FAQ
CTU
ARY

FAQCTUARY 2021-1

ICC, DEVIANCE, AUC : WHAT ARE THE DIFFERENCES ?

By Louise d'Oultremont, Michel Denuit and Julien Trufin

April 2021

Detralytics

DISCLAIMER

The content of the FAQctuary’s is for a pedagogical use only. Each business case is so specific that a careful analysis of the situation is needed before implementing a possible solution. Therefore, Detralytics does not accept any liability for any commercial use of the present document. Of course, the entire team remain available if the techniques presented in this FAQctuary required your attention.

Detralytics
Rue Belliard/Belliardstraat 2
1040 Brussels
www.detralytics.com
info@detralytics.eu

Contents

1 Introduction

Assessing performance of different models to determine which one gives the best estimation is essential in modelling processes. Different measures are widely used in order to compare models. Some of these widely used measures are for example the Akaike Information Criterion, the Bayesian Information Criterion and the deviance. For binary models, the area under the ROC curve (AUC) is another widely used measure. There exists a huge amount of other more or less known measures that can be used to the same purpose. An example is the integrated concentration curve (ICC), another measure also used to compare models. In this FAQctuary, the focus is put on the deviance, the AUC and the ICC. A comparison of these three measures is made on two data sets to observe the different behaviours of each measure and answer questions like the followings: ICC is not yet very spread in the practice but is it different from the others? Does all these measures behave the same way or is it better to rely more on one than the others?

2 Definition of the measures

In this section, we define the three measures under consideration.

2.1 The Integrated Concentration Curve

As defined in Denuit et al. (2019), the integrated concentration curve is the area under the concentration curve over the whole interval $[0, 1]$. Formally, the concentration curve (CC) and the ICC are given by the following expressions:

$$CC[\mu(X), \pi(X); \xi] = \frac{\mathbb{E}\left[\mu(X) I\left[\pi(X) \leq F_{\pi}^{-1}(\xi)\right]\right]}{\mathbb{E}[\mu(X)]}$$

and

$$ICC[\mu(X), \pi(X)] = \int_0^1 CC[\mu(X), \pi(X); \xi] \mathrm{d}\xi$$

where $\mu(X) = \mathbb{E}[Y|X]$ is the true premium, $\pi(X)$ is the estimation predicted by the model under consideration given the information X and F_π is the distribution function of π.

Meaning of this measure

The concentration curve as expressed here is the proportion of the total true premium income corresponding to $100\xi\%$ of contracts with the smallest premium π.

The independence between the response and the predictors is represented by the 45° line concentration curve which leads to an ICC of 0.5. The more the concentration curve is far from this independent line, the smaller becomes the ICC and the more the response and predictors are dependant. The further the curve is from the independent line, the more discriminant the model is. This means that the smaller the ICC, the more discriminant the model.

2.2 The Deviance

The deviance can be defined as follows

$$D = 2\left(ln(L_s) - ln(L_\pi)\right)$$

where L_s is the likelihood of the saturated model, the more precise model and L_π is the likelihood of the fitted model.

Meaning of this measure

Theoretically, this indicator measures the deviation of the estimations from the observations. The smaller this deviation, the more appropriate the model.

2.3 The Area Under the ROC curve

The ROC curve is a representation of the True positive rate (TPR) on the y-axis in function of the False positive rate (FPR) on the x-axis. This curve aims at representing the proportion of observations well estimated in a binary model. The classical area under the curve is defined as the integral of this ROC curve over the unit interval.

The variable has 2 levels, 0 and 1. There exists four scenarios:

- The estimation is 0 and the real value too, this value is considered as true positive (TP).
- The estimation is 0 but the real value is 1, this is called false positive (FP).
- The estimation is 1 but the real value is 0, this is a false negative (FN).
- The estimation is 1 and the real value too, this is a true negative (TN).

Formally, the true positive rate is defined as $TPR = \frac{TP}{(TP+FN)}$ and the false positive rate as $FPR = \frac{FP}{(TN+FP)}$. As estimations are probabilities, a threshold determines if the estimate goes into the 0 or the 1 class. Computing the TPR and FPR for different thresholds gives the ROC curve. The Figure 2.1 represent an example of this curve. The AUC is the integral under this curve on the unit interval.

In the present case, the response is Poisson distributed and has more than two classes. This is handled as in Hand & Till (2001) by taking the mean of the different AUC computed for each pair of class of the variable. For a variable taking values from 0 to 4, the global AUC can be expressed as

$$AUC = \frac{1}{10}\sum_{i<j} AUC(i,j) \text{ for } i,j \in \{0,1,2,3,4\}.$$

Results in the following are computed using the *multiclass.roc* R function from the *pROC* package which uses this method. Note that the uniform weighting of AUC may not be appropriate if the number of observations in each class varies a lot.

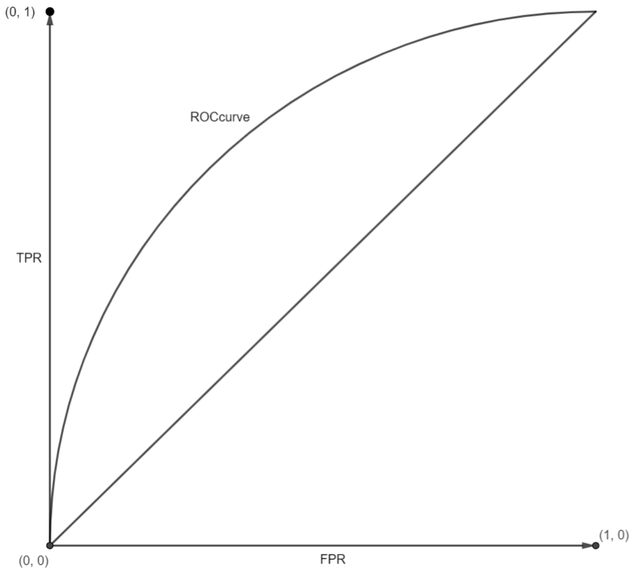

Figure 2.1: Representation of the ROC curve

Meaning of this measure

This indicator measures the fact that the estimated value corresponds or not to the initial response value. As only two values exist, it allows to determine the part of effective estimations and of failed estimations. This number indicates "how much" the estimations are well classified regarding the observations. In the case of a Poisson variable, the multiclass AUC reflects how estimations are on average well classified two by two, which is a little bit less intuitive.

2.4 Discussion on the three measures

These three measures are useful to evaluate the goodness of a fit. Nevertheless, each measure considers different elements to determine the better fit on the data. The ICC considers that the best model is the one discriminating the more the data. The deviance selects the best model as the one deviating the less from the saturated model. The multiclass AUC on its side looks at how the model classes the different modalities of the response variable two by two. The model that on average classes the better the different classes will be considered as the best. This is the model with the largest AUC.

3 First example : A simulated data set

3.1 Data set

The structure of the first data set is known, the data is generated on constructed features. Four different variables $X = (X_1, X_2, X_3, X_4)$ are available to impact the annual claim frequency of an individual :

- X_1 is the policyholder's gender (male or female);
- X_2 is the policyholder's age (from 18 to 65 years);
- X_3 is the premium split (paid annually or not);
- X_4 is the car a sport model or not (yes or no).

All variables are supposed independent and distributed as follows :

$$\begin{aligned}
&\mathrm{P}\left[X_1 = female\right] = \mathrm{P}\left[X_1 = male\right] = 0.5;\\
&\mathrm{P}\left[X_2 = 18\right] = \mathrm{P}\left[X_2 = 19\right] = \ldots = \mathrm{P}\left[X_2 = 65\right] = 1/48;\\
&\mathrm{P}\left[X_3 = yes\right] = \mathrm{P}\left[X_3 = no\right] = 0.5;\\
&\mathrm{P}\left[X_4 = yes\right] = \mathrm{P}\left[X_4 = no\right] = 0.5.
\end{aligned}$$

The annual number of claims is denoted by Y and is assumed Poisson distributed with the following value for the parameter λ:

$$\lambda(x) = 0.1\times\big(1 + 0.1I_{\{x_1=male\}}\big)\times\left(1 + \frac{1}{\sqrt{x_2 - 17}}\right)\times\big(1 + I_{\{x_4=yes\}}\left(0.5I_{\{x_2\in[18,35[\}} - 0.5I_{\{x_2\in[45,65[\}}\right)\big)$$

where $I_E = 1$ if the event E is realized, 0 otherwise.

Men have a claim frequency 10% higher than women. The claim frequency decreases with the age of the policy holder. The premium split does not have an impact on the claim frequency. An interaction exists between the policyholder's age and the fact that the car is a sport model or not. In fact, if the policyholder has a sport car and is between 18 and 35 years, the claim frequency increases of 50%. Similarly, if the policyholder has a sport car but is between 45 and 65 years, the claim frequency decreases from 50%. A data set of 500.000 policies is generated in this example.

3.2 Models fitting

The true annual claim frequency $\lambda(x)$ is known for each policyholder x in the simulated data set. In order to estimate this annual claim frequency, the data is divided into five folds. Each of these is going to be the validation set while the four others are the training sets. The four following models are fitted on each training set:

1. A GLM without any variables (a simple mean of the response).
2. A GLM including all the principal effects : removing the non significant variables, only the gender is included in this model.

3. A GAM including all the principal effects : removing the non significant variables, the gender and the age remain in the model.

4. A GBM including all the variables in order to capture the interaction included in the data.

The models from 1 to 4 should be ranked from the furthest to the closest of the true claims frequency values. In fact, the GBM is the only model able to account for the interactions, we expect it to have a better performance than the others.

3.3 Comparison of the models

Figure 3.1 displays the values of the different measures on each validation set. From the top to the bottom, the ICC, the deviance and the AUC can be observed. The behaviour of each of them is commented in the following.

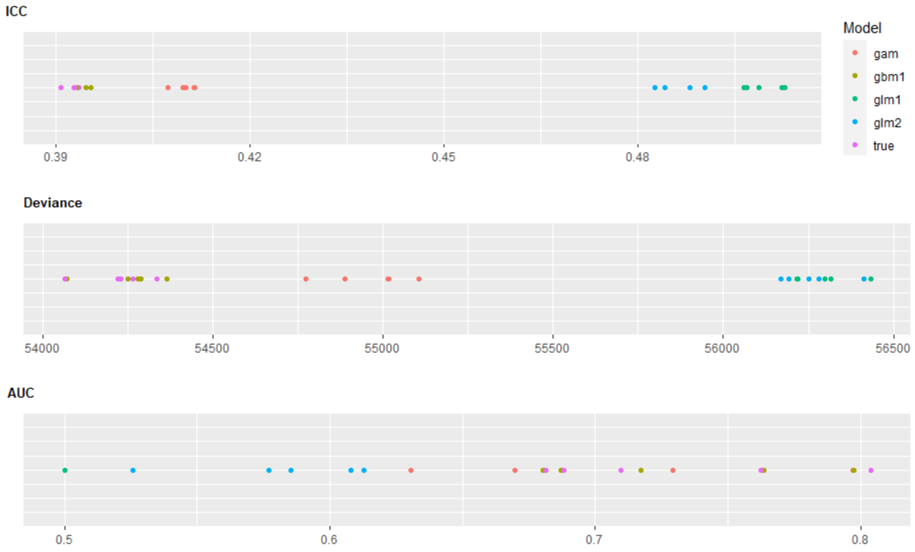

Figure 3.1: Values of the ICC, deviance and AUC on the simulated data.

Looking at the ICC, we observe a clear difference between the considered models. This means that the discrimination of policyholders varies from one model to another. The results of all models can be interpreted and classified easily. The models are ranked as expected. The worse model is the intercept only GLM, all policyholders are considered to have the same risk, it is the furthest from the true model. Then comes the second GLM, making only a difference between men and women, followed by the GAM, the GBM and finally the true values. This seems logical with the known structure of the frequency. In fact, the GAM allows to consider the continuous variable Age that cannot be handled correctly by the GLM but do not capture the interaction effect. In the different models fitted here, only the GBM can account for this effect, considering all variables, this explains why it has the best performance and we observe that the GBM is closed to the true results.

As far as the deviance is concerned, the ranking between the different models remains the same. The difference is that both GLM are not clearly differentiated. Even if it seems clear that the GLM with the variable "Gender" should be a better estimation than the one without variables given the constructed data, only looking at deviances does not allow to give such a conclusion. Conclusions are the same as from the ICC. The difference is the meaning, here the GBM is not better because it makes more differentiation between policyholders but because the estimations are nearer of the observations than for the other models.

Finally, the AUC does not allow to class so clearly the different models. The intercept only GLM is clearly the worse as it does not make any difference between the different classes. Everyone has the same expected claim frequency. The second GLM has the following better performances but as far as the GAM is concerned, results are similar to the one given by the GBM or the true results. Differentiation between these three models/values cannot be done based on these results. Even the true scenario is not classified as the better one. This measure is much more volatile than the two other ones and does not allow to rank the models only by looking at its results. Also, it is not really a measure built to compare this kind of models.

Nevertheless, looking at the average results on the different folds represented on Figure 3.2, AUC ranking is the same as for the other measures, even if the volatility observed does not lead to a reliable trust in the result given by this measure. Note that results for the averaged AUC are the same for the true model and the GBM model, this is why we only see the pink point on the graph and not the brown one. This observation means that the true model is considered as good performing as the GBM but the GBM do not give as good results as the true model as it is not equal, even if it is really closed. This measure should be considered very carefully.

From all these observations, the GBM model can be set as the one giving the best performance. The estimations of this model should be the closest to the true values.

4 Second example : A real data set

4.1 Data set

The second example is based on the *freMTPLfreq* data set coming from the *CASdatasets* in R. This data is composed of 413.169 french motor third-part liability policies. We have the information on the number of claims occurred on a given exposure period. The different variables available to explain this claim frequency are the following :

- Power : The power of the car (categorized);

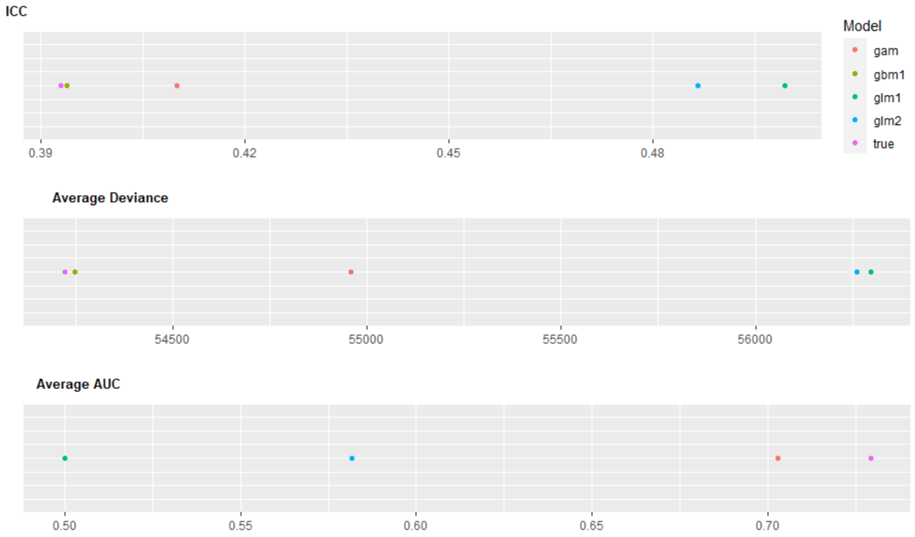

Figure 3.2: Mean values of the ICC, deviance and AUC on the simulated data.

- CarAge : The vehicle age;
- DriverAge : The driver age;
- Brand : The car brand (divided in different groups);
- Gas : The car gas (Diesel / Regular);
- Region : The policy region ;
- Density : The density of inhabitants in the city of the driver.

4.2 Models fitting

As for the simulated data, four types of models are fitted on each training set.

1. A GLM without any variables (a simple mean of the response).
2. A GLM including all the principal effects : removing the non significant variables, only the power (grouping the modalities), CarAge, DriverAge, Gas ans Density are included in this model.
3. A GAM including all the principal effects : all the variables remain in the model. All continuous variables are smoothed.
4. A GBM including all the variables in order to capt interactions included in the data.

The main difference is that we do not have the real value of the data. We also do not know the real structure of the data.

4.3 Comparison of the models

Figure 4.1 represents the results for the ICC, deviance and AUC on these data.

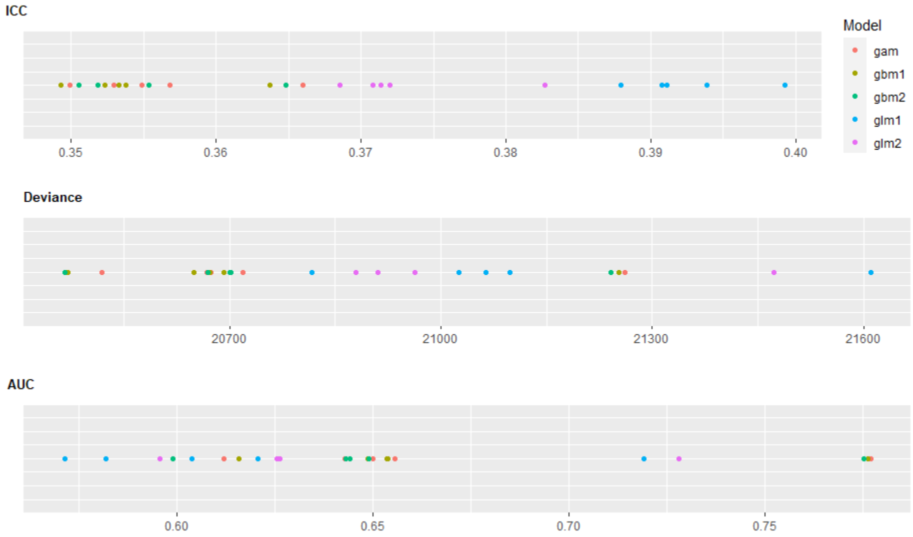

Figure 4.1: Values of the ICC, deviance and AUC for the freMTPLfreq data.

The observations based on the different measures are not so obvious as on the other data. As the data is not clearly structured, the models have more difficulties to capture correctly all impacts of the variables as these impacts are more volatile. There is already a certain variability in the ICC. Only the differentiation between both GLM's and the other models is clear. The GAM and GBM have similar performance based on the ICC. In fact, as all variables are used in both models, the discrimination of both models is similar. The performance would have been more different if the variables in both models were not exactly the same.

Looking at the deviance, the conclusion is not very different. No model appears to have clearly a better performance than all the others. This means that the distances between estimations and observations are not so different from one model to another. Similarly, the AUC is again very volatile. This means that for a particular model, the classes are on average well determined on some sets and not on other.

As far as the means are concerned, results can be found on Figure 4.2. The deviance and the AUC give different rankings. This does not allow to determine unanimously which model performs the best compared to the other. Nevertheless, as each indicator measures the performance of the model according to different criteria, both results can be considered depending on which aspect of the model we want to evaluate.

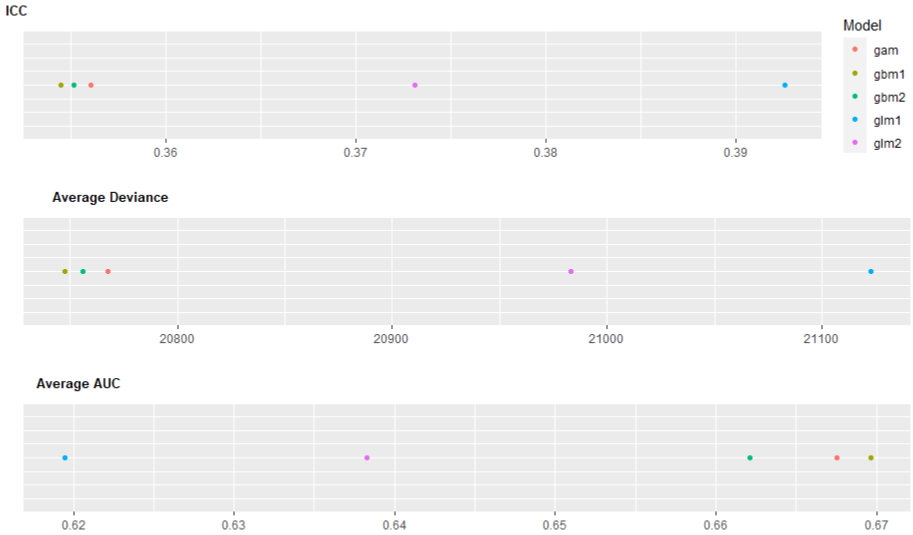

Figure 4.2: Mean values of the ICC, deviance and AUC for the freMTPLfreq data.

The first GBM1 seems to be the best model on these data, even if the analysis is not as obvious as on the other data.

5 Conclusion

Model's performance measures must be used cautiously. Each measure gives different information. In fact, they do not focus on the same features of the model: ICC highlights discrimination of policyholders, deviance prefers the ones closer to the observation and the AUC focus on the way estimations are classified. You should choose the measure depending on the aim of your performance's evaluation. Given its high variability on simple data, the AUC should be avoided to compute performance of Poisson models. ICC on the other hand seems to be the less volatile of the three measures.

6 References

- Denuit M., Sznajder D., Trufin J.(2019). Model selection based on Lorenz and concentration curves, Gini indices and convex order. Insurance: Mathematics and Economics 89, 128–139.
- Hand D.J., Till R.J.(2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45, 171–186.

7 About the serie and the authors...

7.1 The FAQctuary's

The FAQctuary's are a series of educational papers dedicated to the insurance sector. Each issue addresses a specific actuarial topic, expressed as a question asked by market players. FAQctuary's are published by members of the Detralytics team and written in a clear and accessible language. The team combines academic expertise and business knowledge. Detralytics was founded to support companies in the advancement of actuarial science and the solving of the profession's future challenges. It is within the scope of this mission that we make our work available through our Detra Notes and FAQctuary's series.

7.2 Authors' biographies

Louise d'Oultremont

Louise is part of the Talent Accelerator Program (TAP) at Detralytics. Prior to joining Detralytics, Louise did an internship at Axa Belgium, where she was in the risk pricing team. She holds two Master's degree in Actuarial Science and Mathematics, both from UCLouvain. Her thesis focused on multi-state individual reserving and consisted in comparing an analytical method and a simulation method using an individual multi-state semi-Markovian model.

Michel Denuit

Michel is Scientific Director at Detralytics, as well as a Professor in Actuarial Science at the Université Catholique de Louvain. Michel has established an international career for some two decades and has promoted many technical projects in collaboration with different actuarial market participants. He has written and co-written various books and publications. A full list of his publications is available at : https://uclouvain.be/en/directories/michel.denuit .

Julien Trufin

Julien is Scientific Director at Detralytics, as well as a Professor in Actuarial Science at the department of mathematics of the Université Libre de Bruxelles. Julien is a qualified actuary ofthe Instute of Actuaries in Belgium (IA|BE) and has experience as a consultant, as well as a compelling academic background developed in prominent universities such as Université Laval (Canada), UCL and ULB (Belgium).He has written and co-written various books and publications. A full list of his publications is available at : http://homepages.ulb.ac.be/~jtrufin/.

Detralytics

Expertise and innovation at the service of your future

www.detralytics.com - info@detralytics.eu