Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/2024.25128

Optimizing Employee Promotions: A Machine Learning Approach Using AdaBoost

Bohua Li[1*], Jie Geng[2], Qiyue Lu[3], Zikuan Xing[4]

[1]*University of Washington, Seattle, USA*
[2]*Huaqiao University, Quanzhou, China*
[3]*Soochow University, Hangzhou, China*
[4]*Shanghai World Foreign Language Academy, Shanghai, China*
**Corresponding Author. Email: bohua.li23@gmail.com*

Abstract. For many companies, deciding on employee promotions is challenging due to subjective judgments and financial constraints. Many employees demand higher salaries and promotions after working for a few years; they are not satisfied with the status quo. However, companies are not willing to give promotions to employees easily, which will lead to a higher expenditure for the enterprise and greater economic pressures. This paper introduces a statistical analytics model to provide objective recommendations for employee promotions, thus minimizing biases and enhancing decision-making processes. The model can help the human resources department to recognize outstanding employees at the same time. With the accumulation of more data and the iteration of models, promotion strategies can be continuously adjusted and optimized to more accurately reflect employees' abilities and performance. Utilizing a comprehensive dataset, including age, service length, awards, and performance ratings, we applied multiple predictive models—Bagging, Random Forest, AdaBoost, Gradient Boosting, and Logistic Regression, to find significant correlations between specific employee attributes and promotion likelihood, highlighting the critical role of training scores. The implementation of automated predictive models has not only reduced the HR department's workload but also improved the overall efficiency of the promotion process by enabling continuous adjustments and optimizations based on data insights.

Keywords: AdaBoost, Employee promotion, Machine learning

1. Introduction

1.1. Business background

In the workplace, promotion often means more authority and higher pay. It affirms an employee's workability, attitude, and efficiency. Employees often work harder for promotion, creating more resources for the company. The Human Resources (HR) department is one of the most important departments in the promotion process. The HR department is often responsible for recruiting employees, adjusting their work, calculating their KPIs, and more. Promoting employees is crucial

DOI: 10.54254/2754-1169/2024.25128

for HR, as a suitable promotion incentive mechanism is vital for the smooth operation of a company. However, this task takes work. The HR department needs to assess employees extensively, recording and collecting personal data. Often, promotions are judged based on these data and metrics by the HR team. In large companies, the data can be extensive and complex, and the key metrics used by HR for promotion decisions may not be fair or directly related to employee promotion. Thus, there is a need for an objective evaluation standard based on data and models, which can be used by various companies and is highly accurate. This will significantly reduce the workload of the HR department, which is the purpose of this research.

We obtained comprehensive data on employee promotion cycles from the HR team at JDM Company, which includes a range of personal and professional information. The dataset comprises unique employee identifiers, department affiliations, and classified and numbered regions indicating employee locations. It also includes details on educational backgrounds, gender (m for male, f for female), and recruitment channels, distinguishing between sourcing and other channels. Additionally, the data captures the number of training sessions employees have completed, their ages, previous year performance ratings, length of service in years, awards received, and average scores from company training sessions. This detailed information is utilized to analyze and understand promotion cycles within the company.

1.2. Brief overview of methods

We selected five machine learning algorithms to determine the best model for predicting employee promotion: Gradient Boosting, Logistic Regression, Adaboost, Random Forest, and Bagging Classifier. Due to the imbalanced nature of this dataset, oversampling and undersampling methods were used. The models were evaluated on the validation set using metrics such as accuracy, precision, recall, and F1-score. As a result, we selected the best model based on performance and identified the most critical features for promotion that can assist the HR team in predicting employee promotions.

1.3. Previous research

In the study titled "Employee Promotion Prediction Using Improved AdaBoost Machine Learning Approach", the authors employed an enhanced AdaBoost machine learning technique to predict employee promotions [1]. The findings demonstrated that the modified AdaBoost model significantly improved the accuracy of promotion predictions. The study highlighted the key benefits of using machine learning techniques, such as Logistic Regression and Random Forest, in predicting employee promotions.

Jafor et al. found that the modified AdaBoost method outperformed several other machine learning models, including Support Vector Machine, Logistic Regression, Artificial Neural Network, Random Forest, and XGBoost [1]. The modified AdaBoost model achieved an accuracy rate of 95.30%, a precision of 92.79%, a recall of 98.93%, and an F1-score of 95.76%. These results indicate that the modified AdaBoost method provided significant improvements in employee promotion prediction, surpassing traditional machine learning models with an accuracy rate of up to 95.30%.

Similarly, Kogila et al. reported that the Hybrid-ABC-AdaBoost model achieved an accuracy of 97.22% in their study; underscores the effectiveness of hybrid approaches in enhancing prediction accuracy [2].

DOI: 10.54254/2754-1169/2024.25128

In another study, "Analysis and Prediction of Employee Promotions Using Machine Learning" by Alqahtani and Almaleh, the Gradient Boosting method was found to perform slightly better than the AdaBoost model, suggesting that different boosting techniques may offer varying levels of effectiveness depending on the dataset and context [3].

However, it is important to note that in a related study investigating employee performance rather than promotion, K-Nearest Neighbors and Artificial Neural Network models did not perform well [4]. This discrepancy may be attributed to differences in the datasets, as the study focused on employee performance rather than promotion. Additionally, the lack of hyperparameter tuning could have impacted the results [4].

The research conducted by Long et al. demonstrated that the Random Forest model performed best among several machine learning models, verifying the validity of features such as personal basic features and post features [5]. They found that post features, including working years, the number of different positions, and the highest department level, had a higher impact on promotion compared to personal basic features [5].

Similarly, Al-Alawi and Albuainain found that employing data balancing techniques like SMOTE significantly improved the accuracy of promotion predictions [6]. Their study achieved a remarkable accuracy rate of 99% using the Random Forest Classifier (RFC) in combination with SMOTE. They also emphasized the importance of addressing imbalanced data to enhance model performance,

Our study aligns with these findings, showing similar improvements when applying the AdaBoost model to the original data for employee promotion prediction. In our research, the accuracy, precision, recall, and F1-score of several models were as follows...

2. Method

2.1. Data collection

The data for this project includes details of each employee: ‘employee_id,’ ‘department,’ ‘region,’ ‘education,’ ‘gender,’ ‘recruitment_channel,’ ‘no_of_trainings,’ ‘age,’ ‘previous_year_rating,’ ‘length_of_service,’ ‘awards_won,’’avg_training_score,’ ‘is_promoted.’

2.2. Exploratory data analysis (EDA)

Exploratory data analysis was conducted to gain initial insights into the datasets as well as relationships between each variable. Univariate analyses were conducted on certain features (‘age’, ‘length_of_service’, ‘avg_training_score’, ‘previous_year_rating’). Bivariate analyses were conducted between certain features (‘Age’, ‘Awards_won’, ‘department’, ‘previous_year_rating’, ‘education’, ‘gender’) with the target variable ‘is_promoted’. A correlation analysis was then conducted in order to identify the top features that are closely related to the target variable ‘is_promoted’.

2.3. Data preprocessing

The data was split into training and validation sets in an 80:20 ratio to facilitate model training and performance evaluation. For missing values in the dataset, we employed the SimpleImputer from scikit-learn, using different strategies for different columns. Specifically, the strategy set to 'most_frequent' was used to impute missing values in the ‘education’ column, as it often has categorical values where the mode is a suitable replacement. For the ‘previous_year_rating’ and ‘avg_training_score’ columns, the 'median' strategy was chosen to handle missing values because

DOI: 10.54254/2754-1169/2024.25128

these columns contain numerical data, and the median is a robust measure that is not overly affected by outliers.

Categorical variables such as 'department', 'region', 'education', 'gender', and 'recruitment_channel' were converted into numerical values using one-hot encoding. This process creates binary columns for each category, allowing these categorical features to be used effectively by machine learning algorithms. Additionally, the 'employee_id' feature was dropped from the dataset. This decision was made because 'employee_id' is a unique identifier and does not contain any information that would help in predicting whether an employee gets promoted.

2.4. Model selection

Several machine learning algorithms were selected to identify the best model for predicting employee promotions, including Bagging classifier, Random Forest, AdaBoost, Gradient boosting, and logistic regression. Given the imbalanced nature of the dataset, where the number of promoted employees was significantly lower than the number of non-promoted employees, oversampling and undersampling methods were used to treat this imbalanced dataset. All of the models were applied to these three types of datasets (original dataset, oversampled dataset, undersampled dataset). This approach allowed us to compare the performance of the models under different conditions and ensure that the chosen model could handle the imbalance effectively. The models were evaluated on the validation set using metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model's performance, with accuracy indicating the overall correctness, precision showing the proportion of true positive predictions among all positive predictions, recall indicating the proportion of actual positives correctly identified, and F1-score providing a balance between precision and recall. To further refine the models, Randomized Search CV and cross-validation techniques were then applied to fine-tune the hyperparameters of all five models and optimize these metrics. The best-performing model was selected in the end based on the performance and prepared for deployment to assist the HR team in predicting employee promotions.

3. Results

3.1. Exploratory data analysis

The age, length of service, and average training score of employees exhibit a normal distribution.

As shown in Figures 1 and 2, the distributions of age and length of service are unimodal and right-skewed. Figure 3 shows a multimodal distribution of the average training score.

DOI: 10.54254/2754-1169/2024.25128

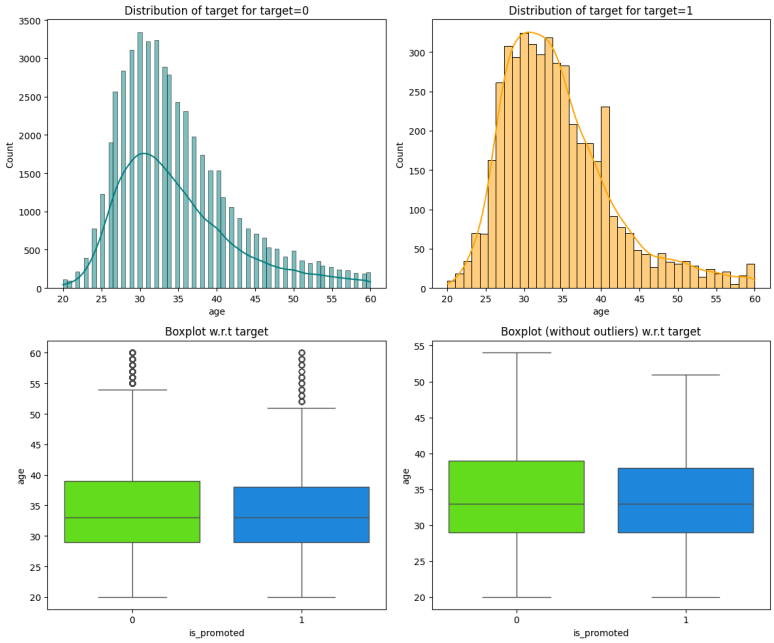

Figure 1: Target variable vs. Age

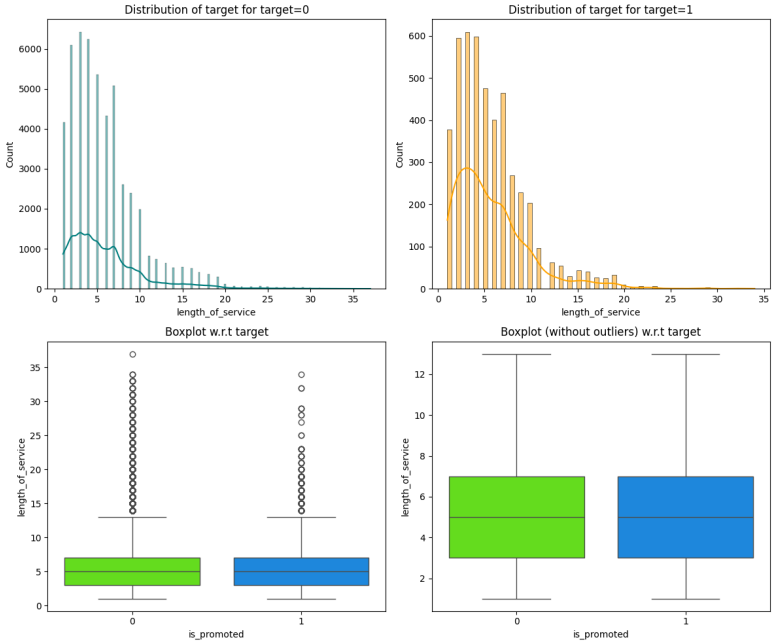

Figure 2: Target variable vs. Length of service

DOI: 10.54254/2754-1169/2024.25128

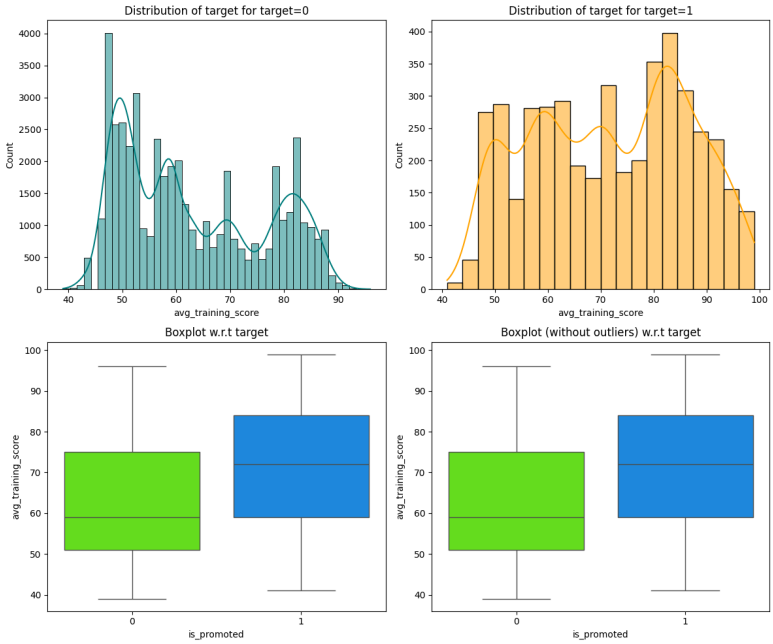

Figure 3: Target variable vs. Average training score

Figure 4 illustrates the correlation coefficients between various features and the target variable 'is_promoted' in the dataset. Features such as 'awards_won', 'avg_training_score, and 'previous_year_rating' show the highest positive correlation with the promotion outcome.

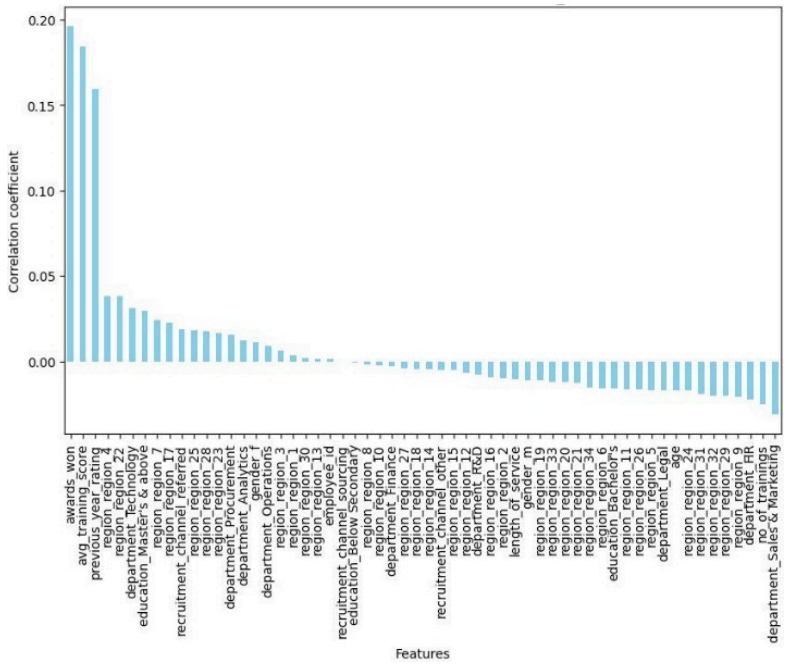

Figure 4: Correlation of features with target variable(is_promoted)

3.2. Data preprocessing

The data set was successfully preprocessed through data replication, dataset segmentation, missing value filling, and encoding classification variables, preparing for model training. Specifically, the

DOI: 10.54254/2754-1169/2024.25128

dataset is reasonably segmented, providing sufficient training and validation data for the model. The missing values were appropriately handled to avoid introducing bias in model training; converting categorical variables into numerical features that the model can better handle can help improve the performance of the model. The entire process ensures the quality and consistency of data, laying a solid foundation for subsequent model training and evaluation. Figure 5 shows successful results after the missing value-filling process that has no missing values.

DOI: 10.54254/2754-1169/2024.25128

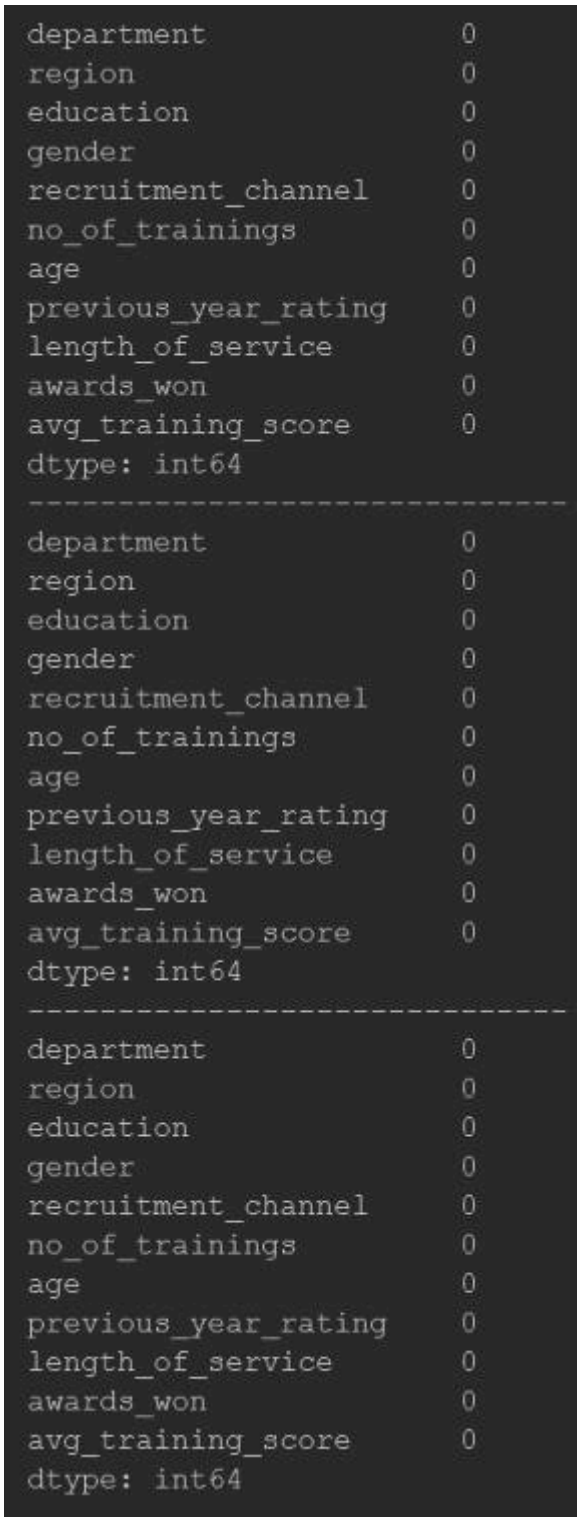

Figure 5: No missing values after data processing

DOI: 10.54254/2754-1169/2024.25128

4. Model building

4.1. Original data model

Initially, we constructed models using original data, including Bagging, Random Forest, AdaBoost, Gradient Boosting, and Logistic Regression. Figure 6 shows us the recall scores for cross-validation are as follows: Bagging (0.336), Random Forest (0.262), AdaBoost (0.165), Gradient Boosting (0.278), and Logistic Regression (0.113). The model that performs best on the validation set is Logistic Regression, with a recall rate of 0.125.

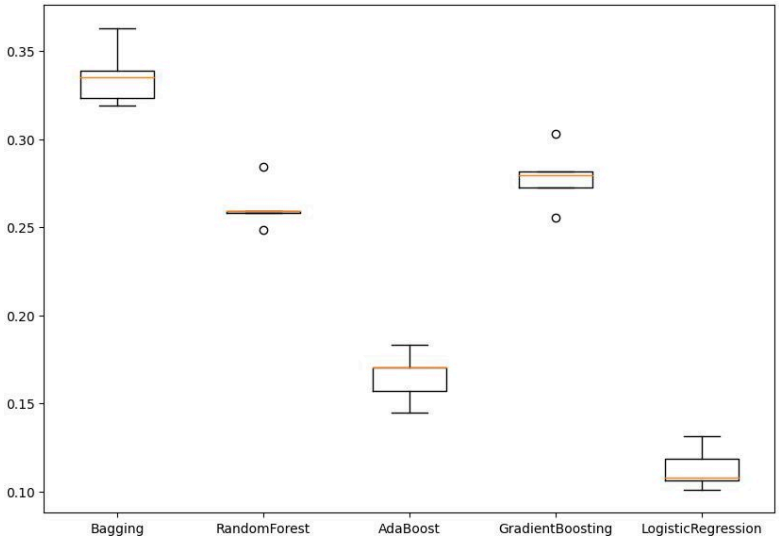

Figure 6: Algorithm comparison

4.2. Oversampled data model

To address the issue of class imbalance, we applied the Synthetic Minority Oversampling (SMOTE) technique to the training data. Figure 7 shows us the cross-validation recall scores of all models significantly improved: Bagging (0.930), Random Forest (0.932), AdaBoost (0.906), Gradient Boosting (0.899), and Logistic Regression (0.759). The Logistic Regression model continues to perform well on the validation set, with a recall rate of 0.371.

DOI: 10.54254/2754-1169/2024.25128

Figure 7: Algorithm comparison

4.3. Undersampled data model

Another way to address the issue of imbalance is to undersample the majority categories. Figure 8 shows us the cross-validation recall scores for undersampled data are Bagging (0.622), Random Forest (0.669), AdaBoost (0.666), Gradient Boosting (0.617), and Logistic Regression (0.667). The AdaBoost model achieved the highest performance on the validation set with a recall rate of 0.780.

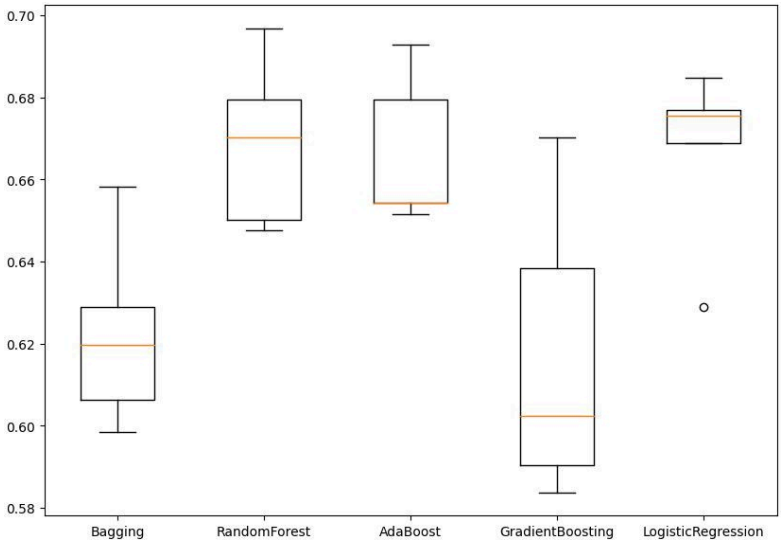

Figure 8: Algorithm comparison

4.4. Hyperparameter tuning

Figures 9 and 10 demonstrated the training performance comparison and validation performance comparison results for five tuned models: AdaBoost, Logistic Regression, Random Forest, Bagging, and Gradient Boosting.

DOI: 10.54254/2754-1169/2024.25128

```
Training performance comparison:
   AdaBoost trained with Undersampled data  \
0                                    0.752

   AdaBoost trained with Original data  \
0                                 0.940

   AdaBoost trained with Oversampled data  \
0                                    0.946

   LogisticRegression trained with Undersampled data  \
0                                              0.721

   LogisticRegression trained with Original data  \
0                                           0.936

   LogisticRegression trained with Oversampled data  \
0                                              0.882

   RandomForest trained with Undersampled data  \
0                                         0.864

   RandomForest trained with Original data  \
0                                    0.973

   RandomForest trained with Oversampled data  \
0                                         0.995

   Bagging trained with Undersampled data  Bagging trained with Original data  \
0                                    0.999                                0.992

   Bagging trained with Oversampled data  \
0                                    0.993

   GradientBoosting trained with Undersampled data  \
0                                              0.774

   GradientBoosting trained with Original data  \
0                                         0.941

   GradientBoosting trained with Oversampled data
0                                           0.921
```

Figure 9: Training performance comparison

DOI: 10.54254/2754-1169/2024.25128

```
Validation performance comparison:
   AdaBoost trained with Undersampled data  \
0                                    0.792

   AdaBoost trained with Original data  \
0                                 0.940

   AdaBoost trained with Oversampled data  \
0                                    0.918

   LogisticRegression trained with Undersampled data  \
0                                              0.757

   LogisticRegression trained with Original data  \
0                                           0.938

   LogisticRegression trained with Oversampled data  \
0                                              0.872

   RandomForest trained with Undersampled data  \
0                                         0.777

   RandomForest trained with Original data  \
0                                    0.935

   RandomForest trained with Oversampled data  \
0                                       0.911

   Bagging trained with Undersampled data  Bagging trained with Original data  \
0                                   0.756                                0.939

   Bagging trained with Oversampled data  \
0                                  0.922

   GradientBoosting trained with Undersampled data  \
0                                              0.795

   GradientBoosting trained with Original data  \
0                                           0.939

   GradientBoosting with Oversampled data
0                                   0.911
```

Figure 10: Validation performance comparison

4.5. Final model selection

Based on the performance score of each model, classification report and confusion matrix, it is evident that the AdaBoost model trained on the original data performs best. In Figure 11, it is shown that the model achieves high precision, with scores of 0.94 for non-promoted employees and 0.92 for promoted employees. The recall for non-promoted employees is perfect at 1.00, indicating all non-promoted employees are accurately identified. The recall for promoted employees is low at 0.31. The F1 Score is 0.97 for non-promoted employees. The overall accuracy of the model is 0.94. The confusion matrix in Figure 12 confirms this high accuracy: 7527 true negatives, 208 true positives, 468 false negatives, and only 18 false positives.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      7545
           1       0.92      0.31      0.46       676

    accuracy                           0.94      8221
   macro avg       0.93      0.65      0.71      8221
weighted avg       0.94      0.94      0.93      8221
```

Figure 11: Classification report

DOI: 10.54254/2754-1169/2024.25128

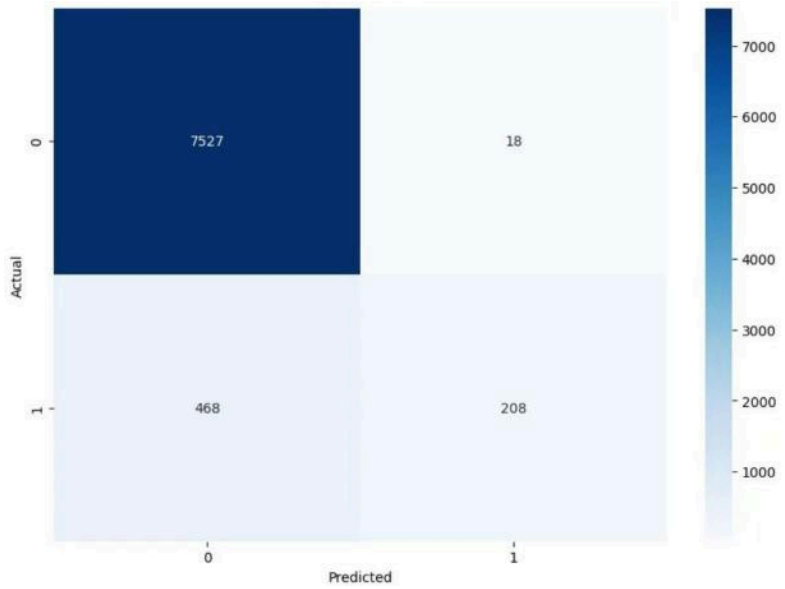

Figure 12: Confusion matrix

4.6. Feature importance

We visualized the importance of the feature for the AdaBoost model trained on original data and revealed the relative contribution of each feature to model prediction. Figure 13 indicates the average training score, age, length of service, previous year rating, and department Finance are features that have strong correlations with the target variable is_promoted, making them key predictors and providing valuable insights for business strategy and decision-making.

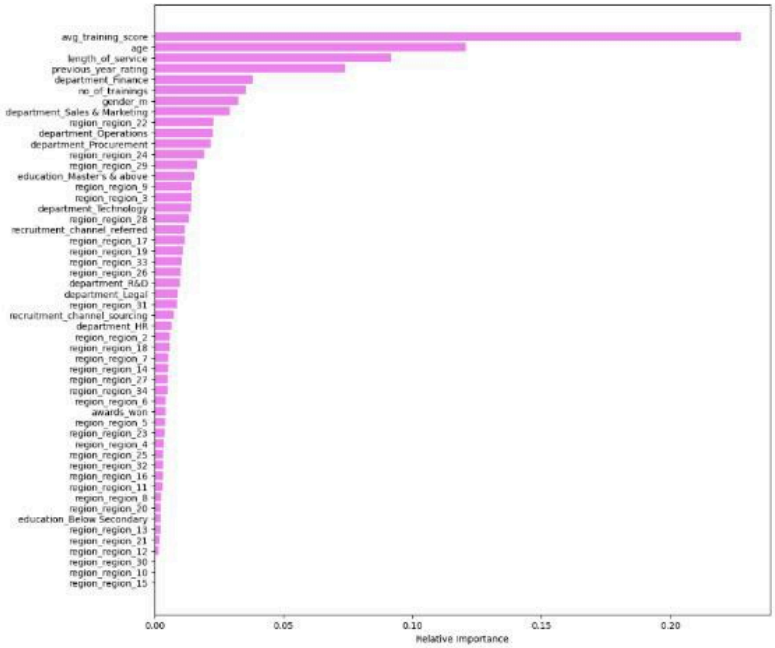

Figure 13: Feature importance

DOI: 10.54254/2754-1169/2024.25128

5. Discussion

The results of this study offer significant insights into the factors influencing employee promotion and the effectiveness of using different machine-learning models to predict employee promotion in a business setting. The findings of this study provide practical implications for promotion dynamics and HR practices in businesses.

5.1. Exploratory data analysis (EDA)

The EDA section revealed crucial information about the dataset. Variables such as age, length of service, and average training score of employees exhibited normal distributions despite different promotion statuses, suggesting a balanced representation across these variables. The distribution of variables such as department, education, and gender is imbalanced with regard to promotion statuses, highlighting potential areas of bias or structural differences.

Age and average training score were positively correlated with promotion, indicating that older employees with higher training scores are more likely to be promoted. In addition, employees who have won awards in the past are also more likely to be promoted, underscoring the value of experience and recognition in promotion decisions.

5.2. Model performance

The study's results demonstrate that the AdaBoost model trained on the original data outperformed other models in predicting employee promotions according to the validation performance comparison (fig.9, fig.10). In addition, the classification report also indicates high precision, achieving scores of 0.94 for non-promoted employees and 0.92 for promoted employees (fig.11). The recall for non-promoted employees was perfect at 1.00. However, the recall for promoted employees is only 0.31, indicating that the model failed to recognize a significant number of employees who should be promoted (fig.11). This suggests that while the model accurately identifies non-promoted employees, it misses a significant number of actual promotions. The F1 Score for non-promoted employees was 0.97, indicating a good balance between precision and recall. The model's overall accuracy was 0.94, with the confusion matrix confirming high accuracy, showing 7527 true negatives, 208 true positives, 468 false negatives, and only 18 false positives (fig.12).

5.3. Feature importance

Based on the feature analysis of this study, The average training score, age, length of service, previous year rating, and the department (Finance) emerged as key predictors for employee promotion. Suggesting HR departments should focus on these areas for employee development and promotion strategies.

5.4. Implications and recommendations

The implications of this study are significant for the HR departments in large organizations. Based on the performance comparison, one can see that the AdaBoost model outperformed other machine learning models; this result is different compared to some of the previous findings. In this study, the AdaBoost model's high precision in predicting promotions can assist HR teams in making more informed and objective promotion decisions, thereby reducing the potential for bias and human error. The high recall for non-promoted employees ensures that almost all non-promotions are

DOI: 10.54254/2754-1169/2024.25128

accurately identified, although the model's lower recall for promoted employees suggests that additional factors or alternative models might be needed to capture all potential promotions accurately.

5.5. Business recommendations

5.5.1. Implementation of the Adaboost model

Based on the high precision and accuracy of the Adaboost model, it should be used as a primary tool for evaluating employee promotion in businesses. The Adaboost model can significantly reduce the workload of HR teams and increase the objectivity of promotion decisions.

5.5.2. Employee training and development

According to insights from the feature importance analysis, focusing on key areas such as training scores, length of service, and performance ratings to enhance the promotion prospects of employees. Businesses could strengthen or invent new employee-training modules in order to increase the promotion rate for outstanding employees, as training score is the number one factor that influences employee promotion status.

5.5.3. Address recall for promotions

To improve the model’s recall for promoted employees, businesses could consider adding additional features that are more indicative of promotion potential or combining the AdaBoost model with other machine learning models that might better capture the nuances of promotion decisions.

5.5.4. Continuous monitoring for improvement

Businesses should continuously monitor the performance of the AdaBoost model, regularly updating and retraining it with new data to ensure it maintains high performance. Additionally, businesses should address any changes in promotion criteria or company dynamics to keep the model relevant and accurate.

6. Conclusion

In conclusion, the study's findings highlight the effectiveness of the AdaBoost model in predicting employee promotions and provide actionable insights for HR departments across businesses to improve their promotion processes. The use of machine learning models like AdaBoost can lead to more data-driven and efficient suggestions for promotion practices, benefiting both employees and the organization.

References

[1] M. A. Jafor, M. A. H. Wadud, K. Nur, and M. M. Rahman, “Employee Promotion Prediction Using Improved AdaBoost Machine Learning Approach”, AJSE, vol. 22, no. 3, pp. 258 - 266, Dec. 2023. doi: https: //doi.org/10.53799/ajse.v22i3.781

[2] N. Kogila, R. Rajkumar, S. Rajesh, and S. Vennila, "A Novel Approach of Ecommerce for Sales Prediction Using Hybrid ABC and AdaBoost Approach, " 2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC), Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/ICECCC61767.2024.10593970.

DOI: 10.54254/2754-1169/2024.25128

[3] F. A. Alqahtani and A. Almaleh, "Analysis and Prediction of Employee Promotions Using Machine Learning, " 2022 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 2022, pp. 01-09, doi: 10.1109/DSIT55514.2022.9943959

[4] L. G. Tanasescu, A. Vines, A. R. Bologa, and O. Vîrgolici, "Data Analytics for Optimizing and Predicting Employee Performance, " Appl. Sci., vol. 14, 3254, 2024, doi: https: //doi.org/10.3390/app14083254.

[5] Yuxi Long, Jiamin Liu, Ming Fang, Tao Wang, and Wei Jiang. 2018. Prediction of Employee Promotion Based on Personal Basic Features and Post Features. In Proceedings of the International Conference on Data Processing and Applications (ICDPA 2018). Association for Computing Machinery, New York, NY, USA, 5–10. https: //doi.org/10.1145/3224207.3224210

[6] A. I. Al-Alawi and M. S. Albuainain, "Machine Learning in Human Resource Analytics: Promotion Classification using Data Balancing Techniques, " 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Manama, Bahrain, 2024, pp. 1-5, doi: 10.1109/ICETSIS61505.2024.10459566.