Stat597: Data Wrangling and Husbandry
Fake Job Posting Data Analysis

Ajinkya Patankar (Netid: - aap256)

Department of Computer Science

Rutgers University, New Brunswick

https://github.com/ajinkyapatankar/Data-Wrangling-Final-Project

Abstract

Fake jobs are posted for various reasons. For example, to trick job seekers (by cyber criminals), to assess available talent pool (by hiring managers), spam email distribution (by random people), unfair hiring practices (by companies) and so on. This leads to employment scams, loss of private information of applicants, resume plagiarism, company's financial and reputation loss among others. This project can be used to overcome these by immediate identification and removal of such jobs from job portals.

Introduction

Job Market

The job market is the market in which employers search for employees and employees search for jobs. The job market is not a physical place as much as a concept demonstrating the competition and interplay between different labor forces. It is also known as the labor market. The job market can grow or shrink depending on the demand for labor and the available supply of workers within the

overall economy. Other factors which impact the market are the needs of a specific industry, the need for a particular education level or skill set and required job functions.

Fake jobs in Market

Fake job posts are everywhere, and they always will be. As a job seeker, your anger is justified. Fake job postings exist so that employers can gauge the current talent pool. The number of applications can be a valuable pointer on what to pay a person for a given job. It also is an indicator of how easy that person is to potentially replace.

Project Description

In this project, we will be using a Kaggle dataset that consists of job descriptions and their meta-information for 18k job postings. A small proportion of these descriptions are fake or scam which can be identified by the column "fraudulent".

Part 1: Dataset

The data consists of the following features: Unique job ID, title of the job ad entry, geographical location of the job ad, corporate department (e.g. sales), salary range (e.g. $50,000-60,000), a brief company description, detailed description of the job ad, enlisted requirements for the job opening, enlisted offered benefits by the employer, position to telecommute, presence of company logo, employment type (Full-type, Part-time, Contract, etc.), required experience (Executive, Entry level, Intern, etc.), required education (Doctorate, Master’s Degree, Bachelor, etc.), industry (Automotive, IT, Health care, Real estate, etc.), function (Consulting, Engineering, Research, Sales etc.), fraudulent or not.

Part 2: Knowledge Discovery from Data (KDD)

Knowledge Discovery from Databases (KDD) can be defined as a recognition and extraction of valuable, genuine, useful, unique, and comprehensible correlations or patterns in the data. The KDD model includes five main phases: Selection of relevant prior knowledge; Acquiring or creating targeted dataset; Preprocessing to handle missing values, noise and errors in the data; Transformation to create dataset form suitable for easily implementing data mining algorithms; Data mining, the decision making activity to define models such as regression, classification or clustering to obtain patterns of interest, representational form, or rule sets and trees; and Interpretation and Evaluation with respect their validity, and visualization of the patterns and models.

Part 3: Data Mining Tools

1. **ggplot**: ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics.
2. **sapply/ lapply**: sapply() function takes list, vector or data frame as input and gives output in vector or matrix.
3. **Corpus**: Corpora are collections of documents containing (natural language) text
4. **tm library**: A framework for text mining applications within R.
5. **VCorpus**: VCorpus in tm refers to "Volatile" corpus which means that the corpus is stored in memory and would be destroyed when the R object containing it is destroyed.
6. **tm_map**: Interface to apply transformation functions (also denoted as mappings) to corpora.
7. **Term-Document Matrix**: Constructs or coerces to a term-document matrix or a document-term matrix.
8. **Document-Term Matrix**: Many existing text mining datasets are in the form of a Document-Term Matrix class (from the tm package).

Part 4: Analysis and Visualisations

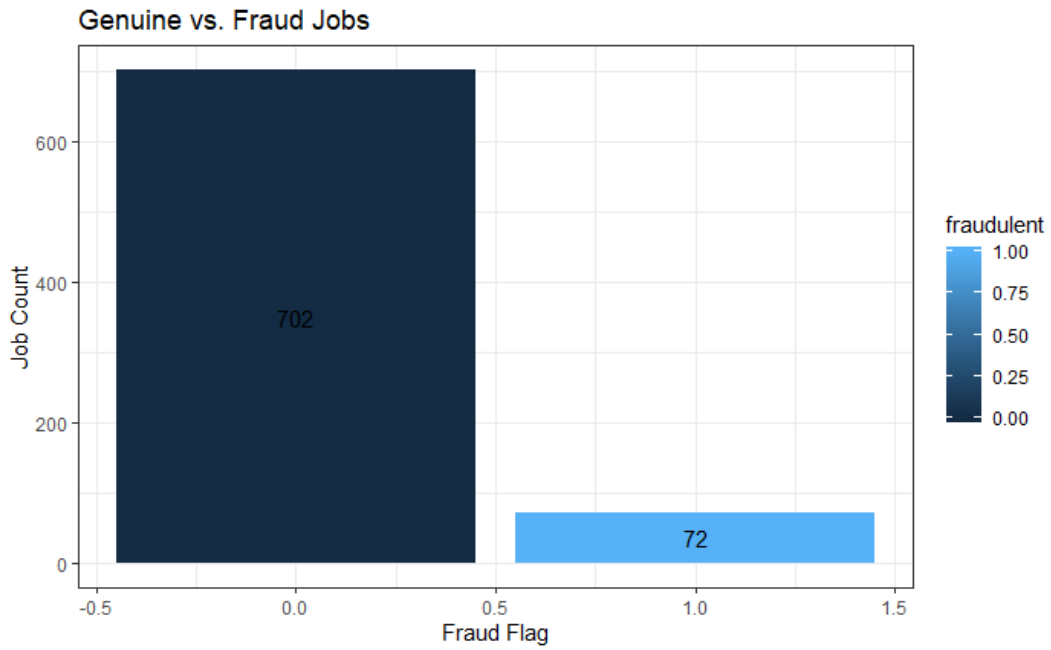

Fig 2: Genuine Vs Fraud Job Postings

In the above plot, jobs types (fraud or genuine) are plotted on the x-axis, while the count of jobs is plotted on y-axis. We can observe that this is a very unbalanced data with only 866 postings as fraud results, while 17014 jobs are genuine.

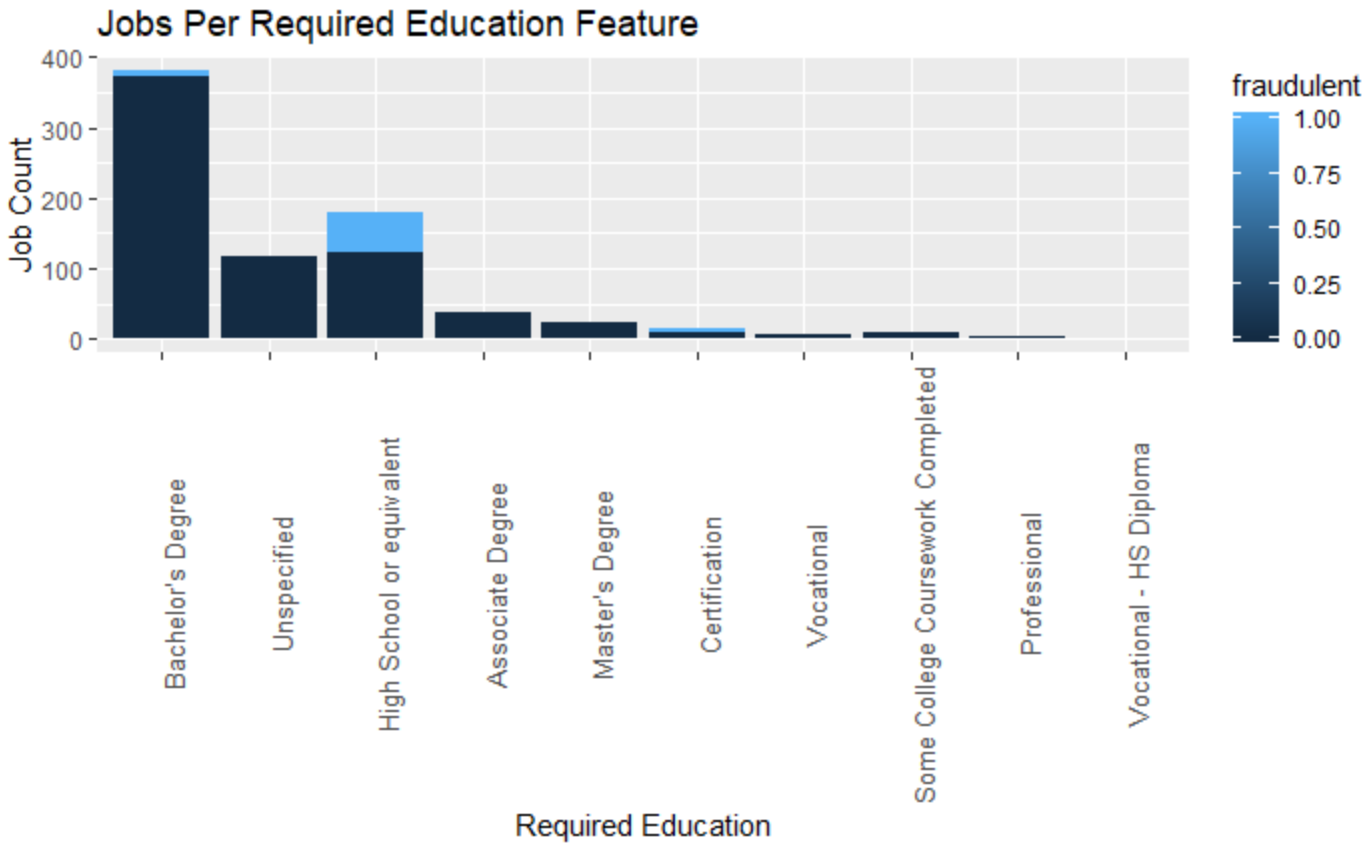

Fig 3: Jobs per Required Education Feature

In the above plot, the kind of education required has been plotted against the count of jobs. We can observe that majority of jobs, both fake and genuine, require a bachelor's degree and vocational degree, while vocational degree requirements are required the least.

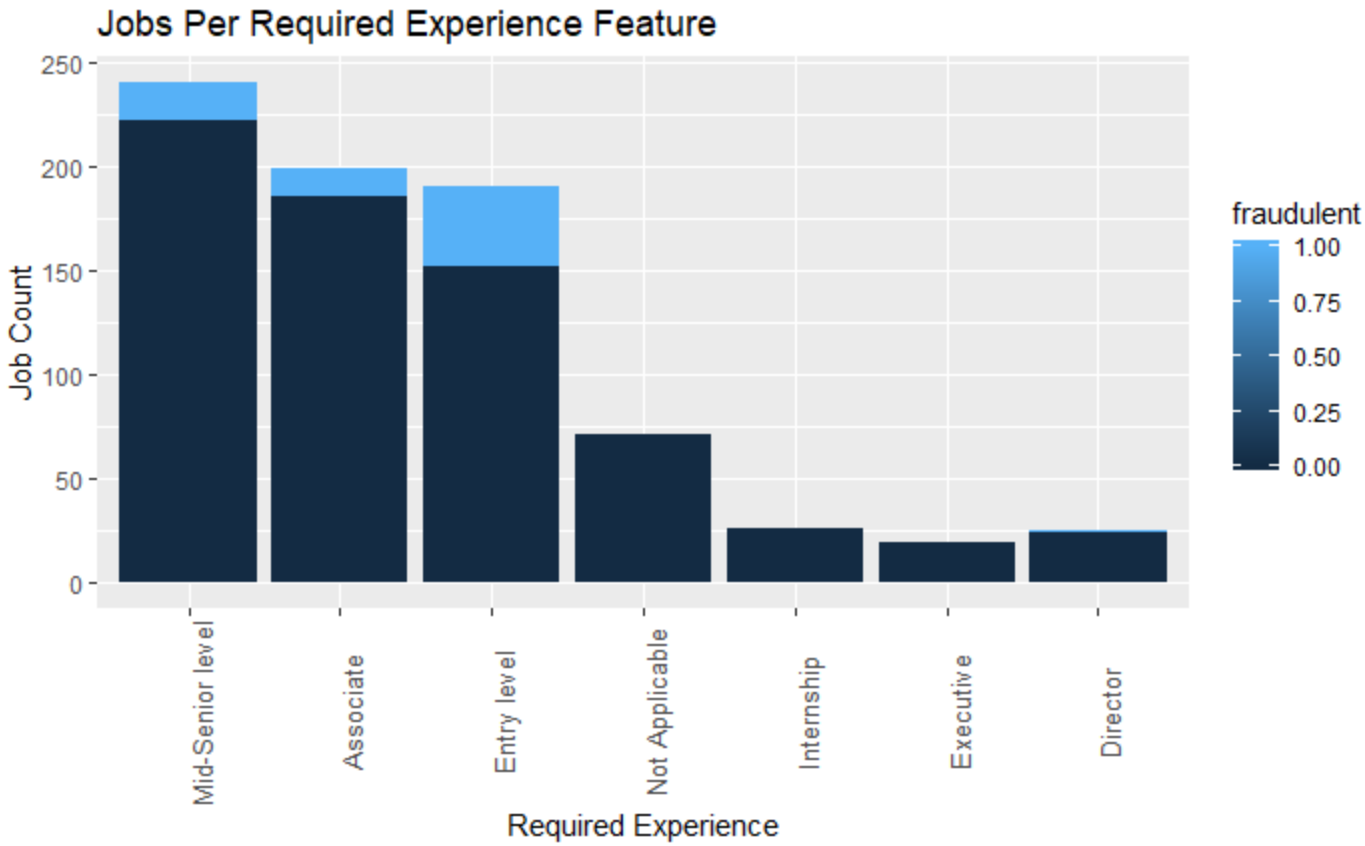

Fig 4: Jobs per Required Experience Feature

In the above plot, the kind of experience required has been plotted against the count of jobs. We can observe that for majority of jobs, both fake and genuine, job posters don't mention the kind of experience they need for the job which creates a lot of confusion, and least job posts are for the executive posts.

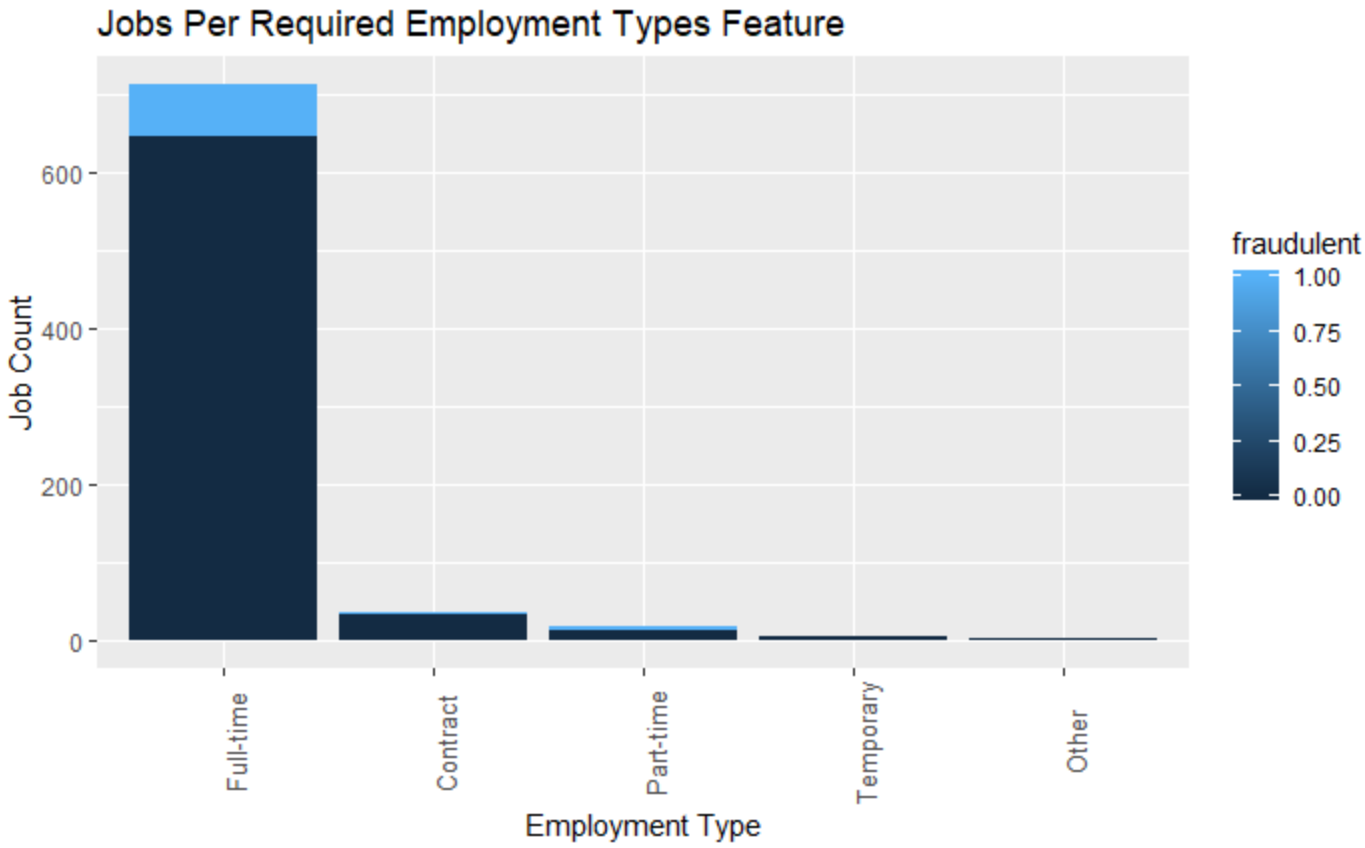

Fig 5: Jobs per Required Employment Feature

In the above plot, the kind of employment (e.g. Full-time or Contract) required has been plotted against the count of jobs. We can observe that majority of jobs, both fake and genuine, are for Full-time requirements, while others i.e. are required the least.

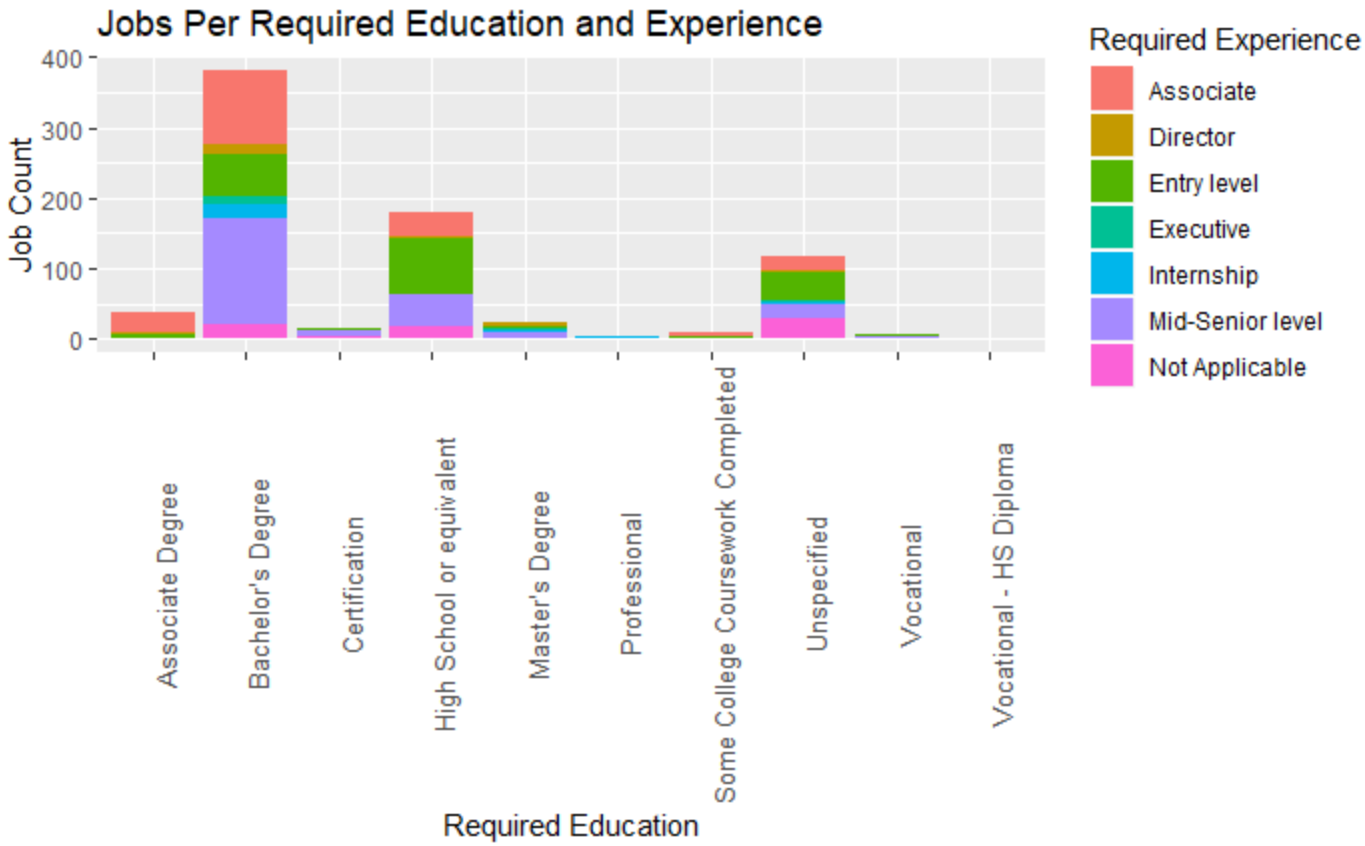

Fig 6: Jobs per Required Education and Experience Feature

In the above plot, the kind of experience required, and education required has been plotted against the count of jobs. We can observe that majority of jobs, are for bachelor's degree, in the internship level while vocational degree requirements are required the least.

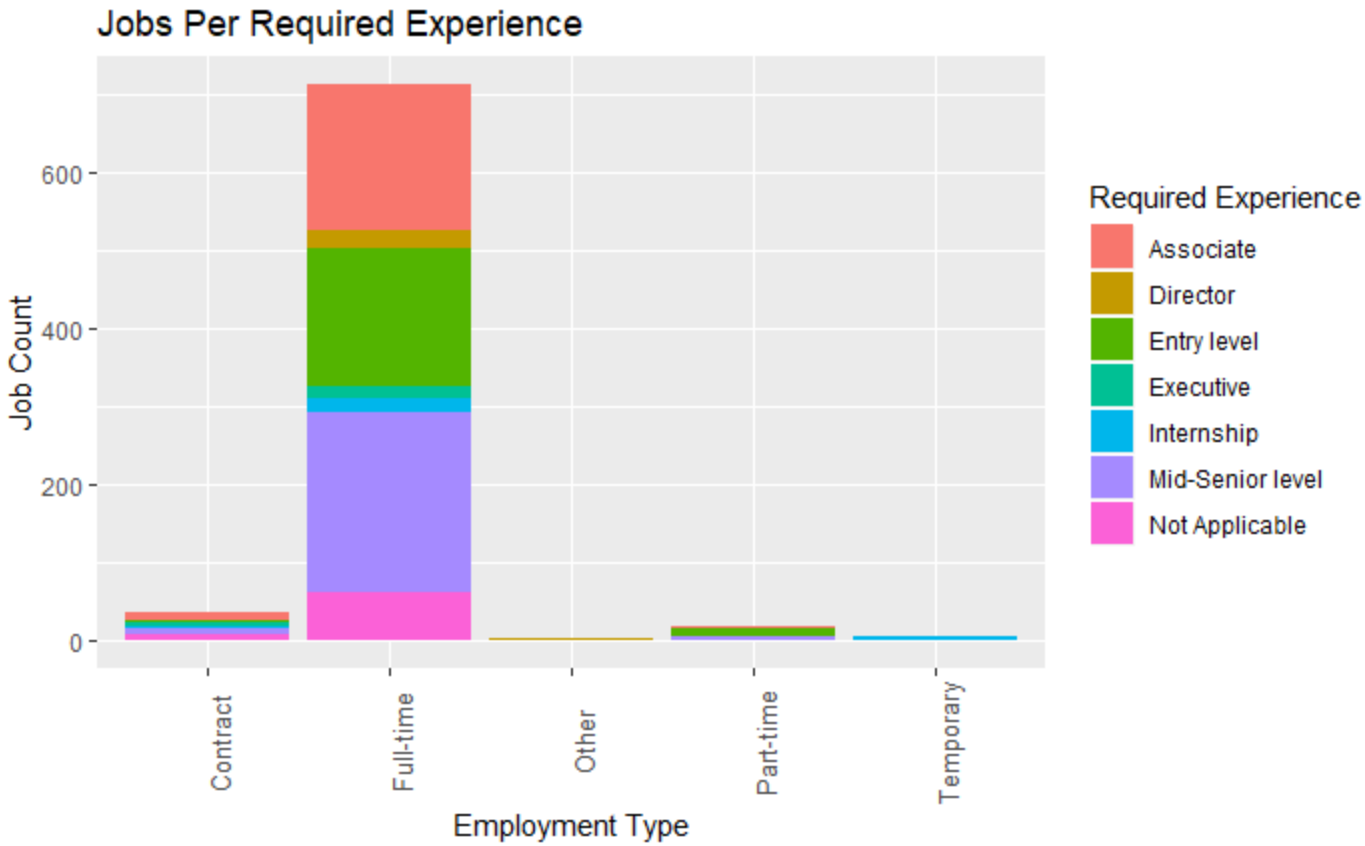

Fig 7: Jobs per Required Employment and Experience Feature

In the above plot, the kind of employment required, and experience required has been plotted against the count of jobs. We can observe that majority of jobs, are for full-time again, in the internship level while temporary and others are required the least.

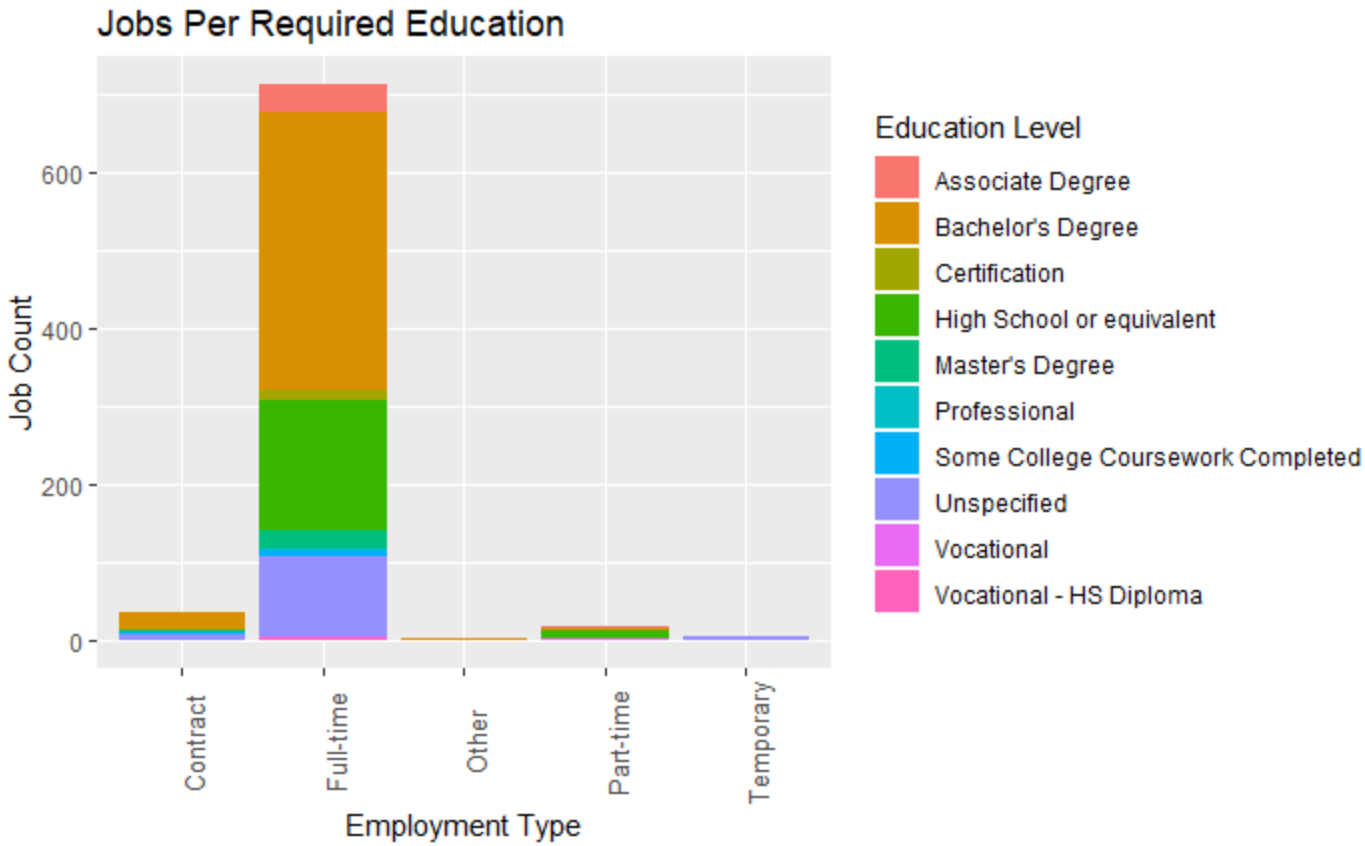

Fig 8: Jobs per Required Employment and Education Feature

In the above plot, the kind of employment required, and education required has been plotted against the count of jobs. We can observe that majority of jobs, are for full-time again, in the Bachelor level of education while temporary and others are required the least.

Fig 9(a): Word-cloud: Job Requirement

Fig 9(b): Word-cloud: Job Description

Here, two-word clouds have been created job postings requirements and descriptions. One difference which we can see from the two word-clouds is that requirements and job descriptions are pretty much like each other, contrary to the popular belief that they are not. Most of the words in the description and requirements have the same count and, many a times, the sentences are repeated more often. Most of these advertisements use wordplay very well.

Conclusion

This is the biggest giveaway in these scams that a job being offered without any screening process. In most cases, a job offer is followed by asking for confidential information like bank or credit card details, which is the primary way these fake job postings scam you. These scammers usually target seekers for entry-level jobs. These are mostly fresh college graduates who are looking for their first role in the industry, and the current economic situation, the easiest targets for the scammers. As we go down the order, the seniority level increases, which indicates that as an individual gains more experience, they can differentiate between the fake and genuine posts.

References

https://ggplot2.tidyverse.org/

https://www.rdocumentation.org/packages/tm/versions/0.7-7/topics/Corpus

https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

https://www.rdocumentation.org/packages/tm/versions/0.7-7

https://www.youtube.com/watch?v=dE10fBCDWQc

https://www.rdocumentation.org/packages/tm/versions/0.7-7/topics/TermDocumentMatrix

http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

https://www.r-graph-gallery.com/