Predicting Fake Job Advertisements

Bharath Janapareddi, Kavya Nagaraju, Priyanka Patil, Tejaswi Gundapaneni, Vidhi Shah

Introduction

In the era of digitalization, most of the recruitment process begins online. This has helped to reduce the hiring cost as well as to increase the reach to a potential job seeker. This makes job listing websites the go-to place for both employers as well as job seekers. This brings along a high cost of maintaining brand equity and viewership. A trusted website can easily lose its reputation due to poor customer satisfaction. One of the key factors for customer satisfaction on an online job portal is not being scammed while searching for a job. Generally, 65% of the consumers are more likely to avoid using a website with fake content.

Challenges were faced in calculating the brand value/brand equity. Therefore, an assumption was made that revenue per view reduction has a direct relation to the brand value. Negative review reach was calculated along with the positive review reach. The average price per ad per viewer was then used to calculate the total revenue impact because of the Fall in brand equity.

A job seeker usually spends 11-15 hours per week looking for a job. On average a person applies to 10 job applications per week. This can result in a critical loss of time for a job seeker if they fall prey to fake advertising. We see from our dataset that approximately 5% of the advertisements are fraudulent. This means it can cost a job seeker an hour of his week.

As part of this project, we tried to analyze the cost of fake ads to brand equity. Recommendations were provided to improve customer satisfaction by safeguarding job seeker's information and saving their time.

Data Cleansing

The Fake job postings dataset was taken from Kaggle[1]. For confidentiality purposes, the company name is masked in the Dataset/ Data source. The dataset has 17,880 records with 18 variables and is in a .csv. file format. As part of Data Cleaning, R studio, Tableau, and Excel software were used. There were lots of missing values that indicated there was no entry from the user posting the advertisement, hence they were replaced with "Not Mentioned". R Studio was used to clean the variables - Salary Range, Company Profile, Requirements, Benefits, Employment Type, Required Experience, Required Education, Industry, and Function. A new column was created in excel to make the Salary variable an integer with the max value from the range specified. Job ID was a numeric field and was not adding any value to the analysis; hence it was removed. For the usage of geo filter appropriately in Tableau and Data studio, the variable Location was split into Country, State, and City. The fraudulent column had 0,1 values that were not working as Boolean values in Tableau, hence new field was created with TRUE/FALSE for insight.

Project Approach

As our first step, we explored the data and provided insights about variables that can identify a fake ad. The identified fake ads will raise red flags which will help the company's internal team to run a review. We used text analytics on a couple of descriptive fields to identify high-frequency words used for fake job ads. Decision Tree Model helped in identifying the probability of an ad being fake given a missing variable field.

Project Analysis and Key Findings

The job listing companies must identify fake job ads to retain their website's reputation. It was observed 4.84% of the listings are fake. Even though this might look insignificant for this set of data but considering the bigger picture, it can hamper the Brand Equity. Also leading to a negative impact on viewership. Our initial analysis using geo maps showed that the majority of fake ads are from the USA. On checking further, we identified 20% of the fake ads were posted for California location followed by New York and Texas.

The company logo had an impact on determining fake ads. Out of a total of 866 fake ads, 583 did not have a logo which sums up to 67.32%. Thus, having a logo("Has_company_logo") in job ads would play a vital role, hence it should be made mandatory.

We noticed that pre-screening questions ("Has_questions") on the job ads have a considerable influence on identifying fake ads. Out of 866 fake ads, 616 ads did not ask questions which is equal to 71.3%. Therefore, we recommend our audience (job posting website) to encourage their clients to ask questions on their ads. However, if the ad is still posted without questions, an internal team should review the authenticity of the ad before publishing it.

"Description" and "Benefits" fields in the dataset had unstructured blob data. Text analytics was performed using R. Word-clouds were used to display high-frequency words in fake job ads. Further in the analysis process, a relation between these words with other variables in the dataset was identified. For instance, the word cloud generated from the description field had "Darren", "Lawson" as the high-frequency words. These words were a part of the description under the company Aptitude staffing which was listed with a location in California. By looking at the observation it was found this ad had multiple spelling errors and no link to the website. It was also found from the benefits word-cloud, "Aker" refers to a staffing company and "subsea" refers to Oil and Energy industry. Therefore, a recommendation was made to run text analytics on the ad's descriptive fields, to highlight the potential fake ads.

It was found 36.76% of fake ads did not have anything listed under industry type. Therefore, a recommendation was made to make industry type a mandatory field. Another key recommendation was to run text analytics on Industry and generate high-frequency words to see a possible relationship with other fields of fake ads. It was identified majority of the fake ads were posted from Oil & Energy and Accounting industries. Deep diving into these industries it was identified same ads were posted multiple times by same companies (example "Aptitude staffing solutions" – 35, "Aker solutions" – 31), and these companies have been the regular publishers of fake ads. These companies should be scrutinized and blacklisted from the website.

Decision Tree Model using “Has_Company_Logo”,” Has questions”,” Required Experience”,” Required Education” variable was modeled. This was built to predict the probability of fraud ads based on the information provided in these variables. The results obtained from this model helped to back our recommendations above.

1) If the company logo was mentioned, then there is only a 2% probability of an ad being fraudulent.
2) If the company logo and employment type are not mentioned, then there is a 27% probability of an ad being fraudulent.
3) If the company logo, employment type is not mentioned and has no questions asked, then there is a 37% probability of an ad being fraudulent.

Conclusion

Considering the above recommendations, the job listing website can maintain and improve brand value by filtering fake ads. Developing a brand image as a trusted website will help the company to improve viewership. This will ultimately result in an increase in popularity leading to command a price premium. Employers should be encouraged to add the details in all the mandatory fields. A complete ad attracts more resumes as a rational candidate ousts out an incomplete ad suspecting it to be fake. On the other hand, this will result in improved customer satisfaction index which would, in turn, attract more resumes and viewership. Reliability being a major concern these days, some of the above suggestions could help retain confidential information without exposing it to scammers.