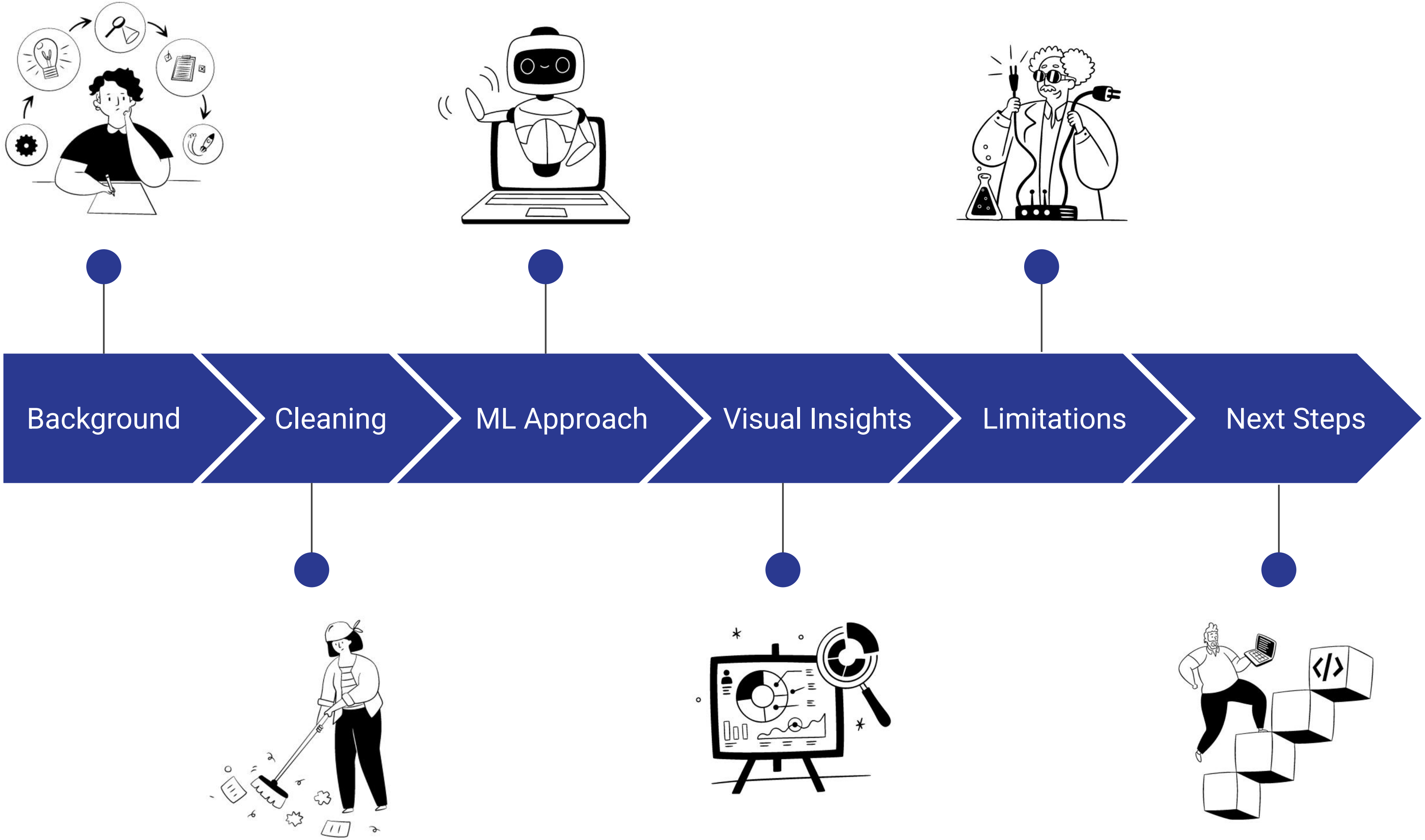
Project 4

Fake Job Postings **Detection**

Group 2

Derek Bates
Ally Eveslage
Jackson Popelka
Erica Wollmering



Background
Cleaning
ML Approach
Visual Insights
Limitations
Next Steps

Background

Problem

Why do fake job postings matter?

Fake job listings can waste applicants' time, expose them to financial scams, and even lead to identity theft.

In a job market where people are already vulnerable, these scams exploit hope and urgency — often leaving real harm in their wake.

The Data

2016 Data

Contains ~18,000 job postings

Each listing is labeled as “real” or fraudulent”

Kaggle does not provide documentation on how these labels were created

2023-2024 Data

Contains recent job postings scraped from LinkedIn

Does not include labels for “fake” vs. real postings

Used to test model performance on more recent job postings

The Data: Shared Columns

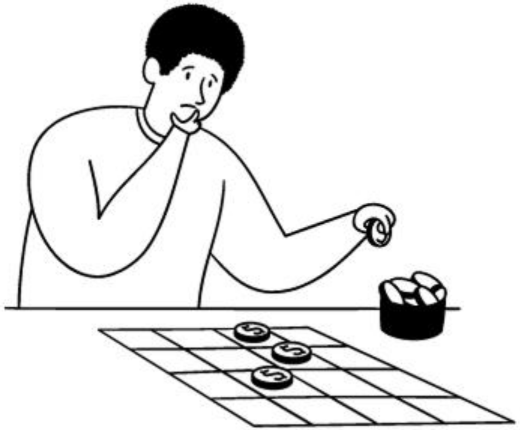
Job Title | Employment Type | Industry

Challenges - Only 4.3% of Postings Are Fake

Challenge

Imbalanced Dataset

- We can't rely on accuracy alone.
- We’ll look at precision, recall, and F1 score.



Cleaning the Data

Preprocessing for Better Results

2016 Data

- Loaded and duplicated raw CSV for safe editing
- Checked value counts and flagged low-variance columns
- Selected relevant features (e.g. title, description, location, education, experience)
- Split location column into country, region, and city
- Renamed columns for clarity and consistency
- Exported cleaned data to new CSV for modeling

Our ML Approach

Machine Learning

approach

- Combined multiple text fields into a single input
- Used TF-IDF vectorization to convert text to numeric format
- Trained a Random Forest Classifier using GridSearchCV
- Tuned n_estimators and max_depth for best F1 score

Model Performance & Metrics

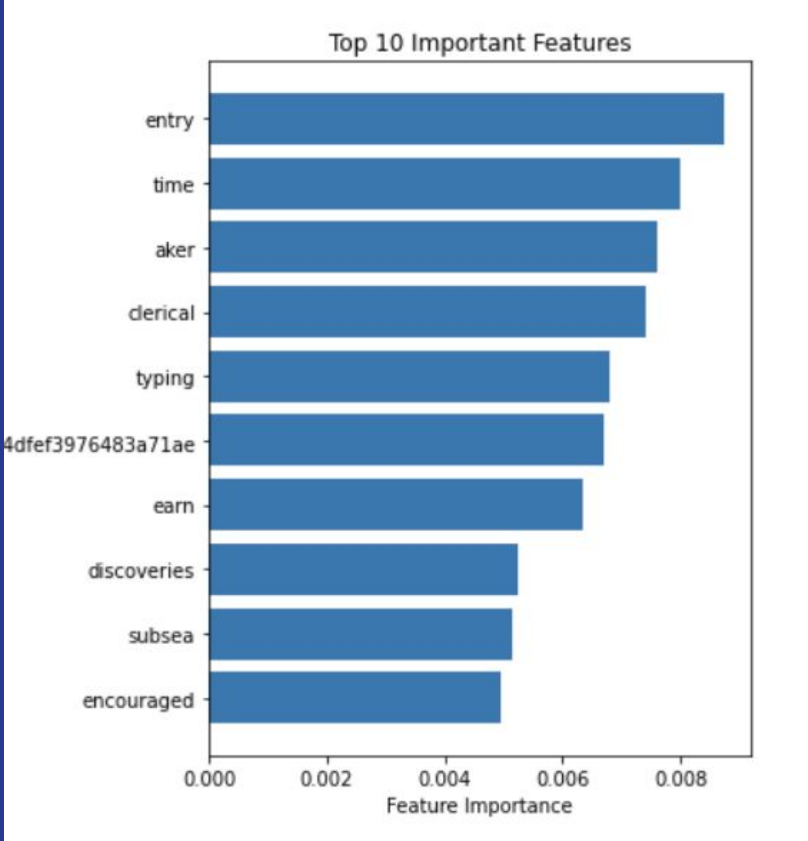
Accuracy, Precision, Recall, F1

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3403
1	1.00	0.60	0.75	173
accuracy			0.98	3576
macro avg	0.99	0.80	0.87	3576
weighted avg	0.98	0.98	0.98	3576

How the Model “Thinks”

- Top features included keywords, company descriptions, and location signals
- Feature importance chart generated from Random Forest model

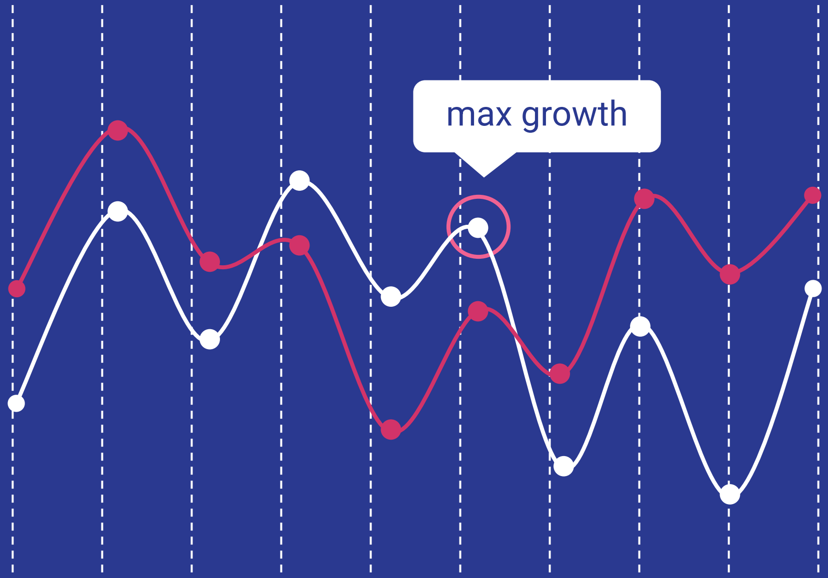


Insights

2016

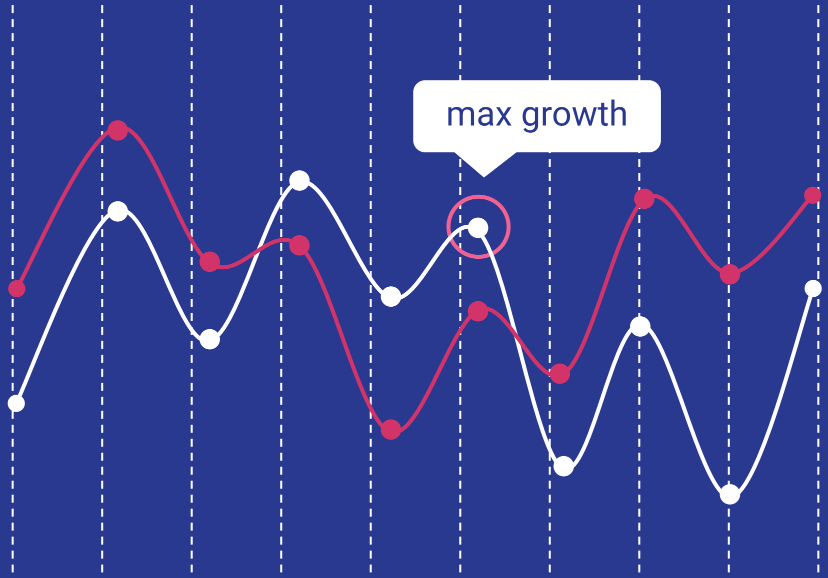
Y/N

Real vs Fake



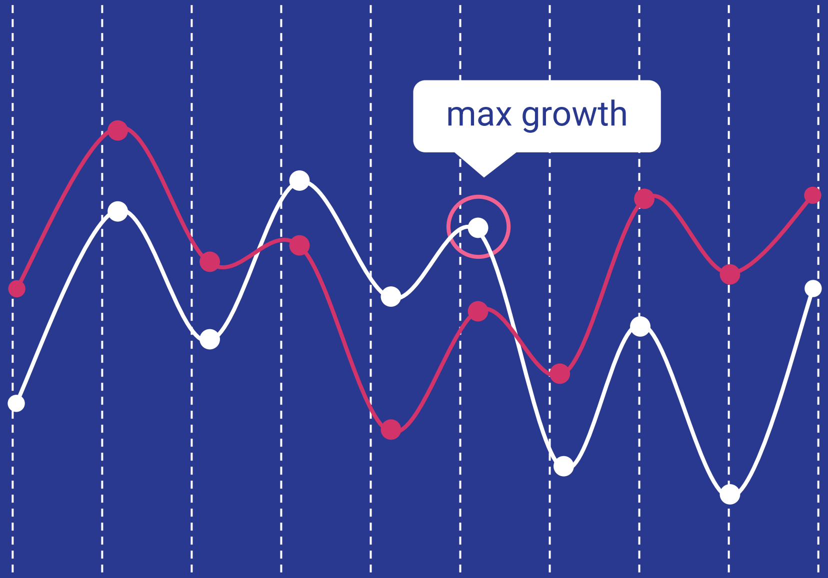
2016 Job Titles

Real vs Fake



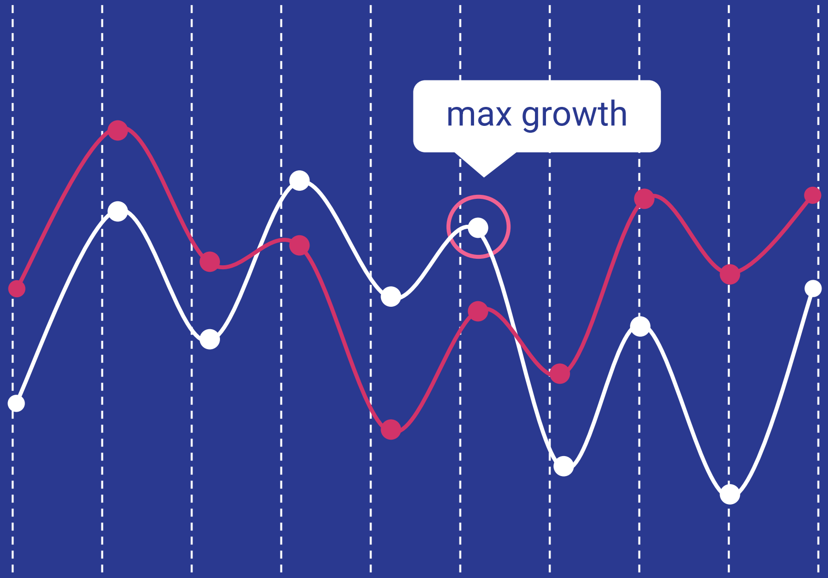
2016 Keywords

Fake vs Real



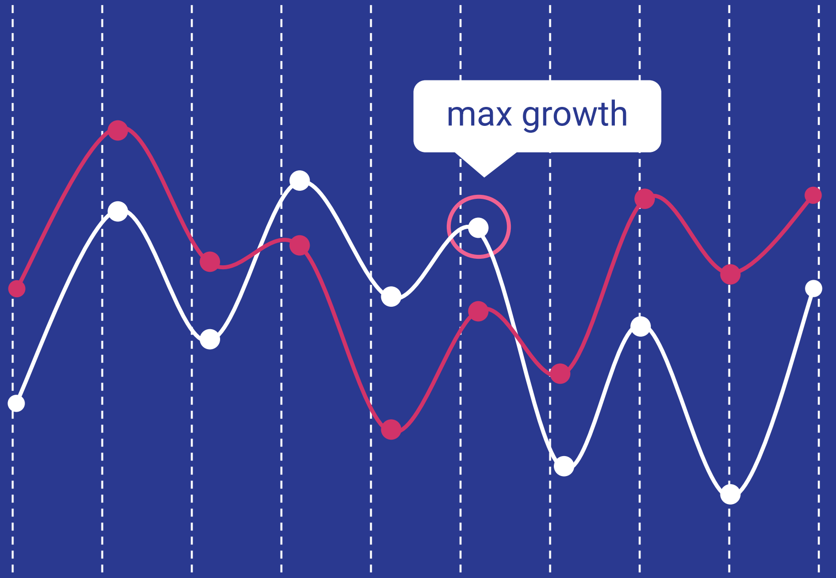
2023-24
Y/N

Real vs Fake



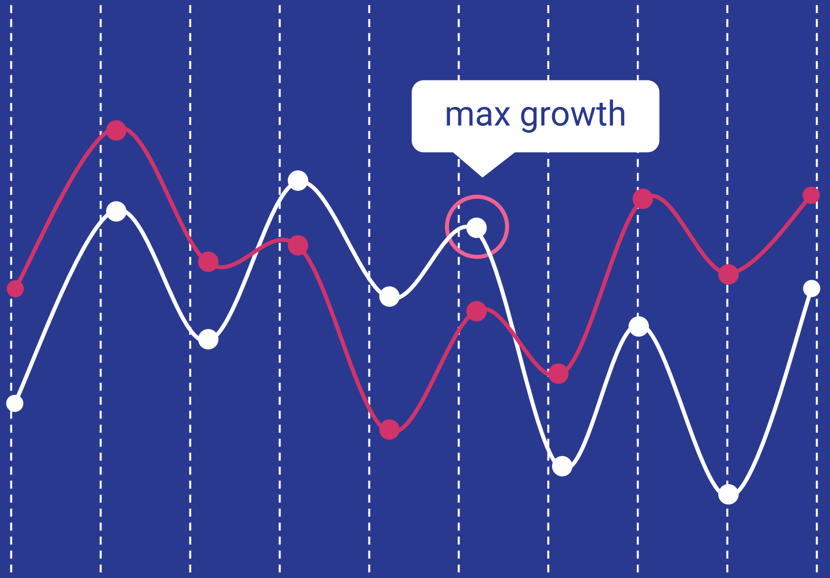
2023-24 Keywords

Real vs Fake



2023-24 Job Titles

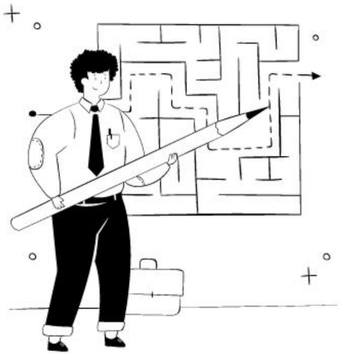
Real vs Fake



Key Differences Between the Datasets

2016 Data	2023-2024 Data
Point 1	Point 1
Point 2	Point 2
Point 3	Point 3

Limitations



Challenges & Caveats

Warning!

- Outdated training data — scammers evolve fast.
- Small proportion of fake listings (~4.3%)
- Manual labeling may introduce bias
- We're building intuition, not a perfect detector

What’s Next/Final Thoughts

Next Steps

Real World Use Cases

- Re-train on updated job datasets
- Explore more advanced NLP (e.g., BERT)
- Integrate as a flagging tool for job platforms ?
- Use for scam-awareness education & training?