EasyVisa
(Machine Learning based solution for VISA approvals)
Business Presentation

By: Sushma Rao

Contents

Business overview Problem and Solution Approach

- 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 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.
- There was a nine percent increase in the overall number of processed applications from 2016.
- The process of reviewing every case is a tedious task as the number of applicants are increasing every year.
- OFLC has hired the firm Easy Visa for data-driven solutions.

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. Easy Visa must 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 Description

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.
- 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

Observations	Variables
25480	12

There are no missing values .
There are no duplicated values.
The Absolute values of the no of employees are taken to treat the negative values.

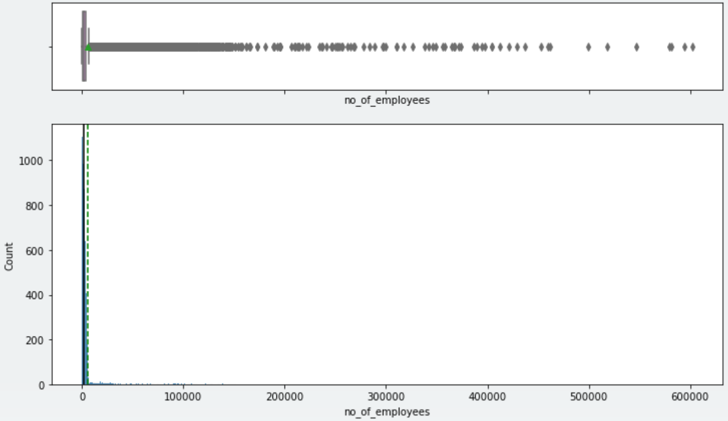
No of employees, years of establishment are of the type integer.

Prevailing wage is of the type float.

Case id, continent, education of the employee, has job experience, requires job training, region of employment, unit of wage ,full time position and case status are of the type object.

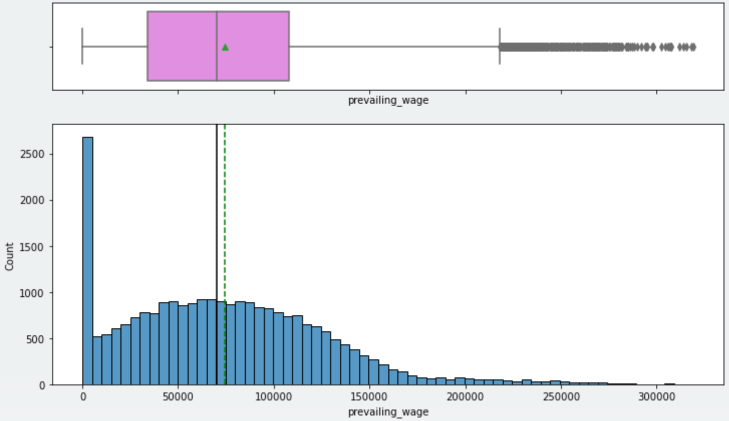
Exploratory Data Analysis-Univariate Analysis

No of employees



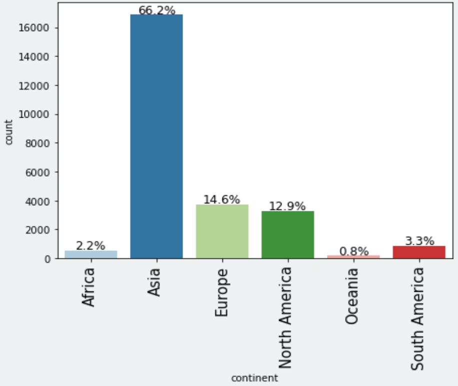
- The “no of employees” attribute is right skewed distribution with a minimum of 11 and maximum of 60,000 plus.
- The mean is greater than the median.
- It is possible to have a large range of employees depending on the size of the hiring companies.
- The graph shows that no of employees above 20,000 are above the 75% quartile range.

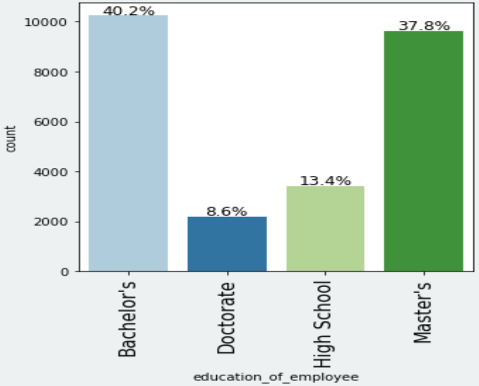
Prevailing wage

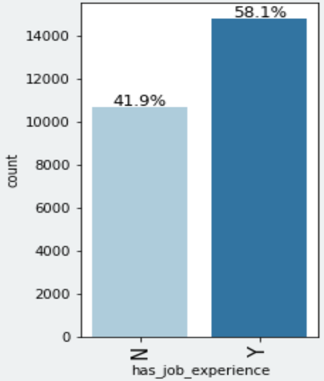


- The prevailing wage is also right skewed distribution.
- We can see there are some significant numbers with minimum wages.
- There are outliers also.
- Mean is greater than the median , but not so far apart.
- If we don’t consider the minimum and outliers, the distribution looks almost normal.

Exploratory Data Analysis-Univariate Analysis







Continent

- Asia has the maximum number of applicants followed by Europe and North America. Nearly 66.2% of the application are from Asia.
- Oceania has the least of 0.8%.

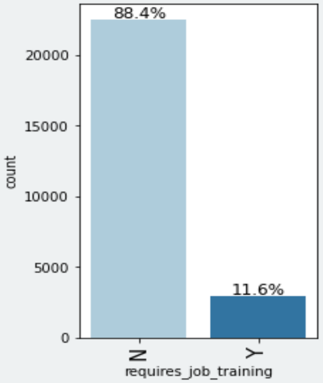
Education of the employees.

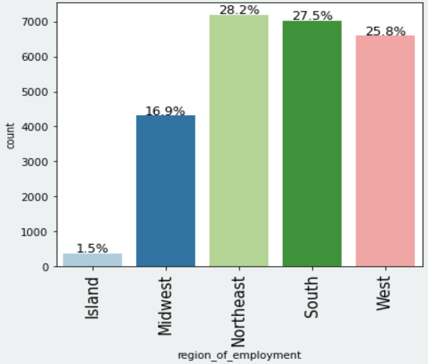
- Maximum number of employees have a Bachelors degree (40.2%) followed by Masters (37.8%).
- There are only 8.6% of the employees with a Doctorate.
- 13.4% of them have education of high school.

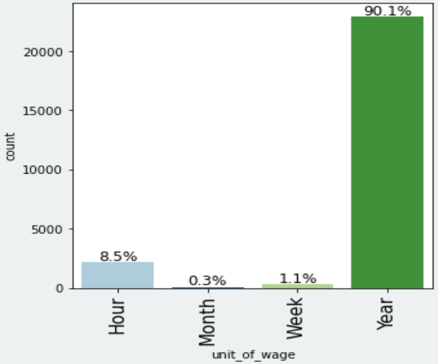
Has job experience

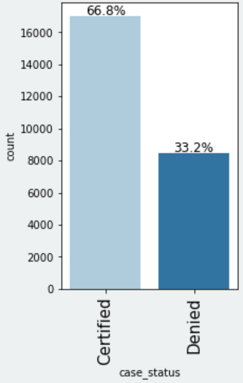
- Maximum number of people have job experience .
- Nearly 59% of the employees who apply for VISA have job experience .
- The number of employees who don’t have job experience is 42% , so there is significantly no much difference between the 2 categories.

Exploratory Data Analysis-Univariate Analysis









Requires job training

- Nearly 88% of the employees don’t need any job training. This can be an important factor for the employer to make decision.
- Only a least of 11.6% need job training.

Region of employment

- Northeast, south and west have the maximum number of employments with northeast being the maximum at 28.2%.
- Midwest has decent number of employments of nearly 17%.
- Islands have the least of 1.5%.

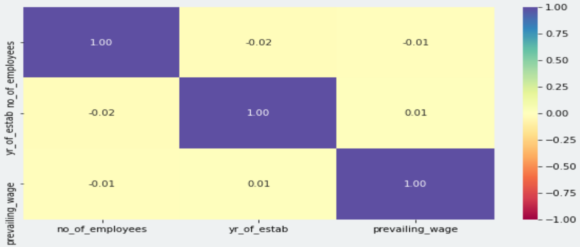
Unit of wage

- Maximum number of employees have yearly wages .
- 90% of employees get an annual package .
- 8.5% of them work on hourly basis.
- Month and week , unit of wage is the least .

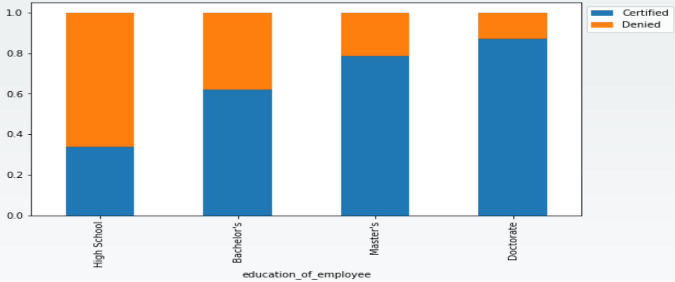
Case Status

- 66.8% of the cases get certified.
- Around 33.2% of the cases get denied.

Bivariate Analysis

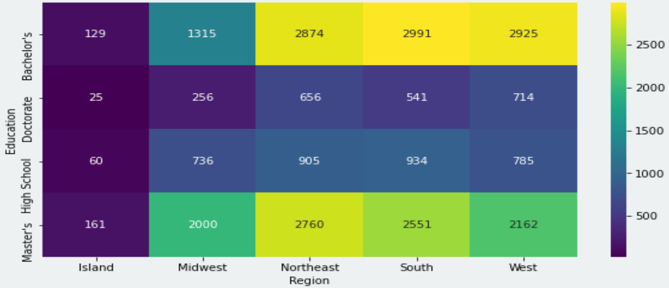


There is no correlation between the numerical columns.no of employees, year of establishment , prevailing wages are not significantly dependent on each other.



- Employees with a Doctorate degree have more chances of getting their VISA followed by Masters and Bachelors.
- Employees with high school education have the least chance of getting the Visa.

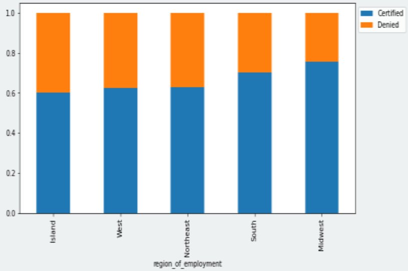
Different regions have different requirements of talent having diverse educational backgrounds. Bivariate Analysis of Education and Region of employment.



- Northeast, South and West have a higher requirement of employees with Bachelor’s degree followed by Masters.
- Midwest seem to require more employees with Masters than Bachelor’s.
- Requirements for Doctorate and high school are in the moderate range in all the regions .
- Overall requirements are less for Island region. Islands may have less opportunities for all types of education and hence the number of cases are also less.

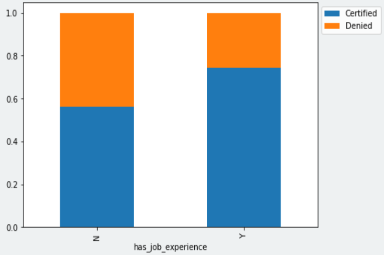
Bivariate Analysis

Analysis of percentage of visa certifications across each region



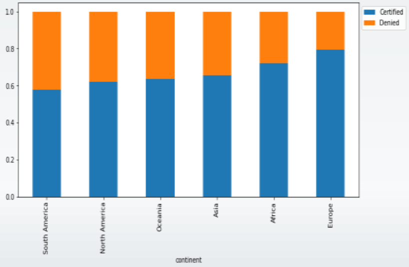
More visas are certified in the Midwest region followed by South and Northeast.

Experienced professionals might look abroad for opportunities to improve their lifestyles and career development. Let's see if having work experience has any influence over visa certification



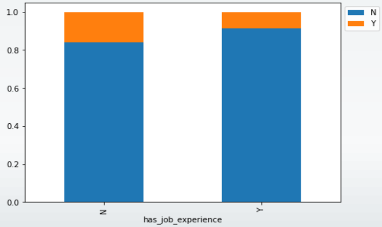
Employees with Experience have higher chances of getting Visa certified compared to the employees with no experience.

Analysis of status of the Visa with different Continents.



- Europe has the highest number of visas certified followed by Africa and Asia.
- Asia has the highest number of cases, but the percentage of visas getting certified is more for Europe.

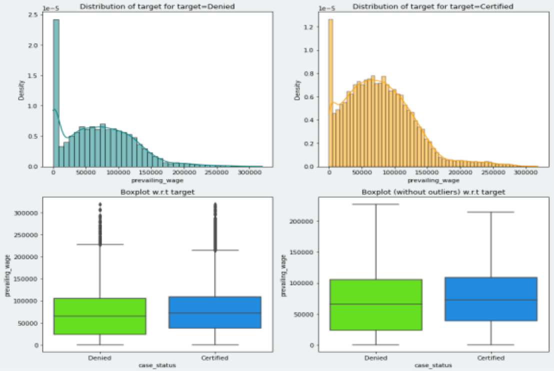
Analyzing to check if the employees who have prior work experience require any job training or not.



Looks like Employees with or without experience does not require job training. Maybe only employees moving into a new field of work may require job training.

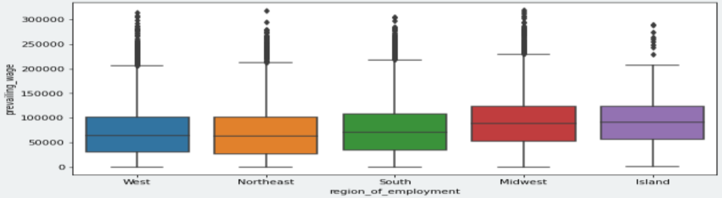
Bivariate Analysis

The US government has established a prevailing wage to protect local talent and foreign workers. Let's analyze the data and see if the visa status changes with the prevailing wage



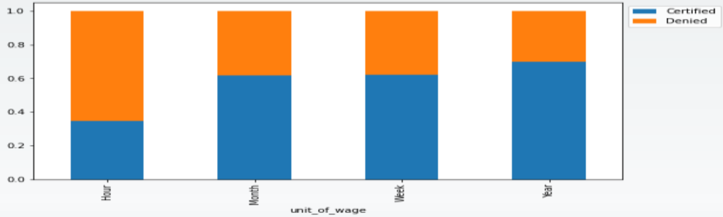
- The chances of getting the visa certified or denied is almost the same.
- As the prevailing wage increases , there is slightly higher chance of getting the visa certified.
- The plot after removing the **outliers** clearly shows that there is a better chance of getting the visa certified with higher prevailing wage.

Checking if the prevailing wage is similar across all the regions of the US



Prevailing wage is higher in the Midwest and Island regions compared to other regions of the US.

The prevailing wage has different units . Let's find out if it has any impact on visa applications getting certified.



There is higher chance of getting visa certified when the prevailing wage has the unit of wage as year compared to hour.

Data preparation for model building

Model evaluation criterion
Model can make wrong predictions as:
1.Model predicts that the visa application will get certified but, the visa application should get denied.
2.Model predicts that the visa application will not get certified but, the visa application should get certified.

Since both the cases are important , we consider **F1 score** to be our evaluation metric , greater the F1 score higher are the chances of minimizing the False positives and False negatives. We use **balanced** class weights so that our model focuses on both the classes.

General steps:

- Check for **the outliers**(we have not done outlier treatment, as the outliers may be essential in effective model building)
- We will encode **categorical** features.
- split the data into **"training**" and "**test**" sets.
- We will build Decision tree without and with **hyperparameter tuning** and check the performance on the test data.

Bagging Ensemble Models

- We will build **Bagging classifier model** with decision tree as the base estimator. Also perform hyperparameter tuning to check the performance.
- We will build **Random forest model** and check the performance with and without hyperparameter tuning .

Boosting Ensemble Models

- We will build **ADA boost model , Gradient Boost Model and XGBOOST models.**
- Perform hyperparameter tuning with different parameters to evaluate the model performance

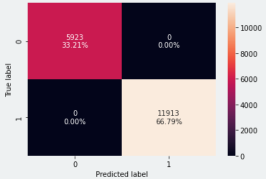
Compare the performances of all the models to decide which model has the **best F1 score**.

Decision Tree Model Evaluation

Model performance of the decision tree before hyperparameter tuning

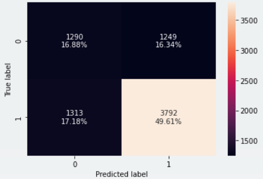
Train data

accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0



Test data

accuracy	Recall	Precision	F1
0.664835	0.742801	0.75223	0.74748

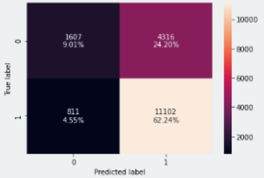


- The decision tree is overfitting on the train data with F1 score 1.
- The performance of the test data is very poor with a F1 score of 0.74748.
- Let's do hyperparameter tuning and check for the model performance.

Model performance of the decision tree after hyperparameter tuning

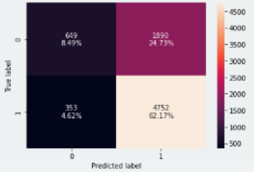
Train data

accuracy	Recall	Precision	F1
0.7125	0.9319	0.7200	0.8124



Test data

accuracy	Recall	Precision	F1
0.7065	0.9308	0.7154	0.8090



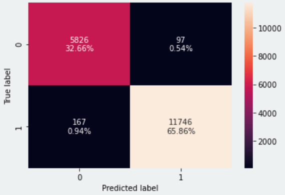
- The F1 score for the train data is decreased, but the performance on the test data is improved to F1 =0.8090.
- Overfitting is reduced.
- Decision tree classifier performs best at the max depth of 10, max leaf nodes of 2, min samples leaf of 3 with min impurity decrease of 0.0001 and class weight "balanced".

Bagging Classifier Model Evaluation

Model performance of the Bagging Classifier before hyperparameter tuning

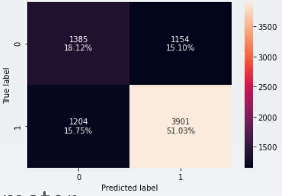
Train Data

accuracy	Recall	Precision	F1
0.9851	0.9859	0.99181	0.9888



Test Data

accuracy	Recall	Precision	F1
0.6915	0.7641	0.7717	0.76791

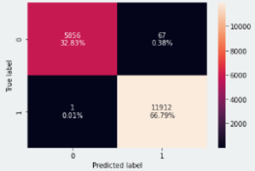


- Decision tree classifier is the base estimator.
- Performance of the Bagging Classifier on the train data is good with F1=0.988, but it is overfitting.
- Performance on the test data clearly shows poor result with F1=0.7679.
- Let’s check if the performance will improve with hyperparameter tuning.

Model performance of the Bagging Classifier after hyperparameter tuning

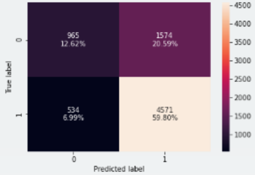
Train Data

accuracy	Recall	Precision	F1
0.9961	0.9999	0.9944	0.997154



Test Data

accuracy	Recall	Precision	F1
0.7242	0.8953	0.74385	0.81262



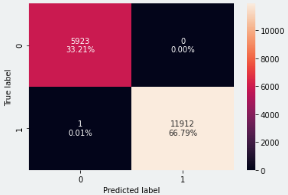
- The Bagging classifier shows best result with max features =0.7, max samples=0.7, estimators=100 and random state 1.
- The performance of the test Sample has increased to F1=0.8126.

Random Forest Classifier Model Performance

Model performance of the Random Forest Classifier before hyperparameter tuning

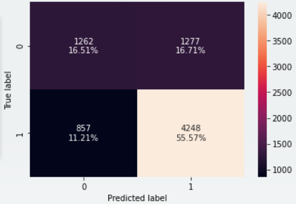
Train data

accuracy	Recall	Precision	F1
0.99994	0.99916	1.0	0.99958



Test data

accuracy	Recall	Precision	F1
0.720827	0.832125	0.768886	0.79924

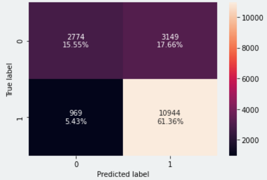


- Random Forest performs well on the Train data with F1=0.99958. The model is overfitting.
- The performance on the Test data is poor with F1=0.79924.
- The confusion Matrix of the test data shows false positives and false negatives.

Model performance of the Random Forest Classifier after hyperparameter tuning

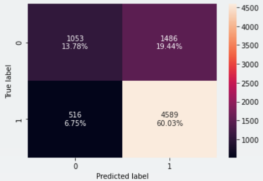
Train data

accuracy	Recall	Precision	F1
0.76911	0.91866	0.77655	0.84165



Test data

accuracy	Recall	Precision	F1
0.73809	0.89892	0.75539	0.82093



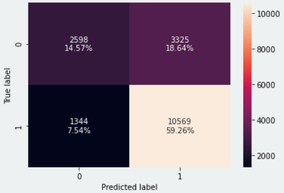
- The model performance on the train data is not overfitting .
- The F1 score of both train and test are almost close to each other.
- Random forest classifier shows the best fit model with max depth =10, max features ='sqrt', min sample split=7, estimators=20, oob score=True and random state 1 .
- Random forests are a strong modeling technique and much more robust than a single decision tree. They **aggregate many decision trees to limit overfitting as well as error due to bias** and therefore yield useful results.

ADABOOST Classifier Model Performance

Model performance of the Ada boost Classifier before hyperparameter tuning

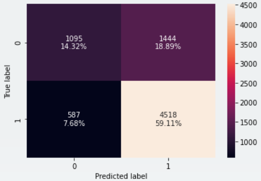
Train data

accuracy	Recall	Precision	F1
0.73822	0.88718	0.76068	0.81908



Test data

accuracy	Recall	Precision	F1
0.73430	0.885015	0.75779	0.81648

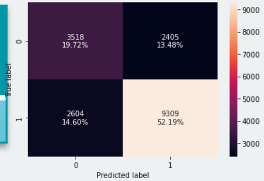


- The performance of the model is better compared to other models, the F1 =0.81648.The model is also not overfitting.
- Performance on both the train and test data is almost close. Let's check if we can improve the F1 score by hyperparameter tuning.

Model performance of the Ada boost Classifier after hyperparameter tuning

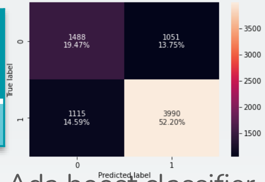
Train data

accuracy	Recall	Precision	F1
0.719163	0.78141	0.79469	0.78799



Test data

accuracy	Recall	Precision	F1
0.71664	0.78158	0.79151	0.78651



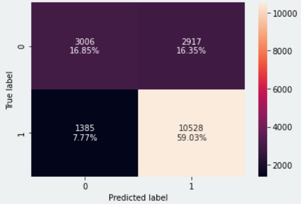
- After running the grid search, the Ada boost classifier shows the best performance with max depth 1, no of estimators=90, class weight =“balanced, learning rate=0.1 and random state =1.
- The F1 score has reduced, and the performance of the model is poor .

Gradient Boosting Classifier Model Performance

Model performance of the Gradient boost Classifier before hyperparameter tuning

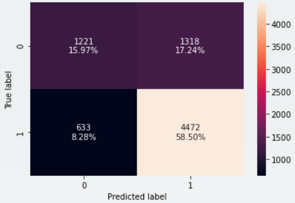
Train data

accuracy	Recall	Precision	F1
0.75880	0.88374	0.783041	0.83034



Test data

accuracy	Recall	Precision	F1
0.744767	0.87600	0.77236	0.82092

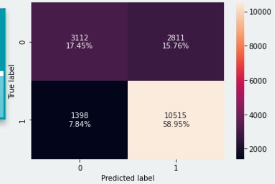


- The Gradient boost Classifier gives a F1 =0.8303 on the train data and F1=0.82092 on the test data, The performance is almost close and not overfitting.
- Let's check if hyperparameter tuning can enhance the model performance.

Model performance of the Gradient boost Classifier after hyperparameter tuning

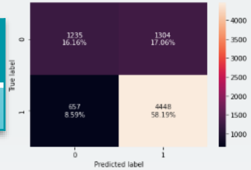
Train data

accuracy	Recall	Precision	F1
0.76401	0.88264	0.78905	0.83323



Test data

accuracy	Recall	Precision	F1
0.74345	0.87130	0.77329	0.81937



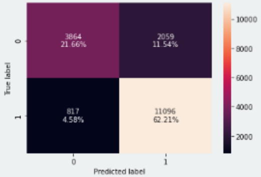
- We have considered ADA boost classifier as our initializer, after the grid search the best models has the max features=0.8, no of estimators=200, subsample=1 and random state=1.
- The F1 score on the test data doesn't seem to improve much with hyperparameter tuning.

XGBOOST Classifier model performance

Model performance of the XGBOOST Classifier before hyperparameter tuning

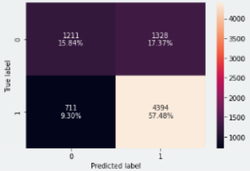
Train data

accuracy	Recall	Precision	F1
0.83875	0.93141	0.843482	0.88527



Test data

accuracy	Recall	Precision	F1
0.73325	0.86072	0.76791	0.81167

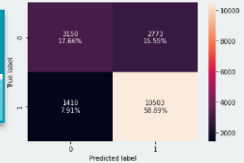


- We have defined the XG boost classifier with random state =1 and evaluation metric as **log loss.**
- The performance on the test data ,F1= 0.8116 is comparatively less compared to that of the train data.
- Let's check the performance after hyperparameter tuning.

Model performance of the XGBOOST Classifier after hyperparameter tuning

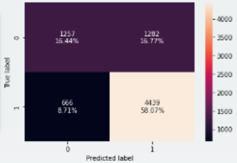
Train data

accuracy	Recall	Precision	F1
0.76547	0.88164	0.791112	0.83393



Test data

accuracy	Recall	Precision	F1
0.74516	0.86954	0.77591	0.820063



- After running the grid search, the tuned xg boost classifier has scale pos weight=1, learning rate=0.1, colsample bytree=0.9, colsample bylevel=1, colsample bynode=1 and random state=1.
- The performance on the test data has definitely increased with F1=0.8200.

Stacking Classifier model performance

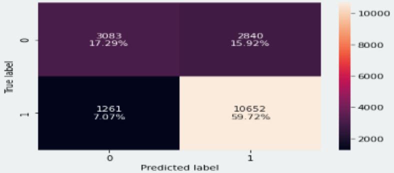
The Stacking classifier was defined using
Ada boost classifier, tuned gradient boost classifier and tuned random boost classifier as estimators was building the model.

Tuned XGBoost model was considered as the final estimator.

- The performance on the train and test data are close and looks like the model is performing good with a F1=0.82128 on the test data.
- The stacking model looks complex with both bagging and boosting classifiers even though the performance is good.
- Let’s compare all the models to check which model can satisfy are business goal.

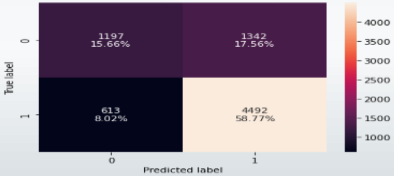
Train data

accuracy	Recall	Precision	F1
0.77007	0.89414	0.78950	0.838575



Test data

accuracy	Recall	Precision	F1
0.74424	0.87992	0.76996	0.82128



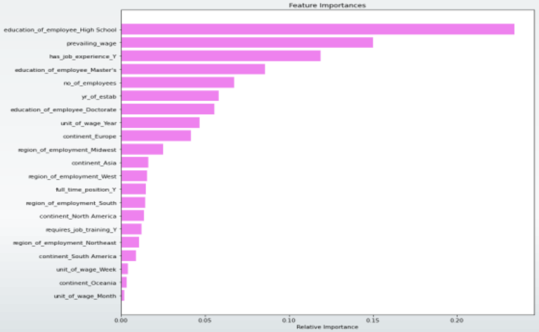
Comparing all the models on the test data

	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	0.706567	0.706567	0.69152	0.72422	0.72082	0.73809	0.734301	0.716641	0.744767	0.743459	0.73325	0.74516	0.74424
Recall	0.930852	0.930852	0.76415	0.89539	0.83212	0.89892	0.885015	0.781587	0.876004	0.871303	0.86072	0.86954	0.87992
Precision	0.715447	0.715447	0.77171	0.74385	0.76886	0.75539	0.757799	0.791510	0.772366	0.773296	0.76791	0.77591	0.76996
F1	0.809058	0.809058	0.76791	0.81262	0.79924	0.82093	0.816481	0.786517	0.820927	0.819379	0.81167	0.82006	0.82128

- Tuned Random forest classifier, gradient boost classifier tuned XG boost classifier and stacking classifier are the models which perform better and has the highest F1 scores.
- Stacking classifier can be considered as our final model, but it can be very time consuming to train. To be effective , we must assemble many base models, which can take a long time. It can be expensive to deploy and maintain.
- Among all the models Tuned Random forest tends to perform best and can be considered as our final model. It is suitable when we have a large dataset like our business problem of VISA approvals

Important features of the final model

Education of the employee – high school , is the most important feature followed by prevailing wage, has job experience-Y and education of the employee- masters are also some of the important features.



Business Insights and Recommendations

As a data scientist at EasyVisa I have analyzed the data provided with the help of a classification model. In order to facilitate the process of visa approvals and to 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.

Based on our analysis ,we can say that the VISA approvals for the employees to work in USA mainly depends on the following features:

- Education of the employee – The feature importance indicates that, the visa can be certified if the education of the employee is high school. May be many work on hourly basis and pursue their graduation, which can be extra revenue for the country.
- Prevailing wage – if the wages are more, then there is more chance of getting the Visa approved. The contribution of such employees may be beneficial to the companies.
- Has job experience- There is more chance of getting the visa certified if the employees have prior experience , companies can hire them without training.
- There are a greater number of applications from employees with Bachelors and Master's degree . Our analysis shows there is more chance of getting the visa approved if one have Masters followed by Doctorate.
- The number of employees in a company and the year of its establishment also plays an important role, may be there is a higher chance of getting the visa approved if the employees apply for a well known and larger companies.
- Yearly wages also play an important role. mostly employees with doctorate, bachelors or masters getting hired do desk jobs with annual salaries and chances of quitting the jobs are less compared to hourly basis jobs.

greatlearning

Power Ahead

Happy Learning !

APPENDIX-Parameters considered for hyperparameter tuning

Decision tree

- "max_depth": np.arange(10, 30, 5),
- "min_samples_leaf": [3, 5, 7],
- "max_leaf_nodes": [2, 3, 5],
- "min_impurity_decrease": [0.0001, 0.001]

Bagging Classifier

- "max_samples": [0.7, 0.8, 0.9]
- "max_features": [0.7, 0.8, 0.9]
- "n_estimators": np.arange(90, 120, 10)

Random Forest

- "max_depth": list(np.arange(5, 15, 5))
- "max_features": ["sqrt", "log2"]
- "min_samples_split": [3, 5, 7]
- "n_estimators": np.arange(10, 40, 10)

Gradient Boost classifier

- "n_estimators": [200, 250, 300],
- "subsample": [0.8, 0.9, 1],
- "max_features": [0.7, 0.8, 0.9, 1],
- "learning_rate": np.arange(0.1, 0.4, 0.1)

AdaBoost Classifier

- "base_estimator": [
- DecisionTreeClassifier(max_depth=1, class_weight="balanced", random_state=1),
- DecisionTreeClassifier(max_depth=2, class_weight="balanced", random_state=1),
- DecisionTreeClassifier(max_depth=3, class_weight="balanced", random_state=1)]
- "n_estimators": np.arange(60, 100, 10)
- "learning_rate": np.arange(0.1, 0.4, 0.1)

XGBoost Classifier

- "n_estimators": np.arange(150, 250, 50)
- "scale_pos_weight": [1, 2]
- "subsample": [0.7, 0.9, 1]
- "learning_rate": np.arange(0.1, 0.4, 0.1)
- "gamma": [1, 3, 5]
- "colsample_bytree": [0.7, 0.8, 0.9]
- "colsample_bylevel": [0.8, 0.9, 1]