Validation of Fintech risk management models

Arianna Agosto and Dave Remue

UNIPV and B-HIVE

1st FIN-TECH Workshop
1 February 2019

Introduction

Dissemination and Validation

- Both dissemination and validation are enacted and made transparent within the project platform of WP6, which includes a communication infrastructure, a dedicated project website, and a Slack social network channel, aimed at engaging all stakeholders, existing and potential.
- The platform is continuosly fed with feedback participation and evaluation to all projects' events: workshops, training sessions, training with coding sessions.
- A key part of WP7 concerns the validation of the fintech risk management models built in the project by European banks, through the European Bank Federation (EBF).
- The validation is performed in line with the risk management regulations valid for the European banks, adapted to the fintech context. The next slides introduce some technical guidelines.

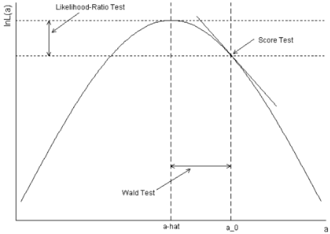
Validation of risk management models

Model evaluation

- The traditional way to choose a model is statistical testing: a sequence of pairwise model comparisons, on the basis of a test statistics whose distribution is known.

```
Likelihood ratio test

Model 1: admit ~ gre + gpa + rank
Model 2: admit ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   6 -145.75
2   1 -165.77 -5 40.048   1.46e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```



- A complete ordering of models can be obtained using scoring likelihood-based methods such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion):

$$AIC = -2logL(\hat{\theta}; x_1 \ldots\ldots, x_n) + 2q$$
$$BIC = -2logL(\hat{\theta}; x_1 \ldots\ldots, x_n) + qlog(n)$$

Predictive accuracy criteria

- Neither statistical tests nor scoring methods are generally applicable to machine learning models, which do not necessarily have an underlying probabilistic model.
- A different approach measures predictive accuracy assuming that one part of the data is not observed, but should be predicted.
- The predicted values can be compared with the observed ones, to measure predictive accuracy.
- The problem with this approach is that accuracy measures differ depending on the type of reponse to be predicted.

Predictive accuracy for binary variables: the confusion matrix

- A confusion matrix contains information about actual and predicted classifications.

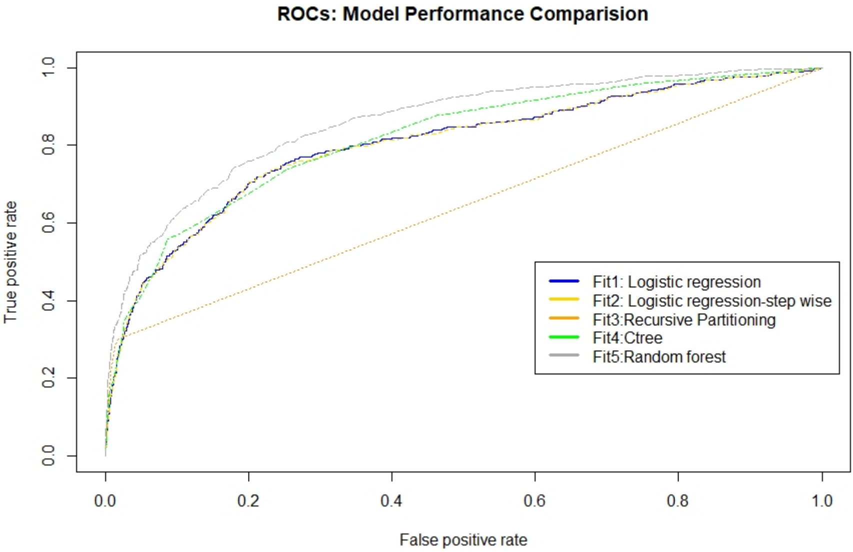
	Predicted: 0	Predicted: 1
Actual: 0	TN (True negative)	FP (False positive)
Actual: 1	FN (False negative)	TP (True positive)

- Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$
- Sensitivity $= \frac{TP}{TP+FN}$
- Specificity $= \frac{TN}{TN+FP}$

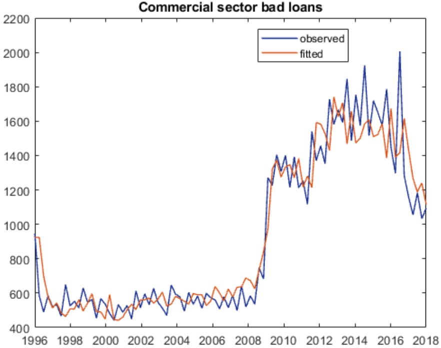
Predictive accuracy for binary variables: the ROC curve - I

- A widely used predictive accuracy measure for binary responses is the Receiver Operating Characteristics (ROC curve).
- The ROC curve displays the relationship between the sensitivity (on the y-axis) and the complement of the specificity (1- specificity, on the x-axis), across a series of predetermined cut-off points.
- The Area Under the ROC curve (AUROC) is the area between the ROC curve and the bisector line, and summaries predictive accuracy into a single statistics.
- The ideal curve coincides with the y-axis between 0 and 1, and the AUROC, in this case, is equal to $1/2$.

Predictive accuracy: the binary variables: the ROC curve - II



Predictive accuracy: continuous case



- When the response variable is continuous, a widely used measure to assess how far the predicted values are from the observed ones is the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{n}}$$

x_i actual values

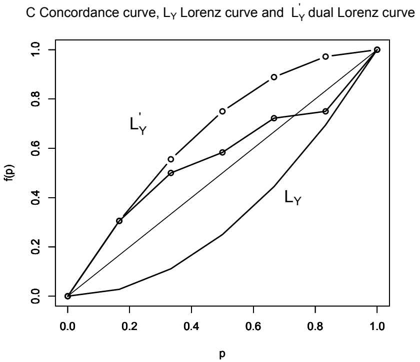
$\hat{x}_i$ fitted values

Predictive accuracy: general case I

- When the response variables are both binary and continuous there is not a unique measure. Giudici and Raffinetti (2019) have worked out one possible solution.
- The proposal is based on the concordance curve, obtained by ordering the Y original values according to the ranks of the corresponding $\hat{Y}$ estimated values.
- Let Y be a (binary or continuous) response variable and let $X_1, \ldots, X_p$ be a set of p explanatory variables. Suppose to apply a model such that $\hat{y} = f(\boldsymbol{X})$.

Predictive accuracy: general case II

- The model predictive accuracy is assessed by measuring the distance between the set of points lying on the C concordance curve $(i/n, (1/(n\mu))\sum_{j=1}^{i} y_{\hat{r}_j})$ and the set of points lying on the bisector curve $(i/n, i/n)$: if the C curve overlaps with the bisector curve, the model has no predictive capability.



Predictive accuracy: general case III

- The proposed predictive accuracy measure, called *RG* (Rank Graduation) index, is defined as:

$$RG = \sum_{i=1}^{n} \frac{\left\{(1/(n\mu))\sum_{j=1}^{i} y_{\hat{r}_j} - i/n\right\}^2}{i/n} = \sum_{i=1}^{n} \frac{\left\{C(y_{\hat{r}_i}) - i/n\right\}^2}{i/n},$$

  where $\mu = (i/n)\sum_{i=1}^{n} y_{\hat{r}_i}$ and $C(y_{\hat{r}_j}) = \frac{\sum_{j=1}^{i} y_{\hat{r}_j}}{\sum_{i=1}^{n} y_{r_i}}$ represents the cumulative values of the response variable.

- The *RG* index takes values between 0 and RG_{max}, which is obtained when the predicted ranks order the response variable values in full concordance (or full discordance) with the observed ranks.

Case study: predictive accuracy of credit scoring models

- Verifying whether the predictive performance of P2P credit scoring models can be improved by network models, using data from a European Credit Assessment Institution (ECAI).
- The available data include information on the status of the companies ([1 = Defaulted] and [0 = Active]).
- Three different models were taken into account: the BLR model, the TNBS (that uses all links) the WNBS (that uses only links between good and bad companies).
- Below the results. The maximum value that the *RG* index can attain is 49.895. The best model is easily recogines as WNBS.

Models	AUROC	*RG*	normalised *RG*
BLR	0.622	4.527	0.09
TNBS	0.836	14.839	0.29
WNBS	0.870	22.027	0.44