Fake Job Posting Detection Using Machine Learning

Final Project Report — Group 8

Binghamton University

Team Members:

Viranchi More
Pallavi Billava
Mahalakshmi Senthil Kumar

May 1, 2025

Contents

1 Motivation

With the surge in online job postings, fake job advertisements have become a serious threat to job seekers. These postings often lead to data breaches, financial scams, or phishing attacks. Manual identification of fake job listings is infeasible due to the scale and subtlety of deceptive practices. This project aims to build an intelligent system that automatically classifies job postings as fake or real using natural language processing and machine learning techniques.

2 Problem Statement

The objective is to develop a machine learning model that can predict whether a given job posting is fake or real. The model is expected to:

- Analyze textual content such as job descriptions, requirements, and benefits.
- Incorporate engineered features such as salary data, remote status, and suspicious keyword frequency.
- Generalize well to new postings through a user-friendly front-end.

3 Dataset Description

The dataset used is the **Fake Job Postings** dataset, which includes 17,880 records across 18 attributes. Each record represents a job posting with a binary label: 0 (Real) or 1 (Fake).

Column	**Description**
`title, description`	Text fields containing job content
`requirements, benefits`	Additional job context
`salary_range`	Offered salary (range format)
`location, department`	Organizational and geographical info
`fraudulent`	Target label (1 = Fake, 0 = Real)

Table 1: Selected dataset fields

4 Data Exploration & Preprocessing

4.1 Handling Missing Data

- Dropped rows missing critical fields (title, description).
- Imputed or dropped less informative attributes.

4.2 Feature Engineering

- `combined_text`: Merged all text fields.
- `suspicious_keyword_count`: Count of scam-prone terms (e.g., *easy money*, *click here*).
- `years_experience`: Extracted from free text using regex.

- `avg_salary`: Mean of parsed min/max salaries.
- `missing_salary_flag` and `is_remote_flag`: Binary flags.

4.3 Text Vectorization

- Used TF-IDF (bi-grams, 7000 features).
- Combined with numeric features using `hstack`.

5 Exploratory Data Analysis (EDA)

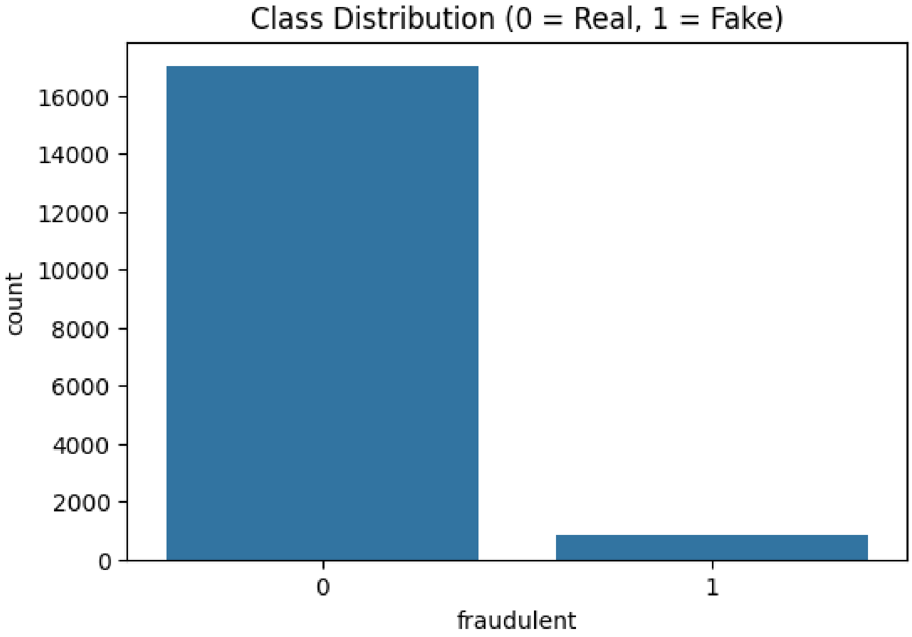


Figure 1: Distribution of Real vs Fake Job Postings in the Dataset

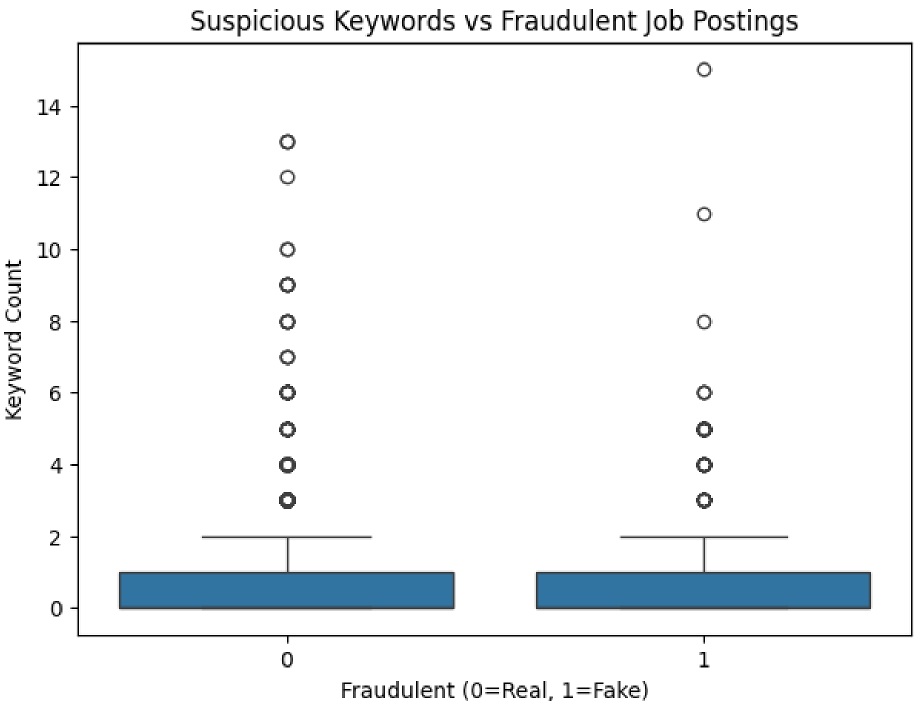


Figure 2: Boxplot of Suspicious Keyword Counts for Real and Fake Jobs

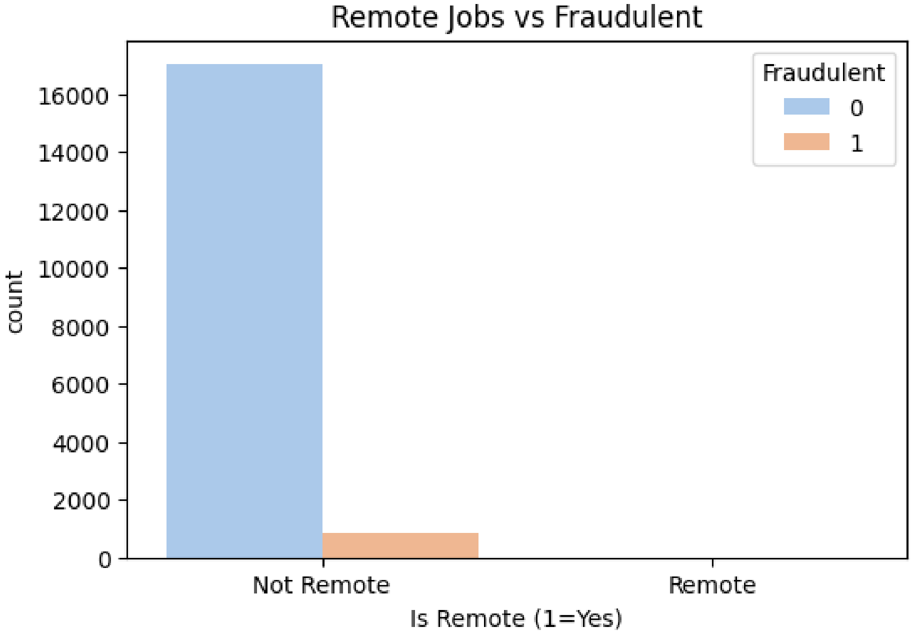


Figure 3: Comparison of Remote vs Non-Remote Jobs with Respect to Fraudulence

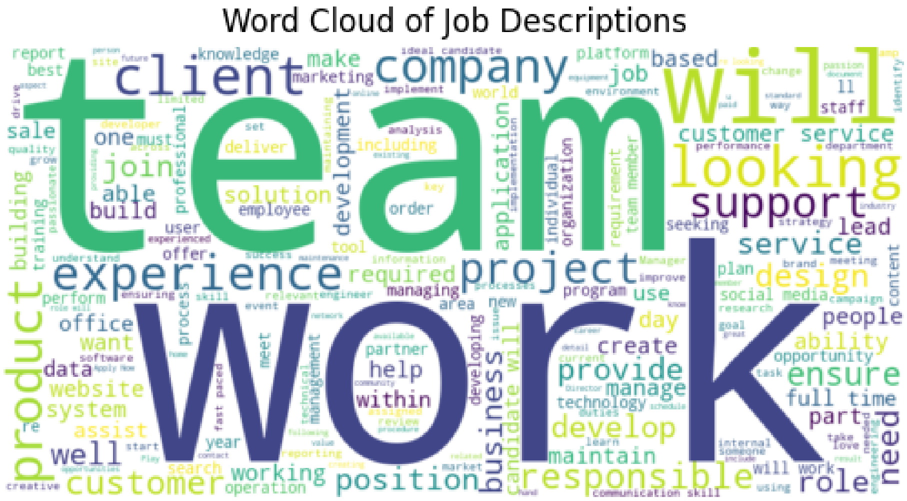


Figure 4: Most Frequent Words in Job Descriptions

6 Model Building & Evaluation

Three models were trained using SMOTE-resampled data:

- **XGBoost** (best performing)
- Logistic Regression
- Random Forest

Model	Accuracy (%)
XGBoost	98.85
Logistic Regression	73.24
Random Forest	98.57

Table 2: Updated Accuracy Comparison Across Models

6.1 Model Accuracy Comparison Visualization

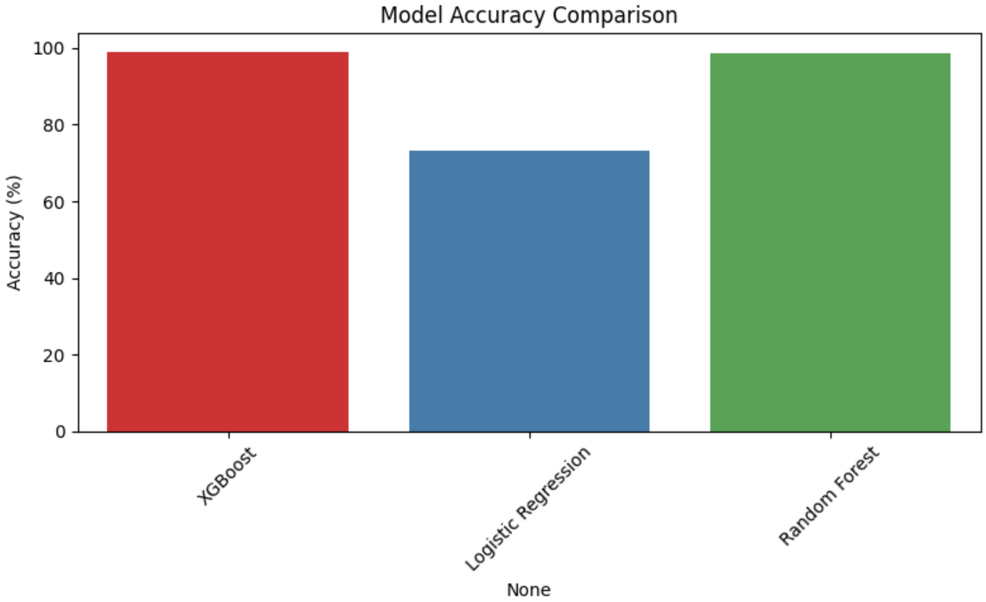


Figure 5: Accuracy Comparison of All Models

6.2 Feature Importance Analysis

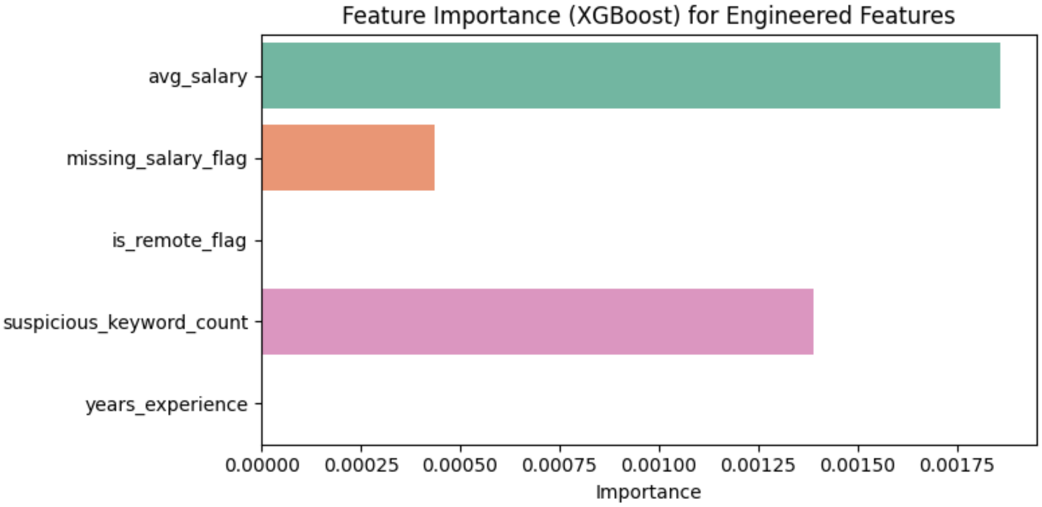
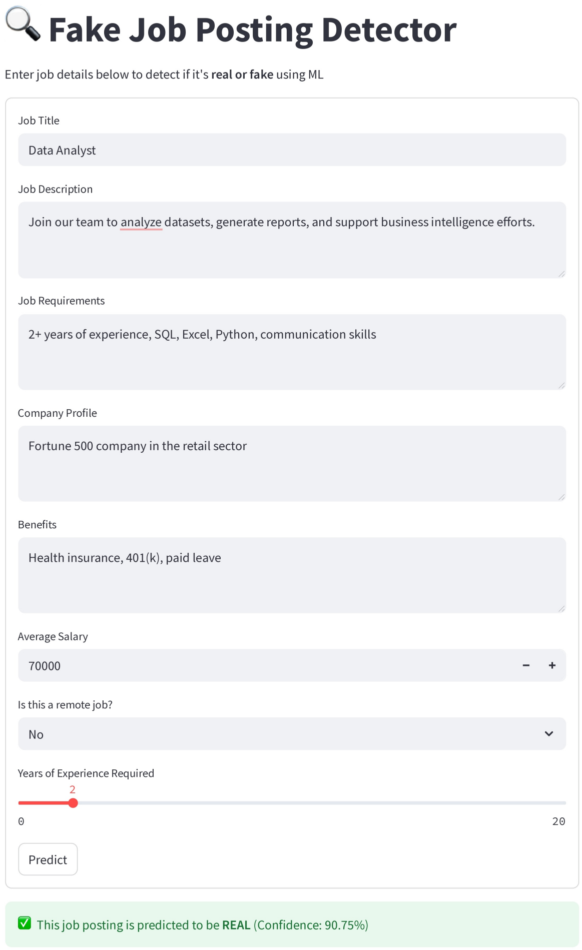


Figure 6: XGBoost Feature Importance for Engineered Features

7 Streamlit Frontend



🔍 Fake Job Posting Detector

Enter job details below to detect if it's **real or fake** using ML

Job Title: Data Analyst

Job Description: Join our team to analyze datasets, generate reports, and support business intelligence efforts.

Job Requirements: 2+ years of experience, SQL, Excel, Python, communication skills

Company Profile: Fortune 500 company in the retail sector

Benefits: Health insurance, 401(k), paid leave

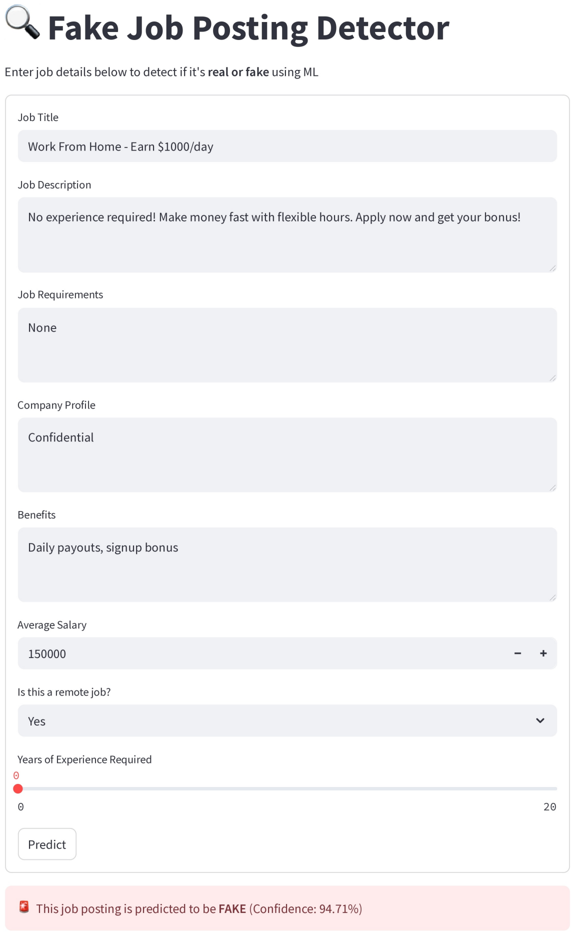
Average Salary: 70000

Is this a remote job?: No

Years of Experience Required: 2

✅ This job posting is predicted to be **REAL** (Confidence: 90.75%)

Figure 7: Sample Prediction for Real Job — Result: REAL (Confidence: 90.75%)



🔍 Fake Job Posting Detector

Enter job details below to detect if it's **real or fake** using ML

Job Title: Work From Home - Earn $1000/day

Job Description: No experience required! Make money fast with flexible hours. Apply now and get your bonus!

Job Requirements: None

Company Profile: Confidential

Benefits: Daily payouts, signup bonus

Average Salary: 150000

Is this a remote job?: Yes

Years of Experience Required: 0

🚨 This job posting is predicted to be **FAKE** (Confidence: 94.71%)

Figure 8: Sample Prediction for Fake Job — Result: FAKE (Confidence: 94.71%)

8 Conclusion

This project successfully demonstrates a robust machine learning pipeline to detect fraudulent job postings.

Key takeaways:

- Text + engineered features improve detection accuracy.
- The model generalizes well on real-world input.
- Streamlit interface adds usability.

Future Enhancements:

- Incorporate advanced NLP techniques (e.g., BERT).
- Continuously train on fresh scraped job data.
- Add explainability features (e.g., SHAP values).

9 Appendix

- Dataset Source: Kaggle — Fake Job Postings Dataset
- Libraries Used: pandas, scikit-learn, xgboost, streamlit, seaborn
- Streamlit App: `streamlit_app.py`
- Notebook: `Final_Project_DM.ipynb`