A Report created on the basis of Fake Job Postings dataset

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#### **Summary**

This report presents a comparative analysis of two machine learning models — Logistic Regression and Random Forest Classifier — aimed at detecting fake job postings. The primary objective of this study is to evaluate the performance of both models in identifying fraudulent job listings based on textual features extracted from job descriptions.

This analysis includes:

* Data preprocessing
* Model training
* Evaluation and comparison of key metrics such as accuracy, precision, recall, and F1-score.

Models

1. Logistic Regression:

It’s a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and probability of specific outcome. Logistic Regression is particularly useful when the relationship between the input features and the output label is approximately linear.

1. Random Forest Classifier:

Random forest is a supervised machine learning algorithm in which the calculations of numerous decision trees are combined to produce one final result. It is an ensemble method — a technique where we take many base-level models and combine them to get improved results.

Key Differences

The comparison of the models is based on the accuracy and robustness in detecting fake job postings. While Logistic Regression provides a simpler linear model, Random Forest leverages the power of ensemble learning to enhance accuracy. A detailed analysis of performance metrics reveals the strengths and weaknesses of each model.

Result

The study shows us that Random Forest outperforms Logistic Regression in terms of accuracy and robustness. However, Logistic Regression remains a valuable option for its simplicity and interpretability.

Libraries and Features Used:

1. Pandas
2. Scikit learn
   1. TF-IDF vectorizer
   2. Train-test split
   3. Logistic regression
   4. Accuracy score
   5. Classification report
3. Imbalanced - learn
   1. SMOTE (Synthetic Minority Over-sampling Technique)
4. Regular Expression

Dataset Used:

Open data source from Kaggle:

[Real / Fake Job Posting Prediction](https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction)