**Employment Scams: How can we predict job fraudulence?**

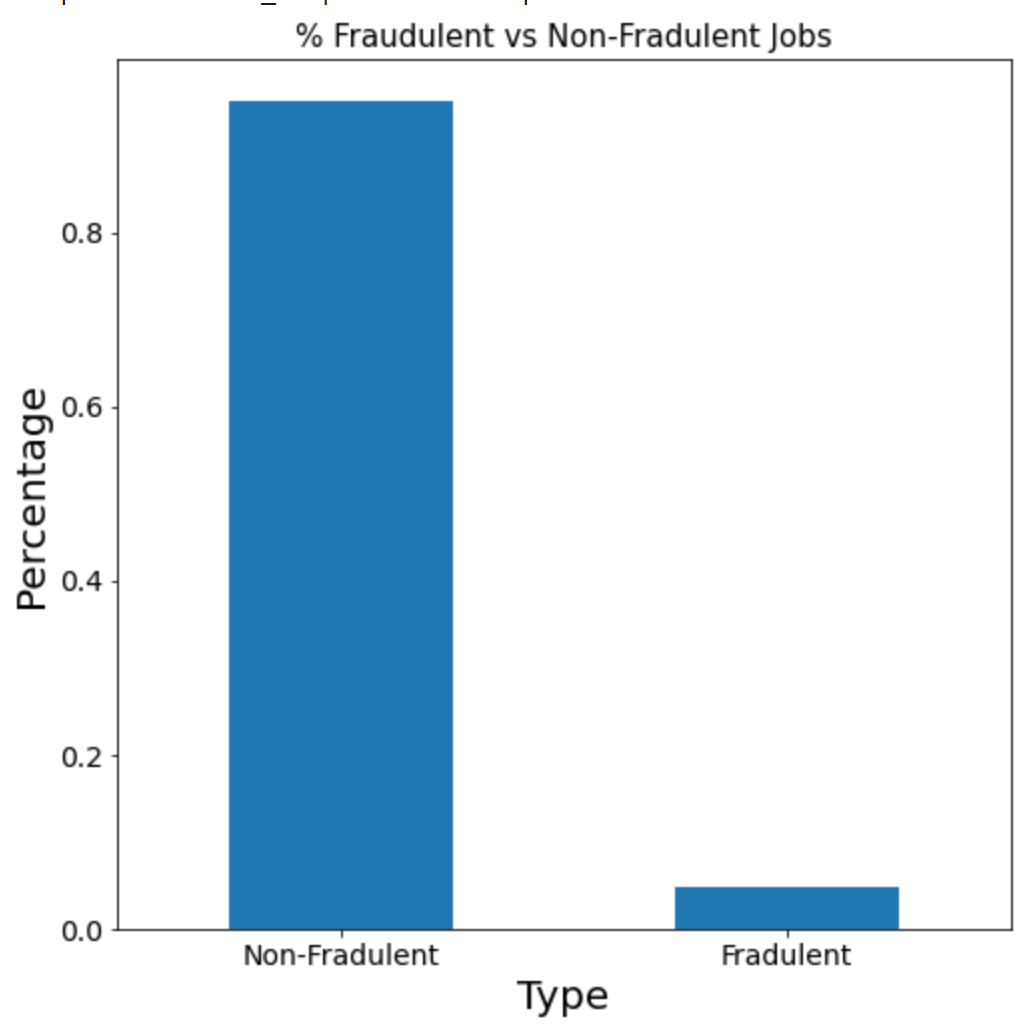
****

Image credited to RapidEye (Pixabay)

By Samantha Mathieu, Shyla Sharma, and Ruben Valdovinos

**Introduction**

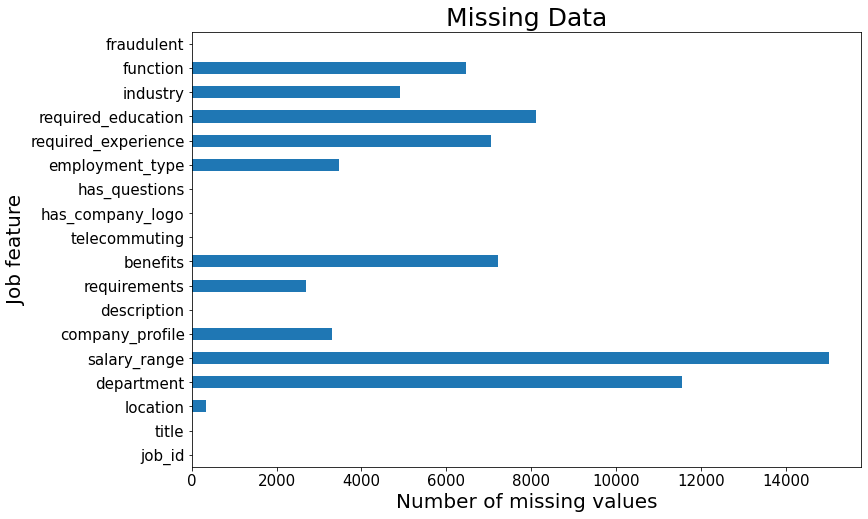
The data used by our team for this data science project was downloaded from Kaggle which includes job descriptions for 18,000 job postings. The jobs are across various industries with factors including salary, employment type, industry, education and experience level, company profile, benefits, and the factor of interest being fraudulence. Eight hundred out of the 18,000 job postings are fraudulent.



Because many of the factors are descriptions, most of the information in the dataset is textual, which means we needed to work with the data so we could predict fraudulent jobs. To do this, we had four main objectives. These objectives were to explore if missing data from the job descriptions could predict fraudulence, what features impact fraudulence, if the length of descriptions could predict fraudulence, and if combining all of these variables together could predict fraudulence.

**Missing Values Analysis**

Missing data from job descriptions seemed to be an important aspect of the dataset when it comes to fraudulent listings. We wanted to see if the amount of NA values from each column with missing data could serve as a good predictor of job fraudulence. First, we looked at the amount of NA values in each variable.



As we can see from the figure, salary range and department had the most missing data where some variables had no missing data. We extracted the columns with NA values and made new variables that indicated if that variable had NA values for each job. Each of these variables were put into the missing values model.

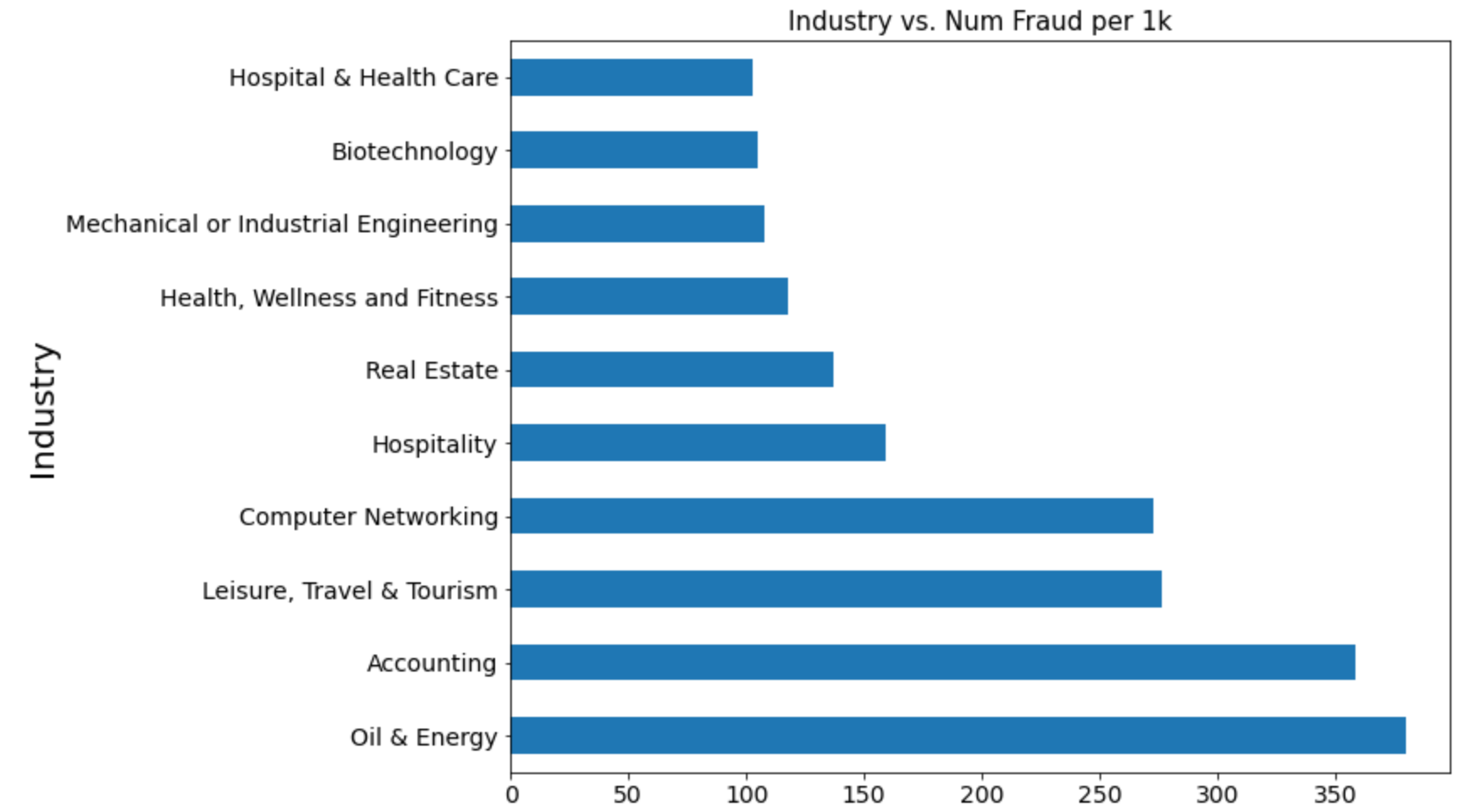
| True Positive Rate  0.83 | False Positive Rate  0.17 |
| --- | --- |
| False Negative Rate  0.36 | True Negative Rate  0.64 |

The confusion matrix above tells us that with just the NA value columns included, the model is able to predict 83% of fraudulent jobs and 64% of non fraudulent jobs correctly. The accuracy of the model was 0.83 and the efficiency (ROC AUC score) was 0.75. Overall, the model did a fairly decent job at predicting fraudulent jobs based off of missing values, but we wanted to explore if other variables would improve the model.

**Features Analysis**

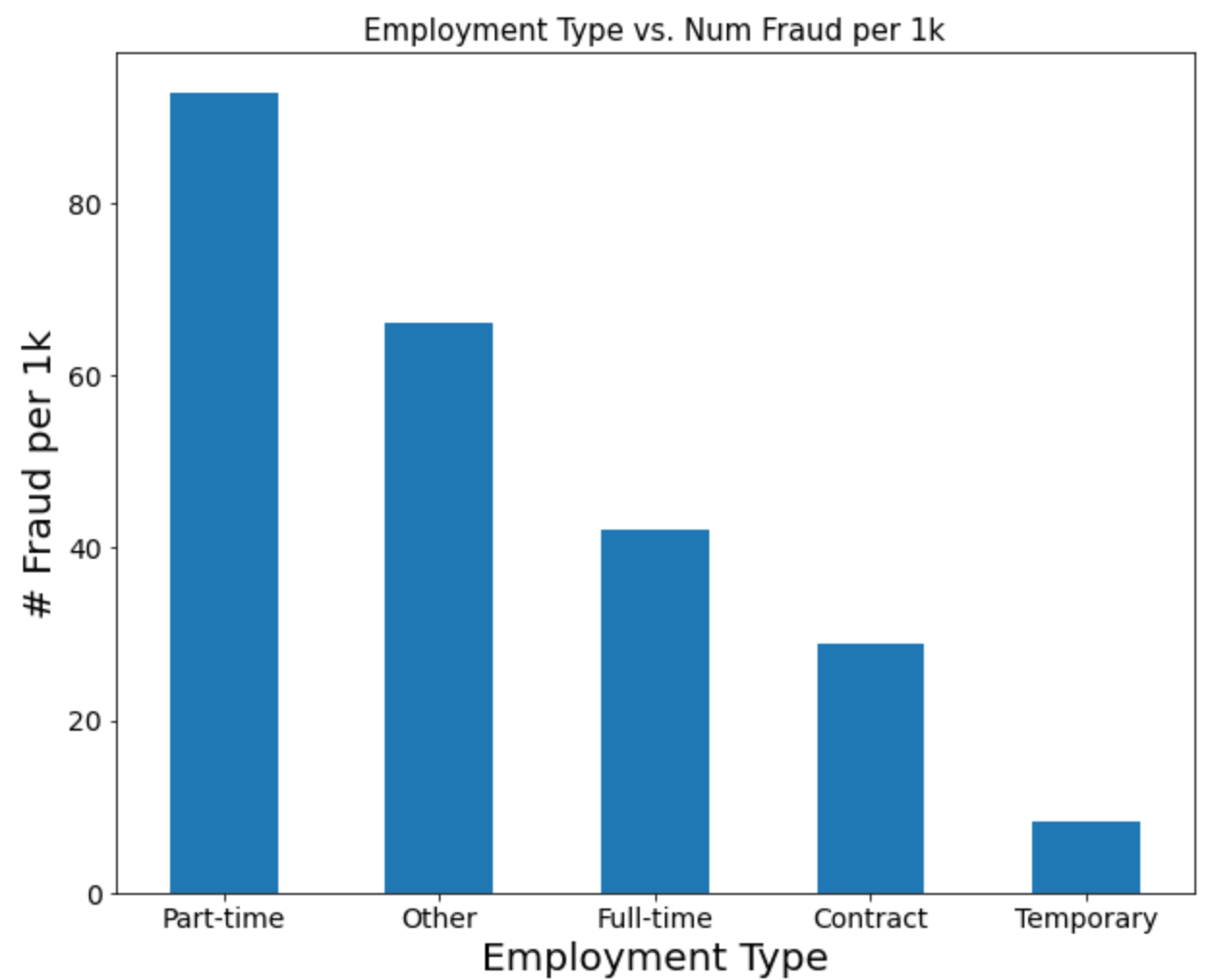
After analyzing how missing data impacts whether or not a job is fraudulent, we wanted to look into which specific features could help predict the fraudulency of a job. The categories we decided to analyze were industry, employment type, required experience, and required education.

We began by determining the top 10 most fraudulent industries.



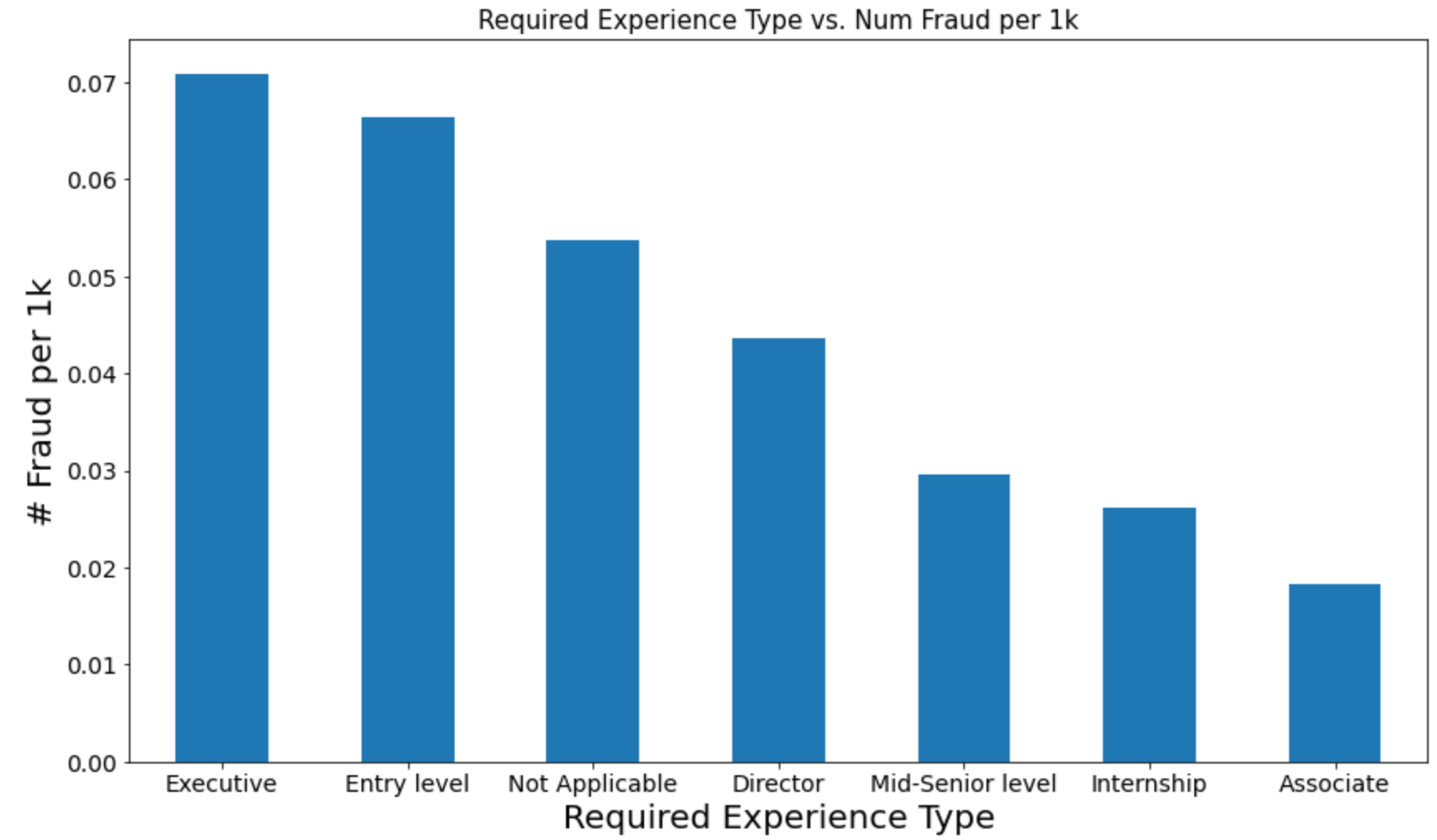
The Oil & Energy industry was determined to be the most fraudulent, with approximately 380 per 1K jobs being fraudulent. Comparatively, the 10th most fraudulent industry, Hospital & Healthcare had approximately 102 per 1K jobs being fraudulent. Because these ten industries all had a relatively high proportion of fraudulent jobs, we decided to include them in our feature model.

The next feature we looked at was employment type.

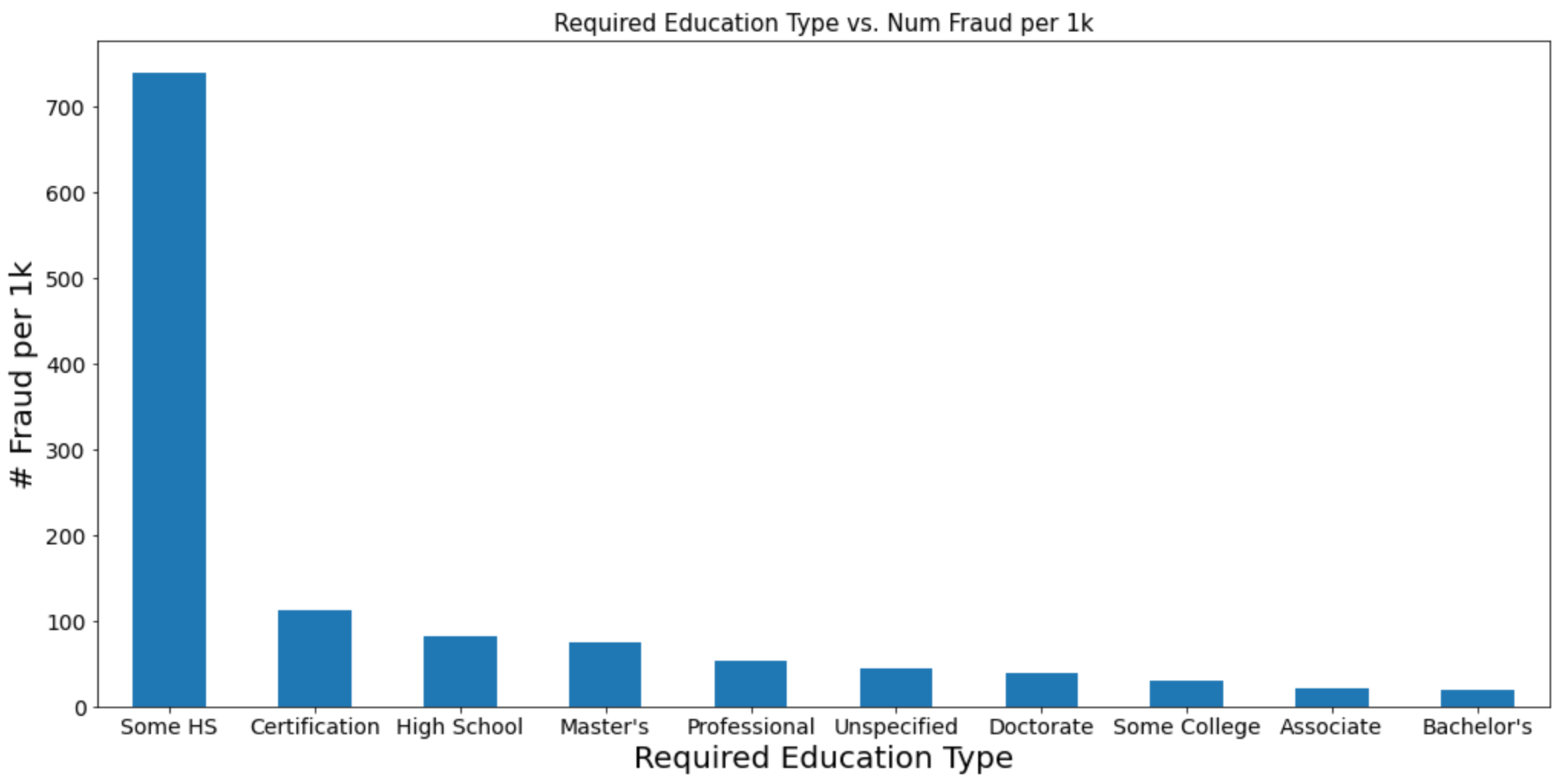


As demonstrated by the graph, part-time jobs had the most fraudulency, with approximately 92 jobs per 1K being fraudulent. For this reason, we included specifically part-time jobs in our feature model.

The third category we analyzed was required experience. For required education, Executive and Entry Level positions had approximately the same number of fraudulent jobs, with 70 fraudulent executive jobs per 1K and 66 fraudulent entry level jobs per 1K. Jobs that required executive or entry level experience were then added into the feature model.



The final feature we looked at was required education. As demonstrated by the graph below, jobs that required only some high school coursework had the highest rate of fraudulence, with approximately 740 fraudulent jobs per 1K jobs. For reference, jobs that require a certification had 111 fraudulent jobs per 1K jobs.



By consolidating certain values from all of these different categories, we created our feature model. The specific values we fed into this model were the top ten fraudulent industries, part-time jobs, jobs that required executive or entry level experience, and jobs that require only some highschool education.

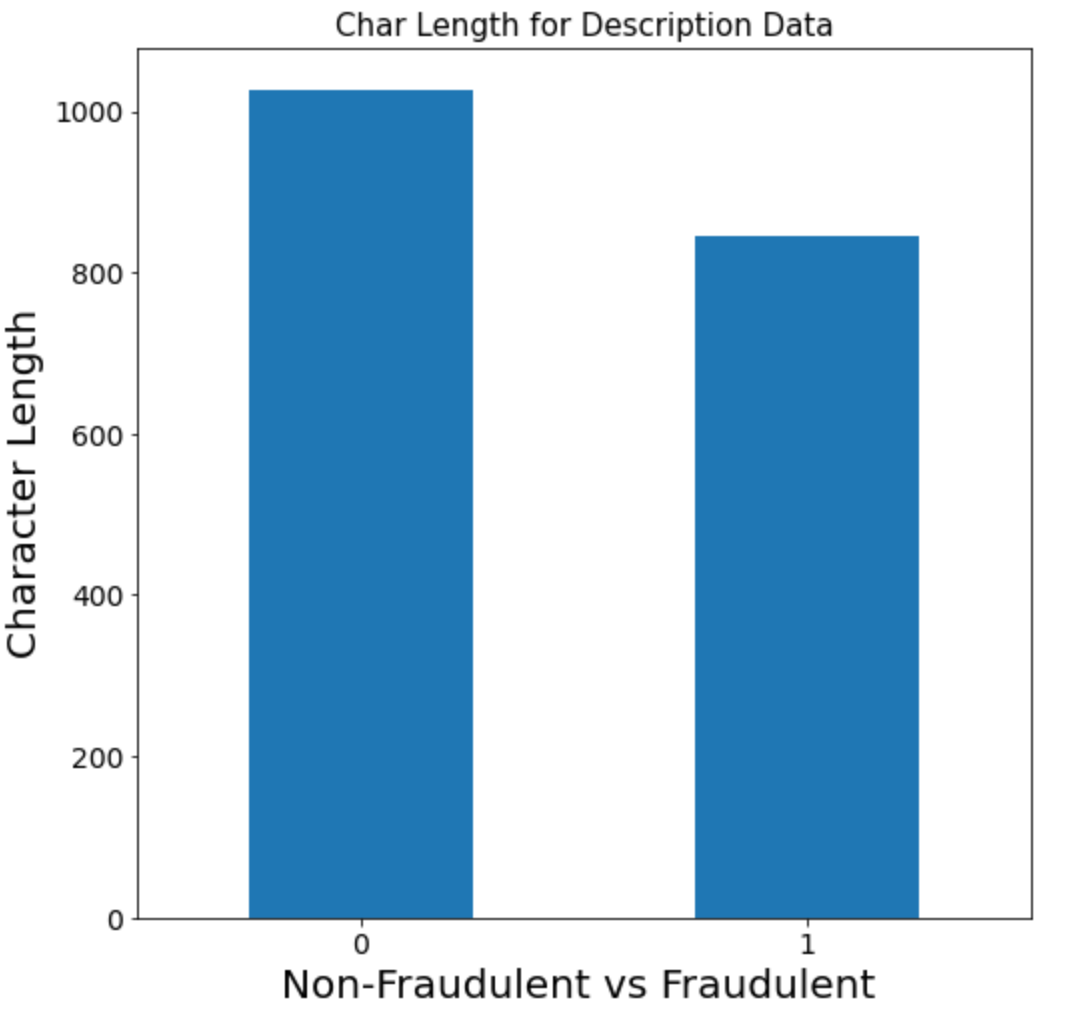
| True Positive Rate  0.89 | False Positive Rate  0.11 |
| --- | --- |
| False Negative Rate  0.51 | True Negative Rate  0.48 |

As demonstrated by the confusion matrix above, this model was able to predict about 89% of fraudulent jobs accurately. However, the true negative value was significantly lower, with the model only predicting 48% of non-fraudulent jobs accurately. This model also had an ROC AUC score of 0.69, meaning that it was decent at differentiating between a fraudulent and non-fraudulent job.

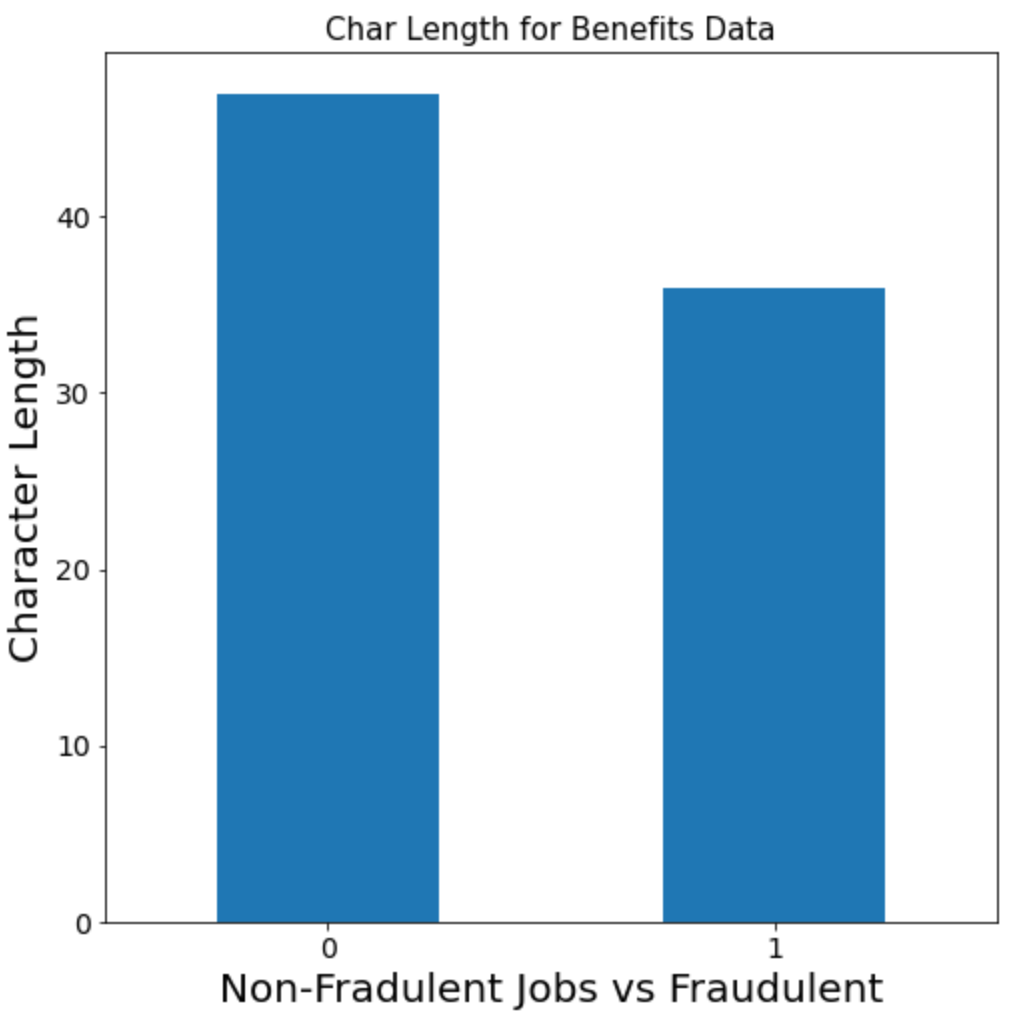
**Character Length Analysis**

After we had looked at certain features we began to look into character length. The columns with textual data were description, benefits, requirements, and company profile. Once we identified these columns, we analyzed the character length of each string and stored that value as a number in a separate column. We took the median of both fraudulent and non-fraudulent character length.

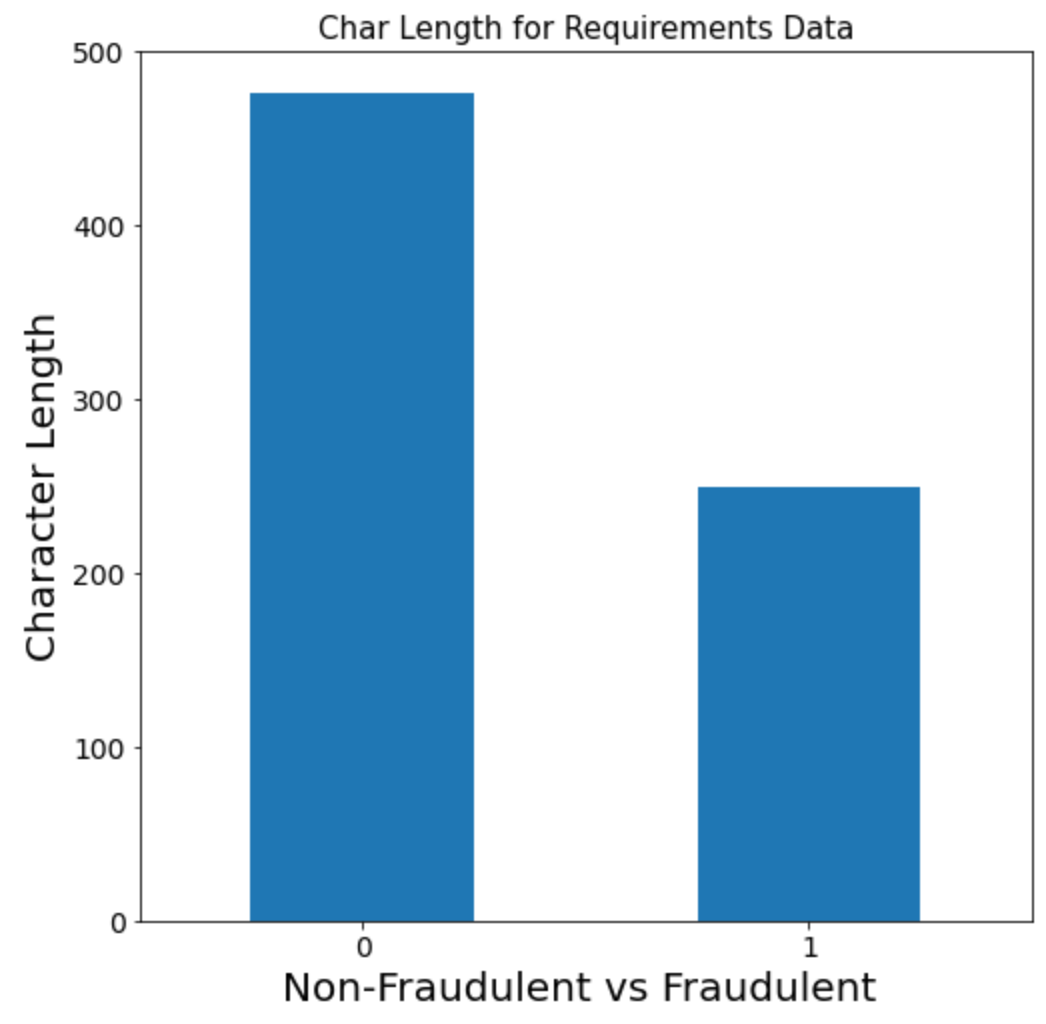
The first thing we looked at was comparing the median character length of the description column. The following graph is that comparison.



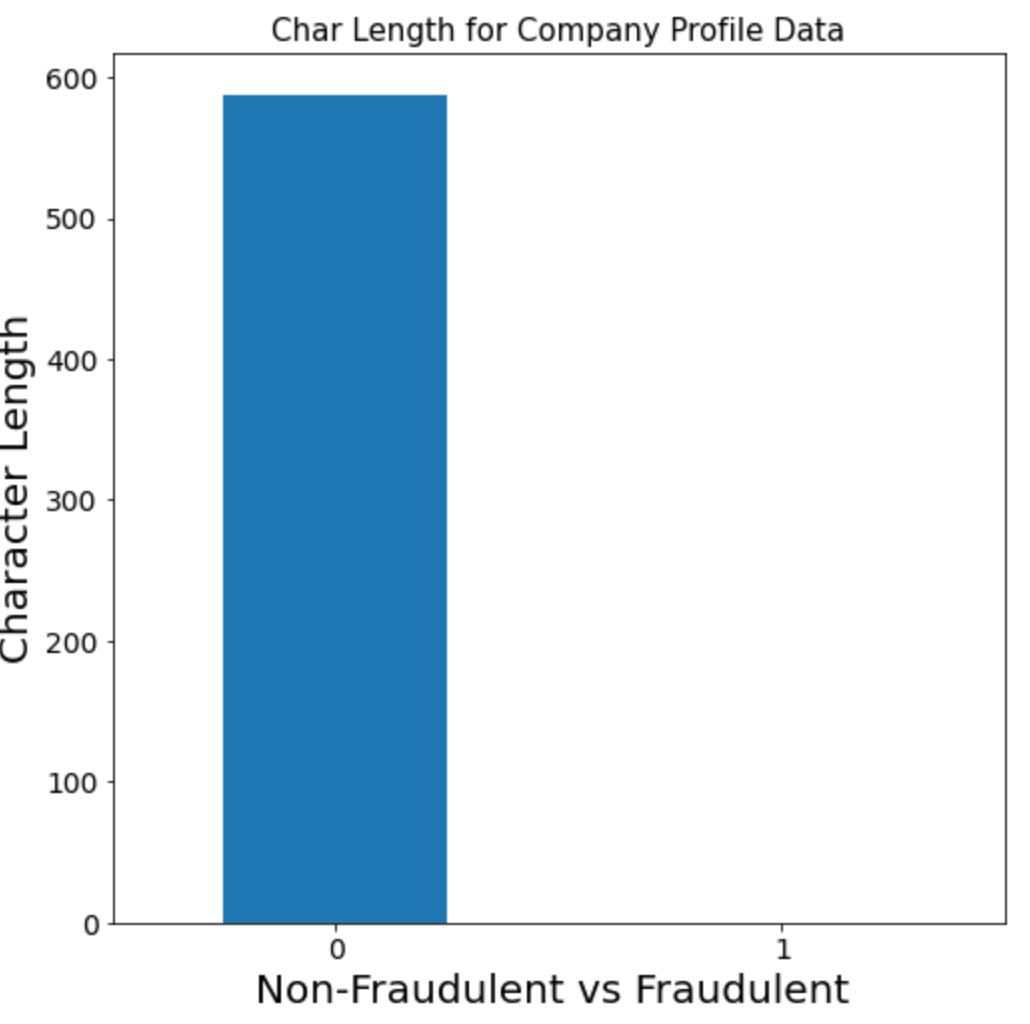
The median character length for Non-fraudulent has roughly 47 characters and for fraudulent job openings roughly 37 characters. After seeing that the fraudulent jobs did a good job on character length we moved on to look at the benefits column. Once again the following graph is



the comparison of the median character length. The non-fraudulent jobs had 1027 characters while the fraudulent jobs had roughly 844 characters. The fake jobs do well once again by getting close enough to the non-fraudulent jobs in character length. However, we kept looking and then examined the graph for the median character length for the requirements column.



Non-fraudulent has roughly 476 characters and fraudulent job openings roughly have 249 characters. This is useful because it shows that legitimate job listings in our data have almost double the character length. Lastly, we looked at the company profile character length and we struck gold.



We found that the non-fraudulent data had a median value of 588 and the median for the fraudulent cases had a median of zero. Meaning that if there is no company profile character length this is a huge red flag.

Finally, we put all of this into a linear regression model taking into account every column previously mentioned.

| True Positive Rate  0.62 | False Positive Rate  0.37 |
| --- | --- |
| False Negative Rate  0.27 | True Negative Rate  0.73 |

Our model gave us a true positive of .62 meaning that out of all our fraudulent jobs, this model successfully predicted the outcome 62% of the time. The model also gave us a true negative rate of 73% meaning that this model could successfully predict non-fraudulent 73% of the time. However, we still wanted to make a better model so we consolidated our results.

**Combined Model**

By consolidating the results of the analyses on missing values, features, and character lengths of textual data, we were able to generate the following model.

| True Positive Rate  0.79 | False Positive Rate  0.21 |
| --- | --- |
| False Negative Rate  0.17 | True Negative Rate  0.83 |

As demonstrated by the confusion matrix, the true positive rate was 79%, meaning that out of all the actual fraudulent jobs, 79% were accurately predicted. The true negative rate was 83%, meaning that out of all the actual non-fraudulent jobs, 83% were accurately predicted. This model had an ROC AUC score of 0.80, which means that it was efficient at distinguishing between fraudulent and non-fraudulent jobs.

**Conclusion**

After looking at our three previous models they all had their strengths and weaknesses. When we put our models together we were able to create a strong model that took everything into account. The main takeaway from the missing values model was that it had a strong true positive. The features model also had a very strong true positive with part-time jobs being the most important feature of its model. Lastly, the character length model had the best true negative and its most important column was the company profile. Taking everything into account and putting it into a model yielded us a good true positive and a good true negative. Now with our model, we can successfully predict when a job is fraudulent 79% of the time and when a job is real 83% of the time.