Data Mining Techniques in Direct Marketing on Imbalanced Data using Tomek Link Combined with Random Under-sampling

Ümit Yılmaz
Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey
umit.yilmaz@agu.edu.tr

Cengiz Gezer
Research & Development Center, adesso Turkey, Istanbul, Turkey
cengiz.gezer@adesso.com.tr

Zafer Aydın
Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey
zafer.aydin@agu.edu.tr

V. ÇaĞrı GÜngÖr
Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey
cagri.gungor@agu.edu.tr

ABSTRACT

Determining the potential customers is very important in direct marketing. Data mining techniques are one of the most important methods for companies to determine potential customers. However, since the number of potential customers is very low compared to the number of non-potential customers, there is a class imbalance problem that significantly affects the performance of data mining techniques. In this paper, different combinations of basic and advanced resampling techniques such as Synthetic Minority Oversampling Technique (SMOTE), Tomek Link, RUS, and ROS were evaluated to improve the performance of customer classification. Different feature selection techniques are used in order the decrease the number of non-informative features from the data such as Information Gain, Gain Ratio, Chi-squared, and Relief. Classification performance was compared and utilized using several data mining techniques, such as LightGBM, XGBoost, Gradient Boost, Random Forest, AdaBoost, ANN, Logistic Regression, Decision Trees, SVC, Bagging Classifier based on ROC AUC and sensitivity metrics. A combination of Tomek Link and Random Under-Sampling as a resampling technique and Chi-squared method as feature selection algorithm showed superior performance among the other combinations. Detailed performance evaluations demonstrated that with the proposed approach, LightGBM, which is a gradient boosting algorithm based on decision tree, gave the best results among the other classifiers with 0.947 sensitivity and 0.896 ROC AUC value.

CCS CONCEPTS

• **Computing methodologies** → Machine learning.

KEYWORDS

Direct Marketing, Data Mining, Tomek Link, Machine Learning, Imbalanced Data

ACM Reference Format:
Ümit Yılmaz, Cengiz Gezer, Zafer Aydın, and V. ÇaĞrı GÜngÖr. 2021. Data Mining Techniques in Direct Marketing on Imbalanced Data using Tomek Link Combined with Random Under-sampling. In *2021 the 5th International Conference on Information System and Data Mining (ICISDM 2021), May 27–29, 2021, Silicon Valley, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3471287.3471299

ICISDM 2021, May 27–29, 2021, Silicon Valley, CA, USA

ACM ISBN 978-1-4503-8954-9/21/05…$15.00
https://doi.org/10.1145/3471287.3471299

1 INTRODUCTION

Class imbalance is a major and significant issue in data mining. Class imbalance occurs when the number of samples in the classification categories is not approximately equal. Especially for direct marketing, the number of potential hot lead customers is generally very low compared to non-potential customers. As a result, there is a very important challenge for companies who want to identify potential customers. Data mining strategies are one of the most powerful tools for companies to detect potential customers however imbalanced data still has a significant negative impact on the performance of data mining techniques. Hence, different kinds of solutions, such as data level and algorithmic level, have been suggested to solve imbalanced data problems in the literature.

Splitting data to train and test without any adjustment will not be enough to solve the imbalanced data problem, since both test and train datasets will still have the same class distribution. As a result, there are some basic and advanced sampling methods that are used to balance the distribution of minor and major classes. Randomly undersampling the major class, randomly oversampling the minor class, and doing both can be considered as basic sampling methods. On the other hand, there are advanced under-sampling methods, such as Tomek Link (T-Link) [2], and One-Sided Selection (OSS) [3]. There are also advanced over-sampling methods, such as Synthetic Minority Oversampling Technique (SMOTE) [4], and ADASYN [5]. In addition, there are different versions of SMOTE that have been developed, such as SMOTENC, bSMOTE [6].

In this paper, we aim to increase the classification performance of the different machine learning classifiers by combining different

advanced and basic resampling strategies. A comparison of combinations was conducted through different classification algorithms, such as LightGBM, XGBoost, Gradient Boosting, Random Forest, AdaBoost, Artificial Neural Network, Logistic Regression, Decision Trees, Support Vector Classifier and Bagging Classifier.

Since there is a class imbalance problem generally in direct marketing campaigns as mentioned above and companies try to find out which customers are most likely to accept the proposed offer, misclassification of "yes" customers who accept the offer is highly costly compared to the misclassification of "no" customers who don't accept the offer. As a result, in this paper, the main classification performance comparison is done based on area under the ROC curve (AUC) and sensitivity value which is the ratio of true positives to all positives. In addition, different feature selection techniques were used in order the decrease the number of non-informative features, needed computational power and increase the performance of classification algorithms. These feature selection algorithms are Information Gain, Gain Ratio, Chi-squared and Relief.

To the best of our knowledge, this is the first study focusing on showing the performance evaluations of different combinations of basic and advanced resampling strategies using different machine learning classifiers and feature selection methods in direct marketing which will be the main contribution of this paper to the literature. Furthermore, comparison of different feature selection algorithms which are tuned using Bayesian Search optimization technique and detailed performance evaluation of proposed approach based on Accuracy, AUC ROC, F1 Score, Specificity and Sensitivity values will be the other contributions of this paper to the literature. In addition, this paper shows that LightGBM classification algorithm has superior performance among the other classifiers.

This paper is organized as follows. Details of existing studies which use the same dataset are explained in Section 2. In section 3, the dataset description is provided. In section 4, data pre-processing steps are explained. In Section 5, the data balancing strategy is explained. In Section 6, performance comparison of data balancing strategies is presented. In Section 7, performance comparison of feature selection algorithms is presented. In section 8, the proposed approach is presented. In section 9, detailed performance comparison of classification algorithms is explained. In section 10, improvement in model training time is presented. Finally, in section 11, paper is concluded.

2 LITERATURE SURVEY

Until now, there have been many attempts to apply different data mining techniques to data imbalance problem. Since there is a class imbalance problem in the dataset, the comparison will be based on the AUC of the ROC curve and the sensitivity values.

Iqbal and Farooqi have applied 5 different common machine learning models including k-nearest neighbor, naive Bayes, artificial neural network, J48 decision tree, and sequential minimal optimization [7]. Their results showed that the overall performance of J48 is the best model compared to others. Amini et al. have tried to use an ensemble classification method by removing imbalance in the data using clustering and an under-sampling method [8]. As an ensemble approach predictions of multiple classifications algorithms have been combined and better performance results are obtained. Makassar et al. proposed an Adaptive Boosted Support Vector Machine method and showed that its sensitivity value is much higher than the ordinary SVM approach [9]. They have randomly under-sampled the majority class to obtain a balanced dataset. Moro et al. proposed a data-driven approach to predict subscribed customers [10]. They have compared four different data mining techniques, and these are Decision Tree, Artificial Neural Networks, Support Vector Machine, and Logistic Regression. Their results showed that the Artificial Neural Network has the best performance results. Lastly, Cherif et al. proposed a new approach by implicitly fostering the most important features and using these features to predict target customers [11].

Although all these existing studies provide valuable foundations to direct marketing, there is no internationally accepted standard approach for imbalanced datasets in direct marketing. This paper aims to fulfill this gap and show the classification performance of the different machine learning classifiers by combining different advanced and basic resampling strategies.

3 THE DATA

The data is taken from the University of California at Irvine (UCI) Machine Learning Repository related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be 'yes' or 'no' subscribed. The classification goal is to predict if the client will subscribe (yes/no) to a term deposit (variable y). The dataset contains a total of 41,188 customers' information however only 4,640 of them have subscribed to term deposit and the remaining 36,548 customers did not subscribe. This situation clearly shows that there is a huge class imbalance problem. Description of each feature in this dataset is provided below directly from (UCI) Machine Learning Repository website:

- age (numeric)
- job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
- education:(categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- loan: has personal loan? (categorical: 'no', 'yes', 'unknown')
- contact: contact communication type (categorical: 'cellular', 'telephone')
- month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')

- duration: last contact duration, in seconds (numeric).
- emp.var.rate: employment variation rate - quarterly indicator (numeric)
- cons.price.idx: consumer price index - monthly indicator (numeric)
- cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- euribor3m: euribor 3-month rate - daily indicator (numeric)
- nr.employed: number of employees - quarterly indicator (numeric)
- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- pdays: number of days passed after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- previous: number of contacts performed before this campaign and for this client (numeric)
- poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

4 DATA PRE-PROCESSING

Actual output, 'y' values are converted from 'yes'-'no' to binary values. Since the dataset contains categorical features, firstly one hot encoding method is applied to the data, which creates new columns using all the unique features of the categorical features. After one hot encoding process, the number of columns in the dataset became 63. Before applying any scaling and normalization for the correct evaluation purposes, the dataset is split into train, validation and test with a 70% train 10% validation and 20% test ratio. Standard scaling is applied by removing the mean and scaling to unit variance. Standard scaling is applied just for the numeric columns, so categorical columns which are encoded with one hot encoding method did not change. New X values are calculated by dividing the difference of X and U by S where U is the mean of each feature and S is the standard deviation of each feature and X is the value of that feature. Scaling is performed independently on each feature by computing the relevant statistics on the samples in the training set. The same U and S values are used for both training and testing splits for fair evaluation purposes.

5 DATA BALANCING STRATEGY

All the under-sampling and oversampling methods were applied after the train, validation and test split and they are applied to the train set only for fair evaluation and to avoid memorized and biased results. Some advanced over-sampling methods such as Synthetic Minority Oversampling Technique (SMOTE) [4], and ADASYN [5] are applied to the dataset and the results show that SMOTE performed better. On the other hand, some advanced under-sampling methods such as Tomek Link (T-Link) [2], and One-Sided Selection (OSS) [3] are applied to the dataset and the results show that Tomek Link performed better.

Noisy samples are close to the class boundaries which makes the job of classification algorithms harder since algorithms try to understand the differences between classes. As a result, removing noisy samples from the data makes the class boundaries clearer. Since Tomek Link is a way of removing the noisy samples, it is combined with other resampling strategies and a total of six resampled data sets are obtained. These are ROS, SMOTE, RUS, Tomek Link, SMOTE/Tomek Link, ROS/Tomek Link, and RUS/Tomek Link. For the combination of resampling algorithms, first Tomek Link is applied to data and noisy samples are removed then other under-sampling methods are applied.

For the comparison of these different resampled datasets, 10 different data mining techniques have been applied. Since the dataset is highly unbalanced in the original dataset while there was a high accuracy, there was a very bad value of area under the ROC curve (AUC) and sensitivity (recall) values. In addition, these values are the most important values for the company because they are directly related to the number of customers who accept the proposed offer and companies do not want to miss them. Consequently, these two values were used for the performance comparison of classification algorithms using the validation data on different versions of resampled datasets.

6 PERFORMANCE COMPARISON OF DATA BALANCING STRATEGIES

AUC values of classification algorithms have increased between 12-19% in T-link/RUS data as compared to the original data (Table 1).

AUC and Sensitivity results showed improved performance when Tomek Link and Random Under-sampling techniques are combined. Only Logistic Regression and AdaBoost showed better sensitivity results and AdaBoost, ANN, Logistic Regression, and Bagging obtained better AUC results using Random Under-sampling technique. Sensitivity values of classification algorithms: LightGBM, XGBoost, Gradient Boost, Random Forest, AdaBoost, ANN, Logistic Regression, Decision Trees, SVC, Bagging Classifier has increased in the T-link/RUS data as compared to original data from 56% to 95%, 51% to 94%, 53% to 94%, 45% to 93%, 41% to 87%, 46% to 85%, 42% to 88%, 50% to 84%, 38% to 92%, 45% to 89%, respectively (Table 2).

7 PERFORMANCE COMPARISON OF FEATURE SELECTION ALGORITHMS

For all feature selection algorithms, Tomek Link combined with Random Under-sampling strategy is applied as resampling method since it has the best results among the other combinations and validation data is used for the performance evaluation. (Table 2 and Table 3)

The best ROC AUC value, 0.895, is obtained using LightGBM classification algorithm with the Chi-squared feature selection method. This result showed superior performance than the ROC AUC value of the best resampling combination (Tomek Link) which was 0.888. (Table 3)

The best ROC AUC value, 0.895, is obtained using LightGBM classification algorithm with Chi-squared feature selection method. (Table 3) This result showed superior performance than the ROC AUC value of the best resampling combination (Tomek Link) which was 0.886. (Table 1)

The best Sensitivity value, 0.947, is obtained using LightGBM classification algorithm with Chi-squared feature selection method.

Table 1: ROC AUC Comparison of Classification Algorithms using different Sampling Techniques. The values in bold are the best results achieved by each of the individual classifiers.

Method	Original	ROS	SMOTE	RUS	T-link	T-link/SMOTE	T-link/ROS	T-link/RUS
LightGBM	0.763	0.886	0.796	0.884	0.782	0.813	0.884	**0.886**
XGBoost	0.738	0.881	0.848	0.878	0.765	0.851	0.881	**0.881**
Gradient Boost	0.747	0.880	0.849	0.880	0.773	0.855	0.881	**0.882**
Random Forest	0.710	0.761	0.770	0.877	0.733	0.787	0.782	**0.880**
AdaBoost	0.691	0.867	0.751	**0.874**	0.707	0.767	0.872	0.869
ANN	0.710	0.753	0.751	**0.841**	0.728	0.769	0.798	0.839
Logistic Reg.	0.696	0.867	0.863	**0.865**	0.721	0.863	0.868	0.863
Decision Trees	0.721	0.722	0.746	0.834	0.735	0.767	0.741	**0.840**
SVC	0.676	0.861	0.835	0.868	0.715	0.839	0.863	**0.874**
Bagging	0.704	0.765	0.773	**0.871**	0.737	0.789	0.779	0.870

Table 2: Sensitivity Comparison of Classification Algorithms using different Sampling Techniques. The values in bold are the best results achieved by each of the individual classifiers.

Method	Original	ROS	SMOTE	RUS	T-link	T-link/SMOTE	T-link/ROS	T-link/RUS
LightGBM	0.565	0.906	0.646	0.927	0.612	0.688	0.906	**0.935**
XGBoost	0.507	0.929	0.792	0.927	0.572	0.803	0.924	**0.931**
Gradient Boost	0.527	0.917	0.792	0.924	0.589	0.810	0.926	**0.927**
Random Forest	0.449	0.571	0.594	0.914	0.503	0.635	0.622	**0.923**
AdaBoost	0.407	0.863	0.548	**0.881**	0.443	0.584	0.876	0.871
ANN	0.464	0.582	0.575	0.844	0.513	0.629	0.697	**0.853**
Logistic Reg.	0.421	0.880	0.864	**0.879**	0.477	0.868	0.885	0.873
Decision Trees	0.509	0.507	0.567	0.825	0.540	0.617	0.546	**0.844**
SVC	0.378	0.862	0.780	0.910	0.467	0.796	0.870	**0.917**
Bagging	0.447	0.587	0.610	0.885	0.519	0.652	0.623	**0.888**

Table 3: ROC AUC comparison of classification algorithms using different sampling techniques. The values in bold are the best results achieved by each of the individual classifiers.

Method	Relief	Information Gain	Chi-squared	Gain Ratio
LightGBM	0.850	0.888	**0.896**	0.885
XGBoost	0.839	0.878	**0.879**	0.878
Gradient Boost	0.841	0.881	0.876	**0.882**
Random Forest	0.842	0.875	**0.880**	0.875
AdaBoost	0.828	**0.872**	0.866	0.871
ANN	0.834	0.865	0.877	**0.880**
Logistic Reg.	0.832	**0.869**	0.863	0.864
Decision Trees	0.796	0.835	0.842	**0.846**
SVC	0.852	0.871	**0.876**	0.872
Bagging	0.837	**0.874**	0.864	0.865

(Table 4) This result showed a similar performance with the Sensitivity value of the best resampling combination (Tomek Link) which was 0.942. (Table 2)

Although these feature selection methods increase the performance of classifiers, their main contribution is decreasing the required time for training, need for computational power and space. For instance, according to Chi-squared algorithm results, features default, job, contact, campaign, marital, education, day_of_week, housing, and loan were removed from the dataset. Dataset has a total of 20 features and 9 of them are removed from the data and after one hot encoding process number of columns in the data was

Table 4: Sensitivity comparison of classification algorithms using different sampling techniques. The values in bold are the best results achieved by each of the individual classifiers.

Method	Relief	Information Gain	Chi-squared	Gain Ratio
LightGBM	0.873	0.928	**0.947**	0.929
XGBoost	0.849	0.902	**0.923**	0.923
Gradient Boost	0.852	**0.926**	0.912	0.925
Random Forest	0.863	0.906	**0.913**	0.905
AdaBoost	0.817	0.882	0.875	**0.883**
ANN	0.839	0.893	0.919	**0.931**
Logistic Reg.	0.808	0.879	**0.886**	0.876
Decision Trees	0.794	0.834	0.846	**0.856**
SVC	0.873	0.918	**0.922**	0.919
Bagging	0.843	**0.899**	0.873	0.876

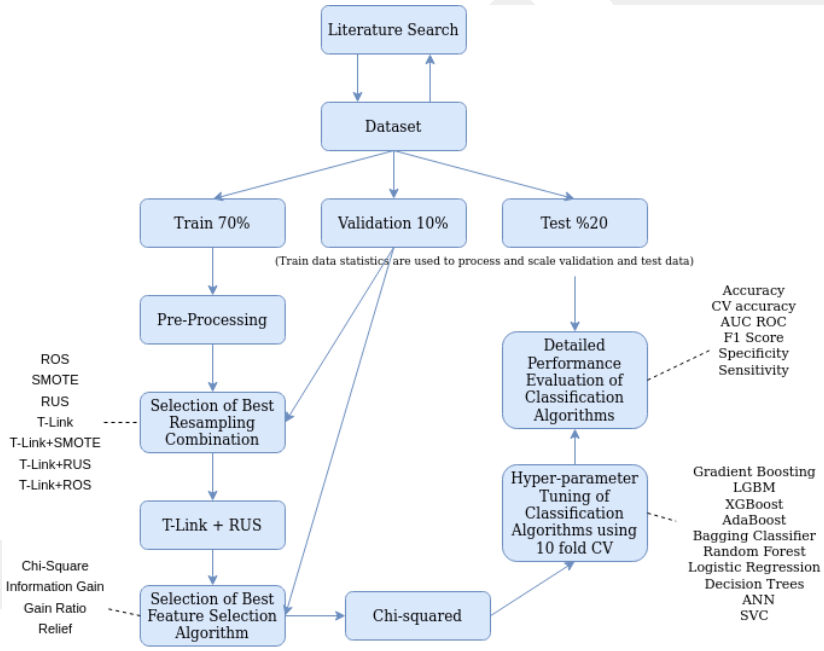

Figure 5: Proposed Approach Diagram.

63, thanks to Chi-squared feature selection algorithm it also has decreased to 22 which means a 65% reduction.

8 PROPOSED APPROACH DIAGRAM

Figure 5 which is the proposed system diagram shows all the stages of our research. Firstly, the literature search was done to see which datasets and approaches are proposed previously for imbalanced direct marketing datasets.

A dataset from UCI Machine Learning Repository was selected and spitted into 3 parts as train, validation and test. Pre-processing steps (one-hot encoding and scaling) are made for the train part of the dataset and statistics of train data were used in order to process and scale the test and validation data. Combination of resampling strategies are evaluated using validation data. Since Tomek Link and Random under-sampling strategy has the best performance results among the other resampling strategy combinations, the Tomek-link under-sampling method was applied to processed train data to create a clearer separation between different classes and balance the data. Then, to balance the data more, samples from the majority class (0, “no” customers) are randomly under-sampled. Feature selection algorithms are evaluated using validation data for all machine learning classifiers and Chi-squared method showed superior performance than the other methods. Later, the dataset which is resampled using Tomek Link and random under-sampling strategy was used for Chi-square feature selection algorithm, and according to Chi-squared algorithm results, features default, job, contact, campaign, marital, education, day_of_week, housing and

Table 15: Performance Evaluation of Classification Algorithms using Tomek Link/RUS data with Chi-squared feature selection algorithm. The values in bold are the best results achieved among all classifiers.

Method	Accuracy	Average accuracy of 10-fold CV	AUC ROC	F1 Score	Specificity	Sensitivity
LightGBM	0.856	0.901	**0.896**	**0.595**	0.843	**0.947**
XGBoost	0.848	**0.903**	0.887	0.582	0.837	0.938
Gradient Boost	0.846	0.903	0.882	0.577	0.835	0.929
Random Forest	0.842	0.902	0.885	0.573	0.829	0.939
AdaBoost	**0.868**	0.890	0.876	0.594	**0.855**	0.887
ANN	0.844	0.896	0.879	0.572	0.833	0.925
Logistic Reg.	0.856	0.896	0.874	0.584	0.850	0.898
Decision Trees	0.820	0.896	0.863	0.535	0.807	0.918
SVC	0.839	0.893	0.877	0.548	0.814	0.929
Bagging	0.844	0.896	0.886	0.577	0.832	0.937

Table 16: Comparison of existing studies on the same UCI Bank Customer dataset. The values in bold are the best results achieved among all existing studies.

Reference	Year	Method	ROC AUC	Sensitivity
Farooqi et al. [7]	2019	J48 Decision Tree	0.884	53.80%
Amini et al. [8]	2015	Ensemble Approach	0.80	83.23%
Makassar et al. [9]	2017	Adaboost SVM	NA	91,65 %
Moro et al. [10]	2014	Artificial Neural Network	0.80	NA
Cherif et al. [11]	2018	Fostering Approach	NA	50.00%
Proposed Approach	**2020**	**T-link + RUS + Chi-squared + LGBM**	**0.896**	**94.73%**

loan were removed from the dataset. Then, parameters of classification algorithms are tuned using the Bayesan Search optimization technique. Finally, a detailed performance comparison is conducted for all classification algorithms in terms of Accuracy, Average accuracy of 10-fold CV, AUC ROC, F1 Score, Sensitivity and Specificity using the test data.

9 TEST PERFORMANCE USING THE OPTIMUM PRE-PROCESSING COMBINATION WITH TUNED HYPER-PARAMETER VALUES

Performance comparison of studied classification algorithms is shown in Table 15. Based on these results, LightGBM classification algorithm demonstrated superior performance compared to other classifiers. The only specificity of AdaBoost algorithm is best among the classifiers, however, since this is an unbalanced dataset and misclassification of "yes" customers who accept the offer is highly costly compared to the misclassification of "no" customers who don't accept the offer. As a result, in this paper, the main performance comparison is done based on sensitivity and ROC AUC values.

10 CONCLUSION

Machine Learning classification algorithms require a balanced class distribution for good performance results because applying standard classification algorithms without data level or algorithmic level adjustment causes a biased classification towards the majority class. Many experiments have been done in the literature both using data level and algorithmic level solutions. However, there is no absolute winner because sometimes resampling strategies performed better while sometimes cost-sensitive and algorithmic level solutions performed better.

Tomek Link which is an advanced under-sampling strategy increased the performance of algorithms. It was removing the noise in the data and making the class boundaries more distinctive however a balanced distribution of the data still was not achieved. Tomek Link technique itself was not enough. As a result, in this paper, our approach proposed a solution by combining Tomek Link with different resampling strategies such as Random Over-Sampling, Synthetic Minority Oversampling Technique (SMOTE) and Random Under-sampling. On this dataset, the combination of Tomek Link with Random Under-Sampling showed better performance results among all studied classifiers. In addition, different feature selection algorithms are applied to eliminate the non-informative features and increase the performance of the model and decrease the computational need and space. Chi-squared algorithm showed better performance among all feature selection algorithms and eliminated 9 features from the dataset. LightGBM classifier showed a

Table 17: Required time before and after proposed pre-processing combination in seconds.

Method	Before proposed pre-processing approach	Average accuracy of 10-fold CV
LightGBM	9.857	3.239
XGBoost	59.012	4.900
Gradient Boosting	477.459	19.195
Random Forest	32.347	6.725
AdaBoost	269.571	44.480
ANN	672.439	94.526
Logistic Regression	642.373	92.294
Decision Trees	801.799	92.325
SVC	468.402	100.276
Bagging	587.177	92.107

superior performance among the other classifiers and its hyperparameters are tuned using the Randomized Search CV optimization technique. As a result, the proposed approach which is a combination of Tomek Link, Random Under-sampling and Chi-squared feature selection algorithms and tuned LightGBM algorithm has increased the performance results significantly. Thanks to the proposed approach, sensitivity value which is the most important value for this kind of imbalanced datasets since it has a very high cost to the companies has increased from 0.565 to 0.947. Also, the ROC AUC value has increased from 0.763 to 0.896 thanks to the proposed approach. Besides, the required time for training has decreased between 60% to %92.

REFERENCES

[1] H. A. Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," Int. J. Comput. Appl., vol. 85, no. 7, pp. 12–22, 201

[2] I. Tomek, "Two modifications of CNN," IEEE Transactions on Systems, Man, and Cybernetics, vol. 6, pp. 769-772, 1976.

[3] M. Kubat, S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," In Proceedings of the 14th International Conference on Machine Learning, vol. 97, pp. 179-186, 1997.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[5] H. He, Y. Bai, E. A. Garcia, S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," In Proceedings of the 5th IEEE International Joint Conference on Neural Networks, pp. 1322-1328, 2008.

[6] H. Han, W.-Y. Wang, B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," In Proceedings of the 1st International Conference on Intelligent Computing, pp. 878-887, 2005.

[7] Iqbal, N., & Farooqi, R. (2019). Performance Evaluation for Competency of Bank Telemarketing Prediction using Data Mining Techniques. International Journal of Recent Technology and Engineering, 8(2), 5666–5674

[8] Amini, Mohammad & Rezaeenour, Jalal & Hadavandi, Esmaeil. (2015). A Cluster-Based Data Balancing Ensemble Classifier for Response Modeling in Bank Direct Marketing. International Journal of Computational Intelligence and Applications. 14. 1550022. 10.1142/S1469026815500224.

[9] Lawi, A., Velayaty, A. A., & Zainuddin, Z. (2017). On identifying potential direct marketing consumers using adaptive boosted support vector machine. 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT).

[10] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

[11] Koumetio Tekouabou, Cédric Stéphane & Cherif, Walid & Hassan, Silkan. (2018). Optimizing the prediction of telemarketing target calls by a classification technique. 1-6. 10.1109/WINCOM.2018.8629675.

[12] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS*.

[13] Chen, Tianqi, and Carlos Guestrin. "XGBoost." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): n. pag. Crossref. Web.

[14] Natekin, Alexey & Knoll, Alois. (2013). Gradient Boosting Machines, A Tutorial. Frontiers in neurorobotics. 7. 21. 10.3389/fnbot.2013.00021.

[15] Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[16] Breiman, L. Bagging Predictors. *Machine Learning* **24,** 123–140 (1996). https://doi.org/10.1023/A:1018054314350