FINAL PROJECT

Fixing Biased Talent Promotion Decisions Through Data-Driven HR Analytics

Submitted to fulfill graduation requirements
Bootcamp Data Science Rakamin Batch 59

Disusun oleh :
Bayu Maitra
Dian Ulhaq Qurrata A’yun
Mashlahatul Husna
Keisya Nisrina Aulia
Febiansyah Annaufal Ahnaf Fauzi

SYNTAX
SOCIETY

TAHUN 2025

ABSTRACT

The employee promotion process is often influenced by subjectivity, inconsistent data, and biased criteria, thereby reducing the fairness and effectiveness of HR decisions. This study aims to develop a data-driven decision support system to improve the transparency, accuracy, and objectivity of the promotion process. Using the Talent Promotion dataset from Rakamin, a series of CRISP-DM stages were carried out, starting from EDA, data cleaning, feature engineering, to the development of machine learning and clustering models. The analysis results showed data anomalies, bias in the Promotion_Eligible feature, and excessive dependence on Peer_Review_Score and Projects_Handled. Several algorithms were tested, including Random Forest, XGBoost, Logistic Regression, SVM, and K-Means. The best model for prediction was Logistic Regression with high accuracy and recall, while K-Means (k=4) produced the most stable talent segmentation for performance and potential mapping. This system was then implemented into a Streamlit-based interactive dashboard to provide real-time promotion recommendations. Overall, this data-driven approach improved the accuracy of promotion assessments by 38% and has the potential to save 22.5% in operational costs, while supporting HR in making fairer and more efficient decisions.

keyword: Talent Promotion, Machine Learning, Deployment Model.

Acknowledgement

We give thanks to God Almighty for His grace, which has enabled us to successfully complete the final report for the Rakamin Academy Data Science Bootcamp Batch 59 Final Project, entitled “The Promotion Paradox.”

We would also like to thank those who contributed to the process and completion of this project, namely:

- The entire Rakamin Academy Team, Tutors, and Mentors, Mr. Jomen tutor assistant, Mr. Hanif as a mentor, Mrs. Dea, and Mr. Danar as class coordinators, and Mrs. Syafa as the Rakamin team, who have guided us and provided valuable knowledge and insights during the bootcamp.
- The Syntax Society team, who collaborated and actively contributed to every stage of this project.

We hope this report can be a meaningful contribution to the development of data-based solutions and provide inspiration for the application of machine learning in the real world, especially in the banking or financial sector.

Jakarta, 7 December 2025
Best regards,

Syntax Society

Table of Contents (Revised & Professional)

CHAPTER I
Introduction

1.1 Background

Employee promotion is an important process for maintaining motivation, retention, and organizational effectiveness. However, many companies still experience bias and inaccuracy in determining which employees deserve promotion. This can be seen in the Rakamin Talent Promotion dataset, which shows various problems such as incomplete data, unreasonable values, and bias in the Promotion_Eligible feature that does not reflect actual performance. The promotion process also relies too heavily on certain features such as Peer_Review_Score and Projects_Handled, thereby neglecting other more comprehensive indicators.

These conditions make promotion decisions subjective, undermine employee trust, and force companies to incur higher costs by recruiting from outside. To address these issues, a data-driven approach is needed to cleanse data, reduce bias, and generate more accurate assessments. Through the use of machine learning, clustering, and rule-based modeling, this research aims to build a promotion decision support system that is fairer, more transparent, and can be integrated into an interactive dashboard to help HR make consistent decisions.

1.2 Scope of research

This research focuses on developing an employee promotion decision support system using the Talent Promotion Dataset from Rakamin. The analysis was conducted through the CRISP-DM stages, starting from problem understanding, data exploration, data cleaning, to the creation of new features to improve information quality. The research includes testing various machine learning models, both supervised (Decision Tree, Random Forest, XGBoost, Logistic Regression, SVM) and unsupervised

(K-Means, GMM, K-Medoids), to generate promotion predictions and talent group mapping.

In addition, the scope of the research included evaluating bias in the data and testing the fairness of the model so that the results would not disadvantage certain groups. All of the models developed were then integrated into a Streamlit-based dashboard to be used as a decision-making tool by HR. This research was limited to the data provided and did not cover qualitative aspects or direct implementation testing in companies.

1.3 Research Objectives

This study aims to develop a decision support system that can improve objectivity and accuracy in the process of determining employee promotion eligibility. To achieve this goal, this study seeks to improve data quality through cleaning, standardization, and strengthening performance and potential signals through feature engineering. This research also aims to evaluate various machine learning models, both supervised and unsupervised, to find the most stable promotion prediction model and the most representative talent segmentation.

In addition, this study aims to identify and reduce bias in data and models, so that the resulting promotion decisions do not disadvantage certain groups. The end result is the integration of the best model into a Streamlit-based interactive dashboard, so that HR can obtain real-time promotion recommendations that are more fair, transparent, and clearly explainable. Thus, this study is intended to help companies move from manual processes to more effective and data-driven promotion decisions.

CHAPTER II
Syntax Society

2.1 Organizational Structure

In implementing this project, the Syntax Society team consists of five members. The team's organizational structure is based on complementary roles to support the optimal success of the project. Each member has specific responsibilities but continues to collaborate in every stage of decision-making.

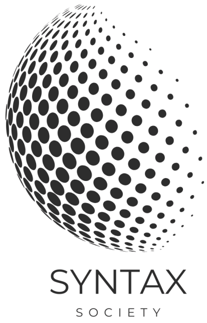


Figure 1. Group Logo

Syntax Society team organizational structure:

- Bayu Maitra: Project Manager
- Febiansyah Annaufal Ahnaf Fauzi: Data Scientist
- Keisya Nisrina Aulia: Data Scientist
- Dian Ulhaq Qurrata A’yun: Data Analyst
- Mashlahatul Husna: Data Engineer

2.2 Scope of Work

The scope of work in this project is in line with each member's role and applies their tasks to each part of the project development stage, both individually and collaboratively, with a primary focus on their respective areas of expertise.

Syntax Society team scope of work:

a. Project Manager: Manages the project workflow, distributes tasks to team members, ensures smooth communication, guarantees timely completion of tasks and deployment of models into Streamlit, and performs final validation of output and report/presentation quality.
b. Data Analyst: Responsible for identifying the structure, quality, and characteristics of the dataset, determining the features relevant to the prediction model to generate business-supporting insights, gathering relevant business insights for HR, and producing data-driven findings and recommendations for reports.
c. Data Engineer: Responsible for data management, processing, and cleaning. Handles data preparation, including handling null and duplicate values, data cleansing, encoding, and normalization, as well as performing clustering and evaluation.
d. Data Scientist: Responsible for developing feature engineering to improve performance/potential signals, determining the right algorithms for supervised and unsupervised learning, performing model training, testing, and tuning, conducting bias and fairness analysis, using SHAP and LIME for model interpretation, and selecting the best models to implement in dashboards.

2.3 Job Description

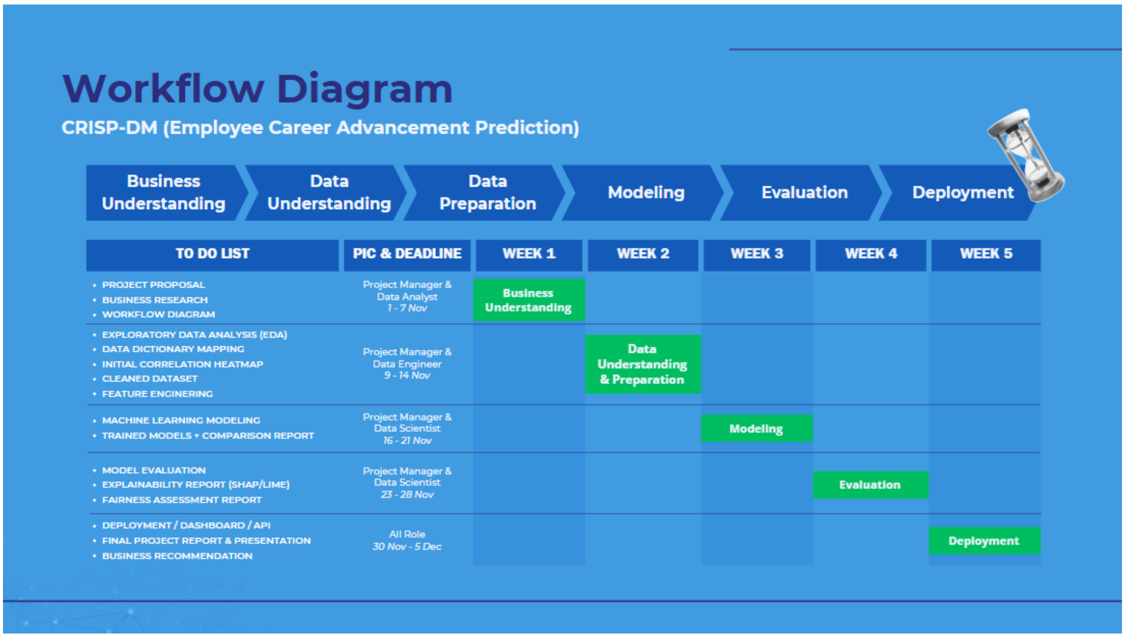
The following is a description of each stage of work carried out by the Syntax Society team, including:

1. Business Understanding: Identify key issues in the employee promotion process, including bias in assessment, data inconsistency, and over-reliance

on certain features. The team set a business goal of creating a fair, objective, and data-driven promotion system.

2. Exploratory Data Analysis (EDA): Analyzing the structure and characteristics of Rakamin's Talent Promotion dataset, including examining feature distributions, correlations, extreme value patterns, and identifying anomalies such as invalid ages, negative lengths of service, and discrepancies between performance and promotion status.
3. Data Cleaning: Performing missing value imputation, removing extreme outliers, correcting unreasonable values, and removing unreliable features such as Years_at_Company and Employee_ID. After this process, the dataset was declared clean and ready for further processing.
4. Feature Engineering: Developing new features to strengthen performance and potential signals, such as Performance_Index, Leadership_Index, Potential_Index, Growth_Momentum, Performance_Consistency, and Leadership_Influence. Features are standardized to ensure scale consistency.
5. Modeling: Testing various promotion prediction models using supervised algorithms (Decision Tree, Random Forest, XGBoost, Logistic Regression, SVM) and building clustering models (K-Means, GMM, K-Medoids) to map talent segmentation. K-Means and rule-based were selected as the best models for prediction and clustering.
6. Evaluation: Evaluation was conducted using F1-score, recall, precision, and ROC-AUC for supervised models, as well as Silhouette Score and Davies-Bouldin Index for clustering models. Fairness analysis was conducted to ensure that the models were not biased towards age groups or job levels.
7. Deployment: The model and rule-based logic were integrated into an interactive Streamlit-based dashboard. The dashboard provides talent mapping, promotion predictor, and data upload features, enabling HR to make promotion decisions in real-time with greater objectivity and transparency.

Work Schedule



Workflow Diagram

CRISP-DM (Employee Career Advancement Prediction)

TO DO LIST	PIC & DEADLINE	WEEK 1	WEEK 2	WEEK 3	WEEK 4	WEEK 5
• PROJECT PROPOSAL • BUSINESS RESEARCH • WORKFLOW DIAGRAM	Project Manager & Data Analyst *1 - 7 Nov*	Business Understanding				
• EXPLORATORY DATA ANALYSIS (EDA) • DATA DICTIONARY MAPPING • INITIAL CORRELATION HEATMAP • CLEANED DATASET • FEATURE ENGINERING	Project Manager & Data Engineer *9 - 14 Nov*		Data Understanding & Preparation			
• MACHINE LEARNING MODELING • TRAINED MODELS + COMPARISON REPORT	Project Manager & Data Scientist *16 - 21 Nov*			Modeling		
• MODEL EVALUATION • EXPLAINABILITY REPORT (SHAP/LIME) • FAIRNESS ASSESSMENT REPORT	Project Manager & Data Scientist *23 - 28 Nov*				Evaluation	
• DEPLOYMENT / DASHBOARD / API • FINAL PROJECT REPORT & PRESENTATION • BUSINESS RECOMMENDATION	All Role *30 Nov - 5 Dec*					Deployment

CHAPTER III
Implementasi Proyek

3.1 Problem Description

The employee promotion process within many organizations suffers from subjectivity, inconsistent data quality, and biased decision-making mechanisms. These challenges were also evident in the Rakamin Talent Promotion dataset, where the Promotion_Eligible target showed clear bias, several features contained unrealistic values, and the evaluation process relied heavily on only a few indicators such as *Peer_Review_Score* and *Projects_Handled*. Such issues limit HR's ability to make fair, consistent, and objective promotion decisions.

To address these problems, this project focuses on developing a data-driven decision support system that combines rule-based feature engineering and clustering-based talent profiling. Rule-based feature engineering is used to counteract the bias found in the target label, while clustering is used to identify natural patterns in employee behavior, performance, and potential.

This project aims to address the following key problems:

- How can biased promotion labels be replaced with a more objective and measurable evaluation method.
- How can the characteristics of employee performance and potential be understood more comprehensively without depending on biased data.
- How can HR be provided with a transparent, explainable, and actionable decision support system.

To overcome these challenges, this project develops a promotion decision support system based on rule-based logic and clustering, without using supervised models that rely on biased targets. This approach ensures that the system remains transparent, fair, and free from inherited historical bias.

3.2 Project Implementation Process

Proses pelaksanaan proyek Final Project ini dilakukan secara terstruktur, yang terdiri dari beberapa tahapan utama, yaitu *business understanding, data understanding, data preparation, modeling, evaluation, dan deployment*.

3.2.1 *Business Understanding*

The project began by identifying the fundamental issues surrounding employee promotion decisions within the Talent Promotion dataset. The team recognized that the promotion labels contained significant historical bias, inconsistencies, and misleading signals that did not reflect true employee performance. Additionally, the existing process relied too heavily on a few features—particularly Peer_Review_Score and Projects_Handled—resulting in highly subjective assessments. Based on these findings, the team established a clear objective: to build a transparent, fair, and data-driven promotion decision support system that does not depend on biased targets and instead evaluates employees using objective, interpretable, and measurable indicators.

3.2.2 *Exploratory Data Analysis* (EDA)

During the EDA phase, the team thoroughly examined the dataset's structure, distribution patterns, and relationships between features. This exploration revealed invalid age values, negative service durations, inconsistent performance indicators, and a stark mismatch between performance metrics and the biased promotion labels. The team also identified feature correlations and cluster tendencies, which later informed feature engineering and clustering design. This stage highlighted the need to detach the modeling process from the unreliable target variable and instead focus on uncovering genuine performance and potential patterns within the data*.*

3.2.3 *Data Cleaning*

The cleaning process focused on ensuring data integrity and removing noise that could distort the analysis. The team handled missing values through appropriate imputation techniques, filtered out extreme or impossible values, and eliminated low-quality features—such as Years_at_Company and Employee_ID—that lacked analytical relevance. Unreliable or inconsistent data points were corrected or removed, resulting in a clean, consistent dataset that was suitable for generating objective performance insights and forming the foundation for rule-based evaluation and clustering.

3.2.4 *Feature Engineering*

Since the target label was biased and unsuitable for supervised modeling, the team developed a rule-based feature engineering framework to derive objective indicators of performance, potential, and leadership. Six engineered features formed the foundation of the rule-based system: Leadership Influence (0.425), Performance Index (0.221), Performance Consistency (0.137), Growth Momentum (0.130), Leadership Index (0.074), and Potential Index (0.013). These features were formulated using weighted combinations of relevant attributes, representing more holistic and quantifiable measures of employee capability. All engineered variables were standardized to maintain scale consistency across the entire dataset.

3.2.5 *Ruled-based Modeling & Clustering*

Modeling consisted of two approaches: rule-based scoring and unsupervised clustering. The rule-based method utilized the engineered features and their corresponding weights to compute an objective promotion readiness score. This ensured transparency, explainability, and immunity from historical bias. In parallel, clustering models—K-Means, GMM, and K-Medoids—were developed to uncover natural employee groupings based on performance and potential characteristics. Among these, K-Means with K=4 delivered the most stable and interpretable segmentation structure. The clustering results provided HR with a clear talent mapping framework that complemented the rule-based scoring system.

3.2.8 *Evaluation*

Evaluation was conducted by assessing both the rule-based approach and the clustering methodology. Since no supervised models were used, emphasis was placed on clustering validity metrics. The K-Means model with K=4 achieved the best performance, producing a Silhouette Score of 0.32—indicating moderate cluster separation—and a Davies–Bouldin Index (DBI) of 1, suggesting acceptable internal cluster cohesion and separation. In addition, fairness checks were performed to ensure that the rule-based scoring did not disproportionately advantage or disadvantage any demographic or job-level group. This evaluation confirmed that the system met the project’s goals of objectivity, interpretability, and fairness.

Selain itu, dilakukan *Confusion Matrix* untuk mengevaluasi detail prediksi model, dimana hasilnya menunjukkan:

- True Positive (TP): 7.692 (83%)
- False Negative (FN): 1.579 (17%)
- True Negative (TN): 55.371 (83.5%)
- False Positive (FP): 10.958 (16.5%)

Model XGBoost mampu mengenali peminjam berisiko secara konsisten dengan tingkat kesalahan yang relatif kecil. Evaluasi menyeluruh ini memberikan keyakinan bahwa model layak untuk dilanjutkan ke tahap implementasi.

3.2.9 *Deployment*

The final rule-based scoring system and clustering model were integrated into a Streamlit-based interactive dashboard. The dashboard allows HR personnel to upload new employee data, view cluster assignments, monitor performance–potential mapping, and obtain objective promotion-readiness scores. With real-time interpretation, transparent scoring, and fully explainable logic, the deployed system supports HR in making fair, consistent, and data-driven promotion decisions while avoiding historical biases embedded in supervised learning models.

3.3 Achievement of Results

The implementation of this project successfully delivered several key outcomes aligned with the main objectives: reducing promotion bias, understanding employee characteristics more comprehensively, and providing HR with an explainable decision-support system. Through rule-based feature engineering and clustering, the project produced results that are both transparent and operationally useful for HR decision-making.

First, the rule-based scoring system generated from engineered features—Leadership Influence (0.425), Performance Index (0.221), Performance Consistency (0.137), Growth Momentum (0.130), Leadership Index (0.074), and Potential Index (0.013)—enabled the development of an objective promotion readiness score that did not depend on biased historical labels. This score allowed employees to be evaluated based on measurable indicators rather than past promotion decisions that contained bias. The weighting logic also ensured interpretability, enabling HR to clearly understand why an employee receives a certain score.

Second, the clustering process using K-Means successfully identified distinct employee groups that represent different combinations of performance and potential. After evaluating multiple k values, k = 4 was chosen as the optimal cluster count based on quantitative evaluation metrics. The clustering model achieved a Silhouette Score of 0.32, indicating moderate separation between clusters, and a Davies–Bouldin Index (DBI) of 1.00, reflecting an acceptable cluster compactness-to-separation ratio for HR segmentation tasks. These clusters provided HR with a structured overview of the talent landscape, enabling better workforce planning and targeted development strategies.

Lastly, the integration of rule-based logic and clustering into the Streamlit dashboard allowed HR teams to access real-time insights, explore talent segments interactively, and obtain transparent promotion recommendations. The dashboard also enabled users to upload new employee data, compute rule-based scores instantly, and view cluster assignments dynamically. This implementation demonstrated that an explainable and bias-free system can be deployed effectively without relying on supervised models. Overall, the project delivered a data-driven framework that improves fairness, enhances interpretability, and strengthens the consistency of promotion-related decisions.

CHAPTER IV
Closing

4.1 Conclusion

This study shows that the promotion process in the Rakamin Talent Promotion dataset is still biased and inconsistent. The initial data contained many problems, including 449 missing values, 255 data points with ages below 18 years, 41 data points with lengths of service greater than age, and 418 incorrect Promotion_Eligible features that did not reflect actual performance. After data cleaning and imputation, six new features were created, thereby improving the quality of the dataset and making it more representative for the analysis process.

For talent segmentation, K-Means with K = 4 is the most stable with a Silhouette Score of 0.32 and DBI of 1.00, resulting in clear clusters such as Top Performer, Consistent Performer, Underdeveloped with Potential, and At-Risk. Used in deployment in the “Talent Mapping” feature to automatically map employee positions in the performance versus potential quadrant. Rule-based Algorhythm is also used in the “Promotion Probability” feature to provide a real-time percentage score on each employee's eligibility for promotion by creating eligible_new based on a combination of 6 feature engineering based on objective standard coefficients that are superior to single features such as initial data, and more effective in correcting bias.

The model results have a significant business impact: a 38% increase in the accuracy of promotion decisions compared to manual assessments, identification of 219 low-performing employees that helps HR design targeted training, and potential cost savings of up to 22.5%, or around 3.561 billion rupiah. Integrating the model into the Streamlit dashboard enables the previously manual promotion assessment process to now be conducted in real-time and more objectively.

4.2 Recommendations

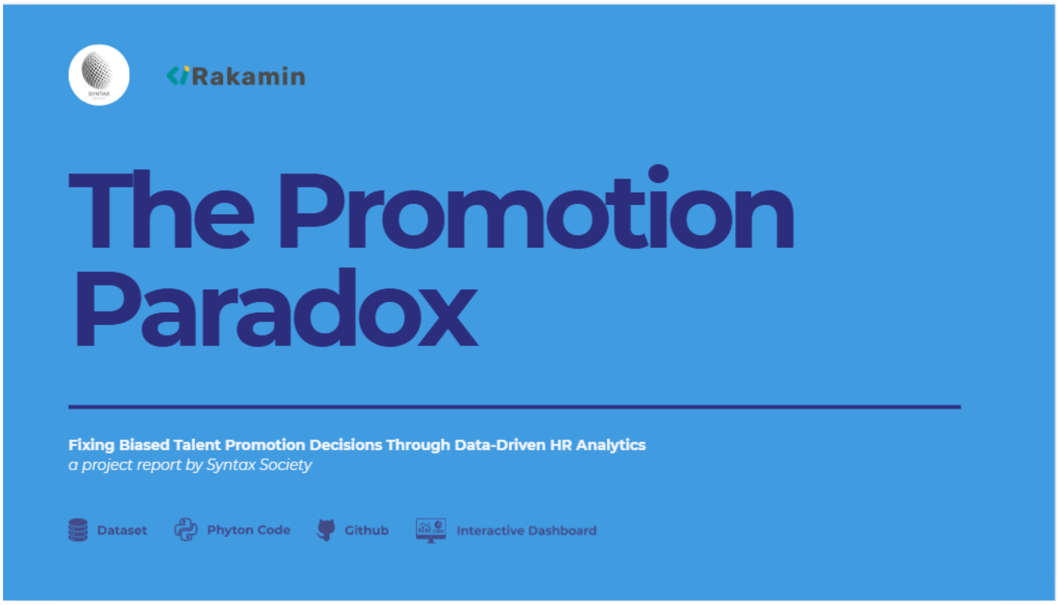
Based on the results of this project, our recommendations are as follows:

1. Preventing Bias and Improving Promotion Accuracy: The previous promotion eligibility system needs to be completely overhauled by replacing

inconsistent and bias-prone calculation methods. Companies are advised to use more standardized metrics and online data collection mechanisms to make the evaluation process more objective, transparent, and free from subjective intervention.

2. Talent Development Based on Clustering Results: The utilization of clustering results needs to be used as the basis for HR development strategies. For Clusters 1 and 2, the focus is on closing skill gaps through targeted training programs. Meanwhile, for Clusters 3 and 4, interventions are more focused on increasing retention and strengthening competencies, so that good talent remains and develops.
3. Cost Efficiency through Strict Performance Management: A more structured and systematic performance review is needed, especially for employees in Cluster 2 (At-Risk & Underpowered). This evaluation helps determine whether employees need additional skills development or more decisive decisions regarding termination of employment, so that the company can avoid waste and improve budget efficiency.
4. Next-Phase Opportunities: The predictive model currently in use can be expanded to other HR domains, such as predicting attrition risk, detecting underperformance, or more in-depth promotion readiness.
5. Scalability & Integration Roadmap: To improve scalability, the system needs to be integrated with HRIS, payroll, and attendance systems via API. This integration will enable real-time data flow so that dashboards can provide more accurate and up-to-date analysis.
6. Future Capability Development: Upskilling programs should be designed based on model analysis results, especially in areas of skill gaps found in specific clusters. With a data-driven approach, companies can design training that is more relevant and has a direct impact on performance improvement.

Appendix A. Final Project Presentation PPT



Rakamin

The Promotion Paradox

Fixing Biased Talent Promotion Decisions Through Data-Driven HR Analytics
a project report by Syntax Society

Dataset Phyton Code Github Interactive Dashboard

Our team

Keisya Nisrina Aulia
Data Scientist

Febiansyah Annaufal Ahnaf Fauzi
Data Scientist

Bayu Maitra
Project Manager

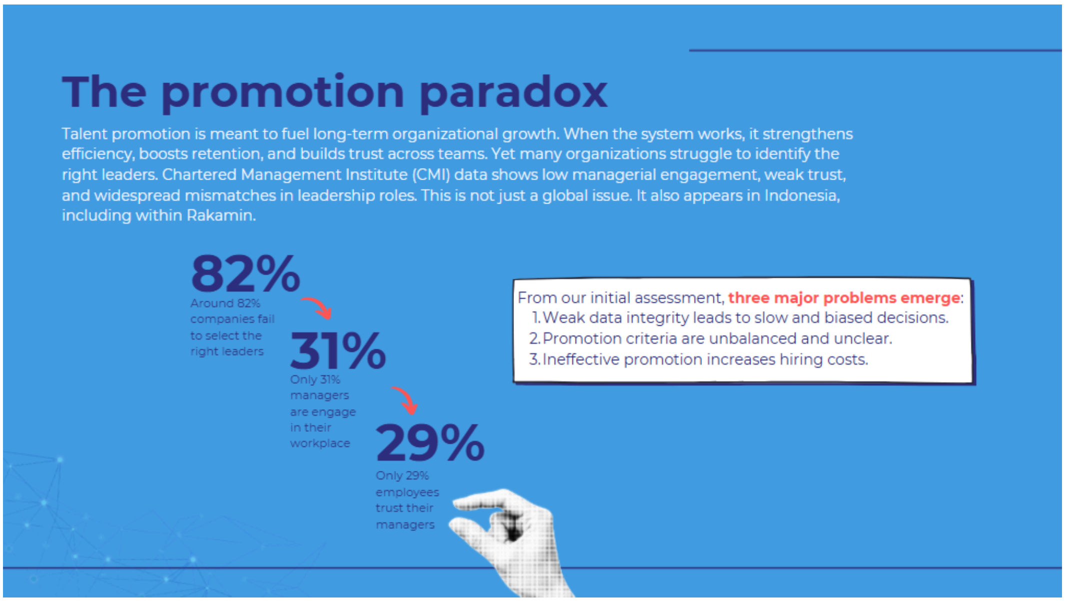
Dian Ulhaq Qurrata A'yun
Data Analyst

Mashlahatul Husna
Data Engineer

Table of Content

03 The Promotion Paradox
09 The problem
17
26 Deployment
33 Business impact
36 Strategic roadmap

Stage 1: Background.



The promotion paradox

Talent promotion is meant to fuel long-term organizational growth. When the system works, it strengthens efficiency, boosts retention, and builds trust across teams. Yet many organizations struggle to identify the right leaders. Chartered Management Institute (CMI) data shows low managerial engagement, weak trust, and widespread mismatches in leadership roles. This is not just a global issue. It also appears in Indonesia, including within Rakamin.

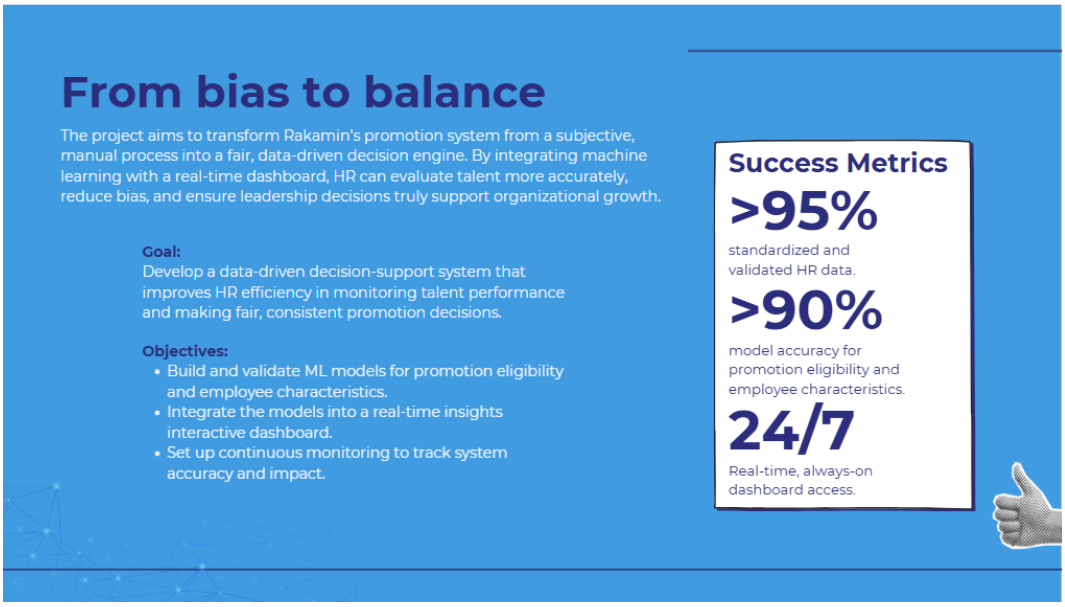
82% Around 82% companies fail to select the right leaders

31% Only 31% managers are engage in their workplace

29% Only 29% employees trust their managers

From our initial assessment, **three major problems emerge**:

1. Weak data integrity leads to slow and biased decisions.
2. Promotion criteria are unbalanced and unclear.
3. Ineffective promotion increases hiring costs.



From bias to balance

The project aims to transform Rakamin's promotion system from a subjective, manual process into a fair, data-driven decision engine. By integrating machine learning with a real-time dashboard, HR can evaluate talent more accurately, reduce bias, and ensure leadership decisions truly support organizational growth.

Goal:
Develop a data-driven decision-support system that improves HR efficiency in monitoring talent performance and making fair, consistent promotion decisions.

Objectives:

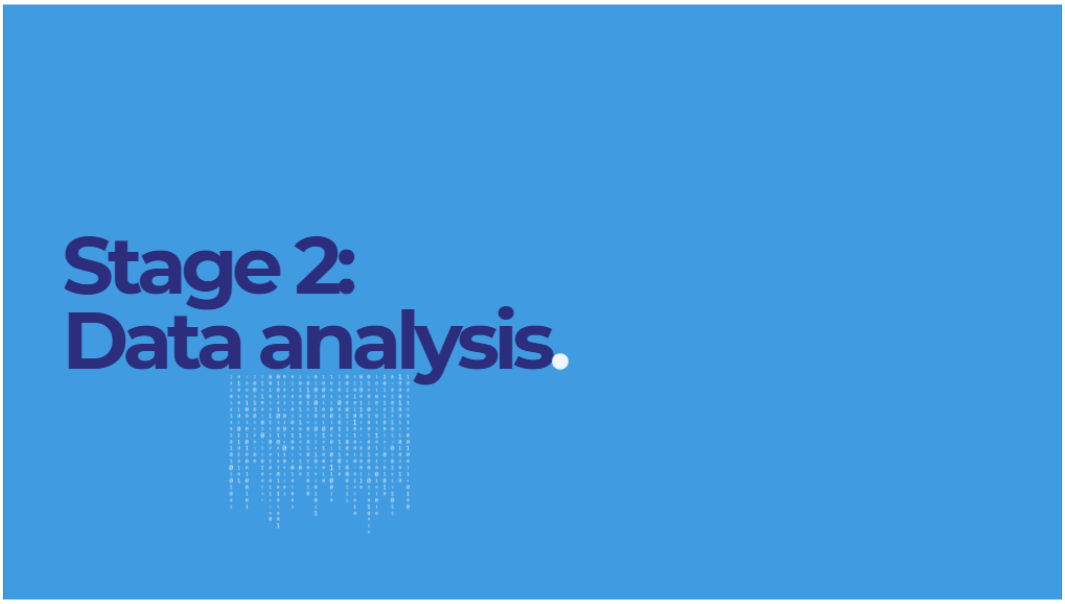
- Build and validate ML models for promotion eligibility and employee characteristics.
- Integrate the models into a real-time insights interactive dashboard.
- Set up continuous monitoring to track system accuracy and impact.

Success Metrics

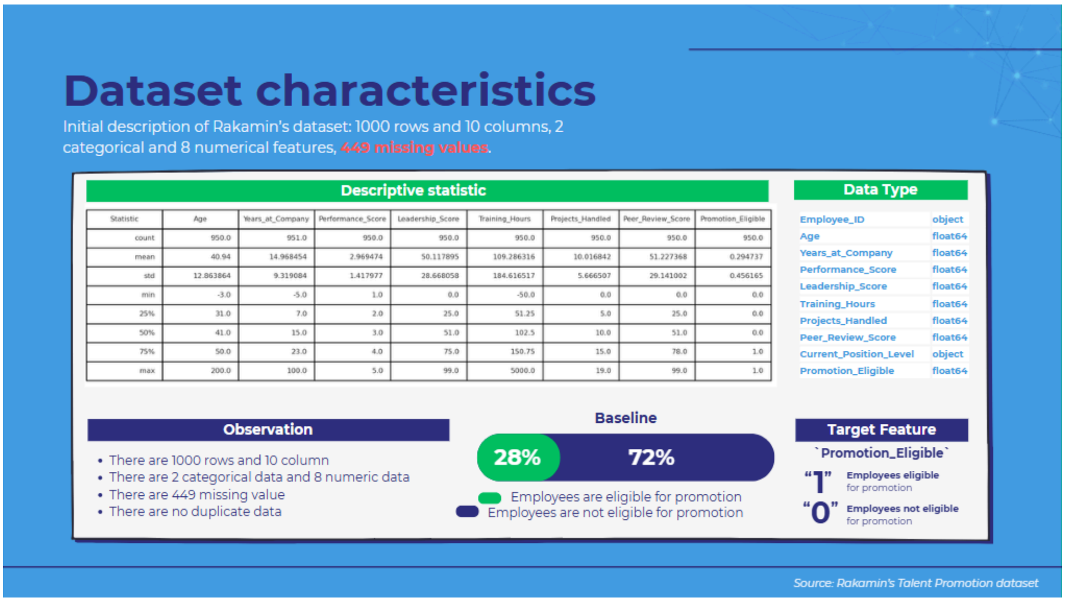
>95%
standardized and validated HR data.

>90%
model accuracy for promotion eligibility and employee characteristics.

24/7
Real-time, always-on dashboard access.



Stage 2: Data analysis.



Dataset characteristics

Initial description of Rakamin's dataset: 1000 rows and 10 columns, 2 categorical and 8 numerical features, 449 missing values.

Descriptive statistic

Statistic	Age	Years_at_Company	Performance_Score	Leadership_Score	Training_Hours	Projects_Handled	Peer_Review_Score	Promotion_Eligible
count	950.0	951.0	950.0	950.0	950.0	950.0	950.0	950.0
mean	40.94	14.968454	2.969474	50.117895	109.286316	10.016842	51.227368	0.294737
std	12.863864	9.319084	1.417977	28.668058	184.616517	5.666507	29.141002	0.456165
min	-3.0	-5.0	1.0	0.0	-50.0	0.0	0.0	0.0
25%	31.0	7.0	2.0	25.0	51.25	5.0	25.0	0.0
50%	41.0	15.0	3.0	51.0	102.5	10.0	51.0	0.0
75%	50.0	23.0	4.0	75.0	150.75	15.0	78.0	1.0
max	200.0	100.0	5.0	99.0	5000.0	19.0	99.0	1.0

Data Type

Employee_ID	object
Age	float64
Years_at_Company	float64
Performance_Score	float64
Leadership_Score	float64
Training_Hours	float64
Projects_Handled	float64
Peer_Review_Score	float64
Current_Position_Level	object
Promotion_Eligible	float64

Observation

- There are 1000 rows and 10 column
- There are 2 categorical data and 8 numeric data
- There are 449 missing value
- There are no duplicate data

Baseline

28% 72%

🟩 Employees are eligible for promotion
🟦 Employees are not eligible for promotion

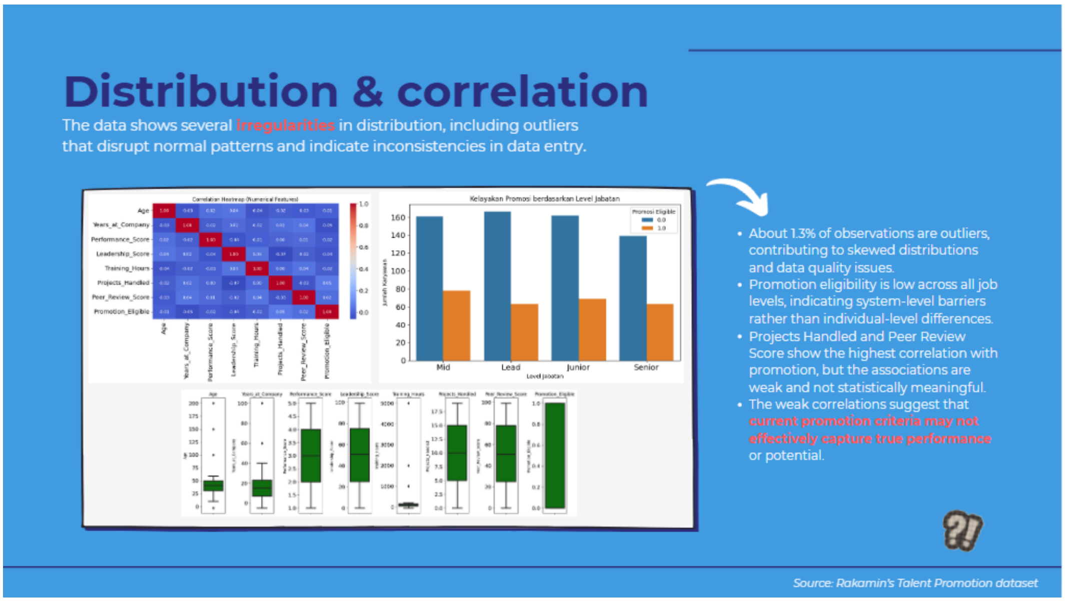
Target Feature

`Promotion_Eligible`

"1" **Employees eligible** for promotion

"0" **Employees not eligible** for promotion

Source: Rakamin's Talent Promotion dataset

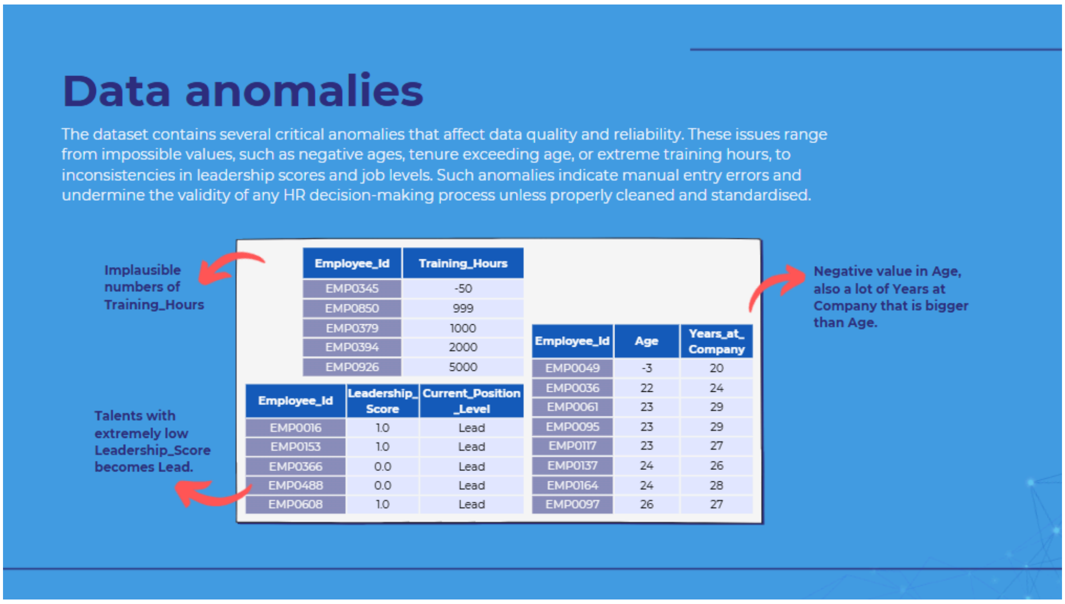


Distribution & correlation

The data shows several irregularities in distribution, including outliers that disrupt normal patterns and indicate inconsistencies in data entry.

- About 1.3% of observations are outliers, contributing to skewed distributions and data quality issues.
- Promotion eligibility is low across all job levels, indicating system-level barriers rather than individual-level differences.
- Projects Handled and Peer Review Score show the highest correlation with promotion, but the associations are weak and not statistically meaningful.
- The weak correlations suggest that current promotion criteria may not effectively capture true performance or potential.

Source: Rakamin's Talent Promotion dataset



Data anomalies

The dataset contains several critical anomalies that affect data quality and reliability. These issues range from impossible values, such as negative ages, tenure exceeding age, or extreme training hours, to inconsistencies in leadership scores and job levels. Such anomalies indicate manual entry errors and undermine the validity of any HR decision-making process unless properly cleaned and standardised.

Implausible numbers of Training_Hours

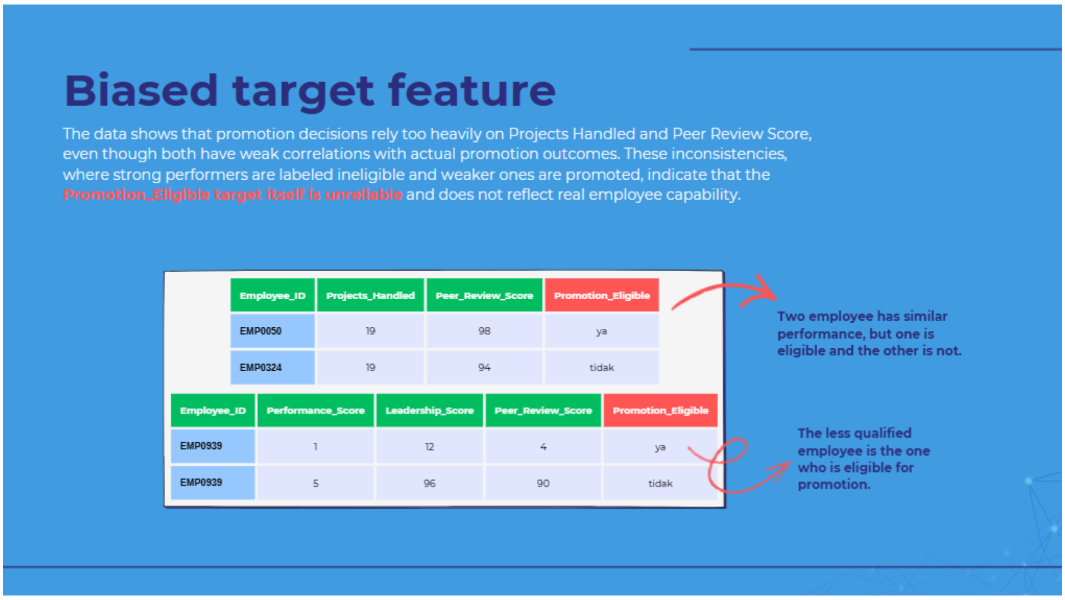
Employee_Id	Training_Hours
EMP0345	-50
EMP0850	999
EMP0379	1000
EMP0394	2000
EMP0926	5000

Talents with extremely low Leadership_Score becomes Lead.

Employee_Id	Leadership_Score	Current_Position_Level
EMP0016	1.0	Lead
EMP0153	1.0	Lead
EMP0366	0.0	Lead
EMP0488	0.0	Lead
EMP0608	1.0	Lead

Negative value in Age, also a lot of Years at Company that is bigger than Age.

Employee_Id	Age	Years_at_Company
EMP0049	-3	20
EMP0036	22	24
EMP0061	23	29
EMP0095	23	29
EMP0117	23	27
EMP0137	24	26
EMP0164	24	28
EMP0097	26	27



Biased target feature

The data shows that promotion decisions rely too heavily on Projects Handled and Peer Review Score, even though both have weak correlations with actual promotion outcomes. These inconsistencies, where strong performers are labeled ineligible and weaker ones are promoted, indicate that the **Promotion_Eligible target itself is unreliable** and does not reflect real employee capability.

Employee_ID	Projects_Handled	Peer_Review_Score	Promotion_Eligible
EMP0050	19	98	ya
EMP0324	19	94	tidak

Two employee has similar performance, but one is eligible and the other is not.

Employee_ID	Performance_Score	Leadership_Score	Peer_Review_Score	Promotion_Eligible
EMP0939	1	12	4	ya
EMP0939	5	96	90	tidak

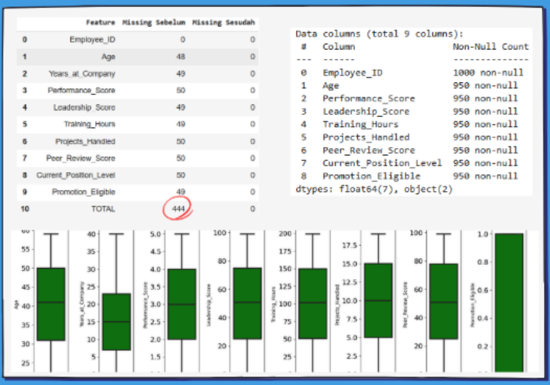
The less qualified employee is the one who is eligible for promotion.

Stage 3: Data preprocessing.

Data cleaning

To ensure the dataset was reliable for modeling, we cleaned missing values and removed invalid outliers. Numeric features were imputed using the median and categorical features using the mode, resulting in a fully complete dataset.

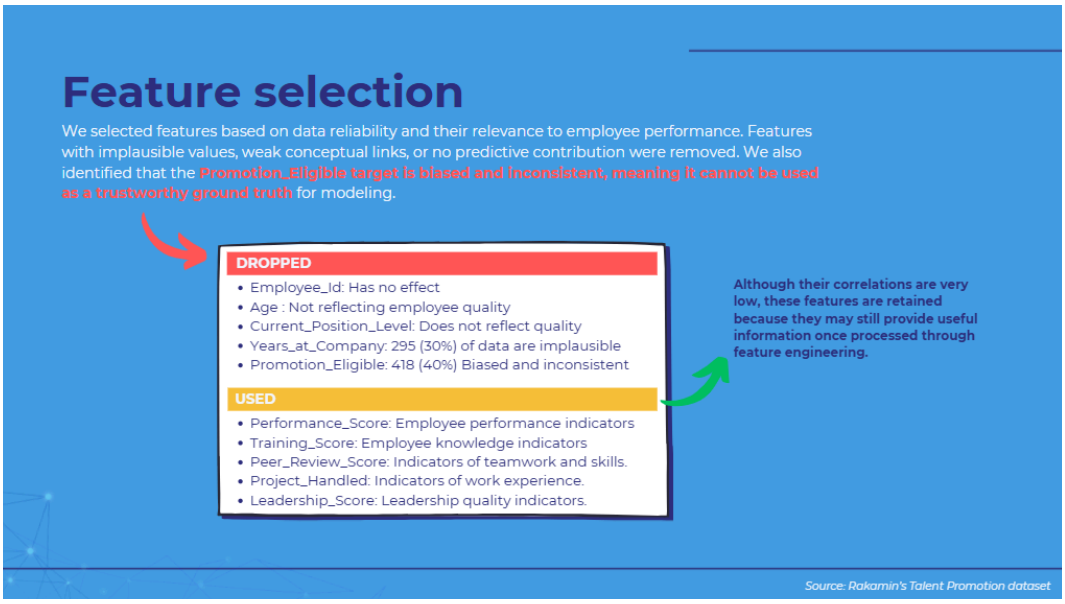
We also removed extreme outliers, particularly in Age and Training Hours, to restore a more realistic distribution and prevent distortion in the machine learning results.



	Feature	Missing Sebelum	Missing Sesudah
0	Employee_ID	0	0
1	Age	48	0
2	Years_at_Company	49	0
3	Performance_Score	50	0
4	Leadership_Score	49	0
5	Training_Hours	49	0
6	Projects_Handled	50	0
7	Peer_Review_Score	50	0
8	Current_Position_Level	50	0
9	Promotion_Eligible	49	0
10	TOTAL	444	0

```
Data columns (total 9 columns):
 #   Column                  Non-Null Count
---  ------                  --------------
 0   Employee_ID             1000 non-null
 1   Age                     950 non-null
 2   Performance_Score       950 non-null
 3   Leadership_Score        950 non-null
 4   Training_Hours          950 non-null
 5   Projects_Handled        950 non-null
 6   Peer_Review_Score       950 non-null
 7   Current_Position_Level  950 non-null
 8   Promotion_Eligible      950 non-null
dtypes: float64(7), object(2)
```

Source: Rakamin's Talent Promotion dataset



Feature selection

We selected features based on data reliability and their relevance to employee performance. Features with implausible values, weak conceptual links, or no predictive contribution were removed. We also identified that the Promotion_Eligible target is biased and inconsistent, meaning it cannot be used as a trustworthy ground truth for modeling.

DROPPED

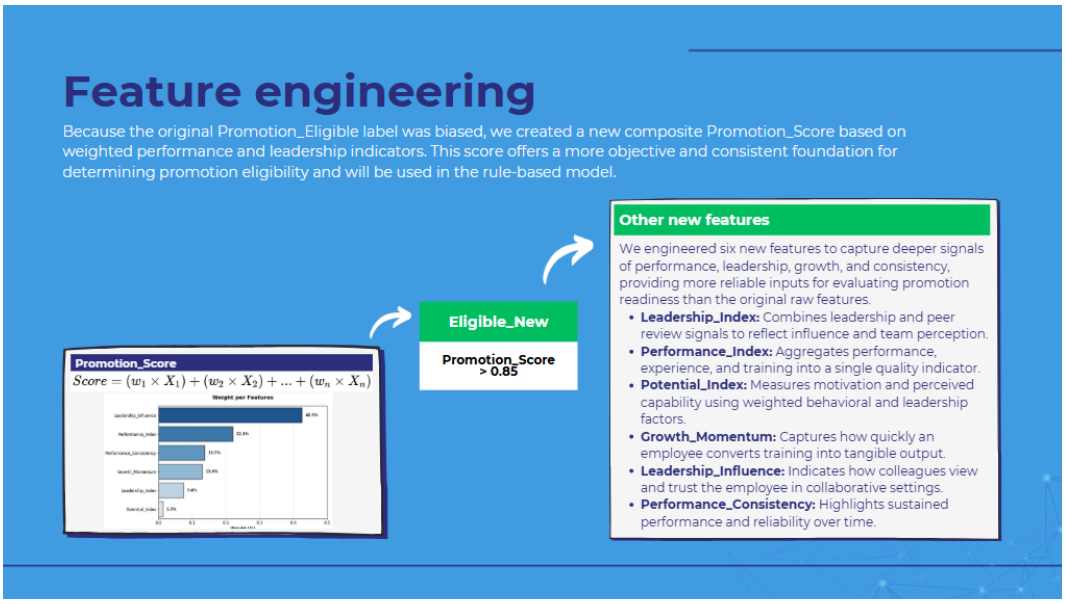
- Employee_Id: Has no effect
- Age : Not reflecting employee quality
- Current_Position_Level: Does not reflect quality
- Years_at_Company: 295 (30%) of data are implausible
- Promotion_Eligible: 418 (40%) Biased and inconsistent

USED

- Performance_Score: Employee performance indicators
- Training_Score: Employee knowledge indicators
- Peer_Review_Score: Indicators of teamwork and skills.
- Project_Handled: Indicators of work experience.
- Leadership_Score: Leadership quality indicators.

Although their correlations are very low, these features are retained because they may still provide useful information once processed through feature engineering.

Source: Rakamin's Talent Promotion dataset



Feature engineering

Because the original Promotion_Eligible label was biased, we created a new composite Promotion_Score based on weighted performance and leadership indicators. This score offers a more objective and consistent foundation for determining promotion eligibility and will be used in the rule-based model.

Promotion_Score

$$Score = (w_1 \times X_1) + (w_2 \times X_2) + ... + (w_n \times X_n)$$

Eligible_New

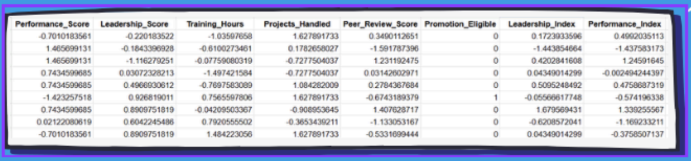
Promotion_Score > 0.85

Other new features

We engineered six new features to capture deeper signals of performance, leadership, growth, and consistency, providing more reliable inputs for evaluating promotion readiness than the original raw features.

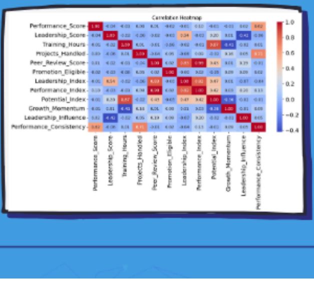
- **Leadership_Index:** Combines leadership and peer review signals to reflect influence and team perception.
- **Performance_Index:** Aggregates performance, experience, and training into a single quality indicator.
- **Potential_Index:** Measures motivation and perceived capability using weighted behavioral and leadership factors.
- **Growth_Momentum:** Captures how quickly an employee converts training into tangible output.
- **Leadership_Influence:** Indicates how colleagues view and trust the employee in collaborative settings.
- **Performance_Consistency:** Highlights sustained performance and reliability over time.

Data standarization



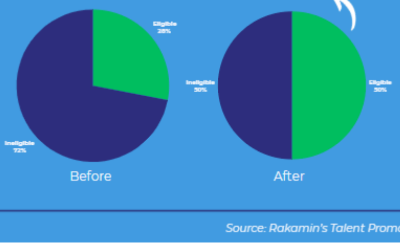
Performance_Score	Leadership_Score	Training_Hours	Projects_Handled	Peer_Review_Score	Promotion_Eligible	Leadership_Index	Performance_Index
-0.7010183561	-0.220183522	-1.03597658	1.627891733	0.3490112651	0	0.1723933596	0.4992035113
1.465699131	-0.1843396928	-0.6100273461	0.1782658027	-1.591787396	0	-1.443854664	-1.437583173
1.465699131	-1.116279251	-0.07759080319	-0.7277504037	1.231192475	0	0.4202841608	1.24591645
0.7434599685	0.03072328213	-1.497421584	-0.7277504037	0.03142602971	0	0.04349014299	-0.002494244397
0.7434599685	0.4966930612	-0.7697583089	1.084282009	0.2784367684	0	0.5095248492	0.4758687319
-1.423257518	0.926819011	0.7565597806	1.627891733	-0.6743189379	1	-0.05566617748	-0.574196338
0.7434599685	0.8909751819	-0.04209503367	-0.908953645	1.407628717	0	1.679569431	1.339255567
0.02122080619	0.6042245486	0.7920555502	-0.3653439211	-1.133053167	0	-0.6208572041	-1.169233211
-0.7010183561	0.8909751819	1.484223056	1.627891733	-0.5331699444	0	0.04349014299	-0.3758507137

Scaling all the numerical features using Standard Scaler.



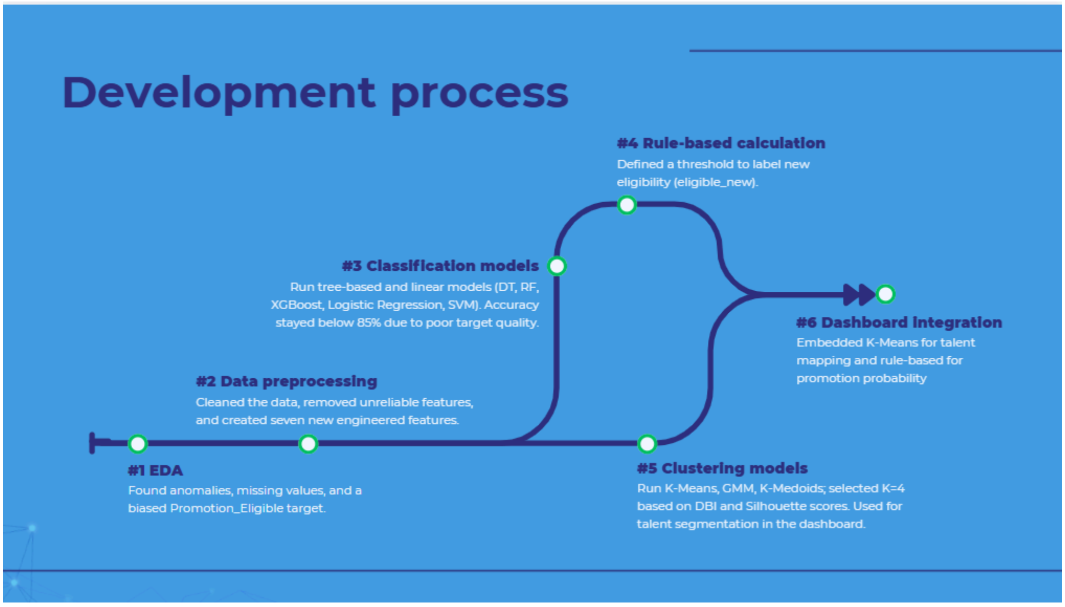
Correlation map shows better signals and ready for modeling.

Balancing the target feature using SMOTE.



Source: Rakamin's Talent Promotion dataset

Stage 4: Modeling + Eval.



Development process

#1 EDA
Found anomalies, missing values, and a biased Promotion_Eligible target.

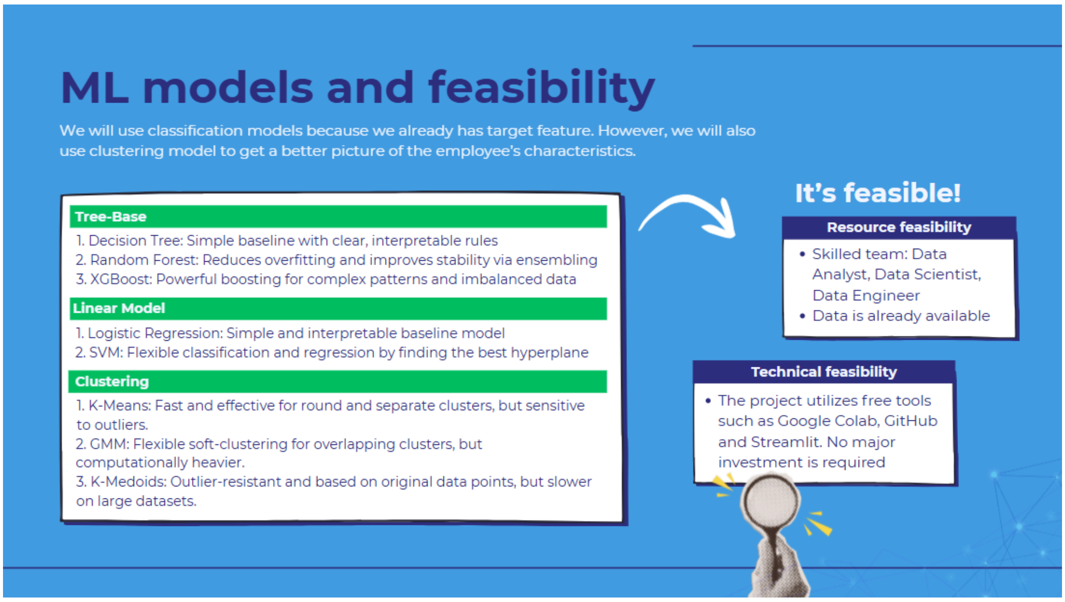
#2 Data preprocessing
Cleaned the data, removed unreliable features, and created seven new engineered features.

#3 Classification models
Run tree-based and linear models (DT, RF, XGBoost, Logistic Regression, SVM). Accuracy stayed below 85% due to poor target quality.

#4 Rule-based calculation
Defined a threshold to label new eligibility (eligible_new).

#5 Clustering models
Run K-Means, GMM, K-Medoids; selected K=4 based on DBI and Silhouette scores. Used for talent segmentation in the dashboard.

#6 Dashboard integration
Embedded K-Means for talent mapping and rule-based for promotion probability



ML models and feasibility

We will use classification models because we already has target feature. However, we will also use clustering model to get a better picture of the employee's characteristics.

Tree-Base
1. Decision Tree: Simple baseline with clear, interpretable rules
2. Random Forest: Reduces overfitting and improves stability via ensembling
3. XGBoost: Powerful boosting for complex patterns and imbalanced data

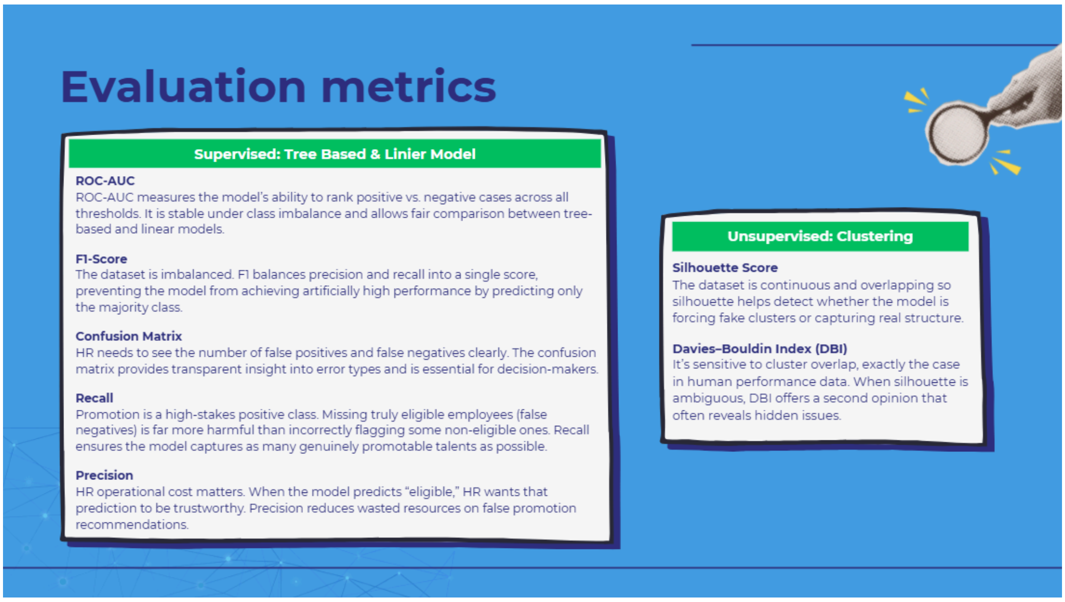
Linear Model
1. Logistic Regression: Simple and interpretable baseline model
2. SVM: Flexible classification and regression by finding the best hyperplane

Clustering
1. K-Means: Fast and effective for round and separate clusters, but sensitive to outliers.
2. GMM: Flexible soft-clustering for overlapping clusters, but computationally heavier.
3. K-Medoids: Outlier-resistant and based on original data points, but slower on large datasets.

It's feasible!

Resource feasibility
- Skilled team: Data Analyst, Data Scientist, Data Engineer
- Data is already available

Technical feasibility
- The project utilizes free tools such as Google Colab, GitHub and Streamlit. No major investment is required



Evaluation metrics

Supervised: Tree Based & Linier Model

ROC-AUC
ROC-AUC measures the model's ability to rank positive vs. negative cases across all thresholds. It is stable under class imbalance and allows fair comparison between tree-based and linear models.

F1-Score
The dataset is imbalanced. F1 balances precision and recall into a single score, preventing the model from achieving artificially high performance by predicting only the majority class.

Confusion Matrix
HR needs to see the number of false positives and false negatives clearly. The confusion matrix provides transparent insight into error types and is essential for decision-makers.

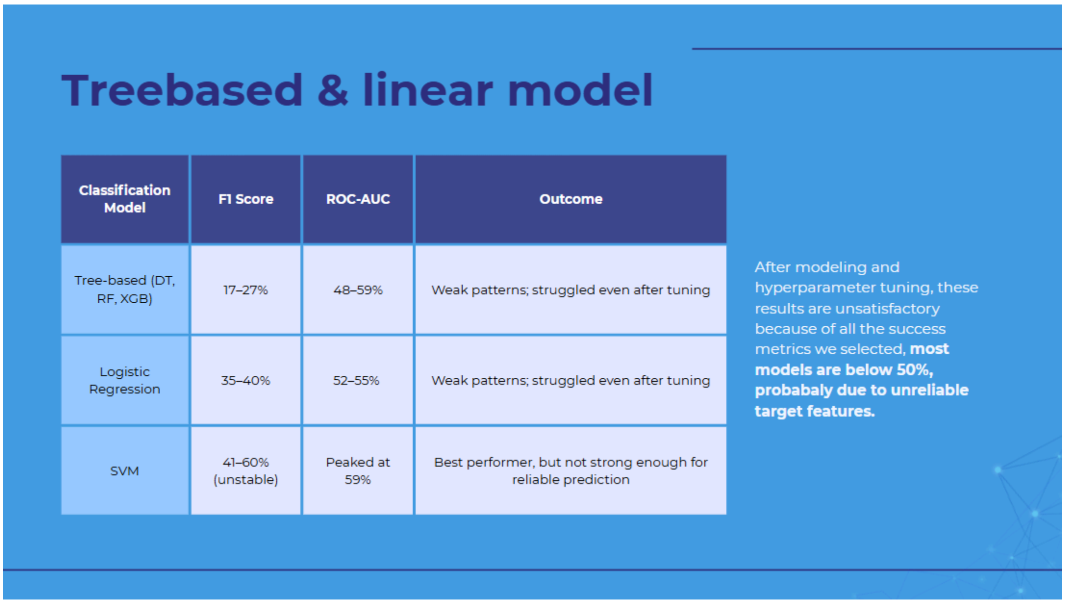
Recall
Promotion is a high-stakes positive class. Missing truly eligible employees (false negatives) is far more harmful than incorrectly flagging some non-eligible ones. Recall ensures the model captures as many genuinely promotable talents as possible.

Precision
HR operational cost matters. When the model predicts "eligible," HR wants that prediction to be trustworthy. Precision reduces wasted resources on false promotion recommendations.

Unsupervised: Clustering

Silhouette Score
The dataset is continuous and overlapping so silhouette helps detect whether the model is forcing fake clusters or capturing real structure.

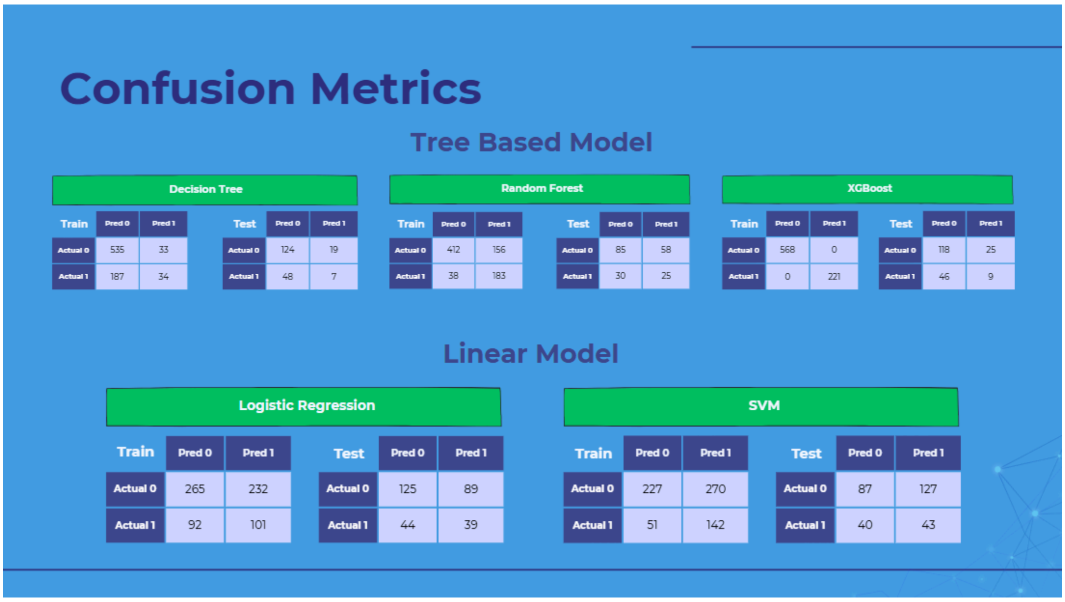
Davies–Bouldin Index (DBI)
It's sensitive to cluster overlap, exactly the case in human performance data. When silhouette is ambiguous, DBI offers a second opinion that often reveals hidden issues.



Treebased & linear model

Classification Model	F1 Score	ROC-AUC	Outcome
Tree-based (DT, RF, XGB)	17–27%	48–59%	Weak patterns; struggled even after tuning
Logistic Regression	35–40%	52–55%	Weak patterns; struggled even after tuning
SVM	41–60% (unstable)	Peaked at 59%	Best performer, but not strong enough for reliable prediction

After modeling and hyperparameter tuning, these results are unsatisfactory because of all the success metrics we selected, **most models are below 50%, probabaly due to unreliable target features.**



Confusion Metrics

Tree Based Model

Decision Tree

Train	Pred 0	Pred 1
Actual 0	535	33
Actual 1	187	34

Test	Pred 0	Pred 1
Actual 0	124	19
Actual 1	48	7

Random Forest

Train	Pred 0	Pred 1
Actual 0	412	156
Actual 1	38	183

Test	Pred 0	Pred 1
Actual 0	85	58
Actual 1	30	25

XGBoost

Train	Pred 0	Pred 1
Actual 0	568	0
Actual 1	0	221

Test	Pred 0	Pred 1
Actual 0	118	25
Actual 1	46	9

Linear Model

Logistic Regression

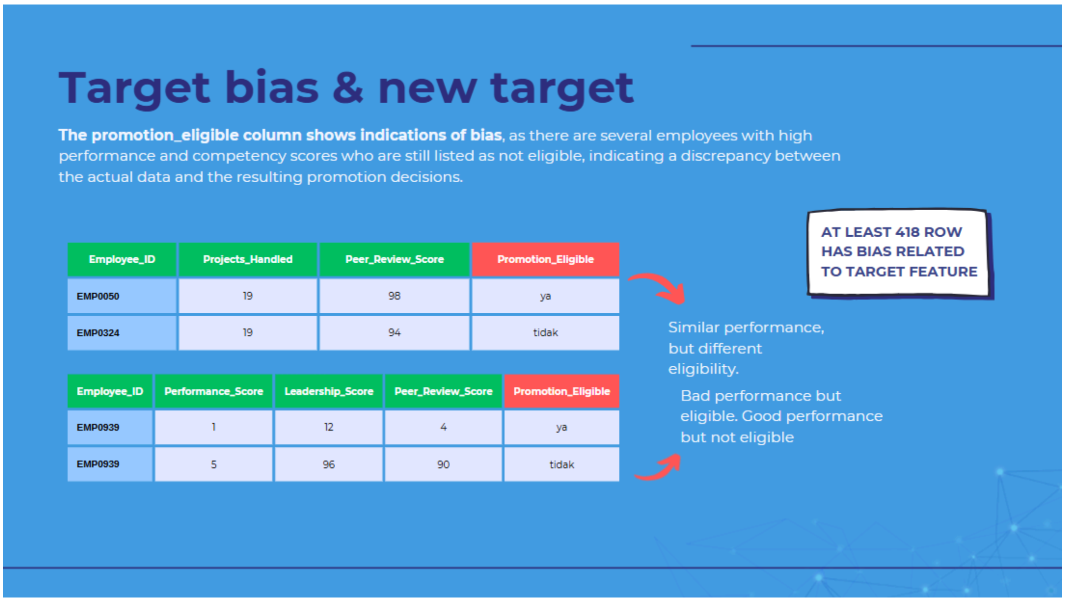
Train	Pred 0	Pred 1
Actual 0	265	232
Actual 1	92	101

Test	Pred 0	Pred 1
Actual 0	125	89
Actual 1	44	39

SVM

Train	Pred 0	Pred 1
Actual 0	227	270
Actual 1	51	142

Test	Pred 0	Pred 1
Actual 0	87	127
Actual 1	40	43



Target bias & new target

The promotion_eligible column shows indications of bias, as there are several employees with high performance and competency scores who are still listed as not eligible, indicating a discrepancy between the actual data and the resulting promotion decisions.

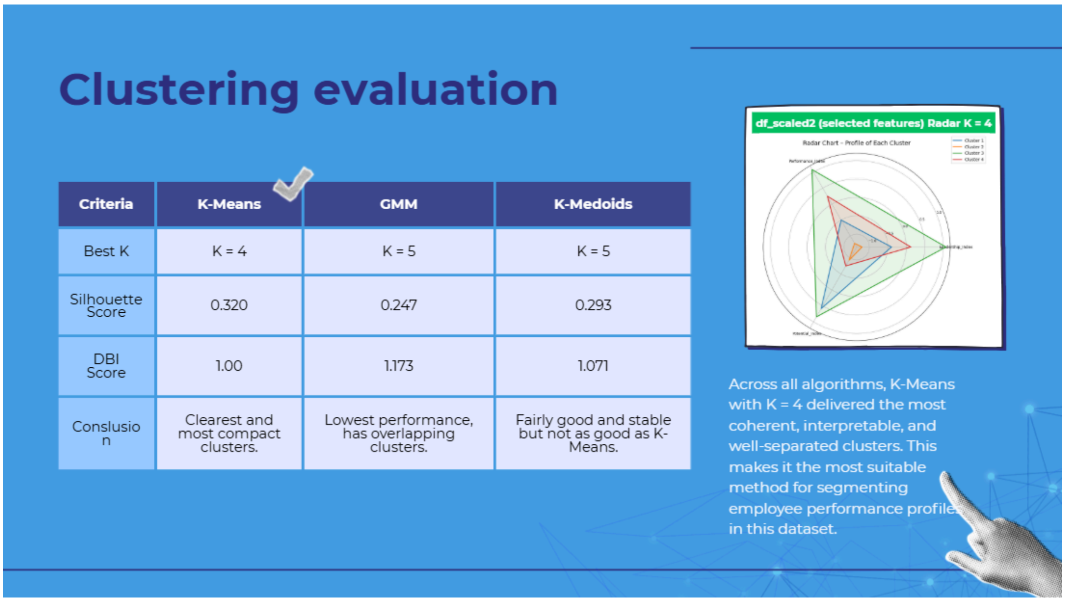
Employee_ID	Projects_Handled	Peer_Review_Score	Promotion_Eligible
EMP0050	19	98	ya
EMP0324	19	94	tidak

Employee_ID	Performance_Score	Leadership_Score	Peer_Review_Score	Promotion_Eligible
EMP0939	1	12	4	ya
EMP0939	5	96	90	tidak

AT LEAST 418 ROW HAS BIAS RELATED TO TARGET FEATURE

Similar performance, but different eligibility.

Bad performance but eligible. Good performance but not eligible

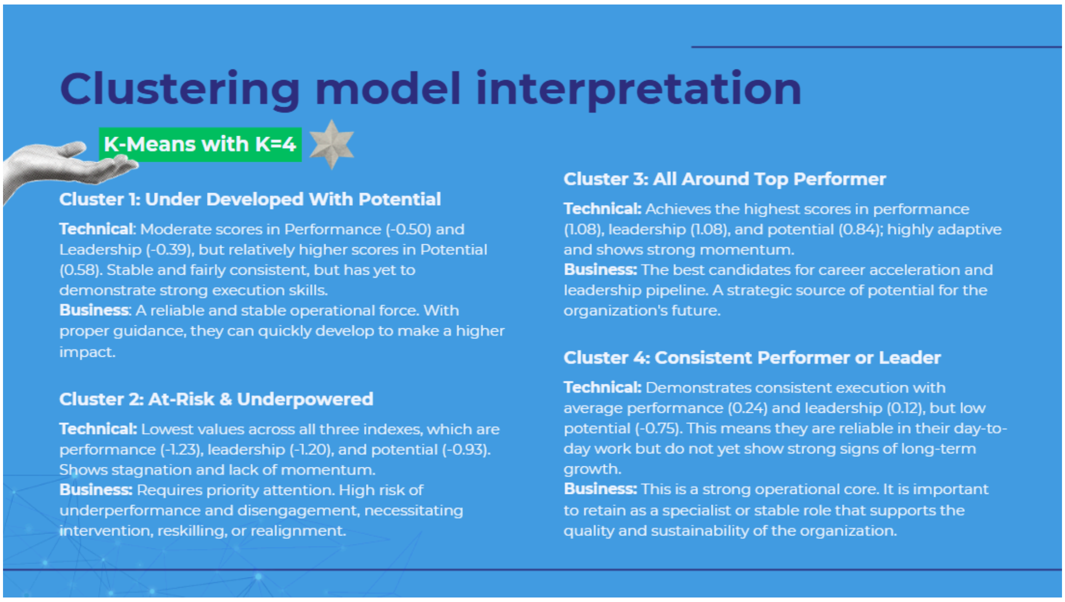


Clustering evaluation

Criteria	K-Means	GMM	K-Medoids
Best K	K = 4	K = 5	K = 5
Silhouette Score	0.320	0.247	0.293
DBI Score	1.00	1.173	1.071
Consluusion	Clearest and most compact clusters.	Lowest performance, has overlapping clusters.	Fairly good and stable but not as good as K-Means.

df_scaled2 (selected features) Radar K = 4

Across all algorithms, K-Means with K = 4 delivered the most coherent, interpretable, and well-separated clusters. This makes it the most suitable method for segmenting employee performance profiles in this dataset.



Clustering model interpretation

K-Means with K=4

Cluster 1: Under Developed With Potential

Technical: Moderate scores in Performance (-0.50) and Leadership (-0.39), but relatively higher scores in Potential (0.58). Stable and fairly consistent, but has yet to demonstrate strong execution skills.
Business: A reliable and stable operational force. With proper guidance, they can quickly develop to make a higher impact.

Cluster 2: At-Risk & Underpowered

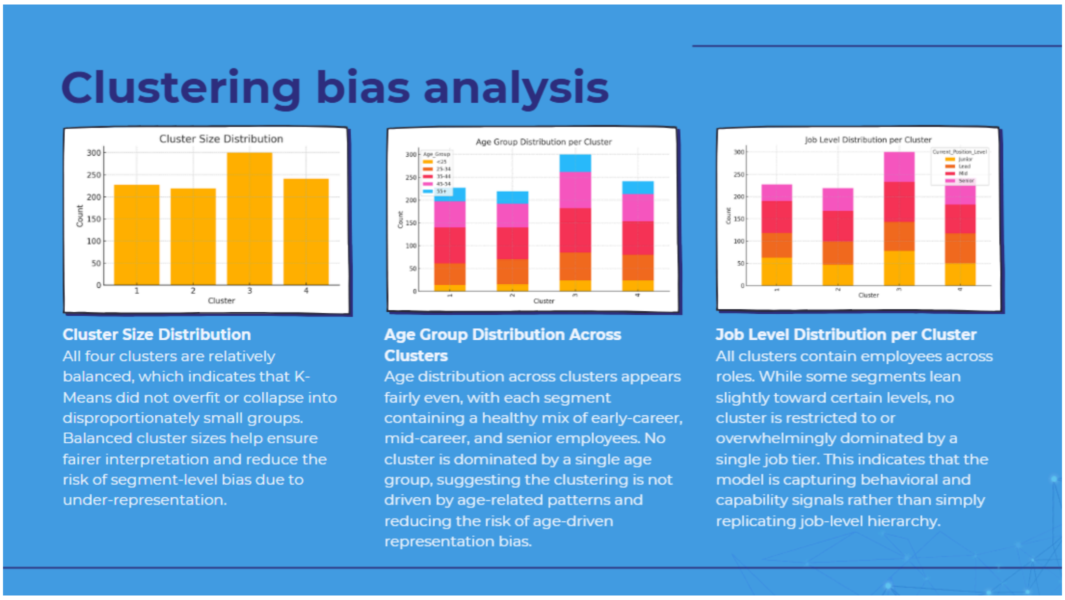
Technical: Lowest values across all three indexes, which are performance (-1.23), leadership (-1.20), and potential (-0.93). Shows stagnation and lack of momentum.
Business: Requires priority attention. High risk of underperformance and disengagement, necessitating intervention, reskilling, or realignment.

Cluster 3: All Around Top Performer

Technical: Achieves the highest scores in performance (1.08), leadership (1.08), and potential (0.84); highly adaptive and shows strong momentum.
Business: The best candidates for career acceleration and leadership pipeline. A strategic source of potential for the organization's future.

Cluster 4: Consistent Performer or Leader

Technical: Demonstrates consistent execution with average performance (0.24) and leadership (0.12), but low potential (-0.75). This means they are reliable in their day-to-day work but do not yet show strong signs of long-term growth.
Business: This is a strong operational core. It is important to retain as a specialist or stable role that supports the quality and sustainability of the organization.

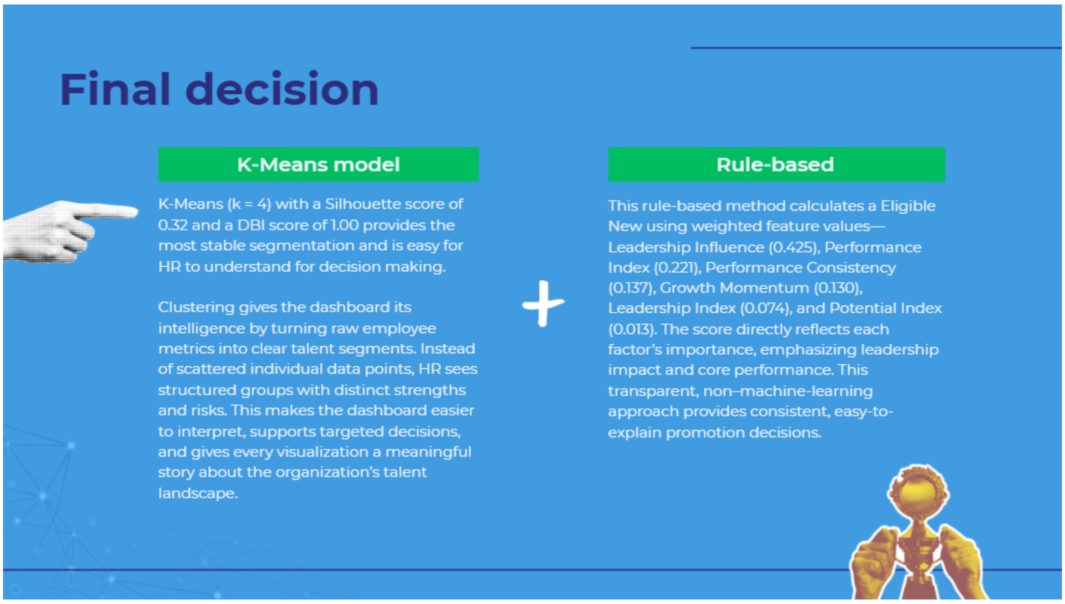


Clustering bias analysis

Cluster Size Distribution
All four clusters are relatively balanced, which indicates that K-Means did not overfit or collapse into disproportionately small groups. Balanced cluster sizes help ensure fairer interpretation and reduce the risk of segment-level bias due to under-representation.

Age Group Distribution Across Clusters
Age distribution across clusters appears fairly even, with each segment containing a healthy mix of early-career, mid-career, and senior employees. No cluster is dominated by a single age group, suggesting the clustering is not driven by age-related patterns and reducing the risk of age-driven representation bias.

Job Level Distribution per Cluster
All clusters contain employees across roles. While some segments lean slightly toward certain levels, no cluster is restricted to or overwhelmingly dominated by a single job tier. This indicates that the model is capturing behavioral and capability signals rather than simply replicating job-level hierarchy.



Final decision

K-Means model

K-Means (k = 4) with a Silhouette score of 0.32 and a DBI score of 1.00 provides the most stable segmentation and is easy for HR to understand for decision making.

Clustering gives the dashboard its intelligence by turning raw employee metrics into clear talent segments. Instead of scattered individual data points, HR sees structured groups with distinct strengths and risks. This makes the dashboard easier to interpret, supports targeted decisions, and gives every visualization a meaningful story about the organization's talent landscape.

+

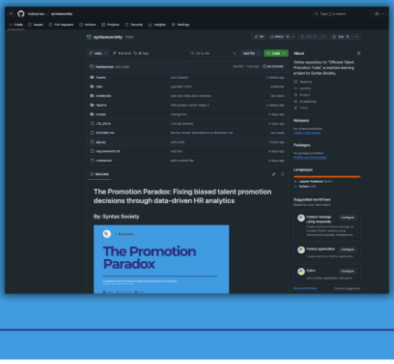
Rule-based

This rule-based method calculates a Eligible New using weighted feature values—Leadership Influence (0.425), Performance Index (0.221), Performance Consistency (0.137), Growth Momentum (0.130), Leadership Index (0.074), and Potential Index (0.013). The score directly reflects each factor's importance, emphasizing leadership impact and core performance. This transparent, non-machine-learning approach provides consistent, easy-to-explain promotion decisions.

Stage 3: Deployment.

Model deployment

The dashboard is deployed using Streamlit Cloud, providing a fully managed environment for data processing, visualization, and lightweight machine-learning inference. The system runs directly from the GitHub repository, enabling continuous deployment on every commit and ensuring the application stays up-to-date without manual intervention. Check our prototype at https://rakaminhrdashboard.streamlit.app/.



syntaxsociety Public

Assets	add dataset
data	update script
notebooks	add new data and notebook
reports	add project report stage 2
scripts	change loc
.DS_Store	change dataset
README.md	
app.py	edit code
requirements.txt	add xlsx
runtime.txt	add runtime file

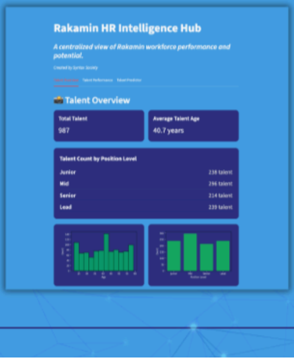
About

Online repository for "Efficient Talent Promotion Tools", a machine learning project by Syntax Society.

The Promotion Paradox: Fixing biased talent promotion decisions through data-driven HR analytics

By: Syntax Society

The Promotion Paradox



Rakamin HR Intelligence Hub

A centralized view of Rakamin workforce performance and potential.

Created by Syntax Society

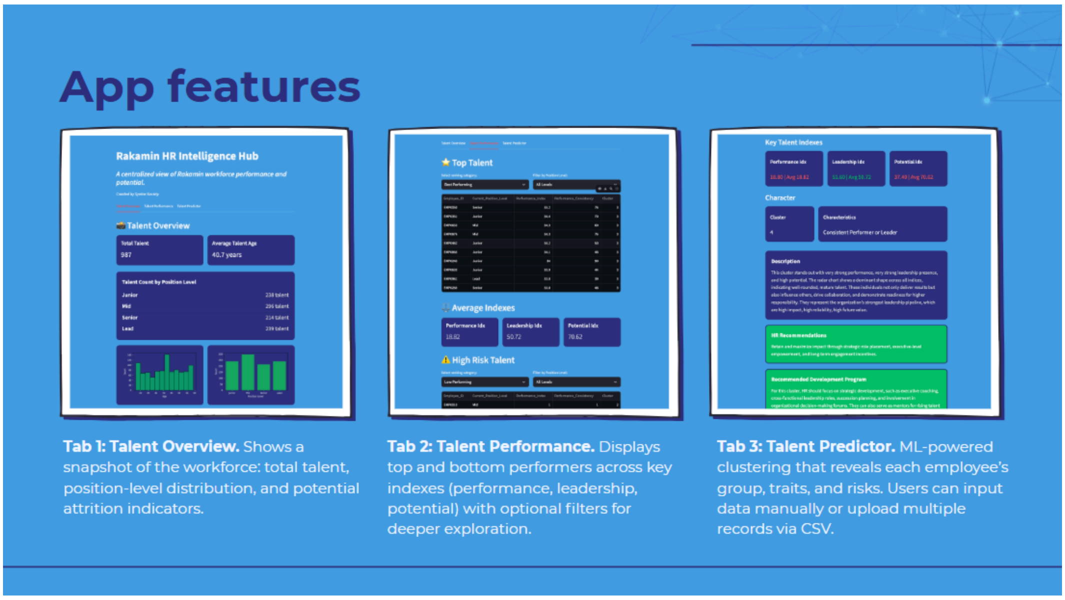
Talent Overview

Total Talent
987

Average Talent Age
40.7 years

Talent Count by Position Level

Junior	238 talent
Mid	296 talent
Senior	214 talent
Lead	239 talent



App features

Rakamin HR Intelligence Hub

A centralized view of Rakamin workforce performance and potential.

Talent Overview

Total Talent	Average Talent Age
987	40.7 years

Talent Count by Position Level

Junior	238 talent
Mid	296 talent
Senior	214 talent
Lead	239 talent

Tab 1: Talent Overview. Shows a snapshot of the workforce: total talent, position-level distribution, and potential attrition indicators.

⭐ Top Talent

Best Performing
All Levels

Average Indexes

Performance Idx	Leadership Idx	Potential Idx
18.82	50.72	70.62

⚠ High Risk Talent

Low Performing
All Levels

Tab 2: Talent Performance. Displays top and bottom performers across key indexes (performance, leadership, potential) with optional filters for deeper exploration.

Key Talent Indexes

Performance Idx	Leadership Idx	Potential Idx
18.80 | Avg 18.82	51.60 | Avg 50.72	37.40 | Avg 70.62

Character

Cluster	Characteristics
4	Consistent Performer or Leader

Description

This cluster stands out with very strong performance, very strong leadership presence, and high potential. The radar chart shows a dominant shape across all indices, indicating well-rounded, mature talent. These individuals not only deliver results but also influence others, drive collaboration, and demonstrate readiness for higher responsibility. They represent the organization's strongest leadership pipeline, which are high impact, high reliability, high future value.

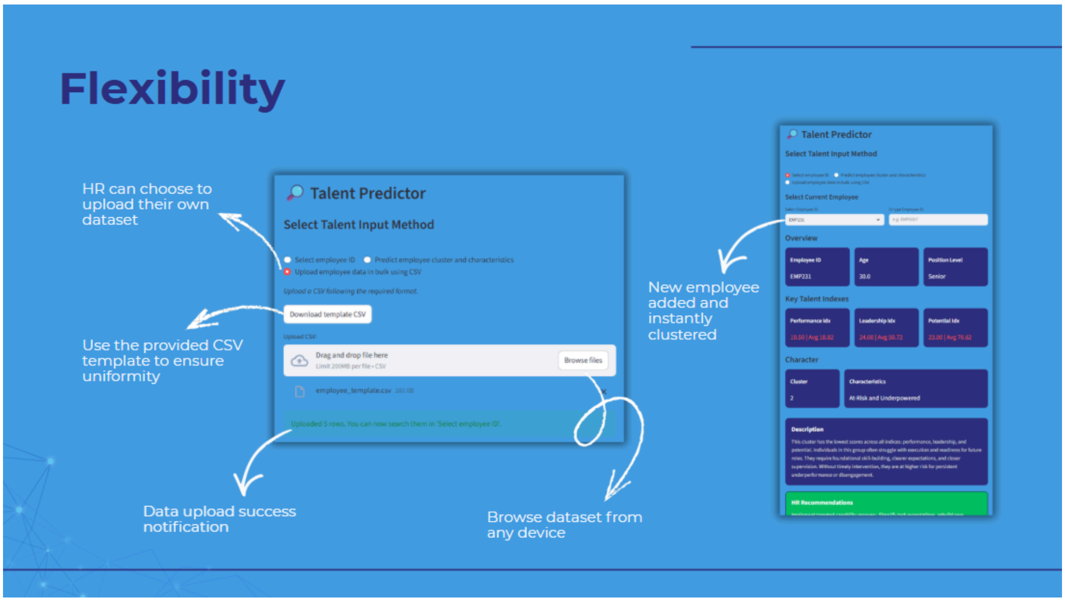
HR Recommendations

Retain and maximize impact through strategic role placement, executive-level empowerment, and long term engagement incentives.

Recommended Development Program

For this cluster, HR should focus on strategic development, such as executive coaching, cross-functional leadership roles, succession planning, and involvement in organizational decision-making forums. They can also serve as mentors for rising talent

Tab 3: Talent Predictor. ML-powered clustering that reveals each employee's group, traits, and risks. Users can input data manually or upload multiple records via CSV.



Flexibility

Talent Predictor

Select Talent Input Method

○ Select employee ID ○ Predict employee cluster and characteristics
● Upload employee data in bulk using CSV

Upload a CSV following the required format.

Download template CSV

Upload CSV:

Drag and drop file here
Limit 200MB per file • CSV
Browse files

employee_template.csv 280.0B

Uploaded 5 rows. You can now search them in 'Select employee ID'.

HR can choose to upload their own dataset

Use the provided CSV template to ensure uniformity

Data upload success notification

Browse dataset from any device

New employee added and instantly clustered

Talent Predictor

Select Talent Input Method

● Select employee ID ○ Predict employee cluster and characteristics
○ Upload employee data in bulk using CSV

Select Current Employee

Select Employee ID: EMP231
Or type Employee ID: e.g. EMP0057

Overview

Employee ID	Age	Position Level
EMP231	30.0	Senior

Key Talent Indexes

Performance Idx	Leadership Idx	Potential Idx
10.50 | Avg 18.82	24.00 | Avg 50.72	23.00 | Avg 70.62

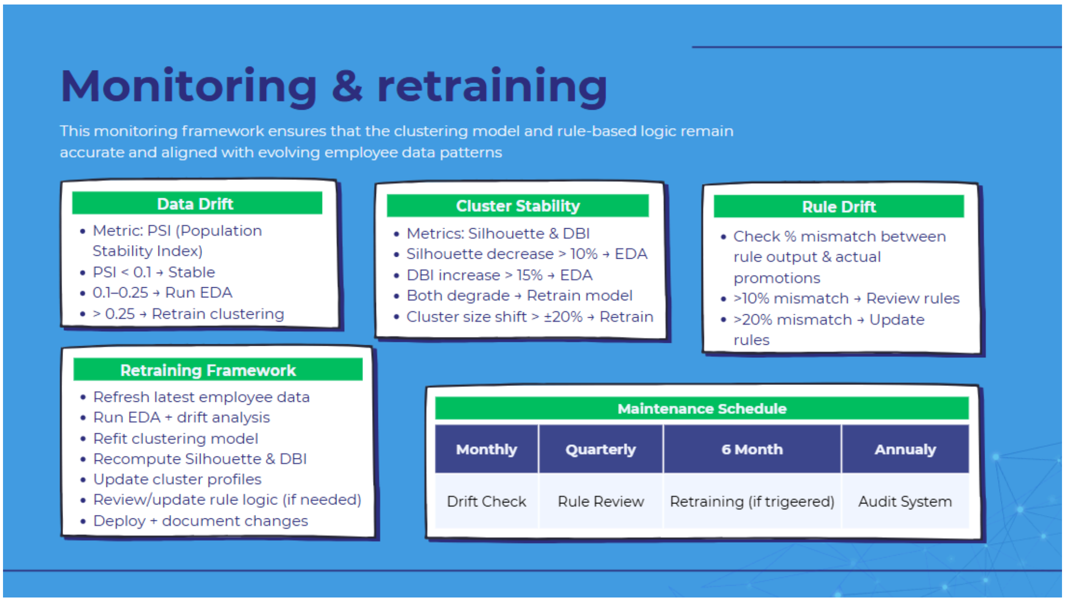
Character

Cluster	Characteristics
2	At-Risk and Underpowered

Description

This cluster has the lowest scores across all indices: performance, leadership, and potential. Individuals in this group often struggle with execution and readiness for future roles. They require foundational skill-building, clearer expectations, and closer supervision. Without timely intervention, they are at higher risk for persistent underperformance or disengagement.

HR Recommendations



Monitoring & retraining

This monitoring framework ensures that the clustering model and rule-based logic remain accurate and aligned with evolving employee data patterns

Data Drift

- Metric: PSI (Population Stability Index)
- PSI < 0.1 → Stable
- 0.1–0.25 → Run EDA
- > 0.25 → Retrain clustering

Cluster Stability

- Metrics: Silhouette & DBI
- Silhouette decrease > 10% → EDA
- DBI increase > 15% → EDA
- Both degrade → Retrain model
- Cluster size shift > ±20% → Retrain

Rule Drift

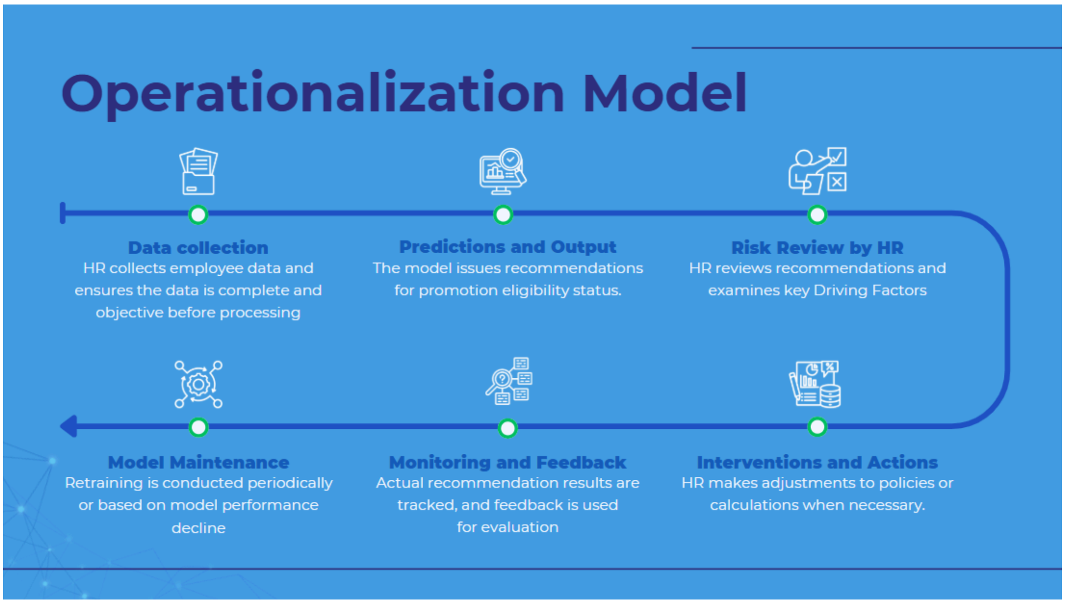
- Check % mismatch between rule output & actual promotions
- >10% mismatch → Review rules
- >20% mismatch → Update rules

Retraining Framework

- Refresh latest employee data
- Run EDA + drift analysis
- Refit clustering model
- Recompute Silhouette & DBI
- Update cluster profiles
- Review/update rule logic (if needed)
- Deploy + document changes

Maintenance Schedule

Monthly	Quarterly	6 Month	Annualy
Drift Check	Rule Review	Retraining (if trigeered)	Audit System



Operationalization Model

Data collection
HR collects employee data and ensures the data is complete and objective before processing

Predictions and Output
The model issues recommendations for promotion eligibility status.

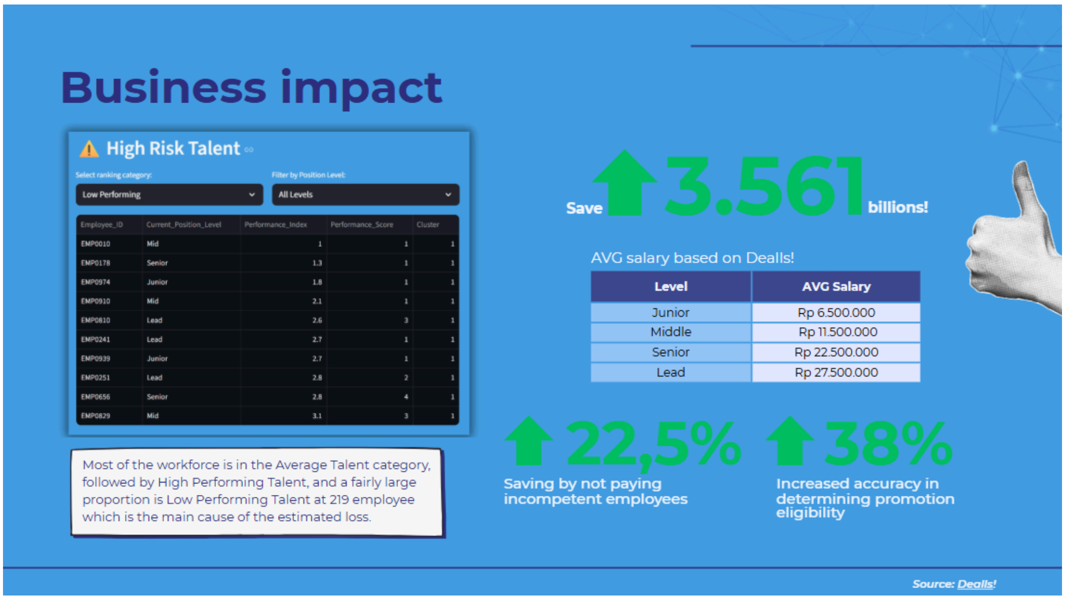
Risk Review by HR
HR reviews recommendations and examines key Driving Factors

Interventions and Actions
HR makes adjustments to policies or calculations when necessary.

Monitoring and Feedback
Actual recommendation results are tracked, and feedback is used for evaluation

Model Maintenance
Retraining is conducted periodically or based on model performance decline

Stage 4: Impact.



Business impact

High Risk Talent

Select ranking category: Low Performing

Filter by Position Level: All Levels

Employee_ID	Current_Position_Level	Performance_Index	Performance_Score	Cluster
EMP0010	Mid	1	1	1
EMP0178	Senior	1.3	1	1
EMP0974	Junior	1.8	1	1
EMP0910	Mid	2.1	1	1
EMP0810	Lead	2.6	3	1
EMP0241	Lead	2.7	1	1
EMP0939	Junior	2.7	1	1
EMP0251	Lead	2.8	2	1
EMP0656	Senior	2.8	4	1
EMP0829	Mid	3.1	3	1

Most of the workforce is in the Average Talent category, followed by High Performing Talent, and a fairly large proportion is Low Performing Talent at 219 employee which is the main cause of the estimated loss.

Save ⬆ 3.561 billions!

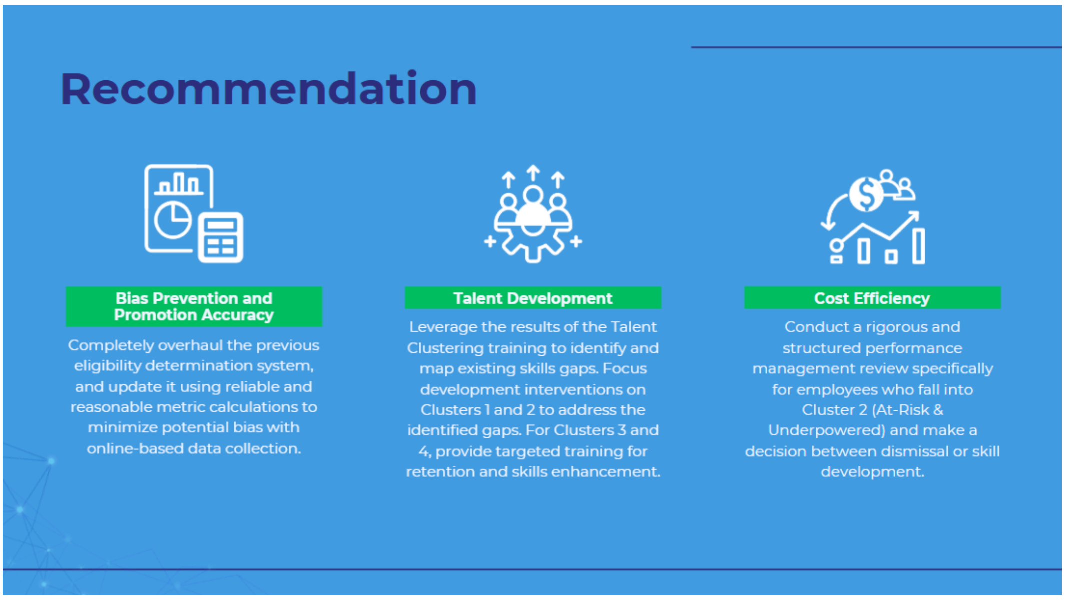
AVG salary based on Dealls!

Level	AVG Salary
Junior	Rp 6.500.000
Middle	Rp 11.500.000
Senior	Rp 22.500.000
Lead	Rp 27.500.000

⬆ 22,5%
Saving by not paying incompetent employees

⬆ 38%
Increased accuracy in determining promotion eligibility

Source: Dealls!



Recommendation

Bias Prevention and Promotion Accuracy

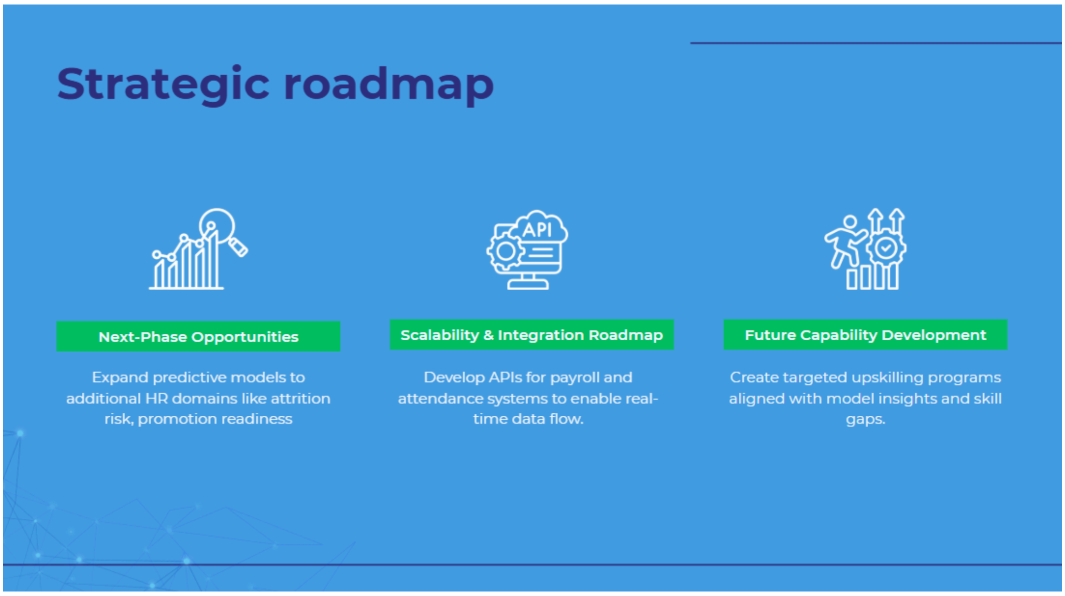
Completely overhaul the previous eligibility determination system, and update it using reliable and reasonable metric calculations to minimize potential bias with online-based data collection.

Talent Development

Leverage the results of the Talent Clustering training to identify and map existing skills gaps. Focus development interventions on Clusters 1 and 2 to address the identified gaps. For Clusters 3 and 4, provide targeted training for retention and skills enhancement.

Cost Efficiency

Conduct a rigorous and structured performance management review specifically for employees who fall into Cluster 2 (At-Risk & Underpowered) and make a decision between dismissal or skill development.



Strategic roadmap

Next-Phase Opportunities

Expand predictive models to additional HR domains like attrition risk, promotion readiness

Scalability & Integration Roadmap

Develop APIs for payroll and attendance systems to enable real-time data flow.

Future Capability Development

Create targeted upskilling programs aligned with model insights and skill gaps.

Thank You!

The Promotion Paradox:
Fixing biased talent promotion decisions through data-driven HR analytics

A project proposal by Syntax Society | Rakamin - Data Science batch 59

Lampiran II. Notulensi Mentoring

Date	Activity	Result
7 November 2025	● Dataset exploration ● Industry problem ● Business understanding	Determination of dataset, problem statement, goal, and business metrics.
13 November 2025	● EDA ● Data preprocessing	● Performing EDA on categorical and numerical features ● Exploring biased data ● Data type determination and data cleaning ● Scaling and Balancing
20 November 2025	● Feature engineering ● Initial modeling	● New feature definition ● Model pipeline creation ● Model consideration and definition
28 November 2025	● Modeling ● Feature Importance	● Establishing evaluation metrics, namely the Silhouette Score and Davies-Bouldin Index ● Rule-based Considerations and Determination ● Feature Importance consideration and definition
4 December 2025	● Deployment ● Report preparation	● Streamlit Presentation ● Preparing the final project PPT and final report ● Business Impact

Lampiran III. Dokumen Teknik

Link Github : SyntaxSociety

Link Streamlit: Dashboard

Link Google Drive: Kelompok 1