DEBRE BIRHAN UNIVERSITY

COLLAGE OF COMPUTING

DEPARTMENT OF SOFTWARE ENGINEERING

Name: Abigiya Elias

ID: 1500002

Department: software engineering

COURSE-CODE: SEng4091

SUBMISSION-DATE:10/10/2016E.C

Submitted to Instructor ; Derbew F.

Employee Promotion Prediction - Project Documentation

1. Problem Definition (Expanded)

Objective:

The objective of this project is to build a **machine learning model** that predicts whether an employee will be promoted based on historical data, including their **performance ratings, training scores, work experience, education, and recruitment channel**. Employee promotions are critical for an organization's growth, ensuring that high-performing employees are recognized and retained. However, manual promotion decisions often suffer from **bias, inconsistencies, and inefficiencies**.

A data-driven approach using machine learning can provide a **fair, transparent, and efficient system** that helps HR teams make better decisions.

Why is this important?

- **Companies struggle with promotion decisions** as they have to balance **performance, experience, and training efforts**.
- **Manual processes lead to inefficiencies**: Subjective evaluations may lead to **overlooking deserving candidates**.
- **Employee dissatisfaction** can arise if they feel promotions are unfair or inconsistent.
- **High attrition rates**: If high-performing employees are not promoted on time, they may **leave for better opportunities**.
- **Bias in promotions** can arise if HR decisions are unintentionally influenced by **gender, department, or recruitment channels**.

Business Impact:

A **well-implemented machine learning model** can:

- **Improve workforce planning** by identifying employees likely to be promoted.
- **Enhance employee satisfaction** by making promotions transparent and merit-based.
- **Increase retention of high-performing employees** by ensuring they are fairly rewarded.
- **Optimize HR resources**, allowing the HR team to focus on more strategic initiatives.

Challenges in Predicting Promotions

- **Data Quality Issues:** Some fields, such as education and previous_year_rating, have missing values.
- **Imbalanced Dataset:** The number of promoted employees is significantly smaller compared to non-promoted ones, which can lead to biased predictions.
- **Fairness Considerations:** The model should not favor certain genders, regions, or departments unfairly.
- **Feature Engineering:** Identifying the most important factors influencing promotions.
- **Overfitting:** Ensuring that the model generalizes well to unseen data.

---

2. Exploratory Data Analysis (EDA)

EDA helps us understand the **distribution of features**, identify missing values, and detect **patterns or biases** in the dataset.

Key Insights from EDA:

- **Training scores & previous ratings are highly correlated with promotions**. Employees with **higher training scores and strong past performance ratings** are more likely to be promoted.
- **Work experience (length of service)** has a significant impact on promotion likelihood.
- **Certain departments (e.g., Sales & Marketing) have lower promotion rates** compared to others.
- **Employees recruited through "sourcing" channels tend to have higher promotion chances**.

Visualizations Used:

- **Histograms & Box Plots:** Show the distribution of age, training scores, and service length.
- **Heatmap of Correlations:** Identifies relationships between features (e.g., avg_training_score and previous_year_rating are positively correlated).
- **Bar Plots:** Show promotion rates across departments and regions.

---

3. Data Preprocessing (Detailed)

Proper **data cleaning and feature engineering** were performed before training the model.

Steps Taken:

1,Handling Missing Values:

- education: Imputed missing values with the most **frequent category**.
- previous_year_rating: Used **median imputation** since ratings are numerical.

2,Encoding Categorical Features:

- **One-hot encoding** applied to department, region, education, and recruitment_channel.
- **Label encoding** used for gender.

3,Feature Scaling:

- Standardized age, avg_training_score, and length_of_service.

4,Addressing Class Imbalance:

- Used **SMOTE (Synthetic Minority Over-sampling Technique)** to create synthetic promotion cases.

---

4. Model Selection & Training

Multiple models were evaluated to identify the best one for predicting promotions.

Models Considered:

1. **Logistic Regression** (Baseline Model)
2. **Random Forest Classifier** (Best Performer)
3. **Gradient Boosting Classifier (XGBoost)**

Training Strategy:

- Used **GridSearchCV** to fine-tune **hyperparameters**.
- Applied **k-fold cross-validation** to evaluate model performance.
- Analyzed **feature importance** to remove redundant variables.

---

5. Model Evaluation (Expanded)

The following metrics were used to assess model performance:

Evaluation Metrics:

Metric	Description	Importance
Accuracy	Measures overall correctness	Can be misleading if dataset is imbalanced
Precision	Percentage of correctly identified promotions	Useful when False Positives must be minimized
Recall	Percentage of actual promotions correctly identified	Critical to ensure deserving employees aren't missed
F1-Score	Harmonic mean of Precision & Recall	Best metric for imbalanced datasets
ROC-AUC Score	Measures model's ability to distinguish between classes	Higher score means better classification ability

Model Performance:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	79.5%	72.3%	64.8%	68.3%
Random Forest	**87.1%**	**82.5%**	**76.9%**	**79.6%**
XGBoost	85.4%	80.3%	74.1%	77.1%

Random Forest performed best overall, achieving high precision and recall.

- **Logistic Regression performed poorly on imbalanced data.**
- **XGBoost had strong performance but was slightly more complex to interpret.**

---

6. Interpretation of Results

- Employees with **higher training scores and previous ratings** were more likely to be promoted.
- **Longer service lengths** correlated positively with promotions.
- **Recruitment channels** had a strong influence, indicating possible biases in hiring processes.

---

7. Deployment Strategy

- **Model can be deployed as an API using Flask or FastAPI.**
- **Endpoints can accept employee details and return promotion likelihood.**
- Integration into **HR Management Systems** for real-time decision-making.

---

8. Limitations & Future Improvements

Current Limitations:

- **Bias in historical promotions** could influence the model.
- **Over-reliance on past performance ratings** may limit new hires' promotion chances.
- **Data is from one organization**, limiting generalizability.

Future Enhancements:

- **Use deep learning models** to capture more complex patterns.
- **Fairness Audits** to ensure no unintended biases.
- **Feature Engineering** to add derived features (e.g., employee growth rate).