EasyVisa Project

Objective

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. 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 having higher chances of VISA approval. OFLC has hired your firm EasyVisa for data-driven solutions. You as a data scientist have to analyze the data provided and, with the help of a classification model:

* Facilitate the process of visa approvals.

* Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

Data Set

The data Set has 25480 samples with 12 columns

No Null values

The following are the columns available in the Data set :

* 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

EDA

Observation:

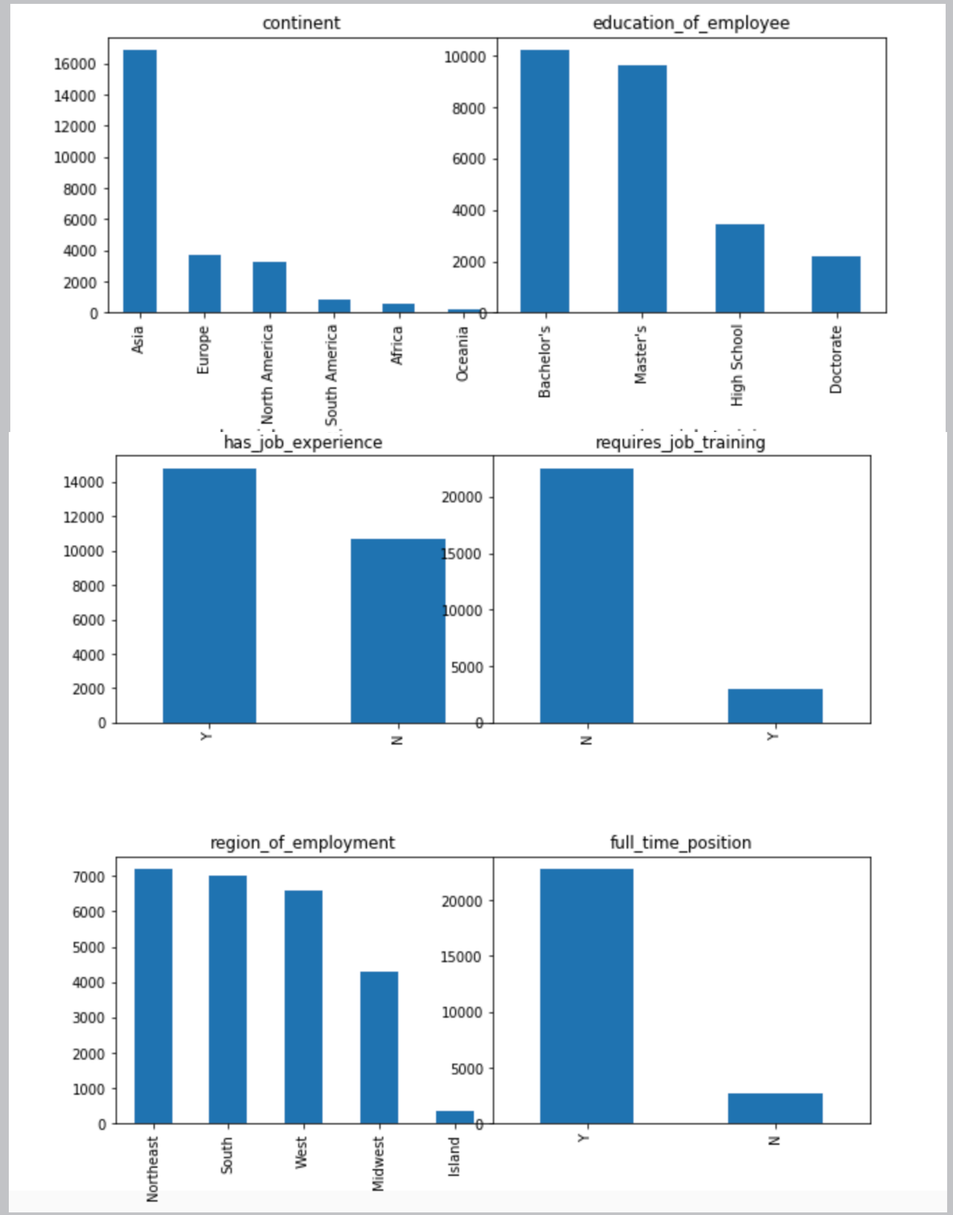
1) High number of applicants from Asia Continent

2) Mostly Educations of applicants are Bachelor & Masters

3) Applicants doent need job training , applicants are with the prior experience for the job

4) The regions of the employer are from Noreast & South

5) Applicants are mostly applied for the full time position

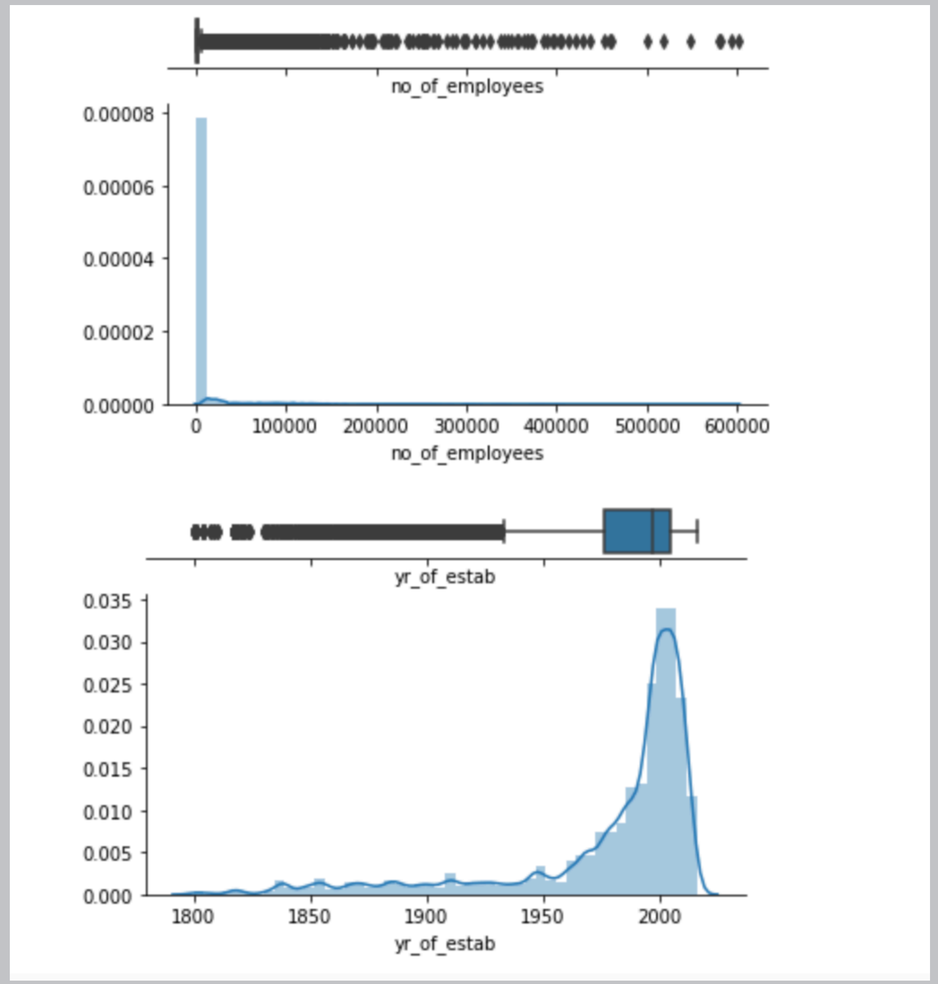


EDA

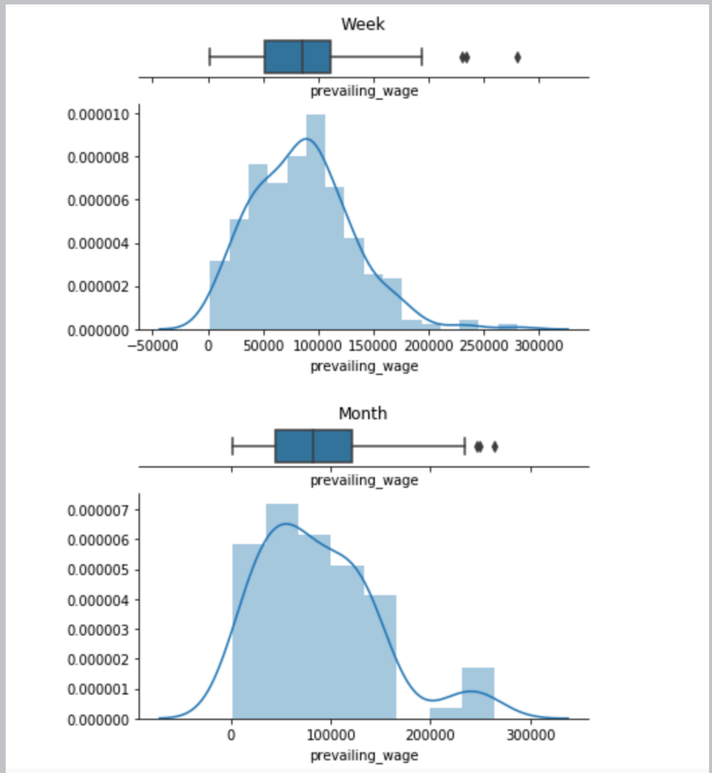
Observation: (Employers Profile)

1) Medain value for Years of establishments 1997

2) No of employees with mean 5667.043210 and median 2109.00



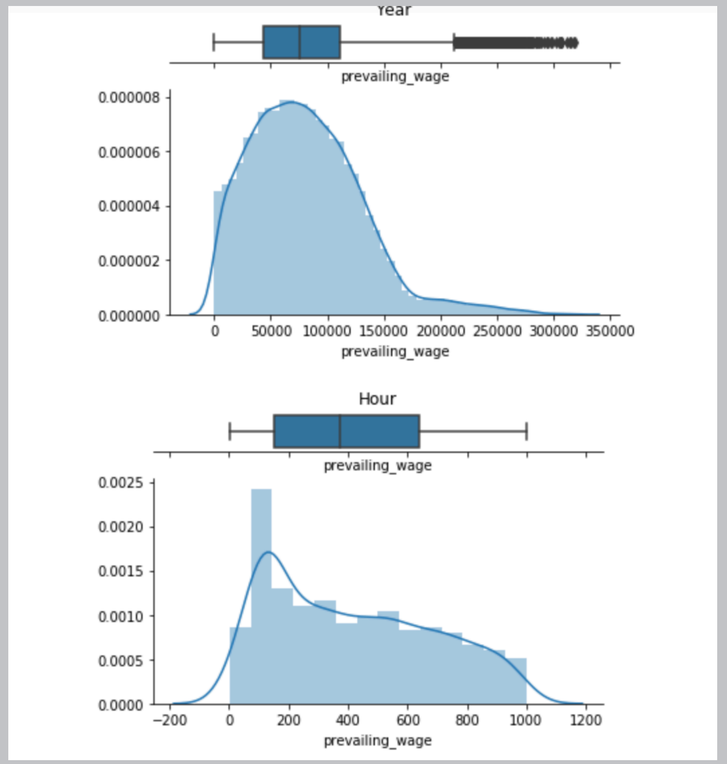
EDA



Observations :

The wages are at different unit. Hours , week , month, Year

The hight number of unit of wages are Year and it follows normal distribution of data. Except the Hourly wage type , others follows normal distribution.

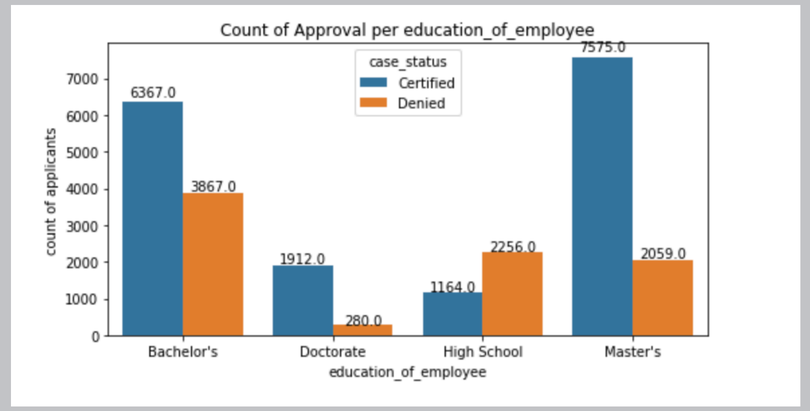


EDA

Observation :

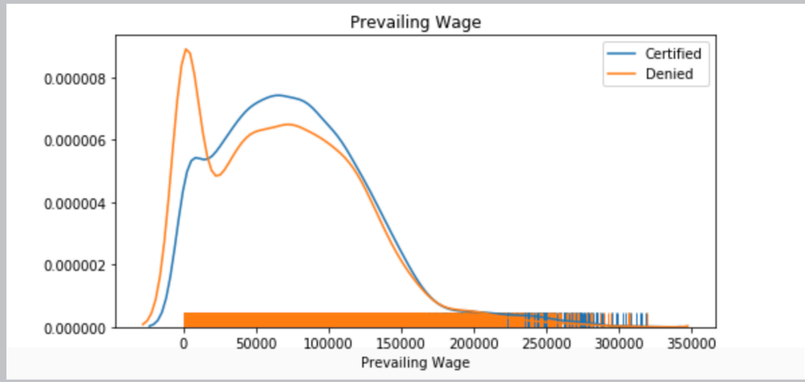
The visa status is rejected mostly with the applicants of High School.

The higher education level has high probability of visa approval

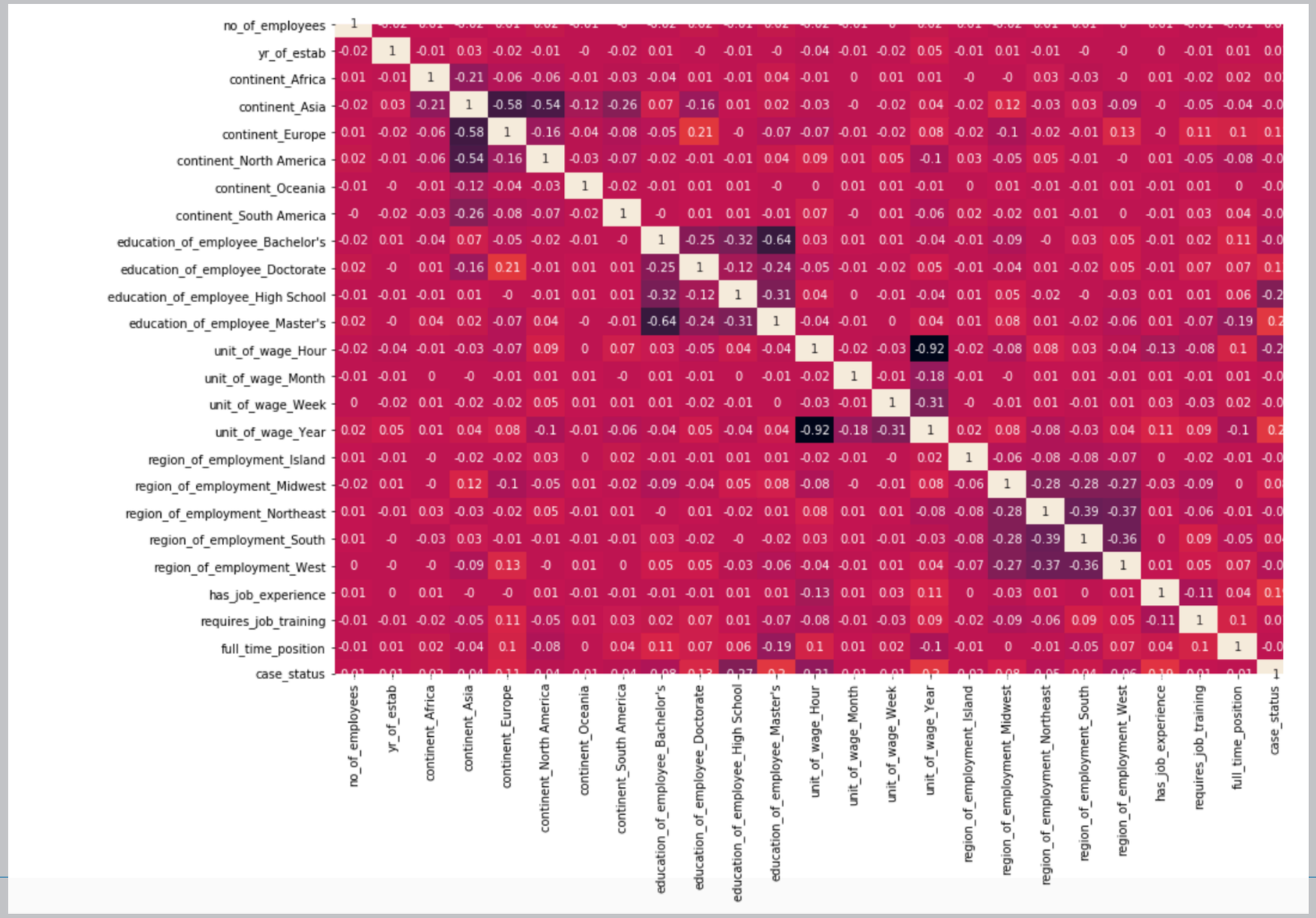


Observation :

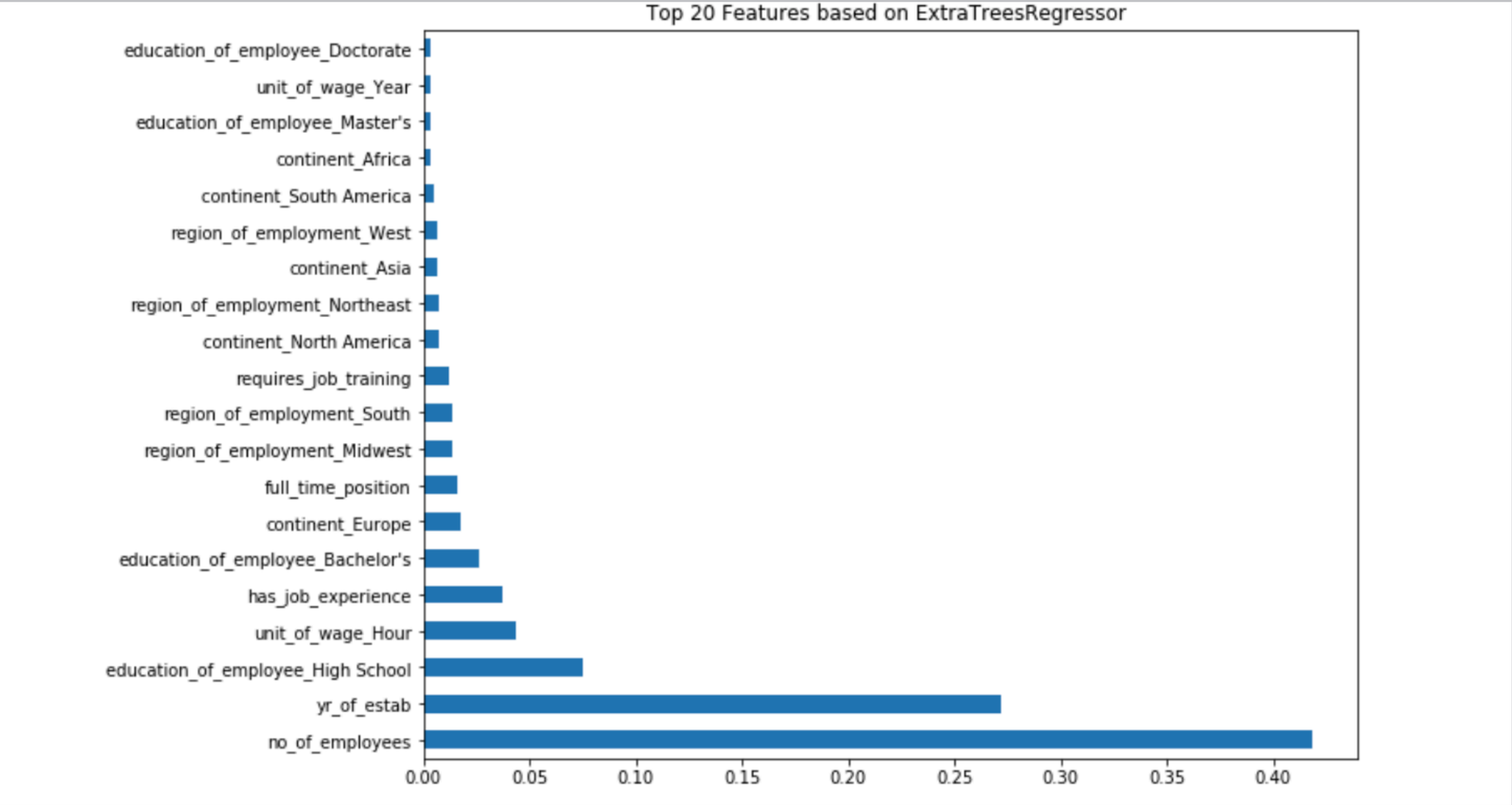
the prevailing wage is more for certified applicants than the Denied applicants



Corerelation Matrix



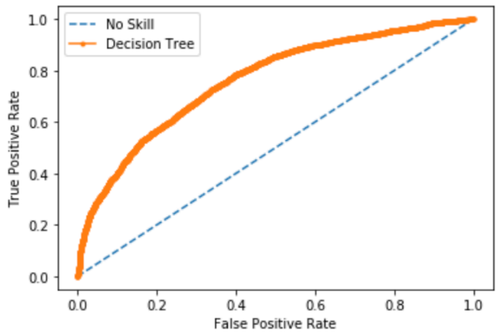
Top 20 Features



Model Performance And Evaluation

Sno	Model	Accuracy	Precision	Recall	F1_Score
1	Random Forest	0.690	0.756	0.790	0.7727229192606072
2	Decision Tree	0.659	0.748	0.738	0.7431952662721892
3	Logistic Regression	0.668	0.668	1.0	0.8008471252647267
4	Gradient Boosting	0.737	0.761	0.883	0.8175473579262214
5	Tuned Random Forest	0.701	0.761	0.804	0.7820952380952382

```
Accuracy: 0.737
F1 Score :  0.8175473579262214
Precision: 0.761
Recall: 0.883
No Skill: ROC AUC=0.500
Gradient Boosting : ROC AUC=0.763
```



Out of the 5 Models the Gradient Boosting model performs better with F1 Score as 0.81

The ROC curve of the final selection Gradient Boosting Model

Conclusion

The Following are the actionable Insights:

-> We can add additional features to the data set like Relevant Years of Experience

-> Field of Study can be included as part of education