Model Report

Job Postings Fraud Detection

Introduction

This report presents an analysis of machine learning models applied to the Job Postings dataset, aiming to predict whether a job posting is **fraudulent** based on various job-related features. The models are evaluated using metrics such as **Accuracy, Precision, Recall, and F1-score**. The goal is to identify fraudulent postings efficiently to improve recruitment integrity and reduce scam incidents.

2. Dataset Overview

- **Number of Records:** 17,880
- **Number of Features:** 18
- **Target Variable:** fraudulent (0 = Not Fraudulent, 1 = Fraudulent)
- **Data Types:**
  - Integer: 5 columns (job_id, telecommuting, has_company_logo, has_questions, fraudulent)
  - Object/Text: 13 columns (title, location, department, salary_range, company_profile, description, requirements, benefits, employment_type, required_experience, required_education, industry, function)

Feature	Description	Data Type	Missing Values
job_id	Unique identifier for each job posting	int64	0
title	Job title	object	0
location	Job location (city/state/country)	object	346
department	Department of the job	object	11,547
salary_range	Offered salary range	object	15,012
company_profile	Company profile description	object	3,308
description	Full job description	object	1
requirements	Job requirements	object	2,696
benefits	Benefits offered by the company	object	7,212
telecommuting	Indicates if telecommuting is allowed (0 = no, 1 = yes)	int64	0

Feature	Description	Data Type	Missing Values
has_company_logo	Indicates if the posting has a company logo (0 = no, 1 = yes)	int64	0
has_questions	Indicates if the posting includes questions (0 = no, 1 = yes)	int64	0
employment_type	Type of employment (Full-time, Part-time, Contract, etc.)	object	3,471
required_experience	Required experience level (e.g., 1-3 years)	object	7,050
required_education	Required education level (e.g., Bachelor's, Master's)	object	8,105
industry	Industry category of the company	object	4,903
function	Job function or role	object	6,455
fraudulent	Target variable indicating fraud (0 = Not Fraudulent, 1 = Fraudulent)	int64	0

4. Data Preprocessing

1. **Handling Missing Values:**
    - Columns with missing values were considered for imputation.
    - Moderate missing values were imputed using suitable strategies
2. **Encoding Categorical Features:**
    - Label Encoding applied for features with low cardinality.
    - One-Hot Encoding applied for high-cardinality categorical features.
3. **Scaling:**
    - Standard Scaling applied to numerical features for models sensitive to feature magnitude.
4. **Target Variable:**
    - fraudulent column used as the target for classification.

5. Models and Evaluation

Models Evaluated

- Logistic Regression (baseline linear model)
- Random Forest Classifier (ensemble method)
- XGBoost (gradient boosting ensemble)
- LightGBM (gradient boosting with optimized performance)

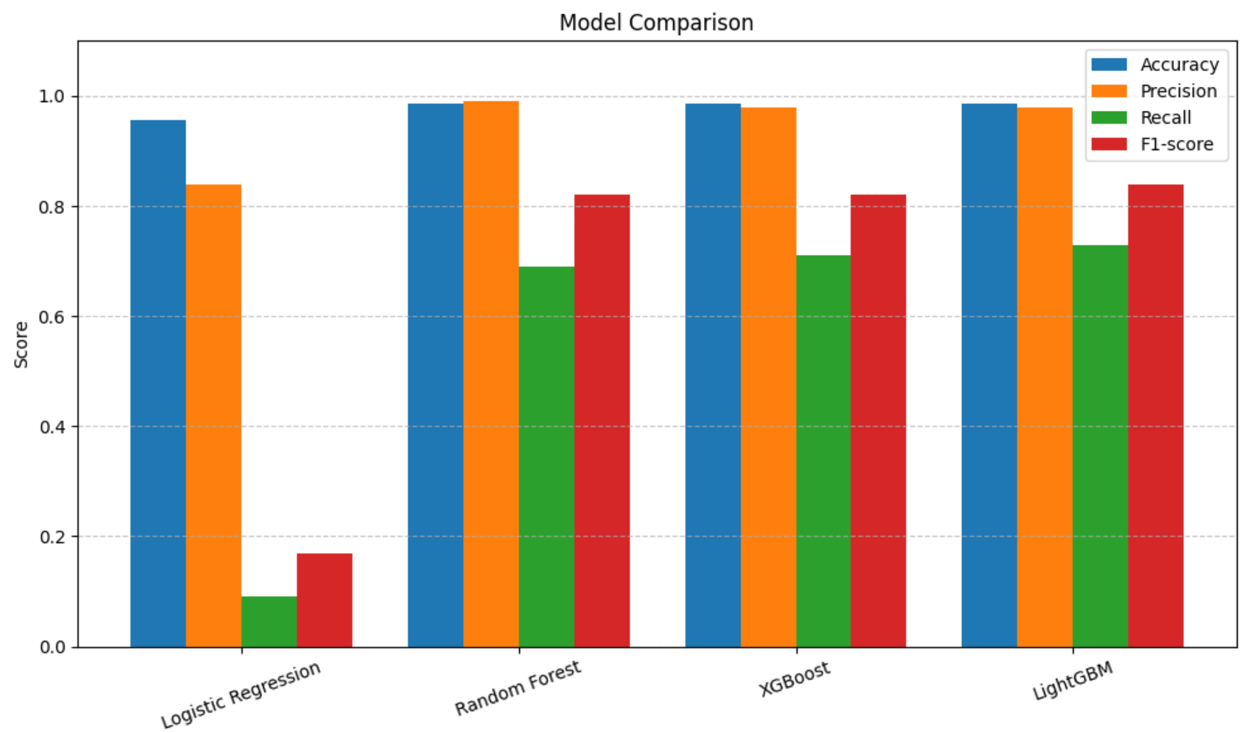
Evaluation Metrics

- **Accuracy:** Proportion of correct predictions.
- **Precision:** Ability to correctly identify fraudulent postings.
- **Recall:** Ability to capture all actual fraudulent postings.
- **F1-score:** Harmonic mean of precision and recall, important for imbalanced data.

6. Results and Model Comparison

Performance Before Hyperparameter Tuning

Model	Accuracy	Precision (fraudulent)	Recall (fraudulent)	F1-score (fraudulent)
Logistic Regression	0.9553	0.84	0.09	0.17
Random Forest	0.9849	0.99	0.69	0.82
XGBoost	0.9852	0.98	0.71	0.82
LightGBM	0.9863	0.98	0.73	0.84

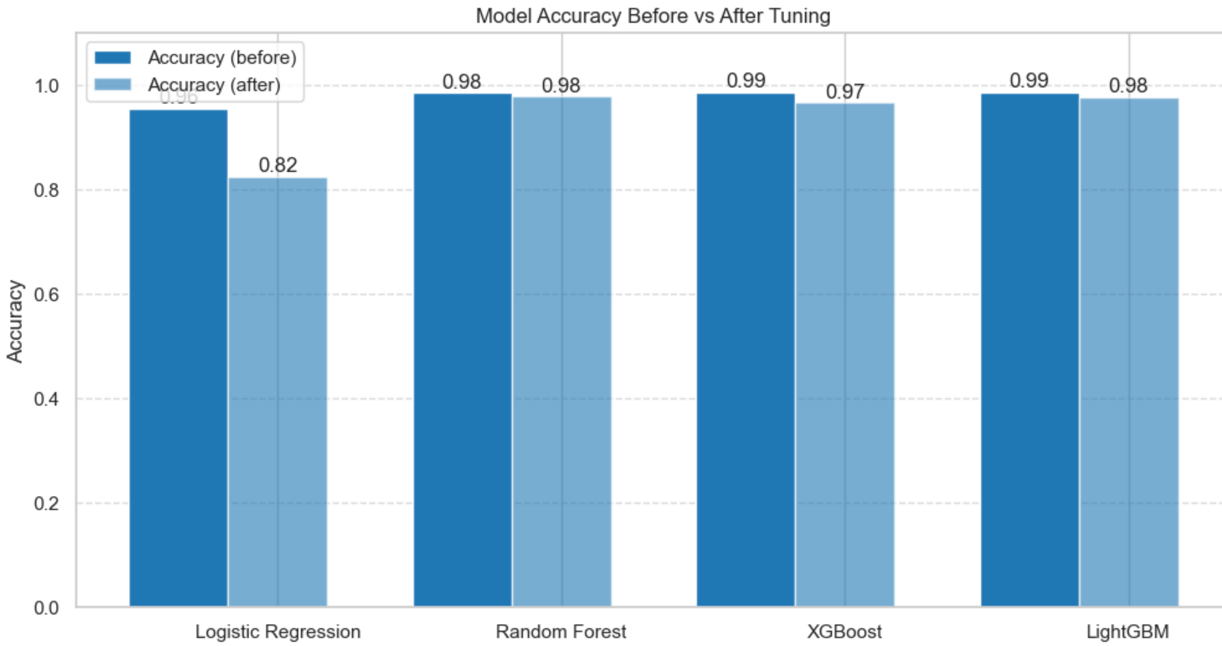


Performance After Hyperparameter Tuning

Model	Accuracy	Precision (fraudulent)	Recall (fraudulent)	F1-score (fraudulent)
Logistic Regression	0.8238	0.18	0.75	0.29
Random Forest	0.9771	0.81	0.68	0.74
XGBoost	0.9659	0.63	0.70	0.66
LightGBM	0.9765	0.78	0.72	0.75

Analysis:

- **Logistic Regression:** Recall improved but precision dropped → many false positives.
- **Random Forest & LightGBM:** Best overall performance with balanced precision and recall.
- **XGBoost:** Good performance, slightly lower than Random Forest and LightGBM.



7. Model Deployment and Endpoint

- The trained models and code have been uploaded to the following GitHub repository for review and future deployment:
  https://github.com/ShodiyAbdulloh

8. Conclusion

- **Logistic Regression:** Simple and interpretable but poor precision after tuning; suitable only for baseline analysis.
- **Random Forest & LightGBM:** High accuracy, good balance of precision and recall; recommended for deployment.
- **XGBoost:** Competitive performance; slightly lower than Random Forest and LightGBM.
- **Recommendation:** Ensemble models, particularly **Random Forest** and **LightGBM**, are optimal for detecting fraudulent job postings.