Dataset: Fake or Real Job Posting

Data description:

Variables of the dataset: [1]

Binary

Telecommuting	True for telecommuting positions.
Company logo	True if the company logo is present.
Questions	True if screening questions are present.
Fraudulent	Classification attribute.
In balanced	Selected for the balanced dataset
String	

Name	Description
Title	The title of the job ad entry.
Location	Geographical location of the job ad.
Department	Corporate department (e.g. sales).
Salary range	Indicative salary range (e.g. \$50,000-\$60,000)

HTML fragment

Company profile	A brief company description.
Description	The details description of the job ad.
Requirements	Enlisted requirements for the job opening.
Benefits	Enlisted offered benefits by the employer.

Nominal

Employment type	Full-type, Part-time, Contract, etc.
Required experience	Executive, Entry level, Intern, etc.
Required education	Doctorate, Master's Degree, Bachelor, etc.
Industry	Automotive, IT, Health care, Real estate, etc.
Function	Consulting, Engineering, Research, Sales, etc.

Dataset: Fake or Real Job Posting

Data Cleaning:

The original dataset contains a 17800 job posting.

The feature job_id is an index and I decide to drop that feature. Then, regarding the percentage of missing values (Figure 1) the 'department', 'salary_range', and 'benefits' were dropped from the dataset.

title	0.000000
location	0.019351
department	0.645805
salary_range	0.839597
company_profile	0.185011
description	0.000056
requirements	0.150727
benefits	0.403244
telecommuting	0.000000
has_company_logo	0.000000
has_questions	0.000000
employment_type	0.194128
required_experience	0.394295
required_education	0.453300
industry	0.274217
function	0.361018
fraudulent	0.000000
dtype: float64	

Figure 1

The next step was counting the NA values and filling them properly. (Figure 2)

title	0
location	346
company_profile	3308
description	1
requirements	2695
telecommuting	0
has_company_logo	0
has_questions	0
employment_type	3471
required_experience	7050
required_education	8105
industry	4903
function	6455
fraudulent	0
dtype: int64	

Figure 2

Dataset: Fake or Real Job Posting

The backward method was used to fill the NAs values for features 'employment_type', 'required_experience', 'required_education', 'industry', and 'function' after sorting them based on the title of job ads. And drop the rest NAs and duplicate rows from the dataset.

In the following step, merging the 'description', 'requirements' and 'company_profile' features were done to have only one 'description' for each job. And based on the location feature the city and country of the job post were split into the different features namely, city and country. The clean data has a 11272 job posting.

The data cleaning codes are available in <u>GitHub</u> for both R and Python language.

Exploratory Analysis:

The distribution of the target variable shows that the dataset is highly imbalanced with 11023 real job posting and 249 fake job advertisements. (Figure 3)

Figure 3

The distribution of other features in this dataset is shown below:

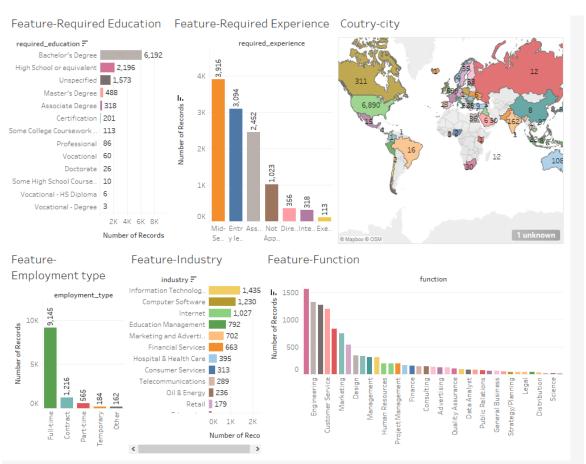


Figure 4

The analysis of fake job postings: Number fake job posting for different industries Number of Fake job advs for different employment types Analysis of Fake job posting industry \mp True Number of Fake job advs for different em. Oil & Energy Financial Services Number of Records 231 Marketing and Advertising 200 13 Real Estate 13 Information Technology a.. 100 10 70 Number fake job posting for different industries Full-time Contract Part-time Other Number fake job posting for different functions Number of Fake job Posting for different countries Analysis of Fake job posting function True Analysis of Fake job posting / country_n. Engineering 75 True 46 Customer Service Number of Records Other 200 Information Technology 19 18 Sales 100 Business Development Administrative 10 United 70 80 Number fake job posting for different functions Number of Fake job posting for Number of Fake job posting Number of fake job different Levels without having Question posting for having question or not. Analysis of Fake job posting / required_experie Analysis of Fake job posting / requir. True True Analysis of Fake job posting. 105 39 ber of Reco. Numberof 40 100 69 of Reco. 150 50 17 10 3 100 Mid-S Entry Asso.. Direc.. Not A Exec.. Inter enior.. level pplic. False

Figure 5

Based on figure 5 scammers are targeting mostly full-time jobs. And the United States has the highest number of fake job postings. And we can compare the other features for different outputs.

References

[1] U. o. t. Aegean, "Employment Scam Aegean Dataset," [Online]. Available: http://emscad.samos.aegean.gr/.