Employee Promotion Prediction – Project Documentation

- **"In many companies, promotion decisions are opaque. Can data predict who gets promoted?"**
- In many companies, promotion decisions are unclear and subjective.
  By analysing employee data, we can find patterns linked to promotions.
  Models like Logistic Regression, RandomForestClassifier or XGBoost can predict likely candidates.
  However, data can't capture everything — like personal traits or bias.
  So, while predictions help, human judgment still plays a key role.

*** Goal : The Goal is to find which features are most helpful for the predicting whether an employee gets promoted or not.**

1. Project Overview

The goal of this project is to predict which employees are likely to be promoted in a given organization using historical employee data(“Train.csv”). This can help HR departments make data-driven decisions and identify key factors influencing promotions.

---

2. Dataset Description

The dataset contains the following columns:

- **employee_id**: Unique identifier for each employee
- **department**: Department where the employee works
- **region**: Employee's region
- **education**: Employee's education level
- **gender**: Employee's gender
- **recruitment_channel**: Channel through which the employee was recruited
- **no_of_trainings**: Number of trainings attended
- **age**: Employee's age
- **previous_year_rating**: Previous year's performance rating from 1 to 5
- **length_of_service**: Years of service in the company
- **KPIs_met >80%**: Whether KPIs were met above 80%
- **awards_won?**: Whether the employee won any awards
- **avg_training_score**: Average score in training
- **is_promoted**: Target variable (1 if promoted, 0 otherwise)

3. Exploratory Data Analysis (EDA)

- **Importing the useful libraries:** For cleaning , EDA, feature engineering , Resampling, model fitting , Evaluation, Model evaluation.
- **Data Cleaning**: Checked for missing values, outliers, and data types. The education column and the previous_year_rating columns have null values.
  Describing the statistical data using the data cleaning methods.
- The data set consists of 54808 rows and 14 columns in which 13 are the features and 1 is the target variable.
- **Feature Engineering**:
  - Created new features such as total_score and service_years labels.
  - Creating the Count plots ,kde and hist plots for checking how each feature is related with the target variable.
  - Creating new_columns and pivot_tables for the data for better understanding.
  - Analysed feature correlations with the target variable. Using "corr" and heatmap to check how each feature is correlated with the target variable("is_promoted_")
- **Visualization**: Used histograms, boxplots, and correlation heatmaps to understand feature distributions and relationships.

4. Model Building

Several machine learning models were trained and evaluated:

- **Logistic Regression**: Used as a baseline model with class balancing.
- **Random Forest Classifier**: Ensemble model to capture non-linear relationships.
- **XGBoost Classifier**: Advanced boosting algorithm for improved performance.

Class Imbalance Handling: Used class_weight='balanced' to address the imbalance in promotion labels.

5. Prediction and Submission

Prediction

The model outputs probabilities for each employee being promoted. To convert these probabilities into binary predictions

Submission File

A submission file is created with only employee_id and is_promoted columns:

6. Feature Importance

- Feature importance was analyzed using model-specific methods (e.g., .feature_importances_ for Random Forest/XGBoost).
- Key features influencing promotion included:
  - previous_year_rating
  - avg_training_score
  - length_of_service
  - KPIs_met >80%
  - awards_won?

7. Results & Evaluation

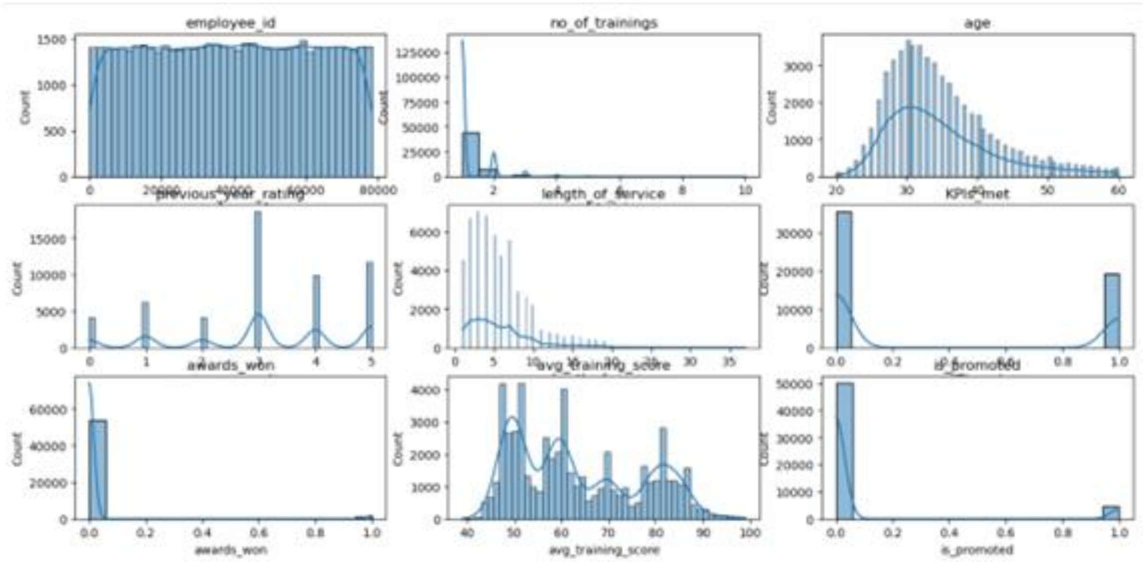
- Models were evaluated using accuracy, precision, recall, and F1-score.
- AUC_ROC curve is plotted in the graph. Each model has the same AUC_ROC value. (80%)
- Cross-validation was used to ensure model robustness.
- The final model achieved satisfactory performance, with recall prioritized due to the business need to identify all potential promotees.

8. Conclusion & Recommendations

- The project successfully identified employees likely to be promoted.
- Feature engineering and handling class imbalance were crucial for model performance.
- The approach can be extended to other HR analytics tasks, such as attrition prediction.

9. Files

1. **Promotion_prediction.ipynb**: Main Jupyter notebook with code and analysis.
2. **submission_file.csv**: Submission file with predictions.
3. **Description of project**: (This documentation)
4. Importing the requires libraires for the data analysis and visualization (NumPy, pandas, matplotlib, seaborn)
5. Loading the dataset and displaying descriptive statistics is a good step for understanding the data.

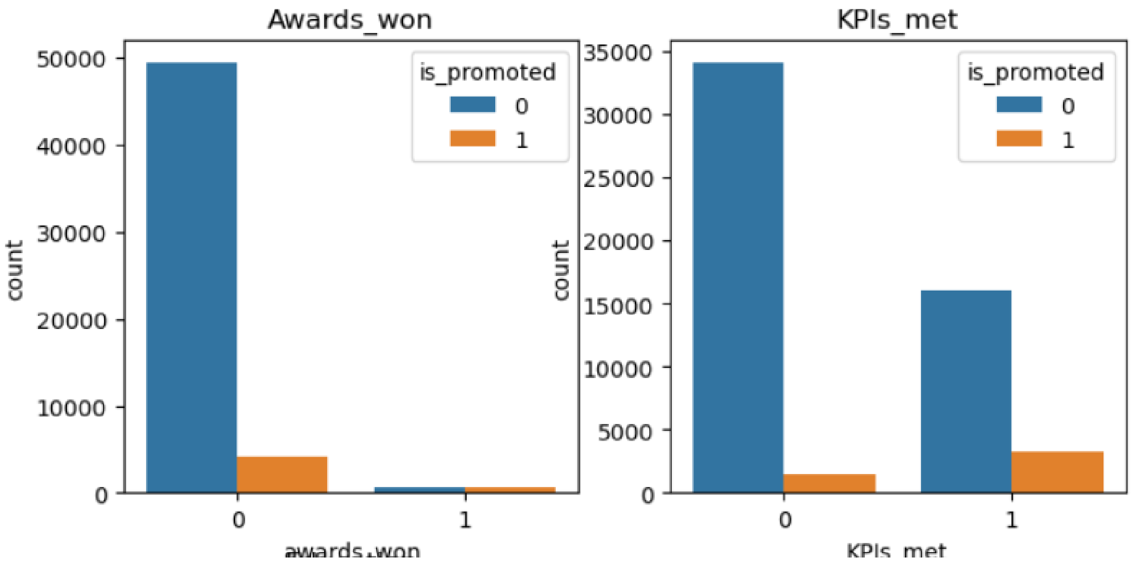


6.

7. **From the graph we can conclude that:**

    1. no_of_trainings: Skewed heavily to the left: Most people took very few trainings (mostly 1).
    2. Age: Bell-shaped distribution centered around 30–35.(close to normal distribution)
    3. Previous_year_rating: A spike around 3 or 4 suggests most employees got mid-to-high ratings.(from 1 to 5)
    4. length_of_service: Skewed right Majority have shorter service durations.
    5. KPIs_met: Most employees did meet their KPIs.
    6. awards_won: Most employees did not win the awards.
    7. is_promoted: it is highly imbalanced.most employees are not employees.
    8. avg_training_score: scores are between 50 and 90..

    9. Imbalanced features like is_promoted, awards_won?, and KPIs_met >80% can affect ML model performance and may need resampling (e.g., SMOTE or undersampling).
    10. Highly skewed features (like no_of_trainings and length_of_service) might need transformation (e.g., log scaling) for some algorithms.
    11. Distribution of avg_training_score could indicate different training impact groups.
    12. age and avg_training_score appear more normally distributed and can be treated differently from binary/skewed ones.



8.

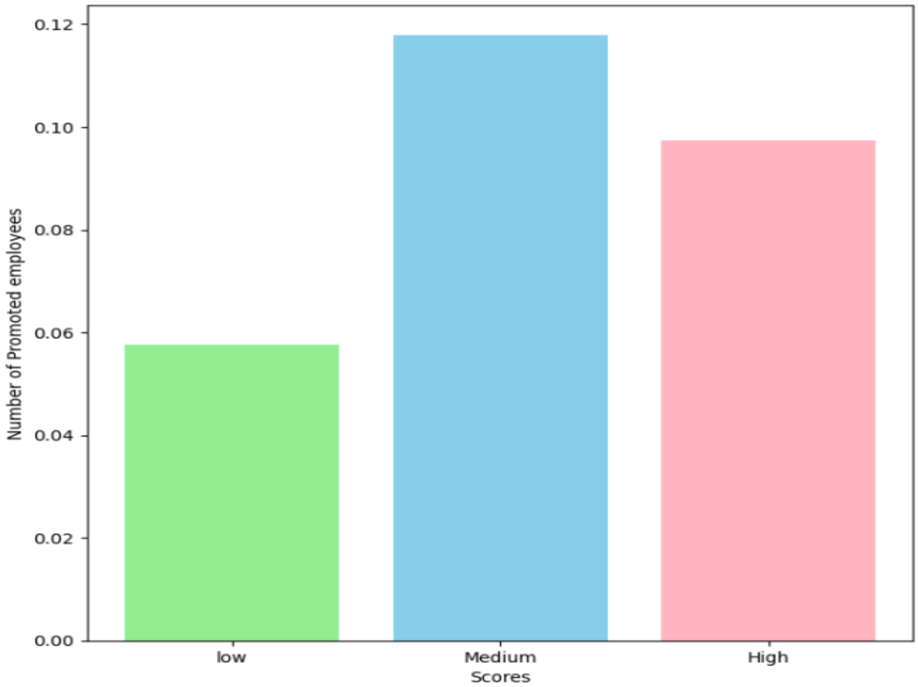
9. **From the countplot we can conclude that:**

   From Awards_Won:

   - Most employees did not win awards.Those who are promoted,a higher proportion won awards compared to those who weren't. so there is slight chance of promotion if employee has won the award

   From KPIs_met:

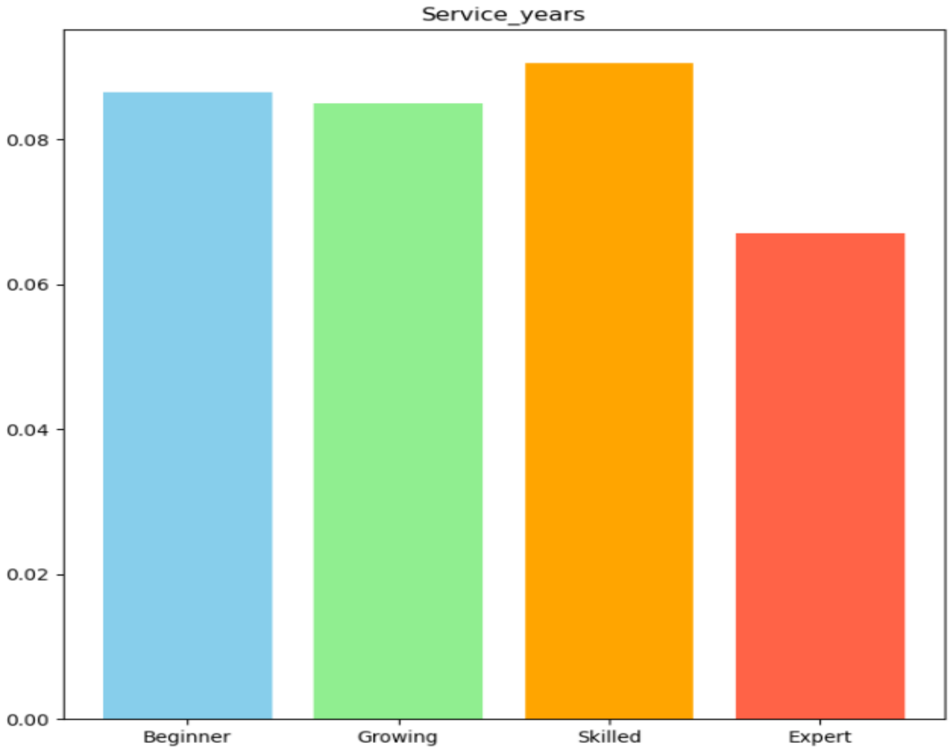
   - A maximum number of employees who were promoted had met KPIs. Most people who haven't met KPIs_met are not promoted. The KPIs_met strongly related with being promoted



10.

Employees promoted have scores in the Medium and High range i.e. 65 and greater Medium score level having the highest percentage of promoted employees means total score is not the only criterion for promotion.

11.



- Skilled employees are likely to get promoted than other categories, due to their experience and understanding of the company Beginner and Growing employees have almost equal likelihood of getting promoted. Expert employees generally get less promotions. Due to their number of years given to the company, majority of them might have reached the success.

12. **After completion of the data cleaning, handling null values, feature engineering. We can conclude that is_promoted is highly dependent on KPIs_mets and also the employee should be the skilled person not a beginner or an expert (because the expert had already seen their success and ready to retire the job/company. So, they are less chance an expert can get promoted.)**

13. **Plotting the correlation heatmap for the data, for the numerical_type to know how each feature is dependent on the target variable and also the on the other features. After checking the values_counts of the is_promoted (Target variable) . We can say that the data is imbalanced and there are less people who are promoted than the people who are not promoted.**
14. **Since the data is imbalanced, we are using the resampling methods for the data to get balanced. Here, the RandomOverSampler is used for sampling the data.44**
15. **The data is splitted into the x features and the y variable. i.e Training set(80%) and the Testing set(20%), which are used to determine how the model performs on new data.**
16. **After splitting, training and testing sets, the training set is used to train the classifier.**
    - **classifier is a type of machine learning algorithm used to categorize data into predefined classes (e.g., Promoted or not promoted).**
    - **The test data is used to evaluate how well the classifier performs on new, unseen data.**

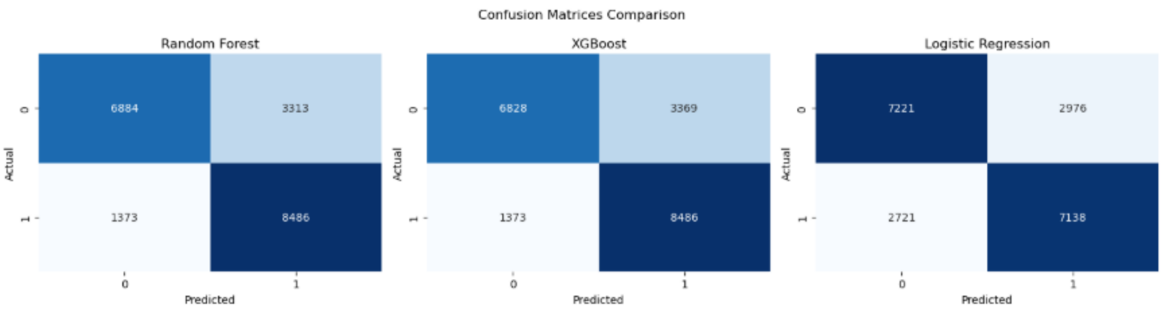
17. **Here, Logistic regression, Random Forest Classifier, XGBoost Classifier are used.**
    - Logistic Regression: Linear model, used to understand basic patterns in the data and compare with complex models.
    - Random Forest Classifier: Ensemble the Bagging method, Suitable when the data has complex relationships that are not captured by the linear models.
    - XGBoost Classifier: Ensemble Boosting method, Effective for handling the imbalanced classes and improving generalization performance.

---------------------------- Model Performance Comparison--

Metric	**Logistic Regression**	**Random Forest**	**XGBoost**
Accuracy	0.7159	0.7664	0.7636
AUC-ROC	0.7963	0.7963	0.7963
Precision (Class 1)	0.71	0.72	0.72
Recall (Class 1)	0.72	0.86	0.86
F1-score (Class 1)	0.71	0.78	0.78

- **Observations:**
  Random Forest and XGBoost outperformed Logistic Regression in terms of accuracy, recall, and F1-score.
  Random Forest and XGBoost provided a better balance of precision and recall, especially for the positive class (label 1).
  Logistic Regression had balanced performance but lacked the power to capture complex patterns in the data.
  So, we can conclude that Random Forest and the XGBoost have high performance then the Logistic Regression.

18. Confusion Matrix Comparison:



- Among the three models, Random Forest and XGBoost outperformed Logistic Regression in terms of accuracy and recall.
  Both ensemble models effectively identified the positive class with high true positive rates.
  Logistic Regression was more conservative but resulted in higher false negatives.
  All models had similar AUC-ROC scores, indicating comparable discrimination ability.
  Overall, Random Forest or XGBoost is recommended for better balanced and robust performance.

19. After fitting and evaluating the values, the are going to create a submission file where the file consists of two columns (i.e. Employee_id, and is_promoted_pred) we use the predicted promoted values for the file. And save the file with a .csv extension.

* RECOMMENDATIONS AND INSIGHTS.

1. Which factors most influence an employee's promotion in company?

Answer:
The most influential factors identified by the model include previous year rating, average training score, length of service, whether key performance indicators (KPIs) were met, and whether the employee has won awards.

2. What actions should we take based on the model results?

Answer:

- Focus on developing employees who are close to the promotion threshold by offering targeted training or mentoring.

3. Are there patterns or trends among employees who are promoted?

Answer:
Yes, analysis shows that employees who are promoted typically have higher previous year ratings, consistently meet or exceed KPIs, participate in more training programs, and often have longer tenure with the company. They are also more likely to have received awards or formal recognition.

4. Are there any groups (departments, regions, demographics) with lower or higher promotion rates?

Answer:
Yes, the analysis found that certain departments and regions have higher promotion rates, while others are underrepresented. For example, employees in the Sales and Marketing department and those based in Region 2 were promoted more frequently than those in other groups. This suggests a need to review promotion practices for fairness and consistency across all groups.

5. What are the most common barriers preventing employees from being promoted?

Answer:
The most common barriers are low previous year ratings, not meeting KPIs, and insufficient participation in training programs. Addressing these areas through targeted support can increase promotion rates.

- **Conclusion:**
  This project used the data driven methids and machine learning models to predict the employee promotions within the organization. Through analysis, we identified that the key factors or features that strongly influence the target variable is previous year ratings, KPIs_met, training scores, Length of service, awards_won.

  The most common mistakes to avoid when seeking a promotion are failing to communicate your achievements, not taking initiative, neglecting professional development, and assuming hard work alone is enough.