Job Postings

Characteristics of real vs. fake job listings

Sujay Chebbi | Alexandre Nicolai | Patricia Schutter | Kaiwen Zhang

Agenda

- Project Goals
- Exploratory Analysis
- Analysis
- Solution & Insights

Project Goals

Description

Based on job postings, try to determine the features that make a job posting fraudulent or non-fraudulent.

Importance

- Optimize efficiency
- Protect against job scams

Exploratory Analysis

Dataset

17,880 job postings with 18 features each

- Job ID
- Job title
- Location
- Department
- Salary Range
- Company Profile

- Description
- Requirements
- Benefits
- Telecommuting
- Company Logo
- Questions

- Employment Type
- Required Experience
- Required Education
- Industry
- Function
- Fraudulent

Dataset Challenges

Categorical Data

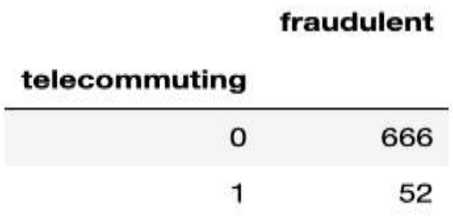
- All 18 columns were categorical
- 4 columns were made up of sentences

Missing Data

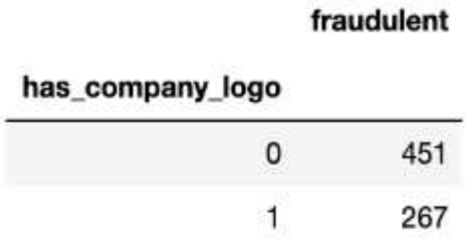
- Dataset included 70,103 missing values across all rows and columns

Analysis

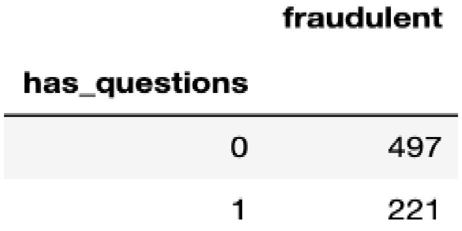
Important Dummy Variables



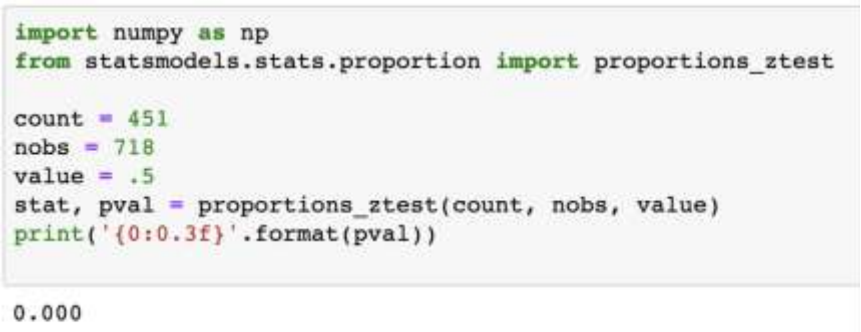
	fraudulent
telecommuting	
0	666
1	52



	fraudulent
has_company_logo	
0	451
1	267



	fraudulent
has_questions	
0	497
1	221



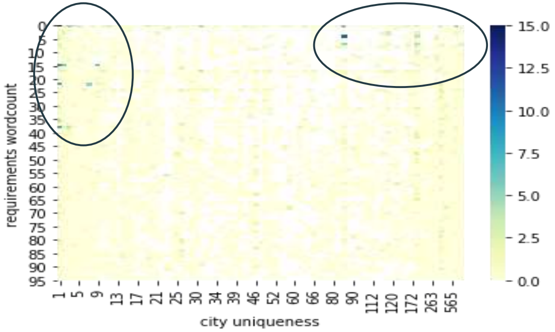
```
import numpy as np
from statsmodels.stats.proportion import proportions_ztest

count = 451
nobs = 718
value = .5
stat, pval = proportions_ztest(count, nobs, value)
print('{0:0.3f}'.format(pval))
```

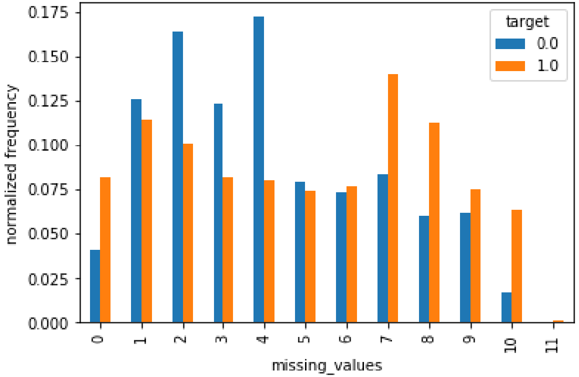
```
0.000
```

- Statistically significant proportion z-test
  - P-values effectively 0
  - Pattern between fraudulence and these characteristics
- Assumptions:
  - Random sample
  - Proportions normally distributed

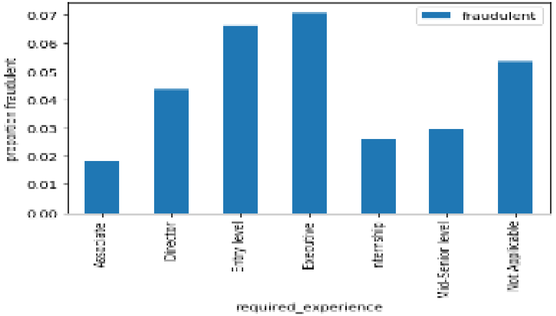
Preference for Extremes



Bogus/misspelled/unusual cities vs. big-market cities

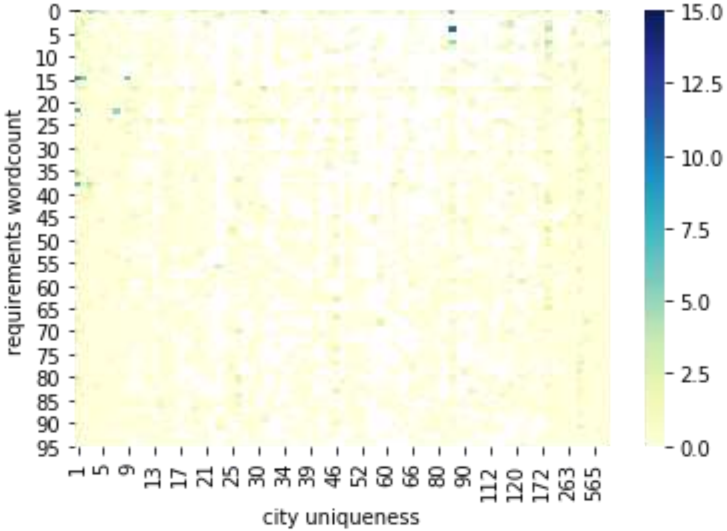


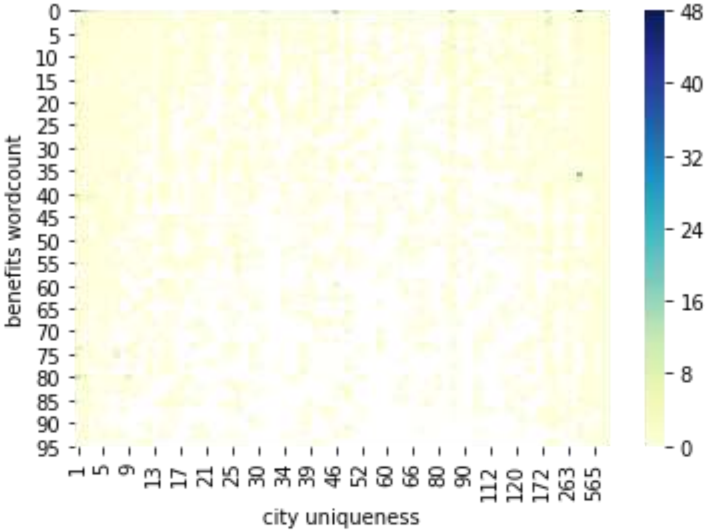
“Be super detailed” vs. “don’t try at all”



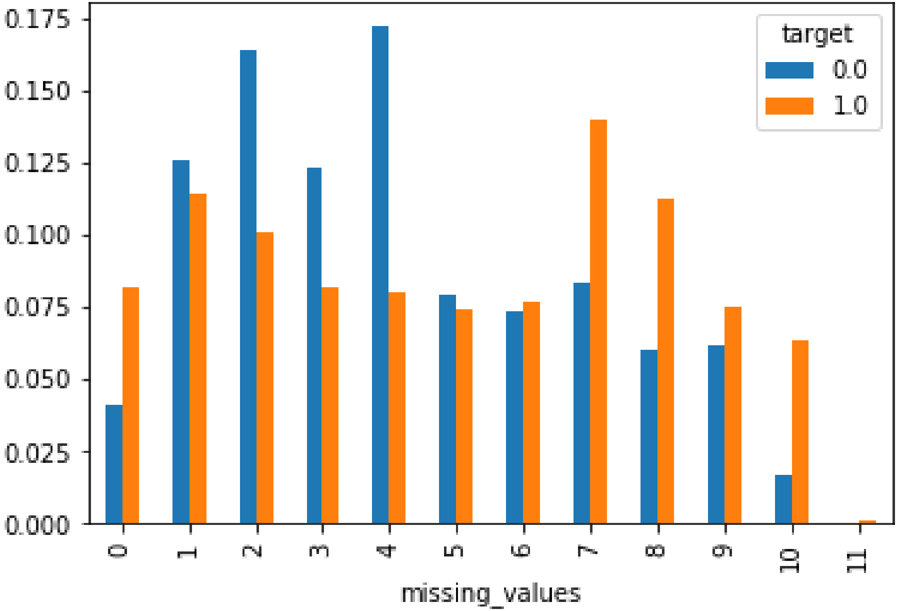
“Too good to be true” vs. “too good to be true”

“Laziness” and word counts are good predictors

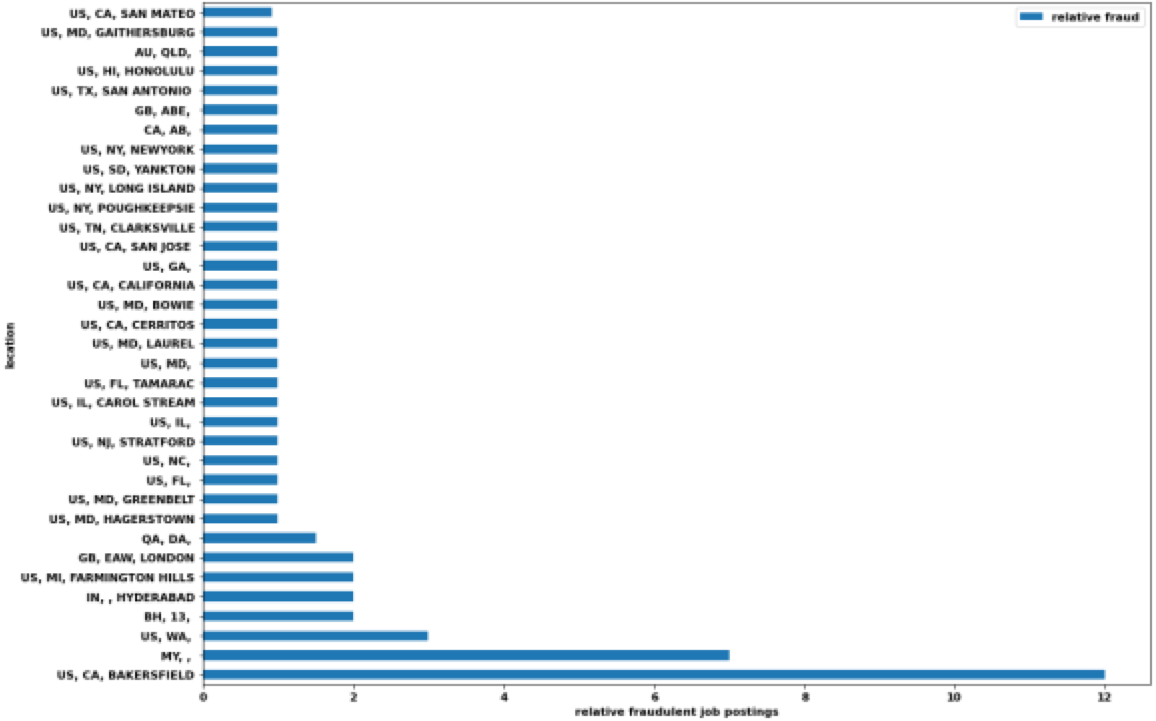




Missing Values

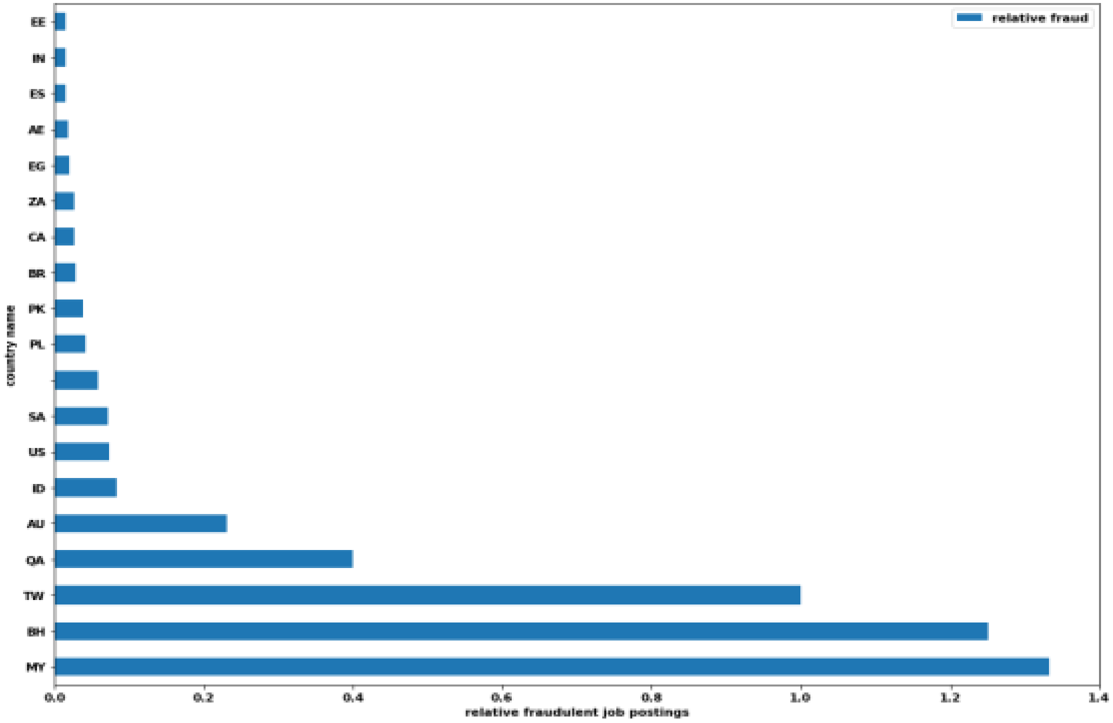


Determining relative fraudulent job postings by location

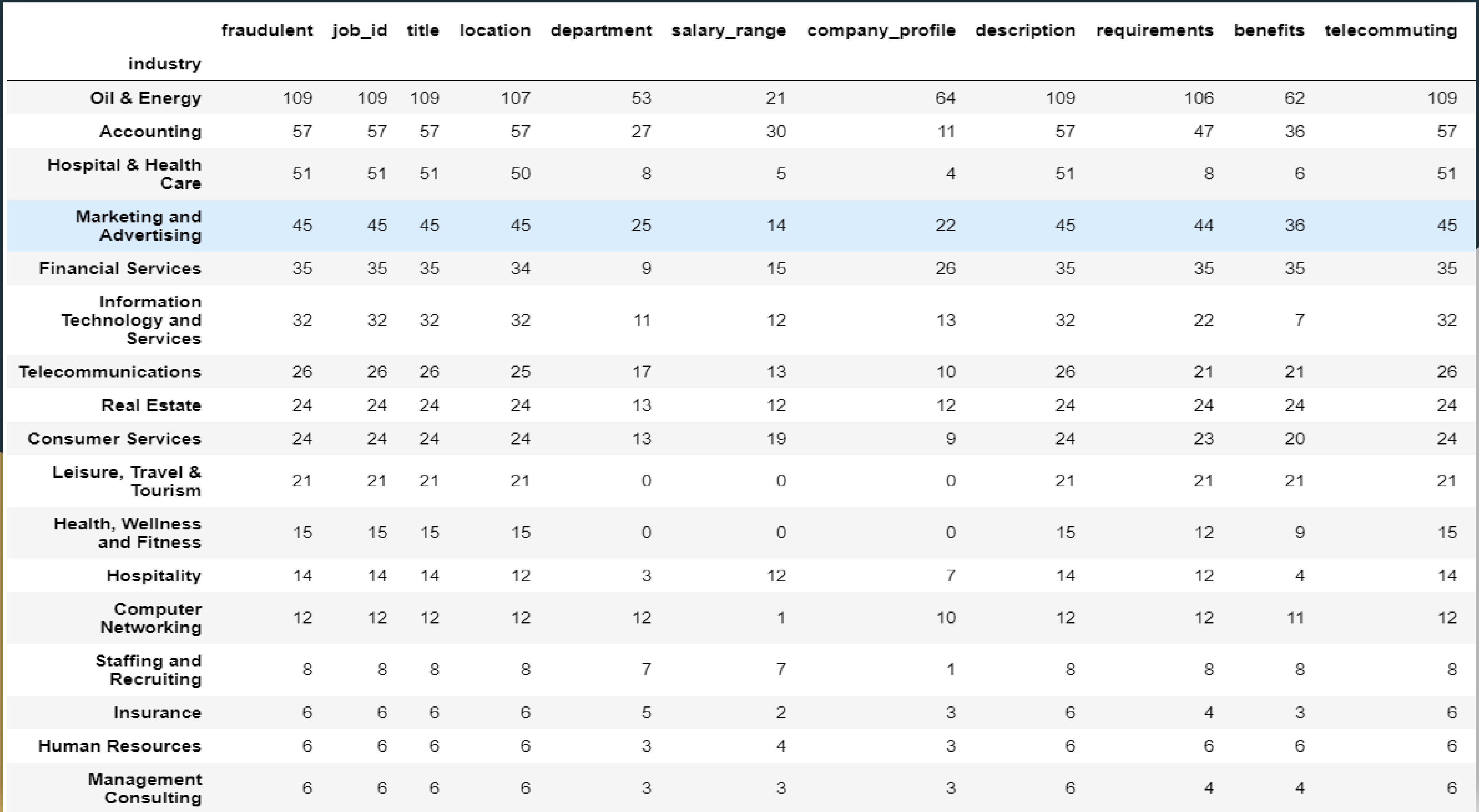


A list of cities with relative fraudulent job postings greater than or equal to 0.9, meaning that these cities are more likely to have fraudulent job postings

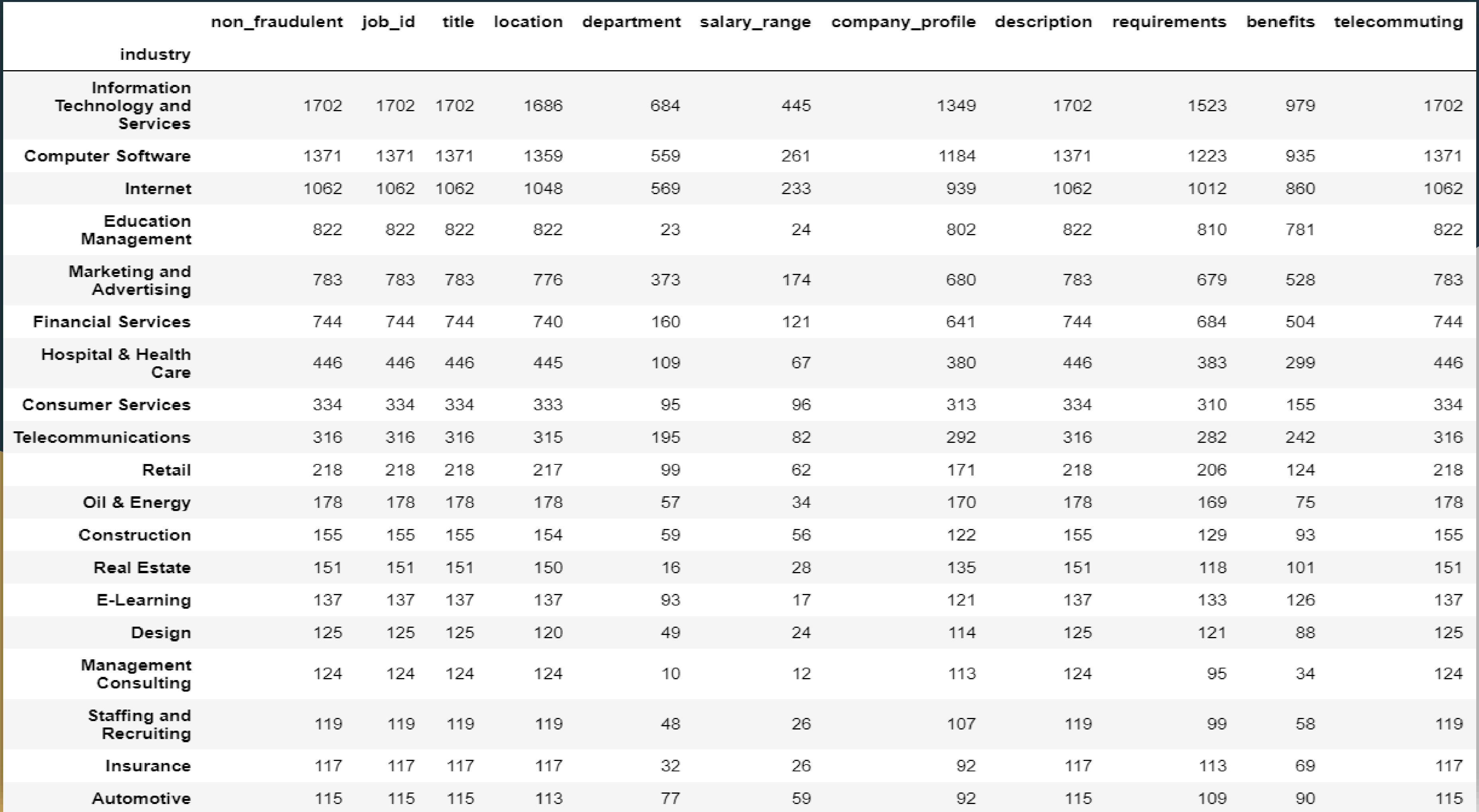
Determining relative fraudulent job postings by country



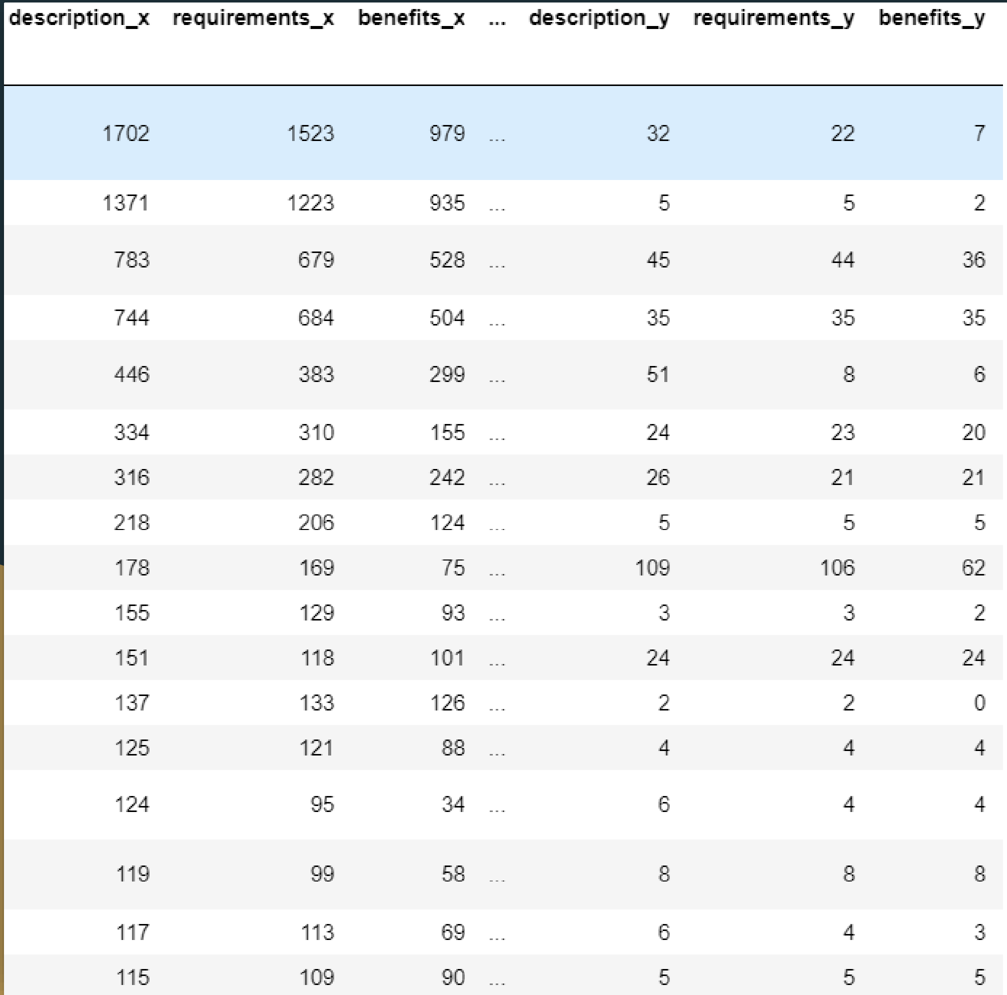
A list of countries with relative fraudulent job postings greater than or equal to 0.01, meaning that these countries are more likely to have fraudulent job postings



	fraudulent	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting
industry											
Oil & Energy	109	109	109	107	53	21	64	109	106	62	109
Accounting	57	57	57	57	27	30	11	57	47	36	57
Hospital & Health Care	51	51	51	50	8	5	4	51	8	6	51
Marketing and Advertising	45	45	45	45	25	14	22	45	44	36	45
Financial Services	35	35	35	34	9	15	26	35	35	35	35
Information Technology and Services	32	32	32	32	11	12	13	32	22	7	32
Telecommunications	26	26	26	25	17	13	10	26	21	21	26
Real Estate	24	24	24	24	13	12	12	24	24	24	24
Consumer Services	24	24	24	24	13	19	9	24	23	20	24
Leisure, Travel & Tourism	21	21	21	21	0	0	0	21	21	21	21
Health, Wellness and Fitness	15	15	15	15	0	0	0	15	12	9	15
Hospitality	14	14	14	12	3	12	7	14	12	4	14
Computer Networking	12	12	12	12	12	1	10	12	12	11	12
Staffing and Recruiting	8	8	8	8	7	7	1	8	8	8	8
Insurance	6	6	6	6	5	2	3	6	4	3	6
Human Resources	6	6	6	6	3	4	3	6	6	6	6
Management Consulting	6	6	6	6	3	3	3	6	4	4	6



industry	non_fraudulent	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting
Information Technology and Services	1702	1702	1702	1686	684	445	1349	1702	1523	979	1702
Computer Software	1371	1371	1371	1359	559	261	1184	1371	1223	935	1371
Internet	1062	1062	1062	1048	569	233	939	1062	1012	860	1062
Education Management	822	822	822	822	23	24	802	822	810	781	822
Marketing and Advertising	783	783	783	776	373	174	680	783	679	528	783
Financial Services	744	744	744	740	160	121	641	744	684	504	744
Hospital & Health Care	446	446	446	445	109	67	380	446	383	299	446
Consumer Services	334	334	334	333	95	96	313	334	310	155	334
Telecommunications	316	316	316	315	195	82	292	316	282	242	316
Retail	218	218	218	217	99	62	171	218	206	124	218
Oil & Energy	178	178	178	178	57	34	170	178	169	75	178
Construction	155	155	155	154	59	56	122	155	129	93	155
Real Estate	151	151	151	150	16	28	135	151	118	101	151
E-Learning	137	137	137	137	93	17	121	137	133	126	137
Design	125	125	125	120	49	24	114	125	121	88	125
Management Consulting	124	124	124	124	10	12	113	124	95	34	124
Staffing and Recruiting	119	119	119	119	48	26	107	119	99	58	119
Insurance	117	117	117	117	32	26	92	117	113	69	117
Automotive	115	115	115	113	77	59	92	115	109	90	115

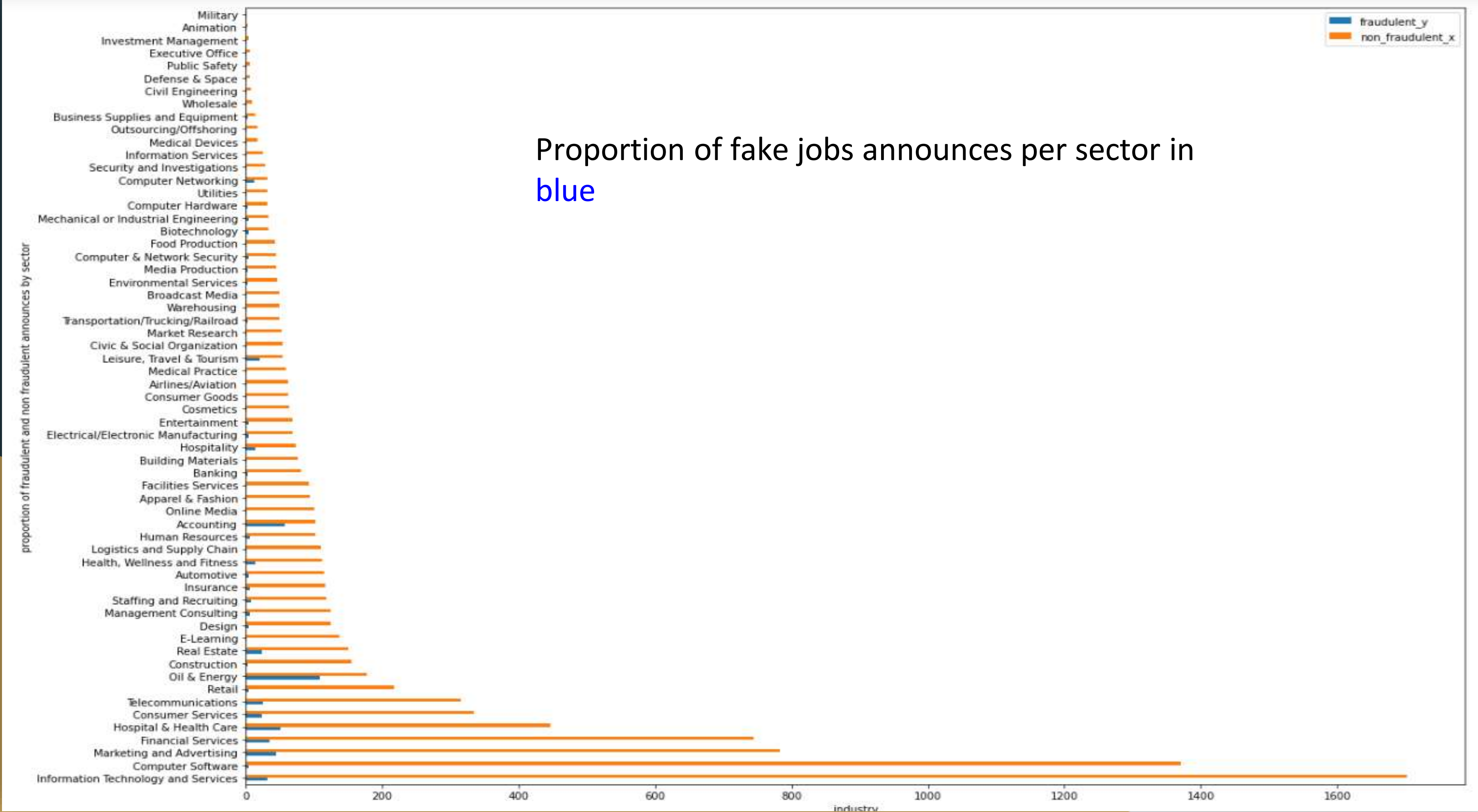


description_x	requirements_x	benefits_x	...	description_y	requirements_y	benefits_y
1702	1523	979	...	32	22	7
1371	1223	935	...	5	5	2
783	679	528	...	45	44	36
744	684	504	...	35	35	35
446	383	299	...	51	8	6
334	310	155	...	24	23	20
316	282	242	...	26	21	21
218	206	124	...	5	5	5
178	169	75	...	109	106	62
155	129	93	...	3	3	2
151	118	101	...	24	24	24
137	133	126	...	2	2	0
125	121	88	...	4	4	4
124	95	34	...	6	4	4
119	99	58	...	8	8	8
117	113	69	...	6	4	3
115	109	90	...	5	5	5

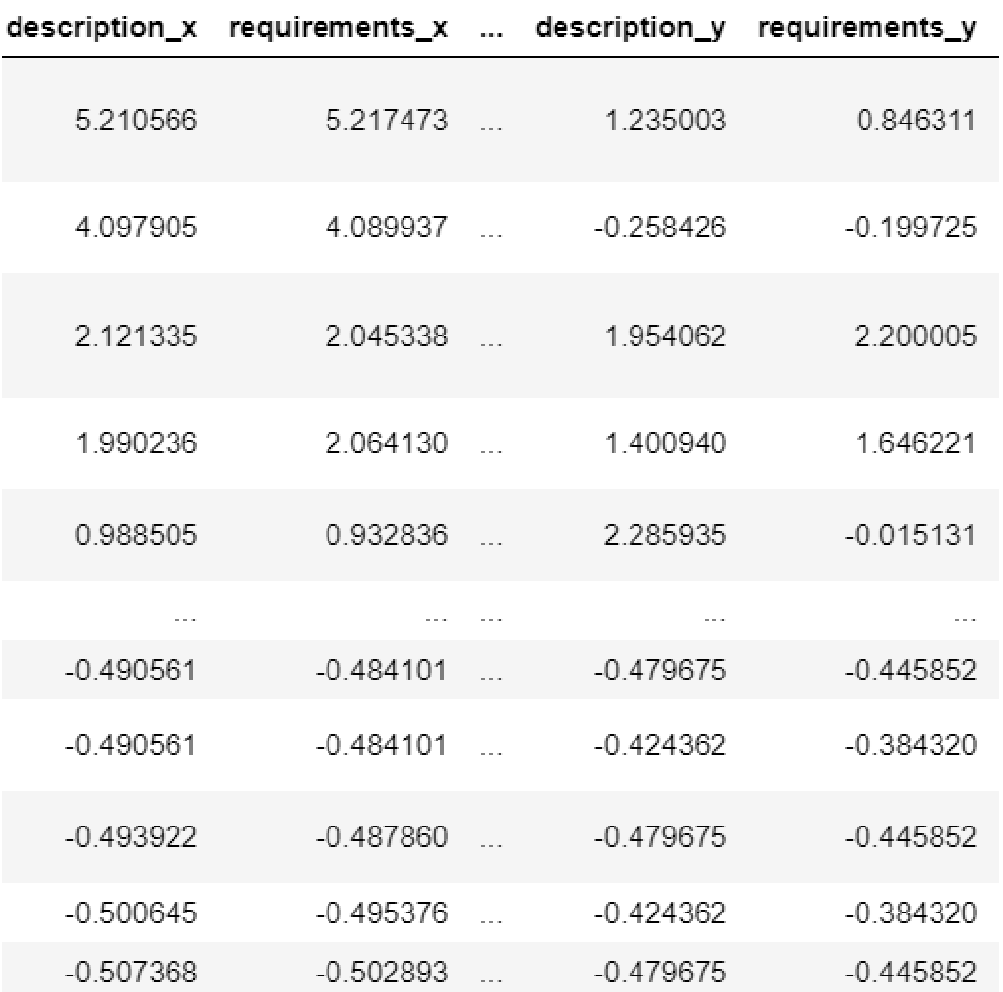
Result merging the two previous tables on their index (the sector).

X = Real job announces

Y = Fake Job announces

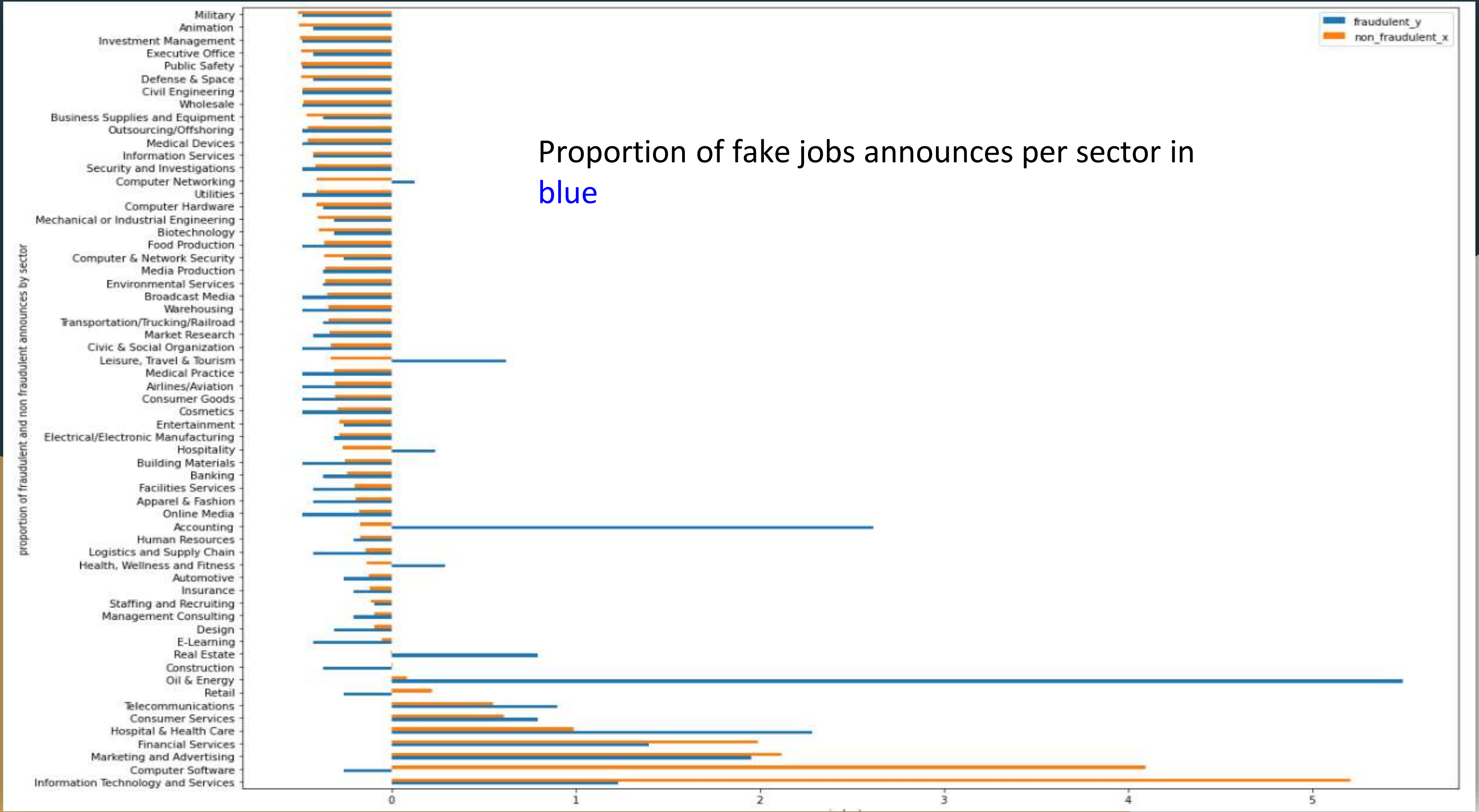


Proportion of fake jobs announces per sector in blue

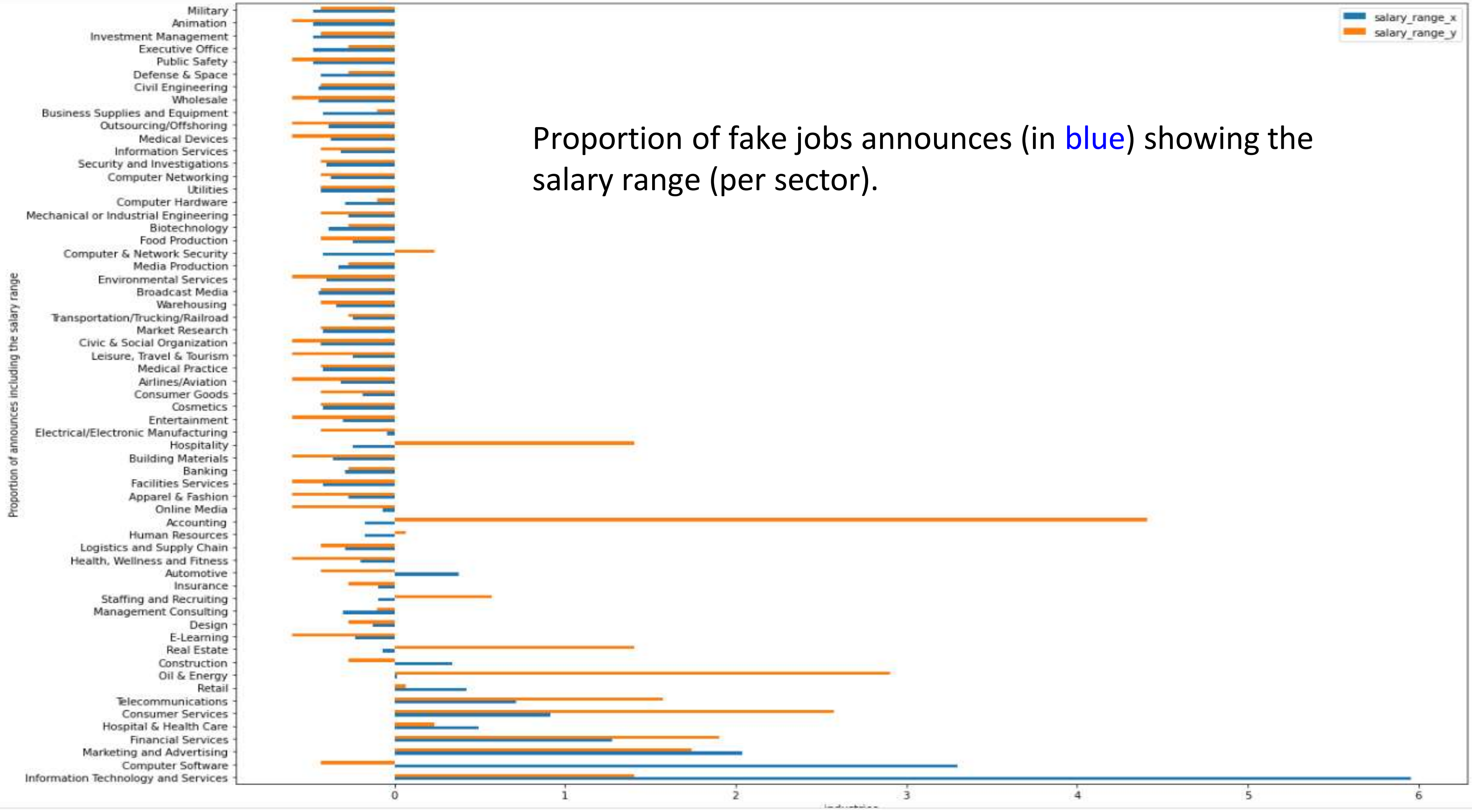


description_x	requirements_x	...	description_y	requirements_y
5.210566	5.217473	...	1.235003	0.846311
4.097905	4.089937	...	-0.258426	-0.199725
2.121335	2.045338	...	1.954062	2.200005
1.990236	2.064130	...	1.400940	1.646221
0.988505	0.932836	...	2.285935	-0.015131
...	...	...	...	...
-0.490561	-0.484101	...	-0.479675	-0.445852
-0.490561	-0.484101	...	-0.424362	-0.384320
-0.493922	-0.487860	...	-0.479675	-0.445852
-0.500645	-0.495376	...	-0.424362	-0.384320
-0.507368	-0.502893	...	-0.479675	-0.445852

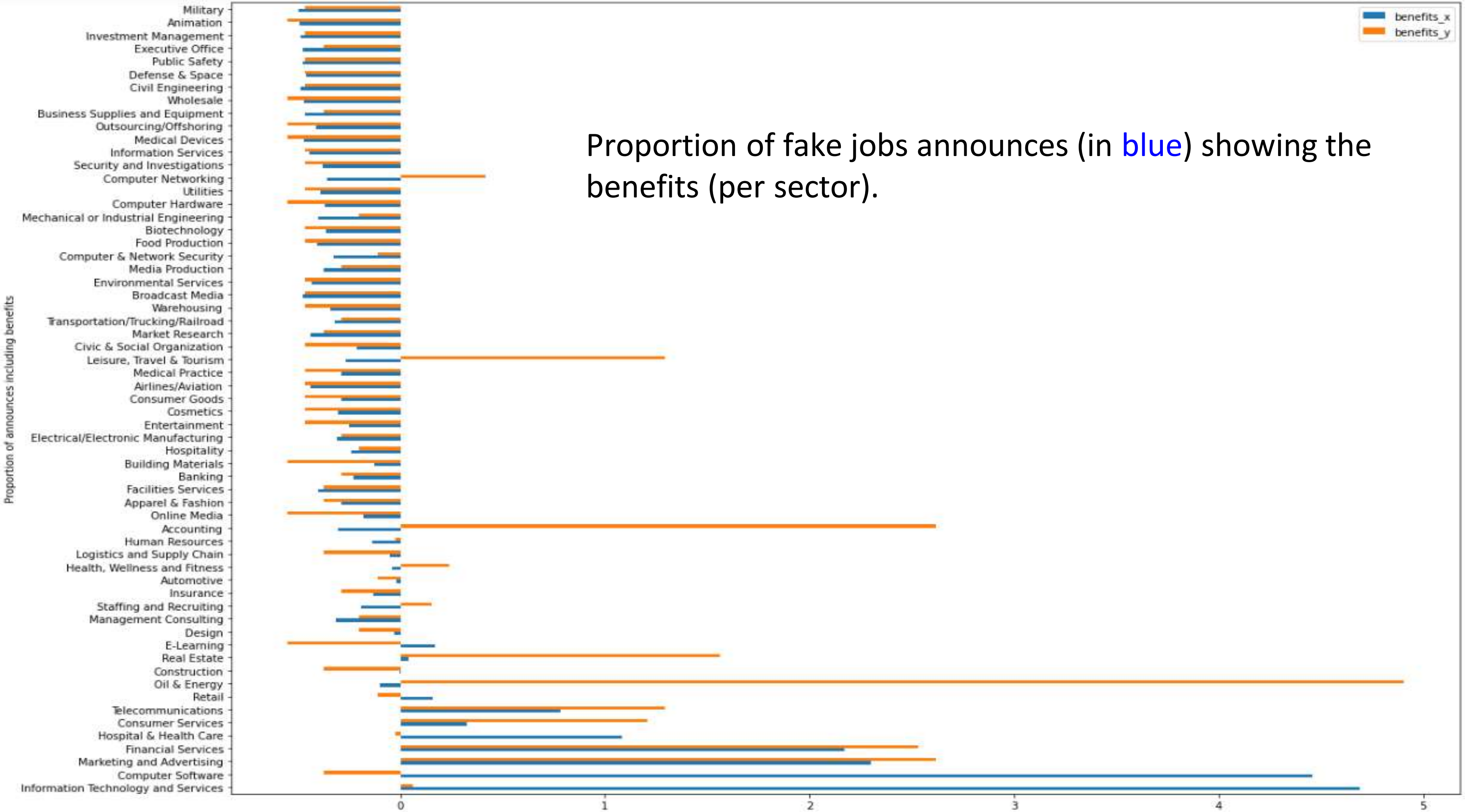
Table normalized using the Z-Score method.



Proportion of fake jobs announces per sector in blue



Proportion of fake jobs announces (in blue) showing the salary range (per sector).



Proportion of fake jobs announces (in blue) showing the benefits (per sector).

Solution
& Insights

Logistic Regression

Features of a legit posting:

- Industry
- Function
- Location
- Salary Range
- Department
- Employment type
- Required Experience
- Company Logo

FEATURE	IMPACT
Industry [Education Management]	-2.453213
Industry [Internet]	-2.186165
Function [Health Care Provider]	-2.02904
Location [Greece]	-1.956881
Salary Range [55,000-75,000]	-1.781989
Industry [Computer Software]	-1.661547
Industry [Restaurants]	-1.599323
Required Experience [Associate]	-1.581123
Location [Germany]	-1.489198
Industry [Insurance]	-1.44227
Department [Operations]	-1.427709
Location [Philippines]	-1.403489
Employment Type [Temporary]	-1.39878
Department [Oil and Gas]	-1.343899
Department [Legal]	-1.310619
Department [Marketing]	-1.302265
Required Experience [Executive]	-1.243289
Department [Department]	-1.208611
Has company Logo [True]	-1.205376
Salary Range [0-0]	-1.081002

Logistic Regression

Features of a fraudulent posting:

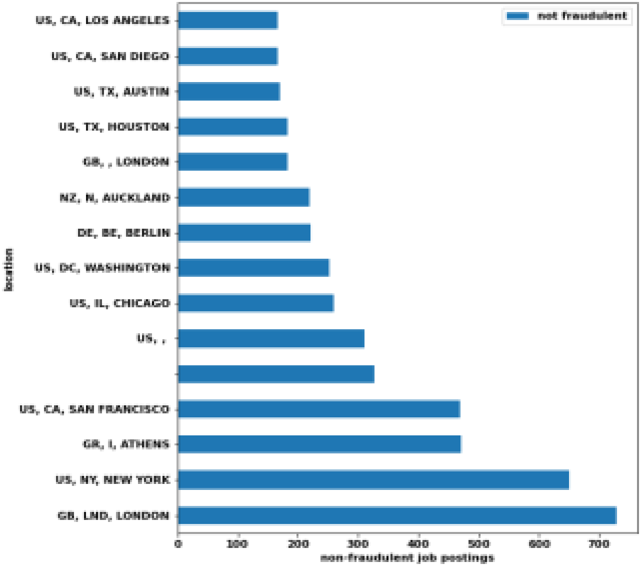
- Department
- Industry
- Location
- Salary Range
- Title

FEATURE	IMPACT
Department [Oil & Energy]	3.441561
Department [Information Technology]	2.89915
Industry [Oil & Energy]	2.886244
Department [Engineering]	2.723301
Location [Malaysia]	2.557432
Location [Australia]	2.239522
Department [Call Center]	2.218432
Department [Accounting/Payroll]	2.20586
Salary Range [7200-1380000]	2.190508
Industry [Leisure, Travel, & Tourism]	2.147251
Industry [Computer Networking]	2.102746
Salary Range [28000-32000]	1.922421
Department [Clerical]	1.914747
Department [CSR]	1.863665
Department [Biotech]	1.695124
Department [Power Plant & Energy]	1.660967
Industry [Hospitality]	1.645928
Title [12]	1.644899
Department [Engineering]	1.631634
Industry [Accounting]	1.557629

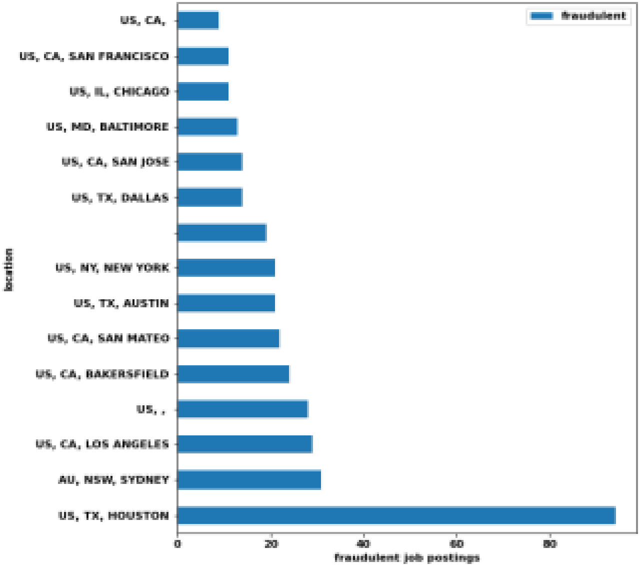
Thank You

Appendix

Determining fraudulent and legitimate job postings by location

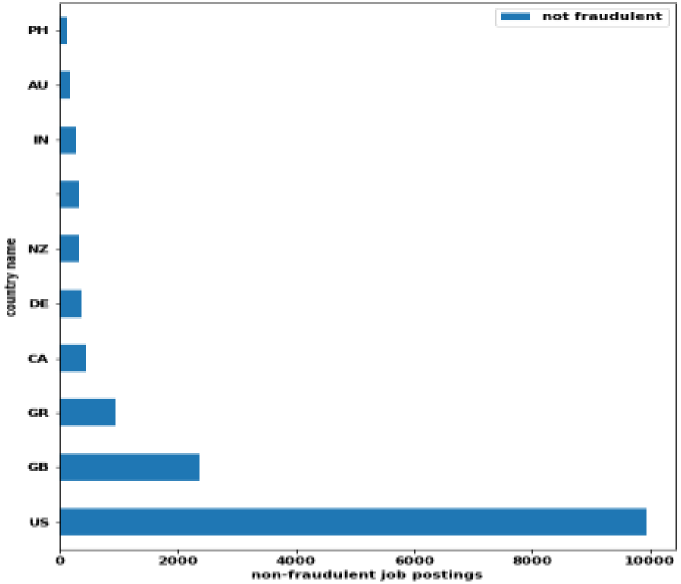


Top 15 cities with most legitimate job postings

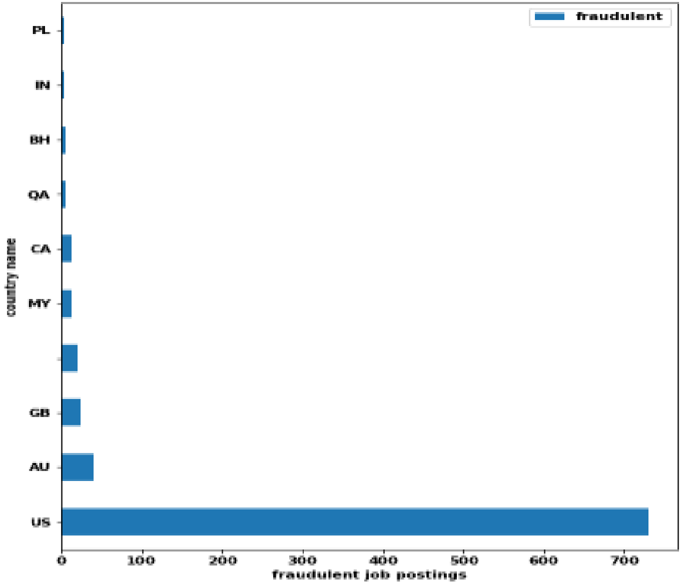


Top 15 cities with most fraudulent job postings

Determining fraudulent and legitimate job postings by country



Top 10 countries with most legitimate job postings



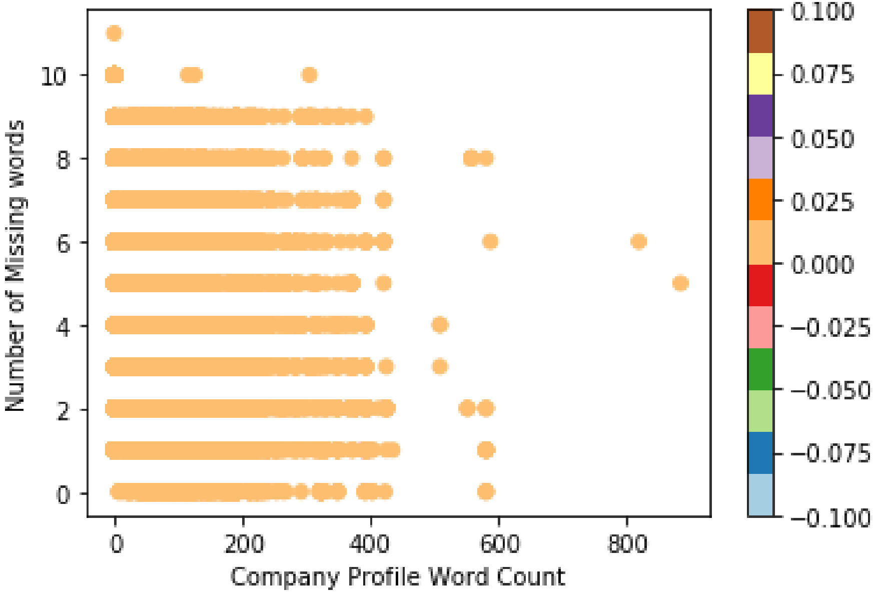
Top 10 countries with most fraudulent job postings

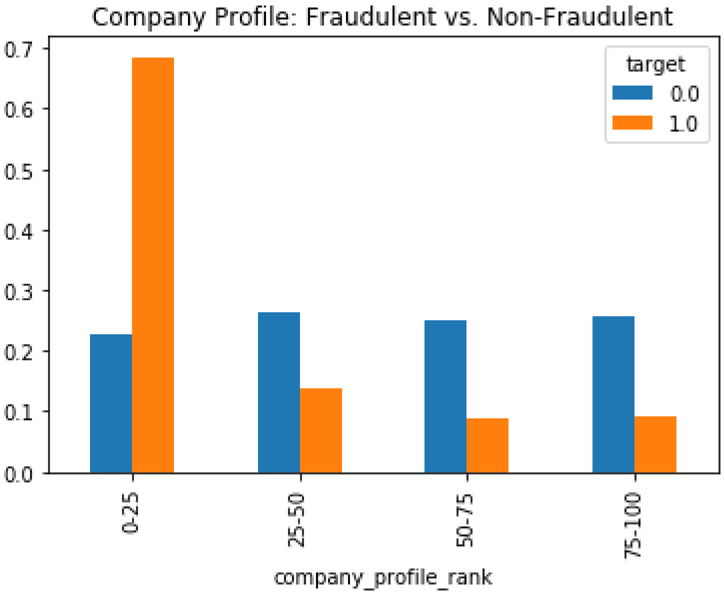
Missing Values By Feature

Feature	Amount
Title	0
Location	0
Department	11547
Salary Range	15012
Company Profile	3353
Description	13
Requirements	7326
Benefits	7326
Telecommuting	0

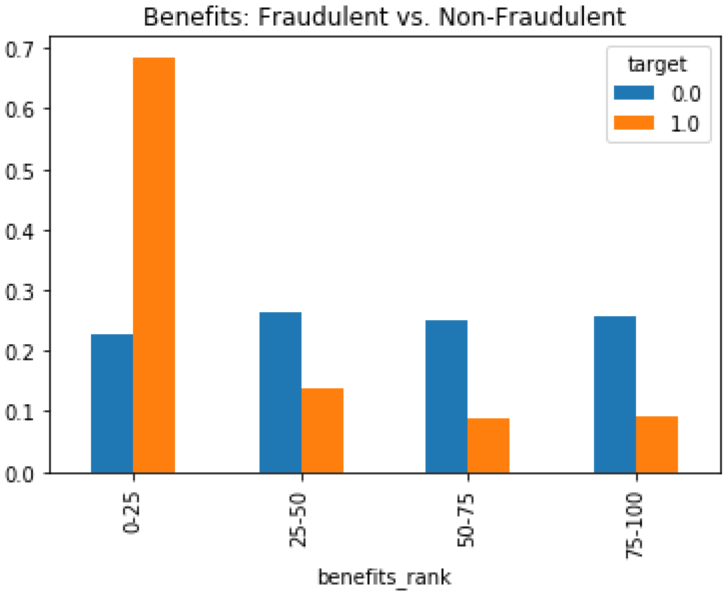
Feature	Amount
Company Logo	0
Questions	0
Employment Type	3471
Required Experience	7050
Required Education	8105
Industry	4903
Function	6455
Fraudulent	0

KNN


Number of Missing words
0
2
4
6
8
10
0
200
400
600
800
Company Profile Word Count
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100



Company Profile: Fraudulent vs. Non-Fraudulent
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
target
0.0
1.0
0-25
25-50
50-75
75-100
company_profile_rank



Benefits: Fraudulent vs. Non-Fraudulent
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
target
0.0
1.0
0-25
25-50
50-75
75-100
benefits_rank