Job Fraud Detector – Full Project Report

1. Project Overview

Job Fraud Detector is an AI-
powered Streamlit web application designed to identify potentially fraudulent job posting
s. The system utilizes both rule-
based heuristics and a machine learning classification model to analyze job data and highl
ight risks. It provides an interactive dashboard for data visualization and actionable insigh
ts, aiming to protect job seekers and assist recruiters or analysts in auditing job listings.

---

2. Objectives

- Detect fraudulent job postings using automated analysis.
- Provide a user-friendly, interactive dashboard for exploring results.
- Enable both rule-based and machine learning-based fraud detection.
- Support multiple data input methods (CSV upload, sample data, manual entry).
- Allow users to train, export, and apply machine learning models for advanced dete
  ction.

---

3. System Architecture & Workflow

3.1. Components

- **User Interface:** Built with Streamlit, providing sidebar controls, data upload, and dashboard tabs.
- **Rule-
  Based Engine:** Applies heuristic checks (keywords, patterns, missing info) to each job posting.
- **Feature Extraction:** Generates features like text lengths, keyword counts, and pat
  tern flags.
- **Machine Learning Model:** Random Forest Classifier trained on rule-
  based outputs.
- **Visualization:** Interactive charts (pie, bar, histogram) and styled tables using Plotl
  y and Streamlit.

3.2. Workflow

1. **Data Input:** User uploads a CSV, uses sample data, or manually enters job data.

2. **Rule-Based Detection:** Heuristic engine assigns fraud risk scores and labels.
3. **Dashboard Display:** Results shown as metrics, charts, and detailed tables.
4. **Model Training (optional):** User trains a Random Forest Classifier on rule-based results.
5. **Prediction (optional):** Trained model predicts on new, unseen data.
6. **Download:** Results and trained models are available for download.

---

4. Features

4.1. Rule-Based Fraud Detection

- Checks for:
    - Fraudulent keywords in title/description/requirements (e.g., "easy money", "urgent hiring").
    - Suspicious patterns (e.g., unrealistic pay, requests for money, missing comp any info).
    - Short or generic descriptions, excessive punctuation.
    - Remote/work-from-home indicators.
- Outputs:
    - Fraud probability (0-1)
    - Prediction label (“Fraudulent”/“Genuine”)
    - Risk level (“High”/“Medium”/“Low”)

4.2. Machine Learning Detection

- Feature extraction from job posts.
- Trains a **Random Forest Classifier** (binary classification).
- Model can be saved and loaded.
- Predicts fraud risk on new/test data.

4.3. Data Visualization Dashboard

- **Metrics:** Total jobs, fraudulent, genuine, high-risk.
- **Pie Chart:** Fraudulent vs Genuine jobs.
- **Bar Chart:** Risk level distribution.

- **Histogram:** Fraud probability distribution.
- **Results Table:** Styled by label and risk, downloadable as CSV.
- **Suspicious Jobs:** Expandable detailed panels for high-risk jobs.
- **Analytics:** Summarizes common fraud indicators and provides safety tips.

4.4. Flexible Data Input

- **CSV Upload:** Accepts user datasets with flexible columns.
- **Sample Data:** Built-in examples for instant testing.
- **Manual Entry:** Single job analysis for quick checks.

---

5. Classification Model

- **Type:** Random Forest Classifier (scikit-learn)
- **Purpose:** Binary classification — predicts whether a job posting is fraudulent or g enuine.
- **Features Used:** Text lengths, keyword counts, pattern flags, company/location inf o, etc.
- **Training:** Uses results of rule-based detection as labels.
- **Evaluation:** Classification report (accuracy, precision, recall, F1 score) on validati on split during training.

---

6. Technologies Used

- **Python 3.x** — Programming language
- **Streamlit** — Web app framework
- **pandas, numpy** — Data handling
- **scikit-learn** — ML model, scaling, evaluation
- **joblib** — Model and scaler serialization
- **plotly** — Interactive graphs
- **re** — Regex for text analysis

---

7. Data Requirements

- **Required columns:** title, description
- **Optional columns:** company, location, requirements
- **Format:** CSV file (UTF-8), or manual entry via UI.

---

8. Example Data

title	company	location	description	requirements
Software Engineer	Tech Corp	San Francisco	Join our team to build scalable apps...	Bachelor's degree...
EASY MONEY!!! Work from home!!!	Confidential	Remote	Make $5000 per week working from home! ...	None! Just send money!

---

9. How to Use

9.1. Setup

bash

```
git clone https://github.com/acharyamohan/job-fraud-detector2.git
cd job-fraud-detector2
pip install -r requirements.txt
# or install needed packages individually
```

9.2. Run the App

bash

```
streamlit run app.py
```

9.3. Usage Flow

- Choose data input (upload, sample, or manual).
- Analyze data via dashboard (rule-based detection).
- Optionally, train an ML model and use it for predictions.
- Explore results, download CSV/model as needed.

10. Results and Outputs

- **Visual dashboard** for quick insights.
- **Detailed tables** for in-depth review.
- **Downloadable CSV** for offline analysis.
- **Downloadable ML model** for reuse or further training.

---

11. Application Areas

- **Job seekers:** Screen postings before applying.
- **Recruiters:** Audit postings for suspicious ads.
- **Analysts/researchers:** Study fraud trends in employment data.

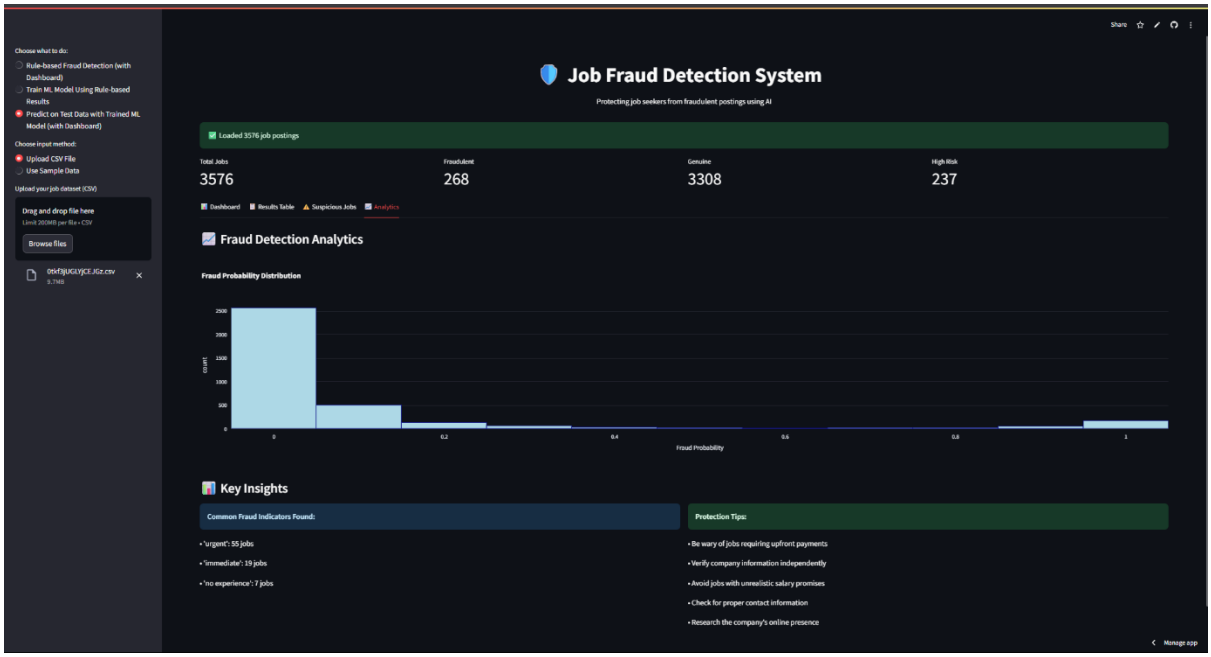
---

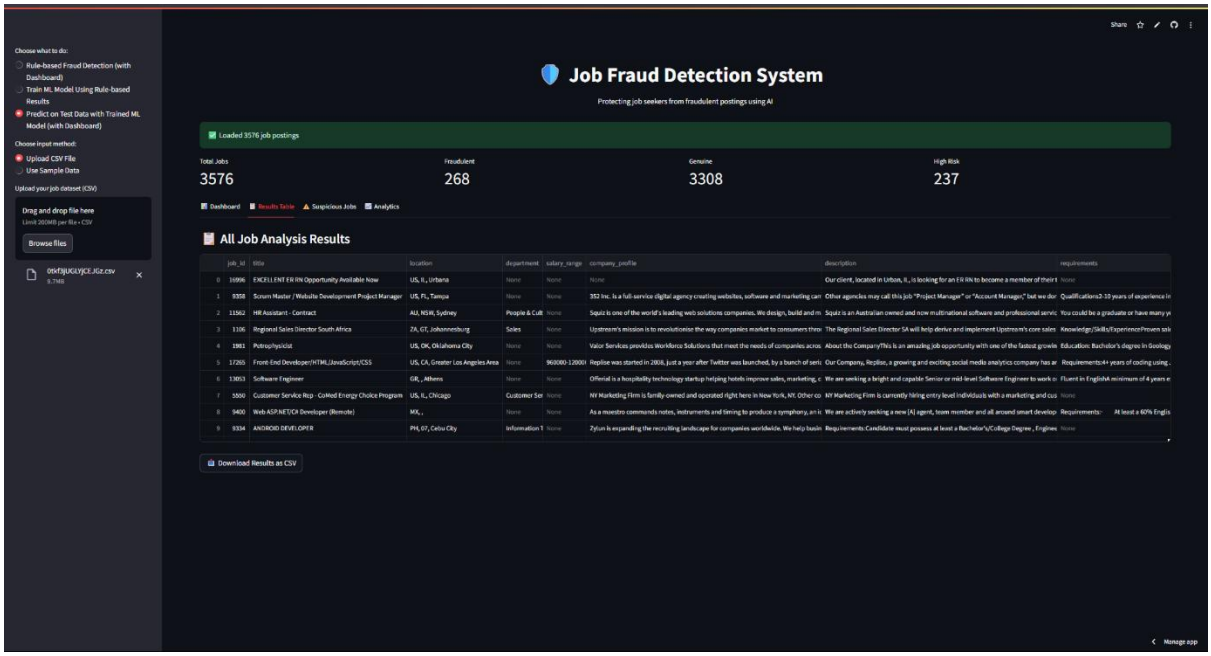
12. Limitations & Future Work

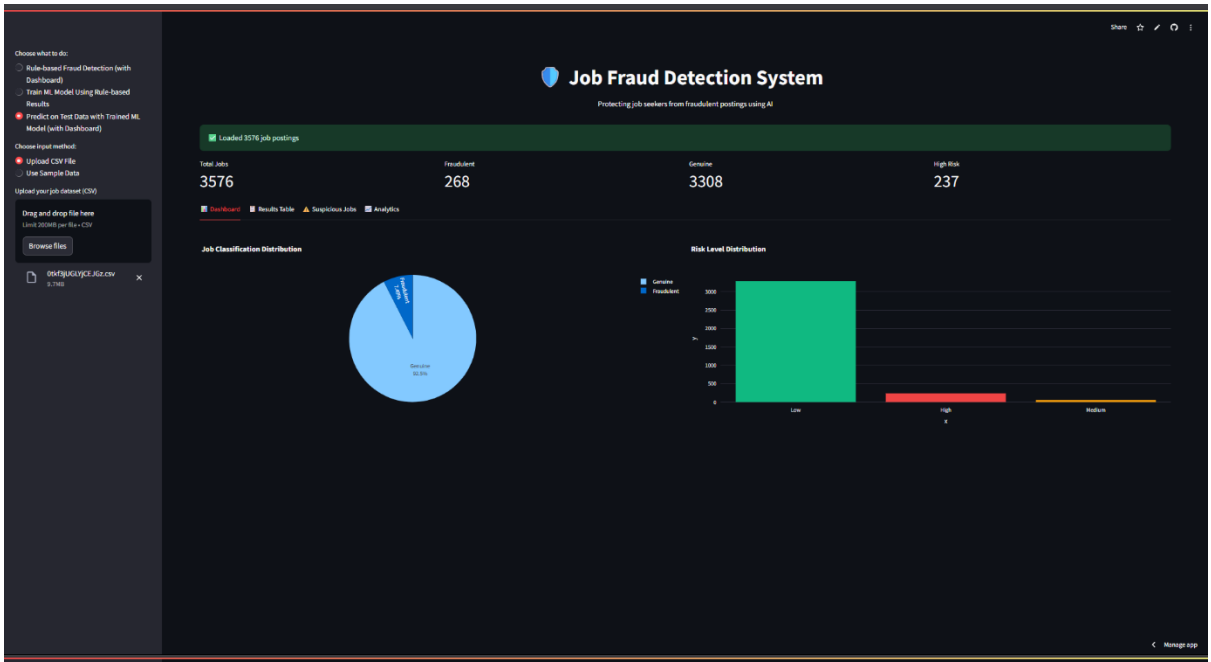
- Rule-based system may not catch subtle/new fraud patterns.
- Machine learning model performance depends on quality/quantity of training data.
- Potential for extension: add NLP analysis, support multi-class classification, deploy as a web service, etc.

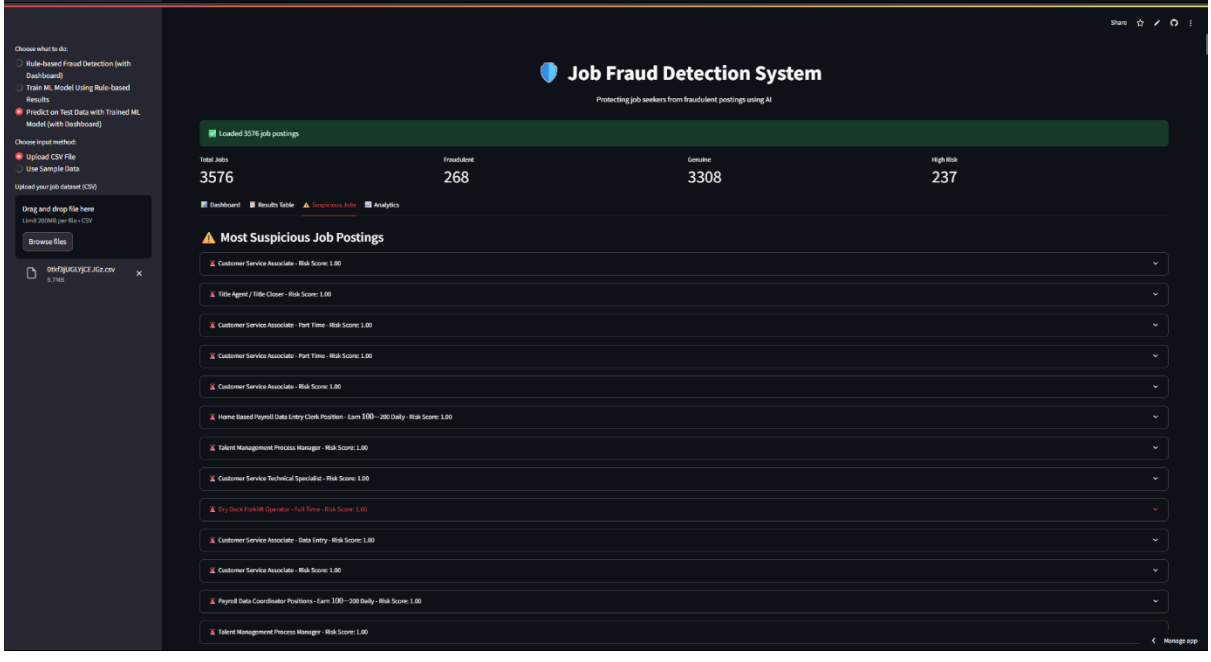
---

13. Screenshots




Choose what to do:
Rule-based Fraud Detection (with Dashboard)
Train ML Model Using Rule-based Results
Predict on Test Data with Trained ML Model (with Dashboard)
Choose input method:
Upload CSV File
Use Sample Data
Upload your job dataset (CSV)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Share
Job Fraud Detection System
Protecting job seekers from fraudulent postings using AI
Loaded 3576 job postings
Total Jobs
3576
Fraudulent
268
Genuine
3308
High Risk
237
Dashboard
Results Table
Suspicious Jobs
Analytics
All Job Analysis Results
job_id
title
location
department
salary_range
company_profile
description
requirements
Download Results as CSV
Manage app


Choose what to do:
Rule-based Fraud Detection (with Dashboard)
Train ML Model Using Rule-based Results
Predict on Test Data with Trained ML Model (with Dashboard)
Choose input method:
Upload CSV File
Use Sample Data
Upload your job dataset (CSV)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Share
Job Fraud Detection System
Protecting job seekers from fraudulent postings using AI
Loaded 3576 job postings
Total Jobs
3576
Fraudulent
268
Genuine
3308
High Risk
237
Dashboard
Results Table
Suspicious Jobs
Analytics
Job Classification Distribution
Risk Level Distribution
Genuine
Fraudulent
Low
High
Medium
Manage app


Choose what to do:
Rule-based Fraud Detection (with Dashboard)
Train ML Model Using Rule-based Results
Predict on Test Data with Trained ML Model (with Dashboard)
Choose input method:
Upload CSV File
Use Sample Data
Upload your job dataset (CSV)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Share
Job Fraud Detection System
Protecting job seekers from fraudulent postings using AI
Loaded 3576 job postings
Total Jobs
3576
Fraudulent
268
Genuine
3308
High Risk
237
Dashboard
Results Table
Suspicious Jobs
Analytics
Most Suspicious Job Postings
Customer Service Associate - Risk Score: 1.00
Title Agent / Title Closer - Risk Score: 1.00
Customer Service Associate - Part Time - Risk Score: 1.00
Customer Service Associate - Part Time - Risk Score: 1.00
Customer Service Associate - Risk Score: 1.00
Home Based Payroll Data Entry Clerk Position - Earn $100 – 200 Daily - Risk Score: 1.00
Talent Management Process Manager - Risk Score: 1.00
Customer Service Technical Specialist - Risk Score: 1.00
Dry Dock Forklift Operator - Full Time - Risk Score: 1.00
Customer Service Associate - Data Entry - Risk Score: 1.00
Customer Service Associate - Risk Score: 1.00
Payroll Data Coordinator Positions - Earn $100 – 200 Daily - Risk Score: 1.00
Talent Management Process Manager - Risk Score: 1.00
Manage app