Fraudulent Job Postings

1. **Hypothesis**: We can predict which job postings are fraudulent or real
2. **Dataset**: ~18,000 job descriptions, containing 800 fakes <https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction?resource=download>
3. **Reshape the Data**:

* Use all features (columns), except job\_id (not significant)
* One-hot coding – represent features as numerical values, 1 or 0
  + Location – by city/country or embedding
  + Department
  + Employment types – full-time, part-time, contract
  + Required experience – entry, mid, senior level
  + Required education – high school, bachelor’s, master’s
  + Industry
  + Function – “marketing,” “engineering”
* Target variable = fraudulent column

1. **Clean the Data:**

* Handle missing values

1. **Error Metric**

* For binary classification:
  + Accuracy - % of predictions that are correct
  + Precision - % of predicted frauds that are truly fraudulent
  + Recall - % of actual frauds the model successfully caught
  + F1 score – harmonic mean of precision and recall
  + ROC-AUC – how well model separates fake vs real overall

1. **Split the Data**

* Training data
* Test data

1. **Train a Model**

* Classification model