FAKE JOB POSTING DETECTION

Aluoch Phanela

Business Background

Fake job postings waste job seekers' time, damage trust in platforms, and are increasingly sophisticated.
They often:

- Collect sensitive data through deception
- Impose fake fees
- Mimic legitimate postings convincingly

Using machine learning, we aim to detect and flag fraudulent listings automatically.

Business Understanding

- Can we predict fake job postings using job content?
- Which features strongly indicate fraud?
- How accurate is our prediction model?
- How can it support moderation workflows?

Objectives: How Can We Hit Our Goals

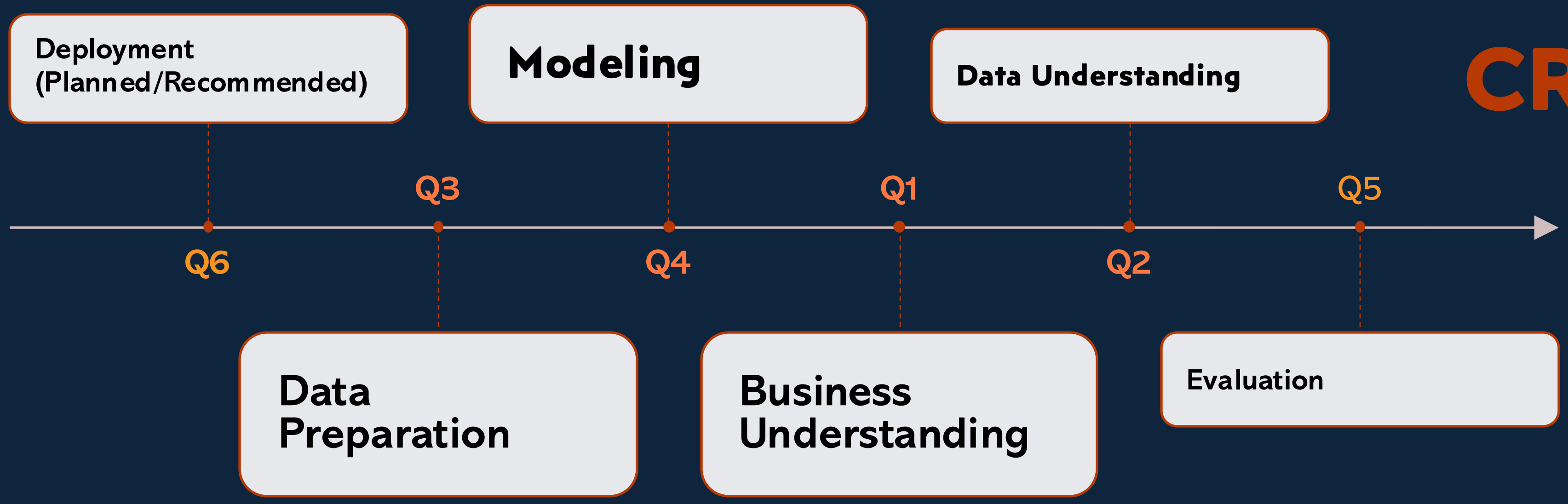
- Explore fraudulent vs real job patterns
- Engineer features that expose hidden fraud signals
- Train and test models to predict job fraud
- Identify the best performing algorithm
- Recommend automation support for moderators
- e on emerging market preferences.

We used the “Fake Job Postings” dataset from Kaggle.
 It includes job attributes like title, description, location, employment type, and a binary `fraudulent` label.

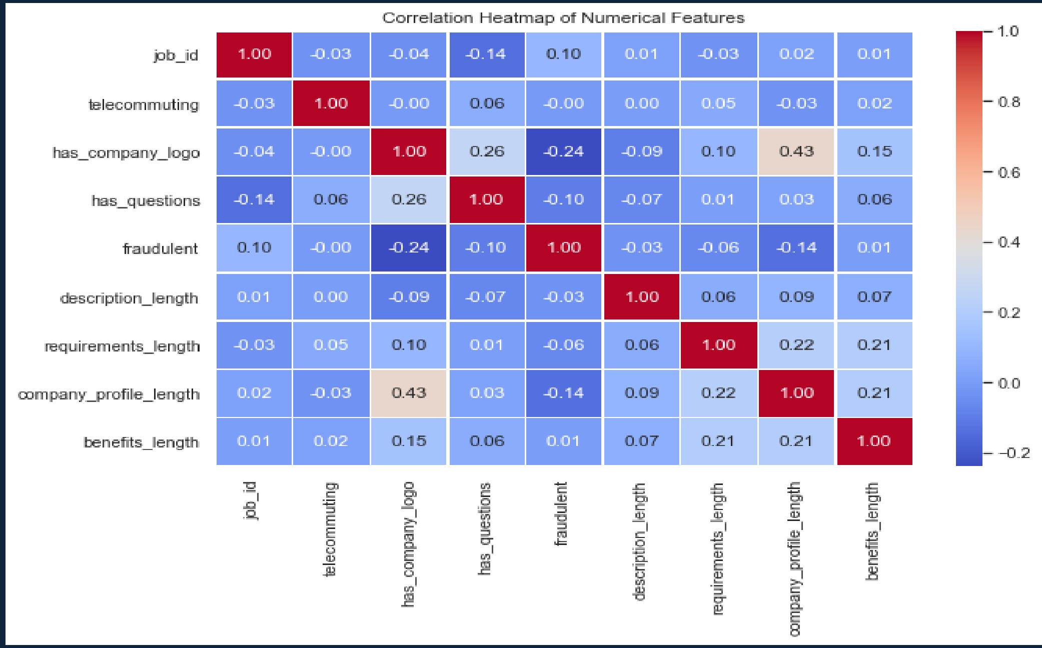
Data Overview

- Fake job postings hurt job seekers and platforms. Goal: Build a model to identify fake postings.
- Loaded dataset from Kaggle. Inspected columns like `title`, `description`, `location`, etc. Target: `fraudulent`.
- Cleaned missing data, engineered features like word count & presence of suspicious keywords, encoded categorical variables.
- Used Random Forest + SMOTE to handle class imbalance. Tuned hyperparameters and evaluated with precision, recall, ROC AUC.
- Final model achieved good precision and recall on fraudulent class. ROC AUC ~0.95.
- Model & notebook ready. Can be integrated into moderation workflows or automated systems.

CRISP-DM Process

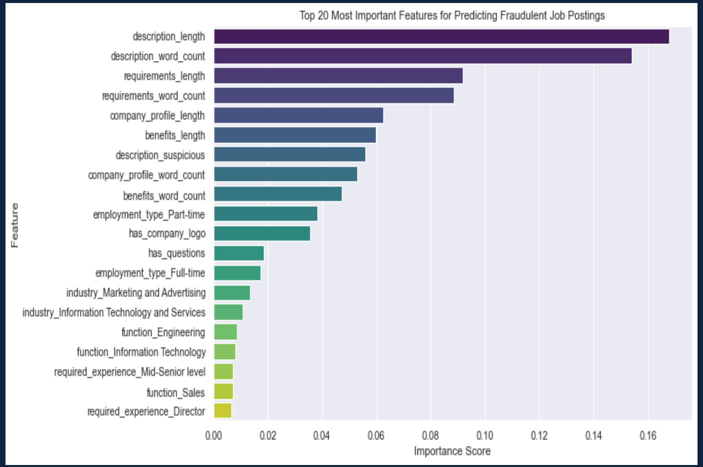


EDA The three strongest predictors



- The heatmap shows how strongly pairs of features are related (values range from -1 to 1).
- High positive values (close to 1) indicate a strong direct relationship.
- High negative values (close to -1) indicate a strong inverse relationship.
- Values near 0 suggest no linear correlation.

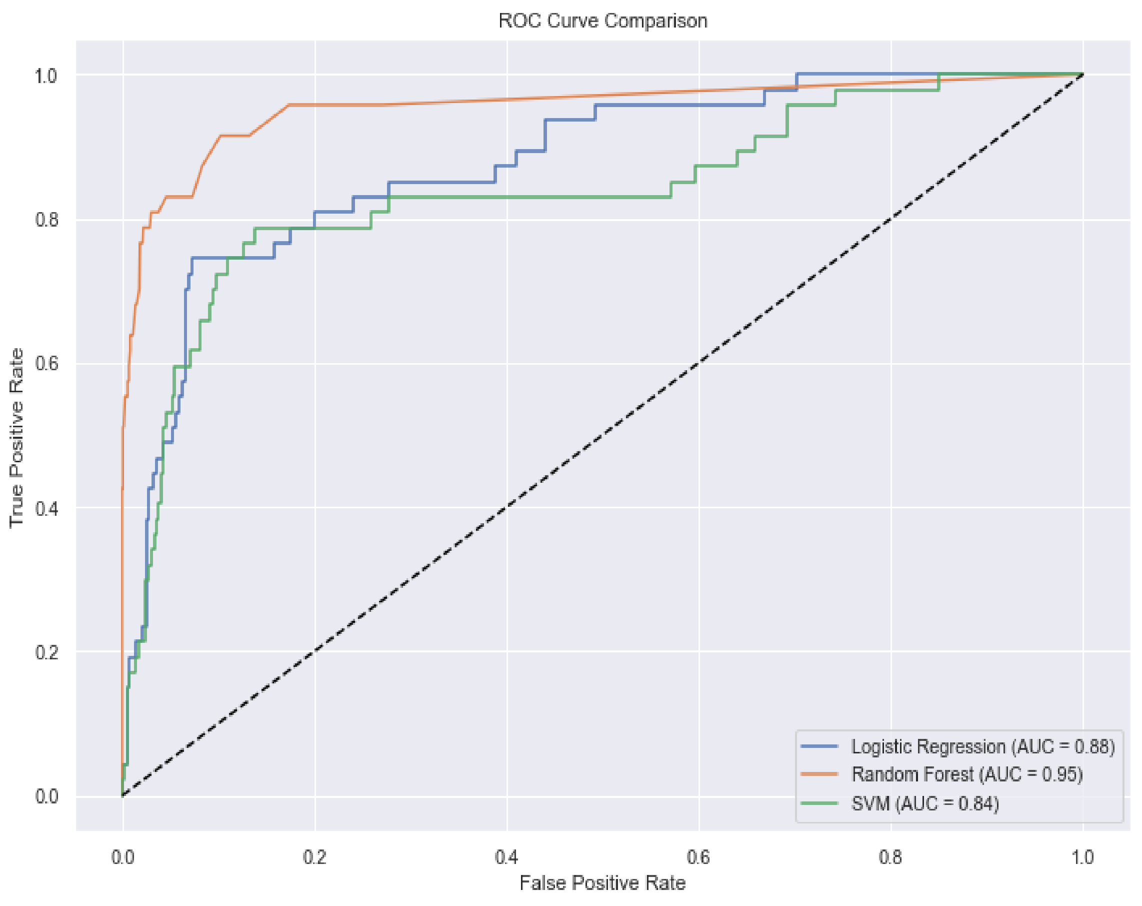
Features that are most predictive of fraudulent listings



The initial Random Forest model identified several features as highly predictive of fraudulent job postings

Model that detects fake job postings without misclassifying

- The ROC curves showed that both models performed better than random guessing (AUC > 0.5), but Random Forest achieved a curve that hugged closer to the top-left corner, suggesting stronger predictive power.



Conclusion

- Fake job postings can be predicted with high accuracy using machine learning.
- Feature importance reveals significant text-based and location-based indicators.

Recommendation

- Deploy model into job posting platforms
- Use predictions to flag suspicious listings in real time
- Continuously retrain with fresh data to catch evolving scams

Future initiatives

1. Deploy model into moderation pipeline.
2. Perform real-time scoring.
3. Improve text preprocessing for future iterations.

Thank you

Phanela Aluoch

Phanelaabigael@yahoo.com

phanelabigaelquotes.blogspot.com