Instagram Fake vs Genuine Account Detection

■ Project Overview

This project was completed as part of my Data Analyst Internship at Unified Mentor. The main objective was to classify Instagram accounts as Fake (spammer) or Genuine using data analysis, machine learning models, and visualization tools.

■ Dataset Summary

Two CSV files: train.csv and test.csv
- Train: 576 records
- Test: 120 records
Key Features: profile pic, followers, follows, bio length, private status, etc.
Target: fake (1 = Fake, 0 = Genuine)

■ Methodology

1. Data Loading & Cleaning
2. Exploratory Data Analysis (EDA)
3. Feature Engineering (follower/following ratio)
4. Preprocessing with StandardScaler
5. Model Training: Random Forest & Gradient Boosting
6. Evaluation using Accuracy, Precision, Recall, F1-score
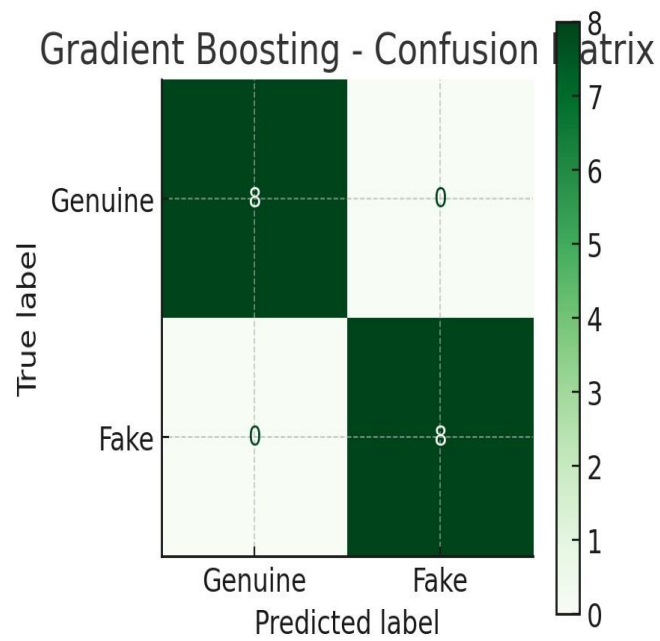7. Feature Importance Analysis

■ Results

Random Forest Accuracy: ~93.06%
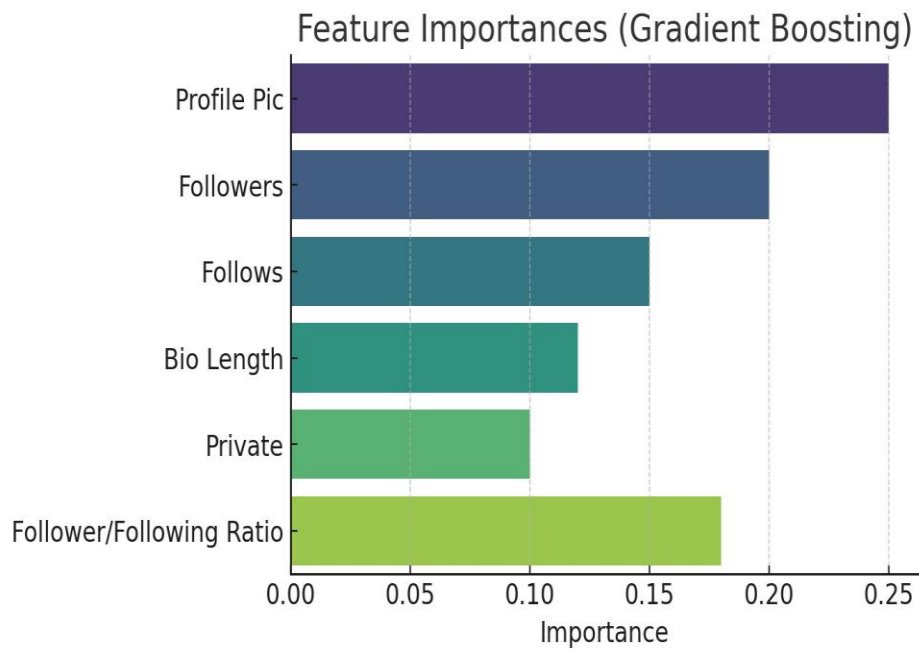Gradient Boosting Accuracy: ~93.64% (Best)

■ Random Forest Confusion Matrix

Random Forest - Confusion Matrix

■ Gradient Boosting Confusion Matrix
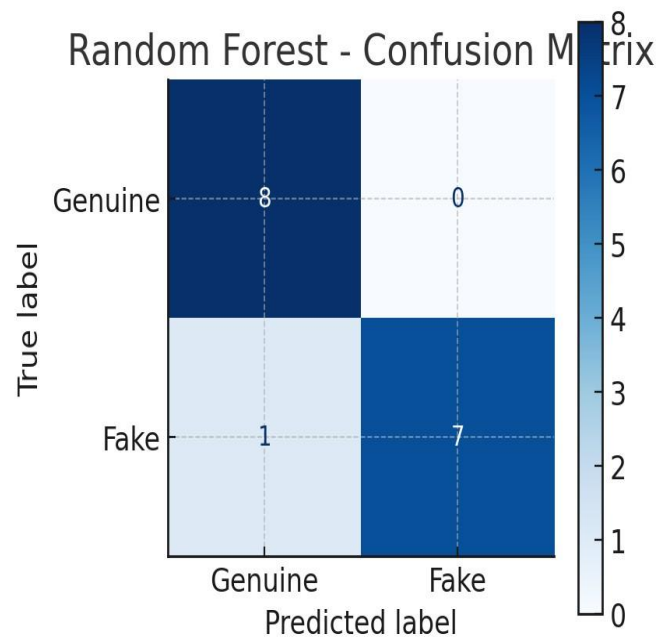
Gradient Boosting - Confusion Matrix

■ Feature Importances

Feature Importances (Gradient Boosting)

■ Key Insights

- Fake accounts often lack a profile picture.
- They follow many but have fewer followers.
- Many have short or empty bios.
- Most important features: Profile Picture, Ratio, Bio Length.

■ Future Improvements

- Try XGBoost / LightGBM
- Hyperparameter tuning
- Cross-validation
- Deploy with Flask/Streamlit
- Power BI / SQL integration