NLP-based Classification of Fraudulent Job Postings on LinkedIn

Oct 2023

Minh Duong
Springboard DSC - Capstone Project 2

Table of contents

Table of contents **2**
Problem Statement **3**
Data Collection **5**
Exploratory Data Analysis **7**
Data Preprocessing **13**
Modeling and Model Selection **14**
Conclusion **17**
Future Works **18**

Problem Statement

In the age of digital recruitment, online platforms like LinkedIn serve as pivotal tools for connecting job seekers with potential employers. However, the rise of fraudulent activities, particularly in the form of misleading or deceptive job postings, poses a significant challenge to the integrity of these platforms. The objective of this project is to develop a robust classification model capable of identifying and flagging potentially fraudulent job postings on LinkedIn.

Background

- Platform Vulnerability: LinkedIn, being a popular professional networking platform, attracts a diverse audience, including both legitimate employers and malicious entities.
- Impact on Users: Fraudulent job postings can mislead job seekers, leading to financial scams, identity theft, or phishing attacks.

Objective

Develop a machine learning model that can automatically classify job postings on LinkedIn as either legitimate or fraudulent based on various features associated with the job descriptions, company profiles, and other relevant information.

Key Challenges

- Imbalanced Data: Fraudulent job postings are likely to be significantly outnumbered by legitimate ones, leading to class imbalance.
- Evolving Tactics: Fraudsters adapt their strategies, making it essential to build a model capable of recognizing new and diverse patterns of fraudulent behavior.
- Multimodal Data: Information for classification is likely to be spread across various features such as textual content, company details, and potentially user interactions.

Scope of the Project

- Data Collection: Gather a diverse dataset of job postings from LinkedIn in the United States, including both legitimate and confirmed fraudulent instances.
- Feature Engineering: Extract relevant features from job descriptions, company profiles, and other associated metadata.
- Model Development: Employ machine learning algorithms, possibly leveraging natural language processing (NLP) techniques for text analysis.
- Evaluation Metrics: Utilize metrics such as precision, recall, and F1 score, considering the imbalanced nature of the data.

Expected Outcomes

- A robust model capable of accurately identifying fraudulent job postings.

- Insights into the key features and patterns associated with fraudulent activity on LinkedIn.
- Recommendations for improving the overall security and trustworthiness of job postings on the platform.

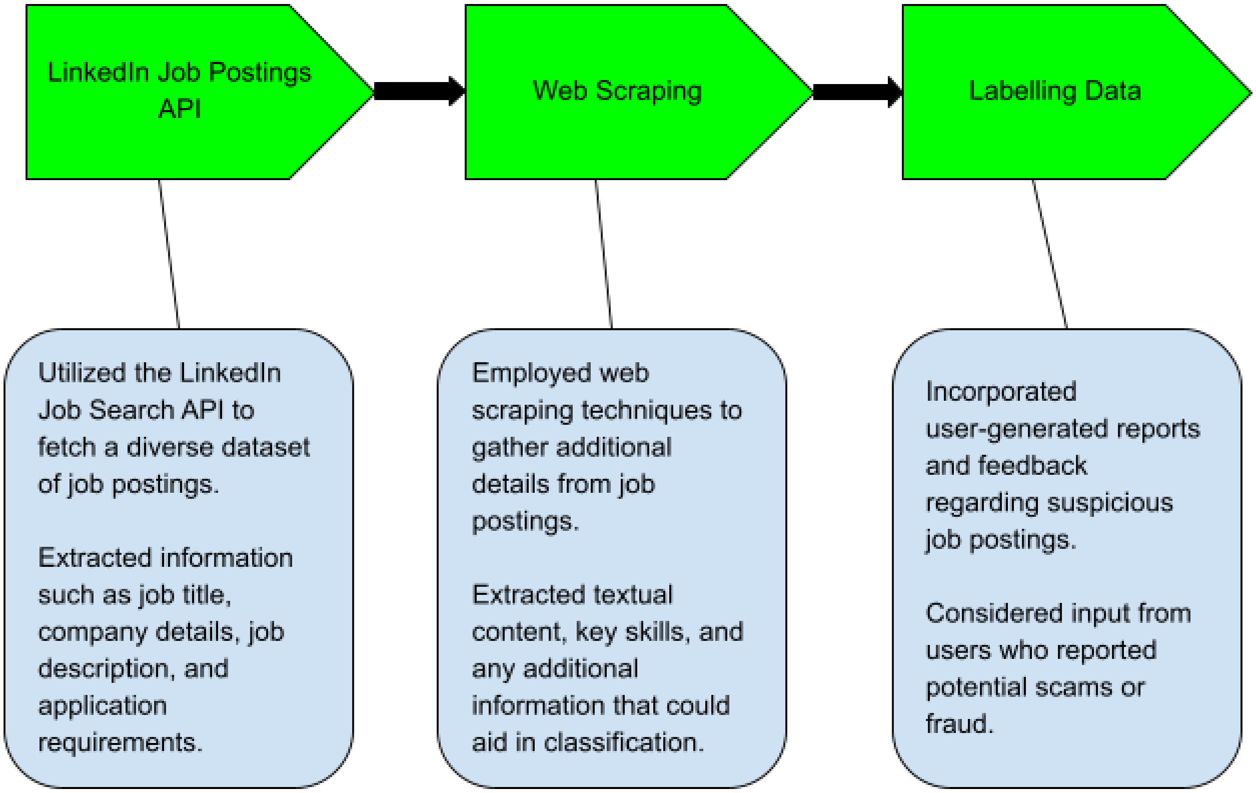
Significance

The successful implementation of this project not only enhances the security of LinkedIn users but also contributes to the broader goal of maintaining the integrity and reliability of online professional networking platforms.

Data Collection

Procedures

The data set for this project in real-life situation could be obtained by the following procedure.



Utilized the LinkedIn Job Search API to fetch a diverse dataset of job postings.

Extracted information such as job title, company details, job description, and application requirements.

Employed web scraping techniques to gather additional details from job postings.

Extracted textual content, key skills, and any additional information that could aid in classification.

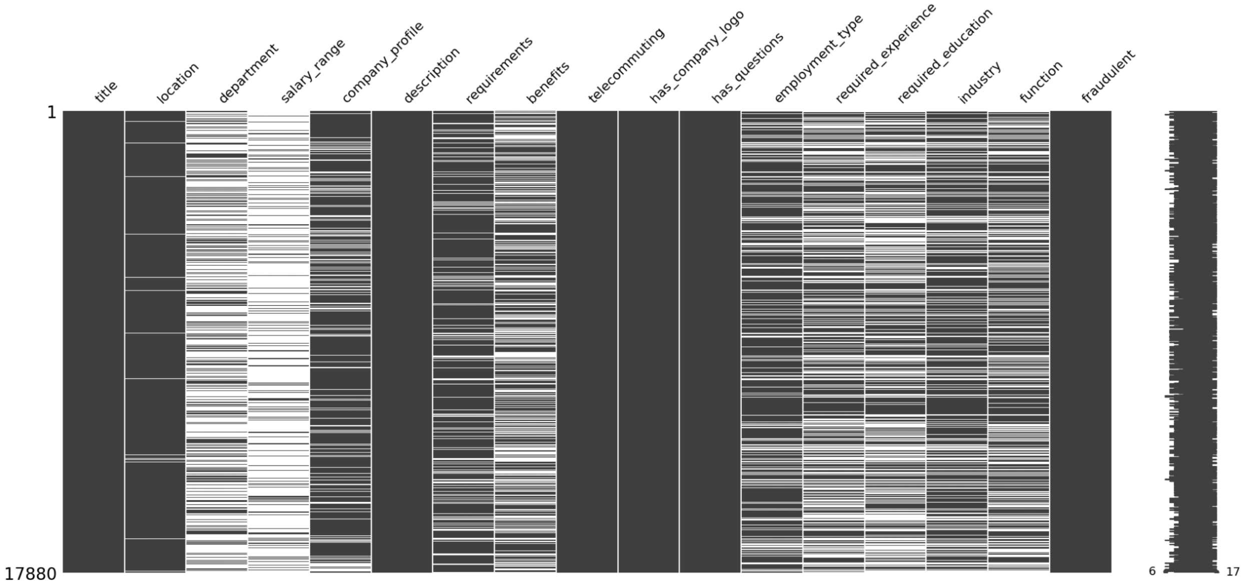
Incorporated user-generated reports and feedback regarding suspicious job postings.

Considered input from users who reported potential scams or fraud.

For the simplicity of this project, the data set is obtained from Kaggle's "Real / Fake Job Posting Prediction" dataset. The dataset has 17 columns in total. The column "fraudulent", which is in binary, labels each entry whether it is a fraudulent job posting or not.

Dealing with missing values

The missingno matrix shows that this dataset has many missing values.

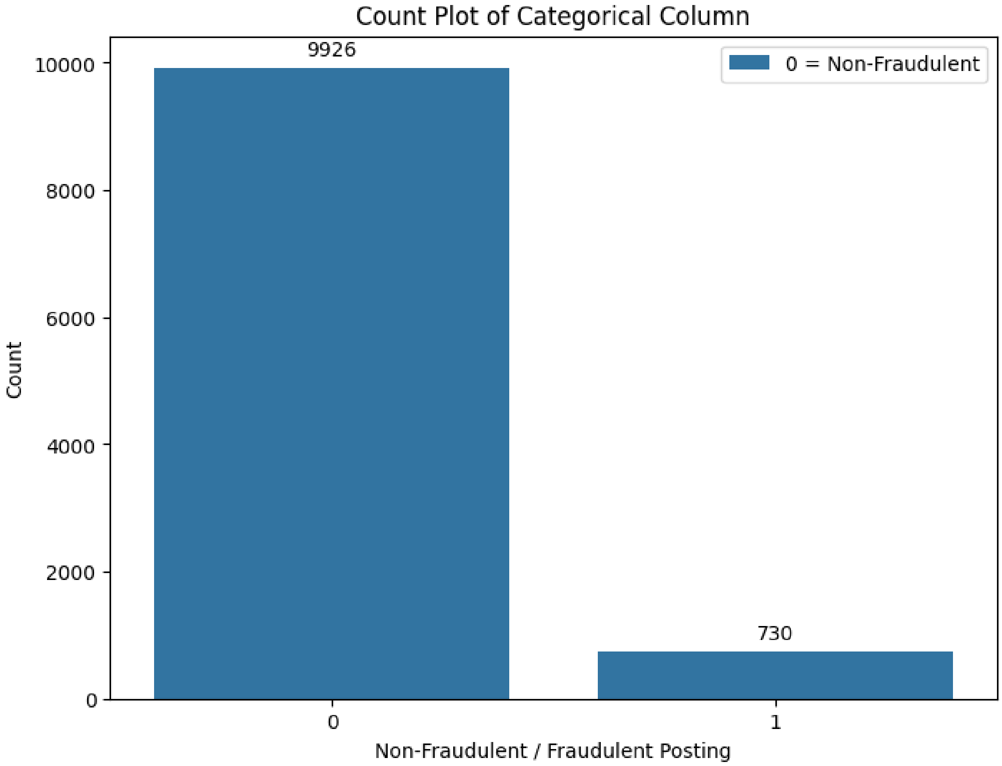


However, looking closely at the columns with missing values, we can simply replace those missing values with a generic “Not specified” value.

Exploratory Data Analysis

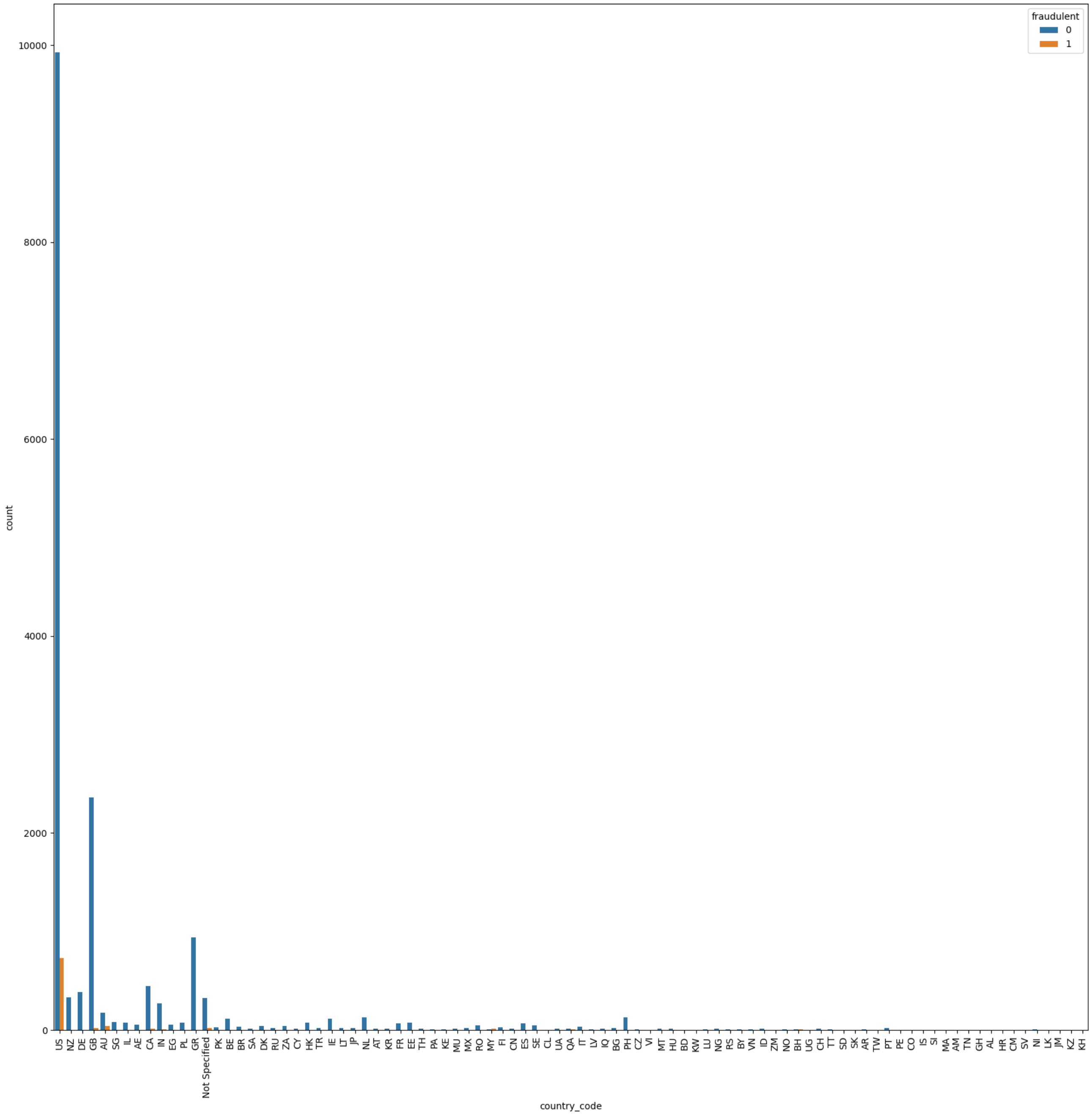
Data Imbalance

The data is very imbalance with 93% of the datapoint is labeled as ‘Non-fraudulent’. This is a typical problem with fraud detection problem. We will stratify the dataset when splitting the dataset into train/test set.



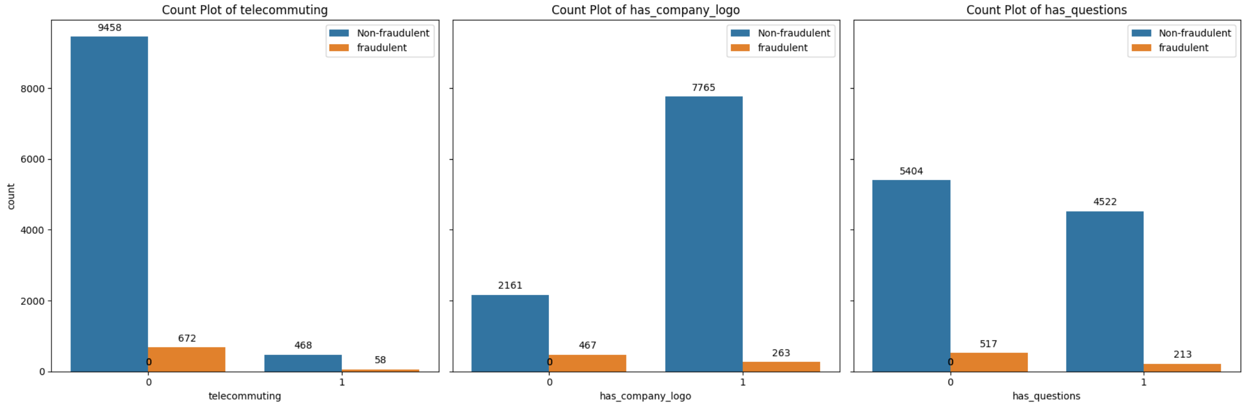
Job Postings by Location

From the count plot that group by country, we see that most of the job postings in this dataset are in the United States. As the scope of the project, we will focus only on those postings that locate in the US.



Numerical Features

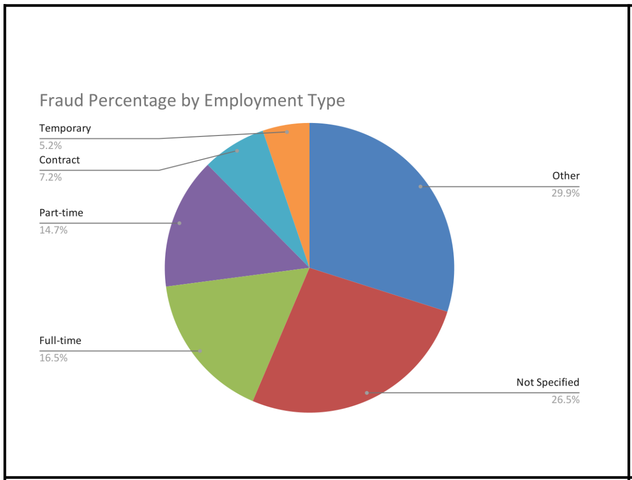
The numerical features, [“telecommuting”, “has_company_logo”, “has_question”, are actually in binary, which indicate whether a job posting has not does not has a certain feature. We use count plots to see the fraud rate for each feature.

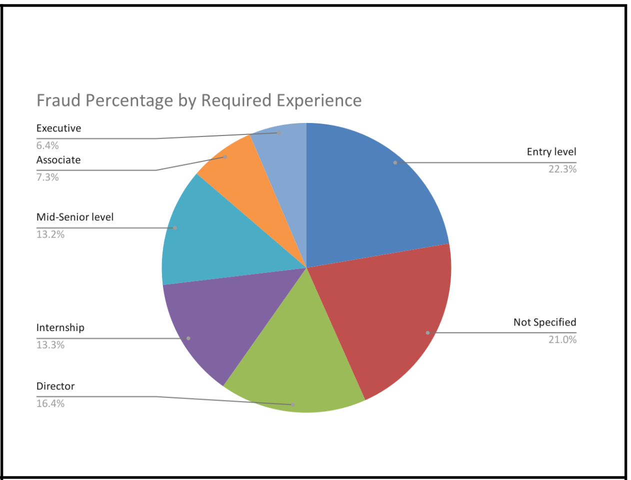


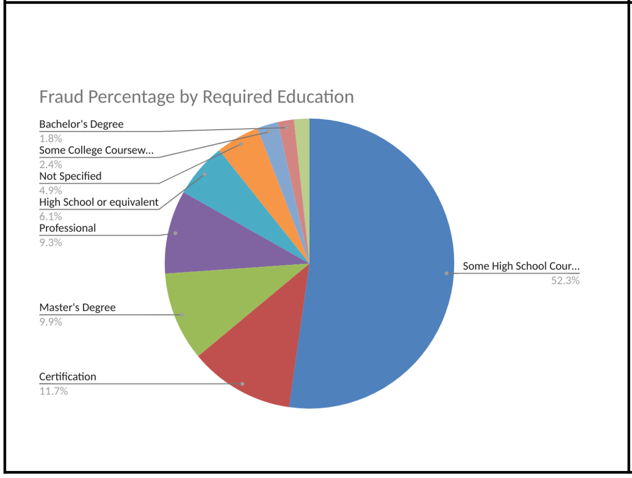
These three features provide a good rule of thumb for predicting fraudulent job postings. A job that provides telecommuting, and/or has no company logo, and/or does not has any questions asked is clearly a red flag.

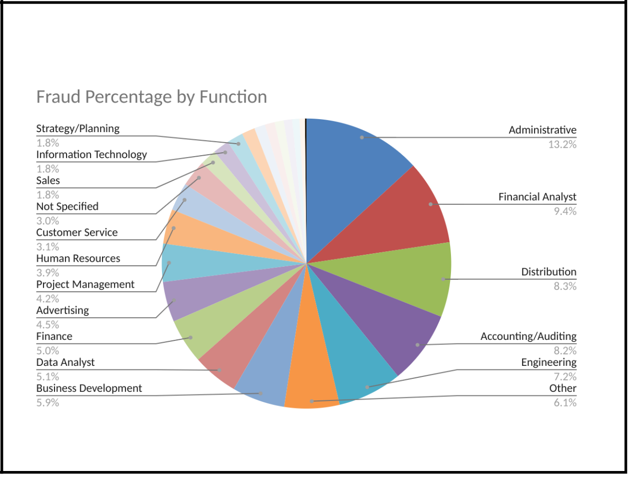
Categorical Features

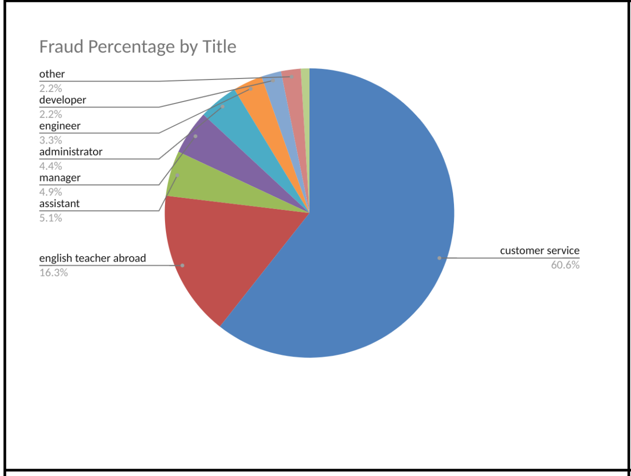
- One of the big problem with the categorical features is that some of them have too many distinct values.
- For example, there are 6294 unique values for a total of 10656 values in the “title” feature, 1744 unique values for a total of 10656 values in “location” feature, or 697 unique values for a total of 10656 values in “department” feature.
- To narrow down the unique values of those features, we use KMean Clustering method to regroup those features into fewer categories.
- We will then calculate the fraud rate for each feature. The result is summarized in the below charts.



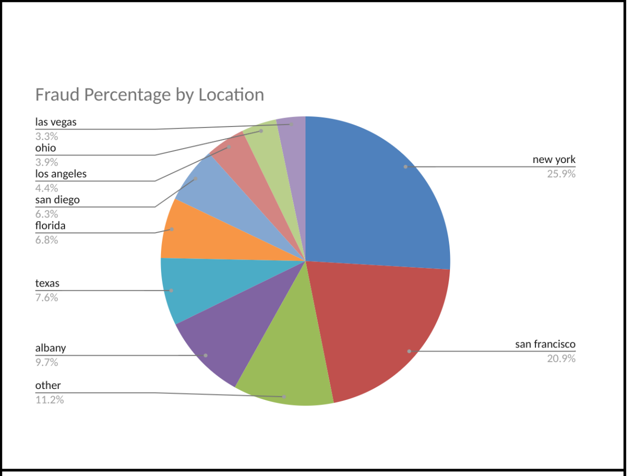




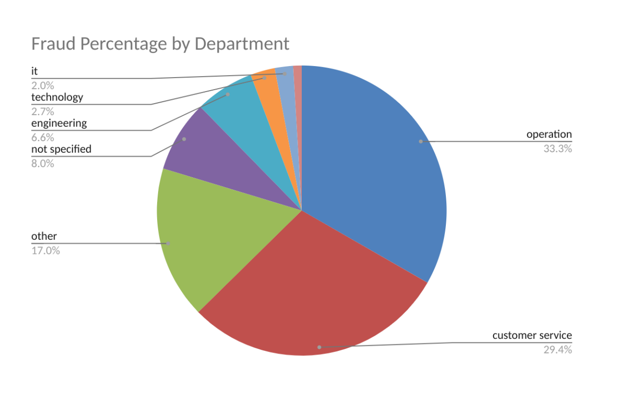




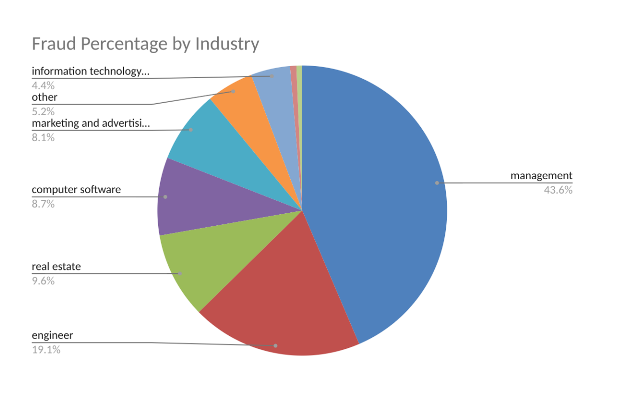
Fraud Percentage by Title



Fraud Percentage by Location



Fraud Percentage by Department



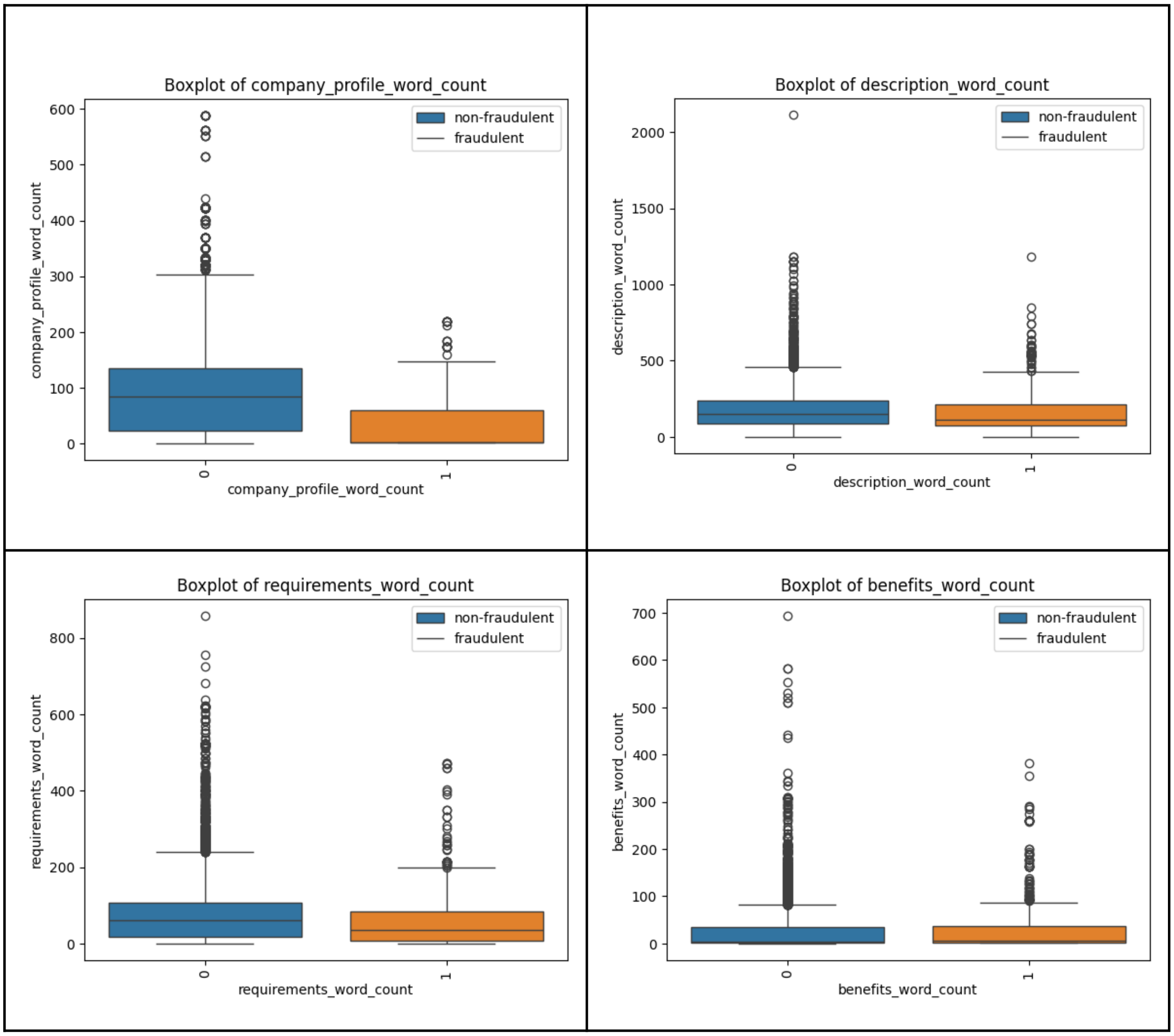
Fraud Percentage by Industry

Key insights

- More than half of the fraudulent job postings do not have a proper description of employment type.
- Most of fraudulent job postings are entry-level positions or not properly defined.
- More than half of fraudulent job postings require only some highschool education.
- Customer service is the type of job that has the highest rate of fraudulent job postings.
- Most of the fraudulent job postings are in New York City or San Francisco, which is quite straight forward since those are among the biggest cities in the US.

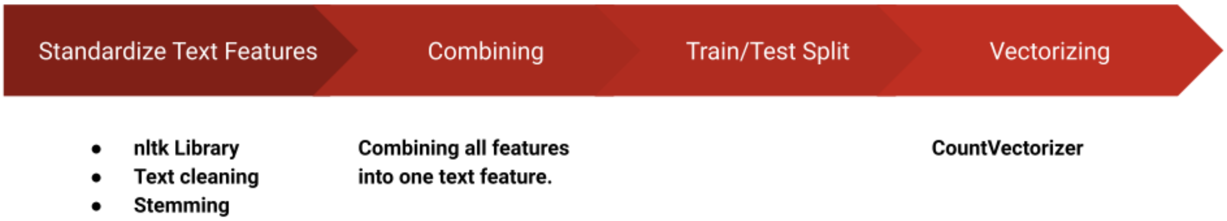
Text Features

- Some features are considered as text due to their distinct characteristics. In this EDA step, we will only consider one of the characteristic of text features, which is the word count.
- We use box plots to see the difference in the word counts of the text features between non-fraudulent job postings and fraudulent ones.
- We can see from the plot that non-fraudulent job postings have more words in each text feature on average. This is quite intuitive since fraudsters usually less careful with their posting.



Data Preprocessing

- We will use Natural Language Processing technique to build a predictive model.
- Target value is the “fraudulent” column which labels 0 or 1 for non-fraudulent/fraudulent.
- We’ll combine other features into one single text feature.
- We will split the dataset into train/test sets.
- We apply text vectorizing method such as CountVectorizer to transform text data to numerical matrix.



Standardize Text Features

- nltk Library
- Text cleaning
- Stemming

Combining

Combining all features into one text feature.

Train/Test Split

Vectorizing

CountVectorizer

Modeling and Model Selection

Models

We employ the following three models for classification.

- Naive Bayes Classifier for Multinomial Models
- Linear Support Vector Machine
- Logistic Regression

Performance Metrics	**NB**	**SVM**	**Logit**
Recall	0.73	0.6	0.84
F1	0.73	0.73	0.87
Precision	0.73	0.95	0.90
Accuracy	0.96	0.97	0.98

For fraud detection problem, we will choose the model that has the highest recall and F1 score. The logistic regression model has the highest target score.

Model Optimization

We employ two cross-validation search method to fine-tune the hyper-parameters of the logistic regression model.

Hyper-parameters for Logistic Regression Model

- C
- Solver
- warm_start

GridSearchCV

Wait time	CPU times: user 15min 55s, sys: 818 ms, total: 15min 56s Wall time: 6min 29s
Optimal Hyperparameters	C: 100 solver: lbfgs warm_start: True
Recall	0.8424657534246576
F1	0.87

RandomSearchCV

Wait time	CPU times: user 27.2 s, sys: 5.72 s, total: 32.9 s Wall time: 17min 22s
Optimal Hyperparameters	warm_start: False solver: liblinear C: 10
Recall	0.8424657534246576
F1	0.87

Final Model

The two hyper-parameter methods give the same result. However, the Random Search method takes more time (nearly 3 time longer) and computing power. We should, therefore, choose the model from the Grid Search method as the best model.
Due to computing power constrain, we are unable to do more intensive hyper-parameter tuning. With more resource, we can perform more intensive parameters search.

Parameters	**Value**
C	100
class_weight	None
dual	False
fit_intercept	True
intercept_scaling	1
l1_ratio	None
max_iter	100
multi_class	auto
n_jobs	None
penalty	l2
random_state	1
solver	lbfgs
tol	0.0001
verbose	0
warm_start	True

Conclusion

This project has the potential to make a significant contribution to the fight against fraudulent job postings on LinkedIn. By developing a robust machine learning model, it can help to protect job seekers from scams and other forms of harm. Additionally, the insights gained from the project can be used to improve the overall security and trustworthiness of job postings on the platform.

I am particularly interested in the challenge of imbalanced data. Fraudulent job postings are likely to be a small fraction of the overall dataset, which could make it difficult for a machine learning model to learn from them. However, there are a number of techniques that can be used to address this challenge, such as oversampling the minority class or using cost-sensitive learning algorithms.

Another challenge that I will need to address is the evolving tactics of fraudsters. Fraudsters are constantly developing new ways to deceive job seekers, so it is important to build a model that is capable of recognizing new and diverse patterns of fraudulent behavior. This can be done by retraining the model on a regular basis and by incorporating new features that are indicative of fraudulent activity.

This project has the potential to make a positive impact on the lives of millions of job seekers and to help maintain the integrity and reliability of online professional networking platforms.

Future Works

- Creating a complete data collection pipeline for real-time data from LinkedIn and other job sites, such as Indeed, Glassdoor etc.
- Employing oversampling/undersampling technique to deal with imbalance data set.
- Employing different text vectorizing methods, such as TF-IDF or word2vec, to see if we can come up with a better model.
- Employing more classification models.
- Employing more hyper-parameters tuning method.
- Running more intensive hyper-parameter tuning.
- Creating a complete user-friendly tool to employ the best model to real-time data.