

Credit: Canva Pro

Spotting a Fake Job Posting

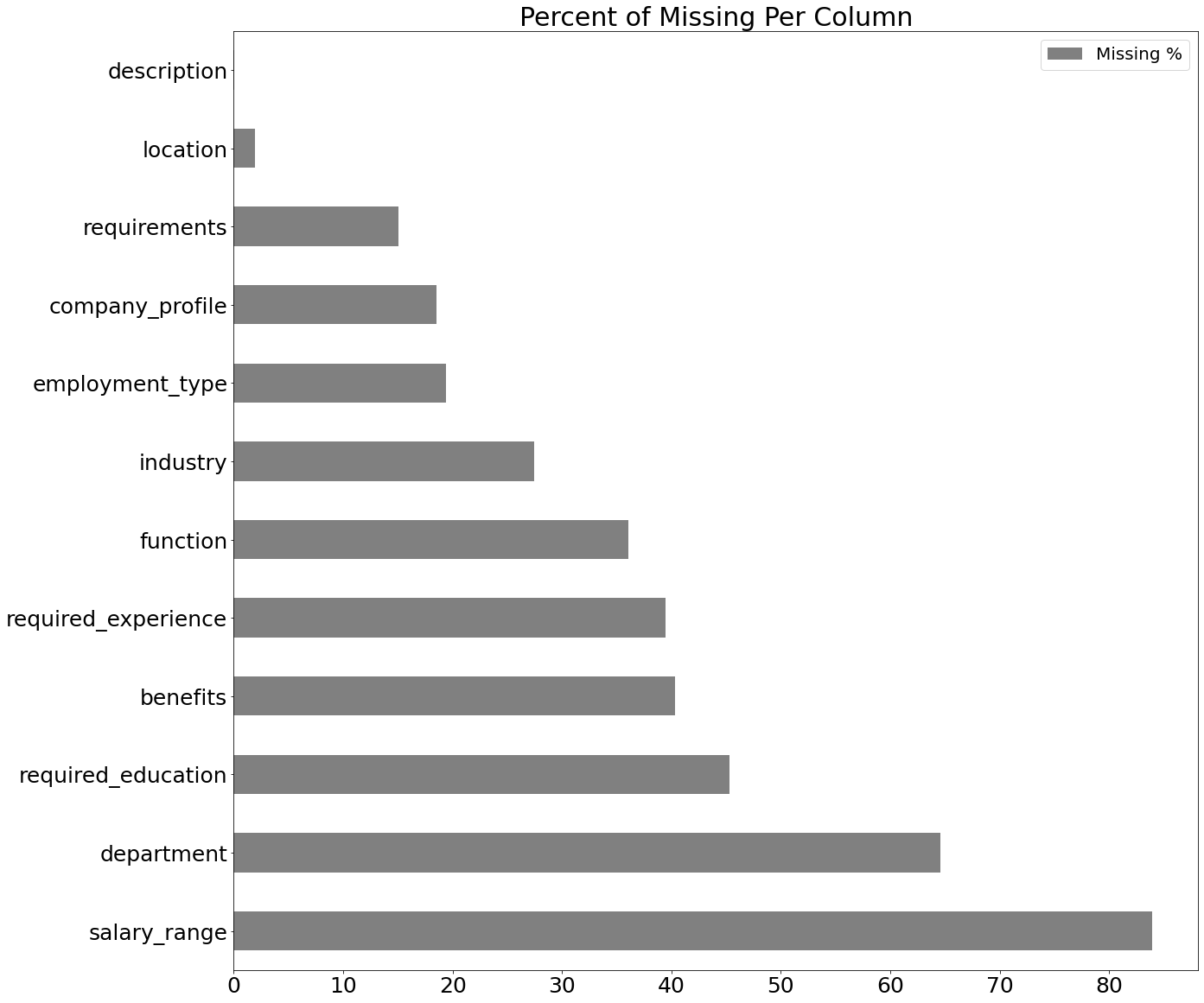
By [Marci Morrell](mailto:amorrell@stedwards.edu), [Pender Fitz Gerald](mailto:pfitzger@stedwards.edu) and Leticia Rinaldini

The job market is booming in current times. With the “great resignation” after the pandemic, people are redirecting their career paths and searching for the perfect position. Virtual job hunting has also become the primary form of searching nowadays, and that leaves a lot of uncertainty for the applicant. Could the job you are applying for be a fake? Could it be a scam to steal your information? Knowing what to look for in a job posting can save you from giving out personal information and falling for these scam job ads.

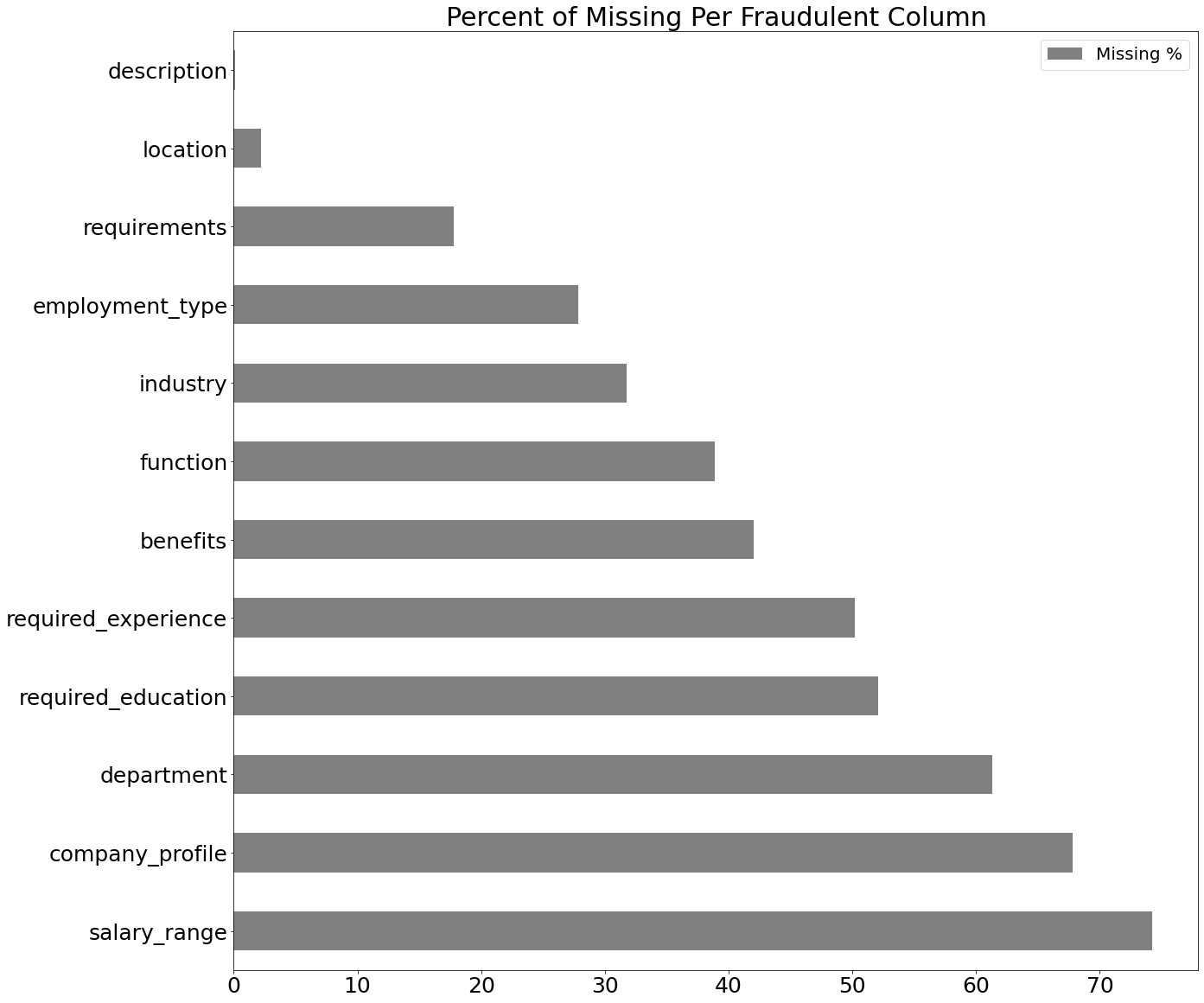
Our group decided to take a deeper look into the realm of job postings and their validity. We specifically analyzed the types of data for each job posting, the effects of missing data, and what types of missing data seem to make the biggest impact. Our overall goal is to create a machine learning model to predict whether a job posting is fraudulent or not.

We started off with a data file from Kaggle of approximately 18,000 job postings. Within the data, only 5% of the total jobs were fraudulent job postings. The data included all sorts of factors about each job posting including job title, location, department, benefits, company profile, job description, job requirements, company logo, employment type, fraudulency status and more.

First, we analyzed the missing values and explored their correlation with job fraudulency. Overall, the column with the most missing values was the salary range. About 84% of the values in the column are missing. Within the fraudulent job postings only, the salary range was also missing the most values. The column with the second-most missing values within the fraudulent postings was the company profile which was missing about 70% of the values. This is a large difference from looking at just the real job postings, which only had about 20% of the company profile column missing. This led us to believe that if a job posting does not include a company profile, it has a higher likelihood of being fraudulent. Fraudulent jobs also had a lower percentage of salary ranges missing compared to the percentage of salary ranges missing in the real posts. A majority of the salary ranges in the fraudulent postings were also unrealistic or a really large or obscure range such as $130000-$145000, $10000-$250000, and $23040-$28800. Fake job ads seem to be attempting to bait the applicant using these salary ranges and the benefits.

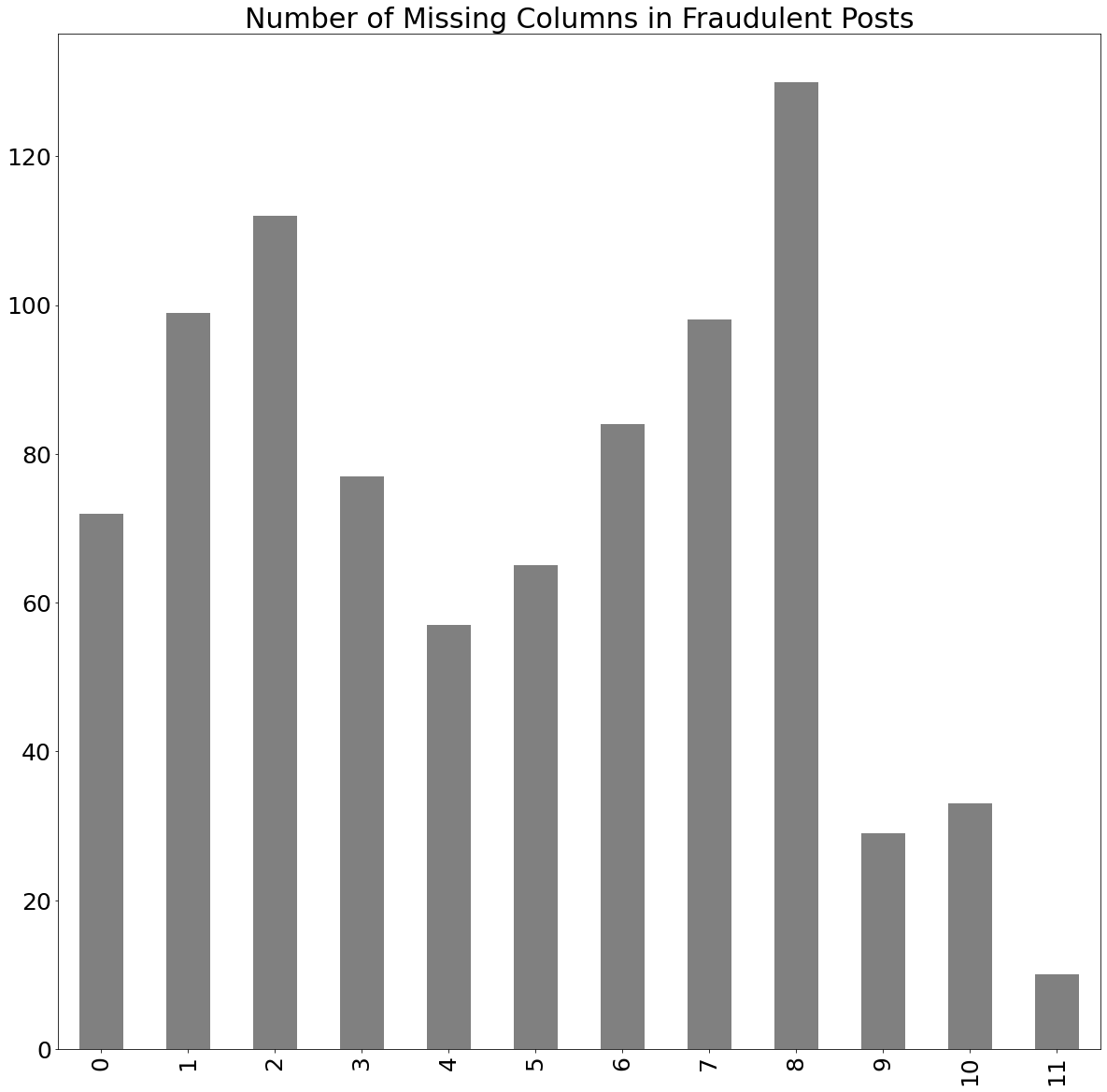


*Figure 1*

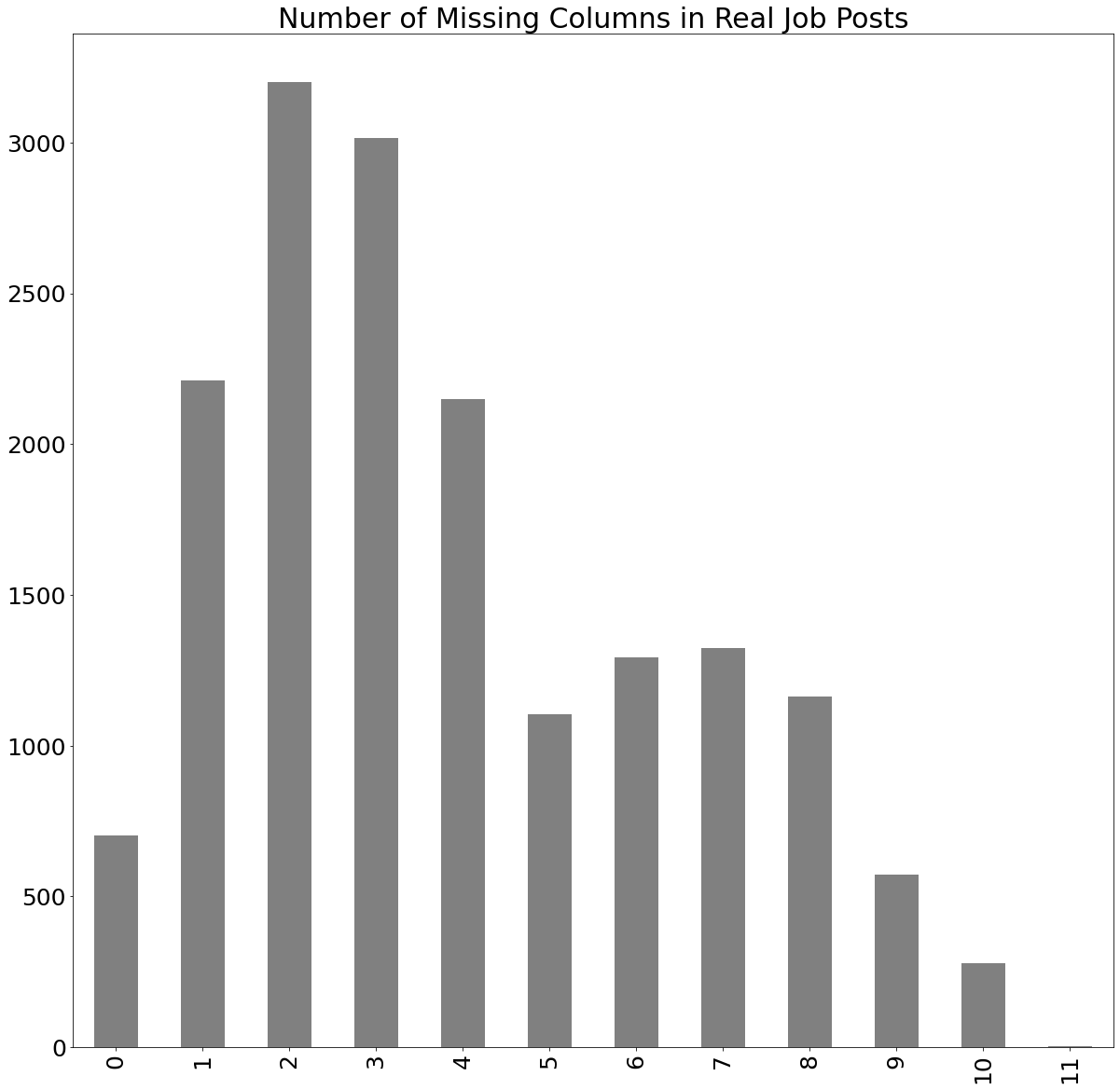


*Figure 2*

Almost none of the legitimate job postings had 11+ missing columns of data as shown in the figures below. We also found that most of the legitimate job postings had between 1 and 4 columns of missing information. We did not find any other staggering correlations when looking at the number of missing columns and fraudulency. This led us to utilize other columns of information when creating our model, in order to get the most accurate predictions.

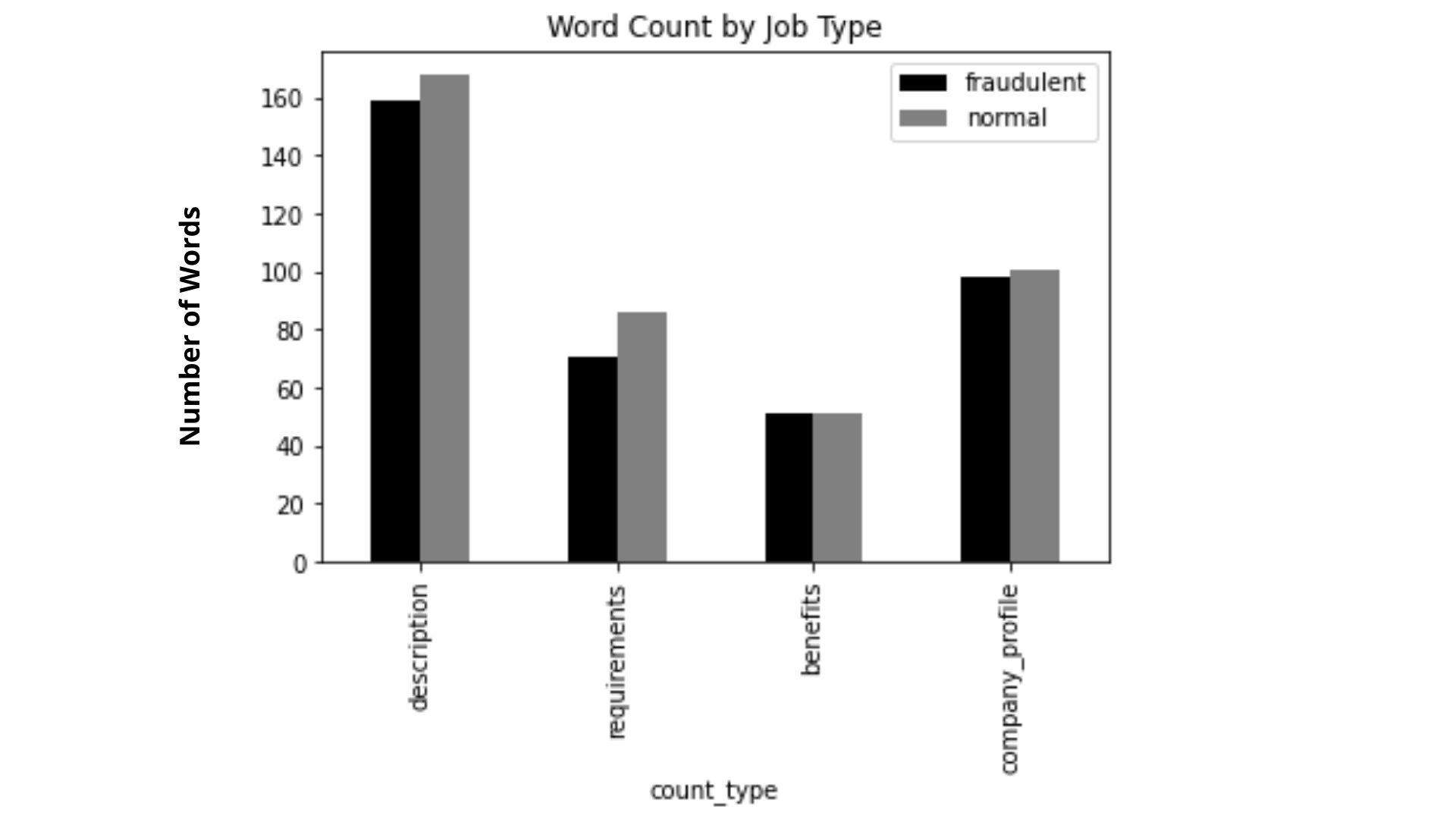


*Figure 3*



*Figure 4*

Next, we looked at the word count of the columns which included short paragraphs of information to see if there was any correlation with fraudulency. Real job posts had higher counts for job description, company profile and job requirements as shown in the figure below. This makes sense because real job posts generally include more detail on the role and company itself, rather than the salary and benefits.



*Figure 5*

Finally, we began creating the model. The data we were working with, as mentioned in the introduction of this article, only included 5% fraudulent job posts overall. This gave us an extremely unbalanced data set to work with. We solved this by undersampling our data which entails creating a new data set that includes the same number of real posts and fake posts. Since we did not find any strong correlation between fraudulency and missing data columns, we decided to find the correlation coefficients for all of the columns in relation to the fraudulent column. The company logo column, indicating whether a company logo was present in the job post, as well as the company profile column were found to have the highest correlation with fraudulency. We created a logistic regression model utilizing these two columns and were able to achieve 74% prediction accuracy. This is not a terrible accuracy percentage, but we still wanted to see if there was a better model out there. Using a power set function, we found a combination of columns in the data that yields the best accuracy. Those columns were the word count of the job description, word count of the company profile, whether salary range was present, whether employment type was present, and department ID. Using those columns we created another logistic regression model which yielded a 75% accuracy score.

Ultimately, what we discovered using our models was that with the data we were given to analyze, the best models can correctly predict the fraudulence of job posts between 75% and 80% of the time. We did find a few other interesting things to look out for when deciphering between a real and fake job posting. If a job post has incredible benefits and the salary range is unrealistic or unusual, it is likely a false or scam post. Job posts with lengthier and more detailed information on the requirements of the role, the role itself, and the company profile are likely safe and real companies. We also found that whether the job posting includes a company logo or not correlates highest with fraudulency. Any job advertisement that seems fishy, or too vague is likely to be fraudulent so be sure to keep your guard up in the world of virtual applications in order to potentially protect yourself from identity theft or falling for an unfortunate false job scam.