*Grenze International Journal of Engineering and Technology, Jan Issue*

# A Comparative Study of Re-Sampling Techniques on Machine Learning Model Performance

Pavitha N[1], Ashwini B P[2], Savithramma R M[3] and R Sumathi[4]
[1]Vishwakarma Institute of Technology, Pune, Maharashtra, India
Email: pavitha.n@vit.edu
[2-4]Siddaganga Institute of Technology, Tumakuru, Karnataka, India
Email: {ashvinibp. savirmrl, rsumathi}@sit.ac.in

***Abstract*—In this data era, gaining insights into available data through machine learning to provide business solutions is the trend. Data in few domains such as banking, disease diagnosis and fraud detection, etc, are imbalanced, and to apply machine learning on these require overcoming the imbalance. Applying re-sampling techniques is one the most followed method to overcome this. Various re-sampling methods are available with the effort of researchers across the world. In this context, an attempt has been made in this article to compare various re-sampling techniques on an imbalanced dataset; a Portuguese bank marketing dataset from UCI Machine learning repository. The impact of various re-sampling techniques on the performance of a classic machine learning model; Logistic regression is analyzed. Binary classification is conducted in-order to forecast the subscription status of a customer for the term deposit. The model is evaluated for various performance evaluation metrics such as AUC-ROC (Area Under Curve-Receiver Operating Characteristic), Accuracy, Precision, Recall and F1-score. The obtained empirical results showed that the K-Means SMOTE is the best (94% accuracy), whereas, SMOTE+ENN is the second best (92% accuracy) suitable model for logistic regression while using the Portuguese banking dataset.**

***Index Terms*— Re-sampling techniques, machine learning, logistic regression, imbalanced dataset, binary classification.**

## I. INTRODUCTION

Data is ubiquitous and analyzing the dataset for gaining insights is the current trend. The amount of data collected per day in 2021 is exceeding 1.145 trillion megabytes[1]. With such a volume of data around, every sector from financial to e-commerce is trying to make the best use of the archived and real-time data to improve their trade, sales, production, marketing, etc. Using Artificial Intelligence (AI) techniques to conduct descriptive, predictive, and prescriptive analysis[2] on data is the most researched topic from the past decade. When a preliminary study is conducted on the data for the applications such as banking fraudulent detection, intrusion detection, student performance predictor, disease diagnosis, etc, seems to be imbalanced[3] and needs to be handled before modeling.

In imbalanced datasets the interesting classes are a minority and non-interesting are the majority, which may lead to a model[4] that is under-fitted or over fit, to overcome this, a re-sampling method has to be applied before modeling the data. Re-sampling is a concept of effectively using sample data to increase the accurateness and

*Grenze ID: 01.GIJET.8.1.509*

quantify the uncertainty of a population variable. Re-sampling methods are categorized as follows[5]:

- Over-sampling
- Under-sampling
- Hybrid re-sampling

Over-sampling is the most used re-sampling type in which data is supplemented with multiple copies of the instances belonging to the minority class. Synthetic Minority Over-Sampling Technique (SMOTE) [6] and its variants, Random over-sampling, etc. are some of the examples of Over-Sampling.

In Random over-sampling, the instances of minority class are randomly selected with replacement, in SMOTE[6] randomly an instance from the minority class say x is selected and it finds k-nearest neighbors of it, one of the neighbors say y is chosen, and a new instance for the minority class is synthesized using x and y and added to the training set whereas in Borderline- SMOTE [7] the synthetic data is created only along the decision boundary of both minority and majority classes to minimize the misclassifications. In Adaptive Synthetic Sampling( ADASYN)[8] unlike Borderline SMOTE, it synthesizes data based on data density, i.e. synthetic minority class data are created only in the area where the minority class is less dense.

All the aforementioned methods suffer from a drawback of generating noise along with data, to overcome these, combinational over-sampling methods are adapted like K-Means SMOTE and Support Vector Machine (SVM) – SMOTE[9]. In K-Means SMOTE data is initially clustered into K clusters, clusters with higher minority classes are chosen as it is, and synthetic instances are added only to the clusters in which the minority classes are sparse. In SVM - SMOTE the focus of generating instances of new minority class is near the borderlines generated with SVM.

Under-sampling is a method in which the instances of the majority class are deleted. Most of the time over-sampling methods are preferred over under-sampling, as removing instances of majority class may lead to missing out on necessary insights. Few under-sampling techniques used extensively in the existing literature are Edited Nearest Neighbor (ENN), Neighbourhood Cleaning Rule, Near Miss, Tomek, Instance Hardness Threshold (IHT), One side selection, etc.

The basic under-sampling method Random under-sampling randomly selects the instances of the majority class and deletes them to prepare the training set whereas in Edited Nearest Neighbor (ENN) the instances misclassifying (using k-NN) the majority class is removed. In the Near Miss method, it selects the instances based on the distance of majority to minority class using K-Nearest Neighbors. In Tomek, the cross-pairs, i.e., the pairs of different binary classes with the least distance are selected because these instances define the class boundary; this link between the cross-pair is called Tomek-link. One-Sided Selection (OSS) combines Tomek-Links and the Condensed Nearest Neighbor (CNN) Rule. In Neighborhood Cleaning Rule (NCR), the CNN Rule that removes redundant examples and the ENN Rule that removes noisy or ambiguous examples are combined.

Hybrid re-sampling is a method in which over-sampling and under-sampling methods are combined to overcome the drawbacks of the individual models. In this technique an over-sampling method is applied to data initially, followed by the under-sampling method. SMOTE-Tomek[10], SMOTE-ENN, etc. are some of the examples of hybrid re-sampling

The process of hybrid re-sampling methods includes two steps, i.e. in step-1, over-sampling methods are applied, whereas, under-sampling is applied in step-2. In the SMOTE-ENN method, SMOTE initially generates minority class synthetic instances followed by ENN, which deletes observations of both minority and majority classes which lead to misclassifications, whereas, in SMOTE-Tomek the step-1 is the same as SMOTE-ENN, but in step-2 Tomek links identified in the majority class is removed.

Modeling data is the core task of Machine Learning (ML). A model is developed based on the training set provided. From linear regression to deep learning, researchers have tried to apply various techniques as per the requirements of the application, for gaining insights from data. And also, researchers have applied re-sampling techniques on the training data to improve the performance of the ML model[5][11]. Logistic regression is a classical statistical model adapted in ML. Similar to linear regression, logistic regression is a linear model, but the predictions made by logistic regression are binary values whereas linear regression can estimate continuous values. The logistic regression model is the most widely used method for binary classification. It is sensitive to noise and outliers[12] and hence, well-prepared data has to be provided to the logistic regression model for better predictions.

In this work, an attempt has been made to highlight the importance of re-sampling, and assess the impact of re-sampling on the performance of a machine learning model i.e. the logistic regression technique. The dataset used is Portuguese bank marketing data[13]. Logistic regression technique is applied in-order to forecast the subscription status of a customer for the term deposit.

The organization of the remaining sections of the paper is as follows; Second section briefs about the literature, highlighting the existing research work in this area, followed by section 3 in which the methodology adopted for the study is briefed. The results and its inferences are discussed in section 4. Finally, the highlights of the study are summarized as the conclusion in section 5.

## II. Literature Review

Artificial Intelligence (AI) is the revolutionary technology ruling all most all types of industries in recent decades. The efficiency of an AI model is mainly depending on the dataset used in building the intended model. Data plays an important role in AI development. Building an AI model is a tedious process involving huge data processing. Also, the process requires high-end computing capabilities. The processing time is directly proportional to the size of the dataset. Therefore, it is necessary to reduce the data size but, without affecting the overall and important characteristics of the dataset. In this context, data re-sampling is an efficient tool that serves the purpose. Data re-sampling further evolved into various re-sampling techniques to enhance the model accuracy. Many researchers across the world have proposed various re-sampling techniques and applications of these techniques in ML as well. The research with respect to the combination of various re-sampling technique and ML applied in different fields are summarized in this section.

[7] Have introduced the impact of imbalance in datasets while applying data mining techniques. Also, the summary of existing methods and evaluation metrics to overcome the dataset imbalance problem is presented in the article, authors have used SMOTE to address the issue. The authors [8] have proposed a novel approach to learn from an imbalanced dataset through ADASYN sampling. The use of weighted distribution sampling of minority classes to improve the learning capability of the model is the basic idea of the proposed approach. [14] Have discussed three statistical re-sampling techniques namely permutation test, the jackknife and the bootstrap in depth.

[15] Have presented an improved particle filter algorithm in two different forms by combining analysis of the partial stratified re-sampling algorithm. One algorithm is with the improvement in weights, whereas, the other is based on adaptive mutation re-sampling. The algorithm capability is demonstrated through simulation using MATLAB. Whereas, [16] have applied the hierarchical re-sampling in two steps for distributed particle filters. The efficiency of the algorithm is presented with the metrics such as execution time, memory usage, and scalability.

The effect of the re-sampling technique on the behaviour of particle filter-based robot localizer is demonstrated by [17]. While, the authors investigated various re-sampling schemes such as Multinomial, Residual, Residual Systematic, Stratified, and Systematic Re-sampling. The comparative analysis is demonstrated through simulation using MATLAB. A novel re-sampling algorithm is developed by [18] for particle filter which is suitable for real-time implementation. The authors have proved the efficiency of the proposed algorithm in-terms of hardware and digital signal processing realization complexity. [19] proposed a metropolis re-sampling method to overcome the problems faced by standard particle filters.

[20] have proposed an online learning algorithm using online binary classification for identifying the signals of interest within a sequence of received signals which are generated by background sift. On the other hand, [21] have applied the re-sampling technique to satellite image processing. Basically, the interpolation function is used to preserve the image quality. And the authors have used different re-sampling methods namely Nearest Neighbour, Bilinear, Cubic Convolution, etc. to determine the best interpolation function. [22] have applied a supervised convolution neural network to automatically detect the image re-sampling patterns based on residual mapping relationship. The authors have proved the robustness of the proposed algorithm against noise.

[23] presented an improved version of SMOTE to overcome the imbalanced dataset by reducing the class noise in it. The proposed method is based on fuzzy rough set theory. Extreme Learning Machine (ELM) is a well-known feed-forward neural network for solving classification problems. But, the drawback of ELM is that it tends to consider the major class and neglects minority class while dealing with the imbalanced dataset. Hence, [24] have addressed this issue and applied a re-sampling technique to improve the accuracy of ELM concerning classification. Various re-sampling techniques used to handle the imbalanced datasets are summarized by [25]. Also, the effect of these techniques on classification is studied and presented by the author.

Modeling data with respect to the student is an important part of the education process. But, the student data will have more imbalanced characteristics with it. This issue is addressed by [26] by applying a re-sampling approach namely Neighborhood Cleaning Rule (NCL) to enhance the accuracy of the student model. [27] have attempted to compare various re-sampling methods including SMOTE, SMOTE-ENN, Borderline SMOTE, SVM-SMOTE, SMOTE-Tomek, and Random Over Sampler in-order to manage the problem of imbalance in training data for

predicting students' performance. [28] have proposed a combined sampling technique to improve the performance of imbalanced classification of university student depression data.
[29] have conducted a comparative study on the effectiveness of re-sampled data while training an ML model. The authors showed that the accuracy of models with respect to prediction is improved with the application of Synthetic Minority Oversampling and Random Oversampling techniques. [30] have investigated the effectiveness of four re-sampling strategies namely random-up-sampling, random-down-sampling, ROSE, and SMOTE in combination with two different ML algorithms namely decision tree and random forest in-order to detect the sleep-wake of Fitbit wristbands. The authors can show that the specificity of Fitbit is improved by 75% in the best case. Misclassification of medical diagnosis by an ML model is a serious issue. The main reason for misclassification is the imbalanced training dataset. Therefore, [31] have attempted to address this issue by proposing a novel hybrid sampling algorithm, a combination of Misclassification oriented SMOTE and ENN with a random forest ML algorithm.

## III. METHODOLOGY

The study is carried out mainly in three phases namely (a) Data phase (b) Machine learning modeling and (c) Comparative analysis. The overview of the study methodology followed in this article is depicted in Fig.1. Google Colab environment and python programming is used to model the data through logistic regression, a widely used ML algorithm in the financial sector.

### *A. Data phase*

The first and most important phase of the current study is the data phase which itself involves three phases such as data collection (raw data), data preprocessing, and data re-sampling as described below.
*Data collection*: The Portuguese bank marketing data from UCI Machine Learning Repository[13] is used in the study. The data is having a direct relation with marketing campaigns (i.e., surveys through phone calls) of a Portuguese banking institution. The collected data has 19 independent variables namely age, job, marital, education, default, housing, loan, contact, month, day_of_week, duration, campaign, pdays, previous, poutcome, emp_var_rate, cons_price_idx, cons_conf_idx, euribor3m, nr_employed and one dependent target varible named y.
*Data preprocessing*: Most of the time, it is not possible to model the raw data directly as the data may have various types of noise in it. Hence, the data is to be restructured as per requirements. In this context, data transformation and feature selection operations are performed on downloaded data. The categorical fields in the data like marital status, education, default, loan, housing, contact mode, etc. are *transformed* into continuous variables. The data is checked for missing values and NAs. Then Recursive Feature Elimination (RFE) one of the widely used feature selection algorithms is used for selecting the best features from the model. The best 18 features suggested by RFE are used for modelling.
*Data re-sampling*: Fouteen different re-sampling techniques including under-sampling, over-sampling, and hybrid re-sampling are used for the study as discussed in section 1.

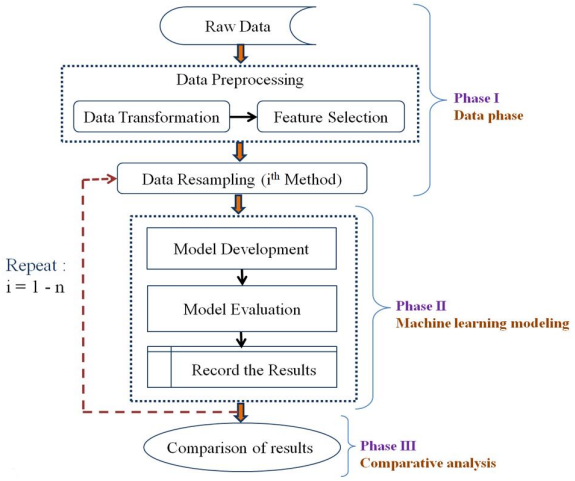


Figure 1. Flow of proposed study methodology

*B. Machine learning modeling*

Machine learning modeling is the process of developing a model to forecast the classes. The process is carried out in mainly two steps. The first step is *model development* (training the model with training dataset), and the second step is model evaluation (performance evaluation through test dataset)

*Model development:* Logistic regression is the ML model selected for the study. The data re-sampled with different re-sampling techniques are modeled using a logistic regression algorithm. From the preprocessed and re-sampled data 70% of records are used to train the model, whereas 30% of records are considered for model testing.

*Model evaluation*: The developed model is evaluated through metrics such as Area Under the Curve Receiver Operating Characteristic (AUC-ROC), Accuracy, Precision, Recall, and F1-score. Finally, the *results* are *recorded* for further analysis.

The process is carried out for each $i^{th}$ re-sampling method and i takes the values from 1 to n. In this article, fourteen (n=14) re-sampling techniques are considered for comparison. Also, 10-fold cross-validation is performed to reduce bias in the data and to evaluate the developed model.

*C. Comparative analysis*

The obtained results for each evaluation metric of each re-sampling technique are recorded in the previous phase are tabulated to compare against each other, in-order to know the best re-sampling technique suitable for logistic regression with the available unbalanced dataset. Along with tabulated metrics, the AUC-ROC obtained for each re-sampling method is compared through graphical representation.

## IV. Results and Discussion

The dataset used for the study is an imbalanced dataset consisting of 41188 records in total, whereas 36548 observations are negative while 4640 are positive observations. The study uses a logistic regression algorithm in-order to forecast the subscription status of a customer for the term deposit. An attempt is made to compare fourteen different re-sampling techniques which include oversampling, under sampling, and hybrid methods. The performance of the logistic regression algorithm while applying each of these re-sampling techniques is recorded. The results are shown in Table 1. The performance is evaluated using performance metrics of a classification problem such as Area Under the Curve Receiver Operating Characteristic (AUC-ROC), Accuracy, Precision, Recall and F1-score.

TABLE I. PERFORMANCE OF LOGISTIC REGRESSION CLASSIFIER ON DIFFERENT RE-SAMPLING TECHNIQUES

| Sl. No. | Method | AUC-ROC | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| Over-sampling | | | | | | |
| 1 | SMOTE | 0.78 | 0.73 | 0.73 | 0.73 | 0.73 |
| 2 | ADASYN | 0.72 | 0.67 | 0.67 | 0.67 | 0.67 |
| 3 | Borderline SMOTE | 0.80 | 0.74 | 0.74 | 0.74 | 0.74 |
| **4** | **K-Means SMOTE** | **0.97** | **0.94** | **0.95** | **0.94** | **0.94** |
| **5** | SVM SMOTE | 0.87 | 0.81 | 0.81 | 0.81 | 0.81 |
| Under-sampling | | | | | | |
| 6 | Random Under Sampling | 0.77 | 0.73 | 0.73 | 0.73 | 0.73 |
| 7 | ENN | 0.82 | 0.89 | 0.88 | 0.89 | 0.87 |
| 8 | Near Miss | 0.80 | 0.74 | 0.74 | 0.74 | 0.74 |
| 9 | Tomek | 0.78 | 0.90 | 0.88 | 0.90 | 0.87 |
| 10 | Instance Hardness Threshold | 0.87 | 0.80 | 0.80 | 0.80 | 0.80 |
| 11 | Neighbourhood Cleaning Rule | 0.79 | 0.90 | 0.88 | 0.90 | 0.88 |
| 12 | One Sided Selection | 0.78 | 0.90 | 0.88 | 0.90 | 0.87 |
| Hybrid-sampling | | | | | | |
| **13** | **SMOTE + ENN** | **0.92** | **0.90** | **0.90** | **0.90** | **0.90** |
| 14 | SMOTE+ Tomek | 0.78 | 0.73 | 0.73 | 0.73 | 0.73 |

For the performance metric AUC-ROC, K-means SMOTE gives a very good result of 97% and SMOTE+ENN gives 92%. K-means SMOTE gives 94% of accuracy and SMOTE+ENN, Neighbourhood cleaning rule, one-sided selection methods giving the accuracy of 90%. A precision of 95% is obtained using K-means SMOTE and 90% using SMOTE+ENN. K-means SMOTE has given a recall of 94% and three methods namely SMOTE+ENN, neighbourhood cleaning rule, the one-sided selection is giving equal recall of 90%. F1- score of 94% is obtained using K-means SMOTE and SMOTE+ENN gives an F1-score of 90%.
The empirical results show that K-means SMOTE re-sampling technique is the best re-sampling technique for logistic regression algorithms while using Portuguese Bank Marketing Data. The second-best choice is SMOTE+ENN for this study. The best two re-sampling techniques are highlighted in bold in Table 1. The graphical representation of AUC-ROC is shown in Fig. 2-16. The false positives are plotted on the x-axis and true positives on the y-axis.

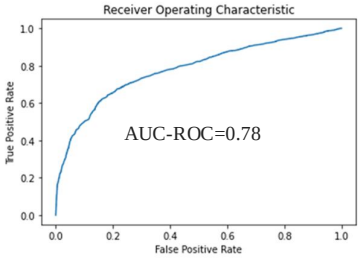


Figure 2. SMOTE

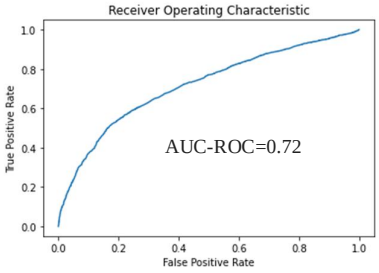


Figure 3. ADASYN

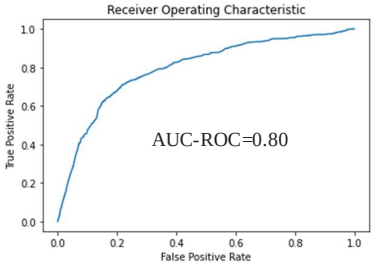


Figure 4. Borderline SMOTE

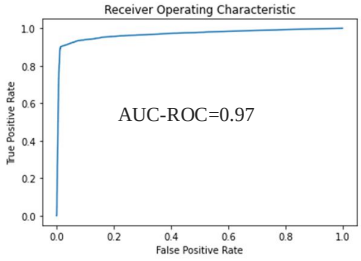


Figure 5. K-Means SMOTE

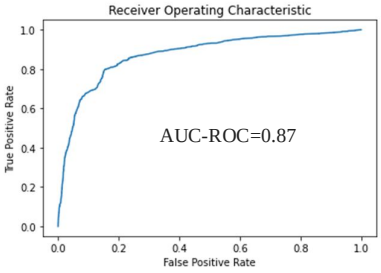


Figure 6. SVM SMOTE

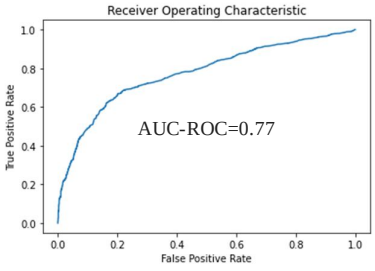


Figure 7. Random Under sampling

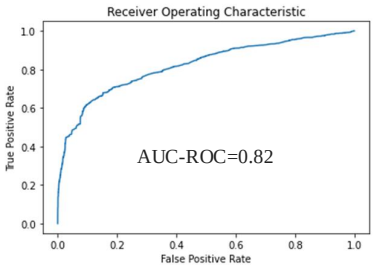


Figure 8. Edited Nearest Neighbor

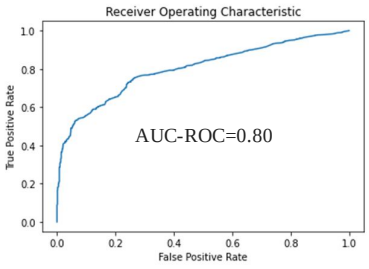


Figure 9. Near Miss

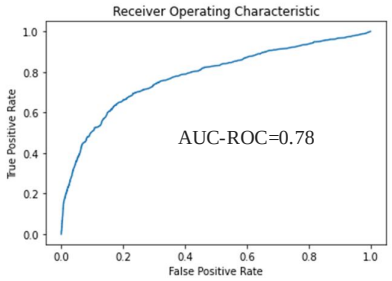


Figure 10. Tomek

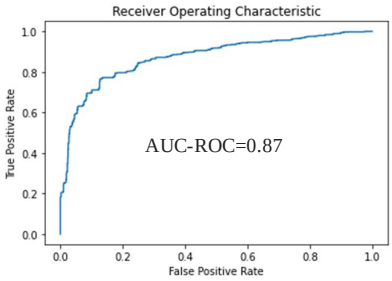


Figure 11. Instance Hardness Threshold

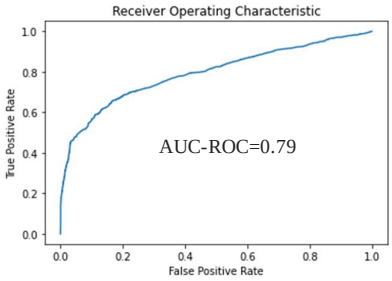


Figure 12. Neighborhood Cleaning Rule

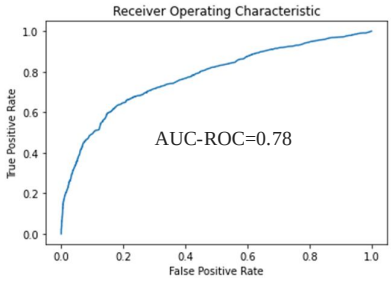


Figure 13. One Sided Selection

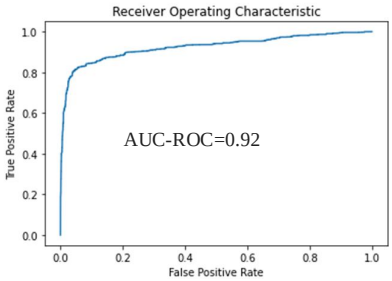


Figure 14. SMOTE+ENN

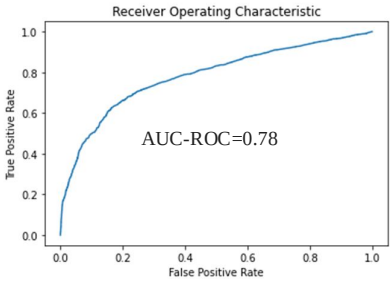


Figure 15. SMOTE+Tomek

## V. Conclusion

This article presents the study carried out to conduct a comparative analysis of fourteen re-sampling techniques. With this objective, Portuguese banking data is obtained from UCI Machine Learning Repository. The data reflects imbalanced nature that is, among 41188 total records, 36548 observations are negative while 4640 are positive observations. In-order to overcome the problem of unbalanced nature, re-sampling has been performed over the data by applying various re-sampling techniques. The logistic regression model is developed for the re-sampled data. The developed models are evaluated through different performance metrics including AUC-ROC, Accuracy, Precision, Recall, and F1-Score. The obtained empirical results showed that the K-means SMOTE is the best suitable and SMOTE+ENN is the second best suitable re-sampling technique for logistic regression algorithm while using Portuguese bank marketing data.

## References

[1] J. Bulao, "How Much Data Is Created Every Day in 2021?," *techjury*, 2021. .

[2] C. El Morr and H. Ali-Hassan, "Descriptive, Predictive, and Prescriptive Analytics," pp. 31–55, 2019, doi: 10.1007/978-3-030-04506-7_3.

[3] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets : A review," *Science (80-. ).*, vol. 30, no. 1, pp. 25–36, 2006.

[4] K. Fujiwara *et al.*, "Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis," *Front. Public Heal.*, vol. 8, no. May, pp. 1–15, 2020, doi: 10.3389/fpubh.2020.00178.

[5] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, no. April, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002.
[7] B.-H. M. Hui Han, Wen-Yuan Wang, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing*, 2005, p. 878, [Online]. Available: https://link.springer.com/chapter/10.1007/11538059_91.
[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proc. Int. Jt. Conf. Neural Networks*, no. March, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.
[9] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, doi: 10.1016/j.ins.2018.06.056.
[10] T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, "A resampling method for imbalanced datasets considering noise and overlap," *Procedia Comput. Sci.*, vol. 176, pp. 420–429, 2020, doi: 10.1016/j.procs.2020.08.043.
[11] ]Y. Qian, Y. Liang, M. Li, G. Feng, and X. Shi, "A resampling ensemble algorithm for classification of imbalance problems," *Neurocomputing*, vol. 143, pp. 57–67, 2014, doi: 10.1016/j.neucom.2014.06.021.
[12] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013, doi: 10.1016/j.jesp.2013.03.013.
[13] "Portuguese Bank Marketing Data Set," *UCI Machine Larning Repository*, 2014. .
[14] W. H. B. and J. L. Rodgers, "Re-Sampling Methods," in *Handbook*, 2009.
[15] J. Yu and W. Liu, "Improved Particle Filter Algorithms Based on Partial Systematic Resampling Jinxia," *IEEE*, no. I, pp. 434–438, 2010.
[16] N. Zheng, Y. Pan, X. Yan, and R. Huan, "Hierarchical resampling architecture for distributed particle filters," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 1565–1568, 2012, doi: 10.1109/ICASSP.2012.6288191.
[17] I. Mahmoud, A. A. El Tawab, M. Salama, and H. A. A. El-Halym, "C38. Appraisal of different particle filter resampling schemes effect in robot localization," *Natl. Radio Sci. Conf. NRSC, Proc.*, no. Nrsc, pp. 477–484, 2012, doi: 10.1109/NRSC.2012.6208556.
[18] M. Bolić, P. M. Djuric, and S. Hong, "Resampling algorithms for particle filters: A computational complexity perspective," *EURASIP J. Appl. Signal Processing*, vol. 2004, no. 15, pp. 2267–2277, 2004, doi: 10.1155/S1110865704405149.
[19] H. O. Ozcan Dulger, "Factors on the Execution Times of Metropolis Resampling and its Variations," pp. 6–9.
[20] M. J. Pekala and A. J. Llorens, "Online learning with minority class resampling," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 2248–2251, 2011, doi: 10.1109/ICASSP.2011.5946929.
[21] S. Porwal and S. K. Katiyar, "Performance evaluation of various resampling techniques on IRS imagery," *2014 7th Int. Conf. Contemp. Comput. IC3 2014*, pp. 489–494, 2014, doi: 10.1109/IC3.2014.6897222.
[22] Y. Liang, Y. Fang, S. Luo, and B. Chen, "Image Resampling Detection Based on Convolutional Neural Network," *Proc. - 2019 15th Int. Conf. Comput. Intell. Secur. CIS 2019*, pp. 257–261, 2019, doi: 10.1109/CIS.2019.00061.
[23] N. Verbiest, E. Ramentol, C. Cornelis, and F. Herrera, "Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection," *Appl. Soft Comput. J.*, vol. 22, pp. 511–517, 2014, doi: 10.1016/j.asoc.2014.05.023.
[24] X. Wang and S. Xing, "The Research of ELM Ensemle Learning on Multi-class Resampling Imbalanced Data," vol. 0, pp. 7–11.
[25] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," no. August, 2016, [Online]. Available: http://arxiv.org/abs/1608.06048.
[26] K. Agustianto and P. Destarianto, "Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling," *Proc. - 2019 Int. Conf. Comput. Sci. Inf. Technol. Electr. Eng. ICOMITEE 2019*, vol. 1, no. July, pp. 86–89, 2019, doi: 10.1109/ICOMITEE.2019.8921159.
[27] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, no. April, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
[28] S. Sawangarreerak and P. Thanathamathee, "Random forest with sampling techniques for handling imbalanced prediction of university student depression," *Inf.*, vol. 11, no. 11, pp. 1–13, 2020, doi: 10.3390/info11110519.
[29] A. D. Chakravarthy, S. Bonthu, Z. Chen, and Q. Zhu, "Predictive models with resampling: A comparative study of machine learning algorithms and their performances on handling imbalanced datasets," *Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019*, pp. 1492–1495, 2019, doi: 10.1109/ICMLA.2019.00245.
[30] Z. Liang and M. A. Chapa-Martell, "Combining Resampling and Machine Learning to Improve Sleep-Wake Detection of Fitbit Wristbands," *2019 IEEE Int. Conf. Healthc. Informatics, ICHI 2019*, pp. 1–3, 2019, doi: 10.1109/ICHI.2019.8904753.
[31] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *J. Biomed. Inform.*, vol. 107, no. June, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.