HR Analytics

Prepared by Younjoo Lee

This report is sectioned as follows:

1) Business Case: Predictive Modeling for Employee Promotion
2) Dataset Overview & Exploratory Data Analysis
3) Preparing Model
4) Models
    a. Ordinary Least Squares Regression
    b. Logistic Regression
    c. Random Forest
    d. ADA Boost on RF
    e. Decision Tree
    f. ADA Boost on DC Tree
5) Conclusion

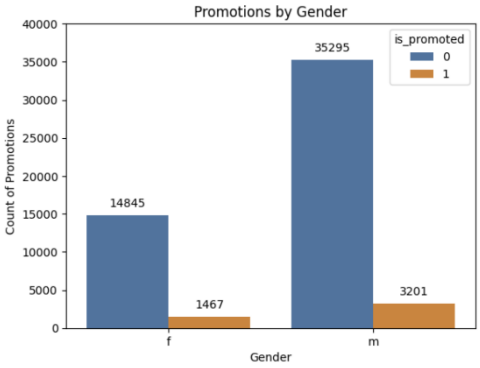
Business Case: Predictive Modeling for Employee Promotion

Bauer is a multinational German corporation with over 100,000 employees. Bauer seeks to identify the right people for leading in managerial positions. Currently, the process is manual and based off historical performances and recommendations from upper-level leadership. Thus, to ensure a fair process for employees and help leadership make decisions by expediting the process for identifying the right people, my goal is to create a transparent, predictive model that generates recommended employees for promotion. To evaluate the

performance of the model which uses a binary classification (1 = recommend for promotion, 0 = do not recommend), I will be using the F1 score[1] and a confusion matrix for a sensitivity and specificity analysis.

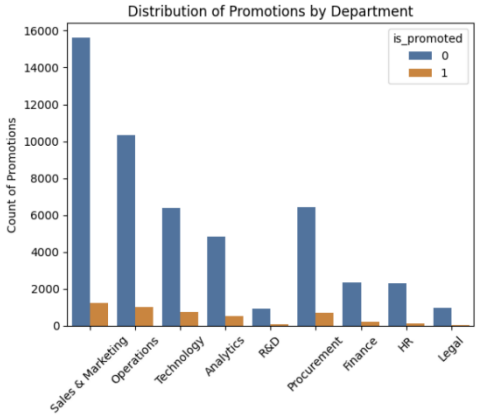
Dataset Overview & Exploratory Data Analysis

There are 54,807 employee records in the training dataset. There are 13 features or variables which include, employee ID, department, region, education level, gender, age, recruitment channel, number of trainings, previous year rating, length of service, KPIs met < 80%, awards won, average training score, which are used to predict whether an employee is recommended for promotion. The overall promotion rate is approximately 8.5%. Furthermore, I explore the promotions by gender, department, and education. I am interested in the demographic variable gender because I want to create awareness of any biases in the training data.

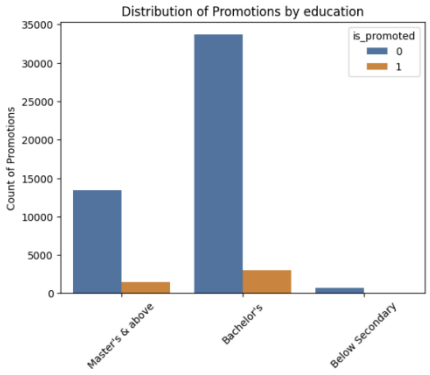


While there are more males than females in the dataset, the proportion of promotions are relatively balanced. The female promotion rate is approximately 9% while the male promotion rate is approximately 8%.

The promotion rate by department ranges from approximately 5% to 11% with the Technology department leading with the highest promotion rate.



department	promotion_rate
Analytics	0.095665
Finance	0.081230
HR	0.056245
Legal	0.051011
Operations	0.090148
Procurement	0.096386
R&D	0.069069
Sales & Marketing	0.072031
Technology	0.107593



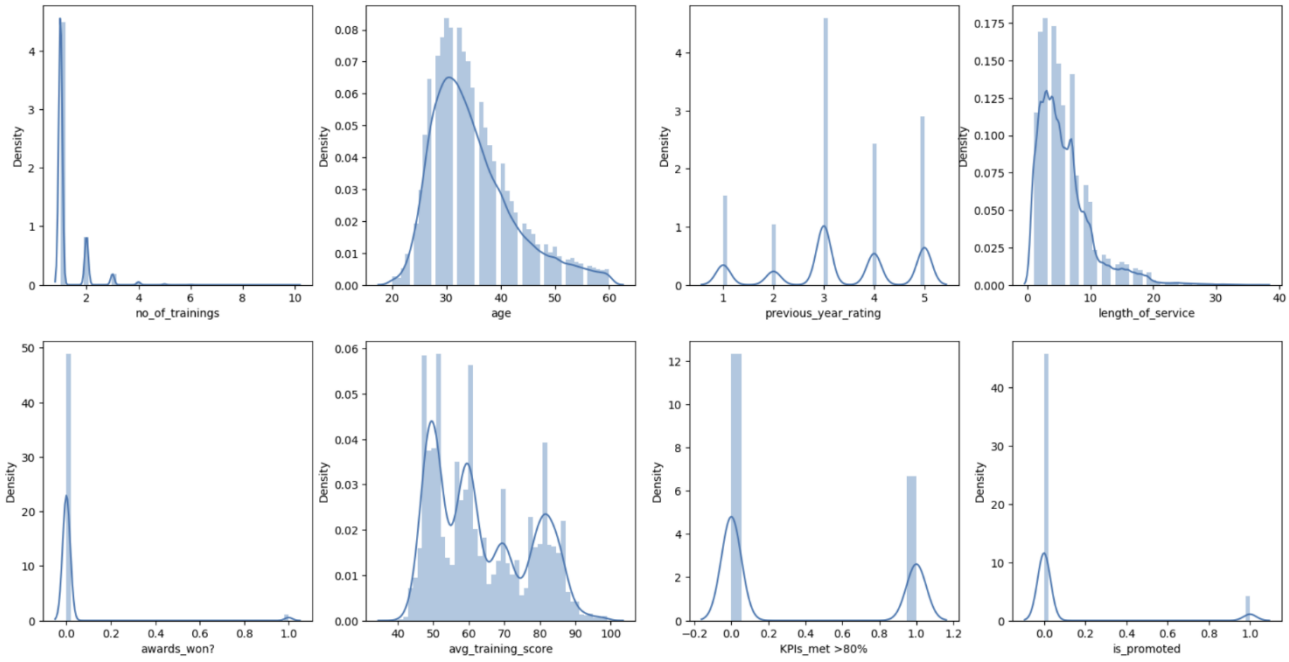
Interestingly, the promotions by education are relatively balanced as well, around the average overall promotion rate of 8%. Although, there is a higher rate of promotion for Master's & above.

	education	promotion_rate
0	Bachelor's	0.082031
1	Below Secondary	0.083230
2	Master's & above	0.098559

[1] F1 Score is a measure of model accuracy, and it is based on the precision and recall scores. https://www.v7labs.com/blog/f1-score-guide#what-is-f1-score

Preparing Model

Let's take a look at the distribution of features.



Based off the skewed distribution for age, length of service, average training score, and KPIs met above 80%, I took the log to help normalize these distributions for model building.

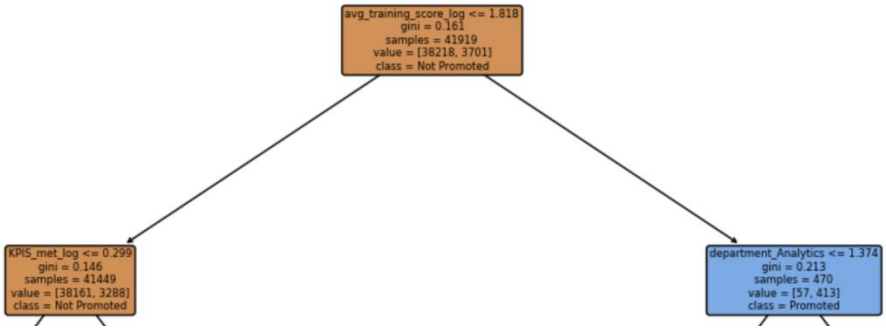
There were missing values for education (2,409) and previous year ratings (4,124). I conducted feature engineering by filling in the missing values for previous year ratings with the average rating (3.3). Next, instead of replacing values for education, I dropped the records without education as this was a relatively small set of data to sacrifice and I did not want to assume the education of the missing values. Finally, I ran dummy variables for the features with categorical values such as department, gender, education, region, and recruitment channel. In total, I prepared my model with 58 x-variables or features and 52,399 employee records which makes this a relatively medium sample size and a relatively medium number of features. I split the training and test set by 80% and 20%.

Models

The simplest base model I chose was the Ordinary Least Squares Regression model. In addition, because this is a binary classification problem, I chose a logistic regression, decision tree, and random forest model with minimal tuning to evaluate as base models. Then, I tuned parameters and ran an ADA Boost with a Random Forest model base and an ADA Boost with a Decision Tree base.

Evaluation Metrics[2]	OLS[3]	Logistic Regression[4]	Random Forest (RF)[5]	ADA Boost on RF[6]	Decision (DC) Tree[7]	ADA Boost on DC Tree[8]
F1 Score	0.17	0.40	0.42	0.47	0.48	0.49
Sensitivity Rate (TPR)	9%	26%	29%	33%	35%	36%
Specificity Rate (TNR)	99%	99%	99%	99%	99%	99%

Based on the specificity rate or the true negative rate, all of the models performed well in terms of predicting employees that do not receive the promotion. The OLS model performed poorly at predicting the right employees for the promotion but was accurate at predicting the employees that do not get promoted. The logistic regression, which is a common base model for binary classification problems for smaller sample sizes with many features, also performed poorly in terms of predicting the promotion of employees. The random forest model performed slightly better than the logistic regression by promoting deeper learning and interaction between the variables. The ADA Boost on RF, Decision Tree, and ADA Boost on DC Tree performed relatively similarly with F1 scores from 0.47-0.49. In terms of the tradeoff between "simplicity," transparency, and a relatively higher performance in terms of accuracy, I chose the decision tree. With the decision tree, the importance of the features



[2] I checked the evaluation metrics by testing several random seeds
[3] After I generated predicted y-values, I converted them into binary outputs by setting the threshold > 0.50 to determine whether the employee should be recommended for a promotion
[4] Solver was liblinear
[5] No tuning, default random forest model
[6] Base model random forest had n estimators = 50, max depth = 10, ADA RF model had n estimators = 25
[7] I used GridSearchCV and tested various depths and sample splits and found best parameters to be max depth = 15, sample split = 20
[8] I used GridSearchCV and tested various n estimators and learning rates and found the best parameters for DC tree to be max depth = 5, ADA DC tree had learning rate = 0.5, n estimators 50

and what goes into each “decision” is more easily interpretable.

For instance, the first decision[9] was based off the average training score which shows that this feature was important in predicting promotions.

Conclusion

While the ADA Boost on the base Decision Tree model performed the best in terms of F1 score, none of the models met a high accuracy rate (over 90% F1 score) and accurately predicted true positive rates or employees who receive the promotion. For future research, I would recommend tuning the model further or developing more advanced modeling such as neural networks; however, it is important to note that the number of features (13 original features) along with the sample size (54,807) may not suffice for a neural network and can lead to overfitting which will not allow the model to perform well outside the dataset.

An important aspect to note in terms of assessing algorithmic bias, especially as it pertains to making decisions regarding humans and their career growth, is that when I analyzed the promotion rate by gender on the testing dataset was that they were relatively even. The rate of promotion for women was about 10% while the rate of promotion for men was about 9%.

[9] This figure only depicts the top portion of the 15 max depth tree