SAS® GLOBAL FORUM 2021

Paper ###-2021

Detecting Online Recruitment Fraud

Matthew Fruge; Noah Heinrich

Oklahoma State University

ABSTRACT

With the advent of the global pandemic caused by the Coronavirus disease, the job market is in flux, with many people currently unemployed. These individuals are seeking employment via job posting websites such as Indeed. Online job posting websites are vulnerable to fraudulent job posts created by scammers in order to socially engineer job seekers out of private or sensitive information. This paper examines job postings and attempts to automatically determine if the job posting is genuine (non-fraudulent) or fraudulent. Predictive models were used on job postings in the United States, labeling each as fraudulent or non-fraudulent based on variables provided in the dataset. By using a text topic node combined with a decision tree model in SAS Enterprise Miner, the team achieved over a 96% accuracy rate at distinguishing whether a job posting is fraudulent or legitimate. This method could be used by sites like Indeed and Monster to flag job postings, warning the job seeker to proceed with caution, until further search quality control teams can assess the posting and take it down.

INTRODUCTION

COVID-19 has impacted the American workforce significantly, with some measures indicating that the unemployment rate reached 16%, the highest since the Great Depression (25%) [1]. The lack of income, coupled with the cost of living rapidly outpacing inflation, has led to a perfect storm of desperate individuals looking for work [2]. With the influx of prospective job seekers, there is a greater opportunity for criminals to prey upon those who are desperate for a job. The Federal Bureau of Investigations states that criminals utilize fake job postings to social engineer their target to divulge important personal information such as bank account information or social security number, with the average reported loss being nearly $3,000 per victim, damaging their credit score [3]. Creating a model that can potentially screen these dangerous and fraudulent job postings can help websites like Indeed and Monster improve their quality of service, driving more job seekers to their platform. The ability to quickly and efficiently differentiate between a real job posting and a fraudulent job posting will enable job websites to increase employer-employee match rate, leading to increased consumer confidence and subsequently higher usage and revenue.

DATA

The dataset used for this analysis was compiled by the University of the Aegean, Laboratory of Information & Communication Systems Security [4]. The dataset consists of approximately 18,000 job postings of which approximately 800 are fraudulent postings. The data was gathered from various job-related websites between the years 2012-2014. The dataset contains information regarding: the location of the job, the job title, the department, prospective salary, education requirements, as well as the nature of the job. The dataset also contains text information with a description of the job, descriptions of the

company from the posting, a description of the job requirements, and a description of what the company is looking for in their prospective employee. Finally, there is a binary value on each posting that indicates whether the job posting in question is fraudulent or real; this will be the target variable for this analysis.

DATA PREPARATION

The data consists of job postings from both domestic and international areas. The team decided to narrow the scope of the analysis to focus on domestic jobs in order to create more accurate and precise models. The first step was to identify the international job postings and remove them from the dataset. The data was then divided by region at a state level. Additionally, the team dummy coded all attributes that had more than 40% missing/null values as binary variable labeling if the posting had that information or not. For the text analytics portion, the SAS Enterprise Miner Text parsing node preprocessed the data by removing certain parts of speech such as articles, stemming the terms utilizing a SAS dictionary, and removed stop words utilizing the SAS Stop word dictionary.

PROJECT APPROACH

Data was imported via the File Import Node in SAS Enterprise Miner, the target variable was identified and the data was partitioned into a training and validation split. The text parsing node prepared the data to be further analyzed using the Text Topic and Text Cluster Nodes. The Text Topics and Text Clusters were then used as inputs to various decision tree based predictive models. Decision tree models were then compared to each other via the Model Comparison Node. This whole process was iterated multiple times throughout the research in order to find what variables were most effective in creating the best models.

ANALYSIS

The analysis of the data was conducted using SAS Enterprise Miner. The data was initially divided via stratified sampling into training and validation subsets at a ratio of 70% training and 30% validation. Multiple models were trained to find an accurate predictive model to predict fraudulent postings based on the given variables. Decision tree modeling was utilized in order to create an accurate predictive model that can handle the frequent null values found within the dataset. Each iteration of the modeling included several decision trees that would subsequently be compared against each other with the "champion" model being the tree with the most accuracy. The trees utilized in each iteration were: a decision tree with a Gini index utilized as the split criterion, a decision tree with a p-value of F test used as the split criterion, and a gradient boosted tree.

The first set of models trained analyzed all the variables except the text-based variables (benefits, company profile, description, and requirements). This was executed to see if an accurate model could be created based purely off the binary and numeric values in the data set. The most accurate model in this iteration was found to be the decision tree that utilized Gini as its split criterion. Precision and sensitivity were lacking (Table 1), which led to a decision to utilize the text variables in order to create more robust models.

In order to properly model the text variables, the team decided to attempt two different text analysis techniques to feed into the decision tree models. The decision was made to utilize text clustering and text topics, each with their own set of three models ran

(as stated before), and then had all six models compared against one another in order to select the most accurate model. This splitting of the data into six comparative models would be iterated many times throughout the research.

Initially, four text variables were combined into a single corpus in order to parse the text into topics and clusters to improve the decision tree models. The four text variables and all the numerical and binary variables were then analyzed by the text topic and text clustering decision tree models. The best model in this iteration utilized text clustering and a split criterion of ProbF for the subsequent decision tree. The precision and sensitivity could also be improved. Variables were also individually analyzed utilizing the same modeling techniques for the variable combinations.

	Champion	Misclass	TP	TN	FP	FN	Sens	Spec	Acc	Prec
Benefits	Cluster/Gini	0.052	70	2261	15	96	42.169	99.341	0.948	82.353
Benefits + DV	Cluster/ProbF	0.055	76	2231	45	90	45.783	98.023	0.945	62.810
Profile	Topic/Gini	0.040	68	2257	19	98	40.964	99.165	0.960	78.161
Profile + DV	Cluster/ProbF	0.046	69	2260	16	97	41.566	99.297	0.954	81.176
Description	Cluster/ProbF	0.057	72	2232	44	94	43.373	98.067	0.943	62.069
Description + DV	Cluster/ProbF	0.055	76	2231	45	90	45.783	98.023	0.945	62.810
Requirements	Cluster/ProbF	0.051	62	2256	20	104	37.349	99.121	0.949	75.610
Requirements + DV	Topic/ProbF	0.055	68	2239	37	98	40.964	98.374	0.945	64.762
All Variables	Cluster/ProbF	0.057	72	2232	44	94	43.373	98.067	0.943	62.069
No Text Variables	Topic/Gini	0.052	59	2255	21	107	35.542	99.077	0.948	73.750

Table 1: Overall Model Comparison Statistics

The team then ran the 6 models against each individual text variable, including a separate iteration for each of the text variables created with dummy variables. This led to 8 additional iterations of the previous models, each with their own "champion model". Table 1 indicates how each iteration performed, and the most accurate model amongst all the iterations was selected to be the overall "supreme champion" model. A decision was made that the best model should to also have good sensitivity defined as greater than 40%. Accuracy was important due to the fact that the team wanted a model that could correctly identify whether a post was fraudulent. Sensitivity was also a factor in the decision since the team wanted to ensure that if a post is marked as fraudulent that it is fraudulent and not a legitimate post. This would ensure that customers would not become disenfranchised if their legitimate posting was flagged as fraudulent.

The most accurate model was found to be the model that analyzed the "Company profile" variable and utilized text topics to generate variables for the decision tree model that had a Gini Index as its split criterion. This model was not only the most accurate, but still demonstrated a good level of sensitivity, which meant that if a post is marked as fraudulent, one can be confident that it is actually fraudulent and not a false positive. Figure 1 has all the variables and their corresponding importance. Figure 2 has the variable importance of the text topics along with the words comprising the topics. The variable importance leads to conclusions that are drawn in a later section of the paper.

The model created in the research can be applied to additional data that fulfills the variable criteria created by the University of the Aegean. Since the data was gathered between the years 2012 and 2014, new data can be filled out that follows the basic information from the team's research dataset, that includes text information about the company profile, benefits, job description, and job requirements, in addition to the general job posting information such as location, salary, and type of Industry. Since the majority of this information is already gathered by job posting websites, this model can be utilized against job postings from job posting websites such as Indeed or Monster.

FUTURE SCOPE

One limitation to the model presented in this paper is that one does not know the severity of false positive or false negative from a cost standpoint for the user or business. This makes it hard to determine what model comparison metric to use and determine the best performing model. For example, the fraud type of each fraudulent case is unknown, so one cannot know how that affects job seekers or corporations on a per case basis. Platforms like Indeed and Monster have quality control teams to ensure fraudulent postings are flagged, with standard methods for determining which job posts are genuine and records of how many applied. The team decided to utilize sensitivity to ensure that a genuine posting would not be flagged as fraudulent to maintain customer trust in the platform. Collecting more data and information held by the platforms would allow for a more accurate representation of the severity per case, and improve model deployment. Finally, with more time, the team would have liked to explore more performance metrics like Cohen's Kappa and ROC Curves, due to the dataset being heavily imbalanced (classes not represented equally) with a ratio of approximately 15:1. With the target category being heavily imbalanced our 96% accuracy is not as appealing at deciphering the fraudulent postings and the non-fraudulent postings, because if the team predicted all the job postings as non-fraudulent, 93% accuracy would still be achieved. Additionally, it is possible that a further study compensating for the "Accuracy paradox" could produce a better result.

CONCLUSION

With COVID-19 creating high levels of unemployment, scammers have a greater opportunity to defraud job seekers who are desperate for work. The analysis created a model with 96% accuracy with sensitivity of 40.9% that can determine whether or not a job posting is fraudulent based on several variables. The industries of Accounting, Oil, and Energy, IT, Leisure, Travel and Tourism, and Biotech are the most susceptible to fraudulent postings, as well as any job posting with an administrative function. Additionally, jobs that do not require experience and pay under 80,000 dollars a year are also more likely to be fraudulent. Postings that fall under these categories on a job posting website can be put under additional scrutiny.

The analysis also revealed certain text topics found in the company description are indicators of a fraudulent post. Words such as "contract", "work from home", "free", "accommodation" "lender" and "retirement" were also often red flags for a fraudulent post. This can indicate that scammers are seeking the "low hanging fruit" people looking for easy

part-time work that allows telecommuting or flexible hours, as well as potentially targeting older workers and retirees.

There were also two indicators of legitimate postings. The first was any posting that contained a company logo, which often indicated legitimacy. The second was a text topic with words such as "encourage", "bright", "colleague", "develop", and "advanced". This typical corporate jargon is indicative of genuine job postings, as scammers seem to avoid using corporate buzzwords and instead attempt to use simple language to attract more potential victims.

Overall, the research model highlighted potential indicators for both legitimate and illegitimate job postings that can be provided to job posting websites such as Indeed and Monster. They can provide additional screening on certain postings that meet the criteria referenced above, making it harder for scammers to create fraudulent postings on their platform. Also, the research model can be further tested by applying it to additional data from a more recent period in order to see if the model will perform as well in a different data environment.

REFERENCES

[1] Kochhar, R. (2020, August 26). Unemployment rose higher in three months of COVID-19 than it did in two years of the Great Recession. Retrieved December 05, 2020, from https://www.pewresearch.org/fact-tank/2020/06/11/unemployment-rose-higher-in-three-months-of-covid-19-than-it-did-in-two-years-of-the-great-recession/

[2] Desilver. (2018, August 7). For most U.S. workers, real wages have barely budged in decades. PewResearch. Retrieved December 05, 2020, from https://www.pewresearch.org/fact-tank/2018/08/07/for-most-us-workers-real-wages-have-barely-budged-for-decades/

[3] Kolko. (2020, January 24) FBI Warning: Scammers using fake job listings to target applicants for info. WinkNews. Retrieved December 05, 2020 from https://www.winknews.com/2020/01/24/fbi-warning-scammers-using-fake-job-listings-to-target-applicants-for-info/

[4] University of Aegean Dataset http://emscad.samos.aegean.gr/

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Matthew Fruge
Matthew.fruge@okstate.edu
Noah Heinrich
Noah.heinrich@okstate.edu

Appendix

Figure 1: All Variable Importance

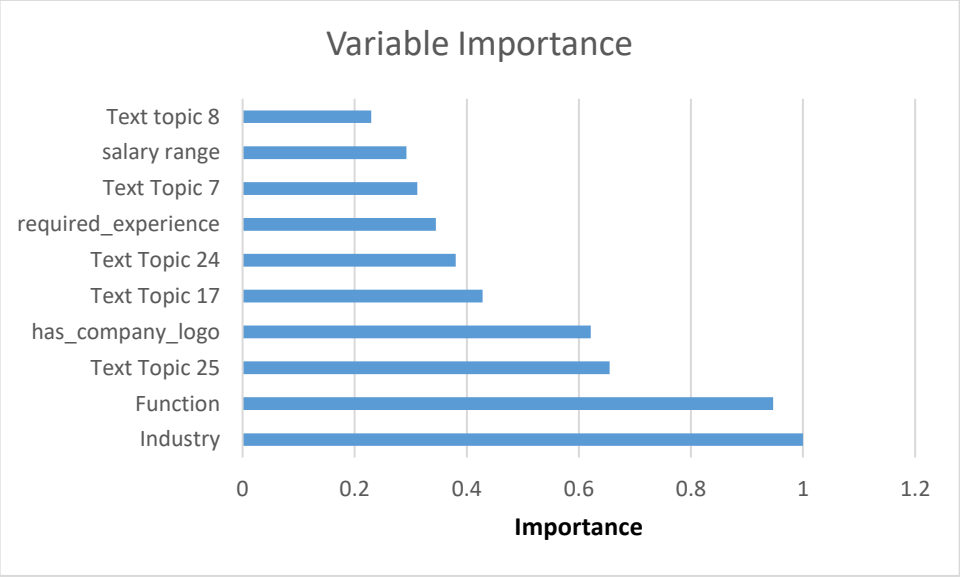

Figure 2: Text Topic Variable Importance