

GenAI Misinformation: A Binomial Model

ADSP 31014

Anisa Dye

Tori Healey

Anais Morales



Introduction



Research Question: Can characteristics of online job postings predict the likelihood of the posting being fraudulent?

- Fraudulent job postings can cause individuals to share personal or sensitive information with ill-intentioned strangers and can result in distress, financial loss, or loss of privacy
- Statistically identifying predictors of fraud is difficult due to the lack of data on fraud/difficulty in identifying fraud
- Predictors of fraudulent job postings may provide hints towards predictors of other types of scams

Data Description

- Data comes from research done by the Aegean University & Workable
- 17,800 observations, each representing a Workable job posting
- Outcome variable: **fraudulent** (1 if fraudulent job posting, 0 if not)
 - Fraud identified by Workable employees based on suspicious activity in their system, false contact or company information, applicant complaints, and periodic analysis of clientele.
- 16 other columns include job description, required education, if the position contains screening questions, the type of employment (part-time vs full-time) and job industry, among others



Data Description: Transformations & Cleaning

Restricting the Data Set

- Observations come from over 90 countries, but having data from different countries may complicate the model
 - Ex. Different currencies for salaries; length of job description may fundamentally vary by language
- **Solution:** Restrict data to the country with the highest number of observations
 - US: 10.6k out of 17.8k observations

Addressing Missing Values

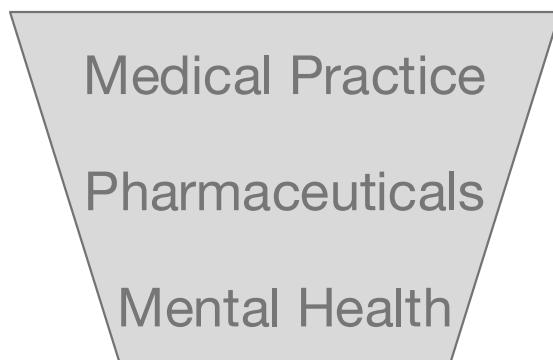
- Variables such as salary range & department are missing for the majority of observations
- **Solution 1:** Discard variables with high number of missing values if there is another similar category that is well-populated
 - Drop department, use function instead
- **Solution 2:** Treat the missing values as a category of their own
 - A missing salary range in a job posting may be predictive of fraud

Data Description: Transformations & Cleaning

Creating Dummies from Categorical Variables

- Categorical variables can be transformed into a series of dummy variables, but some columns contain > 30 categorical values
- **Solution:** bin categories into larger buckets based on their similarity & size

Healthcare Bucket



Creating & Standardizing Quantitative Variables

- Variables such as job description or benefits are long strings and unique to each job posting
- **Solution:** To make them useful in a regression, we can derive quantitative values, such as the length of the job posting
- To avoid feature dominance, we apply z-score normalization to measures of length

Model Specification: Logistic Regression

The **logistic regression** is a Generalized Linear Model that assumes the log-odds of an event are a linear combination of independent variables.

$$\ln(\text{fraud}_i / 1 - \text{fraud}_i) = \beta_0 + \beta_{1,i}x_{1,i} + \dots + \beta_{k,i}x_{k,i} + \varepsilon_i$$

The equation is annotated with three curly braces below it. The first brace groups the term $\ln(\text{fraud}_i / 1 - \text{fraud}_i)$ and is labeled "link function (logit)". The second brace groups the terms $\beta_0 + \beta_{1,i}x_{1,i} + \dots + \beta_{k,i}x_{k,i}$ and is labeled "linear predictor". The third brace groups the term ε_i and is labeled "error".

Why a logistic regression?: Our outcome variable is binary, so a binomial regression is appropriate, and the logit link function lends itself to ease of interpretation.

Model Specification: Logistic Regression

The **logistic regression** uses the Maximum Likelihood Estimation method to estimate parameters.

Key MLE Equations

- Identifies the parameters which maximize the likelihood function:

$$l(\mathbf{y}; \boldsymbol{\theta}) = \sum_{i=1}^n \log f(y_i; \boldsymbol{\theta})$$

- Substituting in the binomial distribution:

$$l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^N \left[y_i \log(p_i) + (n_i - y_i) \log(1 - p_i) + \log \binom{n_i}{y_i} \right]$$

MLE Assumptions

- ✓ Independent observations
- ✓ Sufficient number of observations
- ✓ Observations assumed to be drawn from identical distributions
- ✓ GLM model accurately reflects underlying relationship

Variable Selection: 53 Independent Variables

Included Variable(s)	(Jointly) Significant at 5%?	Domain relevance
Salary Dummies	Yes	Fraudulent postings potentially lure applicants with seemingly high salaries
Industry Dummies	Yes	Fraudulent postings may be less common in industries such as government
Experience (Work or Leadership)	Yes	Fraudulent posting may target those with less experience
Job Type Dummies (Full vs Part Time)	Yes	Fraudulent postings could target those looking for entry level, full-time opportunities
Education Dummies	Yes	Fraudulent postings could target those with lower education levels
Has a logo	Yes	Fraudulent postings might not have a logo
Has screening questions	Yes	Fraudulent posting possibly don't include screenings
Normalized posting lengths	Yes	Fraudulent postings possibly have shorter lengths
Education x Experience Interactions	Yes	Fraudulent postings could require less experience with less education to seem more desirable
Normalized Lengths x Job Types	Yes	Fraudulent postings could have short descriptions for full time roles

Variable Selection: Excluded Variables

Some variables were excluded, for reasons such as:

- Lack of statistical significance (ex. telecommute variable)
- Being too granular of a category (ex. salary x industry interaction terms)
 - See “model limitations” for further details on complete separation
- Creating multicollinearity (ex. industry and function dummies)
 - See “model limitations” for further details on multicollinearity

Results & Interpretation: Model Fit

	Deviance	Degrees of freedom
Intercept Only Model	5322.8	0
Full Model	3532.6	52

- Likelihood Ratio: $5322.8 - 3532.6 = 1790.2$
- p-value from Chi-sq distribution: 0.000
 - Indicates that the full model is statistically different from the intercept-only model

Results & Interpretation

Higher likelihood of fraud:

Variable	Coefficient, p-value	Odds Ratio Interpretation
Salary < \$50,000	1.4, p = 0.000	~ 4x as likely vs missing salary
Above a Bachelor's Degree	1.5, p = 0.012	~ 4.5x as likely vs missing education
Engineering/Manufacturing/Trades Industry	1.9, p = 0.000	~ 6.7x as likely vs missing industry
Finance Industry	0.59, p = 0.000	~ 1.8x as likely vs missing industry
Leisure/Luxury Industry	0.5, p = 0.011	~ 1.6x as likely vs missing industry
Real Estate Industry	0.7, p = 0.005	~ 2x as likely vs missing industry
Benefits Length	0.62, p = 0.000	1 S.D. increase in length = ~1.9x as likely

Results & Interpretation

Higher likelihood of fraud (interactions):

Variable	Coefficient, p-value	Odds Ratio
Bachelor's * Leadership Experience	2.1, p = 0.000	~ 8x as likely vs missing x no experience
Below college * Leadership Experience	2, p = 0.001	~ 7.4x as likely vs missing x no experience
Vocational * Leadership Experience	3, p = 0.003	~ 20x as likely vs missing x no experience
Below college * Work Experience	0.73, p = 0.026	~ 2x as likely vs missing x no experience
Full-time * Profile Length	1.4, p = 0.000	~1.9x as likely ¹
Full-time * Description Length	0.27, p = 0.034	~ 1.3x as likely ¹
Full-time * Requirements Length	0.56, p = 0.002	~ 1.75x as likely ¹
Temporary Position * Profile Length	1.2, p = 0.002	~ 3.3x as likely ²
Temporary position * Requirements Length	0.84, p = 0.003	~ 2.3x as likely ²

1: Specifically, a 1 S.D. increase in length is an additional Y times as likely when the job posting is full-time vs missing type (where Y = odds ratio)

1: Specifically, a 1 S.D. increase in length is an additional Y times as likely when the job posting is temporary vs missing type (where Y = odds ratio)

Results & Interpretation

Lower likelihood of fraud:

Variable	Coefficient, p-value	Odds Ratio
Leadership Experience	-1.6, p = 0.000	~ 0.2x as likely vs no experience
Part-time positions	-1.9, p = 0.010	~ 0.15x as likely vs missing type
Education/Research Industry	-1.9, p = 0.017	~ 0.15x as likely vs missing industry
Technology Industry	-0.38, p = 0.040	~ 0.68x as likely vs missing industry
Has a logo	-1.3, p = 0.000	~ 0.27x as likely vs no logo
Has screening questions	-0.23, p = 0.029	~ 0.79x as likely vs no questions
Normalized Profile Length	-1.8, p = 0.000	1 z-score increase in length = ~.17x as likely
Normalized Description Length	-0.32, p = 0.002	1 z-score increase in length = ~.73x as likely
Normalized Requirements Length	-0.69, p = 0.000	1 z-score increase in length = ~.50x as likely

Results & Interpretation

Lower likelihood of fraud (interactions) :

Variable	Coefficient, p-value	Odds Ratio
Full-time * Benefits Length	-0.3, p = 0.046	~ 0.74x as likely ¹
Temporary position * Benefits Length	-1.4, p = 0.022	~ 0.25x as likely ²

1: Specifically, a 1 S.D. increase in length is an additional Y times as likely when the job posting is full-time vs missing type (where Y = odds ratio)

1: Specifically, a 1 S.D. increase in length is an additional Y times as likely when the job posting is temporary vs missing type (where Y = odds ratio)

Model Limitations: Complete Separation

- Complete separation occurs when an independent variable perfectly predicts the outcome variable (ie. if X is 1, then Y is always 1).
 - Quasi- complete separation: **near** perfect prediction
- Can lead to problems like a lack of convergence, large standard errors, and large coefficient estimates
- Due to the low number of fraudulent job postings, we must be careful when defining independent variables to avoid separation in our model

Cross-tabulated Count of Observations

	fraudulent=0	fraudulent=1
(Low Salary & Engineering Industry) = 0	9920	730
(Low Salary & Engineering Industry) = 1	4	0

Engineering jobs with a low salary have no fraudulent observations – this is an example of complete separation and why we must be careful in creating in adding interaction terms.

Addressing Other Limitations

Multicollinearity: Increases Standard Errors

- 44/52 variables have a low (<5) VIF score, but there is some multicollinearity between certain terms
 - Benefits length & benefits length x full-time (**.88** correlation)
 - Bachelor's & Bachelor's x leadership experience (**.59** correlation)
- We ultimately kept these variables due to their statistical significance & the fact that standard errors were a reasonable size

Outliers: Bias Coefficients & Increase Standard Errors

- Our independent variables are mostly dummy variables: with values of only 0 and 1, there are no outliers
- For quantitative variables such as normalized description length, we removed values with a z score of over 3
 - This removed around 700 observations (around 6%) but limited potential bias

Predictions (80/20 Test/Train Split)

ROC_AUC: 0.86

- ❖ 86% probability the model assigns a higher fraud score to a fraudulent posting than an authentic posting

Confusion Matrix

		Predicted	
		0	1
Actual	0	1973	13
	1	125	21

Sensitivity Analysis

Scenario

(Predicted P/
Actual P)

Postulated 3 categorical
match “scenarios” with a
mixture of top features
(informed by XGBoost)
and their data points i.e.
“has_logo: 0”, “salary:
low”, “country: US”,
“type: temporary”
for the predicted
probability of fraud vs.
actually observed.

Baseline Fraudulent Posting

Predicted Probability of Fraud:

0.0080 (0.80%)

Actual Observed Fraud Rate:

0.0224 (2.24%) - 715 matches

Non - Fraudulent Posting

Predicted Probability of Fraud:

0.0134 (1.34%)

Actual Observed Fraud Rate:

0.0692 (6.92%)- 130 matches

Mixed Characteristic Posting

Predicted Probability of Fraud:

0.0172 (1.72%)

Actual Observed Fraud Rate

0.0560 (5.60%)- 5603 matches

Industry Variable

Modeled Probability of
Risk/Safety in type of
Industry listed in the job
description.

Engineering	28.75%
Real Estate	11.92%
Finance	9.61%
Education	0.45%
Healthcare	0.95%
Govt/Nonprofit	1.17%

Conclusions

Takeaways

- XGBoost techniques show the strongest indicators of an authentic job posting are having a company logo, being in the finance or tech industries, and requiring leadership experience
- In contrast, strong indicators of a fraudulent posting are Engineering jobs or having a low posted salary

Practical Relevance

- Results can assist job seekers in detecting fraudulent postings by encouraging them to be careful looking into job listings when applying, and be weary of things such as industry, salary, and the existence of a company logo

Possible Extensions

- Re-running the model on new job postings which could identify ways that fraudsters are changing their techniques over time
- An algorithm could be developed that reports the likelihood of a job posting being fraudulent in real-time

Questions?

Thank you

