A Comparison of Classical Statistical & Machine Learning Techniques in Binary Classification

KVU Perera[1#] and SD Viswakula[1]

[1] *Department of Statistics, University of Colombo, Sri Lanka*
[#] vijiniup15@gmail.com

Abstract *— Predicting a precise response for previously unseen input variables is a vital and challenging task, as precise predictions can minimize the risks related to different domains by making correct decisions. The main objective of this study was to compare the performance of several classical statistical and machine learning techniques by considering the prediction task as a binary classification. The classification techniques; Logistic Regression (LR) and Linear Discriminant Analysis (LDA) were considered under classical statistical techniques while Random Forest (RF), Naïve Bayes (NB), Boosting (BT) and Bagging (BA) were considered under machine learning techniques. The performance of those techniques were compared under the two different aspects by using five real datasets. In one aspect, class imbalance was artificially introduced to the datasets by resampling. In the other aspect sampling approaches such as undersampling, oversampling and hybrid approach (mix of both undersampling and oversampling) were considered, to overcome class imbalance in the training set. Several evaluation methods such as accuracy, precision, F-measure, G-mean and Receiver Operating Characteristics Area Under Curve (ROC AUC) were considered to evaluate the performance of the classification techniques. The results indicated that the performance of Random Forest and boosting are better than the performance of other techniques in both resampling and overcoming class imbalance aspects. In many cases when the training set was balanced, not only the machine learning techniques but also the statistical techniques had better performance.*

Keywords— **Statistics, Machine Learning, Classification, Resampling, Class Imbalance**

I. INTRODUCTION

Predicting a response for future input variables (prediction) is one of the intentions in analyzing data. In order to approach this intention, the classical statistical techniques as well as the machine learning techniques can be used. When considering the areas such as health, education, finance, etc., predictions play a major role in decision making. Precise predictions can avoid uncertainties in decision making and then the risks can be minimized. Therefore, the most vital and challenging task is to predict responses correctly as much as possible for input variables which have never seen before with the aid of existing data. Obviously, this task highly depends on the technique (either classical statistical or machine learning) which is used for prediction. So, this study concerns in comparing several classical statistical and machine learning techniques by considering binary classification as the prediction task. Simply, binary classification is the task in allocating the observations of a given dataset into two categories according to a particular classification rule. The classification techniques Logistic Regression and Linear Discriminant Analysis are considered under classical statistical techniques while Random Forest, Naïve Bayes, Bagging and Boosting are considered under machine learning techniques in this study (Hastie et al., 2008).

In general, both classical statistical and machine learning techniques have the same concern, which is basically learning from data, but in a different manner, where classical statistical techniques check for assumptions while machine learning techniques are used as black boxes (Breiman, 2001). According to the literature, the amount of data involved and the contribution from humans to build models differ for these two techniques.

When considering a particular scenario, different classification techniques have different performance in classifying the response. Sometimes classical statistical techniques might classify the response precisely by providing a high predictive accuracy than the machine learning techniques or in converse manner. In some other cases, both techniques might provide the same predictive accuracy. Therefore it is worth to identify when to use what.

The main objective of this study is to compare the performance of several classification methods (classical statistical and machine learning techniques) by considering following aspects.

- Data resampling aspect
- Overcoming class imbalance aspect

The other objective is to identify whether there is an effect from the above mentioned two aspects for the performance of classification methods.

II. METHODOLOGY & EXPERIMENTAL DESIGN

A. Experimental Aspects

This study concerned on the two experimental aspects which were depended in resampling and overcoming class imbalance phenomena. The software R was used for this study. The theories of used methods can be explained as below.

1) Resampling: The method in drawing repeated samples from the original data called as resampling. In resampling, the randomized cases are selected with replacement from original data in such a way that the each sample size is equal to the number of observations in original dataset. Due to replacement of observations, the drawn samples by using resampling methods consist of repetitive cases.

2) Overcoming Class Imbalance Phenomena: The class imbalance phenomena can be defined as, "A dataset is imbalanced if the classification categories are not approximately equally represented" (Chawla, 2005, p. 853). In an imbalance dataset, the class having a low number of occurrences is named as the minority class while the class having comparatively more number of occurrences is named as the majority class.

According to Brownlee (2015), the problem with the imbalance phenomena is that most of the machine learning algorithms may favor majority class instances. That means there is a high chance in classifying minority class instances as majority class instances (misclassifying the minority class instances). Hence achieve a high accuracy. Among the approaches available to deal with class imbalance, the sampling methods such as undersampling, oversampling and hybrid (mix of both undersampling and oversampling) approaches were considered in this study.

- Undersampling - This approach works with the majority class. In order to balance the distribution of classes in a dataset, this method removes a number of observations randomly from the majority class. Discarding valuable information is the major drawback of this method.
- Oversampling - This method works with the minority class. In order to balance the distribution of classes in a dataset, this method replicates the observations randomly from minority class. Over fitting to data is the major drawback of this method.
- Hybrid Approach - This method is the combination of undersampling and oversampling approaches.

B. *Cross Validation Technique*

Cross validation is a model evaluation method in determining how the outcome of the considered model will generalize to a new dataset. The basic idea of cross validation is to separate the dataset into two non-overlapping sets where one is used for training the model and the other set is used for testing the performance of learned model. The hold-out and the k-fold cross validation methods are the two well-known cross validation techniques (Schneider, 1997). According to the literature, many comparative studies were carried out by using the hold out method. Therefore by considering that aspect as well as the advantages, the k-fold (k=10) cross validation technique was used in this study.

1) k-fold Cross Validation Technique: In this method, the dataset is randomly divided into *k* mutually exclusive, approximately equal size of subsets (folds). One fold is reserved for testing while all the other *(k-1)* folds (all together) are used for training. Then this process is repeated for *k* times, by using each fold exactly once as the testing set. Finally, the average of the evaluation measures obtained from each time is used to assess the performance of model. The mostly used value for k is 10. Since there is a guarantee of using each and every data point to train the model, the accurate performance information can be obtained. Also the variance of resulting measures is reduced when the *k* is increased. The high computational time due to repetition process can be considered as a disadvantage of this method.

C. *Evaluation Measures*

Since this study was a comparison between classical statistical and machine learning techniques in binary classification task, the best technique was determined by comparing the performance between the techniques. The performance of evolved models and methods was evaluated by using certain evaluation criterions such as the accuracy and the predictability of the predicted model. Since the class imbalance phenomena was existed in some of the datasets used for this study, not only the individual evaluation measures but also the combined and the graphical evaluation measures are considered.

1) Individual Evaluation Measures: Precision, accuracy, sensitivity and specificity are the individual evaluation measures which can be obtained from the confusion matrix (Kohavi & Provost, 1998). A two-by-two contingency table which is called as confusion matrix can be shown as in Table 1.

Table 1. Confusion Matrix

		Predicted Class	
		Positive (P)	**Negative (N)**
True Class	**Positive (P)**	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$Precision = \frac{TP}{TP+FP}$ (1). $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$ (2).

$Sensitivity = \frac{TP}{TP+FN}$ (3). $Specificity = \frac{TN}{FP+TN}$ (4).

2) Combined Evaluation Measures: The F-measure and the G-mean are the combined evaluation measures, considered in this study.

$$F - measure = 2\times\frac{(Precision\times Sensitivity)}{(Precision+Sensitivity)}$$
(5).

$$G - mean = \sqrt{(Sensitivity\times Specificity)}$$
(6).

3) Graphical Evaluation Measures: In this study, the evaluation measure, Area Under Curve (AUC) in Receiver Operating Characteristics (ROC) graph was considered. A ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance. Each point on the ROC plot represents a sensitivity – false positive rate (1 - specificity) pair corresponding to a decision threshold. A test with perfect discrimination (no overlap in the two classes) has a ROC plot that passes through the upper left corner (100% sensitivity and 100% specificity). Therefore, the overall accuracy of a particular test is higher when the ROC plot is closer to upper left corner. In most of the cases, ROC AUC values lie between the 0.5 and 1 where 0.5 represents a worthless test and 1 represents a perfect test.

D. Datasets

The five different datasets retrieved from well-known database called UCI Machine Learning Repository were used for this study. They contain real and generated data with respect to different domains. The characteristics of a dataset such as number of observations, number of variables and types of variables (continuous, categorical and mix of both) are varied for considered datasets. Since the prediction task is binary classification, the response variables in all five datasets contain only two categories. The relevant links of the datasets are provided in following Table 2.

Table 2. Downloaded links of the datasets

Dataset	Downloaded Link
Gamma Telescope	<https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope>
Wilt	<https://archive.ics.uci.edu/ml/datasets/Wilt>
Bank Marketing	<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>
German Credit	<https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)>
Tic-Tac-Toe	<https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame>

E. Evaluation Procedures

1) Procedure in Resampling Aspect: In this aspect classification techniques were compared while introducing the class imbalance phenomena artificially to the dataset. Initially, the dataset was separated into two groups by considering the two classes in the response variable. Then the observations were selected randomly with replacement from the two groups to create a new dataset which the size of it was equal to the numbers of observations in the original dataset. When considering the response variable in a dataset with two classes, one of them could be considered as 'success'(positive) class and the other one as 'failure' (negative) class according to the problem domain. The number of observations that should select randomly from each group was decided as below.

$$p_i = \frac{n_s}{N}$$
(7).

$$n_f = N - n_s$$
(8).

p_i: probability of success
N: total number of observation in original dataset
n_s: number of observations select from 'success' group
n_f: number of observations select from 'failure' group

Equation 7 was used to determine the number of observations that should be chosen from 'success' group while the equation 8 was used to decide the number of observations that should be taken from 'failure' group. Then a new dataset was formed by combining the randomly chosen two sets of observations which contained the same number of variables and same number of observations as in original dataset. Since the class distribution was not random in newly created dataset, random class distribution was introduced by shuffling the data row wise. When shuffling the data, the order of observations (entire row) was changed but the values relevant to an observation (values within a row) were not affected.

Then the 10-fold cross validation method was used. The above mentioned classification techniques were trained by using training set and the class outcomes were predicted by using testing set separately for each technique. When repetition was done for 10 times, each time all the classification techniques were trained by using same training set and tested on same testing set. Then by comparing actual class outcomes and predicted class outcomes, the evaluation measures were obtained for each classification technique separately. Finally, the best classification algorithm was determined by analyzing the achieved results for evaluation measures.

This whole process was repeated for 9 times by varying the probability of success (p_i) from 0.1 to 0.9.

2) Procedure in Overcoming Class Imbalance Aspect: In this aspect, the classification techniques were compared under different sampling approaches to overcome the class imbalance phenomena. The main focus was to balance the training set. Since, the class distribution was not random in most of the original datasets, the randomness of class distribution was introduced by shuffling the dataset row wise. As in previous aspect, in here also the order of observations (entire row) was changed while the values relevant to an observation (values within a row) remain unchanged.

Then as same as in previous aspect, the 10-fold cross validation method was applied. The approaches undersampling, oversampling and hybrid (mix of undersampling and oversampling) were applied to balance the training set derived from shuffled dataset. The same modified dataset was used under each technique separately. Finally the best classification technique was determined according to three sampling approaches. The only difference in here was that the use of balanced training set for the training process of classification techniques.

F. *Special Considerations*

All the classification techniques were tested by considering the five original datasets separately. LDA is one of the statistical classification technique used in this study. The theory of LDA suggests that it is suitable when the predictor variables are in interval scale. Therefore LDA was applied only for the datasets which contained continuous variables except the response variable (Wilt and Gamma Telescope datasets). Since bagging and boosting are ensemble based methods (not stand alone methods), the classification tree was used as the single classifier. When considering on machine learning techniques, some of them use several parameters. The default values suggested by the functions in R is used for those techniques. Also feature selection and parameter tuning are important when training a machine learning algorithm. Even in logistic regression, after fitting the model the best model can be selected using stepwise methods. These scenarios were ignored in order to treat both equally for classical statistical and machine learning techniques. That means when training the considered techniques, all the predictor variables were used as well as no parameter tuning was done to the machine learning techniques.

III. RESULTS

A. *Resampling Aspect*

The results obtained for resampling aspect are summarized in Table 3.

Table 3. Classification techniques which have highest and lowest performance in Resampling Aspect

	Precision	
Dataset	**Highest**	**Lowest**
Gamma Tel.	RF	LR
Wilt	RF & Boosting	NB, *LDA (0.1)*
Bank Mark.	RF	NB, *Bagging (0.1)* *LR (0.2)*
German Cr.	Boosting & *RF* ↑ *(0.1,0.2,0.3)*	LR
Tic-Tac-Toe	Boosting *LR* ↑*(0.1)*	NB
	Accuracy	
Dataset	**Highest**	**Lowest**
Gamma Tel.	RF	LR
Wilt	RF & Boosting	NB
Bank Mark.	RF	NB
German Cr.	Boosting & *RF*↓*(0.1, 0.2)*	LR
Tic-Tac-Toe	Boosting *LR* ↑*(0.1)*	NB
	F-measure	
Dataset	**Highest**	**Lowest**
Gamma Tel.	RF	LR
Wilt	RF & Boosting	NB
Bank Mark.	RF	NB
German Cr.	Boosting & *RF* ↓ *(0.1, 0.2, 0.3)*	LR
Tic-Tac-Toe	Boosting, *LR* ↑*(0.9)*	NB
	G-mean	
Dataset	**Highest**	**Lowest**
Gamma Tel.	RF	LR, *LDA (0.1)*
Wilt	RF & Boosting	NB, *LDA (0.1)*
Bank Mark.	RF	NB, *Bagging (0.1),* *LR (0.2)*
German Cr.	Boosting & *RF* ↓ *(0.1, 0.2, 0,3, 0.6, 0.7, 0.8, 0.9)*	LR
Tic-Tac-Toe	Boosting *LR* ↑*(0.1, 0.9)*	NB
	ROC AUC	
Dataset	**Highest**	**Lowest**
Gamma Tel.	RF	LR
Wilt	RF &Boosting	NB, *LDA (0.1)*
Bank Mark.	RF	NB, *Bagging (0.1),* *LR (0.2)*
German Cr.	Boosting & *RF* ↓ *(0.1, 0.2, 0,3, 0.6, 0.7, 0.8, 0.9)*	LR
Tic-Tac-Toe	Boosting *LR* ↑*(0.1, 0.9)*	NB

Up arrow (↑) - represents higher values
Down arrow (↓) - represents lower values

In Gamma Telescope and Bank Marketing datasets, the technique RF has the highest performance for all probabilities of success. The techniques, RF and boosting have almost the same highest performance for all success probabilities in Wilt dataset. In German Credit dataset, both RF and boosting have almost the same highest performance for all success probabilities except the probabilities mentioned in brackets and for those probabilities RF deviates from boosting as shown by up-arrow and down-arrow. In Tic-Tac-Toe dataset, LR has the highest performance for the probabilities mentioned in brackets and for all the other probabilities boosting has the highest performance.

According to the lowest measures, the techniques which are in italic form has the lowest measures at relevant probabilities mentioned in brackets and for all the other probabilities, mainly mentioned technique (which are in non-italic form) has the lowest measures in each dataset.

B. Overcoming Class Imbalance Aspect

The results obtained for overcoming class imbalance aspect are summarized in Table 4.

Table 4. Classification techniques which have highest & lowest performance in Overcoming Class Imbalance Aspect

Precision								
Dataset	**Highest**				**Lowest**			
	R	**U**	**O**	**H**	**R**	**U**	**O**	**H**
Gamma Tel.	RF	RF	BT	RF	LR			
Wilt	BT	LDA	LDA	LDA	LDA	NB	NB	NB
Bank Mark.	NB	RF	BT	BT	LR	NB	RF	NB
German Cre.	BT	RF	BA	NB	LR			
Tic-Tac-Toe	BT	RF BT	BT	BT	LR			

Accuracy								
Dataset	**Highest**				**Lowest**			
	R	**U**	**O**	**H**	**R**	**U**	**O**	**H**
Gamma Tel.	RF				LR			
Wilt	BT	BT	RF BT	RF BT	NB			
Bank Mark.	RF	BT	RF	RF	NB	BA	BA	BA
German Cre.	RF	NB	BT	NB	LR			
Tic-Tac-Toe	BT				LR			

F-measure								
Dataset	**Highest**				**Lowest**			
	R	**U**	**O**	**H**	**R**	**U**	**O**	**H**
Gamma Tel.	RF				LR			
Wilt	BA	BT	RF BT	RF BT	NB			
Bank Mark.	RF BT	BT	RF	RF	NB	BA	BA	BA
German Cre.	RF	NB	RF BT	BT NB	LR			
Tic-Tac-Toe	BT				LR			

G-mean								
Dataset	**Highest**				**Lowest**			
	R	**U**	**O**	**H**	**R**	**U**	**O**	**H**
Gamma Tel.	RF				LR			
Wilt	BT	BT	LR BA	LR BA	LDA	NB	NB	NB
Bank Mark.	NB	RF	BT	BT ↑	LR	NB	RF	NB
German Cre.	BT	RF	NB BA	NB	LR			
Tic-Tac-Toe	BT				LR			

ROC AUC								
Dataset	**Highest**				**Lowest**			
	R	**U**	**O**	**H**	**R**	**U**	**O**	**H**
Gamma Tel.	RF				LR			
Wilt	BT	BT	LR BA	LR BA	LDA	NB	NB	NB
Bank Mark.	RF	RF	BT	BT	LR	NB	NB	NB
German Cre.	BT	RF NB	NB	NB	LR			
Tic-Tac-Toe	BT				LR			

The used symbols in these tables are as follow.
R – Before applying sampling approaches (Real case)
U – After applying undersampling approach
O – After applying oversampling approach
H – After applying hybrid approach

According to the Table 4, the techniques RF or boosting have highest performances for the most of the cases in overcoming class imbalance aspect.

IV. DISCUSSION & CONCLUSIONS

When considering the resampling aspect, resampling was done to the whole dataset by changing the probability of success from 0.1 to 0.9. Though the dataset was balanced when success probability was equal to 0.5, there was no guarantee about the balance of training set. Here, for Gamma Telescope and Bank Marketing datasets the technique RF had the highest performance for all success probabilities. When considering all success probabilities, two techniques (RF and boosting) had the same highest performance for the Wilt dataset. In some datasets (German Credit and Tic-Tac-Toe), there was no one such method which could be considered as the best method for all success probabilities. The obtained results suggested that either RF and boosting separately or both of them as the best methods for all success probabilities.

Also it was noticed that in resampling aspect, the G-mean and ROC AUC values, were approximately similar to each other in all five scenarios respectively. According to those two evaluation measures, almost all classification techniques had higher performance at middle probabilities (eg: 0.5) than the extreme probabilities (eg: 0.1 and 0.9).

In overcoming class imbalance aspect, the main focus was balancing only the training set not the whole dataset. Not like in previous aspect, there was an assurance that the classification techniques were trained by using a balanced training set. The sampling approaches such as undersampling, oversampling and hybrid approach (mix of both undersampling and oversampling) were used to balance the training set. The performance of classification techniques were analyzed individually before and after applying the sampling approaches to make the training set balanced under each dataset separately. The techniques RF and boosting had the highest performance for many cases in this aspect. Since the sampling approach which maximize the performance of classification techniques was depended on domain, the best sampling approach couldn't specifically mentioned.

Also majority of evaluation measures (precision, G-mean, ROC AUC) suggested that the performance of most classification techniques were higher after applying the sampling approaches. The evaluation measures, accuracy and F-measure had lower values after applying sampling approaches for most of the classification techniques. The performance of positive class may low due to issues such as loss of valuable information and over fitting in sampling approaches. Since the accuracy and F-measure are relied only on positive class, they might provide lower results after applying the sampling approaches.

Though LDA and LR are classical statistical techniques, they were sensible to sampling approaches. However for most of the cases, the performance of LR also had been improved after applying sampling approaches but sometimes it was considerably very low when compared to the performance of other techniques. That was probably due to the characteristic of domains. Not only the machine learning techniques, but also statistical techniques had higher performance by using the balanced training set.

Not only after training classification algorithms by using the balanced training set but also in resampling aspect, RF and boosting performed well for most of the cases. It might happened probably due to the sampling structure in their classification algorithm. When considering the obtained results, it is difficult to make a general conclusion about the best classification algorithm due to the dependency in performance of classification technique on problem domain.

When the classification classes are not equally represented (imbalanced) in dataset approximately, the accuracy of machine learning techniques are biased towards the majority class. That means if higher proportion (about 90% or above) of data belongs to one particular class (called majority class), then classification algorithm tends to predict the majority class often. Hence achieves highest accuracy while the decisions can be misled. Since the class imbalance was considered in both aspects, relying only on accuracy was not appropriate. That was the reason for using several evaluation measures such as precision, F-measure, G-mean and ROC AUC. When considering the precision, it is based on high performance of only one class. According to equation 5 and 6, F-measure is combination of precision and sensitivity while G-mean is combination of sensitivity and specificity. ROC AUC gives the area under the curve which represents sensitivity – (1-specificity) pairs. Though the F-measure was suggested as an alternative for the class imbalance, it was relying only on success (positive) class. The evaluation measures, G-mean and ROC AUC can handle the class imbalance scenario well as they consider the performances of both classes at once. Therefore, use of several evaluation measures was important as some of them can handle the issues like class imbalance and can provide much confident conclusions.

REFERENCES

Aha D (1991). *UCI Machine Learning Repository: Tic-Tac-Toe Endgame Data Set*. [online] Archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame.

Bock R and Savicky P (2007). *UCI Machine Learning Repository: MAGIC Gamma Telescope Data Set*. [online] Archive.ics.uci.edu. Available at:https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope.

Breiman L (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), pp.199-231.

Brownlee J (2015). *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*. [online] Machine Learning Mastery. Available at: http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learningdataset/.

Chawla NV (2005). Data Mining for Imbalanced Datasets: An Overview. In: O. Maimon and L. Rokach, ed., *Data Mining and Knowledge Discovery Handbook*, 1st ed. Springer US, pp.853-867.

Hastie T, Tibshirani R and Friedman J (2008). *The Elements of Statistical Learning*. 2nd ed. California: Springer Series in Statistics.

Hofmann H (1994). *UCI Machine Learning Repository: Statlog (German Credit Data) Data Set*. [online] Archive.ics.uci.edu.

Available at: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).

Johnson B, Tateishi R and Hoan N (2013). A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International Journal of Remote Sensing*, [online] 34(20), pp.6969-6982. Available at: https://archive.ics.uci.edu/ml/datasets/Wilt.

Kohavi R and Provost F (1998). Glossary of Terms. *Machine Learning*, 30, pp.271-274.

Moro S, Cortez P and Rita P (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, [online] 62, pp.22-31. Availableat: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Schneider J (1997). *A Locally Weighted Learning Tutorial using Vizier 1.0*. [online] Available at: https://www.ri.cmu.edu/pub_files/pub2/schneider_jeff_2000_1/schneider_jeff_2000_1.pdf.

Acknowledgment

We wish to express our sincere gratitude to the colleagues at Department of Statistics, University of Colombo and University of Colombo School of Computing who helped us in numerous ways to make this research a success.

BIOGRAPHY OF AUTHORS

K. V. U. Perera is currently working as a Temporary Instructor at Department of Statistics, Faculty of Science, University of Colombo. (BSc. in Statistics with Computer Science special degree with 2nd class – Upper division)

Dr. S. D. Viswakula is a Senior Lecturer (Grade II) at Department of Statistics, Faculty of Science, University of Colombo. His research interests are Biostatistics, Computational Statistics and Probabilistic Modeling.