INF 1340: Programming for Data Science

Final Project
Detecting Fraudulent Job Postings:
An Exploratory and Predictive Analysis

Yifei Zhang
1011618466

Introduction

Online job fraud has become a growing concern in recent years, as cybercriminals exploit digital recruitment platforms to target job seekers. Fraudulent job postings can lead to financial loss, identity theft, and emotional distress for applicants (Indeed Editorial Team, 2023). Research has shown that fake job ads often share common patterns, such as incomplete company information, unusually short or vague job descriptions, and language that appears overly informal or exaggerated (Jain et al., 2018; Chiraratanasopha & Chay-intr, 2022). Identifying these patterns can help platforms and users detect scams before they cause harm.

This study uses the *Real or Fake Job Posting Prediction* dataset, publicly available on Kaggle and originally compiled by Shivam Bansal. The dataset contains 17,880 job postings collected from an online job portal, each labeled as either fraudulent or legitimate. It includes both structured fields (e.g., employment type, required experience, remote status) and unstructured text fields (e.g., job description, company profile, requirements), making it suitable for exploring how different types of features can be combined to detect fraudulent postings.

By analyzing this dataset, the project aims to identify the key characteristics that distinguish fake postings from real ones, providing insights that may help improve online job safety and guide future fraud detection systems.

Research Question

This project aims to identify the key features that most strongly differentiate fraudulent job postings from legitimate ones. The central research questions are:

1. Are postings with short or missing company descriptions more likely to be fraudulent?
2. Are postings with vague or unusually short job descriptions more likely to be fraudulent?
3. Do remote jobs (telecommuting positions) exhibit a higher rate of fraud compared to on-site jobs?
4. Does the textual content of fraudulent job postings contain more informal or exaggerated language than legitimate ones?

These questions are well-defined and directly measurable within the dataset, which contains both structured fields (e.g., company profile, telecommuting) and unstructured text fields (e.g., job description). They are grounded in prior literature and common patterns in online recruitment scams, ensuring relevance. The questions are also feasible to investigate with the available features and the course-covered methods, including exploratory data analysis, categorical variable comparison, and introductory natural language processing for text content.

Methods and Exploratory Data Analysis

The analysis began with data cleaning and organization using Pandas in Python,

following TidyData principles to ensure each variable formed a column, each observation a row, and each type of observational unit a table. Missing values were inspected using conditional statements and loops to identify incomplete records, with binary indicator variables created for missing company_profile entries and computed description lengths. These preprocessing steps were essential for enabling subsequent exploratory and predictive analyses.

Exploratory data analysis was then conducted to understand the structure and patterns within the dataset. Descriptive statistics and cross-tabulations were used to examine relationships between categorical predictors and the target variable (fraudulent). For example, a cross-tabulation and bar chart of telecommuting against fraudulent revealed that remote positions had a higher proportion of fraudulent postings compared to on-site roles.

fraudulent	0	1
telecommuting		
0	0.953135	0.046865
1	0.916558	0.083442

Table 1 - telecommuting vs Fraudulent

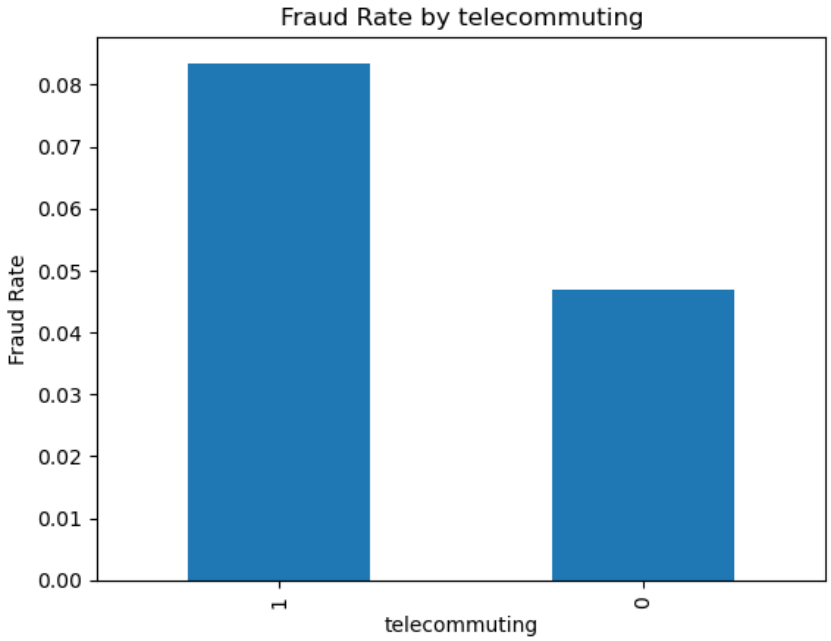

Figure 1 - telecommuting vs fraudulent bar chart

Building on this, further analysis examined textual completeness, showing that missing or extremely short company profiles and job descriptions were more prevalent among fraudulent postings.

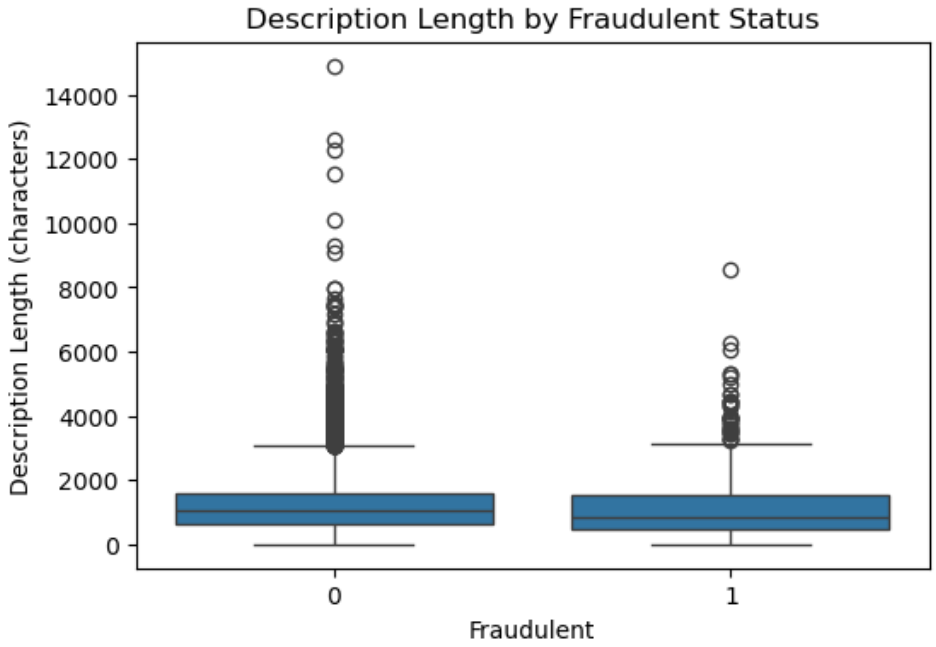

Figure 2 - Description length vs Fraudulent box plot

Categorical variables also revealed noticeable differences between fraudulent and legitimate postings. Bar charts comparing employment_type and required_experience categories against the proportion of fraudulent postings show distinct patterns. Part-time exhibited a higher share of fraudulent postings, whereas some employment types, such as temporary roles, showed lower fraud prevalence. Similarly, required experience levels at the extremes, either “Entry level” or “executive” tended to have higher fraud rates than mid-level roles. These patterns provide further evidence that job type and stated experience requirements can serve as indicators of potential fraud.

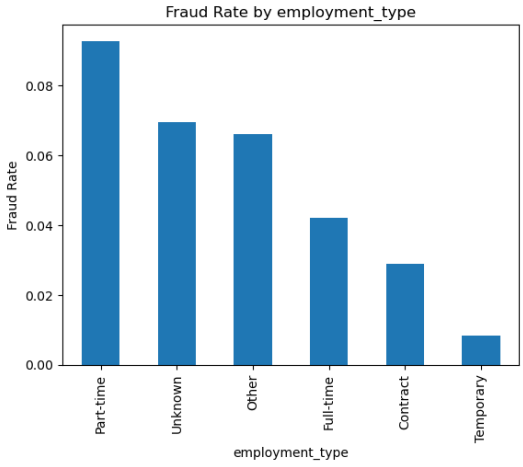

Figure 3 - employment_type vs Fraudulent bar chart

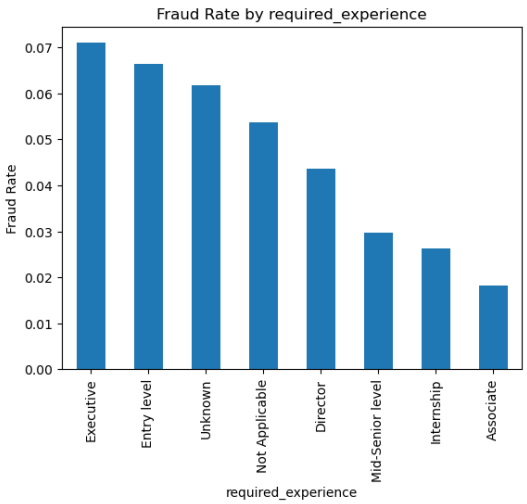

Figure 4 - required_experience vs Fraudulent bar chart

For text-based fields such as description, introductory natural language processing techniques were applied, specifically sentiment analysis from the NLTK library, to assess whether fraudulent postings tend to contain more informal or exaggerated language. This method was chosen because it allowed quantifying stylistic differences without requiring complex linguistic models. While NLP sentiment scores can capture general tone, they may overlook domain-specific scam language, which is a limitation. This directly addresses the research question on whether fraudulent postings contain more informal or exaggerated language.

Finally, to assess the predictive power of these features, a classification model was built using logistic regression from scikit-learn. Predictor variables included both numeric features (e.g., description length, binary missing indicators) and encoded categorical features (e.g., employment type, required experience). The data was split into training and test sets (75%/25%), and model performance was evaluated using a confusion matrix and classification report. Logistic regression was selected because it provides interpretable coefficient estimates, allowing identification of features most strongly associated with fraudulent postings. To address the dataset's class imbalance, a second model was trained with class_weight='balanced', improving recall for the fraudulent class. While logistic regression offers clear interpretability and is well-suited for binary classification, its performance can be sensitive to multicollinearity and may not capture complex non-linear relationships.

Results

The analysis identified several patterns that distinguish fraudulent job postings from legitimate ones.

First, telecommuting status showed a clear relationship with fraud likelihood. Remote positions had a higher proportion of fraudulent postings (8.3%) compared to on-site roles (4.7%). This pattern was evident in both the cross-tabulation and the bar chart. (Please see Table 1 and Figure 1)

Second, fraudulent postings were more likely to lack a company profile and to feature shorter job descriptions. The average description length for fraudulent postings was 1,154 characters, compared to 1,221 characters for legitimate postings. These differences were visualized in the box plot of description length versus fraudulent status. (Please see Figure 2)

fraudulent	Description length
0	1221.219701
1	1154.834873

Table 2 - Description length vs Fraudulent

Third, categorical variables also revealed strong patterns. Fraud rates varied substantially by employment type, with part-time positions having the highest proportion of fraudulent postings, while temporary roles were overwhelmingly

legitimate. Required experience followed a U-shaped pattern, where both entry-level and executive-level positions showed higher fraud rates than mid-level roles.
(Please see Figure 3 and 4)

Fourth, sentiment analysis using the VADER model suggested modest stylistic differences between fraudulent and legitimate postings. Fraudulent job descriptions had slightly lower average positive sentiment (0.139) and higher negative sentiment (0.0135) compared to legitimate postings (positive 0.154, negative 0.0107). The average compound sentiment score, which summarizes overall sentiment on a scale from -1 (most negative) to +1 (most positive), was also lower for fraudulent postings (0.793) than for legitimate ones (0.850). Although these differences were small, the consistent pattern across sentiment metrics suggests a potential, but limited, role for sentiment in fraud detection.

fraudulent	positive	negative	neu	compound
0	0.153887	0.010717	0.835392	0.849566
1	0.139435	0.013527	0.844758	0.793105

Table 3 - Sentiment Analysis

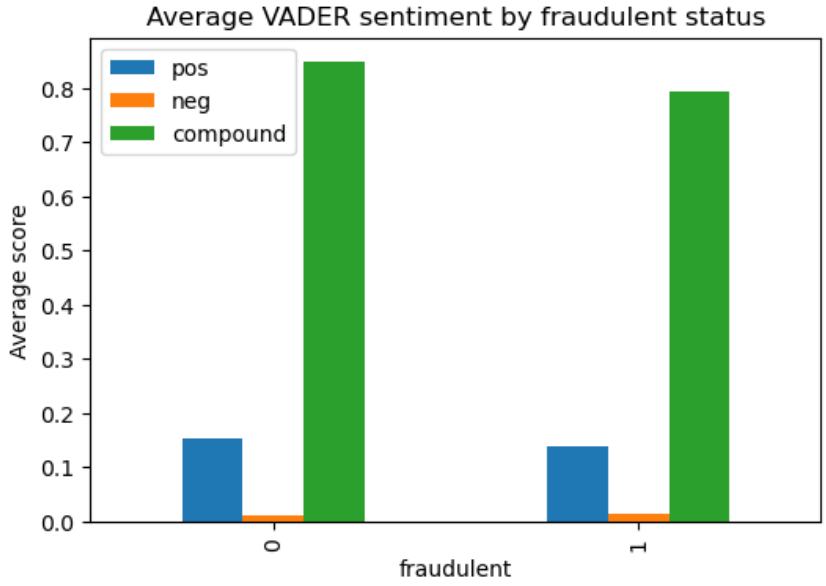

Figure 5 – Ave VADER sentiment by fraudulent status

Fifth, word frequency analysis revealed distinctive lexical patterns. Fraudulent postings frequently used terms such as “work,” “service,” and “customer,” whereas legitimate postings often emphasized “team,” “customer,” and “service,” but with substantially higher absolute frequencies. These differences indicate that word usage patterns, while overlapping, may reflect subtle distinctions in tone and emphasis between fraudulent and legitimate job ads.

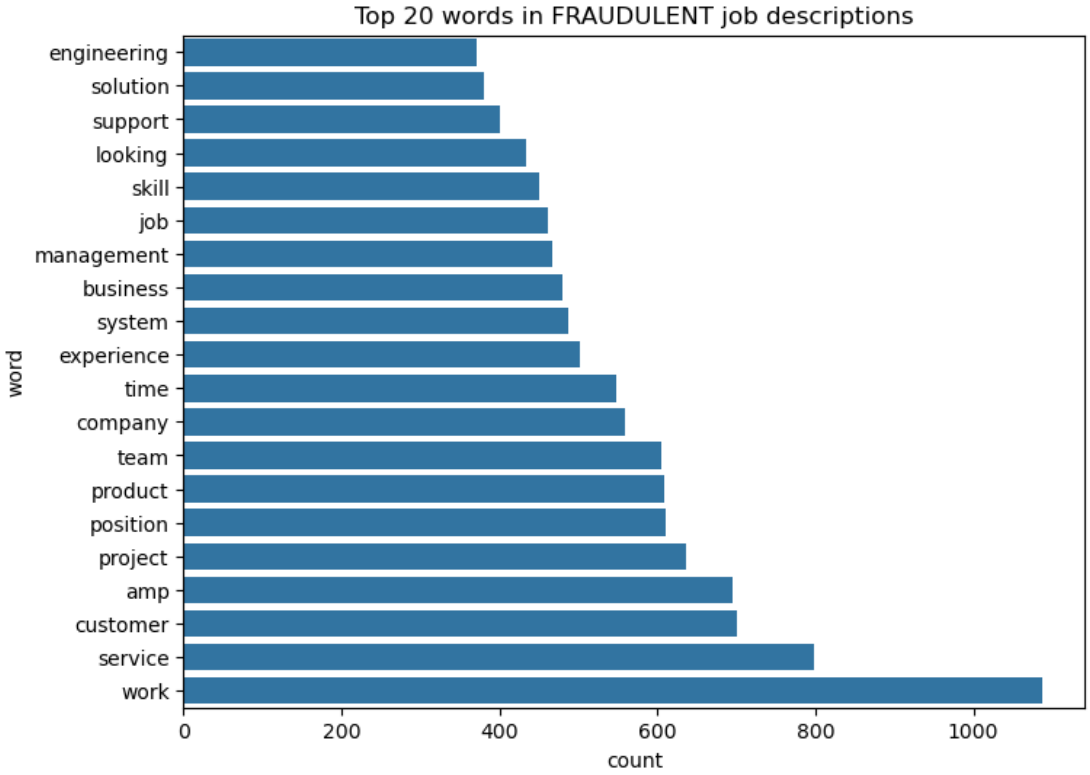

Figure 6 – Top 20 word in fraudulent job descriptions

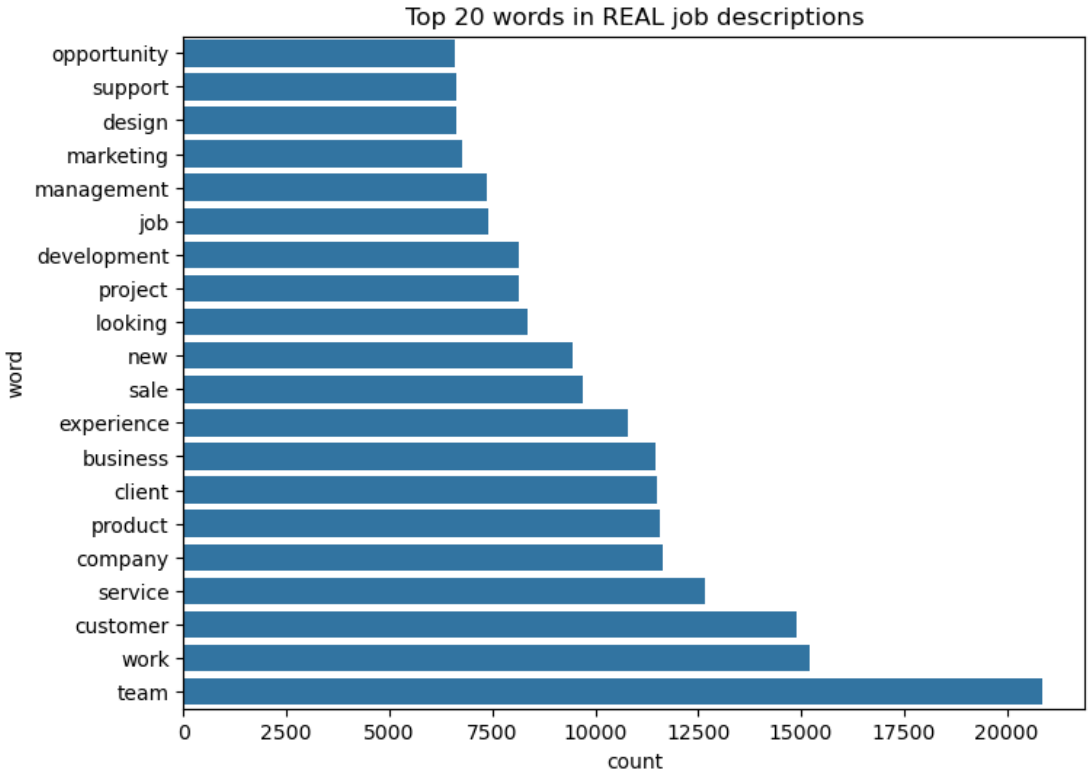

Figure 7 – Top 20 word in real job descriptions

Finally, predictive modeling was conducted using logistic regression. An initial model without class balancing failed to identify any fraudulent postings due to class imbalance. After applying class_weight='balanced', the model achieved an accuracy of 0.83 and a recall of 0.65 for the fraudulent class. Feature coefficient analysis revealed that missing company profiles (2.52), entry-level (0.83) and director-level experience requirements (0.52) were most strongly associated with higher fraud likelihood, whereas temporary employment (-0.53) and internships (-0.52) were negatively associated.

Confusion Matrix:				
4273	0			
197	0			
Classification Report:				
	precision	recall	f1-score	support
0	0.96	1.00	0.98	4273
1	0.00	0.00	0.00	197
accuracy			0.96	4470
macro avg	0.48	0.50	0.49	4470
weighted avg	0.91	0.96	0.93	4470

Table 4 – Model 1

Confusion Matrix:				
3579	694			
69	128			
Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.84	0.90	4273
1	0.16	0.65	0.25	197
accuracy			0.83	4470
macro avg	0.57	0.74	0.58	4470
weighted avg	0.94	0.83	0.87	4470

Table 5 – Model 2

Discussion

The results provide several insights into the features that can separate fraudulent job postings from real ones.

First, remote roles had a higher share of fraudulent postings. This matches earlier research showing that scammers often use the appeal of working from home to reach more job seekers (Indeed Editorial Team, 2023). Remote jobs may be easier for scammers to post because applicants cannot easily verify the employer.

Second, many fraudulent postings had missing or very short company profiles and descriptions. This supports the finding from Chiraratanasopha and Chay-intr (2022) that fake postings often avoid giving clear company information. This means that job boards could review postings with limited company details more closely.

Third, some categories of employment type and required experience showed clear fraud patterns. Part-time jobs and postings for entry-level or director-level applicants had higher fraud rates. Scammers may target less experienced job seekers who are more likely to trust an offer, and they may also use high-level job titles with vague requirements to attract people interested in quick promotion.

Fourth, sentiment analysis showed small differences in style between fraudulent and real postings. Fraudulent postings had slightly lower positive sentiment, higher negative sentiment, and lower compound scores. These differences were not large enough to be useful on their own, but they may help when combined with other features.

Finally, the logistic regression model showed that a small group of features such as missing company profiles, remote status, part-time jobs, and extreme experience levels can help detect possible scams. Using class_weight="balanced" improved the recall for fraudulent postings, which is important for finding rare cases like scams. However, the low precision means the model also predicts more false positives, which could be a challenge if used for automatic job posting checks.

These findings lead to the conclusion that fraud detection benefits from a combination of multiple indicators rather than relying on a single feature.

Conclusion

This project examined the characteristics of fraudulent job postings using exploratory data analysis, sentiment analysis, and a logistic regression model. The results showed that missing or short company information, remote work status, certain employment types, and extreme required experience levels are linked to higher fraud risk. Sentiment differences were small and not effective on their own but may be useful when combined with other features. The classification model was able to improve fraud detection when class imbalance was addressed, though false positives remain an issue. Overall, the study suggests that a mix of text-based, categorical, and numerical features can help identify suspicious postings, and that automated systems should be supported by human review to ensure accuracy.

References

Indeed Editorial Team. (2023, August 31). *How to know if a job is a scam*. Indeed. https://www.indeed.com/career-advice/finding-a-job/how-to-know-if-a-job-is-a-scam?utm_source=chatgpt.com

Jain, A., Gupta, V., & Bhutani, A. (2018). Detection of fraudulent job postings using hybrid approach. *Future Internet, 9*(1), 6. https://doi.org/10.3390/fi9010006

Chiraratanasopha, B., & Chay-intr, T. (2022). Detecting fraud job recruitment using features reflecting from real-world knowledge of fraud. *Current Applied Science and Technology, 22*(6). https://doi.org/10.55003/cast.2022.06.22.008

Shivam Bansal. Real or Fake Job Posting Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction (Accessed June 2025)