Presentation Summary

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

Slide1 :

Title Page

Slide2 : About the Dataset

For the confidentiality purpose, company name is masked in the Dataset/ Data source.

We first identified the problem.
Developed a plan of how we will analyze and derive conclusions.

1. Data cleaning was the first step – Software's R, Tableau and Excel were used.
2. Care – Secondary reading was done on the loss incurred.
3. Do – Tableau and Data studio was used.
4. Impact – R was used, and a Logistic regression model was made to make predictions.

Slide3 : Data Cleansing

Data cleaning was performed on above variables.

1. New column was created to remove the effect of Salary range.
2. Missing values indicated no entry from the user posting the advertisement, hence they were replaced with "Not Mentioned"
3. Job ID was a numeric field which was not adding any value to our predictive model; hence it was removed.
4. Fraudulent had 0,1 values which were not working as Boolean values in Tableau, hence new field was created with TRUE/FALSE.
5. Description, Requirements, Benefits variables had unstructured blob data, hence text analytics and data mining techniques were performed.
6. To use the geo filter correctly in Tableau the variable Country, state and city was split into individual columns.

Slide4 : Care

"41% of viewership will be lost to 5% of fake ads accounting to an annual revenue loss of approximately $2.1M"

Our Audience is the Job listing website.

We first ran a quick EDA to check the number of Fake ads in the dataset. For better understanding we took a percentage.

Even though the 5% might look small, but if it interacts with lot of users it can damage a websites reputation and reviews.

Hence it is essential to do something.

Calculation -

A person spends 11 hours a week, spends 1 hour per job application approximately. Each corporate ad attracts 250 resumes.

Nearly two-thirds(65%) of the consumers would stop using the website, if they see false content.

Total ads	17880	
Average resume per ad	250	
Total resume	4,470,000	
Fake ads	866	
resume sent to fake ads	200	(assuming 50 people will be rational and would verify before applying)
Total resume sent to fake ads	173200	
WOM of negative review	10	
Total reach of negative review	1,732,000	
Real ads	17014	
Average resume per ad	250	
Total real ad resume	4,253,500	
only 1/10 of customer leave positive review	425,350	
Total reach of positive review	4,253,500	Assumed 1 positive review will effect 10 positive
Negative ratio	41%	
	705,260	
Average ad price	64.2	
Average ad price per viewer	0.2568	
Total revenue loss	181,110.80	
Revenue loss Annually	2,173,329.64	

Source :

https://www.globenewswire.com/news-release/2019/06/17/1869379/0/en/Study-Consumers-Reject-Brands-That-Advertise-on-Fake-News-and-Objectionable-Content-Online.html
https://www.inc.com/andrew-thomas/the-hidden-ratio-that-could-make-or-break-your-company.html
https://www.smartjobboard.com/blog/monetize-your-job-board-website/
https://goingclear.com/how-much-advertising-revenue-will-website-make/
https://learn.g2.com/customer-reviews-statistics
https://www.flexjobs.com/blog/post/common-job-search-scams-how-to-protect-yourself-v2/
https://www.globenewswire.com/news-release/2019/06/17/1869379/0/en/Study-Consumers-Reject-Brands-That-Advertise-on-Fake-News-and-Objectionable-Content-Online.html

Slide5 : Which location has majority of fraud Ads ?

In the superficial EDA we saw majority of the fake job postings are from USA. Hence, we decided to conduct in-depth analysis on USA. Since the company is masked, it is a possible that the Job listing website is having majority of the ad listed for USA only. We analyzed fake ads and found out that majority of them are posted from Texas -152 ads. (Texas was chosen over NY, because count of fakes ads was highest) But this could be because the total number of ads posted are more from Texas as compared to other states. Hence it was important to take a percentage of it to see which state is leading. And as we can see 20% of the ads posted for California(143) are fake.

Slide 6 : What is the effect of company logo and pre-screening questions on job ads ?

It was observed out of 866 fake ads, we see that 583 ads of the total fake ads do not have a logo, which is equal to 67.32%. And it equals to 3.26% of the total ads posted. “If we recall total of 4.85% of the total ads are fake”. Thus, having a logo in job ads would play a vital role for Job seekers as they will not be scammed and for job posters too, as rational candidates will not oust it out thinking it is a fake ad. Therefore, company logo should be made mandatory. It was also observed out of 866 fake ads, we see 616 ads of the total fake ads do not ask questions, which is equal to 71.13%. And it equals to 3.44% of the total ads posted. Thus, we recommend our audience (job posting website) to encourage their clients to ask questions on their ads. They can say that an ad with questions attracts more viewership and eliminates incompetent applicants. However, if the ad is still posted without questions, an internal team should review the authenticity of the ad before publishing it.

Slide 7 : What is the effect of pre-screening questions on job ads ?

Text analytics was performed on Description and Benefits fields in R studio. For Text mining we replaced alphanumeric chars with spaces and then deleted stop words from the field. Punctuation’s and the extra spaces were removed to account for the actual words that we might want to consider. Word cloud was created for the words that appear only for the Fraud jobs in our dataset for these unstructured variables. There were many repeated words observed in the different fields. Therefore, we made sure to pick distinct words relative to the respective fields. Example Darren Lawson is a person from Aptitude staffing, which is a company based in California. This connects to all our findings so far which can be used on monthly basis to identify such fraudulent advertise.

Slide 8 : What is the industry wise classification of fake ads posted ?

36.76% of the Fake ads do not have anything listed under the industry type. Therefore, make industry type a mandatory field and use text analytics model to generate high frequency words in this ad and compare it with the fraud ads result. Further analyzing the top 3 industries from the figure on left we were able to pin-point companies with maximum fraudulent ads. Therefore, make company name a mandatory field and create

a vector of identified fake ads, which will put the new ads directly for review. On further merging the analysis from EDA and Text Analytics, few ads were highlighted. Looking individually into these ads, it was observed that few specific company names were used to post majority of the fake ads. For Example – Aptitude Staffing Solutions, Aker solutions, Aptitude staffing Solutions + url.

Slide 9 : Model

RStudio was used to run a classification model for cross validating our EDA results. Training data did not have descriptive variables, and only variables with low VIF values were used. Text analytics was performed on descriptive string variables. We would recommend the company to make the highlighted fields mandatory as this would possibly oust out the fake/scam ads posted by bots and scamsters. R code is attached for reference.

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.

Slide 10 : Impact

An average person applies to 10 jobs per week and spends 11 hrs. in total per week. On average a person spends 1.1hrs behind each add this person is losing 1hr in a week on fake/scam ads.

Improve viewership and attract more resumes, as on an average a job seeker spends only 72 seconds on an ad, thus increasing the chances of handling more genuine Resumes.

As reliability increases website can help the company to attract more than the average 250 resumes per ad

More viewership will ultimately result to an increase in popularity, leading to command a price premium as the demand for the website increases.

Improved Customer Satisfaction Index by reducing 1 hour/week spent on fake ads by job seekers.

Prevent customers from exposing PII (Personally Identifiable Information) to scammers.

Maintain and Improve brand value by filtering fake ads and save $2.1M of revenue which could have been lost because of 5% of fake Ads. (Assumption)

source -
https://zety.com/blog/hr-statistics
https://www.zipjob.com/blog/how-many-jobs-applications-should-you-be-sending-out/

Google Data studio Summary

Link : Datastudio Dashboard

https://datastudio.google.com/reporting/70153237-cc9b-46ce-82c5-aaa08b2cfe30/page/hFYOB

Overview (Page 1)

Job listing websites would first want to see a brief overview of the company, hence the first page contains a brief overview of Fraudulent ads(0- Real/1- Fake) which is an interactive filter.

1. Scorecards are used to show total ads in the data for selected filter control.
2. Geo map was created on the variable Location to identify from where most of the fake ads were posted.
3. Bar plot was made to show real vs fake ads posted under an industry.
4. Interactive Pie charts were created on Logo Present/Absent, Has Questions Present/Absent, Telecommuting Required/Not Required for EDA analysis under fake ads.
5. The salary was analyzed using the pie chart and was found to not have a significant impact on fake ads.

Requirements vs Industry classification (Page 2)

This was created to identify the relationship between Industry and Education Requirements, Employment Type, Required Experience. These bar plots will help to identify a combination amongst these 4 variables which will lead to identifying fake ads.

With Industry “Not mentioned” and fraudulent=1, we have analyzed and concluded our Do’s.