Final Report of BIOS 611 - Employee Promotion Evaluation Analysis

Qiyao Qin

2022-11-19

Contents

1 Introduction

1.1 Research Background

Nowadays, employee promotion is of great significance for both company and employees, since this will directly impact the efficiency and performance of the organization. One of the greatest challenges in employee promotion is to identify the right person for promotion and prepare him/her in time.

In the promotion process, the company should first identify a set of employees based on recommendations or past performance. Selected employees go through the separate training and evaluation program. At the end of the program, based on the evaluations of various factors, the final promotions can be announced, which means that this may lead to delay in transition to new roles. Hence, company does need help in identifying the eligible candidates at a particular checkpoint so that they can expedite the entire promotion cycle.

Besides, for employees, promotions are quite important to their career development. However, most of employees know little about what they can do at a specific point in time to get a promotion chance. Thus, it's of great necessity to analyze the main factors will impact their career promotion and what the most important factors are for them to get the promotion.

1.2 Research Aim

On the one hand, I plan to predict whether a potential candidate at checkpoint in the test set will be promoted or not after the evaluation process to help company recognize the brilliant employees quickly.

On the other hand, I intend to develop a system for employees to see the factors impacting their promotions at any time to help them know what they should do to improve their career development.

2 Data Description

2.1 Variables Description

The data set is about employee promotion from Kaggle (https://www.kaggle.com/datasets/arashnic/hr-ana). It consists of 13 variables including employee_id, department, region, education, gender, recruitment_channel, no_of_trainings, age, previous_year_rating, length_of_service, awards_won, avg_training_score and is_promoted. And is_promoted is the target variable. The detailed features' descriptions are as follows:

- employee_id: Unique ID for employee
- department: Department of employee
- region: Region of employment
- education: Education Level
- gender: Gender of Employee
- recruitment_channel: Channel of recruitment for employee
- no_of_trainings: Number of other trainings completed in previous year on skills etc.
- age: Age of Employee
- previous_year_rating: Employee rating for the previous year
- length_of_service: Length of service in years
- awards_won: If awards won during previous year then 1 else 0
- avg_training_score: Average score in current training evaluations
- is_promoted: If recommended for promotion then 1 else 0

2.2 Data Visualization

Firstly, we can analyze the probable impacting factors by visualization. Limited by the required number of words, only parts of plots revealing relationships among variables in this data set are shown in this report.

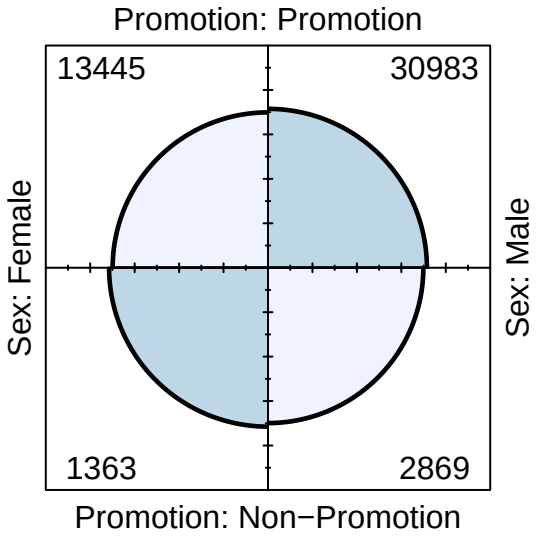

Figure 1: Promotion and Sex

From the four-fold plot between sex and whether an employee is promoted (Figure 1), we can see that the proportion of female in promoted employees is smaller than that of male in promoted employees. And we can further conclude that male employees may be more likely to have promotion chances than female employees.

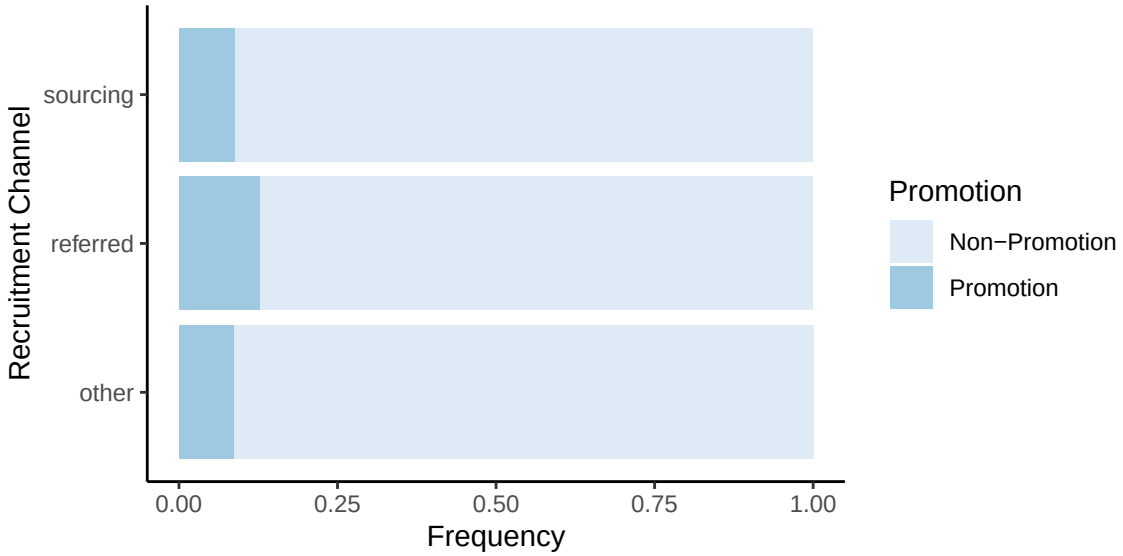

Figure 2: Promotion and Recruitment Channel

Meanwhile, we can find that the proportion of promoted employees who are recruited by others' referring is much larger than that of promoted employees from other two recruitment channels in Figure 2, which indicates that employees recruited by others' recommendation have a greater chance to be promoted.

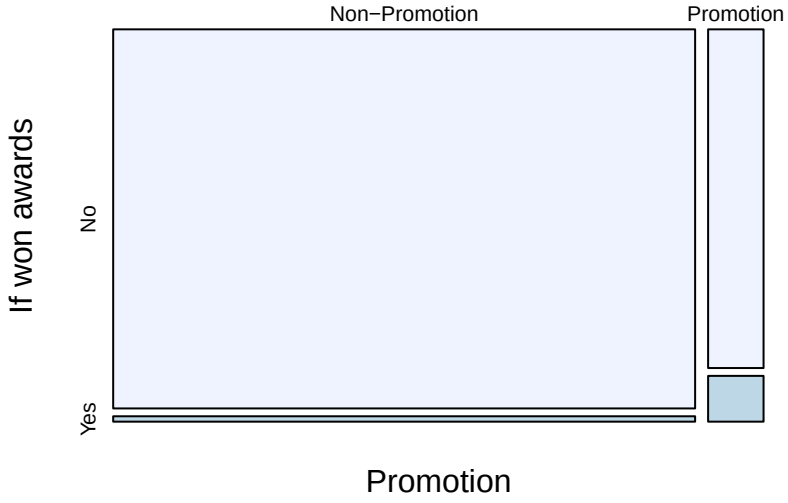

Figure 3: Promotion and Whether won awards

Besides, mosaic plot above shows that employees with awards before are more likely to have promotion chances. And following box plot reveals that the more average training score one employee obtain, the more probability of promotion the employee will have.

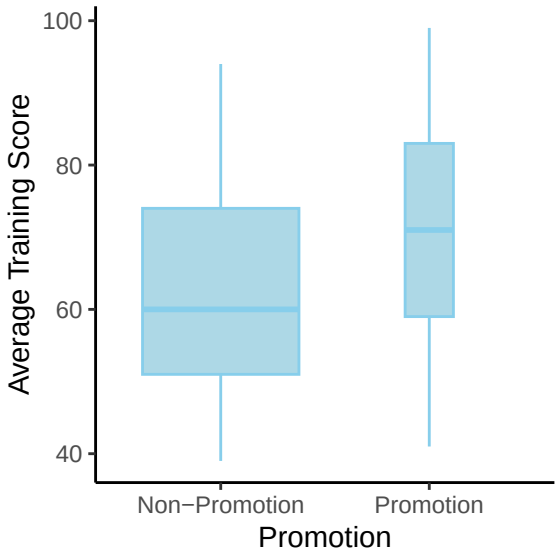

Figure 4: Promotion and Average Training Score

3 Statistical Model

3.1 Random Forest Model

After analyzing data roughly by visualization, statistical model is needed to further complete the research aim. Random forest model, consisting of a large number of individual decision trees that operate as an ensemble, is used in this report to predict employee promotion. Each individual tree in the random forest model spits out a class prediction and the class with the most votes becomes our model's prediction.

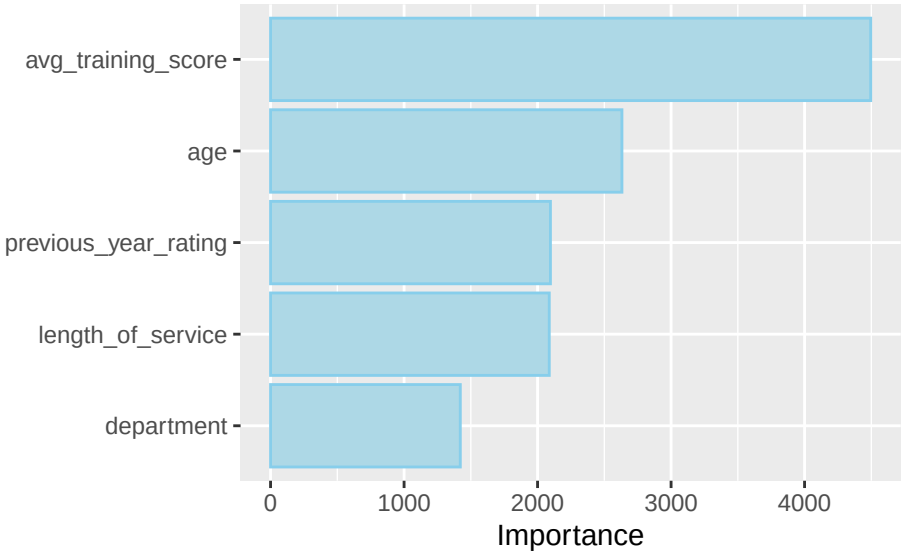

Figure 5: Importance of Top 5 Variables

Firstly, I split data into training and testing sets. Since the data is imbalanced, I used a combination of over-sampling minority examples and under-sampling majority examples with the re-sampling from the rare class probability of 0.5 for training set. After fitting the model by R, we can get Figure 5 showing the most

important variables for this model and following accuracy-related statistics based on testing data. Lastly, we can plot ROC curve based on testing data with the AUC of 0.72, which also illustrates that this random forest model works well.

Table 1: Accuracy-related Statistics

Statistics	Value
Accuracy	0.91
Sensitivity	0.69
Specificity	0.93
Balanced Accuracy	0.81

3.2 LIME Algorithm

LIME stands for Local Interpretable Model-agnostic Explanations. It is a method for explaining predictions of Machine Learning models, developed by Marco Ribeiro in 2016.

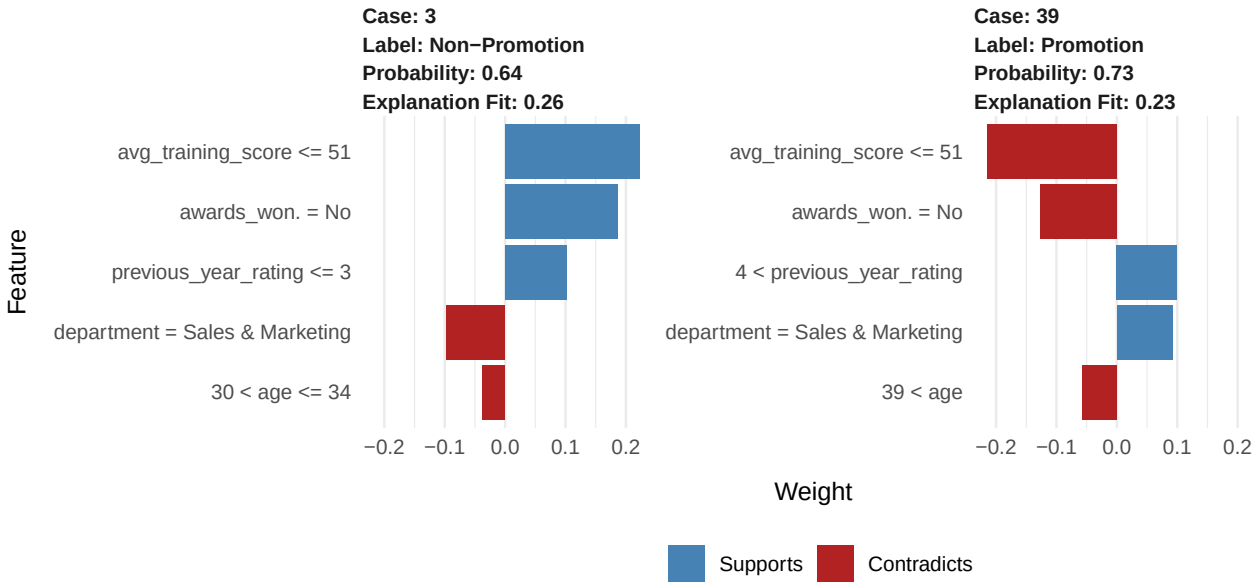

Figure 6: LIME Plots

After choosing the machine learning model (random forest model) and a reference point to be explained, LIME generates points all over the $\mathbb{R}^p$ space and predicts the Y coordinate of the sampled points using the random forest model. Then, it will assign weights based on the closeness to the chosen point by Gaussian kernel and train linear ridge regression on the generated weighted data set. Then, the coefficients in this regression model are regarded as LIME explanation, which generates LIME plots (Figure 6) for each new sample. LIME ensures that we can explain the importance of variables to the promotion result for each employee individually.

4 Employee Promotion Evaluation System

Based on Section 3, I set up an online employee promotion evaluation system by shiny. One can input information of each employee, and the system will output relative prediction results as a LIME plot which can perfectly accomplish research aim in Section 1 without requiring users to understand statistics or programming.

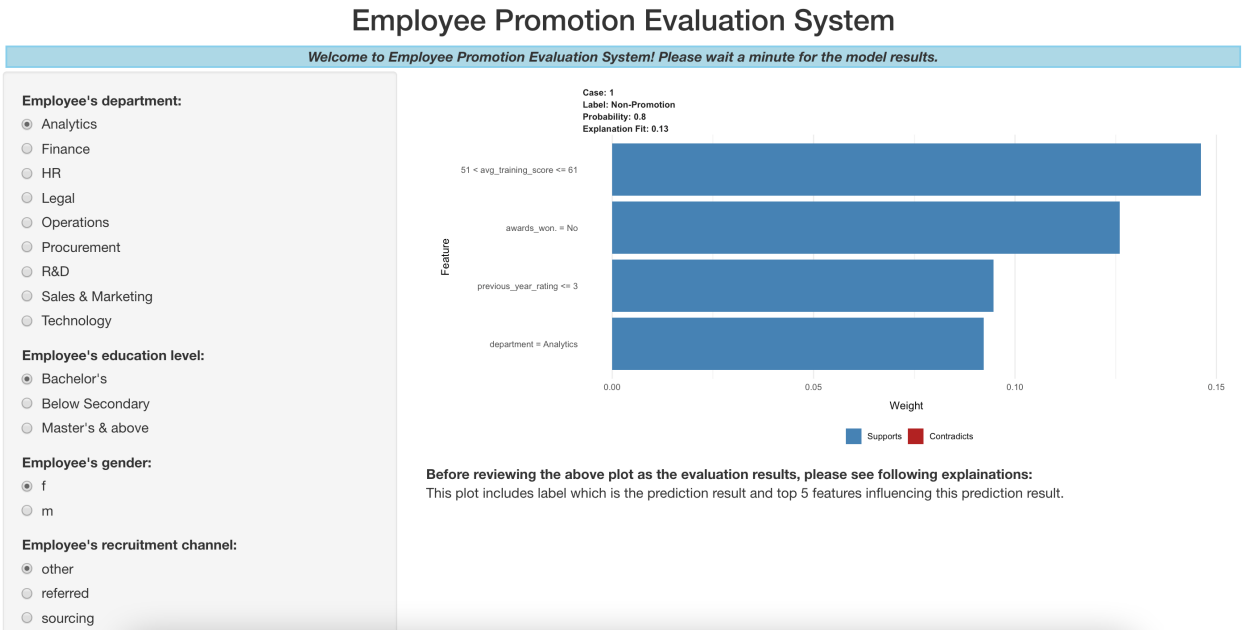

Figure 7: Online Employee Promotion Evaluation System

5 Reference

[1] Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1-17. doi: 10.18637/jss.v077.i01.

[2] Sandri, M. & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. J Comput Graph Stat, 17:611-628. doi: 10.1198/106186008X344522.

[3] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " Why should I trust you?" Explaining the predictions of any classifier. SIGKDD

[4] Visani, G., Bagli, E., Chesani, F., 2020. OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms. AIMLAI Workshop @ CIKM