HR ANALYTICS: PREDICTING EMPLOYEE PROMOTIONS

Data Driven Insights for fair and effective Promotion Decisions

Our Team

Bedan Kibunja

Team Head
Deployment Lead

Monica Onyango

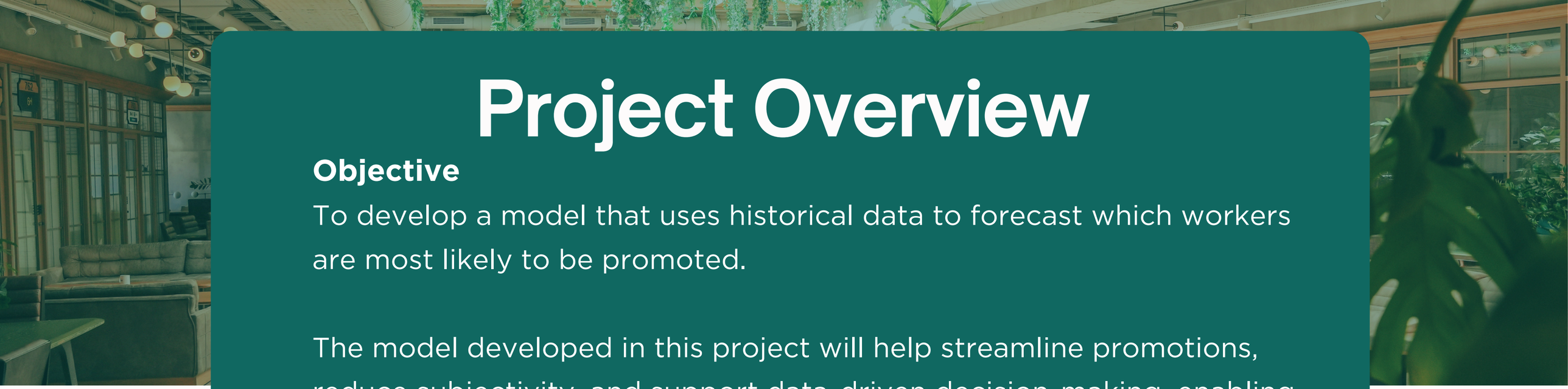
Data Analysis and
Feature
Engineering Lead

Grace Wacheke

Modelling and
Presentation Lead

Martin Kabare

Graphic Design



Project Overview

Objective

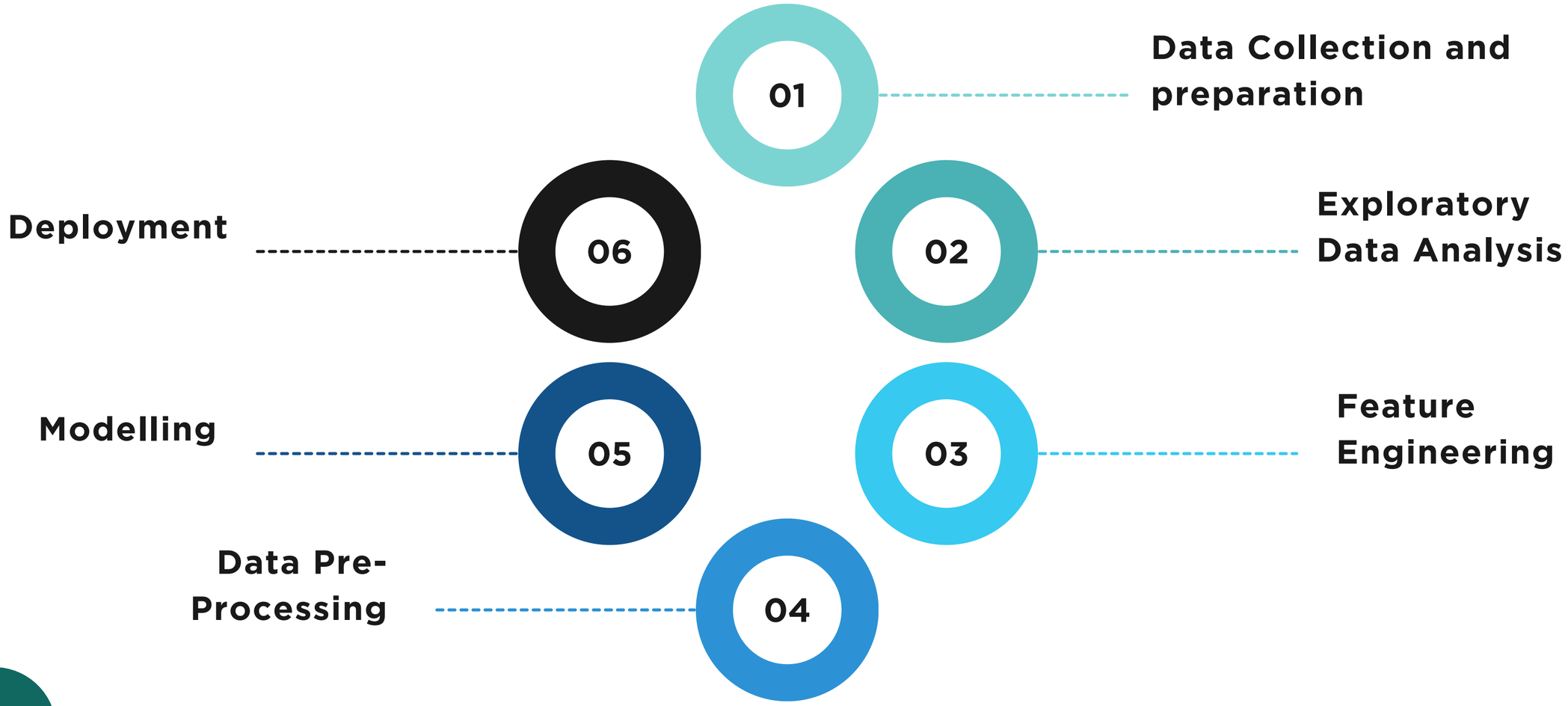
To develop a model that uses historical data to forecast which workers are most likely to be promoted.

The model developed in this project will help streamline promotions, reduce subjectivity, and support data-driven decision-making, enabling timely and equitable identification of high-potential candidates.

Stakeholders

- HR Department
- Department Heads and Team Managers
- Executive Leadership

WORKFLOW OVERVIEW



Business Problem

The current promotion process for managerial positions and below is manual, time-consuming, and perceived as biased, with concerns about favoritism affecting advancement opportunities.

To improve fairness and efficiency, the HR department aims to implement a predictive model that uses demographic and performance data to assess promotion eligibility objectively.

"The best way to predict the future is to create it."

— Peter Drucker

Data Understanding

The dataset used for this project was sourced from an ongoing hackathon hosted on Analytics Vidhya website.

The target variable is the **is_promoted** feature indicating if the employee was recommended for promotion (1) or not (0).

Training Data

Contains information on current employees and a target variable indicating if they were promoted. Contains 54,808 records and 14 columns

Testing/Validation Data

Similar dataset to the traing data, without the target label, used for model evaluation. Comprises 23,490 records and 13 columns

Data Preparation

Missing Values

Records with missing values in the **previous_year_rating** and **education** features were dropped

Duplicates

The dataset had no duplicate records

Outliers

Outliers were identified in the **age** and **length_of_service** columns, but they represent plausible values within the employee data context.

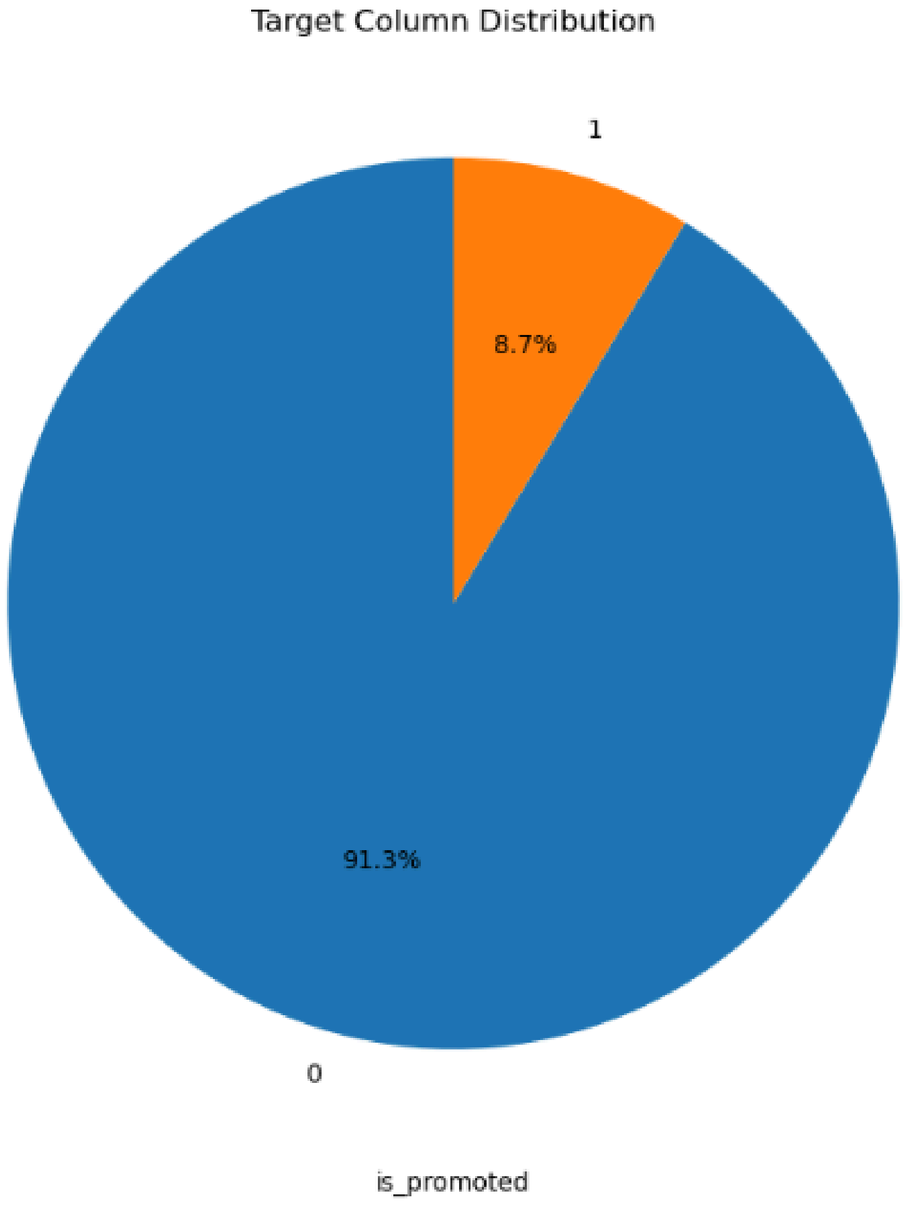
Exploratory Data Analysis

Target Column analysis

The dataset displayed a substantial ***imbalance*** in promotion eligibility, with only 8.7% of employees receiving promotions and 91.3% not being promoted.

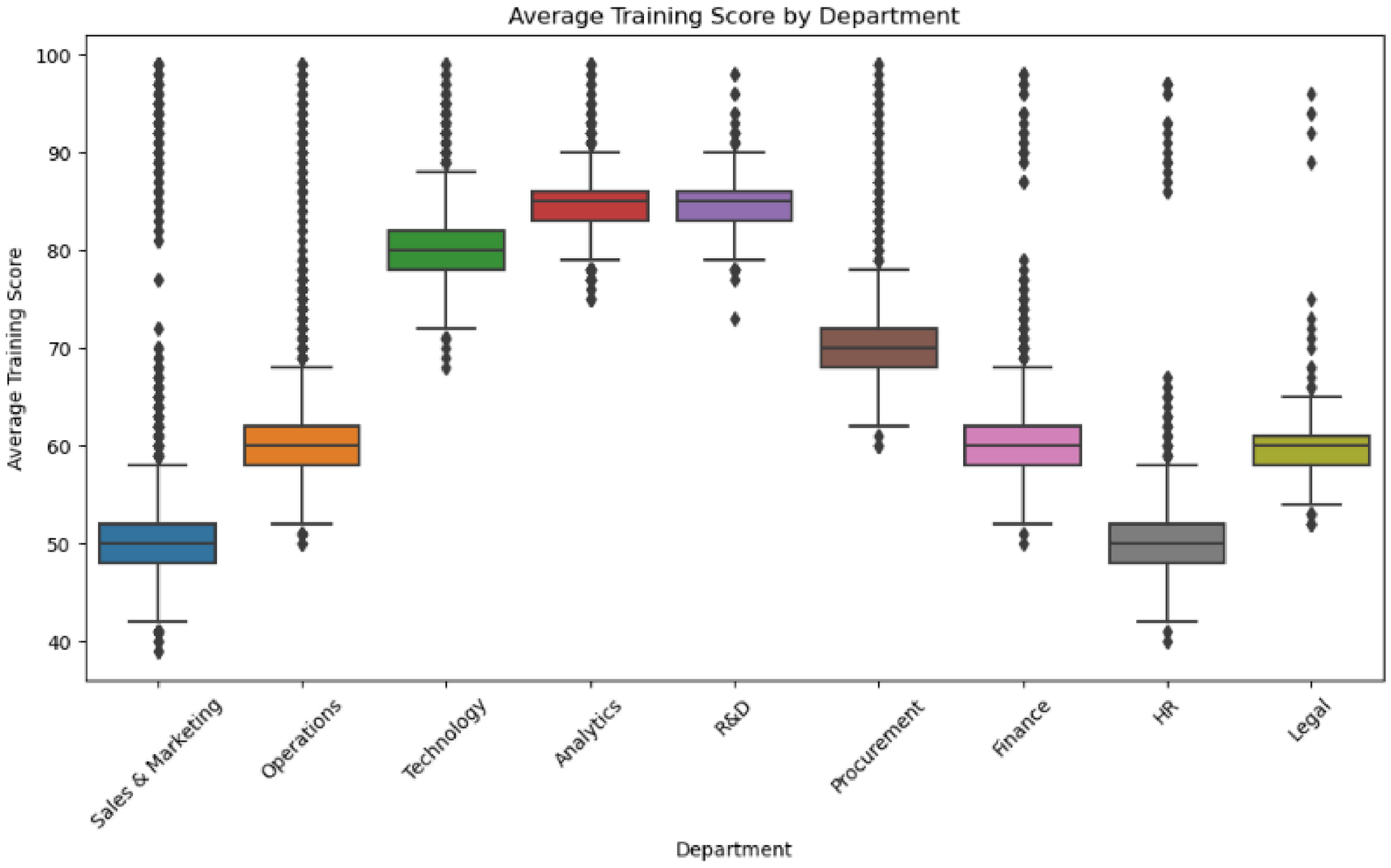
Class Balancing Techniques applied

- ADASYN balancing technique (variation of SMOTE)
- Weighted Models



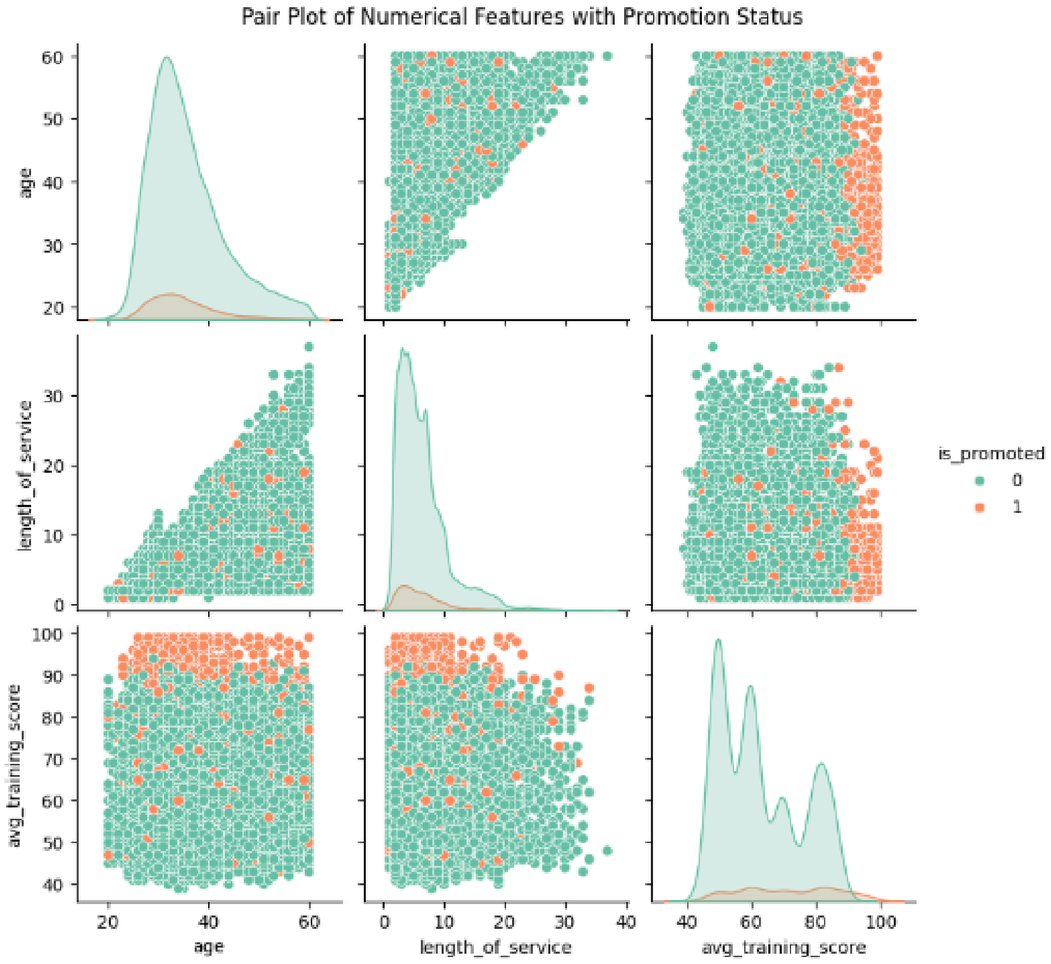
Training Score Comparison Across Departments

The Technology and Analytics departments have higher median training scores, generally between 80 and 90, while HR and Operations have lower medians around 60.



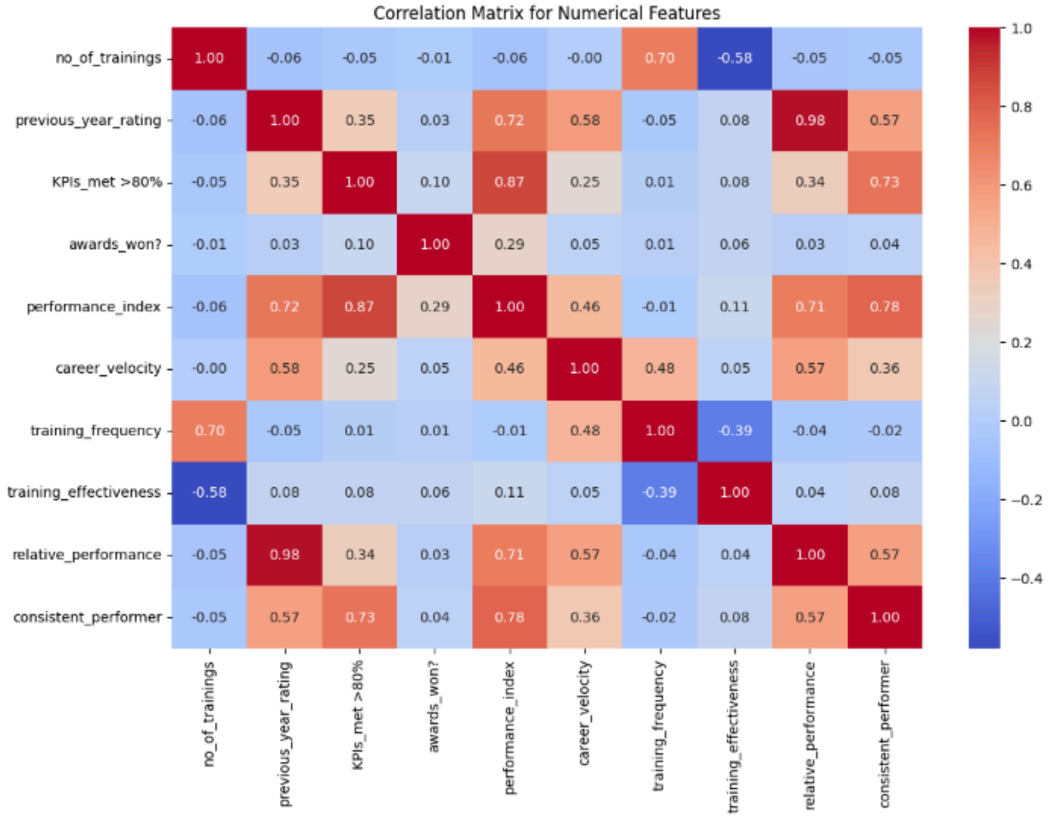
Pair Plot of Numerical Features with Promotion Status

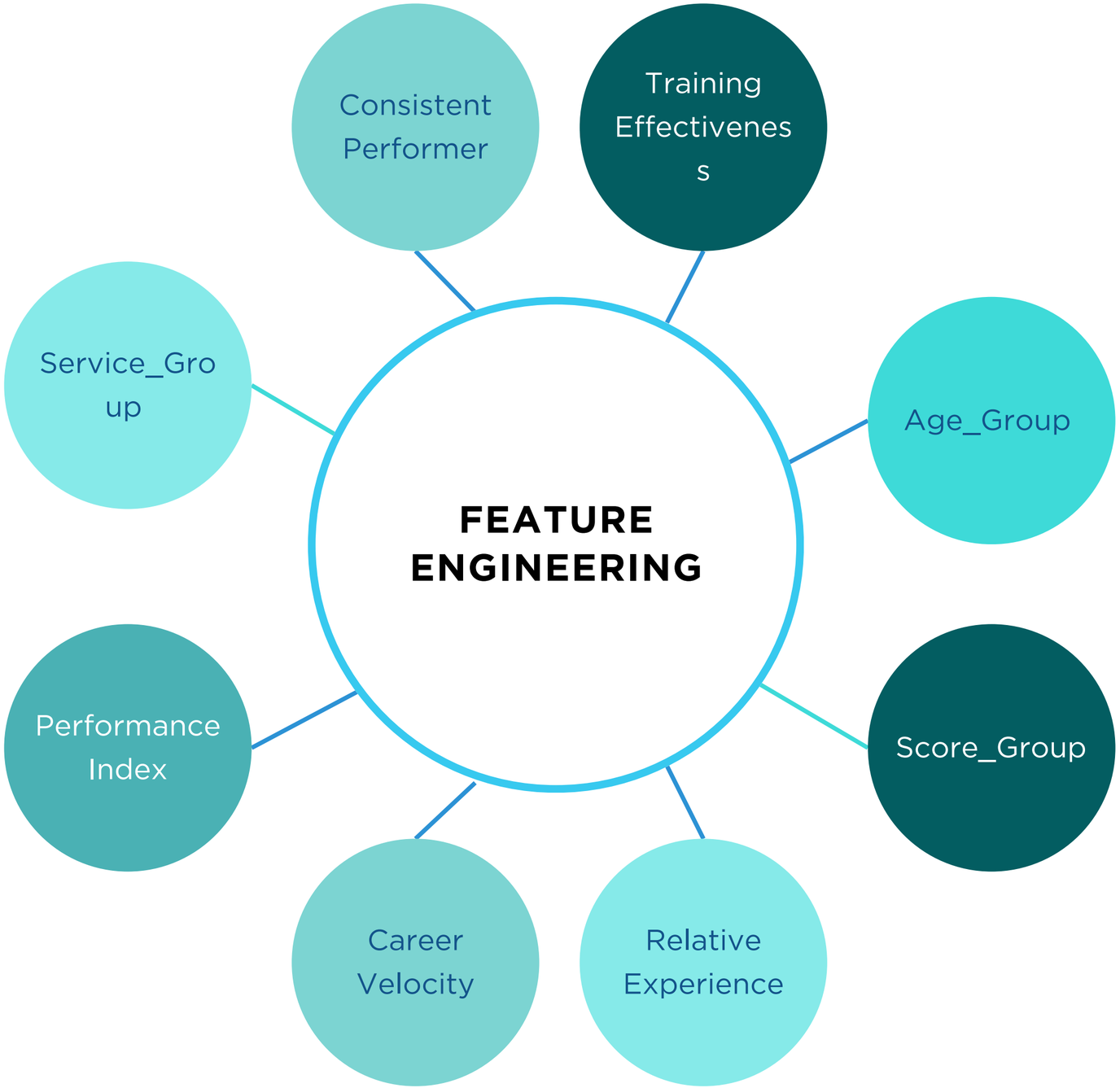
- There’s a strong positive correlation between age and length of service
- Average training score shows no significant correlation with either age or length of service,
- Regarding promotion, promoted employees (orange points) appear distributed similarly across the age and length of service ranges but are more concentrated in higher average training scores.
- Age and tenure show less of an impact on promotion likelihood.



Feature Correlation Matrix

- **previous_year_rating** and **relative_performance**: These features have a very high correlation (≈ 0.98), suggesting that they capture nearly identical information.
- **KPIs_met >80%** and **performance_index**: With a correlation of around 0.87, these two features are also highly correlated.
- **performance_index** and **consistent_performer**: There is a moderate correlation (≈ 0.78) between these features, which may indicate some overlap in the information they provide.




Consistent Performer
Training Effectiveness
Service_Group
Age_Group
FEATURE ENGINEERING
Performance Index
Score_Group
Career Velocity
Relative Experience

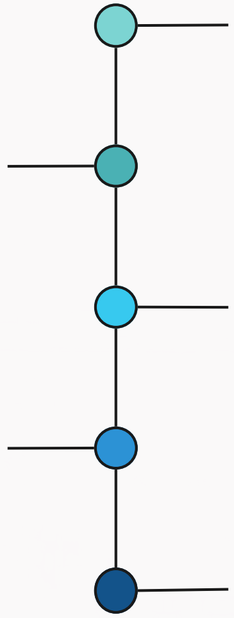
Data Pre-Processing

Splitting traing dataset into the predictor and target variables

train test split was used

Balancing the target variable

ADASYN (Adaptive Synthetic Sampling) is used



Encoding Categorical Features

LabelEncoder was used

Feature Importance Assessment

Scaling the data

Ensuring that each feature contributes equally to model performance.

Modelling

Base Bagging and Boosting Models

- XGBClassifier - 0.481
- RandomForest - 0.426
- DecisionTree - 0.420
- GradientBoosting - 0.3340

Tuned Bagging and Boosting Models

- XGBClassifier - 0.491
- RandomForest - 0.460
- DecisionTree - 0.410
- GradientBoosting - 0.483

Model ensembling

- XGBClassifier, GradientBoosting, and RandomForest - 0.479
- GradientBoosting and XGBClassifier - 0.485

Weighted Models

- LGBMClassifier - 0.452
- BalancedBagging - 0.415
- XGBClassifier with scale_pos_weight - 0.439

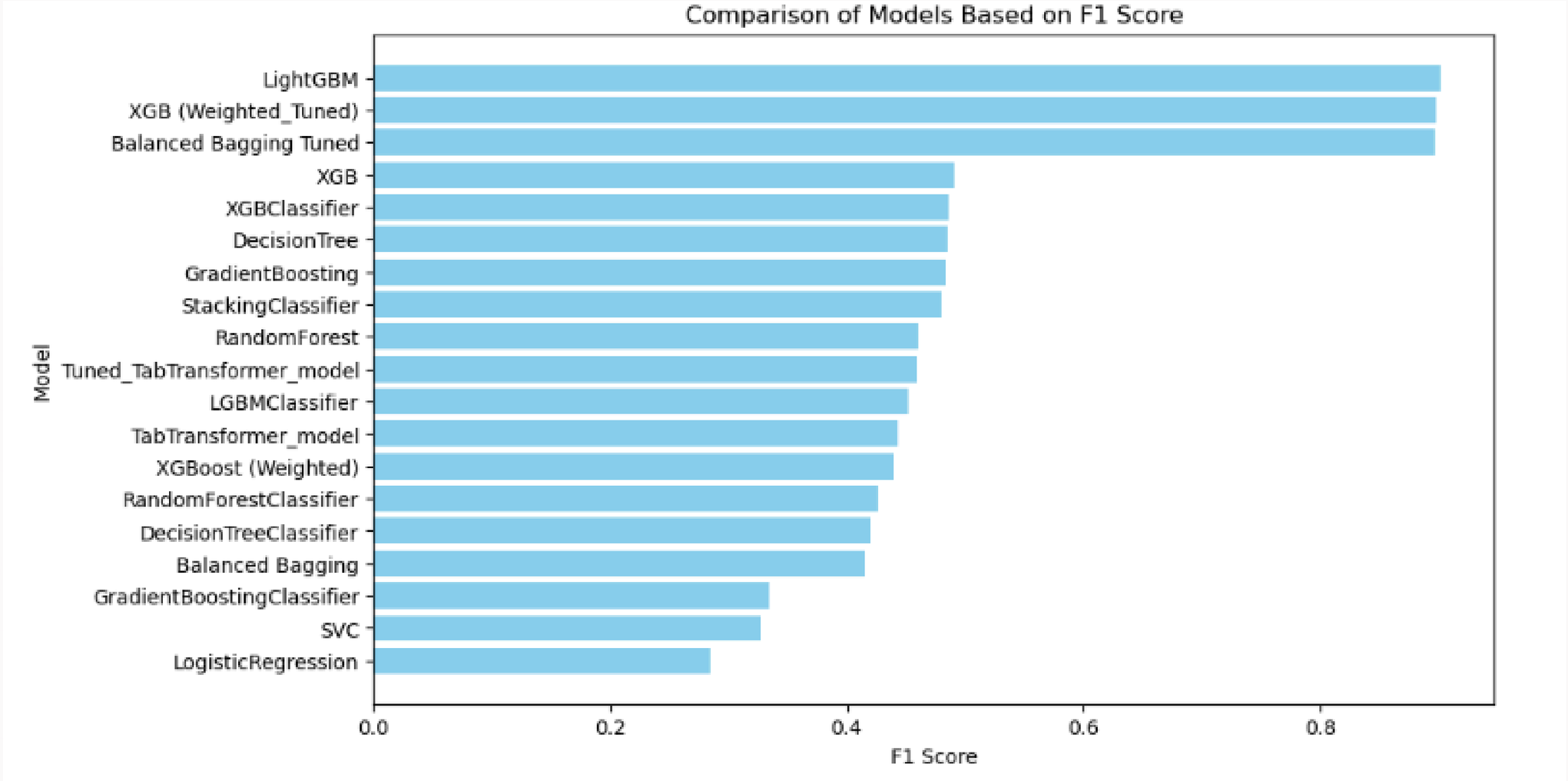
Tuned Weighted Models

- LGBMClassifier - 0.902
  BalancedBagging - 0.896
- XGBClassifier with scale_pos_weight - 0.897

Neural Network

- TabTransformer - 0.443
- Tuned TabTransformer - 0.459

Modelling


Comparison of Models Based on F1 Score
LightGBM
XGB (Weighted_Tuned)
Balanced Bagging Tuned
XGB
XGBClassifier
DecisionTree
GradientBoosting
StackingClassifier
RandomForest
Tuned_TabTransformer_model
LGBMClassifier
TabTransformer_model
XGBoost (Weighted)
RandomForestClassifier
DecisionTreeClassifier
Balanced Bagging
GradientBoostingClassifier
SVC
LogisticRegression
Model
0.0
0.2
0.4
0.6
0.8
F1 Score

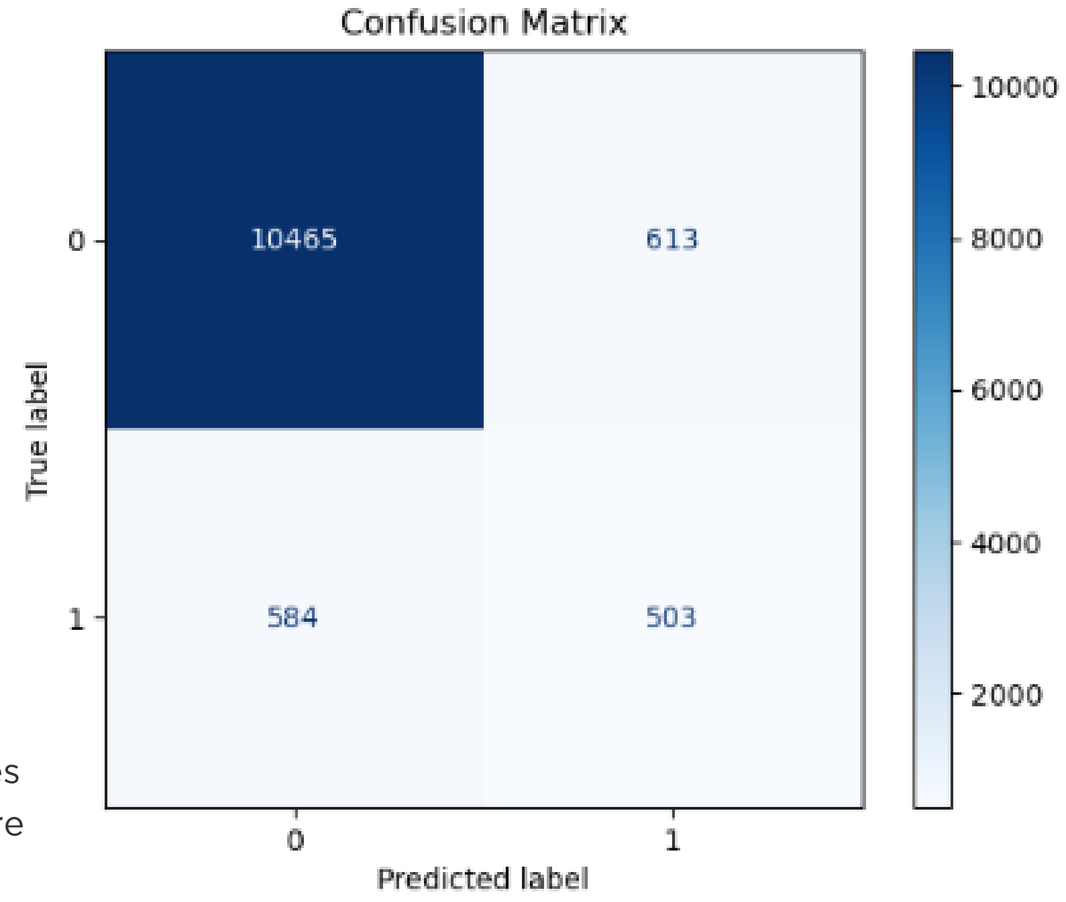
Evaluating the Best Model

LightGBM Model

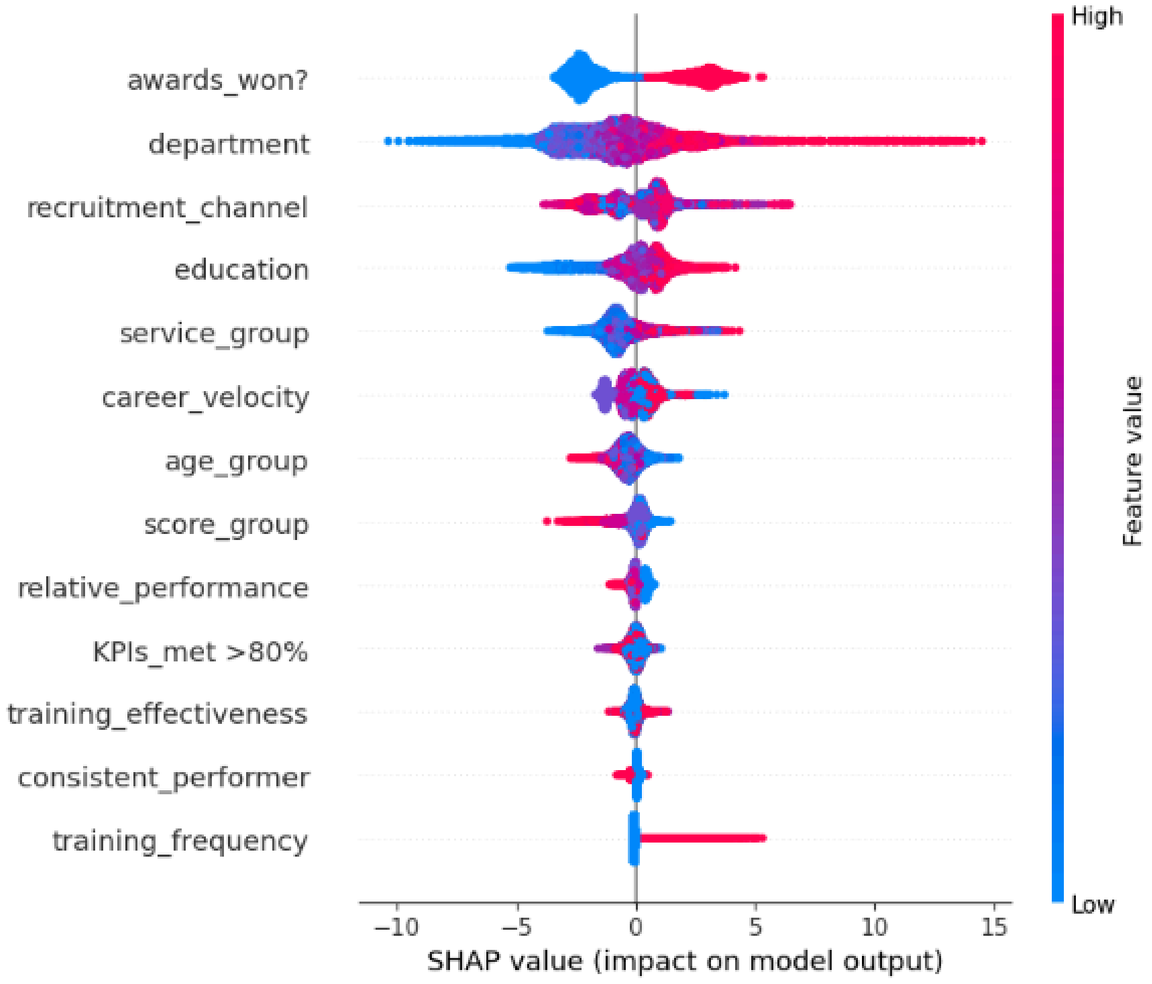
The LightGBM model was selected for multiple reasons: it demonstrated improvement in the critical F1 score metric (0.9), was straightforward to implement, and provided a means to address the minority class by giving it greater emphasis in the loss function.

Evaluation

- The model performs well in identifying the majority class (0), with a high count of true negatives (10,476) compared to false positives (602). However, it struggles more with the minority class (1), where there are a notable number of false negatives (591).



Feature Importance



- The plot shows that features like **awards_won**?, **department**, and **recruitment_channel** have a significant impact on the model's predictions.

- High values for features like **awards_won**? and **consistent_performer** (shown in red) tend to push predictions towards the positive outcome.

Deployment

The model was deployed using Streamlit, creating an interactive web application that HR teams can use to input employee data and receive promotion eligibility predictions in real-time.

Challenges

TARGET VARIABLE CLASS IMBALANCE

The main challenge encountered in this project is dealing with the imbalance in the is_promoted column

All the models strained with the minority class

Conclusions

- The data highlights a significant imbalance, with a small percentage of employees eligible for promotion.
- Extensive feature engineering and careful feature selection was done to boost model performance
- Initial Model Performance and Tuning Needs: Early model tests indicated low F1 scores, highlighting the need to balance precision (identifying true promotion candidates) and recall (minimizing false negatives).

Recommendations

- Expand Feature Engineering for Richer Insights
- Establish Regular Model Audits
- Implement Techniques to Address Data Imbalance
- Explore advanced hyper-parameter tuning

THANK YOU