Automation in Job Promotion Processes through Machine Learning Model

Jun Chuan Quah[1*], Muhammad Azhar Aizad Asfarizailin[2], Justin Yuen Phin Lai[3], Nur Qistina Imani[4], Sharifah Nurul Izzah[5], and Vimala Balakrishnan[6]
[1] Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Federal Territory of Kuala Lumpur, Malaysia
[1]22004851@siswa.um.edu.my
[2]u2100687@siswa.um.edu.my
[3]s2172692@siswa.um.edu.my
[4]u201068@siswa.um.edu.my
[5]u2100665@siswa.um.edu.my
[6]vimala.balakrishnan@um.edu.my

Abstract – This research aims to develop a machine learning classification model to automate job promotion and enhance the transparency and efficiency of talent management processes. The dataset used was retrieved from Analytics Vidhya where it consists of 54,808 rows and 14 columns, including the target variable "is_promoted." A total of four classification models were built: Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors. Random Forest was selected as the most suitable model as it shows the highest accuracy, precision, recall, and F1 score across all the models, making it suitable for classifying non-promoted and promoted employees based on their performance and abilities. Additionally, a dashboard was built to facilitate the visualization of the company's employee performance. The insights generated from this research offer valuable guidance for HR professionals seeking to adopt an automated merit-based advancement system for promotion practices within their organizations.

Keywords – Human resource, job promotion, automation, classification, visualisation

1 Introduction

The Human Resource (HR) department plays a crucial role in managing the workforce of an organization. Typically, HR is involved in various operations such as recruitment, employee training and development, fostering relationships with employees, and conducting performance appraisals. A key aspect of HR's mandate includes overseeing promotions within the company, as promotions and demotions can have a significant impact on morale, retention, and productivity. According to Indeed, an employee who has been promoted and changes jobs may expect an average pay raise between 10% and 20% [1].

Position promotion is an integral part of career development, as it rewards workers with more responsibilities and better pay. Based on organizational policies, these promotion-based decisions are taken on different aspects. These can be the length of service, experience, seniority, performance, etc [2]. However, the challenge of identifying high-potential employees based on their performance and abilities is exacerbated by the presence of biases in the Malaysian working environment. This scenario often results in discrimination and prejudice persisting between employees and higher management with different ethnicities, religions, skin colours, and gender. This bias can lead to negative outcomes, including a toxic culture, reduced employee satisfaction, and a high turnover rate. The repercussions of this prejudice are significant, as the issue not only affects the employees themselves but also impacts the entire enterprise. To further support this statement quantitatively, the State of Discrimination Survey Malaysia conducted by Architects of Diversity in 2023 revealed that discriminative experiences among Malaysians ranked second (30%) during the job search process and third, with 29% experiencing discrimination at work. These findings underscore the prevalence of discrimination and its detrimental impact on fostering toxic working environments in our multicultural country [3].

Through this research, we hope to enhance the efficiency and transparency of talent management processes by automating the promotion prediction process to identify employees most likely to be promoted. Specifically, we aim to develop a classification model to identify potential workers for position promotion based on their abilities and performances using machine learning algorithms. The research also aims to identify the most important factors that contribute to position promotion for the company. We hope to mitigate the impact of discrimination and biases in the Malaysian workforce, fostering a more inclusive and merit-based system for talent recognition and advancement.

2 Methodology

2.1 Data Collection

The dataset was retrieved from a public datahack contest held on the Analytics Vidhya website [4]. It contains various features related to employee details, performance, training evaluations and the state of whether the employee is recommended for promotion as the target variable. The dataset has 54,808 rows and 13 columns, providing a large sample size for robust statistical analysis and model training.

2.2 Data Cleaning

The dataset was cleaned by checking for null values and any duplicated rows with Python. Two columns were identified with null values: "previous_year_rating" and "education." To handle these, the "previous_year_rating" column was filled with the value 0, indicating that the employee is working for the first year and does not have any previous year rating. For the "education" column, a clustering algorithm was employed. Specifically, Principal Component Analysis (PCA) was applied for dimensionality reduction on all numerical features and then K-means clustering was used to impute the missing values in the "education" column.

2.3 Data Visualisation

We utilized PowerBI to visualize employee performance. Its intuitive drag-and-drop interface made it accessible for users of all technical skill levels to create insightful dashboards for easier navigation and comprehensive analysis. Descriptive analysis was conducted on summarized employee statistics, distribution of variables, and examining the relationship between features and the target variable "is_promoted" to analyze employee performance within the company. PowerBI's real-time data connections and dynamic visualizations provided immediate insights into workforce dynamics, allowing HR teams to monitor performance metrics consistently. Through visualization, we had a clearer understanding of the dataset's characteristics which allowed us to identify patterns and trends in employee performance and provide further insights for analytics and model training.

2.4 Data Preprocessing

Several preprocessing steps were carried out using Python Library to prepare the dataset for classification tasks. First, extreme outliers were removed using the z-score method to ensure the robustness of subsequent analyses. Following this, label encoding was applied to the "education" column to map categorical values to numerical representations, while one-hot encoding was implemented on the "recruitment_channel" column to transform categorical data into binary vectors. Feature selection was conducted to enhance the predictive performance of the model by dropping columns deemed irrelevant for the classification task, which are "employee_id", "region", "department", "gender" and "age".

To address the issue of class imbalance in the classification task, two separate dataframes were created to compare the performance of different approaches in handling this issue. Firstly, an oversampled dataframe was generated by up-sampling the class with lower samples using the SMOTE (Syntheticers to Minority Over-sampling Technique) library, specifically by up-sampling class '1' to match the sample size of class '0' (where "0" ref not promoted and "1" refers to promoted). SMOTE works by selecting minority observations that are similar to each other and drawing a line between the examples to create new synthetic samples (Hoffman,

2021).[5] Conversely, an undersampled dataframe was created by down-sampling the class with higher samples using the Pandas library, in which class '0' was down-sampled to match the sample size of class '1'. These two approaches allow for a comparative analysis of handling class imbalance and its impact on classification performance. The dataset will then be split into 80:20 ratio for training and testing.

2.5 Data Modelling

For model training, we implemented four different classification algorithm models using Python for each dataframe which are "Logistic Regression", "Decision Tree", "Random Forest" and "K-nearest Neighbors". The objective was to evaluate the performance of different models on the same task and identify the most suitable model based on their performance metrics generalized on unseen testing data. Besides comparing internally between different models, we also compared the overall performance of classification models on the oversampled and undersampled data to identify any significant differences in the results.

4.5 Model Evaluation

For each model trained on both datasets, we evaluated their performance on testing data using key evaluation metrics such as accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model by calculating the proportion of correctly classified instances (both true positives and true negatives) out of the total instances. It is defined as:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Instances} \quad (1)$$

Precision measures the accuracy of the positive predictions by calculating the proportion of true positive instances out of all instances predicted as positive. It is defined as:

$$Precision\ = \frac{True\ Positive}{Total\ Positive + False\ Positive} \quad (2)$$

Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify all relevant instances (true positives) out of the actual positives. It is defined as:

$$Recall\ = \frac{True\ Positive}{Total\ Positive + False\ Negative} \quad (3)$$

F1-score is the harmonic mean of precision and recall, providing a balance between the two. It is useful when you need to take both false positives and false negatives into account. It is defined as:

$$F1\ score\ = \frac{Precision \times Recall}{Precision + Recall}\ (4)$$

By comparing the results of each model across both dataframes, we aimed to identify the model with the best performance for further hyperparameter tuning. This comprehensive evaluation process enabled us to make informed decisions regarding the selection of the most effective classifier for the classification task at hand. Finally, we visualized the feature importance of the tuned model to identify the most significant factors contributing to position promotion within the company.

3 Results and Discussion

3.1 Visualisation on Employee Performance

Fig. 1. PowerBI Dashboard for Overview Performance

Based on the visualization in Fig. 1, it's evident that 8.52% of the 54.81K employees across all departments and channels have received promotions. The average training score is 63.39 out of 100, indicating moderate effectiveness of training programs for the company. The average previous year rating is 3.08 out of 10, which is relatively low, suggesting room for improvement compared to the recent training score. Most of the employees are in the 30-40 age range, with fewer employees in the 20-30 and 50+ age ranges. Most employees have a length of service between 0-10 years, and the count of promotions is also higher within this range. This infers that new employees are more easily promoted. Additionally, employees with higher education levels (Bachelor's and Master's) are more likely to be promoted as they likely possess more skills and knowledge in their respective fields. Moreover, employees who meet 80% or more of their KPIs have a higher chance of being

promoted, highlighting the importance of KPI performance in promotion decisions. The ratio of awarded employees promoted to awarded employees not promoted is higher than the ratio of non-awarded employees promoted to non-awarded employees not promoted. This suggests that receiving awards is a strong indicator of promotion likelihood. The dashboard can also be filtered by department or recruitment channel using the slicer above.

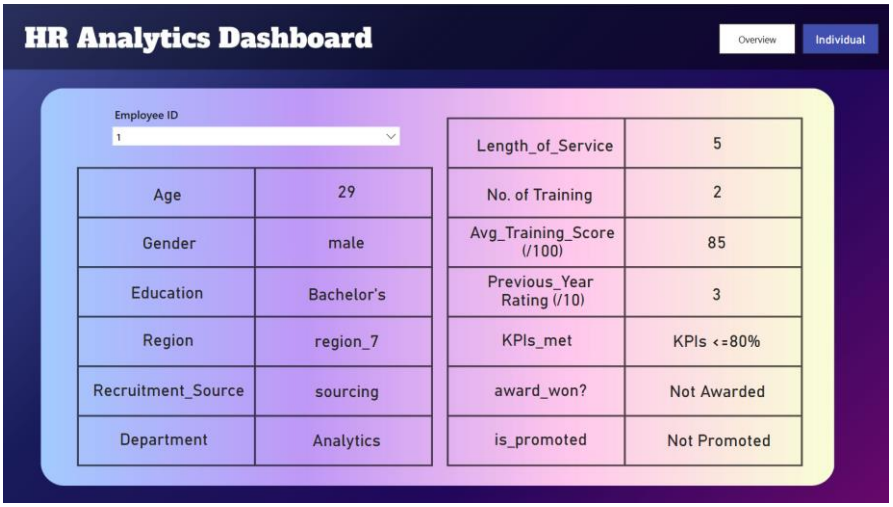

Fig. 2. PowerBI Dashboard for Individual Performance

In addition to providing an overview of the performance of all employees, there is also a dashboard interface for HR to view the details and performance of an employee by searching using their unique employee ID, as shown in Fig. 2.

3.2 Evaluation of Classification Model

The initial exploration of the dataset revealed a significant imbalance in the 'is_promoted' column, with promoted employees (indicated by '1') being notably underrepresented. This imbalance necessitates the use of sampling techniques to address the disparity. Both oversampling and undersampling techniques have been applied to the dataset to manage the imbalance. The four machine learning algorithms have been evaluated on both the oversampled and undersampled datasets to determine the most effective method for improving prediction accuracy.

Combined Evaluation Metrics Table:

	Oversampled				Undersampled			
	Accuracy	Precision	Recall	F1_Score	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.7125	0.7140	0.7123	0.7132	0.7253	0.7166	0.7420	0.7291
Decision Tree	0.8636	0.8403	0.8990	0.8687	0.6713	0.6787	0.6456	0.6617
Random Forest	0.8688	0.8413	0.9104	0.8745	0.7016	0.6926	0.7209	0.7065
K-nearest Neighbors	0.8328	0.8106	0.8701	0.8393	0.7215	0.7113	0.7420	0.7263

Fig. 3. Comparison Result of Evaluation Metrics on Oversampled and Undersampled Data

Based on the evaluation metrics shown in Fig.3., it is evident that the overall model performance is better on the oversampled data compared to the undersampled data when generalised on unseen testing data. The main factor contributing to this better performance is that oversampling increases the number of instances in the minority class while preserving all the information in the majority class, leading to a balanced dataset. Moreover, oversampling allows for better generalization by utilizing more data in the model and avoids introducing bias against the minority class, which in this case is the promoted employees. This is essential for achieving fair and accurate predictions, as the model is less likely to overlook or misclassify minority class instances.

In contrast, undersampling reduces the number of instances in the majority class, potentially leading to the loss of valuable information. This reduction can be detrimental because it limits the data the model has to learn from, potentially weakening its ability to predict outcomes for the majority class accurately. Additionally, this reduction can introduce bias against the majority class and result in poor performance on new, unseen data in the imbalanced dataset. Furthermore, undersampling can lead to overfitting if the model fails to generalize well to new instances that reflect the true distribution of the data.

Based on the results above, the best-performing model selected among all classifiers is the Random Forest when trained on the oversampled data as it achieves the highest accuracy (0.8688), precision (0.8413), recall (0.9104), and F1 score (0.8745).

Evaluation Metrics Table for Oversampled Data:

	Accuracy	Precision	Recall	F1_Score
Random Forest Classifier	0.8688	0.8413	0.9104	0.8745
Fine-Tuned Random Forest Classifier	0.8706	0.8464	0.9068	0.8756

Fig. 4. Comparison result of evaluation metrics of random forest classifier after tuning

After hyperparameter tuning on the random forest classifier, as shown in Fig. 4. The results improved slightly. The accuracy increased to 0.8706, the precision to 0.8464, and the F1 score to 0.8756 for the fine-tuned Random Forest model.

Evaluation Metrics Table for Oversampled Data:

	Accuracy	Precision	Recall	F1_Score
Random Forest Classifier	0.8688	0.8413	0.9104	0.8745
Fine-Tuned Random Forest Classifier	0.8706	0.8464	0.9068	0.8756

Fig. 4. Comparison result of evaluation metrics of random forest classifier after tuning

3.3 Identify the key factor that contributes to position promotion

We used the feature importance method for the fine-tuned Random Forest model to identify the key factors for promoting employees. The importance of a feature is measured by how much it reduces impurity (e.g., Gini impurity or entropy) when it is used to split a node. The importance scores from all the trees are averaged or summed to get a final importance score for each feature.

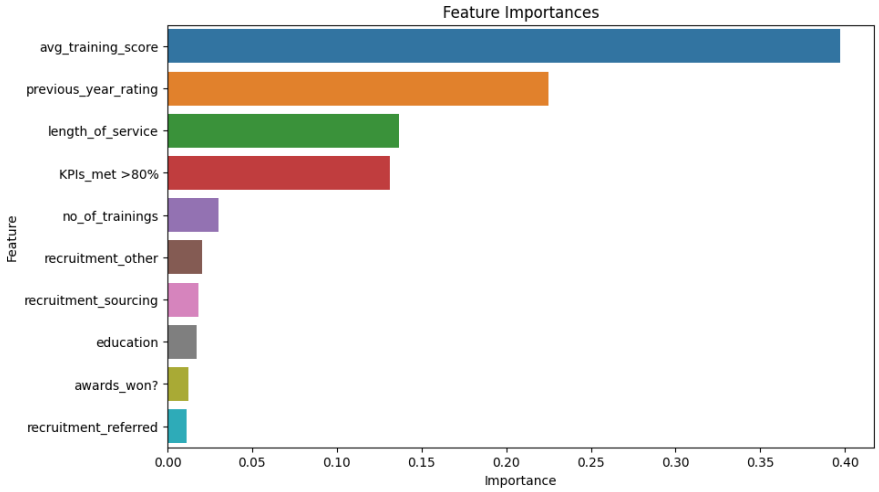

Fig. 5. Feature Importance for Fine-tuned Random Forest Classifier

In Fig.5., it is evident that the average training score, with an importance level of up to approximately 40%, is the major factor that HR considers when deciding on employee promotions. This may be due to the reason that employees with higher training scores have a higher level of skill development and knowledge enhancement, making them more capable in their roles and deserving of position promotion. The ranking continues with employee ratings from the previous year at

22%, followed by the length of employee service at 14%, and employees with higher KPIs at 13%, in descending order of significance, prior to considering any other features.

4 Conclusion

In conclusion, the purpose of this study was to develop a machine learning classification model to automate job promotion decisions. By using the dataset obtained from Analytics Vidhya, we built and evaluated four classification models: Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors. Based on the evaluation metrics, the Random Forest model was selected as the best-performing classifier, particularly when trained on oversampled data. It achieved the highest accuracy (0.8688), precision (0.8413), recall (0.9104), and F1 score (0.8745). The key factors influencing promotion decisions, as identified by the model were the average training score which scores 40% of importance. The employee performance can also be viewed in the PowerBI dashboard to generate useful insights. One limitation of this study is that it relies heavily on historical data, which may contain inherent biases that the model could inadvertently learn and propagate. Despite this limitation, the insights generated offer valuable guidance for HR professionals aiming to adopt an automated, merit-based system for promotion practices within their organizations.

References

1. Bert, J. (2024, February 2). *How To Effectively Negotiate a Promotion Salary Increase*. Indeed. https://www.indeed.com/career-advice/pay-salary/negotiate-promotion-salary
2. Barman, J. P. (2024, April 1). *Employee Promotion: The Types, Benefits, & Whom to Promote*. Vantage Circle. https://blog.vantagecircle.com/employee-promotion/
3. STATE OF DISCRIMINATION SURVEY MALAYSIA 2023. (2023, September). Persatuan Pendidikan Diversiti. https://www.aodmalaysia.org/_files/ugd/15355c_9d8e35607f5a440995263f255a2b0e93.pdf
4. HR Analytics. (2019, January 28). Analytics Vidhya. https://datahack.analyticsvidhya.com/contest/wns-analytics-hackathon-2018-1/#ProblemStatement
5. Hoffman, K. (2021, February 14). *Machine Learning: How to Handle Class Imbalance.* Medium. https://medium.com/analytics-vidhya/machine-learning-how-to-handle-class-imbalance-920e48c3e970