**Busted: Analyzing Job Posting Patterns and Predicting Fraudulent Job**

**SI 618 Final Project**

**Jing Cao, School of Information, University of Michigan**

**Motivation**

This is a tough time for getting jobs. The last thing we want to see is we are so happy about an available position, but it turns out a fake one. In fact, it is common to encounter a fake job on the job seeking websites. For example, people complained about the fake jobs they’ve applied on Indeed ([link](https://www.indeed.com/community/interview-tips/what-percentage-of-jobs-posted-online-are-fake/td-p/2135682)). So, when I saw the dataset about fake job postings ([data](https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction)) on Kaggle.com. I decided to play around with it.

Based on the dataset, I proposed four questions as following.

1. Is there a relationship between fraudulent job and if the job is telecommuting, has logo or has screening questions?
2. What's the most possible job title for fake jobs?
3. What's the most words said in the description and benefit?
4. How can I predict whether a job posting is fraudulent or not?

**Data Source**

The [dataset](https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction) is found on Kaggle.com. There are 17880 records in total and 866 of them are fake jobs. The data includes both textual information like job title, location, department, company file, description and requirements (string) and other information like job\_id (integer), telecommuting (integer), has\_company\_logo (integer), has\_questions (integer), fraudulent (integer) about the jobs.

In the analysis, I’m focus on “title”(str), “description”(str), “benefits”(str), “telecommuting”(int), “has\_company\_logo”(int), “has\_questions”(int), and most important prediction y, “fraudulent”(int). For the analysis of ‘’title”, “telecommuting”, “has\_company\_logo”, “has\_questions”, and “benefits”, I will use all of the 17880 records. For the analysis of “description”, since there is only one missing value, I simply dropped that record.

**Methods**

Before I answered any of the questions, I looked into my dataset and checked the shape of it and null values in the variables I’m interested in.

**Question1:**

Since there is no missing value in those three variables, there is no need to perform further data cleaning. In order to get the answer, first, I needed to check the number of records(job postings) for real and fake jobs based on if the company is telecommuting, has a company logo, or has screening questions. Then I used t-test to check if there is a significant difference between real and fake jobs based on those factors.

**Question2:**

The values under “title” is short text and pretty clean, and I wanted to leave the title be whatever it should be instead of tokenizing it, so I simply joined all the values, generated a word cloud using the library “wordcloud” and visualized the word cloud using matplotlib.

At first, I tried to tokenize the title strings and remove the stopwords, but I realized that it is better to keep the title as a whole because even though some of the titles look similar or contain the same word, the job responsibilities are totally different, such as “Data Entry” and “Data Scientist”.

**Question3:**

Since “description” and “benefits” are both sentences, I cleaned the strings(lowering the characters, removing character like “&amp”, etc.), tokenized them and removed the stopwords. ). In order to compare if there is any difference between the common words of fake and real jobs, I separated the data set to real\_job and fake\_job. And then used Counter() from collections library to see the results.

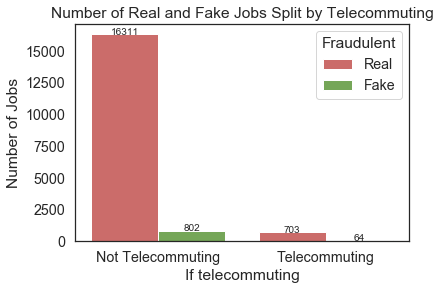
The challenge I encountered in this question is how to handle the null values. For the “description”, there is only one null value, so I simply dropped that record (even though the record is a fake job). For the “benefits”, null values take a big portion of the data. So, I checked the number of null values for both real job and fake job first, and found out both real jobs and fake jobs have about 40% of null values, no significant difference. Therefore, I kept the null values as a token.

**Question4:**

I wanted to predict whether a job is fraudulent or not based on the above variables. So, I created a dataset with dummies for textual variables, then separated the data to training and testing datasets (test\_size = 0.3). I run the random forest classification to generate a model, found the most important features. I also tried two kinds of Naïve Bayes algorithms to generate the model.

**Analysis and Results**

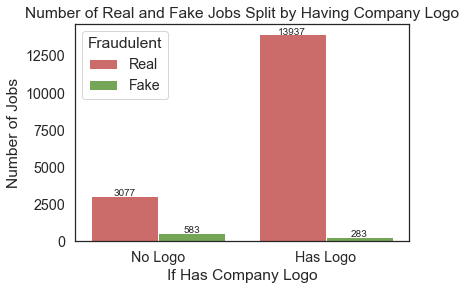
1. If company is telecommuting is a factor of if the company is fraudulent.



For jobs that are not telecommuting, the number of real jobs is more than 20 times of the number of fake jobs. However, for jobs that are telecommuting, the number of real jobs is only about 11 times of the number of fake jobs.

Since the p-value for telecommuting is really small (<0.05), there is almost no likelihood to have the difference by change, which means there is significant difference between real and fake jobs on if the company is telecommuting

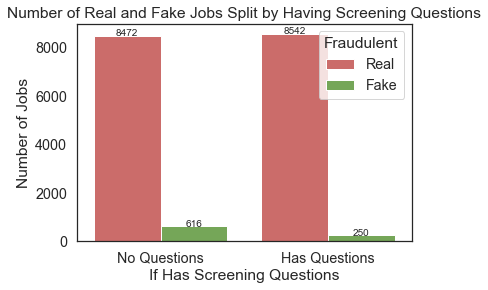
1. If company has a logo is a factor of if the company is fraudulent.



If the company has a logo, the number of real jobs is almost 50 times of the number of fake jobs. However, if the company doesn't have a logo, the number of real jobs is only about 5 times of the number of fake jobs.

Since the p-value for has\_company\_logo is really small (<0.05), there is almost no likelihood to have the difference by change, which means there is significant difference between real and fake jobs on if the company has a logo.

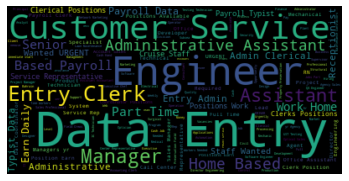
1. If company has screening questions is another factor of if the company is fraudulent.



If the company provides screening questions, the number of real jobs is more 34 times of the number of fake jobs. If the company doesn't provide screening questions, the number of real jobs is only about 14 times of the number of fake jobs.

Since the p-value for has\_screening\_questions is really small (<0.05), there is almost no likelihood to have the difference by change, which means there is significant difference between real and fake jobs on if the company has screening questions.

1. The top 3 possible job titles for fake jobs are Data Entry, Customer Service and Engineer.



1. In the description of jobs, fake jobs mentioned “team”, “work”, “experience”, “company”, and “customer” a lot, same as real job. But fake jobs also mentioned a lot “time”, “position”, “management”, “project” and “service”. However, real jobs have common words like “business”, “new”, “sales”, “looking”, “development”. Actually, there are a lot of repeated description both in fake and real jobs because 1) the same jobs are posted on the different job seeking websites; 2) the same positions are available in several locations; 3) similar job titles share the same description. So, there is definitely a bias in this result. Further analysis should try to eliminate this bias.
2. In the benefits part, fake jobs mentioned “nan”(null value), “benefits”, “company”, “work”, “time”, “competitive”, and “paid” the most, same as real jobs. However, fake jobs also mentioned “training”, “environment”, and “full”, and real job mentioned a lot “salary”, “paid”, “dental”, which is very interesting. It looks like fake jobs don’t talk about monetary benefits like real job (because they can’t pay it). This analysis has the same issue as the analysis about description.
3. Using Random Forest Classifier to generate prediction models, the accuracy is more than 97%, which is pretty good. And the top 3 important features are description, title and benefits.
4. I also used Naïve Bayes Classifier to generate prediction. For the method of Gaussian, the accuracy is only 57.16%, which is low because Gaussian Naïve Bayes is based on a continuous distribution. However, when I run Bernoulli Naïve Bayes, the accuracy is 95.77%, which is high since Bernoulli Naïve Bayes is useful when a feature can be present or not.
5. There are a lot of other interesting variables like “required\_experience”, “required\_degree” for the future analysis. But there more missing data in those variables, so how to handle them is a problem.