# A Classification Model for Employee Promotion Prediction

DATA 1030 Midterm Project Report - Ziyin Li

## 1 Introduction

Companies nowadays can utilize machine learning tools not only to improve external processes like number of sales or customer satisfaction but also to optimize an internal process like the promotion of employees [1]. Promotion is the focus of human resource management research and it will be much helpful for the HR department to use a model to get prediction of whether an employee should be promoted or not by simply feeding the previous records of the employee into the model.

The objective of this project is to build such a model to predict if an employee is eligible for promotion or not based on the stored data of the promotion cycle in the previous year in order to save the time and effort of the HR team on making decisions.

The dataset used for this project called Employees Evaluation for Promotion came from the Kaggle datasets. The first column contains the unique employee ids and the last column called is_promoted is a categorical variable indicating whether an employee got promoted or not last year. Hence, this project is to solve a classification problem with is_promoted as the target variable. Records in the dataset are from a total of 54808 employees and there are 11 employee evaluation features.

Several authors have already done some previous work on this dataset and published their work on Kaggle. They have performed different machine learning models including Logistic Regression, Random Forest, Gradient Boosting etc. and all of the models have achieved great accuracy on this specific problem. Among these, it shows that XGBoost, an efficient and scalable implementation of gradient boosting framework, could achieve the highest accuracy of 94% [2]. However, for this issue concerned with an individual's career development, the closer the accuracy is to perfect the better. Therefore, the goal of this project is to further improve the accuracy of the previously mentioned models.

## 2 Exploratory Data Analysis

Several figures have been created during the exploratory data analysis as shown below.

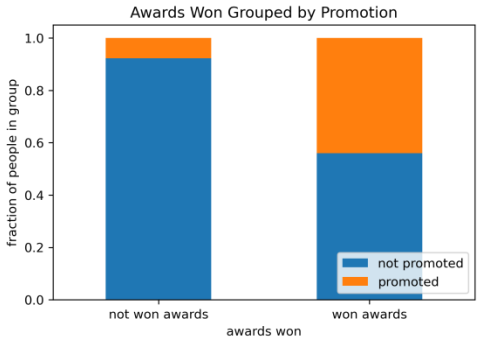


**Figure 1** Promoted employee ratio distribution over awards-won

The stacked bar plot above shows that employees who have won awards have a significantly higher fraction of promotions compared to those without awards. It could suggest that winning awards means more chance to get promoted.

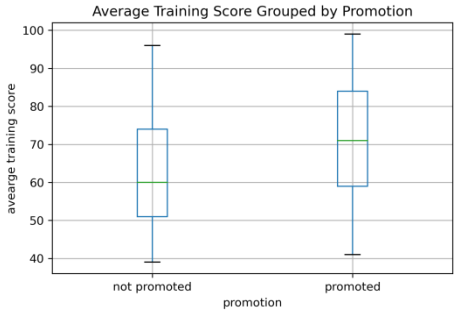


**Figure 2** Average training score distribution over promotion

The boxplot above shows that promoted employees have achieved higher training scores than those who did not get promoted. It is reasonable to conclude that an employee's

training performance could play an important role in the evaluation for promotion.

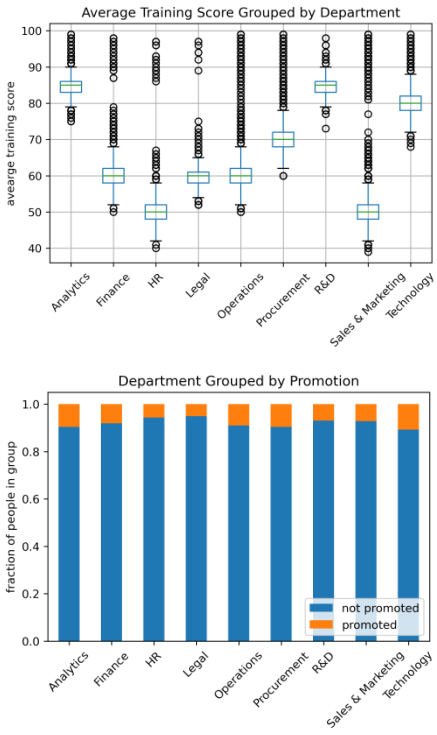


**Figure 3** Relationships of promotion ratio and average training score for different department

The average training score varies on departments, as shown in the boxplot on the top. The stacked bar plot on the bottom shows the distribution of the promoted employee ratio over different departments. It can be seen that employees in Sales & Marketing, Operations, Finance, HR and Legal generally have lower scores while those in Technology, Analytics and R&D have higher scores. However, the promotion ratio of department Operations and Finance is relatively high while that of department R&D is low. The explanation could be that each department grades their employees' training performance separately so the standard of grading can be different to some extent. For example, perhaps the department Operations tends to give employees lower scores due to a strict grading standard but there are many employees in the department indeed qualified for promotion. Actually the Operation department and the Sales & Marketing department have the greatest number of employees – that intense competition could be the reason for a strict grading standard.

## 3 Methods

### 3.1 Data Splitting and Preprocessing

Since the dataset is imbalanced, for splitting, first 20% of the data is randomly chosen for test and then the stratified K-fold method is used for the other 80% of the data for a 4 fold cross-validation.

For preprocessing, OrdinalEncoder is applied on the education feature because it is an ordered categorical feature with ranked categories and on the previous_year_rating feature since there are only 5 categories of ratings from 1.0 to 5.0 as well as a category of 0 indicating the null values. OneHotEncoder is applied on other categorical features. MinMaxEncoder is applied on the continuous features since they are all bounded.

The data is independent and identically distributed without any group structure and is not time-series data. The data has 56 features after preprocessing.

### 3.2 Machine Learning Models

Four different ML algorithms have been selected on this dataset: Logistic Regression, Random Forest, Gradient Boosting and XGBoost. For Logistic Regression, four different conditions are considered: no penalty (without regularization), L1 regularization, L2 regularization and ElasticNet. All the models are trained and compared.

For hyper-parameter tuning on the models, GridSearchCV, a brute-force grid search method is used to find the best parameter combination specified with the param_grid. The process is repeated on 5 different random states for 5 different splits.

Parameters tuned and values tried for each model are as below:

| Model | Parameters |
|---|---|

| | |
|---|---|
| Logistic Regression (no penalty) | solver:['newton-cg','lbfgs','sag','saga'] |
| Logistic Regression (L1) | solver:['liblinear','saga']<br>C:[5,10,30] |
| Logistic Regression (L2) | solver:['newton-cg','lbfgs','saga']<br>C:[10,20,30] |
| Logistic Regression (ElasticNet) | C:[50,75,100]<br>l1_ratio:[1e-7,1e-6] |
| Random Forest | max_depth:[10,15,20]<br>max_features:[40,50,60] |
| Gradient Boosting | max_depth:[3,4,5]<br>learning_rate:[0.5,0.6,0.75] |
| XGBoost | max_depth:[3,4,5]<br>gamma:[0.1,0.5] |

**Table 1** Parameters used for tuning of each model

Accuracy is chosen to be the models evaluation metric in this classification problem because the goal of this project is to further improve the models accuracy of previous work conducted by other authors.

## 4 Results

In this classification problem, the baseline accuracy is 91.44% for the random state where all the models reach the best score by using the simple majority estimation method.

From the results, the XGBoost model is the most predictive with the accuracy of 94.22% and a standard deviation of 2.78% above the baseline.

The performance of all the ML models is summarized in the table below:

| Model | Accuracy | Std above baseline |
|---|---|---|
| Logistic Regression | 93.93% | 2.49% |
| Random Forest | 94.09% | 2.65% |
| Gradient Boosting | 93.79% | 2.35% |
| XGBoost | 94.22% | 2.78% |

**Table 2** ML models performance

An interesting finding is that all the fours Logistic Regression models have achieved the same best accuracy though the parameter combinations are different.

The avg_training_score feature is the most important feature from the findings of the global feature importance plots for the Logistic Regression model, the Random Forest model and the Gradient Boosting model. Department is also of great importance. The feature recruitment_channel and length_of_service tend to have the least importance.

However, age should not be of as such great importance as shown in the plot of Random Forest global feature importances. Hence, the method of using the feature_importances_ metric of the Random Forest model is not as accurate as the permutation method and the SHAP method.

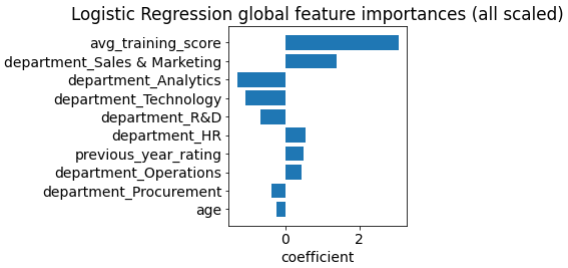


**Figure 4** Logistic Regression global feature importances (all scaled) using coefficients

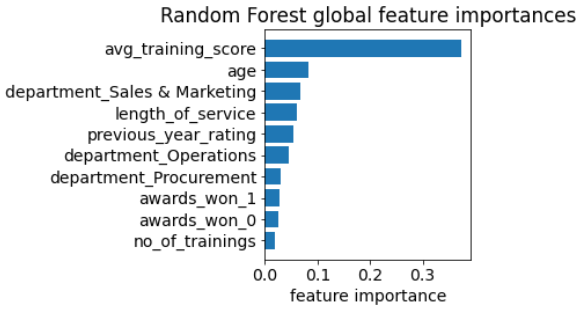


**Figure 5** Random Forest global feature importances using feature_importances_ metric

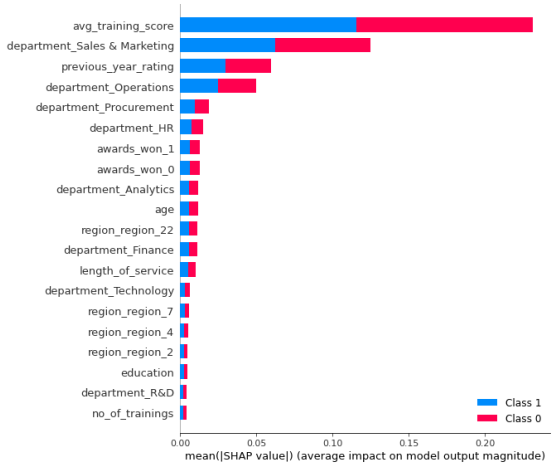


**Figure 6** Random Forest global feature importances using SHAP

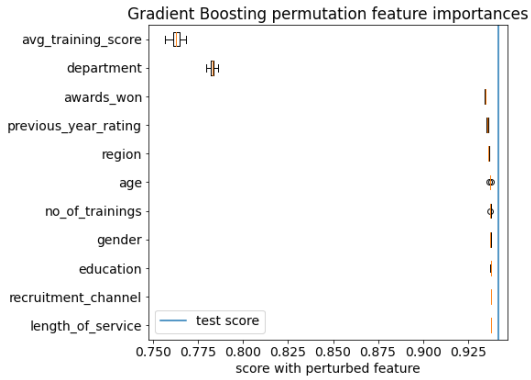


**Figure 7** Gradient Boosting global feature importances using permutation method

## 5 Outlook

## 6 References

[1] Стефан (2019, June 21). How to optimize the promotion process in your company using data science? Medium.

[2] M. I. Zaman (2021, Sep 26). Employee Promotion end-to-end Solution. Kaggle. https://www.kaggle.com/muhammadimran112233/employee-promotion-end-to-end-solution

**Github repository link:**

https://github.com/Ziyin-Li/finproject-Ziyin-Li.git