Real or Fake Job Postings



John Jefferson, Stephen Babcock

IST 707 Project Report

Dec 5, 2022

Contents

[Introduction 3](#_Toc122033022)

[Motives 4](#_Toc122033023)

[Problem Statement 4](#_Toc122033024)

[About the Data 4](#_Toc122033025)

[Data Loading 5](#_Toc122033026)

[Fake and Real Posting Count 5](#_Toc122033027)

[Data Preparation and Cleaning 6](#_Toc122033028)

[Creation of Fake and Real Data Frames 7](#_Toc122033029)

[NA Count, Real vs Fake Postings 7](#_Toc122033030)

[Binary Variables 8](#_Toc122033031)

[Final Structure for Analysis 9](#_Toc122033032)

[Models and Methods of Analysis 11](#_Toc122033033)

[K-Means Clustering 11](#_Toc122033034)

[HAC Analysis 15](#_Toc122033035)

[Decision Tree 17](#_Toc122033036)

[Association Rule Mining 18](#_Toc122033037)

[Word Clouds 19](#_Toc122033038)

[Naïve Bayes 22](#_Toc122033039)

[K Nearest Neighbors 22](#_Toc122033040)

[Support Vector Machine 24](#_Toc122033041)

[Random Forest 25](#_Toc122033042)

[Conclusion 26](#_Toc122033043)

[References 27](#_Toc122033044)

## Introduction

Identity theft has become a growing threat in our society over the past couple of decades. Scammers are becoming increasingly clever with new schemes to scam people on the internet. A call about “your car’s extended warranty” or a text stating “your amazon account was locked, click this link” are both common enough to have spawned an entire genre of meme and short video humor. These types of scams are improving as time advances through the 21st century with new vectors popping up in every facet of daily life, including in job postings.

Cyber criminals advertise jobs through posting in the same way an actual company does. They lure the applicant in with a job and leverage their assumed position as an employer to lull the applicant into a false sense of trust. The cyber criminal’s goal is to obtain personal information that may be used to access accounts, defraud the applicant, or they may even trick the applicant into providing money.

Fake job postings have been around for decades, but discipline in the fraudulent activity, combined with new technology have made it harder than ever for applicants to differentiate between an authentic job posting and a fake one. The cyber criminals may create job postings using fake company logos, names, locations, requirements, benefits, or more to deceive their victims. There are few tools available to the applicant to validate the authenticity of a job posting.

The Covid-19 pandemic has provided a significantly lucrative opportunity for the fake job scammers because of the large amount of churn across the work force, with tons of people looking for new jobs. The problem has become a large enough threat to cause the FBI to issue warnings about the activity with their most recent official warning on this activity being issued in 2021.

### Motives

Combating the growing prevalence of fake job postings requires more than mere warnings from government agencies. Job boards and employment agencies must take care to ensure the safety of their service and the applicants that rely on it by installing security controls that limit the presence of fraudulent job postings. Applying to a job posting exposes the job applicant to potential social engineering and such social engineering leverages normal social conventions to collect information that may be used for fraudulent purposes such as identity theft, or to gain account access for property theft.

### Problem Statement

To assist in combating fraudulent job postings, this project seeks to train an algorithm to detect fake job postings accurately and reliably. This algorithm would become a candidate tool for screening future job postings for authenticity.

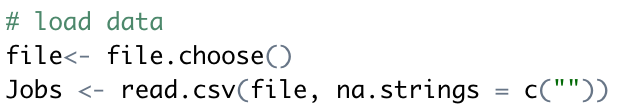
## About the Data

The data from this dataset includes columns which represent the different attributes on a particular job posting. The rows are the individual job postings. There are 18 columns in this dataset and 17,880 rows. The variables are a mix of strings and integer, with the integer variables being binary (yes or no) for those items. The data is 95% real jobs; 17,014 real jobs and 866 fake jobs.

The data set is provided by Kaggle. This is a newer data set, uploaded approximately three months prior to the date of this report. The Kaggle site does not include any background information about the data, so it is unknown where the data originated from.

<https://www.kaggle.com/datasets/whenamancodes/real-or-fake-jobs>

### Data Loading

This is how we loaded in the data:

### Fake and Real Posting Count

Confirming the data load matches the description provided by Kaggle, the data set shows a large number of real jobs, and a much smaller proportion of fake jobs. One hopes this is relational to the reality of proportion of fake jobs in the wild.

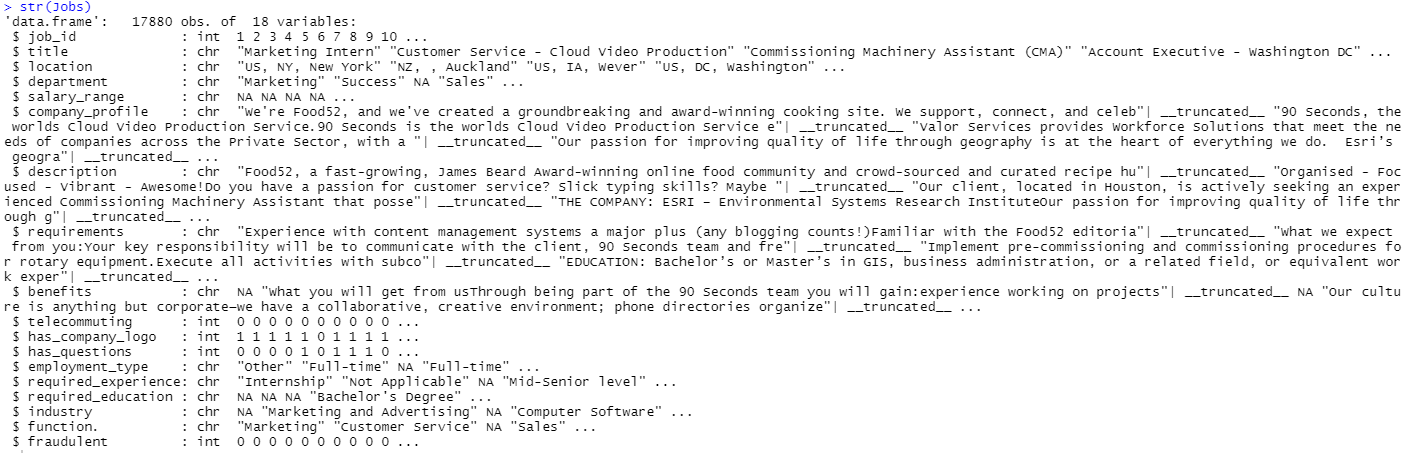
Graphical user interface, text

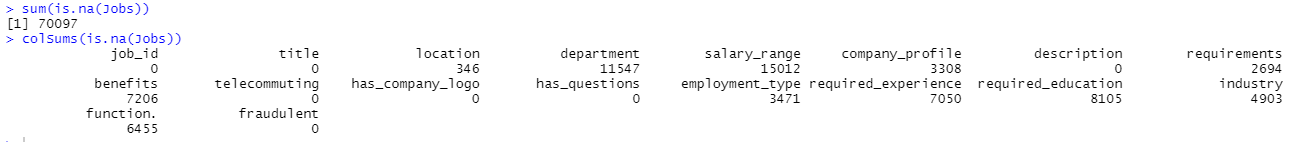
Description automatically generated

Chart

Description automatically generated

### Data Preparation and Cleaning

Preparation and cleaning of the data begins with the “str” command and reviewing the completeness of the data set by identifying the presence of NAs.



Table

Description automatically generated with medium confidence

### Creation of Fake and Real Data Frames

#### The initial review of the data set revealing a large volume of NAs in different variables prompts deeper analysis to identify if the absence of data defines the authenticity of the job posting. To further this evaluation, the data set is broken into two data frames- one data frame containing all of the real job postings (fraudulent = 0), and a second containing all of the fraudulent job postings (fraudulent = 1).

#### 

### NA Count, Real vs Fake Postings

#### The graphics below display the volume of NAs per variable for the real and fake job postings. The review of the variables finds that the rate of missing data between fake and real job postings is similar, except regarding company profiles. Only 16% of real job postings are missing a company profile whereas 68% of fake jobs are missing this detail.

#### Chart, bar chart Description automatically generated Chart, bar chart Description automatically generated

### Binary Variables

In the binary variables, there is variation in the posting containing questions and company logo. About half of the real job postings have questions, and only a third of the fake jobs do. Most of the real jobs include a company logo, but only about a third of the fake postings do. There is also a slight variation in word count between the fake jobs and real job postings- the fake job postings are 13.23 words shorter, on average, than the real jobs. This difference might not be noticeable to a casual reader of job postings, but it might be enough for an algorithm to notice when considering the difference in other variables as well.

Shape

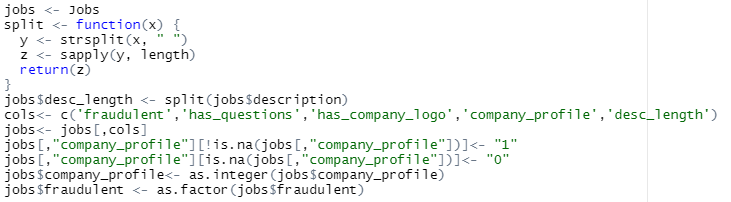
Description automatically generated with medium confidence

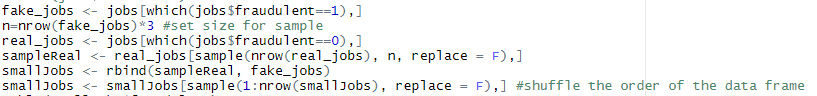
#### Chart, box and whisker chart Description automatically generated

### Final Structure for Analysis

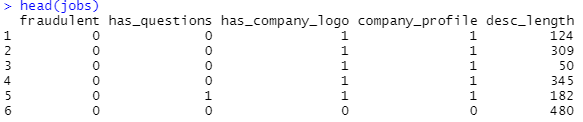
The ability to differentiate between authentic and fraudulent job postings will depend on identifying the differences in the postings. For that reason, modeling will leverage the variables in the data set that show the greatest amount of variation between the two outcomes. The company profile variable is transformed to a binary variable depicting the presence or absence of a profile in the posting. The variables used for analysis become: fraudulent, company\_profile, description length, has\_questions, and has\_company\_logo. The resulting data set for analysis contains no NAs and is of good, uniform nature.

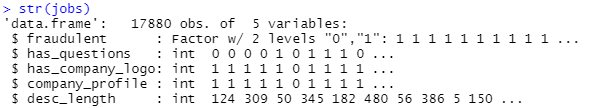
Analysis was originally performed against the full set of data. The data being heavily skewed towards real jobs presented difficulties as predication models had a 95% accuracy rate if they classified all jobs as real. The disparity in samples also made visualizing the kmeans clusters difficult. This led to developing a subset of real rules leveraging a random sample three times the size of the fake jobs.



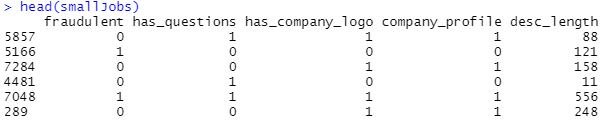


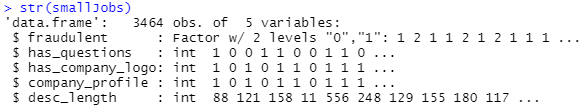
Full data set:





Reduced sample set:





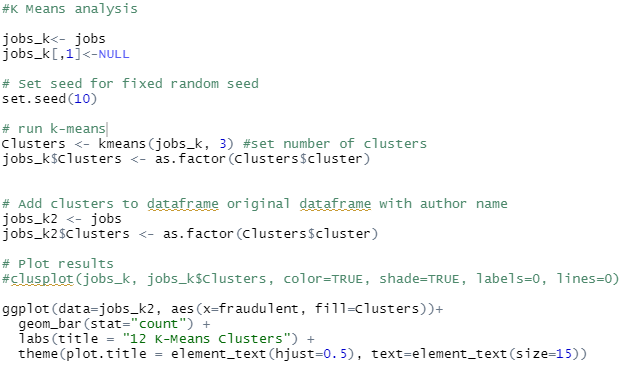
## Models and Methods of Analysis

#### With the data set defined and cleaned, analysis can begin towards identifying if an algorithm can differentiate between real and fake job postings. Multiple machine learning methods will be attempted to find the best candidate algorithm.

### K-Means Clustering

Analysis throughout the report leverages the reduced sample size of real jobs. However, kMeans clustering from the full data set are included below for the benefit of the reader to understand the disparity and degree of problem the skew in the data presented.

K Means clustering is a method to help identify patterns within the data set. The algorithm is tuned by providing the number of clusters to for, and a seed that randomizes the data and provides a measure for repeatable results. The data set was run through clustering with 3, 6, and 12 clusters. Multiple seeds were attempted, but each provided similar results. The results below leverage a seed of 10.



Full data:

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Reduced data set:

Chart, bar chart

Description automatically generated

Chart, bar chart

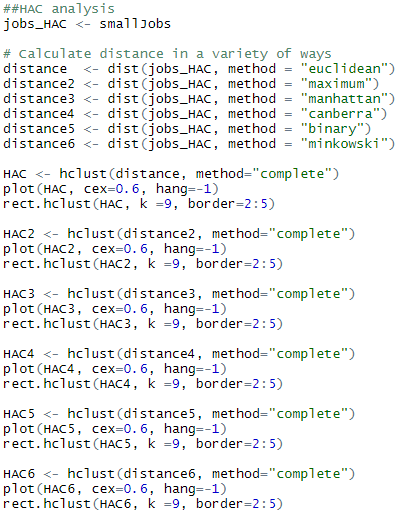
Description automatically generated

Chart, bar chart

Description automatically generated

### HAC Analysis

Hierarchical Agglomerative Clustering (HAC) is another clustering technique that results in a dendrogram visualization. This visualization can be helpful in understanding how the data is clustering. HAC analysis was performed on this data set, but the scale of the data created a dendrogram too large to be of much value. The visualization shows the observations are clustering on certain variables, but it is difficult to ascertain what those variables are.



.A picture containing diagram

Description automatically generated

### Decision Tree

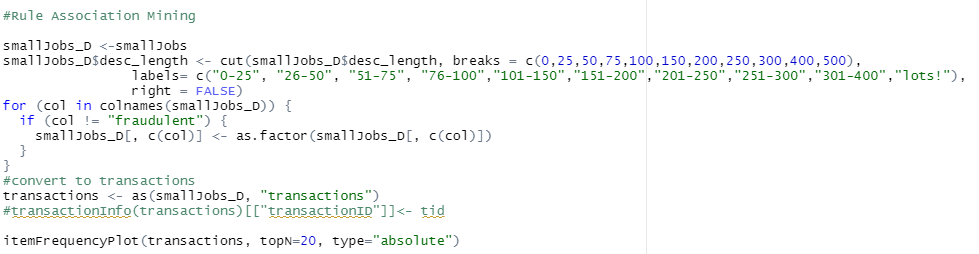
To determine the variables that are defining the clusters, the data set was converted to ARFF format and processed in Weka, j48 decision tree modeling. The resulting decision tree shows the observations clustering on presence of company profile, then presence of company logo.

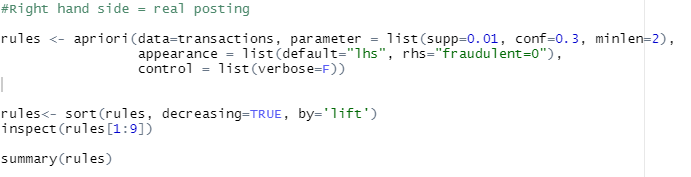
Diagram

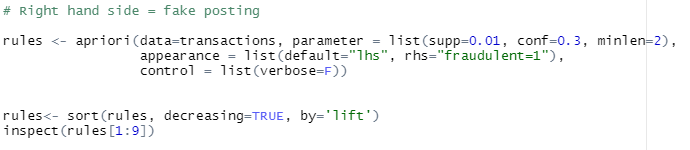
Description automatically generated

### Association Rule Mining

To support rule association mining, the description length was discretized. When the description length was left as raw numbers, the mining was returning poor results. Discretizing the description lengths led to improved metrics on the rules.







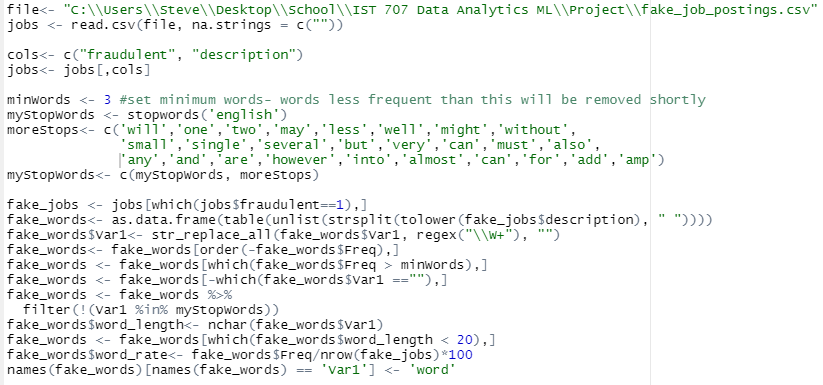
Text

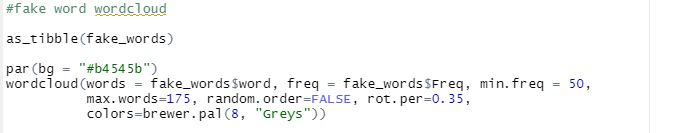
Description automatically generated with medium confidence

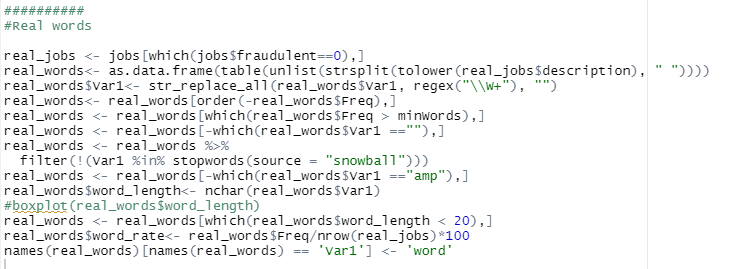
The rule association mining mirrors the decision tree, finding that the fake jobs are more likely to be absent of a company profile and company logo. The rules also show indication of the observation that fake rules have a lower average description length.

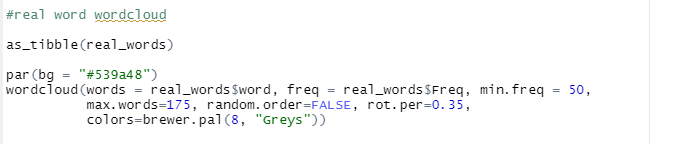
### Word Clouds

  Word clouds were constructed to identify if there is a disparity in word utilization between real and fake jobs. Two data sets are developed, each being a list of words used in the job descriptions with a count of how many times that word is present across all of the postings. One of the data sets contains the words from the real job postings, the second being from fake job postings. The words are normalized be making them all lower case, removing punctuation, numbers, and non-English words. A list of stop words was used to remove common words.









Fake Word Cloud

Text

Description automatically generated

Real Word Cloud

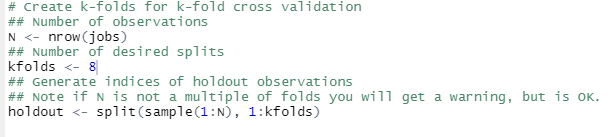
A picture containing text

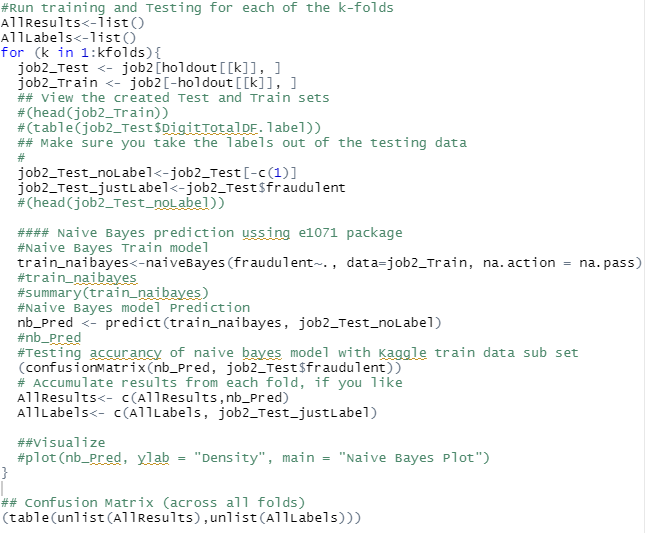
Description automatically generated

### Naïve Bayes

Naïve Bayes is a predictive classification algorithm based on Bayes Theorem. A notable feature of this algorithm is an assumption of independence within the variables of the data set. This algorithm and the preceding algorithms leverage analysis across a number of folds and there is a subset of the data that is held out for cross-validation testing.

The best achieved accuracy was 60.1%. with 8 folds. This is interesting considering if the model had determined every observation was a real posting, the result would have been 75%.





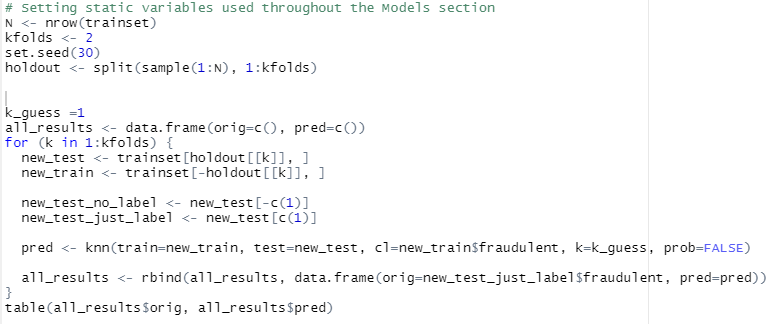
#### Naïve Bayes Confusion Matrix:

#### A picture containing text Description automatically generated

### K Nearest Neighbors

The K Nearest Neighbors (KNN) algorithm classifies a test observation based on the highest vote of the observations’ neighbors. Tuned with a value for k, k represents the number of neighbors that are given a vote on the identity of the test observation. The observations are plotted in space, with coordinates calculated based on the variable attributes of each observation. If k is set to 5, the 5 points nearest the test observation are considered for assessing the identity of the test observation. The identity with the highest presence among those 5 points defines the test observation. For example, if 3 of those points are fake job postings, then the test observation is identified as a fake posting as well, taking the identity of that majority.

Best performance of this algorithm was achieved when k is set to 1. This setting means that an observation is classified based on the nearest single neighboring observation. Best accuracy was 89.5%. This was the best performing algorithm for classifying the job postings.



#### K Nearest Neighbors Confusion Matrix:

A picture containing text

Description automatically generated

### Support Vector Machine

Support Vector Machines (SVM) are another supervised learning model that plot the observations in space. This algorithm seeks to define a linear segregation between the classifications. Different kernels are used to help define that line between real and fake jobs. If tuned properly, a test observation is defined based on what side of that line it falls on.

Initial analysis on the data set in this algorithm resulted in all kernels and settings classifying all observations as real job postings. Accuracy was 75%. Following the in-class presentation of this project, the SVM analysis was repeated with different results. In this re-analysis, the SVM kernels provided the following results:

Baseline: 83.55%

Polynomial: 83.57%

Radial: 83.545

Sigmoid: 67.55%

#### 

#### To set the kernel, add the kernel flag in the svm() command.

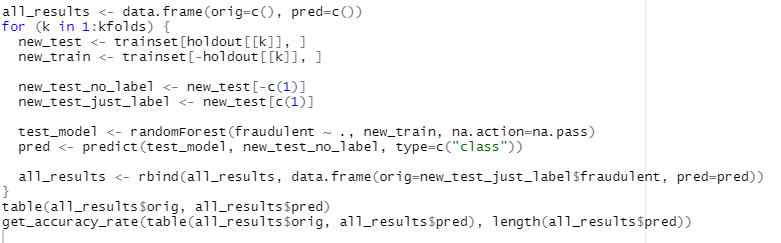
#### Example:

#### Support Vector Machine, Polynomial Kernel Confusion Matrix

#### 

### Random Forest

Random Forest algorithm is an advanced decision tree model that produces multiple trees and classifies a test observation based on the majority class of the observations on the same branch as the test observation. This model leveraging the majority vote is similar to k Nearest Neighbor, but deploys this technique in a decision tree format instead of plotting the observations. The algorithm is tuned by providing the number of trees to form. A loop can be used to test the algorithm across a range of trees and provide the best performing forest. This project identified the best performance coming from a forest of three trees with 83.8% accuracy.



**Random Forest Confusion Matrix:**

Text

Description automatically generated

## Conclusion

Initial review of the data showed variation between the real and the fake job postings that were expected to provide opportunity for machine learning algorithms to properly classify the job postings. Thorough analysis across multiple algorithms revealed top performance from the kNN algorithm with an accuracy rate of 89.5%. This accuracy rate is not significant enough to declare the algorithm as a candidate for professional use in classifying the authenticity of job postings. A candidate algorithm is expected to have an initial accuracy rate near 95%.

One clue to the lack of performance of the algorithms may be viewed in the k Means Clustering. The job postings did not cluster in line with the authenticity of the posting. The clusters spanned both classifications, and in equal proportions because the job postings are too similar. There are variations in the variables used by this study, but they are not strong enough to properly classify all the jobs.

Reviewing word usage within the job description of the posting through word clouds showed some variation in word usage, but this, too, was not viewed as a significant difference. The frequency of certain words did vary between the classification, but a word in the top 20 of one class, is still in the top 50 of the other. More sophisticated language processing algorithms might perform better in classifying the postings.

It is unsettling how similar the fake postings are to the real postings. It is unfortunate that this study did not find an algorithm to solve the problem, but some items were identified that may help the job seeker in keeping themselves, and their personal information safe throughout the application process. It was noted that fraudulent job postings were missing company information at a much higher rate than the real postings. If the applicant is unable to verify the company they are applying to, it is advised to not apply- this may be a fraudulent posting, and if it is a real posting, lack of company information in the posting may be a red flag for the culture of that company regardless of the authenticity of the post.

With the absence of protective algorithms screening job postings, it is recommended to reduce personal information on your resume where possible. A best practice to safeguard your accounts is to use passwords that are not easily guessed- your passwords should not include, or be based on any details that are contained in your resume. Lastly, the applicant should trust their gut- if the job posting, or interview process feels off, it probably is- don’t apply, or end the conversation.

### References

<https://www.fbi.gov/contact-us/field-offices/elpaso/news/press-releases/fbi-warns-cyber-criminals-are-using-fake-job-listings-to-target-applicants-personally-identifiable-information>

<https://www.kaggle.com/datasets/whenamancodes/real-or-fake-jobs>