PREDICTION ACCURACY OF FINANCIAL DATA - APPLYING SEVERAL RESAMPLING TECHNIQUES

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Mohammad Reza Ali

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

October 2020

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Prediction Accuracy of Financial Data - Applying Several Resampling Techniques

By

Mohammad Reza Ali

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Simone Ludwig
Chair

Dr. Oksana Myronovych

Dr. Ying Huang

Approved:

10/31/2020
Date

Dr. Simone Ludwig
Department Chair

ABSTRACT

With the help of Data Mining and Machine Learning, prediction has been a very popular and demanding instrument to plan and accomplish a future goal. The financial sector is one of the crucial sectors of present human society. Predicting the correct outcome is a pivotal matter in this sector. In this work, an assessment was done to the prediction efficiency by applying several Machine Learning Classification Algorithms and resampling methods. These techniques were applied to financial data, more specifically to Bank Marketing in order to predict the tendency of clients to subscribe to a bank term deposit. For the correct prediction of the outcome, imbalance in the data set affects the results greatly. Consequently, the prediction becomes inaccurate. Researchers are working this issue and many investigators are using different methods. This research paper uses some sampling techniques together with several conventional Machine Learning algorithms to improve the prediction precision.

ACKNOWLEDGEMENTS

I would like to thank my research advisor Dr. Simone Ludwig and others who were surrounded around me in order to get through process of each step of my successes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Machine learning is a way of analyzing data based on a model built to get an informative result for justifying the data. With the high-tech devices, day by day people have trillions of various data in their control. To make efficient use by extracting some in-depth insight information machine learning techniques are used quite rigorously.

Data Mining is the non-trivial extraction of implicit, previously unknown and potentially useful information from data. Data Mining describes the whole process from data preparation, the actual mining of the data to infer patterns, and then the post-processing of the results. Machine learning describes the different algorithms that can be used during the data mining process.

Finance is dealing with money in so many aspects such as management, creation, study, investment, future, business, profit etc. on a large scale. A Bank is one of the major financial services in the society. For interacting with the banking sector, planning and managing, as well as prediction is an inevitable part. Data Mining is a very demanding and popular solution to meet the need of this sector.

Banks are using almost all sort of advanced technology. There is a huge amount of data which are generated and managed by the Banks. Thus, identifying future assumption by applying Machine Learning Techniques to those data will help in the area of finance.

In helping the overall financial system, applying data mining helps to deliver better service, increase operational efficiency, enhance security, etc. To offer their customers new products, recommend experiencing personalized services, animate chatbots, and so many demanding and encouraging features, the Bank also leverages predictive analytics and machine learning.

The data that has been used was collected from the real world. It was used in a bank for their marketing purpose to entice its customer to subscribe to a new financial service making a term deposit. The intention is to predict how many customers are going to avail this service. This

research attempted to use some techniques in machine learning to identify which techniques are performing better in the case of prediction. Some of the data was used for training and remainder for testing.

Traditional algorithms used before are different from applying Machine Learning algorithms. These ML (Machine Learning) algorithms are used to train a set of data and build a model. It is the core engine or actor to build a mathematical model from data. The algorithms are improved by themselves based on experiences acquired based on the provided data. This experience is nothing but getting more data for training or learning to adjust/improve the model. The experiments are implemented such that several different algorithms are applied for future hypothesis of the data.

Some Classification techniques were going to be used for identifying if the bank customers are going to subscribe to the service or not. That means, it classifies clients as new service subscriber or not. For this reason, several supervised learning algorithms were used such as Decision Tree, Random Forest, XGBoost to compare the performance of the learning task. For the classification or categorization purpose, supervised learning approaches have been used.

Due to the imbalance in the data, resampling techniques provide better and more accurate results. As the bank data, what is imbalance data, several resampling techniques were used to avoid miss interpretation and to achieve more precise result.

For the measure of performance, this research uses some statistical measures such as F1 score, Precision, Recall, Accuracy and AUC. All the results were observed after training the data and how each algorithm is performing on actual test data and after resampling was applied.

2. RELATED WORK

The banking sector is one of the crucial sectors of business. The competition increases day by day. However, whoever will have the better prediction about the future will be in a favorable position, and thus, different technologies are in use. For the prediction task, several machine learning techniques are used [1]-[8].

One of the approaches used in [9] applies both the clustering and classification method. For clustering, the K-means algorithm was used and for classification task Decision Tree has been used. The approach shows good results, however, the combination may not always become effective for banking analysis in particular when imbalanced data is involved.

Patil et al. [10] mentioned about Decision Tree, K-Means, Naïve Bayes, Support Vector Machine algorithms but used Artificial Neural Network and thus a comparative analysis of Artificial Neural Network with other algorithms is needed.

A more closely related work was done by Valarmathi et al. [11]. The authors are working on imbalanced data set in the banking sector using the technique of dimensionality reduction. They were using Naïve Bayes, J48, KNN and the Bayesnet algorithms.

Like the banking sector, machine learning is going to be implemented in so many fields. One of the mentionable fields is medical science and bioinformatics. This research attempt has been encouraged by one of the approaches which was used in the study of Breast Cancer. Kabir et al. [12] have used some Resampling Techniques for identifying the risk factor of Breast Cancer. Thus, their approach of applying the Resampling Techniques had been copied. Though, the authors used 6 resampling techniques and this paper 10 resampling techniques are used.

3. DATA SET

The data set was collected from the UCI Machine Learning Repository [13]. It is a free data set available to the public. The data is actually from a Portuguese Bank Institution and was used for telemarketing purposes. The collection of data was for a period of several years. The purpose was to encourage the clients to make use of term deposit services of that bank, which was accomplished through phone calls.

There are 4 version of the same data set provided. This investigation has chosen the full data set (bank-additional-full.csv) with all examples (41,188) and 20 input variables. The goal was to forecast whether the client is going to subscribe (Yes/No) to the term deposit (variable y).

The data set has a total of 20 attributes as input. Some attributes are focusing on the client's own information, some are for contacts of the current related campaigns, and the socio economic context has been used for some attributes. In addition, other previous campaigns related information has also been used. The 1st input variable represents the age of the client involved. The 2nd input is for the type of job of the client, and the 3rd input represents the marital status. All the variables have been explained in Table 1.

Table 1: Description of Input Variables

No	Variable Name	Explanation
1	age	numeric
2	job	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3	marital	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4	education	(categorical: basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5	default	has credit in default? (categorical: 'no','yes','unknown')
6	housing	has housing loan? (categorical: 'no','yes','unknown')
7	loan	has personal loan? (categorical: 'no','yes','unknown')
8	contact	contact communication type (categorical: 'cellular','telephone')
9	month	last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10	day_of_week	last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11	duration	last contact duration, in seconds (numeric)
12	campaign	number of contacts performed during this campaign and for this client (numeric, includes last contact)
13	pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14	previous	number of contacts performed before this campaign and for this client (numeric)
15	poutcome	outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
16	emp.var.rate	employment variation rate - quarterly indicator (numeric)
17	cons.price.idx	consumer price index - monthly indicator (numeric)
18	cons.conf.idx	consumer confidence index - monthly indicator (numeric)
19	euribor3m	euribor 3 month rate - daily indicator (numeric)
20	nr.employed	number of employees - quarterly indicator (numeric)

This data set has a total of 41,188 records. There are several missing attributes in some category of data. All the missing information are labeled as Unknown Data.

4. APPROACH

The experiments were done in multiple stages. Firstly, the preprocessing of the data was done. Afterwards, the data had to be analyzed to check whether it was imbalanced. Then, the algorithm to be applied had to be chosen. As a matter of comparison and to seek better performance, research was conducted to compare techniques that could be used for the classification task. Lastly, several resampling techniques were selected. Both the actual data and the resampled data was prepared, and a final analysis was done. In terms of result evaluation, some metrics have been selected, which are precision, recall, F1-score, accuracy and AUC.

The data set is multi-valued with some values being of text type. For example, if the attribute is 'job', the possible values are housemaid, retired, un-employed, self-employed, technician, etc. But for the case of evaluation those entries may not be recognized. Thus, all text labels were converted to numeric values. Some of the attributes had values like Yes or No. An instance of this Boolean attribute is 'Has Loan', and the Yes and No labels were converted to 1 or 0, respectively. Thus, the preprocessing took care of converting the text labels to numeric values.

The total number of records in this data set is 41,188, but due to some missing values all the records with unknown label were removed. Because, unknown or missing values can lead to wrong results. After the removal, the total number of records was 30,488. Exploring this data set after preprocessing, the output variable is of type boolean with yes and no labels. Thus, binary classification-based machine learning algorithms were chosen. Also, the investigation of the data set found that 3,859 out of 30,488 records had the value yes. That means, 12.66% of customers had subscribed to the term deposit service. This shows that data set is an imbalance data set where the minority class is attributed to successful term deposit subscriptions.

When a set of elements is divided in the two groups then the process is called a Binary Classification [14]. The data set investigated has two labels for the output variable, Yes and No.

Thus, Binary Classification algorithms have to be used. Most of the popular Binary Classification algorithms are Decision Tree, Bayesian Network, Neural Networks, Support Vector Machines, K-Nearest Neighbor, Logistic Regression, Random Forests, Gradient Boosting. There are also many Binary Classification algorithms and new algorithms being proposed. Among them, there was a plan to use Decision Tree, Random Forests, XGBoost (Open Source Software Library for Gradient Boosting Algorithm), etc.

Decision Tree is the most widely used classifier. Decision Tree is a tree structure for helping to make decisions. Its structure is like a flowchart [15]. As an algorithm it shows conditional control statements. As a tree, its different parts express separate functional behavior. Such as, internal nodes show a test attribute, leaf node presents class labels, branches display the outcome of the tests. All the paths are seen as classification rules and start from the root and ends at leaf. One of the reasons to use this algorithm is because of its simplicity and the other is that it can be combined with other decision techniques.

Random Forest is another algorithm for binary classifications but can also be used for regression [16]. That is why the algorithm is also called ensembled learning technique. It constructs a multitude of decision trees and takes a mean to predict an individual tree avoiding overfitting. The reason for using this method is that in many cases the accuracy of Decision Trees are very high. Also, the decision tree algorithm is suitable to be applied to comparatively large size data sets. Furthermore, the algorithm has feature of handling missing data and also can be used for future use on other data sets. Another important fact is it can control imbalanced data set such as one that is used for this investigation.

XGBoost is an open source software library implementation of the Gradient Boosting Machine Learning algorithm. Gradient Boosting is an optimization algorithm [17] whereby the

optimization is based on a differentiable and/or loss function. It is actually an ensemble form of the weak prediction model. Generally for Machine Learning purposes, Decision Tree is used as a weak model and thus is used for the Gradient Boosting Algorithm. XGBoost has a feature of Regularization which prevents overfitting [18] and also utilizes parallel processing. This algorithm implementation use tree pruning, so in many cases it gives better performance by reducing unwanted processing steps. In addition, it also has the capability to deal with missing values.

In terms of binary classification, all tree algorithm techniques are good choices. Since only 12.6% of our data set has the positive response, the imbalance has to be handled. Thus, resampling techniques are used to improve the classification accuracy.

For the experiments, an 80:20 ratio of the data set was chosen, i.e., the training data uses 80% of full data set, and 20% is used for testing. Below is the overall statistics for the division among the train vs test sets at Table 2.

Table 2: Summary of Prepared Data Set for Experiments

Train-Test Ratio	Class = Yes	Class = No	Total
Total	3,859	26,629	30,488
Train 80%	3,087	21,303	24,390
Test 20%	772	5,326	6,098

The class imbalance influences the models for its disproportionate number of different class instances in practice [19]. Thus, to deal with this there are several ways such as cost functions and sampling. For this experiment different resampling techniques were used to get a higher classification accuracy. Two basic types of resampling were used, which are Random Over Sampling and Random Under Sampling. There are also several hybrid combinations of those over and under sampling techniques. The name of sampling techniques that were used are Random Under Sampling (RUS), Random Over Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), Extended Nearest Neighbor (ENN), Hybrid of SMOTE and ENN, Hybrid

of SMOTE and Tomek Link, Near Miss, Hybrid of K-Means and SMOTE, ADASYN, and Cluster Centroid.

Among the minority class, duplicating examples randomly and including those examples in the training data set is called Random Over Sampling [20]. Examples from the minority class are selected and added to the new improved balanced training data set and can be used for multiple times. This technique is much more suitable for those machine learning algorithms, which are affected by a skew distribution. It is also effective for multiple duplicate examples of a class that can influence the fit of the model. As an example, if the data set has 90 yes entires and 10 no entries, then the ROS will add 80 more no entries to achieve a balance among the data set.

Random Under Sampling actually deletes some majority class data from the data set. But this delete selection is chosen randomly [20]. The approach is iterative and runs until the expected class distribution is reached. Both Over and Under sampling is same but it is done in the opposite way. The intention is to induce bias of the specific class to neutralize the imbalance in the data. As an example, if there are 90 yes entries and 10 no entries, then RUS will decrease the number of yes entries to 80.

SMOTE is an over sampling technique where the over sampling is done in a different way than over-sampling with replacement [7]. The technique is taking each minority class sample and introduces synthetic example along with the K minority class nearest neighbor. The details will be discussed in the following paragraph. In SMOTE, the Kth nearest neighbor is found for the same class [21]. K difference vectors are obtained and these vectors are multiplied by a random number between 0 and 1. After multiplication, those are added to the feature vectors. For binary classification, SMOTE sampling has proven in the past to be a good choice.

Extended Nearest Neighbor is the one of the enhanced versions of Kth Nearest Neighbor. The basic principle is that similar things are near to each other. The algorithm targets one specific datum and finds the distance among this datum compared to the rest of the data records and list them. Then, it sorts the list of distances and selects the value from the lowest to Kth position. The ENN method makes a prediction via a two-way communication [22]. That means, like KNN, first it will find its nearest neighbor and then similarly, the neighbor will also find who is in the list of its neighbors. That is why the algorithm has a better chance to perform well using cross verification for finding neighbors.

Tomek link is an under sampling technique. It removes undesired overlap among classes [23]. Until all minimum distanced nearest neighbor pair exist for the same class, the major class links continues to be removed. The combination of SMOTE and Tomek Link were used together for the experiments.

Another under sampling technique is Near Miss, which is very good for extremely imbalanced data set [24]. The basic technique works as follow. First it calculates all the majority classes and all the minority classes, then it sorts the distances and selects K number of short distances between the majority and minority class. If the minority class has n numbers, then the majority class will have K*n number of majority classes. This sampling technique was used to in this investigation.

K-means is a cluster algorithm [25], but it can also be used for classification. It divides the data set into K number of clusters. Each of the datum belongs to the cluster of the nearest mean of the center. The hybrid of K-Means and SMOTE had been used. It is a combination of Over and Under sampling since K-Means is an under sampling algorithm. Another hybrid that was used is

the combination of SMOTE with ENN. Both the SMOTE and ENN was described. The combination of those sampling techniques were used.

Adaptive Synthetic Sampling Method (ADASYN) is an oversampling method. It is similar to SMOTE [26]. The difference is that after creating the samples it imposes some random values on it, which are not linearly correlated [27] and is seen as an improvement over SMOTE.

Cluster Centroid could be considered as the multi-dimensional average of a cluster [28]. It is a clustering technique, which is well established in the classification area. This technique is considered as an under sampling technique. It is close to K-Means but simpler.

For the measurements, Precision, Recall, F1-Score, Accuracy and AUC were used. Precision and Recall is important if the target is to measure the positive class [12]. For example, in this experiment, the number of yes among client was used, which is the positive class.

Before explaining these parameters, some relevant terms TP, FP, TN, FN are introduced. True Positive (TP) is number of positive samples correctly classified. False Positive (FP) is number of positive samples incorrectly classified. True Negative (TN) is number of negative examples correctly classified, and False Negative (FN) is the number of negative samples wrongly classified.

Precision is the positive predictive rate:

$$\text{Precision} = TP/(TP+FP) \tag{1}$$

Recall is the true positive rate:

$$\text{Recall} = TP/(TP+FN) \tag{2}$$

F-Measure or F1 score is a measure of test accuracy. It is accomplished by the weighted harmonic mean:

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \tag{3}$$

F1 is more effective for accuracy when the data set has imbalance classes, and there is a need to measure the accuracy of minority class.

Receiver Operating Characteristic Curve (ROC Curve) is a curve in a graph with the value of TPR versus FPR at different classification thresholds. The area under the ROC curve is called AUC [12]. It is used to measure how well the predictions are ranked by means of threshold. It also measures the quality of the model's predictions irrespective of what classification threshold is chosen. The value of the threshold is generally between 0 and 1. The larger the threshold the better the prediction. But with respect to the increase of threshold, not only the True Positives should be measured but also the False Positives. Thus, it is a tradeoff.

The number of correctly predicted data out of all is called the Accuracy measure of the model [29]:

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN) \quad (4)$$

This equation tells how well the model would predict the outcome.

All the measures here uses a value between 0 to 1, where 1 is best score.

5. RESULTS

This section will describe the different results that had been obtained while using different models and resampling techniques for the classification of the data set.

5.1. Experiment 1

For the first experiment, the overall performance has been measured of the data set applying three classification techniques. The purpose is to check the overall performance of the prediction. As mentioned earlier, 5 evaluation measures precision, recall, F1-Score, accuracy, and AUC have been used. The result at Table 3 shows that, both Recall and Accuracy have the highest values. Comparing among the three classifier models, Random Forest provides the better result than the other two (DT and XGBoost).

Table 3: Overall Performance on Original Data

Methods	Precision	Recall	F1-Score	Accuracy	AUC
DT	0.87	0.87	0.87	0.875	0.721
RF	0.90	0.91	0.90	0.910	0.747
XGBoost	0.90	0.91	0.90	0.906	0.735

Figure 1 shows the overall performance graphically below.

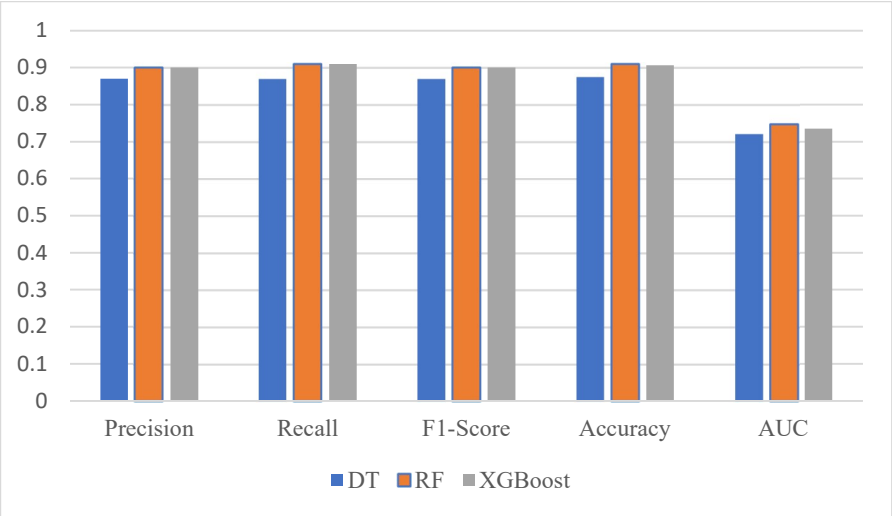

Figure 1: Overall Performance Original Data

5.2. Experiment 2

For the second experiment, we focus only on the minority class to verify its accuracy. Precision, Recall, and F1-Score were used. For this experiment, the raw data set was used to focus on the minority class. The result of the experiment is as shown in Table 4.

Table 4: Minority Class Performance on Original Data

Methods	Precision	Recall	F1-Score
DT without sampling	0.53	0.51	0.52
RF without sampling	0.69	0.53	0.60
XGBoost without sampling	0.67	0.51	0.58

From the table it can be seen that most of the values are close to 0.5, which is a poor score. The maximum value or best performance is achieved by the Random Forest classifier. Looking at precision, the best score is achieved by Random Forest with a value of 0.69. The worst value is presented by Recall for both the Decision Tree and the XGBoost Classifier. As can be seen from the table, the overall performance score is poor for the minority class, and thus this experiment shows the reason why resampling techniques were needed to obtain better accuracy results for an imbalanced data set. Figure 2 is the chart presentation of these data.

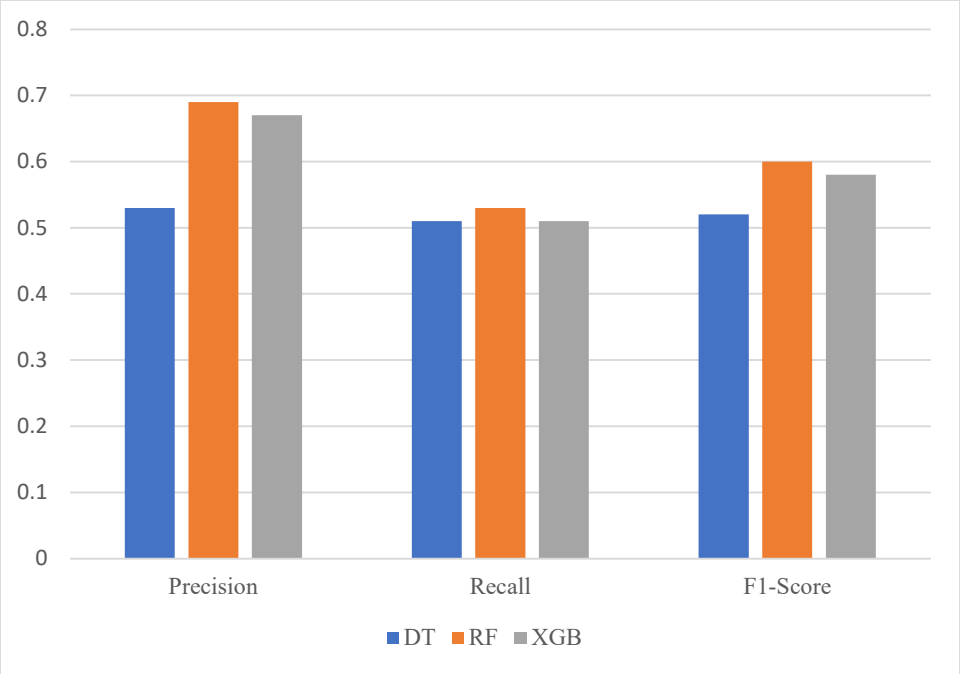

Figure 2: Minority Class Original Data

5.3. Experiment 3

In this experiment ten different sampling techniques were used to create a balanced in data set. Then decision tree classifier was applied to measure the performances. It should be noted that the experiment is done using all the data. Both the majority class and minority class are present in the data. The result of this experiment is given in Table 5.

Table 5: Overall Performance on Resampled Data (Decision Tree)

Methods	Precision	Recall	F1-Score	Accuracy	AUC
DT with RUS	0.83	0.83	0.83	0.825	0.825
DT with ROS	0.97	0.96	0.96	0.962	0.962
DT with SMOTE	0.92	0.92	0.92	0.915	0.915
DT with ENN	0.94	0.94	0.94	0.936	0.873
DT with SMOTE+ENN	0.97	0.97	0.97	0.975	0.974
DT with SMOTE+Tomek Link	0.92	0.92	0.92	0.925	0.925
DT with Near Miss	0.79	0.79	0.79	0.786	0.787
DT with KMeans_SMOTE	0.93	0.93	0.93	0.931	0.931
DT with ADASYN	0.92	0.92	0.92	0.923	0.923
DT with Cluster Centroid	0.79	0.79	0.79	0.786	0.786

After using several resampling techniques that were mentioned earlier, the overall performance is better compared to not using any resampling technique. Thus, it is desired use a balanced data set. Comparing different sampling techniques, Random Over Sampling provides better performance looking at the unique sampling techniques. On the other hand, with hybrid sampling, SMOTE with ENN provide the best results. Finally, if unique versus hybrid sampling is compared, then the hybrid sampling shows overall better results than the unique sampling technique, and the best value is achieved by SMOTE with ENN with an accuracy given as 0.975. Figure 3 represents the glance survey of the data.

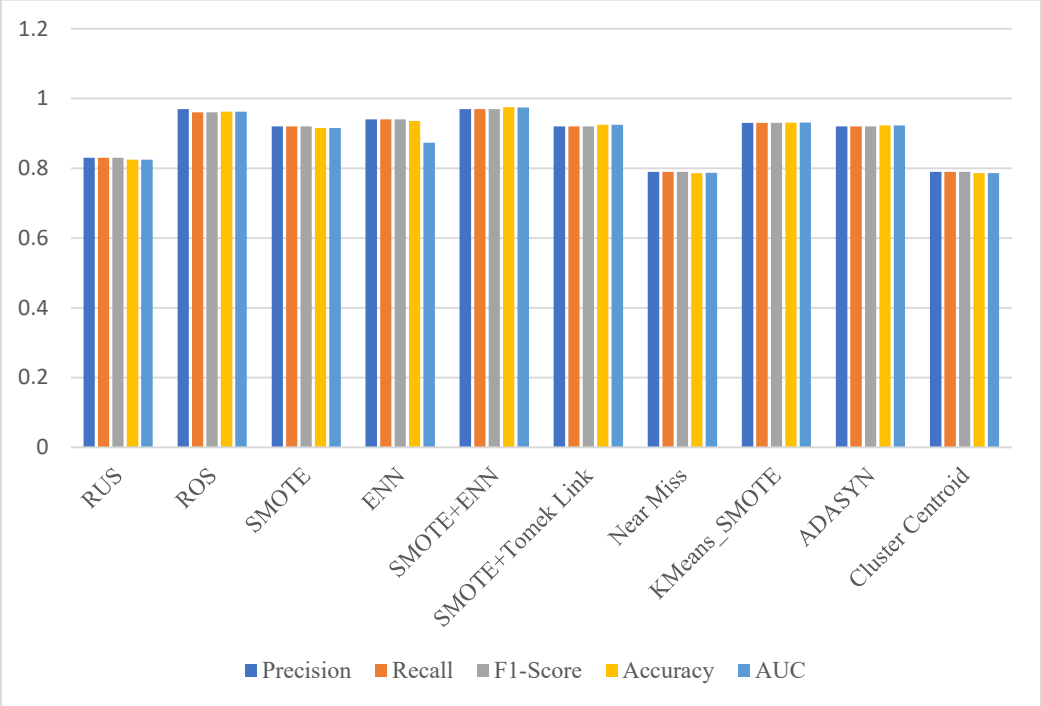

Figure 3: Decision Tree Overall Performance

5.4. Experiment 4

Though the research was made for the complete data set with sampling just to recognize that sampling with Decision Tree achieves better results, but focusing on the minority class need to be verified. The expectation was to find the prediction of the minority class. Thus, the objective

of this experiment was to use sampling techniques for the minority class only and observe the outcome. Table 6 shows the results.

Table 6: Minority Class Performance on Resampled Data (Decision Tree)

Methods	Precision	Recall	F1-Score
DT with RUS	0.83	0.83	0.83
DT with ROS	0.93	1.00	0.96
DT with SMOTE	0.91	0.92	0.92
DT with ENN	0.75	0.79	0.77
DT with SMOTE+ENN	0.97	0.98	0.98
DT with SMOTE+Tomek Link	0.92	0.93	0.92
DT with Near Miss	0.81	0.77	0.79
DT with KMeans_SMOTE	0.93	0.93	0.93
DT with ADASYN	0.92	0.93	0.92
DT with Cluster Centroid	0.78	0.78	0.78

The results are significance. For the Decision Tree with sampling, the best value was 0.53 with precision. But now, the minimum value by any sampling technique looking at precision is 0.78. Similarly, for rest of the results are also better such as Recall that is 1.00, which means absolute correct prediction, and for the F1-Score the value is 0.98. Relating to the performance of data without sampling, it performs much better. On the other hand, if the sampling techniques are compared, SMOTE with ENN gives the best result for both Precision and F1-Score, whereas Random Over Sampling (ROS) is the best for Recall. Thus, this does not show that only hybrid sampling performs best in all cases. Figure 4 is the concise form of these data description.

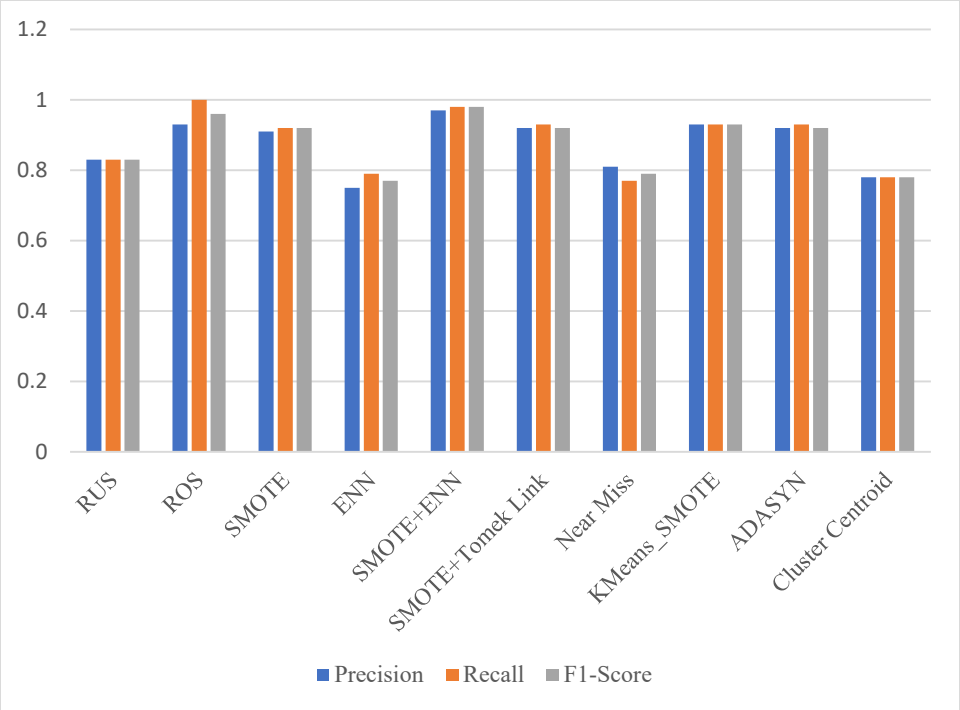

Figure 4: Decision Tree Minority Class Performance

5.5. Experiment 5

The second classifier, Random Forest, was used to do the fifth experiment with the same number of sampling techniques. After using each of the ten sampling techniques, the Random Forest machine learning algorithm was applied. Table 7 shows the results.

Table 7: Overall Performance on Resampled Data (Random Forest)

Methods	Precision	Recall	F1-Score	Accuracy	AUC
RF with RUS	0.90	0.89	0.89	0.894	0.894
RF with ROS	0.96	0.96	0.96	0.960	0.959
RF with SMOTE	0.94	0.94	0.94	0.943	0.943
RF with ENN	0.96	0.96	0.96	0.961	0.904
RF with SMOTE+ENN	0.98	0.98	0.98	0.980	0.979
RF with SMOTE+Tomek Link	0.95	0.95	0.95	0.949	0.950
RF with Near Miss	0.85	0.85	0.85	0.854	0.854
RF with KMeans_SMOTE	0.95	0.95	0.95	0.950	0.950
RF with ADASYN	0.94	0.94	0.94	0.941	0.942
RF with Cluster Centroid	0.86	0.85	0.85	0.851	0.852

This experiment is using the complete data. That means, both the binary labels are included. It is similar to Experiment 3 where Decision Tree was used with the same number of sampling techniques. However, it is unlike Experiment 1, where the raw data set had been used without any sampling techniques. On the other hand, for all three experiments, Experiment 1, Experiment 3 and Experiment 5, the AUC provides the least score. If all the sampling techniques are compared, hybrid sampling SMOTE with ENN link provides the best score. It is the same case when Decision Tree was used. But not following the same as the pervious Experiment 3 of Decision Tree, Random Under Sampling is showing the least performance compared to the others with Random Forest. SMOTE with ENN has the best performance with a Precision, Recall and F1-Score of 0.98. Random Under Sampling on the other hand achieves the least score of 0.89 for Recall and F1-Score. Figure 5 shows the data in below format.

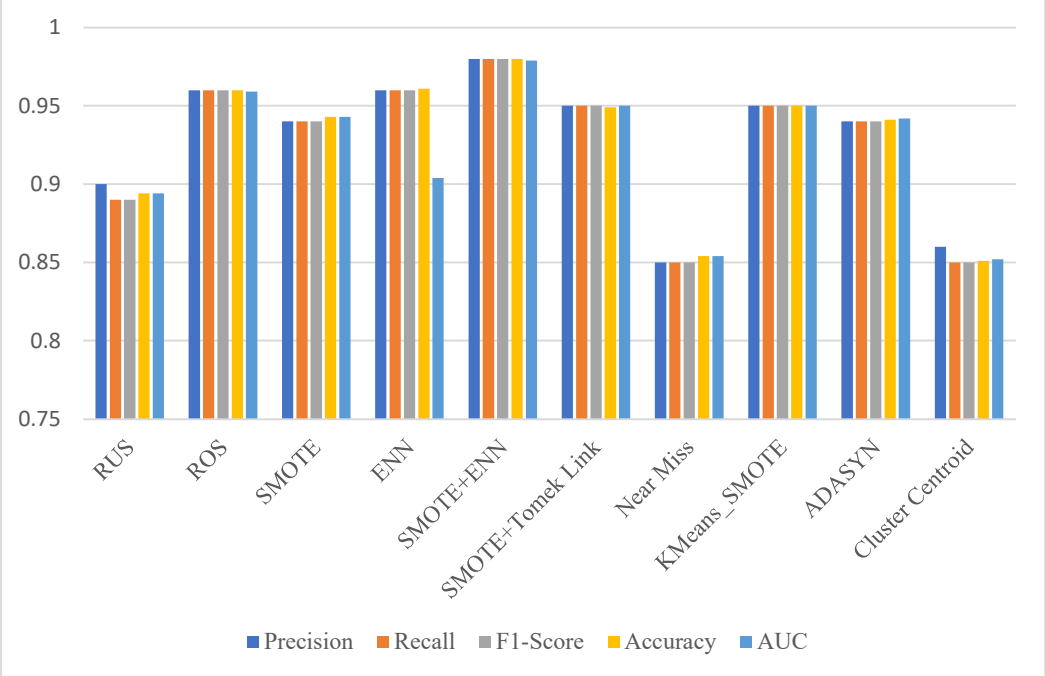

Figure 5: Random Forest Overall Performance

5.6. Experiment 6

Second testing for the minority class classifier with ten sampling techniques is going to be the Experiment 6 where the classifier is Random Forest. The plan for this experiment is almost the same as Experiment 4 where after applying each sampling techniques, Random Forest was used in place of Decision Tree. Due to the similarity of the experiment there is a good opportunity for a direct comparison. Table 8 shows the results.

Table 8: Minority Class Performance on Resampled Data (Random Forest)

Methods	Precision	Recall	F1-Score
RF with RUS	0.86	0.94	0.90
RF with ROS	0.93	1.00	0.96
RF with SMOTE	0.93	0.96	0.94
RF with ENN	0.89	0.82	0.86
RF with SMOTE+ENN	0.98	0.99	0.98
RF with SMOTE+Tomek Link	0.93	0.97	0.95
RF with Near Miss	0.87	0.84	0.85
RF with KMeans_SMOTE	0.96	0.94	0.95
RF with ADASYN	0.92	0.97	0.94
RF with Cluster Centroid	0.79	0.94	0.86

If the general performance is considered, then for the minority class, Random Forest has the better performance than Decision Tree. Though, the best score is the same for both algorithms, but Random Forest is slightly better. The best score for this experiment is Recall with a value of 1.0 and least score is Precision with a value of 0.79. But those two scores are using different resampling techniques. The former is Random Over Sampling and latter is Cluster Centroid. If it were compared without the resampled data, then it most likely would be much better. Without sampling, it showed that precision achieves a better score and Recall has a lesser score, but after sampling is applied, Recall was improved whereas Precision had declined. Another interesting fact is comparing with the overall data with both binary classes, the minority class result shows a better score which has significantly improved. It was not similar when Random Forest was used

with the data set without sampling. Among all resampling approaches, again the combination of SMOTE and ENN has given the best result. Another point to note is that for this experiment, when only ENN sampling has been used, it gave worse results. Figure 6 is the comparative expression of the table.

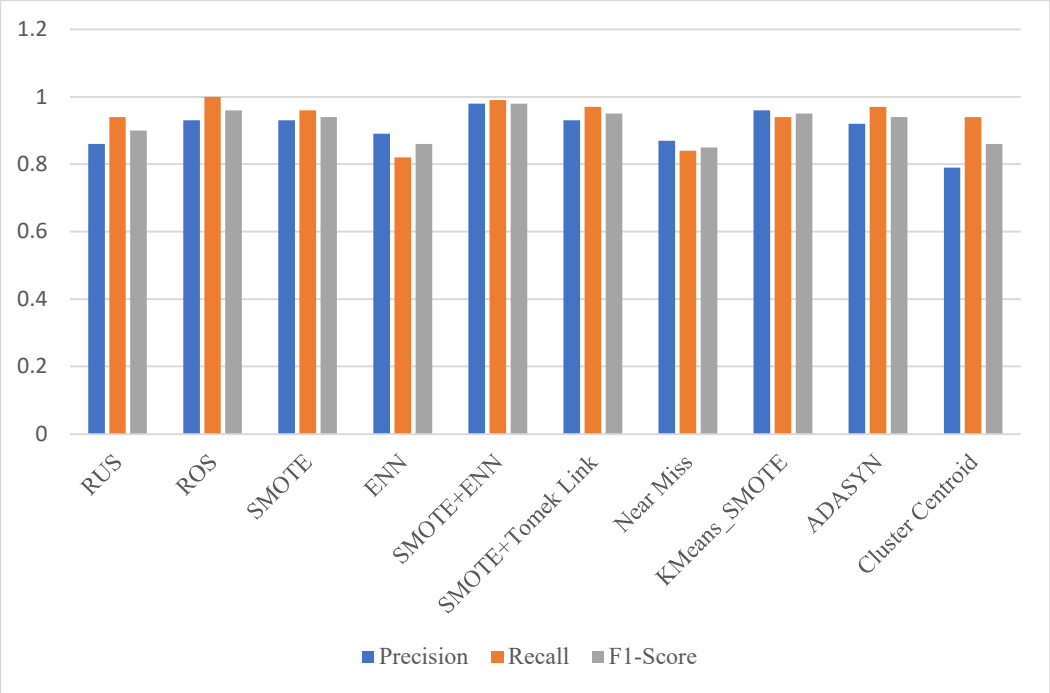

Figure 6: Random Forest Minority Class Performance

5.7. Experiment 7

The last algorithm used in this research is the XGBoost algorithm, which has been used for this experiment. Like the other experiments, except the first two, the same ten sampling algorithms were used. Each sampling procedure had been applied first followed by XGBoost. The values are provided in Table 9.

Table 9: Overall Performance on Resampled Data (XGBoost)

Methods	Precision	Recall	F1-Score	Accuracy	AUC
XGB with RUS	0.87	0.87	0.87	0.869	0.868
XGB with ROS	0.89	0.88	0.88	0.881	0.881
XGB with SMOTE	0.92	0.91	0.91	0.914	0.914
XGB with ENN	0.95	0.95	0.95	0.953	0.884
XGB with SMOTE+ENN	0.97	0.97	0.97	0.965	0.964
XGB with SMOTE+Tomek Link	0.91	0.91	0.91	0.913	0.913
XGB with Near Miss	0.85	0.85	0.85	0.851	0.851
XGB with KMeans_SMOTE	0.94	0.94	0.94	0.944	0.944
XGB with ADASYN	0.91	0.91	0.91	0.909	0.909
XGB with Cluster Centroid	0.86	0.85	0.85	0.853	0.852

Decision Tree with sampling, if compared with DT, it gave better results. However, Random Forest with sampling provides even better result than the XGBoost algorithm. Being unique from the other two algorithms, XGBoost's best value was Precision with a score of 0.97. But for the other two, it was Precision, Recall and F1-Score. The least score obtained by XGBoost algorithm was for Precision, Recall and F1-Score with a value of 0.85. In the case of the highest score, XGBoost is same as Decision Tree, but Random Forest is even better than those two. For the least value, it is same as Random Forest but the value is better than DT. Considering all ten sampling techniques, SMOTE and ENN combination is best, and Near Miss is the least valued sampling approach. In all three algorithms, SMOTE and ENN combination has given the best results. Figure 7 is another impression of the given results.

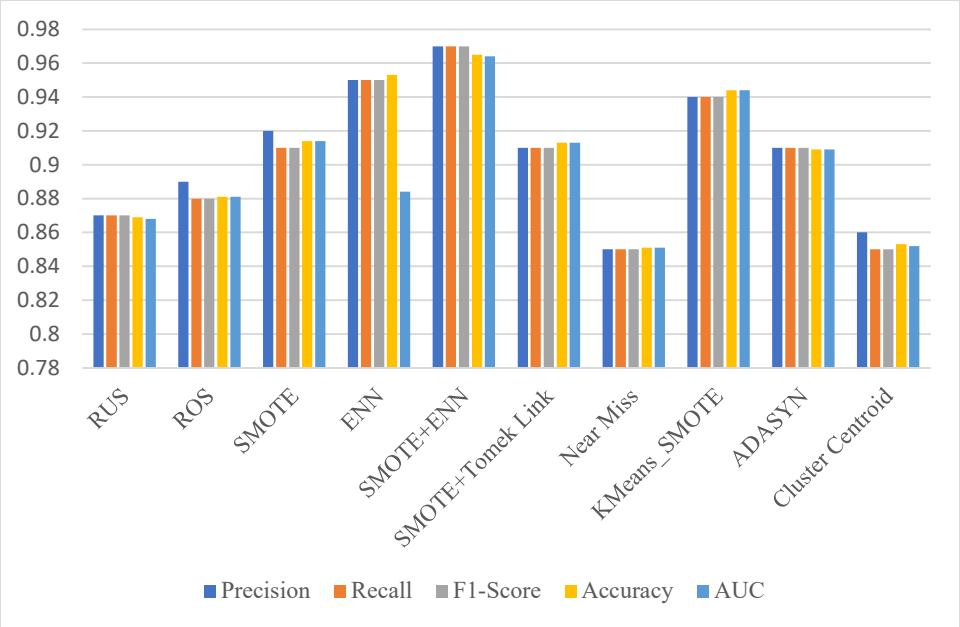

Figure 7: XGBoost Overall Performance

5.8. Experiment 8

It is the third and last experiment for the minority class. In the same way, the second and last experiment using the XGBoost classifier. For all the minority class experiments, the sampling techniques are the same but after sampling the classifier algorithm was different such as this experiment represents the XGBoost algorithm. Table 10 shows the results for this experiment, and the observations made are commented below.

Table 10: Minority Class Performance on Resampled Data (XGBoost)

Methods	Precision	Recall	F1-Score
XGB with RUS	0.84	0.91	0.88
XGB with ROS	0.85	0.93	0.89
XGB with SMOTE	0.88	0.95	0.92
XGB with ENN	0.87	0.79	0.83
XGB with SMOTE+ENN	0.96	0.98	0.97
XGB with SMOTE+Tomek Link	0.89	0.95	0.92
XGB with Near Miss	0.85	0.84	0.85
XGB with KMeans_SMOTE	0.96	0.93	0.94
XGB with ADASYN	0.87	0.96	0.91
XGB with Cluster Centroid	0.82	0.91	0.86

The XGBoost minority class experiment was as a whole better than the DT minority class experiment, but not better than the Random Forest minority class experiment. However, like the other sampling techniques applied it delivers better result than without sampling. Comparing within the sampling techniques, it is not different than the previous two algorithms. Hybrid SMOTE and ENN combined gives the best performance. Like the Random Forest, Near Miss sampling technique had the least accuracy value. When the comparison is between both binary class and minority class for the same XGBoost algorithm, the minority class performance is not better than for the overall class data set. It was the same for DT but not for RF. Figure 8 helps to get the idea by another outlook of calculated values.

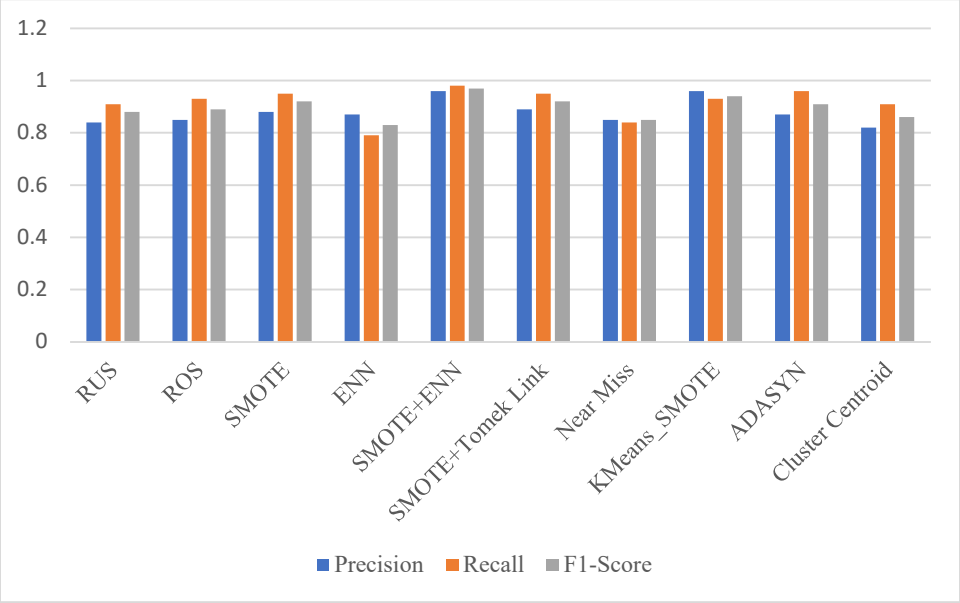

Figure 8: XGBoost Minority Class Performance

6. CONCLUSION

The intention of this investigation was to find out the prediction performance for banking data. Different resampling techniques have been applied to an unbalanced banking data set. The result shows that if the overall prediction performance was considered using different sampling techniques, it most likely will give better performance over an original, balanced data set without sampling though the difference is not too high. But if the data set is imbalanced and if the objective is to identify the prediction for the minority class, then the outcome is completely different.

Among the classifier algorithms, Random Forest shows better performance than others, though it was not much improved compared to the others. The larger improvements are only achieved after the resampling techniques are applied. Ten sampling techniques were used but for all classifier and for both experiment of all classes versus minority class SMOTE with combination of ENN techniques was the best algorithm. There were not only one specific sampling techniques which had the poorest result for all experiments.

For future work, it is better to do experiments with a variety of data sets of the bank and financial sector. The experiment was just done on a single set of data. Thus, lots of experiments for financial data can provide more accurate results. Also, experimenting with more data sets and also with more classifiers will enhance this investigation.

7. REFERENCES

[1] J. Han, M. Kamber. “Data mining concept and technology.” Publishing House of Mechanism Industry: 70-72, 2001.

[2] J. R. Quinlan, “Constructing decision tree.” C4 5, 17-26, 1993.

[3] M. F. Kabir, S. Aziz, S. Ahmmed, and C. M. Rahman. “Information theoretic SOP expression minimization technique.” Computer and information technology, 2007. iccit 2007. 10th international conference on. IEEE, 2007.

[4] J. Vanerio, and P. Casas. “Ensemble-learning approaches for network security and anomaly detection.” Proceedings of theWorkshop on Big Data Analytics and Machine Learning for Data Communication Networks. ACM, 2017.

[5] T. Chen, and C. Guestrin. “Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.

[6] G. Batista, R. C. Prati, and M. C. Monard. “A study of the behavior of several methods for balancing machine learning training data” ACM SIGKDD explorations newsletter 6.1 (2004) : 20 – 29

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. “SMOTE: synthetic minority over-sampling technique.” Journal of artificial intelligence research, 16: 321-357, 2002.

[8] T. Fawcett, “An introduction to ROC analysis.” Pattern recognition letters 27.8 (2006): 861-874.

[9] A. Çaliş, A. Boyaci, K Baynal. “Data mining application in banking sector with clustering and classification methods”, Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management, Dubai, United Arab Emirates (UAE), March 3 – 5, 2015, (978-1-4799-6065-1/15©2015 IEEE)

[10] P. S. Patil, N. V. Dharwadkar, “Analysis of Banking Data Using Machine Learning”, International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017), (978-1-5090-3243-3/17©2017 IEEE)

[11] B. Valarmathi, T. Chellatamilan, H. Mittal, Jagrit, and Shubham. “Classification of Imbalanced Banking Dataset using Dimensionality Reduction” , Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019), IEEE Xplore Part Number: CFP19K34-ART; ISBN:978-1-5386-8113-8

[12] M. F. Kabir, S. A. Ludwig, “Classification of Breast Cancer Risk Factors Using Several Resampling Approaches”, 2018 17th IEEE International Conference on Machine Learning and Applications, (978-1-5386-6805-4/18/ ©2018 IEEE)

[13] “UCI Machine Learning Repository: Bank Marketing Data Set”, https://archive.ics.uci.edu/ml/datasets/Bank+Marketing, last retrieved: September 2020

[14] “Binary classification - Wikipedia”, https://en.wikipedia.org/wiki/Binary_classification, last retrieved: September 2020

[15] “Decision tree - Wikipedia”, https://en.wikipedia.org/wiki/Decision_tree, last retrieved: September 2020

[16] “Random forest - Wikipedia”, https://en.wikipedia.org/wiki/Random_forest, last retrieved: September 2020

[17] “Gradient boosting - Wikipedia”, https://en.wikipedia.org/wiki/Gradient_boosting, last retrieved: September 2020

[18] “The Professionals Point: Advantages of XGBoost Algorithm in Machine Learning”, http://theprofessionalspoint.blogspot.com/2019/03/advantages-of-xgboost-algorithm-in.html, last retrieved: September 2020

[19] J. Mathew, C. K. Pang, M. Luo, and W. H. Leong “Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines.” IEEE transactions on neural networks and learning systems, 2017

[20] “Random Oversampling and Undersampling for Imbalanced Classification”, https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/, last retrieved: September 2020

[21] “Application of Synthetic Minority Over-sampling Technique (SMOTe) for Imbalanced Datasets | by Navoneel Chakrabarty | Towards AI—Multidisciplinary Science Journal | Medium”, https://medium.com/towards-artificial-intelligence/application-of-synthetic-minority-over-sampling-technique-smote-for-imbalanced-data-sets-509ab55cfdaf, last retrieved: September 2020

[22] B. Tang, and H. He, "ENN:Extended Nearest Neighbor Method for Pattern Recognition", IEEE Computational Intelligence Magazine, vol.10, no.3, pp.52--60, Aug,2015

[23] “Oversampling and undersampling in data analysis - Wikipedia”, https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis, last retrieved: September 2020

[24] “Using Under-Sampling Techniques for Extremely Imbalanced Data | by Dr. Dataman | Towards Data Science”, https://towardsdatascience.com/sampling-techniques-for-extremely-imbalanced-data-part-i-under-sampling-a8dbc3d8d6d8, last retrieved: September 2020

[25] “k-means clustering - Wikipedia”, https://en.wikipedia.org/wiki/K-means_clustering, last retrieved: September 2020

[26] “Oversampling with SMOTE and ADASYN | Kaggle”, https://www.kaggle.com/residentmario/oversampling-with-smote-and-adasyn, last retrieved: September 2020

[27] “SMOTE and ADASYN (Handling Imbalanced Data Set) | by Indresh Bhattacharyya | Coinmonks | Medium”, https://medium.com/coinmonks/smote-and-adasyn-handling-imbalanced-data-set-34f5223e167, last retrieved: September 2020

[28] “Interpret all statistics and graphs for Cluster K-Means - Minitab”, https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs, last retrieved: September 2020

[29] “Accuracy (error rate) Definition | DeepAI”, https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate, last retrieved: September 2020