EasyVisa

(Factors influencing the process of visa approvals in the United States)

EasyVisa– Data analysis of factors influencing Visa approvals
The University of Texas at Austin
McCombs School of Business

Itu Mukherjee
Date : 13.01.2023

Contents / Agenda

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

Executive Summary

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 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. We analyzed data with the help of classification models to:

- 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.

Executive Summary

Recommendations

- To prioritize limited resources towards screening a batch of applications for those most likely to be approved, the OFLC can:
  - Sort applications by level of education and review the higher levels of education first.
  - Sort applications by previous job experience and review those with experience first.
  - Divide applications for jobs into those with an hourly wage and those with an annual wage, sort each group by the prevailing wage, then review applications for salaried jobs first from highest to lowest wage.

Executive Summary

- As stated previously, the Gradient Boosting classifier performs the best of all the models created. However, as shown above, the tuned Decision-Tree model performs barely worse by F1 score and is a far simpler model. This model may be preferable if post-hoc explanations of OFLC decision-making is expected to be required.
  - Furthermore, OFLC should examine more thoroughly why whether an application will be certified or denied can be very well predicted through just three nodes as shown above.
  - For those in less skilled, entry-level, and/or hourly jobs, the system would appear to be biased against these applications being certified.

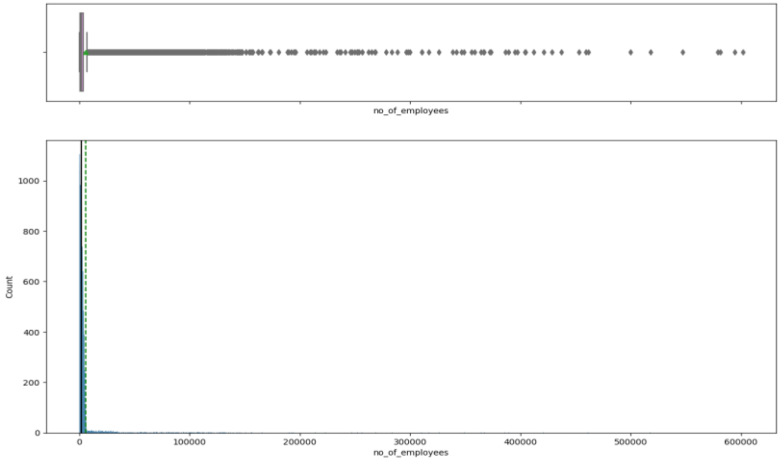
Business Problem Overview and Solution Approach

- Our objective is to identify the factors that have a high impact for approvals of visas and build a model that can predict that are influencing the criteria for visa approvals.

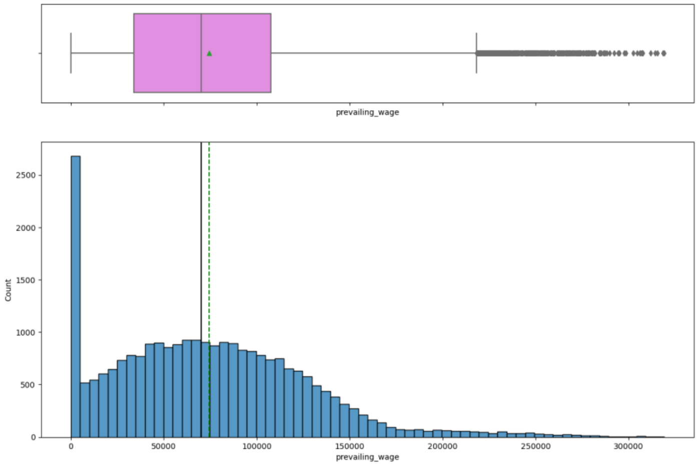
- Solution approach and methodology
  - EDA (bivariate and univariate analysis), duplicate value check, missing value treatment, outlier check (treatment if needed)
  - Model building and Hyperparameter Tuning
    - Decision Tree, Random Forest, Bagging, Boosting Classifiers (AdaBoost , Gradient Boosting, XGBoost), Stacking Classifier
  - Model Performance Comparison and Final Model Selection
  - Important features of the final model

Exploratory Data Analysis

Number of employees

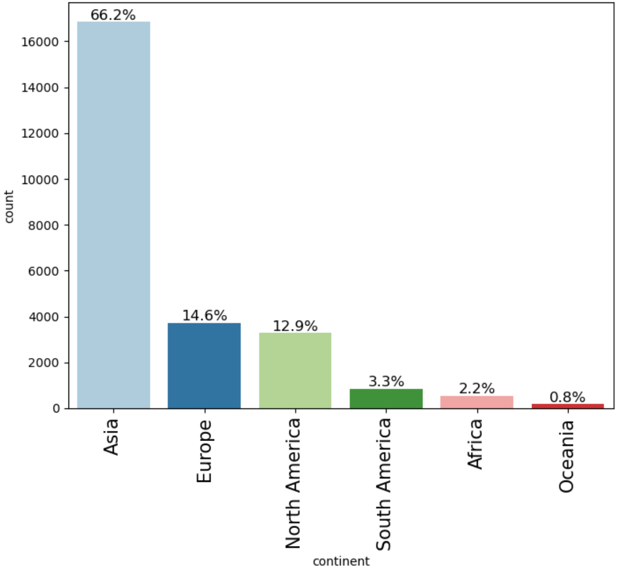


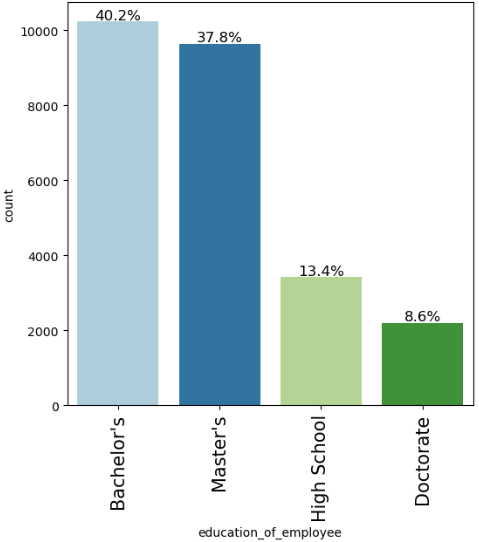
Number of employees

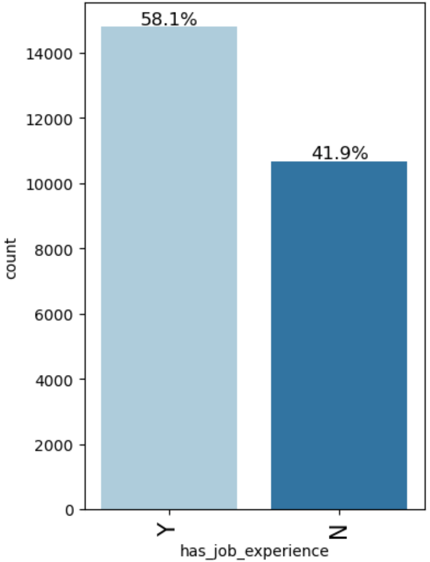


- The distribution for number of employees for employers is heavily skewed right
- The average and median annual salary is approx. USD 70,000 which seems accurate
- The trend appears correct with outliers in the higher income bracket between USD 200,000 to USD 300,000
- There are several very low salaries as well, which appears incorrect and requires further investigation

Exploratory Data Analysis

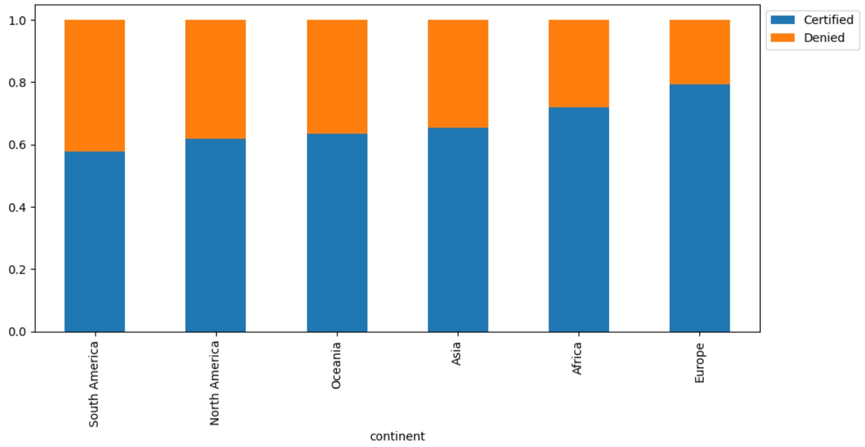


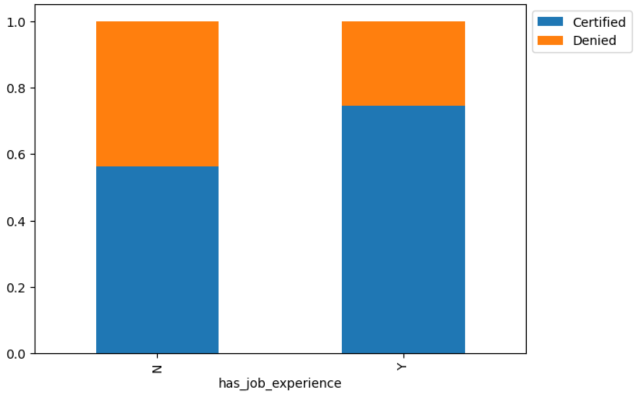




- Majority of employees >50% are from Asia
- Majority of employees have either a bachelor's 40% or a master's 38% and minority of applicants have either a doctorate 8% or only a high school diploma 13%
- Around 58% employees have prior job experience and 42% employees do not

Exploratory Data Analysis

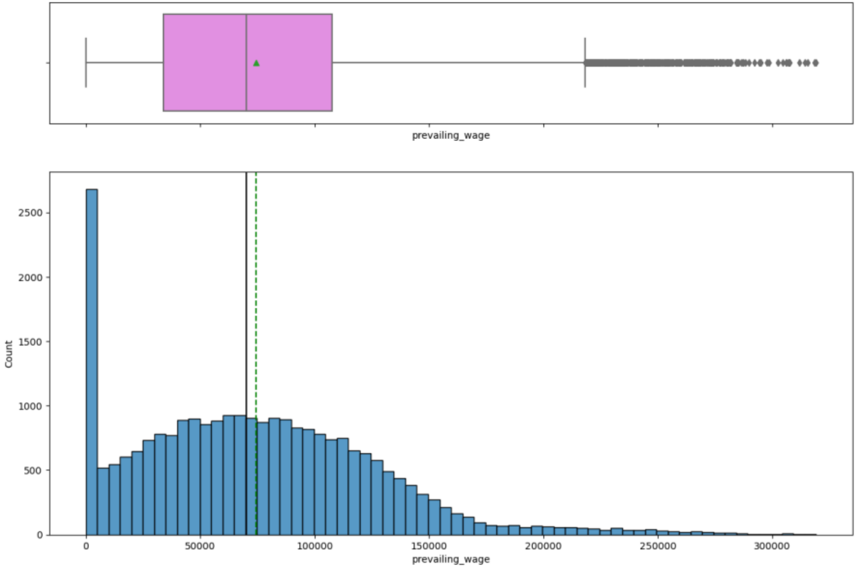




- Irrespective of the continent the employee is from, more cases are certified than denied
- The trend observed w.r.t % certification for continents is Europe > Africa > Asia > Oceania > North America & South America
- As expected, the trend observed w.r.t % visa certifications with having job experience

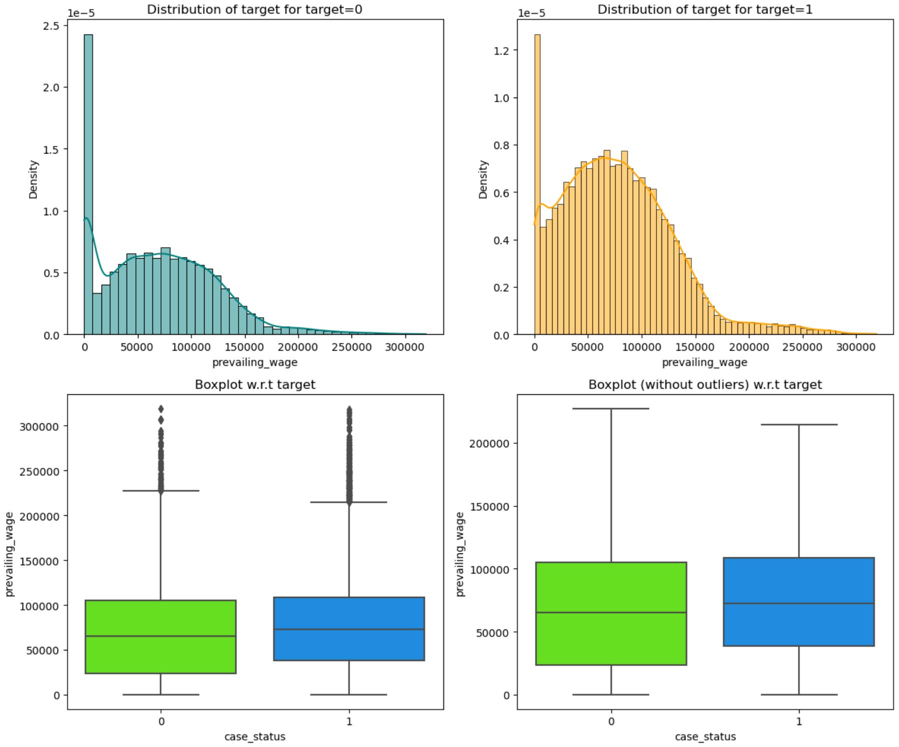
Exploratory Data Analysis

Prevailing wage & unit of wage



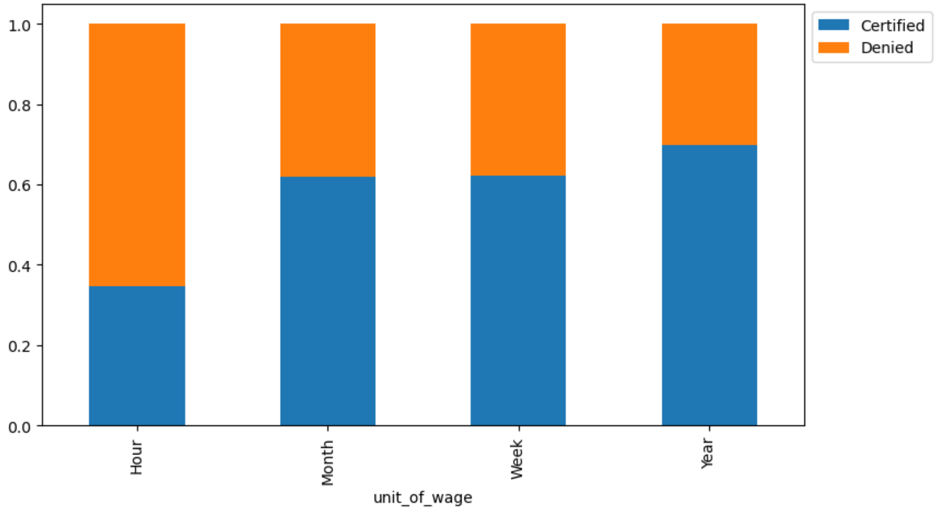
- Count & % certification for visa statuses by unit of wage
- Unit of wage is assumed to be not hourly, when employee receives fixed salary irrespective of number of hours worked (i.e.,Week, Month, or Year) & hourly otherwise (i.e., Hour)

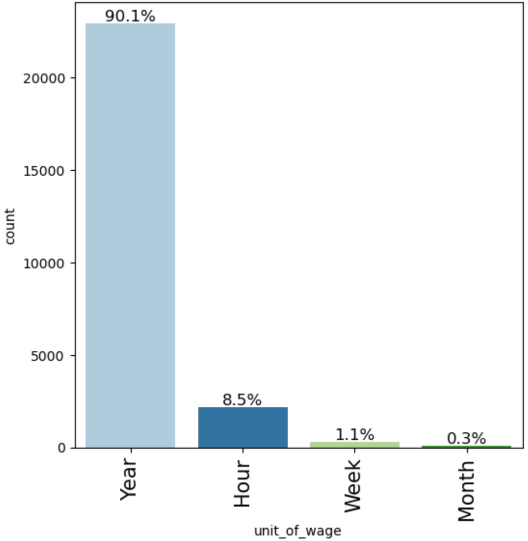
Exploratory Data Analysis



- Almost 92% of all entries are with unit_of_wage as not hourly & only 8% entries as hourly
- 70% of cases are certified when the unit_of_wage is not hourly, and only 35% cases are certified when the unit_of_wage is hourly • Prevailing_wage was cleaned up to only contain annual wages

Exploratory Data Analysis





- Values on the lower end (US$200K). These are not treated further as decision tree ML model is robust to outliers
- General trend that ~ twice the cases are certified more than denied, dropping slightly & increasing slightly on lower & upper end of wages respectively

Exploratory Data Analysis

- More than twice the number of cases were certified than denied irrespective of the number of employees in the employer's organization & the year of establishment of the employer's organization. These attributes are hence, not thought to have an impact on case statuses
- From the EDA, we infer 58% of all cases were for smaller organizations (<2500 employees) and 61% of all cases were for employer's established after 1990
- Only 35% of the cases were certified when the unit of wage is hourly but 70% were certified when the unit of wage are not hourly (i.e, Weekly, Monthly or Yearly). This indicates unit of wage is an important attribute that can influence case statuses
- Only 8.5% of all cases were for unit of wage hourly and the remaining 91.5% of all cases were for unit of wage not hourly (i.e, Weekly, Monthly or Yearly)

Exploratory Data Analysis

- Majority of applicants have a bachelor's (40%) or a master's degree (37.87%). A small number have only high school certification (13.4%) or are very highly educated/ doctorate (8.6%). However, cases getting certified is highest for doctorate degree (>86%),followed by master degree (>76%), then bachelor's (~62%). The cases getting certified is very low for those applicants with only a high school certification (<35%). The trend observed is intuitive and one can expect attributes having a doctorate degress & having only a high school certification to significantly contribute to a case being certified and denied respectively
- From the EDA, we infer that 58% of all applicants have prior job experience and 42% do not. The cases getting certified is high for applicants with prior job experience (75% of such cases) and low for applicants without prior job experience (~56% of such cases). This is again an important attribute with an applicant having prior job experience significantly contributing to a case being certified
- Majority do not require the employee to receive any additional job training. This attribute was not found to have an impact on the case statuses

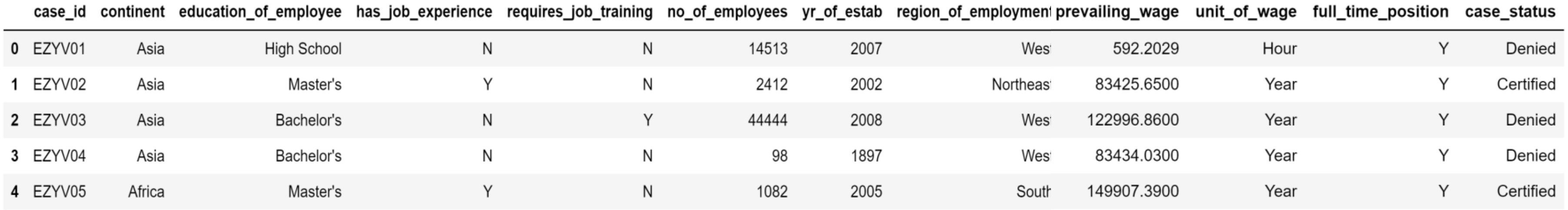
Exploratory Data Analysis

- Majority of the applications are to Northeast (28.3%), then South (27.5%), then West (25.8%), Midwest (16.9%) and least to Island (1.5%) regions of the US. However, the cases certified follows the trend Midwest (75% of such cases), then South (70% of such cases), then Northeast, West, & Island (60% of such cases). Region of employment being Midwest hence is an important attribute contributing positively to a case being certified

- Majority of the jobs are full time rather than part time. This attribute was not found to have an impact on the case statuses

- Majority of cases are from applicants in Asia (66%), then Europe (15%), N.America (13%) & S.America (3%). However, cases getting certified is highest for Europe then Africa then Asia & least for S.America & N.America. Being from Europe is though to be an important attribute to have an impact on case statuses

Data Overview



	case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employmen	prevailing_wage	unit_of_wage	full_time_position	case_status
0	EZYV01	Asia	High School	N	N	14513	2007	Wes	592.2029	Hour	Y	Denied
1	EZYV02	Asia	Master's	Y	N	2412	2002	Northeas	83425.6500	Year	Y	Certified
2	EZYV03	Asia	Bachelor's	N	Y	44444	2008	Wes	122996.8600	Year	Y	Denied
3	EZYV04	Asia	Bachelor's	N	N	98	1897	Wes	83434.0300	Year	Y	Denied
4	EZYV05	Africa	Master's	Y	N	1082	2005	Sout	149907.3900	Year	Y	Certified

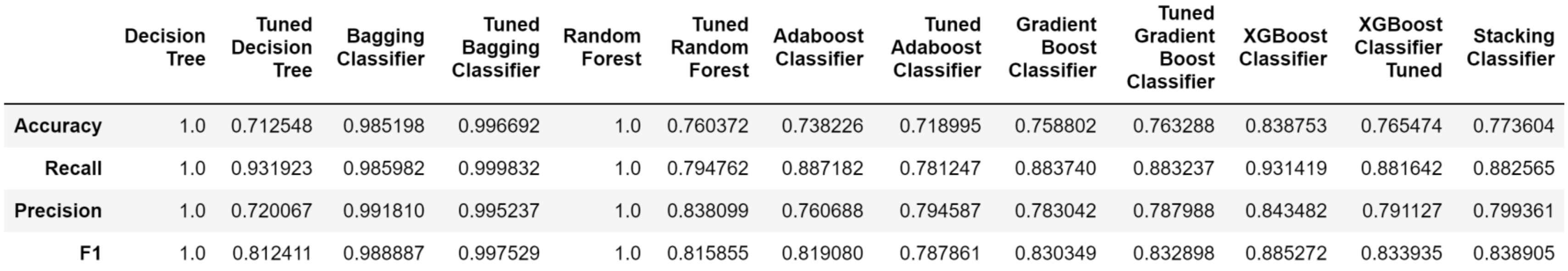
- The average number of employees in the employer's organization are 5667 while the median number of employees in the employer's organization are 2109. This implies the attribute has a right skewed distribution with several positive outliers.
- The minimum number is negative which does not appear to be a valid data point
- There are companies in the dataset with years of establishment from 1800 to 2016
- The average prevailing wage for occupation is united states is USD 74,455 while the median (~50th percentile of wages) is USD 70,308. This indicates, slight right skewness in the data set. The minimum value of USD 2.1367 does not appear to be a valid data point. The attribute has to be studied in union with unit_of_wage to gather further insight

Data Preprocessing

- Duplicate value check - There are no duplicate values in the Data Frame
- Missing value treatment - There are no missing values
- Outlier check (treatment if needed) - No insignificant outliers found
- Data preparation for modeling
    - To evaluate the model, we split the data into train and test in the ratio of 70-30.
    - Built models using Decision tree, Randam forest, Bagging Classifier, Boosting Classifier (AdaBoost, Gradient Boost, XGBoost), Stacking Classifier the train data and then checked its performance.
    - Checking model performance on training and test sets

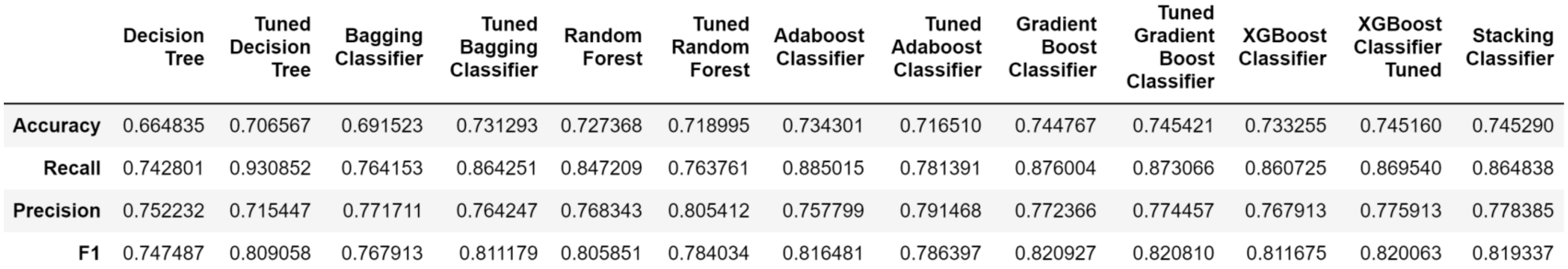
Model Performance Summary

Training performance comparison



	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	0.712548	0.985198	0.996692	1.0	0.760372	0.738226	0.718995	0.758802	0.763288	0.838753	0.765474	0.773604
Recall	1.0	0.931923	0.985982	0.999832	1.0	0.794762	0.887182	0.781247	0.883740	0.883237	0.931419	0.881642	0.882565
Precision	1.0	0.720067	0.991810	0.995237	1.0	0.838099	0.760688	0.794587	0.783042	0.787988	0.843482	0.791127	0.799361
F1	1.0	0.812411	0.988887	0.997529	1.0	0.815855	0.819080	0.787861	0.830349	0.832898	0.885272	0.833935	0.838905

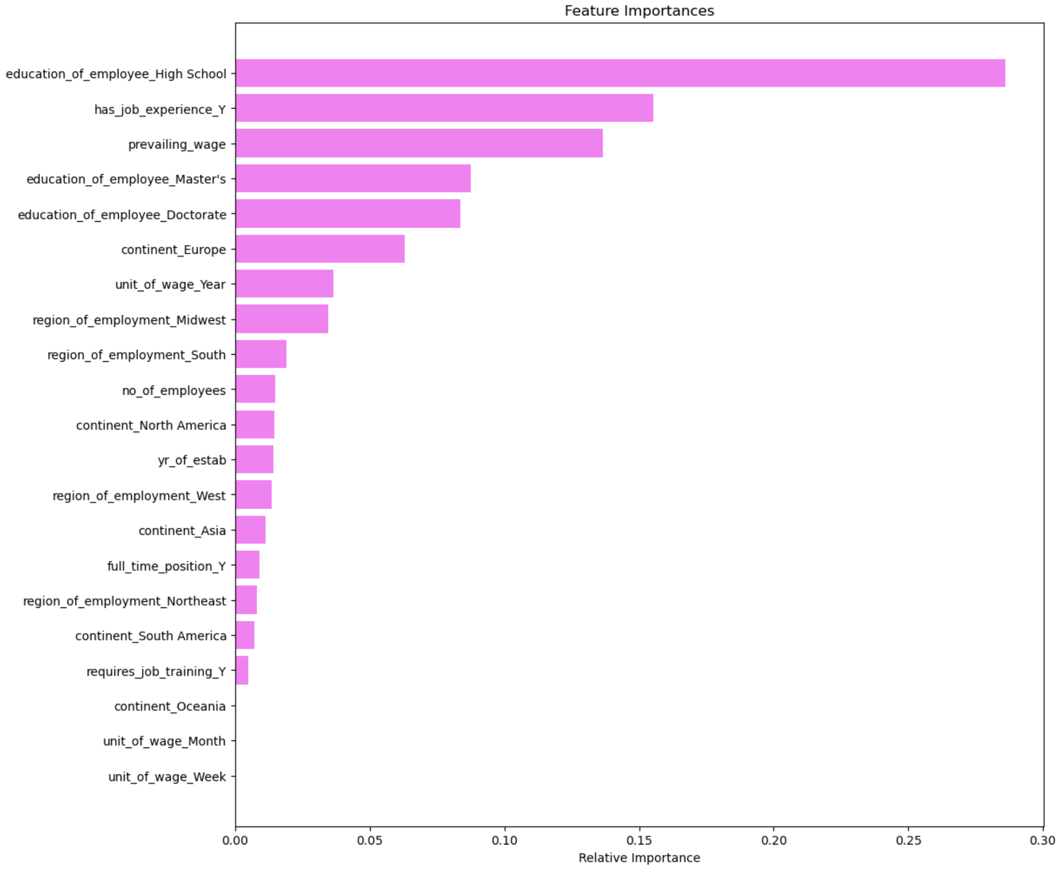
Testing performance comparison



	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.664835	0.706567	0.691523	0.731293	0.727368	0.718995	0.734301	0.716510	0.744767	0.745421	0.733255	0.745160	0.745290
Recall	0.742801	0.930852	0.764153	0.864251	0.847209	0.763761	0.885015	0.781391	0.876004	0.873066	0.860725	0.869540	0.864838
Precision	0.752232	0.715447	0.771711	0.764247	0.768343	0.805412	0.757799	0.791468	0.772366	0.774457	0.767913	0.775913	0.778385
F1	0.747487	0.809058	0.767913	0.811179	0.805851	0.784034	0.816481	0.786397	0.820927	0.820810	0.811675	0.820063	0.819337

Model Performance Summary

Important features of the final model



Model Performance Summary

Recommendations

•To prioritize limited resources towards screening a batch of applications for those most likely to be approved, the OFLC can:

- Sort applications by level of education and review the higher levels of education first.
- Sort applications by previous job experience and review those with experience first.
- Divide applications for jobs into those with an hourly wage and those with an annual wage, sort each group by the prevailing wage, then review applications for salaried jobs first from highest to lowest wage.

Model Performance Summary

•As stated previously, the Gradient Boosting classifier performs the best of all the models created. However, as shown above, the tuned Decision-Tree model performs barely worse by F1 score and is a far simpler model. This model may be preferable if post-hoc explanations of OFLC decision-making is expected to be required.

- Furthermore, OFLC should examine more thoroughly why whether an application will be certified or denied can be very well predicted through just three nodes as shown above.
- For those in less skilled, entry-level, and/or hourly jobs, the system would appear to be biased against these applications being certified.

Happy Learning !