Exploratory Methods for Imbalanced Data Classification in Online Recruitment Fraud Detection: A Comparative Analysis

Jiaxu Li
College of Systems Engineering, National University of Defense Technology

Yunxuan Li
College of Systems Engineering, National University of Defense Technology

Hongjian Han
College of Systems Engineering, National University of Defense Technology

Xin Lu*
College of Systems Engineering, National University of Defense Technology

ABSTRACT

Online recruitment platforms have been increasingly used by companies and applicants. However, there have been a growing number of online recruitment frauds (ORFs) in recent years, seriously affecting the company's reputation and accounting for personal information leakage. On the other hand, identifying ORF with classification models is challenging, as the ORF datasets are typically highly imbalanced such that the accuracy in predicting the minority class in practical recruitment systems is not satisfiable. To overcome these limitations, with empirical Employment Scam Aegean Dataset (EMSCAD), we implement a comprehensive comparative evaluation on processing approaches for imbalanced data, including data sampling techniques, cost-sensitive learning, and ensemble learning. And we design a LightGBM ORF detection model based on hybrid sampling. The results indicate that our model has a higher value in F1-measure, precision and recall of 0.93, 0.93 and 0.94, respectively, and that comparative analysis shows that the model with a combination of data sampling and ensemble learning can achieve improved performance in finding frauds in ORF datasets.

CCS CONCEPTS

• **Social and professional topics**; • **Computing and business**; • **Employment issues**;

KEYWORDS

Online recruitment fraud detection, Highly imbalanced ORF datasets, Ensemble learning, Sampling

ACM Reference Format:
Jiaxu Li, Yunxuan Li, Hongjian Han, and Xin Lu*. 2021. Exploratory Methods for Imbalanced Data Classification in Online Recruitment Fraud Detection: A Comparative Analysis. In *2021 4th International Conference on Computing and Big Data (ICCBD 2021), November 27–29, 2021, Wuhan, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3507524.3507537

1 INTRODUCTION

Internet techniques have been increasingly used for human resource management, which has quickly shift from offline to online, particularly during the recent year as the pandemic of COVID-19 causes recruitment difficulties. Online platforms provide more detailed and comprehensive recruitment channels and opportunities, which is of great significance in optimizing the recruitment market and reducing the unemployment rate [1]. The online platforms significantly improve the efficiency and economics of recruitment [2-3]. However, for various purposes, there are a certain number or scammers posing as professional recruiters post non-existent job openings on online platforms. Such fake information regarding recruitment, hereafter called online recruitment fraud (ORF), is one of the most critical challenges for online human resource management. ORFs present a direct risk to company's reputations and personal information of job applicants. Therefore, it is important to find a method to identify recruitment frauds (hereafter called ORF detection). Machine learning techniques [4-5] offers a series of algorithms to handle such binary classification task, i.e., Logistic Regression (LR), Decision Tree, Support Vector Machine (SVM) and XGBoost.

However, for most online recruitment systems, the class distribution of authentic and fake recruitment information is highly imbalanced, i.e., there are usually few collected samples about fake recruitment information (minority class). As a classifier tends to classify almost all samples as majority class to achieve high training accuracy in the imbalanced dataset (IDS), fake recruitment samples in online recruitment systems are difficult to be identified [6-7]. Thus, it is critical to develop approaches to deal with the imbalanced problem of IDS.

In recent years, a few studies have been conducted to address the problem of online recruitment frauds detection. Vidros et al. collected 17880 manually annotated advertisements (Employment Scam Aegean Dataset, EMSCAD) and proposed a random forest classifier with 89.5% accuracy to detect fraud jobs [8]. Later in 2018, Mahbub and Pardede added some contextual characteristics in feature engineering [9]. Using voting, Lal et al. built an ORFDetector framework by WEKA and achieved an accuracy of as high as 95.4% [10]. However, the limitation of these studies is that there is a lack of methods to address the class-imbalance problem in ORF dataset. To fill in this gap of knowledge, our work aims to present a comprehensive comparative analysis of different methodologies applied on imbalanced dataset for improving the effectiveness of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICCBD 2021, November 27–29, 2021, Wuhan, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8719-4/21/11…$15.00
https://doi.org/10.1145/3507524.3507537

ORF detection and propose a LightGBM ORF detection model based on hybrid sampling designed to identify fraudulent offers.

The class imbalance problem is expected in the diagnosis of rare diseases [11], detection of credit card fraud [12], prediction of loan risk [13], and so on. In recent years, many methods have been proposed to address the class-imbalance problem in the dataset, which could be categorized as data sampling methods, cost-sensitive learning techniques, and boosting strategies. Data sampling methods draw samples from the original data based on acknowledged rules, such as random over-sampling and under-sampling [14-18]. In addition to these methods, Chawla et al. proposed a Synthetic Minority Over-sampling Technique (SMOTE) to create synthetic minority class examples with KNN algorithm rather than random over-sampling from minority class [16]. Elkan et al. proved that cost-sensitive learning could effectively boost the accuracy of different classifications by taking the misclassification cost into account [19-20]. Chan and Stolfo found that ensemble classifiers using non-overlapping subsets from the majority class could improve the capacity of the training model [21].

Based on the context of online recruitment fraud analytics and three different methodologies that address the class-imbalance problem in the dataset, we present a comprehensive comparative evaluation on processing approaches for imbalanced data with EMSCAD, including data sampling techniques, cost-sensitive learning, and ensemble learning. Then, we propose a LightGBM ORF detection model based on hybrid sampling, which could improve the effectiveness of detecting ORFs obviously.

The rest of the paper is organized as follows. The methods used to handle imbalanced datasets in this study are explained in Section 2. In Section 3, we present the empirical online recruitment advertisements dataset, EMSCAD, and data processing techniques. Moreover, we describe performance metrics for evaluation of detection models, and design nine experimental cases that encompass data sampling, cost-sensitive learning, and ensemble learning techniques in this research. In Section 4, we present the results from each of these nine cases and the discussions of our findings. We finally conclude our work and discuss future directions of this study in Section 5.

2 METHODS

To comprehensive evaluate the effectiveness of approaches in handling imbalanced data for ORF detection, we systematically compare three types of methods with ESCAD, including data sampling techniques, cost-sensitive learning, and ensemble learning, for their performance in five binary classifiers, i.e., Logistic Regression (LR), Decision Tree, AdaBoost, XGBoost and LightGBM. Additionally, we propose a LightGBM ORF detection model based on hybrid sampling, which could improve the effectiveness of detecting ORFs significantly.

2.1 Data Sampling

Several over-sampling and under-sampling techniques are applied to tackle the imbalance problem in the training dataset, such as random over-sampling, SMOTE, and random under-sampling. Over-sampling keeps the majority class and generates new samples for the minority class. Apparently, random under-sampling keeps the minority class and drops some examples from the majority class randomly [22]. Instead of obtaining samples from the original dataset, SMOTE creates synthetic minority class samples. K nearest neighbors of a minority class sample x_i are selected based on KNN algorithm. Then, a synthetic sample x_{syn} is calculated by equation 1) based on a randomly chosen sample among K nearest neighbors x_{knn} [16].

$$x_{syn} = x_i + (x_{knn} - x_i) \times \alpha \tag{1}$$

where α is a random number between 0 and 1.

Table 1: The cost matrix

		Prediction	
		Label 0	Label 1
True	Label 0	0	c(0, 1)
	Label 1	c(1, 0)	0

2.2 Cost-sensitive Learning

Most classifiers take the assumption that the misclassification costs of false negative and false positive are the same. However, the assumption does not hold in datasets with highly-skewed class distributions. For example, in ORF detection, the cost of misclassifying a piece of fraudulent recruitment information as a piece of legitimate recruitment information is much higher than the cost of misclassifying legitimate information.

Cost-sensitive learning is a type of algorithm that takes the misclassification costs of some classes into account, and its goal is to minimize the total cost. Here we denote the positive class (+1) as the minority, and the negative class (0) as the majority. $c(i, j)$ is the cost of predicting a data point pertaining to class j when it is actually in class i (see Table 1). We develop a rule that the cost of misclassification of the minority examples is usually much higher than the cost for the majority class, where $c(1, 0)$ is bigger than $c(1, 0)$ [19-20]. The total cost of the classification model is calculated as follows:

$$TotalCost = C\,(1, 0) \times \#FN + C\,(0, 1) \times \#FP \tag{2}$$

where *#FN* and *#FP* are the number of false negative and false positive samples separately.

2.3 Ensemble Learning

Ensemble learning is the algorithm by which multiple sub-classifiers are generated on training sets and aggregated to classify the test examples. There are two commonly used ensemble learning algorithms, including bagging and boosting.

Breiman proposed the bagging (bootstrap aggregating) algorithm. The diversity of classifiers in bagging is obtained by using randomly drawn subsets from the entire training set [23]. For a given test instance, the class chosen by the most significant number of classifiers is the ensemble decision. The sampling processes in bagging are parallel, so the training processes on each subset are also parallel. We use the BaggingClassifier algorithm in the following comparative analysis [24].

Boosting combines weak classifiers to form a single strong learner, such as Adaboost and Gradient Boosting Decision Tree

(GBDT). Compared with bagging, each sub-classifier in Adaboost is serial and each iteration depends on the classification results of the previous round. EasyEnsemble algorithm is an ensemble of AdaBoost learners trained on different balanced bootstrap samples [25]. Compared with AdaBoost, Gradient Boosting improves a combined learner by optimizing the derivative objective function continuously in the function space [26]. In 2016, a scalable end-to-end tree boosting system, eXtreme Gradient Boosting (XGBoost) was proposed by Chen et al [27]. The main differences between XGBoost and other gradient boosting are the definition of the objective function and regularization techniques used to control the overfitting. XGBoost is an additive model composed of k base learners (e.g., decision trees). For a given dataset with n examples and m features D = $\{(x_i, y_i)\}$, (|D| = n, $x_i \in R^m, y_i \in$ R) the prediction of example i after t-th iteration is $\widehat{\mathrm{y_i}}^{(\mathrm{t})}$.

$$\widehat{y_i}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \widehat{y_i}^{(t-1)} + f_t(x_i) \tag{3}$$

where $f_t(x_i)$ is the tree model to be trained in the t-th iteration and $\widehat{y_i}^{(t-1)}$ is the prediction of the first $t-1$ trees. The loss function is represented by the prediction $\widehat{y_i}$ and the true value y_i.

$$L(f) = \sum_{i=1}^{n} L(\widehat{y_i}, y_i) \tag{4}$$

In order to prevent XGBoost from overfitting, we introduce a regularized term, which penalizes the model's complexity. The objective function is defined as follows:

$$Obj = \sum_{i=1}^{n} L(\widehat{y_i}, y_i) + \sum_{i=1}^{t} \Omega(f_i) \tag{5}$$

where $\sum_{i=1}^{t} \Omega(f_i)$ is the sum of the complexity of t trees. The complexity of a decision tree is determined by the number of leaves and the L_2 norm of weights of all leaves.

$$\Omega(f_t) = \alpha\, |T| + 0.5\beta \sum_{j=1}^{T} w_j^2 \tag{6}$$

where $|T|$ is the number of leaves in a decision tree and w_j is the weight of each leaf node.

2.4 The LightGBM ORF Detection Model Based on Hybrid Resampling

As discussed in Section 2.1, 2.2 and 2.3, there are two main perspectives for improving the learning effectiveness of classifiers in imbalanced dataset, including optimizations of data sampling techniques and classification algorithms. Thus, we propose a LightGBM ORF detection model based on hybrid resampling (see Figure 1), which is an optimized combination of data preprocessing using hybrid sampling techniques and modeling with a light gradient boosting machine (LightGBM) algorithm.

2.4.1 Hybrid Resampling. As discussed in Section 2.1, for random under-sampling, since many samples in the majority class are removed from our dataset, the classifier cannot be fully trained from samples of the majority class, which reduces low accuracy in predicting the majority class. However, random over-sampling increases the number of samples in the minority class and improves the classifier's effectiveness in identifying the minority class. Given that these two algorithms both have their advantages and disadvantages, we combine them into a hybrid resampling method which could balance the majority and minority classes and improve the classification accuracy of the two types of samples.

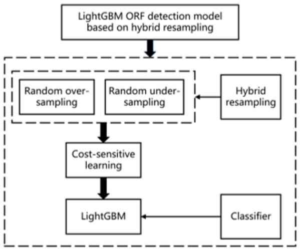

Figure 1: LightGMB ORF detection model based on hybrid resampling

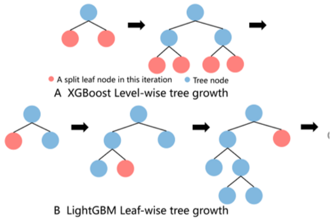
● A split leaf node in this iteration
● Tree node
A XGBoost Level-wise tree growth
B LightGBM Leaf-wise tree growth

Figure 2: A dection tree growth in XGBoost and LightGBM

2.4.2 LightGBM.. LightGBM is a gradient boosting framework originally developed by Microsoft. In comparison with conventional GBDT, LightGBM has faster training speed, lower memory usage and better accuracy. Given the superiorities of LightGBM, we choose LightGBM as the classification algorithm in this model. Compared with XGBoost, the decision trees in LightGBM are grown leaf-wise, instead of checking all previous leaves for each new leaf node, as shown in Figure 2, so that each split could greatly reduce the cost of storage and calculation [28].

3 EXPERIMENTS

In this section, we present the dataset description, evaluation metrics, data processing, and nine experimental cases.

3.1 Dataset

The data used in this study comes from Employment Scam Aegean Dataset (EMSCAD), including 17,880 annotated job ads, of which 866 are fraudulent (label=1) and 17014 are legitimate (label=0). Figure 3 demonstrates the proportion of two types of job ads in EMSCAD. As shown in Fig.3, legitimate job ads comprise 95.16% of all data, whereas fraudulent jobs only accounts for 4.84%.

3.2 Evaluation Metrics

We use five metrics to assess the performance of prediction algorithms with different methods used in imbalanced data processing.

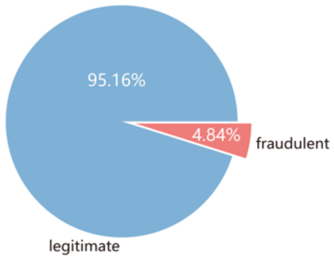

Figure 3: The distribution of jobs in EMSCAD

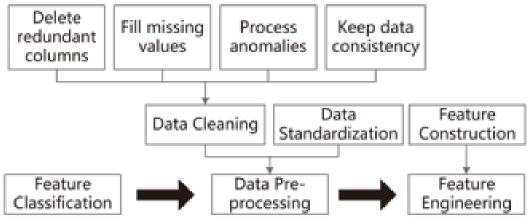

Figure 4: The flowchart of data processing

Precision is the ratio of the number of true positives (*TP*) to the number of true positives plus the number of false positives (*FP*).

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

where *TP* represents the number of fraudulent samples correctly predicted, and *FP* represents the number of samples misclassified as fraudulent jobs.

Recall is defined as the number of true positives over the number of true positives plus the number of false negatives (*FN*).

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

where *FN* represents the number of fraudulent samples classified as legitimate, and *TP+FN* represents the number of all actual fraudulent samples.

The formula for the standard F1-measure (F_1) is the harmonic mean of the precision and recall. It is a useful measure of success of prediction when the classes are highly imbalanced [29]. A perfect model has an F1-measure of 1.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

An ROC curve is a graph which plots TPR (recall) vs. FPR at different classification thresholds. AUC (Area Under the ROC Cure) measures the entire two-dimensional area underneath the entire ROC curve (integral calculus) from (0,0) to (1,1).

$$\text{FPR} = \frac{FP}{TN + FP} \tag{10}$$

where *TN* represents the number of correctly predicted as legitimate samples.

Accuracy is the ratio of the number of correctly categorized samples, positively or negatively, to the total samples. Since a high accuracy is achieved by a model that only predicts the majority class, accuracy is inappropriate for imbalanced classification. Thus, in the context of online recruitment fraud analytics, we do not take Accuracy as principal evaluation metrics.

In our comparative experiments, we use these five evaluation metrics to measure the performance of different prediction models.

3.3 Data Processing

Before training the prediction models, we implement a three-stage pre-processing process for the EMSCAD dataset, including feature classification, data pre-processing and feature engineering (Figure 4). The first two stages aim to convert raw data in EMSCAD into more structured data. Feature engineering then constructs new features expected by ORF detection machine learning models from prepared data. We firstly classify all features in the original dataset into four categories, binary features (bin_features), categorical features (cat_features), text features (text_features) and complex features (complex_features). Then we carry data pre-processing and feature engineering on the dataset.

3.3.1 Data pre-processing. Data pre-processing includes two steps, data cleaning and data standardization. Data cleaning aims to delete redundant fields, fill missing values, and process anomalies. The missing value distributions of all data fields are represented in Figure 5, with black rows representing non-null values. The missing values are replaced by specified values depending on values of different fields. We use Z-score normalization to standardize the data in our work, as shown in Equation 11).

$$X_z = \frac{X - \mu}{\sigma} \tag{11}$$

where X_z represents the normalized value, μ is the mean value and σ is the standard deviation of a feature.

3.3.2 Feature engineering. Table 2 shows the 18 features used in our model after feature engineering, are placed into three categories: text features, numeric features, and categorical features.

3.4 Predictive Models with EMSCAD

To present a comprehensive comparative evaluation on processing approaches for imbalanced data, including data sampling techniques, cost-sensitive learning, and ensemble learning, we modify five different machine learning models, Logistic Regression (LR), Decision Tree, AdaBoost, XGBoost and LightGBM, to detect frauds in the online recruitment data of EMSCAD. Nine developed experimental cases are listed in Table 3. Each of these cases represents a combination of a binary classifier and a method of dealing with imbalance (discussed in Section 2) with 10-fold cross validation method. After comparisons, the better models would be found based on five achieved evaluation metrics, precision, recall, F1-measure, AUC, and accuracy (discussed in Section 3.2).

4 RESULTS AND DISCUSSION

In this section, first, we discuss and analyze the results of nine experimental cases listed in Section 3.4, represented in Table 3.

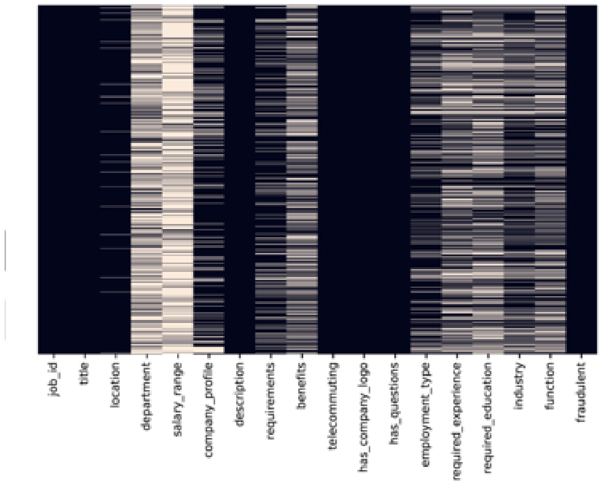

Figure 5: The missing value distributions of all data fields

Table 2: Features used in our model

S. No	Feature	Description	Type
1	Title	The job advertisement header	text_features
2	company_profile	A brief description of the company	text_features
3	Description	Advertised job details	text_features
4	requirements	Required list for job	text_features
5	Benefits	Benefits list offered by employer	text_features
6	min_salary	Suggested maximum salary	num_features
7	max_salary	Suggested minimum salary	num_features
8	company_profile_count_of_words	Feature count of the country description	num_features
9	requirements_count_of_words	Feature count of the job required list	num_features
10	Department	Job relevant department like sales	cat_features
11	employment_type	Full-type, Part-time and Contract	cat_features
12	required_expreience	Executive, Entry level, Intern, etc.	cat_features
13	required_education	Doctorate, Master’s degree, Bachelor’s, etc.	cat_features
14	Industry	Automotive, IT, Health care, etc.	cat_features
15	Function	Consulting, Engineering, Research, Sales etc.	cat_features
16	Country	The country of the job adviser	cat_features
17	State	The state of the job adviser	cat_features
18	City	The city of the job adviser	cat_features

Then, we choose a baseline model [8] for comparison to highlight the effectiveness of our model (Table 4).

Referring to ORF-related work, we find that the Logistic Regression model is widely applied on the binary classification. To improve the training efficiency, we choose the Logistic Regression model as one of the baseline classifiers. Firstly, we test performances of SMOTE, random over-sampling, random under-sampling and cost-sensitive learning on the LR model (see cases 1, 2, 3 and 4). Random over-sampling in case 2 performs better in accuracy and ROC with 0.98 and 0.93 respectively, and random under-sampling in case 3 performs better in precision, recall and F1-measure with 0.90, 0.93 and 0.91.

As discussed in Section 3.2, accuracy is a misleading metric on imbalanced dataset in which the negative samples (majority class) dominate. Thus, the classifier of imbalanced datasets tends to classify almost all samples as majority class, leading to high training accuracy in predicting the majority class. The ROC reflects accuracy for the entire test result outcomes and is insignificant to imbalanced data. Since F1-measure is the harmonic mean of precision and recall and includes false positives and negatives, it is a better-chosen tool for evaluating the performance of different machine learning classifiers on imbalanced data [22]. Since the metrics values not including accuracy in case 2 are higher than in case 1 and 4, it indicates that the random over-sampling achieves a

Table 3: Different experimental cases

Case	Methods of dealing with imbalanced dataset	Classifier	Precision	Recall	F1-measure	ROC	Accuracy
Case 1	SMOTE	Logistic Regression	0.84	0.84	0.84	0.84	0.84
Case 2	Random over-sampling	Logistic Regression	0.88	0.87	0.88	0.93	0.98
Case 3	**Random under-sampling**	**Logistic Regression**	**0.90**	**0.93**	**0.91**	**0.91**	**0.91**
Case 4	cost-sensitive learning	Logistic Regression	0.84	0.87	0.86	0.93	0.99
Case 5	Random under-sampling	XGBoost	0.90	0.92	0.91	0.94	0.93
Case 6	**Random under-sampling**	**LightGBM**	**0.91**	**0.94**	**0.92**	**0.95**	**0.94**
Case 7	Bagging	Decision Tree	0.95	0.80	0.87	0.92	0.98
Case 8	Bagging and boosting	AdaBoost	0.38	0.96	0.55	0.94	0.92
Case 9	**Random over-sampling and under-sampling**	**LightGBM**	**0.93**	**0.94**	**0.93**	**0.95**	**0.95**

Table 4: Comparison results

Model	Precision	Recall	F1-measure	Accuracy
Baseline	0.28	0.75	0.41	0.89
Proposed	**0.93**	**0.94**	**0.93**	**0.95**

better result than SMOTE and the cost-sensitive learning. Compared with over-sampling, random under-sampling is more reliable for ORF detection. Furthermore, we could combine over-sampling and under-sampling methods with improved model performance.

The metrics values in case 6 are higher than in case 5 and 3, in F1-measure with 0.92. The LightGBM model improves the result of our experiment. This outcome further proves our claim in Section 2.4 that the gradient boosting framework using tree-based learning algorithms has better accuracy.

Then we present a comparative analysis of several ensemble methods (cases 7, 8, 9). In case 7, we use a bagging classifier with the base estimator as a decision tree. We modify a bagging classifier with AdaBoostClassifier as learners and a LightGBM model based on hybrid resampling, respectively, in case 8 and 9. The model in case 7 and 8 results in biased learning with higher accuracy and lower F1-measure. Our LightGBM model based on hybrid resampling has a higher value in F1-measure and ROC for 0.93 and 0.95, respectively. The value of precision is 0.93 in case 9, which illustrates that there are 93% of fraudulent job postings that are correctly predicted. A recall of 0.94 presents that our model identifies 94% of all fraudulent job postings. It is noticeable that precision and recall are more appropriate metrics while optimizing an ORF detection classifier.

Table 4 shows the comparison results between our model in case 9 and baseline. The baseline used the EMSCAD and extracted 16 features. Compared with the baseline, our model performs better in following metrics. The effectiveness in feature engineering and modeling is highlighted.

Taking into account all the above observations, it can be stated that our model in case 9 using the hybrid data sampling techniques and LightGBM can give us a higher value in F1-measure, precision and recall in a comparative analysis. These results show that a model with a combination of data sampling and ensemble learning has better effectiveness in ORF detection. To achieve better classification results on empirical imbalanced datasets, we could optimize data sampling techniques and the boosting system.

5 CONCLUSION AND FUTURE WORK

Online recruitment has become an important way for companies to have high-efficiency interactions with applicants. However, there have been numerous instances where scammers have been posing corporate recruiters and making fake job offers called online recruitment frauds, consequently increasing the risk of personal information misuse. It is critical to find a way to detect online recruitment frauds effectively. However, previous studies are mostly limited in lacking of processing highly imbalanced ORF datasets, leading to low accuracy in detecting fraudulent job postings in practical recruitment systems. In this work, using Employment Scam Aegean Dataset, we have presented a comparative evaluation on processing approaches for imbalanced data, including data sampling techniques, cost-sensitive learning, and ensemble learning. Our study aims to construct an effective model for imbalanced datasets on online recruitment fraud detection.

We design a LightGBM ORF detection model based on hybrid resampling with a combination of bagging, Gradient Boosting Decision Tree, random over-sampling and under-sampling. The results indicate that our model has a higher value in F1-measure, precision and recall with 0.93, 0.93 and 0.94. Our comparative analysis shows that a model with a combination of data sampling and ensemble learning has better performance in finding frauds (the minority class) in ORF datasets.

For future works, we are interested in several novel and interesting directions. First, we would collect more recruitment information from various online recruitment websites and other social media to create a larger-scale and more representative dataset and then develop a system integration model to deploy on a real-time monitor platform of ORFs. Furthermore, it is desirable to conduct a more robust hybrid algorithm of processing highly imbalanced datasets, and to verify its versatility on general imbalanced data to improve the model's accuracy for identifying minority class samples.

ACKNOWLEDGMENTS

This study is supported by the National Nature Science Foundation of China (72025405, 82041020, 71771213), the Hunan Science and Technology Plan Project (2020TP1013, 2019GK2131, 2020JJ4673), and the Shenzhen Basic Research Project for Development of Science and Technology (JCYJ20200109141218676).

REFERENCES

[1] Jacobi L and Kluve J, Before and After the Hartz Reforms: The Performance of Active Labour Market Policy in Germany, Journal for Labour Market Research, vol. 40, no. 1, pp. 45-54, 2007.

[2] Kyriakidou N, The Handbook of Human Resource Management Practice - By Michael Armstrong, International Journal of Training and Development, vol. 14, no. 1, pp. 77-79, 2010.

[3] Bandar Alghamdi,Fahad Alharby. An Intelligent Model for Online Recruitment Fraud Detection. Journal of Information Security, vol. 10, no. 3, pp. 155-176, 2019.

[4] M.P.S.-B. Almeida, Classification for Fraud Detection with Social Network Analysis, Engenharia Informática e de Computadores, 2009.

[5] Jeongrae Kim,Han-Joon Kim, Hyoungrae Kim. Fraud detection for job placement using hierarchical clusters-based deep neural networks, Applied Intelligence, vol. 49, no. 8, pp. 2842-2861, 2019.

[6] Lin Zhiyong, Hao Zhifeng and Yang Xiaowei, Current state of research on imbalanced data sets classification learning, Application Research of Computers, vol. 02, pp. 332-336, 2020.

[7] Fathy, Yasmin, Mona Jaber, and Alexandra Brintrup, "Learning with imbalanced data in smart manufacturing: a comparative analysis." IEEE Access 9, pp. 2734-2757, 2020.

[8] Vidros, Sokratis, *et al.*, "Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset." Future Internet 9.1, vol. 6, 2017.

[9] Mahbub, Syed, and Eric Pardede, "Using contextual features for online recruitment fraud detection", 2018.

[10] Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A. and Mourya, R., 2019, August. ORFDetector: ensemble learning based online recruitment fraud detection, In 2019 Twelfth International Conference on Contemporary Computing (IC3), IEEE, pp. 1-5, 2019.

[11] Devarriya, Divyaansh, *et al.*, "Unbalanced breast cancer data classification using novel fitness functions in genetic programming." Expert Systems with Applications, vol. 140, pp. 112-866, 2020.

[12] Bhattacharyya, Siddhartha, *et al.*, "Data mining for credit card fraud: A comparative study." Decision support systems, vol. 50, no. 3, pp. 602-613, 2011.

[13] Hamid, Aboobyda Jafar, and Tarig Mohammed Ahmed, "Developing prediction model of loan risk in banks using data mining." Machine Learning and Applications: An International Journal (MLAIJ), vol. 3, no .1, 2016.

[14] Prusa, Joseph, *et al.*, "Using random undersampling to alleviate class imbalance on tweet sentiment data." 2015 IEEE international conference on information reuse and integration, IEEE, 2015.

[15] Estabrooks, Andrew, Taeho Jo, and Nathalie Japkowicz, "A multiple resampling method for learning from imbalanced data sets." Computational intelligence, vol. 20, no. 1, pp. 18-36, 2004.

[16] Chawla, Nitesh V., *et al.*, "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[17] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." International conference on intelligent computing, Springer, Berlin, Heidelberg, 2005.

[18] Kubat, Miroslav, and Stan Matwin, "Addressing the curse of imbalanced training sets: one-sided selection." Icml, vol. 97, 1997.

[19] Ling, Charles X., and Victor S. Sheng. "Cost-sensitive learning and the class imbalance problem." Encyclopedia of machine learning, vol. 2011 pp. 231-235, 2008.

[20] Elkan and Charles, "The foundations of cost-sensitive learning." International joint conference on artificial intelligence, vol. 17, no. 1, Lawrence Erlbaum Associates Ltd, 2001.

[21] Philip, K., and S. J. S. Chan, "Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection." Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining, 1998.

[22] Kerwin, Kathleen R., and Nathaniel D. Bastian, "Stacked generalizations in imbalanced fraud data sets using resampling methods." The Journal of Defense Modeling and Simulation, vol. 18, no. 3, pp. 175-192, 2021.

[23] Yin, Qing-Yan, *et al.*, "An empirical study on the performance of cost-sensitive boosting algorithms with different levels of class imbalance." Mathematical Problems in Engineering 2013, 2013.

[24] Breiman and Leo, "Bagging predictors." Machine learning, vol. 24, no. 2, pp. 123-140, 1996.

[25] Liu and Tian-Yu, "Easyensemble and feature selection for imbalance data sets." 2009 international joint conference on bioinformatics, systems biology and intelligent computing, IEEE, 2009.

[26] Natekin, Alexey, and Alois Knoll, "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7, vol. 21, 2013.

[27] Chen, Tianqi, and Carlos Guestrin, "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016.

[28] Ke, Guolin, *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems, vol. 30, pp. 3146-3154, 2017.

[29] Wardhani, Ni Wayan Surya, *et al.*, "Cross-validation metrics for evaluating classification performance on imbalanced data." 2019 international conference on computer, control, informatics and its applications (ic3ina), IEEE, 2019.