ALGORITHMIC STAFF PROMOTION

1.0 Overview

In a recent global survey of HR leaders worldwide, it was found that only 2% of employee in organizations believe that their company current performance-management systems are very effective [1]. The use of Machine learning algorithms in HR of organization can unarguably simplify and improve integrity and objectivity in Human Resources promotion process.

To instill objectivity to promotion process, machine learning methods are to be employed to study the pattern of previous staff promotion process of YAKUB TRADING GROUP, the gained insight is to be used to predict promotion eligibility. Also, the insight from the prediction algorithm is to be used to understand the important features among available features for the future prediction of promotion eligibility.

Problem Statements (Ask)

1. What is the promotion pattern of YAKUB TRADING GROUP based on the available data?
2. From the insight above, what is the important features among available features that can be used to predict promotion eligibility?

2.0 Data Preparation

Data Source

The dataset used in the project was copied from open source in Kaggle repository. The dataset was used for Data Science Nigeria AI bootcamp 2019 pre-qualification competition. The dataset represented information about the list of employees in a private company in Nigeria. Each instance represented an employee unique information (features) as it relates to the outcome of if the employee is promoted or not. There are 38312 employees in the dataset with 19 attributes. Each of the features are described in the table 1.

The first 18 attributes are the features relating to the employee qualities that counts to the final decision of if the employee should be promoted or not. The qualities comprise of of 6 numerical values, 2 dates and 10 categorical features. The 19^{th} feature represents the class if the employee is promoted or not. The 19^{th} feature is represented with 1 or 0 with 1 and 0 representing promoted and not promoted respectively. The description of each of the features of the dataset is represented in the table below.

Table 2.1 Dataset Variable description

Variable Name	Description	Variable type
EmployeeNo	Employee unique staff ID	Categorical and Numeric
Division	Employee Department	Categorical
Qualification	Employee Highest qualification	Categorical
Gender	Male or Female	Categorical
Channel*of* Recruitment	How the staff was recruited	Categorical
Trainings_Attended	trainings attended by employee	Numeric
Year_of_birth	Birth year of employee	Date
Last_performance_score	Prior year overall performance on a scale of 0-14	Numeric
Year_of_recruitment	Employee Year s recruited	Date
Targets_met	If employees met set target or not: If met=1; not=0.	Numeric
Previous_Award	previous award won. If yes = 1; No= 0.	Numeric
Training_score_average	Feedback score on training attended	Numeric
State_Of_Origin	Employee state of origin	Categorical
Foreign_schooled	Employee with any post-secondary education abroad	Categorical
Marital_Status	Marital status of employee	Categorical
Past_Disciplinary_Action	staff with past disciplinary action	Categorical
Previous_IntraDepartmental_Movement	staff who that have worked in more departments	Categorical
No_of_previous_employers	Past employers before joinning present employer	Numeric
Promoted_or_Not	The outcome	Numeric

The dataset was downloaded from Kaggle and saved on my laptop memory for easy access. The dataset was upload for analysis.

3.0 Data Cleaning and Processing

We did a few data cleaning exercise on the downloaded HR dataset to ensure we achieve a good performance of the model. The under listed cleaning exercise were performed on the dataset.

Checking for duplicate: We established that no staff ID appeared twice.

Identification and filling of missing values: We identified that Qualification variable has 1679 null values. To refill the missing values, we computed the qualification value with the highest frequency to replace the missing value.

	Total	%
Qualification	1679	4.4
EmployeeNo	0	0.0
Previous_Award	0	0.0
No_of_previous_employers	0	0.0
Previous_IntraDepartmental_Movement	0	0.0

Fig 1. The sorted value and percentage of the variable with missing values

```
count                      36633
unique                         3
top        First Degree or HND
freq                       25578
Name: Qualification, dtype: object
```

Fig 2. The qualification value with the highest frequency

Data Manipulation: The Age and number of years spent by each of the employees were extracted from date of birth and year of recruitment.

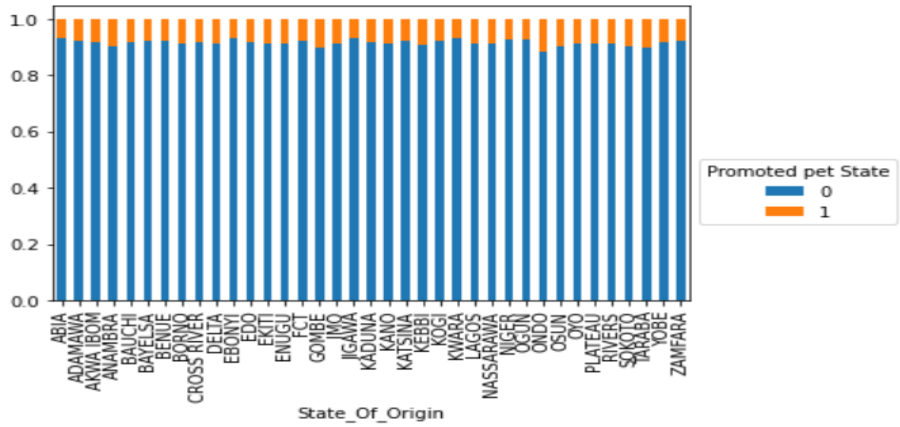
Age = present year - year of recruitment

years_of_service = present year – year of recruitment

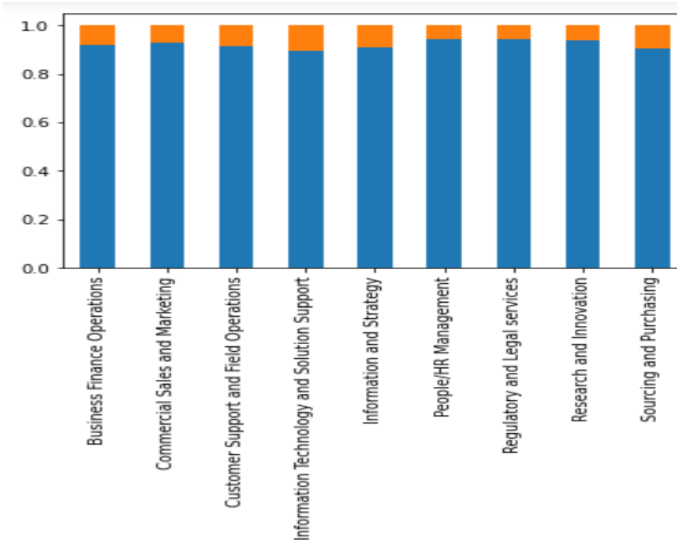
4.0 Insight and Exploratory Data Analysis

To provide more insight in the dataset, a few explanatory analyses were performed. The promotion based on different variables (state of origin, departments, sex, qualification type, gender, meeting the target and Last performance score) in dataset were investigated. This is done to determine the rate of promotion based on the variables in the dataset.

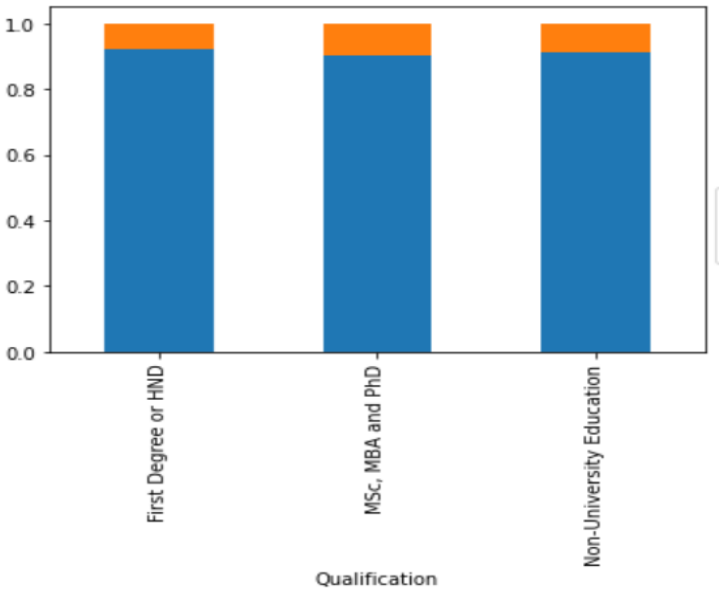
A. We observed that promoted and not promoted across state of origin, departments, qualification type, and gender have no significant difference. It can be deduced that promotion is not based on these variables. See the figures below for more graphical explanation.



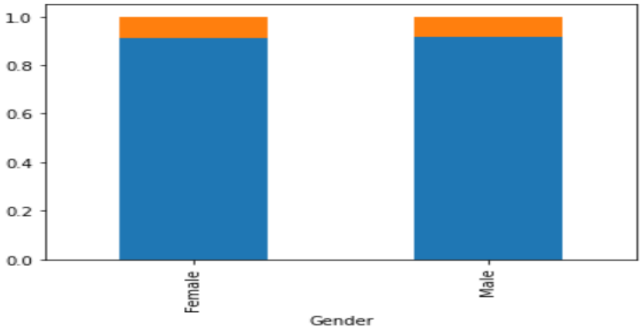
i. Promotion per state of origin plot



ii. Promotion per department plot

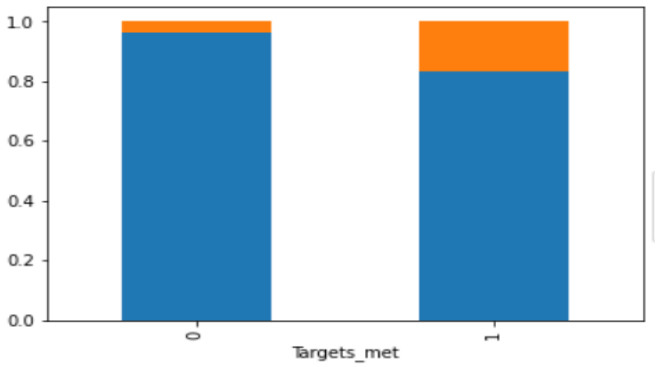


iii. Promotion per Qualification plot

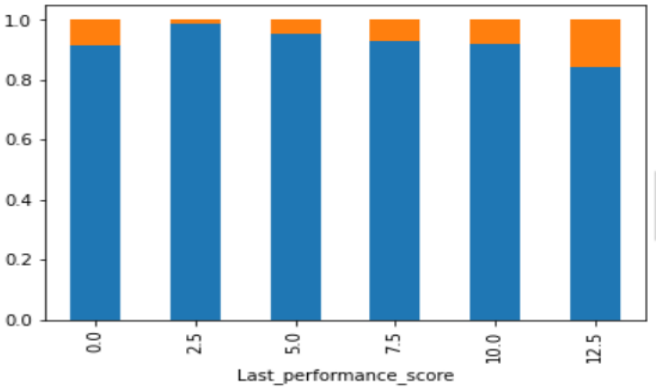


iv. Promotion per Gender plot

B. Variables with significant difference in number of promoted employee: It was observed that number promoted are significant among employee that met target and those with high score in their performance score. It can be deduced that staff that met target and have high score in performance score has higher probability of getting a promotion. See the figures below for more graphical explanation.



i. Employee promotion and employee that met target plot



ii. Employee promotion based on last performance

5.0 Dataset Split, models development and Promotion Prediction

The cleaned and pre-processed dataset was split in ratio of 80:20 for training and holdout respectively using sklearn library in python. The 20% holdout was set aside for validation/test of model. The 80% dataset (training dataset) was trained using three different algorithms The algorithms trained were validated by using them to predict the holdout (20%) sample.

Three different classification models namely Decision Tree, KNN and Random Forest Classifications were used, and the percentage accuracies were compared to suggest the best method to develop the algorithms for the predictions.

The comparison of the three models employed showed that Random Forest algorithm has the highest performance. The diagram below has the performance records for each of the three models are summarized below.

Model type	Training % Accuracy	Validation % Accuracy	Rating
Random Forrest	99%	93%	1st
Decision Tree	99%	89%	2nd
KNeighbour	94%	91%	3rd

Table 2. performance records for each of the three models

6.0 feature importance and Conclusion

To find the feature importance of each of the variables, the most performed algorithm (Random Forest) was used for the tabulation and plot.

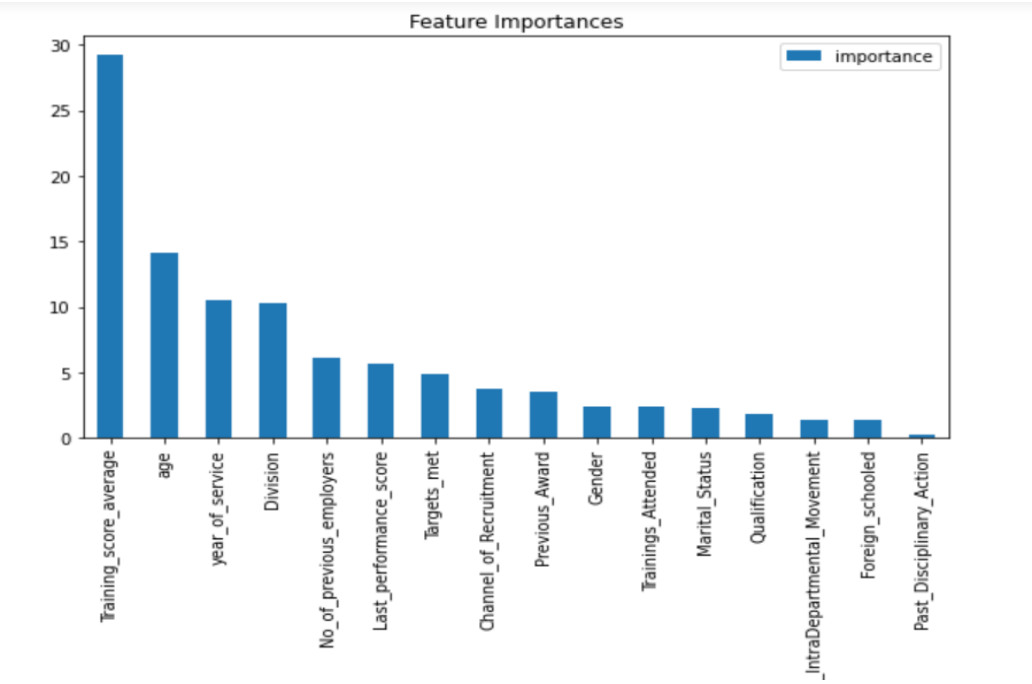


Fig 6.1 Feature Importance plot

In conclusion, with the percentage accuracy of 99% and 94% for training and validation dataset respectively, it is established that Random Forest model has the best performance when compared with other employed algorithm for promotion prediction for YAKUB TRADING GROUP. In addiction Training score average, age, years of service in the company, division, last performance score and targets achievement are the most contributory factors to the promotion of employee.