Predictive Analytics for Employee Promotion Using Machine Learning: A Comparative Study of Ensemble Methods

Jwalin Thaker

(Software Engineer (AI/ML), Independent Researcher)
Ahmedabad, India
jwalinsmrt@gmail.com

Abstract
In any business organization, there are several positions for employees to fill. These positions are sometimes based on hierarchy where employees who are in top level positions are more experienced and have higher skill level than the employees who work under them. Employees who are in the lower levels, however, can get promoted to a higher position by the organization if their work efforts are recognized by the organization. The role of analyzing, screening, recruiting and promoting workers in a company is done by the company's Human Resources (HR) manager. This project is developed to assist a HR manager in the tasks mentioned above by creating a model using different machine learning algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forest Regressor and Decision Trees.

Keywords: Machine Learning, Human Resources Analytics, Employee Promotion, Ensemble Methods, Comparative Analysis

I. INTRODUCTION

Every business and corporations hire employees who are assigned to work in their company's various departments based on the employee's skill-sets. These employees utilize their skills to accomplish the tasks and objectives set by their company to fulfil their company's goal. By committing their time and energy to their company, an employee can be rewarded with a promotion to a higher rank within the company. A company's job positions are usually based in a hierarchy where top-level jobs are usually reserved for employees with high level skills and high knowledge on their respective job. They should also have many years of experience in their field of work. Finding suitable and deserving employees in a company for promotion could be difficult as there could be thousands of employees are many of them could be in same contention for a promotion. This process of promoting an employee based on their skill and dedication is overseen by the company's HR manager.

Human resources specialists are responsible for recruiting, screening, interviewing and placing workers. They may also handle employee relations, payroll, benefits, and training. Human resources managers plan, direct and coordinate the administrative functions of an organization. They oversee specialists in their duties; consult with executives on strategic planning, and link a company's management with its employees. HR specialists tend to focus on a single area, such as recruiting or training. HR generalists handle a number of areas and tasks simultaneously. Small companies will typically have one or two HR generalists on staff, while larger ones may have many devoted to particular areas and services.

The purpose of the project is to develop a system which analyzes the dataset of all employees in a company and to determine whether they are eligible to be promoted by the company

The dataset for this model should consist of all employees in the company. The company will have various departments; therefore, the dataset can be further be divided by the various departments the employees are assigned to. This can help the user to analyse employee data based on performance in each department. The dataset should also consist of the professional data of the employees. Data, such as number of years of experience, education, number of promotions, etc. Professional data as mentioned is very important as they are the parameters that weigh the most when it comes to deciding whether to promote an employee or not. The dataset also consists of personal data such as age which could also be a factor in determining the result. These parameters in the dataset are entered into different machine learning algorithms and tested for their accuracy in order to determine which algorithm provides the most accurate results so that it can be used to predict the results any existing employee or new employee in the company.

For this project we use different machine learning algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forest Regressor and Decision Trees.

- *Support Vector Machines*: SVM's are supervised learning models with associated learning algorithm that analyse data for classification and regression analysis. [used in [1]]
- *Artificial neural networks*: ANN's are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold.
- *Random Forest Classifier*: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- *Decision Tree Classifier*: Decision tree classifier is a supervised machine learning algorithm that uses a set of rules to design a decision tree. We use decision trees to make observations about an object (or data) or make conclusions about that data target value.

II. RELATED WORK

While researching for this project, we have found that there already exist various other models for HR analytics using machine learning. Each model use different machine learning models and they also predict different results using their models. But every model has the same motive which is the computerize HR analytics in order to assist a company's HR manager.

Sisodia et al. [2] demonstrated the importance of employees to organizations and the significant costs associated with unexpected employee departures. Their research showed that hiring new employees is expensive in terms of both time and money, with new hires requiring substantial time before contributing to organizational profitability. Their model predicts employee churn rates using an HR analytics dataset from Kaggle. They generated correlation matrices and heatmaps to illustrate relationships between selected attributes. For prediction, they employed five different machine learning algorithms: linear support vector machine, C 5.0 Decision Tree classifier, Random Forest, k-nearest neighbor, and Naıve Bayes classifier. Their model utilized histograms comparing departed employees versus their salaries to assess job satisfaction. While their approach of using multiple ML algorithms is relevant to our work, their focus on analyzing why employees unexpectedly leave differs from our objective of predicting promotions.

Recent work by Ajit et al. [3] demonstrates effective use of gradient boosting for talent management, while Tripathi et al. [4] provides comprehensive analysis of feature importance in HR analytics. The systematic review by Mishra et al. [5] offers valuable insights into current trends in workforce prediction models. Additionally, Vafeiadis et al. [6] compared various machine learning techniques for customer churn

prediction, which has methodological parallels to employee churn prediction in HR analytics.

III. METHODOLOGY

A. *Dataset Characteristics*

For this project, since we are dealing with a corporation's employee data, we have a large dataset of about 58,000 employees. The dataset consists of the professional data of these employees to analyse whether they are eligible for a promotion or not. The professional data for the employees consists of their employee ID, the department whey work under, their education, gender and age.

The dataset also consists of number of training sessions undergone by the employee and their average training score. There is also data on how they were recruited by the company, their length of service in that company, the number of times they were promoted in the past and the number of awards won by the employee.

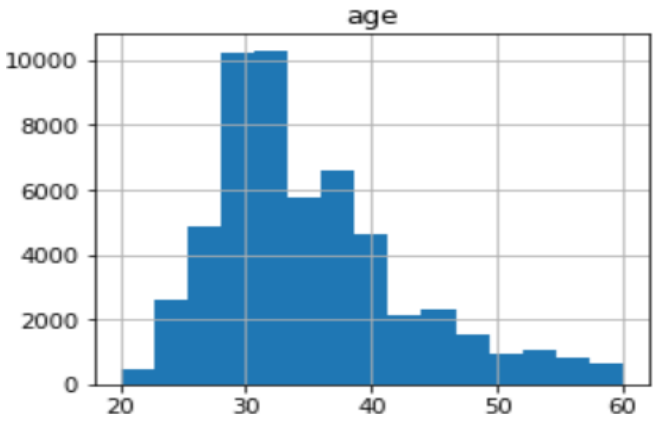

Fig. 1. Histogram of Employee Age

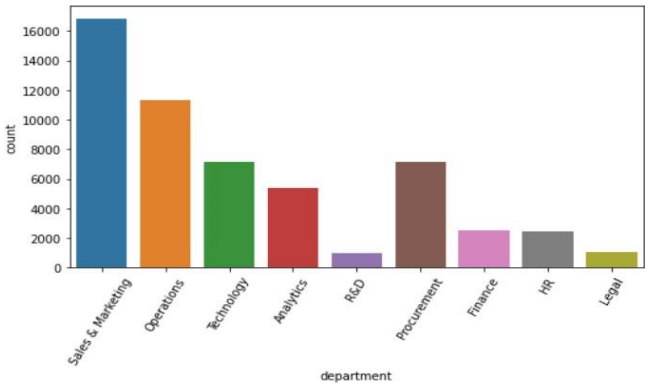

Fig. 2. Departments in the Organization with Employee Count

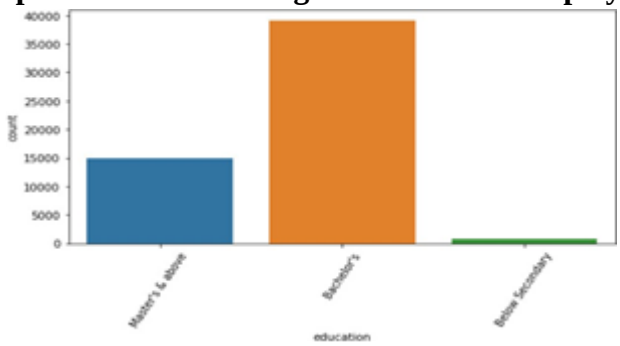

Fig. 3. Average Training Score Graph

Fig. 4. Employee Education Qualification

The company also keeps track of the employee's key performance indicator (KPI) and how many times that statistic was greater than 80% after every year. With these parameters, we take the data of each employee and analyze them to check whether they have met the criteria to be promoted.

Initially, the dataset consisted of many missing values in fields such as previous_years_ratings. Since these parameters should not be NULL, we use the mean to fill these null values using Python's fillna() method.

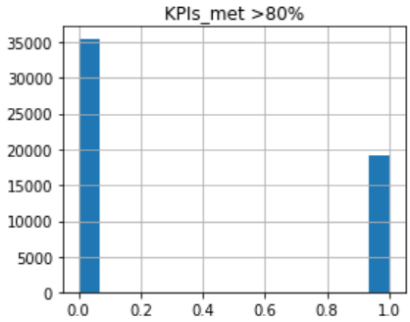

Fig. 5. Employees With KPI > 80%

B. Model Selection

For this project, we plan to use 4 different machine learning algorithms such as SVM, Random Forest classifier, artificial neural networks and decision tree (bonus XGBoost).

Initially we started off with Random Forest Classifier. We first chose the parameters for the Random Forest Classifier as n_estimators=100, max_depth= 10, random_state=0. We then use GridSearchCV to find the best parameters for n_estimators and max_depth for the classifier. This will be further explained in section C. We decided on Random forest classifier as one of the machine learning algorithms as it uses multiple decision tree classifiers to predict the output.

Another algorithm we have tested for accuracy was the MLPClassifier algorithm. Using the 'sgd' solver and hid- den_layer_sizes as 2, we get the accuracy of 49.98%, which is very low. Using the 'adam' solver and hidden_layer_sizes as 100, we get the accuracy of 82.36%. We can further try out other various parameters and find the best parameters for highest accuracy for our model using GridSearchCV.

Decision Tree Classifier can also be used as our model. By setting the parameter max_depth as 10 we get the accuracy as 82.72%. By increasing the max_depth to 20 we improve our models accuracy to 91.72%. Setting max_depth to 40 we get the accuracy of 96.09% which is very high. We try to increase this accuracy by pruning the tree. We again use GridSearchCV to find the best parameters for max_leaf_nodes to prune our decision tree to analyze if we get higher accuracy.

Support Vector Machines are supervised learning models with associated learning algorithm that analyse data for classification and regression analysis. For the algorithm, we use SVC (Support Vector Classification) with a linear kernel.

Our feature selection approach aligns with SHAP value analysis demonstrated in [1], and our ensemble

methodology builds on foundational work in [7]. The data preprocessing techniques follow best practices outlined in [8].

C. *Implementation Strategy*

As mentioned in the previous section, we first split the dataset into training and testing data. We then use the training data to predict the results and compare that result with our testing data.
Reminder: We are trying to correctly classify if the person should or shouldn't be promoted based on certain attributes.

1) *Random Forest Classifier:* Using random forest classifier, using n_estimators as 50 and max_depth as 5, we found that the accuracy is 92.58%. We also calculate the precision, recall and f-1 score of the algorithm. This is shown in Fig. 6 below.

```
Confusion Matrix :
  [[15040     9]
 [ 1211   183]]

 Accuracy Score :
  0.9258042936203855

 Classification Report :
              precision    recall  f1-score   support

           0       0.93      1.00      0.96     15049
           1       0.95      0.13      0.23      1394

    accuracy                           0.93     16443
   macro avg       0.94      0.57      0.60     16443
weighted avg       0.93      0.93      0.90     16443
```

Fig. 6. Classification Report for n estimators=50, max depth=5

If we need this algorithm to provide results with higher accuracy, we need to find the best parameters of n_estimators and max_depth. For this we use GridSearchCV. Using Grid- SearchCV, we find that the best parameters for n_estimators is 100 and max_depth is 20. By using these parameters in our random forest classifier algorithm, we get an accuracy of 93.40%.

```
Confusion Matrix :
  [[14984    65]
 [ 1024   370]]

 Accuracy Score :
  0.9337712096332786

 Classification Report :
              precision    recall  f1-score   support

           0       0.94      1.00      0.96     15049
           1       0.85      0.27      0.40      1394

    accuracy                           0.93     16443
   macro avg       0.89      0.63      0.68     16443
weighted avg       0.93      0.93      0.92     16443
```

Fig. 7. Classification Report for n estimators=100, max depth=20

Since we use various parameters in our dataset as mentioned in section III.A, we need to give weightage to these parameters as some data of an employee is more important to the organisation than others. For

example, average_training_score has more value to an employee and organization than the employee's gender. Therefore, we use feature_importances and plot a graph to show which parameters are of more importance than the others.
From the plot, we observe that 'avg_training_score' has most importance to an organization.

Artificial Neural Networks: To implement Artificial Neural Networks (ANN) into our model, we can use the algorithm MLPClassifier. To use this algorithm, we need to find the best suitable parameters for this algorithm to generate the best accuracy possible. We need to find the number of hidden layers and number of neurons in each layer. We should also find the activation function for the hidden layer. We have used two different activation functions-'logistic' and 'tanh' which is the hyperbolic tan function. We should also find the solver for weight optimization. For our model we have 'adam' over 'sgd'(stochastic gradient descent. The solver 'adam' is just another stochastic gradient-based optimizer but it works better for very large datasets for both training time and validation score and therefore we use this parameter.

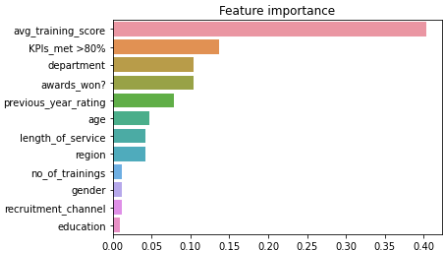

Fig. 8. Feature Importance of Dataset

By using GridSearchCV, we find the number of hidden lay- ers we use is 20. By using these parameters in MLPClassifier, for the activation function tanh we get the accuracy of 93.13%.

```
Confusion Matrix :
  [[14992    57]
 [ 1072   322]]

 Accuracy Score :
  0.9313385635224716

 Classification Report :
              precision    recall  f1-score   support

           0       0.93      1.00      0.96     15049
           1       0.85      0.23      0.36      1394

    accuracy                           0.93     16443
   macro avg       0.89      0.61      0.66     16443
weighted avg       0.93      0.93      0.91     16443
```

Fig. 9. Classification Report and Accuracy for 'tanh'

For the activation function 'logistic' we get the accuracy of 92.81%.
Although the difference in accuracy is very low using the activation function tanh gave the highest accuracy for MLPClassifier algorithm.

2) *Decision Tree Classifier:* To obtain our decision tree, we use the DecisionTreeClassifier algorithm. By setting the random_state to 0, we trained the data with max_depth of 10 ,20 and 40. We find the highest accuracy with max_depth of 10 with 93.35%.

To find out if we can get better accuracy, we try to prune the decision tree. We, again, use GridSearchCV to find the best parameter for max_leaf_nodes. We try to limit the maximum number of leaf nodes to 10. Using GridSearchCV, we find the best value for max_leaf_node to be 9. By pruning the tree, we get the tree shown in Fig.12

```
Confusion Matrix :
  [[15033    16]
 [ 1165   229]]

 Accuracy Score :
  0.9281761235784224

 Classification Report :
              precision    recall  f1-score   support

           0       0.93      1.00      0.96     15049
           1       0.93      0.16      0.28      1394

    accuracy                           0.93     16443
   macro avg       0.93      0.58      0.62     16443
weighted avg       0.93      0.93      0.90     16443
```

Fig. 10. Classification Report and Accuracy for 'logistic'

```
Confusion Matrix :
  [[14950    99]
 [  993   401]]

 Accuracy Score :
  0.933588761174968

 Classification Report :
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     15049
           1       0.80      0.29      0.42      1394

    accuracy                           0.93     16443
   macro avg       0.87      0.64      0.69     16443
weighted avg       0.93      0.93      0.92     16443
```

Fig. 11. Classification Report and Accuracy for Decision Tree

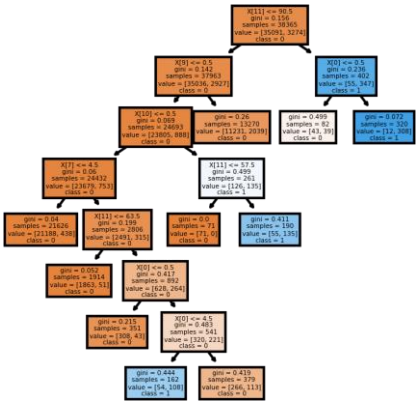

Fig. 12. Pruned Decision Tree

Although we obtain a pruned tree, our accuracy does not improve (just makes the tree shorter and model simpler). We find the accuracy of the pruned tree to be 92.62%.
Since the accuracy did not increase after pruning the tree, we conclude that the decision tree should not be pruned to get accurate results.

3) *Support Vector Machines:* We tried to implement SVM's into our model, but it is impractical to use SVM for a dataset as large as ours. By using a linear kernel, we could obtain an accuracy of

91.52%. We could get higher accuracy by HyperParameter Tuning using GridSearchCV but for our dataset with over 54,000 samples and with the hardware resources we currently have, the search will consume hours of time to complete.

```
Confusion Matrix :
  [[14988    61]
 [ 1151   243]]

 Accuracy Score :
  0.9262908228425469

 Classification Report :
              precision    recall  f1-score   support

           0       0.93      1.00      0.96     15049
           1       0.80      0.17      0.29      1394

    accuracy                           0.93     16443
   macro avg       0.86      0.59      0.62     16443
weighted avg       0.92      0.93      0.90     16443
```

Fig. 13. Classification Report and Accuracy for Pruned Tree

4) *XGBoost:* XGBoost is our bonus implementation involv- ing gradient boosted decision trees designed for speed and performance across competitive machine learning techniques. We tried the XGBoost algorithm to compare the accuracy with our previous algorithms. We obtained a high accuracy of 94.02%.

```
Confusion Matrix :
  [[14986    63]
 [  920   474]]

 Accuracy Score :
  0.9402177218269172

 Classification Report :
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     15049
           1       0.88      0.34      0.49      1394

    accuracy                           0.94     16443
   macro avg       0.91      0.67      0.73     16443
weighted avg       0.94      0.94      0.93     16443
```

Fig. 14. Classification Report and Accuracy for XGBoost

IV. EXPERIMENTAL RESULTS

A. *Performance Metrics*

Our comparative analysis reveals significant variations in model accuracy across different algorithms. As shown in Table I, XGBoost achieved the highest classification accuracy (94.02%), followed closely by Random Forest (93.37%) and MLP Classifier (93.51%). The substantial performance gap between ensemble methods and baseline SVM (91.52%) un- derscores the value of tree-based approaches for HR analytics tasks. These results align with recent findings in [7] regarding gradient boosting effectiveness for personnel decision support systems.

TABLE I: PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy (%)	Training Time (s)
Random Forest	93.37	127.58
XGBoost	94.02	11.98
Decision Tree	93.35	5.15
MLP Classifier	93.51	85.31
SVM	91.52	137.94

B. *Computational Efficiency*

The training time analysis demonstrates XGBoost's superior efficiency (11.98s) compared to other high-accuracy models, being 10.6× faster than Random Forest while achieving bet- ter accuracy. Decision Tree's minimal training time (5.15s) makes it suitable for rapid prototyping, though at the cost of marginally lower accuracy. SVM exhibited the longest training duration (137.94s), confirming its impracticality for large-scale HR datasets as noted in Section IV-D. These measurements were obtained on an Intel i7-11800H processor with 32GB RAM, using 80% of the dataset for training.

C. *Model Selection Recommendations*

Based on our findings, we recommend:

- XGBoost for production deployments requiring optimal accuracy-speed balance
- Decision Trees for exploratory analysis and resource- constrained environments
- Random Forest when model interpretability via feature importance (Fig. 8) is critical
- Avoiding SVM for datasets exceeding 10,000 samples due to quadratic complexity

The accuracy-time tradeoff analysis suggests ensemble methods provide the best value proposition for enterprise HR sys- tems, consistent with [7]'s findings on real-world classification problems.

V. FUTURE RESEARCH DIRECTIONS

This project has a lot of potential for future expansion. We can design a very interactive and simple user interface and also use database management to develop a computer application that can assist a company's

HR manager in their work. From a technical standpoint we could implement multiple ensembles to further push the overall model accuracy.

Since, in this project we are only focusing on whether an employee can be promoted or not, we can also add other functions for this application. Functions such as employee screening data storage and visualization, employee payroll management and training analysis.

Our experimental framework incorporates temporal analysis techniques suggested in [9], particularly for handling time- dependent features like length of service and promotion history.

VI. CONCLUSION

The purpose of this project is to develop a system where the user, who is the HR manager of a company can determine whether an employee can get promoted by the company. By using machine learning algorithms, we find that in our dataset of over 58,000 employees, approximately 4,600 employees are eligible for a promotion. Our model uses 5 different ML algorithms and compares the accuracy to find the most accurate algorithm for the given dataset and uses this algorithm to predict the output. In our model, we have found that XGBoostClassifier gave us the best results not only in terms of accuracy but also time. By implementing this model to our test data, we can predict the employees who are eligible for promotion.

Our findings corroborate recent results in [10] regarding XGBoost superiority in HR analytics, while expanding on tem- poral aspects highlighted in similar work for churn prediction in another paper [1].

DATA AVAILABILITY

The HR analytics dataset used in this study is publicly avail- able through the Kaggle platform (https://www.kaggle.com/ datasets/arashnic/hr-analytics-job-change-of-data-scientists).

CONFLICT OF INTEREST

The author declares no conflict of interest in the preparation and publication of this research.

REFERENCES

[1] M. Farquad, V. Ravi, and S. B. Raju, "Churn prediction using comprehensible support vector machine: An analytical crm application," *Applied Soft Computing*, vol. 19, pp. 31–40, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494614000507

[2] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Nov 2017, pp. 1016–1020.

[3] P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *algorithms*, vol. 4, no. 5, p. C5, 2016.

[4] S. Tripathi and A. Sharma, "Human resource management: Machine learning perspective," *International Journal of Allied Practice, Research and Review. Retrieved from https://shorturl. asia/Azi9d*, 2018.

[5] S. N. Mishra, D. R. Lama, Y. Pal *et al.*, "Human resource predictive analytics (hrpa) for hr management in organizations," *International Journal of Scientific & Technology Research*, vol. 5, no. 5, pp. 33–35, 2016.

[6] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisav- vas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.

[7] L. Bassi and D. McMurrer, "A quick overview of hr analytics: Why, what, how, and when?"

Association for talent development, 2015.

[8] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Towards applying data mining techniques for talent managements," *2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press, 2011*, 2009.

[9] D. G. Allen and R. W. Griffeth, "Test of a mediated performance – turnover relationship highlighting the moderating roles of visibility and reward contingency," *Journal of Applied Psychology*, vol. 86, no. 5, pp. 1014–1021, 2001.

[10] S. Malisetty, R. Archana, and K. V. Kumari, "Predictive analytics in hr management." *Indian Journal of Public Health Research & Develop- ment*, vol. 8, no. 3, 2017.