EasyVisa

EasyVisa – Ensamble Techniques

22/03/2022

Contents / Agenda

- Business Problem Overview and Solution Approach
- Executive Summary
- General Information
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Business Problem Overview and Solution Approach

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 temporary and permanent labor certifications positions. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year. The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates who have higher chances of VISA approval.

Executive Summary

As EasyVisa we were requested to analyze the data provided and, with the help of a classification model:

- A model was developed consisting of the different variables shared by the OFLC.
- The results of the study show that the most common level of rejection is for applicants with a high school diploma. Even though the applicants with a professional degree are the most common applicants the ones with a high school diploma are rejected at a higher rate compared to the rest of the candidates.
- The study shows that the European candidates are the ones with the highest chances to be accepted even though this group consists of 14.6% of the total applicants, but the Asian group which constitutes over the 60% has lower chances based on the mode.
- The model also shows that the people who apply for a more stable job, those with a yearly base income rather than those of other formats are the ones preferred.

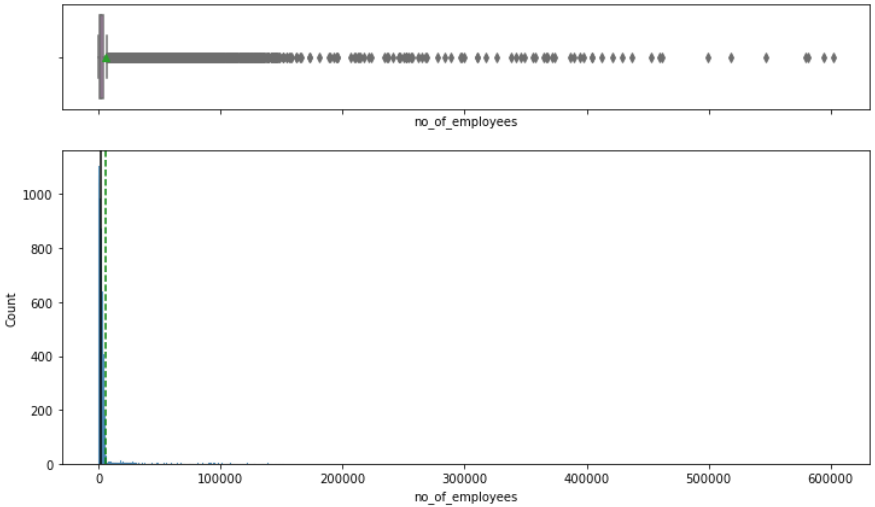
General Information

- The shared data set contains 12 variables and 25480 rows of information.
- The variables consisted mostly of objects, and 3numeric variables, no_of_employees, prevailing_wage and yr_of_estab.
- This data set didn't contain any duplicated information.

Link to Appendix slide on data background check

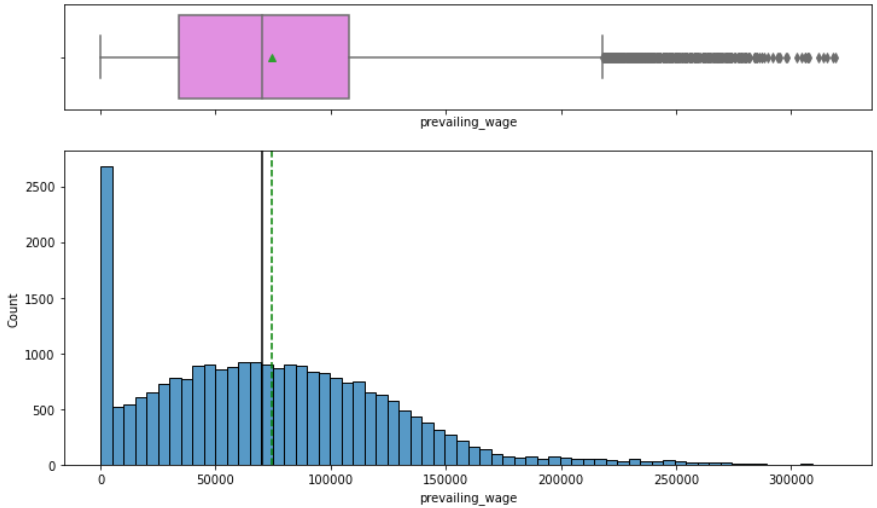
EDA Results – Univariate Analysis

- Observations on number of employees



- This variable shows a large distribution of the data as different types of companies accept applicants.

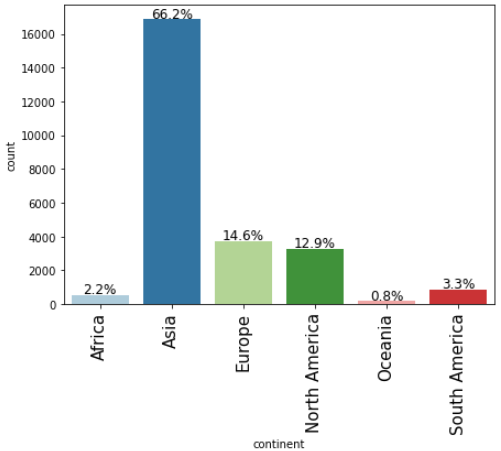
- Observations on prevailing wage



- This variable shows that the prevailing wage has a significant amount of outlier.

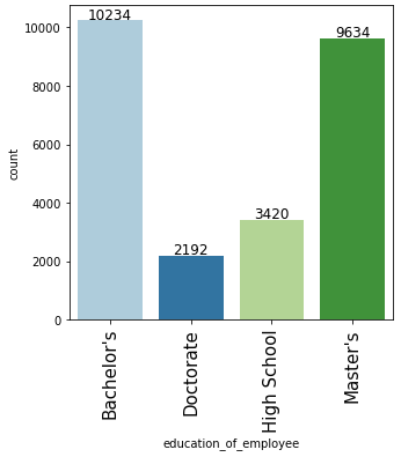
EDA Results – Univariate Analysis

- Observation on Continent



- The most common continent from where the applicants come from is Asia.

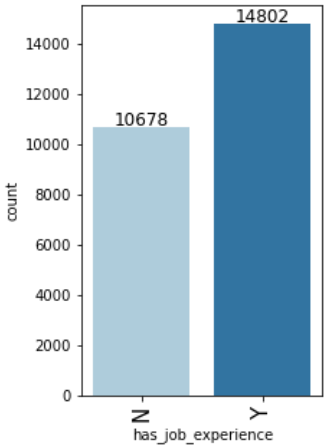
- Observations on education of employee



- The employee usually has at least a bachelor's degree. Which indicates that professionals are the ones who seek to apply to different jobs.

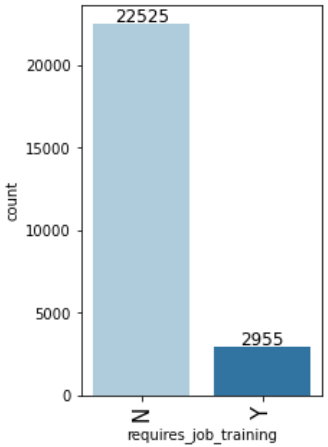
EDA Results – Univariate Analysis

- Observations on the experience



- The distribution of experience shows that most of the candidates have a previous experience in a job.

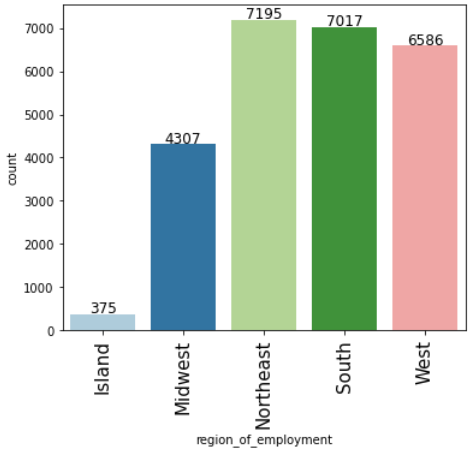
- Observations on training



- The jobs that the candidates are applying for requires no training from the different companies.

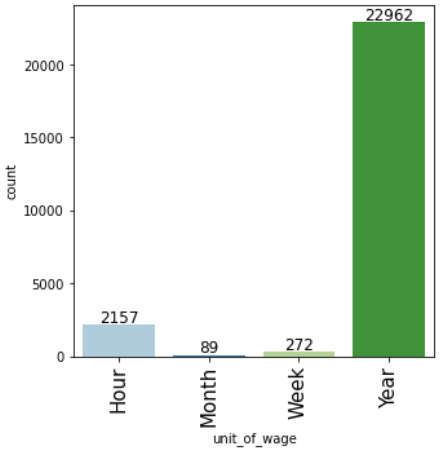
EDA Results – Univariate Analysis

- Observations on region of employment



- The regions of employment show a similar requirement of personal. Except for the islands that the requirement is a fraction of the rest of the regions.

- Observations on unit of wage

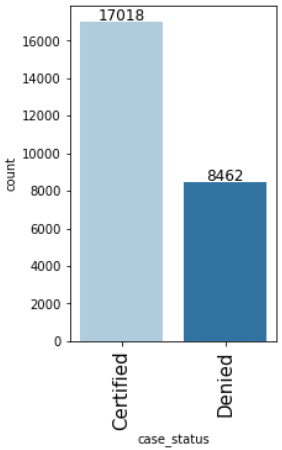


- The employees show with this variable that seek a stable candidate

EDA Results – Univariate Analysis

- Observations on Case Status

- The cases of most applications were accepted. Around 66% of all applications were accepted.

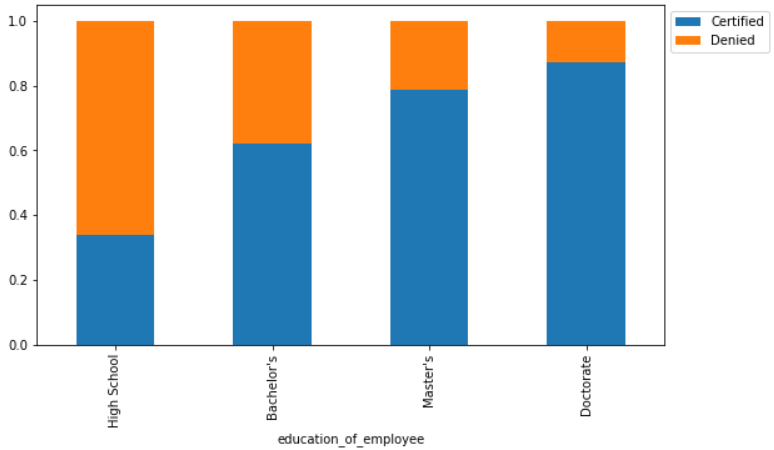


EDA Results – Bivariate Analysis

- Correlation
- The correlation between the numeric variables shows that there isn't a correlation between the variables.

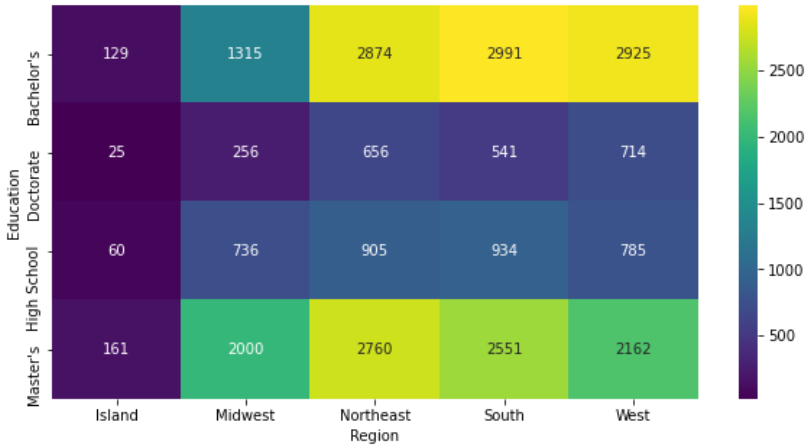
EDA Results – Bivariate Analysis

- Education vs Status
- The employees with a higher level of education have a higher level of acceptance to obtain a visa. The high school applicants are the least desirable for the visa applicants.



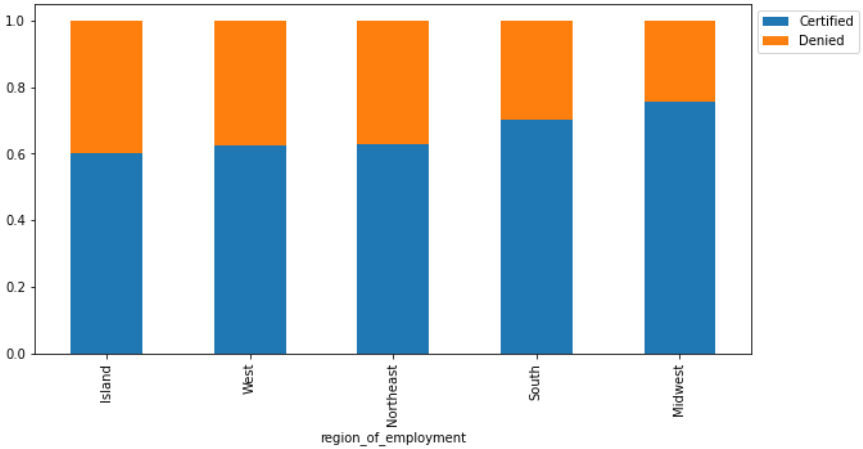
EDA Results – Bivariate Analysis

- Regions vs Education
- The regions show a similar distribution of requirements of high school applicants. The Midwest, which is the one with the least number of required employees shows a similar distribution of employees in the different education levels.



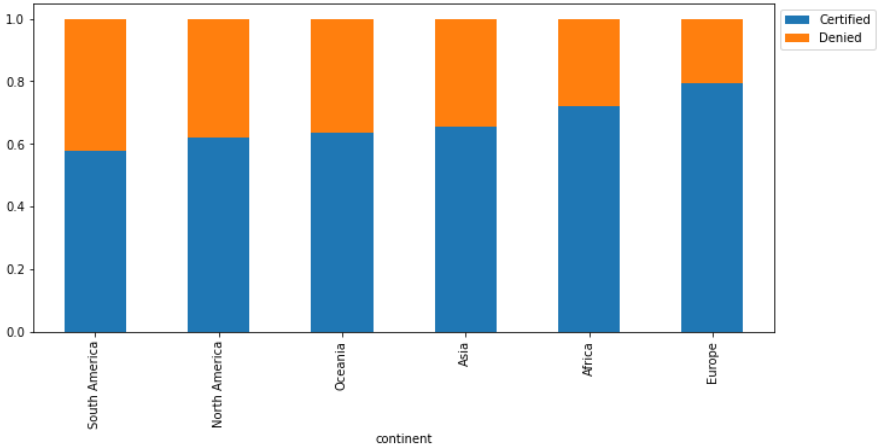
EDA Results – Bivariate Analysis

- Regions vs Status



- The region with the highest rate of certified status is the Midwest. The rest of the regions shows a similar level of certified visas.

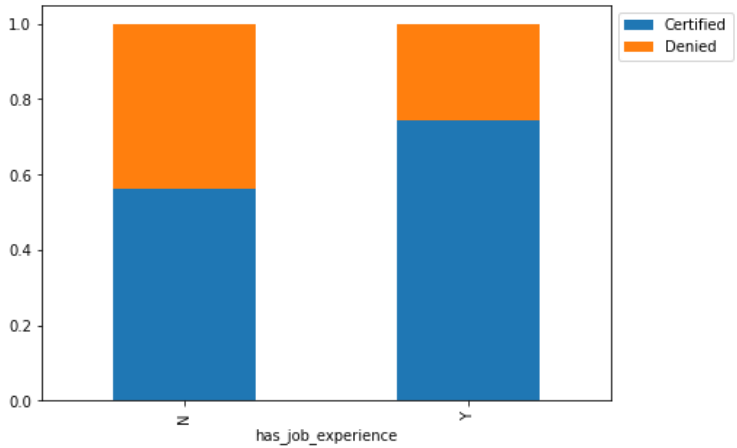
- Continent vs Status



- The European applicants are the ones with the highest success rate. The ones from Latin America are the ones with the least level of certified visas.

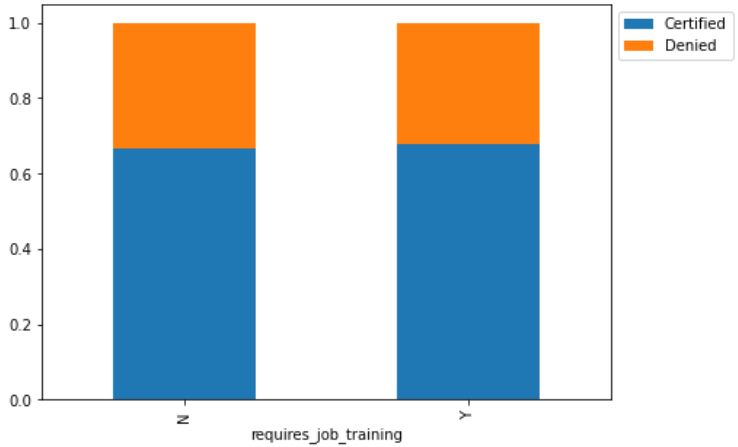
EDA Results – Bivariate Analysis

- Job experience vs Status



- The employee applicants that have previous job experience show that are the ones with a higher rate of certified visa status.

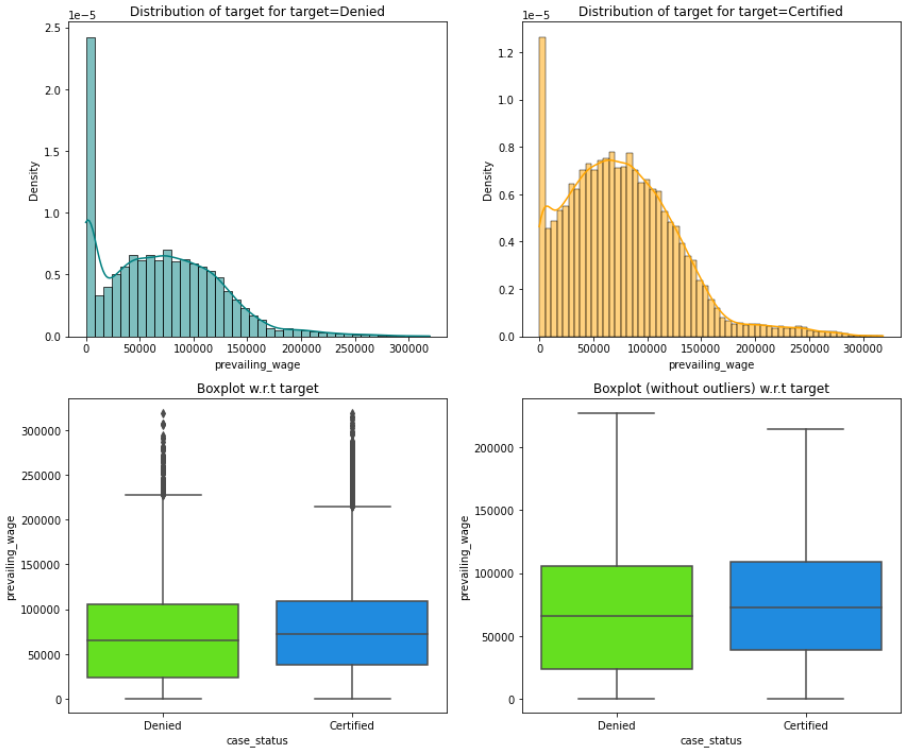
- Job training vs Status



- The distribution between training and non-training for a certified visa is practically the same.

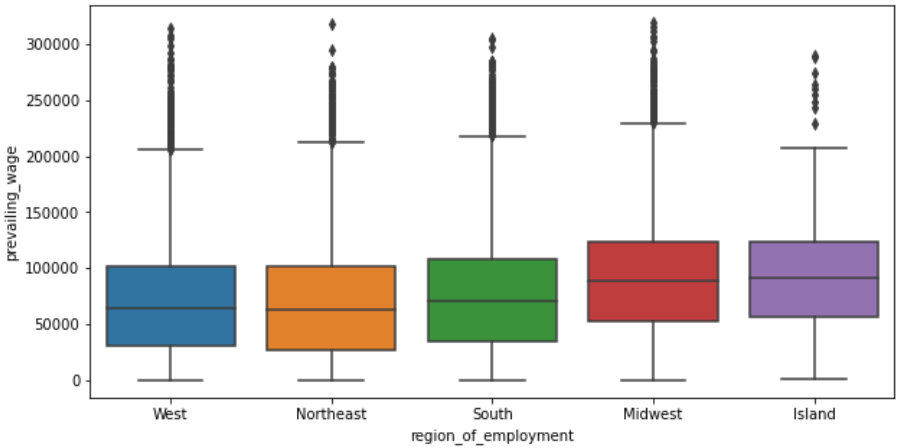
EDA Results – Bivariate Analysis

- Case status vs Prevailing wage
- The prevailing wage of the certified status shows a similar range of wages. The US government wants to protect its citizens by having similar paying jobs for visa applicants. If the applicants have a lower wage the companies will have an incentive to hire specialized employees at lower wages affecting the US citizens.



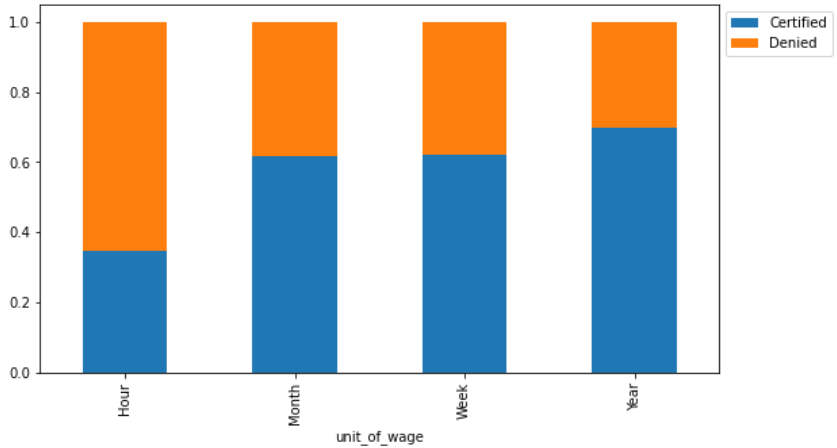
EDA Results – Bivariate Analysis

- Prevailing Wage vs Region of employment



- The prevailing wage shows a similar distribution between the different regions.

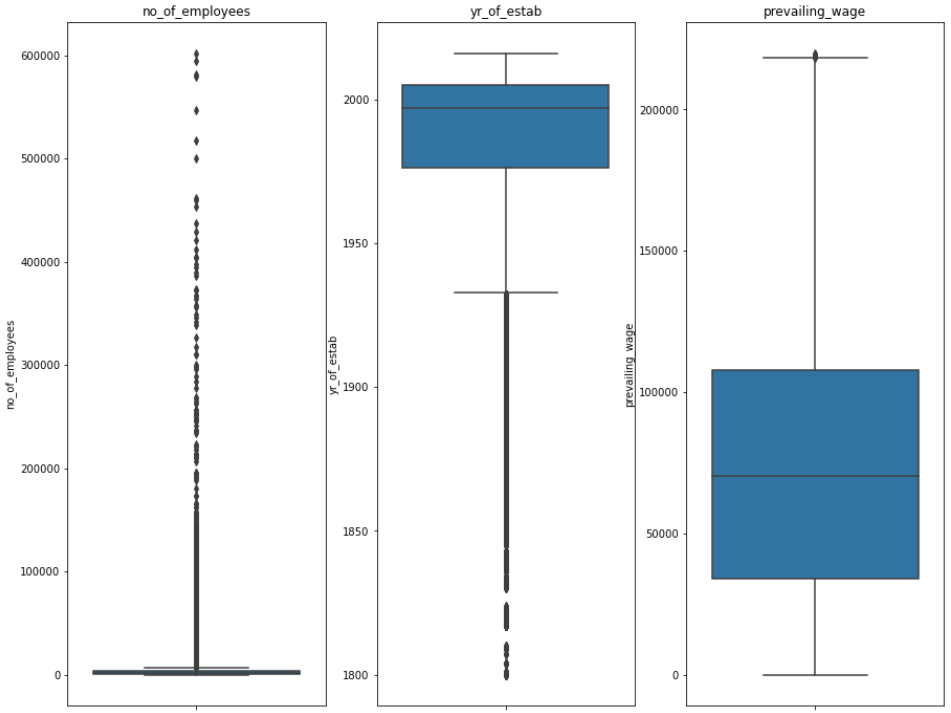
- Unit of Wage vs Status



- The hourly rate is the one with the least status certified. Perhaps that variable is related to the high school required jobs.

Data Preprocessing

- A duplicate value was realized, and it showed that no entry was duplicated.
- A missing value check was realized, and it showed that the dataset doesn't have any missing value.
- Based on the outlier results even though it shows results with outliers the number of employees is variable as the size of the companies changes significantly. The year is an established value of the companies. Finally, the prevailing wage was adjusted to change all the values above $ 220,000 with an upper whisky of 1.5 times the interquartile range plus the quantile 75.



Data Preprocessing

- For the preparation of the analysis the data set was dived into training test and test set. The first one will consist of 70% of the values and the test one will consist of the reaming 30%. Also, the values will be stratified in order to maintain the categories in a correct distribution.

Model Performance Summary

The model can make wrong predictions as:

- Model predicts that the visa application will get certified but in reality, the visa application should get denied.
- Model predicts that the visa application will not get certified but in reality, the visa application should get certified.

Which case is more important? Both the cases are important as:

- If a visa is certified when it had to be denied the wrong employee will get the job position while US citizens will miss the opportunity to work in that position.
- If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy.

How to reduce the losses?

- F1 Score can be used as the metric for evaluation of the model, the greater the F1 score higher the chances of minimizing False Negatives and False Positives.
- We will use balanced class weights so that model focuses equally on both classes.

Link to Appendix slide on model assumptions

Model Performance Summary -

Training Model

	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	0.985255	0.995963	1.0	0.771922	0.737497	0.718995	0.757793	0.757793	0.837295	0.837295	0.773099
Recall	1.0	0.931923	0.986066	0.999916	1.0	0.922102	0.888441	0.781247	0.881222	0.881222	0.930748	0.930748	0.891547
Precision	1.0	0.720067	0.991810	0.994075	1.0	0.777699	0.759417	0.794587	0.783257	0.783257	0.842233	0.842233	0.794034
F1	1.0	0.812411	0.988930	0.996987	1.0	0.843767	0.818878	0.787861	0.829357	0.829357	0.884281	0.884281	0.839970

Test Model

	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	0.663658	0.706567	0.693223	0.726975	0.721088	0.740319	0.734040	0.716510	0.746075	0.746075	0.725667	0.725667	0.743721
Recall	0.745739	0.930852	0.767483	0.896572	0.835064	0.902253	0.886582	0.781391	0.875416	0.875416	0.852106	0.852106	0.872478
Precision	0.749409	0.715447	0.771868	0.745926	0.767693	0.756074	0.756856	0.791468	0.773987	0.773987	0.764230	0.764230	0.772995
F1	0.747570	0.809058	0.769669	0.814340	0.799962	0.822720	0.816599	0.786397	0.821583	0.821583	0.805779	0.805779	0.819729

ns

Model Performance Summary

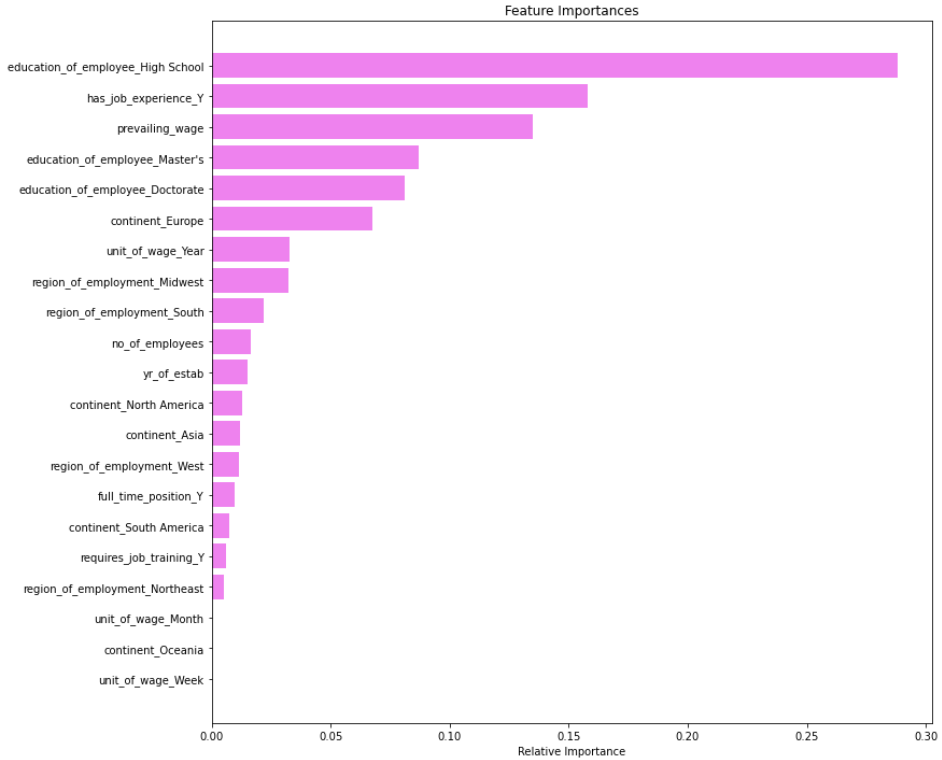
- The model selected based on the information presented is the Gradient Boost Classifier. This model is selected as the results of the F1 score were the closest ones.

- The XGBoost Classifier Tuned and the Stacking Classifier were not selected based on the results of the F1 score. As both showed better performance on the training set but showed a significant difference against the Gradient Boost Classifier. For the XGBoost Classifier Tuned shows a difference of over 0.08 points and the Stacking Classifier nearly 0.02 point, but he difference bewtween the Gradient Boost Classifier is under 0.008 points.

Link to Appendix slide on model assumptions

Model Performance Summary – Gradient Boosting Classifier

- The models give the highest importance to education if it has a high school diploma. This shows that the most common application for a visa comes from people with a high school diploma. This variable also indicates the jobs applications perhaps are for nonqualified jobs.
- The second variable of importance is the experience of the job. This can be understood as this reduces the cost of training the personnel.
- The following variables are considered if the applicants have either a master's or a doctorate. These variables show similar importance for the categorization.
- Following these variables one important consideration the model takes higher importance for where the applicant comes from, giving Europeans a higher change the those of North America and Asia.



APPENDIX

Data Background and Contents - Dictionary

- 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

Data Background and Contents

- The variable no_of_employees contained negative values, which are against the logic of the variable. These 33 values were adjusted to a positive value.

- A count check was realized for the categorical variables to understand the most common values for the different variables, and to seek any abnormal value in the data set.

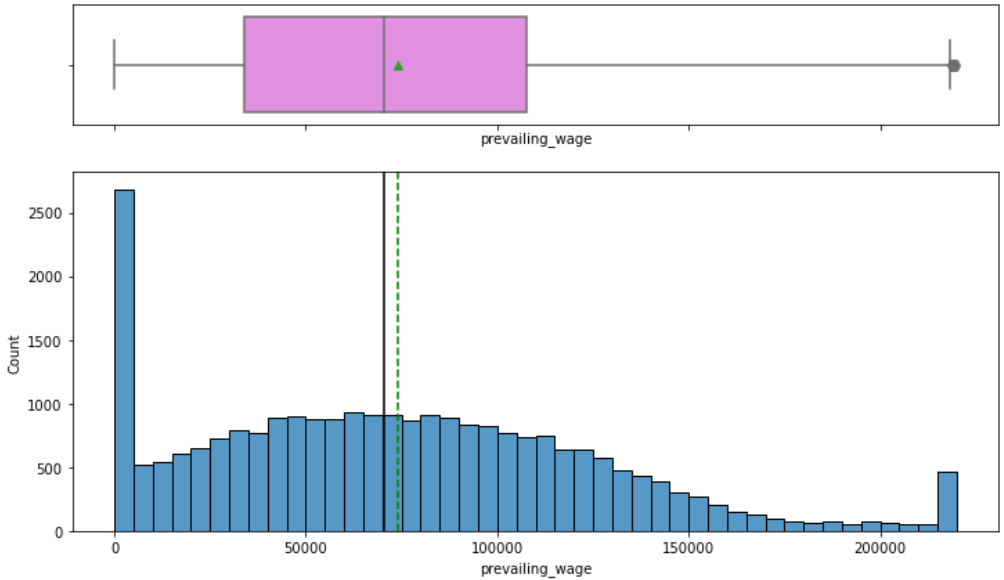
- The case_id variable was dropped from the data set as it was it served as the ID and it could affect the development of the model.

Data Background and Contents

Object	Float64	int64
• Case_id • continent • Education_of_employee • Has_job_experienies • Requieres_job_training • Region_iof_employment • Unit_of_wage • Full_time_position • Case_status	• Prevaling_wage	• No_of_employees • Yr_of_estab

Model Assumptions

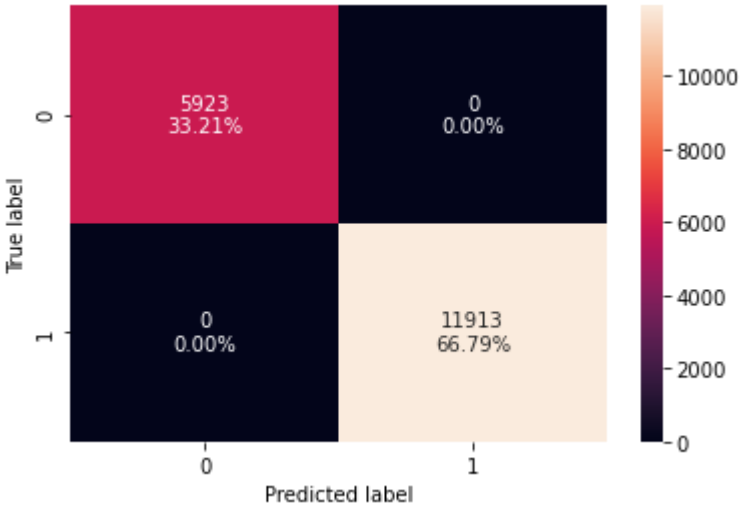
An change was realized for the treatment of the outliers of the outliers of the prevailing wage.



Decision Tree Model

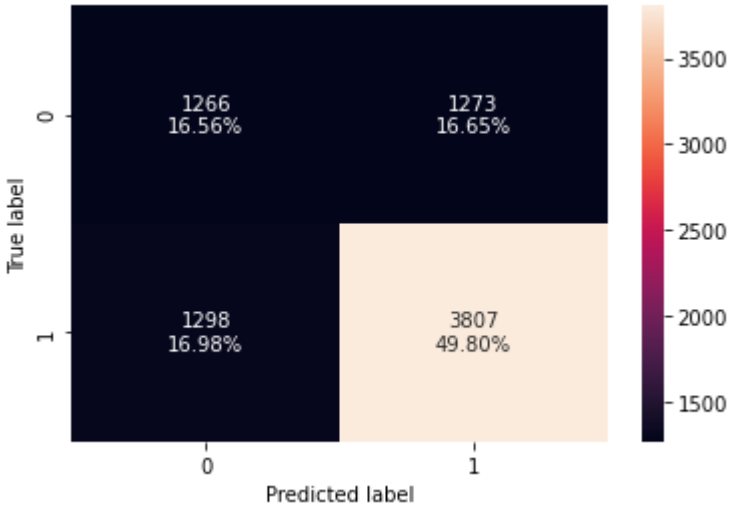
- Training Set

Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0



- Test Set

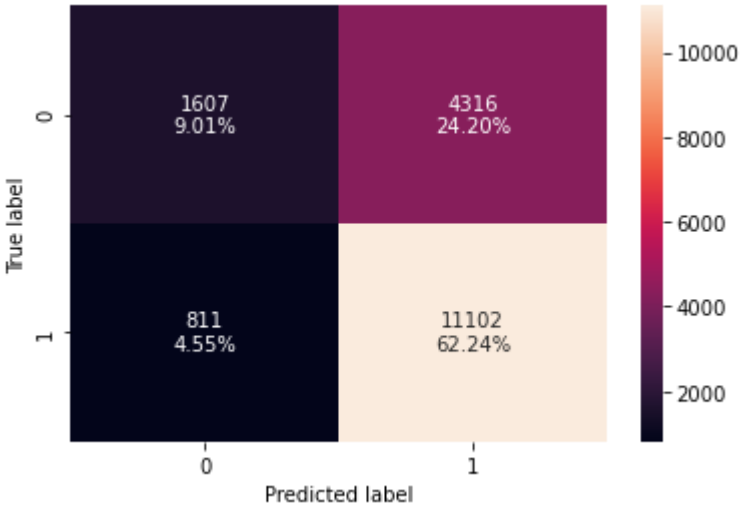
Accuracy	Recall	Precision	F1
0.663658	0.745739	0.749409	0.74757



Hyperparameter Tuning - Decision Tree

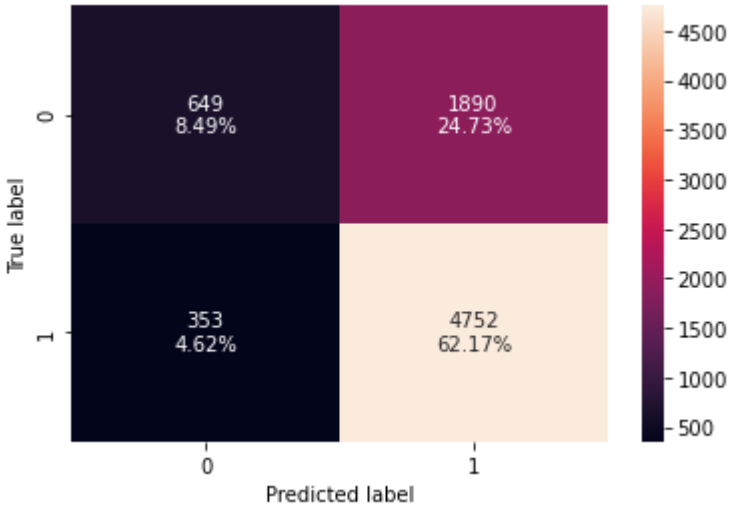
- Training Set

Accuracy	Recall	Precision	F1
0.712548	0.931923	0.720067	0.812411



- Test Set

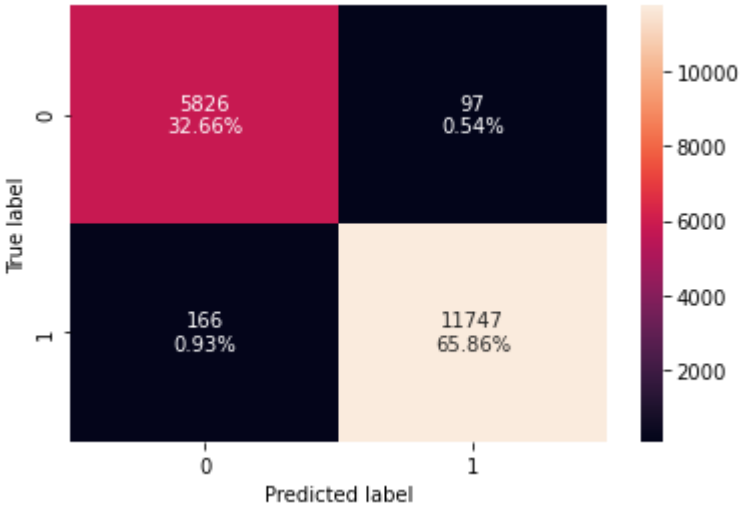
Accuracy	Recall	Precision	F1
0.706567	0.930852	0.715447	0.809058



Bagging Classifier

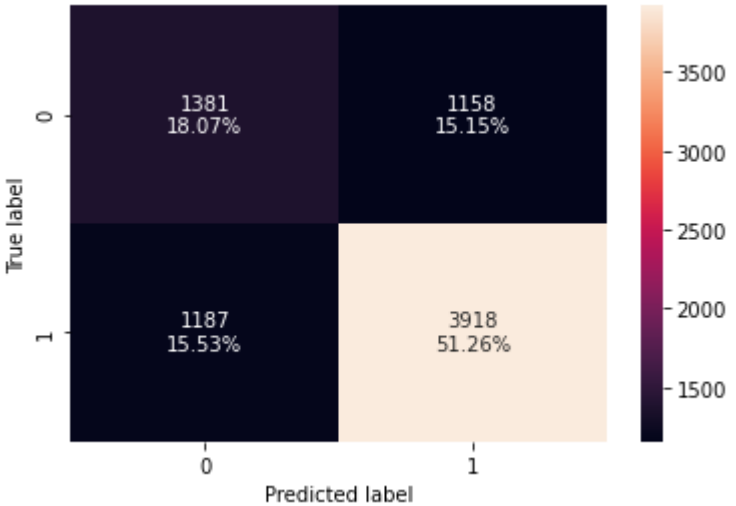
- Training Set

Accuracy	Recall	Precision	F1
0.985255	0.986066	0.99181	0.98893



- Test Set

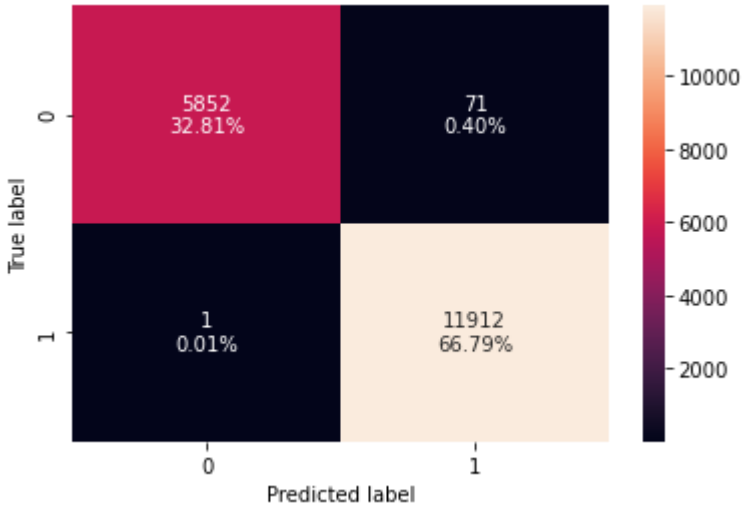
Accuracy	Recall	Precision	F1
0.693223	0.767483	0.771868	0.769669



Hyperparameter Tuning - Bagging Classifier

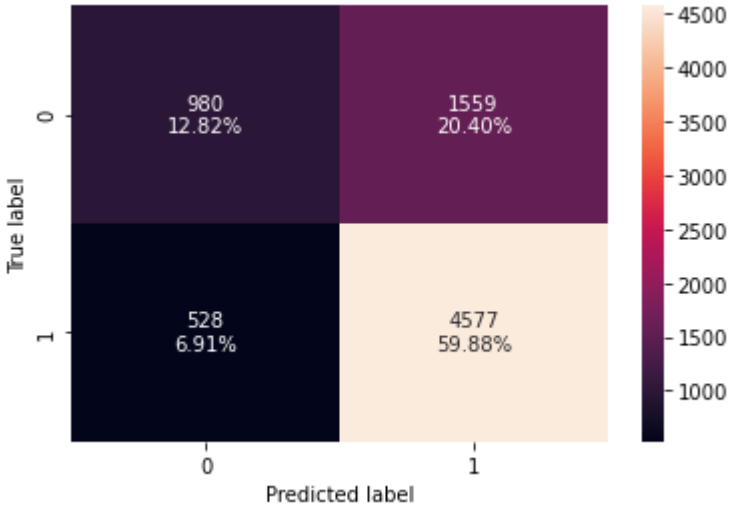
- Training Set

Accuracy	Recall	Precision	F1
0.995963	0.999916	0.994075	0.996987



- Test Set

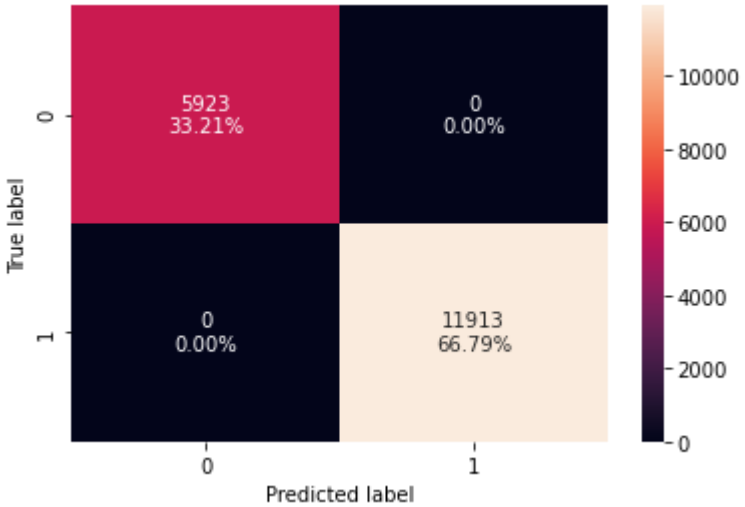
Accuracy	Recall	Precision	F1
0.726975	0.896572	0.745926	0.81434



Random Forest

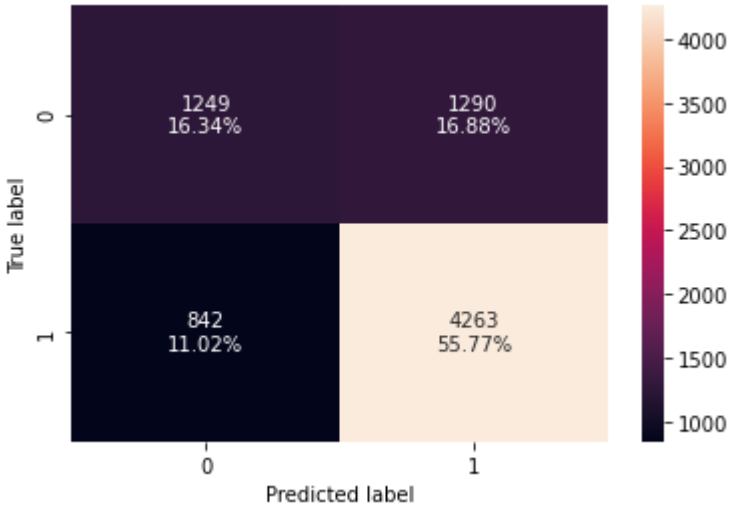
- Training Set

Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0



- Test Set

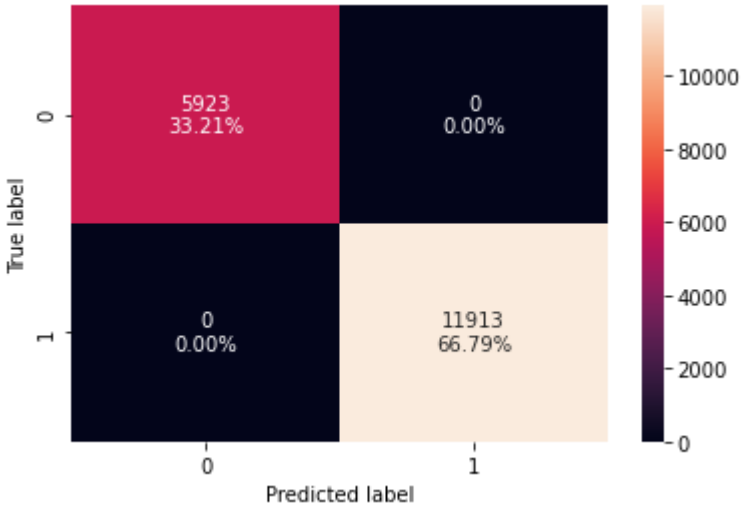
Accuracy	Recall	Precision	F1
0.721088	0.835064	0.767693	0.799962



Random Forest

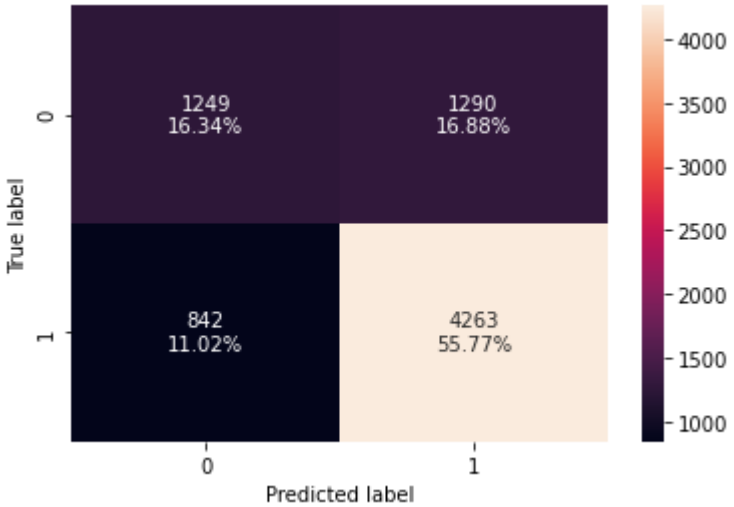
- Training Set

Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0



- Test Set

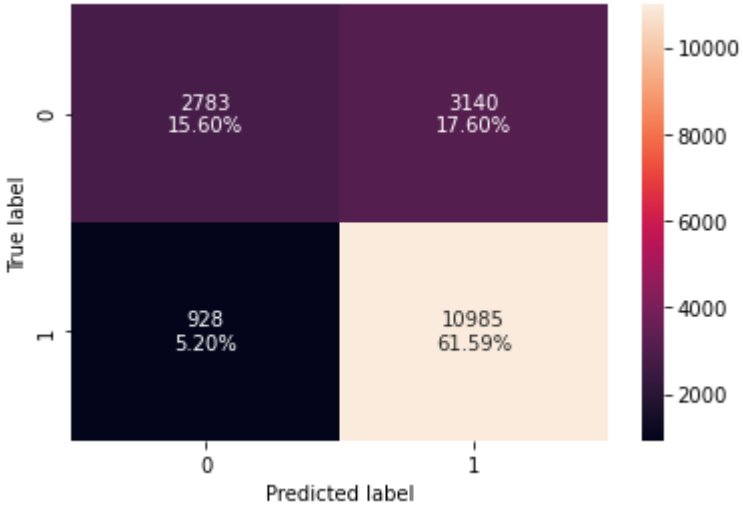
Accuracy	Recall	Precision	F1
0.721088	0.835064	0.767693	0.799962



Hyperparameter Tuning - Random Forest

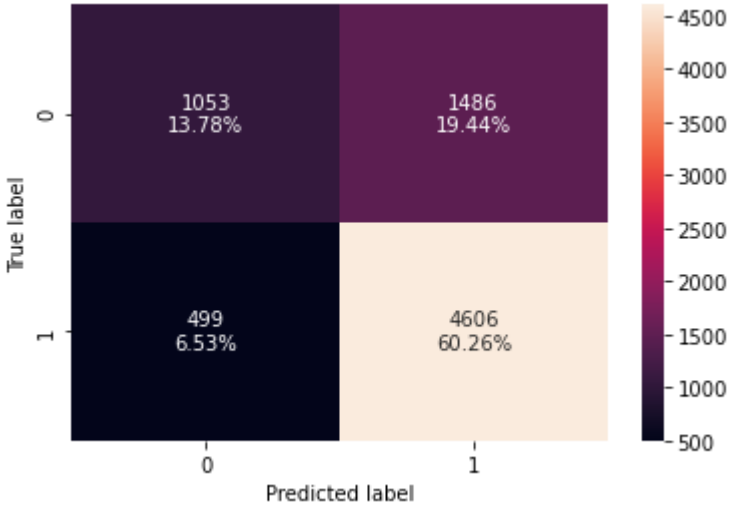
- Training Set

Accuracy	Recall	Precision	F1
0.771922	0.922102	0.777699	0.843767



- Test Set

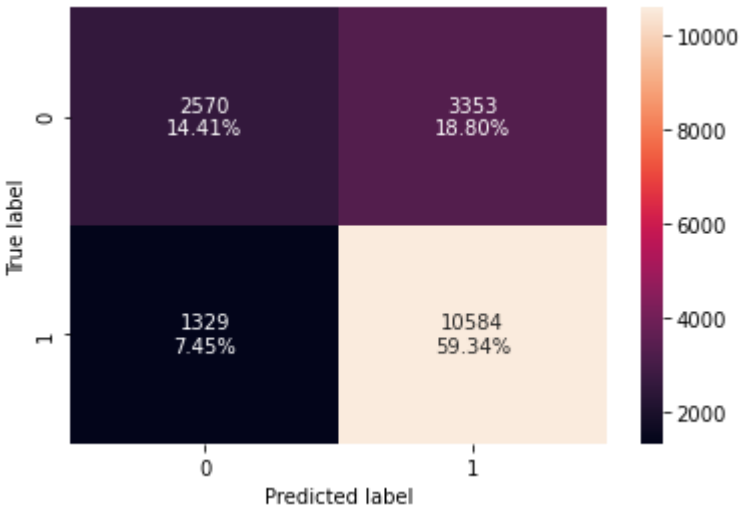
Accuracy	Recall	Precision	F1
0.740319	0.902253	0.756074	0.82272



AdaBoost Classifier

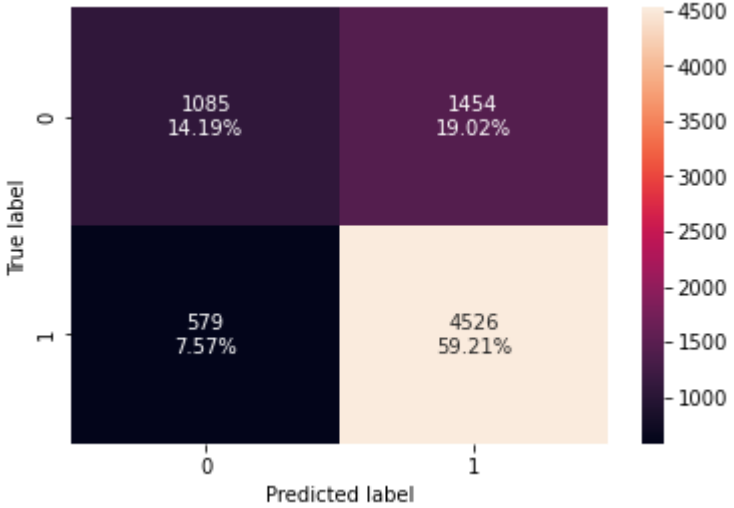
- Training Set

Accuracy	Recall	Precision	F1
0.737497	0.888441	0.759417	0.818878



- Test Set

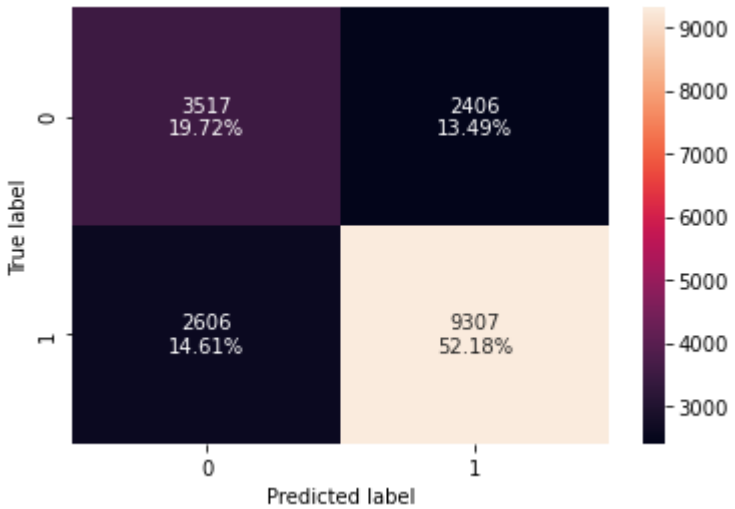
Accuracy	Recall	Precision	F1
0.73404	0.886582	0.756856	0.816599



Hyperparameter Tuning - AdaBoost Classifier

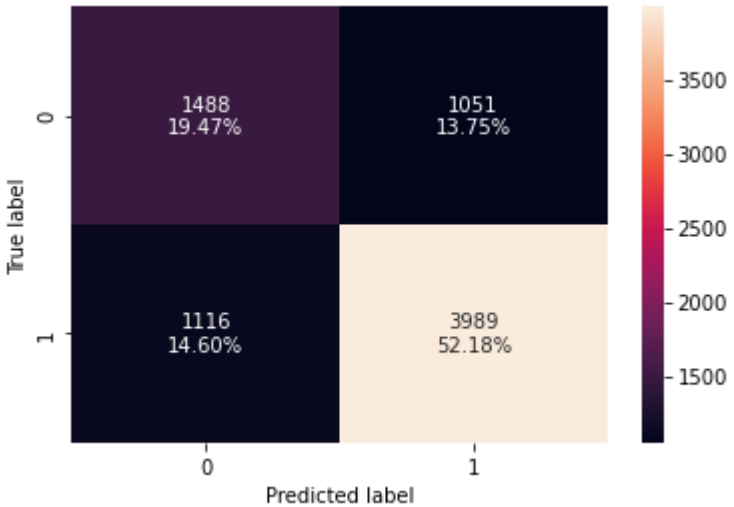
- Training Set

Accuracy	Recall	Precision	F1
0.718995	0.781247	0.794587	0.787861



- Test Set

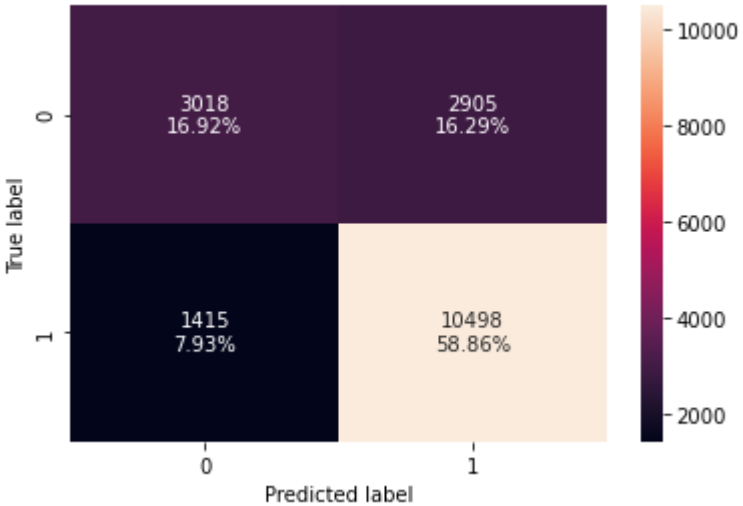
Accuracy	Recall	Precision	F1
0.71651	0.781391	0.791468	0.786397



Gradient Boosting Classifier

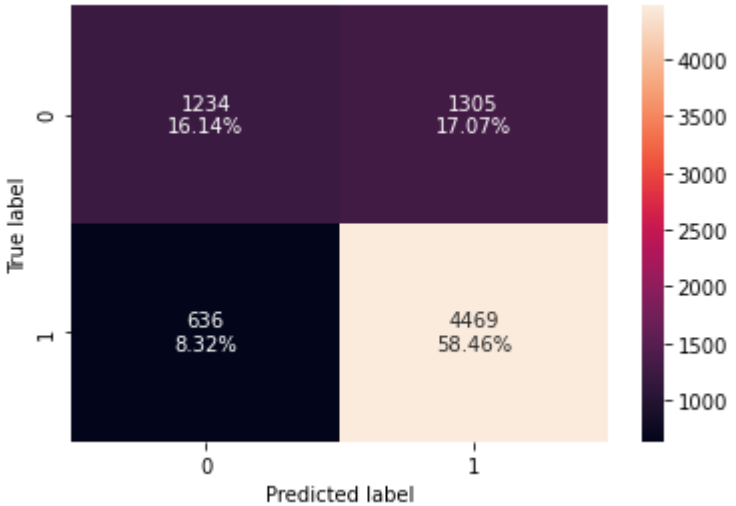
- Training Set

Accuracy	Recall	Precision	F1
0.757793	0.881222	0.783257	0.829357



- Test Set

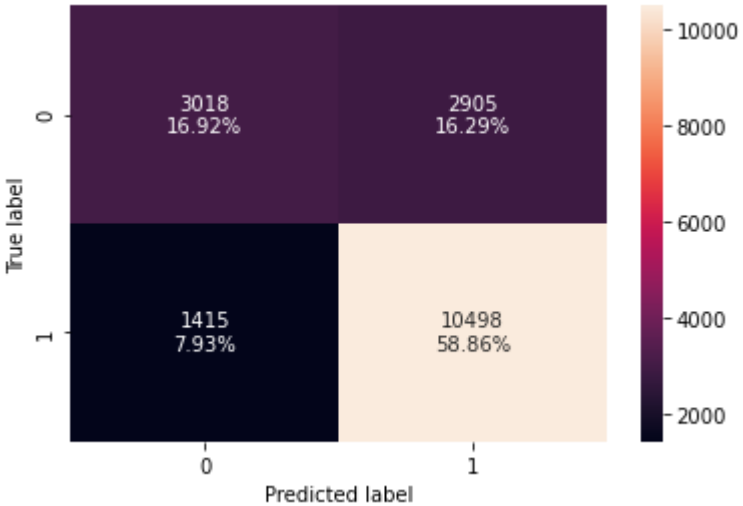
Accuracy	Recall	Precision	F1
0.746075	0.875416	0.773987	0.821583



Hyperparameter Tuning - Gradient Boosting Classifier

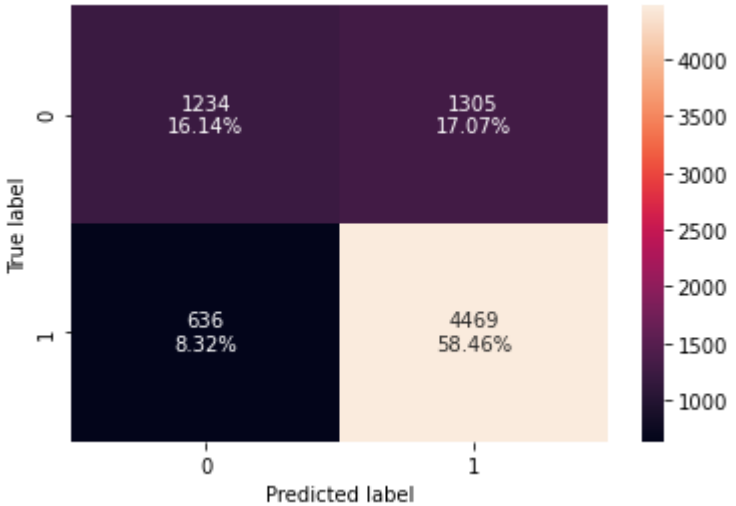
- Training Set

Accuracy	Recall	Precision	F1
0.757793	0.881222	0.783257	0.829357



- Test Set

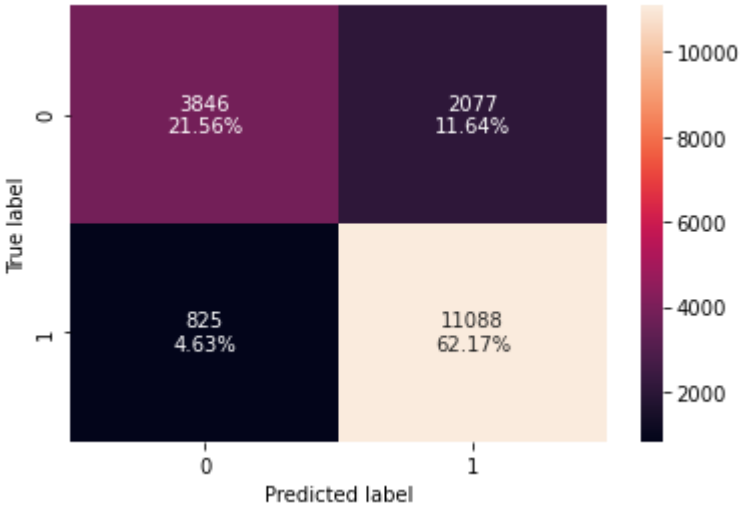
Accuracy	Recall	Precision	F1
0.746075	0.875416	0.773987	0.821583



XGBoost Classifier

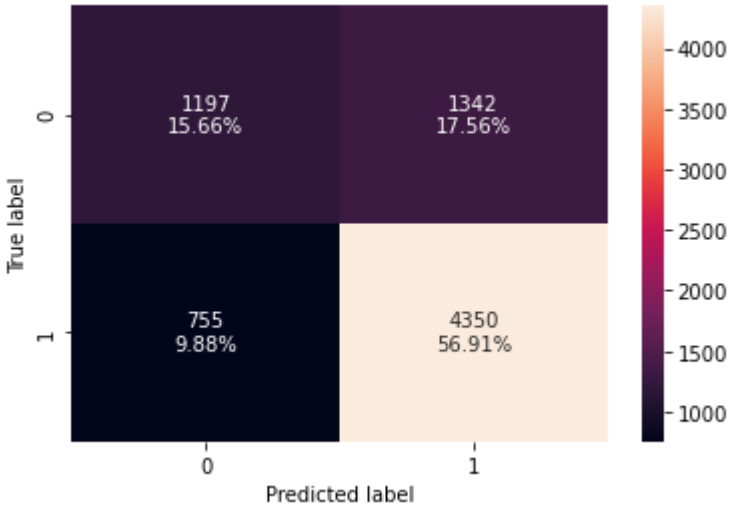
- Training Set

Accuracy	Recall	Precision	F1
0.837295	0.930748	0.842233	0.884281



- Test Set

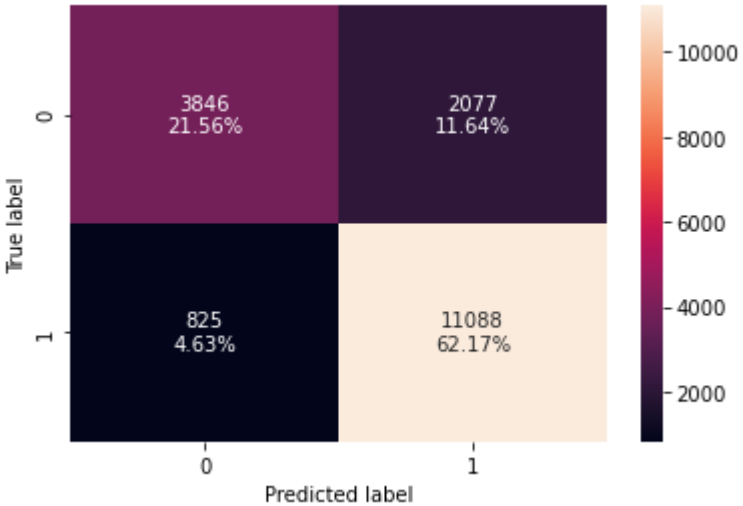
Accuracy	Recall	Precision	F1
0.725667	0.852106	0.76423	0.805779



Hyperparameter Tuning - XGBoost Classifier

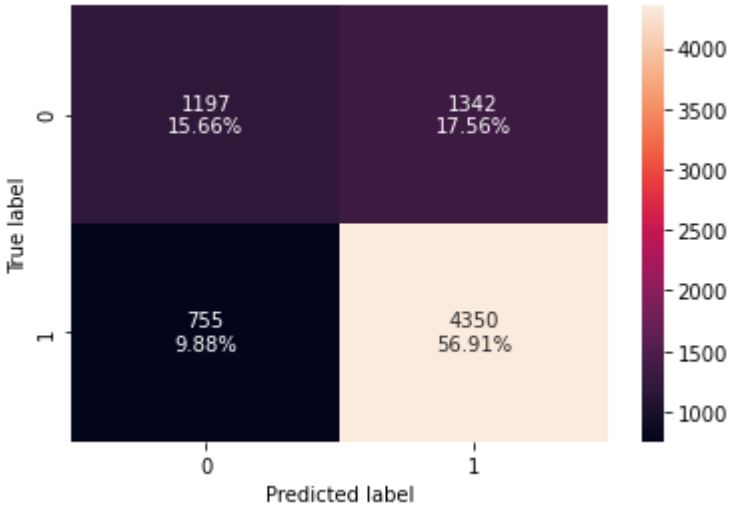
- Training Set

Accuracy	Recall	Precision	F1
0.837295	0.930748	0.842233	0.884281



- Test Set

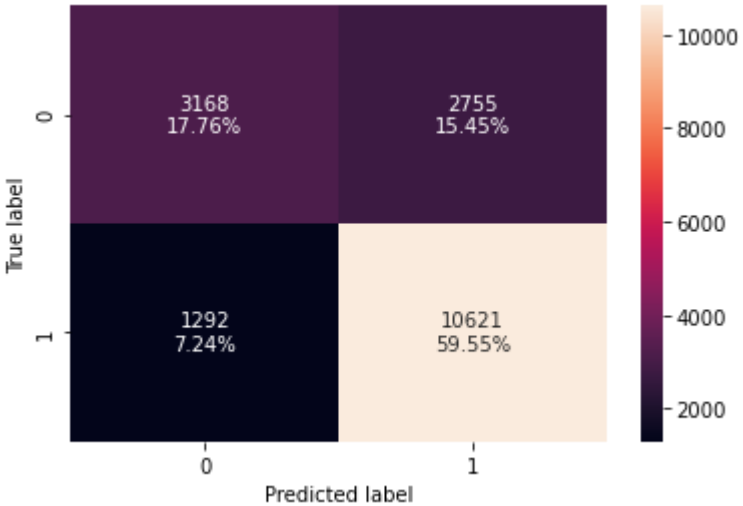
Accuracy	Recall	Precision	F1
0.725667	0.852106	0.76423	0.805779



Stacking Classifier

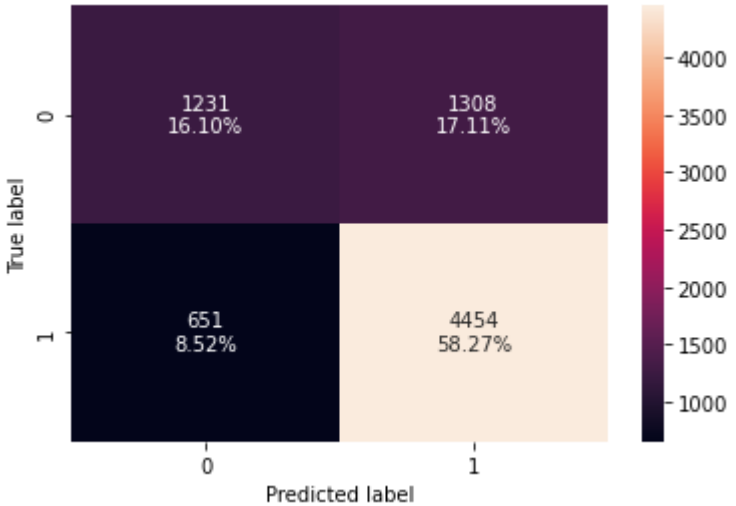
- Training Set

Accuracy	Recall	Precision	F1
0.773099	0.891547	0.794034	0.83997



- Test Set

Accuracy	Recall	Precision	F1
0.743721	0.872478	0.772995	0.819729



Great Learning

Happy Learning !