Roman Dzadzic

02/04/2022

Presentation to The Data Science Lead

EasyVisa

Contents

We will cover the following topics

- Business Problem Overview
- Data Overview
- Exploratory Data Analysis (EDA)
- About the Model
- Business Insights and Recommendations:
- Appendix: various data points

Business Problem Overview

Context

One of the continued challenges for the United States business communities, who wish to remain competitive, is in identifying and attracting the right talent. Companies in the United States are looking for talented, hard-working, and qualified individuals in U.S. and abroad. The U.S. Immigration and Nationality Act (INA), administered by the Office of Foreign Labor Certification (OFLC), allows foreign workers to come to the United States to work on temporary or permanent basis. And it also protects U.S. workers against negative impacts on their wages or on working conditions by making sure that U.S. businesses are in compliance with statutory requirements when hiring foreign workers to fill workforce shortages. OFLC processes job certification applications for employers seeking to bring foreign workers into U.S., and it grants certifications in those cases where employers can demonstrate that there are not enough skilled workers available to perform the work in U.S. at the wage levels that meet or exceed the wage paid for the occupation in the area of intended employment.

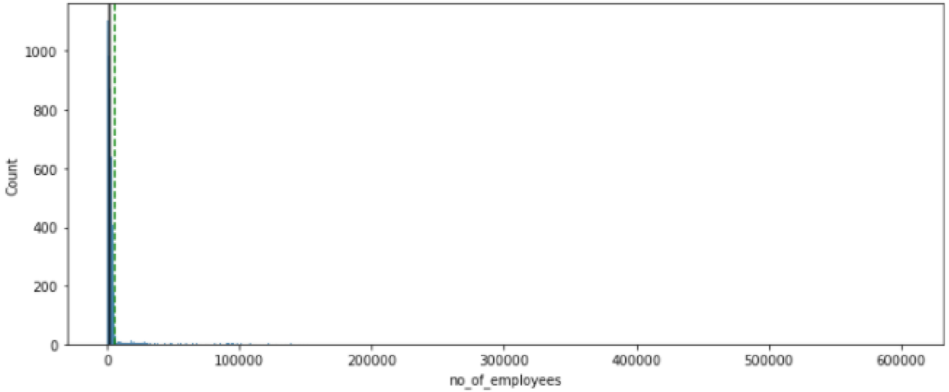
Objective

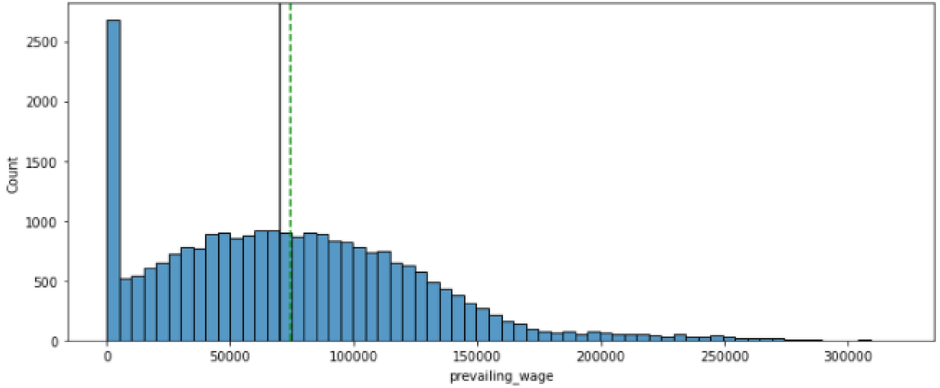
In FY 2016, the OFLC processed 775,979 employer applications for as many as 1,699,957 temp and permanent labor certifications. This was a 9% jump from the previous year. Logically, the certification review process is become more tedious as the number of applications increase each year. This processing bottleneck calls for a Machine Learning solution to help with shortlisting those candidates that have better visa approval chances. OFLC has hired EasyVisa to provide the data science solutions. My job is to analyze the provided data and help with data classification models and framework, which would help to facilitate the visa approvals process, while recommending the best possible applicant profile that is most likely to be certified with the case status approved.

Data Overview

- The dataset has 25,480 rows and 12 columns of data with cases, their statuses, and their various other characteristics/details.
- Case characteristics include data such as case_ID, employee's continent, employee's education, yes/no for employee's job experience, yes/no for whether the employee requires job training, employer size by number of employees, year of company establishment, worker's intended region of employment, prevailing wage, wage unit (e.g., hourly, weekly, monthly or annual), position's full-time status (full-time or part-time), and most importantly the case status for whether the visa is certified or denied.
- Data types are 3 numeric (1 float & 2 integer), and 9 object data points.
- No discrepancies in the available records for each of the columns, so no missing values, and there are no duplicate values.
- The most requested region of employment is Northeast, and most of the positions are for full-time.
- Year is the most common wage unit. Wage statistics are harder to read right off the bat, because of different wage units. However, the min wage is 2.13 units, while max value is as high as 319,210 units.
- Oldest established business is as old as 1800 while the most recent year of establishment is 2016.
- Lastly, the employer size by the number of employees is heavily right skewed with min of -26 employees (likely negative because those job positions need to be filled) all the way to over 602k employees.

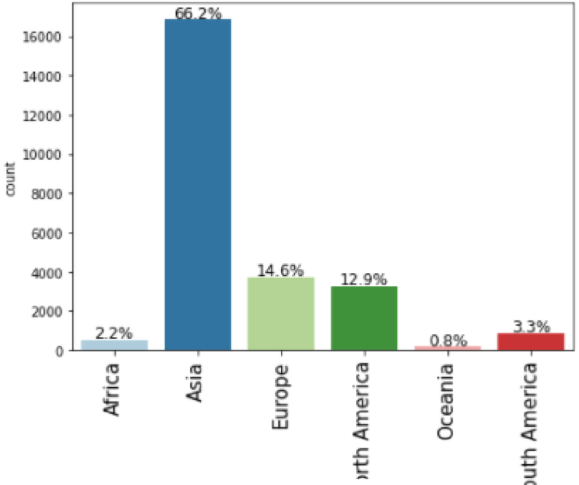
Exploratory Data Analysis

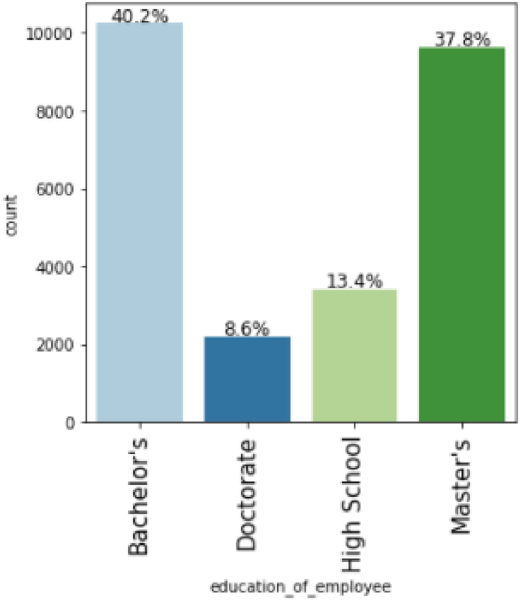


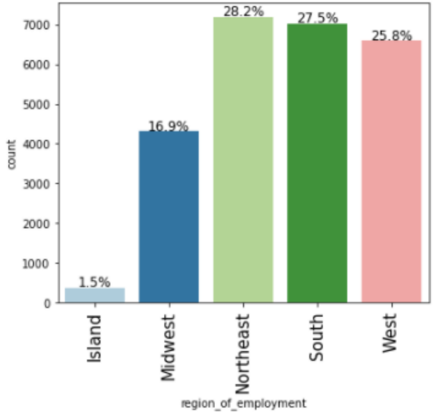


- Majority of companies have smaller number of employees – heavily right-skewed data. Only a few companies have close to 600k. However, from the outlier perspective, I think it's important to keep all this data to realize the impact in further data observations and analysis.
- Prevailing wage seems to be right skewed with mean around $75k
- One exception is the very large count of wages below $100. In further data analysis, it becomes apparent that large count of wages below $100 can be attributed to the unite_of_wage = hour, which is why the data looks offset at that wage

Exploratory Data Analysis

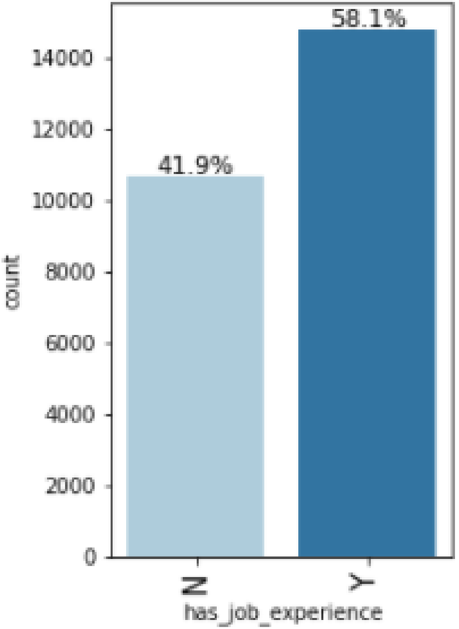


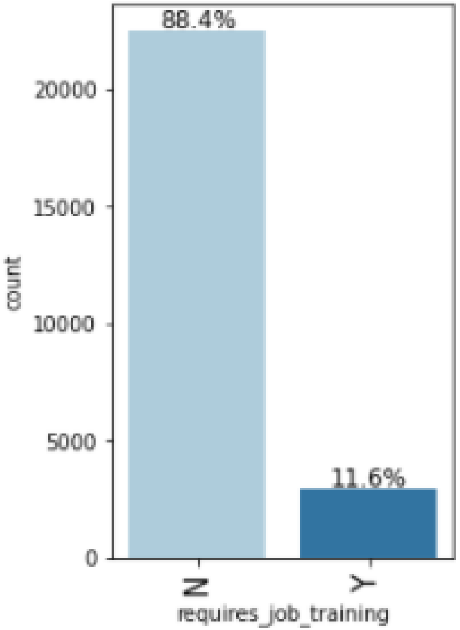




- We can see that 66% of all applicants have Asian continent heritage, with EU at nearly 15%, and North America at almost 13%.
- Almost 9% of all visa applicants have a Doctorate degree, Bachelor's and Master’s are close together, and assumption is that all have completed high school, but only 13.4% have completed just high-school within the given time period.
- Lastly, Northeast, South, and West seem to share the largest areas of regional employment interest amongst the foreign workers, while Midwest is at 17%, and U.S. islands only around 1.5%.

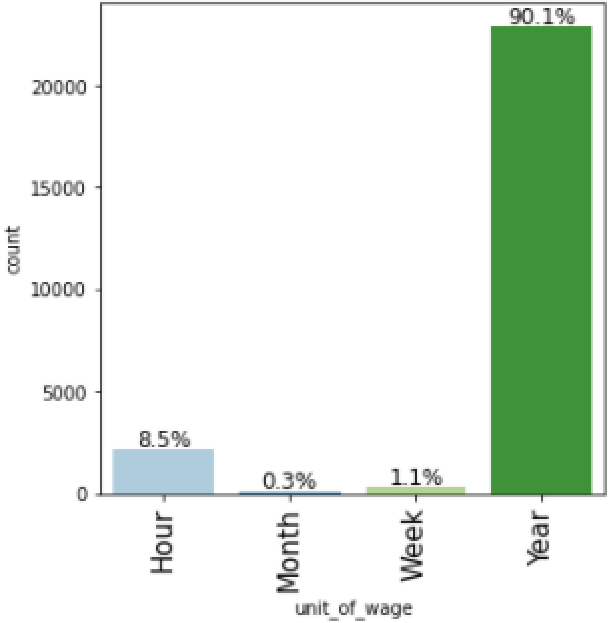
Exploratory Data Analysis

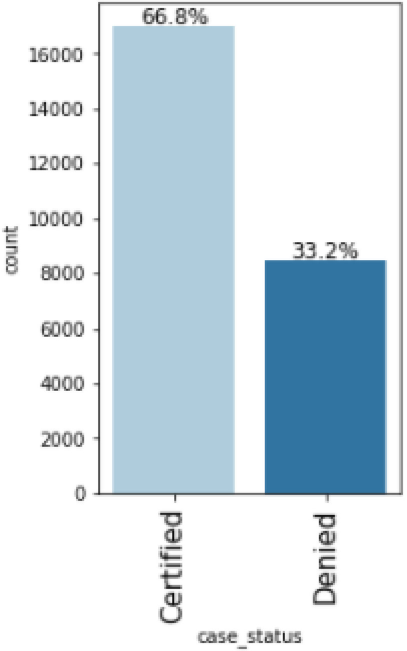




- 58.1% of all applicants have had job experience, while 41.9% have not had a prior experience.
- Almost 90% of all applicants do not require job training, while the smaller minority (the rest) do require some form of job training

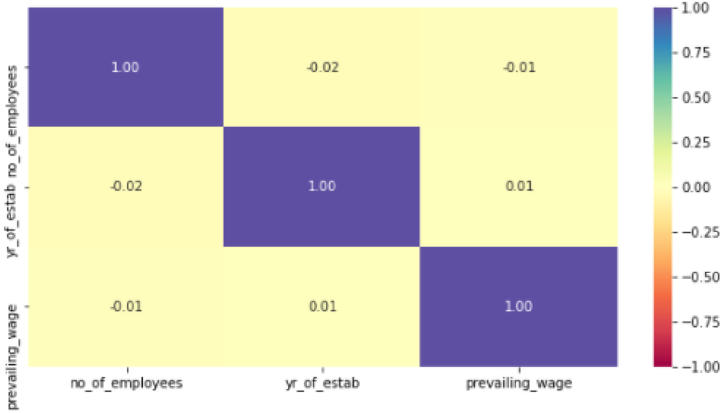
Exploratory Data Analysis

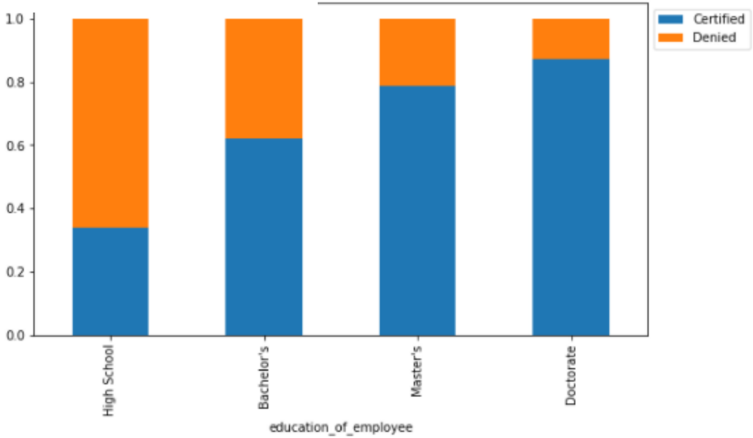




- Most unit of wage salary types are either annual at whopping 90% or hourly rate at around 9%. The rest of the units are either monthly or weekly.
- And lastly the most important data is the case status with 2/3 of all applicants passing the certification while 1/3 of applicants is denied the visa status.

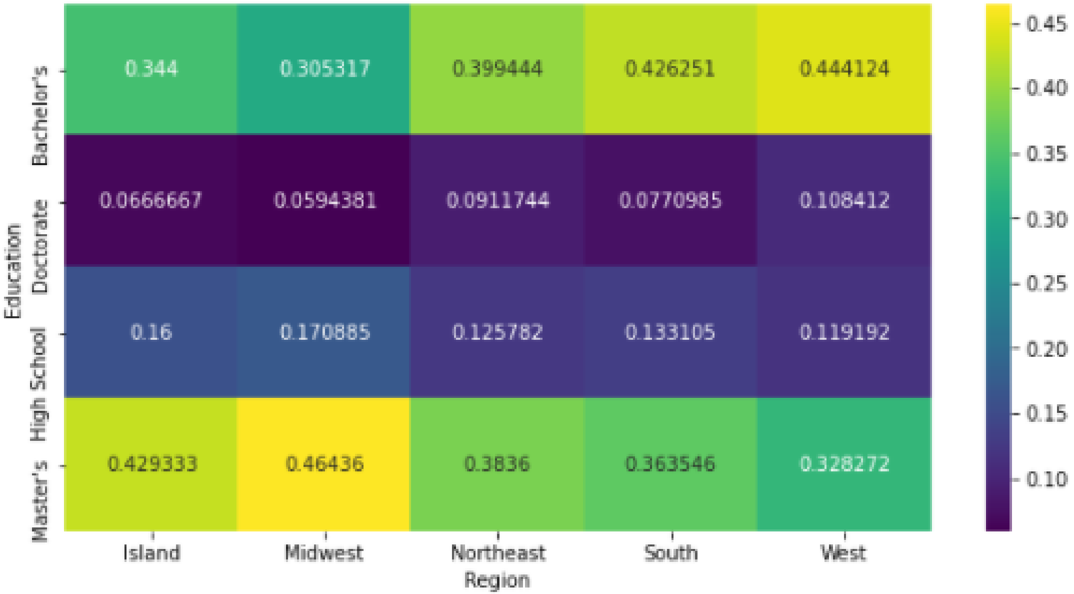
Exploratory Data Analysis





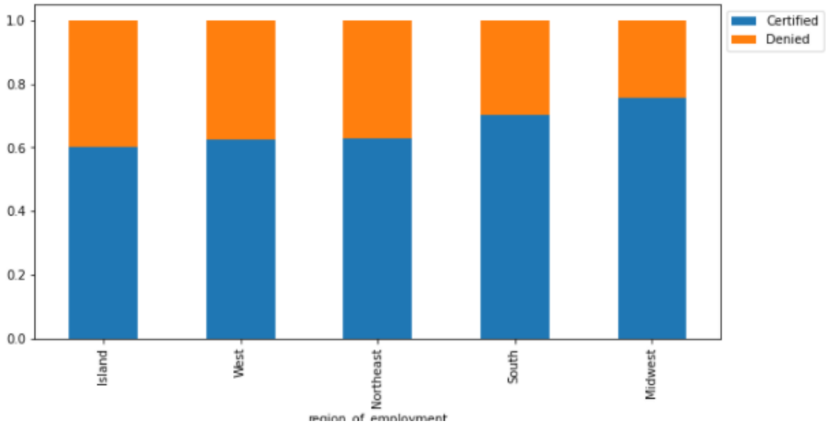
- The heatmap correlation for the available numeric data types shows no interesting information.
- However, looking at education vs. status shows an interesting pattern. We can see that certifications tend to be awarded in larger quantities to those applicants who have higher levels of education.

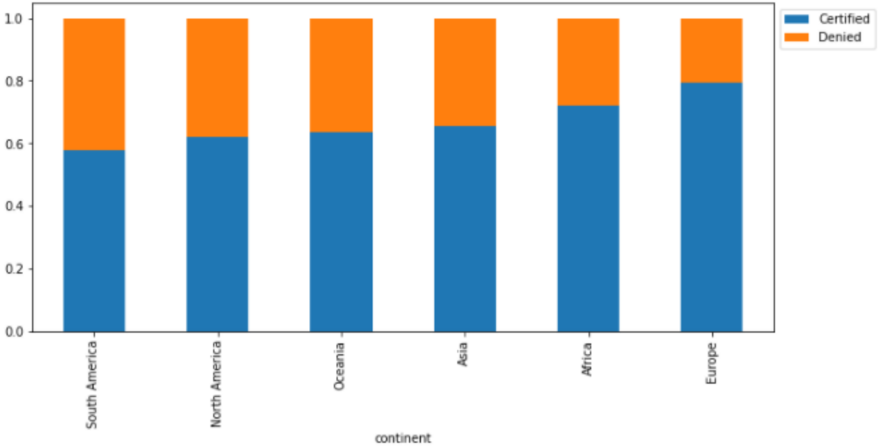
Exploratory Data Analysis



- Looking at the heatmap of the education mapped against different regions where applicants have expressed interest in working, we see that PhD and high school have nearly equally lower representation across all regions.
- Both islands and Midwest have similar higher needs into visa applicants with the higher Master's programs.
- And Bachelor's education seem to be best represented across the West and the South.

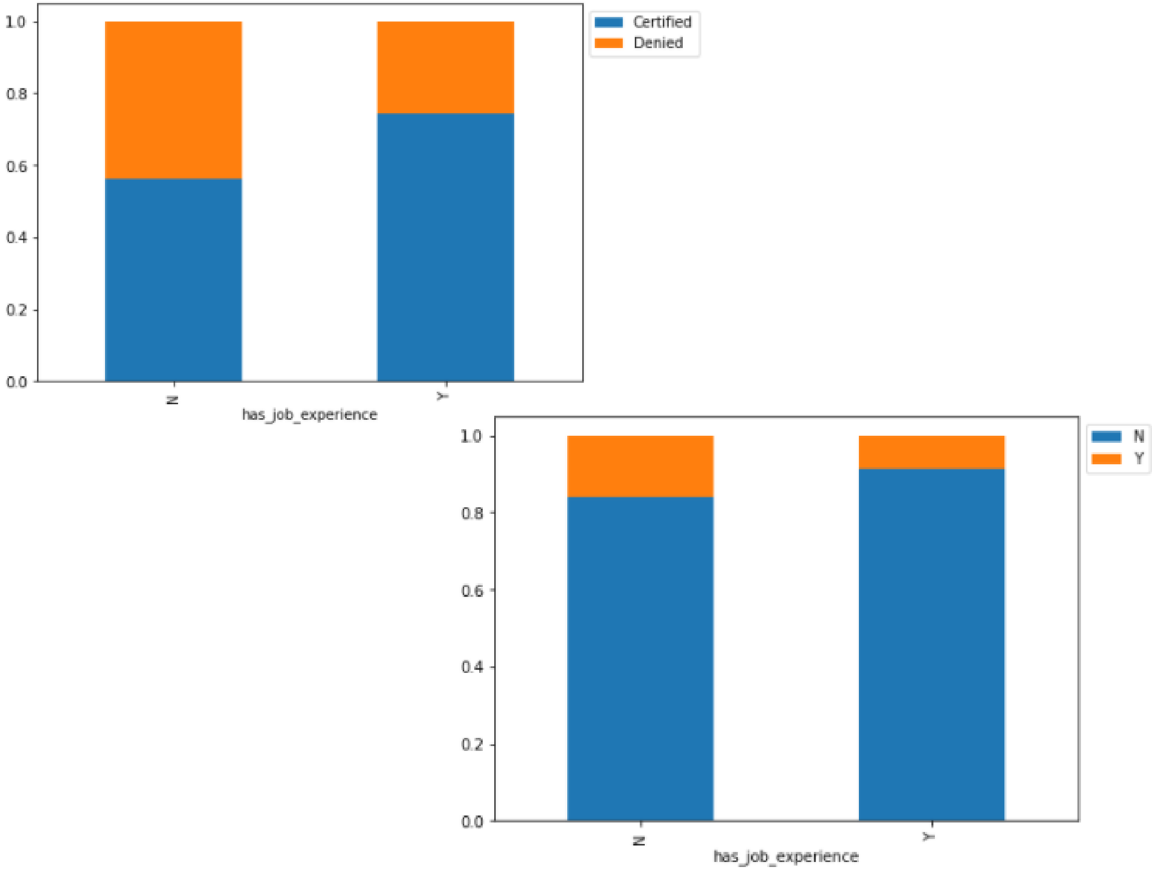
Exploratory Data Analysis





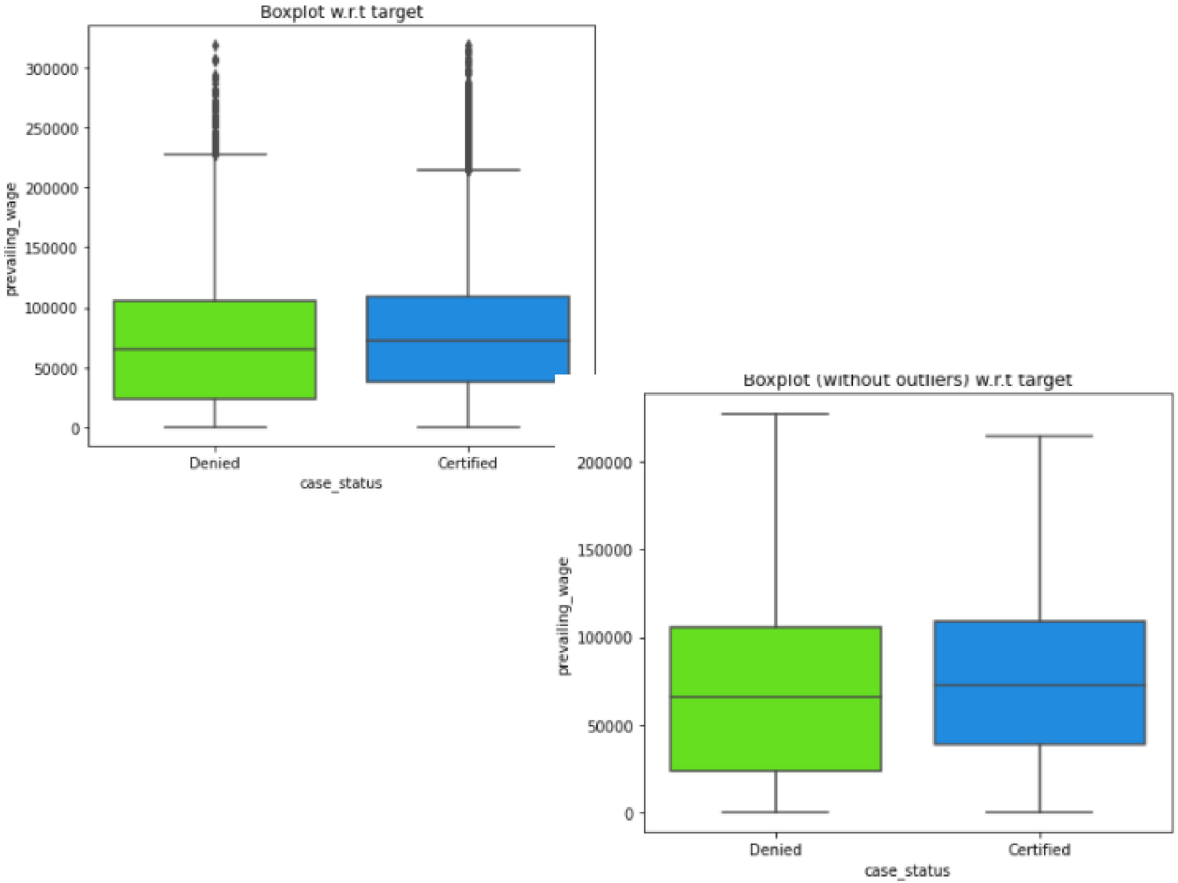
- We can see from the left chart that South and Midwest have the least denied/most certified cases, while Northeast, West, and Islands have similar lower rates of certification.
- From the bottom chart, we see that Europe and Africa candidates have the most certifications. The next level down in the number of certifications are Asia, Oceania, and North America. Lastly, the smallest number of certifications come from South America.

Exploratory Data Analysis



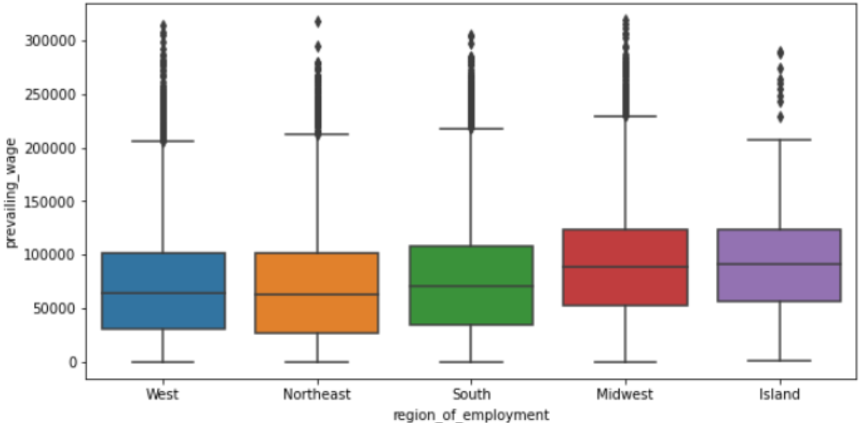
- From the chart on the left, we see that the lack of job experience has more declines. Nearly half of all submissions is declined.
- Meanwhile, in the lower chart, we see that those who have job experience almost have no job training, and even those applicants who do not have a job experience have less job training, which is surprising.

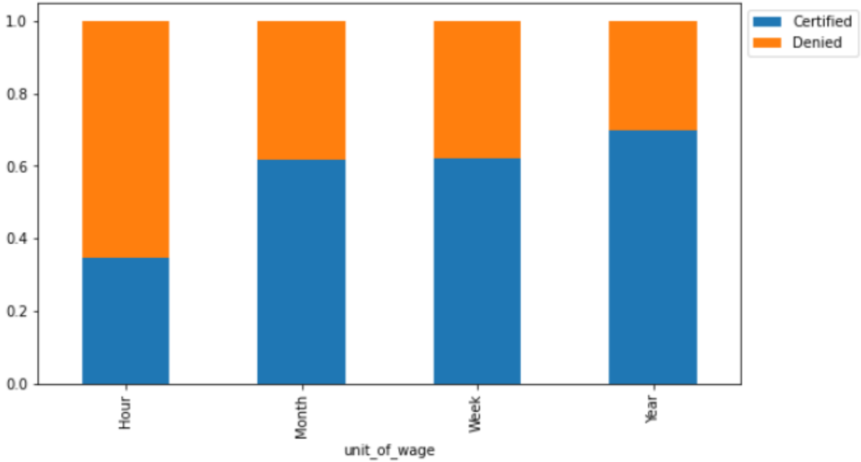
Exploratory Data Analysis



- Plot on the left shows that top quartile for IQR (75%) for both ‘denied’ and ‘certified’ cases has a prevailing wage of around $100k.
- The bottom IQR (25%) is a little more interesting, because for the Certified cases it is just about at $50k, while Denied cases lower IQR is nearly at $25k.
- We see somewhat of a similar situation in the same data even without the outliers.

Exploratory Data Analysis





- In the top chart, it's interesting to see that the prevailing wage IQRs are strongest for the Midwest and Islands (likely between $55k - $120k, while IQRs for the West, Northeast, and South are similarly positioned from around $30k+ through just around $100k
- We see that Annual/Year wage has the most certified applications, while Hourly wage has the most denied statuses. This could also likely be a reflection of the fact that typically hourly wages are associated with lower education (at least in U.S.). And we have seen in previous slides that lower education also has lower passes in certifications.

About the Model

Data Pre-Processing and Preparation for Modeling

- There were no missing values or duplicate values.
- We checked for outliers (see screenshot in Appendix). The no_of_employees data has the most outliers, followed by the prevailing_wage. I didn't treat the outliers due to me not understanding (yet) the Visa certification model, so wanted to check for how outliers impact the overall data and models.
- I wanted to predict which visa applications will be certified.
- Before building the model, I encoded categorical features.
- The case_id data was dropped at the beginning of the work. With all values unique, it didn't bring any value to the research at the moment.
- I created the dummy data.
- Data was split into train and test at 70:30 to be able to evaluate the model built on the train data. The case_status data was dropped from X.
- Lambda function was applied to the case_status to transition it to 1s and 0s, where the label Certified was moved to 1s.
- Percentage of classes were confirmed in the training and testing set.

About the Model – initial model evaluation

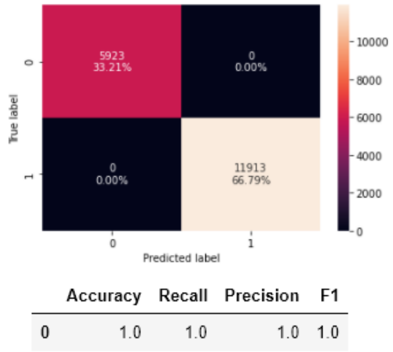
- We want to predict which visa cases (applicants) have the most chances of being selected/certified so that the future workload on case reviews can be reduced by picking the right cases to work on through the use of the ML algorithm, and as based on the various data categories/characteristics provided to us.
- The model can make wrong predictions if:
  - A) we predict that the visa application WILL NOT get certified but in reality the visa application should get certified. AND…
  - B) we predict that the visa application WILL get certified but in reality it should be denied/not certified.
- Both of these cases are nearly equally important to the agency, because in the case A) U.S. and its employers will miss on the opportunity to have/hire good or suitable international employees. And in the case B) there is a chance that we will allow an inadequate employee match, which is further likely to cause a loss of good job position for a U.S. citizen.
- To reduce the losses, the EasyVisa would want their model's F1 Score to be maximized, because higher F1 scores will minimize both cases, False Negatives and False Positives.
- To calculate this, we first created functions to calculate different metrics and confusion matrix so as not to repeat the same code for each model. We created the model_performance_classification_sklearn function to check the model performance, and the confusion_matrix_sklearn function to plot the confusion matrix.
- We also used balanced class weights so as to focus the model building equally on both classes of data.

About the Model – Model Building - Bagging

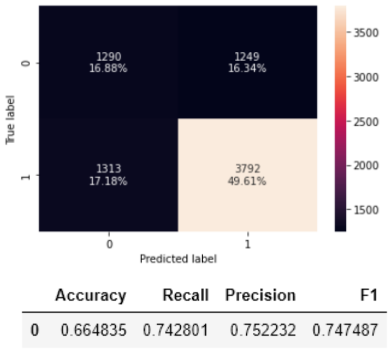
Train data

Test data

Pre-tuning

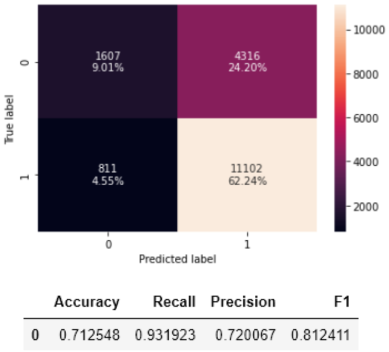


	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

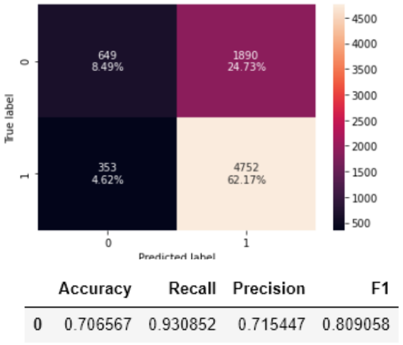


	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487

Post-tuning



	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411



	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

Decision Tree (pre- & post-tuning)

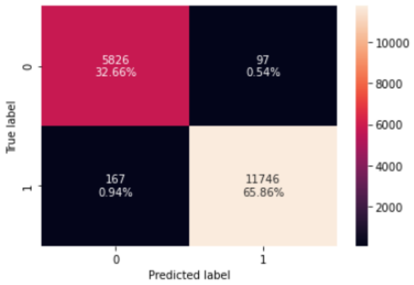
- We first defined a model_performance_classification_sklearn function to compute different metrics to check performance of a classification model built using sklearn.
- Another confusion_matrix_sklearn function was built to plot the confusion matrix.
- Then DecisionTreeClassifier was built using random_state=1, and fit to the training data.
- Checking the training data prior to tuning, we can see that the model is overfitting (all 1s). Testing data prior to tuning show F1=0.747
- We then tuned the model using:
  - ○ "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],
- Both Training and Testing data became more balanced (less overfitted and more similar), and F1 increased along with increase in Accuracy

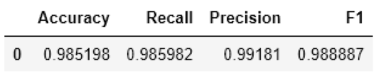
About the Model – Model Building - Bagging

Train data

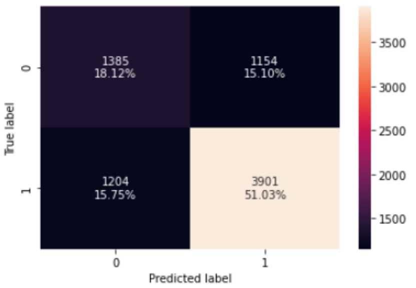
Test data

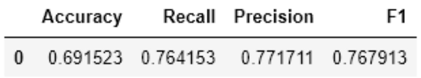
Pre-tuning





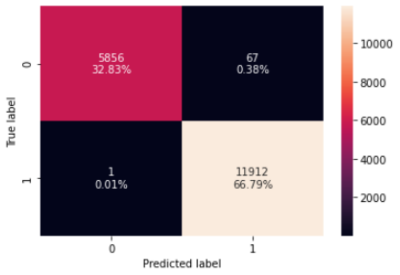
	Accuracy	Recall	Precision	F1
0	0.985198	0.985982	0.99181	0.988887

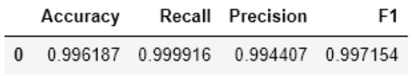




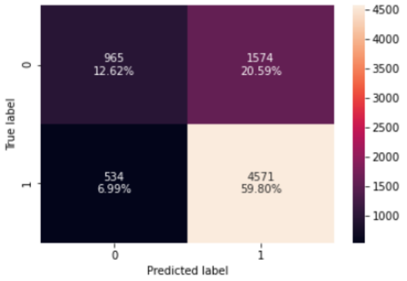
	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913

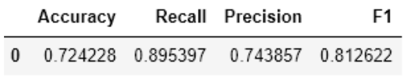
Post-tuning





	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154





	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

Bagging Classifier (pre- & post-tuning)

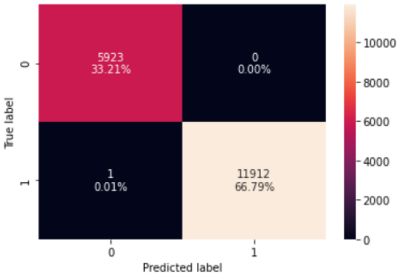
- Then BaggingClassifier was built using random_state=1 and fit to the training data.
- Checking the training data prior to tuning, we can see that the training data is much higher than the test data, so likely some overfitting again (all train data is close to .099). Testing data prior to tuning shows F1=0.767 and this is a very light improvement compared to the decision tree before this.
- We then tuned the model using:
  - "max_samples": [0.7, 0.8, 0.9],
  - "max_features": [0.7, 0.8, 0.9],
  - "n_estimators": np.arange(90, 120, 10)
- Post tuning, the Training data still showed much higher values than the Testing data so likely also overfitted.
- F1 score increased over both, the test data for the Bagging Classifier, and it increased compared to the F1 score of the post-tuning test data of the Decision Tree.
- Accuracy is still lower, but not too low.

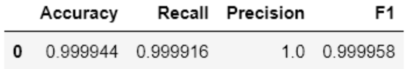
About the Model – Model Building - Bagging

Train data

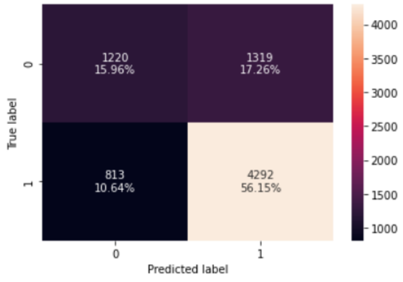
Test data

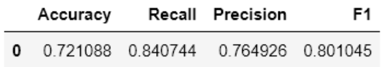
Pre-tuning





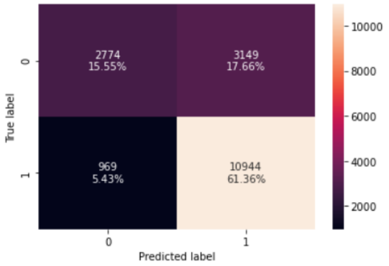
	Accuracy	Recall	Precision	F1
0	0.999944	0.999916	1.0	0.999958

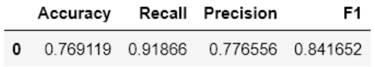




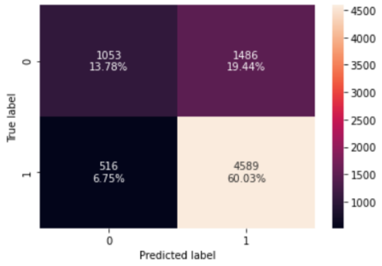
	Accuracy	Recall	Precision	F1
0	0.721088	0.840744	0.764926	0.801045

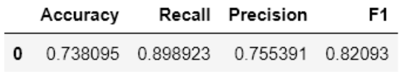
Post-tuning





	Accuracy	Recall	Precision	F1
0	0.769119	0.91866	0.776556	0.841652





	Accuracy	Recall	Precision	F1
0	0.738095	0.898923	0.755391	0.82093

Random Forest (pre- & post-tuning)

- RandomForestClassifier was built using random_state=1, class_weight="balanced", and was fit to the training data.
- Observing the training data prior to tuning, we can see that the training data again appears overfitted and much higher than the test data.
- Testing data prior to tuning shows higher F1 score than prior models pre-tuning (F1=0.801).
- We then tuned the model using:
  - "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),

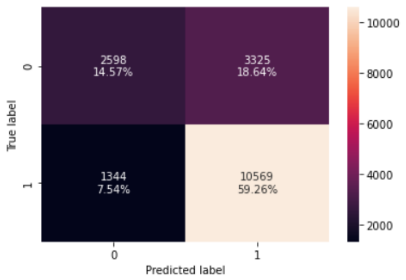
- Post tuning, both Training and Testing data was well-balanced (similar numbers and no overfitting).
- Post tuning in this case gave us best yet F1 score in the testing data F1=0.82, and though still lower, the Accuracy improved too, becoming the highest accuracy thus far.
- Let's explore Boosting in the next set of slides.

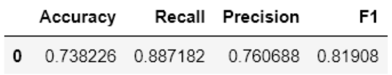
About the Model – Model Building - Boosting

Train data

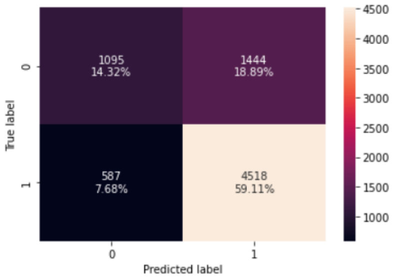
Test data

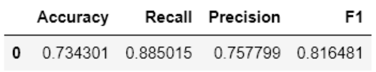
Pre-tuning





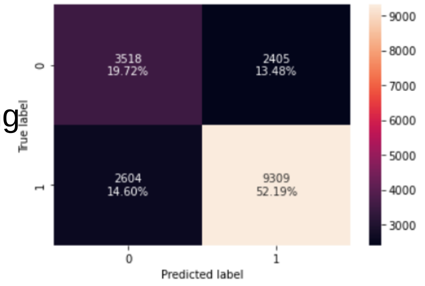
	Accuracy	Recall	Precision	F1
0	0.738226	0.887182	0.760688	0.81908

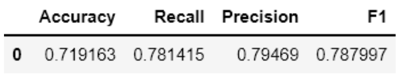




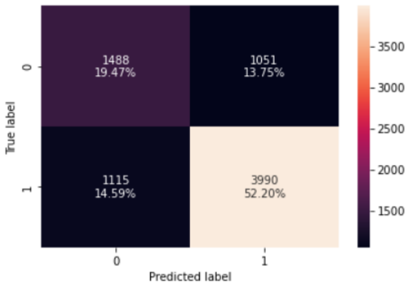
	Accuracy	Recall	Precision	F1
0	0.734301	0.885015	0.757799	0.816481

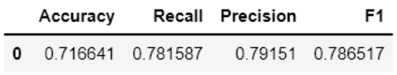
Post-tuning





	Accuracy	Recall	Precision	F1
0	0.719163	0.781415	0.79469	0.787997





	Accuracy	Recall	Precision	F1
0	0.716641	0.781587	0.79151	0.786517

AdaBoost Classifier (pre- & post-tuning)

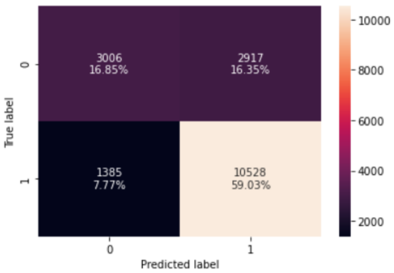
- After exploring the 3 Bagging models and their pre- and post-tuning results in the prior 3 slides, let's now check the Boosting models in the next few slides.
- AdaBostClassifier was built using random_state=1 and was fit to the training data.
- Observing the training data prior to tuning, we can see that the training data appears to be reasonably balanced and matching with the testing data.
- Testing data prior to tuning shows higher F1 score than prior Bagging models pre-tuning (F1=0.816).
- We then tuned the model using:
  - ○ "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),
- Post tuning, both Training and Testing data was well-balanced (similar numbers and no overfitting).
- However, in the post-tuning phase, we can see that both F1 score and Accuracy dropped for both, Training and Testing data.

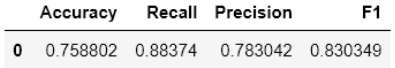
About the Model – Model Building - Boosting

Train data

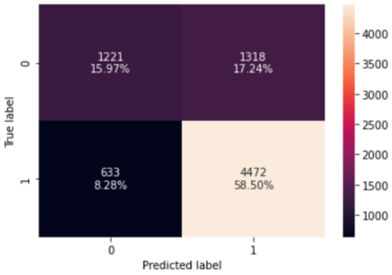
Test data

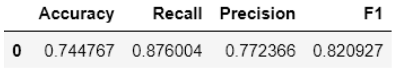
Pre-tuning





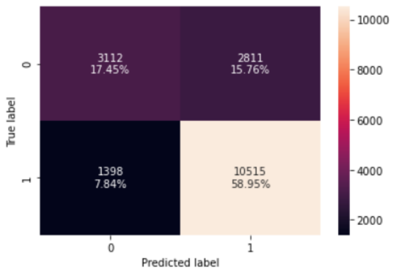
	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

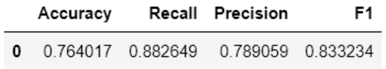




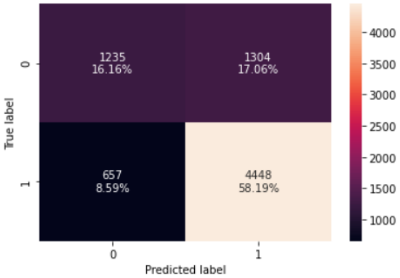
	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

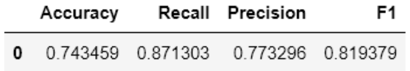
Post-tuning





	Accuracy	Recall	Precision	F1
0	0.764017	0.882649	0.789059	0.833234





	Accuracy	Recall	Precision	F1
0	0.743459	0.871303	0.773296	0.819379

Gradient Boosting Class. (pre- & post-tuning)

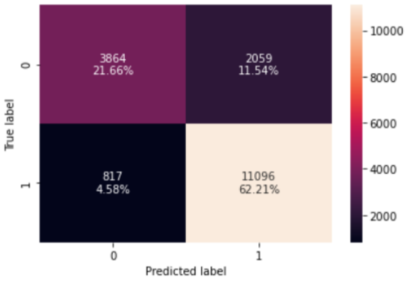
- GradientBoostingClassifier was built using random_state=1 and was fit to the training data.
- Observing the training data prior to tuning, we can see that the training data appears to be reasonably balanced and matching with the testing data.
- Testing data prior to tuning shows the highest F1 score yet than all the prior Bagging & Boosting models pre-tuning (F1=0.820).
- We then tuned the model using:
  - ○ "n_estimators": [200, 250, 300],
  - ○ "subsample": [0.8, 0.9, 1],
  - ○ "max_features": [0.7, 0.8, 0.9],
  - ○ "learning_rate": np.arange(0.1, 0.4, 0.1),
- Post tuning, both Training and Testing data was well-balanced (similar numbers and no overfitting).
- However, post tuning in this case gave us slightly lower F1 score in the testing data F1=0.819, though not much lower.
- Accuracy also dropped by very little.
- Let’s now look at the last Boosting model.

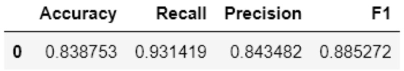
About the Model – Model Building - Boosting

Train data

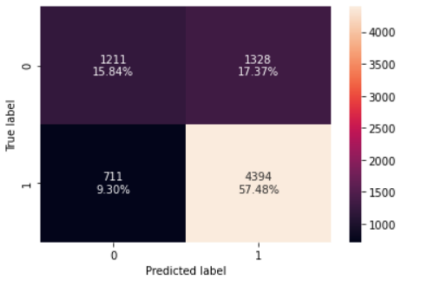
Test data

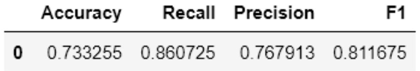
Pre-tuning





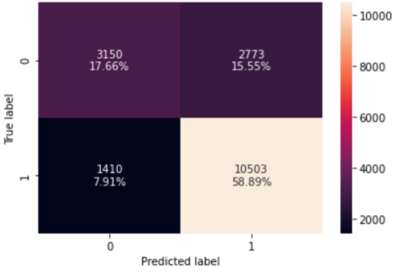
	Accuracy	Recall	Precision	F1
0	0.838753	0.931419	0.843482	0.885272

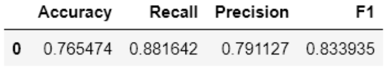




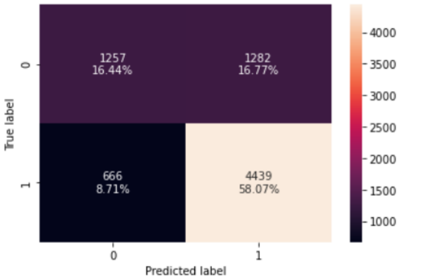
	Accuracy	Recall	Precision	F1
0	0.733255	0.860725	0.767913	0.811675

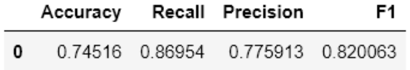
Post-tuning





	Accuracy	Recall	Precision	F1
0	0.765474	0.881642	0.791127	0.833935





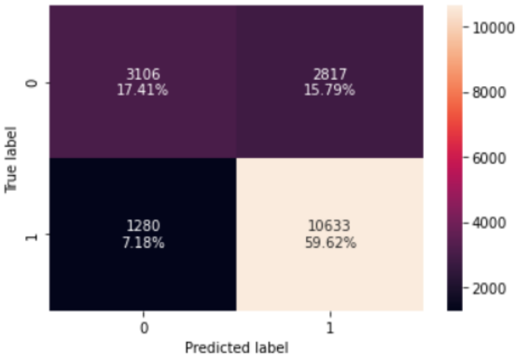
	Accuracy	Recall	Precision	F1
0	0.74516	0.86954	0.775913	0.820063

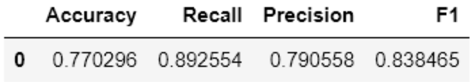
XGBoost Classifier (pre- & post-tuning)

- XGBClassifier was built using random_state=1, eval_metric="logloss", and was fit to the training data.
- Checking the training data prior to tuning, we can see that the training data appears possibly overfitted and higher than the test data.
- Testing data prior to tuning shows good F1 score (F1=0.811), however, not better than Gradient Boosting.
- We then tuned the model using:
  - ○ "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],
- Post tuning, both Training and Testing data was well-balanced (similar numbers and no overfitting).
- Post tuning in this case gave us great F1 score in the testing data F1=0.833, and pretty good though slightly lower F1=0.820 in the Testing data.
- Accuracy is lingering around the usual 0.745 – similar to other Boosting models post-tuning in this case.

About the Model – Model Building - Boosting

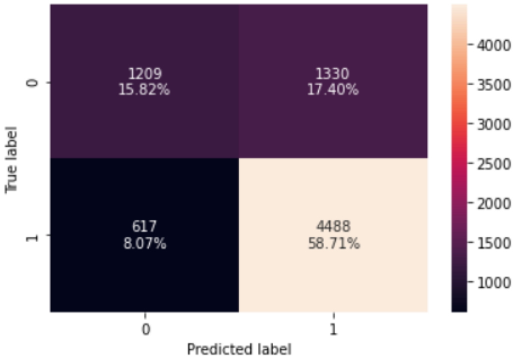
Train data

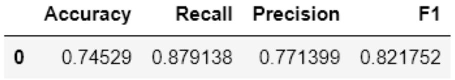




	Accuracy	Recall	Precision	F1
0	0.770296	0.892554	0.790558	0.838465

Test data





	Accuracy	Recall	Precision	F1
0	0.74529	0.879138	0.771399	0.821752

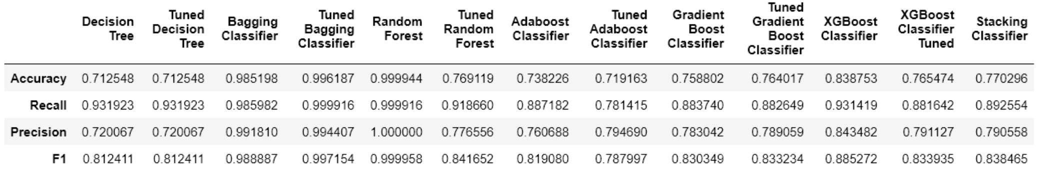
Stacking Classifier

- StackingClassifier was built using estimators (prior built models AdaBoost, GradientBoost and Random Forest), and using the final_estimator=xgb_tuned. It was then fit to the training data.
- Observing the training and testing data, we see that they're pretty close together in numbers – balanced.
- Train data F1=0.838, while Test data F1=0.821; in both cases, it's a relatively good F1 score compared to prior built models in this research case.
- Accuracy, Recall, and Precision all appear pretty good compared to prior models.

- **Thus far, you have seen me build and display all Bagging and all Boosting models and compare each model individually pre- and post-tuning. Let's now compare all models together in the next slide.**

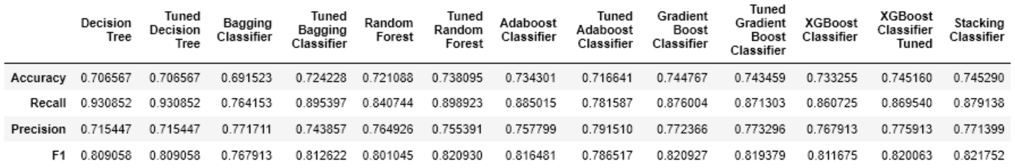
About the Model – Final Comparison

Train 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	0.712548	0.712548	0.985198	0.996187	0.999944	0.769119	0.738226	0.719163	0.758802	0.764017	0.838753	0.765474	0.770296
Recall	0.931923	0.931923	0.985982	0.999916	0.999916	0.918660	0.887182	0.781415	0.883740	0.882649	0.931419	0.881642	0.892554
Precision	0.720067	0.720067	0.991810	0.994407	1.000000	0.776556	0.760688	0.794690	0.783042	0.789059	0.843482	0.791127	0.790558
F1	0.812411	0.812411	0.988887	0.997154	0.999958	0.841652	0.819080	0.787997	0.830349	0.833234	0.885272	0.833935	0.838465

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	0.691523	0.724228	0.721088	0.738095	0.734301	0.716641	0.744767	0.743459	0.733255	0.745160	0.745290
Recall	0.930852	0.930852	0.764153	0.895397	0.840744	0.898923	0.885015	0.781587	0.876004	0.871303	0.860725	0.869540	0.879138
Precision	0.715447	0.715447	0.771711	0.743857	0.764926	0.755391	0.757799	0.791510	0.772366	0.773296	0.767913	0.775913	0.771399
F1	0.809058	0.809058	0.767913	0.812622	0.801045	0.820930	0.816481	0.786517	0.820927	0.819379	0.811675	0.820063	0.821752

Comparing all models

- Thus far, you have seen me build and display all Bagging and all Boosting models and compare each model individually pre- and post-tuning. Let's now compare all models together in this slide.
- Observing all the training data, we see that Bagging Classifier, Tuned Bagging Classifier, and Random Forest have the highest F1 scores. However, at the same time, those 3 models also have the largest differences with their testing data counterparts, hence the models appear to be overfitted, so I wouldn't necessarily choose those for our ML prediction modeling.
- Looking at the highest F1 score beyond the ones in the overfitted data, I would likely choose the Stacking Classifier, because its F1 score appears high in Training (0.833) and Testing data (0.821), and Accuracy is among the highest in Training (0.770) and Testing (0.745) [again, not comparing to the overfitted data].
- Generally speaking, Boosting models seem to have done better in not overfitting the data while achieving higher F1 scores.
- Lastly, the most important features that we can use as predictors (see Appendix):
  - ○ High school education
  - ○ Prevailing wage
  - ○ Having job experience (Y)
  - ○ Master's education

Business Insights and Recommendations

Actionable Insights

- We analyzed the EasyVisa case classification (certified y/n) models and framework, which would help to facilitate the visa approvals process, while recommending the best possible applicant profile that is most likely to be certified and case status getting approved.
- We used 3 Bagging and 4 Boosting models to gather some insights and compare F1 scores in order to get the best combination of the false positive and false negative predictions, which are both important in this case.
- They are both important because U.S. and its employers will miss on the opportunity to hire good or suitable international employees if the certification process does not pick an adequate case. And if the certified employee is inadequate, this is further likely to cause a loss of good job position for a U.S. citizens.
- We were able to build predictive models that can be used by EasyVisa to predict the risk factor of the abovementioned predictors in getting candidates approved/certified. This would help to reduce the cost of the process while making the process more expedient and efficient, and all along removing the process bottleneck so that more (hopefully good) cases can be processed and certified.
- We found the Highschool education, Prevailing wage, Having job experience (Y), and Master's education to be the most important variables (in that order) in predicting the positive outcome of the case certification status (case certification approved). More about this on the next page...

Business Insights and Recommendations

Actionable Insights Continued…

- Our target candidates must have high school education and if they go as far as master's education, we've seen in EDA that master's education candidates – next to PhD candidates – have the least number of denials. Beyond just sorting the applications based on these education factors, the company can establish connection with certain international education institutions and colleges to recruit more favorable candidates.
- While nearly 42% of all candidates do not have job experience, it appears that having job experience is an important predictor of success. So while going directly to colleges would be good in terms of finding educated candidates, getting EasyVisa to establish certain work credentials and possibly get candidates with at least 1-2 years worth of work experience would be even more favorable.
- Prevailing wage is another important factor. The company can ensure that their filtering and certification process considers foreign workers not getting underpaid compared to other workers offering the same or similar service in the same area of employment. As seen in EDA, for example, Midwest and Islands regions tolerate higher wages, while other regions closely align on somewhat lower wages. EasyVisa should definitely match and overlay regional wage data over the candidates' prevailing wage data to ensure that there is a match tailored to each of the U.S. regions.
- Another specific action point for EasyVisa is to consider not picking anyone wanting to make less than $50k for the process of certification, because EDA clearly shows that the lower IQR (25th percentile) for successfully certified candidates is right around/above $50k.
- Lastly, EasyVisa could establish connections with other similar agencies in Europe and Asia continents which seem to have smaller number of declines, but Asia in particular which has 66% of all processed cases.

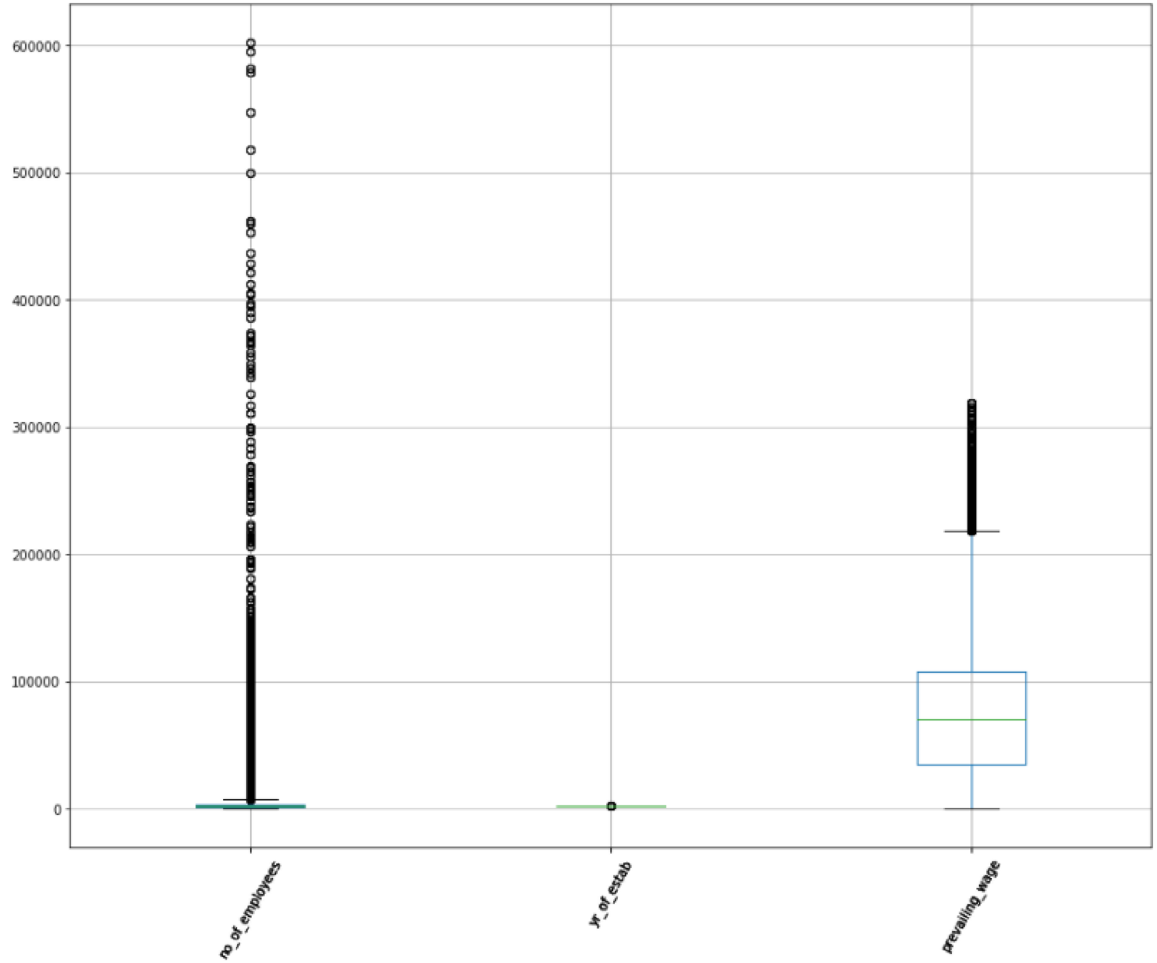
greatlearning
Power Ahead

Happy Learning !

Appendix – Data Dictionary

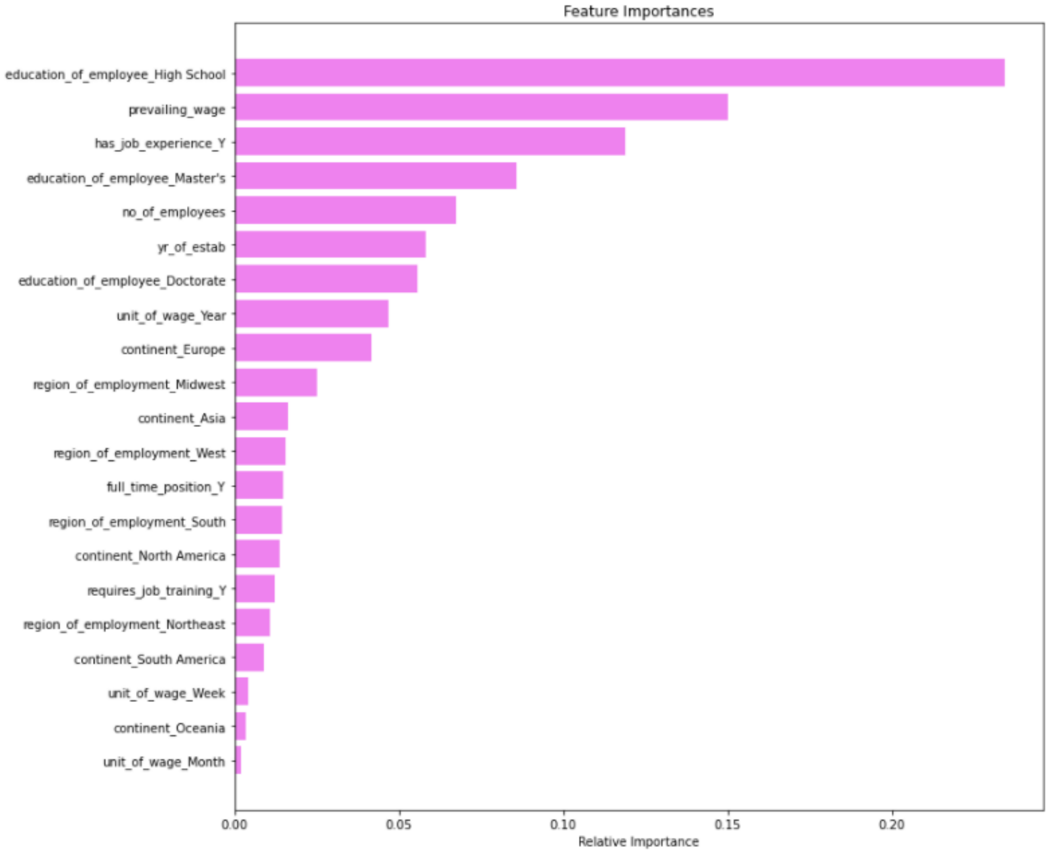
- Booking_ID: the unique identifier of each booking
- case_id: ID of each visa application
- continent: Information of continent the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee has any job experience? Y= Yes; N = No
- requires_job_training: 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

Outliers



600000
500000
400000
300000
200000
100000
0
no_of_employees
yr_of_estab
prevailing_wage

Feature Importances


Feature Importances
education_of_employee_High School
prevailing_wage
has_job_experience_Y
education_of_employee_Master's
no_of_employees
yr_of_estab
education_of_employee_Doctorate
unit_of_wage_Year
continent_Europe
region_of_employment_Midwest
continent_Asia
region_of_employment_West
full_time_position_Y
region_of_employment_South
continent_North America
requires_job_training_Y
region_of_employment_Northeast
continent_South America
unit_of_wage_Week
continent_Oceania
unit_of_wage_Month
Relative Importance