International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.085 ISSN: 2581-4621

Prediction of Deceitful Jobs in Social Websites

S Athinarayanan[1], S Bhargavi[2], B Bhavitha[3], P Dileep[4], C Chandra Mouli[5]
Professor, Student, Department of CSE, AITS, Tirupati

Abstract

In recent years, due to advancement in modern technology and social communication, advertising new posts has become very common issue in the present world Like many other classification tasks, fake posting prediction leaves a lot of challenges to face. With the advancement in technology and rigorous use of social media platforms, many job seekers and recruiters are actively working online. However, due to data and privacy breaches, one can become the target of perilous activates. The agencies and fraudsters entice the job seekers by using numerous methods, sources coming from virtual job-supplying websites. We aim to reduce the quantity of such fake and fraudulent attempts. To design this system we use different data mining techniques and classification algorithm like such as XG Boost classifier, Random Forest classifier and Decision Tree in a based Python environment.
Keywords: XG Boost classifier, Random Forest classifier and Decision Tree

Introduction

Employment scams are one of the more important concerns that have recently been addressed in the realm of Online Recruitment fraud. We are living in unprecedented times as a result of the COVID-19 epidemic, which is wreaking havoc on economies around the globe. Unemployment rates are rising daily, with the United States reporting over 26 million people, the most recorded in the country's long history. In its most recent World Economic Outlook report, the IMF (International Monetary Fund) forecasts unemployment in Pakistan at 13% in 2020, up from 7.3 percent in 2019 and 3.9 percent in 2018. Many organizations now choose to list their job openings online so that job searchers can access them readily and quickly. However, this might be a form of scam perpetrated by fraudsters who promise work to job seekers in exchange for money. To undermine a reputable company's credibility, fraudulent job adverts might be issued. These fraudulent job post detections pique people's interest in acquiring an automated tool for recognizing bogus jobs and alerting them to individuals to avoid applying for such positions. A machine learning technique is used for this goal, which utilizes numerous classification algorithms for detecting bogus postings. In this scenario, a classification tool detects and warns the user when it detects bogus job postings among a bigger set of job adverts.

A machine learning technique is used for this goal, which utilizes numerous classification algorithms for detecting false postings. In this instance. A classification technique separates bogus job postings from real ones a broader range of employment adverts and notifies the user. To address the issue of recognizing job frauds posting, as well as supervised learning algorithm initially, classification approaches are investigated. By mapping input variables to target classes, the classifier taking into account training data Classifiers mentioned in the document is used for detecting bogus job postings.

Related Work

1. Material and Methods:

We have decided to use the dataset published by EMSCAD (EMployment SCam Aegean Dataset) which contains real life job ads posted by workable [2]. EMSCAD contains 17,014 legitimate and 866 fraudulent job ads. Data contains various information like job ID, job title, name of the organization, location, company profile, employment type, jobdescription, job requirements, benefits, required education, type, whether job posting is fraudulent or not, etc. This dataset contains both categorical and description format which is pre-processed to make it useful in training the model.

To give more relevancy and real life market exposure we extracted data from major online job portals, LinkedIn. This was done by web scrapping using Beautiful Soup library. Thedata was obtained in .json format which was then converted to.csv format. The data base had 138 job posts with details like

job title, name of the organisation, location, employment type, job description, job function, type of the organisation, whether job posting. This data was manually analysed to identify if theposting is fake or not. This updated database was used to test model and evaluate the results which are more realistic. Several machine learning models have
been proposed to classify whether job postings are fake, but none adequately address this misdiagnosis problem. Similar studies proposing
models to evaluate such performance classifications generally did not take into account the heterogeneity and the size of the data. Therefore we propose RandomForest, XGBoostclassifiers and decision tree to predict false hire detection.

2. Algorithm

2.1 XG Boost

XGBoost classifier combines the predictions of several decision trees to make a final prediction. Extreme Gradient Boosting (XGBoost) is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning framework. It is the top machine learning package for regression, classification, and ranking tasks, and it supports parallel tree boosting. XG Boost is **a** popular and efficient open-source implementation of the gradient-boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

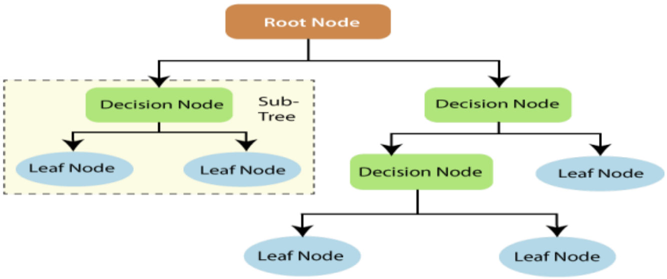
```
Input: Training data X, target variable y, hyper parameters.
Output: Fitted XGBoost Classifier
1. Split the data into training and validation sets using a pre-defined split or cross-validation
2. Initialize an XG Boost classifier with the given hyper parameters
3. Train the classifier on the training set using the  XGBoost fit() method with early stopping criteria
4. Evaluate the classifier on the validation set using the XGBoost predict() method
5. Calculate the classification metrics of interest (such as accuracy, F1 score, or AUC) on the validation set
6. Adjust hyperparameters based on validation metrics and repeat steps 3-5 until desired metrics are achieved
7. Train the final XGBoost classifier on the entire training set using the optimal hyperparameters found in step 6
8. Output the fitted XGBoost classifier.
```

This is how the working of the XG Boost classifier in detectiong the deceitful jobs.

2.2 Decision Trees

Decision trees work by dividing input variables into subsets based on their values and recursively dividing the data into smaller subsets until a stopping condition is met. Each partition is based on the input variable that provides the greatest information gain, i.e. the reduction in entropy or impurity of the data. In crop yield forecasting, decision trees can be used to identify the environmental factors that have the greatest impact on crop yield and how they interact. Decision trees have several advantages in predicting crop yield. It can handle both categorical and continuous input variables, can handle missing values, and can automatically identify interactions between input variables. Additionally, decision trees are interpretable and the structure of the tree provides insight into the relationship between input variables and output variables. Overall, decision trees are a powerful and interpretable machine learning algorithm for crop yield prediction.

It can handle both categorical and continuous input variables and can identify interactions between input variables.



2.3 Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like Scikit-learn).

Entropy and information gain are the building blocks of decision trees. An overview of these fundamental concepts will improve our understanding of how decision trees are built.

Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables.The information gain concept involves using independent variables (features) to gain information about a target variable (class). The entropy of the target variable (Y) and the conditional entropy of Y (given X) are used to estimate the information gain. In this case, the conditional entropy is subtracted from the entropy of Y.Information gain is used in the training of decision trees. It helps in reducing uncertainty in these trees. It helps in reducing uncertainty in these trees. A high information gain means that a high degree of uncertainty (information entropy) has been removed.

4. Result:

In order to establish a baseline for comparison with the proposed cost-sensitive AdaBoost (CS AdaBoost), this research also implemented several well-known classifiers, including logistic regression, decision tree, XGBoost, random forest, and SVM, along with the traditional AdaBoost presented in Algorithm 1. Both the complete and reduced feature sets were used to train the classifiers, in order to demonstrate the impact of feature selection.

Table 2: Performance of the Classifiers without Feature Selection

Classifier	**ACC**	**SEN**	**SPE**	**AUC**
Logistic regression	0.940	0.948	0.935	0.940
Decision tree	0.902	0.932	0.790	0.910
XGBoost	0.963	0.974	0.942	0.960
Random forest	0.952	0.955	0.940	0.960
SVM	0.927	0.943	0.930	0.940
AdaBoost	0.940	0.941	0.935	0.950

Proposed CS AdaBoost	0.965	0.965	0.950	0.980

5. Conclusion:

In this paper, a new method was suggested to improve the detection of deceitful jobs using machine learning. The proposed method uses XG Boost classifier,Random Forest and Decision trees. The algorithms compared the performance of this new method with six other machine learning classifiers, including logistic regression, decision tree, random forest, SVM, XGBoost, and the traditional AdaBoost. Then, the machine learning classifiers were trained using both the important and all the available attributes. The results showed that using the important attributes improved the performance of the classifiers. This new method proposed in this research paper can help improve the accuracy of detecting fake jobs, which is essential for the job seekers.

6. References:

[1]. B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection," J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009

[2]. I. Rish, ―An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,‖ no. January 2001, pp. 41–46, 2014.

[3]. D. E. Walters, ―Bayes's Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[4]. F. Murtagh, ―Multilayer perceptrons for classification and regression,‖ Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5]. P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6]. H. Sharma and S. Kumar, ―A Survey on Decision Tree Algorithms of Classification in Data Mining,‖ Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[7]. E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems,‖ Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[8]. L. Breiman, ―ST4_Method_Random_Forest,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[9] B. Biggio, I. Corona, G. Fumera, G. Giacinto,   and F. Roli, ―Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37

[10] A. Natekin and A. Knoll, ―Gradient boosting machines, a tutorial,‖ Front. Neurorobot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[11] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, ―Spam review detection techniques: A systematic literature review,‖ Appl. Sci., vol. 9, no. 5, pp. 1–26, 2019, doi: 10.3390/app9050987.

[11] H. M and S. M.N, ―A Review on Evaluation Metrics for Data Classification Evaluations,‖ Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[12] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, ―Cohen's kappa coefficient as a performance measure for feature selection," 2010 IEEE World Congr. Comput. Intell. WCCI 2010, no. May 2016, 2010, doi: 10.1109/FUZZY.2010.5584447.