JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Study on Fake Job Predictor

Mr. Aharta Dudhe, Mr. Vivek Khobragade, Miss Saylee Nagdeote, Miss Rani Saden

Prof. Rupa Lichode

Department Of Computer Science and Engineering, Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur (MH)

***Abstract*— In the past few years, because of advancement of latest technology, advertisement of job posts has become head ace in present world. So, fake job prediction posts task is going to be challenging for all of us.**

We use Random Forest Classifier to predict, whether job post is real or fraudulent. This experiment is done on Employment Scam Aegean Dataset (EMSCAD) which containing 1800 samples. This trained classifier has given us accuracy of 95% when it comes to predict a fake job post.

Keywords—Fake Job Prediction, Random Forest Classifier, Employment Scam.

1.INTRODUCTION

Due to the pandemic situation, employment scam is one of the serious issues in recent times addressed in the Online Recruitment Frauds (ORF) domain.
So, to prevent such feasible traps, we are going to create a fake job detection model and for the detection of jobs, we are going to use Random Forest Algorithm.

1.1 What is Machine Learning?

Machine Learning is a system of computer algorithms that can learn from illustration through tone- enhancement without being explicitly enciphered by a programmer. Machine literacy is a part of artificial Intelligence which combines data with statistical tools to prognosticate an affair which can be used to make practicable perceptivity. [1]
The advance comes with the idea that a machine can singularly learn from the data (i.e., illustration) to produce accurate results. Machine literacy is nearly related to data mining and Bayesian prophetic modelling.

A typical machine literacy tasks are to give a recommendation. For those who have a Netflix account, all recommendations of pictures or series are grounded on the stoner's literal data. [2] Tech companies are using unsupervised literacy to ameliorate the stoner experience with bodying recommendation.

Machine literacy is also used for a variety of tasks like fraud discovery, prophetic conservation, portfolio optimization, automatize task and so on.

1.2 Machine Learning vs. Traditional Programming

Traditional programming differs significantly from machine literacy. In traditional programming, a programmer law all the rules in discussion with an expert in the assiduity for which software is being developed. Each rule is grounded on a logical foundation; the machine will execute an affair following the logical statement. [4]
Traditional programming differs significantly from machine literacy. In traditional programming, a programmer law all the rules in discussion with an expert in the assiduity for which software is being developed.[2] Each rule is grounded on a logical foundation; the machine will execute an affair following the logical statement. When the system tends to grows complex, more rules are needed to be written. It can easily become unsustainable to maintain.
Machine Literacy is supposed to overcome this issue. The machine learns how the input and affair data are identified and it writes a rule. The programmers don't need to write new rules each time there's new data. The algorithms acclimatize in response to new data and gests to ameliorate efficacity over time.[7]

1.3 How does Machine Learning Work?

Machine studying is the mind wherein all of the studying takes place. The manner the gadget learns is much like the human being. Humans research from experience. The greater we know, the greater effortlessly we are able to predict. By analogy, whilst we are facing an unknown situation, the probability of achievement is decrease than the regarded situation. Machines are skilled the same.[6] To make a correct prediction, the gadget sees an example. When we provide the gadget a comparable example, it may discern out the outcome. However, like a human, if it's feed a formerly unseen example, the gadget has problems to predict.

The center goal of gadget studying is the studying and inference. First of all, the gadget learns thru the invention of patterns. This discovery is made way to the information. One important a part of the information scientist is to select cautiously which information to offer to the gadget. The listing of attributes used to remedy a trouble is referred to as a function vector.[5] You can think about a function vector as a subset of information this is used to address a trouble.
The gadget makes use of a few fancy algorithms to simplify the truth and remodel this discovery right into a version. Therefore, the studying level is used to explain the information and summarize it right into a version.[8]

1.4 Inferring

When the version is constructed, it's far feasible to check how effective it's far on never-seen-earlier than information. The new information is converted right into a capabilities vector, undergo the version and provide a prediction. This is all of the stunning a part of gadget studying. There isn't any want to replace the regulations or teach once more the version. You can use the version formerly skilled to make inference on new information.[1]

The existence of Machine Learning applications is simple and may be summarized withinside the following points:
1. Define a question
2. Collect information
3. Visualize information
4. Train set of rules
5. Test the Algorithm
6. Collect feedback
7. Refine the set of rules
8. Loop 4-7 till the outcomes are satisfying
9. Use the version to make a prediction
Once the set of rules receives right at drawing the proper conclusions, it applies that expertise to new units of information.

1.5 Random Forest

The set of rules is constructed upon a selection tree to enhance the accuracy drastically. Random wooded area generates generally easy selection timber and makes use of the 'majority vote' technique to determine on which label to return. For the category task, the very last prediction might be the only with the maximum vote; at the same time as for the regression task, the common prediction of all of the trees is the very last prediction.[3]

2.TIMELINE OF TECHNOLOGY

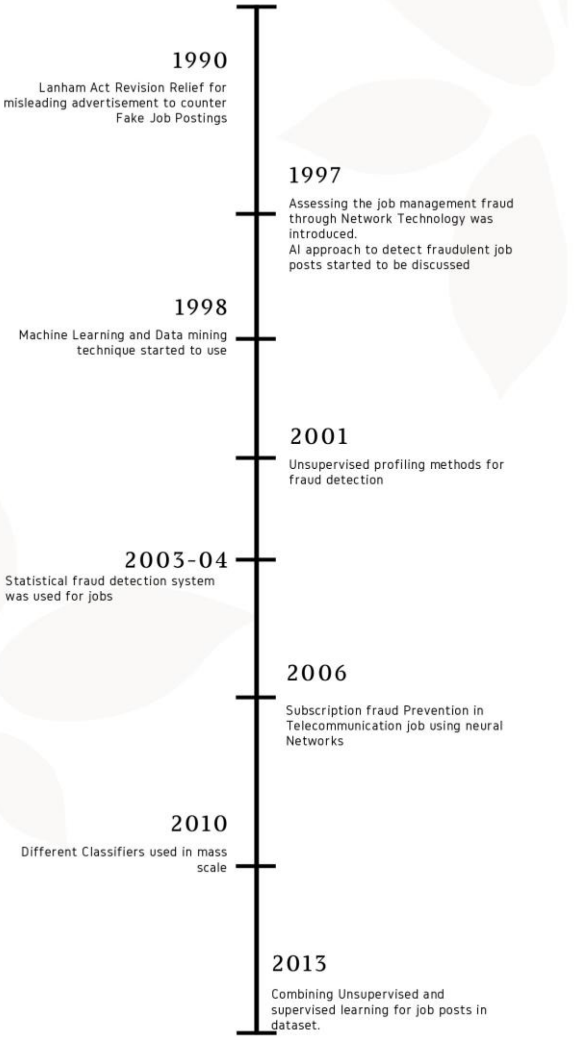
1990
Lanham Act Revision Relief for misleading advertisement to counter Fake Job Postings
1997
Assessing the job management fraud through Network Technology was introduced.
AI approach to detect fraudulent job posts started to be discussed
1998
Machine Learning and Data mining technique started to use
2001
Unsupervised profiling methods for fraud detection
2003-04
Statistical fraud detection system was used for jobs
2006
Subscription fraud Prevention in Telecommunication job using neural Networks
2010
Different Classifiers used in mass scale
2013
Combining Unsupervised and supervised learning for job posts in dataset.

Figure 1. The Timeline of various Technology

3. METHODOLOGY

3.1 MODULES

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set
- Saving the Trained Model

3.2 MODULES_DESCRIPTION

3.2.1 Data Collection

This is the primary actual step toward the actual improvement of a device mastering version. This is a vital step with a view to cascade in how excellent the version will be, the greater and higher records that we get, the higher our version will perform. There are numerous strategies to gather the records, like net scraping, guide interventions and etc. A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques.[6]

3.2.2 Dataset

The dataset consists of 17880 individual data. There are 18 columns in the dataset, which are described below.
1. job_id - unique vacancy identifier
2. title - headline
3. location - the geographical location of the job advertisement
4. department - corporate department (for example, sales)
5. salary_range - indicative salary range (eg 50,000-60,000)
6. company_profile - a short description of the company
7. description - detailed description of the job advertisement
8. requirements - the requirements for the vacancy are listed
9. benefits - the proposed benefits are listed;
10. telecommuting - true for remote posts
11. has_company_logo - true if the company logo is present;
12. has_questions - true if test questions are present
13. employment_type - type of employment;
14. required_experience - necessary experience
15. required_education - necessary education
16. industry - industry
17. function - function to be performed
18. fraudulent - indicates whether the job is fraudulent

3.2.3 Data Preparation

Wrangle data and prepare it for training. Removes unnecessary data which we do not require.
Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
Split into training and evaluation sets[5]

Model Selection:
We used Random Forest Classifier algorithm, We got a accuracy of 94.7% on test set so we implemented this algorithm.

3.2.4 Random Forest Algorithm

The Random Forest Algorithm Let's understand the algorithm in simple terms. Suppose you want to take a trip and travel to a place that you enjoy. So what do you do to find a place you like? You can search online, read reviews on blogs and travel sites, or ask your friends. Suppose you decided to ask your friends and talk to them about their previous travel experiences to different places. You will get some recommendations from each friend. Now you need to make a list of these recommended places. Then ask them to vote (or choose the best place to travel) from the list of recommended places you created. The place with the most votes is your final choice for the trip. The above decision-making process consists of two parts. Start by asking your friends about their individual travel experiences and get recommendations on different places they have visited. This part is like using the decision tree algorithm. Here each friend makes a choice. of the places you have visited so far. The second part, after collecting all the recommendations, is the voting process to choose the best place from the list of recommendations. This whole process of getting recommendations from friends and voting for them to find the best spot is known as the random forest algorithm. Technically, it is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly divided dataset. This collection of decision tree classifiers is also known as the forest. Individual decision trees are generated based on an attribute selection indicator such as information gain, win rate, and Gini index for each attribute
. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In regression, the final result is the average of all tree outputs. It is simpler and more powerful compared to the other nonlinear ones. sorting algorithms.[7]

3.2.5 How does the algorithm work?

It works in four steps: Choose random samples from a given dataset. Build a decision tree for each sample and get a prediction result from each decision tree. Vote for each predicted outcome. Select the prediction result with the most votes as the final prediction.

3.2.6 System Design

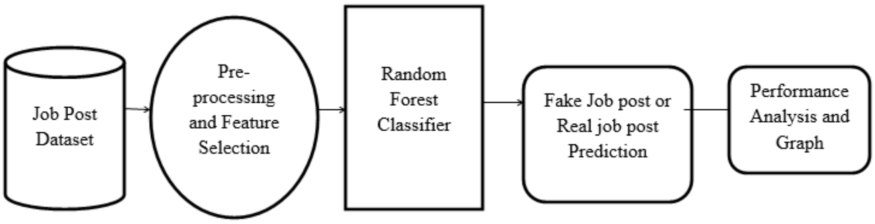
Job Post Dataset
Pre-processing and Feature Selection
Random Forest Classifier
Fake Job post or Real job post Prediction
Performance Analysis and Graph

Figure 2. Diagram of System Design

3.2.7 Advantages

• Random forests are considered to be a very accurate and robust method due to the number of decision trees involved in the process.
• The main reason is that it takes the average of all the predictions, removing the bias.
• The algorithm can be used for both classification and regression problems.
• Random forests can also handle missing values. There are two ways to handle this: use medians to substitute for continuous variables, and calculate the proximity-weighted average of missing values. Obtain the relative importance of the features, which helps to select the features that contribute the most to the classifier. [4]

3.2.8 Disadvantages

• Random forests are slow to make predictions because they have multiple decision trees.
Each time you make a prediction, all trees in the forest must make a prediction for the same given entry and then vote on it.
• The model is difficult to interpret compared to a decision tree where you can easily make a decision by following the path of the tree.[4]

3.2.9 Finding Important Features

Random forests also provide a good indicator of feature selection. Scikitlearn provides the model with an additional variable that shows the relative importance or contribution of each feature to the prediction. Automatically calculates the relevance score for each feature in the training phase. It then reduces the relevance so that the sum of all scores is 1.

This score will help you select the most important features and discard the less important ones for model building.

The random forest uses Gini importance or mean impurity decrease (MDI) to calculate the importance of each feature.
Gini significance is also known as the overall node contamination decrease. This is how much the fit or accuracy of the model decreases when you remove a variable. The greater the decrease, the more important the variable becomes. The mean decrease is an important parameter for variable selection. The Gini index can describe the general significance of the variables[2]

3.2.10 Random Forests and Decision Trees

• Random forests is a set of multiple decision trees.
• Deep decision trees may suffer from overfitting, but random forests prevents overfitting by creating trees on random subsets.
• Decision trees are computationally faster.
• Random forests is difficult to interpret, while a decision tree is easily interpretable and can be converted to rules.

Analyze and Prediction:
In the actual dataset, we chose only 8 features
1. telecommuting - true for remote posts
2. has_company_logo - true if the company logo is present;
3. has_questions - true if test questions are present
4. employment_type - type of employment;
5. required_experience - necessary experience
6. required_education - necessary education
7. industry - industry
8. function - function to be performed
result: indicates whether the job is fraudulent

Accuracy on test set:

We got an accuracy of 97.80% on test set.

Saving the Trained Model:

Once this trained and tested model is ready in production environment. After this first thing you have to do is to save into .pkl file using pickle library.

Accuracy on test set:

We got an accuracy of 95.02% on test set.

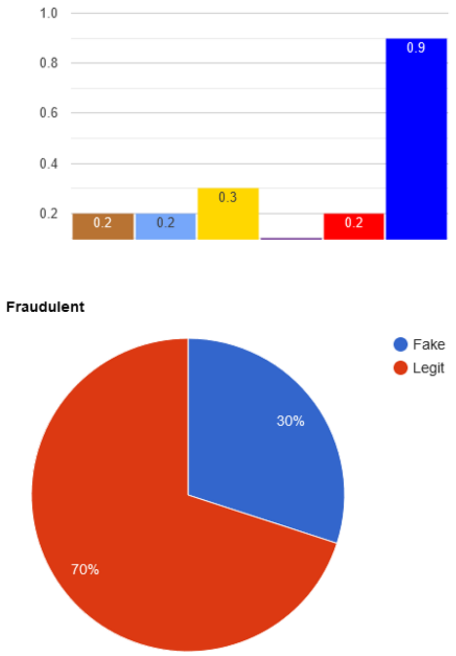

Figure 3. Output result showing percentage of Fake Job and Legit Jobs

4.CONCLUSION

Job scam detection has grown to be a extraordinary difficulty everywhere in the international at present. In this paper, we've analyzed the influences of activity rip-off which may be a completely rich vicinity in studies filed developing loads of demanding situations to hit upon fraudulent activity posts. We have experimented with EMSCAD dataset which includes actual existence faux activity posts. This painting indicates a comparative examine at the assessment of conventional gadget mastering primarily based totally classifiers. We have located maximum category accuracy for Random Forest Classifier amongst conventional gadget mastering algorithms.

Recruitment fraud can certainly come as a wish crushing second in determined instances whilst university placements can't be of any help. However, you need to usually be aware of the truth that no recruiter might ask you for any bills and any reputed employer might now no longer provide you a skyrocket or median package deal simply on a perusal of your CV except it's far very incredible and you return back from any Institutes of National importance.

These fraudsters are capable of live to tell the tale and amplify their operations extra unexpectedly due to the desperation to land a activity and clean availability of your info as an employable candidate at the Internet. It is in no way an amazing concept to click on classified ads pop-ups that ask you to sign-up and to add your resumes. It is usually an amazing concept to speak for your peers, and relatives, faculty, placement mobiliary earlier than leaping to make any fee whatsoever.

Do query yourself, in case you are a capacity organization for the recruiter, and you're decided on the premise of your CV, then why might a recruiter reject you for the activity in case you do now no longer make the bills designated below the recruitment method upfront? In case you're going via way of means of the recruitment enterprise, they will ask you for a few fees for the service, however, they may be simply a facilitator in taking your CV to the proper employment as according to your profile. Even a recruitment enterprise can't assure you approximately the activity and in which you sign on with the recruitment enterprise, you need to be receiving activity gives from the employer thru them and now no longer directly. In case you're a sufferer of such frauds, do now no longer hesitate in reporting, wondering which you had been fooled. Come out brazenly and document the fraud.

5.REFERENCES

[1] S. Vidros, C. Kolias , G. Kambourakis ,and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset", Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

[2] B. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", Journal of Information Security, 2019, Vol 10, pp. 155176, https://doi.org/10.4236/iis.2019.103009.

[3] Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, "Job Prediction: From Deep Neural Network Models to Applications", RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

[4] Jiawei Zhang, Bowen Dong, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network", IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[5] Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s13388-014-0005-5

[6] Y. Kim, "Convolutional neural networks for sentence classification," arXiv Prepr. arXiv1408.5882, 2014.

[7] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.-T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model," arXiv Prepr. arXiv1911.03644, 2019.

[8] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, pp. 806814, 2016.