CRISP-DM Report: Fake Job Postings Detection

1. Business Understanding

Problem Statement

Fake job postings are a rising problem on job platforms. These scams:
- Waste applicants' time and resources.
- Risk identity theft or financial loss.
- Damage platform reputation.

Project Goal:

Build a predictive model to automatically detect fraudulent job postings.

Business Objectives:

- Accurately classify job posts as fraudulent (1) or legitimate (0).
- Minimize false positives (real jobs flagged as fake).
- Support job moderation teams through semi-automation and intelligent alerts.

2. Data Understanding

Dataset Source:

Kaggle's Fake Job Postings Dataset

Dataset Overview:

- ~17,000 job postings
- Key fields: title, location, department, description, requirements, telecommuting, fraudulent, etc.

Key Insights:

- Imbalanced dataset: Only ~5% are fraudulent.
- Text-heavy features (e.g. description, company_profile) hold hidden signals.
- Several missing values in optional fields.

3. Data Preparation

Cleaning Steps:

- Dropped irrelevant columns (job_id, logo, url).
- Filled missing values in text fields with "Not Specified".
- Converted categorical text fields to lowercase.

Feature Engineering:

Created the following features:

- description_word_count, requirements_word_count
- Suspicious keyword flags (e.g. if description contains “investment”, “money”, “urgent” → flag = 1)
- Encoded categorical columns (e.g. location, employment_type) using One-Hot Encoding.

Class Balancing:
Used SMOTE (Synthetic Minority Oversampling Technique) to balance the training set and improve model sensitivity to fraud cases.

4. Modeling

Models Tested:
- Logistic Regression: Baseline, fast & interpretable
- Decision Tree: Tuned for depth and splitting
- Random Forest: ⭐ Best performing ensemble model

Final Pipeline:
Feature Engineering → SMOTE → Random Forest Classifier

Performance Summary:
- Accuracy: ~96%
- ROC AUC: ~0.95
- Fake Recall: ~0.73

5. Evaluation

Key Metrics:
- Confusion Matrix
- Precision / Recall / F1-Score
- ROC AUC Curve

Interpretation:
- Legitimate jobs classified with high accuracy.
- Post-balancing + feature engineering drastically improved fake listing recall.
- Small trade-off in false positives to maximize fraud detection.

6. Deployment (Future Work)

Deployment Possibilities:
- Integrate into job board moderation systems.
- Flag suspicious listings for manual review.
- Power HR dashboards with fraud risk insights.

Next Steps:

- Use TF-IDF or BERT embeddings for better text understanding.
- Expand dataset with real-world labeled samples.
- Build interactive dashboards for moderators.

Final Summary Table

CRISP-DM Phase	Highlights
Business Understanding	Automate fake job detection to protect users & platforms
Data Understanding	Text-heavy, real-world dataset with severe class imbalance
Data Preparation	Filled missing data, created text & keyword features, SMOTE
Modeling	Random Forest + balanced data + engineered features = results
Evaluation	0.95 AUC, 0.73 recall for fake class, strong model performance
Deployment	Model ready for real-world job site integration