JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 15 No. 1 (2024) 15-30
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Assessing the Models with Resampled Data Using Explainable Artificial Intelligence Techniques

Rose Mary Mathew[1]*, R.Gunasundari[2], Sujesh P Lal[1]

[1] *Department of Computer Applications*
Federal Institute of Science and Technology, Angamaly, 683576, INDIA

[2] *Department of Computer Applications*
Karpagam Academy of Higher Education, Coimbatore, 641021, INDIA

*Corresponding Author: rosem.mathew@gmail.com

DOI: https://doi.org/10.30880/jscdm.2024.05.01.002

Article Info

Received: 1 December 2023
Accepted: 25 April 2024
Available online: 21 June 2024

Keywords

Data imbalance, interpretation, multiclass, oversampling, XAI

Abstract

In various real-world domains, the problem of imbalanced data poses a significant challenge since it affects the efficiency and trustworthiness of machine learning models. This article investigates Explainable Artificial Intelligence (XAI) methods for studying models created on imbalanced datasets. The main objective of this paper is to assess models trained on DOSMOTE resampled balanced datasets. Using XAI techniques, the study seeks to understand better inner processes that lead to model decisions. The methodology involves combining DOSMOTE resampling with XAI to provide holistic evaluation through both qualitative and quantitative analysis. It should be noted that F1-Scores of balanced datasets improve significantly: from 76% to 87% for Web-Phishing; and from 58% to 73% for Hayes-Roth. This research highlights the need for XAI in enhancing interpretability of models trained on resampled imbalanced data sets. It also shows how resampling affects decision making in a model while performing and recommends investigating other resampling techniques or combinations with XAI methods aimed at improving model interpretability and transparency.

1. Introduction

Data imbalance is a pervasive issue encountered across diverse real-world applications as it impacts both binary classification problems and multi-class classification tasks [1]. Consequently, researchers have made considerable efforts within the area of imbalanced learning where they strive towards addressing skewed distributions by creating more balanced datasets. In binary datasets, most classes are represented by majority class instances while minority class samples are relatively scarce [2]. On the other hand, skewness in multiclass datasets can take two forms namely multi majority which refers to situations where many classes have high frequencies except one with very few examples; or multi minority when some classes lack sufficient data while others enjoy significant popularity [2]. Therefore, achieving balance sheet data sets is an important goal because such trained models perform better and have increased robustness against any form of data imbalance.

To handle problems associated with skewedness of data, several techniques have been proposed by the research community including resampling and cost-sensitive learning. Resampling methods involve modifying the distribution of data across classes through oversampling, under sampling or hybrid approaches that combine both [3]. Oversampling duplicates samples for minority groups while under sampling deletes instances from majority categories to achieve equal representation between different groupings [4]. Alternatively, cost-sensitive learning considers prediction errors and incorporates costs during training phase so as to mitigate imbalanced nature of input features [5]. However, machine learning algorithms are often considered opaque leading to issues such as trustworthiness and bias. Thus, Explainable AI (XAI) is an emerging field which aims at demystifying black box models used in machine learning. It provides insights into how a model arrived at its decisions thereby enhancing debugging and bias removal capabilities [6]. This discipline makes use of text based and visual explanations that unveil the inner workings behind otherwise hidden layers of deep neural networks thus promoting comprehensibility of such architectures [6]. Therefore, combination these methods with XAI can greatly improve performance on complex tasks while minimizing unfairness induced by biases inherent in highly sophisticated ML systems.

This paper attempts to fill a significant research gap by examining the combined use of DOSMOTE resampling method and Explainable AI (XAI) in dealing with imbalanced datasets. Although individually explored, this has not been done together. Therefore, the main objective of this study is to bring new insights on how XAI can help us understand model behavior on resampled data better. In order to determine which resampling algorithm works well for interpretation of models used in different situations; two unique types of data sets were taken into consideration during an experimental study where such methods were rigorously evaluated according to their performance levels. Section 2 provides a comprehensive literature review while section three explains various applied methodologies. Section 4 gives detailed steps followed during experiments while section 5 represents findings obtained from these tests for discussion. The summit of this article is in Section 6 where they summarize key findings and give important recommendations for future studies. Indeed, this research illuminates the transformative power of resampling methods with XAI in dealing with imbalanced datasets complexities thus making significant contributions towards AI interpretability and model performance.

2. Literature Review

The aim of the article is to synthesize imbalanced learning with Explainable AI (XAI) techniques for handling skewed data sets. This article mainly focuses on using an oversampling method to create synthetic samples for the minority class. Then a model is built using the oversampled data and different XAI techniques are applied to explain and understand how the model behaves.

Regarding oversampling paradigms, various novel approaches such as SMOTE by Chawla [7], SMOTEBoost which combines boosting techniques with SMOTE [8] and safe-level-SMOTE that gives priority to safer objects [9] are extensively discussed in this paper. In addition, LN-SMOTE which uses neighborhood information [10] as well as MWMOTE that modifies synthetic generation process through clustering approach are investigated here. An example safety evaluation method was proposed by Napierała [11] while ADASYN suggests density distribution based on class concept [12]. Moreover, local attributes identification with SPIDER algorithm coupled with local oversampling by Fernandez-Navarro et al. [13] extending multi-class scenarios capability of SMOTE are highlighted.

Model interpretability plays a major role in understanding machine learning models' parameters and hyperparameters when working with imbalanced data as outlined by this article. It cites Dablain et al.'s work that presents a unified framework for explaining models on imbalanced data [14]. Additionally, Zilke et al. proposed a model which extracts learned rules [15], Gilpin et al. developed a model where individual filters or neurons represent specific features [16]and Papernot & McDaniel introduced interpretable model incorporating K-nearest neighbor (KNN) algorithm [17]. The computational requirements of LIME and Shapley values were discussed by Achtibat et al. [18]. Resampling techniques are combined with XAI for imbalanced datasets thus effectively bridging the research gap; it investigates oversampling methods comprehensively using different algorithms thereby contributing significantly to AI through enhancing model transparency and decision-making in practical applications. Dablain et. al. discusses that the disparity between current methods for explaining deep learning models on imbalanced image data and the needs of the imbalanced learning community underscores the necessity for an adaptable framework integrating explainable Artificial Intelligence (XAI) techniques and balancing methods, catering to the complexities of modern deep networks and large, multi-class datasets [19].

3. Methods

To examine how well machine learning models, work when dealing with imbalanced data, a study was carried out. In this study, the results from interpreting models using resampled datasets were analysed. The steps involved in this study are illustrated in Fig 1. The study follows a set of distinct steps, namely:

i. Dataset Acquisition: For classification purposes, two 3-class datasets were selected from KEEL Repository.
ii. Data Pre-processing: Selected datasets went through pre-processing stage where any noise or irrelevant data was eliminated.
iii. Dataset Splitting: An 80:20 ratio split was done on the dataset into training dataset and testing dataset.

iv. Resampling Technique: To deal with class imbalance, the training data set has been processed by means of the DOSMOTE algorithm, and after that model was trained.

v. Classifier Performance Evaluation: We have checked how good KNN, Decision Tree and Random Forest classifiers perform.

vi. Model Interpretation: To interpret models created from it, test data were fed into them which resulted in generation of charts for model interpretation.

vii. Analysis of Model Interpretation: Each model's interpretation diagrams were examined in detail across various models, evaluating how well each model performed on the original and resampled data.

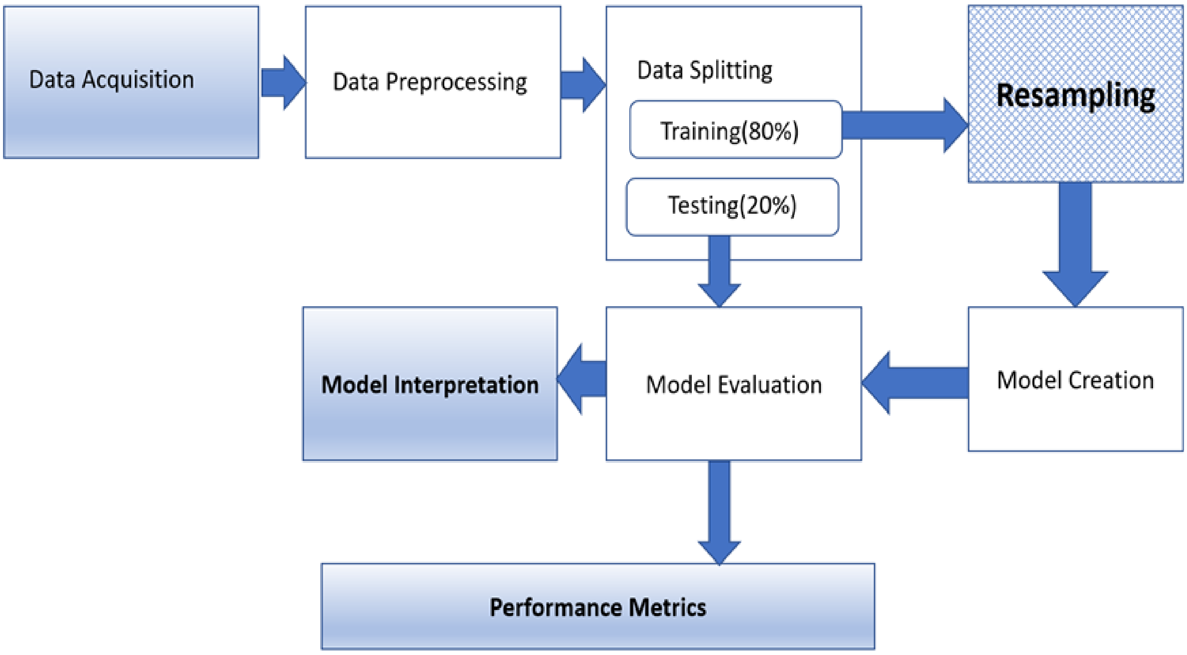


Fig. 1 *Various stages that are performed in the experimental study*

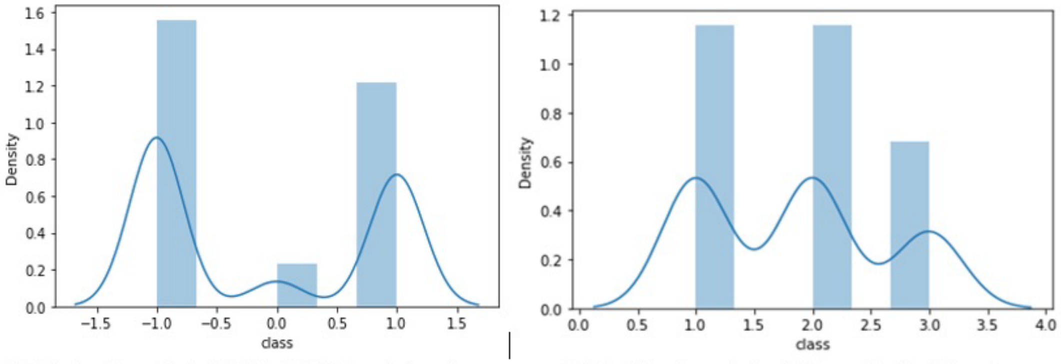
This section of the paper describes the datasets and methods used in the experimental study. This includes details about the dataset used, how it was resampled, as well as what tools for model interpretation were employed.

3.1 Data Acquisition

In this research, KEEL and UCI repositories [20][21] are used to source for datasets. These are Web-Phishing and Hayes-Roth datasets which are both multi-class imbalanced with three class labels [22]. Distribution of data is shown in Fig.2 while detailed descriptions of datasets are provided in Table1.

Table 1 *Description of data distribution in the datasets*

Name of dataset	No. of Instances	Imbalance Ratio	No. of Attributes	Class Label	No. of records
Web-Phishing	1353	6	11	0	702
				1	103
				2	548
Hayes-Roth	160	2.1	5	0	51
				1	51
				2	30



i. Distribution plot of Web-Phishing dataset

ii. Distribution plot of Hayes-Roth dataset

Fig. 2 *Data distribution of different classes in (i) Web_Phishing dataset (ii) Hayes-Roth dataset*

3.2 Resampling Method

The chosen dataset has a problem of being imbalanced which means that if not taken care of during training it can result in biased models. To deal with this issue different resampling techniques were applied where DOSMOTE procedure was used which is combination of SMOTE with Darwinian Particle Swarm Optimization. DOSMOTE is an over-sampling method that is utilized when dealing multiclass problems where majority classes are overrepresented by creating synthetic samples using SMOTE [22]. However, there might be overlap between synthetic points generated from different classes and real ones thus confusing learning process by a model designed to handle such situations especially when dealing with many classes representing small portions within dataset as they might get mixed up during training process leading into wrong predictions being made afterwards hence could not achieve good results, so some improvements needed to be done. To fix this DPSO technique was applied on minority class data so that each created sample can be evaluated individually for optimization purposes otherwise it may overlap with other categories which will make final output become unbalanced again because only those optimized synthetic samples are selected thereby balancing out dataset. Thus, by combining SMOTE together with DPSO we get an approach called DOSMOTE whereby artificial examples adjusted carefully towards enhancing performance of models [22]. Use of DOSMOTE therefore attempts at improving how well machine learning systems perform on multi-class imbalanced datasets. This technique optimizes generation of synthetic samples, prevents overlapping and confusion during model training caused by overlapping between synthetic points from different classes as well as addresses challenges posed by imbalanced data making models' results more reliable.

3.3 Classifiers Used for Model Creation

Creation of a model involves building a machine learning system that can detect patterns and make predictions based on supplied information. In this case study, three different classifiers which are K-Nearest Neighbors (K-NN), Support Vector Machines (SVM) and Random Forests, will be used to construct models.

3.4 Tools Used for Model Interpretation

Model interpretation is an important part of machine learning research, and it is achieved through Explainable AI (XAI) techniques. The goal of XAI is to understand how a model makes decisions or predictions. It involves identifying features that have strong impacts on the model’s outputs as well as their relationships among other influential variables. There are several methods available for explaining the models such as feature importance which helps in pointing out most influential attributes: partial dependencies plots showing marginal effect of one variable on prediction while holding others constant.

3.5 Model Interpretation

This study uses Explainable AI in the application area by employing both SHAP and LIME packages. In this research, a single framework called SHAP (SHapley Additive exPlanations) is used to explain machine learning models. A value is given to each characteristic by SHAP depending on how much it affects predictions compared with some reference point known as baseline. These figures are then merged systematically until we arrive at our final prediction. There are various good things about SHAP values such as being fair, consistent and locally accurate; they can be computed for any kind of model even if it’s a black-box and also provide insights into feature importance together with directionality. It shows which features have a positive or negative effect on predictions and their strength according to SHAP values. There are summary visualizations like dependence plots that illustrate the relationship between overall feature impact and predictions made by LIME which is another method used in this paper (Local Interpretable Model-agnostic Explanations). The main idea behind LIME is to explain single prediction made by complex black-box models. LIME is a model-agnostic technique that explains any single prediction made by complex black-box models. It does so by approximating the decision boundary around the prediction with an interpretable model around it. The original instance is sampled and perturbed to generate a dataset for which an interpretable model is fit to explain predictions locally. This surrogate model helps in understanding feature influence on the black-box model's decision-making process for that specific instance. Such explanations are both locally faithful and interpretable, thus giving human-understandable insights into inherently opaque systems of reasoning like these complicated ones without built-in interpretability as offered by LIME (Local Interpretable Model-agnostic Explanations). SHAP and LIME together provide useful explanations about how models work: SHAP offers global insights about what features matter most globally across many data points while LIME provides localized explanations per individual example considered. These methods can also be employed to identify biases, understand strengths/weaknesses of models, or give clear explanations for predictions in areas where they are needed most.

Two imbalanced datasets were used during evaluation process of this study; they were subjected through extensive pre-processing techniques aimed at dealing with problems associated with missing values/outliers/noise etcetera before being split into train test sets using an 80:20 ratio whereby 80% went towards training hyperparameter tuning purposes whereas rest served as evaluation set for measuring performance of different classifiers considered here; DOSMOTE resampling was applied on training data due to imbalance; various models were built based on KNN, Decision Tree and Random Forest classifiers using original & resampled data with F1-Score, Precision, Recall & Accuracy being some of metrics used to gauge performance of these classifier models against test sets plus finally AI techniques were employed ensure interpretation comprehensiveness surrounding outputs from them thus enabling better understanding classification abilities exhibited by each learning algorithm as well relationship between resampling methods adopted vis-a-vis decision making processes across all considered models.

4. Results

This study created 12 models from 2 datasets, six of which were trained on original data while another six were trained on resampled data. Three classifiers were used in the experiments and the datasets had three class labels making them multiclass. The confusion matrices for these multiclass datasets were 3x3 giving performance metrics for all combinations of classes. Average performance metrics were calculated to give a summary view.

The Web Phishing dataset was used with three classes to check how well classifiers performed on both original and resampled data. DOSMOTE resampling technique was able to improve accuracy when dealing with imbalanced dataset. F1-Score was important in reflecting imbalance nature of the dataset use and it improved the results when used together with DOSMOTE sampling. Hayes-Roth has three classes also. The table shows how different classifiers perform on this dataset considering both original and resampled data. Table2 clearly shows that resampled data gives more accurate results compared to original imbalanced dataset.

Table 2 *Performance metric values for different datasets*

Dataset: Web_Phishing						
Metrics	**KNN**		**Decision tree**		**Random Forest**	
	Original	**Resampled**	**Original**	**Resampled**	**Original**	**Resampled**
Accuracy	0.41	0.51	0.82	0.87	0.79	0.76
Precision	0.34	0.38	0.77	0.88	0.52	0.66
Recall	0.36	0.33	0.76	0.88	0.57	0.69
F1-Score	0.33	**0.34**	0.76	**0.87**	0.55	**0.66**
Dataset: Hayes-Roth						
Accuracy	**0.56**	0.7	0.85	0.85	**0.85**	0.85
Precision	**0.66**	0.74	0.9	0.9	**0.9**	0.88
Recall	**0.56**	0.73	0.87	0.87	**0.87**	0.89
F1-Score	**0.58**	**0.73**	0.86	**0.86**	**0.86**	**0.87**

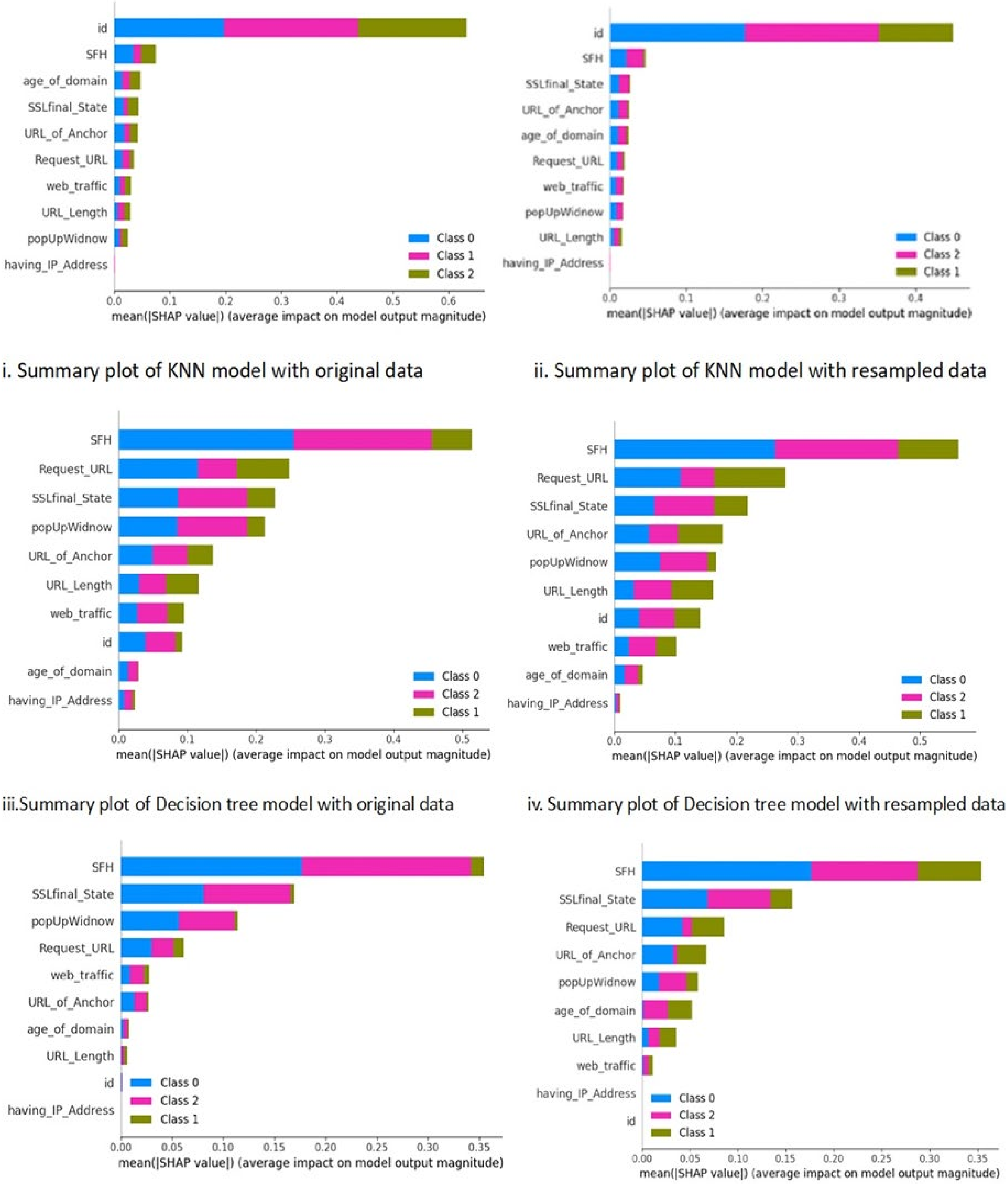
5. Discussions

For model interpretation, Explainable AI techniques were employed where SHAP values describe the models. Summary plot visually presents importance and contributions of features in each model thus showing how different features influenced predictions [23]. This makes the decision-making process by the model more transparent and trustworthy SHAP (SHapley Additive exPlanations).

5.1 Web-Phishing Dataset

Applying Explainable Artificial Intelligence (XAI) techniques such as SHAP values and LIME has greatly increased the transparency and accountability of how the model makes decisions. One significant thing that should not go unnoticed is that, through summary plot in Fig.3 which provided an all-round feature ranking of every model used in this study, it became evident that there exist interesting differences between originality and re-sampling [24].

Furthermore, the figures show that beeswarm plots were of great importance in representing visually the contribution of features to different instances. This enabled easy identification of recognizable patterns as well as outliers in data thus eventually leading to more refined interpretation about how the model behaves. The beeswarm plots were effective in showing how much power any given characteristic had over making predictions by a model thereby illuminating complicated relationships within datasets. Such visualizations not only helped with understanding models but also gave insights into complex interactions between features and their predictions according to these models' outputs. They made it possible for us to dig deep down into what drives decisions made by our machines so that we can know why they think one thing is better than another or why some factors affect outcomes more strongly.



i. Summary plot of KNN model with original data

ii. Summary plot of KNN model with resampled data

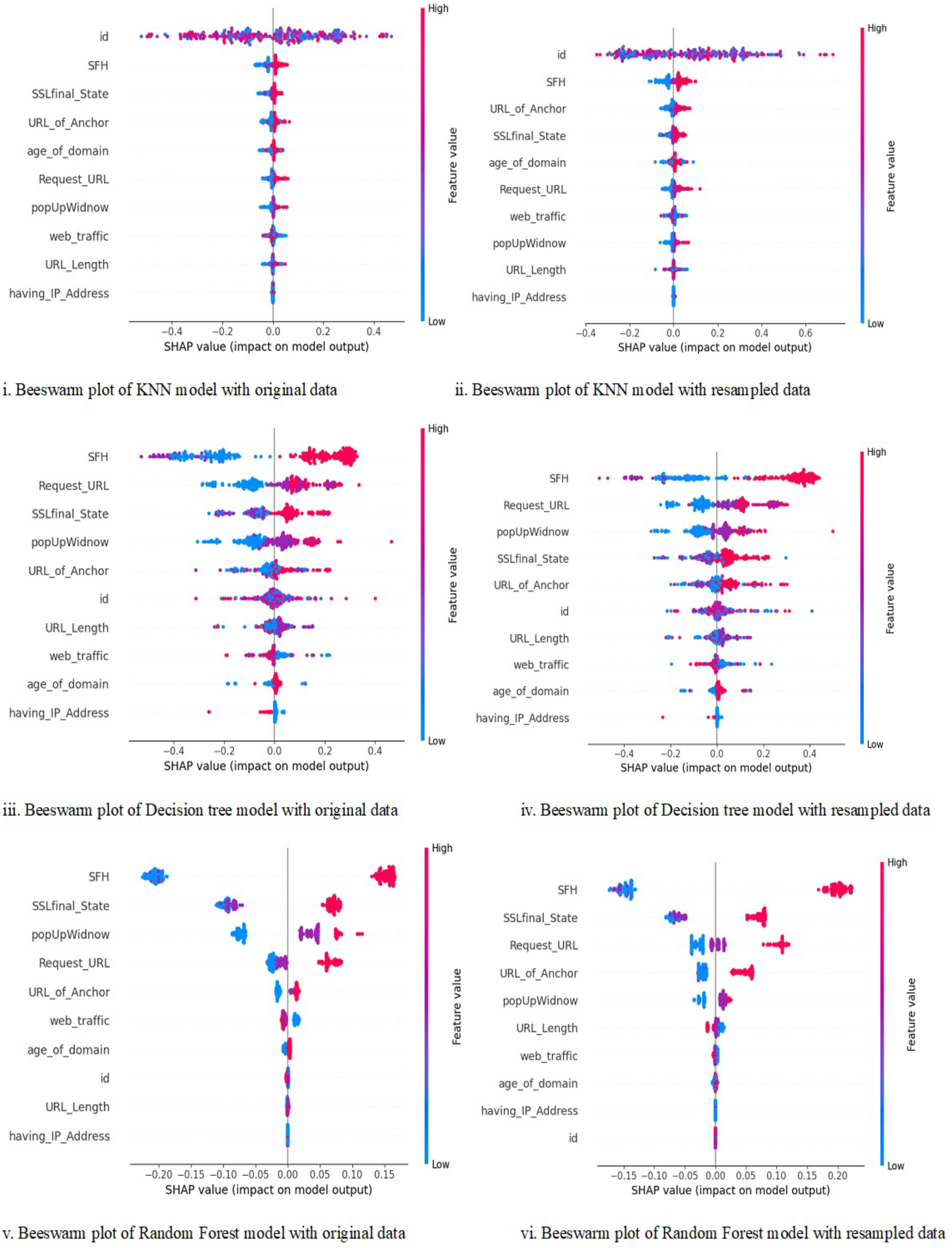
iii.Summary plot of Decision tree model with original data

iv. Summary plot of Decision tree model with resampled data

v.Summary plot of Random Forest model with original data

vi. Summary plot of Random Forest model with resampled data

Fig. 3 *Summary plot of models in SHAP created with original dataset and resampled dataset with different classifiers*



i. Beeswarm plot of KNN model with original data

ii. Beeswarm plot of KNN model with resampled data

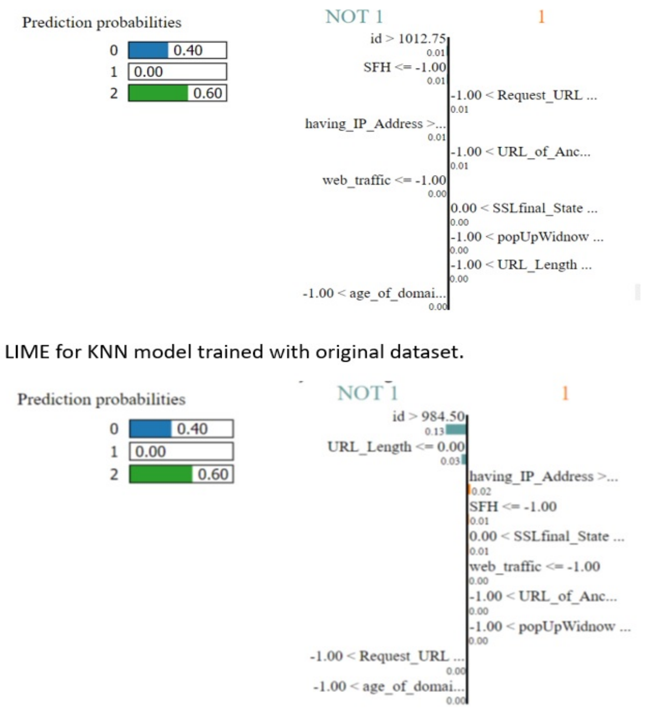
iii. Beeswarm plot of Decision tree model with original data

iv. Beeswarm plot of Decision tree model with resampled data

v. Beeswarm plot of Random Forest model with original data

vi. Beeswarm plot of Random Forest model with resampled data

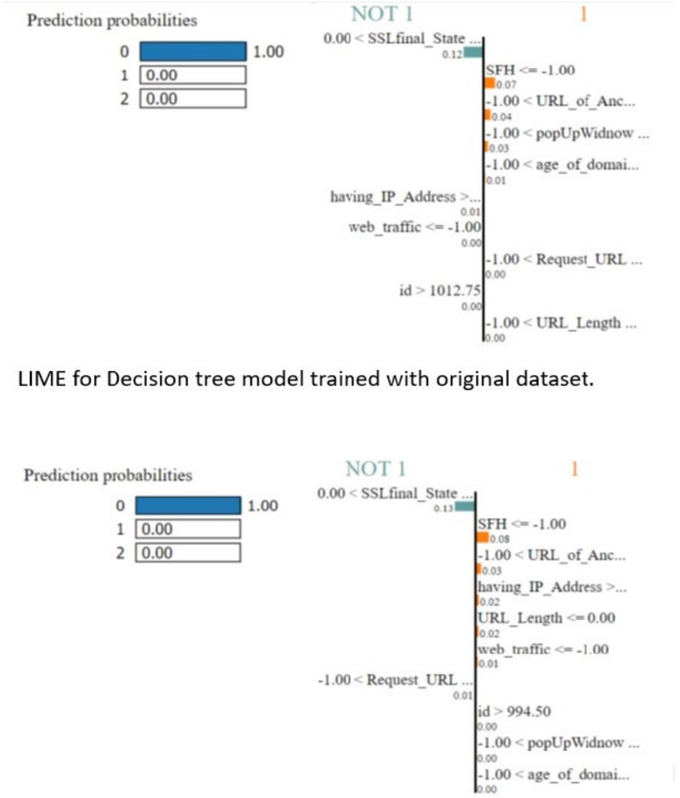
Fig. 4 *Beeswarm plot of different models in SHAP*



LIME for KNN model trained with original dataset.

LIME for KNN model trained with resampled dataset.

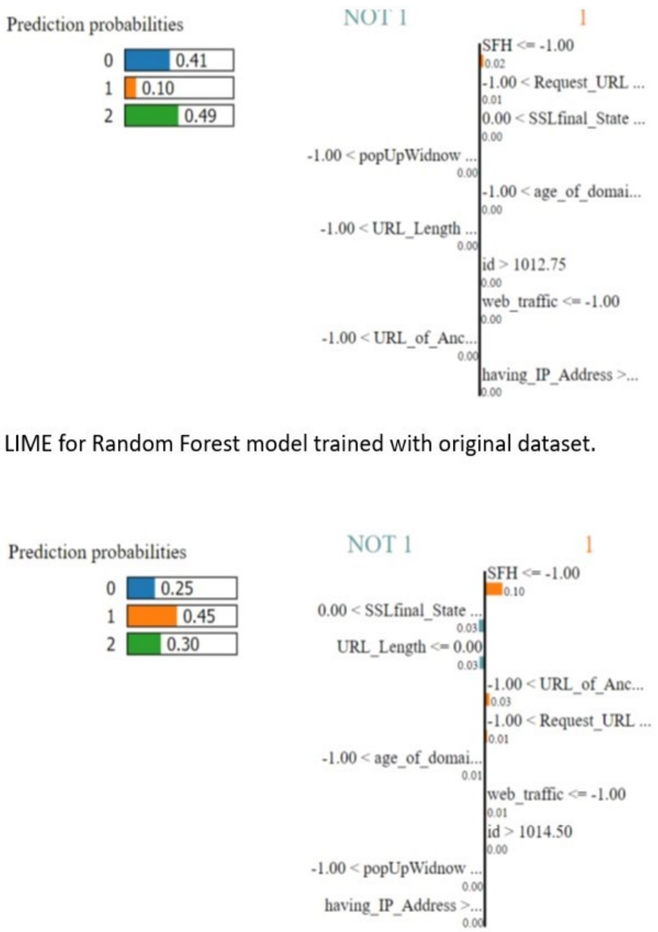
Fig. 5 *LIME prediction for test data 5 of Web_Phishing dataset using KNN model.*



LIME for Decision tree model trained with original dataset.

LIME for Decision tree model trained with resampled dataset.

Fig. 6 *LIME prediction for test data 5 of Web_Phishing dataset using Decision tree model*



LIME for Random Forest model trained with original dataset.

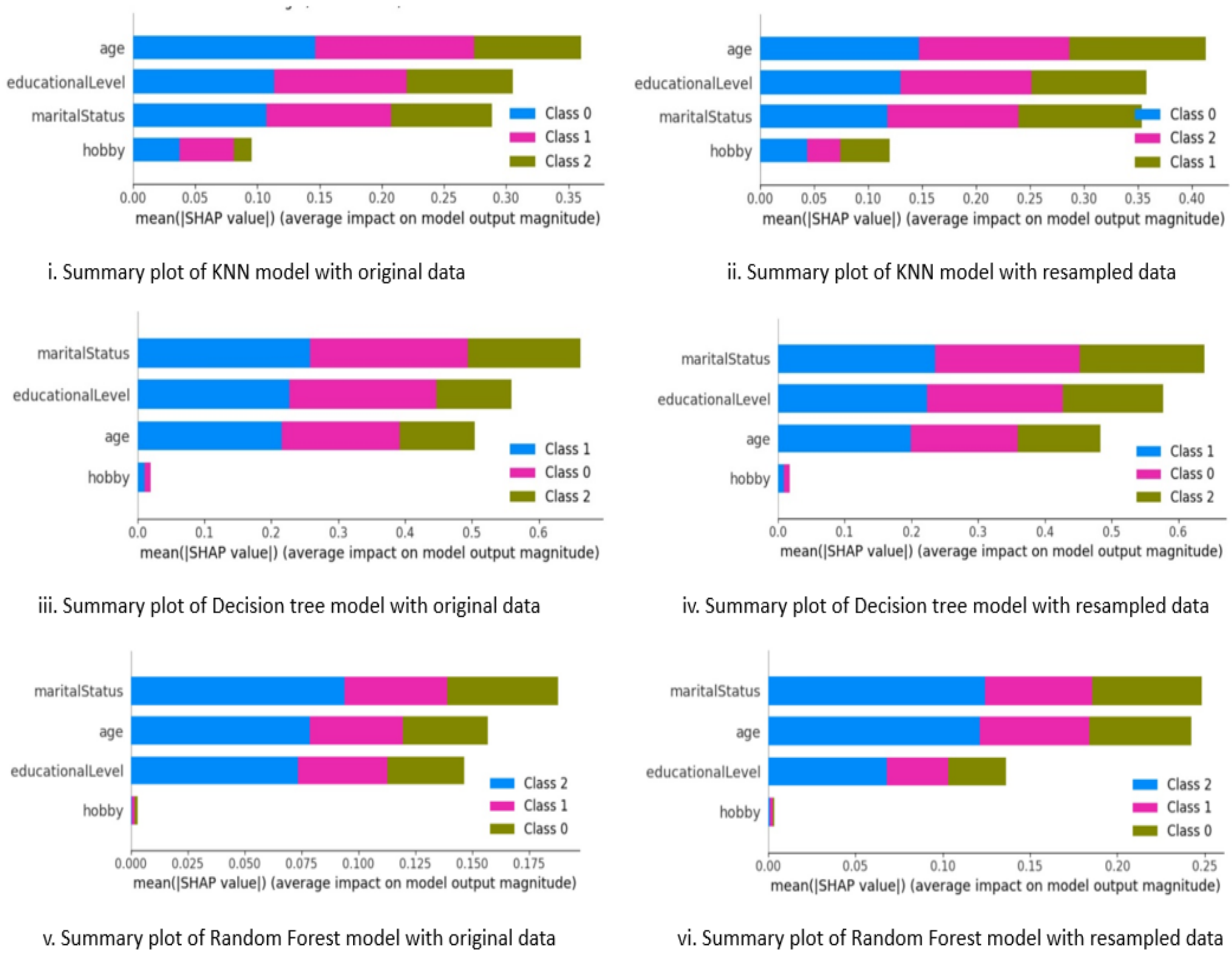
LIME for Random Forest model trained with resampled dataset.

Fig. 7 *LIME prediction for test data 5 of Web Phishing dataset using Random Forest model*

LIME was used in this study as an accurate Local Interpretable Model-agnostic Explanation technique [25]. Using LIME visualization plot helped the team to give insights into specific test data instances revealing key classification features required for correct prediction by each model used and their respective impacts on decisions made by them during classification process. Detailed visualizations showed how far those models went before settling on labels; hence providing clear views on influencing factors behind various predictions. Classification details about probability values considered useful for decision trees were shown alongside random forest probabilities which brought out notable differences between two algorithms' choice of critical attributes used when classifying given samples at hand (see Fig5). Besides, different approaches adopted while making decisions became apparent through this analysis thus indicating hidden complexities affecting predictive performance measures employed within these models. Also, both original resampled parts are illustrated separately throughout Fig7 depicting fifth location explanation under KNN algorithm within random forests model context considering all other possible iterations involving this data set up until convergence occurred. With the help of LIME visualization plot, to present insights into specific test data instances where they identified the relevant features important for classification and how each feature impacted on decision made by model. Detailed visualizations gave a finer grain view of what led different models to their classifications thereby providing more information about factors that influenced them. In addition, this study revealed through decision tree and random forest models which probability values were considered for determining classification details on test data points in Fig7 as well as key features used during classification process according to these two algorithms. These findings point out diverse ways through which decisions can be arrived at when using various predictive methods thus showing complexities behind outcomes obtained from such approaches. Furthermore, LIME visualization plot showed interpretation for fifth location of test data by KNN model considering both original and resampled parts (Fig5). Also, classifying test data 5 using decision tree classifier (Fig6) and giving classification details of test data 5 by random forest model (Fig7) helped in demonstrating differences between these models regarding probability values assigned against critical attributes considered during prediction generation stage. Therefore, it can be said that these results indicate multifaceted nature surrounding decision-making processes employed by different learning algorithms like K-Nearest Neighbors while dealing with imbalanced datasets. The broader understanding regarding relationship between input variables or features on one hand; output labels or predictions made by machine learning system may be achieved due to comprehensive evaluation based on such techniques provided by LIME visualization method [25]. Detailed investigations like those carried out within this research work are therefore expected to enhance transparency around reliability and intelligibility levels associated with ML models hence contributing towards wider XAI advances.

5.2 Hayes-Roth Dataset

The chart given in Figure 8 is an overview of the six predictive models' performances on the Hayes-Roth dataset. This visual representation shows which features are most important in each model, allowing us to see how different inputs affect a model's ability to make predictions. The SHAP framework explains this summary plot by providing information about what matters more than something else and why that might be so. In doing this, it helps users understand why certain things were ranked higher or lower and therefore understand decisions made by various machine learning systems better. It also increases transparency which can lead people into placing trust on these systems as being valid sources for decision making pertaining to their outputs. Looking at this summary graphic we can tell which feature was highest ranked (numbered as 1) across all models besides indicating relative orderings among them according to importance level. For example, among others hobby maritalStatus, educationalLevel age consistently topped the list for both original and resampled KNN models while DT ranked marital status before education then age followed by hobby showing that there was significant difference between first positions of KNNs and decision tree's number one feature variable choice.



i. Summary plot of KNN model with original data

ii. Summary plot of KNN model with resampled data

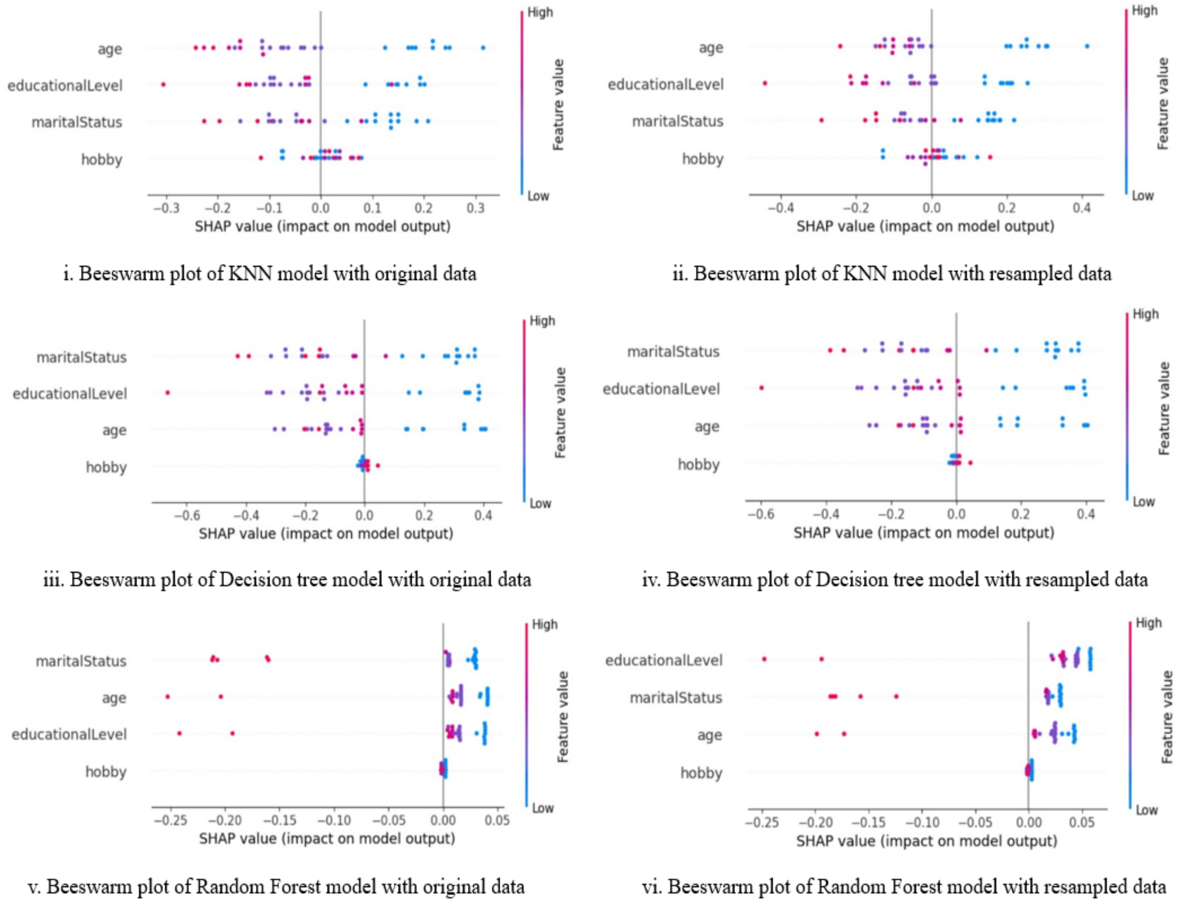
iii. Summary plot of Decision tree model with original data

iv. Summary plot of Decision tree model with resampled data

v. Summary plot of Random Forest model with original data

vi. Summary plot of Random Forest model with resampled data

Fig. 8 *Summary plot of Hayes-Roth dataset for various models by SHAP*



i. Beeswarm plot of KNN model with original data

ii. Beeswarm plot of KNN model with resampled data

iii. Beeswarm plot of Decision tree model with original data

iv. Beeswarm plot of Decision tree model with resampled data

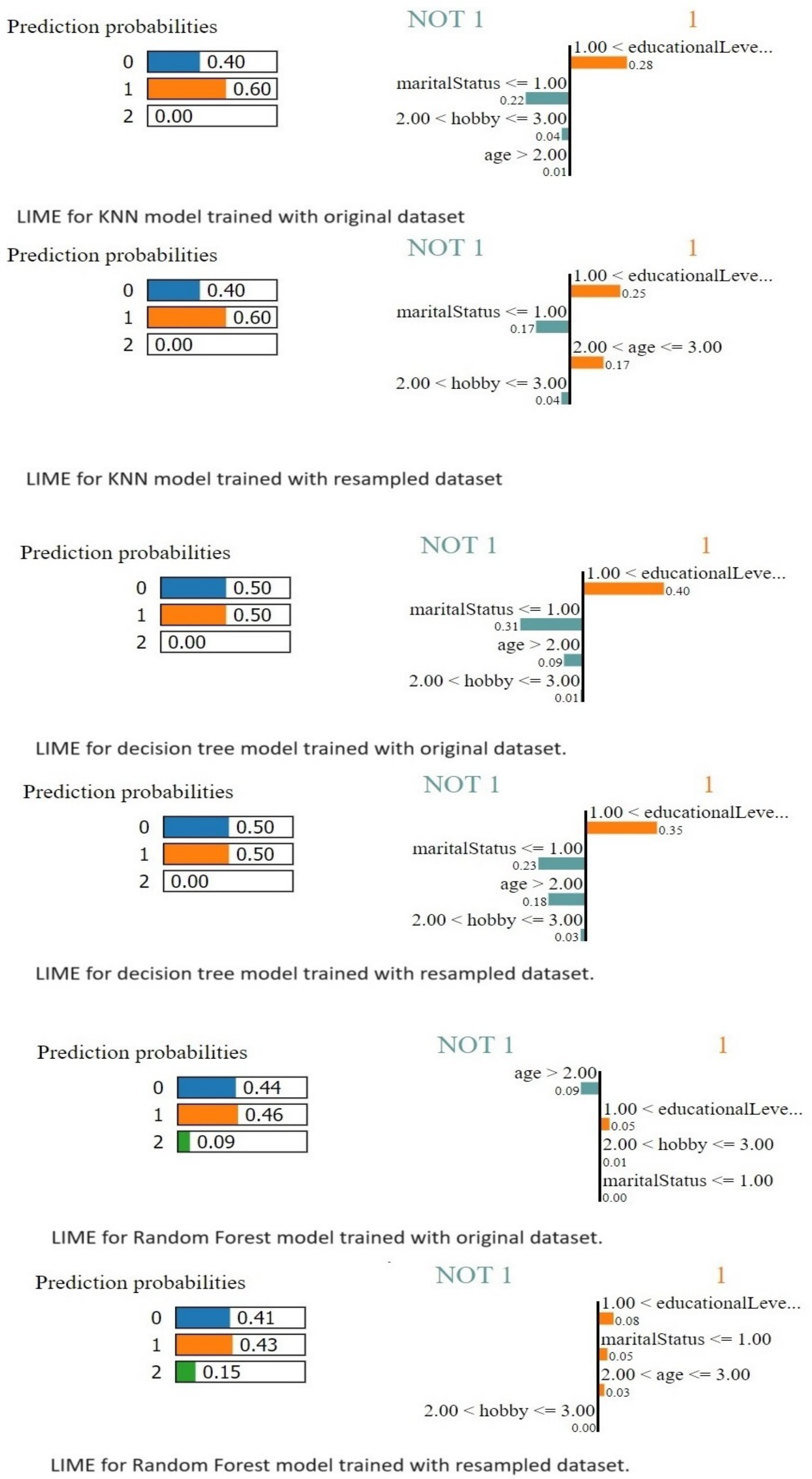
v. Beeswarm plot of Random Forest model with original data

vi. Beeswarm plot of Random Forest model with resampled data

Fig. 9 *Beeswarm plot of Hayes-Roth dataset for various models by SHAP*

The Hayes-Roth dataset's individual feature contributions are effectively revealed with a beeswarm plot as shown in Figure 9. On the horizontal axis, data instances are arranged without overlap so that each dot is placed vertically according to how much did the feature contributed. This will enable different features' importance to be easily understood since they are ranked by their overall importance visually while looking at this chart where also most important ones can be identified by analysts through which some useful information may be found about what these do across different situations thus helping expose model's underlying behavior and giving comprehensive understanding about intricacies within datasets themselves.

Figure 10 illustrates LIME – another way of interpretation which was applied for on fifteenth location from testing data among various models using both original as well resampled data. The first part represents prediction probabilities initially showing that class one is predicted by KNN classifiers with probability equaling 0.6. Then the next part shows feature contributions separating relevant ones' right side from irrelevant left side except decision tree produces same results for both models while random forest gives different outcomes between original and resampled data. Afterward summary type plot displays feature rankings per model thereby indicating how much training influences relevance of each attribute followed with description beeswarm plot which presents contribution per instance relative importance Moreover visualization through lime helps understand what factors are important predicting specific outputs individual wise methods such like these greatly enhance interpretability trustworthiness any black box predictive system hence overall transparency decision making process behind them gets improved thus users become more likely believe in appreciate such systems better.



LIME for KNN model trained with original dataset

LIME for KNN model trained with resampled dataset

LIME for decision tree model trained with original dataset.

LIME for decision tree model trained with resampled dataset.

LIME for Random Forest model trained with original dataset.

LIME for Random Forest model trained with resampled dataset.

Fig. 10 *LIME prediction for test data 15 of Hayes Roth dataset using (a)KNN model with original dataset, (b)KNN model with resampled dataset, (c)Decision tree model with original dataset, (d) Decision tree model with resampled dataset, (e)Random Forest model with original dataset, (f)Random Forest model with resampled dataset.*

Integration of SHAP values and LIME techniques under Explainable Artificial Intelligence (XAI) helped shed light into the decision-making processes employed by our models during this study period. These were successfully applied to improve transparency interpretability trust in predictions made by models used throughout Our methodology consisted creating summary plots along with beeswarm plots further we supplemented these findings using visualizations produced from LIME algorithm While considering an alternative approach called LIME for testing fifteenth location among several models that employed both original and resampled data points were selected based on their highest probabilities achieved under different classifiers such as K nearest neighbors (KNN) Random forests (RF) Decision trees (DT) Moreover it was found out that there exists some inconsistency between results obtained with respect to random forest prediction made using either original or resampling technique except other parts which gave same output Also summary type plot displayed feature rankings per model thereby indicating how much training influences relevance of each attribute followed with description beeswarm plot which presents contribution per instance relative importance Additionally Lime visualization serves increase understanding predictions made any given instance while at same time highlighting those specific features which are most important individual wise These interpretation methods play a big role towards increasing overall transparency explainability such systems thus they should always be used alongside black box predictive models because this helps users gain trust in appreciate them better

Moreover, the different models' ranks of feature importance demonstrated that training data had a huge effect on model behavior. The KNN and decision tree models ranked features in such different ways that it shows how crucial it is to choose algorithms with care according to dataset characteristics. This knowledge adds to what we already know about machine learning by pointing out the importance of choosing features and models for them as well as interpreting or making these machines perform better Despite their substantial contributions, there are some limitations inherent in this research. The study was only conducted on one particular dataset; therefore, any attempt at generalizing findings ought to be approached with caution when applied across other datasets too Furthermore, even though they offered valuable insights into the working of some ML models; explainable artificial intelligence techniques cannot simplify all complex machine learning systems entirely Future investigations need to consider applying these methods in various fields and contexts so as determine their applicability beyond doubt.

According to this study, we propose that future researchers should use resampling techniques in combination with XAI methods to make their models interpretable and trustworthy especially when working with imbalanced datasets. Moreover, there is a need for more investigations into the implementation of these approaches in different domains and real-life situations to ascertain their effectiveness and applicability in practical settings which will contribute towards establishing multi-field transparent, accountable, and interpretable AI models.

6. Conclusion

In this paper resampled data have been shown to be effective when used together with various classification models as it consistently improves performance across different classifiers and datasets. Balanced datasets produced by the DOSMOTE algorithm record significant F1-Score improvements: 76% to 87% for Web_Phishing dataset and 58% to 73% for Hayes-Roth dataset. Resampled data plots show where SHAP package helps us understand how features affect learning on model level while ensuring accurate classification takes place too. This means that interpretation plots clearly display difference between what happens during original vs resample induced model learning stages thereby indicating that preprocessing steps significantly influence outcomes of our models. Pertinently also, LIME package is employed by the research in assessing these models using test data thus providing broader view about same instances' multiple results which are reached by them. Thus through such an in-depth analysis we realize that all along it has been possible indeed applicable even here within imbalanced learning context where traditionally speaking many people believed otherwise; hence making useful one's understanding not only about what but why behind each learnt lesson as well. It therefore follows from this explanation given above that both these findings point towards successful adoption XAI techniques among other things enable drawing valid conclusions about performance of models. Relevance was established by emphasizing transparency during evaluation process whereby trust predictive capabilities were enhanced while remaining true-to-its-word concerning reliability. Finally, another way forward may involve integration advanced resampling methods with complex classification model interpretability through sophisticated XAI methodologies in future studies. Additionally, the use of such techniques should be expanded into real-world situations and diverse fields so that more robust predictive models can be developed based on their practical implications.

Acknowledgement

The authors would like to thank the Faculty of Computer Applications in Federal Institute of Science and Technology and Karpagam Academy of Higher Education.

Conflict of Interest

There is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Rose Mary Mathew;* ***data collection:*** *Rose Mary Mathew;* ***analysis and interpretation of results:*** *Rose Mary Mathew;* ***draft manuscript preparation:*** *R Gunasundari, Sujesh P Lal. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Werner de Vargas, V., Schneider Aranda, J. A., Dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and information systems*, *65*(1), 31–57. https://doi.org/10.1007/s10115-022-01772-8.

[2] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. In *Expert Systems with Applications* (Vol. 73, pp. 220–239). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2016.12.035.

[3] Singh, R., & Raut, R. (2015). Review on Class Imbalance Learning: Binary and Multiclass. *International Journal of Computer Applications*, *131*(16), 4–8. https://doi.org/10.5120/ijca2015907573.

[4] Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, *57*, 164–178. https://doi.org/https://doi.org/10.1016/j.patcog.2016.03.012

[5] Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-Sensitive Learning of Deep Feature Representations from Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3573–3587. https://doi.org/10.1109/TNNLS.2017.2732482

[6] Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11839 LNAI*(September), 563–574. https://doi.org/10.1007/978-3-030-32236-6_51

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 321–357.

[8] Chawla, N.V., Lazarevic, A., Hall, L.O. and Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. *European Conference on Principles of Data Mining and Knowledge Discovery, Cavtat/Dubrovnik, Croatia*, 107–119.

[9] Bunkhumpornpat, C., Sinapiromsaran, K. and Lursinsap, C. (2009). Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. *Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand*, 475–482.

[10] Napierała, K. and Stefanowski, J. (2012). Identification of different types of minority class examples in imbalanced data. *International Conference on Hybrid Artificial Intelligence Systems, Salamanca, Spain*, 139–150.

[11] He, H., Bai, Y., Garcia, E.A. and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEEWorld Congress on Computational Intelligence), Hong Kong, China*, 1322–1328.

[12] Stefanowski, J. and Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performancee. *International Conference on Data Warehousing and Knowledge Discovery, Turin, Italy*, 183–192

[13] Fernández-Navarro, F., Hervás-Martínez, C., & Antonio Gutiérrez, P. (2011). A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, *44*(8), 1821–1833. https://doi.org/10.1016/j.patcog.2011.02.019

[14] Dablain, D., Bellinger, C., Aha, W., Dablain, D. A., Aha, D. W., & Chawla, N. V. (2022). *Understanding Imbalanced Data: XAI & Interpretable ML Framework. December.* https://doi.org/10.13140/RG.2.2.14645.96489

[15] Zilke, J. R., Loza Menc´ıa, E., & Janssen, F. (2016). Deepred–rule extraction from deep neural networks. *International Conference on Discovery Science, Springer*, 457–473.

[16] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89.

[17] Papernot, N., & McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *ArXiv Preprint ArXiv:1803.04765*.

[18] Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., & L., & S. (2022). From" where" to" what": Towards human-understandable explanations through concept relevance propagation. *. . ArXiv Preprint ArXiv:2206.03208,*.

[19] Dablain, D., Bellinger, C., Krawczyk, B., Aha, D. W., & Chawla, N. (2024). Understanding imbalanced data: XAI & interpretable ML framework. *Machine Learning*. https://doi.org/10.1007/s10994-023-06414-w

[20] Al, J. A.-F. et. (2009). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307–318.

[21] Al, J. A.-F. et. (2011). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Log. Soft Comput*, *17*(2–3), 255–287.

[22] Mathew, R. M., & Gunasundari, R. (2023). An Oversampling Mechanism for Multimajority Datasets using SMOTE and Darwinian Particle Swarm Optimisation. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(2), 143–153. https://doi.org/10.17762/ijritcc.v11i2.6139

[23] Das, A., & Rad, P. (2020). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*. 1–24. http://arxiv.org/abs/2006.11371

[24] Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Frontiers in Artificial Intelligence*, *4*(May), 1–15. https://doi.org/10.3389/frai.2021.550030

[25] Urja, P., Donna, O., Susan, R., & O'Reilly Ruairi. (2020). Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain. *CEUR Workshop Proceedings*, *2771*, 169–180.