Transactions on Machine Learning
and Data Mining
Vol. 12, No 1 (2019) 3-22
© Print ISSN: 1865-6781, Online ISSN: 2509-9337
ISBN: 978-3-942952-67-5

ibai Publishing
www.ibai-publishing.org

Measuring Classifier Performance with Risk and Error Matrix Charts

H.K. Koesmarno

Data Science and Engineering, ATO, Canberra, Australia
hari.koesmarno@ato.gov.au

Abstract. Measuring classifier performance is important in machine learning. Risk charts and error matrix charts have been developed for this purpose. The strengths and weaknesses of using these charts are outlined. Challenges with using these charts are discussed including how base rates and using prevalence data for building models and incidence data for evaluating models affect model performance. A number of solutions for overcoming these challenges are covered.

Keywords: Error Matrix, Confusion Matrix, Cumulative Gain Chart, Risk Chart

1 Introduction

The main objectives of this paper are (i) to illustrate the development and the usage of risk charts and error matrix charts for measuring model performance, (ii) to show how the performance of models are affected by the samples used to develop the models compared to the samples used to evaluate the models and (iii) to outline solutions that can be employed to improve the evaluation of models.

1.1 Measuring classifier performance

The performance of binary classifier will be illustrated using risk chart which was developed from cumulative gain chart [16] and error matrix charts which was developed from Proportion Score Function charts [9,10], for two types of target variables, i.e. binary and continuous variables. The binary variable, for example, distinguishes risk cases from non-risk cases, while the continuous variable represents the relative magnitude of the risk. For example, risk to revenue with tax collections has dollar amount while risk of rain has the magnitude of precipitation in either inches or millimetres. In some applications the magnitude of the risk variable can have "negative or positive" values or "debit or credit" in accounting applications. When developing a

supervised learning model, the priority is frequently aimed at the ranked order of the relative magnitude of the risk, e.g. revenue. Hence the modelling process should take into consideration both the classification of the risk and the relative magnitude of the risk.

In order to measure and visualise the performance of the classifiers [12] using both the binary and continuous target variables, a risk chart and error matrix chart are proposed in this paper. An example of risk chart is given in Fig. 1, while an example of error matrix chart is given in Fig. 3. There are a number of reasons for using these charts. Firstly, they are useful for evaluating the effects of the weighted classification problem [13]. Some classification problems can be weighted based on the importance of the cases. For example, with a tax evasion detection model, some cases are likely to provide greater revenue than the others and hence can be given greater weight. In some cases, there will be not much difference in terms of their strike rates, but there can be significant differences in their risk to revenue. The risk to revenue is particularly useful for analysis or exploring the benefit if only a portion of the population will be actioned for recovering the revenue because of limited audit and investigatory resources.

Both Risk charts and Error Matrix charts are sensitive to classifiers performance when compared to receiver operating curve (ROC) charts [4]. One challenge with measuring the performance of classifiers is class imbalance [1]. Recent use of risk charts and error matrix charts indicate that they are very sensitive to class imbalance when compared with ROC. However, ROC charts cannot be used to evaluate the relative magnitude of the risk where risk chart and error matrix charts can.

The risk charts and error matrix charts will also be used for (i) measuring classifier performance including risk which is a measure of the relative size of the gain or loss associated with target variable of each observation; (ii) comparing classifier performance prior and post intervention. Further improvement of risk charts and error matrix charts for measuring classifier performance will be discussed.

1.2 Base-rate variation with prevalence and incidence data

If a sample of size n is drawn for a binary classification problem, then the numbers of sample instances, n_0 and n_1 are respectively in class 0 and 1, $n_0 + n_1 = n$. The base rate is the ratio of n_1 and n_0, BR = n_1 / n_0. When the base rate is not 1, then there is class imbalance. One of the challenges with assessing classifier performance is on sample selection bias. This refers to differences in the proportion of cases selected for prevalence data when compared to incidence data. The prevalence data is used for model building, while the incidence data contains the cases which were actioned. Selection bias can distort the assessment of the classifier using several known methods such as misclassification rate and cumulative gain chart. Base rates can affect how well a classifier performs with identifying positive and negative cases. If the base rate is low, then the classifier will have a low strike rate although the misclassification

rate is high. If the base rate is high, then the classifier will have a high strike rate although the misclassification rate is low. These will be demonstrated in section 2 and 3.

The study of class imbalance data for model development is often done by three different methods, the algorithm approach, the resampling approach and feature selection approach [1,5,7,15]. However, this depends on the application, data distribution and modelling requirements. As an example, if the binary classification comprises of 4 different strata (S) distribution, S_1, S_2, S_3, S_4. In training set, two subsets of binary target variable 0 and 1, are formed based on stratified distribution $\{S_1, S_2, S_3\}$ and $\{S_4\}$ respectively. If each strata has equal number of population, then the two subsets in training set are not equally distributed. In many cases, the over sampling of S_4 to form the base rate (= n_1 / n_0) = 1 is required for model building in order to improve the accuracy or to avoid over fitting to class 0. If incident data was used for model evaluation especially when the information on sampling methods for model building or prevalence dataset is not available, then several issues can be encountered in practice.

The base rate of the prevalence dataset and incidence dataset can be very different, and these will cause issues in obtaining accurate measures of comparative model performance. Unlike ROC charts, which are not affected by base-rates, risk charts and error matrix charts can be misinterpreted when the base-rate changes from the data used to develop a model compared to the data employed to evaluate the performance of a model. These changes can arise because:

1. Each modeller has a tendency to use base-rate 1 from prevalence for sampling prior model building unless a modeller need to use class imbalance data in some applications. Having understood the characteristics of risk charts and error matrix charts, it is likely that the modeller who used smallest base-rate in a sample, will produce smaller error or bigger AUC, although their model performances are the same.
2. Once the model has been built, new data is used to obtain risk score for independent evaluation. Cases being selected for intervention are generally those which are high risk with those that are either low risk or no risk being excluded from consideration when it comes to evaluation of model performance. This distorts the results obtained using risk charts and error matrix charts.

Hence, the incidence data for evaluating model performance needs to be corrected for this bias. Solutions for doing this are proposed in Section 4.

2 Risk chart measures

Risk chart was developed from cumulative gain chart [16] in order to aleviate the bias from the class imbalance data, so the measures produced by both charts are different.

Risk chart is produced by modifying the measure obtained from cumulative gain chart, such as introducing the boundary and limit as illustrated in section 2.1. and standardising AUC of cumulative gain chart in section 2.2 using the geometry as illustrated in fig 2. The properties of risk chart is illustrated in section 2.3.

The risk chart involves plotting two variables, i.e. target variable (being 1 or 0) and risk variable (see Fig. 1). An example is where the data set has two class target-variable, e.g. adjusted or not adjusted cases when it comes to revenue collection; and the risk variable, e.g. the magnitude of the adjustment if made to recovery of revenue. The adjustment value is a measure of the size of the risk associated with each observation. Cases which have no adjustment following an intervention will of course have no risk associated with them (i.e. Adjustment $= 0$). Cases that do have an adjustment will have a risk associated with them, and for convenience the value of the adjustment is viewed as the magnitude of the risk.

Gain is a measure of the effectiveness of a classification model calculated as the percentage of correct predictions obtained with the model, versus the percentage of correct predictions obtained without a model. It shows the percentage of positive predictions that the model gains with each slice of the population. A higher overall gain indicates better performance. A risk chart (see Fig. 1) helps visualize the benefit of using a predictive model. It also allows the effectiveness of different predictive models to be compared. The information from the risk chart can be applied to determine which portion of the overall population or segment of population is to be targeted.
The advantages of using these charts include to:

i. Investigate why models improve when error increases due to the changes on base-rate of prevalence and incidence data
ii. Understand the characteristics, strength and weaknesses of the tools for measuring classifier performance.
iii. Identify the methods to be used for comparing performance prior and post modelling, especially when the base-rate changes

An example of a risk chart is shown in Fig. 1(a). If the lowest scores (i.e. the least risky cases) were removed from the sample, then the results as shown in Fig. 1(b) could be obtained. Hence, the area under curve for the risk chart cannot be used for measuring the model accuracy unless further factors are taken onto consideration. A more realistic measure for visualising the risk chart is proposed as in Fig. 2: this shows upper and lower limits of maximum area under curve for the risk chart.

Three curves in the risk chart in Fig. 1(a) are of interest. The first is the strike rate for each risk scored population, with the score going from high to low (i.e. left to right). The second shows the cumulative relative revenue based on the risk scores. The third is the cumulative cases based on the risk scores. Fig. 1(a) is the performance of a

classifier for prevalence data. It is assumed that the performance of this model is reliable and when new data is scored, it still produces same performance. As noted previously, in practice only the high risk-cases are usually selected for targeting to minimize costs. Hence, by reducing the potential true negative cases (as incidence data), the area under risk curves reduces (see Fig. 1(b)). In fact, the performances are the same, but the area is relative to the upper and lower limit (trapezoidal shape) of the risk charts, which are also consistently dependent on the base-rate. Hence, the proposed standardisation of the AUC measures is proposed in section 2.2.

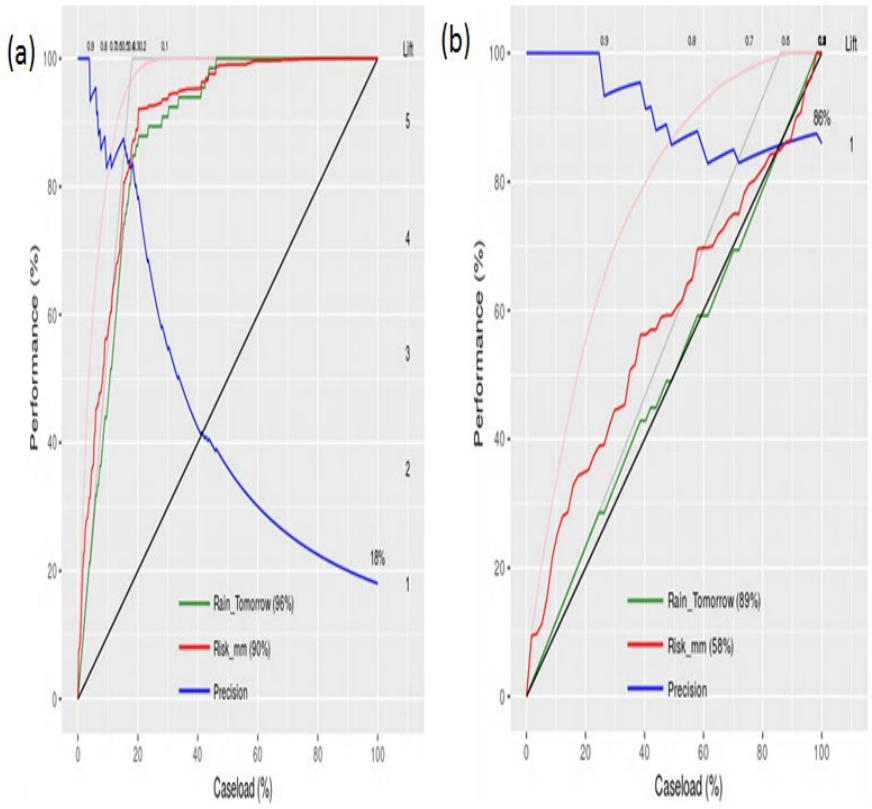

Fig. 1. Risk charts of classifier model performance prior (a) and after selecting score ≥ 0.5 (b)

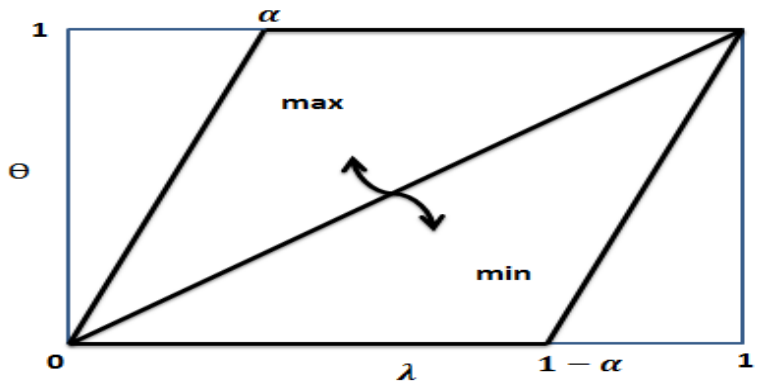

Fig. 2. Upper and Lower limit of Maximum Area Under Curve of Risk Chart where α is the base rate of binary classification

Fig. 2 illustrates how class imbalance affects the risk chart. The slope of the dashed line shows the percentage of positive cases. The higher the percentage, the more the gradient of the line decreases. If it was 100 percent positive cases, the line would be a diagonal going from bottom left to top right of the chart. The line would be vertical if there were very small or no positive cases. Fig. 2 is also used to assist for evaluating the risk. The x-axis is the case load which can be sorted either (1) high-to-low positive scores from 1 to 0 or (2) low-to-high for negative scores from -1 to 0. The curves will be reversed if the caseload was sorted from (i) low-to-high positive scores from 0 to 1 or (ii) high-to-low negative scores from 0 to -1.

Risk chart is produced by modifying the measure obtained from cumulative gain chart in order to obtain more reliable model evaluation and comparison of classifier performances, the methods and characteristics are illustrated below:

2.1 Boundary and limit

Let's defined λ = caseload or percentile of population sorted by its ranked scores,

$0 \leq \lambda \leq 1$:

$$\lambda_i = \sum_{i=1}^{N} \frac{1}{N} \tag{1}$$

Let's also define Θ = The cumulative gain or risk, $0 \leq \Theta \leq 1$,

(a) For r_i is binary (1 or 0) then the following formula applies:

$$\Theta(\lambda_i) = \frac{1}{n} \sum_{i=1}^{N} r_i \quad \text{where I = 1,…., N and n = count of } r_i \text{ when } r_i = 1 \tag{2}$$

(b) For quantifying the magnitude of r_i , m(r_i) continuous variable is used:

$\Theta(\lambda_i) = \frac{1}{M}\sum_{i=1}^{N} m(r_i)$ where I = 1,…., N and M = $\sum_{i=1}^{N} m(r_i)$ (3)

Let's define α is the base rate, α = n/N; where n = count of (r_i) when $r_i = 1$ and N = count of (r_i) when $r_i = 1$ or $r_i = 0$ (N = total number of instances).

The risk chart limit consists of

(a) upper boundary of the instances which are ranked from highest to the lowest.

(b) lower boundary of the instances which are ranked from lowest to the highest.

2.2 Standardizing AUC

In order to obtain consistent measure of AUC for risk chart, the standardised AUC (Ω) is proposed as :

AUC – min(AUC)/[max(AUC) – min(AUC)] (4)

this will give the range of standardised AUC between 0 and 1. In classification, the risk chart limit of the binary target variable has the following upper boundaries are:

(a) $\Theta = \frac{\lambda}{\alpha}$ for $\lambda < \alpha$ (5a)

(b) $\Theta = 1$ for $\lambda \geq \alpha$ (5b)

And lower boundaries are:

(a) $\Theta = 0$ for $\lambda < 1-\alpha$ (6a)

(b) $\Theta = \frac{\lambda}{\alpha} + \left(1 - \frac{1}{\alpha}\right)$ for $\lambda \geq 1-\alpha$ (6b)

The performance measure of classifier with binary target variable can be simply expressed as the standardized AUC:

$$\Omega = \frac{(AUC - \frac{\alpha}{2})}{1-\alpha} = \frac{2\,AUC - \alpha}{2(1-\alpha)} \quad (7)$$

It must satisfy $0 \leq \Omega \leq 1$ (8)

Where Θ = performance, λ = Caseload, Ω = Standardised AUC (Area Under Curve).

2.3 Properties

There are two properties can be derived from equation (7) and (8):

$2\,AUC - \alpha > 0$, so $\alpha < 2\,AUC$ (9)

$2(1-\alpha) > 0$, so $\alpha < 1$ (10)

The performance measure of classifier with binary target variable for balance class distribution can be derived by substituting $\alpha = 0.5$ in equation (7), to give:

$$\Omega = 2AUC - 0.5. \quad (11)$$

and for random performance where the original AUC is the lower triangle. The standardized AUC can be obtained by substituting AUC=0.5 to equation (7), to give $\Omega = 0.5$. Hence, both AUC and Ω are symmetrical at the diagonal: $\Omega = AUC = 0.5$. The performance measure of classifier with binary target variable for class Imbalance, in particular applying to rare case problems:

As $\alpha \rightarrow 0$, the equation (7) gives $\Omega \approx AUC$

As $\alpha \rightarrow 1$, the performance becomes less reliable, as it does not satisfy the condition in equation (7). Whenever possible, it is suggested to consider the conversion of α and scores by using (1-α) and (1-scores) if the condition in equation (8) and (9) cannot be achieved.

3 Risk and error matrix charts

A confusion matrix [11] or also known as an error matrix contains information about actual and predicted classifications provided by a classification model. Performance of such models is commonly evaluated using the data in the matrix. The construction of the error matrix chart is based on the generation of proportion score function (PSF) which was developed from [9]. The algorithm for generating a PSF is in Algorithm 1.

Error matrix chart is, as indicated previously, illustrated in Fig. 3. It is called by this name because of the characteristics of the charts in which the area can be represented as an error matrix. The vertical dash lines which illustrates the cut-off points and the horizontal curve line which represent as PSF. They are used to divide these charts onto four regions of the upper right hand of the chart containing the false positives (FP) and the lower right hand of the chart containing the true positives (TP). The upper left hand of the charts contains the false negatives (FN), the lower left hand of the chart contains the true negatives (TN). The error matrix can be represented as:

$$\begin{bmatrix} FN & FP \\ TN & TP \end{bmatrix}$$

Let‘s consider introducing low, medium and high risk by the low risk vertical line and the high risk vertical line.

Algorithm 1 : Generation of PSF

1. Input(score, predictedClass, trueClass, numberBin)
2. rankedScore ← rank(score, by numberBin)
3. For i = 1 to numberBin
4. sortedRS[i] ← get(rankedScore,i)
5. binSize[i] ← count(sortedRS[i])
6. correct[i] ← count(sortedRS[i], if predClass = trueClass)

```
7.      psf[i] ← correct[i]/binSize[i]
8.      lambda[i] ← i/numberBin;
9.    End;
10.  plot(psf,lambda)
```

Let's consider the y-axis Θ, and the x-axis λ. The four quadrant which formed by proportion score function, Θ(λ) and the cut-off point c, in the Fig. 3, represent the error matrix, hence it is called as error matrix chart, where:

$$\text{TP} = (1 - \lambda) - \int_c^1 \Theta(\lambda)\, d\lambda \quad (11)$$

$$\text{FP} = \int_c^1 \Theta(\lambda)\, d\lambda \quad (12)$$

$$\text{FN} = \int_0^c \Theta(\lambda)\, d\lambda \quad (13)$$

$$\text{TN} = \lambda - \int_0^c \Theta(\lambda)\, d\lambda \quad (14)$$

Hence other characteristics such NPV and PPV can be derived:

$$\text{NPV} = 1 - \frac{1}{\lambda}\int_0^c \Theta(\lambda)\, d\lambda \quad (15)$$

$$\text{PPV} = 1 - \frac{1}{1-\lambda}\int_c^1 \Theta(\lambda)\, d\lambda \quad (16)$$

In order to obtain error matrix decomposition, low, medium and high risk lines were introduced in Fig. 3 and the error matrix decomposition was obtained. The objective of the error matrix decomposition is to enable local classifier performance analysis of for example either high, medium or low risk cases.

Error matrix chart enables the examination of classification performance and misclassification rate. It provides different measures from AUC in ROC or risk chart. Some of the measures produced in the error matrix chart and its segments can be useful in certain applications. An example is where the predictive model was intended to identify high risk cases of non-compliance. If the targeting was based on the overall model performance, then it maximises the strike rate of the non-risk (compliant) cases as they are the majority in the population. The error matrix chart can also be divided into several segments of population and these segments in the error matrix chart can be compared based on the true positive or negative cases for selecting the model with low, medium and high non-compliance for example, refer to the two solid vertical line with caseload 0.2 and 0.8 inside the error matrix chart in Fig. 3. The two solid lines divides the population onto 3 segments, i.e. low, medium and high risk of non-compliance.

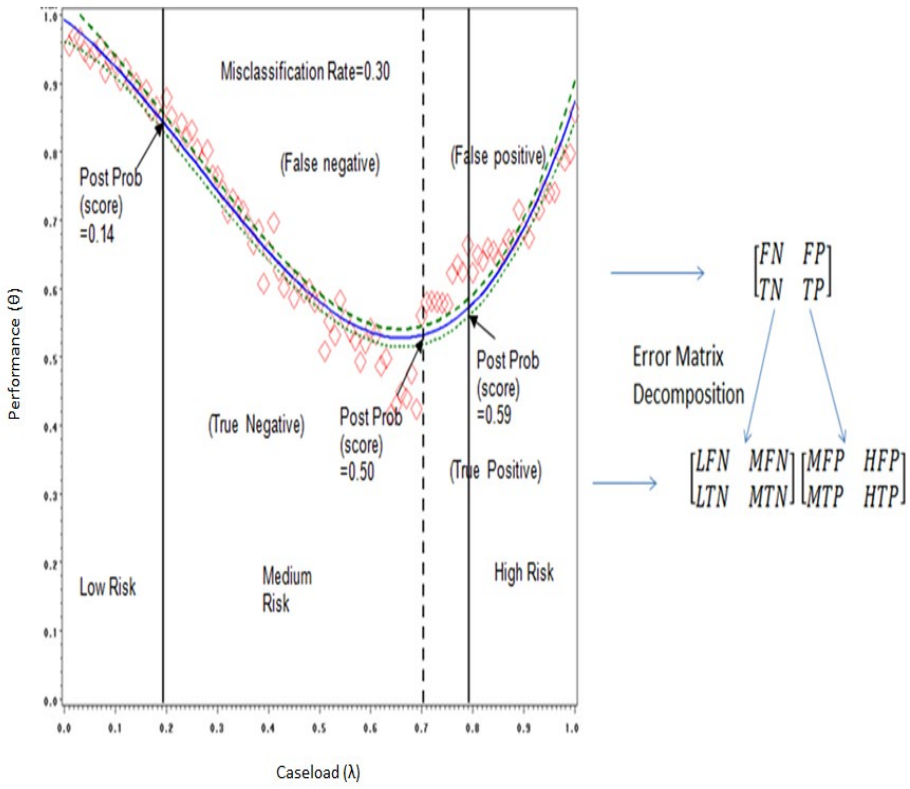

Fig. 3. Error Matrix charts with low and high risk cut-off points and their matrix representation.

There are many binary classification models which can be measured its classification 'gain' and 'loss'. Gain risk variable is the relative magnitude of the risks when the prediction is correct and has positive impact or value, while loss risk variable is the relative magnitude of the risk when the prediction is incorrect and has negative impact or value. When an instance is predicted positive, the actual can be either (a) positive, then it has gain risk variable and (b) negative, then it has loss risk variable. Similarly, when the instance is predicted negative, the actual can be either (a) positive, then it has loss risk variable and (b) negative, then it has gain risk variable.

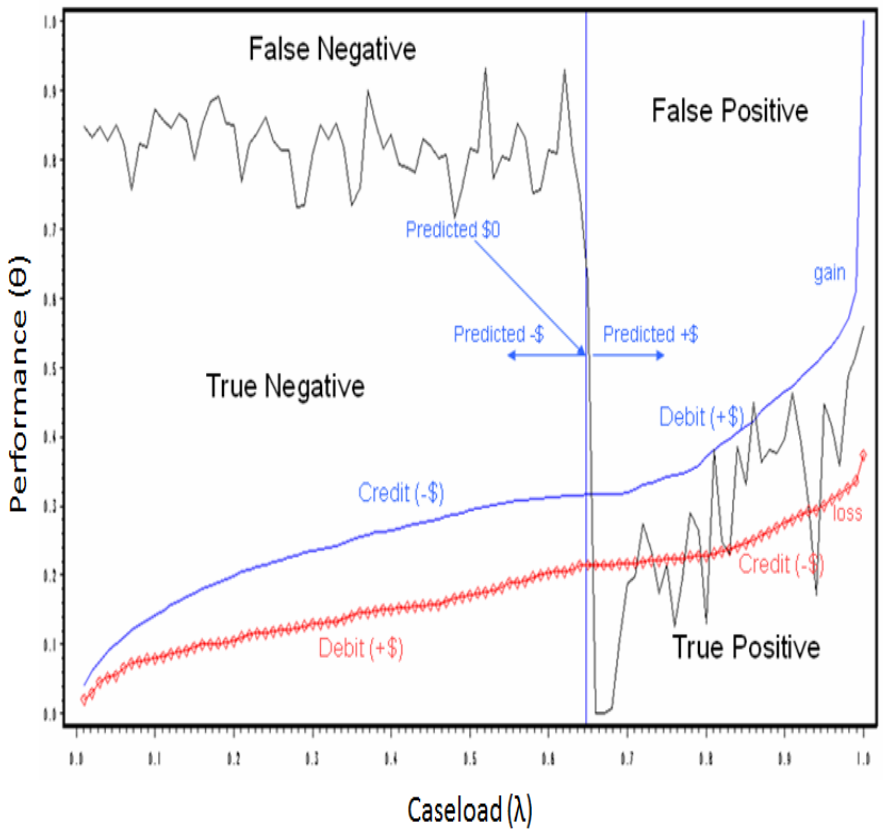

Fig. 4. **Fig. 4**. Error Matrix with gain and loss risk for describing two stage model.

A risk detection model can be used to illustrate the gain and loss risk in revenue. Each outcome of the detection would produce positive or negative relative magnitute of revenue. This problem can also be considered as two-stage modelling [6,8]. The first stage is to predict if a case will result positive or negative outcome. The second stage is to predict the relative magnitute revenue gain for both positive and negative outcomes. PSF has been used to demonstrate the first stage, i.e. the measure for false positive, true positive, false negative and false negative as in Fig. 4. In order to provide a more comprehensive view of the classifier performance, the 'gain' and 'loss' chart should be part of the error matrix chart as demonstrated in Fig. 4.

An example of misleading or biased results is where the sample of prevalence and the sample of incidence cases are different:

i. Let's consider sample with 41 is true negative, 5 false positive, 3 false negative, 5 true positive.
ii. In order to minimise the intervention cost, the true negative cases being reduced, by reducing the true non-risk cases from 41 to 5, it saves 35/58 = 35.185% resources.
iii. The representation of error matrix is changing as shown below:

$$\begin{bmatrix} 3 & 5 \\ 41 & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 5 \\ 5 & 5 \end{bmatrix}$$

Their initial misclassification error change from e = 0.14814815 to $e = 0.444444$.

The approach to deal with these issues will be discussed on next section.

4 Performance of models using prevalence and incidence data

When comparing classifier performance of prevalence and incidence data are required, it is important to make sure the results are comparable. There are several issues when the sampling used for model building and/or sampling of model evaluation [12] are not randomly drawn. These issues are illustrated next.

4.1 Reasons

Comparisons of classifier performance utilising prevalence and incidence data is necessary for several reasons:

Improving model deployment. Constructing the risk and error matrix charts using the incidence data are required for analysing the effects of changing the threshold/cut-off points and case-load selection for model deployment.

Monitoring model performance. One question that often needs resolution is "Has there been any concept drift with model performance where for example it strays from detecting fraud?" If there is concept drift and the model performance is not at an acceptable level, then the model should be rebuilt.

Business Intelligent. Model/classifier performance using incidence data is frequently requested for business performance analysis and reporting.

4.2 Prevalence and incidence sampling

In order to achieve the objectives for comparing model performance using prevalence and incidence data, the sampling selected for both types of data needs to be from the same distribution. For example, if the prevalence sampling is drawn from accidental

sampling (see below), then the incidence sampling should be the same as used in prevalence sampling. The focus of this paper is on measuring classifier performance where the base-rate of prevalence and incidence data is significantly different. This issue is generally due to the method of sampling used to build the model (prevalence) and the sampling used to analyse the modelling outcome (incidence) are frequently different in practice. In order to compare the performance of prevalence and incidence data, the sampling used for building the model should be the same distribution as that used for model evaluation. Generally, a model can be constructed using:

i. **Accidental sampling.** This is the most applicable solution for many data mining applications especially for detecting fraud. The known cases of fraud are usually rare in terms of their occurrence and can be expensive to obtain. Hence the need to maximise the data set used for training purposes. The sample used will often be what is readily available and convenient. This is known as grab, convenience or opportunity sampling. It involves the sample being drawn from that part of the population which is close at hand. The model developer using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative. This type of sampling can be useful for initial model building.

ii. **Non-Accidental sampling.** Most common forms of non-accidental samplings are random sampling, systematic sampling, stratified sampling, cluster sampling and probability-proportional-to-size sampling. While these are the preferred methods for building models, they can have the disadvantages that the positive cases included in these samples may not be readily apparent to those who develop models. That is, those who have this responsibility may not identify all the true positive cases. This is another way of saying some true positive cases remain invisible in the selected sample. If the non-accidental sample contains a limited number of positive cases, this can undermine model performance.

As has been emphasized incidence data usually has cases which have high risk scores and have been actioned. Therefore, the outcomes with these cases are known. Hence, this accidental sample is very different from the sample used to develop the model.

Here the distribution of incidence data has significantly changed from the distribution of prevalence data. There are three possible methods for dealing this challenge. They are:

i. ***Oversampling*** – where all the cells/clusters/strata and scoring percentiles have at least 'minimum' required number of data, while several others have more data than what is required. The "correction sampling incidence data"

proposed in this paper can be utilised and this should provide a reliable correction sampling.

ii. ***Under Sampling*** – There are two scenarios: (i) One or more of the cells/strata/clusters have less data than what are required by the threshold of the sampling. The correction sampling incidence data can be employed. However, the result may generally be less reliable than the one with over-sampling. (ii) One or more of the cells/strata/clusters have no samples or missing data. Here the accuracy of the corrected sampling for these entries depends on the accuracy of the assumptions applied about the distribution they were drawn.

iii. ***Same sampling*** – This sampling usually occurs when the prevalence and incidence data are drawn using the same methods.

There are two possible methods with same sampling to select the incidence data for model evaluation: (a) Non-Accidental sampling such as random sampling can be used for measuring classifier performance; (b) Accidental sampling, this is not recommended for model evaluation as it will be inaccurate if the distribution of the data is different from the prevalence data.

If prevalence data is drawn using accidental sampling and is used for building the model, then there is a need to reconstruct the incidence data prior measuring model performance. This can be called 'corrected sampling incidence data". The reconstruction or correction of the incidence data can be done by "substitution sampling". Substitution sampling is a sampling algorithm used to reconstruct the prevalence data using the incidence data. The main characteristics of substitution sampling is "drawing a random sample" from prevalence data, then substituting each instance using incidence data. The substitution of the prevalence instances which are the same strata or cluster or cell as the incidence data is being substituted. The sample size of prevalence data is not the same as incidence data in practice. There are three possible scenarios of sampling being over, under or the same size with the 'random sample' drawn from prevalence data. If the data in each strata or cluster or cell are either over or under sampling, then bootstrap or jackknife method [3,14] can be utilised for substituting instances in each strata or cluster or cell, until all instance from "substitution sampling" comes from incidence data. The main advantage with substitution sampling is how it captures key population characteristics in prevalence data, the sample collected for model building and the data drawn from 'accidental sampling'. This method of sampling produces characteristics in the sample that are proportional to the prevalence data. The detail is provided in next section.

4.3 Corrected Sampling Incidence Data

Substitution sampling is a method of sampling that involves the substitution and division of a population into smaller groups known as strata or cluster or cell. The strata

and cluster are formed based on members' shared attributes or characteristics. A random sample from each stratum or cluster is taken in a number proportional to the stratum's or cluster's size when compared to the population. These subsets of the strata or clusters are then pooled to form a random sample. Fig. 5 illustrates the description of "substitution sampling" when the sample has only two strata or cluster. The bigger data set (LHS) indicates the sample drawn from prevalence data, while the smaller data set (RHS) is the sample belongs to incidence data.

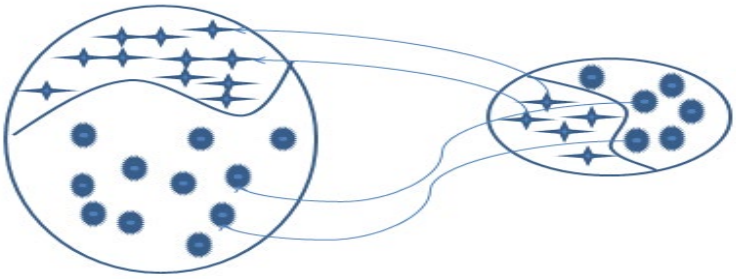

Fig. 5. Substitution sampling, the bigger data set (LHS) indicates the sample drawn from prevalence data, while the smaller one (RHS) is the sample belongs to incidence data.

There are two substitution sampling strategies which are described below:

Mixed Resampling procedure

Let's define the prevalence stratified data is $x_1, x_2, \dots, x_n$, where x_n is the number of cell size at nth cell. The incidence stratified data is $y_1, y_2, \dots, y_n$, where y_n is the number of cell size at nth cell. The stratified sampling need to be carried out and the incidence data should be added by a number of sample in order to match with some proportion of prevalence data which can be formulated as:

$$y_i + \Delta_i = \alpha x_i \tag{17}$$

In order to minimise the increase of the overall sample size:

$$\text{Minimise } \sum_{i=1}^{n} \Delta_i \tag{18}$$

$$\sum_{i=1}^{n} \Delta_i \geq 0 \text{ for increasing the overall sample size.} \tag{19}$$

Equation (17) can be expressed as:

$$\Delta_i = \alpha x_i - y_i \tag{20}$$

Substituting (20) onto expression (19) and (18)

$$\text{Minimise } \sum_{i=1}^{n} \alpha x_i - y_i \text{ and } \sum_{i=1}^{n} \alpha x_i - y_i \geq 0. \quad (21)$$

Let us minimise $f(a) = a \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \; and \; f(a) \geq 0$

$$\text{Hence } a = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \text{ and } \Delta_i = \left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right) x_i - y_i \quad (22)$$

There are 3 possibilities of cell sampling required:

If $y_i < x_i$ then use y_i plus additional re-sampling $\Delta_i$with replacement from y_i
If $y_i = x_i$ then use y_i
If $y_i > x_i$ then use sampling without replacement from y_i

Over Re-sampling procedure

For over re-sampling procedure application, the following condition applies: $\vee i : \Delta_i \geq 0$ then we need to introduce β adjustment, so that all sample are not being reduced, but being increased. We need to substitute Δ_i with $(\beta + \delta_i)$ where $\vee i : \Delta_i \geq 0$, substituting this to equation (17) in order to get

$$y_i + (\beta + \delta_i) = \alpha x_i \quad (23)$$

Equation (23) is used the same way as in expression (18) to (21) in order to obtain

$$a = \frac{(\sum_{i=1}^{n} y_i) + \beta}{\sum_{i=1}^{n} x_i} \quad (24)$$

and substituting α onto equation (23) to give:

$$\delta_i = \left(\frac{(\sum_{i=1}^{n} y_i) + \beta}{\sum_{i=1}^{n} x_i}\right) x_i - y_i - \beta \quad (25)$$

Substituting equation (24) to $\Delta_i = (\beta + \delta_i)$ gives:

$$\Delta_i = \left(\frac{(\sum_{i=1}^{n} y_i) + \beta}{\sum_{i=1}^{n} x_i}\right) x_i - y_i \quad (26)$$

Hence, we need to minimise Δ with the following constraint:

$$\vee i : \Delta(\beta) = \Delta_i \geq 0 \quad (27)$$

The search of the value β is required in order to
minimise $\Delta_i \; and \; \Delta_i \geq 0$ for i = 1, …, n
where n is the number of stratified cells as in Algorithm 2.

The white wine data from UCI data repository [2] was used for the experiment using mixed resampling procedure and over resampling procedure. The data was clustered into seven clusters. One of the clusters consists of only one instance and was removed. The random sample of 200 instances was selected as incident data, while the remaining 4697 instances was selected as prevalence data. The results of experimentation using the methods illustrated above for optimized mixture sample is shown in Table 1 while optimized over sampling is shown in Table 2. For comparison purposes, the Balance Bootstrap resampling B=10 [3] on oversampling of incidence data is illustrated in Table 3.

Algorithm 2: Corrected Sampling

1. β ← abs ($\sum_{i=1}^{n} f(\Delta_i)$); where $f(\Delta_i) = \Delta_i \; if \Delta_i < 0 \; and \; f(\Delta_i) = 0 \; if \; \Delta_i \geq 0$
2. g0 = 0; β0 = 0; Θ = 0; r = (1+sqrt(5))/2; converge = false;
3. Evaluate: $\Delta(\beta)$; if $\Delta(\beta) < 0$ then g = 0; else g = 1;
4. While convergence eq false then
5. Ω = (1-r) * (β-β0);
6. if g eq 1 then β01 = β0 + Ω; $\beta_{11} = \beta - \Omega$;
7. Evaluate: $\Delta(\beta 01)$; if : $\Delta(\beta 01) < 0$ then g01 = 0; else g01 = 1;
8. Evaluate: $\Delta(\beta 11)$; if : $\Delta(\beta 11) < 0$ then g11 = 0; else g11 = 1;
9. if g01 eq 1 and g11 eq 1 then
10. Diff = β01 – β0; β = β01; g=g01;
11. If g01 eq 0 and g11 eq 1 then
12. Diff = β11 – β01; β0 = β01; β = β11;g0=g01;g=g11;
13. If g01 eq 0 and g11 eq 0 then
14. Diff = β – β11; β0 = β11;g0=g11;
15. If diff < 3 then converge = true
16. else $\beta 0 = \beta$; $\beta = \beta + \Omega + \Theta$; Θ = Ω;
17. Evaluate: $\Delta(\beta)$; if : $\Delta(\beta) < 0$ then g = 0; else g = 1;
18. EndWhile;
19. β = round(β); Δ = $\Delta(\beta)$
20. While Δ < 0
21. β = β +1; Δ = $\Delta(\beta)$
22. endWhile;
23. Output(β)
24.

Table 1. Optimised mixture sampling of incidence data

	Prevalence		Incidence		Adjusted Incidence Sample		
Cluster	n_0	%	n	%	Δ	Δ_1	(n+Δ)/n
1	675	14.3678	26	13.0	2.7356	3	1.10522
2	1227	26.1175	56	28.0	-3.7650	-4	0.93277
3	101	2.1499	5	2.5	-0.7003	-1	0.85994
4	1309	27.8629	66	33.0	-10.2742	-10	0.84433
5	948	20.1788	35	17.5	5.3576	5	1.15307
6	437	9.3018	12	6.0	6.6037	7	1.55031

Table 2. Optimised over sampling of incidence data

	Prevalence		Incidence		Adjusted Incidence Sample			
Cluster	n_0	%	n	%	Δ	Δ_1	(n+Δ)/n	%
1	675	14.3678	26	13.0	7.9080	8	1.30416	17.0
2	1227	26.1175	56	28.0	5.6373	6	1.10067	31.0
3	101	2.1499	5	2.5	0.0736	0	1.01473	2.5
4	1309	27.8629	66	33.0	-0.2435	0	0.99631	33.0
5	948	20.1788	35	17.5	12.6220	13	1.36063	24.0
6	437	9.3018	12	6.0	9.9523	10	1.82936	11.0

Table 3. Bootstrapping (B=10) on oversampling of incidence data

Cluster	n	%
1	351	15.54
2	567	25.10
3	45	1.99
4	594	26.29
5	486	21.51
6	216	9.56

4.4 Application to Big Data

Recent technologies in social media, multimedia and online transactions can be used to generate big data in the form of structure and unstructured data. The data mining can be applied to the big data for predictive modelling and knowledge discovery [5,12[. The sampling procedures illustrated earlier can be also useful for getting the right distribution in training, validation, testing and scoring sets in the big data. The prevalence data, i.e. training, validation and testing set of big data can be generally obtained by under sampling. However, if there is existence of concept drift which affect the distribution of some features in the big data, then mixed or over resampling need to be utilized. Evaluation of a model using incident data may require mixed or over resampling procedures: The incident data can be sourced from the results of intervention and the distribution of the incident data may not be the same as the prevalence data.

The weakness of using overall accuracy measures extracted from error matrix and cumulative gain chart in class imbalance data are not optimal metric, since the outcomes will be strongly biased towards the majority class. Some solutions utilising risk and error matrix charts have been illustrated. The sampling procedures to make the two distributions of training and scoring set equal, can be also utilized to remove or alleviate the bias. However, when come to evaluation of the “intervention” or “model performance” using the incident data, if the domain experts use the results without resampling, then the overall accuracy of error matrix may be compromised due to imbalance data. Instead of using an overall measure, the relevant information from several segments of the risk and error matrix charts from the classifier performance of imbalance data set is recommended. The risk and error matrix charts can also be used to segment the population of the big data to identify several subpopulation or strata based on different segments’ performances even when they are on the same class. This information can be used for business intelligent and knowledge discovery in the big data mining.

5 Conclusion and Future Directions

Error Matrix charts enable the visualisation of classification errors and their composition. It provides different measures from AUC in ROC or AUC in risk chart. The measures from error matrix chart and its composition can be very useful for many applications especially class imbalance and rare cases where the overall measure such as the AUC in ROC may not be useful. Both risk chart and error matrix charts are very sensitive to base-rates which usually occur when class-imbalance data are used for modelling. Two approaches have been suggested for comparing classifier performance with risk and error matrix charts as both approaches provides different types of measures of model performance.

When evaluating model performance of prior and post interventions, it is important to make sure the same sampling strategy is applied to both prevalence and incidentence datasets, otherwise it can bias the measure of model performance. Although the sampling of incidence data can be corrected with the algorithm proposed in this paper; the severe under-sampling of incidence data still cannot be solved with any re-sampling methods. This is due to mainly the sample size being too small or alternatively due to data being missing in each cell. Future research of the proposed methods need to be directed towards understanding further the properties and characteristics of risk charts, error matrix charts and their comparative performances with respect to sampling for prevalence and incidence data and their application for an imbalance big data mining [5].

Acknowledgement

The author is very grateful to Graham Williams for implementing some of the proposed earlier risk chart revision to Rattle and Warwick Graco for assistance editing this paper. No real ATO data was used in the paper due to privacy, legal and security requirements.

References

1. Chawla N.V. (2009) Data Mining for Imbalanced Datasets: An Overview. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA (2009)
2. Cortez, P, Cerdeira, A., Almeida, F., Matos, T. and Reis, J.: Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236. (2009)
3. Efron, B, Tibshirani, R. An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall. ISBN 0-412-04231-2 (1993)

4. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters, 27, 861-874. (2006)
5. Hassib, E.M, El-Desouky, A.I, El-kenawy, E.M, Elghamrawy, S.M. An Imbalanced: Big Data Mining Framework for Improving Optimization Algorithms Performance. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2955983 (2019)
6. Heckman, J. J.: The Common Structure of Statistical Models of Truncation Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models, *Annals of Economic and Social Measurement,* 5 (4), 475-492 (1976)
7. Japkowicz, N, and Stephen, S.: The class imbalance problem: A systematic study. *Intelligent data analysis* 6(5), 429-449 (2002)
8. Kapitula LR.: When Two Are Better Than One: Fitting Two-Part Models Using SAS®. In Proceedings of the SAS Global Forum (2015)
9. Koesmarno, H.K.: Class-size percentile transformation for reconstructing a distribution function. Journal of Applied Statistics, 23 (4): 423-434. (1996)
10. Koesmarno, H.K.: Risk and Error Matrix Charts. In Advances in Data Mining, Poster Proceedings of 19th Industrial Conference on Data Mining, Petra Perner (Ed.). p. 16 – 30 (2019).
11. Kohavi, R., Provost, F: Glossary of terms, Machine Learning, Vol. 30, No. 2/3, pp. 271-274 (1998)
12. Perner, P. *Data mining on multimedia data*. Vol. 2558. Springer Science & Business Media (2002)
13. Polo, J. L., Berzal, F., and Cubero, J. C.: Taking class importance into account. In *Hybrid Information Technology, 2006. ICHIT'06. International Conference on* (Vol. 1, pp. 1-6). IEEE (2006)
14. Shao, J. and Tu, D.: The Jackknife and Bootstrap. Springer-Verlag, Inc. (1995)
15. Weiss, G. Mining with rarity: A unifying framework. SIGKDD Explorations 6(1):7{19, 2004
16. Williams, G.J.: Data mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer (2011)