Project Report: Fake Job Postings Detection

1. Introduction: Fake job postings have become a prevalent issue in online job platforms, leading to potential job seekers falling victim to scams. This project aims to develop a machine learning model capable of identifying fraudulent job postings. The dataset, stored in a CSV file named 'fake_job_postings.csv', includes various features such as job titles, locations, and company profiles, with the target variable 'fraudulent' indicating whether a job posting is fraudulent (1) or not fraudulent (0).

2. Data Loading and Exploration: The project begins by loading the dataset and exploring its structure to gain insights into the distribution of fraudulent and non-fraudulent instances. Visualizations such as count plots and bar charts help to understand the dataset's characteristics, including missing values in different columns.

3. Data Preprocessing: Missing values are addressed by filling appropriate values for text and categorical features. Columns such as 'title', 'location', 'employment_type', 'required_experience', and 'required_education' are selected as relevant features for model training. The text data is combined into a new feature, 'text', to be used in the model.

4. Model Development: Several machine learning models are considered, including Gradient Boosting, Random Forest, and XGBoost classifiers. A pipeline is constructed, incorporating preprocessing steps such as one-hot encoding for categorical features and vectorization for text features. Cross-validation is employed to evaluate model performance, and the XGBoost Classifier emerges as the best performer based on accuracy and F1 score.

5. Model Evaluation: The selected models, including Gradient Boosting and Random Forest, are evaluated on the test set. The XGBoost Classifier exhibits the highest accuracy and F1 score, making it the preferred choice for further analysis.

6. Visualization: Confusion matrices and classification reports are visualized to provide insights into the model's behaviour. The visualizations aid in understanding the model's precision, recall, and F1 score at different threshold levels.

7. Imbalance Ratio: The class imbalance in the dataset is calculated and highlighted, emphasizing the need for robust evaluation metrics in imbalanced datasets.

8. Threshold Analysis: A thorough analysis of the model's performance at different probability thresholds is conducted, providing a nuanced understanding of its behaviour in predicting fraudulent job postings.

9. NLP Model Addition: To enhance the model's capability to understand textual data, an NLP model is introduced. The addition involves using a TfidfVectorizer for text feature

extraction. Cross-validation and evaluation metrics demonstrate improved performance with the XGBoost Classifier using NLP features.

10. Conclusion:

In conclusion, the project successfully addresses the challenge of identifying fraudulent job postings using a systematic approach, from data preprocessing to model development and evaluation.

Further Recommendations:

- Consider exploring feature importance to gain insights into the features contributing most to the model's predictions.
- Experiment with more advanced techniques such as ensembling models for potential performance improvement.
- Continue refining the model based on additional data or insights gained from further analysis.