VISA Approval – EasyVisa

Applicant Certification and Processing Approvals

Machine Learning (Ensemble Techniques): Module 5

March 2024

AGENDA

- ❑ Executive Summary
- ❑ Objective & Solution Approach
- ❑ Data Definitions & Details
- ❑ Data Preprocessing – Univariate & Bivariate Analysis

  Outliers- Feature Engineering
- ❑ Model Building & Improvement - Bagging
- ❑ Model Building & Improvement – Boosting

  (AdaBoost, Gradient Boosting, XGBoost & Stacking)
- ❑ Insights & Recommendations

Executive Summary

Staying competitive in the marketplace revolves around having the best talent pool of employees. Within the U.S. companies are vying for the best of the best in terms of hard-working, creative and very qualified people. The Immigration and Nationality Act (INA) was enacted to allow us to look abroad for this level of talent. This act permits foreign workers to come to the U.S. to work on a temporary or permanent basis.

The act also protects U.S. workers in terms of wages and working conditions by ensuring compliance and statutory requirements when foreign workers are hired. Immigration Programs are overseen by the Office of Foreign Labor Certification (OFLC).

The OFCL grants certifications to employers looking to bring foreign workers to the U.S. The employers must show that there are not sufficient U.S. workers available to perform the job as well as meeting wage requirements for the type of position they are attempting to fill.

Approach

EasyVisa was hired to provide a data-driven solution to assist in the growing number of OFCL applications. The overall number of applications received in FY 2016 was 1,669,957 and the amount processed was 775,979. The nine percent increase in applications over one year has made the task of reviewing candidates overwhelming.

EasyVisa will provide a Machine Learning solution to assist in creating a shortlist of candidates that have a higher chance of VISA approval.

Objectives:

Using a Classification Model we will Analyze the Data Provided:

1. Facilitate the process of Visa Approvals

2. Recommend a suitable profile for the applicants that a VISA should be Certified or Denied, based on the Feature Importances of Case Status

Data Definitions:

- **case_id:** ID of each visa application
- **continent:** Continent the employee
- **education_of_employee:** 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:** 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

Dataset Details:

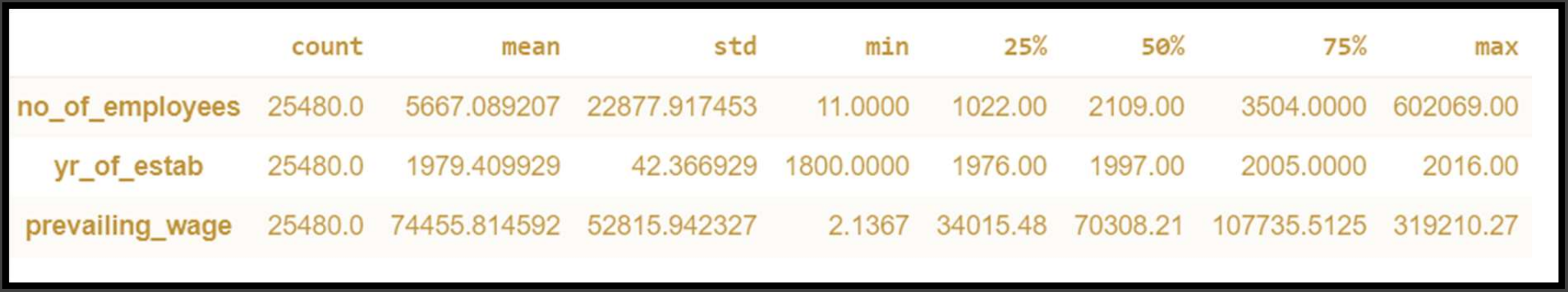
- Size: 12 Categories - 25,480 Dataset
- Criteria: Suitable Profile for applicants Certification or Denial of Work Visa and the Significant Factors
- Missing or Duplicate Values: No Duplicate or Null Values within the dataset
- EDA Analysis: Univariate, Bivariate, Outlier Check, Feature Engineering, Data Preparation
- Bagging: Decision Tree, Bagging Classifier, and Random Forest – Commentary
- Boosting: Adaboost & Gradient Boost (No XGBoost) – Commentary
- Model Improvement & Performance: Commentary on Post-Tuning of AdaBoost & Gradient Boosting Classifier on metrics to improve model performance – Building Stacking Classifier
- Compare Model Performance on a variety of different Metrics

EDA - Univariate Analysis

- When Averaging out the Number of Employees within each company's we see it is 5,667
- On the Max end of the spectrum some come companies employee over 600,000

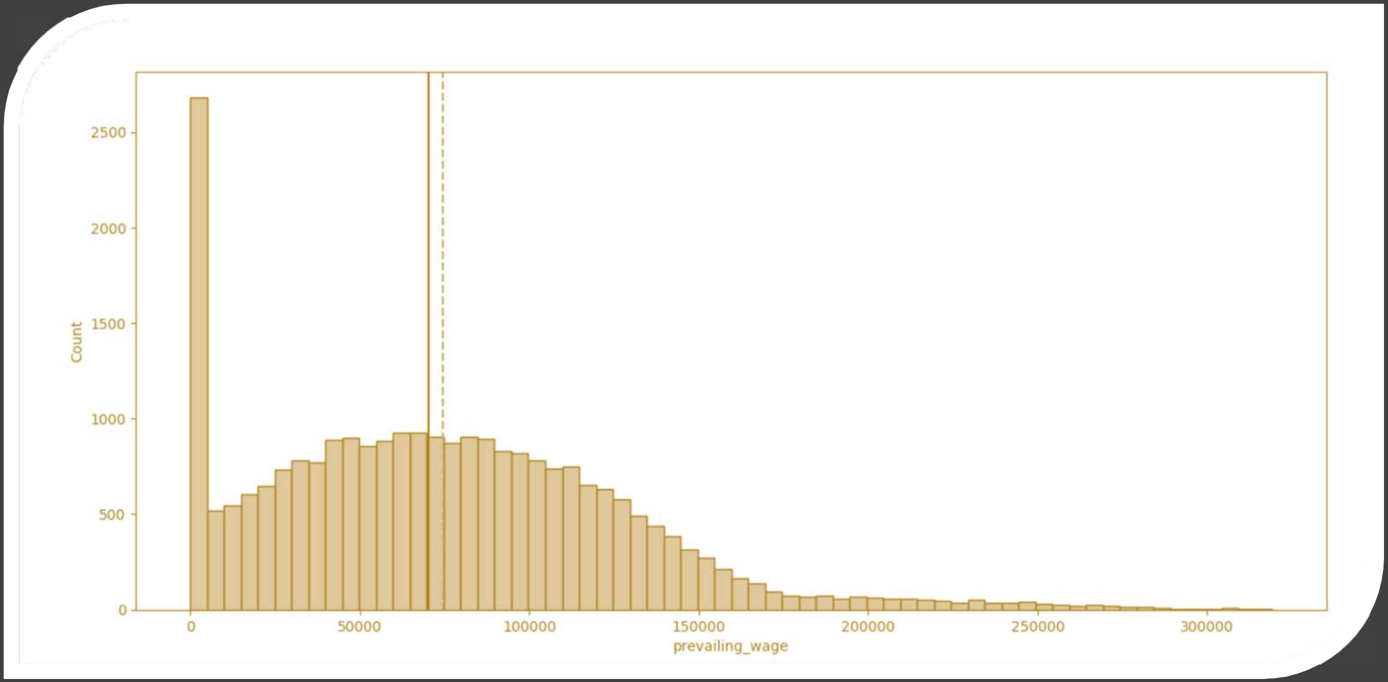
- There is also a wide range between the Year the Company was Established the Average was in 1979 and 75% of the Companies were from 2005

- Prevailing Wage will show to be one the most important features as we move further along
- The Average Wage was $74,455 and the Max Wage was $312,210
- This puts the Standard Wage amount at $52,815



	count	mean	std	min	25%	50%	75%	max
no_of_employees	25480.0	5667.089207	22877.917453	11.0000	1022.00	2109.00	3504.0000	602069.00
yr_of_estab	25480.0	1979.409929	42.366929	1800.0000	1976.00	1997.00	2005.0000	2016.00
prevailing_wage	25480.0	74455.814592	52815.942327	2.1367	34015.48	70308.21	107735.5125	319210.27

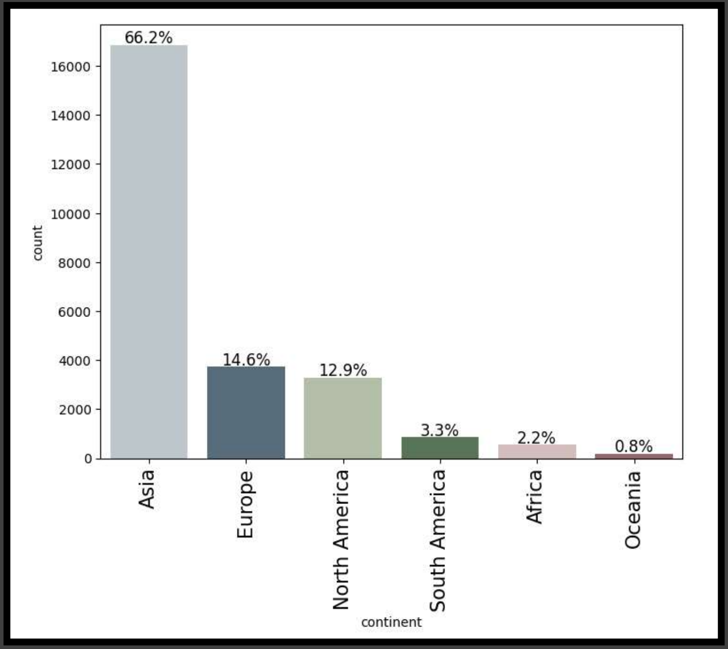
Univariate Analysis

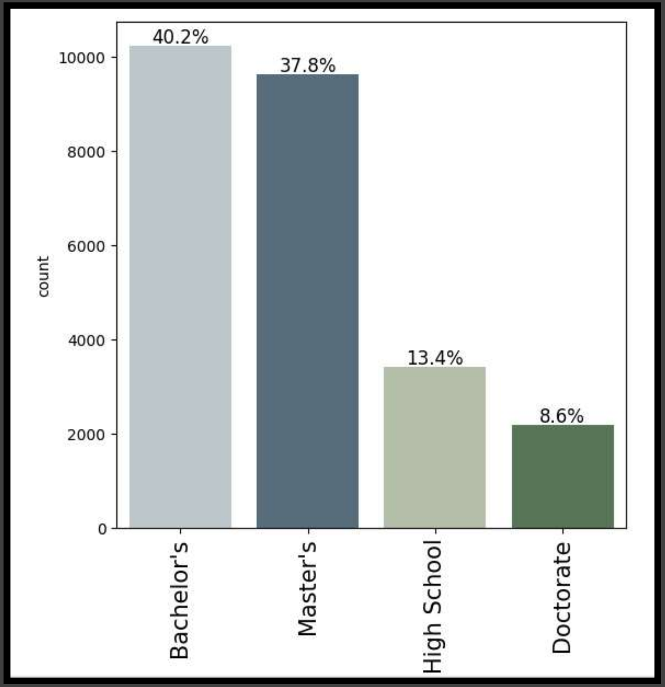


- Barplot on Prevailing Wage shows the Mean Wage of $74,445, and how we have a right-skewed distribution. There are man Outliers in this Variable, which we will account for and use a actual values.

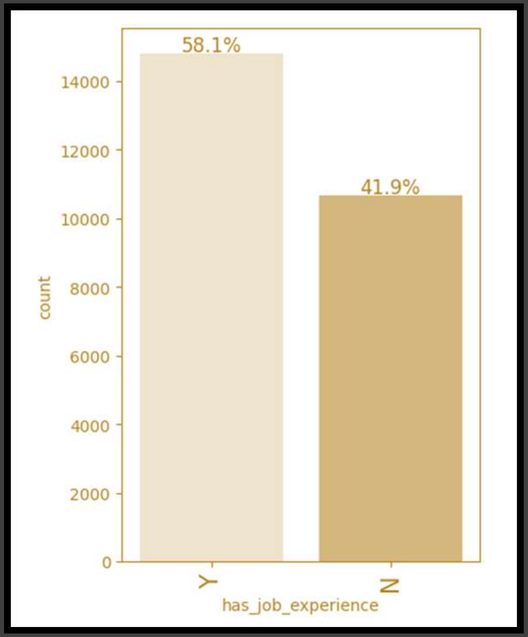
Univariate Analysis

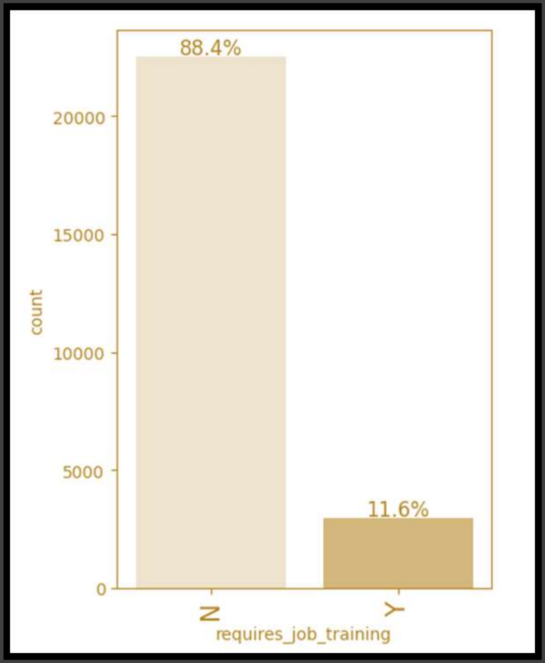
There were many categories that were taken into account throughout our analysis of the data. A series of BarPlots will show factors that we determined were important in determining Visa Certification or Denial

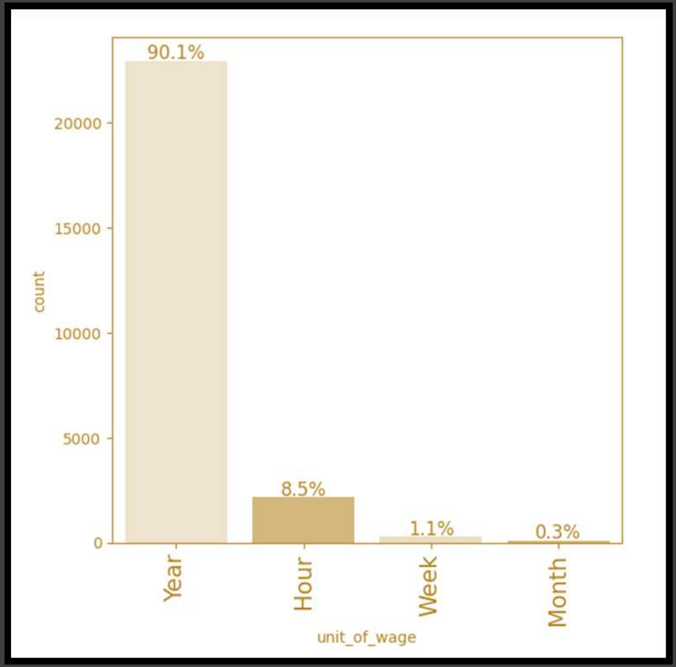




Univariate Analysis







Summary - Univariate Analysis

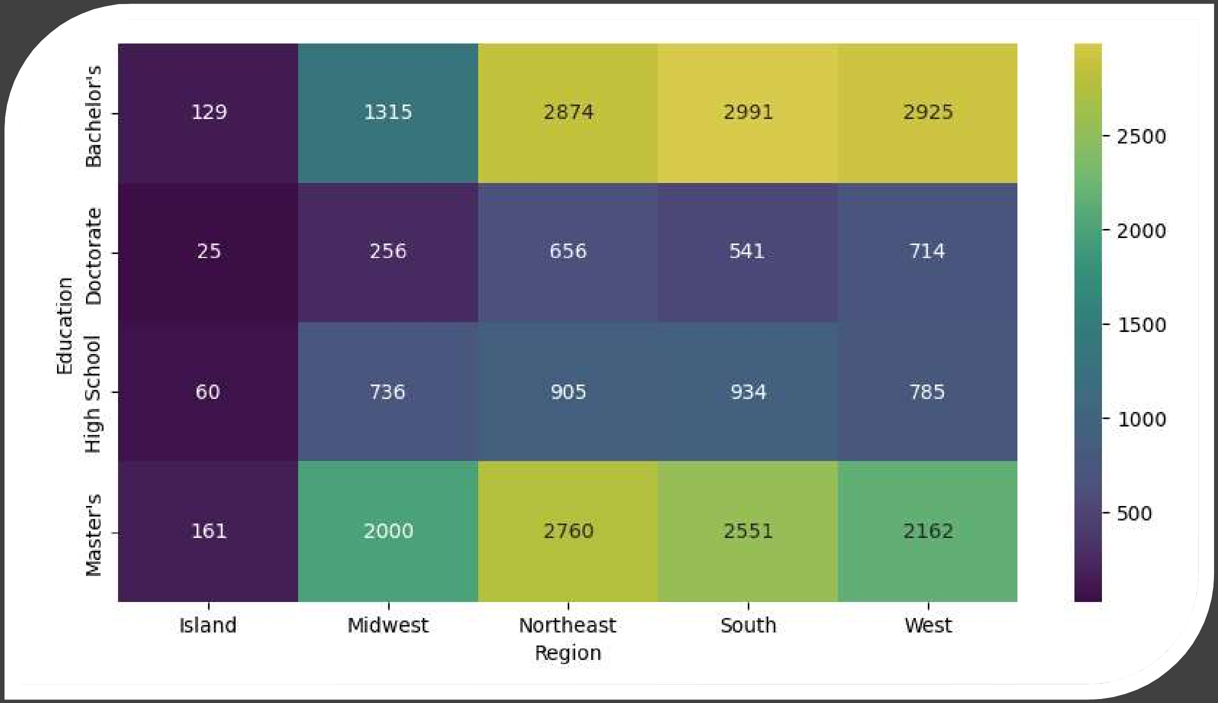
- Asia, by far had the most employee applications for Visa at 66.2% and accounted for over 16,000 people. Europe and North America show a second and third but are less than 50% less applications than Asia

- Employee Education is vital to gaining employment within the United States. Higher Education of Bachelor's and Master's Degrees were the majority of what companies desired in looking for candidates outside the U.S.
    - Approximately 10,000 applicants had a Bachelor's Degree totaling 40.2%
    - A Master's Degree was a very close 37.8%
    - Less than 2,500 applicants had Doctorate Degrees, which may factor into Wage

- Comparing the two categories of Having Job Experience or Requiring Job Experience
    - Has Experience could mean that they work in the field and would only need to learn the company way of doing the job. This category shows that over 14,000 or 58.1% of applicants were within the field they were applying for in the U.S
    - Applicants that did not require job training was 88.4% and over 20,000 people. This definitely matches up in proving that qualified applicants are applying for positions in the U.S. They need minimal training to get started and are expected to rely on their workplace knowledge.

- Unit of Wage shows that 90.1% of employers offer an Annual Salary. Only 8.5% or approximately 2,000 companies offer an Hourly Wage. These types of positions are usually less permanent.

EDA - Bivariate Analysis



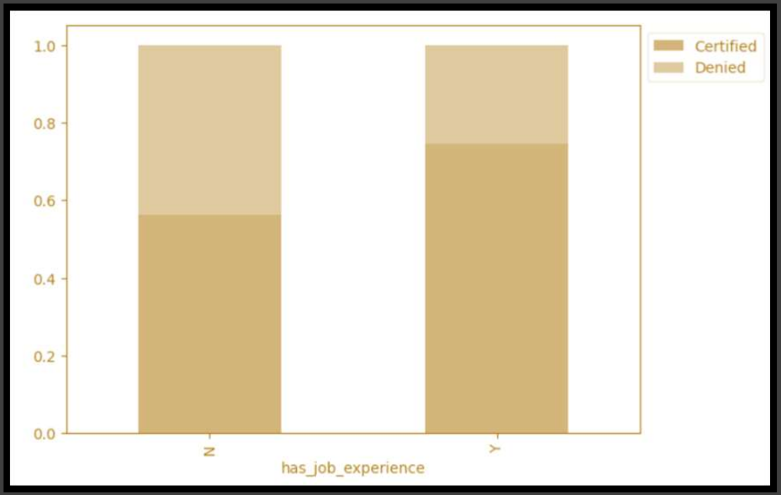
Correlation to Applicants

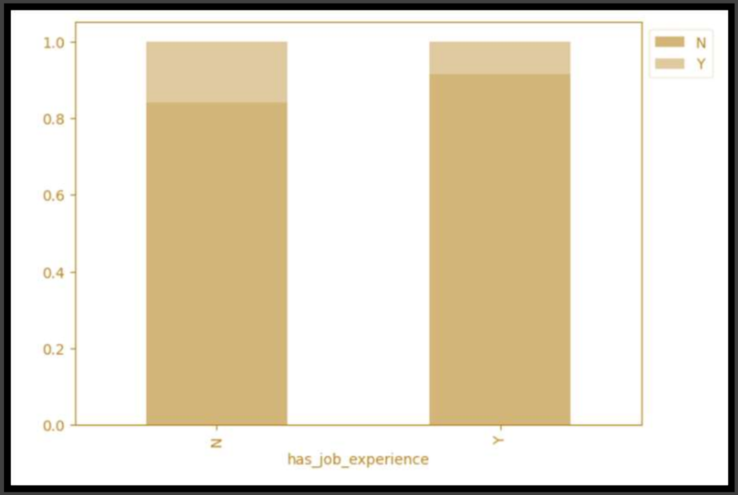
- 2,500 & Above: Bachelor & Master's Degree: South, West & Northeast
- 2,000 – 2,400: Master's Degree all regions
- 800 - 1,500: High School & Doctorate: All Regions
- 750 & Under: Approx. 256 with Doctorate in the Midwest and on an Island most had Bachelor's or Master's Degree

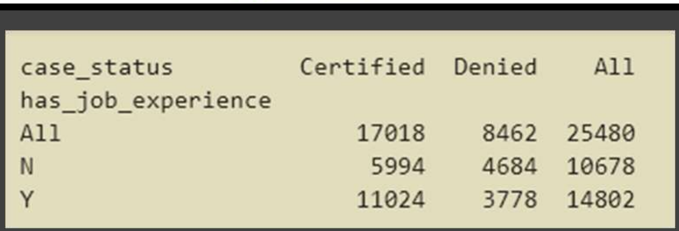
- Heatmap - Correlation of Education and Regions within the U.S. There are different requirements that match expected Education Standards.

EDA - Bivariate Analysis

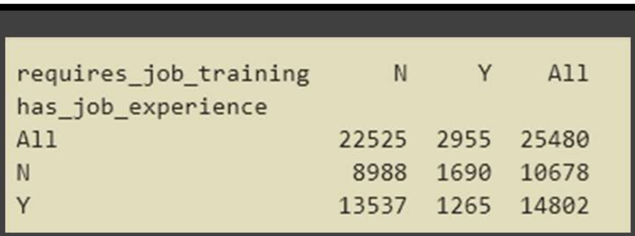
- Case Status is our Dependent Variable. Creating a Machine Learning Model requires us to compare attributes against one another to see how they help to determine Certification or Denial of Visa Applications. Some of these attributes are compared directly to Case Status and others are compared against a direct counterpart.





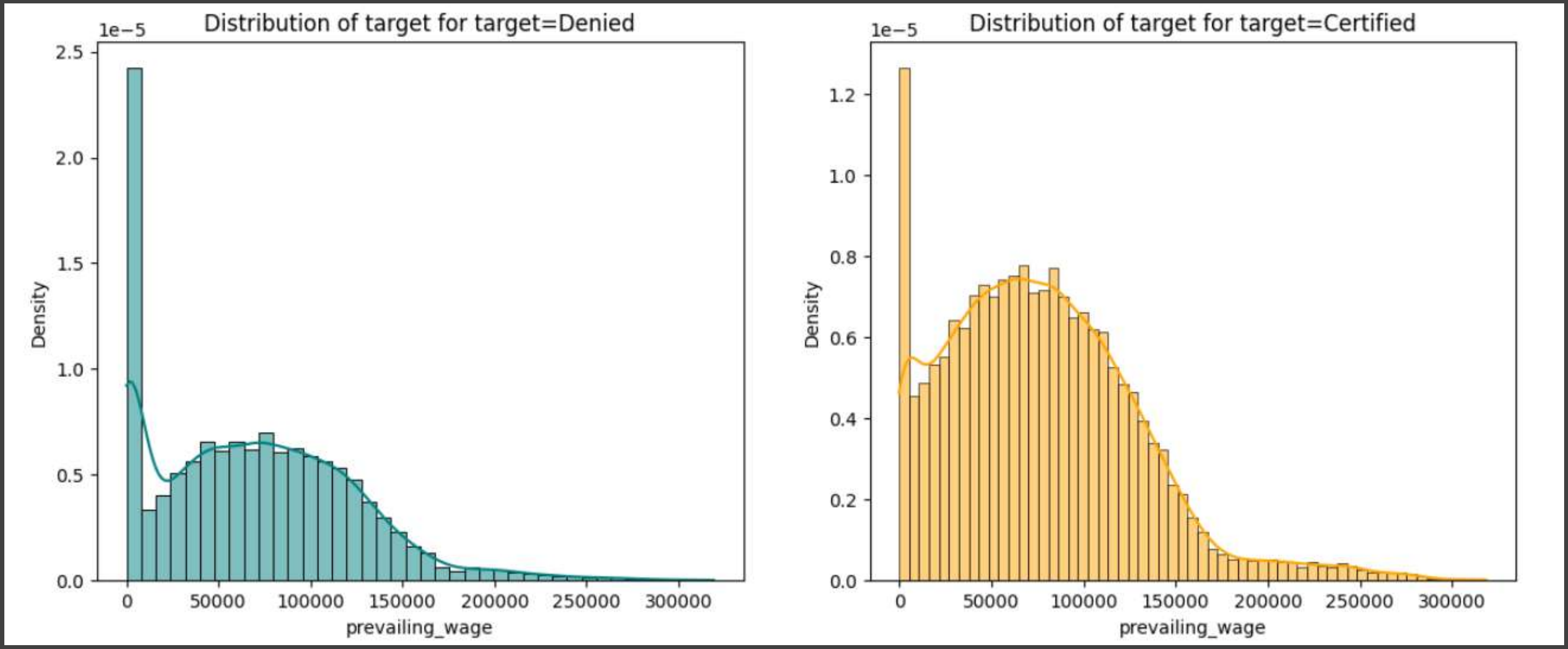


```
case_status          Certified  Denied    All
has_job_experience
All                      17018    8462  25480
N                         5994    4684  10678
Y                        11024    3778  14802
```



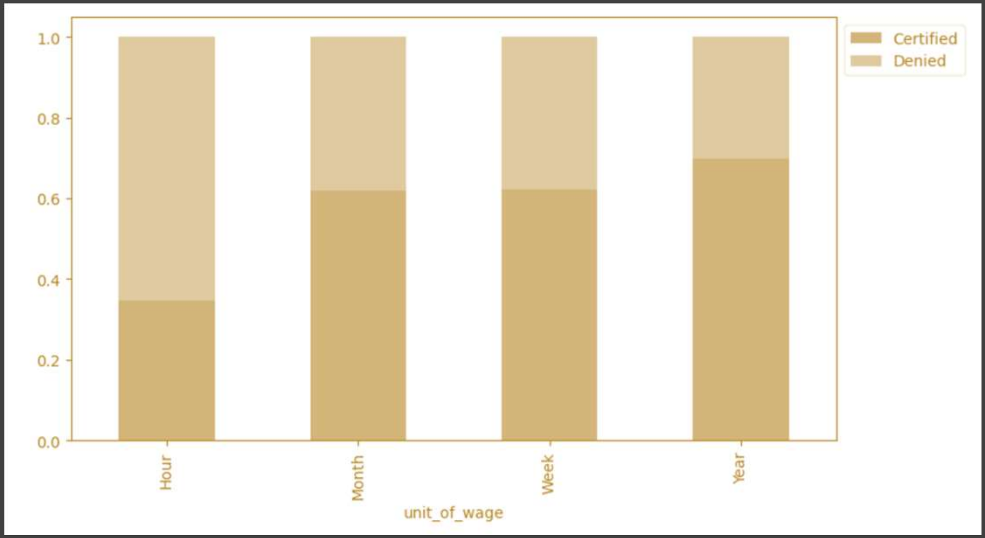
```
requires_job_training      N     Y    All
has_job_experience
All                    22525  2955  25480
N                       8988  1690  10678
Y                      13537  1265  14802
```

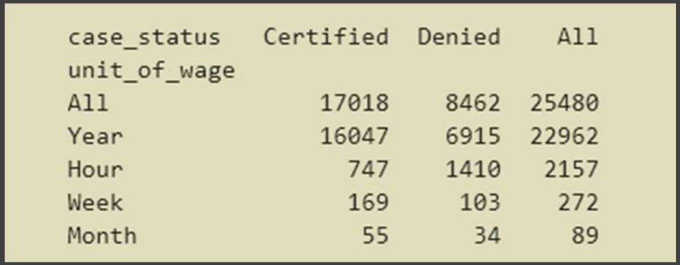
Bivariate Analysis



- The U.S. Government established a Prevailing Wage to protect our Citizens and Foreign Workers.
- Kernel Density Curve has been included, the Theoretical Probability is almost directly in-line with the Experimental Probability

Bivariate Analysis



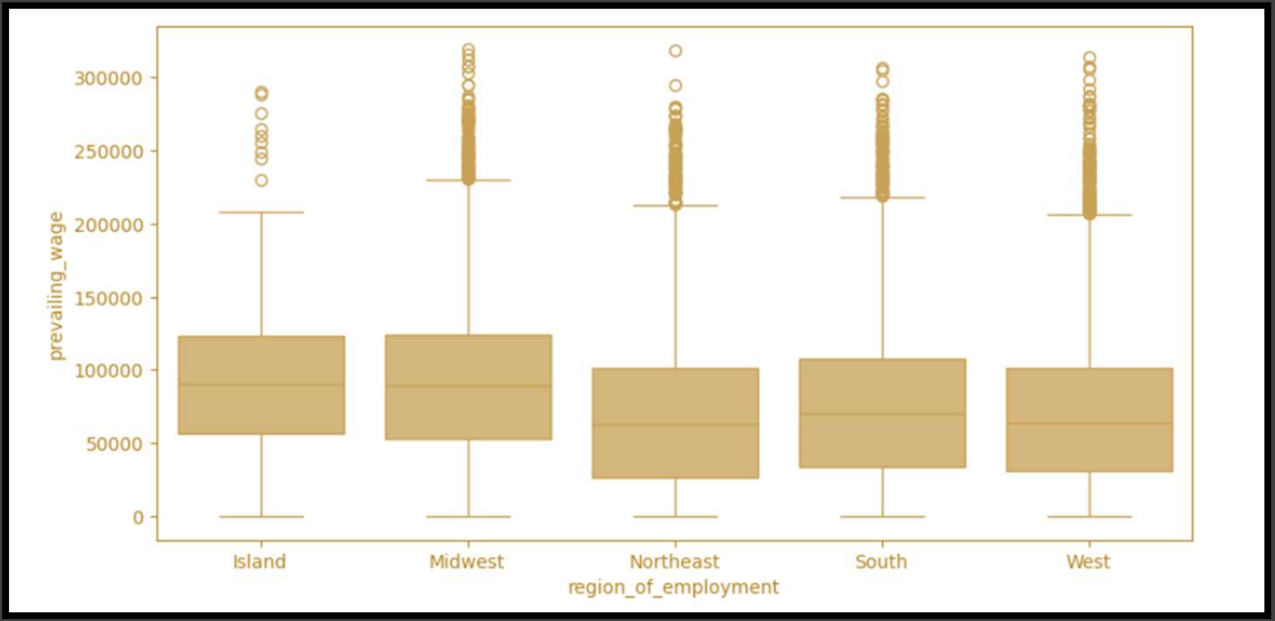


```
case_status    Certified  Denied    All
unit_of_wage
All                17018    8462  25480
Year               16047    6915  22962
Hour                 747    1410   2157
Week                 169     103    272
Month                 55      34     89
```

Unit of Wage in conjunction with Certification or Denial of Visa Applications

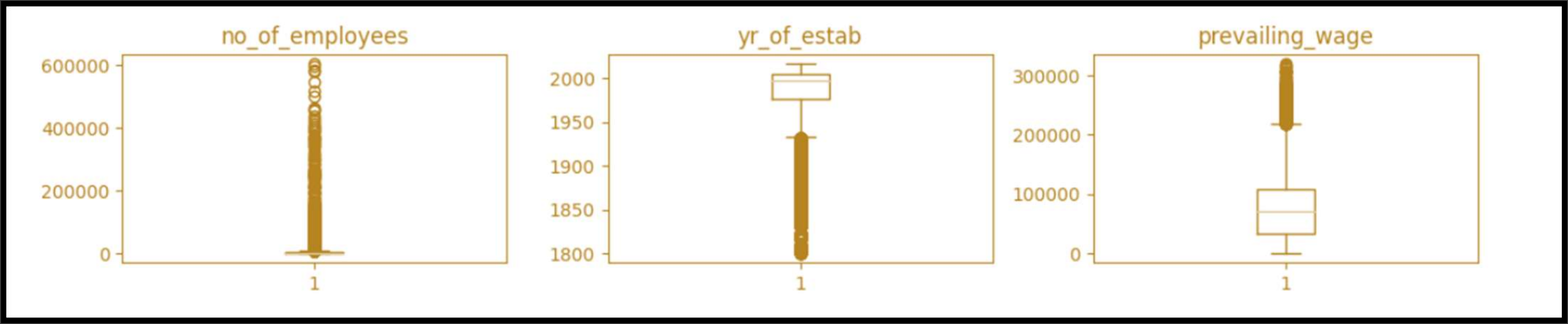
- Hourly Wage: Only 30% of Visa's Certified
- Weekly - Monthly - Yearly: All came in at over 60% of Certified Visa's
- Yearly - The highest percentage of Certified Applications at approx. 75%

Outliers - Region vs Wage



- We did not treat the Outliers, they were left as actual values
- The Prevailing Wage Mean was very close for the Northeast - South - West at approx. $60 - $70,000
- Midwest and Island Wage Mean were equal at approx. $90,000

Outliers - Actual Values



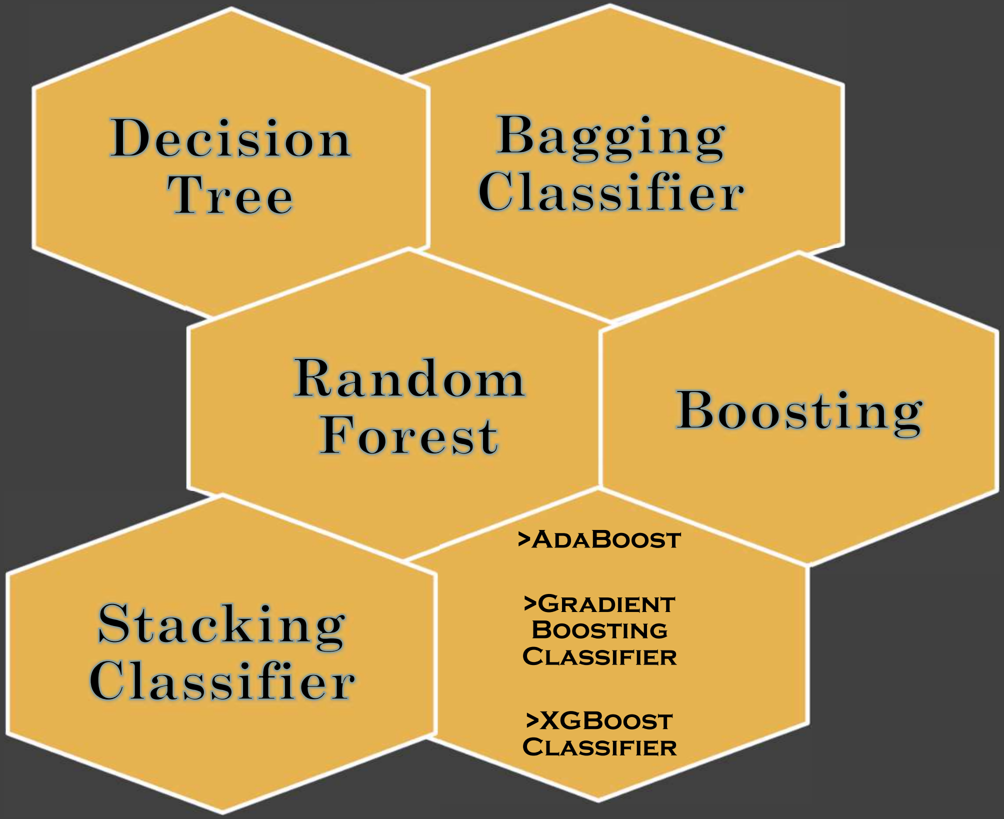
- There were many Outliers in all of the Numerical Valued Columns. We will treat them as actual values because the range of values is so significant.

  ❑ Number of Employees: Minimum = 11 - Maximum = 602k - Mean = 5,667

  ❑ Year of Establishment: Minimum = est. 1800 - Maximum = est. 2016 - Mean = est. 1979/1980

  ❑ Prevailing Wage: Minimum = $2.13 - Maximum = $31,921 - Mean = $74,455

Model Building - Types & Date Preparation



Decision Tree

Bagging Classifier

Random Forest

Boosting

Stacking Classifier

>AdaBoost

>Gradient Boosting Classifier

>XGBoost Classifier

DATA PREPARATION

- Case Status: This Column will be dropped because it is our ultimate goal in determining Visa Certification or Denial
- Encoding Categorical Features
- Splitting the Data into Train & Test to evaluate the model we build against the Train Data
- Stratified Sampling: Classification exhibits a significant imbalance in the distribution of the target classes. This ensures relative class frequencies are within a parameter and preserved in Train and Test Sets

Model Building - Criterion & Overall Process

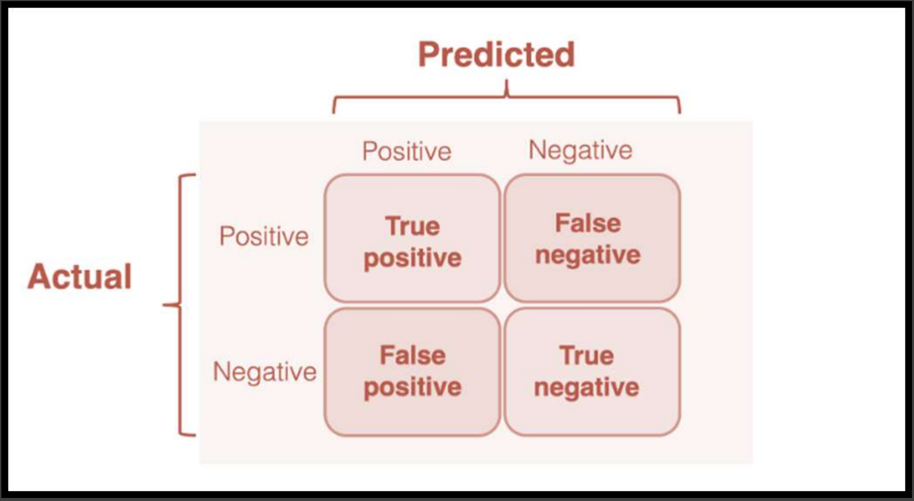
- Model Predictions:
  a. False Positives: Predicts application will get Certified, but should get Denied
  b. False Negative: Predicts application will NOT get Certified, but it should have been Certified

- Case Importance:
  a. If a Visa is Certified when it should have been denied a brought in worker would get the position, while a U.S. Citizen will miss the opportunity
  b. If a Visa is Denied when it should have been certified the Company and the U.S. will miss out on a suitable employee that can contribute to our economy

Reducing the Loss

- F1 Score: The Metric for our Model. The greater the F1 Score, the higher the chances of minimizing False Positives and False Negatives

- Balanced Weights: The model focuses equally on both classes

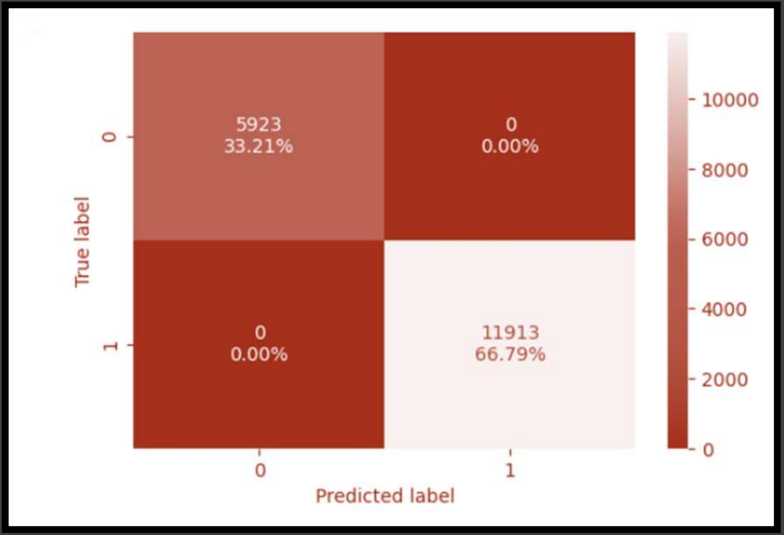


		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

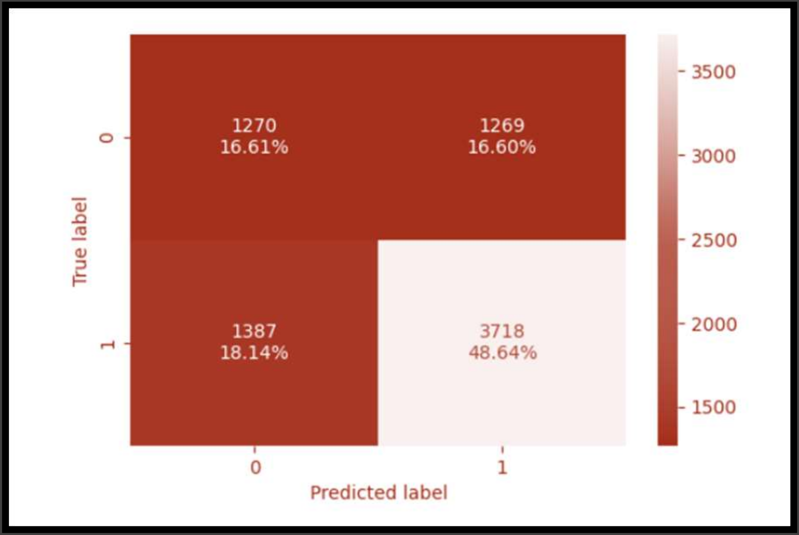
Model: Decision Tree

- Decision Tree Classifier Function
- Gini Criterion to split the Data, and to calculate the probability of a specific randomly selected attribute that was incorrectly classified. Scaled from 0 to 1: Higher Values = Increased Inequality
- 10% - Class A frequency & 90% - Class B frequency
- " B " will become Dominate and the Decision Tree will lean toward this class
- In this case, we will apply class_weight to this model and give more weightage to Class A

Training Model:



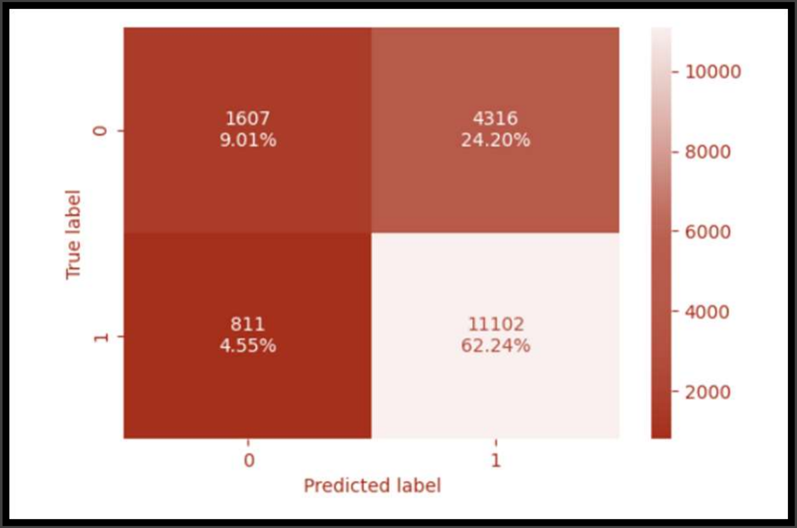
Test Model:

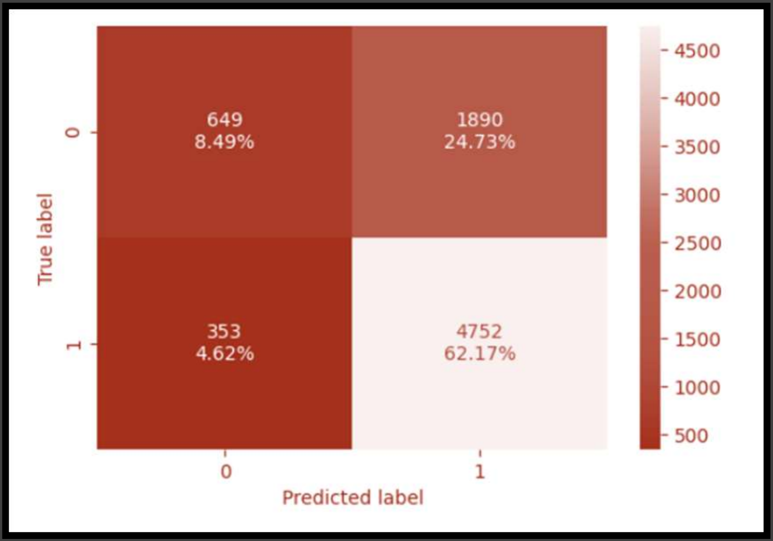


Hyperparameter Tuned - Decision Tree

- F1 score on Initial Test Set = 74% After Hypertuning Test = 81%

Training Model: **Test Model:**



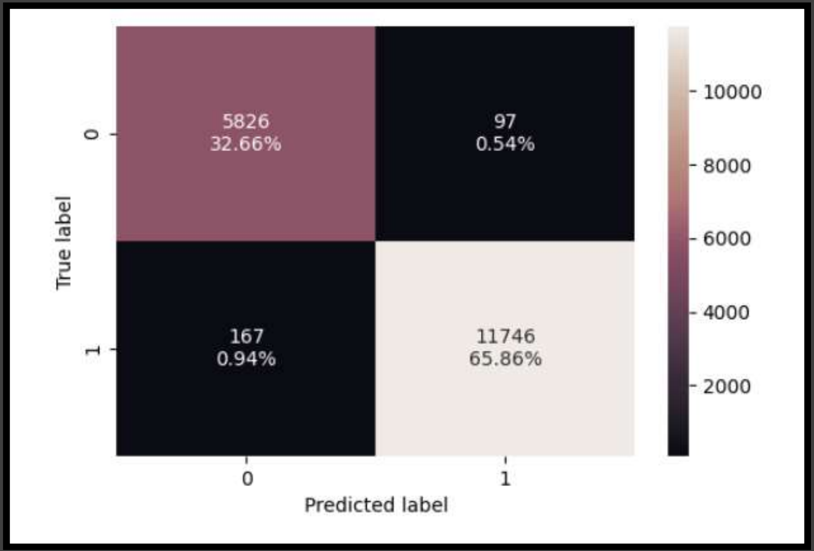


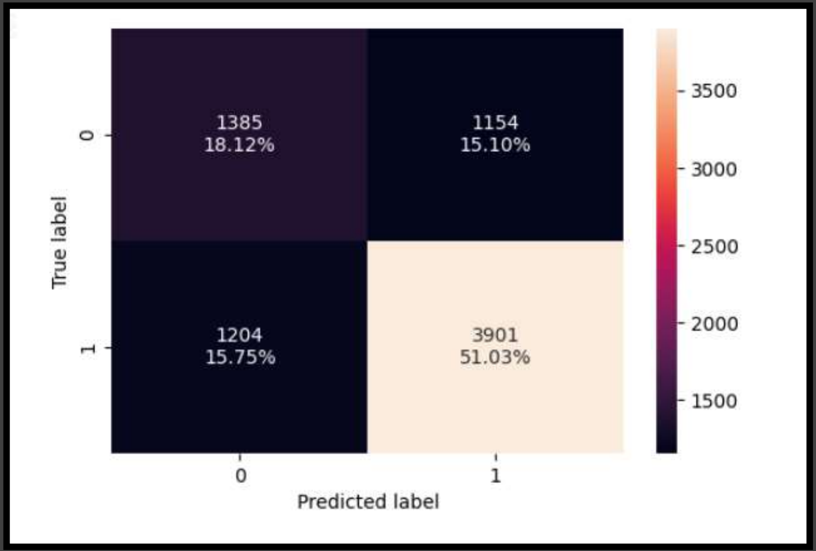
Model: Bagging Classifier

- Bagging is building an independent model on random samples and combining their predictions
- Replacement Sampling helps make the set more independent and uncorrelated
- Increased size of the data allows for diversity throughout the samples
- Guarantee of 63% with replacement sampling

Training Model:

Test Model:



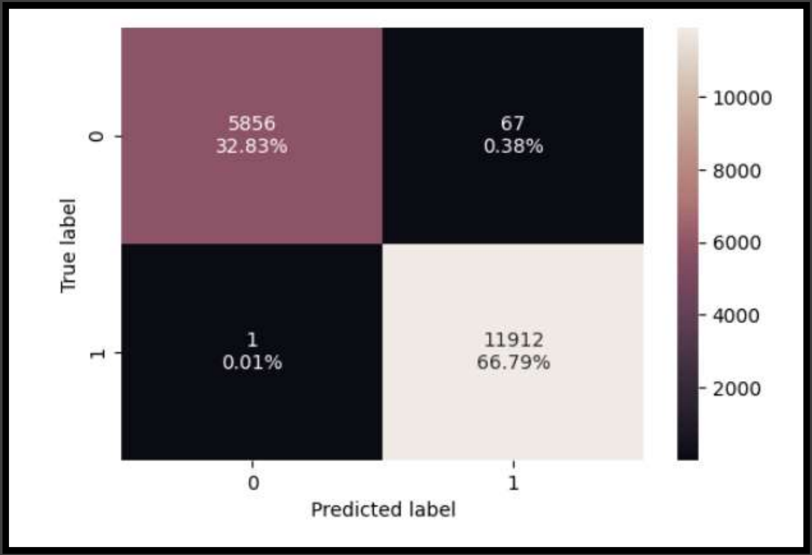


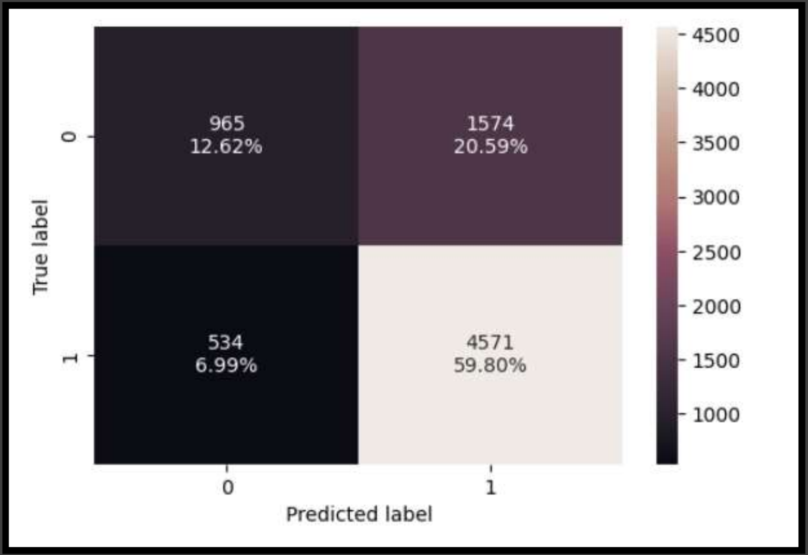
Hyperparameter Tuned - Bagging

- F1 score on Initial Test Set = 77%   After Hypertuning Test = 81%

Training Model:

Test Model:

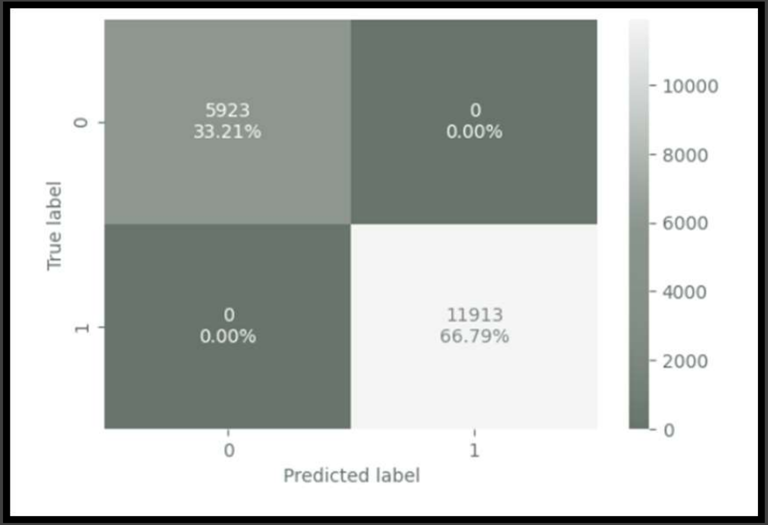




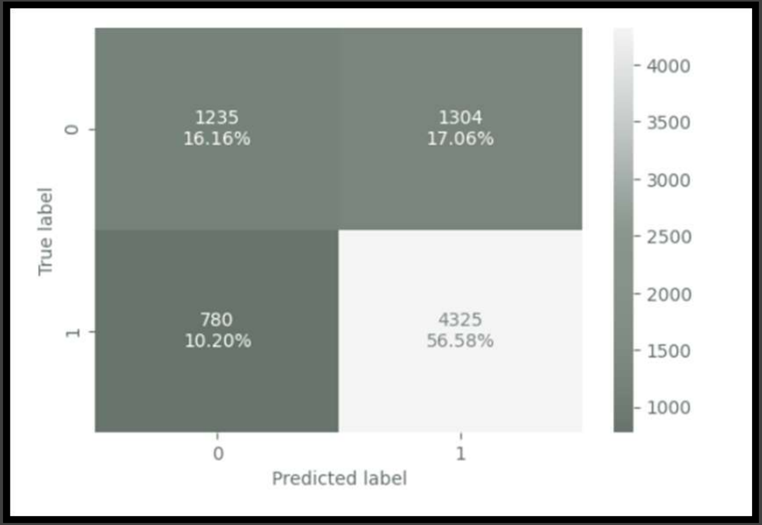
Model: Random Forest

- Random Forest - additional random variation added to Bagging - creating diversity
- Subsets of features are selected at random, then the split feature from that subset is used to split each node
- Trees grow as large as possible or to a determined depth & new data is predicted by aggregating the predictions from every tree
- Class_Weight - Specification of weight given to the non-dominant class

Training Model:



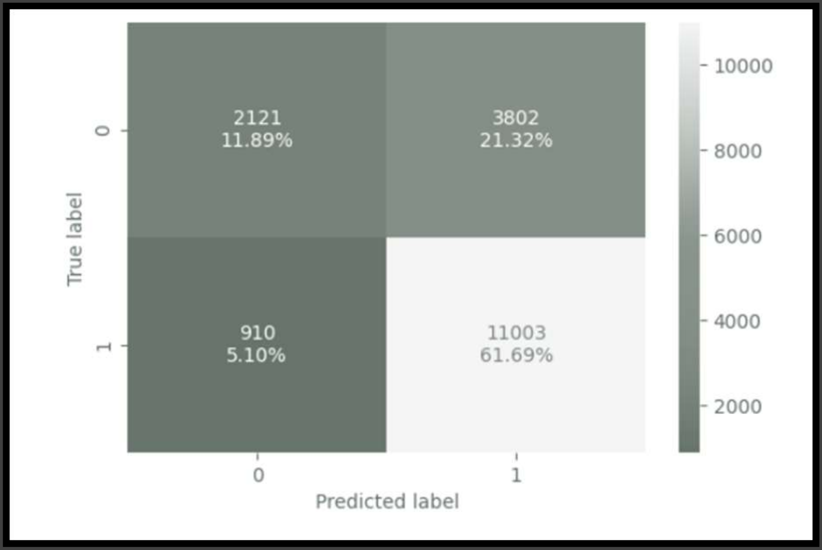
Test Model:



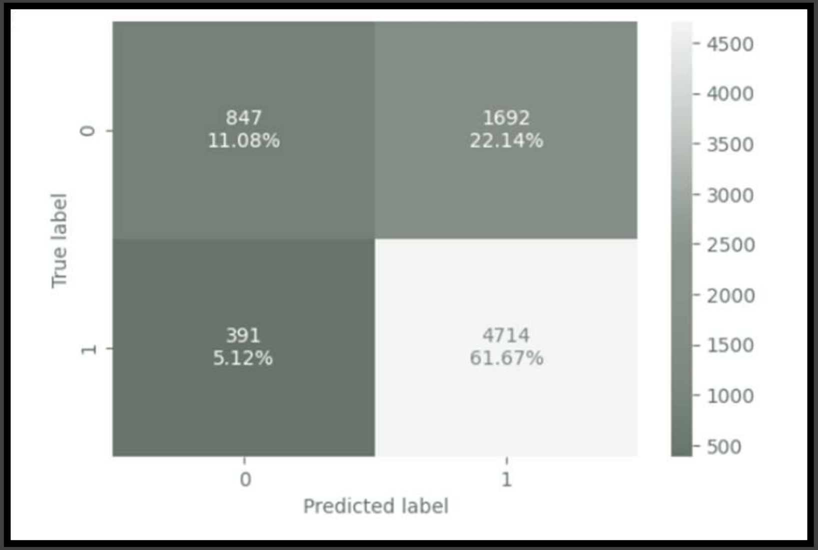
Hyperparameter Tuned - Random Forest

- F1 score on Initial Test Set = 80% After Hypertuning Test =82 %

Training Model:



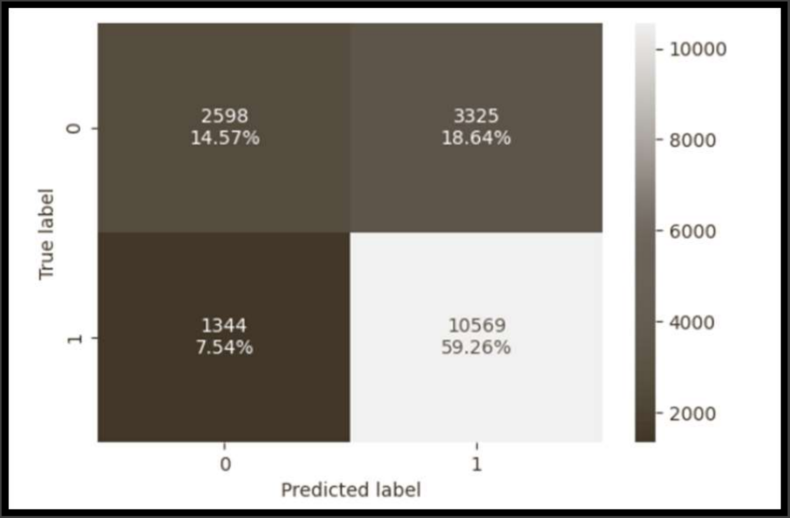
Test Model:



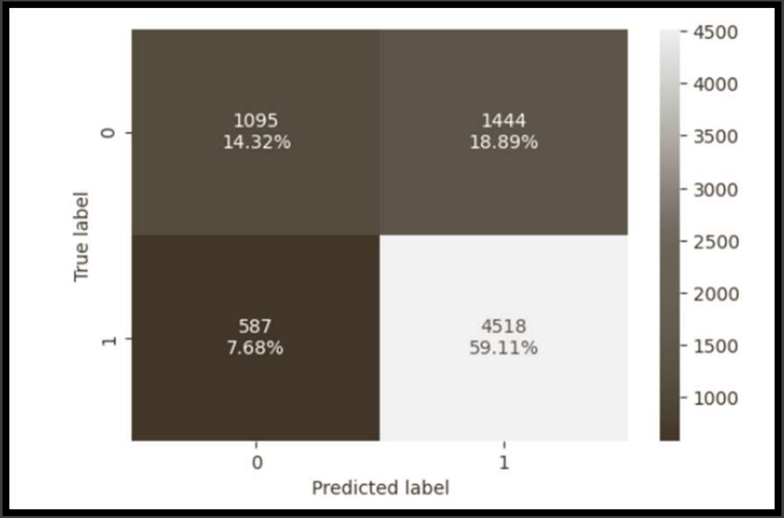
Model: AdaBoost Classifier

- The Weak Learner must first be built, then we measure the importance of a Weak Learner based on error
- A new sample weight is created based on correct & incorrect predictions
- New data set is created with the odds of each sample being chosen on the new weights, this is Bootstrapping
- Learning_Rate: Decreases the contribution of each classifier, this lessens the amount of n_eastimators
- Repeat this process n_number of times and the final predictions a weighted vote of every weak learner

Training Model:



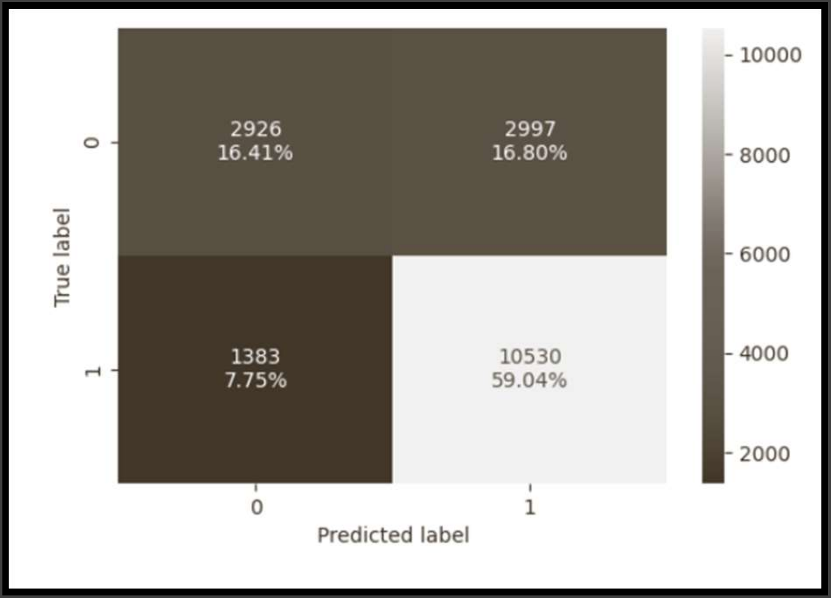
Test Model:

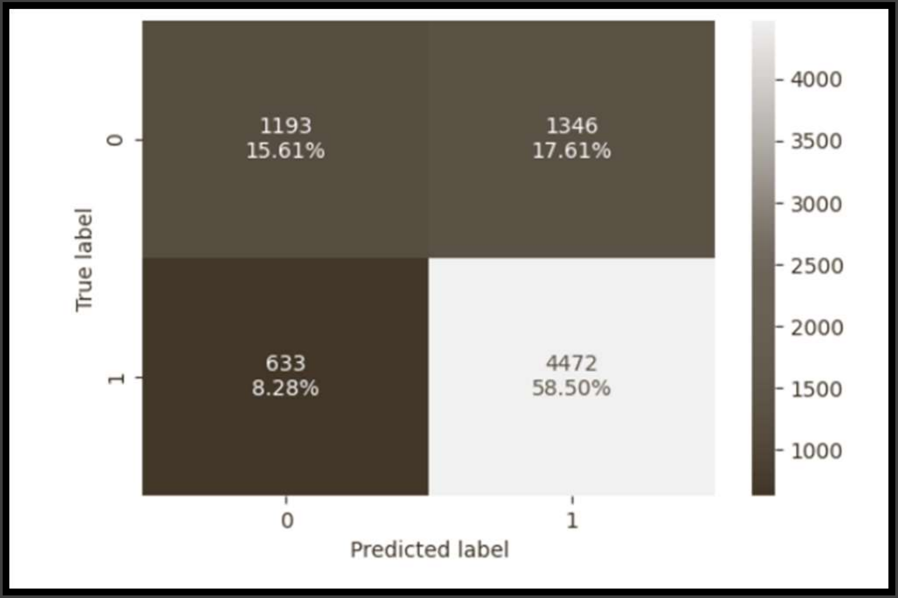


Hyperparameter Tuned - AdaBoost

- F1 score on Initial Test Set = 82 % After Hypertuning Test = 83 %

Training Model: **Test Model:**

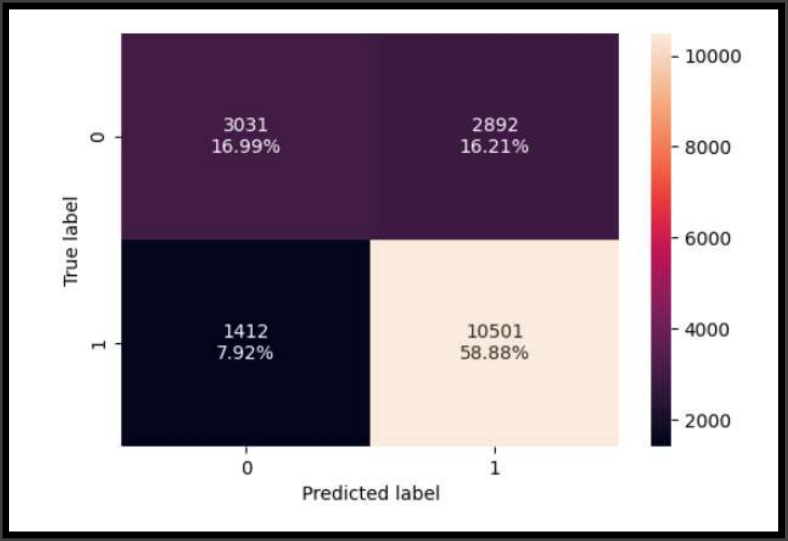




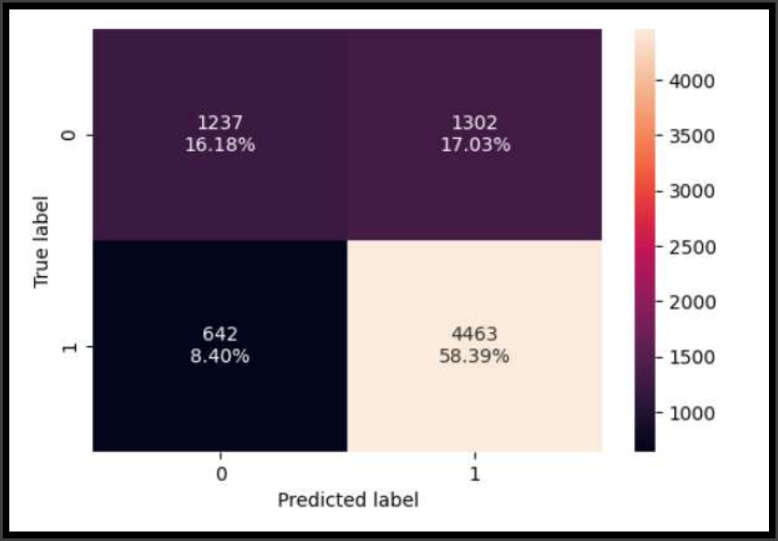
Model: Gradient Boosting Classifier

- Weak Learner is built on a subset of the original data
- Predictions made on Residuals and errors are calculated by comparing those predictions and the actual values
- Gradient Boosting fits the next weak learner to residuals of the previous one
- These residuals now become the target variable for the new weak learner and minimize errors
- Process is repeated until there is no residual reduction or the number of estimators is acheived

Training Model:



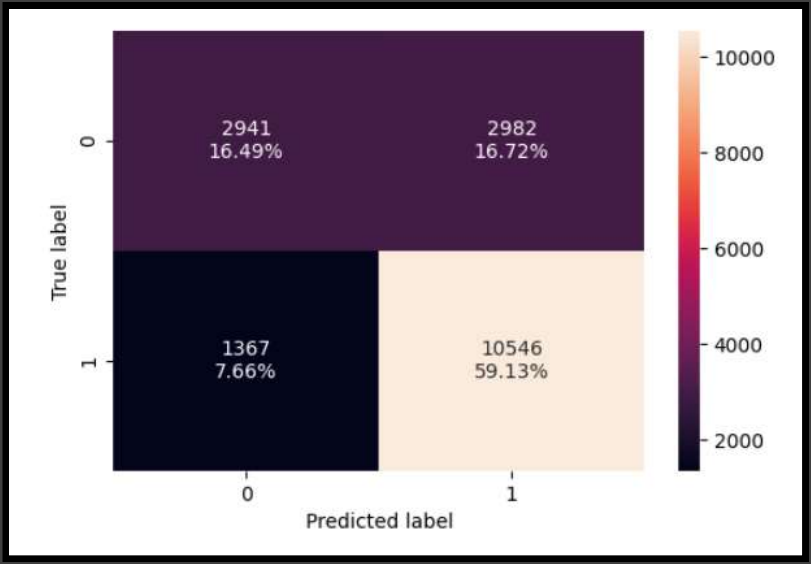
Test Model:

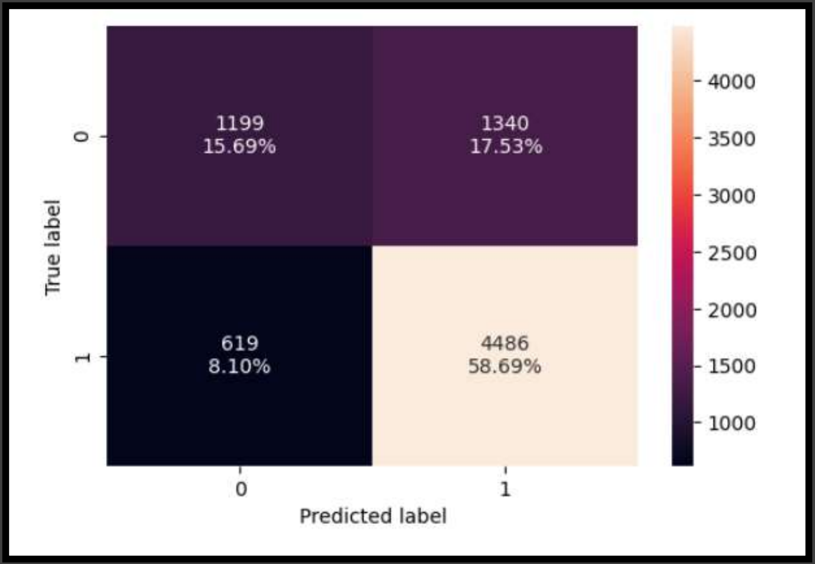


Hyperparameter Tuned - Gradient Boost

- F1 score on Initial Test Set = 82 % After Hypertuning Test = 82 %

Training Model: **Test Model:**



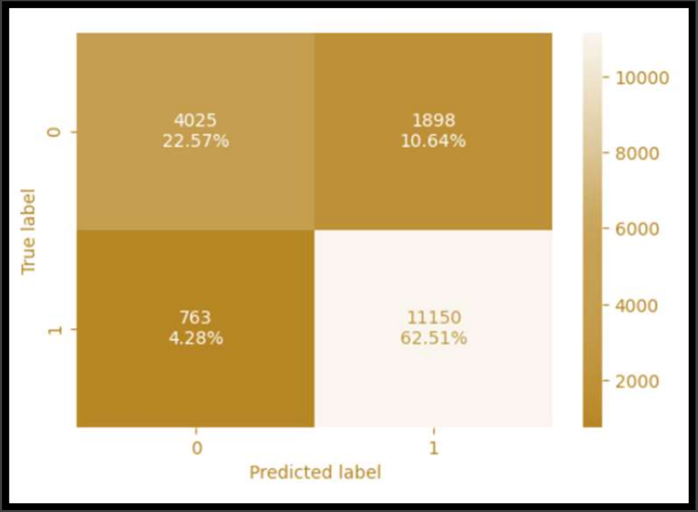


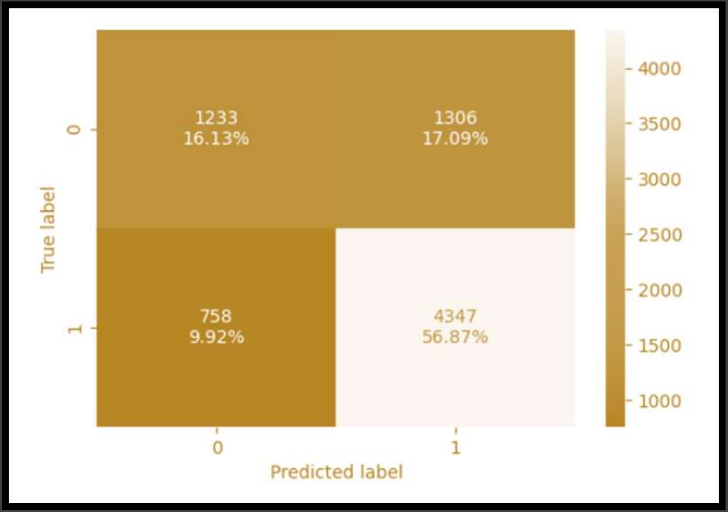
Model: XGBoost Classifier

- A very popular algorithm, based on Gradient Boost, but with super calculating power
- Sequentially built Gradient Boosted Trees
- It has implemental advantages: Regularization, Missing Value Treatment, Cache Optimization, distributed computing, ect.
- Computational Speed is one of the main benefits of XGBoost

Training Model:

Test Model:



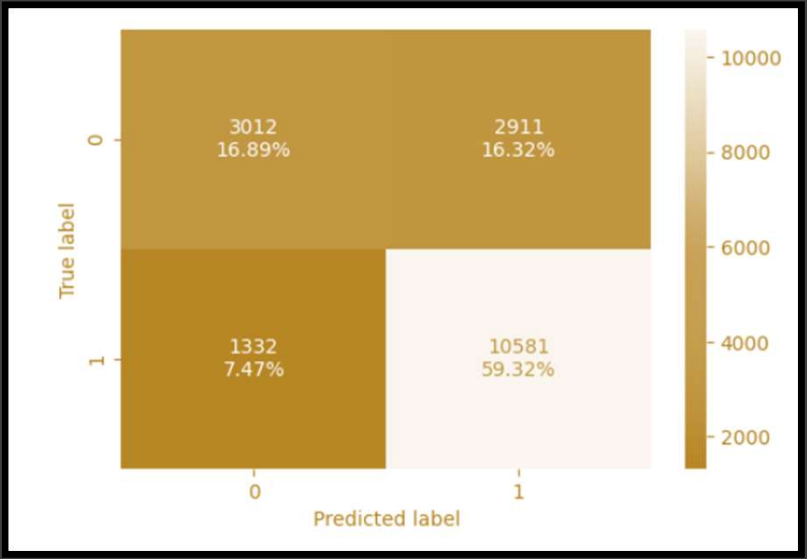


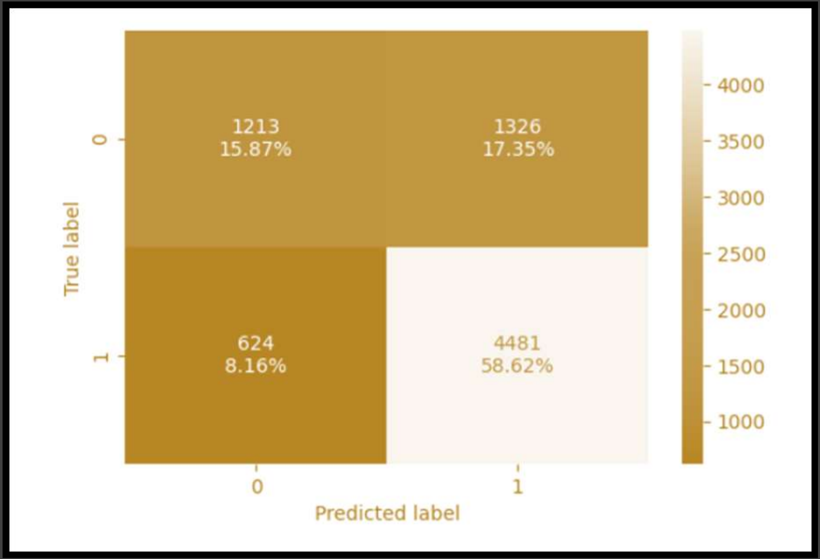
Hyperparameter Tuned - XGBoost

- F1 score on Initial Test Set = 81% After Hypertuning Test = 82%

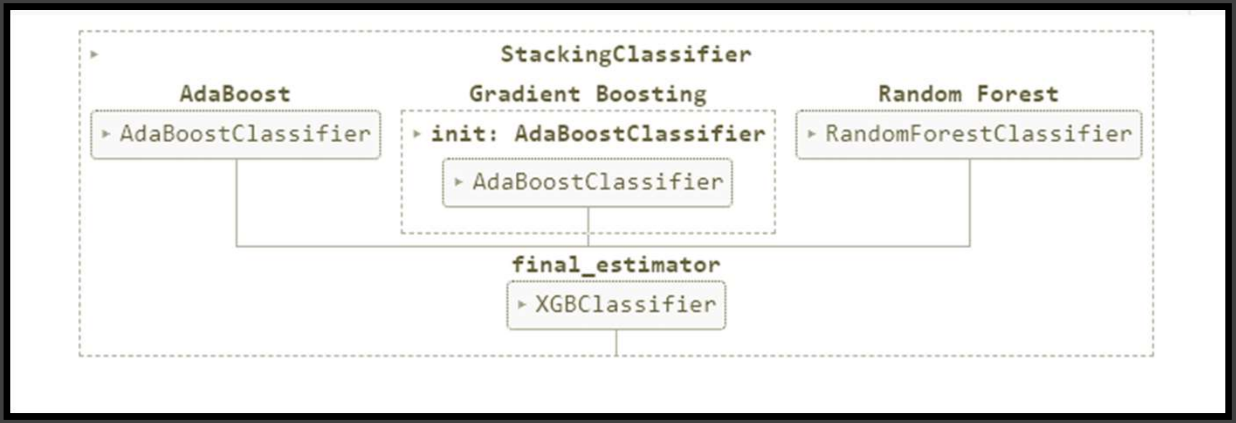
Training Model:

Test Model:





Stacking Classifier



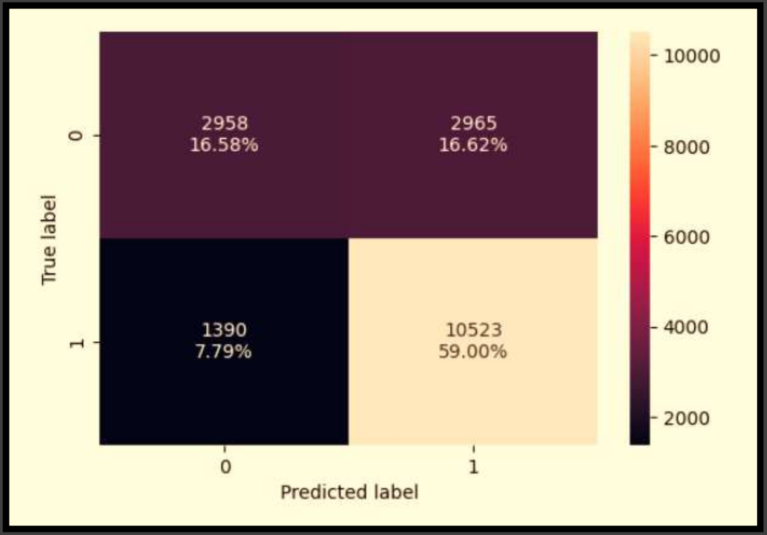
- Builds a Heterogeneous Model
- Combines Bagging and Boosting to create a Meta-Model
- Models are Stacked, the First Set of models built initial predictors and the Second Set combines the predictors
- Train Data is separated into Two or more Folds
- Splitting the Training Data does create a Disadvantage

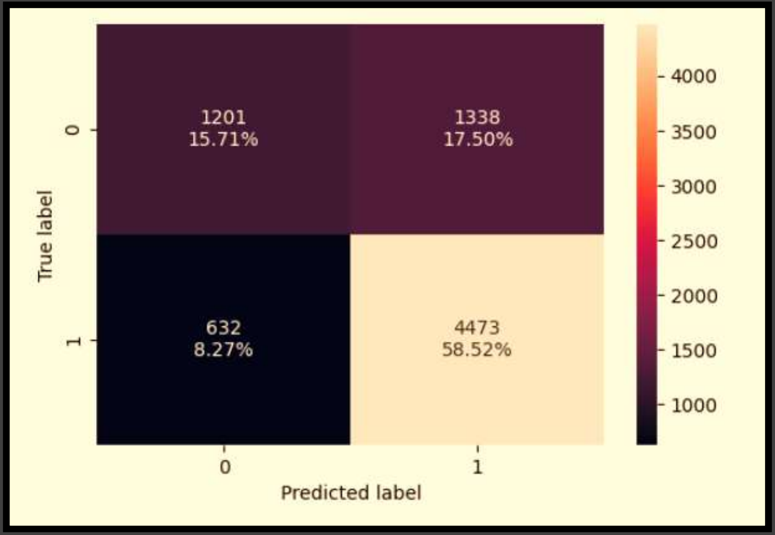
Model: Stacking Classifier

- F1 score on Train Set = 83%   F1 score on Test Set= 82%

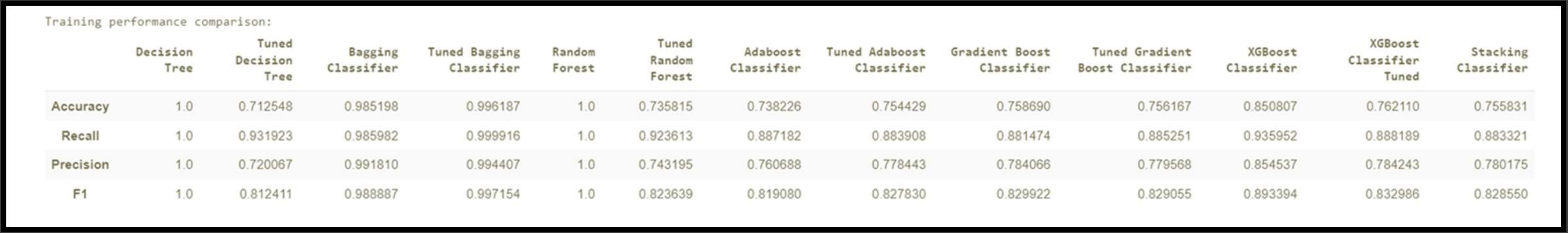
Training Model:

Test Model:



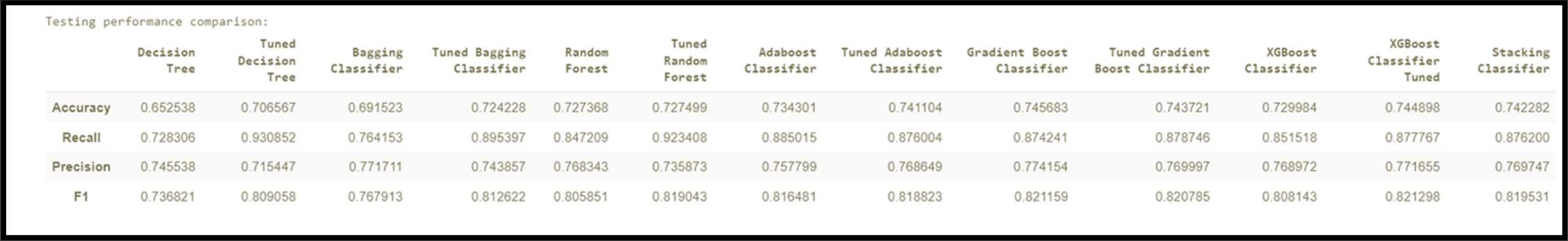


Comparison of Models



Training performance comparison:

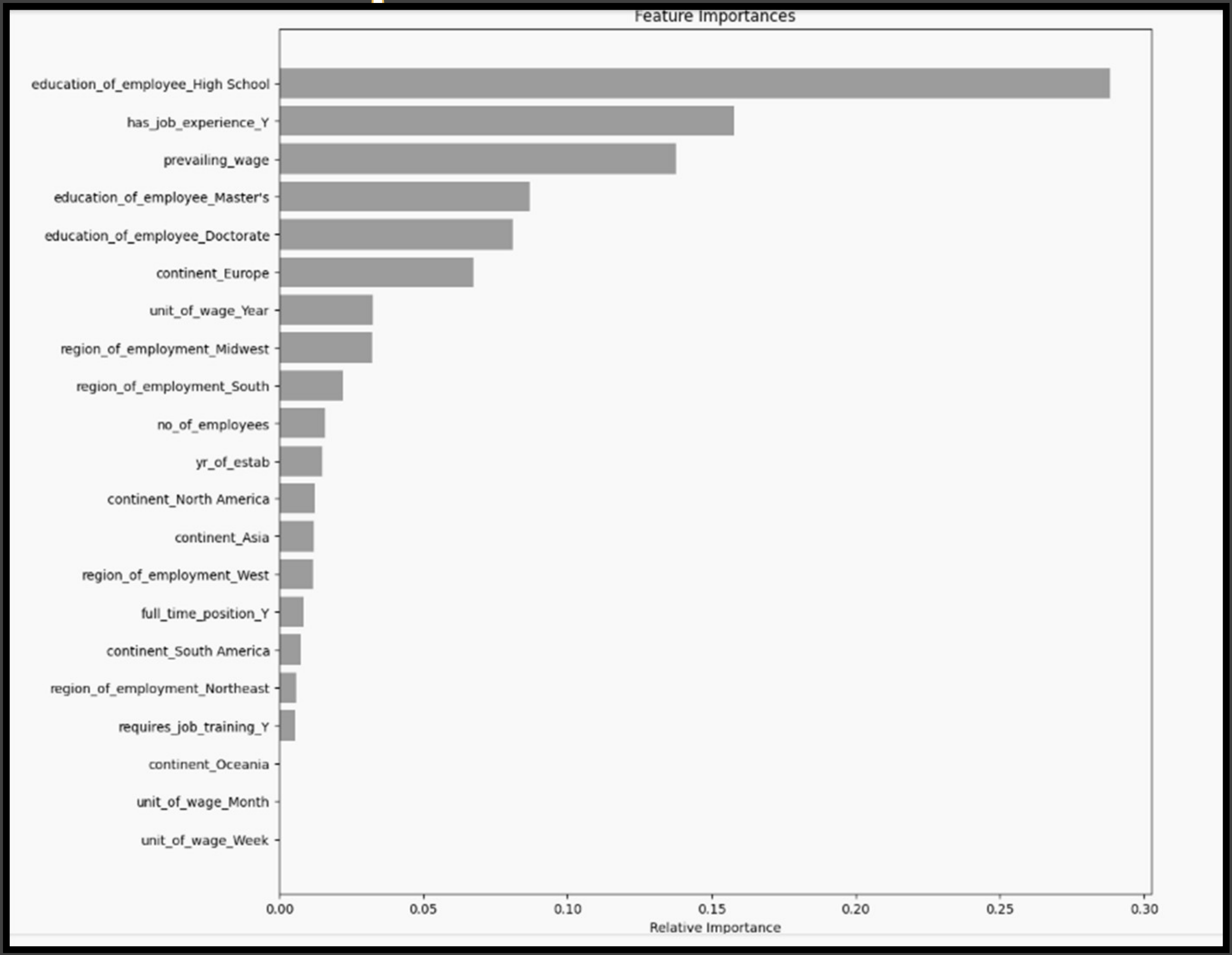
	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	0.712548	0.985198	0.996187	1.0	0.735815	0.738226	0.754429	0.758690	0.756167	0.850807	0.762110	0.755831
Recall	1.0	0.931923	0.985982	0.999916	1.0	0.923613	0.887182	0.883908	0.881474	0.885251	0.935952	0.888189	0.883321
Precision	1.0	0.720067	0.991810	0.994407	1.0	0.743195	0.760688	0.778443	0.784066	0.779568	0.854537	0.784243	0.780175
F1	1.0	0.812411	0.988887	0.997154	1.0	0.823639	0.819080	0.827830	0.829922	0.829055	0.893394	0.832986	0.828550



Testing performance comparison:

	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	0.652538	0.706567	0.691523	0.724228	0.727368	0.727499	0.734301	0.741104	0.745683	0.743721	0.729984	0.744898	0.742282
Recall	0.728306	0.930852	0.764153	0.895397	0.847209	0.923408	0.885015	0.876004	0.874241	0.878746	0.851518	0.877767	0.876200
Precision	0.745538	0.715447	0.771711	0.743857	0.768343	0.735873	0.757799	0.768649	0.774154	0.769997	0.768972	0.771655	0.769747
F1	0.736821	0.809058	0.767913	0.812622	0.805851	0.819043	0.816481	0.818823	0.821159	0.820785	0.808143	0.821298	0.819531

Feature Importance's: Final Model



INSIGHTS:

- A predictive model has been built that can assist in shortlisting candidates with higher Visa Approval odds
- In our final set of Featured Importance's there were only 5 factors that were significant in Certification of Visa Approvals
- F1 Score was the metric for determining our Model Comparison. All of our F1 scores were around 81% which minimized our False Negatives

RECOMMENDATIONS:

- Facilitating the process of Visa Approvals will rely on certain factors that will make the candidate stand out. These factors are Education, Prior Job Experience, and Prevailing Wage
- Based on these factors foreign workers applying for jobs in the U.S. must provide Professional References and go through a of two interview process with pre-determined questions specific to situations and direct knowledge of the industry. Education being the most important factor, proof will need to be provided
- There were many factors that did not have an impact on Visa Certification or Denial
- With a 9% increase in applicants in one year, this process must be streamlined and intentional Questionnaires and Interviews need to be top priority