

# Evaluation Measures for Models Assessment over Imbalanced Data Sets

Mohamed Bekkar<sup>1</sup>, Dr.Hassiba Khelouane Djemaa<sup>2</sup>, Dr.Taklit Akrouf Alitouche<sup>1</sup>

1. ENSSEA, National School of Statistics and Applied Economics, Algiers, Algeria,

2. EHEC, Ecole des Hautes Etudes Commerciales, Algiers, Algeria,

\* E-mail of the corresponding author: moh.bekkar@gmail.com

## Abstract

Imbalanced data learning is one of the challenging problems in data mining; among this matter, founding the right model assessment measures is almost a primary research issue. Skewed class distribution causes a misreading of common evaluation measures as well it lead a biased classification. This article presents a set of alternative for imbalanced data learning assessment, using a combined measures (G-means, likelihood ratios, Discriminant power, F-Measure Balanced Accuracy, Youden index, Matthews correlation coefficient), and graphical performance assessment (ROC curve, Area Under Curve, Partial AUC, Weighted AUC, Cumulative Gains Curve and lift chart, Area Under Lift AUL), that aim to provide a more credible evaluation. We analyze the applications of these measures in churn prediction models evaluation, a well known application of imbalanced data

**Keywords:** imbalanced data, Model assessment, accuracy , G-means, likelihood ratios, F-Measure, Youden index, Matthews correlation coefficient, ROC, AUC, P-AUC,W-AUC, Lift, AUL

## 1. Introduction:

The problem of mining imbalanced data sets receive much interest in recent years[16] [21] [31][34][45][57]; considered as one of the top 10 challenges for data mining [64], the imbalanced data is encounter in several real world application such as social sciences, credit card fraud detection, customer retention, churn prediction, segmentation. Even, in medical diagnostic and fraud detection the imbalanced data sets are the norms and not exception [65]

A data set is considered imbalanced if one of classes (minority class) contains much smaller number of examples than the remaining classes (majority classes); While the minority class is usually the main interest in applications cases. This will lead in focus of learning algorithm in majority classes. Considering a case of fraud detection where 1% of examples are fraudulent and 99% are not (truthful); a naïve model that will predict all base as not fraudulent will reach an accuracy of 99/100=99%, which is excellent as evaluation measure, but almost useless model in real application. In this context, it appears clearly that overall classification accuracy is not an appropriate assessment measure.

There are a number of problems that arise when mining imbalanced data sets; other than improve approaches or models to handle imbalanced data sets [8], Weiss [66] consider that found the proper evaluation measures for model assessment as one the most complex issue faced on imbalanced data learning context.

We will focus on this paper on model assessment measures with imbalanced data framework. The rest of the paper is organized as following: Section 2 will describe the fundamental evaluation measures based on confusion matrix; the combined evaluation measures are detailed on Section 3; while the section 4 will focus on graphical performance evaluation; at the end, we perform applications cases on section 5 with discussion

## 2. Fundamental evaluation measures:

In machine learning, the classifier is basically evaluated by a confusion matrix. For a binary class problem a matrix is a square of 2×2 as shown in Table 1; column represents the classifier prediction; while the row is the real value of class label. In imbalanced data context, by convention, the observations of minority class are labelled as positive, whilst and the class label of the majority class observations are labelled negative.

Table 1. Confusion matrix for two classes classification

	Predicted Positive	Predicted Negative
Actual positive	TP (number of True Positive)	FN (number of False Negative)
Actual Negative	FP (number of False Positive)	TN (number of True Positive)

The acronym TP, FN, FP, and TN of the confusion matrix cells refers to the following:

TP = true positive, the number of positive cases that are correctly identified as positive,

FN = false negative, the number of positive cases that are misclassified as negative cases,

.FP = false positive, the number of negative cases that are incorrectly identified as positive cases,

TN = true negative, the number of negative cases that are correctly identified as negative cases

Table 2 presents the most well known fundamental evaluation metrics

Table 2. Fundamental evaluation metrics based on confusion matrix analysis

Measure	Formula	interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	
Error rate = 1-Accuracy	$\frac{FP + FN}{TP + TN + FP + FN}$	
Sensitivity (or Recall)	$\frac{TP}{TP + FN}$	Accuracy of positive examples
Specificity	$\frac{TN}{TN + FP}$	Accuracy of Negative examples
Precision	$\frac{TP}{TP + FP}$	

Accuracy, the most common metric for classifier evaluation, it assesses the overall effectiveness of the algorithm by estimating the probability of the true value of the class label. On the other hand the error rate=1-accuracy is an estimation of misclassification probability according to model prediction.

Intuitively, precision is a measure of correctness (i.e., out of positive labeled examples, how many are really a positive examples), while Sensitivity (or Recall) is a measure of completeness or accuracy of positive examples (i.e., how many examples of the positive class were labeled correctly). These two metrics, share an inverse relationship between each other. However, unlike accuracy and error, precision and recall are not sensitive to changes in data distributions. A perfect model will capture all positive examples (Recall = 1), and score as only the examples that are in fact (Precision = 1), from an analytical point of view it is desirable to increase recall without sacrificing accuracy.

Specificity is the conditional probability of true negatives given secondary class, it approximates the probability of the negative label being true; in other words, Sensitivity and Specificity assesses the effectiveness of the algorithm on a single class, positive and negative respectively.

Considering the case of imbalanced data assessment; Accuracy places more weight on the common classes than on rare classes, which makes it difficult for a classifier to perform well on the rare classes, it become a misleading indicator; Because of this, additional metrics are developed by combining initially sensitivity and specificity

### 3. Combined evaluation measures:

#### 3.1 G-means:

The geometric mean G-mean was suggested in [35] as the product of the prediction accuracies for both classes, i.e sensitivity: accuracy on the positives examples, and Specificity: accuracy on the negative examples. This metric indicates the balance between classification performances on the majority and minority class. a poor performance in prediction of the positive examples will lead to a low G-mean value, even if the negative examples are correctly classified per the model [31]

Indeed, the G-mean is quite important to measure the avoidance of the overfitting to the negative class and the degree to which the positive class is marginalized.

$$G = \sqrt{\text{Sensitivity} \times \text{Specificity}}$$

The G-mean has been used per several researcher for classifier assessment over imbalanced data set, as the publication of [26] [33] [51] [57] [52]

#### 3.2 The likelihood ratios:

We can distinguish positive likelihood ratio and negative likelihood ratio. The positive likelihood ratio L (noted also  $\rho^+$  or LR(+)) represents the ratio between probability of predict an example as positive when it is truly positive, and the probability of predict example as positive when actually it is not positive.

$$L = \rho^+ = \frac{P(\text{Positive}/\text{Positive})}{P(\text{Positive}/\text{Negative})} = \frac{TP/(TP + FN)}{FP/(FP + TN)} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

While the negative likelihood ratio  $\lambda$  (noted also  $\rho^-$  or LR(-)) is the ratio between the probability of predict an example as negative when it is actually positive, and the probability to predict an example as negative when it is truly negative

$$\lambda = \rho^- = \frac{P(\text{Negative}/\text{Positive})}{P(\text{Negative}/\text{Negative})} = \frac{FN/(TP + FN)}{TN/(FP + TN)} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$

Regarding their interpretation, higher positive likelihood ratio and a lower negative likelihood mean better performance on positive and negative classes respectively. We can refer to the following thresholds table:

Table 3. Thresholds for positive likelihood ratio interpretation

L value	Model contribution
1	Negligible
1-5	Poor
5-10	Fair
> 10	Good

The likelihood ratios are used particularly in medical diagnosis prediction [18] [40], learning from pharmaceutical data sets [2]

### 3.3 Discriminant power:

Discriminant power (DP) is a measure that summarizes sensitivity and specificity, calculated as per formula [23]:

$$DP = \frac{\sqrt{3}}{\pi} (\log X + \log Y)$$

Where : X = sensitivity / (1- sensitivity) , and Y= specificity / (1- specificity)

DP evaluates how well an algorithm distinguishes between positive and negative examples according to the following table:

Table 4. Thresholds for Discriminant power interpretation

DP value	Model contribution
<1	Poor
>1 and <2	Limited
>2 and <3	Fair
> 3	Good

The DP has been used mainly in feature selection over imbalanced data [37]

### 3.4 F-Measure and $\beta$ varied F-Measure:

F-measure is defined as the harmonic mean of precision and recall [27]

$$F = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

This measure which is a harmonic mean of Precision and Recall is null whenever one of the two indices is null, the value of F increases proportionally to the increase of precision and Recall, a high value of F-Measure indicates that the model performs better on the positive class.

On the other hand, the F-Measure is derived from a more general relationship called  $\beta$  varied F-Measure, the formula is written as:

$$F_{\beta} = \frac{(1 + \beta^2) \times \text{Recall} \times \text{Precision}}{\beta^2 \times \text{Recall} + \text{Precision}}$$

Where  $\beta$  is a coefficient to adjust the relative importance of precision versus recall, decreasing  $\beta$  leads a reduction of precision importance; in the general case  $\beta$  is considered equal to 1.

Chawla et al suggest to exploit  $\beta$  varied f-measure instead of basic F-measure in the context of imbalanced data; and in order to adapt the measure to cost-sensitive learning methods, the  $\beta$  value is defined according to formula:

$$\beta = \frac{C(1,0)}{C(0,1)}$$

Where:

C (1.0) is the cost associated to the prediction of a False Negative,

C (0.1) is the cost associated to the prediction of a False positive

The rationale behind  $\beta$  varied F-measure is that the misclassification within minorities class is often more expensive than misclassification of majority examples, consequently improving the Recall will deeply affect the f-measure than the Precision

This indicator is more appropriate to the context of imbalanced data, however, it is difficult to implement outside of costs sensitive learning, where we use generally the basic version of F-measure

F-measure provides more insight into the functionality of a classifier than the accuracy metric; it has been used on the assessment of models with imbalanced data in text classification [36], bioinformatics [48] manufacturing fault detection [34], churn prediction [16] and fraud detection [41]

### 3.5 Balanced Accuracy:

The balanced accuracy is the average of Sensitivity and specificity can be defined also as the average accuracy obtained on either class.

$$\text{Balanced Accuracy} = \frac{1}{2} (\text{sensitivity} + \text{Specificity}) = \frac{TP}{P} + \frac{TN}{N}$$

If the classifier performs equally well on either class, this term reduces to the conventional accuracy detailed in table1. In contrast, if the conventional accuracy is high only because the classifier takes advantage good

prediction on the majority class, then the balanced accuracy will drop [11]. The balanced accuracy has been used in several publications as on statistical patterns of epitasis learning [20], Text mining and video indexing [61]

### 3.6 Youden index :

Youden's index  $\gamma$  [58] evaluates the algorithm's ability to avoid failure; it's derived from sensitivity and specificity and denotes a linear correspondence balanced accuracy:

$$\gamma = \text{sensitivity} - (1 - \text{specificity})$$

$$\gamma = 2 * \text{Balanced Accuracy} - 1$$

as Youden's index is a linear transformation of the mean sensitivity and specificity, its values are difficult to interpret [39], we retain that a higher value of  $\gamma$  indicates better ability to avoid failure.

Youden's index has been conventionally used to evaluate tests diagnostic [9] [1], improve efficiency of Telemedical prevention [45]

### 3.7 Matthews correlation coefficient, MCC :

MCC is a single performance measure less influenced by imbalanced test sets since it considers mutually accuracies and error rates on both classes, and involve all values of confusion matrix; MCC is based on Kh-2 statistics over the confusion matrix

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC ranges from 1 for a perfect prediction to -1 for the worst possible prediction. MCC close to 0 indicate a model that performs randomly [6]. MCC is considered by some authors as the best singular assessment metric [21], especially suitable to the case of imbalanced data learning [55]

## 4. Graphical performance evaluation

### 4.1 ROC curve:

Graphically, we often represent ROC (Receiver Operating Characteristic) as a curve that gives the true positive rate as a function of false positive rate for the same group; Specifically, the ROC approach involves representing the value of the sensitivity as a function of (1-specificity) for all possible threshold values, and join the points with a curve. The more inclined the curve is toward the upper left corner, the better is the classifier's ability to discriminate between positive and negative class

Provost and Fawcett [47] was the leaders in the development of the use of ROC curve, they advice the uses as an alternative to accuracy rate in the case of imbalanced data learning. since, this approach has become widely used, with devoted workshops [28], and application in several research as [24] [56] [28]

#### 4.1.1 Comparison of ROC curves:

Several ROC curves can be represented in the same space to compare the results of different models, in the simplest case, shown below (left), a curve dominates the other and modeling associated with the dominant curve is considered more efficient, we are talking about a simple instead qualitative interpretation of the ROC curve based on the curve shape.

However, in more complex cases, which are more frequent, the two curves intersect, and it is more difficult to identify the top model.

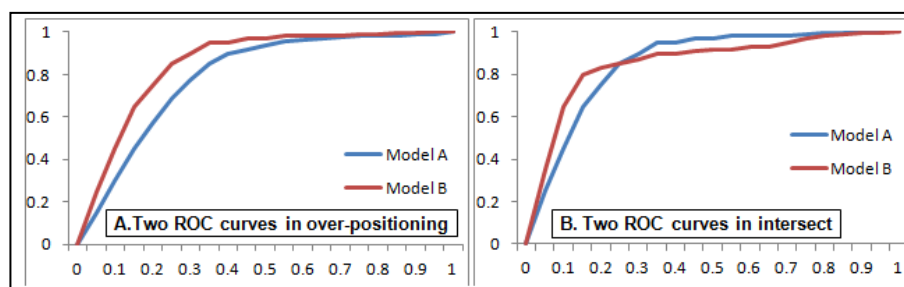


Figure 1. Roc Curves comparison cases

Various ideas was proposed to solve the ROC intersect situation; essentially, the use of decision theory as proposed by [46], and [13] in order to reach an optimal classifier between both who gets the efficient area of each model.

Another innovative approach was proposed by [13] in developing B-ROC as an alternative to the traditional ROC, that generate a new type of curve most suitable to the comparison incase of intersection as shown in figure below

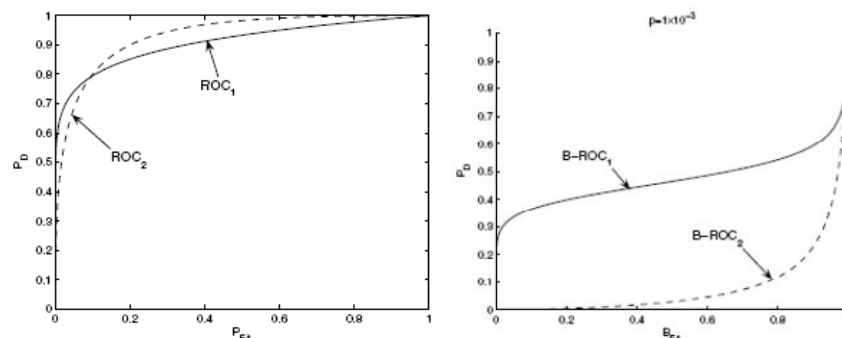


Figure 2. B-ROC curves illustration

The ROC curves are a good way to compare models or sets of models, however, as stated Drummond and Holte [22] they have three major disadvantages for evaluating imbalanced data models:

- Costs and the prior distribution of the classes are not taken into account;
- The decision thresholds are not explicitly represented in the ROC curve;
- The difference between the two models is difficult to quantify.

#### 4.1.2 Area Under Curve, AUC :

The area under the ROC curve (Area under curve, AUC) is a summary indicator of ROC curve performance that can summarize the performance of a classifier into a single metric. Unlike difficulties encountered in the comparison of different ROC curve especially in intersection case, the AUC can sort models by overall performance, as a result, the AUC is more considered in models assessment [7].

The AUC is estimated through various techniques, the most used is the trapezoidal method, which is a geometrical method based on linear interpolation between each point on the ROC curve. Even simpler, some authors [50][4] propose to make the approximation of AUC, in the case of binary learning by Balanced Accuracy

$$\text{Balanced Accuracy} = \frac{1}{2} (\text{sensitivity} + \text{Specificity}) = \frac{TP}{p} + \frac{TN}{N}$$

The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [27].

In practice, the value of AUC varies between 0.5 and 1; Allaire [3] suggest the following scale for the interpretation of AUC value:

Table 5. AUC Value interpretation

AUC Value	Model performance
0.5 - 0.6	Poor
0.6 - 0.7	Fair
0.7 - 0.8	Good
0.8 - 0.9	Very Good
0.9 - 1.0	Excellent

A chance-standardized variant of the AUC is given by Gini coefficient, taking values between 0 (no difference between the score distributions of the two classes) and 1 (complete separation between the two distributions). Gini coefficient is widespread use metric in imbalanced data learning [25], calculated as following :

$$G = 2 \times \text{AUC} - 1$$

Some authors considers that the AUC can give a misleading of model performance, especially in case of imbalanced data learning, as it covers a part of the prediction range uselessness in practical. [10]; Other alternatives were proposed in the literature to achieve a more clear assessments, we can mention

#### 4.1.3 Partial AUC:

Partial AUC (PAUC) originally proposed by McClish [42], is getting increasingly used in recent publications [62]. The logic behind the PAUC as described by Dodd and Pepe[19] is to estimate the AUC on a specific area of the decision threshold; thus the PAUC can compare different models for the same benchmark of decision threshold, the authors illustrate the PAUC through the following graph:

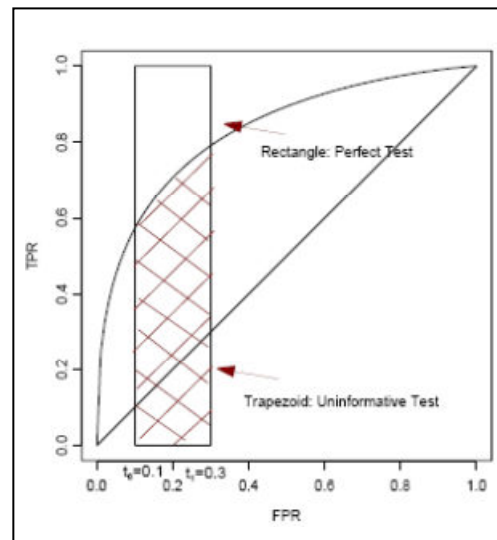


Figure 3. The Partial AUC illustration

#### 4.1.4 Weighted AUC

Weighted AUC is a variant of AUC that fits better the imbalanced data learning case [15]. The rationale behind WAUC is knowing that classifier which perform well in the higher TP region is preferred over ones that does not; so instead of summing up area under curve with equal weights, we want to give more importance to the area near to the top of graph, so we create a skew weight vector by distributing more weights towards the top of the ROC curve as shown in the following figure:

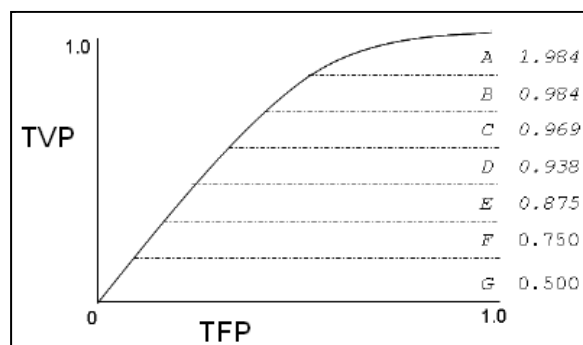


Figure 4. Weighted AUC estimation approach

The WAUC was considered as alternative to basic AUC in the publication of Gohore Bi [29], and Yuanfang and al [59].

#### 4.2 Cumulative Gains Curve and lift chart:

The cumulative gains curve represents the percentage of positive relative to the percentage of targeted population according to score deciles. the lift chart is derived from cumulative gains curve where on each point the lift represents the ratio between the percentage positive to the percentage of targeted population; so it tells how much better a classifier predicts compared to a random selection

within the same graph by setting the % of the target population to 20% , the point B on the lift curve (57% positive) is above the random position (point A with 20% positive) and a lower that the ideal situation (C with 80% positive)



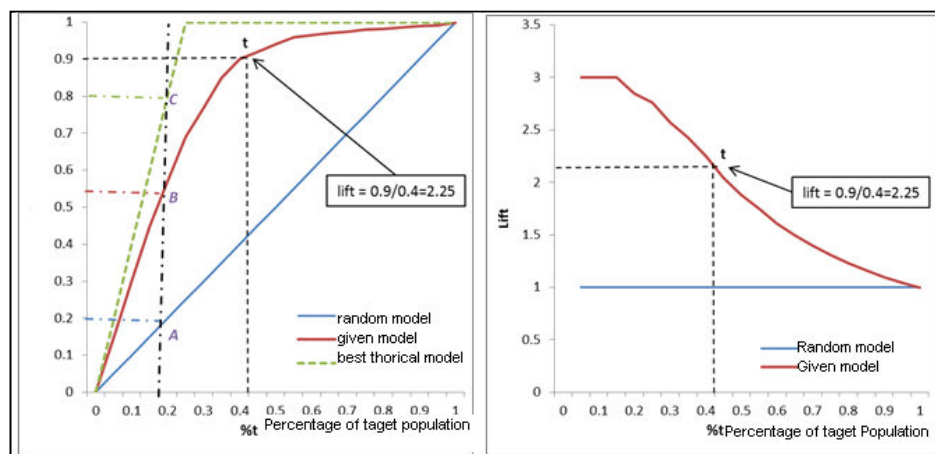


Figure 5. Cumulative Gains Curve and lift Chart

Lift is nearly related to accuracy, and has the advantage of being well used in marketing practice [39]; as on imbalanced data learning assessment [53].

#### 4.2.1 Area Under Lift, AUL:

Similarly to the AUC associated to ROC Curve, we define the AUL (Area under lift); Tuffery [63] demonstrate that the AUL can be estimated through the following equation

$$AUL = \frac{P}{2} + (1 - P) \times AUC$$

Where P = Prior probability of positive observation on the population

We can conclude that the AUC is always greater than AUL, provided that  $AUC > 0.5$ , in other words that ROC curve is above the diagonal. This formula also shows that:

- If  $AUC = 1$  (i.e score separated perfectly) then :  $AUL = p/2 + (1 - p) = 1 - p/2$
- If  $AUC = 0.5$  (i.e Random prediction model) then :  $AUL = p/2 + \frac{1}{2}(1 - P) = 0.5$
- If the probability P is very small, the areas under the two curves are very close

In all cases, to deduce that a model is superior to another, it is equivalent to measuring the area under the curve lift AUL or the area under the ROC curve AUC , ie: if  $AUC_1 > AUC_2$  So  $AUL_1 > AUL_2$

## 5. Application cases and discussion:

### 5.1 Applications Cases area:

One of the most common cases of imbalanced data set learning is the churn prediction. Considered in various industries such as telecom [16] [32] [49] banking and finance [54] [60] retail [43] & insurance[17]. The churn prediction has been also a fruitful field to develop and assess a new approaches for handling imbalanced data [8] [44] [49] or combine the known data mining techniques and methods [32] [30].

Following the same trend, we perform applications cases of churn prediction based data sets issued from two wireless telecom operators. Within this exercise we use the decision trees as prediction method, mainly C5.0 and CHAID, coupled with under-sampling on negative class to adjust imbalanced data; as we combine additional techniques (boosting, cost sensitive learning ) to enhance learning on imbalanced data.

### 5.2 Discussion of 1<sup>st</sup> Data Set results:

The application on first data sets present the following results:

Table 6. Results of churn prediction Models performance on 1<sup>st</sup> data Set

Data Set 1	Model 1A: C5 Basic	Model 1B: C5 cost sensitive	Model 1C: C5 Expert boosting	Model 1D: CHAID Basic	Model 1E: CHAID Cost sensitivity
TP	4698	8568	4602	5749	8090
FP	15980	38585	15684	23180	36668
TN	294590	271985	294886	287390	273902
FN	16503	12633	16599	15452	13111
Accuracy	90.2%	<b>84.6%</b>	90.3%	88.4%	85.0%
Precision	22.7%	18.2%	22.7%	19.9%	18.1%
specificity	94.9%	87.6%	94.9%	92.5%	88.2%
sensitivity, recall	22.2%	40.4%	21.7%	27.1%	38.2%
AUC	0.59	<b>0.64</b>	0.58	0.60	0.63
AUC/accuracy	64.9%	75.7%	64.6%	67.7%	74.3%
lift	3.47	<b>6.32</b>	3.40	4.24	5.97
F-measure	0.224	0.251	0.222	0.229	0.245
MCC	0.172	0.196	0.170	0.170	0.189
Discriminant power	0.397	0.375	0.395	0.366	0.366
Youden's index	0.170	0.280	0.167	0.197	0.264

Based on the traditional accuracy we can state that the best model is Model 1C (C5 with boosting) that reach the highest accuracy at 90.3%, followed per Model 1A with Accuracy ~90.2%, while the weak one is observed on the model 1B (C5 with Cost sensitive learning) with lowest accuracy of 84.6% ; however once we analyze deeply the other evaluation measures (AUC, lift) we observe that model 1B is almost the best as it show the highest AUC~0.64 and best lift 6.32.

The F-measure, MCC, and Youden's Index assigns also the best value to Model 1B , but with slight difference across the other results compared to lift that double from lowest to highest model.

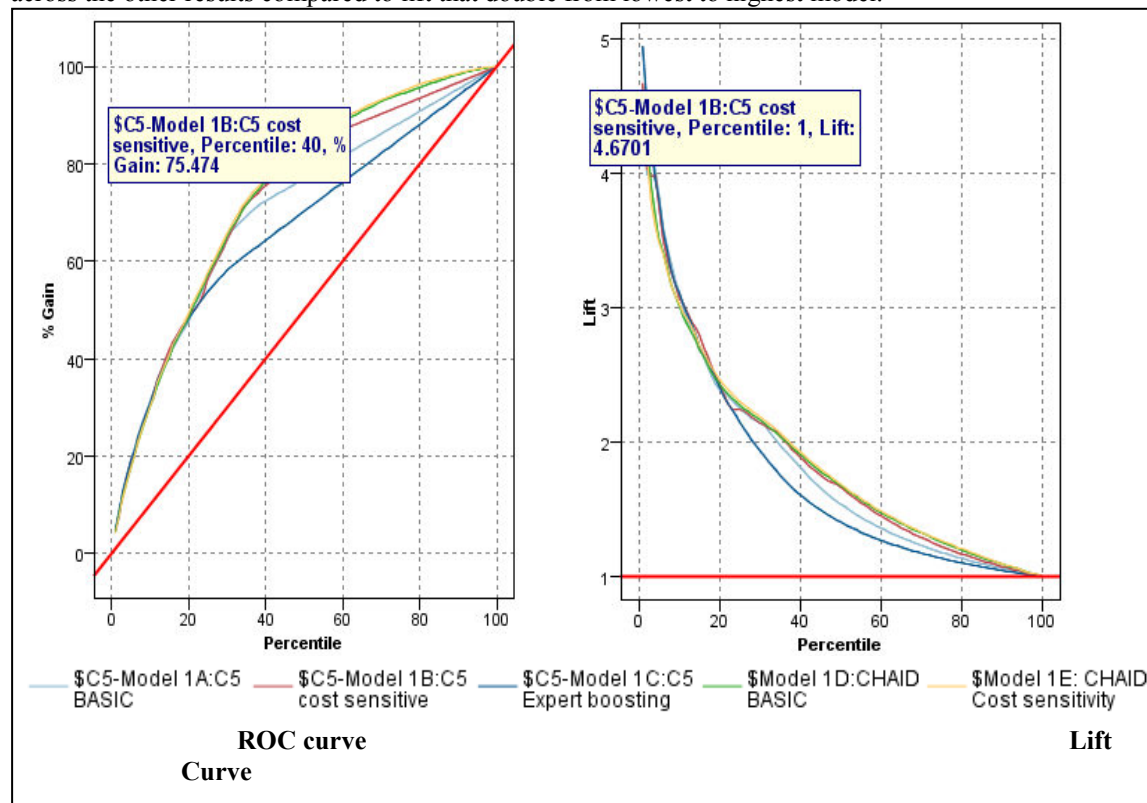


Figure 6. ROC and lift Curves for models generated on 1<sup>st</sup> data Set



The ROC and lift curves confirms the initial results, the models 1B and 1D perform better than 1C specially on the upper limits of percentile (40~100). Following this assessment we can affirm that model 1B is the best in class in this case in spite of the lowest Accuracy that it has

### 5.3 Discussion of 2<sup>nd</sup> Data Set results:

The applications on 2<sup>nd</sup> data set provide the results detailed on the next table

Table 7. Results of churn prediction Models performance on 2<sup>nd</sup> data Set

Data Set 2	Model 2A: C5 Basic	Model 2B: C5 cost sensitive	Model 2C: C5 Expert boosting	Model 2D: CHAID Basic	Model 2E: CHAID Cost sensitivity
TP	12806	13182	12842	12411	12685
FP	26332	30516	21568	14160	16736
TN	299777	295593	304541	311949	309373
FN	5484	5108	5448	5879	5605
accuracy	90.76%	<b>89.66%</b>	92.16%	<b>94.18%</b>	<b>93.51%</b>
Precision	32.72%	30.17%	37.32%	46.71%	43.12%
specificity	91.93%	90.64%	93.39%	95.66%	94.87%
sensitivity, recall	70.02%	72.07%	70.21%	67.86%	69.35%
AUC	0.81	0.81	0.82	0.82	0.82
AUC/accuracy	89.21%	90.74%	88.76%	86.81%	87.81%
lift	13.18	<b>13.57</b>	13.22	<b>12.78</b>	<b>13.06</b>
F-measure	0.446	0.425	0.487	0.553	0.532
MCC	0.438	0.423	0.476	0.534	0.515
Discriminant power	0.785	0.771	0.839	0.919	0.894
Youden's index	0.619	0.627	0.636	0.635	0.642

on the 2nd data set, the Model 2D Basic CHAID reveal the highest KPI in accuracy, Precision and Specificity; AUC are close over different models; while the lift brings a low preference to Model 2B. The Discriminant power and MCC values presents a correlation with Accuracy value, where we observe the lowest value of DP in model 2B, and highest one associated to model 2D which is the same trend of accuracy value.

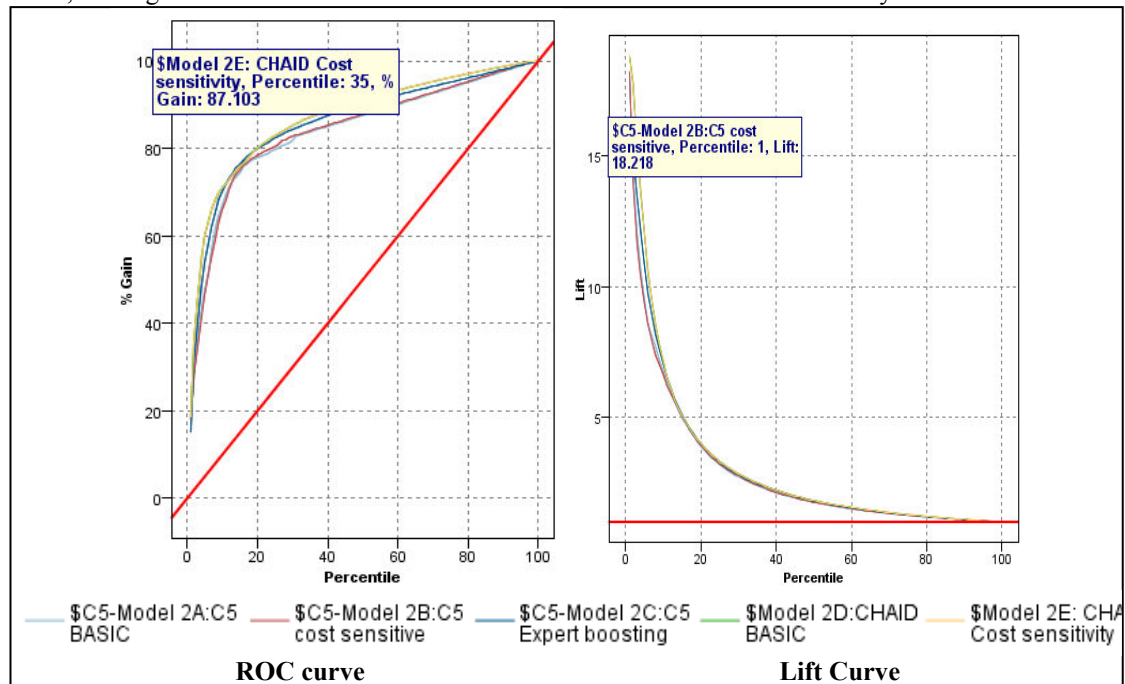


Figure 7. ROC and lift Curves for models generated on 2<sup>nd</sup> data Set.

In this case that appear more complex for decision making than the previous one, the final decision should consider a combination of different measures instead of relying on one measure. Following an approach of models ranking we can observe that Model 2E has one of highest accuracy rate, coupled with excellent AUC, lift and DP value, sustained with highest Youden index. The strong performance of Model 2E is confirmed also through ROC and Lift Curves that show a slight preference to Model 2E despite of the close charts observed on ROC and Lift.

## 5. Conclusion:

The model assessment measure is a key factor in data mining process success; within imbalanced data context, this subject become more complex as the most frequently adopted evaluation metrics share some drawbacks, that don't meet a credible assessment; the research focusing on this concern has been fruitful and several contribution have been achieved. In this paper we have come through the different evaluation measure involved till date on imbalanced data learning assessment, as we have performed an exercise of churn prediction models assessment, that is a common usual case of imbalanced data learning, we demonstrate how the traditional assessment with overall accuracy generate a misreading of model performance.

## References

- [1] Afina, S., Glasa, J. G., Lijmerb., Martin. H., Prinsc, G. J., Bonseld., Patrick, M.M., Bossuyta, (2003) ,“The diagnostic odds ratio: a single indicator of test performance”, *Journal of Clinical Epidemiology* 56 1129–1135
- [2] Aijun An, Nick Cercone, Xiangji Huang, (2001), "A Case Study for Learning from Imbalanced Data Sets", *Proceedings of the 14th Canadian Conference on Artificial Intelligence (CAI-03)*, Ottawa, Canada, June 7-9. Lecture Notes in Computer Science (LNCS) 2056: 1-15. Springer-Verlag Publisher.
- [3] Allaire JF, (2006), « introduction à l'analyse ROC Receiver Operating Characteristic », *Centre de recherche Institut Philippe-Pinel de Montréal, école d'été*.
- [4] Aussem Alexandre, (2010), « Supprot Cours MIF24 : Apprentissage Statistique », Université Claude Bernard, Lyon1.
- [5] Au, W., Chan, C. C., Yao, X., (2003), “A Novel evolutionary data mining algorithm with applications to churn prediction”, *IEEE Transactions on evolutionary computation*, Vol. 7, No. 6.
- [6] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C A F, Nielsen H, (2000), “Assessing the accuracy of prediction algorithms for classification: an overview”, *Bioinformatics vol 16*; P412–424.
- [7] Batista G, Ronaldo C, Monard M C, (2004), “A study of the behavior of several methods for balancing machine learning training data”, *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets*, Volume 6 Issue 1.
- [8] Bekkar Mohamed, Akrouf Alitouche Taklit, (2013), “Imbalanced Data Learning Approaches Review”, *International Journal of Data Mining & Knowledge Management Process (IJDMP)*, 03(04), 15 - 33.
- [9] Biggerstaff, B., (2000), "Comparing diagnostic tests: a simple graphic using likelihood ratios", *Statistics in Medicine* 19(5).
- [10] Briggs W M, Zaretski R, (2008), “The skill plot: a graphical technique for evaluating continuous diagnostic tests”, *Biometrics*, 63, 250-261.
- [11] Broderseny Kay, H., Ong Cheng, S., Stephany K, E., Buhmann J, M., (2010), “The balanced accuracy and its posterior distribution”, *International Conference on Pattern Recognition*, IEEE computer society.
- [12] Cardenas A, Baras J S, Seamon K, (2006), “A framework for the evaluation of intrusion detection systems”, *In Proceedings of the IEEE Symposium on Security and Privacy*.
- [13] Cardenas A, Baras J S, (2006), “B-ROC Curves for the Assessment of Classifiers over Imbalanced Data Sets”, *21st National Conference on Artificial Intelligence*, Boston, Massachusetts.
- [14] Chawla Nitesh V, Cieslak David A, Hall Lawrence O, Joshi Ajay, (2008), “Automatically countering imbalance and its empirical relationship to cost”, *Springer Science & Business Media*, LLC.
- [15] Cheng G, Poon W J, (2008), “A New Evaluation Measure for Imbalanced Datasets”, *conference 7th Australasian Data Mining Conference (AusDM)*, Glenelg, Australia.
- [16] Clement Kirui, Li Hong, Edgar Kirui, (2013), “Handling Class Imbalance in Mobile Telecoms Customer Churn Prediction”, *International Journal of Computer Applications* 72(23):7-13.
- [17] Danso Samuel Odei, (2006), “An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain”, Msc thesis, Bournemouth University.
- [18] Delacoura, H., François, N., Servonneta, A., Gentileb, A., Rochea, B., (2009), « Les rapports de vraisemblance : un outil de choix pour l'interprétation des tests biologiques », *Immuno-analyse & Biologie Spécialisée*, Vol 24, Issue 2.

- [19] Dodd, L. E., Pepe, M. S., (2003), "Partial AUC Estimation and Regression", *UW Biostatistics Working Paper Series*, National Institute of Health University of Washington; Paper 181.
- [20] Digna, R., Velez, Bill, C., White, A. A., Motsinger, W. S., Bush, M. D., Ritchie, S. M., Williams, J. H., (2007), "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction", *Genetic Epidemiology*; Volume 31, Issue 4, pages 306–315.
- [21] Ding, Z., (2011), "Diversified Ensemble Classifier for Highly imbalanced Data Learning and their application in Bioinformatics", Ph. D thesis, College of Arts and science, Department of Computer Science, Georgia State University
- [22] Drummond Chris, Holte Robert C, (2000), "Explicitly representing expected cost : an alternative to ROC representation", *In KDD*, pages 198–207.
- [23] Duda Richard O, Hart Peter E, Stork David G, (2000), "Pattern Classification", 2nd Edition, Wiley-Interscience.
- [24] Elazmeh W, Japkowicz N, Matwin S, (2006), "Evaluating misclassifications in imbalanced data", *17th European conference on machine learning*, Germany, ECML.
- [25] Engler R, Guisan A, Rechsteiner L, (2004), "An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data", *Journal of Applied Ecology* 41, 263–274.
- [26] Ertekin S., Huang J., Bottou L., Giles C L., (2007), "Learning on the border: active learning in imbalanced data classification", *Proceedings of the 6th ACM Conference*, NY, USA.
- [27] Fawcett T, (2006), "An introduction to ROC analysis", *Pattern Recognition Letter* 27(8), 861–874.
- [28] Fem C, Flach P, Orallo J, Lachice N, (2004), "First Workshop on ROC Analysis in AI", *The European Conference on Artificial Intelligence*, ECAI' 2004.
- [29] Gohore Bi Goue Denis, (2010), « Évaluation et contrôle de l'irrégularité de la prise médicamenteuse : Proposition et développement de stratégies rationnelles fondées sur une démarche de modélisations pharmacocinétiques et pharmacodynamiques », Thèse de Philosophie Doctor (Ph.D.) en sciences pharmaceutiques option technologie pharmaceutique, Université de Montréal.
- [30] Hadden John, (2008), "A Customer Profiling Methodology for Churn Prediction", *PhD thesis*, School of Applied Sciences, Cranfield University.
- [31] Hido Shohei, Kashima Hisashi, Takahashi Yutaka, (2009), "Roughly balanced bagging for imbalanced data", *Statistical Analysis and Data Mining*, Volume 2, Issue 5-6.
- [32] Hwang, H., Jung, T., Suh, E., (2004), "An LTV Model and Customer Segmentation Based on Customer Value: A Case Study on the Wireless Telecommunications Industry", *Expert systems with applications*, 26, 181–188.
- [33] Karagiannopoulos M G., Anyfantis D S., Kotsiantis S B., Pintelas P E., (2007), "Local cost sensitive learning for handling imbalanced data sets", *Mediterranean Conf on Control & Automation*, MED '07.
- [34] Kittisak, K., Nittaya, K., (2011), "A Data Mining Approach to Automate Fault Detection Model Development in the Semiconductor Manufacturing Process", *International Journal of Mechanics*, 4, Vol 5.
- [35] Kubat, M., & Matwin, S. (1997), "Addressing the curse of imbalanced training sets: One-sided selection", *In Douglas H. Fisher, editor, ICML*, pages 179–186. Morgan Kaufmann.
- [36] Lewis D, Gale W, (1998), "Training text classifiers by uncertainty sampling". *In Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information*, NY.
- [37] Li, M., Sleep, M., (2004), "Improving melody classification by discriminant feature extraction and fusion". *In Proc International Symposium on Music Information Retrieval ISMIR '04*, 11–14.
- [38] Linnet K., (1988), "A review on the methodology for assessing diagnostic tests", *Clin Chem*; 34(7):1379–86.
- [39] Ling, C., Li, C. (1998), "Data Mining for direct marketing problems and solutions", *In Proceedings of the 4th international conference on knowledge discovery and data mining (KDD)*. New York, AAAI Press
- [40] Marcellin Simon, (2008), « Arbres de décision en situation d'asymétrie » ; Thèse Docteur en informatique, Université Lumière Lyon II, laboratoire ERIC, sous la direction de Zighed, D. A.
- [41] Matias Di Martino, Federico Decia, Juan Molinelli, Alicia Fernandez, (2012), "Improving Electric Fraud Detection Using Class Imbalance Strategies", *International Conference on Pattern Recognition Applications and Methods, ICPRAM*.
- [42] McClish D, (1989), "Analyzing a portion of the ROC curve", *Medical Decision Making*, 190–195.
- [43] Miguéis, V.L., Van den Poel, D., Camanho, A.S., Cunha J, F., (2012), "Modeling partial customer churn: On the value of first product-category purchase sequences", *Expert Systems with Applications* 39
- [44] Pendharkar P, (2009), "Genetic algorithm based neural network approaches for predicting churn in cellular wireless networks service", *Expert Systems with Applications*, 36.
- [45] Petr Nálevka, Vojtěch Svátek, (2012), "Improving Efficiency of Telemedical Prevention Programs through Data-mining on Diagnostic Data", *4th International Conference on Bioinformatics and Biomedical Technology, IPCBEE* vol.29

- [46] Provost F, Fawcett T, (2001), "Robust classification for imprecise environments", *Machine Learning* 42(3):203–231.
- [47] Provost F, Fawcett T, (1997), "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions", *In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97)*, California, USA.
- [48] Rukshan Batuwita, Vasile Palade, (2009), "A New Performance Measure for Class Imbalance Learning. Application to Bioinformatics Problems", *ICMLA '09 Proceed of International Conference on Machine Learning and Applications*.
- [49] Seo, d, Ranganathan, c., Babad, Y., (2008), "Two-level model of customer retention in the US mobile telecommunications service market", *Telecommunications Policy*, In Press, Corrected Proof.
- [50] Sokolova M, Japkowicz N, Szpakowicz S, (2006), "Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation", *in the Proceedings of the 2006 Australian Conference*
- [51] Su C., Hasio Y., (2007), "An evaluation of the robustness of MTS for imbalanced data", *IEEE transactions on knowledge and data engineering*, vol. 19, no10, pp. 1321-1332.
- [52] Sukarna Barua, Md. Monirul Islam, Xin Yao, Kazuyuki Murase, (2012), "MWMOTE - Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning", *IEEE Transactions on Knowledge and Data Engineering*.
- [53] Jonathan Burez Dirk, Van den Poel, (2009), « Handling class imbalance in customer churn prediction », *expert systems with applications*. 36(3). p.4626-4636
- [54] Van Den Poel, D., B. Larivière., (2003), « Customer Attrition Analysis For Financial Services Using Proportional Hazard Models », *Working Papers of Faculty of Economics and Business Administration*, Ghent University, Belgium.
- [55] Weiss, G, M., Provost, F., (2003), "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction", *Journal of Artificial Intelligence Research*, AI Access Foundation and Morgan Kaufmann Publishers.
- [56] Xue Jing-Hao, Titterington D Michael, (2008), "Do unbalanced data have a negative effect on LDA?", *Pattern Recognition* N° 41 p 1558 – 1571.
- [57] Yong Zhang, Dapeng Wang, (2013), "A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets", *Abstract and Applied Analysis*; Article ID 196256.
- [58] Youden W, (1950), « Index for rating diagnostic tests », *Cancer*, 3 :32–35,
- [59] Yuanfang, D., Xiongfei, L., Jun, L., Haiying, Z., (2012), "Analysis on Weighted AUC for Imbalanced Data Learning Through Isometrics", *Journal of Computational Information Systems* 8: 1, 371–378
- [60] Yuan Bo, Ma Xiaoli, (2012), "Sampling + Reweighting: Boosting the Performance of AdaBoost on Imbalanced Datasets", *WCCI IEEE World Congress on Computational Intelligence*, June, 10-15, Brisbane, Australia.
- [61] Zhang, D., Lee, W, S., (2008), "Learning classifiers without negative examples: A reduction approach", *In 3rd International Conference on Digital Information Management*, ICDIM 2008, pages 638 –643.
- [62] Zhenqiu Liu, Hyslop T (2010), "Partial AUC for Differentiated Gene Detection", *Conference on BioInformatics and BioEngineering (BIBE)*, Philadelphia, IEEE International.
- [63] Tufféry Stéphane (2005), « Data Mining et statistique décisionnelle », Edition TECHNIP, Paris.
- [64] Qiang Yang, Xindong Wu, (2006), "10 Challenging Problems in Data Mining Research", *International Journal of Information Technology & Decision Making*, Vol. 5, No. 4 597–604.
- [65] Verhein F, Chawla S (2007), "Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets", *In ICDM*, pages 679–684, IEEE Computer Society.
- [66] Weiss G M, (2004) "Mining with rarity: A unifying framework", *SIGKDD Explorations*, 6:7-9

This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage:

<http://www.iiste.org>

## CALL FOR JOURNAL PAPERS

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. There's no deadline for submission. **Prospective authors of IISTE journals can find the submission instruction on the following page:** <http://www.iiste.org/journals/> The IISTE editorial team promises to the review and publish all the qualified submissions in a **fast** manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request of readers and authors.

## MORE RESOURCES

Book publication information: <http://www.iiste.org/book/>

Recent conferences: <http://www.iiste.org/conference/>

## IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digital Library, NewJour, Google Scholar

