Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

Employee Promotion Evaluation and Prediction using Machine Learning

Nareen Ansari[1], Neha Vora[2]

[1]Student, [2]Assistant Professor, SVKM'S Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, Maharashtra

E-mail: [1]nareenansari@gmail.com, [2]nehavora2501@gmail.com

Abstract

Promoting an employee is an important responsibility of the HR department. Various factors contribute to an employee's promotion, such as age, recruitment channel, number of training, academic qualifications, and length of service for the employee. These factors majorly affect the promotion. This research explores employee promotion evaluation and aims to predict whether an employee will be promoted. The dataset used is a primary dataset, which has been gathered through surveys from employees asking for their information. In this study, predictive analysis will be studied based on the criteria estimated for the employees in the promotion process by machine learning algorithms such as logistic regression, random forest classifier, gradient boosting classifier, and decision tree classifier. Logistic regression achieved the highest performance with 86% accuracy and 86% precision. This research can be beneficial to HR and managers in evaluating and predicting employee promotions.

Keywords: Employee Promotion, Prediction, Promoted, Machine learning Algorithms, Promotion Evaluation.

1. Introduction

Employee promotion is a challenging and time-consuming task. An organization doesn't only consist of a single office or department; rather, it consists of many departments and enterprises, especially in MNCs. In every organization, there are various sections and departments, each with its specific responsibilities and functions. The HR department looks

DOI: https://doi.org/10.36548/jitdw.2024.4.001
Received: 28-08-2024. received in revised form: 21-09-2024, accepted: 06-10-2024, published: 17-10-2024

after management tasks, the IT department performs IT-related tasks, and the marketing department is for advertisement. There are various departments like these in an organization. Promotion is an honor for an employee that also indicates their appreciation for their hard work and seeing them capable of taking on more responsibilities while increasing their pay. When promotion evaluation is done accurately, it leads not only to an organization's growth but also to an employee's career growth.

According to previous research, an employee's performance is significantly impacted by promotions. When an employee gets a promotion, their responsibilities increase, and their work performance increases too. The fact that there is a status increment of employees in the organization can be a motivation to work hard. Several factors affect the employee's promotion. These factors have been included in the dataset. They are their age, gender, academic qualification, department, recruitment channel, number of trainings undertaken, previous year's performance ratings, and length of service in the company. The study utilizes machine learning algorithms, such as logistic regression, random forest classifier, gradient boosting classifier, and decision tree classifier to analyse the criteria for the employees' promotion process and predict outcomes. This approach is more beneficial for HR and managers in evaluating and predicting employee promotions

2. Related Work

Employee performance analysis, taking into account all factors and evaluating promotion, has been studied and applied many times before by human resource management. This analysis and evaluation can be called predictive analytics, and this has been used as a common tool. Employee promotion has been calculated using many machine learning algorithms, like decision trees, which are quite popular among HR for their interpretability [1]. Logistics regression, which is used for classification tasks, is used for predictive analysis as it performs well when the relationship between dependent and independent variables is linear. Random forest and gradient-boosting classifiers improve upon decision trees by combining multiple trees to enhance predictive performance.

There was a study conducted in [2] that was based on the factors that were taken into consideration for an employee's promotion, such as their age. As told before, by knowing the age of an employee, their level of experience, and expertise, the responsibilities they can handle after promotion can be predicted. The recruitment channel is considered for two reasons: to

know their performance metrics for HR to optimize their hiring strategies and to align their strategies with employee promotion. Previous year ratings of an employee are considered to know the performance of an employee in the previous year [3]. The number of trainings is considered to indicate the employee's willingness to grow in the respective field and their commitment to development. Academic qualifications help to understand employee's level of expertise and their theoretical and practical knowledge. Length of service mostly indicates an employee's loyalty to the organization.

Some studies were based on machine learning applications that the human resources department uses to evaluate and predict many things, such as turnover prediction, which mainly predicts how likely it is for an employee to leave the organization [4]. And then there is significant literature on automating performance reviews using natural language processing (NLP) or machine learning to assess qualitative data from performance reviews, which can be a precursor to promotions. Comparative studies in HR analytics often evaluate the performance of different machine learning models to predict outcomes like promotions, job performance, or turnover [5].

When an employee gains more experience over time it becomes valuable for an organization. Enterprises usually quarrel over these kinds of valuable employees. So, they advertise various attractive offers to such valuable employees. This results in employees getting attracted to these offers and leaving their current organizations. HR must identify such a situation taking place in the organization and inform the responsible people of the organization [6]. To deal with such a situation, the logistic regression model was implemented here. There are 2 functions used for 2 different needs: the cross-entropy method as the objective function and Newton’s method and regularization to optimize the model.

3. Methodology

3.1 Dataset Preparation

In this study, primary data was collected from the various organizations through a questionnaire administered by the employees. The information acquired comprised age, gender, academic qualifications, department, recruitment channel, number of trainings, previous year's performance ratings, and length of service, all with the assurance that the data were to be used only for research purposes. The final dataset contained 999 records, which

contained the above information. The dataset had a variance in age, departments, length of services, and whether they were promoted or not.

Table 1. The Overview of the Dataset

Feature Name	Description	Data Type	Number Of Samples
Age	Age of the employee	Integer	999
Gender	Gender of the employee	Categorical	999
Highest academic qualification	Highest Academic qualification of the employee	Categorical	999
Department	Department of Employees where they work	Categorical	999
Recruitment channel	Channel which was used to recruit employee	Categorical	999
No. of training	Number of trainings that an employee attended	Integer	999
Previous year ratings	Ratings they got in their previous year	Integer	999
Length of service	The number of years an employee has been working in the company	Integer	999
Is promoted	Whether the employee was promoted or not	Categorical	999

Table 1 shows the overview of the employee promotion evaluation and prediction dataset, containing 999 samples, in which each feature's description and datatype have been specified.

3.2 Dataset Preprocessing

The dataset was preprocessed to ensure its accuracy for analysis before training the machine learning model. For preprocessing, feature engineering and feature selection methods were applied.

3.2.1Feature Engineering

- Label Encoding: In Label Encoding, categorical features such as “Select Gender”, “Select Your Highest Academic Qualification”, and “What is Your Department?’ have been encoded into numeric values.
- One-Hot Encoding: This was used for features like "What is your department?" and "What was your recruitment channel?" to eliminate ordinality from categorical variables.
- Imputation: The missing values in the “Previous Year Ratings” and in “Select Your Highest Academic Qualification” were imputed through median and mode, respectively.
- Scaling: The feature values of continuous variables like “Enter No. of training” and “Enter your length of service” are scaled using MinMaxScaler to normalize the feature values.

3.2.2 Feature Selection

- Correlation Heatmap: it determines the relationships between numerical features, such as age, number of trainings, previous year ratings, length of service, and the influence it holds on promotion. Low correlation features can be reduced for simplicity.
- Manual Feature Selection: Features such as target variable (is promoted) were dropped before training the models to avoid leakage, and only input variables were considered.

No data augmentation techniques were applied as the primary dataset was sufficient for modeling. The dataset was divided into two sets, namely the training set and testing set, and in the ratio of 80% and 20%, respectively. This assures that 80% of the data is for training, while the remaining 20% is used in tests for generalization evaluation of unseen models.

3.3 Experimental Setup

In this experiment, Python with its libraries such as Pandas, NumPy, and Scikit-learn for machine learning models including logistic regression, decision trees, random forests, and gradient boosting were used for data preprocessing and manipulation, Matplotlib and Seaborn for data visualization, and Jupyter Notebook for the development.

3.4 Model Used

This study evaluates and predicts employee promotion using machine learning models. The models were chosen based on their classification ability and handling of imbalanced data. Four models have been trained and evaluated in this study: Logistic Regression, Decision Tree, Gradient Boosting Classifier, and Random Forest.

Logistic regression is used in this study for its baseline model, as it is quite popular for its simplicity and interpretability [7]. It is used to understand the relationship between the dependent variable, which is other features, and the independent variable, which is whether an employee is promoted or not [8].

Decision trees can easily handle non-linear relationships between features and target variables. It can work with both categorical and numerical data, which is what our data is all about [9]. Decision trees can highlight the most important feature for prediction.

Gradient Boosting Classifier is used for its excellent performance in classification tasks [10]. A gradient-boosting classifier is used for its ability to deal with large datasets and its ability to handle imbalanced datasets.

Random Forest is an ensemble learning method that combines many decision trees [11]. It is used for prediction accuracy and to handle non-linear relationships, as well as for its robustness against overfitting. It is also selected for its ability to handle categorical and continuous data [12].

3.4.1 Hyperparameter Tuning

Table 2. Hyperparameter Tuning for Each Model Used

Model	Hyperparameter	Tuned Values	Optimal Value
Logistic Regression	Penalty	l2	l2

	Solver	liblinear	liblinear
	C (Inverse of Regularization)	0.01 – 1.0	1.0
	Max Iterations (Epochs)	50 – 200	100
Random Forest Classifier	Number of Estimators (n_estimators)	50 – 200	100
	Max Depth	5 – 20	12
	Min Samples Split	2 – 10	2
	Max Features	auto (sqrt of features)	auto
Gradient Boosting Classifier	Number of Estimators (n_estimators)	50 – 200	100
	Learning Rate	0.01 – 0.2	0.1
	Max Depth	3 – 7	3
	Min Samples Split	2 – 10	2
	Subsample	0.5 – 1.0	1.0
Decision Tree Classifier	Max Depth	5 – 20	12
	Min Samples Split	2 – 10	2
	Criterion	Gini / entropy	Gini

Table 2 reports the main hyperparameters tuned for each of the models, along with a range of values explored and the final optimum value selected based on performance. Therefore, the hyperparameter tuning for every one of the above models tries to find this perfect balancing point between model complexity on one side and performance on the other.

3.5 Model Evaluation Methods and Metrics

The following are the methods and metrics that are used in evaluating the machine learning models:

3.5.1 Methods

- Train-Test Split: The dataset was divided into an 80% training and 20% testing configuration to evaluate the model's performance on new data.
- Actual values: The real values from the dataset tell whether there was a promotion or not for any employee, yes or no.
- Predicted Values The machine learning models, like logistic regression, decision trees, etc., predict whether an employee is promoted or not using the training data.

3.5.2 Performance Metrics:

- Accuracy: The number of correctly classified outcomes, whether positive or negative, out of the total predictions in a dataset is called accuracy. It gives one an overview of how well the model works. The formula for the accuracy is as follows:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

- Precision: Precision is the number of correct positive results over the total number of positive predictions. That means, out of the total predictions of promotion, it tells us how many of them are valid. The formula for precision is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Recall: Recall is the proportion measure of actual positives. Actual positives are the employees who received promotions. It shows the degree to which the model captures all the true promotions. The formula is as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- F1-Score: The F1-Score is the harmonic mean of precision and recall. In this way, it offers a balanced measure of the performance of a model, particularly in situations where class distribution is uneven, such as promotion being less common. The formula is as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The actual values in the dataset were Yes or No to indicate whether an employee was promoted or not. To make predictions regarding the performances of each model, the actual values in the dataset have been used to train models such as logistic regression, decision trees, random forests, and gradient boosting, all of which have been tasked with predicting the outcomes of promotions. A comparison of predicted values generated by each model with the actual values is done [13,14]. The metrics of performance here include accuracy, precision, recall, and F1-score, all calculated based on this comparison.

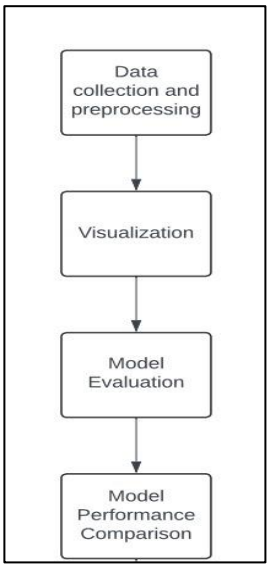

Figure 1. Employee Promotion Evaluation Model Flowchart

Figure 1 illustrates the whole process for employee promotion evaluation and prediction.

4. Results and Discussion

4.1 Correlation Heatmap

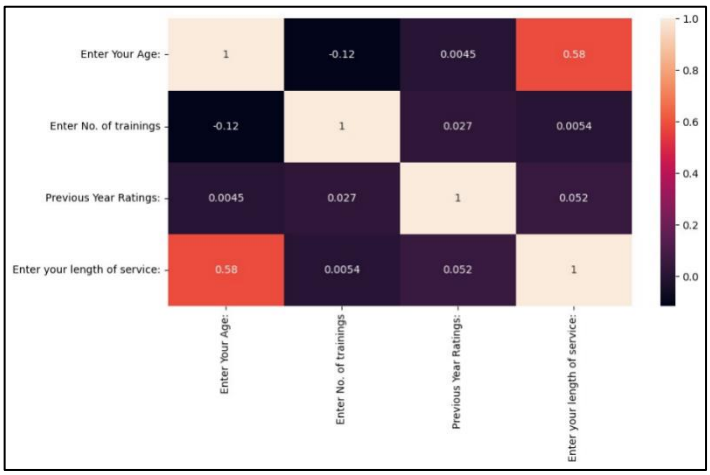

Figure 2. Correlation Heatmap Between Numeric Values

Figure 2 is a correlation heatmap of numeric variables, namely age, the number of training, previous year ratings, and length of service. The coefficients are measured between -1 and 1. The highest positive coefficient of 0.58 exists between age and tenure, and evidence exists that the older employees were those with longer tenures. There is a weak negative correlation (-0.12) of age with training where younger staff might attend more. Ratings of the previous year relate less to any other variable.

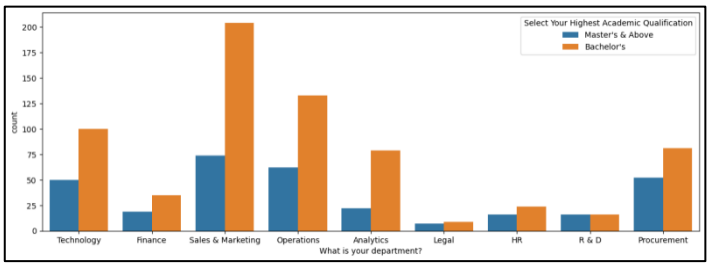

Figure 3. Distribution of Employees by Department and Academic Qualifications.

Figure 3 depicts employee distribution by highest qualification. Most employees from the Sales and Marketing and Operations departments have a BSc, while fewer have an MSc. In R&D and legal departments, though small in strength, practically the number of BSc and MSc graduates is equal.

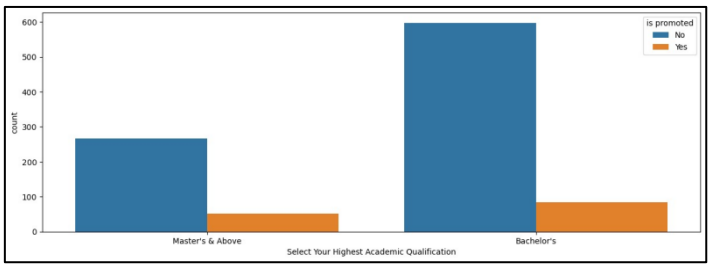

Figure 4. Promotion Status of Employees based on Academic Qualifications.

Figure 4 depicts the relationship between an employee's highest academic qualification and their promotion status. As we can interpret from the graph, a small portion of employees from both qualifications have received promotions, but bachelor's degree holders have fewer chances to receive promotions as compared to their total population.

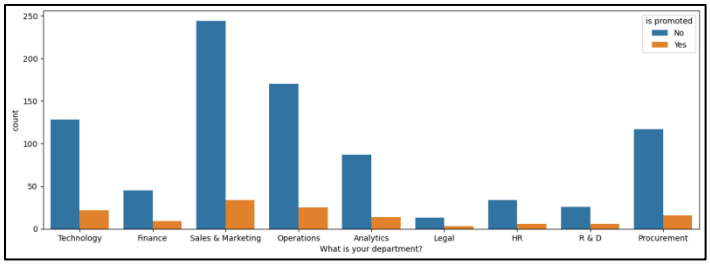

Figure 5. Distribution of Employee Promotions Across Various Departments.

Figure 5 illustrates the distribution of employee promotion across various departments. It can be seen that the sales and marketing department and operations department have the highest count of employees that have been promoted. Legal, HR, and R&D have the lowest count of employees who have been promoted.

```
Accuracy: 0.86
Classification Report:
              precision    recall  f1-score   support

          No       0.86      1.00      0.92       172
         Yes       0.00      0.00      0.00        28

    accuracy                           0.86       200
   macro avg       0.43      0.50      0.46       200
weighted avg       0.74      0.86      0.80       200

Confusion Matrix:
 [[172   0]
 [ 28   0]]
```

Figure 6. Performance Results for Logistic Regression

In Figure 6, logistic regression actually performed very well, achieving an accuracy rate of 86% and a perfect recall of 1.00, meaning no promotions were missed, although it had some overprediction with a precision of 0.86.

```
Accuracy: 0.78
Classification Report:
              precision    recall  f1-score   support

          No       0.87      0.87      0.87       172
         Yes       0.21      0.21      0.21        28

    accuracy                           0.78       200
   macro avg       0.54      0.54      0.54       200
weighted avg       0.78      0.78      0.78       200

Confusion Matrix:
 [[150  22]
 [ 22   6]]
```

Figure 7. Performance Results for Decision Tree

In Figure 7, the decision tree showed a moderate accuracy of 78% and a precision of 0.87, but a lower recall and a tendency to overfit.

```
Accuracy: 0.845
Classification Report:
              precision    recall  f1-score   support

          No       0.86      0.98      0.92       172
         Yes       0.00      0.00      0.00        28

    accuracy                           0.84       200
   macro avg       0.43      0.49      0.46       200
weighted avg       0.74      0.84      0.79       200

Confusion Matrix:
 [[169   3]
 [ 28   0]]
```

Figure 8. Performance Results for Random Forest

In Figure 8, random forest performed extremely well at 84.5% precision, 0.98 recall, and an F1-score of 0.92. It handles complex relationships but is computationally intensive.

```
Accuracy: 0.825
Classification Report:
              precision    recall  f1-score   support

          No       0.86      0.95      0.90       172
         Yes       0.11      0.04      0.05        28

    accuracy                           0.82       200
   macro avg       0.48      0.49      0.48       200
weighted avg       0.75      0.82      0.78       200

Confusion Matrix:
 [[164   8]
 [ 27   1]]
```

Figure 9. Performance Results for Gradient Boosting

In Figure 9, the accuracy of gradient boosting [15] was 82.5%, with recall at 0.95 and an F1-score at 0.90, both of which have good predictive power but appear to require good tuning to avoid overfitting.

Table 3. Overall Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86%	0.86	1.0	0.92
Decision Tree	78%	0.87	0.87	0.87
Random Forest	84%	0.86	0.98	0.92
Gradient Boosting	82%	0.86	0.95	0.90

Table 3 contains the overall results of all the machine learning models.

5. Conclusion

Promotion, thus, is an important part of career development and therefore needs valid predictive analysis. The research was conducted by using a primary dataset collected from surveying people and applying machine learning models such as logistic regression, decision trees, random forests, and gradient boosting. The applied machine learning model achieved an

accuracy score of 86% by logistic regression, 84.5% by random forest, 82.5% by gradient boosting classifier, and 77.5% by decision tree classifier.

It can be said that many other features that might significantly affect employee promotion prediction have not been considered. This can be addressed in future by incorporating additional features or creating new features based on specific domain knowledge. The machine learning models, including decision trees, might be prone to overfitting, which could have impacted the generalization of a new dataset. To enhance model performance, cross-validation can be implemented in the future.

References

[1] Şahinbaş, Kevser. "Employee promotion prediction by using machine learning algorithms for imbalanced dataset." In 2022 2nd International Conference on Computing and Machine Intelligence (ICMI),Istanbul, Turkey. IEEE, 2022.1-5

[2] Zhang, Chang, Ting-jie Lv, Chun-hui Yuan, Yuan-yuan Ren, and Shuo Wang. "The influence of demographic characteristics on employee promotion: research based on data mining and game theory." Wireless Communications and Mobile Computing 2020, no. 1 (2020): 8814733.

[3] Жанузаков, М. Б., and Г. Т. Балакаева. "Prediction of Employee Promotion Based on Ratings Using Machine-Learning Algorithms." Bulletin of Abai KazNPU. Series of Physical and mathematical sciences 77, no. 1 (2022): 106-111.

[4] Rubel, Mohammad Rabiul Basher, and Daisy Mui Hung Kee. "Perceived fairness of performance appraisal, promotion opportunity and nurses turnover intention: The role of organizational commitment." Asian Social Science 11, no. 9 (2015): 183-197.

[5] Araki, Shota, Daiji Kawaguchi, and Yuki Onozuka. "University prestige, performance evaluation, and promotion: Estimating the employer learning model using personnel datasets." Labour Economics 41 (2016): 135-148.

[6] Liu, Jiamin, Tao Wang, Jiting Li, Jingbo Huang, Feng Yao, and Renjie He. "A data-driven analysis of employee promotion: the role of the position of organization." In 2019 IEEE international conference on systems, man and cybernetics (SMC),Bari, Italy, IEEE, 2019. 4056-4062.

[7] AGRAWAL, POONAM, and SHIKHA GOYAL. "Employee Promotion Prediction Using Improved Additive Regression Classifier Along with ANN (Artificial Neural Networks)." African Journal of Biological Science, 2024, 3204-3211.

[8] Alqahtani, Fatma Ayed, and Abdulaziz Almaleh. "Analysis and Prediction of Employee Promotions Using Machine Learning." In 2022 5th International Conference on Data Science and Information Technology (DSIT),Shanghai, China. IEEE, 2022. 01-09

[9] Sarker, Ananya, S. M. Shamim, Md Shahiduz Zama, and Md Mustafizur Rahman. "Employee's performance analysis and prediction using K-means clustering & decision tree algorithm." Global Journal of Computer Science and Technology 18, no. 1 (2018): 1-5.

[10] Vishal Balaji, D., and J. Arunnehru. "Predictive analysis on HRM data: determining employee promotion factors using random forest and XGBoost." In Proceedings of International Conference on Deep Learning, Computing and Intelligence: ICDCI 2021, Singapore: Springer Nature Singapore, 2022. 179-189.

[11] Tuhairwe, Jackson, Jonathan Kalibbala Mukisa, and Ahmed A. Al-Absi. "Web-Based Employee Promotion Prediction Using Random Forest Classifier." In International conference on smart computing and cyber security: strategic foresight, security challenges and innovation, Singapore: Springer Nature Singapore, 2023. 459-475.

[12] Asuquo, Daniel E., Uduak A. Umoh, Francis B. Osang, and Edikan W. Okokon. "Performance evaluation of c4. 5, random forest and naïve bayes classifiers in employee performance and promotion prediction." African Journal of Management Information System 2, no. 4 (2020): 41-55.

[13] Bandyopadhyay, Nilasha, and Anil Jadhav. "Churn prediction of employees using machine learning techniques." Technical Journal 15, no. 1 (2021): 51-59.

[14] Zhang, Zhiyuan, Kevin T. McDonnell, Erez Zadok, and Klaus Mueller. "Visual correlation analysis of numerical and categorical data on the correlation map." IEEE transactions on visualization and computer graphics 21, no. 2 (2014): 289-303.

[15]Chinnapan, Natcha, and Waraporn Viyanon. "Employee Promotion Prediction Using Machine Learning." (2023). http://irithesis.swu.ac.th/dspace/bitstream/123456789/2226/%201/gs641130042.pdf

Author's Biography

Nareen Ansari holds a BSc in Information Technology with a CGPA of 9.0 and is currently pursuing a master's in Information Technology. With strong skills in Python programming, SQL, and Machine learning, she is aiming to become a data scientist. Their academic journey is driven by a keen interest in data science and its applications, with a focus on harnessing data to drive insightful decisions and innovations.

Neha Vora is currently pursuing her Ph.D. in Computer Science and holds a master's in computer applications (MCA). She is qualified in NET, SET, and GATE, and brings over 9 years of teaching experience, along with 1 year of industry experience. Her primary research areas include computer vision, image processing, machine learning, object detection, and artificial intelligence.