Comprehensive Detection of Fake Job Postings: Applying Advanced Data Balancing and Machine Learning Techniques to Combat Online Frauds

Mohan Allam
School of Computer Science and Engineering, VIT-AP,
Amaravati, India
mohan.allam@vitap.ac.in

Vandana Chintala
School of Computer Science and Engineering, VIT-AP,
Amaravati, India
vandanachintala2004@gmail.com

Gayatri Akshaya V
School of Computer Science and Engineering, VIT-AP,
Amaravati, India
akshayavobbilisetti@gmail.com

Varshitha Appikonda
School of Computer Science and Engineering, VIT-AP,
Amaravati, India
varshi.raj04@gmail.com

Ruchitha Gandi
School of Computer Science and Engineering, VIT-AP,
Amaravati, India
gandiruchitha@gmail.com

2025 6th International Conference on Recent Advances in Information Technology (RAIT) | 979-8-3315-2480-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/RAIT65068.2025.11089425

***Abstract*— Fake job postings have become an increasingly common online crisis now-a-days, exploiting the money and identities of individuals seeking employment via online platforms. Genuine job postings are more in number compared to fake job postings; this imbalanced nature of the data makes it even harder to detect the fake job postings. This paper addresses the challenge of identifying fraudulent job postings by applying various data balancing techniques to counteract the class imbalance inherent in the dataset in combination with various state-of-art machine learning techniques. Methods like TF-IDF, categorical data are utilized for feature extraction and the impact of multiple balancing strategies, including SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous), Cluster-Based Oversampling (Cluster- SMOTE) for data augmentation, and cost-sensitive learning is analyzed. After applying these techniques, we use several types of machine learning models on obtained balanced datasets. A comparative analysis of various machine learning models is performed alongside their corresponding data balancing techniques. Our results demonstrate the effectiveness of various balancing techniques in improving model performance, offering a more robust solution to combat online scams like fake job postings.**

***Keywords*— Fake Job Postings, Online Recruitment Fraud, Class Imbalance Handling, Data Balancing Techniques, Machine Learning Models, Fraud Detection Systems**

I. Introduction

The rise of online job platforms has provided both opportunities and challenges for people who are seeking jobs and employers. While these platforms offer convenience of quick search for jobs from comfort of home, they also serve as potential platform for fraudulent job postings that deceive vulnerable job applicants. Fake job postings can lead to financial loss, identity theft, and exploitation of personal data. Traditional machine learning methods often struggle with the imbalanced nature of this problem, where genuine job postings vastly outnumber fake ones. This is where data balancing comes into light.

There is a remarkable 30% increase in instances of online job fraud in 2022 alone, explaining the underlying expanding danger of fraudulent job postings. Approximately 40 percent of job applicants have come across fraudulent job advertising while searching, according to recent data [1]. In some situations, job applicants were requested to give valuable personal information and pay some amount of money to secure a job position all while scamming people with no guaranteed job after the payment and wrongly utilizing the personal information of job applicants. These unfortunate scenarios and dreadful statistics are the main concerns of modern-world job postings. Combating fake job postings should include creating awareness among applicants regarding fraudulent job postings and introducing efficient prediction models. It is everyone's responsibility in creating more trustworthy job-hunting environment for the upcoming generations as we all move towards digital world.

This paper explores a comprehensive approach to detecting fake job postings by applying various advanced data balancing techniques on imbalanced datasets and applying robust machine learning models to enhance detection accuracy and reduce false positives along with a comparative analysis of existing methodologies.

II. Literature Survey

To increase the efficiency of the machine learning model in predicting the fraudulence of job postings, it is vital to perform any data balancing technique on the dataset and then apply the proposed machine learning or deep learning model. Bauder et al. [2] aim to detect Medicare fraud cases in a highly imbalanced dataset where the ratio of righteous to fraud cases is very small. They created a combined dataset from various medical equipment datasets and mapped real-world fraud labels from Database "LEIE", where 0.062% are labelled as fraud, which significantly shows class imbalance. To handle this, they experimented with various sampling techniques like Random Under Sampling (RUS), Random Over Sampling (ROS), SMOTE (Synthetic Minority Oversampling Technique) and used three machine learning models (Logistic Regression, Random Forest, and Gradient Boosting Trees) on Apache Spark to evaluate model performance. Results show that RUS performed well with Random Forest particularly, and their study concludes that under sampling can be effective when there is a severe imbalance in large datasets compared to oversampling which may not improve efficiency in fraud detection.

Lokku et al. [3] addresses the problem of bogus job advertisements, which are increasing day by day and attempt to steal sensitive information. This study also employs the Employment Scam Aegean Dataset(EMSCAD) dataset from Kaggle. Data is pre-processed using tokenization, stopwords removal, and lemmatization. TF-IDF vectorization was used to extract features from text, followed by SMOTE data resampling and Random Forest Classification. Results showed that RF achieved 99% accuracy in predicting fake job postings, and SMOTE and TF-IDF significantly improved model performance by addressing data imbalance.

Hanif et al. [4], conducted an experiment on the EMSCAD dataset by implementing some machine learning models like DT, LR, SVM, and Naive Bayes that can effectively detect fraudulent job postings. The experiment was conducted on an imbalanced and balanced dataset. They handled class imbalance using RUS. The process started off by preprocessing the data, which is followed by vectorization using three techniques (TF-IDF, BoW, and Hashing) and implementing and evaluating four classifiers (DT, LR, SVM, NB). The DT with Bow performed well on a balanced dataset with accuracy of 0.705, precision of 0.73, recall of 0.70, F1-score of 0.71, and an AUC of 0.68. While coming to an imbalanced dataset, it outperformed LR and SVM by performing well in multiple metrics. This study concludes that DT combined with BoW is effective in detecting fraudulent job postings and suggests applying advanced text processing and different resampling strategies to increase prediction accuracy.

Zeno Gantner et al. [5] used methodologies like S-CSL, which combines resampling methods (like SMOTE, random under sampling/oversampling) with Cost Sensitive Learning Method (CSL) using Support Vector Machines (SVM) as the base classifier. This approach reduced misclassification costs on most datasets, improving the classifier’s focus on the minority class. Another methodology is CSL-OCRL (Optimizing Cost Ratio Locally), which optimizes the cost ratio dynamically as a hyperparameter rather than fixing it, training a model with an optimized balance between false positives and false negatives. This method enhanced the GMean metric, especially beneficial in datasets with severe imbalance. Experimental results on 18 datasets demonstrated that both methods reduced misclassification costs and improved performance metrics such as GMean.

DeepSMOTE outperformed state-of-the-art oversampling and GAN-based (Generative Adversarial Networks) methods across multiple imbalanced datasets (MNIST, CIFAR-10, etc.), improving metrics like average class accuracy and F1-score. It demonstrated robustness in highly imbalanced scenarios, especially in image-based datasets, according to [6]. They combined an encoder-decoder framework with the Synthetic Minority Oversampling Technique (SMOTE). This approach enables DeepSMOTE to generate high-quality synthetic samples for minority classes without needing a GAN discriminator. It consists of three main components: an encoder/decoder model, SMOTE-based oversampling, and a specialized loss function with a penalty term.

A Review on Handling Imbalanced Data [7] showcased that when using various Data-level methods, Algorithmic-level methods, Hybrid methods, and techniques like oversampling (e.g., SMOTE), under sampling, and feature selection, oversampling adds synthetic samples to the minority class, while under sampling removes samples from the majority class. SMOTE is highlighted for its effectiveness, though it risks overfitting. Internal techniques, such as cost-sensitive learning and ensemble methods like boosting, are used to adapt algorithms and decrease class imbalance biases. Data and algorithmic techniques are used to enhance classification performance, with hybrid models demonstrating promise in both generalization and adaptability. Precisely indicating that SMOTE often performs well than other algorithmic-level and hybrid methods.

The study in [8] evaluates ensemble-based models like Random Forest, AdaBoost, and Gradient Boosting with other classifiers, such as Naive Bayes, Multi-layer Perceptron (MLP), K-Nearest Neighbour (KNN), and Decision Tree classifiers. Among individual classifiers, Decision Tree achieved the highest accuracy (96.95%) and F1 score (0.98), while the Random Forest classifier outperformed others in the ensemble methods with a notable accuracy of 98.27%. The study also experimented with the Stochastic Gradient Descent classifier, finding it suitable for their pipeline and achieving high precision and recall. Random Forest classifier showed the highest performance with an accuracy of 98.27%, followed closely by AdaBoost and Gradient Boosting. The final model was able to achieve 99.50% accuracy in identifying real jobs and 85.56% accuracy in detecting fake jobs within the validation set.

Numerous deep learning techniques are also used for detecting fake job postings in many researches. Aljedaani et al. recommended a methodology for detecting fraudulent job posts utilizing natural language processing and several supervised machine learning algorithms, following feature engineering with methods such as TF-IDF and Bag-of-Words (BoW). Using the Adaptive Synthetic Sampling Approach (ADASYN) for over-sampling and TF-IDF for feature extraction, ”ETC” produced astonishing results of 99.9% accuracy. Furthermore, this work conducts an in-depth comparative examination of state-of-the-art deep learning models and various re-sampling strategies [9].

In [10], various deep neural networks like TextCNN, Bi-GRU-LSTM-CNN, and Bi-GRU-CNN with various pre-trained word embeddings on an IT job dataset are used for identification of bogus job postings. An ensemble model is also proposed, which achieved an F1 score of 72.21%. Multinomial Naive Bayes, Support Vector Machine (SVM), Decision Tree, K Nearest Neighbors, and Random Forest techniques are employed to the job postings dataset and the results obtained are weighed against each other in [11]. Bandar et al. [12] proposed a reliable intelligent detection model that used SVM method for feature selection and for classification and detection, ensemble classifier using Random Forest is employed. The EMSCAD dataset yielded an accuracy of 97.41%. They also highlighted the most significant elements and variables in detection, such as having a company profile, a corporate logo, and an industry feature. The study [13] examines the use of machine learning, specifically deep neural networks, to predict employment outcomes using a variety of data sources. In the context of your work on detecting bogus job posts, the literature review can emphasize machine learning approaches for classification and prediction, particularly those that focus on text data analysis. We can look at ways that use feature extraction, classification models, and deep learning to find patterns that indicate the validity of job posts.

J. Zhang et al. [14] investigate a unique strategy for identifying false news using a deep learning model, with an emphasis on the use of a Deep Diffusive Neural Net- work (D-DNN). The study provides a model that combines many variables for successful classification and emphasizes the use of deep learning in spotting patterns that indicate fraudulent material. This document might help with comparative analysis paper on fake job posting detection by explaining deep learning approaches for classification and feature extraction in identifying fraudulent material.

The paper [15] focuses on detecting bogus job advertisements with machine learning approaches. It discusses the growing problem of online recruitment fraud (ORF) and the difficulties in spotting fake employment offers, particularly those with class imbalances. The study focuses on the use of classifiers, both single and ensemble techniques, to detect bogus postings, with the Random Forest algorithm demon- starting the best accuracy. This technique may be compared to bogus job posting detection algorithms, particularly those relying on machine learning classifiers.

III. METHODOLOGY

Our workflow for fraudulent detection starts with importing the necessary libraries and datasets. It then goes through data preprocessing and data cleaning to deal with noise, missing values, and inconsistencies. The majority and minority instances are determined by analyzing the dataset. A balanced dataset is produced by applying data balancing techniques to rectify class imbalances. The model is implemented using this balanced data, and its performance is then evaluated through training and evaluation. Result and Analysis, which documents insights and performance data, marks the process's conclusion. The workflow of our proposed research is mentioned in Fig. 1.

A. Data Collection

"Kaggle" provides the dataset of fake job postings. The collection includes information from 17881 job listings and parameters like company profile, industry, salary range, location, company description, employment type, required experience, and company logo details. The dataset is very imbalanced, with only 4% of total job postings being fraudulent. The density distribution of job postings in the dataset can be observed in Fig. 2. This level imbalance in data if unresolved leads to inefficient detection and tampers with the accuracy of the machine learning model. In this context, data balancing is used to obtain accurate results and to improve efficiency of our model. Data is thoroughly cleaned and pre-processed before applying various data balancing techniques.

B. Data Balancing

Data balancing is very crucial in handling the imbalanced dataset, which leads to poor performance of the model on the minority instances, and the model gets more biased towards the majority class. Balancing the class weights addresses this problem and ensures more prediction accuracy in all classes, which is crucial in fields like medical diagnoses and fraudulent detection.

1) SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous)

SMOTE-NC, an extension of the SMOTE approach, generates synthetic samples to handle unbalanced datasets that contain both categorical and numerical variables. For numerical features, it computes weighted differences between original samples and their nearest neighbours, whereas for categorical features, it assigns values based on the nearest neighbour mode. This method successfully balances class distribution, resulting in realistic synthetic examples that decrease overfit- ting, increase generalization, and improve model accuracy, especially for binary and multi-class classification applications. It is notably useful for datasets including fake job advertisements, where class imbalance and diverse feature types pose substantial issues.

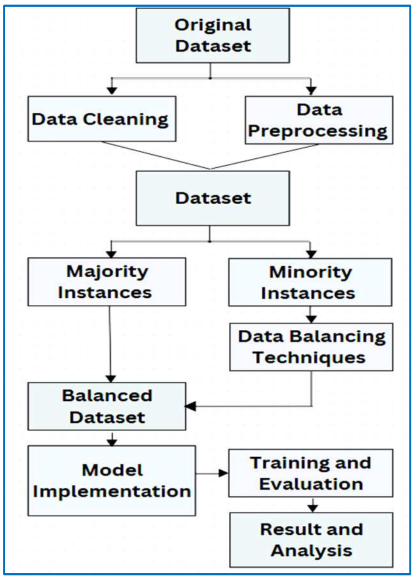

Fig. 1. Data Processing and Model Workflow

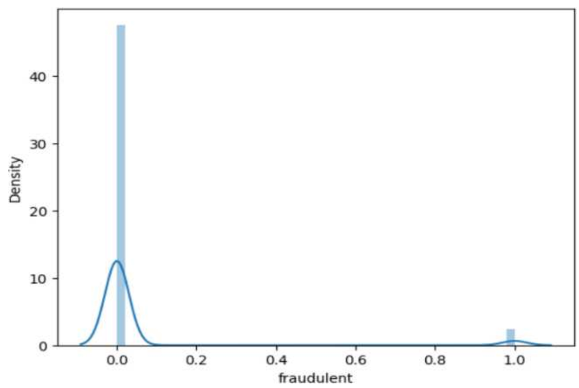

Fig. 2. Density distribution of job postings legitimate (0) fraudulent (1)

2) Cost-sensitive learning

Cost-sensitive learning (CSL) rectifies class imbalance by integrating misclassification costs directly into the learning process. In real-world applications, misclassifying minority class instances is more costly than errors in the majority class.

Cost-sensitive methods adjusts the learning algorithm by either changing the algorithm itself or changing the training data via re- weighting or resampling based on misclassification costs. This allows models to prioritize eliminating costly errors, resulting in improved performance on imbalanced datasets. CSL is especially useful in areas such as healthcare diagnosis and fraud detection, where false negatives can be significantly worse than false positives. By focusing on correctly identifying minority occurrences, CSL develops models that seek high overall accuracy while also reducing certain types of costly misclassification, which is critical when minimizing overall misclassification cost.

3) Ensemble methods with balanced bagging or balanced boosting

Ensemble methods that use balanced bagging and boosting maximize classification performance on im- balanced datasets by integrating different models and focusing on the minority class. Bootstrap samples are constructed using over-sampled minority and under- sampled majority cases, resulting in a balanced subset for each model. In balanced boosting, minority class instances are given more weight during training, in- structing each subsequent model to pay more attention to difficult-to-classify minority samples and reducing prediction bias. These techniques serve to counterbalance model's tendency to favor the majority class, which improves the detection of minority patterns. Balanced bagging and boosting are useful in applications like fraud detection and medical diagnostics because they produce far more precise and unbiased predictions on imbalanced datasets.

4) Deep learning-based data augmentation

Deep learning-based data augmentation creates synthetic data for the minority class in imbalanced datasets using deep learning models like Generative Adversarial Networks (GANs) or auto-encoders, which capture the minority class's underlying distribution and provide realistic data samples. Unlike traditional methods that use transformations such as rotation or scaling, deep learning-based techniques generate new instances that are similar to real data. This method is especially useful in complicated domains like image or text classification, where it enhances minority class representation, hence enhancing model performance.

5) Cluster-based oversampling (Cluster-SMOTE)

Cluster-based oversampling (Cluster-SMOTE) combines clustering and SMOTE to rectify class imbalance by creating synthetic samples from clusters of minority class occurrences. Minority samples are clustered using clustering methods such as k-means, and SMOTE generates synthetic samples inside each cluster while retaining the local data structure and providing realistic instances. This strategy is especially useful when the minority class has a complicated or scattered distribution since it improves the model's capacity to generalize to the minority class, decreases overfitting risk, and gives more representative samples, particularly in high-dimensional datasets. Cluster-SMOTE successfully balances precision and recall in difficult categorization problems.

6) Hybrid approach

When dealing with unbalanced datasets, the hybrid approach combines several techniques to maximize their individual strengths and enhance classification performance. It frequently combines ensemble approaches, cost-sensitive learning, under-sampling methods, and oversampling techniques like SMOTE. By integrating these methods, the hybrid approach overcomes the drawbacks of each technique alone. To create a balanced training set and modify the learning algorithm to concentrate more on the minority class, it might, for example, create synthetic samples for the minority class while decreasing the size of the majority class. This method balances the dataset, maximizes learning, and improves the classifier's capacity to correctly anticipate minority instances, making it particularly useful when no one methodology is adequate to handle class imbalance. The hybrid approach increases the diversity of training samples. Data balancing and model evaluation pipeline can be seen under Fig.3

C. Applying Machine Learning Models

Machine learning models ensure better performance over time, which are automated and constantly enhanced as new data becomes available. These models are perfect for applications that need great accuracy and efficiency in decision- making processes since they are also very scalable and can process enormous volumes of data in real-time without the need for human interaction.

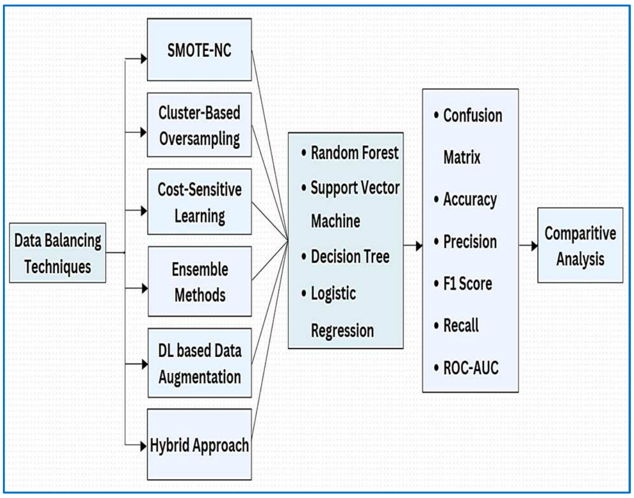

Fig. 3. Data Balancing and Model Evaluation Pipeline

1) Random Forest

Random Forest is an ensemble learning technique. Several decision trees are constructed using the Random Forest ensemble learning technique, which then com- bines them to increase classification accuracy. Through bootstrapping, or random sampling with replacement, a set of decision trees is created, and predictions are based on the trees' majority vote. A random selection of characteristics is used to train each tree, which enhances model generalization and lessens overfitting. Because Random Forest is resistant to overfitting and can produce accurate predictions even in cases of severe class imbalance, it is very useful for addressing unbalanced datasets. Because Random Forest excels at managing intricate, non-linear relationships in the data, it is perfect for identifying minute patterns that differentiate authentic job posts from fake ones. It provides a strong classification model that works well.

2) SUPPORT VECTOR MACHINE (SVM)

Support vector machine (SVM) is a potent supervised learning approach for classification and regression applications. SVM's capacity to handle non-linearly separable data using kernel functions and perform well in high-dimensional areas is one of its main advantages. Finding the best hyperplane to maximize the margin between classes is how SVM operates. By using kernel functions, it can efficiently handle both linear and non- linear classification problems. By applying varying penalties (costs) to instances of misclassification from the minority class, SVM can be modified to operate better on skewed datasets and become more sensitive to class imbalance. SVM excels in exact classification, outperforming other approaches with high-dimensional or non-linear data, making it perfect for unbalanced and complicated problems.

3) DECISION TREES

Decision trees are a basic yet effective classification technique that divides data into subsets depending on feature values, resulting in a tree-like structure. Each node represents a choice based on a trait, whereas the branches indicate potential outcomes. Decision trees are simple to understand and can process both numerical and categorical input. However, they are prone to overfitting, particularly in unbalanced datasets. Pruning, cost-sensitive learning, and altering class weights may all be used to improve performance on unbalanced data. Furthermore, Decision Trees are non-parametric models, which means they do not make assumptions about the underlying data distribution, allowing them to accurately describe complicated relationships.

4) LOGISTIC REGRESSION

Logistic Regression is a binary classification statistical model that uses predictor variables and a logistic function to determine the likelihood of a binary outcome, with values ranging from 0 to 1. While it is simple, interpretable, and computationally efficient, it can be affected by class imbalance. Techniques like changing class weights and regularization can help it perform better on imbalanced datasets. Logistic regression works effectively when there is a linear connection between features and the target variable, making it a dependable and successful baseline model in machine learning.

The Random Forest Classifier was a standout choice, integrating numerous Decision Trees to boost accuracy and robustness while handling unbalanced datasets efficiently. It also supplied feature significance scores, which improved interpretability. The Support Vector Machine (SVM) was chosen for its capacity to segregate classes using ideal hyperplanes, which makes it successful in high-dimensional spaces but computationally expensive for huge datasets. The Decision Tree Classifier provided simplicity and visual understanding, making it ideal for investigating data patterns, but it was prone to overfitting when used alone. Finally, Logistic Regression, a simple statistical model, was added due to its ease of use and good performance in linear connections, especially on smaller datasets. Among them, the Random Forest model proved to be the most successful, with the best mix of accuracy and generalizability for detecting fake job posts.

IV. RESULTS

Evaluation metrics like F1, Precision, Recall, ROC AUC are used to evaluate the performance of various machine learning models when combined with various data balancing techniques. The comparative analysis of evaluation metrics of various ML models along with type of data balancing technique used is showcased in Table I.

Combination of Random Forest with Deep Learning-Based Data Augmentation delivers the highest performance metrics, with 98.6% precision, 97.1% recall, and an F1- score of 84.4%. This combination stands out because Deep Learning-Based Data Augmentation synthesizes new, realistic samples by leveraging neural networks, which captures complex patterns in both the minority and majority classes. This method enriches the training data, reducing bias toward the majority class and enabling the Random Forest model to distinguish between classes more accurately.

V. CONCLUSION

Techniques like Deep Learning-Based Data Augmentation and the Hybrid approach (SMOTE, ADASYN) paired with Random Forest yield the best results, with balanced precision, recall, and F1-scores. Decision Tree also performs well, especially with Cost-Sensitive Learning and Deep Learning- Based Data Augmentation, though its scores are slightly lower than Random Forest's. In terms of oversampling techniques, Deep Learning-Based Data Augmentation and the Hybrid Approach are the most effective, providing consistent improvements across models and balancing precision with recall. Meanwhile, SMOTE-NC and Cluster-Based Over- sampling provide gains, though they fall short of the top- performing methods. Based on these observations, Random Forest paired with Deep Learning-Based Data Augmentation, or the Hybrid Approach is recommended as the most reliable choice, achieving high precision and recall suitable for applications requiring balanced class prediction accuracy.

TABLE I. METRICS EVALUATION

Data Balancing Technique	ML Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous)	Random Forest	0.983781	0.851429	0.823204	0.837079	0.989670
	SVM	0.823266	0.196501	0.806630	0.316017	0.909461
	Decision Tree	0.960291	0.578313	0.795580	0.669767	0.882326
	Logistic Regression	0.747483	0.148833	0.845304	0.253102	0.841551
Cluster-Based Oversampling (Cluster-SMOTE)	Random Forest	0.979866	0.804469	0.795580	0.800000	0.981290
	SVM	0.894855	0.288503	0.734807	0.414330	0.882468
	Decision Tree	0.961130	0.586066	0.790055	0.672941	0.880153
	Logistic Regression	0.766499	0.143791	0.729282	0.240218	0.815162
Cost-Sensitive Learning	Random Forest	0.984899	0.956835	0.734807	0.831250	0.976977
	SVM	0.949385	0.000000	0.000000	0.000000	0.000000
	Decision Tree	0.974553	0.729592	0.790055	0.758621	0.887222
	Logistic Regression	0.949385	0.000000	0.000000	0.000000	0.492948
Ensemble Methods with Balanced Bagging or bagging	Random Forest	0.985136	0.189231	0.845856	0.254430	0.893884
	SVM	0.814038	0.191327	0.828729	0.310881	0.908868
	Decision Tree	0.976790	0.760638	0.790055	0.775068	0.888400
	Logistic Regression	0.743848	0.147651	0.850829	0.251634	0.844760
Deep Learning-Based Data Augmentation	Random Forest	0.986018	0.971223	0.745856	0.843750	0.979634
	SVM	0.961689	0.940000	0.259669	0.406926	0.884306
	Decision Tree	0.972036	0.695652	0.795580	0.742268	0.888512
	Logistic Regression	0.949385	0.500000	0.027624	0.052356	0.809826
Hybrid Approach	Random Forest	0.985738	0.873563	0.839779	0.856338	0.990526
	SVM	0.824105	0.197297	0.806630	0.317047	0.909352
	Decision Tree	0.963926	0.607438	0.812155	0.695035	0.892086
	Logistic Regression	0.791107	0.166274	0.779006	0.274052	0.853826

REFERENCES

[1] Agarwal, Arunima, Arushi Anand, Yuvansh Saini, S. A. Sajidha, and A. Sheik Abdullah. "A Novel Online Job Scam Detection of Imbalanced Data Using ML and NLP Models." In Practical Applications of Data Processing, Algorithms, and Modeling, pp. 82-99. IGI Global, 2024.

[2] Bauder, Richard A., Taghi M. Khoshgoftaar, and Tawfiq Hasanin. "Data sampling approaches with severely imbalanced big data for medicare fraud detection." 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI).

[3] Lokku, Charan, Kesava Naga Sivaram Kolli, and Santhosh Puganuru. "Classification of Genuinity in Job Posting Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology 9.12 (2021): 1569-1575.

[4] Hanif, A. H. M., Maarop, N., Kamaruddin, N., & Samy, G. N. (2024). "Machine Learning Approach in Predicting Fraudulent Job Advertisement". International Journal of Academic Research in Business & Social Sciences, 14(1),1182–1193.

[5] N. Thai-Nghe, Z. Gantner and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 2010, pp. 1-8, doi: 10.1109/IJCNN.2010.5596486.

[6] D. Dablain, B. Krawczyk and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data" in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 9, pp. 6390-6404, Sept. 2023,

[7] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data" 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1- 11, doi: 10.1109/ICCTCT.2018.8551020.

[8] Anbarasu, V., S. Selvakani, and Mrs K. Vasumathi. "Fake Job Predic- tion Using Machine Learning" ubiquity 13.1 (2024): 12-20.

[9] A. Amaar, W. Aljedaani, F. Rustam, S. Ullah, V.Rupapara, and S. Ludi, "Detection of Fake Job Postings by Utilizing Machine Learning and Natural language Processing Approaches" Neural Processing Letters, vol. 54, no. 3, pp. 219–2247, Jun. 2022, doi: 10.1007/s11063-021- 10727-z.

[10] T. Van Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, Job Prediction: From Deep Neural Network Models to Applications," 2019, doi: 10.48550/ARXIV.1912.12214.

[11] Nasser, Ibrahim & Alzaanin, Amjad. (2020). Machine Learning and Job Posting Classification: A Comparative Study. 4. 6-14.

[12] F. Shibly, U. Sharma, and H. Naleer, "Performance Comparison of Two Class Boosted Decision Tree and Two Class Decision Forest Algorithms in Predicting Fake Job Postings," Annals of the Romanian Society for Cell Biology, pp. 2462–2472, Apr. 2021.

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

[14] F. Shibly, U. Sharma, and H. Naleer, "Performance Comparison of Two Class Boosted Decision Tree and Two Class Decision Forest Algorithms in Predicting Fake Job Postings," Annals of the Romanian Society for Cell Biology, pp. 2462–2472, Apr. 2021.

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