Project Title

Job Postings Fraud Detection Using Machine Learning

Team Members

1. Adan Akbar (BSDSF22M006)
2. Muhammad Bilal Qaisar (BSDSF22M046)

Problem Statement

With the increasing digitalization of recruitment processes, fraudulent job postings have become more common on online platforms. These deceptive posts can lead to financial loss, identity theft, and wasted time for job seekers. Therefore, there is a need for an automated, intelligent system that can detect and flag potentially fraudulent job postings.

Objectives

- To develop a machine learning model that can classify job postings as ***fraudulent*** or ***genuine*.**
- To identify the most important features that help in detecting fraudulent postings. • To evaluate and compare the performance of different machine learning algorithms for this classification task.
- To build a user-friendly dashboard or visual representation of fraud prediction results (optional if time allows).

Proposed Methodology

1. **Data Preprocessing**
    - Handle missing values and inconsistent formats.
    - Encode categorical variables (e.g., label encoding, one-hot encoding).
    - Normalize/standardize data where required.
    - Text preprocessing (cleaning `description`, `requirements`, etc.).

2. **Exploratory Data Analysis (EDA)**
    - Analyze class distribution (fraud vs. not fraud).
    - Visualize important categorical and numerical features.
    - Word cloud and text pattern analysis for `description`, `benefits`, etc.

3. **Feature Engineering**
    - Create new features like post length, keyword counts, etc.
    - TF-IDF/word embeddings for textual fields.
    - Merge features from multiple text columns.

4. **Model Selection**
    - Try models like Logistic Regression, Random Forest, SVM, XGBoost, and Naive Bayes.
    - Train-test split and K-Fold Cross-validation.
    - Use grid search for hyperparameter tuning.

5. **Model Evaluation** ○
    - Use accuracy, precision, recall, F1-score, and ROC-AUC.
    - Confusion matrix to understand misclassifications.

6. **Model Deployment**
    - Deploy a basic interface (streamlit or Flask) for real-time fraud prediction.

Dataset Description

The dataset includes job postings with the following fields:

- Textual fields like **`description`**, **`requirements`**, **`benefits`**, etc.
- Categorical fields like **`location`**, **`department`**, **`employment_type`**, etc.
- Binary flags such as **`has_company_logo`**, **`telecommuting`**, etc.
- The target variable is **`fraudulent`**, indicating whether the post is fake or genuine.

Dataset Link:

https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

Expected Outcomes

- A trained machine learning model capable of predicting fraudulent job postings with high accuracy.
- A ranked list of features contributing most to fraud detection.
- Insightful visualizations and model interpretation.
- A deployable tool or dashboard for demonstrating the model.