Project Report: Fake Job Postings Detection

Objective: The primary objective of this project is to develop a machine learning model capable of identifying fake job postings. The model utilizes various features such as job titles, locations, company profiles, and other relevant information to classify job listings as either fraudulent or legitimate.

Dataset: The dataset used for this project is stored in a CSV file named 'fake_job_postings.csv'. It includes columns such as job_id, title, location, department, salary_range, and others. The target variable, 'fraudulent,' indicates whether a job posting is fraudulent (1) or not fraudulent (0).

Data Loading and Exploration: The dataset was loaded using the pandas library, and an initial exploration was conducted to understand its structure. Visualizations were created to depict the distribution of fraudulent and non-fraudulent instances. The count of missing values was analysed for each column, providing insights into the dataset's completeness.

Data Pre-processing: Missing values were addressed by employing appropriate strategies for each column. Categorical features were handled using techniques like one-hot encoding. A new feature, 'text,' was created by combining 'title' and 'description' to facilitate text-based analysis.

Model Development: Various machine learning models were considered for classification, including Gradient Boosting, Random Forest, and XGBoost. The scikit-learn library was utilized for building pipelines that include pre-processing steps and encoding of categorical features. Cross-validation was performed to assess the models' performance, and XGBoost emerged as the best performer.

Model Evaluation: The XGBoost model was evaluated on the test set, yielding impressive results in terms of accuracy and F1 score. The confusion matrix and classification report provided detailed insights into the model's behaviour, highlighting its ability to correctly identify both fraudulent and non-fraudulent instances.

Visualization: Confusion matrices and classification reports were visualized to enhance the interpretability of the results. The imbalance ratio of the dataset was acknowledged, and threshold analysis was conducted to understand the model's performance at different decision thresholds.

Recommendations and Future Work:

1. Further explore advanced techniques such as ensemble methods and stacking to potentially improve model performance.
2. Consider deploying the trained model for real-time identification of fake job postings.
3. Regularly update the model with new data to maintain its accuracy over time.
4. Investigate the impact of different text pre-processing techniques on model performance.

Conclusion: The project successfully developed a machine learning model for detecting fake job postings. XGBoost emerged as the best-performing model based on accuracy and F1 score. The thorough exploration, pre-processing, and evaluation steps contribute to the robustness of the model. Future work could involve deploying the model for practical use and exploring additional techniques to enhance its performance further.