A Predicative Analysis on Employee Promotion

Preseneted By
Biswajit Roy

Problem Statement & Goal:

Problem Statement: To predict from the given dataset with different attributes of Employees in an organization whether the employee is eligible for promotion or not

Goal: To achieve with Supervised Machine Learning to give an reliable platform for HR of the organization to smoothen his/her task and also automated.

Description of Dataset:

- Categorical columns are

'department','region','education','gender','recruitment_channel','no_of_trainings','previous_year_rating', 'age','KPIs_met >80%‘

- Continuous column are 'avg_training_score‘
- Target is 'is_promoted'

Distribution of Variables

•The avg of avg_training score is more in terms of promoted employes than not promotes employess also the awards won,KPI_met is high in terms of promoted employess though dataset is imabalanced.

•in terms of promoted employess avg of education score/items is higher than not promoted employess .

•in terms of Gender male employees are more in not promoted category than promoted.

•no_of_trainings and age are probably equally distributed and provided eaually to both the category (not_promoted/promoted)

•length_of_service avg score is merely high for not prormoted employees than promoted employess but our dataset is imbalanced

•As we can observe no such biassness from recruitment_channel avg score

•previous_year_rating avg score is more in terms of Promoted employess and Not prormoted employess

•in terms of eduction specfic education mattered probably most in terms of promotion.

•Region wise with education probably has an slight impact in terms of promotion

Assumptions from Density plot: Promoted Vs Non Promoted

•in terms of no of trainings the Not promoted employess got more no of trainings but keeping in mind the dataset is imbalanced over here.

•there is such significant change can be visualised in terms of age in non_promoted or promoted employess.

•in terms of previous year rating no of promoted employess got significant rise in rating '5'.

•with in the 2 to 5 years of service the no of non promoted employess is more than promoted employess,most number of promoted employess are scattered with in 7 to 33 years. And No promoted employess served after than 35 years.

•in terms of KPI there is marginal changes as the mean score of KPI of promoted employess is slightly higher than non promoted employess.

•in awards won promoted no of empoyess got slightly more no of awards with respect to non promoted employess.

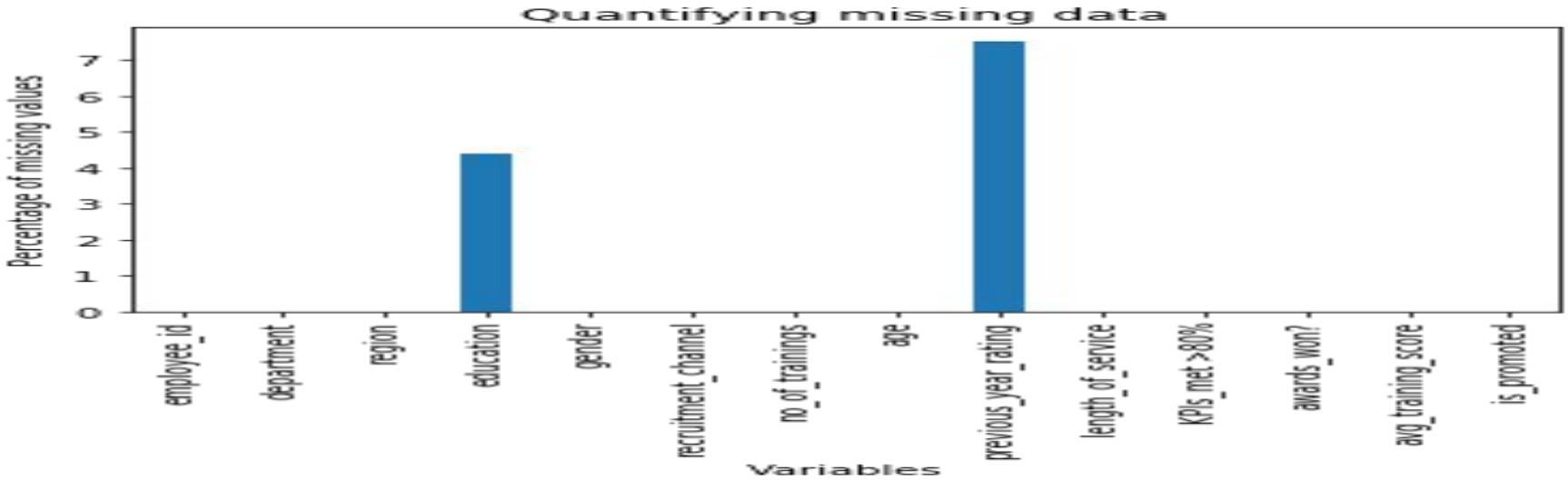
•. in avg_training_score between 42 to 60 non promoted employees got maximum no of score still did not got promotion but avg score of training is more for promoted employess than non promoted employess.

Study of Independent variables with Dependent Variable

- region 7,2,13,27 has the most score of avg training score of promoted employees region 33 & 21 has low avg training score.region 3 has low nof of observation in terms of avg training score.
- Most of Promoted employees won the award.
- Most of promoted employees met kpi and their avg score of training is more.
- Most no of promoted employees are in the range of age between 24 and 42 .promoted employees are less in the range of age between 20 to 22 & 50 to 60.
- promoted employees got no of trainings with in the range of 1,2,3 mostly.
- reffered recruitment channel has least observations , as in other channel recruited employees got promotion.
- Male employees got more promotion than female employees.
- in terms of all category in Education variable ,promoted employees are scoring more though with less samples ,Below secondary category has least no of samples

Moving To Predicative Analysis

- Treating Missing values with a Business logic.
- Encoding the categorical columns for the interpretation of the Model Building.



Study between Independent variables

- Female employees got less no of trainings.
- Most no of training done with bachelor's Educated employees and less no of training recieved by Below Secondary's educated employee.
- referred employees got least no of trainings and Sourcing employees got more no of trainings.
- procurement department empoyees got more no of trainings and also sales&Marketing , but HR, Finance, Analytics & Leagl recieved less no of trainigs.
-
- sales&Marketing,operations &procurement got avg training scoreand also promoted more and Legal,Hr & R&d department got less no of promotions.
- previous rating score is distributed over all department sectors

Model Testing

I have used CatBoost Model As because classification report came good compared to other model (Random Forrest,XGBoost,LightBGM,Naïve Bayes and VotingClassifier) .The 3 Fold StratifiedKfold Observation is given below

****

The roc_auc_score on Train 0.7634695359917438
The roc_auc_score on Test 0.71373414826298
The Precision Score 0.908857910524456
The Recall Score 0.9078847291472999
The F1 Score 0.8934756917217709

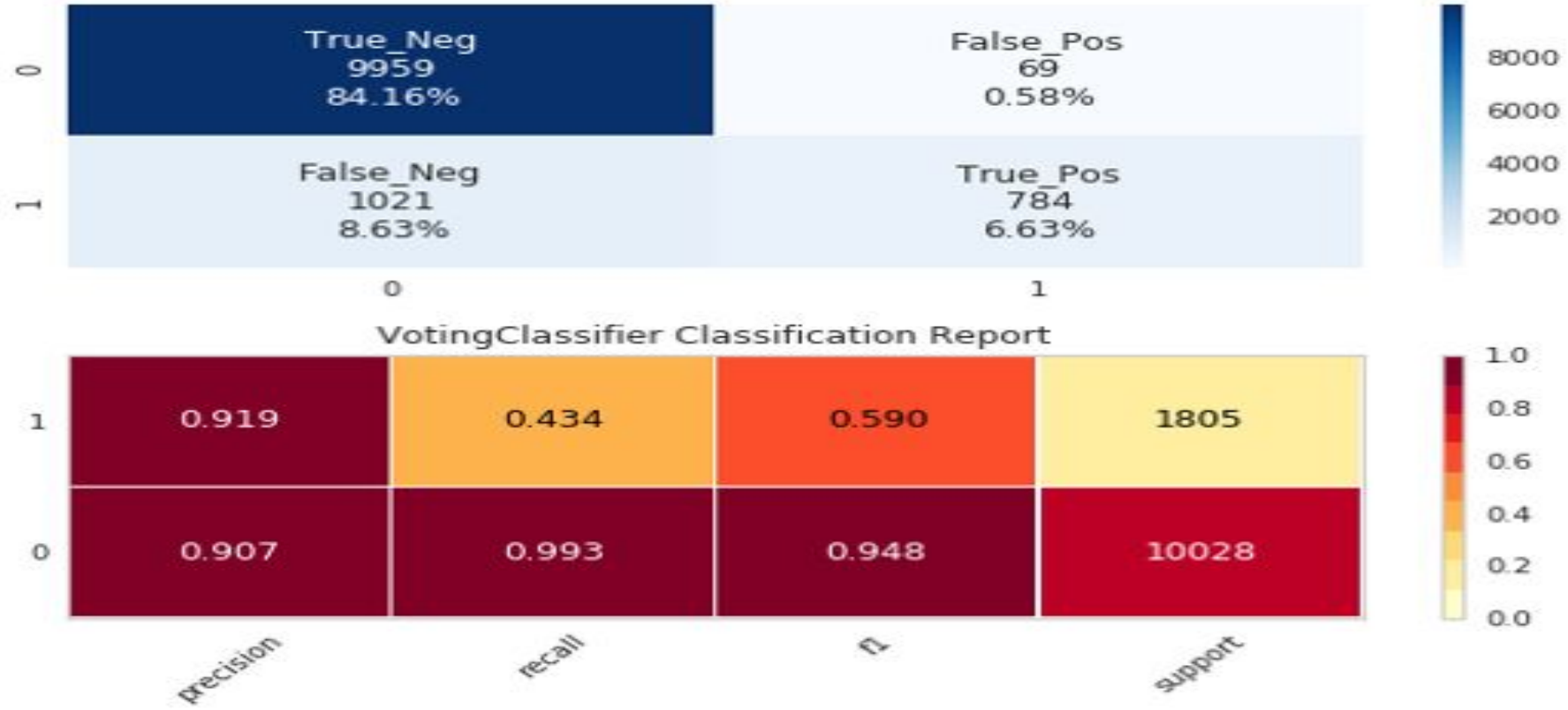
****

The roc_auc_score on Train 0.8228698149336982
The roc_auc_score on Test 0.7699169748526841
The Precision Score 0.9255678043244916
The Recall Score 0.9249556325530297
The F1 Score 0.9164980921477223

****

The roc_auc_score on Train 0.7599875404269707
The roc_auc_score on Test 0.71373414826298
The Precision Score 0.908857910524456
The Recall Score 0.9078847291472999
The F1 Score 0.8934756917217709

Confusion Matrix & Classification Report



CONCLUSION:

As we can see from the test data amd model validation that feautres those were important to classify an employee to be promoted/Not.

we can see that avg_training_score , department,KPIs_met >80% & previous_year_rating are the key factors for promotion