Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com,
ISSN 0973-2861VolumeXVII, Issue II, July-Dec2023

ANALYSIS OF RESAMPLING TECHNIQUES ON CLASSIFICATION MODELS- A CASE STUDY

Kashika Arora[1], Ms. Yogita Punjabi[2], Dr. Ruchi Mathur[3]

[1] Student, JECRC University, Jaipur, India

[2]Assistant Professor, Jaipur Engineering College and Research Centre, Jaipur, India

[3]Professor, Jaipur Engineering College and Research Centre, Jaipur, India

ABSTRACT

Supervised learning is one of the machine learning techniques which is broadly classified into Regression and Classification Algorithms. This paper focuses on two models Logistic Regression and Random Forest Classifier. In order to understand the effectiveness of the models, model validation, performance estimation, and addressing imbalanced datasets resampling methods can be used in conjunction with the Logistic Regression and Random Forest Classifier. Choosing and comparing the performances of various resampling methods using different data sets on both the classification models. The need for analysing the different resampling method arises as in the real-world scenarios, due to the advancement of the technology the number of variables in regression model is often too large and imbalanced. The study involves the analysis on both balance class distribution and training sets that are unbalanced.

Keywords - Supervised learning, Machine Learning, Classification, Logistic Regression, Random Forest Classifier, Resampling

1. INTRODUCTION

Machine Learning is training the computer system based on the past and historical data that helps in predicting the output for the new data, there are various ways to train the machine with the help of data.

One of the approaches is supervised learning where the model learns from the labelled data, which is further divided into Regression and Classification which deals with predicting continuous and categorical values respectively.

Here we are focusing on the Classification models, Logistic Regression and Random Forest Classifier.

1.1 Classification Models

1) Logistic Regression: It's one of the most basic and popular machine learning algorithms, which comes under supervised machine learning. Logistic Regression is a linear model and helps in predicting the categorical dependent variable with the help of given set of independent variables.
2) Random Forest Classification: It is a non-linear machine learning algorithm which is the combination of ensemble learning and decision tree.It is a powerful model that is known for its high accuracy and robustness.

1.2 Resampling Techniques

As discussed, resampling is use to gain useful insights from the datasets which is not possible if we only fit the data once on the given data set. It involves splitting the dataset into training and test data set.

There are various resampling techniques used here are:

1) K Fold Cross Validation: The original data set is divided into k folds or can say k subsamples. From all the fold one-fold is retained for the testing the model and the training is performed on the rest of the folds.

2) Undersampling: This type of technique is used when the available to us is imbalanced. In Undersampling the number of samples present in the majority class is reduced in order to balance the given dataset.

3) SMOTE (Synthetic Minority Oversampling Technique): SMOTE is used to address the problem of class imbalance. It creates synthetic data points for the minority class with the help of existing data points.

4) SMOTE-Tomek: It is a hybrid Undersampling and oversampling technique. This technique combines the strength of both SMOTE and Tomek links

SMOTE creates synthetic points for the minority class using existing data points and Tomek links helps to reduce the noise by removing the majority class samples that is closest to minority class samples.

5) SMOTE-ENN (Synthetic Minority Over-sampling Technique - Edited Nearest Neighbors): Is also one a hybrid Undersampling and oversampling. combination of Smote and ENN (Edited Nearest Neighbor)

First oversampling is done using smote i.e., creating synthetic data points with help of the existing ones for the minority class and then using ENN the noisy samples are removed from the majority class by comparing then to their k neighbor and removing those class labels which differs from the neighbor.

2. CASE STUDY

Using breast cancer dataset in order to evaluate and analyze the performance of different resampling techniques on Logistic Regression model and Random Forest Classifier.There are various parameters to analyze the model some of them are: - Accuracy, Precision, Recall, F-1 score, Area under the ROC, Confusion Matrix etc. Here recall is our main concern as it is a useful metric in the case where False Negative trumps False Positive in accordance with our breast cancer data set.

2.1 Dataset

The breast cancer data set contains two categories namely “B” and “M”. Here, “B” stands for "Benign" which is interpreted as the tumour is non-cancerous and not harmful. “M” stands for "Malignant", this category classifies tumour as cancerous and harmful.

The dataset contains 569 records in which 357 is categorized as Benign and the rest 212 as Malignant, which depicts that are data set is imbalanced.

2.2 K fold cross validation

After fitting the models on the dataset and applying Kfold cross validation the results are as follows: -

Parameters	Logistic Regression	Random Forest Classifier
Accuracy	**57%**	**95%**
Precision	**37%**	**96%**

Recall	**20%**	**93%**
F1-score	**26%**	**94%**

Logistic Regression model has the worst performance when compared to Random Forest Classifier using K fold Cross validation. Logistic Regression correctly identifies 20% of all Malignant tumours whereas Random Forest Classifier correctly identifies 93% of all Malignant tumours. As the data is imbalanced both the models fails to effectively predict the data.

2.3 Undersampling

Parameters	**Logistic Regression**	**Random Forest Classifiers**
Accuracy	**38%**	**99%**
Precision	**38%**	**97%**
Recall	**100%**	**100%**
F1-Score	**55%**	**99%**

The classification models performed in the desired manner with respect to recall parameter where they identify 100% of all the malignant tumors. But as depicted by the accuracy is still low while using logistic regression.

2.4 SMOTE

Parameters	**Logistic Regression**	**Random Forest Classifiers**
Accuracy	**38%**	**96%**
Precision	**38%**	**95%**

Recall	**100%**	**95%**
F1-Score	**55%**	**95%**

Logistic Regression performed in a similar manner when used with undersampling method. Random forest classifier performance declined comparatively with respect to undersampling.

2.5 SMOTE – Tomek

Parameters	Logistic Regression	Random Forest Classifiers
Accuracy	**38%**	**99%**
Precision	**38%**	**100%**
Recall	**100%**	**99%**
F1-Score	**55%**	**99%**

2.6 SMOTE- ENN

Parameters	Logistic Regression	Random Forest Classifiers
Accuracy	**37%**	**100%**
Precision	**37%**	**100%**
Recall	**100%**	**100%**
F1-Score	**54%**	**100%**

3. Conclusion

In this paper, we analyse various resampling techniques. Logistic Regression predicts 20% of all malignant tumours when the data is imbalanced. After treating imbalanced data with the help of resampling techniques such as undersamling, smote, Smote-Tomek and smoke-ENN though it identifies a remarkable 100% of all malignant tumours. Its overall accuracy and other parameters show unsatisfactory outcomes.

The analysis exhibits that as logistic regression is a linear classifier. If the imbalanced categorial data comprises of complex relationships or overlaps significally, logistic regression may struggle.

On the other hand, Random Forest classifier predicts 93% of all malignant tumors and the result advances as we use various resampling techniques to address the imbalanced and complex data set. It showed the finest result when implemented with SMOTE-ENN, a hybrid technique which helps in extensive data cleaning by integrating ENN an undersampling technique with oversampling done by SMOTE.

4. REFERENCES

[1] Fahad Alahmari, "A Comparison of Resampling Techniques for Medical Data Using Machine learning", Journa of Information and Knowlegde Management, Vol.19(1),2020, doi:10.1142.

[2] J.R Fieberg, K. Vitense and D.H. Johnson, "Resampling-based methods for biologists", The Open Access journal for life and Environment research, 2020, doi:10.7717/peerj.9089.

[3] Jin Xiao, Yadong Wang,Jing Chen, Ling Xie and Jing Huang, "Impact of resampling methods and classification models on the imbalanced credit scoring problems", Information Sciences, Vol. 569, 2021, 505-526, doi:10.1016.

[4] W.H. Beasley and Joe Rodgers, "Re-Sampling Methods", The SAGE handbook of Quantitative Methods in Psychology, 2009.

[5] Chong Ho Yu, "Resampling Methods: Concepts, Applications, and Justification, Practical Assessment, Research & Evaluation", Vol. 8(19), 2003, doi: 10.7275.

[6] M.S. Kraiem, F.Sanchez-Hernandez and M.N. Moreno-Garcia, "Selecting the Suitable Resampling Strategy for Imbalanced Data Classification Regarding Dataset Properties. An Approach Based on Association Models", Appl. Sci. 2021, Vol. 11(18), doi: 10.3390.

[7] M. Khadar, V. Ranjith, K Varalakshmi (2021), "Iot Integrated Forest Fire Detection and Prediction using NodeMCU" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 28—35, doi 10.30696/IJEEA.IX.I.2021.28-35

[8] Gayathri. M, Poorviga. A and Mr. Vasantha Raja S.S, (2021), "Prediction Of Breast Cancer Stages Using Machine Learning" Int. J. of Electronics Engineering and Applications, Vol. 7, No. 1, pp. 36-42, doi 10.30696/IJEEA.IX.I.2021.36-42

[9] Karthikeyen, N. Ramya, M. Sai Priya and C. Yuvalakshmi, (2021), "Novel Method Of Real Time Fire Detection And Video Alerting System Using Open-CV Techniques" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 43-50, doi 10.30696/IJEEA.IX.I.2021.43-50

[10] L.Prinslin, M.A.Srenivasan and R.Naveen (2021), "Secure Online Transaction With User Authentication" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 51-57, doi 10.30696/IJEEA.IX.I.2021.51-57

[11] S Lokewar, A Hemaranjanee and V. Narayanee (2021), "Edge Based Ecosystem For Internet Of Things (EBEFIOT)" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 58-67, doi 10.30696/IJEEA.IX.I.2021.58-67

[12] Prof. K. Phani Srinivas and Dr. P. S. Aithal, (2000) . "Practical Oriented Analysis On The Signal Processing Using FFT Algorithm", Int. J. of Electronics Engineering and Applications, Vol. 8, Issue II, July-Dec. 2020. pp 01-10, doi 10.30696/IJEEA.VIII.II.2020.01-10

[13] Onintra Poobrasert, Sirilak Luxsameevanich, Sarinya Chompoobutr, Natcha Satsutthi, Sakda Phaykrew and Paweena Meekanon, (2000), "Heuristic-based Usability Evaluation on Mobile Application for Reading Disability ", Int. J. of Electronics Engineering and Applications, Vol. 8, Issue II, July- Dec. 2020, PP- 11-21, doi 10.30696/IJEEA.VIII.II.2020.11-21

[14] Rajeev Ranjan Kumar and S. P. Singh, (2020), "Variation Of Capacitive Reactance Of Coupled Microstrip Line Structure With Width Of The Similar Metal Strips" Int. J. of Electronics Engineering and Applications, Vol. 8, No. 2, pp. 22-28, DOI- 10.30696/IJEEA.VIII.II.2020.22.28

[15] Sunita Swain and Rajesh Kumar Tiwari, (2020), "Cloud Security Research- A Comprehensive Survey" Int. J. of Electronics Engineering and Applications, Vol. 8, No. 2, pp. 29-39, DOI- 10.30696/IJEEA.VIII.II.2020.29.39

[16] Ritesh Kumar Thakur, Rajesh Kumar Tiwari (2020), 'Security on IoT: A Review", Int. J. of Electronics Engineering and Applications, Vol. 8, No.2, July-Dec 2020, pp-40-48, DOI- 10.30696/IJEEA.VIII.II.2020.40.48.