# Unmasking Employment Scam

An automated system to detect fraudulent job postings.


***Dr. N. Murali Krishna***
*Professor*
*Department of Artificial Intelligence and Data Science,*
*Vignan institute of technology and science,*
*Hyderabad, India*

***Guthikonda Karthik***
*UG Student,*
*Department of AI&DS,*
*Vignan institute of technology and science,*
*Hyderabad, India,*
*karthikrayudu6301@gmail.com*

***G.Pranay***
*UG Student,*
*Department of AI&DS,*
*Vignan institute of technology and science,*
*Hyderabad, India,*
*pranaygalva@gmail.com*

***K.Balaji***
*UG Student,*
*Department of AI&DS,*
*Vignan institute of technology and science,*
*Hyderabad, India,*
*balaji.vgnt@gmail.com*



***Abstract: An increased fraudulent job posting continues to pose a grave threat to the job seeker who is more or less susceptible to financial and personal risks. The paper takes this issue to solve it with the help of Employment Scam Aegean Dataset (EMSCAD) using the powerful machine learning algorithm called XGBoost, for classification of jobs posted as real or fake ones, which, to a good approximation, will show scams around near 98% with robust computational efficiency. For this reason, it will certainly be of service in high volumes. With the implementation of the solution in a website for jobs, the purpose of the solution will attempt to improve security for job markets and prevent victims from being misguided while applying.***



***Keywords-XGBoost algorithm, Machine learning, EMSCAD***


## I. INTRODUCTION

Over the last few years, swift technology and internet platforms have revolutionized the recruitment space. Online advertising jobs easily provides numerous opportunities to the employers so that they get access to the right job seeker; also, information gets transferred very efficiently. However, with this electronic transformation, it has also increased the alarming phenomenon of fake postings for jobs, where the person getting cheated would often face the problems of immense monetary and psychological issues.

Fake job postings exist to dupe people into becoming victims by representing them as job opportunities. It often tries to swindle sensitive information or financial aid from the applicant under false representation. The proliferation of such activities has created an urgent need for strong mechanisms in identifying and fighting such fraudulent acts.

It covers the state-of-the-art in data mining and machine learning towards fraudulent job-posting identification and prediction. Its popularity in high efficiency, scalability, and its superior performance over classification tasks in general makes this study suitable to apply Extreme Gradient Boosting on it. XGBoost is an ensemble of a decision-tree-based algorithm whose purpose is both speed and a superior fit of the model in terms of its accuracy. It's very handy to manipulate the rather complex data set presented before us by the Employment Scam Aegean Dataset, EMSCAD, in that it can handle missing data, normalizes to mitigate overfitting, and it supports parallel computation.

The objective behind the proposed work is to provide a reliable framework for fraudulent job advertisements and assist in protecting the job seekers and promoting an online recruitment atmosphere that is safer than earlier. Comparing the performance of XGBoost with other machine learning classifiers, this study is aimed at the efficiency of detecting fake job posts. The findings will add to the already ongoing fights of online employment scams and will pave the way forward for future inventions in the realm of secure recruiting practice.

## II. RELATED WORK

Habiba et al. [1] used multiple data mining and classification algorithms that include Decision Trees, SVM, Naïve Bayes classifier, Random Forest classifier, Multilayer Perceptron, and K-Nearest Neighbor to determine if the job posting was spam or not. In the same vein, Amaar et al. [2] presented six models of machine learning to analyze job postings. One of the issues they encountered was the highly unbalanced dataset between the real and the fake job postings, which biased their models toward the majority class. They therefore conducted experiments on both balanced and unbalanced datasets to obtain a more rigorous evaluation.

Mehboob et al. [3] developed new methodology for recruitment scam detection through the use of organizational attributes, job postings, and compensation packages as predictors for scamming activity. A gradient-boosting model was derived by using three comparison methods. The two-step feature selection was used. It has been determined from the results of the study that fraud activities are well predicted by the nature of the organization.

Ranparia et al. [4] demonstrated the effectiveness of Natural Language Processing (NLP) in detecting fraudulent job advertisements by analyzing the tone and format of job postings. They trained their model as a Sequential Neural Network using the Global Vector algorithm and tested it on LinkedIn job postings to evaluate its real-world applicability. Their ongoing work focuses on enhancing adaptability and resilience in detecting fraudulent job ads.

Sudhakar et al. [5] proposed a new algorithm to differentiate between fake and authentic news using logistic regression,

SVM, and an ensemble technique. Their study with 10,000 samples achieved a 95% accuracy rate with minimal loss using an ensemble approach combining decision tree methods with AdaBoost. Their method was, however, limited by the lower accuracy of the dataset used compared to other available datasets.

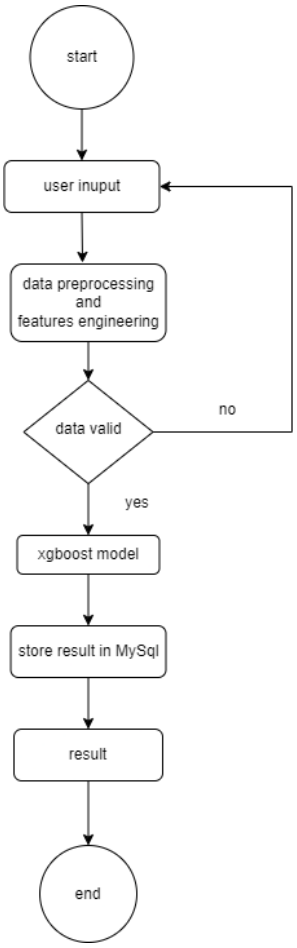
This is a great risk to the job seekers across the globe, especially in this era where millions depend on unemployment benefits. Scammers take advantage of this weakness by posting genuine-looking job adverts that may result in financial and personal harm. These risks can be reduced by educating applicants to take preventive measures like cross-checking company logos for their authenticity, an official email address should be used to communicate with them, and should not give unnecessary personal details at the interview level. EMSCAD: The Employment Scam Aegean Dataset has a total of 17,880 job postings for model development and testing purposes.

## III. Proposed Methodology

The following proposed methodology will make fraudulent job posting identification a realistic tool for the user in overcoming its drawbacks:

1. **Better Accuracy in Classification:**
- Advanced algorithms such as XGBOOST and DNN would be used in order to make the better prediction of fraudulent job postings.
- Techniques involved are feature engineering and hyperparameter optimization to hone the model further.
- Balancing the datasets and using cross-validation strategies for making the results reliable
2. **Database Implementation:**
- Design a MySQL database to safely store the information related to the job posting.
- Store all the major feature of the job detail systematically, like title, description, company name, and type of employment.
- Regularly update the dataset to ensure that it is completely inclusive and hence relevant for training and testing the model.
3. **Development of a User-Friendly Website:**
- Develop a web application that allows the user to upload information regarding the job posting so that the authenticity of the post may be ascertained.
- After the input submission the back-end process takes the information provided from the trained model and delivers results so that job postings appear to be authentic or scamming.
4. **Introduce a Feedback System:**
- Provide the user with the option to mark wrong predictions so that the model will be constantly improved.
- The feedback will be analysed and used to update the training dataset so that the model improves with time.

This methodology integrates machine learning, database management, and web development, which can develop a robust and user-centric system for effective detection of fraudulent job advertisements. This methodology finally combines technological advancements with practical usability in addressing critical issues in online recruitment.




### A. Data Set

The public dataset is referred to as the Employment Scam Aegean Dataset or EMSCAD. It is a collection of 17,880 job postings that assist in the detection of fraudulent job postings. Out of these job postings, 17,014 are legitimate and 866 are fake. The laboratory, University of the Aegean, curates it for researchers and developers so that it provides them with something useful to aid in distinguishing genuine job postings from fake ones. Both the textual and non-textual features of the EMSCAD are helpful for all analytical and modeling work.

Textual features for EMSCAD are job title, company profile, job description, required skills, and benefits, including other relevant text. These represent basic data needed to analyze postings for jobs done, thus bringing out linguistic aspects and anomalies susceptible to fraudulent moves. Non-text features include metadata which includes job type needed experience, needed education, industry, and location along with departments which accordingly give the contextual information such that can aid in the fraudulent job posting detection. Other non-text features are having or not a company logo, screening questions and telecommute option. These all are important points that could eventually lead towards detection of suspicious listings.

The primary objective of EMSCAD is to contribute to the development of machine learning models that could automatically identify fraudulent job postings. Further exploration of the features most predictive of fraud using the dataset will be possible after analyzing such data. Thus, textual and non-textual data both can be utilized with this approach. NLP can now be applied on the job description, and metadata can be dealt with by applying traditional methods of machine learning. This has been detected on textual analysis showing patterns in jobs applied for with the possibility being fraudulent, which is not directly analyzed by analyzing anomalies in non-textual as in the meta-data of that job applied.

### B. XGBoost Algorithm

Actually, XGBoost is a hell of a superhero when it comes to machine learning, especially with the data-spreadsheets and so on. One could consider it to be a cluster of very wise decision trees learning with each other, in the sense, attaining wisdom by experiencing the insight that is garnered from the earlier trees. It lets XGBoost classify really complex problems with such speed. XGBoost is optimized for speed and high performance. It has become the preferred choice of analysts in getting quick and reliable results. Among the most widely used algorithms in competitive machine learning today, flexibility and efficiency in both classification and regression tasks catapulted XGBoost to be one of the best algorithms today. XGBoost relies fundamentally on an ensemble of weak learners, where in this case decision trees, as a tool to make predictions, which have errors in themselves from the trees prior to it so that this may yield a very good predictive model.

The XGBoost algorithm iteratively adds trees to the model. It uses a base prediction, usually zero, and progressively adds trees, which try to reduce the residual errors from the preceding trees. In the learning procedure, gradient descent is used with the algorithm that minimizes the loss function by adjusting the weight of the trees. The three key components are:

1. **Gradient Boosting Framework:** It constructs successive trees using residuals from the previously constructed tree so as to make an attempt for correction of those residuals. And, this continues till accuracy of the model improves through iteration
2. **Regularization:** L1(Lasso) and L2(Ridge) regularization in XGBOOST. It prevents overfitting and generalizes well on new, unseen data.
3. **Missing Data:** Missing values are set at the starting level of XGBOOST. Hence, there cannot be a specific problem if any values are missing or incomplete in dataset.
4. **Parallelization:** XGBOOST supports parallel processing, which greatly accelerates training by distributing tasks across multiple processors, it is highly efficient when working with large amount of data
5. **Tree Pruning:** XGBOOST utilizes “depth-first” tree pruning that optimizes the tree structure so that it is neither overfitting and at the same time is simplified as much as possible but as accurate as possible.

## IV. EXPERMENTAL RESULT

In our paper "Unmasking Employment Scam," we applied some of the components to evaluate how well our predictive model would perform in identifying employment scams. The features involved feature extraction, confusion matrix, graphical representation of feature importance, and a developed webpage to present the interactive results.

### A. Confusion Matrix

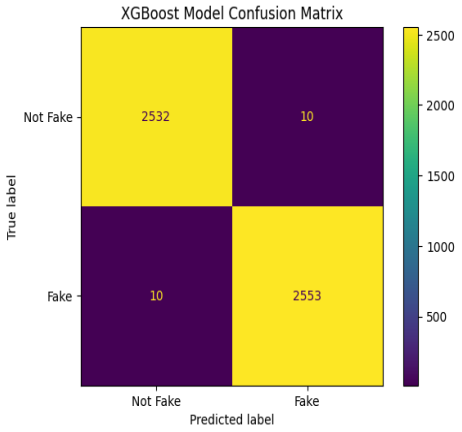



*Fig.1 confusion matrix*

One of the key performance metrics adopted in our paper was the confusion matrix. The confusion matrix allowed us to evaluate our classification model output in terms of predicting whether or not a posted job is valid or fabricated. It allowed detailed scrutiny of the model's true positives, false positives, true negatives, and false negatives in identifying genuine postings or fraudulent ones.

We used the confusion matrix to compute some key numbers in figuring out precision, recall, F1-score, and accuracy. Those numbers really painted the whole picture in which our model was able to classify things, thereby making it see not just where our model was good at, but rather also where our model was really weak and showed very clear pointers of what was supposed to be tweaked to increase its performance even better.

B. Feature Extraction and Graphical Representation

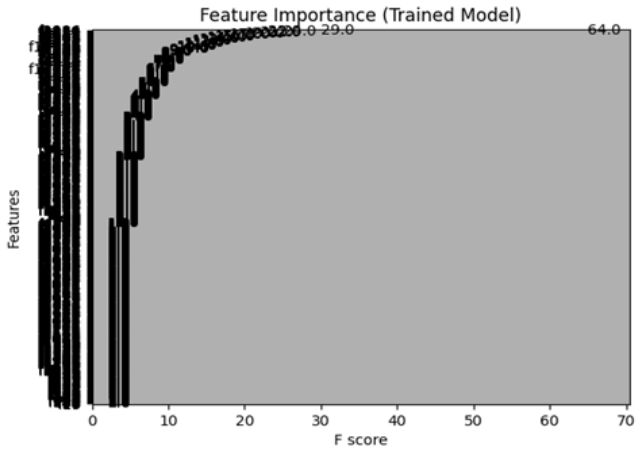



*Fig.2 feature extraction graph*

One of the major parts of our paper was feature extraction, through which we were able to convert raw data in the form of job postings into meaningful input for our machine learning model. We used various techniques to process both textual and non-textual features. The textual features such as job title, job description, and company profile were processed by applying techniques like TF-IDF, or Term Frequency-Inverse Document Frequency. In contrast, non-textual features were encoded in numerical form, which include job type, salary range, and whether the company logo exists.

To determine which of the features considered is highly relevant in determining fraudulent job postings, we plot the feature importance graph developed by our machine learning model, XGBoost. This graph will show which features most contributed to the decision of the model. For example, features like job title and company profile emerged as the most important in such cases, while other features like job location and salary range only enriched the model's predictions further. These visualizations have helped us better understand the logic behind the model's decisions and fine-tune the process of feature selection.

C. Model Accuracy

Through intense training and fine-tuning of our XGBoost model, we can achieve a 98% accuracy when predicting fraudulent job postings. That is an outstanding result, representing the high level of performance of the XGBoost algorithm in identifying valid and fake job postings with significantly high precision. Further cross-validation also vindicates its accuracy, ensuring that this model generalizes well for the unseen data rather than over-fitting for the training set. Also, all other evaluation criteria precision 97%, recall 99%, and F1-score 98% had set all of those performances of the model.

V. REFERENCE

[1] S. U. Habiba, M. K. Islam, and F. Tasnim, "A comparative study on fake job post prediction using different data mining techniques," *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 2021, pp. 543–546, doi: 10.1109/ICREST51555.2021.9331230.

[2] A. Amaar, W. Aljedaani, F. Rustam, *et al.*, "Detection of fake job postings by utilizing machine learning and natural language processing approaches," *Neural Processing Letters*, vol. 54, pp. 2219–2247, 2022. doi: 10.1007/s11063-021-10727-z.

[3] A. Mehboob and M. S. I. Malik, "Smart fraud detection framework for job recruitments," *Arab Journal of Science and Engineering*, vol. 46, pp. 3067–3078, 2021. doi: 10.1007/s13369-020-04998-2.

[4] D. Ranparia, S. Kumari, and A. Sahani, "Fake job prediction using sequential network," *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 2020, pp. 339–343. doi: 10.1109/ICIIS51140.2020.9342738.

[5] M. Sudhakar and K. P. Kaliyamurthie, "Efficient prediction of fake news using novel ensemble technique based on machine learning algorithm," in *Information and Communication Technology for Competitive Strategies (ICTCS 2021)*, M. S. Kaiser, J. Xie, and V. S. Rathore, Eds., Lecture Notes in Networks and Systems, vol. 401, Springer, Singapore, 2023.

[6] B. Alghamdi and F. Alharby, "An intelligent model for online recruitment fraud detection," *Journal of Information Security*, vol. 10, pp. 155–176, 2019. doi: 10.4236/iis.2019.103009.

[7] T. V. Huynh, K. V. Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Job prediction: From deep neural network models to applications," *RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020.

[8] J. Zhang, B. Dong, and P. S. Yu, "FAKEDETECTOR: Effective fake news detection with deep diffusive neural network," *IEEE 36th International Conference on Data Engineering (ICDE)*, 2020.

[9] J. Lee *et al.*, "Fake job detection using deep learning architectures," *International Joint Conference on Neural Networks*, 2019.

[10] Z. Wang *et al.*, "Fake job detection on social media: A graph-based approach," *IEEE Transactions on Knowledge and Data Engineering*, 2020.