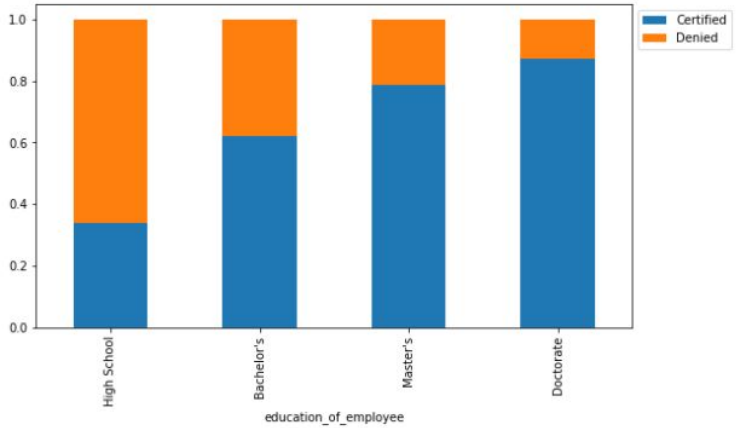
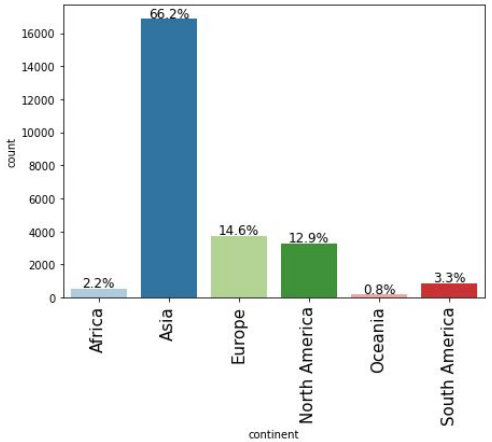
EasyVisa Study

for the Process of OFLC VISA Approvals





Objectives:

1. Using past OFLC data, construct and tune Models to predict VISA acceptance and rejection.
2. With a combination of our Data Analysis and Models, facilitate the process of visa approvals by providing a list of the most important factors, positive and negative.
    a. We will be conducting Exploratory Data Analysis, as well as constructing multiple tuned models (Bagging, Boosting, Stacking) to find the most important and influential factors.
3. Provide the OFLC with a final model which can be used for VISA predictions.

A quick look at our tuned models before we begin…

Testing performance comparison:

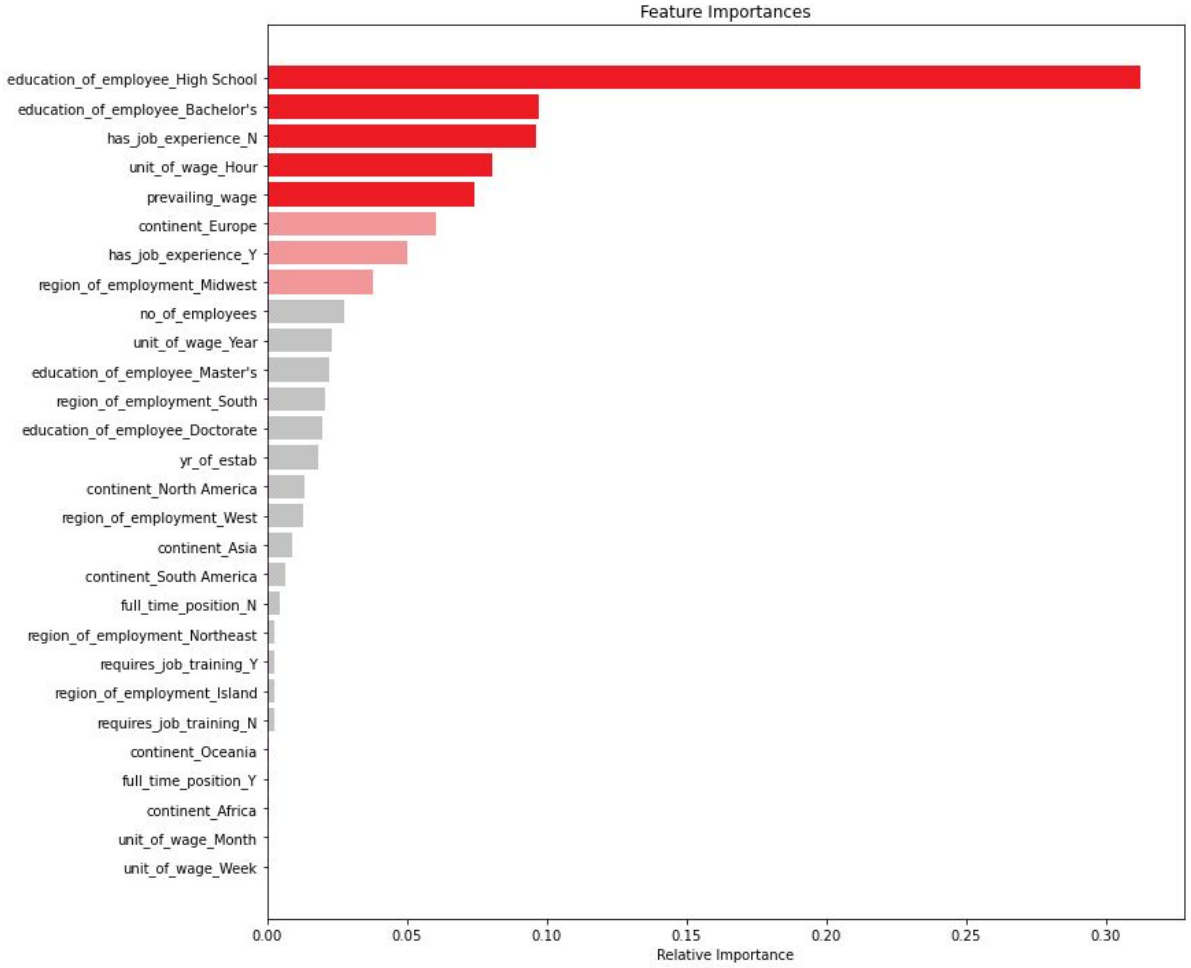
	Tuned Decision Tree	Tuned Bagging Classifier	Tuned Random Forest	Tuned Adaboost Classifier	Tuned GradientBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.706567	0.731293	0.745029	0.734040	0.747384	0.743590	0.745814
Recall	0.930852	0.874241	0.882468	0.887953	0.875416	0.866993	0.861312
Precision	0.715447	0.759660	0.769559	0.756256	0.775330	0.775539	0.780717
F1	0.809058	0.812933	0.822155	0.816830	0.822339	0.818720	0.819037

- Our best models ended up being a **tuned Gradient Boost** and our **tuned Random Forest**.
- It is important to identify both VISA approvals and denials correctly, so the **most important metric is** the Harmonic Mean of Recall and Precision, otherwise known as **F1**.
    - The **higher the F1**, the **greater the chances of reducing false predictions**.
    - However, **Accuracy is also important**. The higher the accuracy, the less the model predicted incorrectly.

Feature Importances

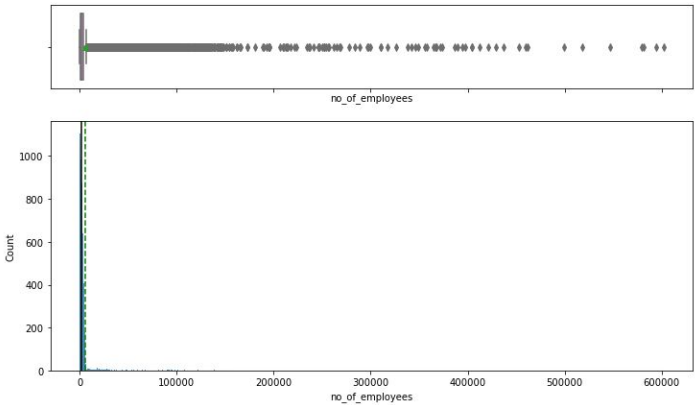
- The 5 greatest influences on VISA certification are **High School Education**, the applicant having a **Bachelor's**, the applicant ***not*** **having job experience**, the applicant's area of employment being **paid hourly**, and **prevailing wage**.
- We will keep this in mind as we begin our Exploratory Data Analysis.

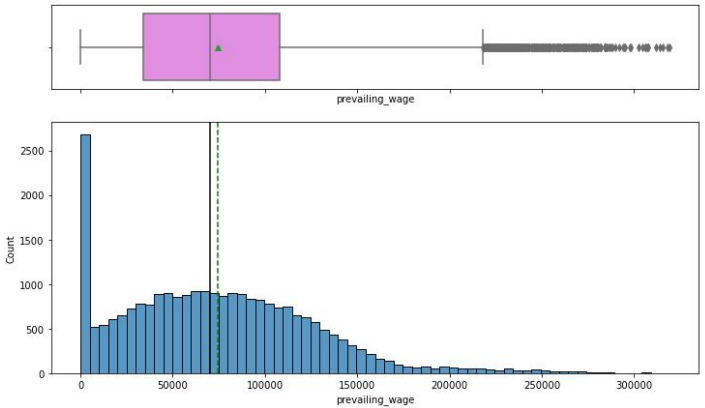
We will revisit the models at the end of the presentation.



Univariate Analysis

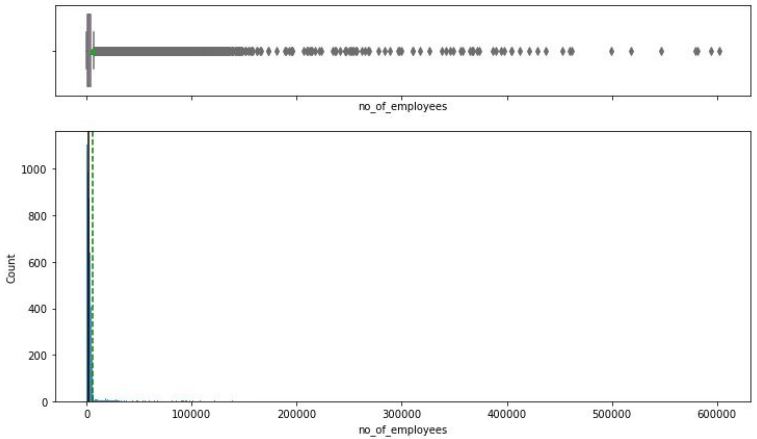
Individual features with notable aspects

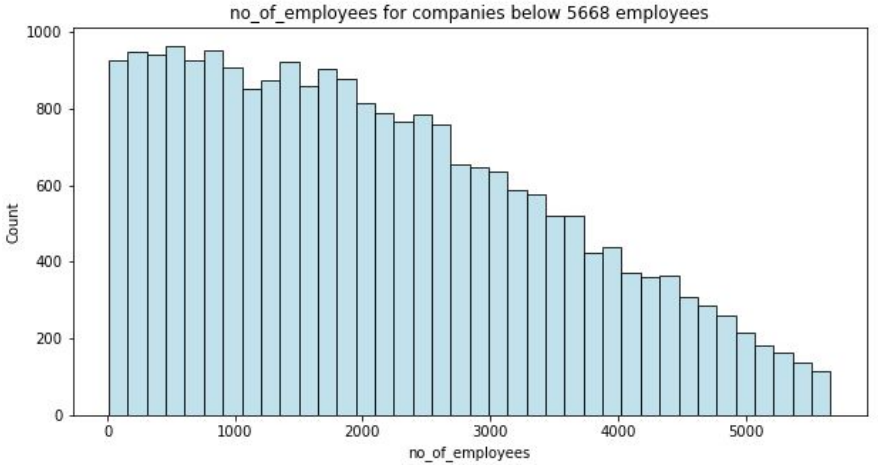




Number of Employees

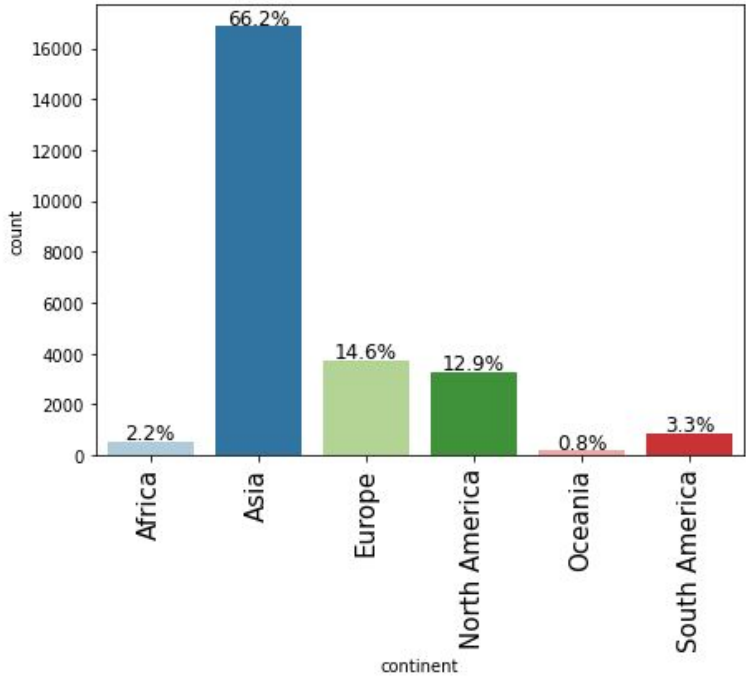
- The average number of employees at US-based companies that hire foreign help is 5668 (based on the data given).
- However, 93% of all companies hiring foreign labor through OFLC have at most 5668 employees. Around 10% of the companies are responsible for employee counts above the mean, up to 600k employees.
- While the final model doesn't list no_of_employees as a particularly important feature, it is important to note the number and range of the outliers.



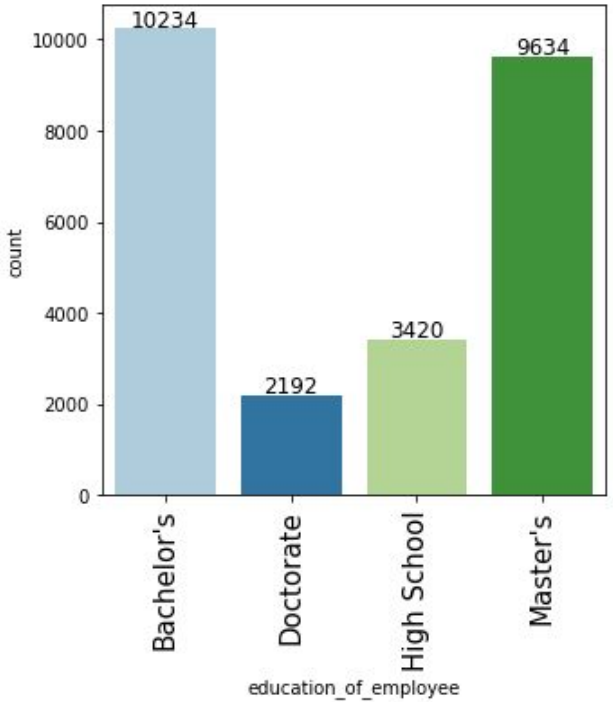


Continent of Applicant

- The majority of applicants come from Asia, with the 2nd and 3rd most applied-from continents being Europe and NA.
- Keep this in mind, as continent_Europe was the 6th most important feature in our final model.



Education of Applicant

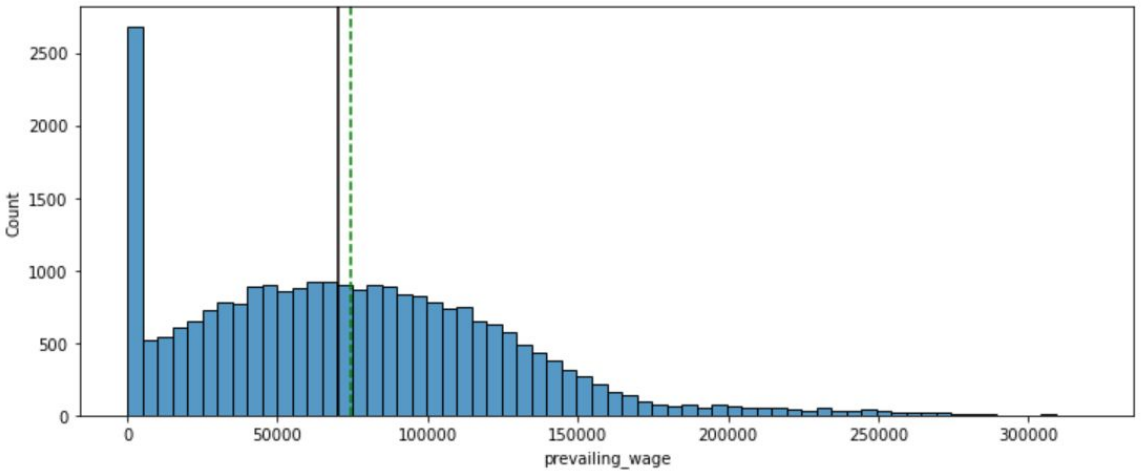


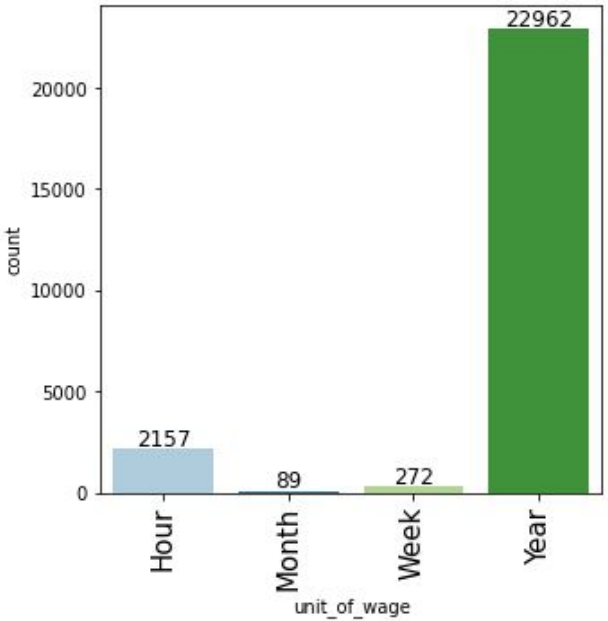
- There are few applicants with only High School education, and there are few applicants with a Doctorate.
- As High School Education is the most important feature in our final model, it is important to note the relatively low percentage of High School applicants present in the data.
- Also take note of the high amount of applicant graduates with Bachelor's degrees, as Bachelor's was the second most important feature.

Wage and Salary

- The majority of foreign employees that apply for a job in the US are applying for an annual salary, not an hourly position.
- This is reflected in the large amount of prevailing_wage entries above 10,000.

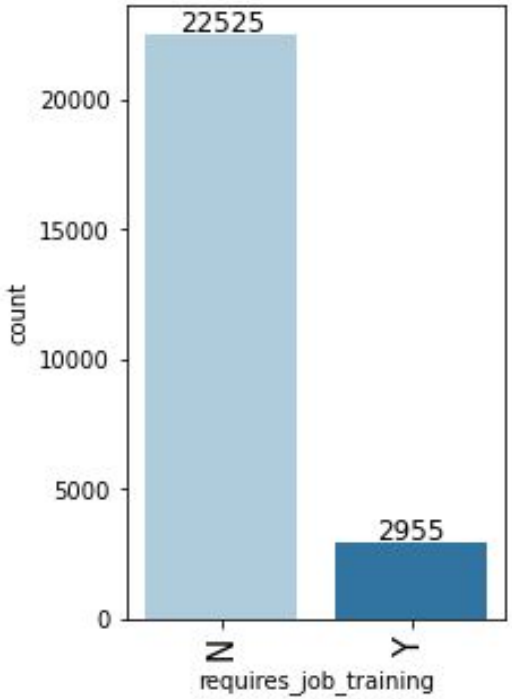
87% of all entries in prevailing_wage are above 10,000, and 98% of those entries are annual salaries. The other 2% are weekly and monthly payments.





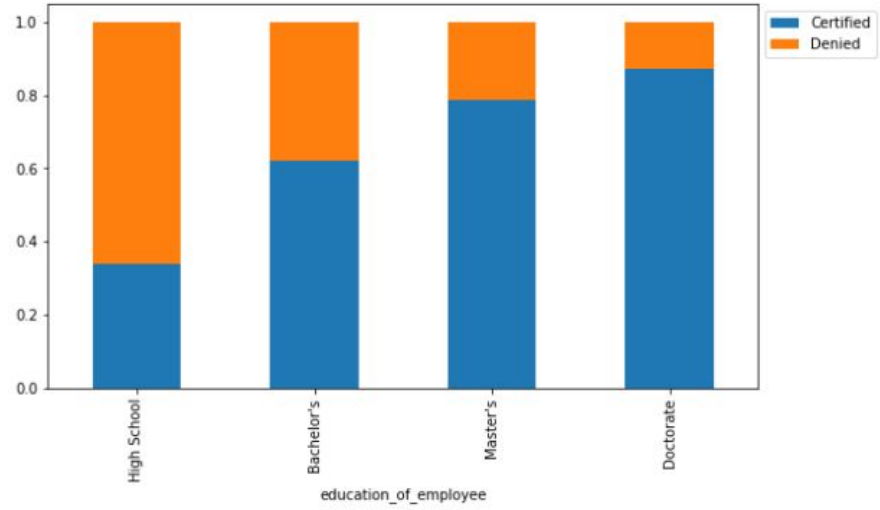
Job Training Required

- This feature was surprisingly low in our final model's Feature Importances.
- Even more surprising is the large difference in counts between those who *do* and *do not* require training.

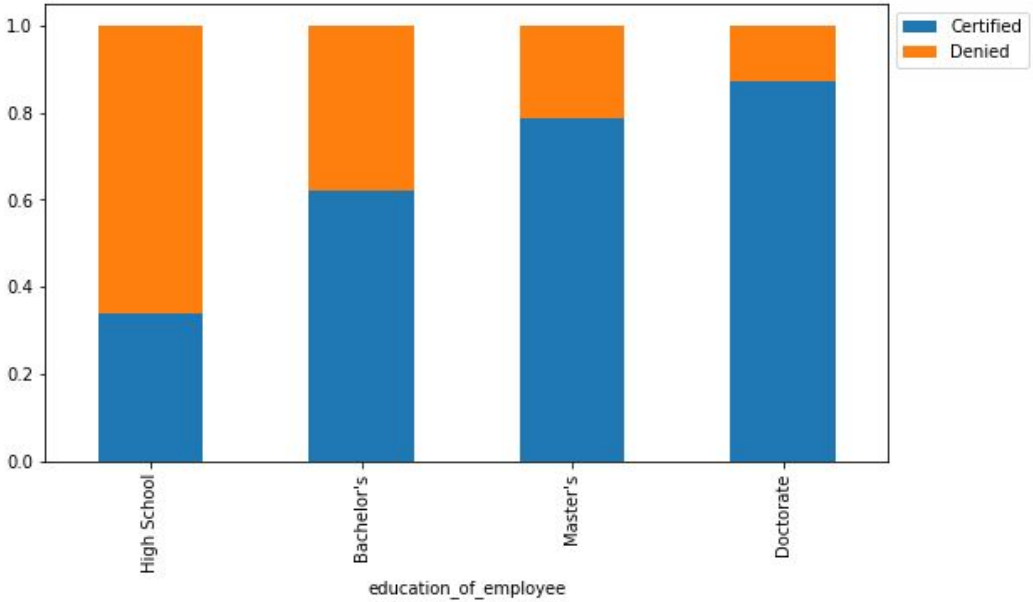


Bivariate Analysis

Features with correlation to others



Education VS Certification Status

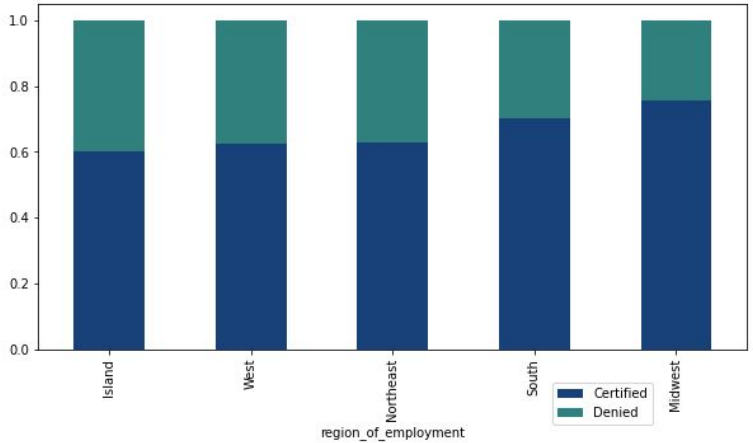


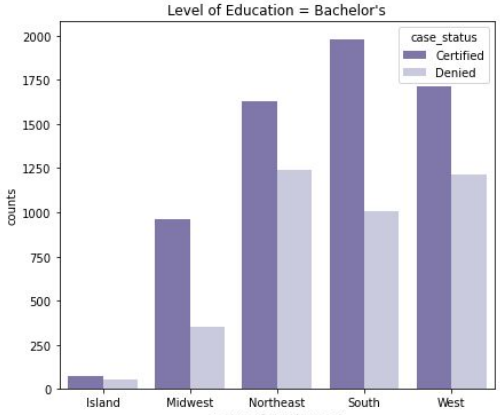
- Seeing as the **two most important features** in the final model are **High School Education** and having a **Bachelor's Degree**, we can put more stock into this area of analysis.
- High School Graduates have below a 40% acceptance rate, and graduates with a Bachelor's Degree have above a 60% acceptance rate.
- Taking this information with the Feature Importances of the model, having only a **High School education has a very significant detriment** to the likeliness of acceptance, and **having at least a Bachelor's has a fairly significant advantage** to foreign workers looking to work in the USA.

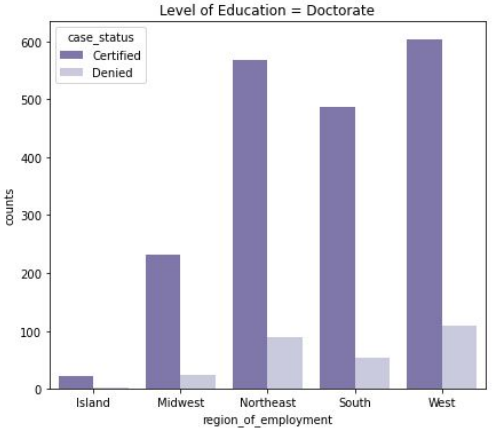
Region
VS Certification Status
VS Level of Education

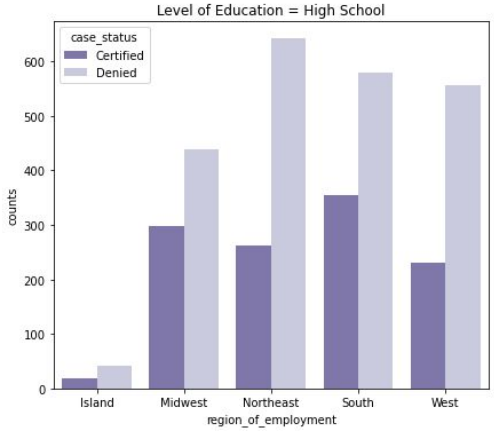
- Across all levels of Education, the Midwest has the highest rates of acceptance.
- As the Midwest is the 8th most important feature in our final model, it's likely that **Midwest applicants are more likely to be accepted than others.**

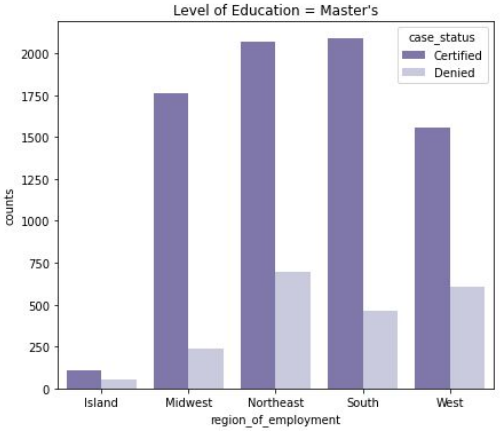
From High School to Doctorate, as education level increases, rate of acceptance in all regions increases.



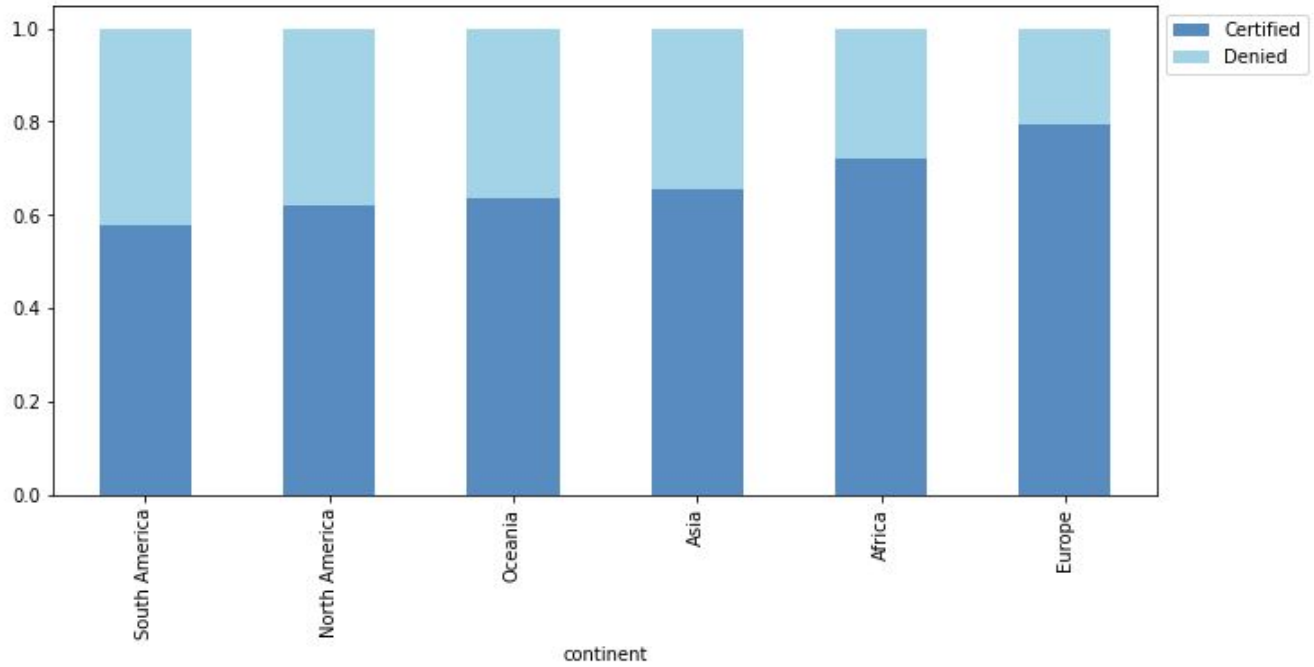








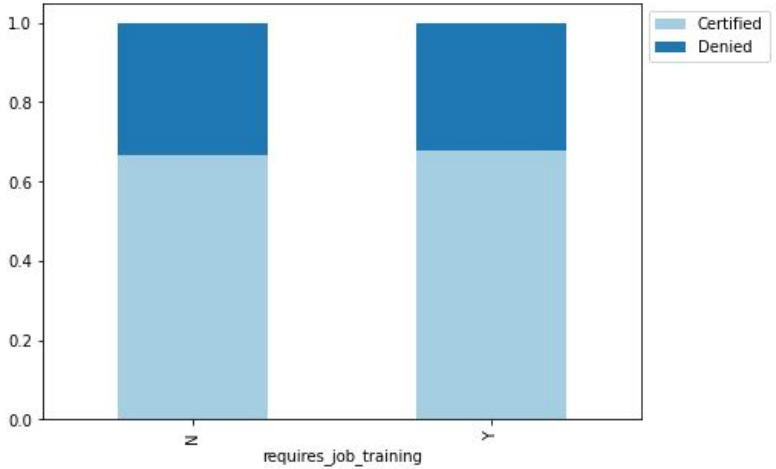
Continent VS Certification Status

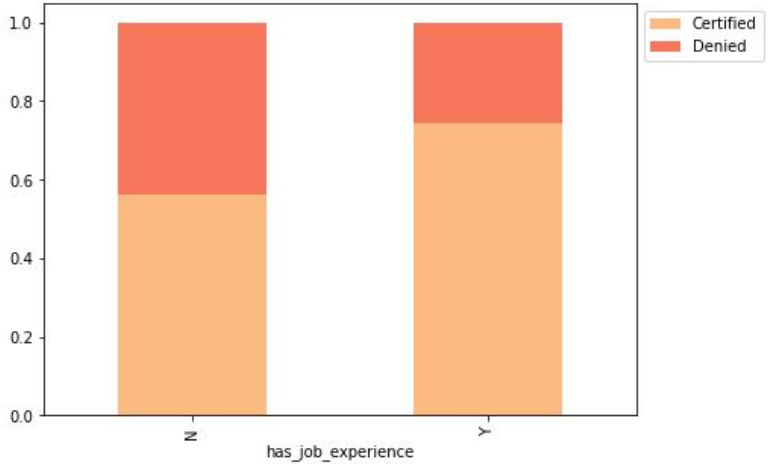


- Europe has the highest rate of application certification.
- As it is the 6th most important feature, it is likely that **European workers are certified more often** than any other continent.

Training Required & Job Experience VS Certification Status

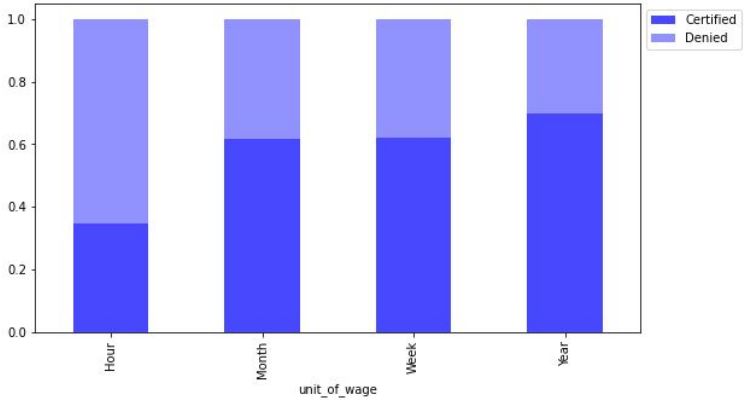
- As we saw earlier, ‘Job Training Required’ had very low significance in our final model, and ‘Previous Job Experience’ had very high significance.
    - Note: ‘No Job Experience’ had a much higher importance than ‘Previous Job Experience’.
- It seems that **‘Job Training Required’ has close to zero impact** over the acceptance of an applicant, both by the raw data and by our final model.
- As Experience has high impact in our final model, it is likely that **‘No Job Experience’ has a significant detriment** on the chances of acceptance, and **‘Previous Job Experience’ has a fairly significant advantage** to foreign applicants looking to work in the USA.

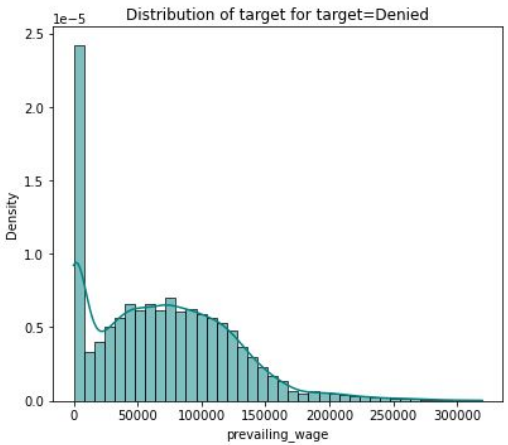


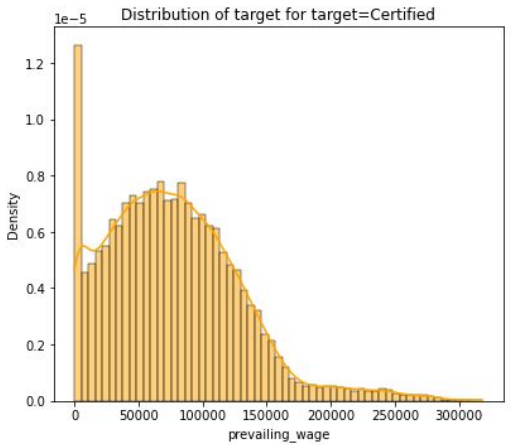


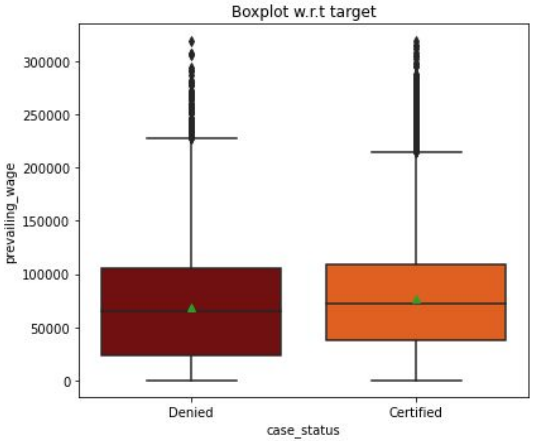
Wage and Salary VS Certification Status

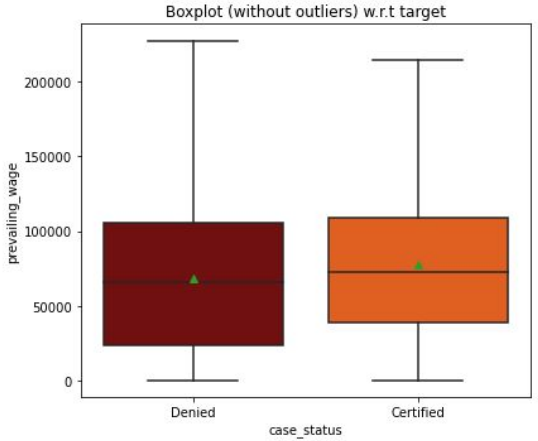
- Prevailing Wage was the 5th most important Feature in our final model.
  - As shown in our boxplots, prevailing wage is slightly higher in areas where foreign workers were certified. Thus, **an application with higher prevailing wages is more likely to be accepted.**
- Applications for jobs with hourly pay is by far the worst in terms of acceptance.
  - Seeing as hourly pay was the 4th most important feature in our model, **jobs for Hourly Pay are more likely to be rejected.**





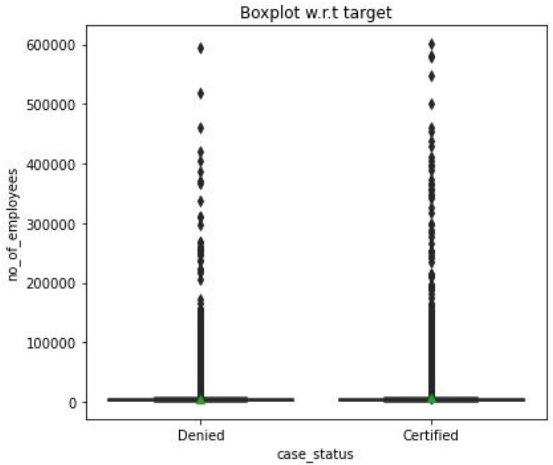


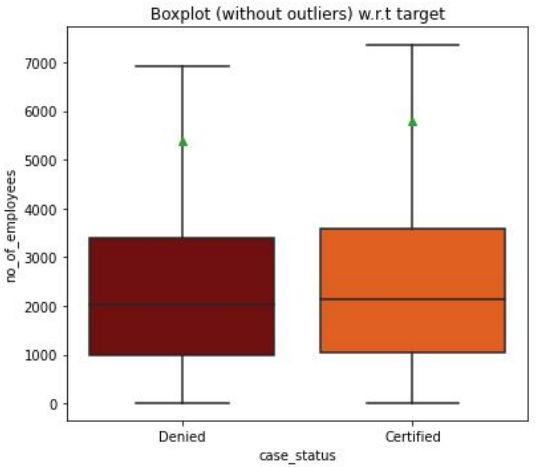




Number of Employees vs Certification Status

- As discussed earlier, the “number of employees” outliers are numerous. The boxplot on the right has all outliers removed and is otherwise unchanged.
- There are cases of application acceptance and denial in the outliers (all the way up to 600k employees), but without the outliers there is a slight trend. **As employee count increases, the likelihood of certification acceptance increases slightly.**
    - This is the 9th most important feature, and has a fairly low relative importance. As such, this isn’t as important as the other features discussed in Bivariate Analysis.





Summary of Exploratory Data Analysis…

- The number of employees at hiring companies has a massive range with many outliers. However, a slight trend can be seen at or below the average: as employee count goes up, certification becomes more likely.
    - The outliers were not removed from the data, as there are many companies above the average, and no companies far above any others.
- Applicants from Asia are the most common, but applicants from Europe are accepted more often.
- Applicants that have only completed High School are far less likely to be accepted than any college-level applicants.
- Applicants looking for a job in Midwest USA are more likely to be accepted than any other hiring region.
    - There are no regions in the USA posing a significant detriment to VISA approval rate.
- Applicants with previous job experience are more likely to be accepted.
- Applicants with no previous job experience are much more likely to be rejected.
- Applicants seeking an hourly pay are less likely to be accepted.

- These observations were made with the final model's Feature Importances in mind.

Model Training

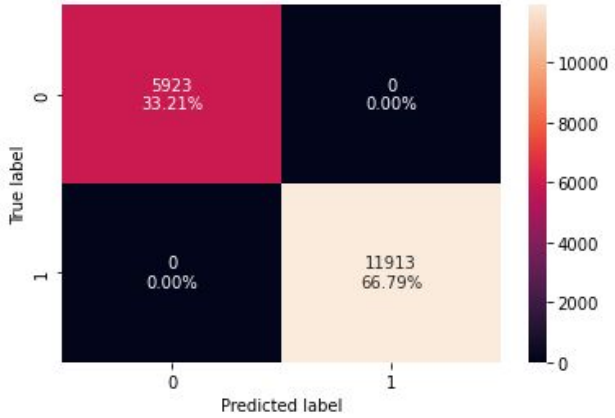
Summaries of the models built in process

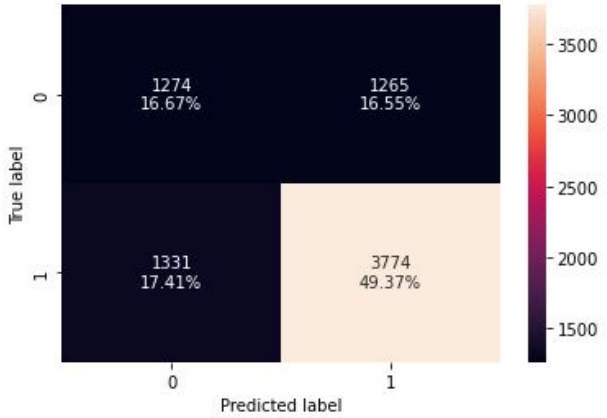
train metrics:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

test metrics:

	Accuracy	Recall	Precision	F1
0	0.660387	0.739275	0.748958	0.744085





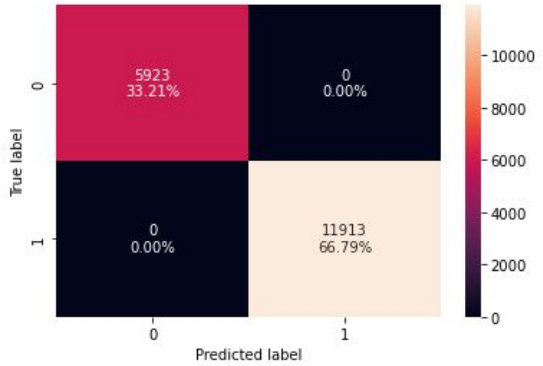
Decision Tree Before Tuning

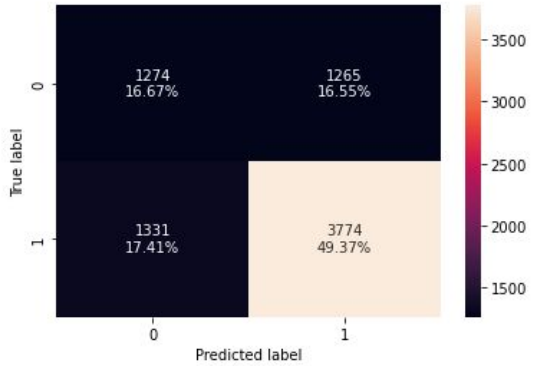
train metrics:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

test metrics:

	Accuracy	Recall	Precision	F1
0	0.660387	0.739275	0.748958	0.744085





- As mentioned at the beginning of the presentation, **F1 score** and **Accuracy** are our most important metrics.
  - The higher the F1, the less likely our model will contain false predictions.
  - The higher the accuracy, the less our model actually predicted incorrectly on the current test set.
- While the train metrics are perfect (100% accurate and 100% F1 score), the test metrics are far from it, which implies high overfitting to the train data.

Decision Tree After Tuning

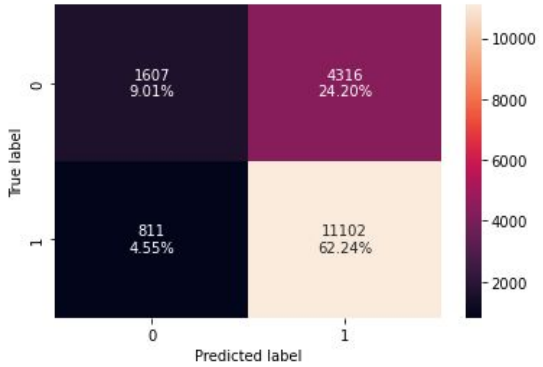
- F1 and Accuracy on the test set has improved post-tuning.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
    - So far, our best model is: Decision Tree Tuned
    - With: F1 of 0.809 on the test set.
    - With: Accuracy of 0.707 on the test set.

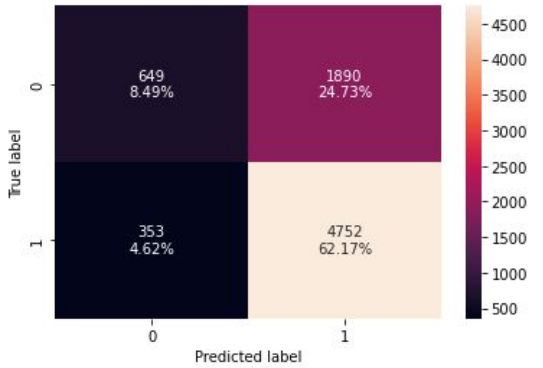
train metrics:

	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411

test metrics:

	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058





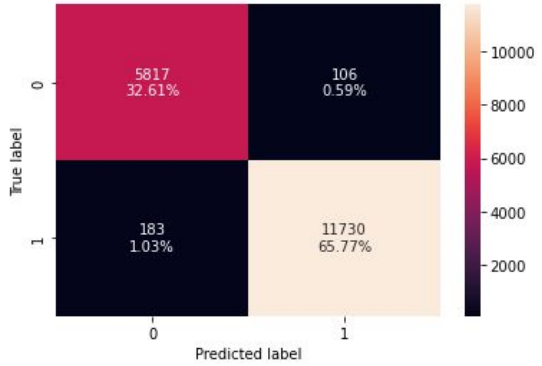
Bagging Classifier Before Tuning

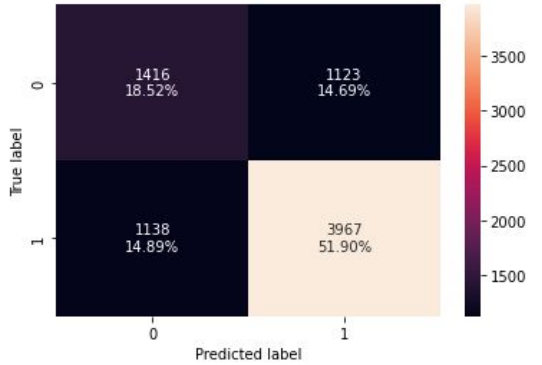
train metrics:

	Accuracy	Recall	Precision	F1
0	0.983797	0.984639	0.991044	0.987831

test metrics:

	Accuracy	Recall	Precision	F1
0	0.704212	0.777081	0.779371	0.778225





- Similar to the Decision Tree, the metrics aren't very good on the test set.
- The Train and Test metrics are far apart, implying a high overfit to the train set.
    - So far, our best model is: Decision Tree Tuned
    - With: F1 of 0.809 on the test set.
    - With: Accuracy of 0.707 on the test set.

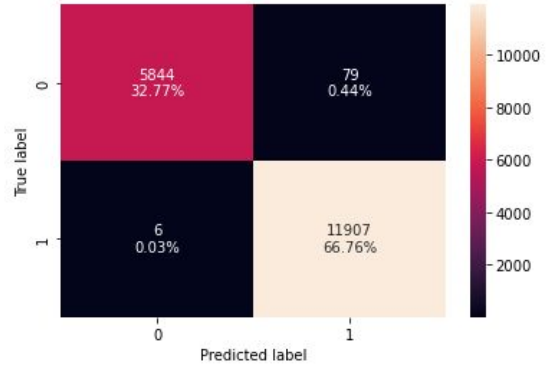
Bagging Classifier After Tuning

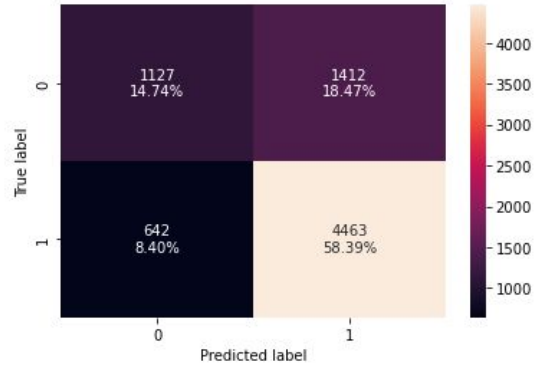
train metrics:

	Accuracy	Recall	Precision	F1
0	0.995234	0.999496	0.993409	0.996443

test metrics:

	Accuracy	Recall	Precision	F1
0	0.731293	0.874241	0.75966	0.812933





- F1 and Accuracy using Bagging is better than the Decision Tree, which was to be expected with a more robust model.
- The Train and Test metrics are far apart, implying a high overfit to the train set.
    - So far, our best model is: Bagging Classifier Tuned
    - With: F1 of 0.813 on the test set.
    - With: Accuracy of 0.731 on the test set.

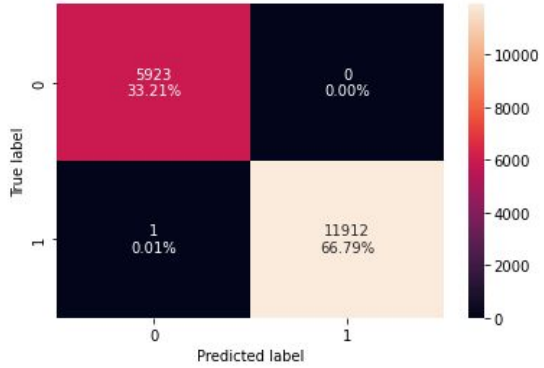
Random Forest Classifier Before Tuning

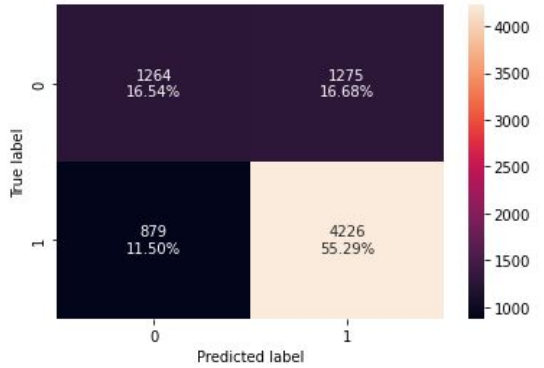
train metrics:

	Accuracy	Recall	Precision	F1
0	0.999944	0.999916	1.0	0.999958

test metrics:

	Accuracy	Recall	Precision	F1
0	0.71821	0.827816	0.768224	0.796907





- Yet again, a massive overfit on the train set. We need to tune the hyperparameters.
- The Train and Test metrics are far apart, implying a high overfit to the train set.
    - So far, our best model is: Bagging Classifier Tuned
    - With: F1 of 0.813 on the test set.
    - With: Accuracy of 0.731 on the test set.

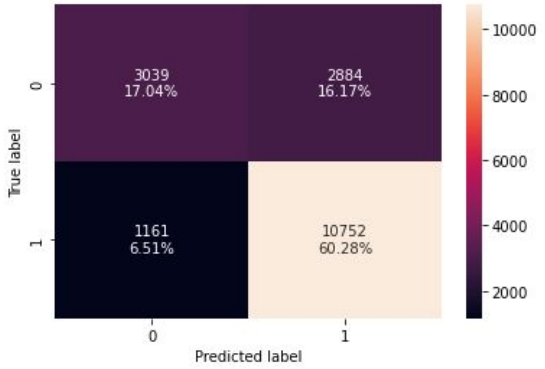
Random Forest Classifier After Tuning

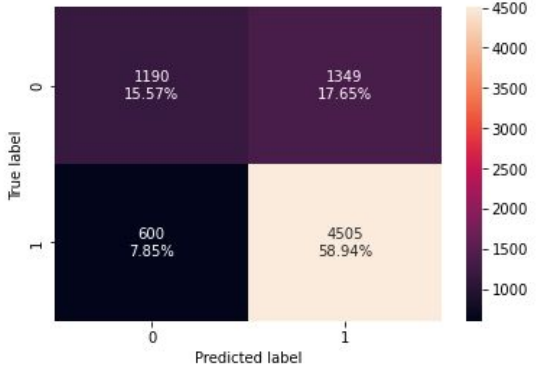
train metrics:

	Accuracy	Recall	Precision	F1
0	0.773211	0.902543	0.788501	0.841677

test metrics:

	Accuracy	Recall	Precision	F1
0	0.745029	0.882468	0.769559	0.822155





- F1 and Accuracy in the Random Forest Classifier is relatively high, higher than the tuned Bagging model.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
    - So far, our best model is: Random Forest Tuned
    - With: F1 of 0.822 on the test set.
    - With: Accuracy of 0.745 on the test set.

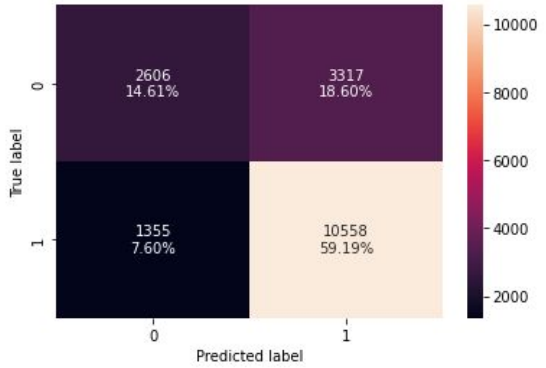
AdaBoost Classifier Before Tuning

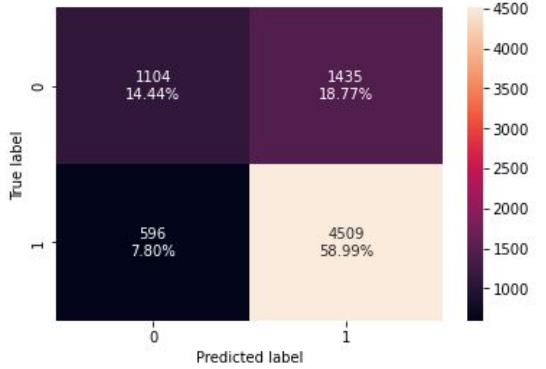
train metrics:

	Accuracy	Recall	Precision	F1
0	0.738058	0.886259	0.760937	0.81883

test metrics:

	Accuracy	Recall	Precision	F1
0	0.734301	0.883252	0.75858	0.816182





- Our F1 and Accuracy is surprisingly good for a fresh model.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
    - So far, our best model is: Random Forest Tuned
    - With: F1 of 0.822 on the test set.
    - With: Accuracy of 0.745 on the test set.

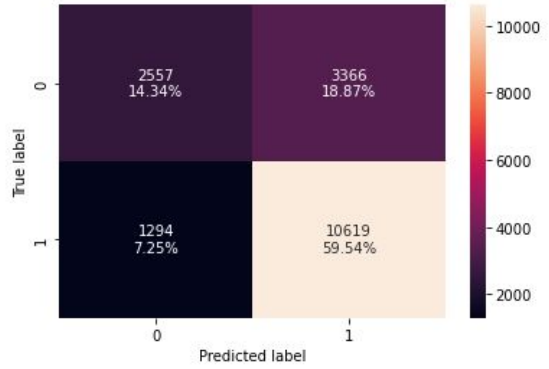
AdaBoost Classifier After Tuning

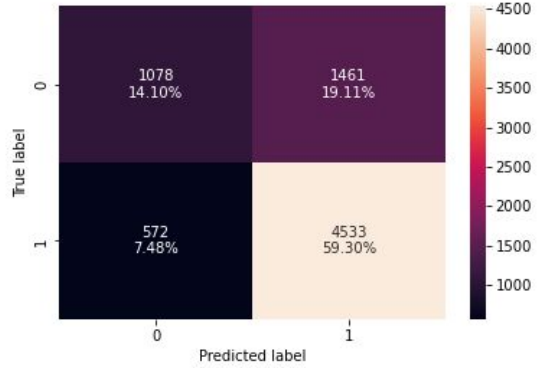
train metrics:

	Accuracy	Recall	Precision	F1
0	0.738731	0.891379	0.759314	0.820063

test metrics:

	Accuracy	Recall	Precision	F1
0	0.73404	0.887953	0.756256	0.81683





- The F1 and Accuracy using our tuned AdaBoost did not improve over our Random Forest Model.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
    - So far, our best model is: Random Forest Tuned
    - With: F1 of 0.822 on the test set.
    - With: Accuracy of 0.745 on the test set.

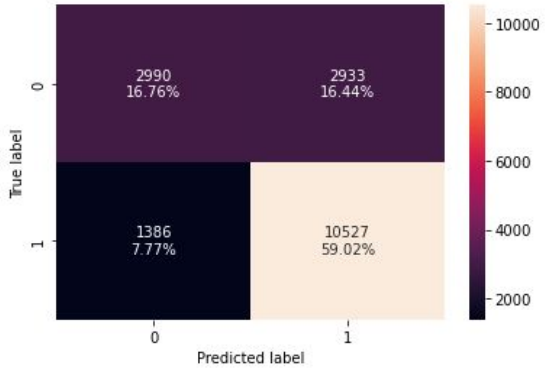
Gradient Boost Classifier Before Tuning

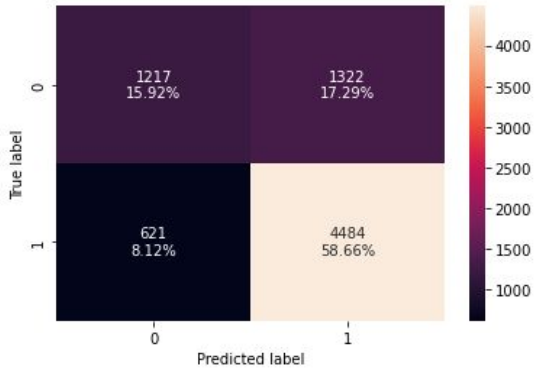
train metrics:

	Accuracy	Recall	Precision	F1
0	0.757849	0.883657	0.782095	0.82978

test metrics:

	Accuracy	Recall	Precision	F1
0	0.745814	0.878355	0.772305	0.821923





- Like AdaBoost, Gradient Boost's untuned parameters have made a fairly generalized model.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
    - So far, our best model is: Random Forest Tuned
    - With: F1 of 0.822 on the test set.
    - With: Accuracy of 0.745 on the test set.

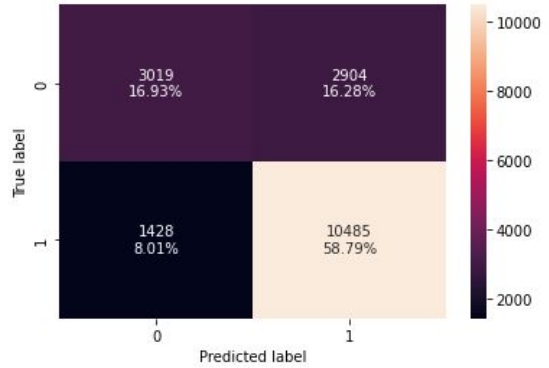
Gradient Boost Classifier After Tuning

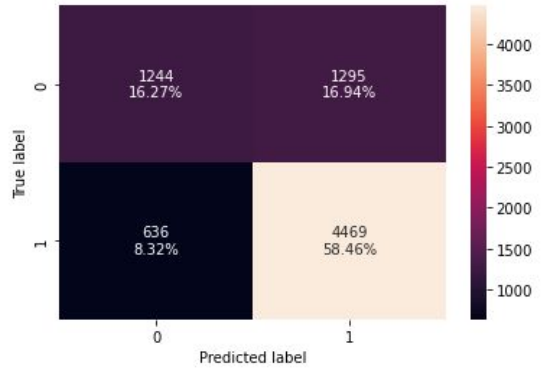
train metrics:

	Accuracy	Recall	Precision	F1
0	0.75712	0.880131	0.783106	0.828788

test metrics:

	Accuracy	Recall	Precision	F1
0	0.747384	0.875416	0.77533	0.822339





- While F1 improved only very slightly, Accuracy improved by 2 thousandths using Gradient Boost.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
    - So far, our best model is: Gradient Boost Tuned
    - With: F1 of 0.822 on the test set.
    - With: Accuracy of 0.747 on the test set.

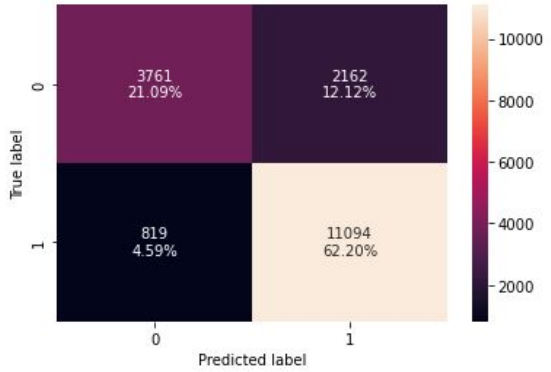
XGBoost Classifier Before Tuning

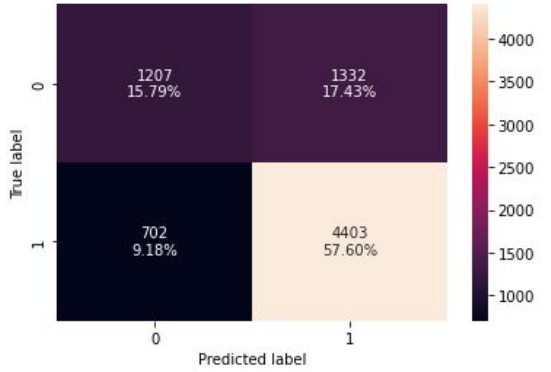
train metrics:

	Accuracy	Recall	Precision	F1
0	0.832866	0.931252	0.836904	0.881561

test metrics:

	Accuracy	Recall	Precision	F1
0	0.733909	0.862488	0.767742	0.812362





- The XGBoost model is slightly more overfit than the previous two Boosting models.
- The Train and Test metrics are far apart, implying a high overfit to the train set.
    - So far, our best model is: Gradient Boost Tuned
    - With: F1 of 0.822 on the test set.
    - With: Accuracy of 0.747 on the test set.

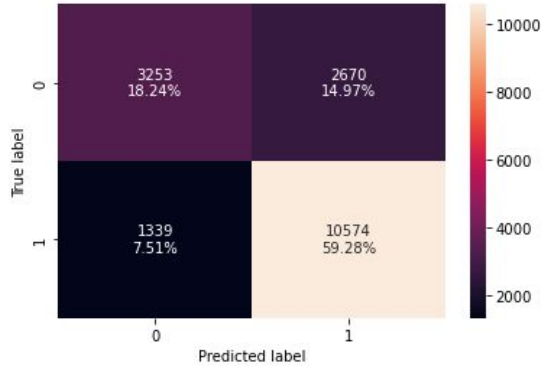
XGBoost Classifier After Tuning

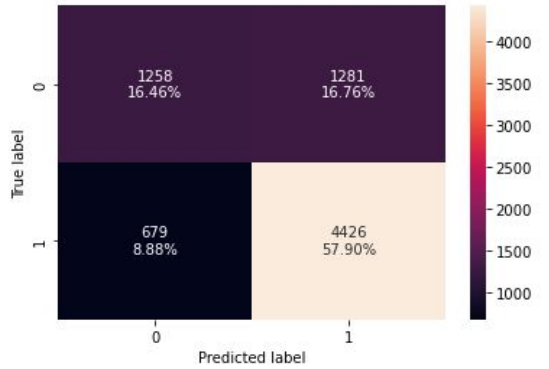
train metrics:

	Accuracy	Recall	Precision	F1
0	0.77523	0.887602	0.798399	0.840641

test metrics:

	Accuracy	Recall	Precision	F1
0	0.74359	0.866993	0.775539	0.81872





- Neither F1 nor Accuracy increased on the tuned XGBoost Classifier.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
    - So far, our best model is: Gradient Boost Tuned
    - With: F1 of 0.822 on the test set.
    - With: Accuracy of 0.747 on the test set.

Stacking Classifier

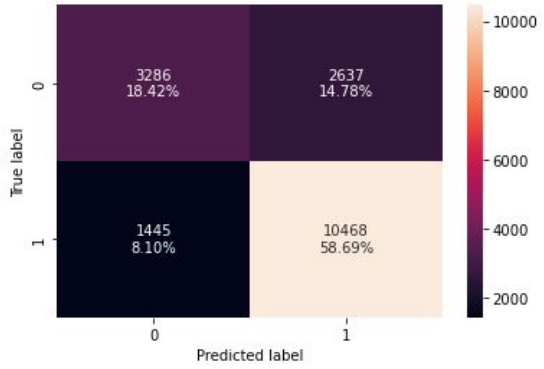
- Neither F1 nor Accuracy increased on our Stacking Classifier.
  - Our estimator models were AdaBoost, Gradient Boost, and Random Forest, all tuned.
  - Our final estimator model was XGBoost tuned.
- The Train and Test metrics are closer together, implying a lower overfit to the train set.
  - So far, our best model is: Gradient Boost Tuned
  - With: F1 of 0.822 on the test set.
  - With: Accuracy of 0.747 on the test set.

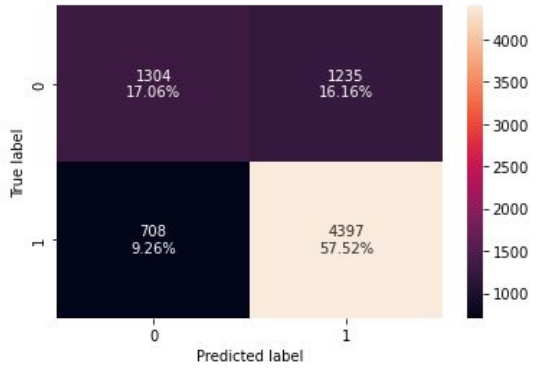
train metrics:

	Accuracy	Recall	Precision	F1
0	0.771137	0.878704	0.798779	0.836837

test metrics:

	Accuracy	Recall	Precision	F1
0	0.745814	0.861312	0.780717	0.819037





Summary of all tuned models (plus our Stacking Model)...

Testing performance comparison:

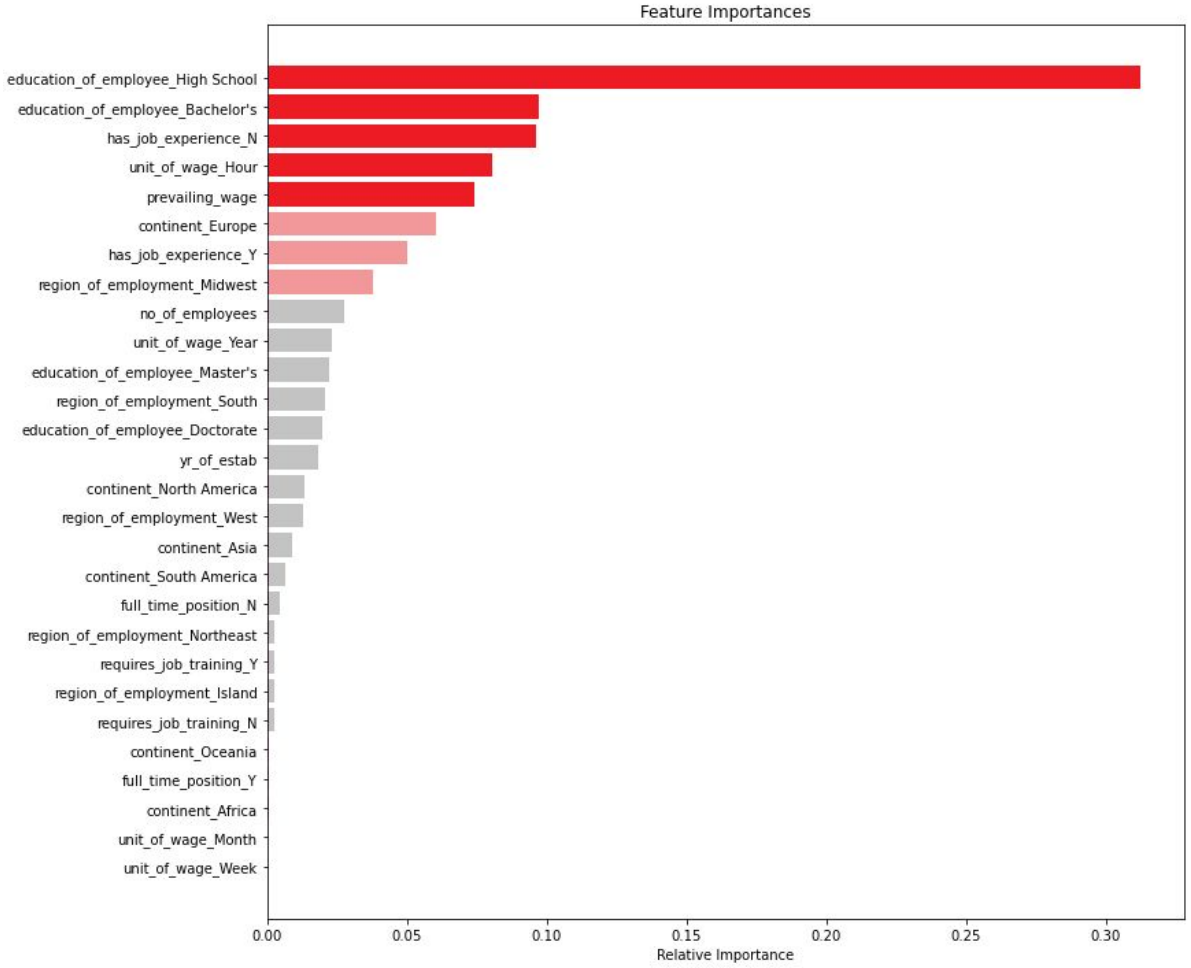
	Tuned Decision Tree	Tuned Bagging Classifier	Tuned Random Forest	Tuned Adaboost Classifier	Tuned GradientBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.706567	0.731293	0.745029	0.734040	0.747384	0.743590	0.745814
Recall	0.930852	0.874241	0.882468	0.887953	0.875416	0.866993	0.861312
Precision	0.715447	0.759660	0.769559	0.756256	0.775330	0.775539	0.780717
F1	0.809058	0.812933	0.822155	0.816830	0.822339	0.818720	0.819037

- Our best models ended up being a **tuned Gradient Boost** and our **tuned Random Forest**.
- Gradient Boost will be used for our Gini/Feature Importances, as it is very slightly better at predicting the certification process.

Feature Importances

(Gradient Boost)

- As we saw in the EDA, all of the top 8 (possibly 9) most important Gini Importance features had influence over the Case Status of foreign workers.
- These top 8 features, High School, Bachelor's, No Job Experience, Hourly Wage, Prevailing Wage, Europe, Previous Job Experience, and Midwest Employers, all have moderate to large effects on the certification acceptance of a applicant.



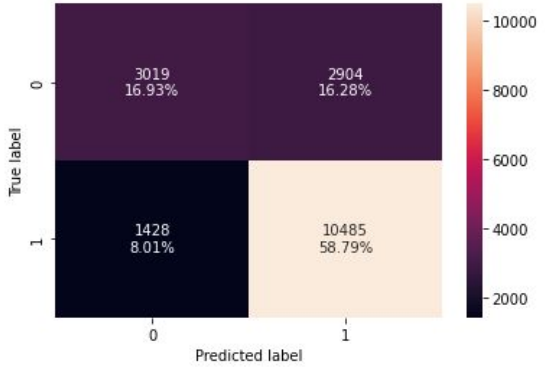
Gradient Boost as our final model

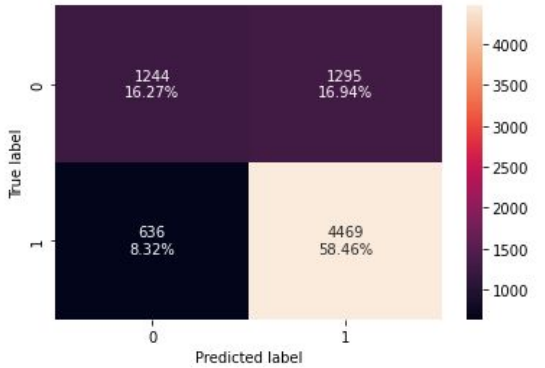
train metrics:

	Accuracy	Recall	Precision	F1
0	0.75712	0.880131	0.783106	0.828788

test metrics:

	Accuracy	Recall	Precision	F1
0	0.747384	0.875416	0.77533	0.822339





- The Recall, or the True Positive Rate, is the number of accepted VISA certifications that the Gradient Boost accurately predicted over the total accepted VISAs.
    - *ie: Out of all of the accepted VISA certifications in our test data, the model correctly predicted 0.875 or 87.5% of them.*
- The Precision is the ratio of correct acceptance predictions to total acceptance predictions.
    - *ie: The model correctly predicted 4469 accepted certifications. It made 5764 total VISA acception predictions, meaning 1295 of its predictions were wrong, and it was only 4469/5764 = 0.775 or 77.5% precise.*
- Because the precision is lower than the recall, this model will accurately predict accepted VISAs more often than rejected VISAs.
    - **Our Gradient Boost will accept more potential rejects than it will reject potential certifications.**
- Our accuracy is lower than our recall, which reinforces that assertion.
- Our F1 is 0.822, which is reflected in our higher recall and lower precision.

Gradient Boost / EDA Summary

- The Gradient Boost model will be accurate around 75% of the time, and favor accepting over rejecting to catch more qualified candidates than it misses.
    - It will predict 87.5% of all qualified candidates for VISA approval, but approximately 22.5% of its VISA approvals will be for candidates that would not have been previously accepted.
- There are two options with this model; Either the model can be used directly for data input, or the stronger Feature Importances can be focused on by personnel to expedite the process of VISA approval manually.
    - For specifics on using the top 9 Feature Importances in our final model, please refer to the EDA summary on slide 18. However, the inferences listed are not set in stone and should be taken with weighted scores rather than accepting or rejecting solely based on one or two features.
    - Using our model will lighten the workload of the OFLC and make it easier to determine potential work VISA candidates.

train metrics:

	Accuracy	Recall	Precision	F1
0	0.75712	0.880131	0.783106	0.828788

test metrics:

	Accuracy	Recall	Precision	F1
0	0.747384	0.875416	0.77533	0.822339