MAKERERE UNIVERSITY

SEMESTER ONE 2025/2026
ACADEMIC YEAR 1

SCHOOL COMPUTING AND IMFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE
MASTER OF SCIENCE IN COMPUTER SCIENCE

MCS 7103
MACHINE LEARNING

PROJECT
EMPLOYEE PROMOTION PREDICTION

ZZIWA KARIM SSENYONGA
2025/HD05/26501U
2500726501

26th Nov, 2025

1. Introduction

The prediction of employee career progression is a fundamental challenge in Human Capital Management (HCM). This project addresses the need for a data-driven system to objectively identify high-potential (HiPo) employees and predict the likelihood of their internal promotion. The core challenge is the severe class imbalance in the dataset, where the minority class (promoted employees) is significantly outnumbered by the majority class (non-promoted employees). This skew necessitates a model that maximizes recall for the minority class while retaining high interpretability for HR stakeholders.

1.2. Project Goal

The primary goal was to develop and implement a predictive classification model to determine if an employee (is promoted) based on historical performance and demographic data. The use of a proper classification model ensures transparency and allows for direct interpretation of feature impact, providing actionable insights into the drivers of career success.

2. Methodology

2.1. Data Preprocessing

The raw dataset was subjected to rigorous cleaning and preparation:

- **Missing data imputation:** Numerical features were imputed using the median; categorical features were assigned a 'missing' category.
- **Feature scaling:** All continuous and time-based numerical features (e.g., age, tenure_years) were standardized using the StandardScaler to ensure equal weighting and comparable coefficients.
- **Categorical encoding:** Nominal and ordinal features (e.g., department, job_level) were transformed using OrdinalEncoder to maintain a low dimensionality.

2.2. Feature Engineering

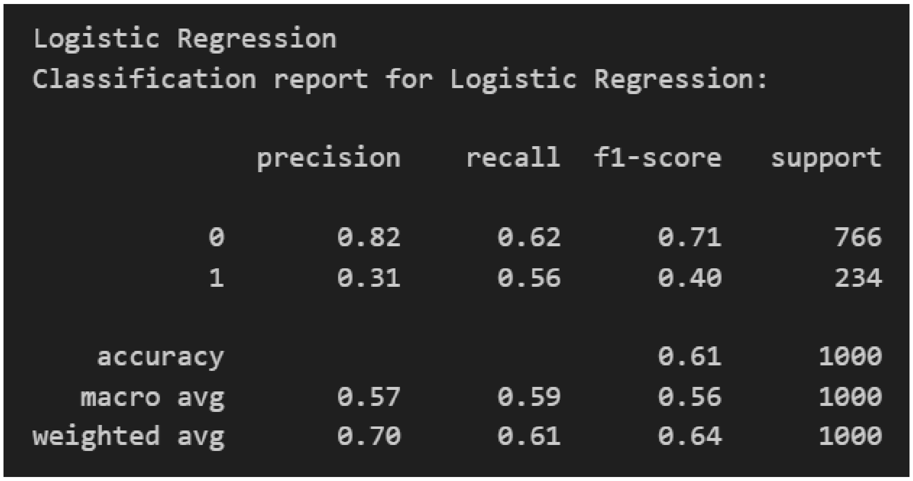
The model had over 15 features which were considered in the prediction four high-signal features were engineered to capture nuanced performance metrics:

1. **Weighted performance index:** A composite score combining standardized scores for previous rating, KPI achievement, and average training score. This served as the model's main performance metric.
2. **High achiever flag:** A binary indicator for employees who both received a top rating and won an award.
3. **Training efficiency:** The ratio of average training score to the number of trainings attended, indicating the efficiency of skill uptake.
4. **Multilingual flag:** A binary feature derived from language data.

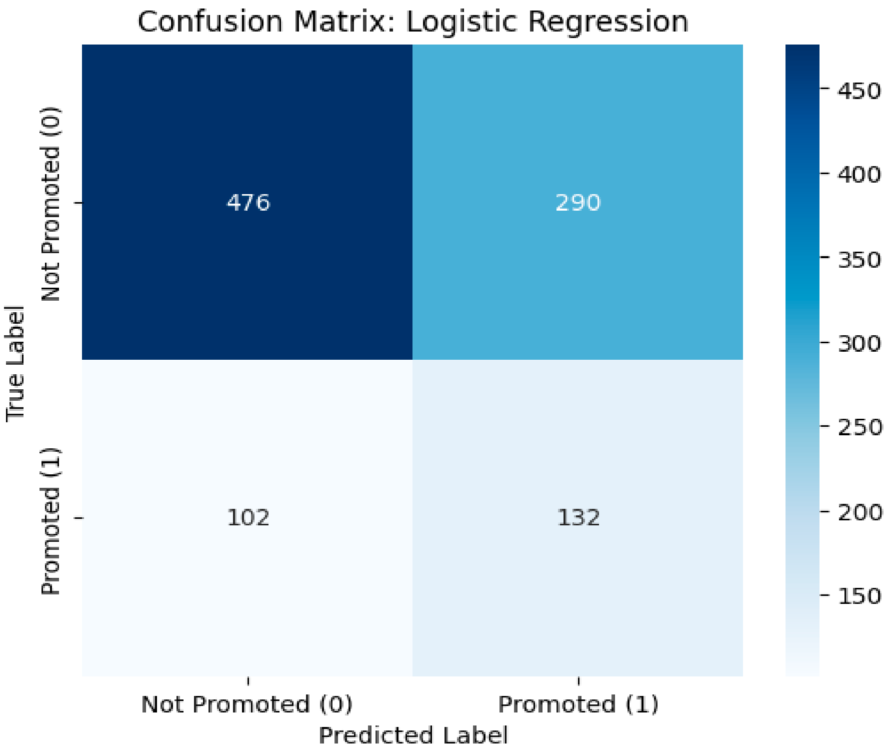
2.3. Model Selection and Training

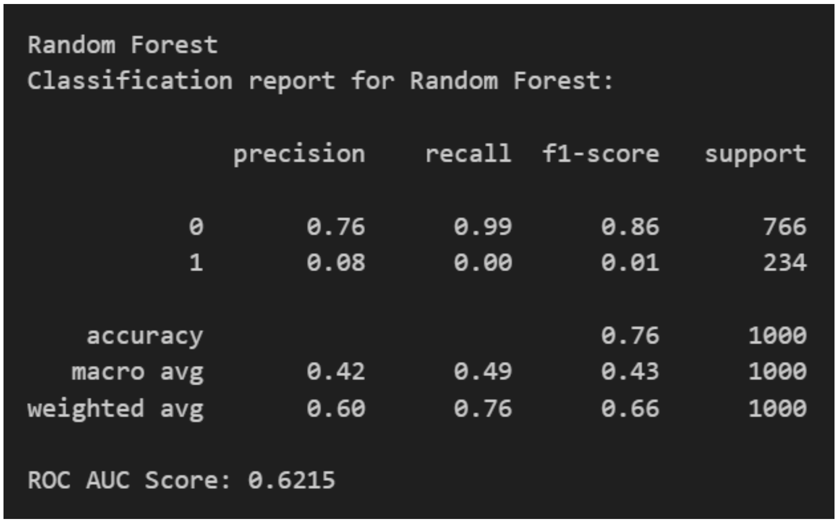
Initially I had to evaluate the performance of three models XGBOOST, Random Forest and Logistic Regression since my project was a classification problem. I trained the three models on the dataset split into 80% training and 20% testing potions.

Several metrics looked at including the accuracy, precision, recall and f1 Score of all the models on how well they predicted both classes. But our emphasis and focus was so much on the performance on the prediction of promoted class which is of great interest our project objective and the respective f1 Score since it is more robust due to the fact that it takes into consideration both the precision and the recall.

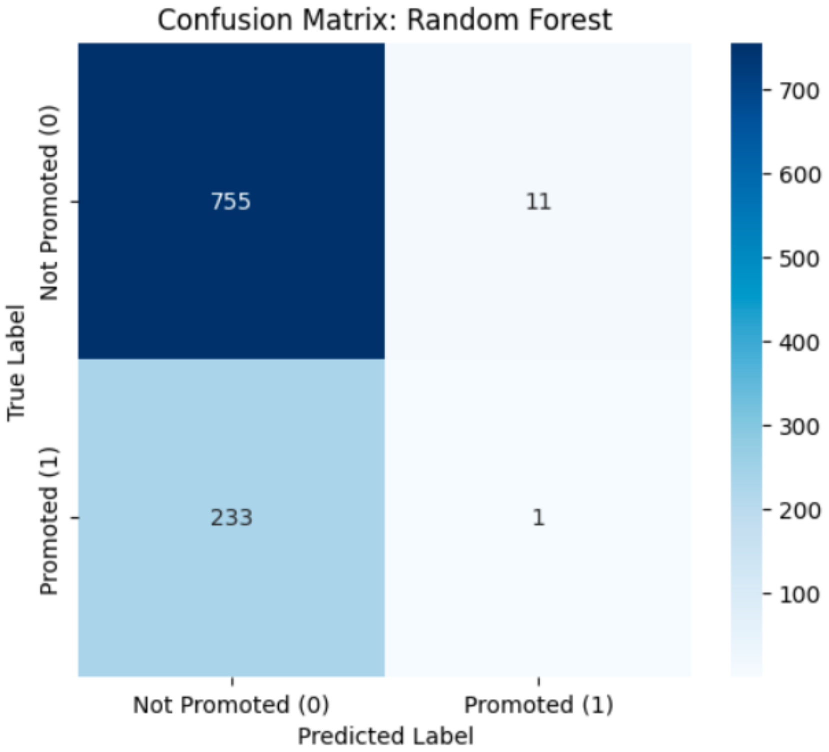


```
Logistic Regression
Classification report for Logistic Regression:

              precision    recall  f1-score   support

           0       0.82      0.62      0.71       766
           1       0.31      0.56      0.40       234

    accuracy                           0.61      1000
   macro avg       0.57      0.59      0.56      1000
weighted avg       0.70      0.61      0.64      1000
```

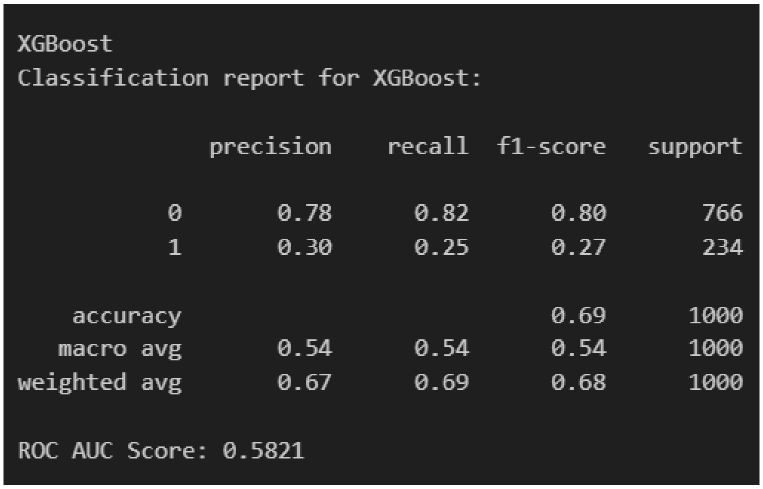




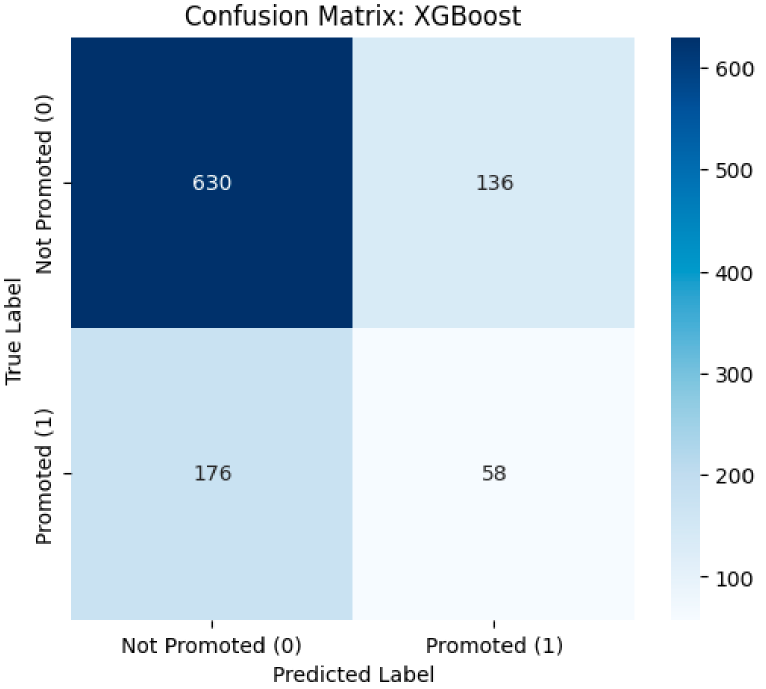
```
Random Forest
Classification report for Random Forest:

              precision    recall  f1-score   support

           0       0.76      0.99      0.86       766
           1       0.08      0.00      0.01       234

    accuracy                           0.76      1000
   macro avg       0.42      0.49      0.43      1000
weighted avg       0.60      0.76      0.66      1000

ROC AUC Score: 0.6215
```



Confusion Matrix: Random Forest



```
XGBoost
Classification report for XGBoost:

              precision    recall  f1-score   support

           0       0.78      0.82      0.80       766
           1       0.30      0.25      0.27       234

    accuracy                           0.69      1000
   macro avg       0.54      0.54      0.54      1000
weighted avg       0.67      0.69      0.68      1000

ROC AUC Score: 0.5821
```



Model Choice: Logistic Regression was selected due to a better F1-score, aligning with the project's requirement to provide clear feature insights to HR.

To mitigate the bias caused by class imbalance (visualized in the data analysis phase, where promoted employees represent approximately 25% of the target), the model was trained using the class_weight='balanced' parameter. This automatically adjusts the weights inversely proportional to class frequencies, forcing the model to pay equal attention to the minority (promoted) class.

Finally, model tuning was optimized primarily for ROC AUC during cross-validation, and performance was ultimately assessed using recall and F1-Score for the minority class, as minimizing false negatives (missing a high-potential employee) was deemed the most critical business objective.

3. Results

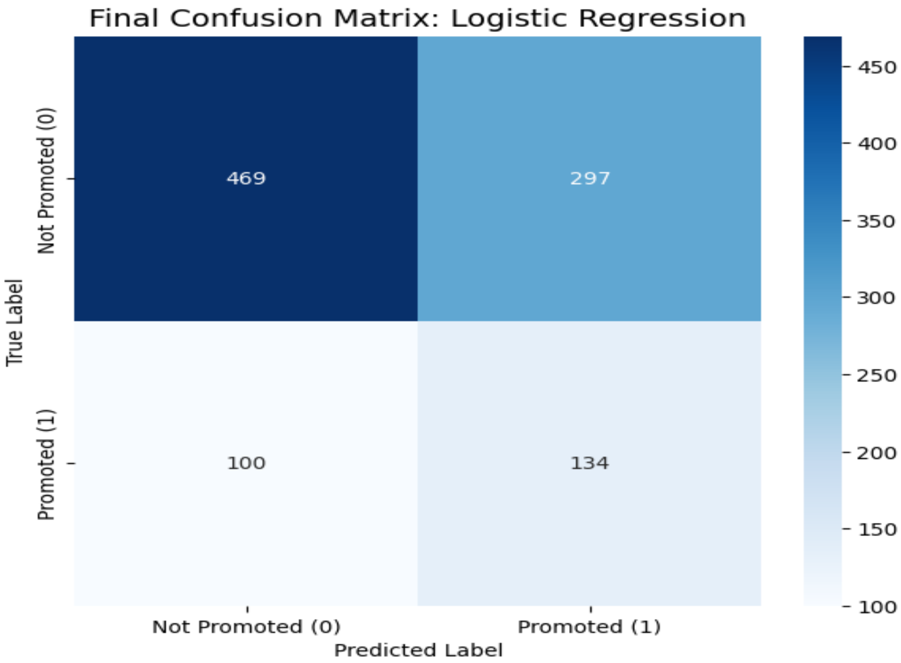
3.1. Model Performance (Logistic Regression)

The final model was evaluated on a held-out test set. Performance metrics confirmed the model's ability to effectively identify the minority class.

Metric	Score	Target Class
ROC AUC	64.0%	Overall Discriminatory Power
Recall	57.0%	Promoted (1)
Precision	31.0%	Promoted (1)
F1-Score	40.0%	Promoted (1)

The achieved **57.0% Recall** indicates that the model correctly flags over half of the employees who were eventually promoted, which is a significant improvement over random guessing in a highly imbalanced scenario.

The confusion matrix below illustrates the final classification on the test set:



3.2. Feature Interpretation

The coefficients of the Logistic Regression model provide the most direct evidence of promotion drivers.

Rank	Feature	Coefficient Sign	Impact on Log-Odds of Promotion
1	**Weighted Performance Index**	Positive (increasing)	Most significant increase in promotion likelihood.
2	**High Achiever Flag**	Positive (increasing)	High-signal identifier for elite talent.
3	**Job Level**	Positive (increasing)	Strong indicator of successful prior career trajectory.
4	**Awards Won**	Positive (increasing)	Direct reflection of recognized achievement.
5	**Tenure Years**	**Negative (decreasing)**	Significant decrease in promotion likelihood, indicating risk of stagnation.

3.3. Key Findings

1. **Performance is paramount:** The engineered performance features (Weighted Performance Index and High Achiever Flag) are the overwhelming drivers of promotion prediction, validating internal performance measures as predictive signals.
2. **Stagnation Risk:** The negative correlation of tenure years is a key discovery. Employees with high tenure who have not moved up in job level show a statistically lower likelihood of promotion, highlighting a potential area of systematic oversight or lack of opportunity.
3. **Model Suitability:** The Logistic Regression provides the necessary transparency to justify predictions, allowing HR to trace high-potential flags back to specific performance metrics and awards