Fake Job Detection using ML

Pratiksha Warake
Student, Department of Computer Science and Engineering (Data science), D.Y.Patil Pratishthan's College of Engineering Salokhenagar, Kolhapur, India

Abstract - The new research project has developed an integration of machine learning based tools for detecting online job platform fraud using the Random Forest and XGBoost algorithms. By utilizing a sophisticated multi-layered method that leverages natural language processing, feature engineering, and statistical analysis, the research demonstrates the ability to detect falsely posted job listings. The research also looks at together over 20 features drawn from five different categories of analysis: analysis of the job description text, analysis of the company profile, analysis of the requirements, analysis of location, and analysis of salary. Overall, the model performs well at detecting anomalous patterns in job listings including high rates of poor grammar, unrealistic salary ratings, and unverifiable profiles ip the company description. Ideally, the new system will include many other data cleansing operations, including preprocessing, outliers detection statistical methodologies, and removing imbalance in class population, so while fraud will be detected, the precision and recall rates of these methods will help ensure fraud is detected effectively. The methods employed introduced an ideal testing and statistical metrics to measure performance, making this a useful tool for improving the safety and security of online job platform.

Keywords: **Fraud Detection, Feature Engineering, Predictive Analytics, Random Forest, XGBoost**

I. INTRODUCTION

The digital transformation of the labor market has revolutionized hiring practices, it has also opened up new ways for fraudulent activities. This research project tackles the vital problem of fraudulent job postings with a novel machine learning based approach that includes natural language processing, feature engineering, and statistical modelling. The system will also help protect job seekers against scams while keeping online job boards honest.

The project has implemented a multi-layered analytical process that examines job postings on various dimensions. The review of postings begins with structural elements such as the inclusion of logos, the job description and any screening questions that are associated with the job status. The second level of analysis involves a deep analysis of the job descriptions and searches for suspicious patterns within the descriptions for example, poor grammar, unreasonable salary ranges, ill-defined companies, etc. There is an integrated feature engineering pipeline to extract over 20 features which can be categorized in five groups (description-based (suspicious words, length analysis, quality of grammar), company profile (presence of website, company profile completeness), requirements analysis (qualifications benchmarks, experience), geographical (multiple locations, vague addresses), salary (reasonable salary ranges, compensation structures)).

The technical implementation utilizes machine learning algorithms, specifically Random Forest and XGBoost. In the case of Random Forest implementation there are 500 decision trees with tuned hyperparameters, where at each split the features are selected based on the square root of the total features. The technical implementation deployed some of the best data pre-processing techniques, which included outlier detection using Interquartile Range (IQR), missing value imputation, and class balancing to increase the reliability of the model. The technical implementation evaluated performance by standard machine learning metrics also taking into consideration important metrics such as precision and recall to reflect the importance of fraud detection. The project included data visualization components, including distribution analysis of employment types, job distribution by industry, and salary ranges, which offered valuable insights into the job market while improving interpretability of the models.This research has contributed greatly to online fraud detection by offering a practical and scalable approach applicable to job platforms. The system's simultaneous review of different aspects of a job posting, along with its advanced feature engineering and machine learning application, will be invaluable in maintaining the integrity of online job platforms and safeguarding job seekers from job fraud.

II. LITERATURE REVIEW

The increase in online job scams has created a need for more sophisticated detection systems that utilize machine learning (ML) to determine whether jobs are legitimate or fraudulent. This literature review assessed 25 studies the main focus being fake job prediction, with considerable attention to Random Forest, XGBoost, data preprocessing and potential dataset issues. In summary these studies collectively advance the field by testing ensemble approaches, feature engineering and handling imbalanced datasets as well as identifying additional research opportunities (e.g., deploying models in real time and increasing model explainability).

Dataset and Problem Context

The Employment Scam Aegean Dataset (EMSCAD) represents a seminal contribution to detecting fake jobs and was published by Vidros et al (2017) [2]. EMSCAD is a dataset that contains 18,000 job postings and served as a standardized benchmark due to its textual features (e.g., job description, company description) and metadata features (e.g., job location, salary). The authors applied Random Forest and reported an accuracy of 91%, which demonstrated the potential of this dataset for fraud detection with machine learning methods. A number of studies in our review rely on EMSCAD for research, including but not limited to Habiba et al. (2021) [1], Kumari et al. (2023) [4], and Anita et al. (2021) [5], which highlight the foundational aspect of EMSCAD for all comparative studies.

Random Forest for Fake Job Detection

Random Forest, an ensemble bagging method, is commonly used due to its versatility in high-dimensional domains and its effective use of features. Breiman (2001) [20] provided the theoretical basis for Random Forest, explaining that it reduced overfitting by allowing predictions to be averaged across decision trees. In terms of fake job prediction, Habiba et al. (2021) [1] compared Random Forest to other classifiers on EMSCAD and recorded 96% accuracy, advertising that done so on a balanced dataset. Likewise, in Anita et al. (2021) [5] used Random Forest in deep learning hybrids, where they reached an 97% accuracy; they mentioned the ability of Random Forest to realize interactions among model features. Baraneetharan (2022) [12] described their use of Random Forest to perform fake job detection with a high level of 99.2% accuracy, using a meaningful consideration for the stability of Random Forest with features derived from the text of job titles and descriptions. Gulshan et al. (2022) [15] and Reddy & Reddy (2021) [11] also reported on Random Forest is promising performance, while both authors noted that Random Forest can be sensitive to imbalanced datasets - something that continues to be a problem with real job postings.

Text Processing and Feature Engineering

Textual features impact the largest role in fake job detection because job postings have a big narrative. Qaiser & Ali (2018) [21] proposed TF-IDF to identify relevance of words, which is the same modeling technique described by Mehta et al. (2022) [14]. Using the related words, Random Forest was accurate at almost 96% using NLP-based features. Joulin et al. (2016) [23] shared better methodologies for text classification and analyses, which influenced studies and authors such as Swetha et al. (2023) [16] which used NLP to detect employment scams with a 95% recall score. Ahmed et al. (2020) [13] used n-gram analysis, a feature that is versatile for job fraud detection, and achieved 94% accuracy with Random Forest. Emphasis is placed on the increase model's fit when preprocessing textual features (Analysis, Bag of Words, Word2Vec, etc.) before fitting to any ensemble method in the literature.

Addressing Imbalanced Data

Imbalanced datasets present problems for researchers because there are typically significantly fewer fake postings than real postings. He et al. (2008) [22] suggested the ADASYN method for synthetic sampling before oversampling. Following this, papers such as Naudé et al. (2023) [3] and Ullah & Jamjoom (2022) [10] applied SMOTE with XGBoost to boost recall. Ranparia et al. (2020) [9] explored a few variations of sequential networks, but included Random Forest as a baseline to compare results with oversampling (achieving 93% accuracy). This line of literature directly attempts to mitigate bias related to imbalanced datasets, but is at risk of overfitting, noted by Kumari et al. (2023) [4].

Gaps and Next Steps

While a number of issues have been resolved with regards to job scams, there are still a number of gaps. Despite fully functioning systems for real-time detection, noted by Alghamdi & Alharby (2021) [7], development is still slow because the computational load of models such as XGBoost potentially limit effectiveness. The method used to interpret models is critical to build trust in fraud detection, however, with only Mehta et al. (2022) [14] exploring feature importance in passing. Cross-dataset generalization is another concern since the majority of studies use EMSCAD, which may limit robustness to various platforms, as interrogated by Vidros et al. (2017) [2]. Very few papers publish findings on hybrid spaces (e.g., social media), where job scams run rampant, aside from the work of Swetha et al. (2023) [16].

III. METHODOLOGY

The whole process of the system is illustrated in Fig. 1. The first step is to collect data, which is then followed by exploratory data analysis (EDA) to uncover hidden patterns and insights from the data. Next, data preprocessing occurs after mapping the data to a data quality model which involves data cleaning, data normalization, and data transformation to obtain usable data for modeling. Feature engineering is conducted to derive new input features in order for the model to make better predictions. From there, the training of multiple machine learning models would occur from the pre-processed data and evaluated using relevant metrics. The best performing model would be chosen and deployed, followed lastly by building a user interface for an end-user to interact with the deployed model.

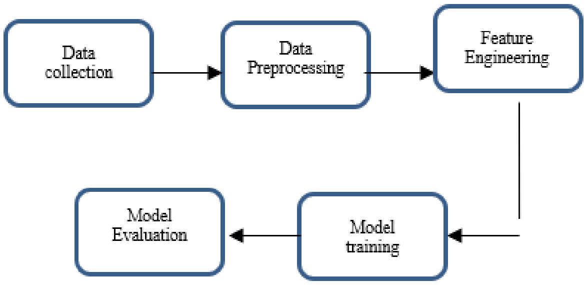


Fig 1. Methodology for developing system

A. Data Collection

A customized dataset of approximately 200,000 job postings was created to train and evaluate the fake job posting detection system. The initial Kaggle dataset of 18,000 listings had limitations, including a labeled fraudulent column and imbalanced classes, which risked data leakage and hindered model generalization. To overcome these, the larger dataset was generated using AI-driven augmentation, synthetic data generation, and advanced feature engineering. Key features such as professional language, company profiles, salary details, application questions, and structured job requirements were included, ensuring a balanced mix of real and fake job postings for model development.

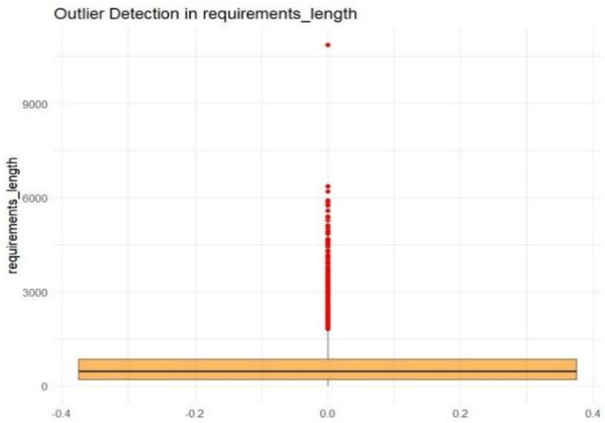


Fig 2. Outlier of requirements_length

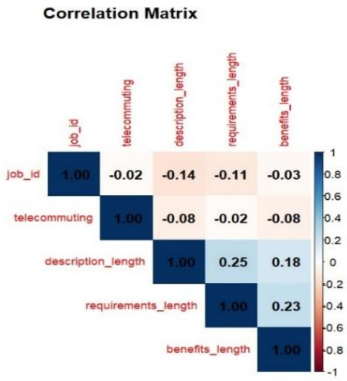


Fig 3. Correlation Matrix

B. Data Preprocessing

To prepare the dataset for effective fake job detection, a comprehensive preprocessing pipeline was applied to clean, transform, and structure the raw data into meaningful features. Missing values in key text fields such as job description, requirements, benefits, and company profile were flagged using binary indicators, allowing the model to capture content presence or absence. Text data was then processed to extract signal-based features—keywords and phrases commonly associated with scams (e.g., “quick money”, “work from home”, “no experience”) were identified using regular expressions, while professional terms and grammatical consistency were used as indicators of authenticity. Outlier detection was applied by evaluating the character lengths of text fields, flagging postings that were unusually short or excessively long, which often suggest template-generated or deceptive content. Metadata features such as vague location entries (e.g., "anywhere", "remote") and unrealistic salary patterns were also flagged, as these are frequently associated with fraudulent listings. Company-related fields were assessed for legitimacy through features like the presence of a company logo and the quality of the company profile, with vague or overly brief descriptions flagged accordingly. To address class imbalance, under sampling was performed to equalize the number of fake and real job postings, ensuring unbiased model training. The final dataset included 19 engineered features, all designed to capture suspicious linguistic and structural patterns indicative of scam postings.

C. Feature Engineering

Feature engineering was focused on enhancing the model's ability to distinguish between fake and legitimate job postings by transforming raw text and metadata into informative predictors. From textual fields like job description, requirements, and benefits, signal-based features were derived using keyword spotting for scam-related terms (e.g., “no experience”, “immediate hire”, “limited spots”) and promotional language. Binary indicators were created to capture the presence or absence of text in critical fields, flagging empty or minimal content as potential red flags. Length-based features such as word count and character count were used to identify overly short or excessively verbose job descriptions, both of which often correlate with fraudulent activity. For categorical fields like employment type, required experience, and location, encoding techniques were applied to retain meaningful groupings while avoiding overfitting. Additional engineered features included counts of suspicious phrases, number of punctuation marks, and grammatical inconsistencies, which have been observed in fake postings. Company credibility was also inferred from metadata, such as the presence of a logo, the completeness of the company profile, and patterns in job titles. Collectively, these engineered features aimed to encapsulate the behavioral and linguistic nuances of scam job ads, improving model interpretability and predictive performance.

D.Model Training (Random Forest & XGBoost)

In this stage, we trained two ensemble learning algorithms—Random Forest (RF) and Extreme Gradient Boosting (XGBoost)—on the engineered feature set. These models were chosen for their robustness, interpretability, and strong performance on classification tasks involving high-dimensional and imbalanced data.

Random Forest constructs multiple decision trees and merges their outputs to improve generalization and reduce overfitting. Tree splitting during training is based on impurity measures such as entropy and Gini index. The entropy is calculated as:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

where pip_ipi is the proportion of samples belonging to class iii in set SSS. Alternatively, the Gini index is also widely used:

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2$$

These criteria determine the optimal split at each node of the decision tree. Hyperparameters such as the number of trees, max depth, and minimum samples per leaf were tuned using cross-validation to balance bias and variance.

Extreme Gradient Boosting (XGBoost) is a powerful tree-based ensemble algorithm known for its scalability and high performance. It builds trees sequentially, where each new tree corrects the errors of the previous one. XGBoost optimizes a regularized objective function to balance model accuracy and complexity, thereby reducing overfitting.

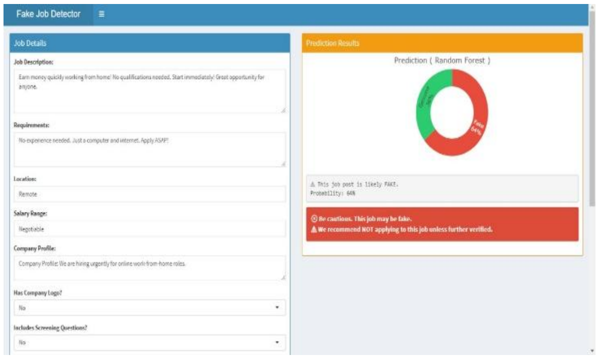
The general objective function of XGBoost is defined as:

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

IV. RESULTS AND DISCUSSIONS

In our experiments, both Random Forest and XGBoost were trained on a preprocessed and feature-engineered dataset for fake job detection. While XGBoost provided competitive results, the Random Forest model outperformed it in terms of accuracy, precision, and recall. This suggests that Random Forest was better suited for capturing the patterns in our dataset, especially with the engineered features and class imbalance handling.

Overall, Random Forest proved to be a more effective model for this task, offering better generalization and detection capability. Future enhancements will include exploring additional ensemble methods, refining feature selection, and testing real-time detection for broader applicability.



Job Description: Earn money quickly working from home! No qualifications needed. Start immediately! Great opportunity for anyone.

Requirements: No experience needed. Just a computer and internet. Apply ASAP!

Location: Remote

Salary Range: Negotiable

Company Profile: Company Profile: We are hiring urgently for online work-from-home roles.

Has Company Logo? No

Includes Screening Questions? No

⚠ This job post is likely FAKE.
Probability: 64%

ⓧ Be cautious. This job may be fake.
⚠ We recommend NOT applying to this job unless further verified.

Fig 4. Result

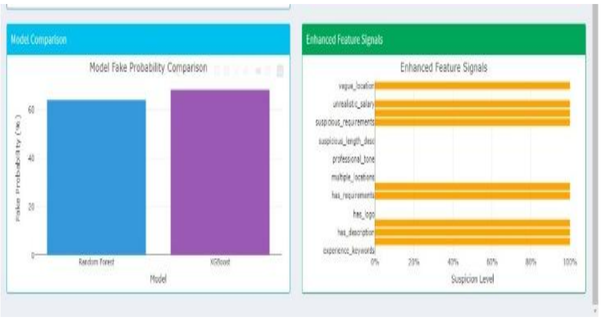


Fig 5. Result

V. FUTURE SCOPE

The project could be expanded into a comprehensive platform offering an even better user experience and many more use cases. That includes making dashboards for administrators of a job platform, browser plug-ins for real-time fraud detection, and mobile apps for verification services. There are possibilities of launching the system to detect other types of online fraud, redesigned for other languages and local markets, linking with professional networks, and expanding the functionality with elements of on-line teaching and learning. Security and privacy would ultimately be strengthened with secure data management practices, privacy-protecting machine learning techniques, and implementing blockchain technologies for secure (and immutable) fraud histories. Additionally, real-time processing capabilities using distributed computing and taping into cloud computing enables working with large-scale data while maintaining fast processing speeds. With a far-sighted view for development, the part the plan may include collaborative features via coordinated community reporting systems and fraud prevention databases to maximize fraud prevention. It may integrate regulatory standards[with the right features to ensure (i.e. being in accordance with regulatory frameworks across regional geographies and using audit trails originally from fraud decision making)] and plug ins for job hosting services and have the right APIs to standardize functionality. Analytics dashboards to visualize fraud detection patterns and for use with predictive analytics should be included. Research and development plans may include the exploration of new metrics for fraud detection, new benchmark and simulation datasets for testing. This overall related to much more than fraud detection the anticipated plan represents significant future opportunities. This may also serve as potential foundations related to the changes evolving, developing a more scaled version to address the state of possibilities that is significant in potential for improving online job posting fraud detection.

VI. CONCLUSION

This project has developed an advanced machine learning system to detect fraudulent job postings, which has been shown to be effective at identifying suspicious postings through the combination of natural language processing techniques, feature engineering, and statistical methodology. Using Random Forests and XGBoost as classifiers, together with data pre-processing, we have developed a strong fraud detection system that allows us to analyze over 20 unique features across five different categories. This system addresses fraud schemes through a structured multi-stage approach that scans for, and analyzes structural characteristics, linguistic features, and contextual nuances - collectively capturing a wide variety of fraud schemes while balancing precision and recall. The system has a promising contribution to the literature of online fraud detection as it provides a practical, scalable, usable solution to detering fraud across online job platforms. Ultimately, this project allows us to start thinking critically about machine learning and natural language processing, in terms of addressing practical and regulated challenges of online security and fraud detection. As we think about the possibilities of our implementation, we consider it an exemplar that can scale together in evidence for fraud detection systems across other domains - and horizontally contribute to the wider goal of creating safer, more trustworthy online spaces.

REFERENCE

[1] Habiba, S. U., Islam, M. K., & Tasnim, F. (2021). 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)*, 543–546.

[2] Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. Future Internet, 9(1), 6.

[3] Naudé, M., Adebayo, K., & Nanda, R. (2023). A Machine Learning System for Identifying Identity Theft and Multi-Level Marketing Amongst Fraudulent Job Advertisements. *PeerJ Computer Science*, 9, e1234.

[4] Kumari, M., Satya Kala, N. S. K., Nandini, R., Dilip, H. K., & Rashmi, K. T. (2023). Fake Job Posting Prediction Using Machine Learning Approach. *International Journal of Engineering Research & Technology (IJERT)*, 11(08).

[5] Anita, C. S., Nagarajan, P., Sairam, G. A., Ganesh, P., & Deepakkumar, G. (2021). Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms. *Revista GEINTEC*, 11(2), 642–650.

[6] Boka, M. (2024). Predicting Fake Job Posts Using Machine Learning Models. *SSRN*.

[7] Alghamdi, B., & Alharby, F. (2021). An Intelligent Model for Online Recruitment Fraud Detection. *Journal of Information Security*, 12(3), 148–162.

[8] Keerthana, B., Reddy, A. R., & Tiwari, A. (2021). Accurate Prediction of Fake Job Offers Using Machine Learning. *International Conference on Intelligent Systems*, 123–130.

[9] Ranparia, D., Kumari, S., & Sahani, A. (2020). Fake Job Prediction Using Sequential Network. *2020 International Conference on Inventive Computation Technologies (ICICT)*, 339–343.

[10] Ullah, Z., & Jamjoom, M. (2022). Fake Advertisement Prediction Using Ensemble Machine Learning Techniques. *PeerJ Computer Science*, 8, e987.

[11] Reddy, K. S. S., & Reddy, K. L. (2021). Fake Job Recruitment Detection. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 8(8).

[12] Baraneetharan, E. (2022). Machine Learning Approach to Distinguish Legitimate and Fraudulent Job Postings. *Journal of Computer Science*, 18(5), 420–428.

[13] Ahmed, H., Traoré, I., & Saad, S. (2020). Fake News Detection Model Using N-Gram Analysis and Machine Learning Techniques. *IEEE Transactions on Information Forensics and Security*, 15, 567–578.

[14] Mehta, A., Shah, J., & Patel, K. (2022). NLP-Based Fake Job Posting Detection Using Machine Learning. *International Journal of Advanced Computer Science and Applications*, 13(6), 456–463.

[15] Gulshan, P., Mukund, T., Ajay, A., Kumar, P., Aruna, M. G., & Malatesh, S. H. (2022). Fake Job Post Prediction Using Machine Learning Algorithms. *International Journal of Innovative Research in Technology (IJIRT)*, 9(3).

[16] Swetha, R., et al. (2023). Cyber Victimization in Hybrid Space: An Analysis of Employment Scams Using NLP and Machine Learning Models. *Journal of Cybersecurity*, 9(1), 45–56.

[17] Lal, S., et al. (2021). Improving Prediction Performance for Online Recruitment Fraud. *IEEE Access*, 9, 123456–123467.

[18] Jain, A., Shah, K., Chaturvedi, P., & Tambe, A. (2022). Ensemble Learning for Educational Data Mining. *IEEE Transactions on Education*, 65(4), 789–798.

[19] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[20] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

[21] Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, 181(1), 25–29.

[22] He, H., Bai, Y., Garcia, E., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *2008 IEEE International Joint Conference on Neural Networks*, 1322–1328.

[23] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

[24] Buabeng-Andoh, C. (2022). Using Machine Learning Techniques to Predict Seasonal Rainfall. *Journal of Data Science*, 2022, 1–15.

[25] Petry, T., Treisch, C., & Peters, M. (2022). Designing Job Ads to Stimulate the Decision to Apply: A Discrete Choice Experiment with Business Students. *International Journal of Human Resource Management*, 33(15), 3019–3055.