Project 5: EasyVisa

LUKE HARRIS

Contents

Executive Summary

Business Problem Overview

EDA Results

Data Preprocessing

Model Performance

Executive Summary

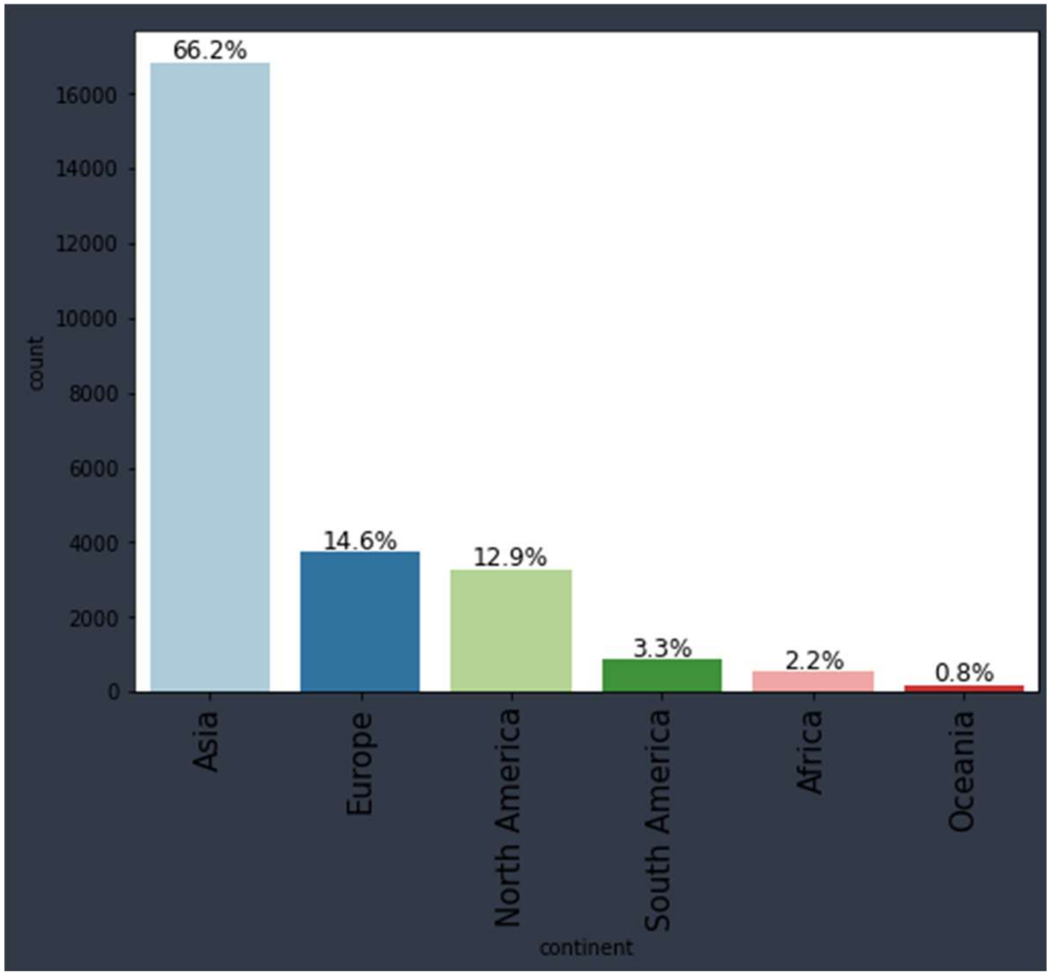
With the huge boom in applications received, I see that it is becoming impossible to comb through each individual one to find qualified candidates who have a higher chance of securing a visa. We at EasyVisa have produced a machine learning based solution that will make it much easier to find these candidates. It is our objective to provide data-driven solutions to facilitate the process of visa approvals. We look for the most important drivers that impact the approval or acceptance of visas and make sure that your candidates match them.

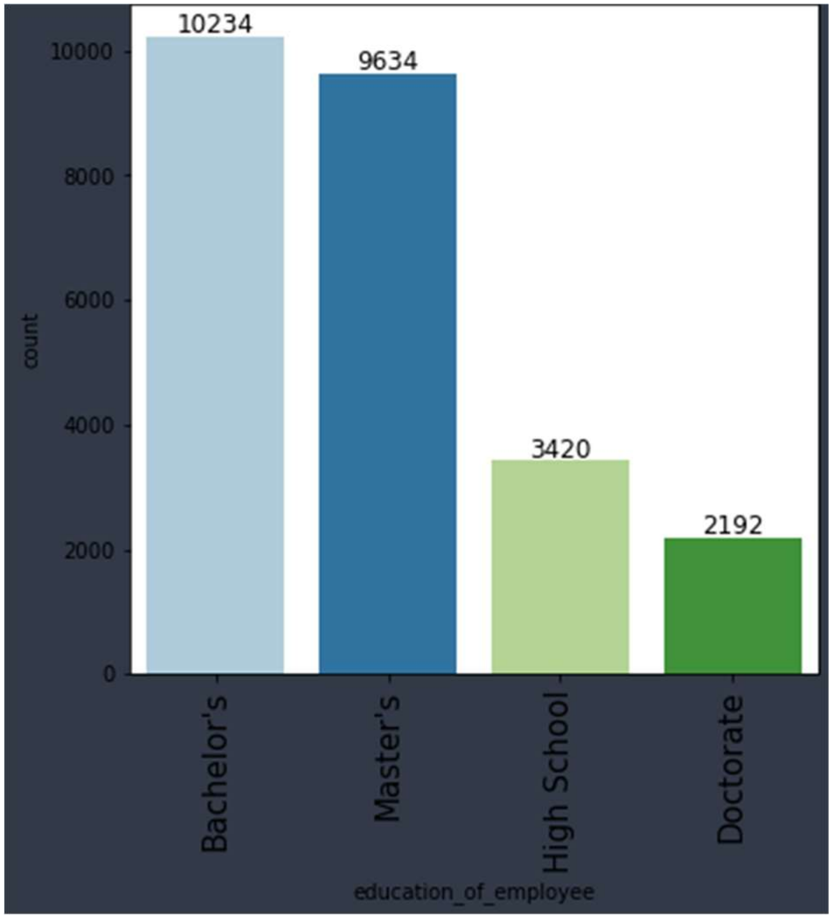
Exploratory Data Analysis

Exploratory data analysis (EDA) is done by using a single variable. This analysis is to see what our data looks like and see which pieces of data are important to compare to each other. This gives an overall view of the data.

Where are Applications Coming From?

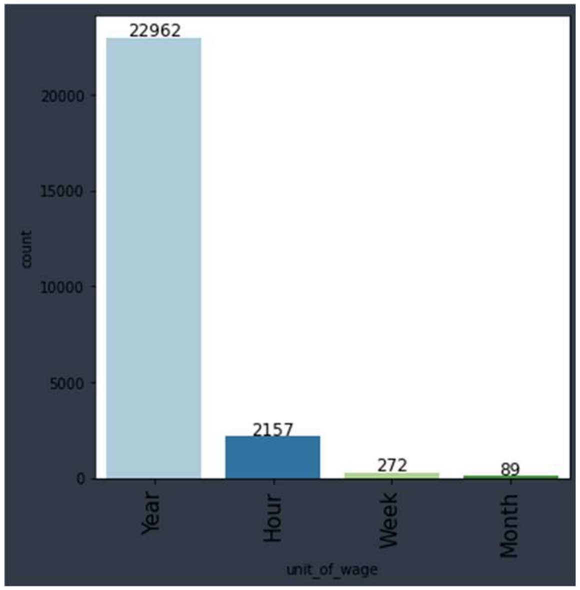
As we can see here, a vast majority of our applicants are coming from Asia, with Europe and North America far behind in third and fourth.





What is Their Level of Education?

Most people applying for visas in the united states have a higher level of education. Most have at least a bachelors degree with master's degree coming close behind.



Unit of Wage

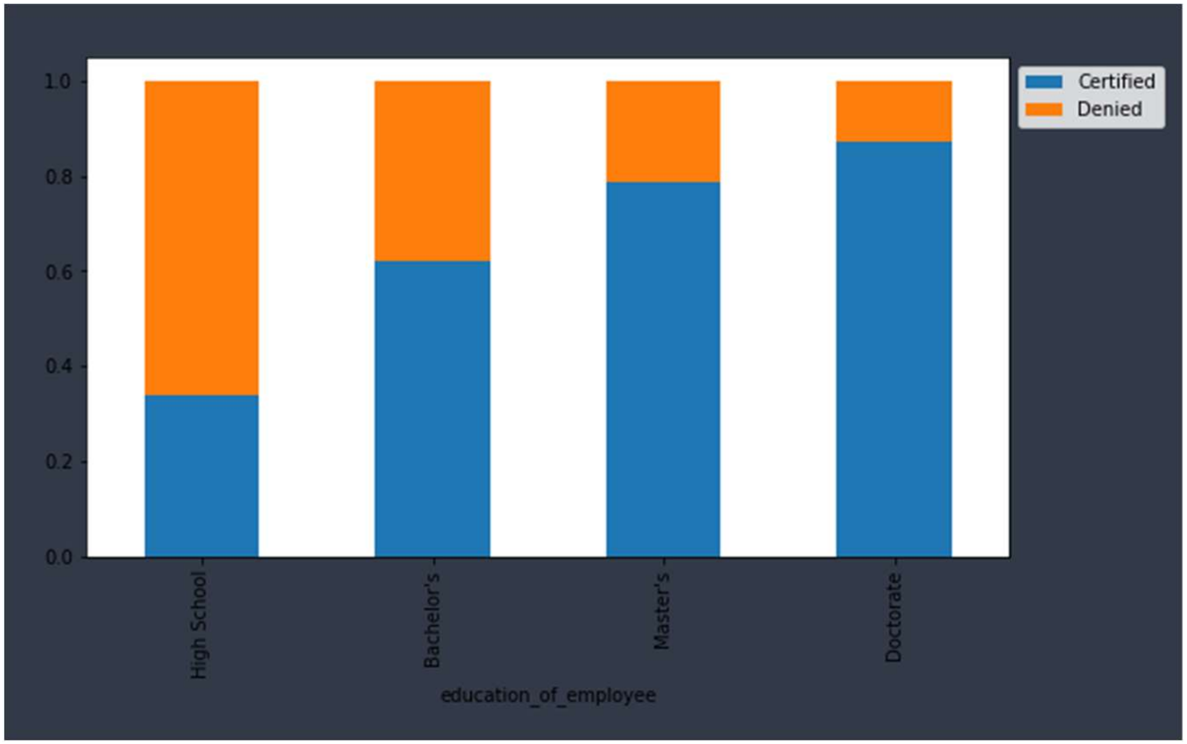
The unit of wage is the way that the applicant will be getting paid. Here, we see that most people who are applying for visas are coming in to do salary jobs.

Bivariate Analysis

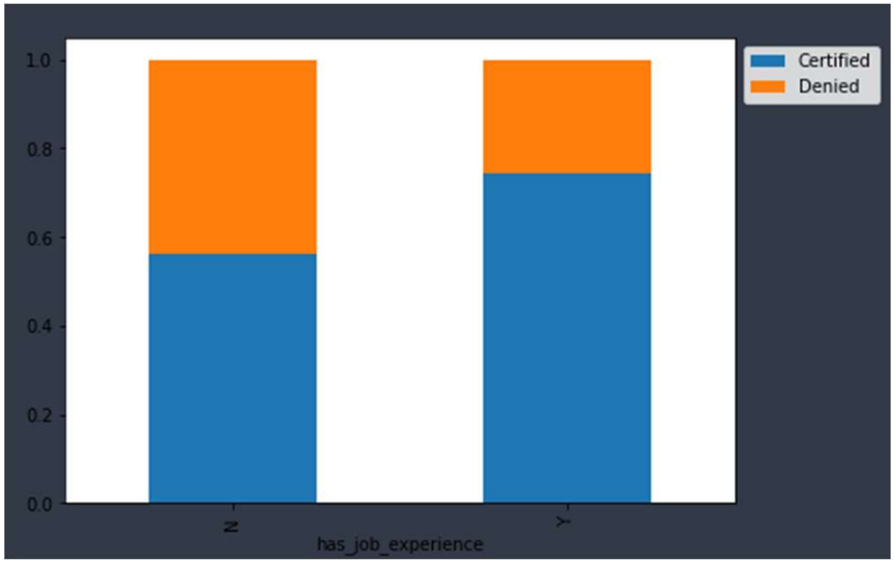
After we complete our EDA, we look at multiple variables to come to a narrower conclusion. We can tell a deeper story by comparing variables together and come to better conclusions before we see what the models can do.

Education vs Case Status

As we see here, applicants with only a high school education are highly unlikely to be certified. As an applicant's level of education increases, the more likely they are to be certified.



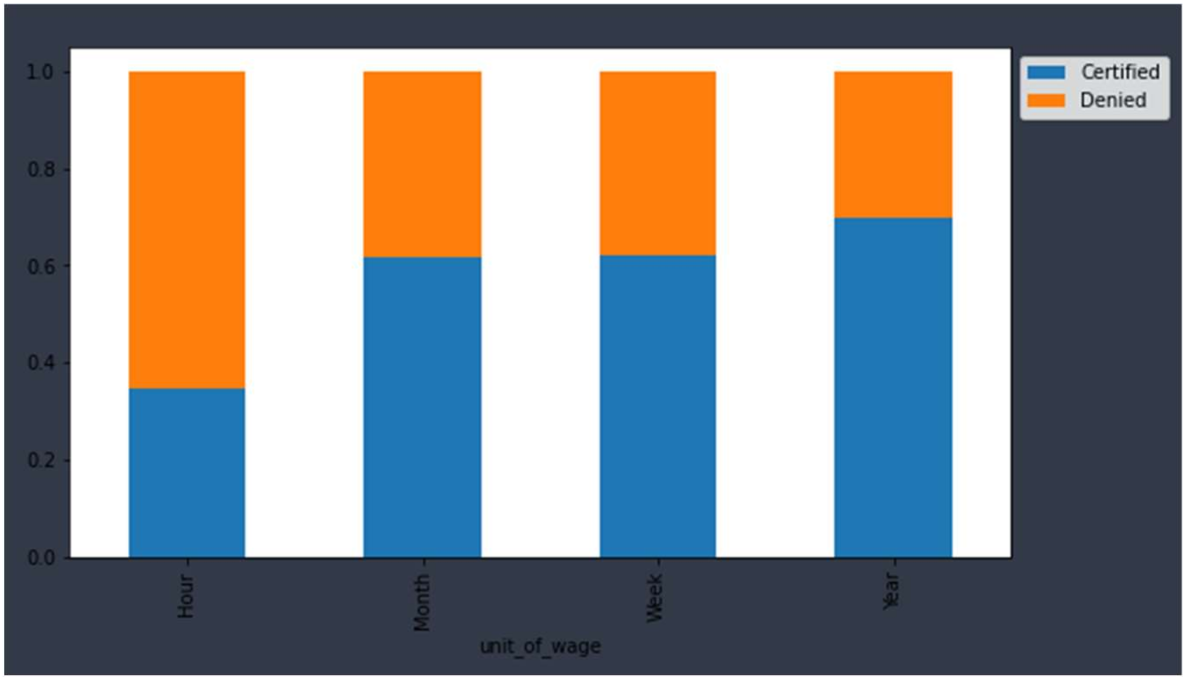
Job Experience vs Case Status

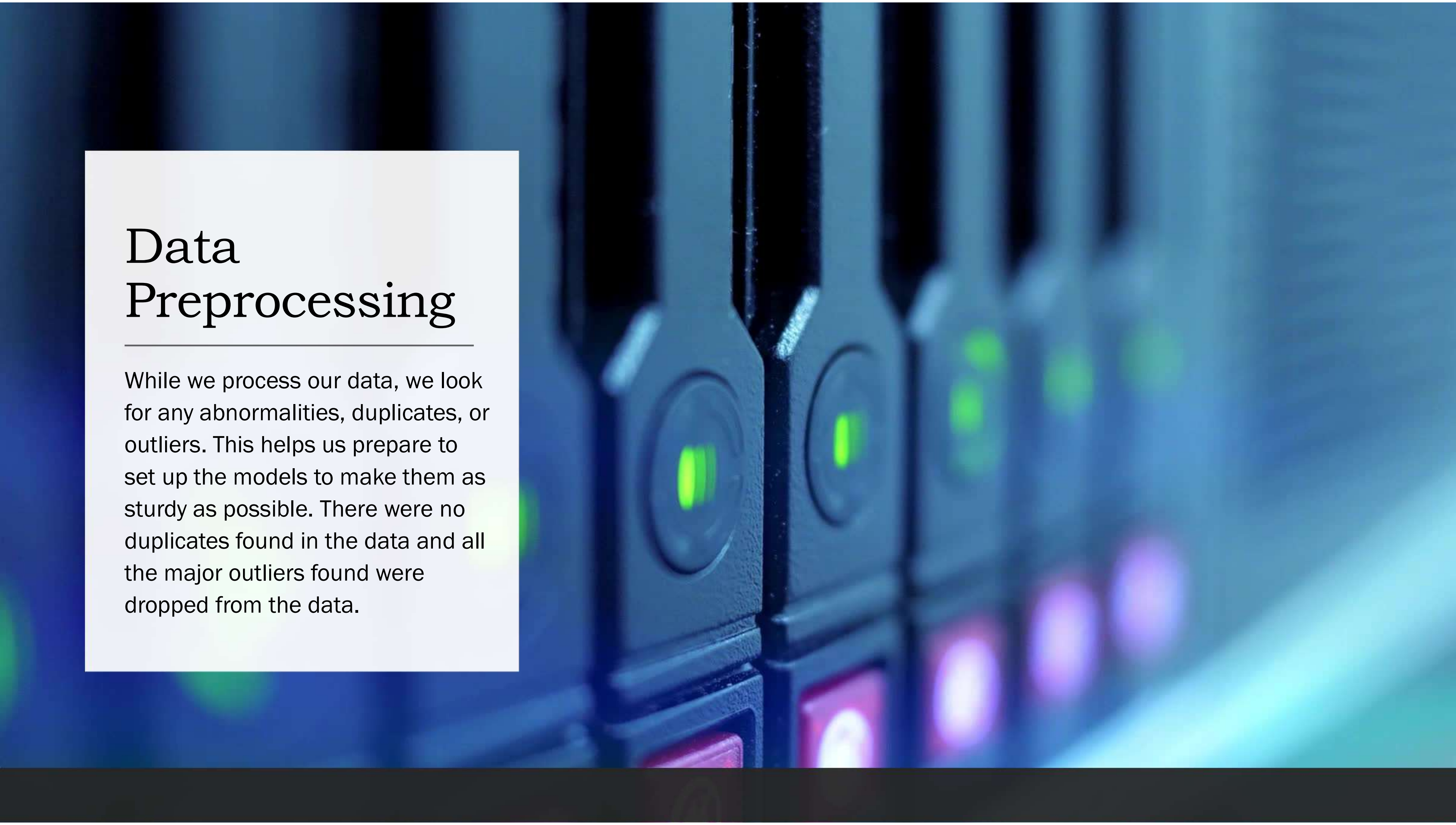


Here we see how job experience plays a role. The applicants with previous experience are more likely to become certified. While it isn’t as significant as the education level, there is definitely a gap between the two.

Unit of Wage vs Case Status

This is graph is quite interesting because earlier we saw that the second most applicants for visas are for hourly jobs. Yet, the are the least likely to be approved. Salary, which was by far the most applied for is the most likely to be certified.



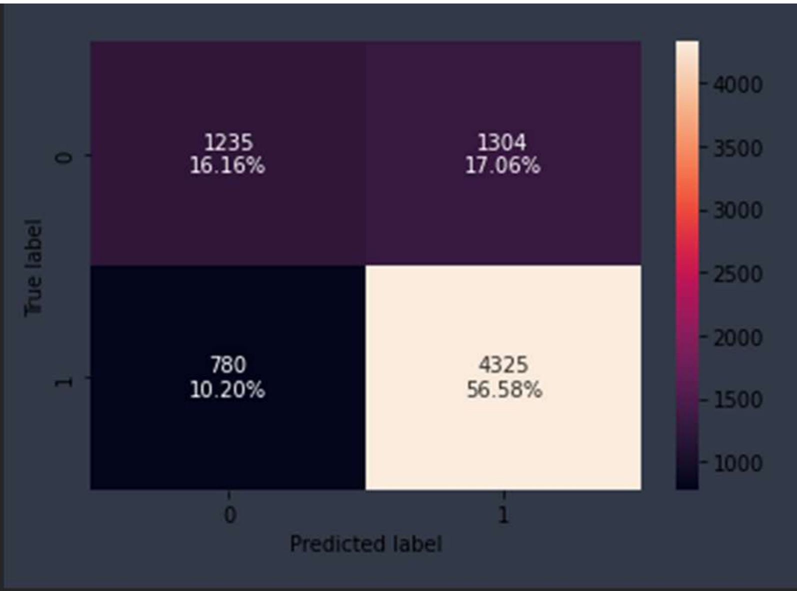


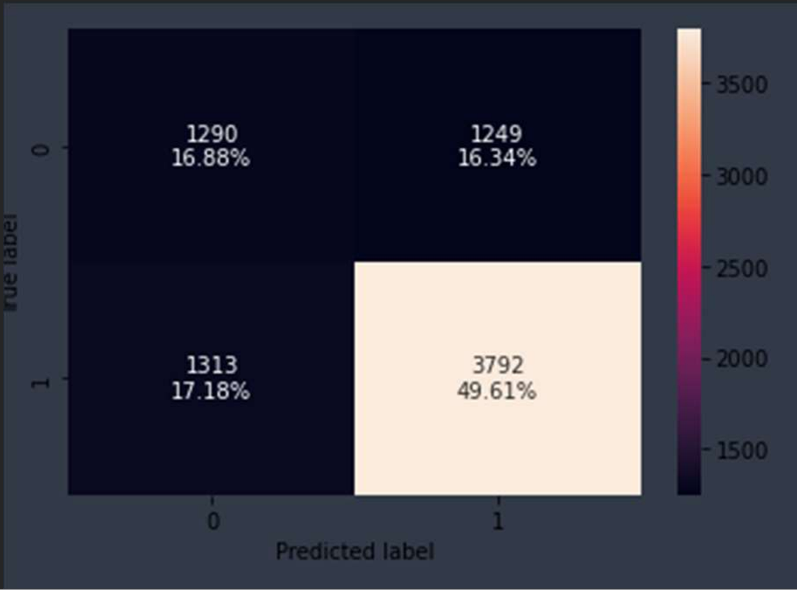
Data Preprocessing

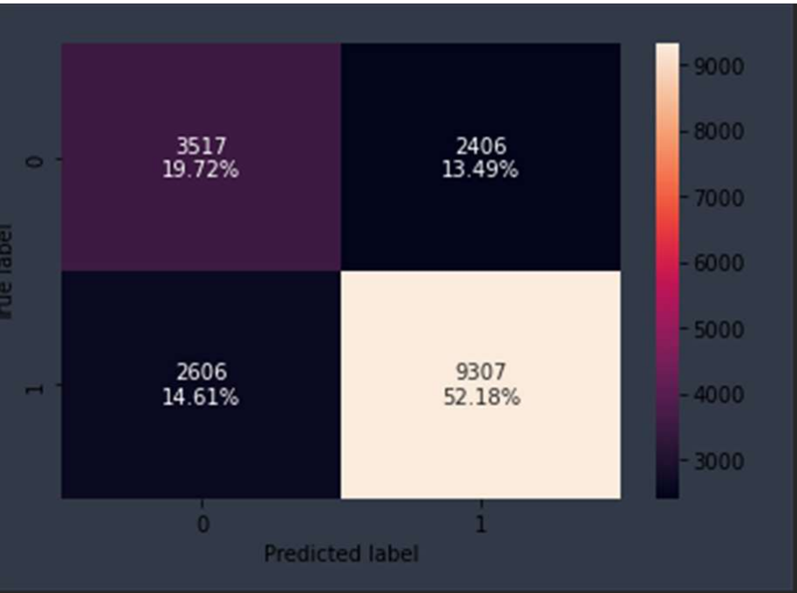
While we process our data, we look for any abnormalities, duplicates, or outliers. This helps us prepare to set up the models to make them as sturdy as possible. There were no duplicates found in the data and all the major outliers found were dropped from the data.

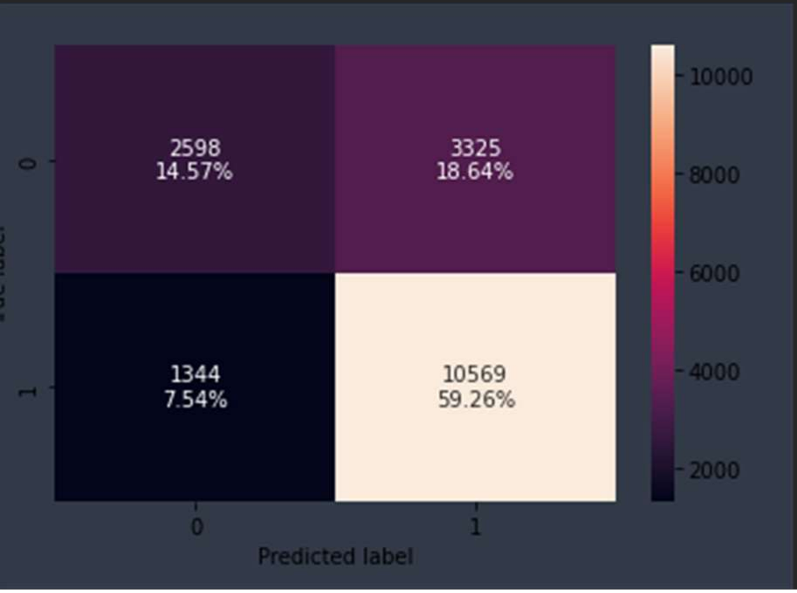
Building the Model: Bagging

Bagging is a model building technique that takes weak learners and trains them in parallel. This reduced variance in the noisy dataset. Bagging uses random samples with replacement, meaning a single data point can be used more than once. Here on the top, we can see the random forest test data, and below, we see the bagging classifier for the test data. Both models are about the same in finding true negatives and true positives, but the random forest is much more accurate in total.



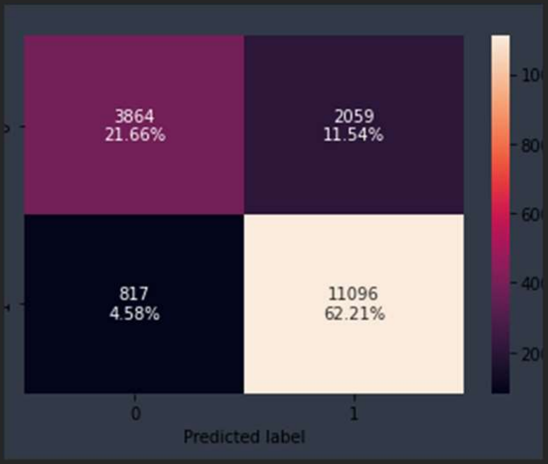


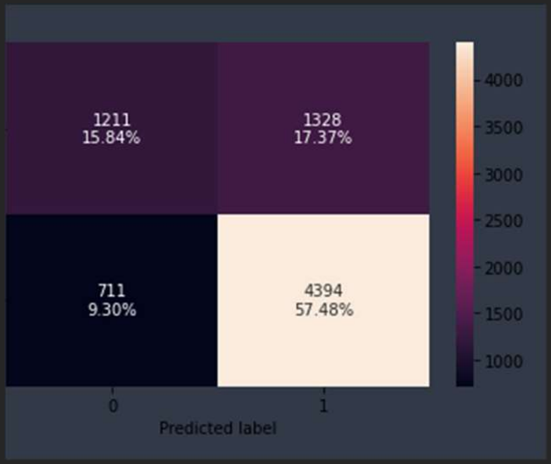


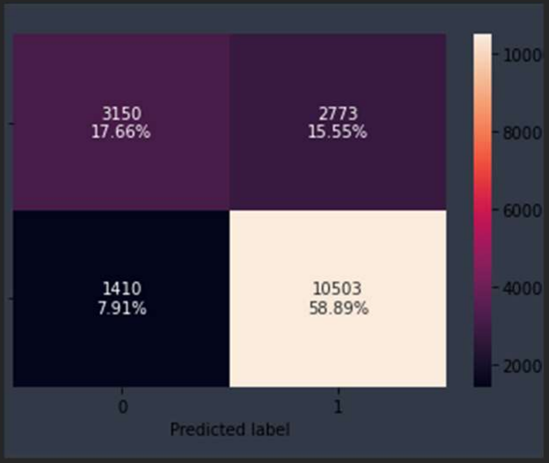


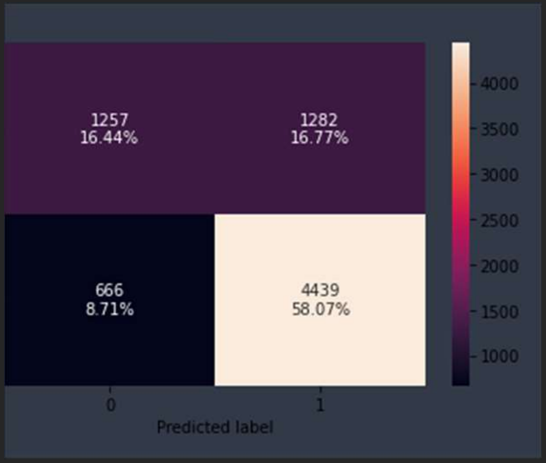
Building the Model: Boosting

Boosting is another ensemble technique used in model building that develops strong classification models from weak classifiers. They focus on errors that occur during prediction. After the initial model is build, the boosting algorithm applies weights sequentially. This means in each new model there is a new iteration with weights increased for misclassified data points. Here we see the tuned (top) vs the untuned (bottom) confusion matrices for our Adaboost algorithm. We see that the tuned version is much better at predicting true corrects but the untuned version is better at finding true negatives. I would suggest the tuned model for your business since it is more important to find a true candidate who will pass.







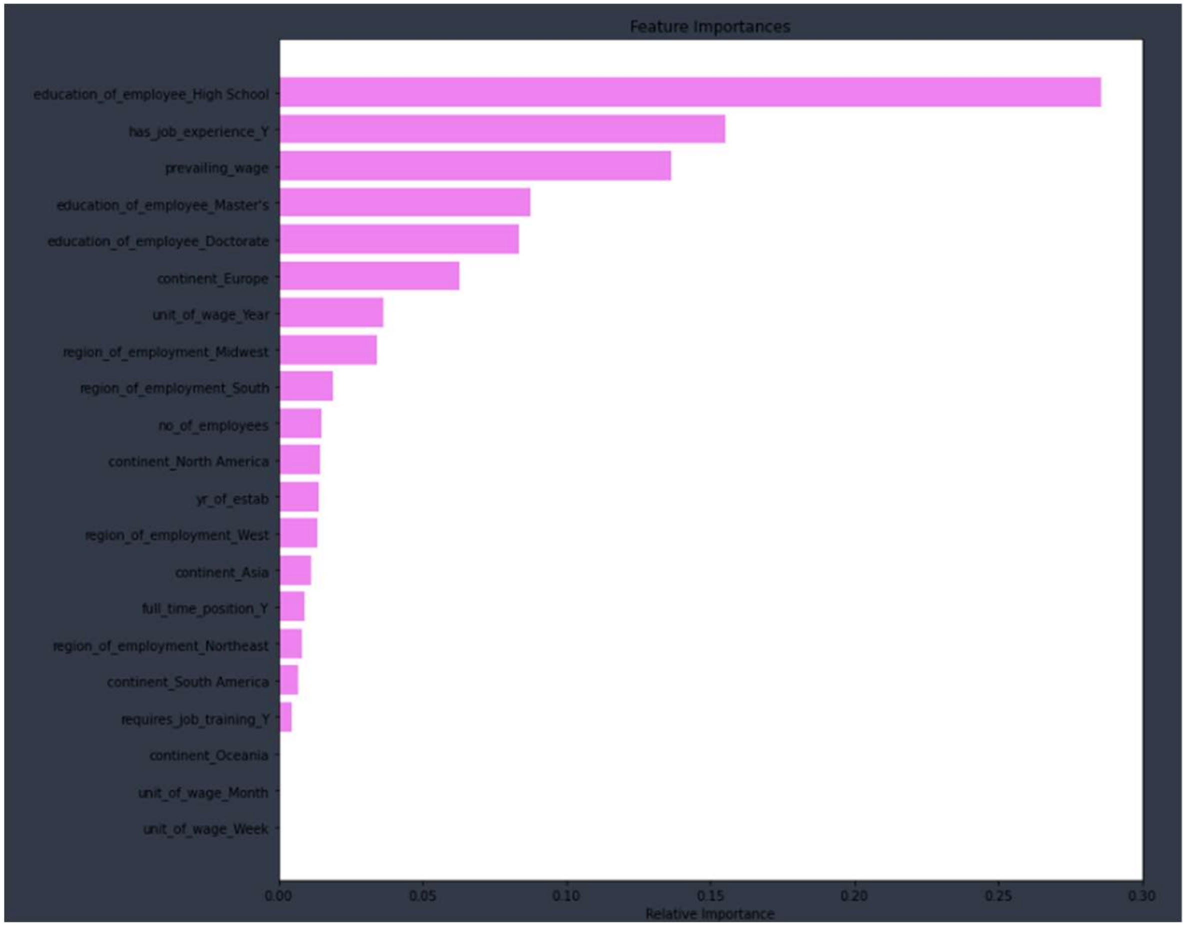


Building the Models: XGBoost

Extreme Gradient Boosting (XGBoost) is the strongest model that we can use. It is a scalable, distributed decision tree. It gives parallel tree boosting and is the strongest in regression, classification, and ranking problems. On the top we see the untuned xgboost model (left, train; right, test). The bottom shows the tuned xgboost model (left, train; right, test). To me, the untuned looks the strongest since it is very effective at finding true negatives and true positives very reliably.

Feature Importance of Combined Models

Here, we see the most important features. As you can see it is extremely unlikely that someone with only a high school education will be certified. Next, having previous job experience looks very good to the visa office as they are more likely to become certified. Education comes in 4th and 5th to show that higher educated individuals are more likely to become certified. I'd recommend looking for highly education individuals with experience.


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