EasyVisa – Streamlining the USA work visa approval process

Machine Learning(Ensemble Techniques – Bagging/Random Forests/Boosting)

Arpan Dinesh

06/16/2023

Contents

- Executive Summary
- Business Problem & Solution Approach
- EDA & Data Cleaning
- Bagging & Random Forests
- Boosting – AdaBoost, Gradient Boost & XG Boost
- Stacking
- Final Model Selection
- Business Recommendations

Executive Summary

The Office of Foreign Labor Certification (OFLC) faces a pressing challenge due to the increasing number of visa applicants in the United States, resulting in a time-consuming case review process.

To address this issue, EasyVisa, a contracted firm, has developed a data-driven solution utilizing machine learning methodologies. By analyzing the provided data and constructing a classification model, EasyVisa aims to streamline the visa approval process by identifying candidates with a higher probability of obtaining a visa.

Key factors such as prevailing wage, education, and job experience are vital in determining visa outcomes. EasyVisa recommends prioritizing these criteria, considering company stability and capacity, and focusing on relevant qualifications and skills rather than job training, employment status or region of employment as factors.

By aligning feature weightage with OFLC's priorities, EasyVisa can enhance decision-making accuracy and offer actionable insights to optimize visa approval outcomes.

Business Problem

The Office of Foreign Labor Certification (OFLC) faces the challenge of a growing number of visa applicants in the United States, resulting in a time-consuming case review process.

To address this, EasyVisa, a hired firm, seeks to develop a data-driven solution using machine learning to analyze provided data and build a classification model. This model will assist in identifying applicants with a higher likelihood of visa approval.

By pinpointing the key factors that significantly influence case status, suitable profiles can be recommended for visa certification or denial, streamlining the OFLC's visa approval process.

Solution Approach

Exploratory Data Analysis: Data Overview, Univariate & Bi-Variate Analysis.

↓

Data Cleaning: Outlier Check and Treatment, Missing/Duplicate Value Check, Feature Engineering.

↓

Bagging: Decision Trees, Bagging Classifiers, Random Forests, Model Performance Evaluation, Model Tuning

↓

Boosting: AdaBoosting, Gradient Boosting, XG Boosting, Stacking, Model Performance Evaluation, Model Tuning

↓

Model Selection: Model Evaluation, Important Features

↓

Business Recommendations: Actionable Insights, Profitable Suggestions.

Data Overview

- The data contains different attributes of an employee and the employer, explored further during EDA.

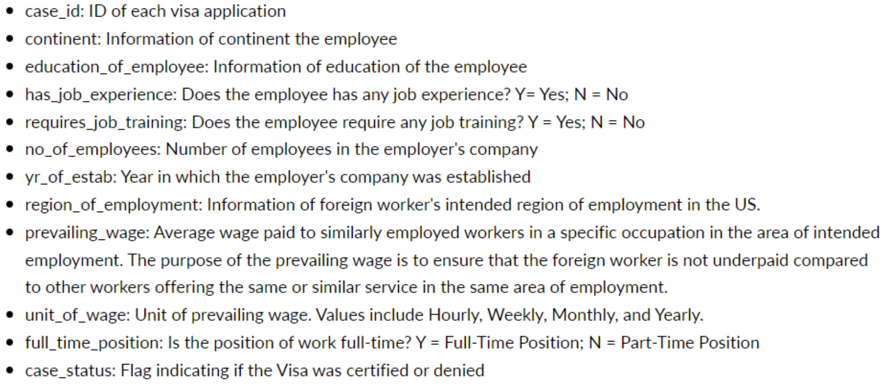
- Each feature is described on the figure to the right.

- The top 5 rows of data can be seen below. There are 12 fields and 25482 records.

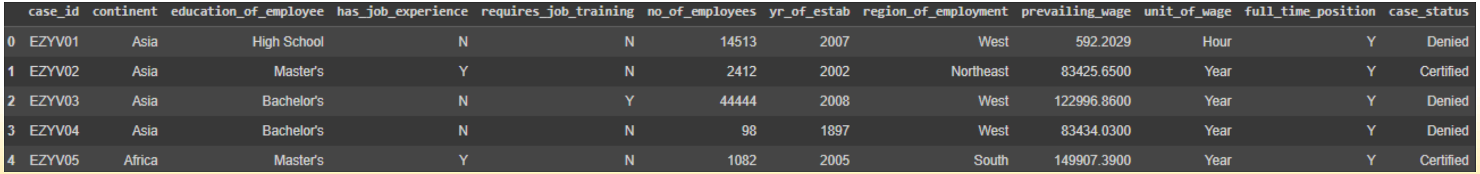
- Only # of employees, year company was established and the wage are of a non object data type.

- There are no duplicated or missing values in the dataset.

- The # of employees field has 33 negative values which needs to be treated.



- 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



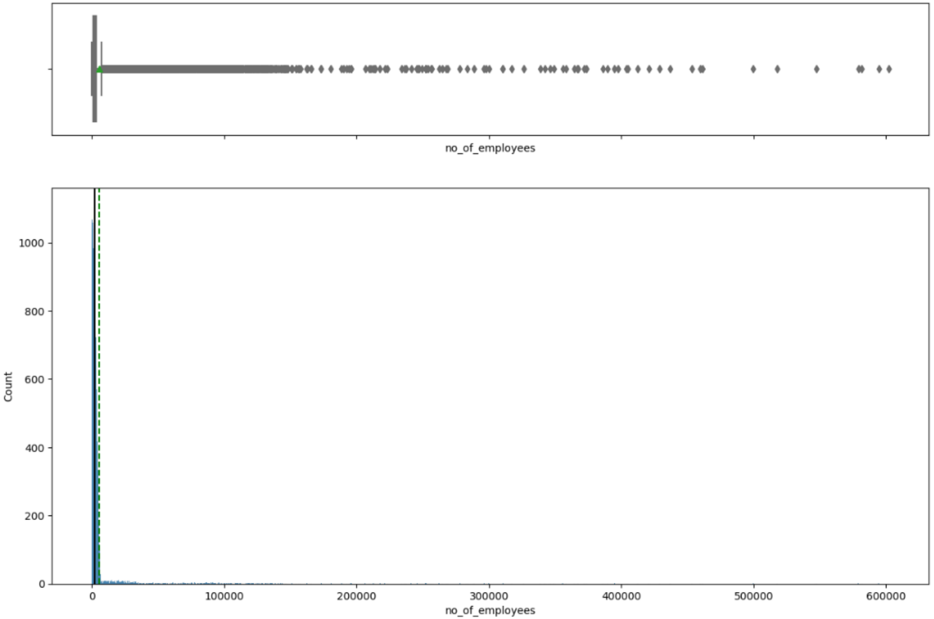
	case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage	unit_of_wage	full_time_position	case_status
0	EZYV01	Asia	High School	N	N	14513	2007	West	592.2029	Hour	Y	Denied
1	EZYV02	Asia	Master's	Y	N	2412	2002	Northeast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

EDA - Univariate(Number of Employees)

- The negative values in the # of employees tend to be a typing error, so they are replaced by their absolute values.

- Expectedly, there is a large variation in the number of employees considering the large sample size of different companies across the United States.

- The outliers are all real data as large corporations have a very large employee count, so they will not be treated.



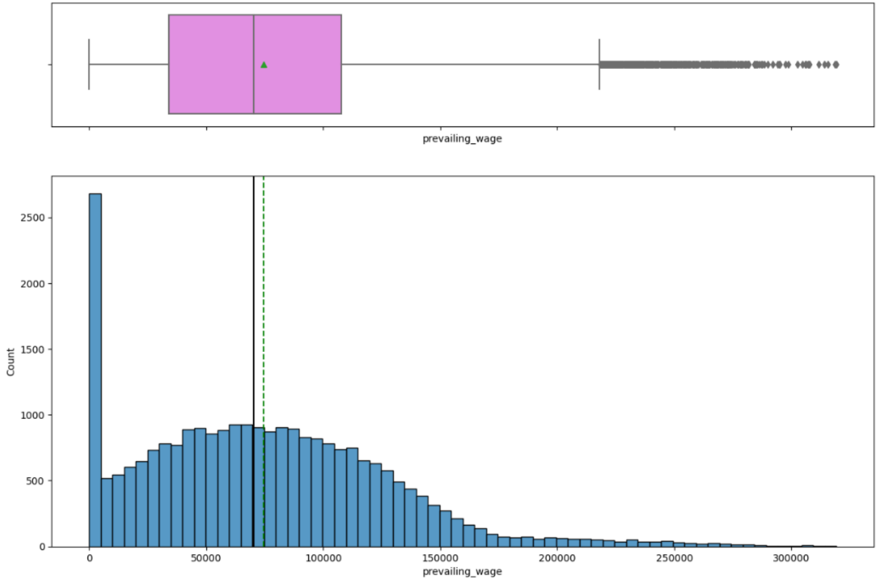
EDA - Univariate(Prevailing Wage)

- 176 of prevailing wage values are on an hourly basis, explaining the <$100 prevailing wages in the data.

- The median and mean are relatively close to each other at around $75,000/year.

- Many small corporations also apply for work visas as seen by the extremely low hourly prevailing wage.

- All outliers are real data considering some regions and corporations set high wages for certain type of work visas.



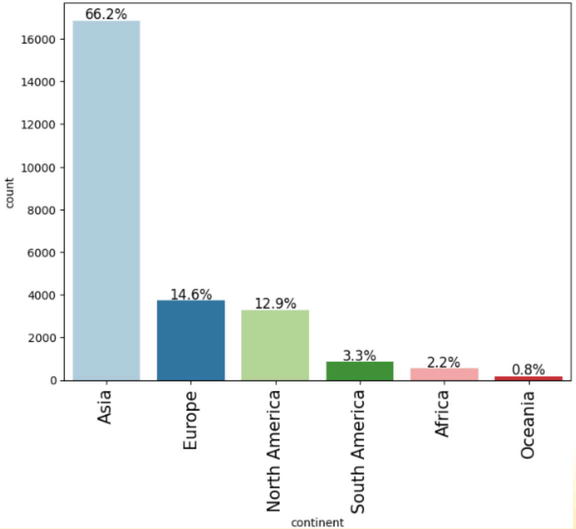
EDA - Univariate(Employee and Employer Region)

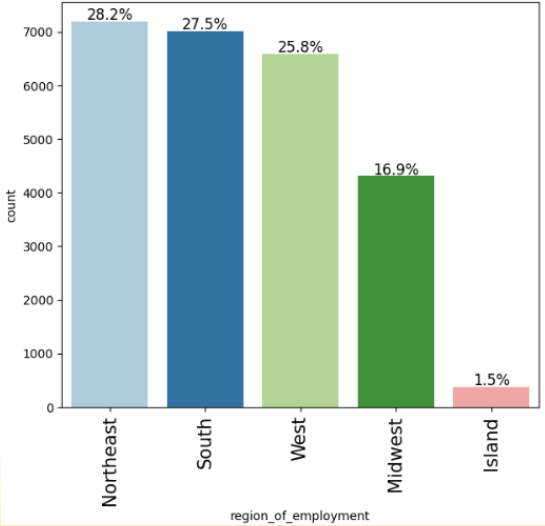
- Almost two thirds of all visa applicants come from Asia, due to the large immigrant populations of India and China.

- Surprisingly, there are a similar number of applicants from Europe and North America(Outside the USA).

- The employees intended employment region is equally split across the USA.

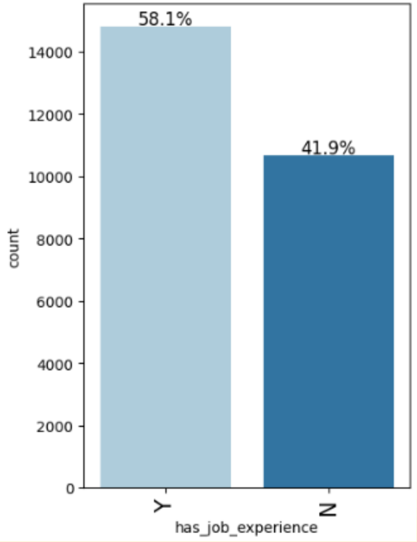
- More than 75% of applicants intend to work in the Northeast, South or West regions of the USA.



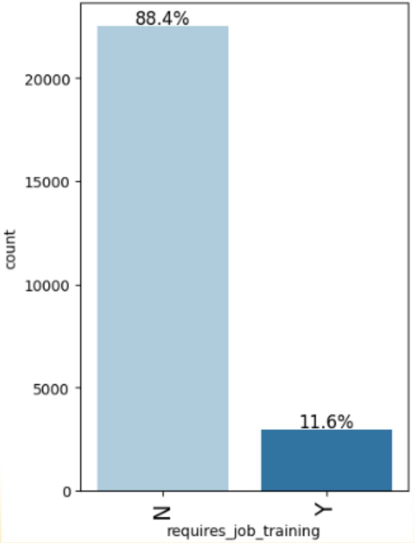


EDA - Univariate(Job Experience and Training)

- A high number of applicants do not have any job experience.
- 58% of applicants do have some job experience.



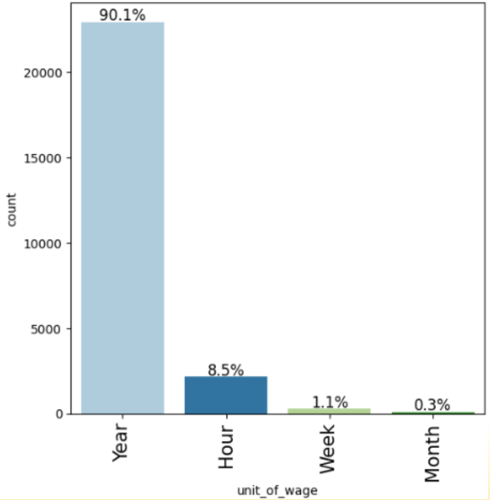
- A large number of applicants do not require any job training.
- Only 11.6% applicants require job training.

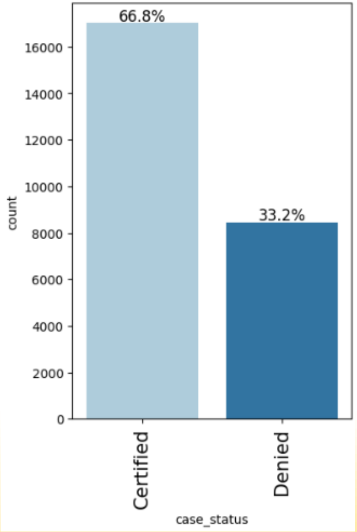


EDA - Univariate(Unit of Wage & Case Status)

- 90% of prevailing wages are in annual salaries.
- 8% of prevailing wages are hourly rates.

- Two thirds of applicants have been certified.
- 33% of applications have been rejected.





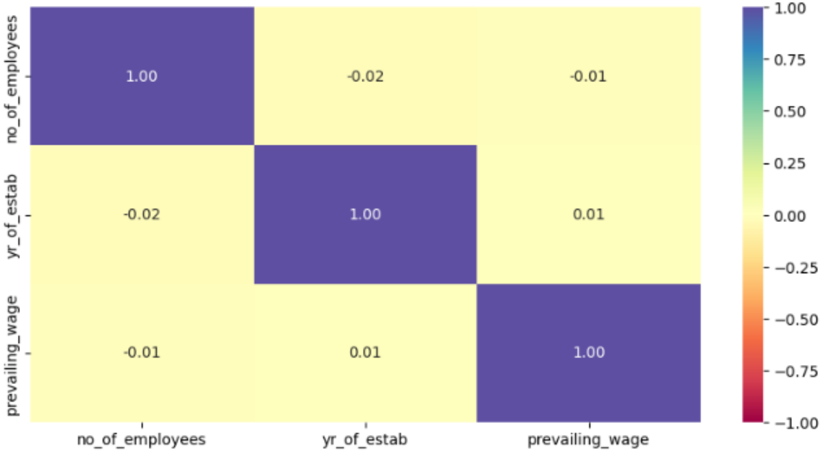
EDA – Bivariate(Correlation)

- None of the numerical variables a correlation with each other.

- The size of the company in terms of number of employees is not correlated to the prevailing wage.

- Older, more established corporations have no correlation to number of employees.

- The prevailing wage is generally based on the current economic status so is not affected by when a company was established.



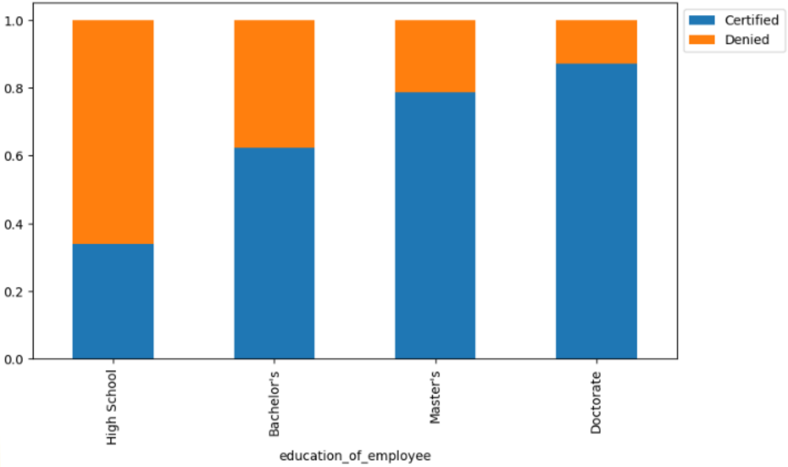
EDA - Bivariate(Education Level and Case Status)

- Expectedly, applicants without at least a Bachelor’s degree have less than a 40% chance of being selected.

- applicants with at least a Bachelor’s degree have more than a 60% chance of being selected.

- The chances go up to almost 80% if applicants hold a master’s degree.

- Not surprisingly, applicants with a doctorate have the greatest chance of getting certified: 90%.



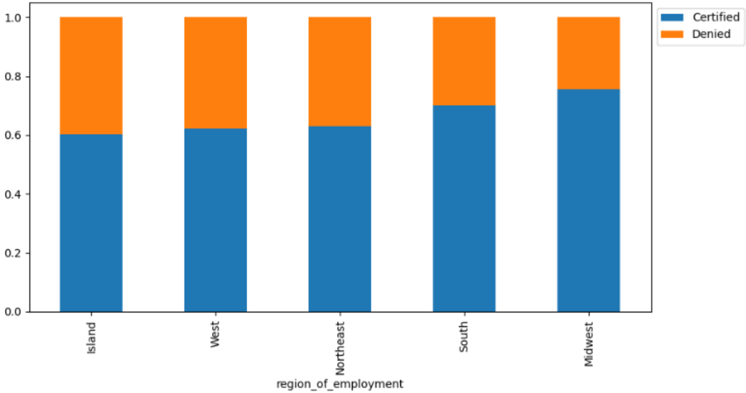
EDA - Bivariate(Visa Certifications Across Regions)

- The Midwest sees the largest fraction of certified applicants: +70%.

- Amongst regions with a large sample size, the South region has the highest fraction of certified applicants.

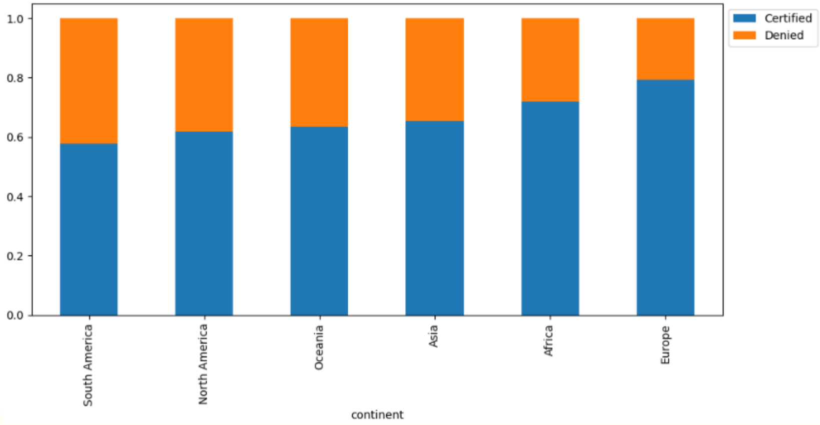
- Amongst regions with a large sample size, the South region has the highest fraction of certified applicants.

- The West and East coasts cannot be separated with fraction of certified applicants: both have around 63%.



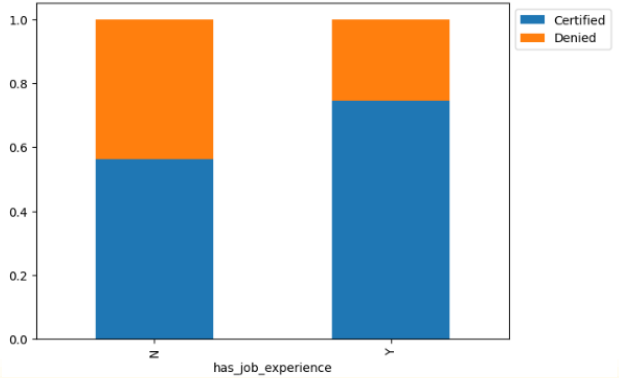
EDA - Bivariate(Visa Certifications Across Continents)

- Albeit a small sample size, applicants from Europe and Africa have the largest fraction of certified applicants.
- More than 40% of applicants from South America have had their visas denied.
- More than 75% of applicants from Europe have had their visas certified.
- Asia has a >60% certification rate even with a large sample size of applications.

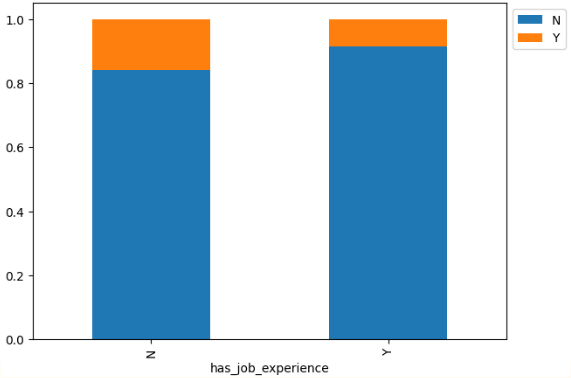


EDA - Bivariate(Job Experience on Case Status and Training)

- Chances of getting a visa certified is 20% higher if the applicant has job experience.
- Surprisingly, more than 50% of applicants with no job experience had certified visas.



- Job Experience does not really affect the requirement for job training.
- Only roughly 20% of applicants who did not have job experience required job training.



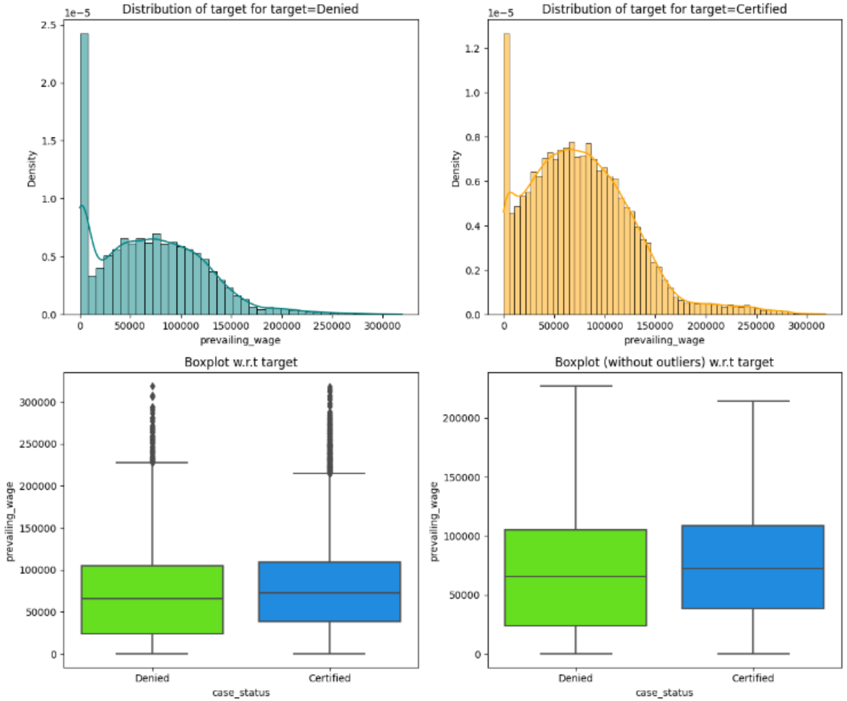
EDA – Bivariate(Prevailing Wage on Case Status)

- The median of the prevailing wages of certified visas is slightly larger than denied visas.

- Denied visas have a larger variation than certified visas.

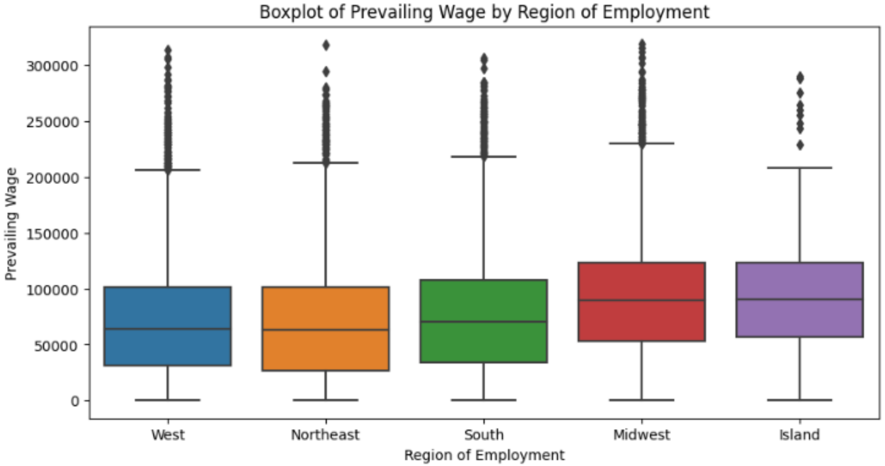
- Certified visas have more outliers than denied visas with some extremely high wages.

- Distributions of prevailing wage for certified and denied visas are rightly skewed due to outliers on the high end.



EDA – Bivariate(Prevailing Wage by Region of Employment)

- The Island and Midwest regions have the highest median of prevailing wages.
- All regions have a similar variation in prevailing wages and outliers on the higher end.
- The West and Northeast regions have the lowest median of prevailing wages.
- Generally, Prevailing Wages do not seem to vary much between the different regions of employment.



Data Cleaning & Feature Engineering

- Outliers can be noticed in many variables, but they will be left untreated as they are all real data.
- There are no missing or duplicate values in the dataset.
- The negative values in the number of employees column are converted to their absolute values considering the high probability of a typo error.
- The prevailing wages are standardized to simplify modelling: all wages in an hourly, weekly, or monthly rate are converted to an annual rate by multiplying the wages by 2080, 52 and 12 respectively.
- The Unit of Wage column is dropped since all wages are standardized to an annual salary amount.
- A new feature named ‘Age of Company’ is computed: 2023 – Year of Establishment, and is added to the dataset since it is a more comprehendible feature to study trends.

Data Pre-Processing & Model Evaluation Criterion

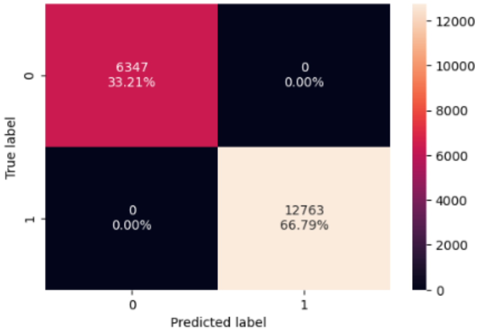
- All categorical variables have been converted to dummy variables with binary units(0 or 1).
- The Case Status(Certified/Denied) is the dependent(response) variable, and is encoded as 1 for certified and 0 for denied.
- Data is split into train and test set with a 70:30 ratio. The classes in each set are very similar(67% certified & 33% denied).
- False Negatives and False Positives are important: If a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work on that position. If a visa is denied when it had to be certified, companies will lose out on employees.
- The F1 score will be the metric to maximize.

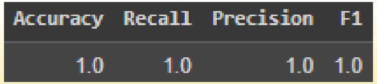
DECISION TREE

- Expectedly, a decision tree allowed to grow to it’s fullest fits the train data perfectly.
- All metric’s are a 100%.

- The tree captures noise as well, resulting in a low F1 score on unseen data.
- The model fits train data perfectly but does not generalize well on test data.

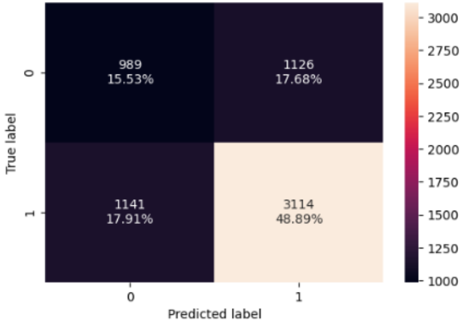
TRAIN

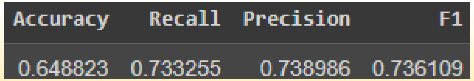




Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0

TEST





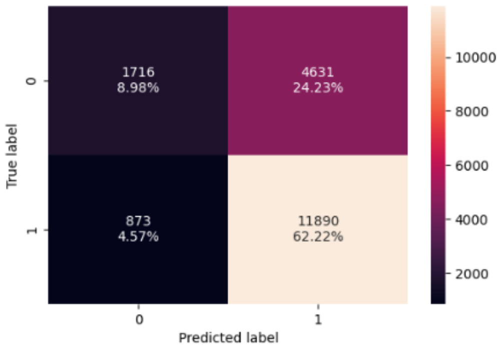
Accuracy	Recall	Precision	F1
0.648823	0.733255	0.738986	0.736109

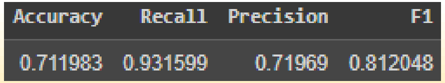
DECISION TREE – Tuned

- The Hypertuned tree generalizes well onto the testing set.
- The F1 score has improved to 0.81 from 0.73 on the test set.

- The tuned tree is not overfitting the training data as much as before.
- Very similar and high F1 scores on both train and test data suggests a good model.

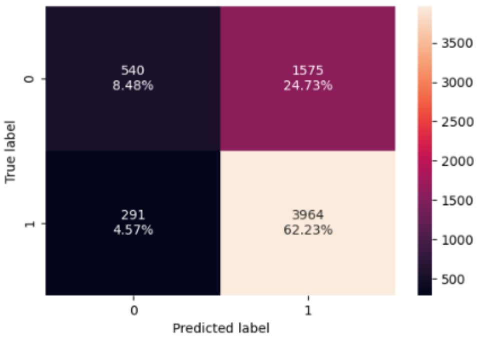
TRAIN

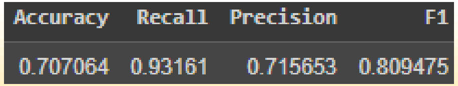




Accuracy	Recall	Precision	F1
0.711983	0.931599	0.71969	0.812048

TEST





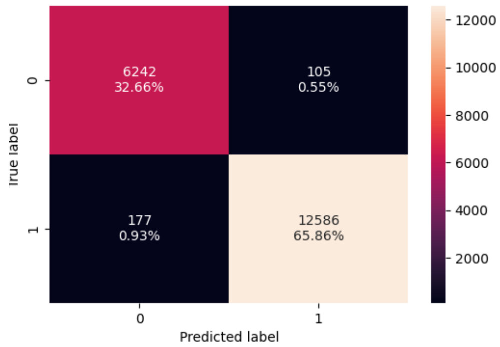
Accuracy	Recall	Precision	F1
0.707064	0.93161	0.715653	0.809475

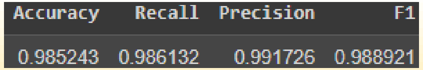
BAGGING CLASSIFIER

- The Bagging Classifier performs extremely well on the training set.

- On Training Data, the Bagging Classifier has a 0.98 F1 Score.

TRAIN



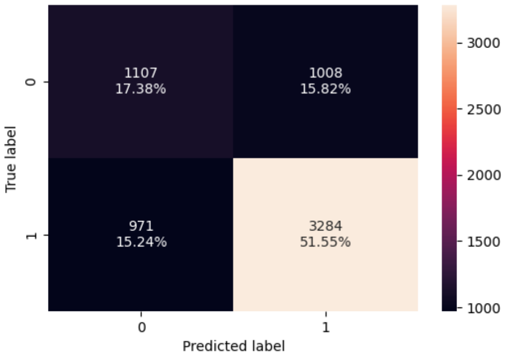


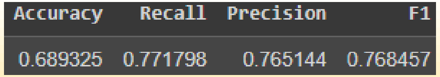
Accuracy	Recall	Precision	F1
0.985243	0.986132	0.991726	0.988921

- The Bagging Classifier does not generalize well on the test set(F1 Score = 0.76).

- The model is definitely overfitting the training data and capturing noise.

TEST





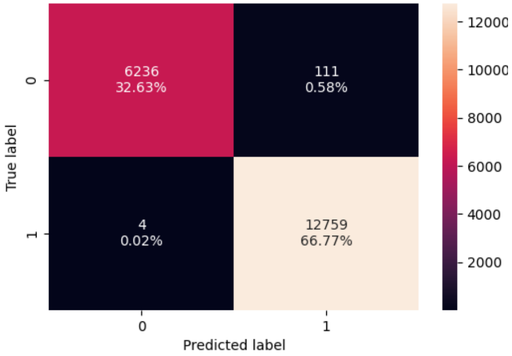
Accuracy	Recall	Precision	F1
0.689325	0.771798	0.765144	0.768457

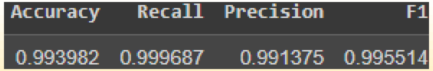
BAGGING CLASSIFIER - Tuned

- The Tuned Bagging Classifier performs better than the non-tuned bagging classifier on the training and testing set with respect to the F1 score.

- The Tuned Bagging Classifier still overfits the data, but generalizes better on unseen data than the non-tuned bagging classifier.

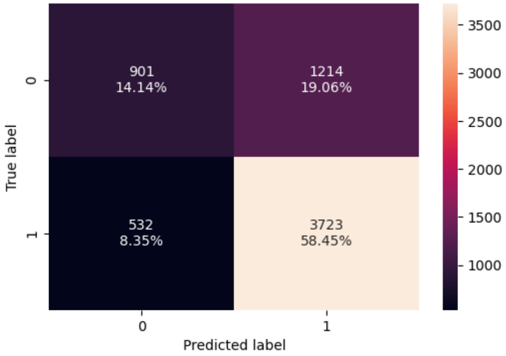
TRAIN

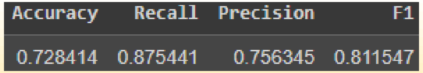




Accuracy	Recall	Precision	F1
0.993982	0.999687	0.991375	0.995514

TEST





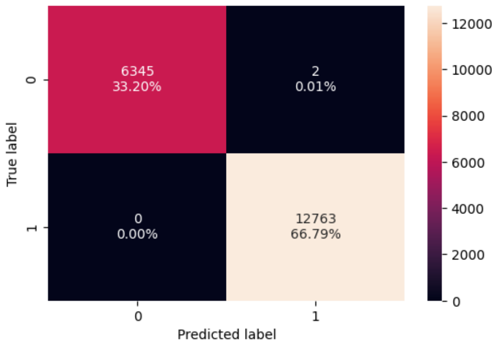
Accuracy	Recall	Precision	F1
0.728414	0.875441	0.756345	0.811547

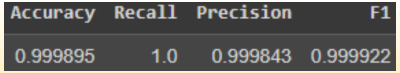
RANDOM FOREST

- The Random Forest has a near perfect score on all metrics on the training set.
- On Training Data, the Random Forest has a 0.99 F1 Score.

- The F1 score on test data is much lower than on training data which suggests the Random Forest is capturing noise.
- The Random Forest metrics need to be analyzed post Hyper tuning.

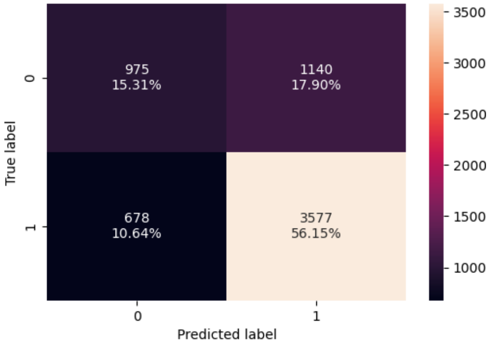
TRAIN

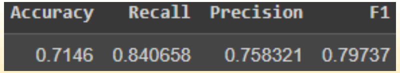




Accuracy	Recall	Precision	F1
0.999895	1.0	0.999843	0.999922

TEST





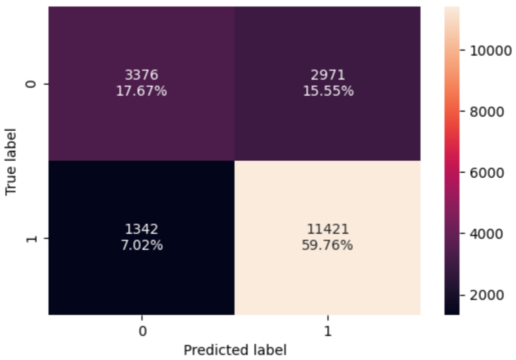
Accuracy	Recall	Precision	F1
0.7146	0.840658	0.758321	0.79737

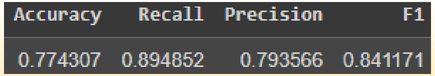
RANDOM FOREST - Tuned

- Tuning has reduced the F1 score on the training data but has increased the F1 score on testing data, suggesting the model is not overfitting the data very much.

- The F1 scores on the train and test data are similar and high, which shows the model is performing well on unseen data as well.

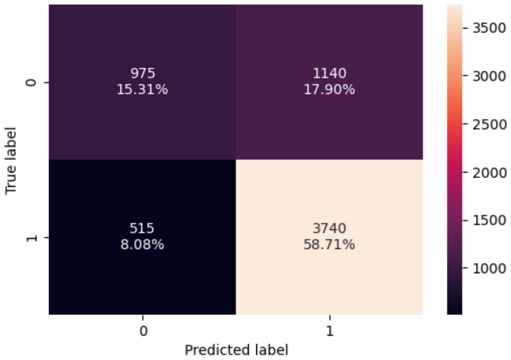
TRAIN

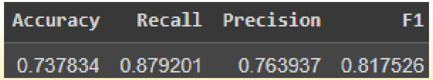




Accuracy	Recall	Precision	F1
0.774307	0.894852	0.793566	0.841171

TEST





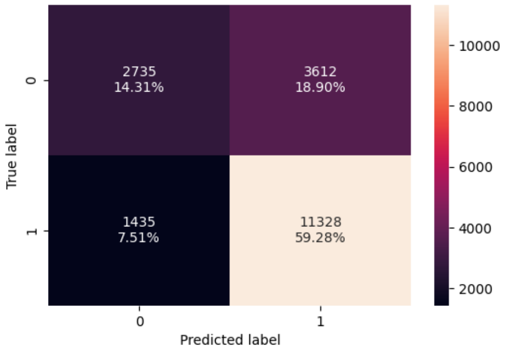
Accuracy	Recall	Precision	F1
0.737834	0.879201	0.763937	0.817526

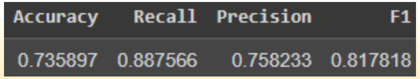
AdaBoost CLASSIFIER

- The AdaBoost Classifier before tuning has an F1 score of 82 on the training data.
- The F1 score could be improved with tuning but it is important to see how it generalizes after tuning as well.

- The AdaBoost Classifier generalizes very well on unseen data with an F1 score of 81.
- All metrics between train and test data are very close indicating an almost perfectly fit model.

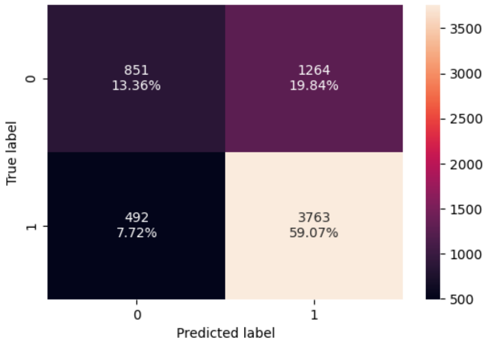
TRAIN

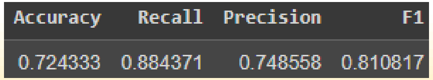




Accuracy	Recall	Precision	F1
0.735897	0.887566	0.758233	0.817818

TEST





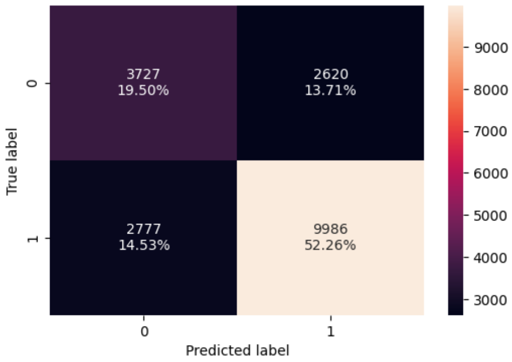
Accuracy	Recall	Precision	F1
0.724333	0.884371	0.748558	0.810817

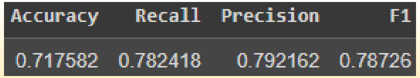
AdaBoost CLASSIFIER - Tuned

- Surprisingly, the tuned AdaBoost Classifier does worse than the non-tuned AdaBoost Classifier on the F1 score on both the training and testing datasets.

- The tuned AdaBoost Classifier generalizes well on unseen data, but again the F1 score is lesser than the non-tuned model on the test data.

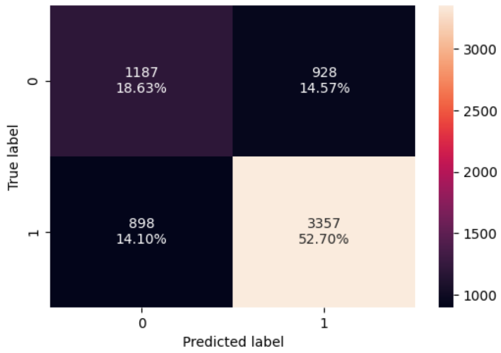
TRAIN

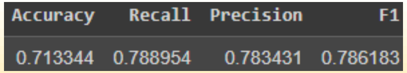




Accuracy	Recall	Precision	F1
0.717582	0.782418	0.792162	0.78726

TEST





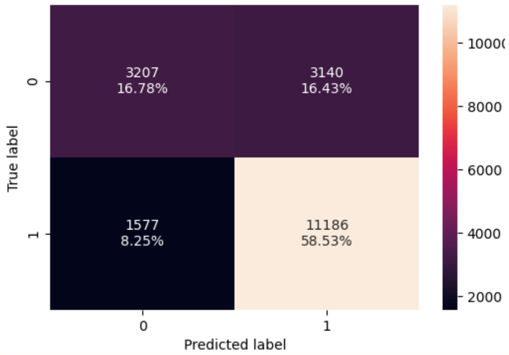
Accuracy	Recall	Precision	F1
0.713344	0.788954	0.783431	0.786183

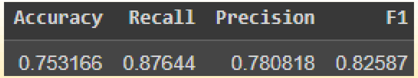
Gradient Boosting

- The Gradient Boosting Model produces an 0.82 F1 score on the training data.
- The accuracy is not that high on training data.

- The Gradient Boosting classifier performs a lot worse on unseen data(F1 score = 0.73), possibly suggesting overfitting by the model.
- Tuning could fix the overfitting problem.

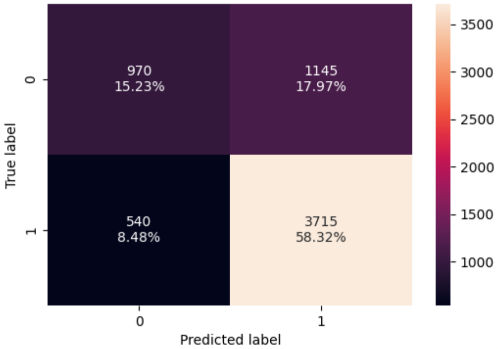
TRAIN

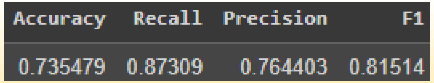




Accuracy	Recall	Precision	F1
0.753166	0.87644	0.780818	0.82587

TEST





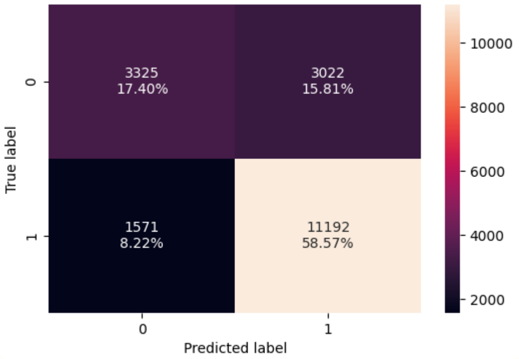
Accuracy	Recall	Precision	F1
0.735479	0.87309	0.764403	0.81514

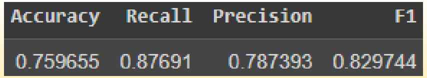
Gradient Boosting - Tuning

- The tuned Gradient Boosting’ performance has not changed on training data but it is generalizing a lot better on unseen data than the non-tuned Gradient Boost.

- The tuned Gradient Boosting classifier performs a lot better on unseen data(F1 score = 0.81) than the non-tuned one , possibly suggesting that tuning got rid of the overfitting and noise.

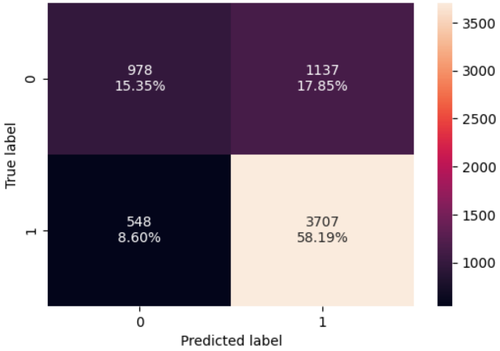
TRAIN

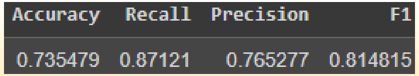




Accuracy	Recall	Precision	F1
0.759655	0.87691	0.787393	0.829744

TEST





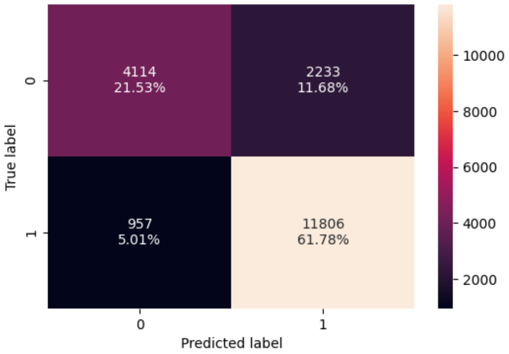
Accuracy	Recall	Precision	F1
0.735479	0.87121	0.765277	0.814815

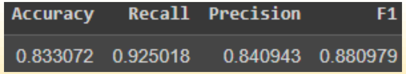
XG Boosting

- The XG Boosted model produces a very high F1 score on the training data(0.88).
- All other metrics are at least 0.8 on the training data.

- On the Test set, the F1 score drops to 0.80 suggesting some overfitting in the model.
- Tuning the XG Boost could reduce the overfitting.

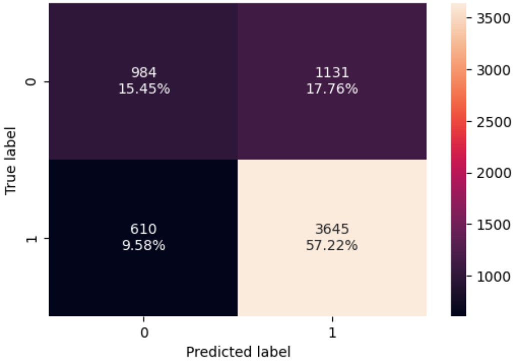
TRAIN

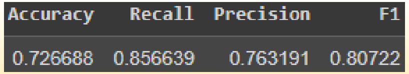




Accuracy	Recall	Precision	F1
0.833072	0.925018	0.840943	0.880979

TEST





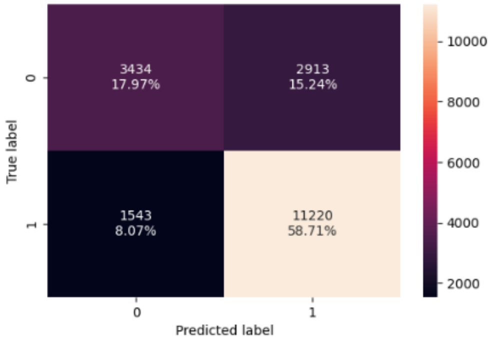
Accuracy	Recall	Precision	F1
0.726688	0.856639	0.763191	0.80722

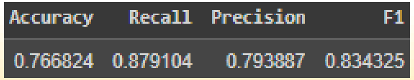
XG Boosting - Tuned

- The tuned XG Boosted model produces a lower F1 score on the training data(0.83) than the non-tuned one(0.88).
- All other metrics have also dropped by roughly 0.05.

- On the Test set, the F1 score does not change much after tuning.
- Tuning the XG Boost has resulted in a closer F1 score between train and test set possibly due to reduction in capturing noise.

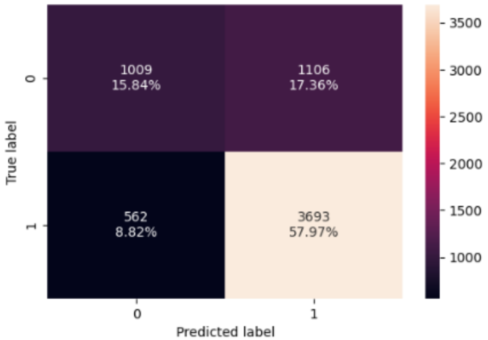
TRAIN

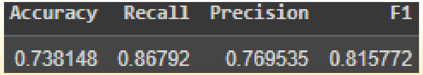




Accuracy	Recall	Precision	F1
0.766824	0.879104	0.793887	0.834325

TEST





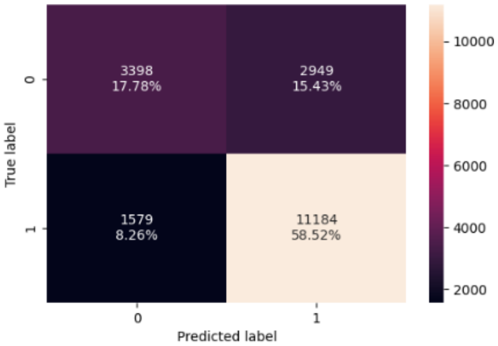
Accuracy	Recall	Precision	F1
0.738148	0.86792	0.769535	0.815772

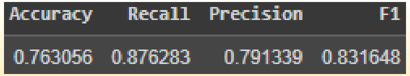
STACKING CLASSIFIER

- The Stacking Classifier has a high F1 score on the training data.
- All other metrics are at least 0.75 on the training data.

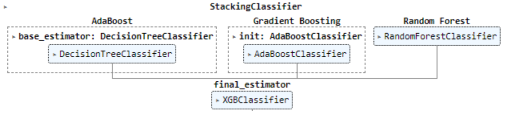
- The Stacking Classifier generalizes well on unseen data: the F1 score drops only by 0.02.
- The base estimator is the tuned Decision Tree Classifier and the final estimator is the tuned XG Boost Classifier.

TRAIN

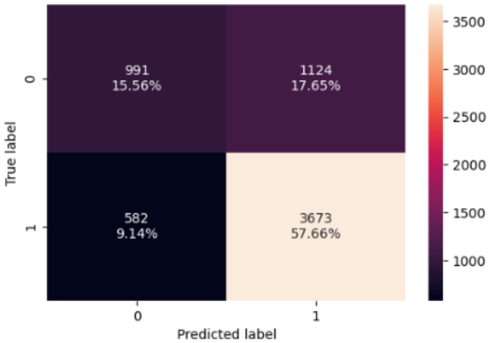


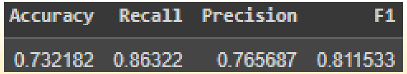


Accuracy	Recall	Precision	F1
0.763056	0.876283	0.791339	0.831648



TEST





Accuracy	Recall	Precision	F1
0.732182	0.86322	0.765687	0.811533

FINAL MODEL SELECTION

- Model Performance(F1 Score) on test data & Generalization (Difference of F1 scores) need to be considered.

- A weight of 0.6 is given to model performance, whilst a weight of 0.4 is given to model generalization. Please note that this is subjective and may be altered based on business application.

- The maximum MPS score will be the best model. Note that to Maximize MPS, Test Set F1 Score needs to be maximized and the Difference of F1 scores needs to be minimized so the equation* makes sense.

- The Maximum MPS is for the Gradient Boost Model, so that may be considered as the best model.*

Training & Testing F1 Score Model Comparison				
		F1 Score		
S.No.	Model	Train	Test	MPS*
1	Decision Tree	1	0.736	0.3361
2	Tuned Decision Tree	0.812	0.809	0.4847
3	Bagging	0.989	0.768	0.3729
4	Tuned Bagging	0.996	0.810	0.4117
5	Random Forest	0.999	0.797	0.3978
6	Tuned Random Forest	0.841	0.819	0.4824
7	AdaBoost	0.818	0.811	0.4836
8	Tuned AdaBoost	0.787	0.786	0.4714
9	Gradient Boost	0.826	0.815	0.4847
10	Tuned Gradient Boost	0.83	0.816	0.4837
11	XG Boost	0.881	0.807	0.4548
12	Tuned XG Boost	0.834	0.816	0.4822
13	Stacking	0.838	0.817	0.4820

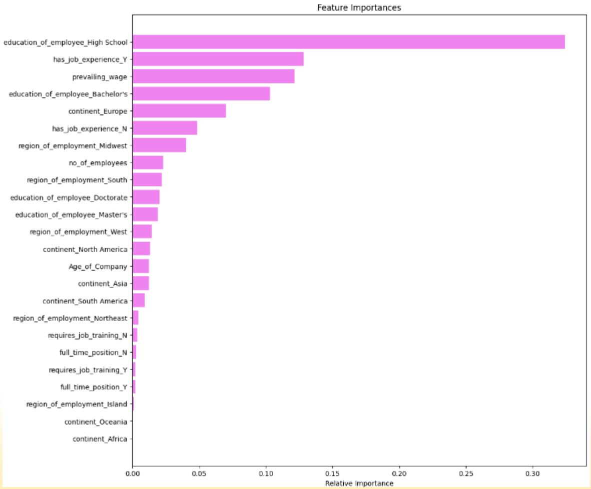
*A new metric MPS(Model Performance Score) is computed:
MPS = 0.6*(Test Set F1 Score) + (-0.4)*(Difference of F1 score between train & test set).

*It is important to note that other model's(Tuned Gradient Boost/AdaBoost) do come very close in terms of performance and generalization, and have a similar processing time so EasyVisa may use them as well.

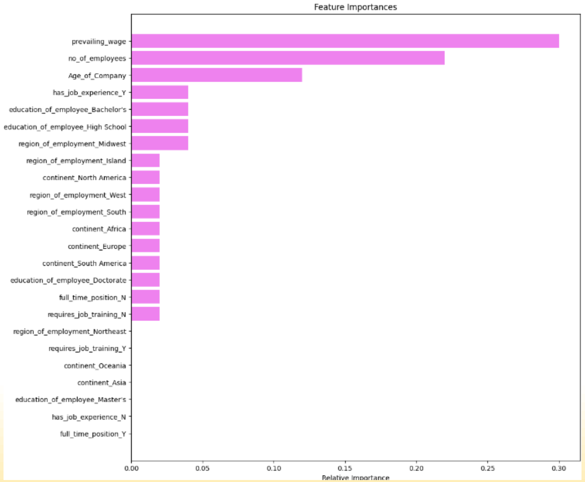
IMPORTANT FEATURES (Gradient Boosting/AdaBoost Models)

- It is important to visualize the important features in two high performing models: Gradient Boosting & AdaBoost Models.
- The prevailing wage, education of an employee and job experience is of high importance in both models.
- Surprisingly, the age of the company and number of employees is very important in the AdaBoost model but not as important in the Gradient Model.
- EasyVisa may also choose the model to use based on which features hold more weightage to them and can be made sense of more easily.

Gradient Boosting



AdaBoosting



BUSINESS RECOMMENDATIONS

- EasyVisa should prioritize prevailing wage, education, and job experience as key criteria for shortlisting candidates, and emphasize relevant qualifications and skills to increase chances of a visa approval.
- EasyVisa must consider company age and number of employees to assess stability and capacity for supporting foreign workers whilst suggesting companies to potential candidates.
- EasyVisa should focus on relevant work experience, skills, and qualifications, and not give much importance to job training whilst judging which profiles have the best chance of visa approval.
- EasyVisa must also evaluate applicants based on qualifications and skills, regardless of full-time or part-time employment status.
- It is highly recommended that EasyVisa assess the feature weightage for model selection, aligning with their and OFLC's priorities to enhance decision-making accuracy.

THANK YOU

Arpan Dinesh