Detecting Fake Job Postings With Machine Learning Techniques

12/8/2025

Content

- Problem at Hand
- The Goal: Proactively Identify Fake Job Postings
- The Dataset
- High-Level Approach
- Results and Findings
- Recommended Next Steps
- Appendix

The Problem at Hand

- Fake job postings lure applicants into sharing sensitive personal information (PII).
- Fraudsters exploit this data for identity theft and financial crimes.
- Victims face reputational damage, financial loss, and emotional stress.

"In the first half of the year, online job scams rose 19% compared to a year earlier and have cost Americans nearly $300 million, with the typical victim losing around $2,000, according to data from the Federal Trade Commission."

Source: "They Were Looking For Work – But Found a Scam Instead" By Shannon Pettypiece, NBC News, 10/20/2025

https://www.nbcnews.com/news/us-news/job-scam-ziprecruiter-linkedin-work-postings-fake-listing-rcna238162

The Goal: Proactively Identify Fake Job Postings Before They Can Do Harm

- There is an opportunity, leveraging machine learning techniques, to proactively flag suspicious job postings based on the content within the postings themselves.
- A machine learning model can potentially detect the characteristics of the postings that are most associated with fakes.
- In production, the solution can produce an algorithm that evaluates the content of a posting and returns a validity assessment for job seekers (e.g., ‘Appears Safe’ or ‘Beware, Likely Fake’)
- Ultimately, the solution can be deployed on job boards, sites, or as a service for job seekers to reduce the threat and harm due to fake postings.
- The scope of this project was the identification of the best performing modeling technique to identify fake job postings on a dataset available on Kaggle.

The Data Used to Build the Algorithm

- Data available on Kaggle @ https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction*
- Contains 17,880 Job Postings, of which 866 are flagged as fraudulent (~4.8% of the postings are fake)
- Includes a mix of 'somewhat' structured categorical data, very text-heavy fields, and binary flags.

'Somewhat' Structured Data

- Location
- Department
- Employment Type
- Required Experience
- Industry
- Function
- Salary Range

Very Text Heavy Fields

- Company Profile
- Job Description
- Job Requirements
- Benefits
- Title

Binary Flags

- Telecommuting?
- Has Company Logo?
- Has Questions?
- Fraudulent Posting Indicator

* Note: It is not clear if the data is actual legitimate/fake job postings or synthetic generated data. There is an acknowledgment of The University of the Aegean | Laboratory of Information & Communication Systems Security but the link to their website is broken.

Data Values from a Sample Record

job_id	title	location	department	salary_range
1	Marketing Intern	US, NY, New York	Marketing	

benefits	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function	fraudulent
	0	1	0	Other	Internship			Marketing	0

company_profile	description	requirements
We're Food52, and we've created a groundbreaking and award-winning cooking site. We support, connect, and celebrate home cooks, and give them everything they need in one place.We have a top editorial, business, and engineering team. We're focused on using technology to find new and better ways to connect people around their specific food interests, and to offer them superb, highly curated information about food and cooking. We attract the most talented home cooks and contributors in the country; we also publish well-known professionals like Mario Batali, Gwyneth Paltrow, and Danny Meyer. And we have partnerships with Whole Foods Market and Random House.Food52 has been named the best food website by the James Beard Foundation and IACP, and has been featured in the New York Times, NPR, Pando Daily, TechCrunch, and on the Today Show.We're located in Chelsea, in New York City.	Food52, a fast-growing, James Beard Award-winning online food community and crowd-sourced and curated recipe hub, is currently interviewing full- and part-time unpaid interns to work in a small team of editors, executives, and developers in its New York City headquarters.Reproducing and/or repackaging existing Food52 content for a number of partner sites, such as Huffington Post, Yahoo, Buzzfeed, and more in their various content management systemsResearching blogs and websites for the Provisions by Food52 Affiliate ProgramAssisting in day-to-day affiliate program support, such as screening affiliates and assisting in any affiliate inquiriesSupporting with PR & Events when neededHelping with office administrative work, such as filing, mailing, and preparing for meetingsWorking with developers to document bugs and suggest improvements to the siteSupporting the marketing and executive staff	Experience with content management systems a major plus (any blogging counts!)Familiar with the Food52 editorial voice and aestheticLoves food, appreciates the importance of home cooking and cooking with the seasonsMeticulous editor, perfectionist, obsessive attention to detail, maddened by typos and broken links, delighted by finding and fixing themCheerful under pressureExcellent communication skillsA+ multi-tasker and juggler of responsibilities big and smallInterested in and engaged with social media like Twitter, Facebook, and PinterestLoves problem-solving and collaborating to drive Food52 forwardThinks big picture but pitches in on the nitty gritty of running a small company (dishes, shopping, administrative support)Comfortable with the realities of working for a startup: being on call on evenings and weekends, and working long hours

High-Level Approach: Finding the Best Algorithm to Detect Fake Postings

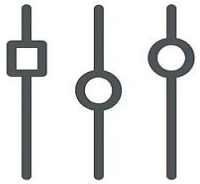
PHASES OF THE MODEL BUILD

DATA EXPLORATION & STANDARDIZATIION

- Missingness throughout (10 fields with1,000+ records with missing values)
- Limited standardization (over 11,000 distinct job titles, 3,100+ locations, etc.)
- Spareness with many ‘levels’ in the data appearing < 50 times

FEATURE ENGINEERI NG

- Standardized where possible (up-case & space removal for categorical fields)
- Parsed the location field to obtain clean ‘country’ indicator
- Parsed salary range to components (low, max, difference)
- Collapsed levels of categorical fields (low count categories combined into ‘other’)
- Turned text heavy fields into into numbers and themes



MODEL FITTING & TUNING

- Fit linear (logistic) and tree-based methods (random forest and xgboost) using the features noted above as inputs; Used weighting to address imbalanced data
- Tuned model parameters for each to find best performing results
- Evaluated alternative cut-off thresholds
- Tested using reduced and expanded feature sets from the detailed text fields

IDENTIFYING CHAMPION SOLUTION & INSIGHT

- Identified the champion solution based on performance statistics on the test or ‘holdout’ dataset
- **Winner = Xgboost methodology using the default cut-off threshold value and the detailed field-specific meaningful word feature set**

Results and Findings

Winning Model: XGBoost with numeric inputs for each text-heavy field (e.g., job description, job requirements, company description). Postings with a predicted probability of being fake > 0.5 were flagged as 'fake'.

Model Performance Results (On Holdout or 'Unseen Data')

Performance Metric	How to Interpret	Value	Meaning
Accuracy	% of all predictions that were correct	0.984	Very high overall correctness
Precision	% of predicted fakes that were truly fake	0.89	Low false positive rate; Limited number of legitimate postings are flagged as fake
Recall	% of actual fakes that were correctly found	0.78	Strong fraud detection coverage; Majority of fakes are identified
F1 Score	Balance of precision and recall	0.83	Solid overall fraud detection performance balancing precision (not too many false positive, with good coverage of fakes)
ROC-AUC	Ability to distinguish fake vs real	0.99	Very high score, reflecting strong discrimination capability
PR-AUC	Precision-recall tradeoff across thresholds	0.91	Very high score on imbalanced data with a 0.04 target incidence rate

Results and Findings (continued)

Top Characteristics or Features of Fake Postings*

- **Industries** such as accounting, computer networking, leisure/travel/tourism, and oil&energy
- **Functions** such as accounting/auditing, administrative, financial analyst, and ‘other’
- **Country** codes such as AU, BH, MY, and QA
- **Company Logo** not present
- **Company Description** less likely to contain ['quality', 'time', 'high', 'provide', 'companies', 'amp', 'people', 'right', 'help', 'job']

* Note: Leveraged SHAP values chart and interpreted most important model features.

Recommended Next Steps

- Deploy the champion model into a controlled production environment (job boards or build a web- or app-based service)
- Monitor performance to validate ongoing accuracy; track precision, recall, and false positive rates
- Retrain periodically with new postings to adapt to evolving fraud tactics
- Expand feature set with additional signals (e.g., posting source, recruiter history, metadata)
- Explore ensemble methods by combining random forest and xgboost into a ‘super’ prediction

Thank You!

Appendix

Detailed Model Results Comparisons and Evaluations

Fake Job Postings Dataset	
Number of Job Postings, Total:	17,880
Number of Fake Postings:	866
Fake Posting (Fraud) Incidence:	4.84%

Model Fitting Results for the Validation/Test Dataset (30% Sample comprised of 5,364 total postings including 260 fakes, a 4.84% fake rate)

Model Iteration	Model Type	TD-IDF Matrix Reduction Scope	Predicted Prob Threshold*	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Model Notes & Interpretation
1	Logistic	Across all TD-IDF Field Matrices	0.5	0.83	0.2	0.88	0.33	0.914	0.4388	Very high recall, but poor precision (resulting in a lot of false positives, flagging valid job postings as 'fake')
2	Logistic	Across all TD-IDF Field Matrices	0.86	0.95	0.5	0.5	0.5	0.914	0.4388	More balanced precision and recall (but still lower than tree-based methods)
3	Random Forest	Across all TD-IDF Field Matrices	0.5	0.97	0.98	0.4	0.56	0.979	0.824	Extremely high precision and solid accuracy, but recall is only 0.40. Excellent at avoiding false positives, but misses many fraudulent postings
4	Random Forest	Across all TD-IDF Field Matrices	0.276	0.975	0.76	0.71	0.74	0.979	0.824	Very stong candidate. High accuracy, balanced precision and recall and high PR-AUC.
5	XGBoost	Across all TD-IDF Field Matrices	0.5	0.98	0.89	0.62	0.73	0.977	0.834	Strong balance across precision, recall, F1, and high PR-AUC. A reliable performer with good fraud detection and manageable false positives.
6	XGBoost	Across all TD-IDF Field Matrices	0.296	0.977	0.79	0.74	0.76	0.977	0.834	Very stong candidate. High accuracy, balanced precision and recall and high PR-AUC.
7	Logistic	For each TD-IDF Field Matrix	0.5	0.822	0.2	0.85	0.32	0.908	0.354	High recall but precision is low (lots of fals positives). Accuracy is modest. Like iteration 1, it's aggressive in catching fraud but at the expense of flagging valid postings as fake.
8	Logistic	For each TD-IDF Field Matrix	0.877	0.9466	0.45	0.46	0.46	0.908	0.354	More balanced precision and recall (but still lower than tree-based methods)
9	Random Forest	For each TD-IDF Field Matrix	0.5	0.979	0.91	0.63	0.75	0.99	0.886	Strong Solution; Excellent balance of precision and recall, producing a solid F1. High accuracy and PR-AUC. One of the strongest performers, especially for minimizing false positives while still catching frauds.
10	Random Forest	For each TD-IDF Field Matrix	0.317	0.979	0.81	0.77	0.79	0.99	0.886	Strongest Random Forest Candidate; Excellent balance of precision and recall, producing a solid F1. High accuracy and PR-AUC. One of the strongest performers, especially for capturing a high volume of all the fraud cases.
11	XGBoost	For each TD-IDF Field Matrix	0.5	0.984	0.89	0.78	0.83	0.99	0.91	Best overall candidate; Excellent balance of precision and recall, excellent PR-AUC. Among top scoring solution in all evaluation metrics
12	XGBoost	For each TD-IDF Field Matrix	0.342	0.982	0.82	0.82	0.82	0.99	0.91	Strong performer, comparable to the high performing xgboost model in row 11.

* Each technique and TD-IDF candidate feature set were fit the default probability threshold of 0.5 and then an 'optimal' threshold identified by finding the threshold tied to the highest estimated F1 score.