AUS | الجامعة الأمريكية في الشارقة
American University of Sharjah

Department of Mathematics and Statistics

STA 401: Introduction to Data Mining

Final Project

Classification Model to Predict Fake and Real Job Postings

Submitted by:

Joel D’Souza b00079296

Sabbir Alam b00079438

Submitted to: Dr. Ayman Alzaatreh

Submission Date: 11th May 2020

Table of Contents

Abstract

In our paper, we create a classification model to predict whether a given online job posting is fake, primarily based on its job description. The dataset we use contains 17,880 observations with 17,014 legitimate and 866 fake job postings. We worked on this dataset primarily with RStudio and used several text analysis techniques, such as Tokenization, N-Grams, Term Frequency - Inverse Document Frequency [TF-IDF], and Singular Value Decomposition [SVD] to convert the unstructured job description into workable structured data. We used this structured data to predict whether a given job posting is fake or not utilizing data mining techniques like Random Forest and achieved a test classification accuracy of 97.74% through this method, which is the best accuracy rate for work done on this data set. We also discovered a few trends within the dataset, which can help better understand how fake job postings deviate from legitimate ones.

Keywords: Text Analysis, Text Mining, Fake Jobs, Classification, Random Forest

Introduction

In this day and age with more people using the internet, companies are shifting their job postings from newspapers and magazines to websites specifically designed for this very purpose. However, since some of these websites are unregulated, it is also possible for fake job offerings to be posted on these websites and this is known as Online Recruitment Fraud [ORF]. Researchers Vidros, Kolias, and Kambourakis argue that the implementation of Applicant Tracking Systems [ATS] has enticed scammers to target these job seekers through fraudulent job offerings aimed at stealing personal information[6]. Therefore, there is an ever-increasing need to filter out these fake postings to protect the average jobseeker from getting scammed.

Objective

Through our work on this dataset, we aim to achieve a better test classification accuracy compared to earlier models that have worked on the same dataset. We also try to provide greater insight into what characterizes a fake or real job posting through various graphs and word clouds. Furthermore, we also try to define which variables present within the dataset have the most significance, when it comes to deciding the prediction accuracy of the model.

Literature Review

Upon going through the literature for this paper, we came across a handful of other individuals who had also worked on the same dataset as us. The first of these was a paper by Bandar Alghamdi and Fahad Alharby, published in the Journal of Information Security in 2019.

In their paper, they predicted the response of this dataset using various classification algorithms and achieved their best classification accuracy of 97.41% using Random Forest[1]. They also reported that the company profile, the company logo, and the type of industry the company is in were highly significant in the decision making of the model.

Next, we came across several online kernels that had worked on this dataset on Kaggle, with the most 'liked' one reporting a classification accuracy of 97.65% and was uploaded in April of this year. The kernel was created by user Madz2000 using Python and he achieved this prediction accuracy using the Multinomial Naive Bayes Classifier algorithm[4]. Finally, there was also a 2 part article written by Sharad Jain on this dataset and published on a data science website, that provided good visualizations and achieved an AUC accuracy of 86% using Neural Net[3].

Based on the previous work, we can see that the required prediction accuracy to beat is 97.62%. These earlier works also provided us with some useful insight into the dataset, which helped us better prepare our own model. Furthermore, the article by Vidros, Kolias, and Kambourakis provided greater insight into why such fake job postings are present and helped us to better understand this subject [6].

Dataset Description

The dataset of Real and Fake Job Posting is a modified version of an original publicly available dataset posted to Kaggle by Shivam Bansal [2]. The original dataset was posted by the University of Aegean by its laboratory of Information & Communication Systems Security department [5].

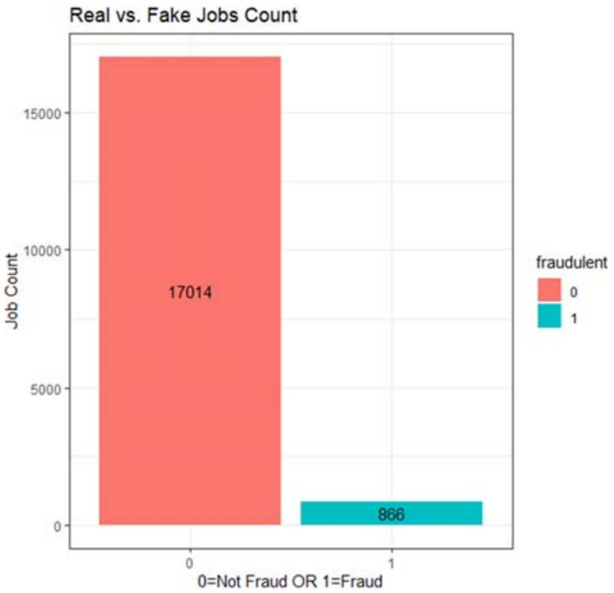
The dataset contains 17,014 legitimate and 866 fraudulent job postings between the years of 2012 and 2014, and contains both textual and meta-information about each job.

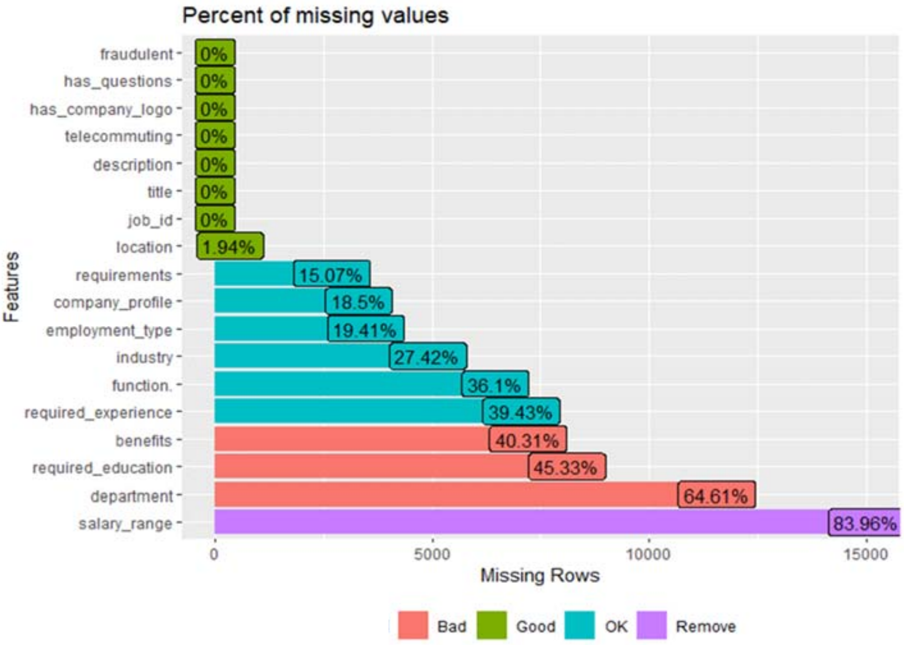
Table 1: Specifications of the dataset

Name	Data Type	Description
job_id	Numerical	A unique ID for each job posting
title	Text	The title of the job ad entry.
location	Text	The geographical location of the job ad.
department	Text	Corporate department (e.g. sales).

salary_range	Range	Indicative salary range (e.g. $50,000-$60,000)
company_profile	Text	A brief company description.
description	Text	The detailed description of the job ad.
requirements	Text	Enlisted requirements for the job opening.
benefits	Text	Enlisted offered benefits by the employer.
telecommuting	Binary	1 if the position requires telecommuting and 0 if it does not
has_company_logo	Binary	1 if the company logo is present and 0 if it is not
has_questions	Binary	1 if the company requires screening questions and 0 if it does not
employment_type	Text - Nominal	The type of employment. Eg. Full-type, Part-time, Contract, etc.
required_experience	Text - Nominal	The experience required for the job. Eg. Executive, Entry level, Intern, etc.
required_education	Text - Nominal	The education level required for the job. Eg. Doctorate, Master’s Degree, Bachelor, etc.
industry	Text - Nominal	The industry that the job listing falls under. Eg. Automotive, IT, Health care, Real estate, etc.
function	Text - Nominal	The function that the job listing falls under. Eg. Consulting, Engineering, Research, Sales, etc
fraudulent	Target - Binary	1 if the job listing is fraudulent and 0 if it is real

Upon analysis of the dataset, we can see that 95.2% of our data are legitimate job postings while the rest 4.8% are fraudulent. This indicates a very high imbalance in our data. Out of the 18 columns present in the data, only 7 of them contain no missing values. The salary_range column has the greatest number of missing values with 83.6% of the column being null. The description, company_profile, benefits, and requirements columns are the only text columns that contain several sentences or paragraphs, making it appropriate for performing text analysis. However, description is the only complete text column from these four with no missing values. Therefore, we decided to concentrate on this column, with a few additions, for our prediction models, but continued to utilize all 4 of these textual columns for visualization and analysis.



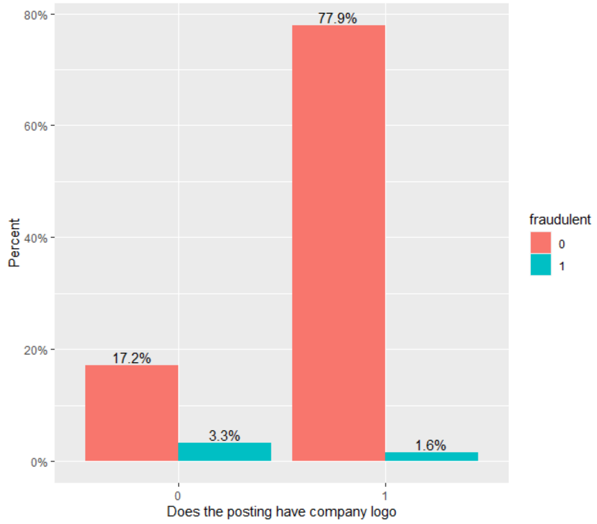


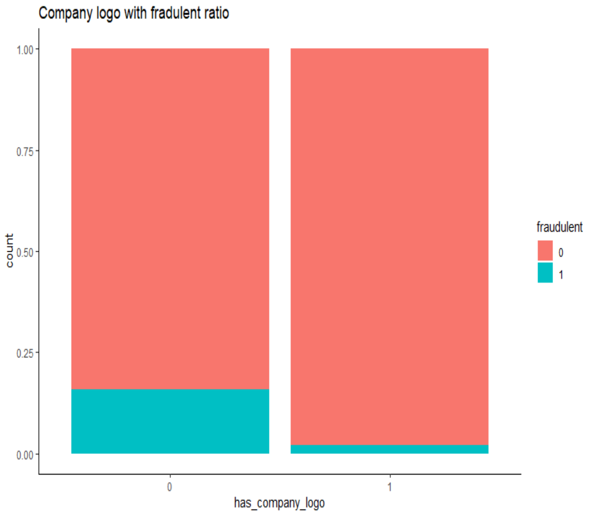
Exploring the Data

We initially performed some descriptive statistics on the model before and after we tokenized our data and noticed some interesting results based on this information.

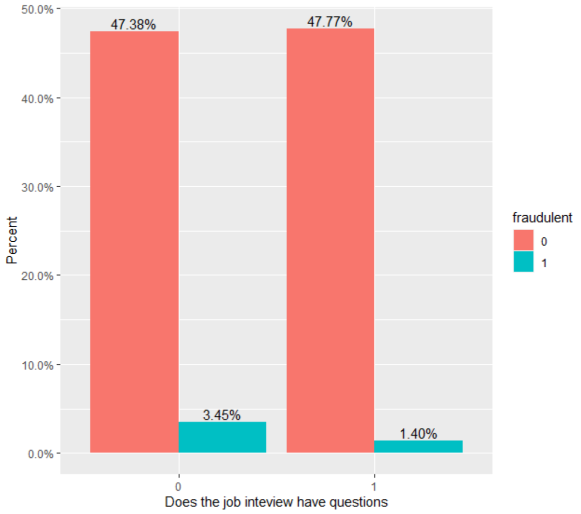
Before Tokenizing

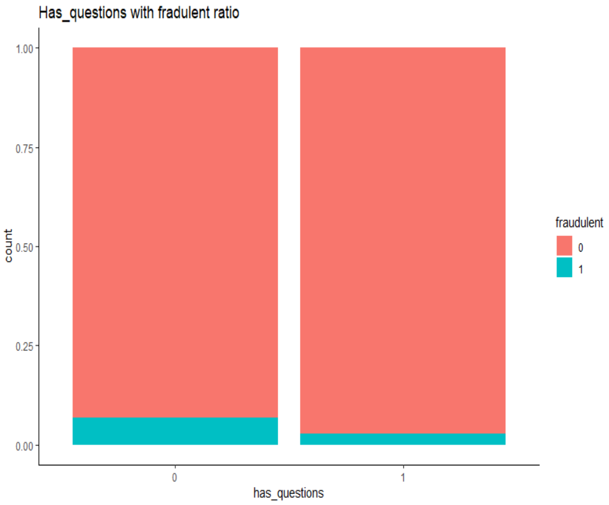
Firstly, we hypothesize that job offers from companies that do not have their company logo are much more likely to be fraudulent, even though such logo-less companies are rare. Out of the 17,880 job postings present within our dataset 3,660 of these were from companies that did not have a logo. However, more than 67% of our fake job postings came from these companies. The percentage of fake to real jobs for companies with no logos is around 16%, compared to a 2% ratio for companies with logos. This ratio indicates that job offerings from companies without a logo are, comparatively, 8 times more likely to be fake.



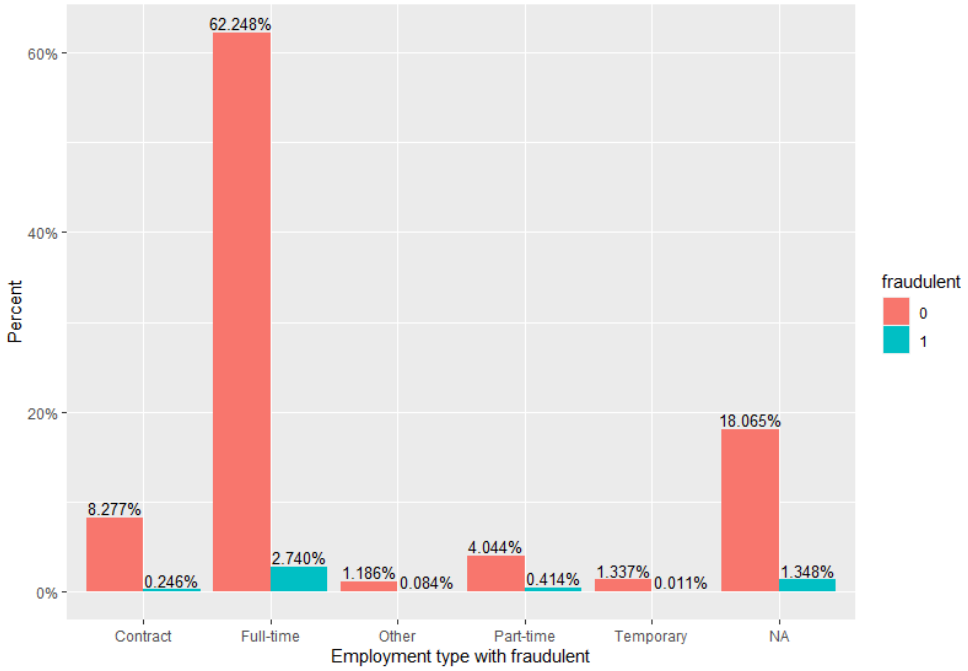


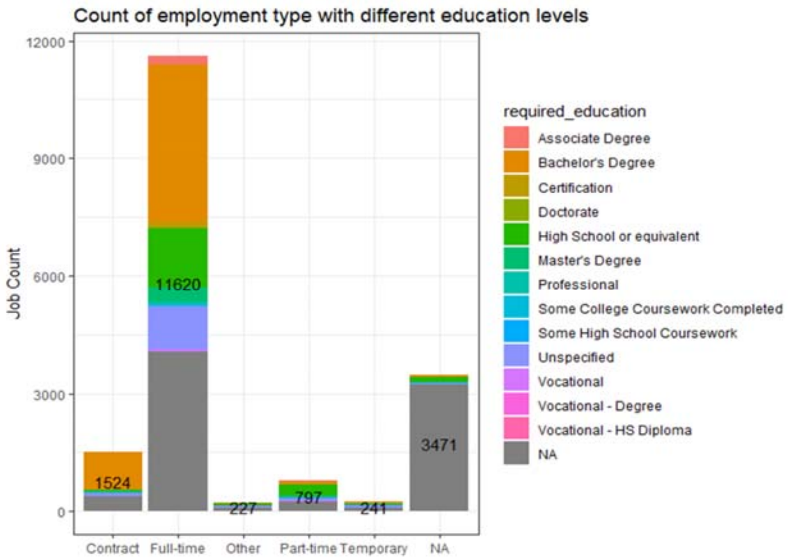
Next, we also hypothesize that job postings without screening questions are slightly more likely to be fake compared to those with screening questions. The ratio of fake to real job offerings were 2.8% when questions were present and 6.8% when there were no screening questions.





Finally, we also learned some more non-predictive features about our dataset, such as most of the job postings are for full-time jobs and that most of the jobs also require a bachelor's degree.



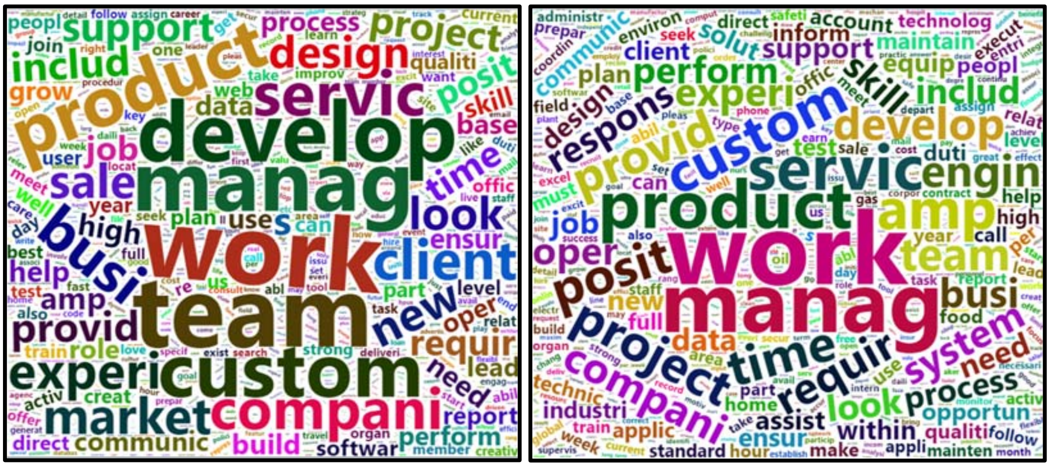


After Tokenization:

Once the textual part of our datasets has been tokenized and stemmed, we created two separate word clouds, one for fraud postings and another for real job postings, for each of the textual columns (description, company_profile, and benefits) and tried to compare the information to see differences in the fraud postings and real postings.

Firstly, we compared the word clouds from the description column of our dataset (shown below). The first trend that stands out is that both fraud and non-fraud descriptions use similar words like work, team, manage, and develop. However, the ratios of some words used are different, for example, team is more often used in non-fraud emails than fraud. We can see that words like system, train, and look are more frequently used in fraudulent emails which could imply that if a job description has those words in it, it may be fraudulent.

Description

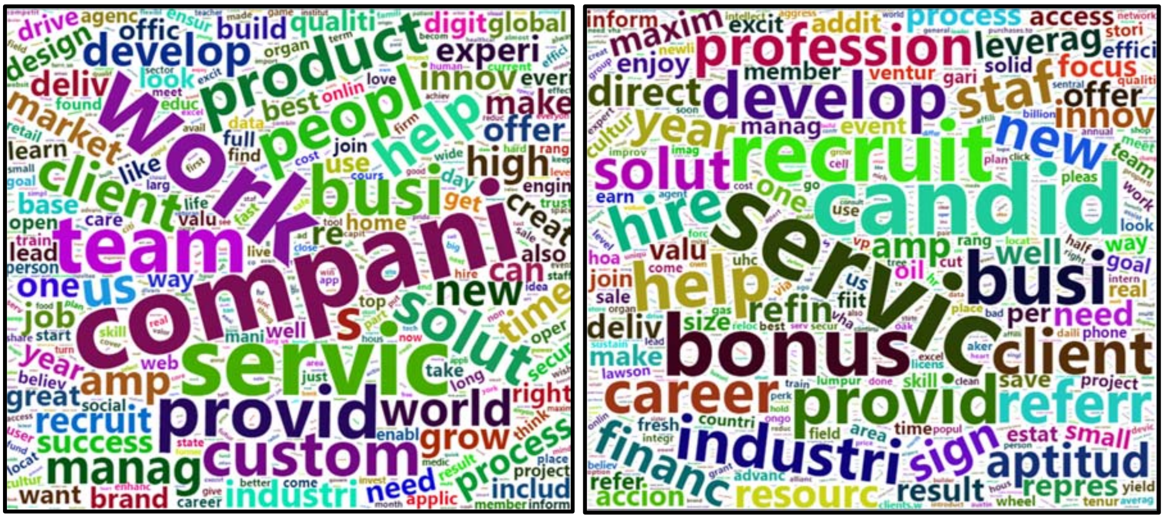


Non-Fraud

Fraud

Next, we compared fraudulent and real emails for the company profile column. Unlike the description section, the company_profile section is slightly more differentiated in their word clouds for real and fraudulent. Stemmed words like compani, custom, brand are more frequently used in non-fraud listings. Other stemmed words like candid, recruit, hire, maxim, profession, and bonus could be signifiers that a company's listing is fake as they are much more frequently used in those types of postings. The use of these words could show that fraudulent emails target job seekers by advertising their company as being more open to recruitment as they used words like "recruit and hire". On the other hand, real companies were more focused on explaining their company in this section rather than focusing specifically on the job.

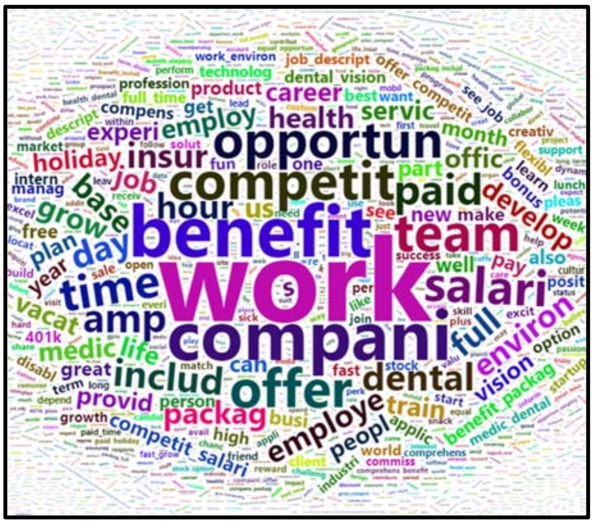
Company profile



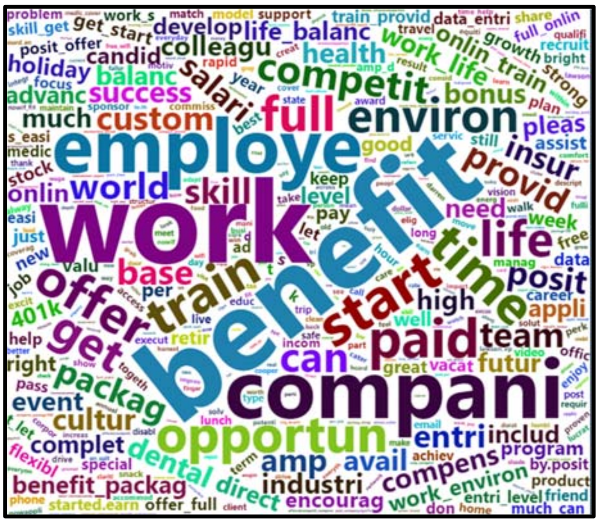
Non-Fraud Fraud

Finally, we analyzed the benefits section of our dataset (shown below) and discovered that words like offer, flexible, start, success, work_life, and online_train are much more used in fraud emails as compared to non-fraud. Furthermore, words like employ, hour, job are more frequently used in real job listings. We could infer that fraudulent job postings are trying to attract job seekers by using positively toned words which could make it seem like the fake jobs have greater benefits than real jobs which use more of a formal tone to describe the job description.

Benefits



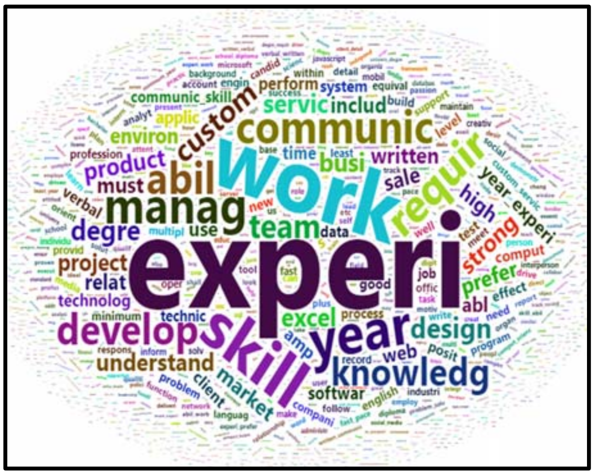
Non-Fraud



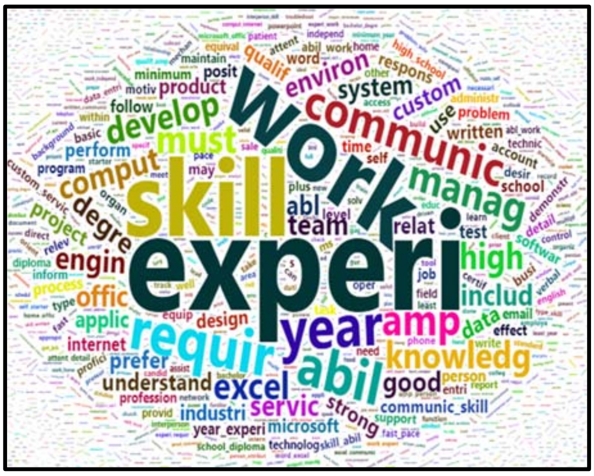
Fraud

The requirements word clouds do not seem to have many differentiating features, although we can say that words such as skill and abil (ability/abilities) occur more frequently in fake job descriptions.

Requirement



Non-Fraud



Fraud

Methodology

<u>Transforming the Data:</u>

We initially split the dataset into 70% train and 30% test by performing a stratified sample with the response and continued to work on the train and test data separately. In order to make the training dataset appropriate for text mining, we had to first clean and tokenize the words present within the description, company_profile, requirements, and benefits columns. We initially tokenized and worked selectively on the description column since it was the only column with several sentences per cell that had no missing values. Since every job must have a job description, we also thought it would keep our model generalized to work on other datasets with similar characteristics. In the end, using only the description for our text analysis part proved to be the best model, giving us our best test classification accuracy. Any other combination of these four columns led to a decrease in test accuracy

The tokenizing process involved some cleaning of the data by removing symbols, numbers, and punctuation. However, we noticed that some special characters such as â, ô, and î were predominantly present within the data even after cleaning, therefore we had to manually clean these characters out ourselves in MS Excel. We then converted all the words into lowercase, removed all the unnecessary stopwords, and stemmed the data to combine all the synonyms. Finally, we tokenized the data through the addition of bi-grams, which led to the creation of two words per token.

After tokenizing the description column, we converted it into a Document-Frequency Matrix (DFM). A DFM is a table in which the rows are the document names (in this case it is Job ID) while the columns are the individual tokens. Each cell contains the frequency of tokens in a particular document. Converting the tokenized data into a DFM allowed us to perform TF-IDF (Term Frequency - Inverse Document Frequency) on our data. The Term Frequency normalizes the frequency of observations while the Inverse Document Frequency penalizes tokens that appear frequently throughout all the documents.

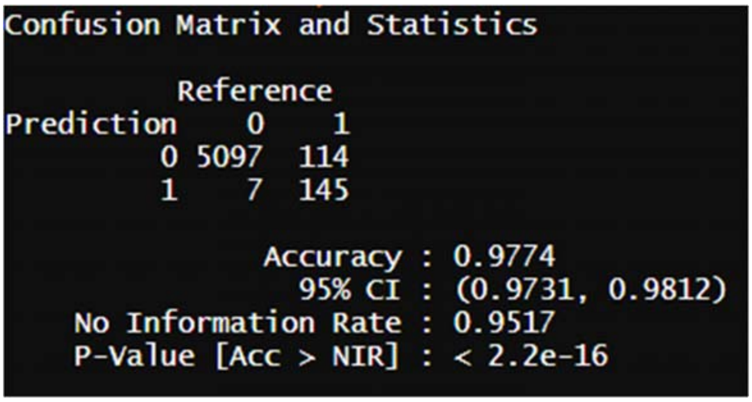
Finally, we used Singular Value Decomposition to reduce the dimensions of our DFM making it feasible to efficiently perform data mining techniques on the data, without much loss of information. We found that reducing the dimensionality of our columns to 300 variables was appropriate for our data. We then selectively appended columns from the original dataset to our SVD matrix, except for company_profile, requirements, and benefits to create our final dataset. We also decided to drop the salary_range model due to its high number of missing values.

Training and Testing the Model:

Having finally reached a dataset that we can perform data mining techniques upon, we utilized classification and prediction techniques such as Random Forest to train and test the accuracy of our model. We created and refined our final test dataset similar to the way we did on our training data, except that we performed SVD Projection on the test data, to project it into the training space, such that it would resemble our training. We also imputed the missing values present within both our test and training datasets.

The Random Forest gave us our best results so far, with a test prediction accuracy of 97.74%. In order to train the model, we utilized the SVD matrix along with the columns function, has_questions, has_company_logo, telecommuting, employment_type, required_experience, and required_education appended to it as factors. According to Alghamdi and Alharby, the company_profile column was considered to be significant, therefore to account for this, we had appended another column called cprof that contained 1 or 0 depending on whether the company profile was present or not. This led to a decrease in the test accuracy by 0.01 % and hence we dropped this column, however, it can still be considered significant as it increased the sensitivity of our model but reduced its specificity. The rest of the columns could not be made as a factor and hence they were dropped as well for the Random Forest.

The model gave us a very high sensitivity of 99.86% but a low specificity of 55.98%, which does indicate further room for improvement. Similar results for the high sensitivity and low specificity were also present within the model done by Alghamdi and Alharby as well as the model created by Madz2000 [1] [4]. Furthermore, according to this model, the variables function and company_logo were the most important within our prediction model.

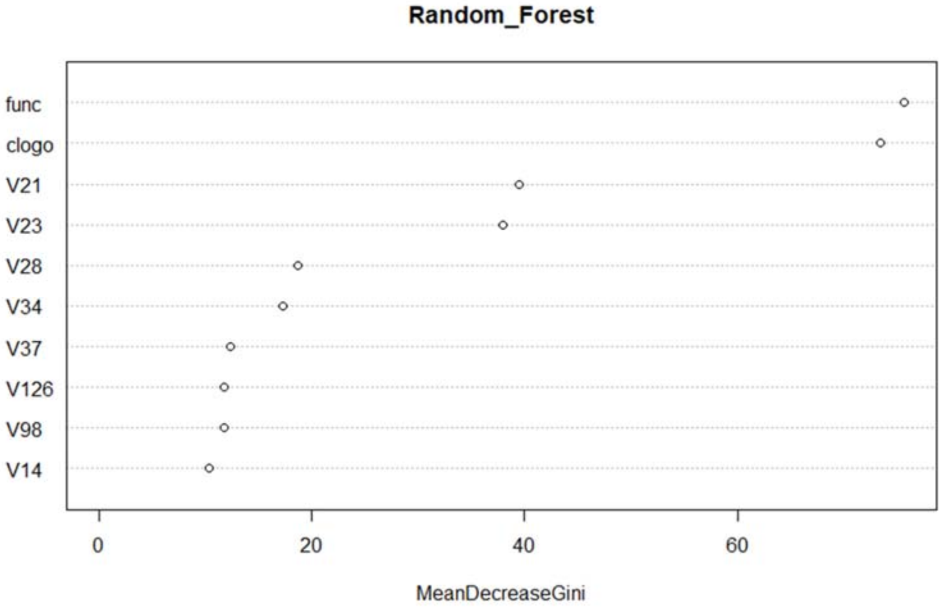


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5097  114
         1    7  145

               Accuracy : 0.9774
                 95% CI : (0.9731, 0.9812)
    No Information Rate : 0.9517
    P-Value [Acc > NIR] : < 2.2e-16
```

Prediction Accuracy of our regression model



Variable Importance Plot

Conclusions and Limitations

We can arrive at a handful of conclusions based on our prediction and analysis model. Firstly, we can see that the description of a job provides significant predictive power in deciding whether a given job posting is fake or not. Furthermore, the function of the job and our initially hypothesized, company logo, provides more information that can help in decision making if the model. The data visualization also indicated that there are noticeable trends that can help differentiate between fake and real company profiles, job descriptions, and job benefits. Finally, we can also see that our earlier hypotheses of a company having screening questions playing a role in the prediction results, turned out to not be true for this dataset.

However, there are also a few limitations that come with our paper. Firstly, there is an imbalance in the ratio of fraudulent to legitimate job postings, which may have been more significant if our test sample size was much bigger. We could have also possibly improved our test accuracy by integrating Rare Event Modelling techniques into our model. The comparatively low specificity of our model is also an area in need of improvement. Finally, even though we achieved our goal of obtaining the highest test accuracy, we only utilized the Random Forest algorithm for our predictions and variable importance and hence we cannot say for certain that the test accuracy we obtained is the best possible one.

Further research could also possibly improve on these shortcomings of our model based on our recommendations. Other text analysis techniques such as sentiment analysis could be used to provide more insightful information on the dataset while also trying to improve the prediction accuracy of the model.

REFERENCES

[1] Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security*, *10*(03), 155-176. doi:10.4236/jis.2019.103009

[2] Bansal, S. (2020, February 29). [Real or fake] fake JobPosting prediction. Retrieved from https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction

[3] Jain, S. (2020, April 26). Predicting fake job postings — (Predictive analysis). Retrieved from https://towardsdatascience.com/predicting-fake-job-postings-part-2-predictive-analysis-3119ba570c35

[4] Madz2000. (2020, April 6). Text classification using Keras/NB(97% accuracy). Retrieved from https://www.kaggle.com/madz2000/text-classification-using-keras-nb-97-accuracy

[5] University of the Aegean, Laboratory of Information & Communication Systems Security. (n.d.). Employment scam Aegean dataset. Retrieved from https://emscad.samos.aegean.gr/

[6] Vidros, S., Kolias, C., & Kambourakis, G. (2016). Online recruitment services: Another playground for fraudsters. *Computer Fraud & Security*, *2016*(3), 8-13. doi:10.1016/s1361-3723(16)30025-2