Characteristics of Real Job Postings Versus Fake Job Postings

Description

The dataset of interest includes 17,880 different job postings and 18 columns/characteristics including 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, and fraudulent. Of the 17,880 job postings, 17,014 are real job postings and 866 are fraudulent job postings. We look for ways to predict fraudulence based on the characteristics of job postings.

Importance

The applications for investigating this dataset involve an improvement of quality for both job sites and job seekers. The analytics ascertained from investigating trends could help job sites develop an algorithm for ranking the quality of job postings. The addition of such an algorithm would give a job site a competitive advantage in the market due to their increased ability to match users with a real job. The addition of this feature would also serve as a form of "public good" as it would optimize the efficiency in terms of the benefits and costs of the online job seeking process, while helping users avoid sending their personal information and work history to fraudulent jobs.

Exploratory Analysis

This dataset comes with a couple of challenges. All 18 columns in the dataset are categorical, with company profile, description, requirements, and benefits given in sentence form. We believed we would be doing a disservice if we disregarded these columns. As we could not compare these four elements between job postings, we chose to get the word count for each of these features and split them among quartiles to more effectively compare them.

Another challenge with this dataset was that it contained 70,103 missing values across all rows and columns. We again believed that we could not disregard these missing values, but rather believed that they were a part of the story in uncovering a fake job posting from a real job posting. To account for this missing data, the number of missing values in each posting was counted up and added to an additional column.

Location seemed to play a real factor in determining if a job posting was fraudulent or not. This was a basic attempt at gaining some insight into

which cities/countries had legitimate or fraudulent job postings. Our first attempt at showing which cities had the most fraudulent and legitimate job postings are seen in **Photos 1 and 2**. As can be seen from these horizontal bar charts, the cities with the most legitimate job postings include London, New York City, Athens, and San Francisco. Cities with the most fraudulent job postings include Houston, Sydney, Los Angeles, job postings with no city specified, and Bakersfield.

We then looked to identify job postings on a more broad scale by looking at which countries had the most legitimate and fraudulent postings. This is seen in the horizontal bar charts in **Photos 3 and 4**. Countries with the most legitimate job postings include the US, Great Britain, Greece, Canada, and Germany. Companies with the most fraudulent job postings include the US, Australia, Great Britain, job postings without a country specified, and Malaysia.

However, even with total fraudulent and legitimate job posting data, it would be more indicative of actual fraudulent postings if we determined relative fraudulent postings. That is, the number of fraudulent postings for every legitimate posting. Larger cities and countries would be bound to have larger volumes of fraudulent postings due to the sheer number of job postings available. We applied a filter for relative fraud in job postings to be greater than 0.9 for cities, and greater than 0.01 for countries. For cities, this can be seen in **Photo 5**, and for countries, this can be seen in **Photo 6**. The cities with the most relative fraudulent postings include Bakersfield, Malaysia with no cities specified, and Washington state with no cities specified. The countries with the most relative fraudulent job postings included Malaysia, Bahrain, Taiwan, Qatar, and Australia.

One pattern that appeared repeatedly throughout our analysis and visualizations was the predominance of "extreme" values. For example, the two required experience levels with the highest normalized frequency of fraudulence were entry level and executive. Likewise, in a pivot-table analysis of fraudulence in the written length of the benefits versus the uniqueness of the city location, the highest frequencies of fraudulence occurred at locations that are very unique in the dataset (reflecting small markets or misspellings) versus that of very big markets (**Photos 7 and 8**). This suggests that there is an advantage for those who post fraudulent job postings to adopt binary strategies such as low effort or high effort, and low bar of entry or high expectations. Knowing that fraudulent job postings often aim to steal personal information gleaned from unsuspecting applicants, this makes sense because a halfway approach may require too much effort and yet is not sufficiently believable either, or too high-expectations to attract applicants of high quantity but not high enough to attract those of high quality.

Clearly, effort was found to be related to fraudulence. One pattern found that later proved helpful for the formal analysis of the problem was the importance of "zero effort" or little effort on predicting the fraudulence of a job posting. From heat maps and bar graphs of word counts of requirements, benefits, and company profile, fraudulence counts drop to insignificant values beyond the 25th percentile of word count. Within the 25th percentile, there is an even higher predominance of fraudulence at the 0th percentile (i.e. word count of 0), suggesting that a quantized measure of word count could be a good predictor to utilize. Additionally, the importance of fraudulence at the 0th percentile prompted this question: If zero effort (not even bothering to write a company profile, job requirements, etc.) can signal a fraudulent job posting, perhaps a measure of the total number of columns/attributes for a job posting left blank could be a good signal as well. In that spirit, we also investigated the viability of "missing_values" as a predictor in our analysis.

From our intuition on the importance of the missing values in the dataset, we began our analysis by creating a plot of normalized missing values per row against whether they are fraudulent or not (**Photo 9**). From **Photo 9**, it can be seen that job postings with 6 or more missing values tend to be fraudulent, while job postings with between 2 and 5 missing values tend to be legitimate job postings. However, a job posting with 0 missing values had a higher chance of being fraudulent as well.

We also tried a different approach regarding the missing values. We know how important they can be to highlight a fake job posting if any. Therefore, we first thought: "What is the most important feature when looking for a job using a tool search?". Since the user generally first input the sector he is interested in, we got rid of the missing value for this feature after separating the fraudulent and non fraudulent announcements. **(Photo 10)**

From these two tables, we could already identify patterns and make hypotheses on the sectors that concentrate the most fraudulent job posting. It appeared that Oil & Energy, Accounting, Hospital & Health Care are the most exposed but to have a more precise idea of the proportion it represents we merged those two tables on their index **(Photo 11)**.

We first plotted this table **(Photo 12)** to have a general idea of the proportion but then for interpretability purposes we decided to normalize the data by subtracting the mean and dividing by the standard deviation each value. **(Photo 13)**

That way we can have a better idea of which sector is the most concerned by the fraudulent announcements. It confirmed our intuition that Oil &

Energy, accounting and Hospital and Health care are more subject to fraud but also showed that Real Estate and Leisure and Travel as well concentrated a high number of fake jobs posting. **(Photo 14)**

We used the merged table normalized to understand the importance of other features like the salary range posted in the announcement. If we focus on the sectors that are subject to fraud we understand that the real jobs announcements tend to show the salary range more often than the fake job announces. **(Photo 15)** It also appears that the real jobs announcements show more often the benefits associated with the position than the fake ones. **(Photo 16)**

Doing this, we got a better idea of the interaction of our features to identify a fake job announcement. They tend to be concentrated in a few sectors. So if you see a job posting in Oil & Energy, Accounting, Hospital and Health care, Real estate or Leisure and Travel without the salary rangy and/or without the benefits associated with the position, you can fairly think that it is a fake one.

Solution & Insights

To understand the dataset as a whole, Logistic Regression was used across all features of our dataset. From **Photo 17**, it can be seen that the features that are most likely to lead to a legitimate job posting are based on industry, function, location, salary range, department, employment type, required experience, and company logo. Out of all of these, a job posting with an industry as education management is the factor that will most likely lead to a legitimate job posting. However, it was interesting to note that a job posting that contains a company logo is among the top 20 factors that indicate that a job posting is legitimate.

On the other hand, the features that tend to increase the odds of a job posting being fraudulent can be seen in **Photo 18**, where again department, industry, location, salary range, and title were among the features that most influenced whether a job posting was fraudulent. Of these, the job postings with departments labeled as Oil & Energy or Information Technology were the top features that indicated that the job posting was fraudulent.

From our findings, job searchers should be concerned about the legitimacy of job postings with departments in oil and energy and information technology, in the industry of oil and energy, in the department of engineering, or located in Malaysia or Australia. Job searchers should feel comfortable applying to jobs with an industry labeled as education management or internet, function as health care provider, and located in Greece.

Appendix

Photo 1 - Top 15 cities with most legitimate job postings

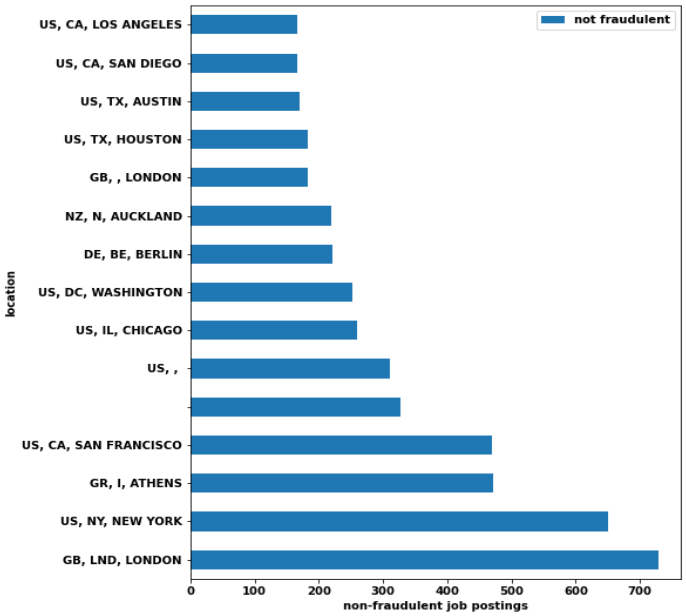

Photo 2 - Top 15 cities with most fraudulent job postings

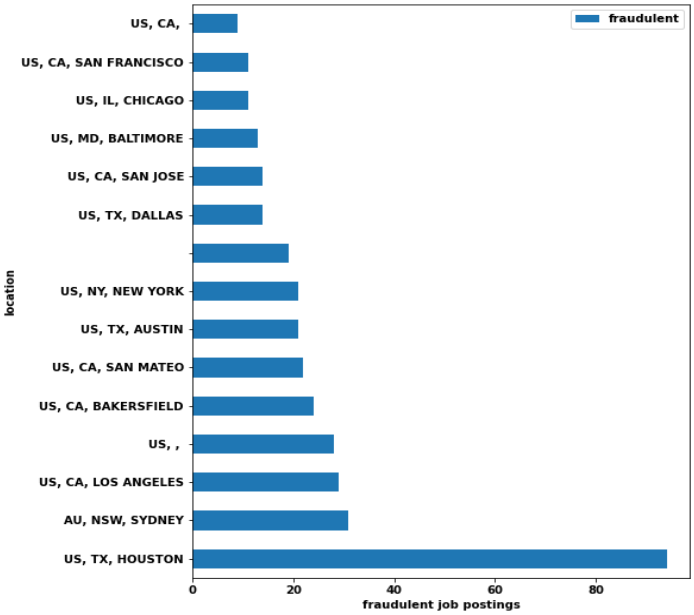

Photo 3 - Top 10 countries with most legitimate job postings

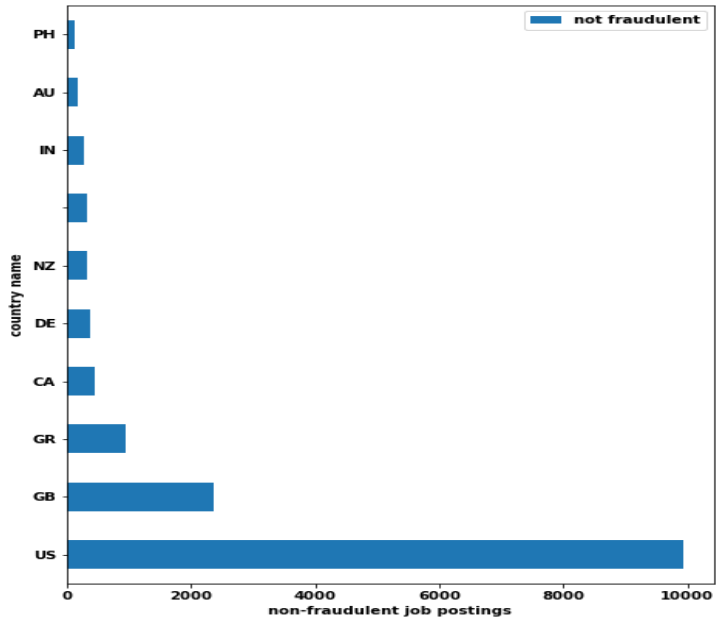

Photo 4 - Top 10 countries with most fraudulent job postings

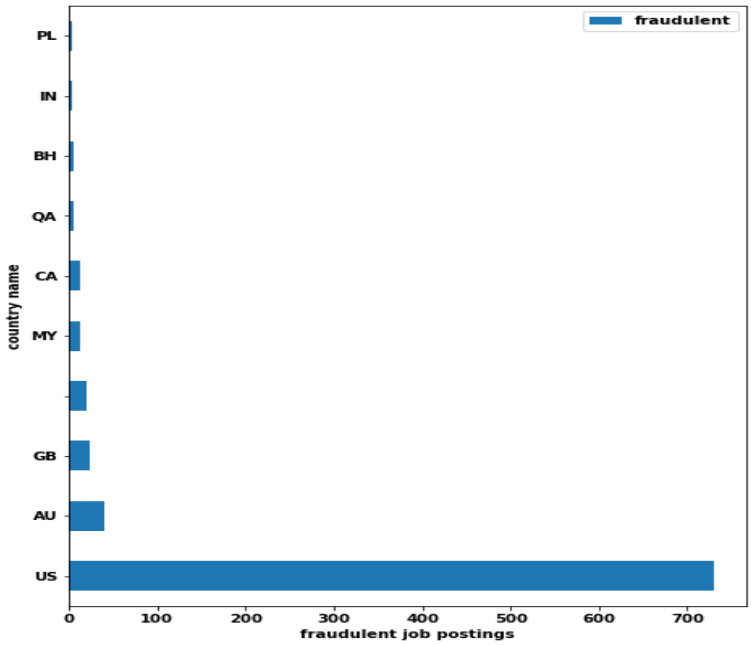

Photo 5 - A list of cities with relative fraudulent job postings greater than or equal to 0.9

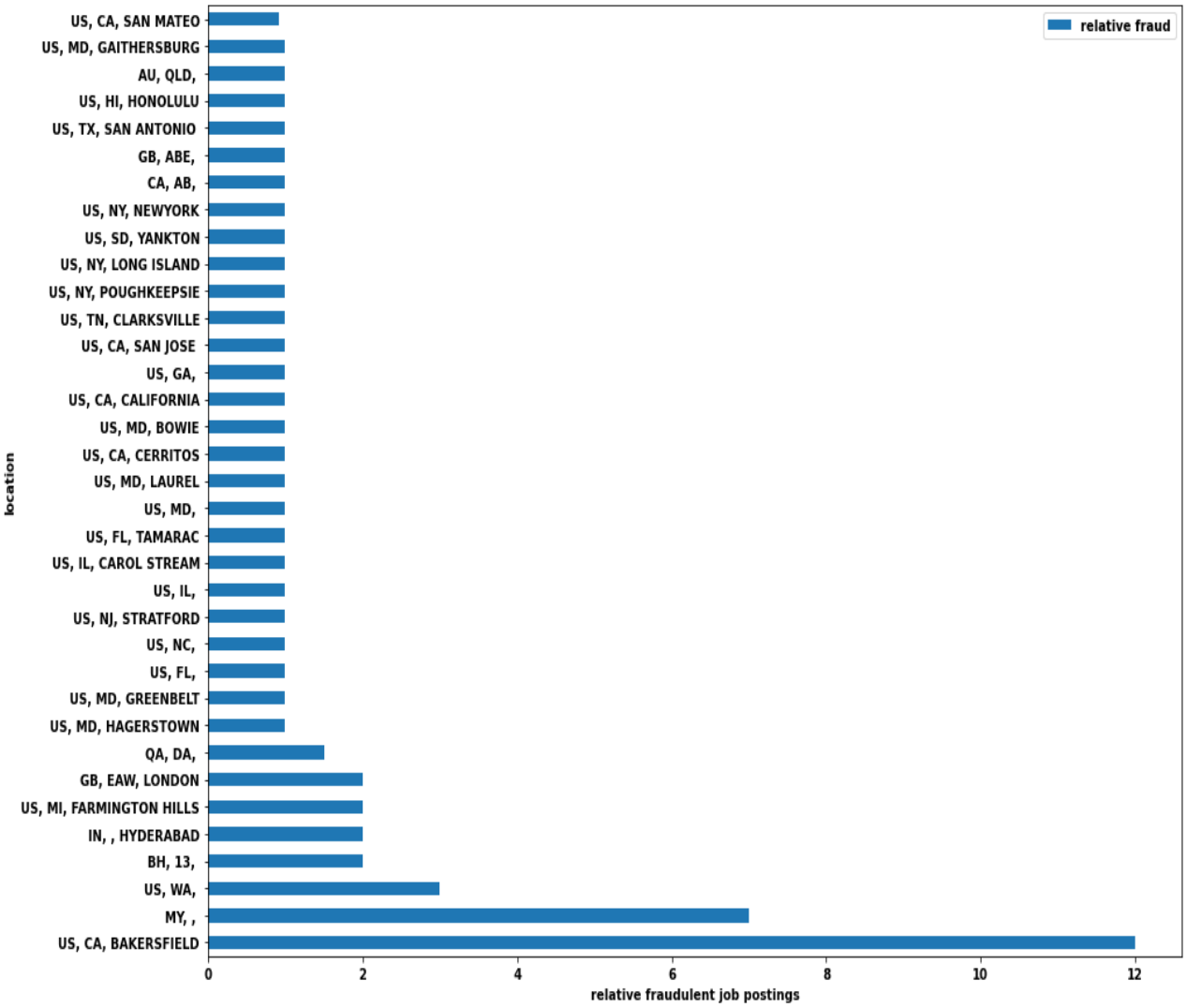

Photo 6 - A list of countries with relative fraudulent job postings greater than or equal to 0.01

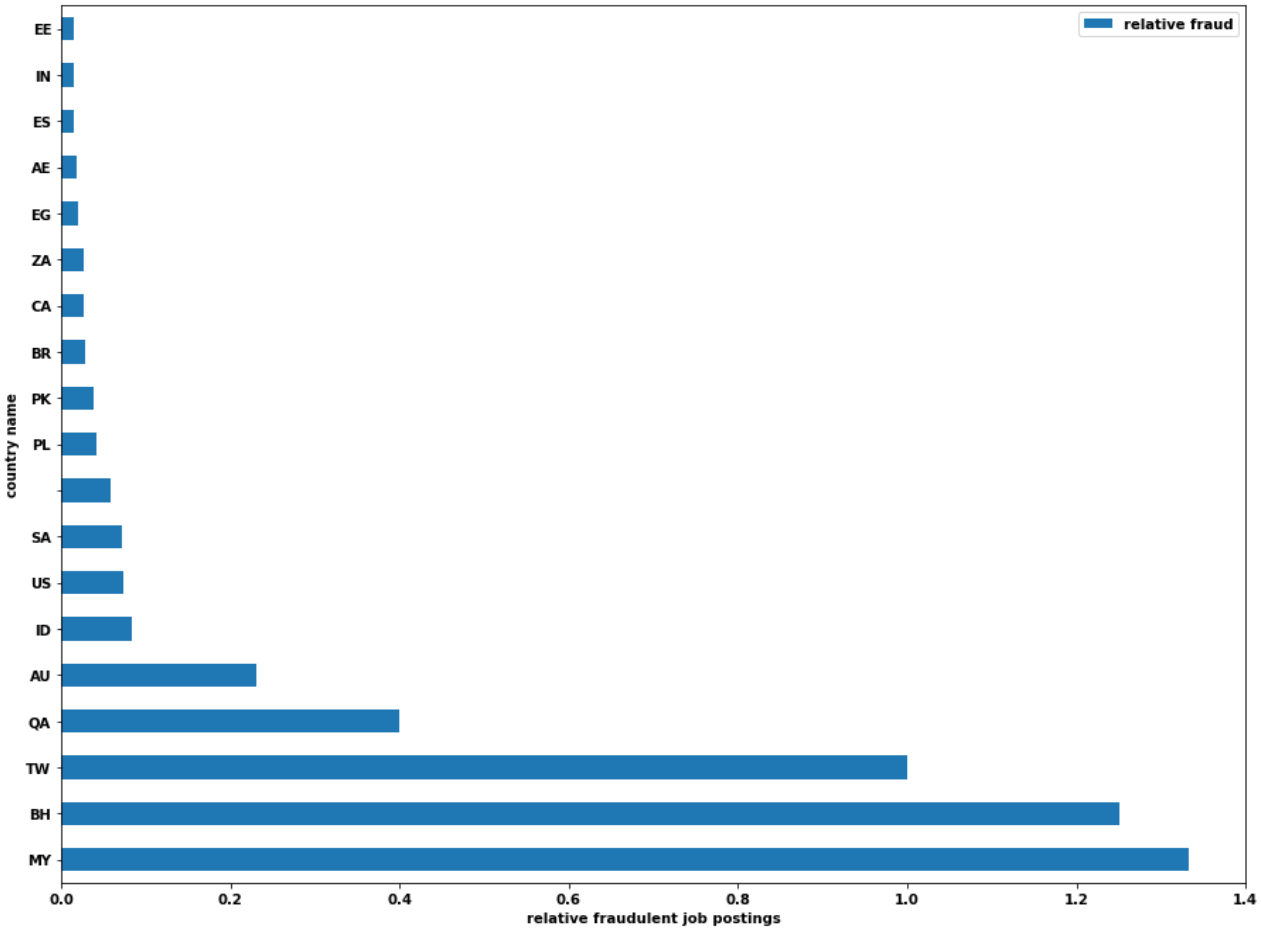

Photo 7 - Heat map of fraudulence counts at different levels of requirements word count vs. city uniqueness

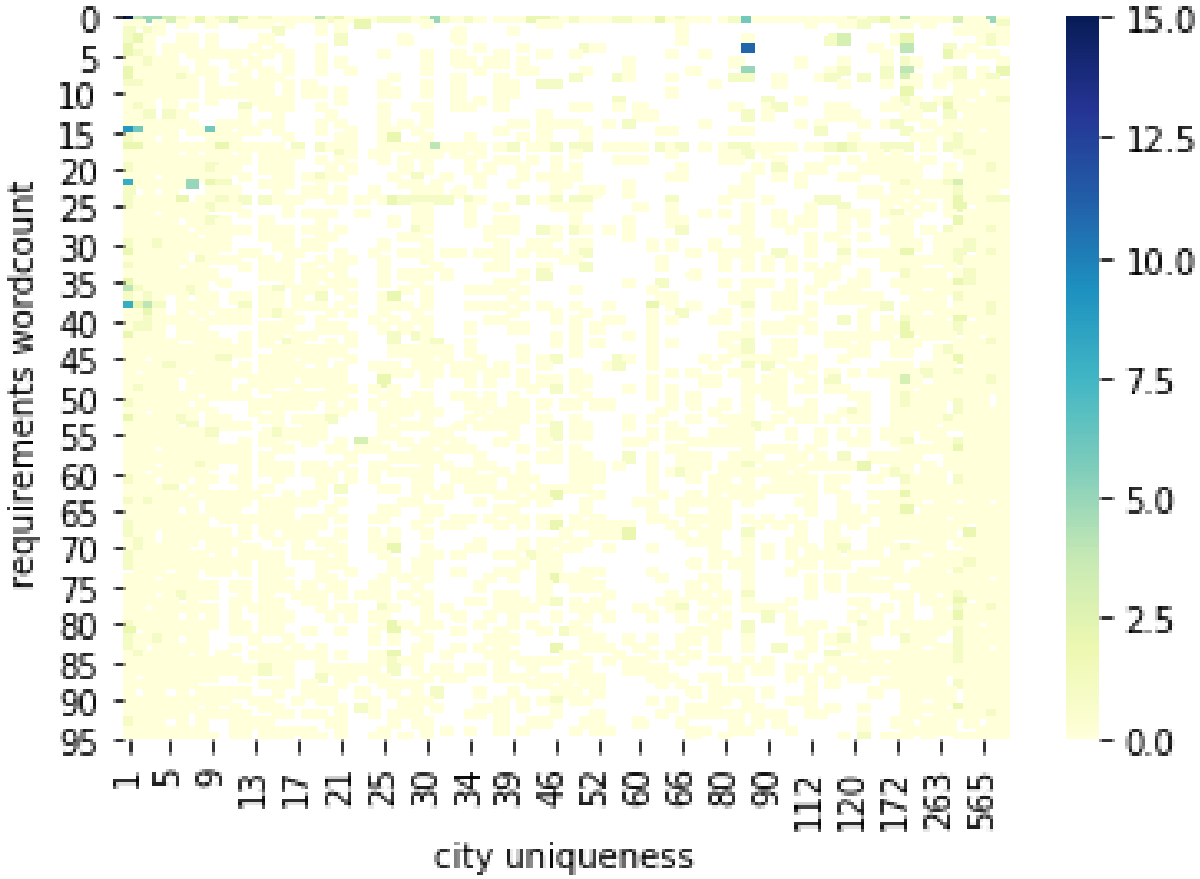

Photo 8 - Heat map of fraudulence counts at different levels of benefits word count vs. city uniqueness

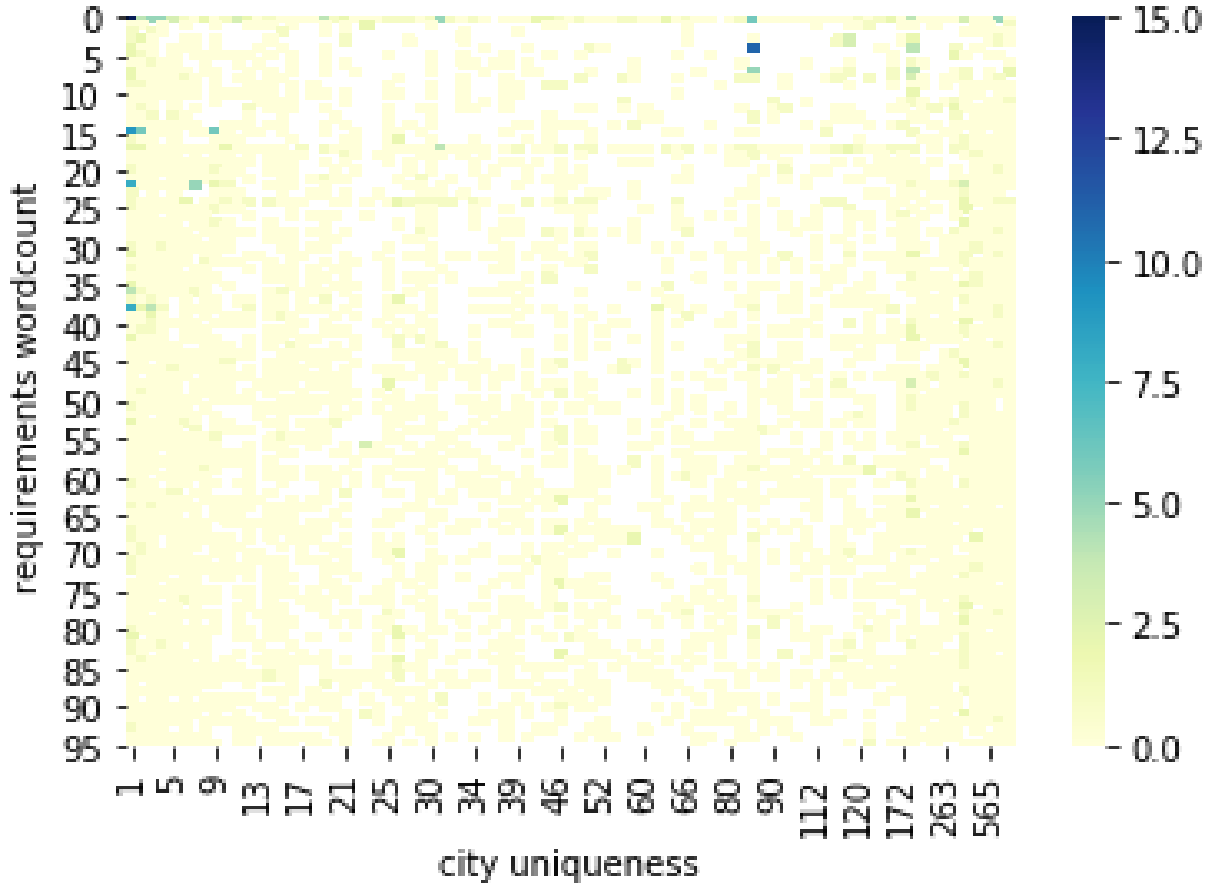

Photo 9 - Missing values per row

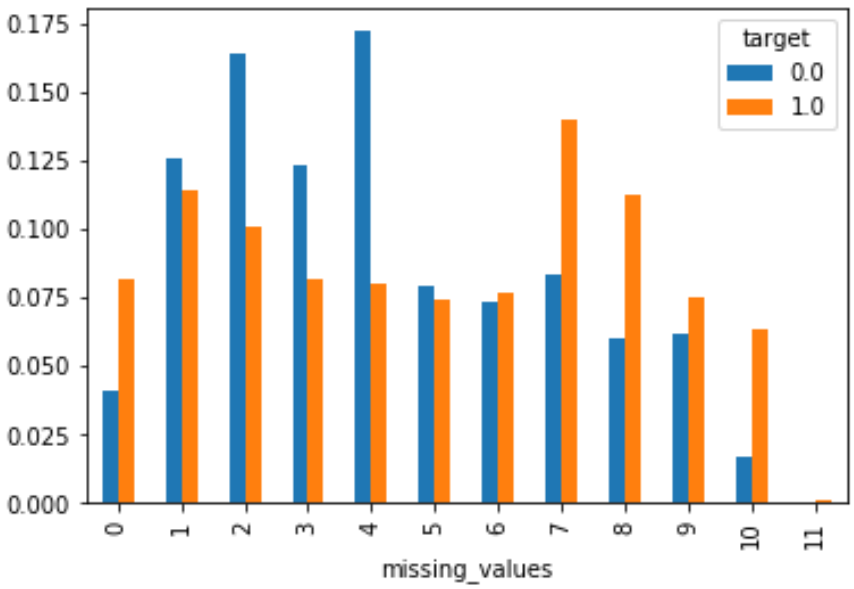

Photo 10 - Number of fraudulent and non fraudulent job postings per sector

a) Fraudulent

industry	fraudulent	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting
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

b) Non fraudulent

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

Photo 11 - Merged table on sector (fraudulent and non fraudulent)

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

Photo 12 - Proportion of fraudulent job posting per sector (non normalized data)

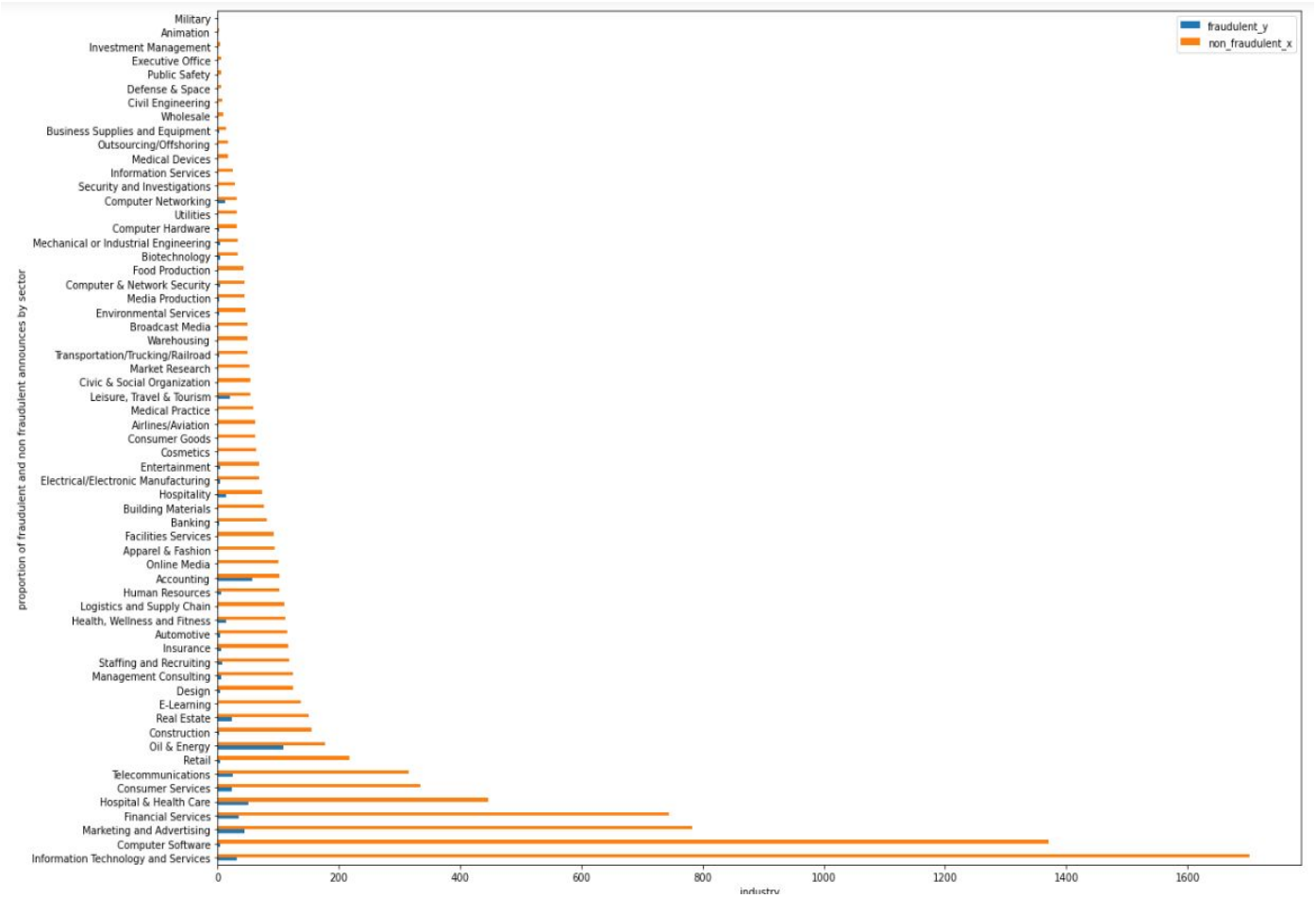

Photo 13 - Normalized data table

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

Photo 14 - Proportion of fraudulent job posting per sector (normalized data)

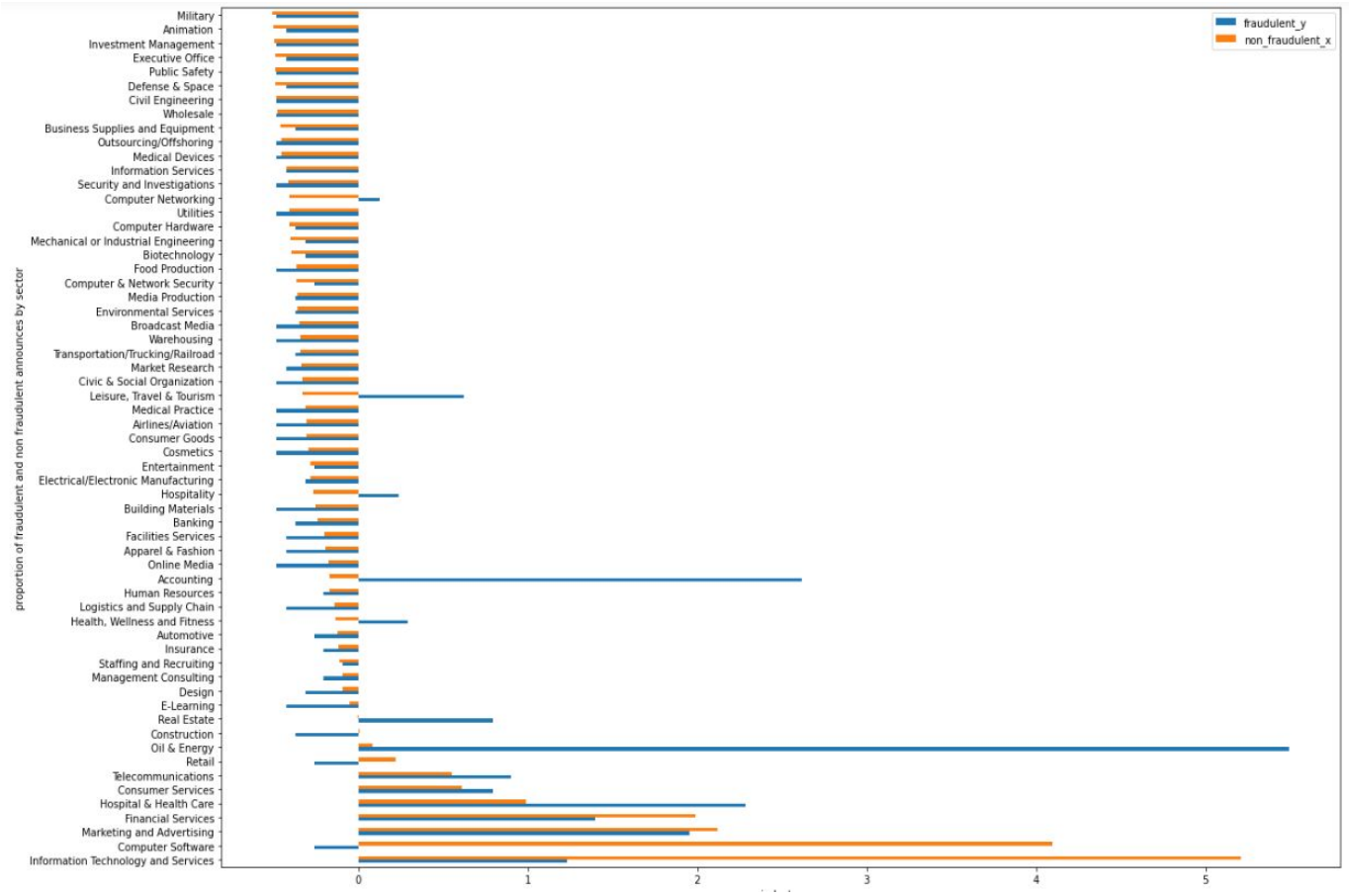

Photo 15 - Proportion of fake and real job posting announces including the salary range per sector

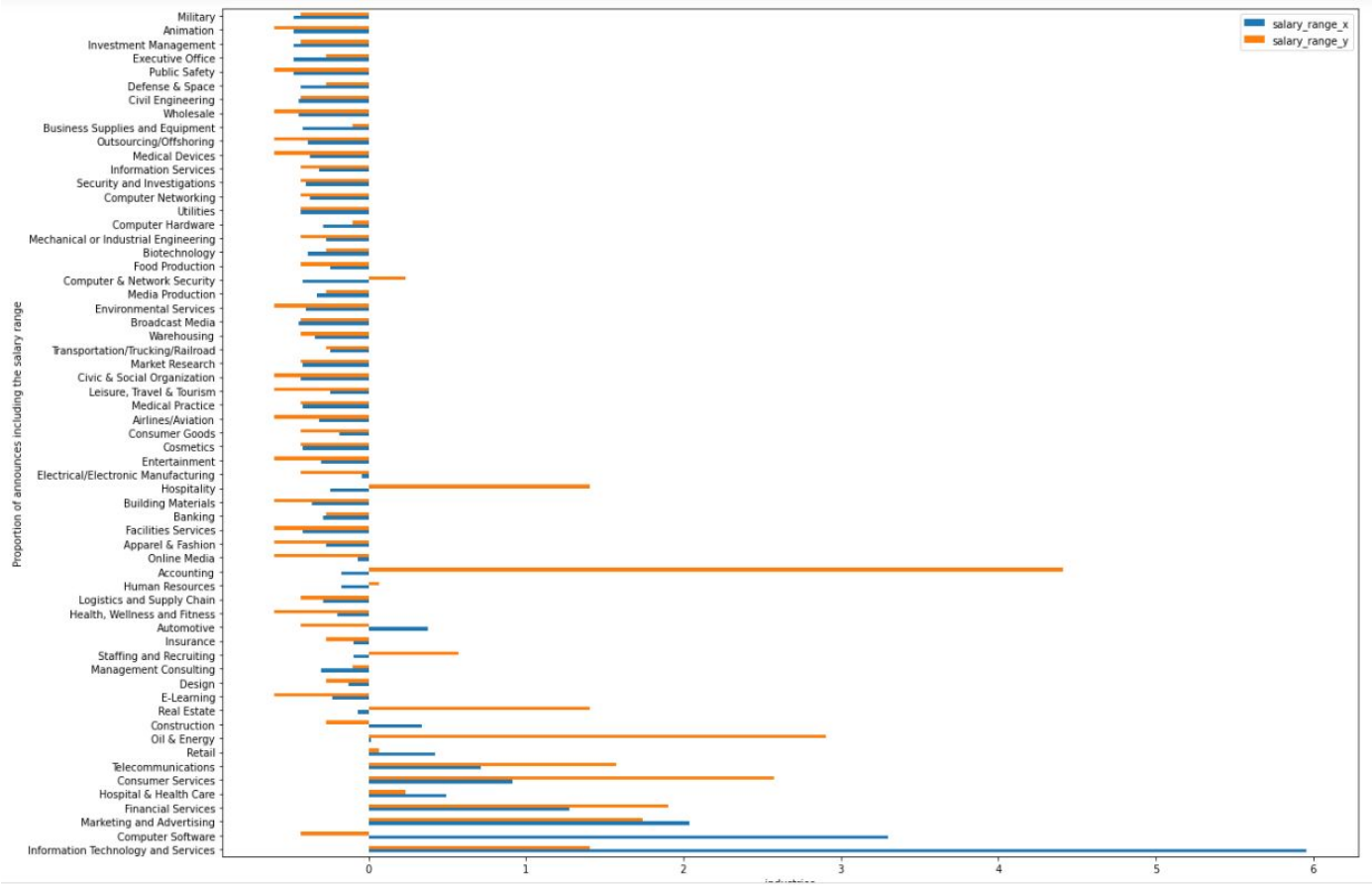

Photo 16 - Proportion of fake and real job posting announces including the benefits associated with the position, per sector

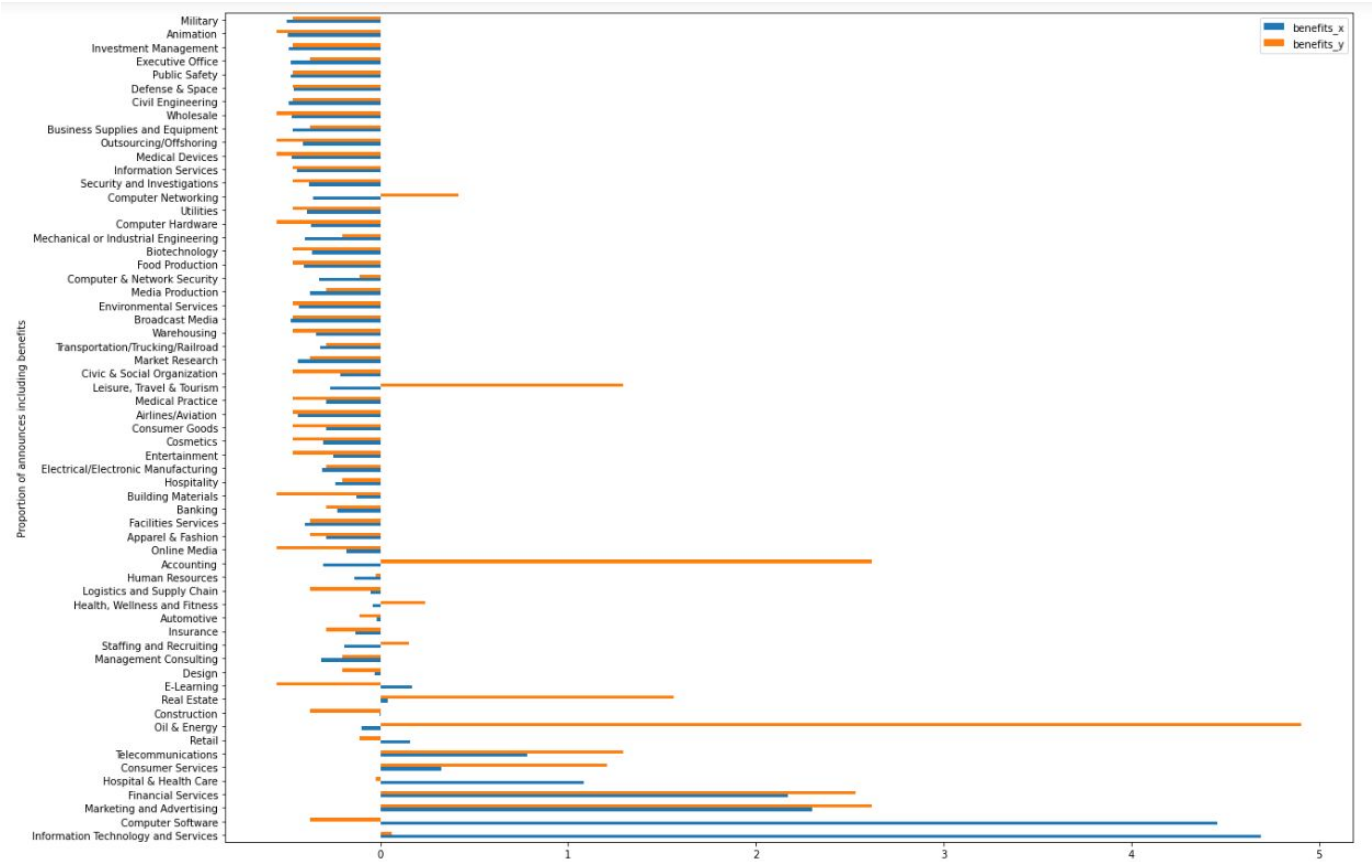

Photo 17 - Logistic Regression showing variables which contribute most to a legitimate job posting

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

Photo 18 - Logistic Regression showing variables which contribute most to a fraudulent job posting

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