Project Report

“Fake Job Posting Prediction Model”

Objective

The objective of this project is to detect fake job postings using machine learning. The model aims to classify job postings as genuine or fraudulent based on the features and patterns in the dataset.

Dataset Overview

Description:

- **Source**: Dataset containing information on job postings.
- **Size**: 17,880 rows and 18 columns.
- **Features**:
    - Textual Data: Job Title, Job Description, Requirements.
    - Categorical Data: Industry, Employment Type.
    - Numeric Data: Salary, Number of Requirements.
- **Target Variable**: Fraudulent Label (1 for fake postings, 0 for genuine).

Data Preprocessing

Steps:

1. **Data Cleaning**:
    - Removed duplicate and irrelevant columns.
    - Handled missing values by imputing or dropping rows where necessary.
2. **Text Preprocessing**:
    - Converted text data (e.g., Job Descriptions) into a machine-readable format using:
        - Tokenization: Splitting text into words.
        - Removal of stop words and punctuation.
        - Vectorization with TF-IDF (Term Frequency-Inverse Document Frequency) to weigh word importance.
3. **Encoding Categorical Variables**:
    - Transformed non-numeric columns like Industry and Employment Type into numeric values using one-hot encoding.
4. **Feature Scaling**:
    - Standardized numeric features such as salary and requirements to normalize their ranges.

Exploratory Data Analysis (EDA)

Key Insights:

- **Fraudulent Job Postings**:
    - Approximately 5% of postings were labelled as fraudulent.
    - Fake postings often lacked specific details (e.g., missing salary or location).
- **Textual Patterns**:
    - Frequent use of phrases like "quick money" or "work from home" in fraudulent postings.

Visualizations:

- Bar chart comparing genuine vs. fraudulent postings.
- Word clouds highlighting common terms in fake postings.

Feature Engineering

Key Features:

1. **Textual Features**:
    - Extracted information from job descriptions using TF-IDF vectorization.
    - Identified key phrases and patterns indicative of fraudulent postings.
2. **Categorical Features**:
    - Encoded fields such as Industry and Employment Type.
3. **Numeric Features**:
    - Included salary range, number of requirements, and job description length.

Model Development

Algorithms Used:

1. **Logistic Regression**:
    - A baseline model with a linear approach.
    - Achieved 91% accuracy but struggled with complex patterns.
2. **Random Forest Classifier**:
    - Utilized an ensemble of decision trees to improve performance.
    - Achieved 97% accuracy with better handling of non-linear relationships.
3. **XGBoost (Gradient Boosting)**:
    - Combined multiple weak learners into a strong predictive model.
    - Achieved the highest accuracy of 98% with excellent precision and recall.

Evaluation Metrics:

- **Accuracy**: Overall correctness of predictions.
- **Precision**: Focused on reducing false positives.
- **Recall**: Measured the ability to detect fraudulent postings.
- **F1-Score**: Balanced metric combining precision and recall.

Model Results

Best Model: XGBoost

- **Metrics**:
  - Precision: 0.96
  - Recall: 0.94
  - F1-Score: 0.95

Insights:

- Fake postings are often flagged due to vague job descriptions, missing details, and suspicious phrases.

To Do:

1. Automate job posting screening using the XGBoost model.
2. Implement additional checks for postings flagged as suspicious.
3. Educate users on identifying and avoiding fraudulent postings.

Conclusion

This project successfully developed a machine learning model to predict fake job postings with high accuracy. The XGBoost classifier demonstrated excellent performance, making it a reliable tool for real-world implementation. Continuous refinement and new data integration can further enhance the model's effectivenes