Analytics Vidhya

DATAHACK.ANALYTICSVINDYA.COM

HR Analytics

Prepared by Younjoo Lee

This report is sectioned as follows:

Business Case: Predictive Modeling for Employee Promotion

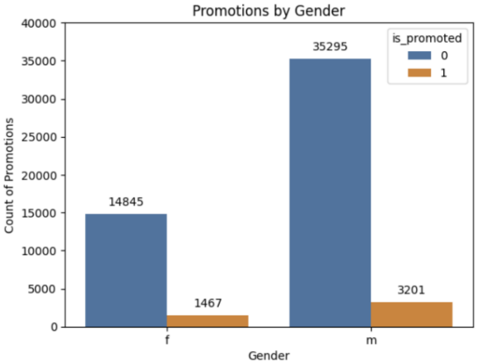
Bauer is a multinational German corporation with over 100,000 employees. Bauer seeks to identify the right people for leading in managerial positions. Currently, the process is manual and based off historical performances and recommendations from upper-level management. To ensure a fair process for employees and help leadership make decisions by expediting the promotion process, my goal is to create a transparent, predictive model that provides recommendations for employee promotion. To evaluate the performance of the model, which uses a binary classification (1 = promoted, 0 = not promoted), I will use the F1 score[1] and

[1] F1 Score is a measure of model accuracy, and it is based on the precision and recall scores.

confusion matrix for sensitivity and specificity analysis. In addition, I evaluate the model, which includes the training data, for gender bias.

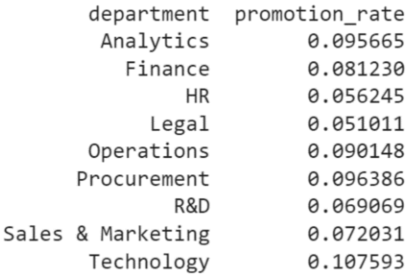
Dataset Overview & Exploratory Data Analysis

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. The overall promotion rate is approximately 8.5%, indicating an imbalanced dataset in terms of the majority class (over 90%) being unpromoted employees. 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.



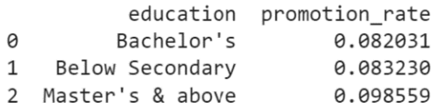
While there are more males (m) than females (f) in the dataset, the proportion of promotions are relatively balanced. The female promotion rate is approximately 9% while the male promotion rate is approximately 8%.

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
```

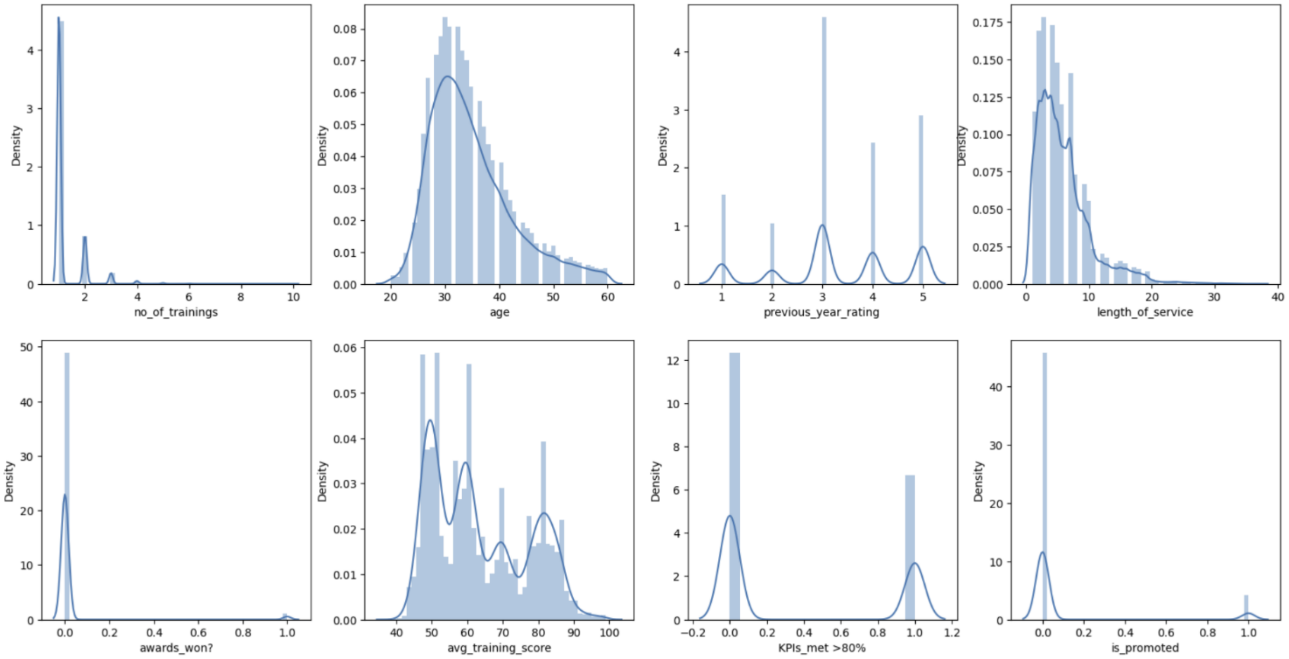
The promotions by education level are relatively balanced as well with the average overall promotion rate of 8%. Although, there is a higher rate of promotion for Master’s & above.



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

Data Wrangling

Let’s 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, I dropped the null records for education because this was a relatively small set of data to sacrifice, and I did not want to assume the education level of the employees. Finally, I ran dummy variables for the features with categorical values: department, gender, education, region, and recruitment channel. In total, I prepared my model with 58 features and 52,399 employee records. I split the training and test set by 80|20.

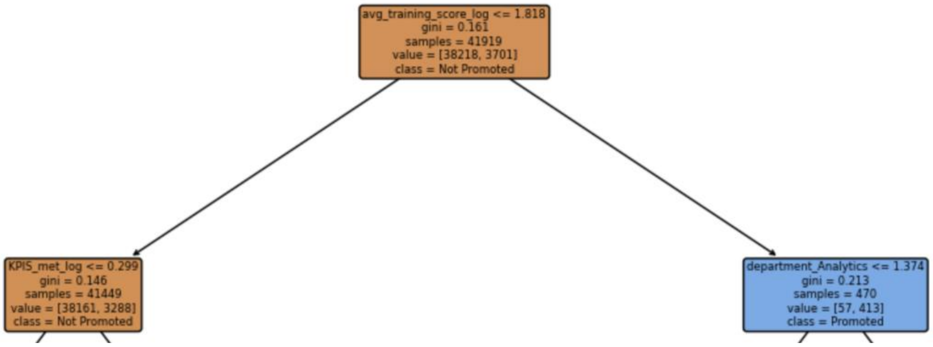
Models

The simplest base model I chose was the Ordinary Least Squares Regression model. I also chose a logistic regression, decision tree, and random forest model with minimal tuning to evaluate as base models. These are common models for binary classification problems. 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 the models performed well in terms of predicting employees that do not receive the promotion. The simplest OLS model had the weakest performance in terms of predicting true positives for employee promotion. The logistic regression also performed poorly in terms of predicting the promotion of employees at 26% for the sensitivity rate. The random forest model performed slightly better than the logistic regression by incorporating 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 (sensitivity and specificity rate), I chose the decision tree. With the decision tree, the importance of the features and what goes into each "decision" for promotion is easier to interpret.

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



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

Conclusion

I recommend using the decision tree model because of the relatively higher accuracy performance and transparency. For future research, I would recommend introducing resampling techniques because the data was imbalanced, meaning the promoted class was heavily underrepresented. One method is to under sample the majority class (no employee promotion) through randomly removing samples from the majority class to balance the dataset. Another method could be using 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.

In terms of assessing the decision tree model for gender bias, the average rate of promotion was 10% for women while the average promotion rate for men was 9%. Compared to the dataset which had a promotion rate of women at 9% with a promotion rate of men at 8%, there is the same 1 percentage point difference between the genders.