Table of Contents

HR Analytics – Promotion Prediction

Rudra Rajeev Grover

Code available at: https://smcse.city.ac.uk/student/acwk480/mynotebook.html

Part 1

1.1 Domain Description

The field of Human Resources Analytics (HRA) encompasses the use of various techniques for analysing HR related processes to aid in organizations' 'data driven decision making' [1]. Such processes relate to wide ranging HR problems, from employee performance analytics, to understanding hiring, training, promotion, and attrition analytics – with respect to the individual capacities of the organization [1]. HRA seeks to perform such analysis, to not just describe employee behaviour in organizations, but also make diagnoses for problem solving, as well as provide a 'predictive' and 'prescriptive' dimension to improve current HR situations [2].

Many authors [1][2] note the relative lack of empirical studies in this domain, seemingly stemming from slow adoption of technology driven HRA within organizations. Within practical HRA in organizations, there also seems to exist a lack of data driven approaches, as smaller proportions of practitioners are well versed in the statistical and technological competencies to best perform HRA tasks at the highest level [2].

Data driven approaches in HRA often rely on a mix of internal and external data, depending on the task to be performed. This study uses anonymised internal data from an organization, to aid in expediting the employee promotion process. The data is sourced from a data science hackathon site, Analytics Vidhya [3]; and consists of metrics mentioned in the data description section [Appendix A]. The problem specifically mentions that the traditional employee promotion cycle, consisting of identifying, training and evaluation, and subsequent promotion causes a loss of work time. A data driven approach is required to identify promotable employees earlier in the pipeline with strong accuracy. Using given employee demographic and performance data, with the outcome variable being promotion (1, 0), this study seeks to perform predictive modelling to ascertain this outcome variable for unseen data (with high accuracy). Model performance is measured by F1 score.

1.2 Analytical Questions

1. What factors are correlated with promotion? Are these performance centric factors? Or is there any bias?

2. How good a machine learning model can be built for this problem?

- What is the effect of outlier treatment on model performance?
- What is the effect of feature engineering on categorical variables on model performance?
- Compare and contrast the two models (Random Forest and Gradient Boosting Machine)

Important Note: As this is a practice Data Science competition, the submission data for this competition will be our test data set and the score returned by the Analytics Vidhya site on our test data will be the test F1 score.

1.3 Analysis Strategy

1. Pull data from competition
2. Exploratory data analysis using tabular outputs and visual plots
    a. Univariate and bivariate analysis
3. Build models in 3 stages:
    a. After simple recoding and missing value treatment
    b. After outlier treatment
    c. After feature engineering to add categorical variables using Count encoding, target encoding and One-hot encoding

d. Note: Use 4-fold cross validation; use grid search on select parameters for hyperparameter turning; use Random Forest and Gradient Boosting Machine models.

Part 2

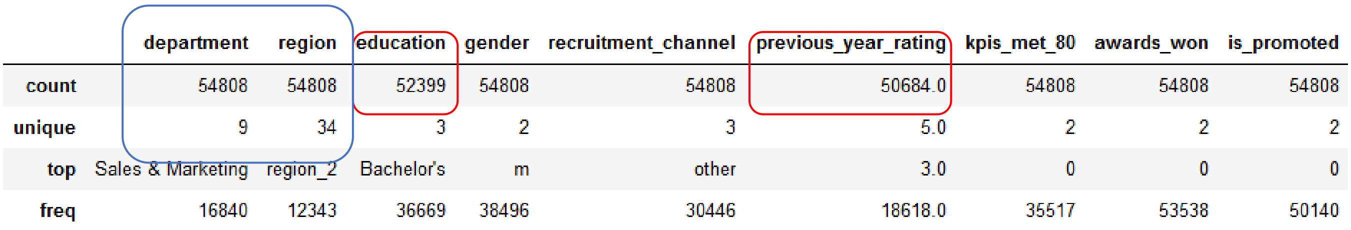
2.1 Exploratory Data Analysis

Univariate Analysis

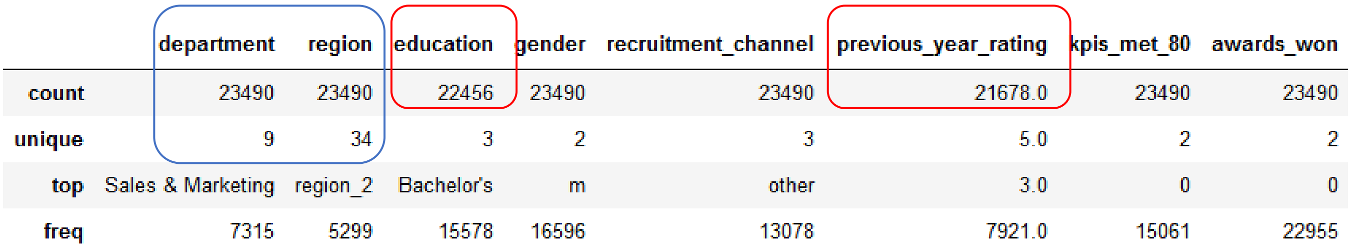
Total rows in training is 54,808 (23,490 in test).

Education and *previous_year_rating* showed missing values. *Education* had approximately 4.4% values, while *previous_year_rating* had 7.52% missing in train and 7.71% missing in.

Two categorical features have a high number of categories: *department* with 9 categories and *region* with 34 unique values.

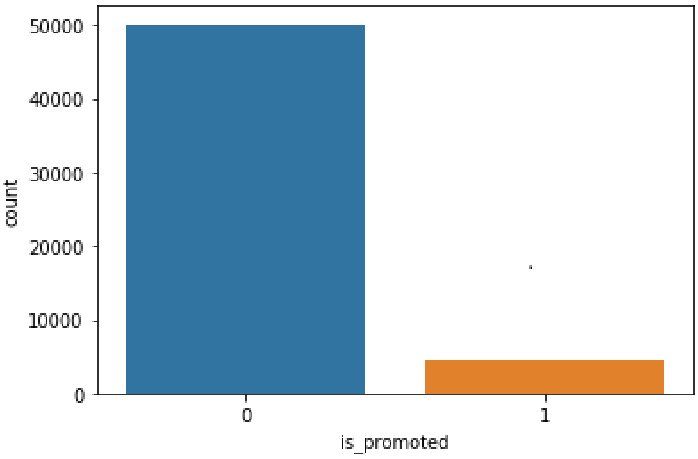


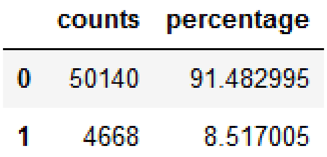
	department	region	education	gender	recruitment_channel	previous_year_rating	kpis_met_80	awards_won	is_promoted
count	54808	54808	52399	54808	54808	50684.0	54808	54808	54808
unique	9	34	3	2	3	5.0	2	2	2
top	Sales & Marketing	region_2	Bachelor's	m	other	3.0	0	0	0
freq	16840	12343	36669	38496	30446	18618.0	35517	53538	50140



	department	region	education	gender	recruitment_channel	previous_year_rating	kpis_met_80	awards_won
count	23490	23490	22456	23490	23490	21678.0	23490	23490
unique	9	34	3	2	3	5.0	2	2
top	Sales & Marketing	region_2	Bachelor's	m	other	3.0	0	0
freq	7315	5299	15578	16596	13078	7921.0	15061	22955

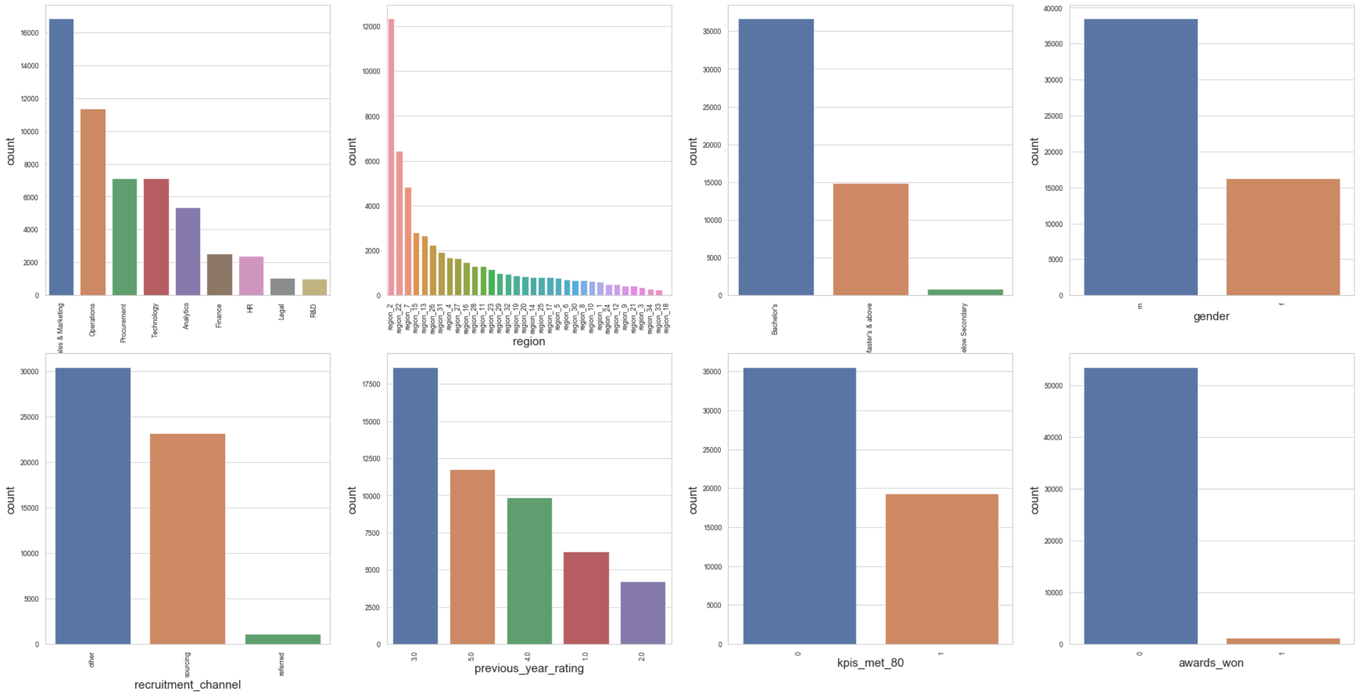
The target variable, *is_promoted,* showed that only 8.52% of employees received promotion, pointing to an unbalanced class problem.



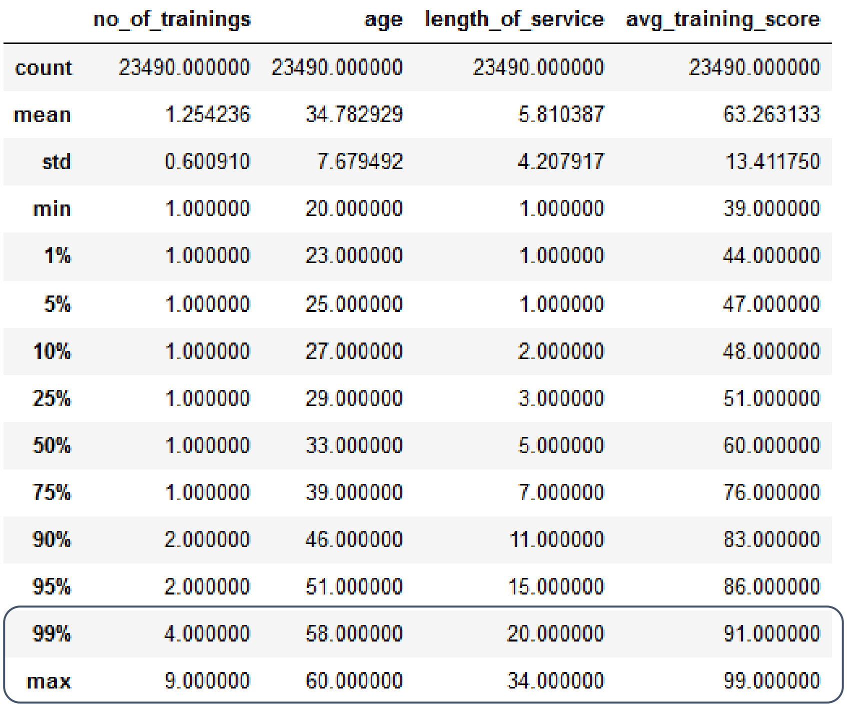


	counts	percentage
0	50140	91.482995
1	4668	8.517005

Three of 9 departments accounted for over 75% of employees, and interestingly Procurement (usually a support function) was third largest (13%). 5 of 34 regions accounted for over 50% of employees. Education consisted of mostly Bachelor's degrees. Low level of recruitment happens from referrals (2%). Male to female was skewed 70:30 in favour of males. Previous year rating had a median value of 3.0 (1 to 5 scale). Only 35% of employees had KPIs over 80%, which could be a good predictor of promotion. Only 2% of employees won awards.



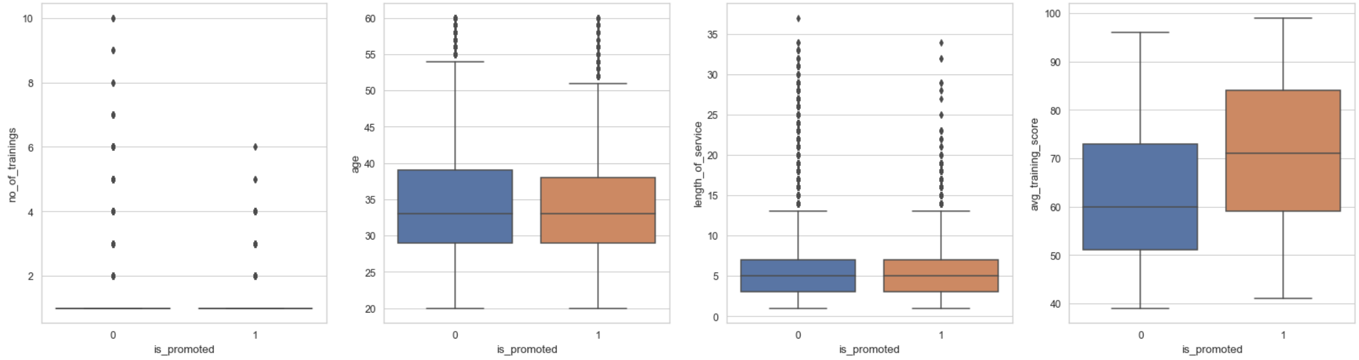
For numerical variables: median number of trainings was only 1, for 75% (or over) of employees i.e. investment in training is required; median length of service of employees is 5 years and more than 75% had been working for 3 years i.e. greater retention; outliers seen for all variables between 99th percentile and maximum.



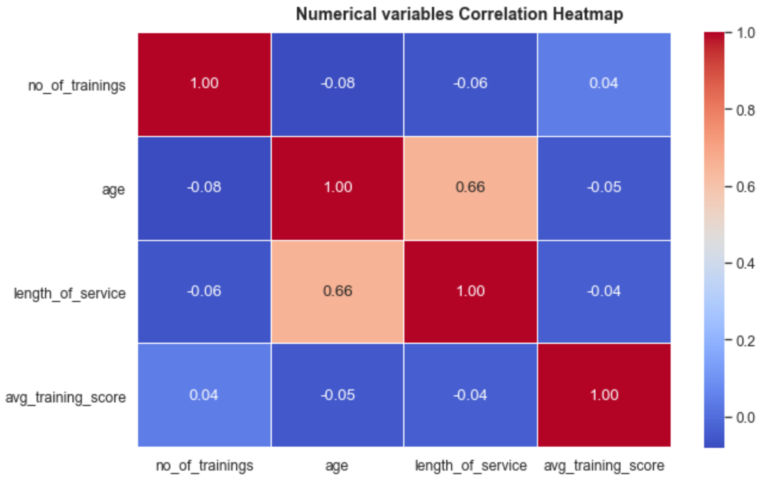
	no_of_trainings	age	length_of_service	avg_training_score
count	23490.000000	23490.000000	23490.000000	23490.000000
mean	1.254236	34.782929	5.810387	63.263133
std	0.600910	7.679492	4.207917	13.411750
min	1.000000	20.000000	1.000000	39.000000
1%	1.000000	23.000000	1.000000	44.000000
5%	1.000000	25.000000	1.000000	47.000000
10%	1.000000	27.000000	2.000000	48.000000
25%	1.000000	29.000000	3.000000	51.000000
50%	1.000000	33.000000	5.000000	60.000000
75%	1.000000	39.000000	7.000000	76.000000
90%	2.000000	46.000000	11.000000	83.000000
95%	2.000000	51.000000	15.000000	86.000000
99%	4.000000	58.000000	20.000000	91.000000
max	9.000000	60.000000	34.000000	99.000000

Bivariate Analysis

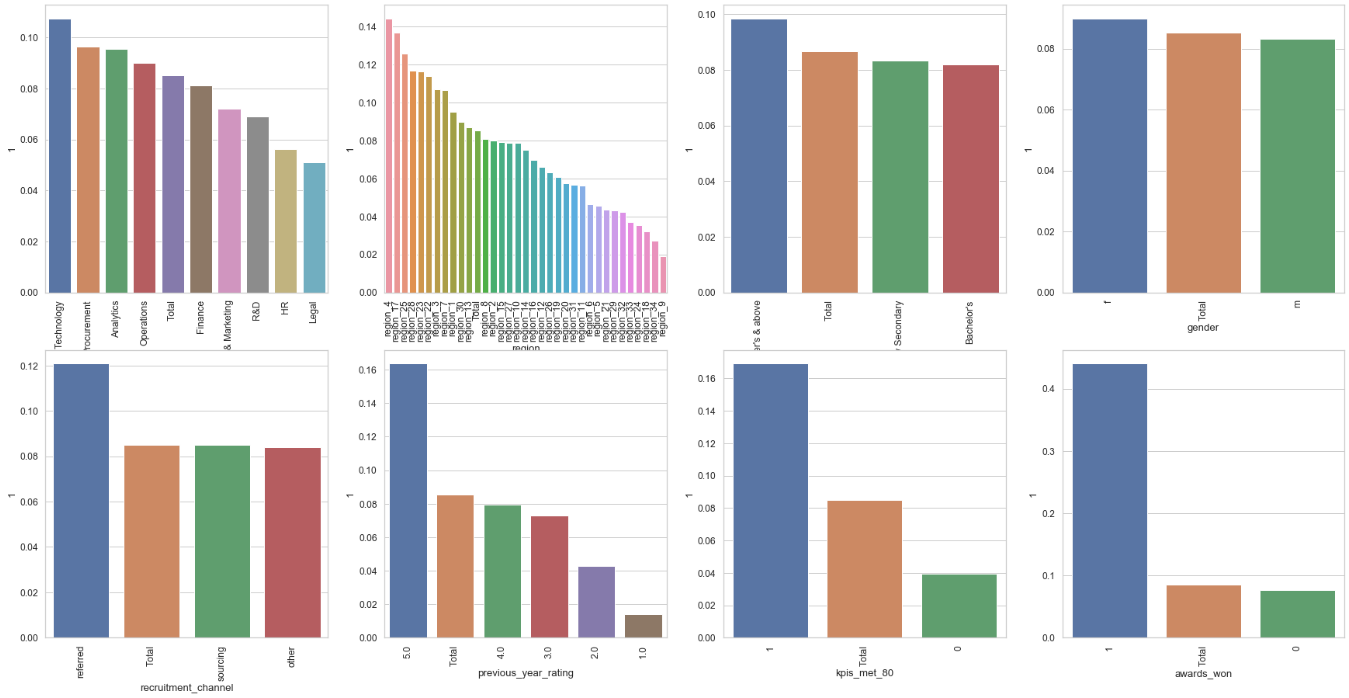
In the box and whisker plots below, for no_of_trainings, age and length_of_service, there is no difference between those who were promoted and those who weren't. However, when we look at the avg_training_score, there is a clear difference between those who were promoted (median above 70) and those who weren't (median around 60). This should be powerful for prediction.



Low correlation between numerical variables, except length of service and age, which are naturally correlated.



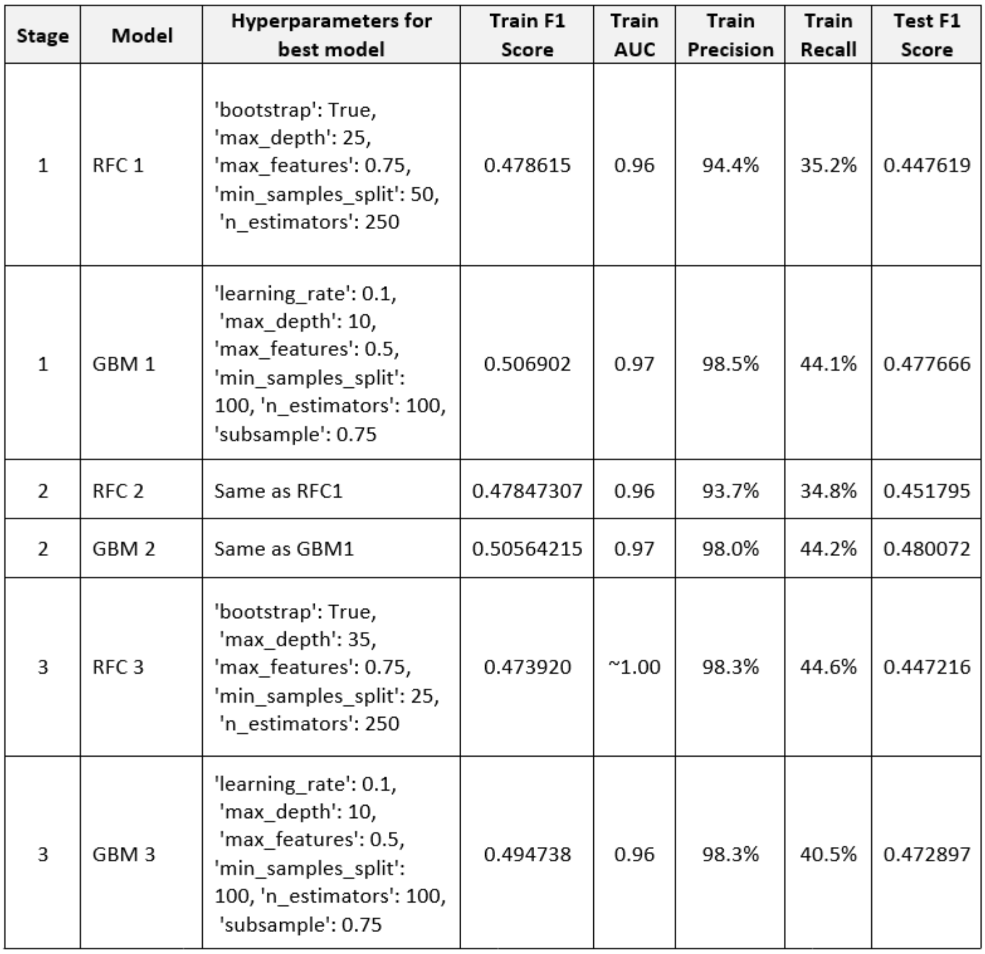
Over 16% of those with *previous_year_rating* 5.0 (possibly highest rating) get promoted while average promotion rate is only 8.5%. More than 16% of people who meet their KPIs get promoted. These two variables can be expected to be strongly predictive of promotion. Even though more than 50% of those who won awards got promoted, they represent a very small portion of the population. Some of the likely newer departments (Technology and Analytics) have a higher promotion rate, as well as Procurement. Certain regions also have a really high rate of promotion but these are not the regions with large number of employees.



2.2 Modeling

Models were built in 3 stages: Simple recoding and missing value treatment; Outlier treatment; and Stage 3 – Feature engineering for categorical variables. Comparison was made between Random Forest and Gradient Boosting Machine (workhorse techniques for Data Science competitions).

Model Performance Summary



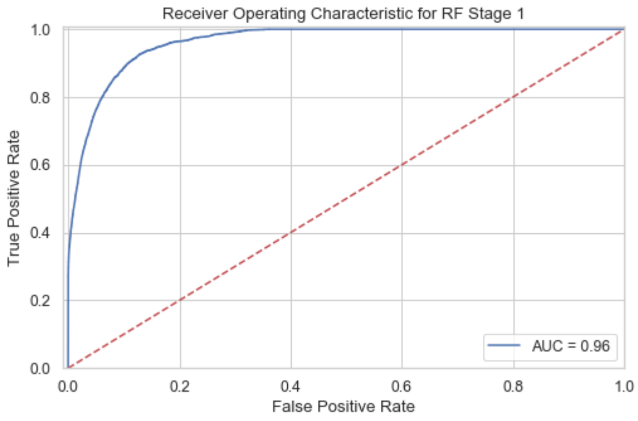
Stage	Model	Hyperparameters for best model	Train F1 Score	Train AUC	Train Precision	Train Recall	Test F1 Score
1	RFC 1	'bootstrap': True, 'max_depth': 25, 'max_features': 0.75, 'min_samples_split': 50, 'n_estimators': 250	0.478615	0.96	94.4%	35.2%	0.447619
1	GBM 1	'learning_rate': 0.1, 'max_depth': 10, 'max_features': 0.5, 'min_samples_split': 100, 'n_estimators': 100, 'subsample': 0.75	0.506902	0.97	98.5%	44.1%	0.477666
2	RFC 2	Same as RFC1	0.47847307	0.96	93.7%	34.8%	0.451795
2	GBM 2	Same as GBM1	0.50564215	0.97	98.0%	44.2%	0.480072
3	RFC 3	'bootstrap': True, 'max_depth': 35, 'max_features': 0.75, 'min_samples_split': 25, 'n_estimators': 250	0.473920	~1.00	98.3%	44.6%	0.447216
3	GBM 3	'learning_rate': 0.1, 'max_depth': 10, 'max_features': 0.5, 'min_samples_split': 100, 'n_estimators': 100, 'subsample': 0.75	0.494738	0.96	98.3%	40.5%	0.472897

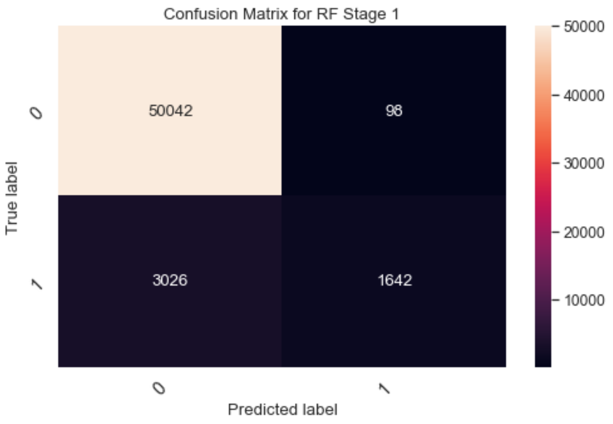
- **All the models are really high performing** as can be seen from the AUC (0.96 and over in all cases) and Precision values (greater than 93% in all cases). However, the Recall is moderate (between 35% and 40%). That is, the models are generally able to identify 2 out of 5 employees who will get promoted with a very high degree of accuracy.

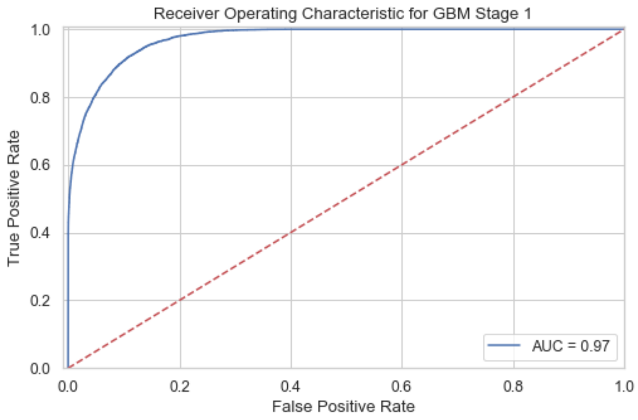
- **GBM model in Stage 2 is the best performing model** as it has the highest F1 Score in test. This model is used to explain factors correlated with promotion; GBM model in Stage 3 is then used for a detailed explanation of the factors.

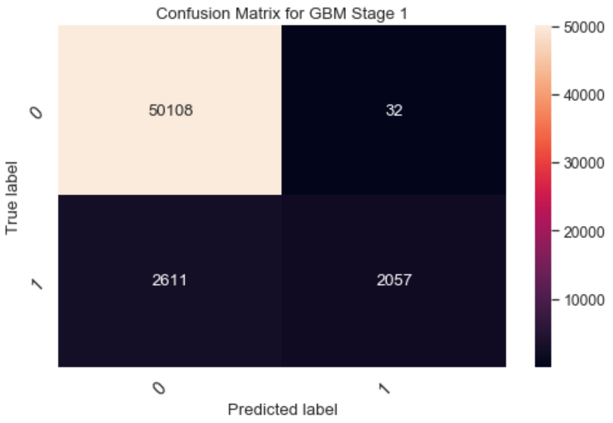
- **Outlier treatment** was fruitful and produced the best model in test (GBM Stage 2).

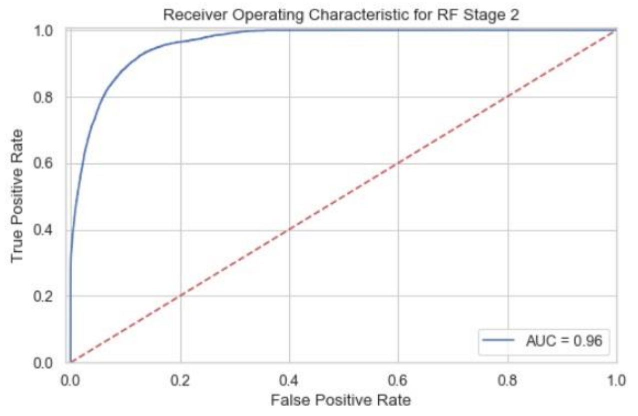
- **Feature engineering through transformation of categorical variables** using count encoding, target encoding and one hot encoding did not improve model performance. This could be because the implementation of the algorithms are able to take care of categorical variables internally, or that further hyperparameter tuning is required.

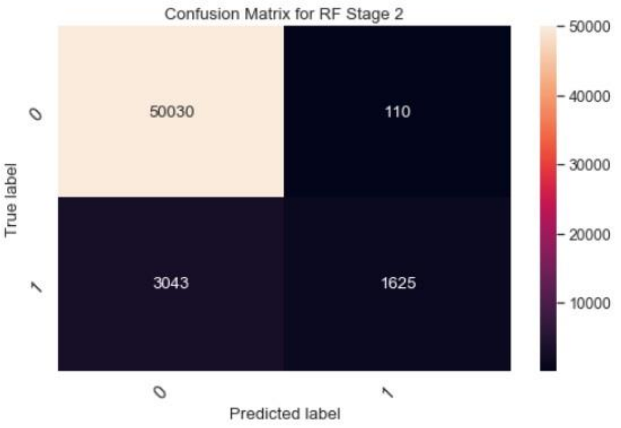

Receiver Operating Characteristic for RF Stage 1
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.96

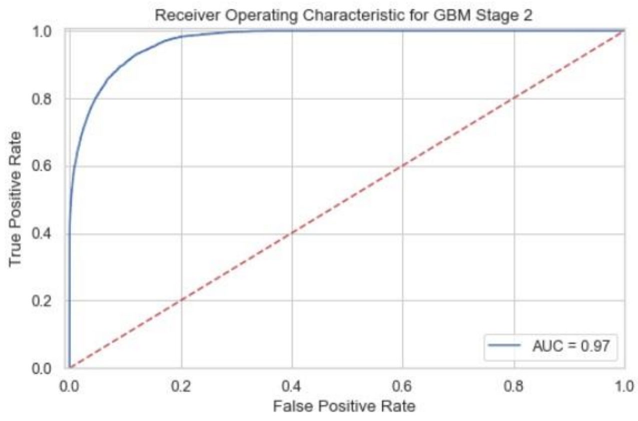

Confusion Matrix for RF Stage 1
True label
Predicted label
0
1
50042
98
3026
1642
50000
40000
30000
20000
10000

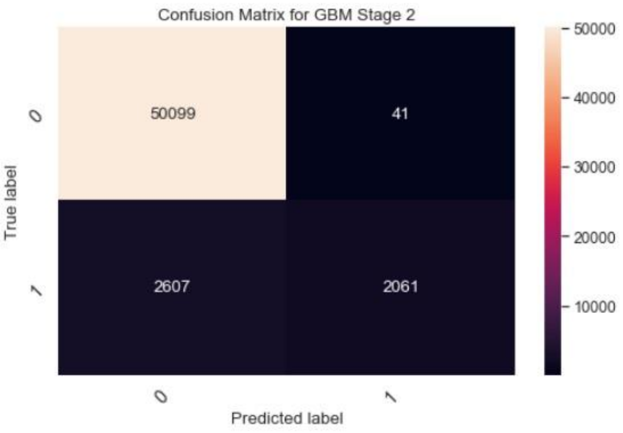

Receiver Operating Characteristic for GBM Stage 1
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.97

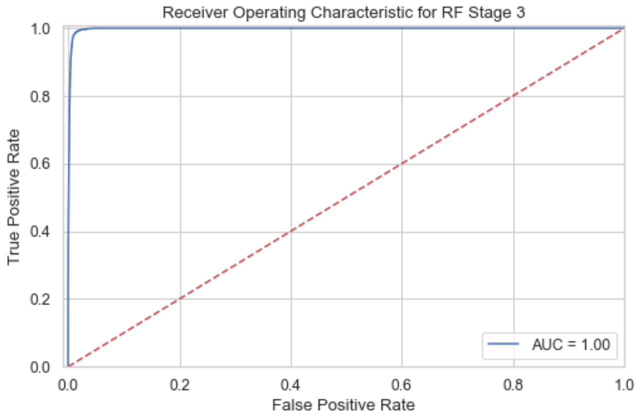

Confusion Matrix for GBM Stage 1
True label
Predicted label
0
1
50108
32
2611
2057
50000
40000
30000
20000
10000

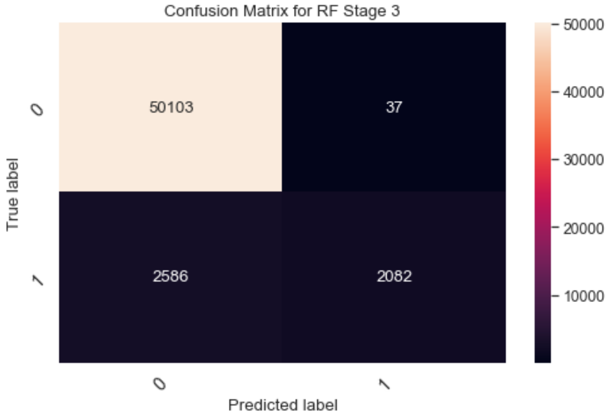

Receiver Operating Characteristic for RF Stage 2
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.96


Confusion Matrix for RF Stage 2
True label
Predicted label
0
1
50030
110
3043
1625
50000
40000
30000
20000
10000

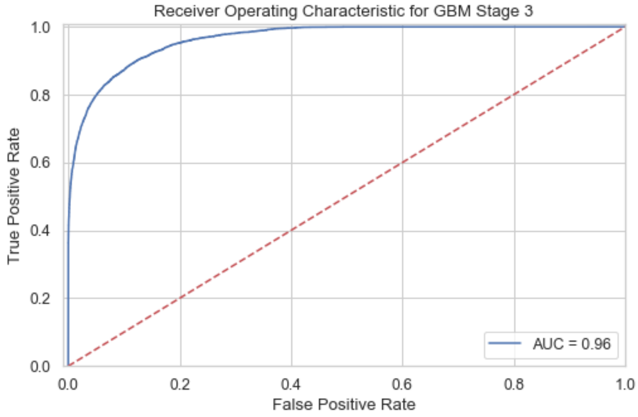

Receiver Operating Characteristic for GBM Stage 2
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.97

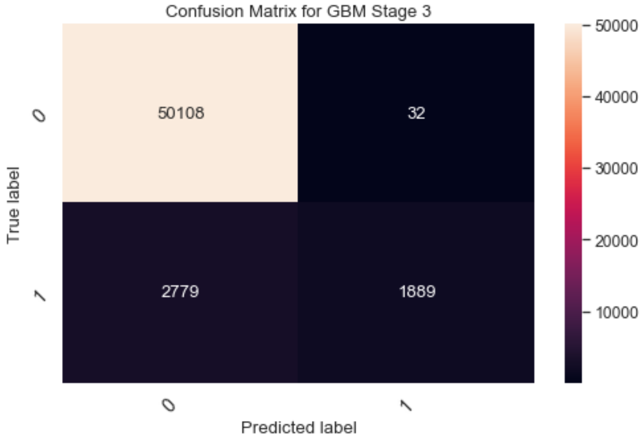

Confusion Matrix for GBM Stage 2
True label
Predicted label
0
1
50099
41
2607
2061
50000
40000
30000
20000
10000





Random Forest model built in Stage 3 had an AUC of very close to 1. However, this model seems to be overfitted judging by its performance on the test data.

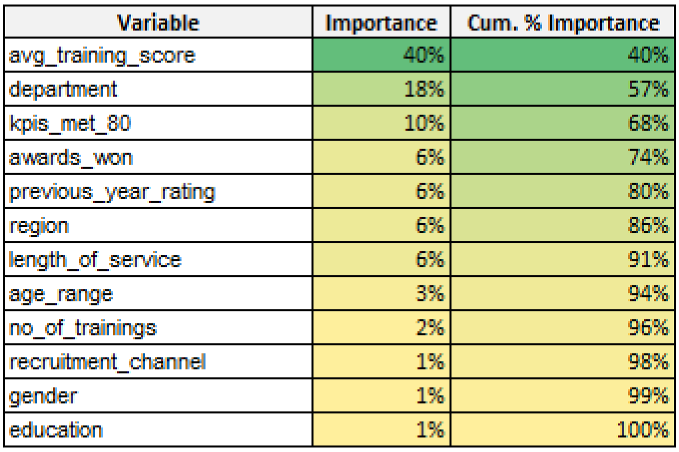




GBM models were consistently better than the Random Forest models (F1 scores 0.49-0.5 vs 0.47-0.48). The RF models were easier to tune and tuning the ‘min_samples_split’ hyperparameter (minimum number of samples required to split an internal node) proved vital in the latter stages, after tuning the ‘max_depth’ (maximum depth of the tree) hyperparameter initially.

Factors connected with promotion: The Variable Importance report for the GBM models in Stages 1 and 2 are used to understand if the ideas from the EDA stage hold up. Note: Variable Importance is not the same as standardized coefficient estimates, like in regression. Machine Learning models thus suffer from an interpretability problem, but the best we can do is to use the Variable Importance reports; and (2) Variable Importance indicates correlation and not causation.

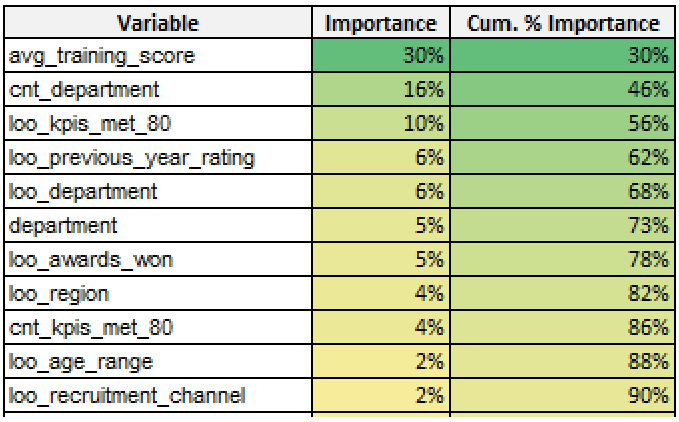
Variable Importance from GBM Stage 2 (Best model)



Variable	Importance	Cum. % Importance
avg_training_score	40%	40%
department	18%	57%
kpis_met_80	10%	68%
awards_won	6%	74%
previous_year_rating	6%	80%
region	6%	86%
length_of_service	6%	91%
age_range	3%	94%
no_of_trainings	2%	96%
recruitment_channel	1%	98%
gender	1%	99%
education	1%	100%

- Average Training Score is the most important factor for promotion.
- Presence of Department as the second most important factor indicates a potential department bias.
- 4 of the top 5 factors are performance based.
- Gender and Age Range rank low in the list and are not highly predictive of promotion (no bias).

Variable Importance from GBM Stage 3 (More detailed breakup of variables)



Variable	Importance	Cum. % Importance
avg_training_score	30%	30%
cnt_department	16%	46%
loo_kpis_met_80	10%	56%
loo_previous_year_rating	6%	62%
loo_department	6%	68%
department	5%	73%
loo_awards_won	5%	78%
loo_region	4%	82%
cnt_kpis_met_80	4%	86%
loo_age_range	2%	88%
loo_recruitment_channel	2%	90%

- 3 of the top 6 variables are related to Department.
- Count of employees in a department (cnt_department) is the second most important factor, indicating that larger departments may receive a higher rate of promotions.

2.3 Final Reflections

Using this data driven approach shows that such methods could be beneficial to organizations, saving time, and thus cost. Supplementing this approach with the domain knowledge of HR professionals could produce further success, such that the approach could be tailored to the individual contexts of organizations. As adoption of similar techniques is low in industry, [1][2] there would be a requirement for human driven analysis as a safety layer to validate results produced by predictive modelling. However, it can be said that the techniques used here for predicting promotions, show promise.

Reference List

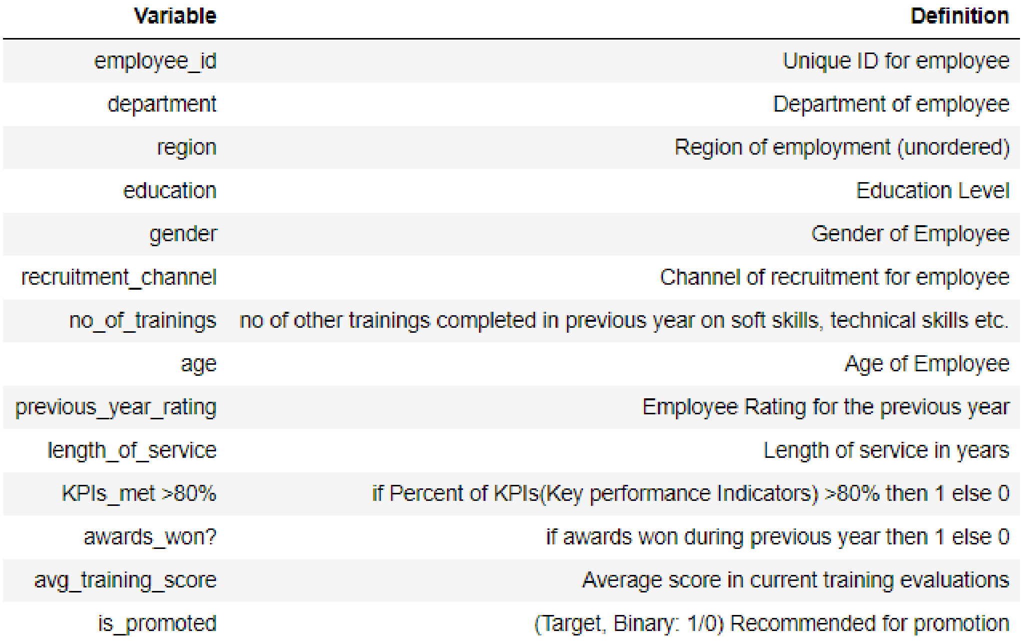
[1] Janet H. Marler & John W. Boudreau, (2017) “An evidence-based review of HR Analytics,” *The International Journal of Human Resource Management*, vol. 28, issue. 1, pp. 3-26, Nov, 2009. Accessed on: Aug, 1, 2019. [Online]. Available: https://www.tandfonline.com/doi/pdf/10.1080/09585192.2016.1244699 (DOI: 10.1080/09585192.2016.1244699)

[2] Witte, L., “We have HR analytics! So what?: an exploratory study into the impact of HR analytics on strategic HRM” Master's thesis. Behavioural, Management and Social Sciences. University of Twente. Enschede, Netherlands, 2016. Accessed on Aug, 2, 2019. [Online]. Available: https://essay.utwente.nl/70301/1/Witte_MA_BMS.pdf

[3] Analytics Vidhya. (2018). *Practice Problem: HR Analytics* [Online]. Available: https://datahack.analyticsvidhya.com/contest/wns-analytics-hackathon-2018-1/

APPENDIX A

Data Description



Variable	Definition
employee_id	Unique ID for employee
department	Department of employee
region	Region of employment (unordered)
education	Education Level
gender	Gender of Employee
recruitment_channel	Channel of recruitment for employee
no_of_trainings	no of other trainings completed in previous year on soft skills, technical skills etc.
age	Age of Employee
previous_year_rating	Employee Rating for the previous year
length_of_service	Length of service in years
KPIs_met >80%	if Percent of KPIs(Key performance Indicators) >80% then 1 else 0
awards_won?	if awards won during previous year then 1 else 0
avg_training_score	Average score in current training evaluations
is_promoted	(Target, Binary: 1/0) Recommended for promotion