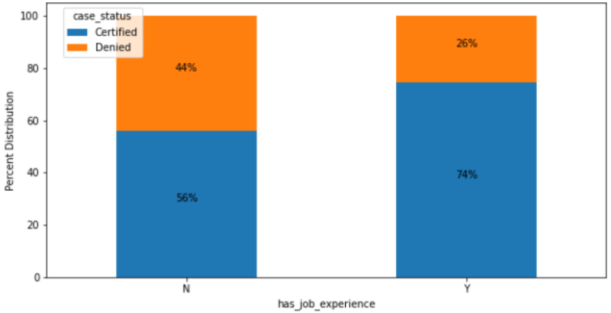


BIVARIATE ANALYSIS

- The three most important features affecting the Certification if case_status are education_of_employee, unit_of_wage and has_job_experience.







Insights Based on EDA

- There are no significant correlation between any of the Numerical variables.
- 75% of the number of employess are below 3505. But there are many companies with huge number of employees, the max value being 602069.
- We do not find any significant relation between case_status and no_of_employees.
- The yr_of_estab varies from 1800 to 2016. More number of comapnies were established after 1970s.
- We do not find any significant relation between case_status and yr_of_estab.
- prevailing_wage varies from 2.14 to 319210. Mean is 74455 and median is 70308. More number of prevailing_wages are on the lower side.
- For prevailing_wage below 50000, the Denial cases are more compared to Certified cases.
- 90.1% prevailing wages given are yearly wages.0.3 % are monthly, 8.5% are Hourly and 1.1% are weekly wages.
- Employees who are paid by hour have only 35% Certification cases. Employees paid by year have 70% Certification cases.
- The highest percent 66.2% employees are from Asia. The least with 0.8% of employees are from Oceania.

- Employees from Europe have highest Certification with 79% cases.Employees from South America has the least Certification with 58% cases.
- 40.2% of applicants have Bachelor's degree followed by 37.8% with Master's. 13.4% employees have High School degree and 8.6% have Doctorate.
- We can see that as the education increases there are higher chances of Visa Certification. Employees with Doctorate have 87% Certification cases. Employees with only High School Education have only 34% cases of Certification.
- 58.1 % employess have job experience.41.9 % employees do not have job experience.
- Job experience plays an important role in Visa Certification. 74% employees with job experience have Visa Certification compared to 56% employees with no job experience.
- 88.4% of employees require job training. 11.6% employess do not require job training.
- Requiring job training does not seem to have any impact on the Visa Certification.
- 28.3% emplyments are in the Northeast region of US. Only 1.5% emplyments are in the Island region of US.
- Companies in Midwest Region have highest Certification cases with 76%. Companies in Island region have the least Certification cases with 60%.
- 89.4% of the positions are full-time. Only 10.6% are not full-time positions.
- If the employment is a Full time position does not seem to have much impact on the Visa Certification.

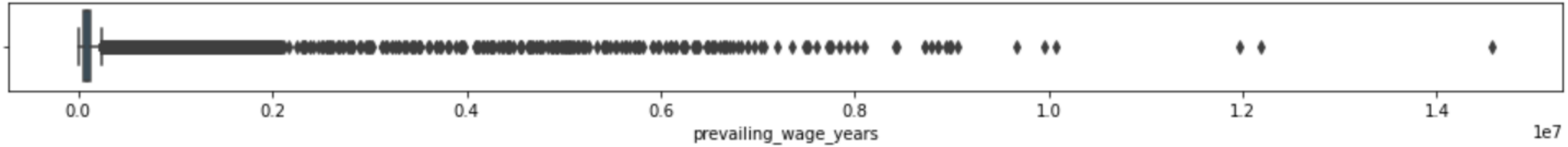
Data Processing Other Steps- Outlier Detection and Treatment

* We convert all the prevailing_wage values into a single unit so that it becomes the yearly wage. We save this in the column prevailing_wage_years.

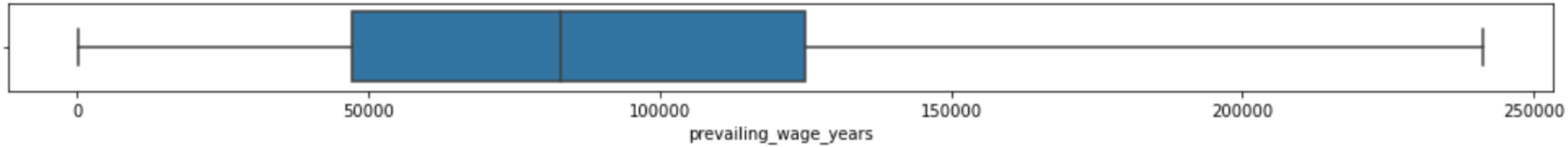
* We see that there are many outliers towards the right for this column prevailing_wage_years.

* We remove these outliers through Outlier Treatment.

Before:



After:



Bagging and Boosting Models

Model evaluation criterion

- Model can make wrong predictions as:

Predicting an employee will not get Visa Certification and the employee gets Visa Certification.

Predicting an employee will get Visa Certification and the employee does not get Visa Certification.

- Which case is more important?

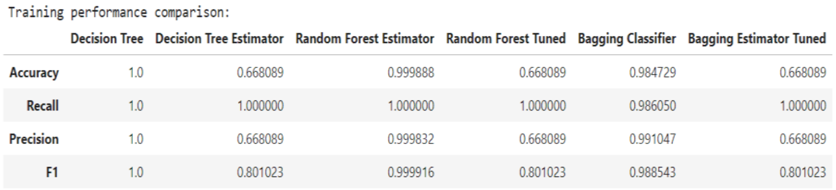
Predicting that employee will get Visa Certification and the employee does not get Visa Certification, i.e. losing on a valuable employee or asset.

- How to reduce this loss i.e need to reduce False Positives?

Company wants Precision to be maximized, greater the Precision higher the chances of minimizing false positives. Hence, the focus should be on increasing Precision or minimizing the false positives so that the company get hard-working, talented, and qualified individuals both locally as well as abroad.

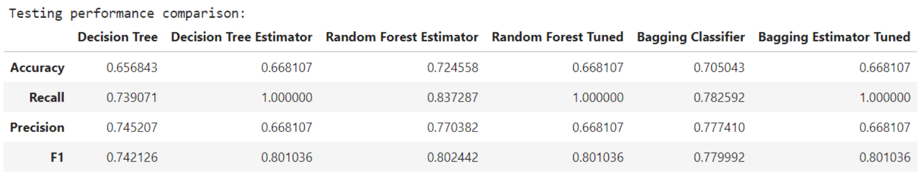
Bagging Models Conclusion

- We tried Decision Tree, Random Forest and Bagging Classifier.
- None of the models seem to be fully meeting all needed criteria.
- Precision values are high ~77 for Random Forest and Baaging Classifier but thier Train models are overfitting.
- Tuned Random Forest and Tuned Bagging Classifier have generalised Train and Test models. But their Precision value is less which is ~66.
- The three most important features to determine Visa Certification as per Tuned Random Forest model are -education_of_employee, has_job_experience and unit_of_wage



```
Training performance comparison:
```

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned
Accuracy	1.0	0.668089	0.999888	0.668089	0.984729	0.668089
Recall	1.0	1.000000	1.000000	1.000000	0.986050	1.000000
Precision	1.0	0.668089	0.999832	0.668089	0.991047	0.668089
F1	1.0	0.801023	0.999916	0.801023	0.988543	0.801023

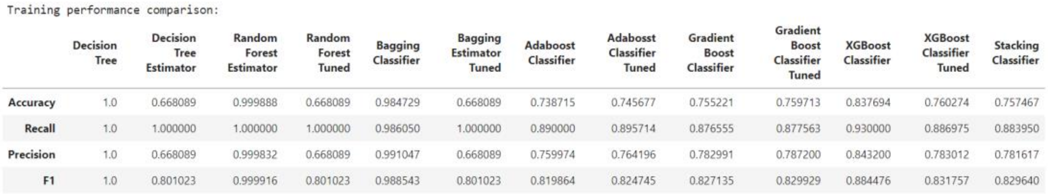


```
Testing performance comparison:
```

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned
Accuracy	0.656843	0.668107	0.724558	0.668107	0.705043	0.668107
Recall	0.739071	1.000000	0.837287	1.000000	0.782592	1.000000
Precision	0.745207	0.668107	0.770382	0.668107	0.777410	0.668107
F1	0.742126	0.801036	0.802442	0.801036	0.779992	0.801036

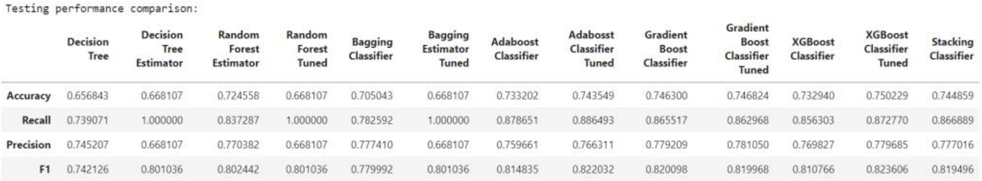
Boosting Models Conclusion

- We tried AdaBoost, Gradient Boost and XGBoost and Stacking Models.
- Both Tuned Models of Gradient Boost Classifier and XGBoostClassifier have done pretty well among all the other given models.
- These models are giving a generalized perfoirmance compared to other models.
- There is also no overfitting of Train data.
- The Precision score is around ~78 which is good.
- The three most important features to determine Visa Certification as per Tuned Random Forest model are -education_of_employee, has_job_experience and unit_of_wage



Training 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	1.0	0.668089	0.999888	0.668089	0.984729	0.668089	0.738715	0.745677	0.755221	0.759713	0.837694	0.760274	0.757467
Recall	1.0	1.000000	1.000000	1.000000	0.986050	1.000000	0.890000	0.895714	0.876555	0.877563	0.930000	0.886975	0.883950
Precision	1.0	0.668089	0.999832	0.668089	0.991047	0.668089	0.759974	0.764196	0.782991	0.787200	0.843200	0.783012	0.781617
F1	1.0	0.801023	0.999916	0.801023	0.988543	0.801023	0.819864	0.824745	0.827135	0.829929	0.884476	0.831757	0.829640



Testing 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	0.656843	0.668107	0.724558	0.668107	0.705043	0.668107	0.733202	0.743549	0.746300	0.746824	0.732940	0.750229	0.744859
Recall	0.739071	1.000000	0.837287	1.000000	0.782592	1.000000	0.878651	0.886493	0.865517	0.862968	0.856303	0.872770	0.866889
Precision	0.745207	0.668107	0.770382	0.668107	0.777410	0.668107	0.759661	0.766311	0.779209	0.781050	0.769827	0.779685	0.777016
F1	0.742126	0.801036	0.802442	0.801036	0.779992	0.801036	0.814835	0.822032	0.820098	0.819968	0.810766	0.823606	0.819496

Business Insights and Recommendations

- The three most important features that we have found affecting the Visa Certification are education_of_employee, has_job_experience and unit_of_wage.
- Companies should Increase wage and keep the minimum wage around 50000. We have found that there are more Denied cases for wage<50000.
- The wage should be paid out yearly and if possible hourly wages should be stopped. Employees with wages paid in year have 70% chances of certification whereas for hourly wages it is only 35%.
- Focus more on recruiting employees from Europe and Africa where there are higher chances of Visa Certification.
- Find Reasons why there might be less certifications for employees from other continents and work upon them.
- Recruit more and more highly educated employees with Doctorate and Master's as they have high chances of visa certifications.
- Mention Doctorate/Masters as one the required qualifications in the job description so that more and more highly education employees can apply.

- Job experience plays an important role in Visa Certification. 74% employees with job experience have Visa Certification compared to 56% employees with no job experience.
- Mention Job experience as one the required qualifications in the job description so that more and more highly qualified employees can apply.
- Companies in Midwest Region have highest Certification cases with 76%. Companies in Island region have the least Certification cases with 60%.
- If companies can open their new offices in Midwest region, they may get more employee with Visa Certification.

greatlearning

Power Ahead

Happy Learning !