Co-financed by the Connecting Europe Facility of the European Union

xAIM

Accuracy of Artificial Intelligence methods

Trustworthy AI - Lecturer: Emanuela Raffinetti; Python instructor: Alex Gramegna
E-mail: emanuela.raffinetti@unipv.it; alex.gramegna01@universitadipavia.it

Premises

Aim: constructing predictive accuracy tools that can evaluate and monitor the quality of the forecasts.

State of the art:

- comparing statistical models within a model selection procedure, in which a model is chosen through a sequence of pairwise comparisons based on the comparison of the likelihoods (or of the posterior probabilities) of the models being compared.

  Problem: these criteria generally not applicable to models whose underlying probabilistic model is not specified.

- comparing the predicted and the actually observed cases, typically within cross-validation methods.

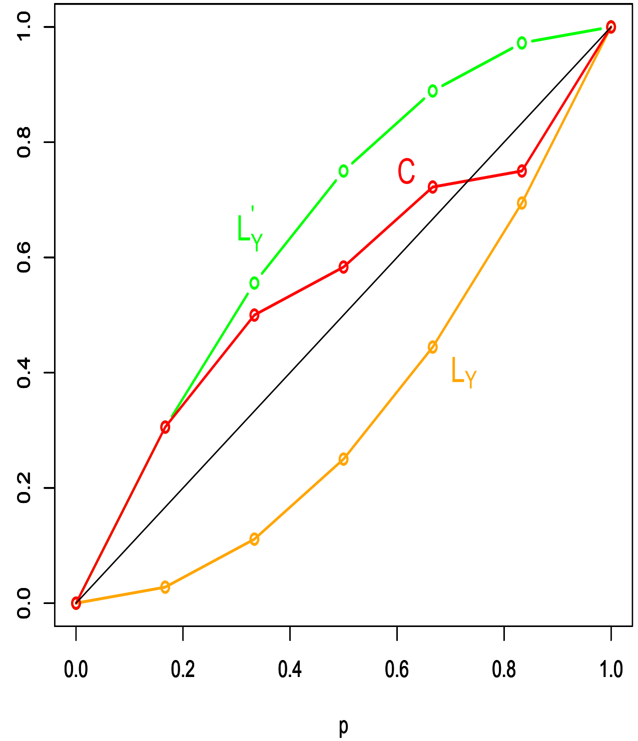
Proposal: a new measure based on the ranks which evaluates the concordance between the ranks of the predicted values and the ranks of the actual values of a series of observations to be forecast.

Co-financed by the Connecting Europe Facility of the European Union

Background

Let: Y be the target variable to be predicted; h be the number of predictors; $\hat{Y}$ be the vector of the predicted values generated by a ML model.

- Build the Y Lorenz curve (L_Y) by re-ordering the Y values in non-decreasing sense, whose coordinates are $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where r_i and $\bar{y}$ indicate the (non-decreasing) ranks of Y and the Y mean value, respectively.
- Build the Y dual Lorenz curve (L'_Y) by re-ordering the Y values in a non-increasing sense, whose coordinates are $(i/n, \sum_{j=1}^{i} y_{d_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where d_i indicates the (non-increasing) ranks of Y.
- Build the concordance curve C by ordering the Y values with respect to the ranks of the predicted $\hat{Y}$ values, whose coordinates are $(i/n, \sum_{j=1}^{i} y_{\hat{r}_j}/(n\bar{y}))$, where $\hat{r}_i$ indicates the (non-decreasing) ranks of $\hat{Y}$.
- Consider the 45-degree line, whose coordinates are $(i/n, i/n)$.



Co-financed by the Connecting Europe Facility of the European Union

Model scenarios

We associate the C curve behavior with the main reference scenarios that occur in model comparison.
It results that:

i) the best case occurs when the ordering of the Y response variable values corresponds to the ordering of the predicted values, with the C curve perfectly overlapping the Lorenz curve L_Y;

ii) the worst case occurs when the ordering of the Y response variable values is in inverse correspondence with the ordering of the predicted values, with the C curve perfectly overlapping the dual Lorenz curve $L_Y^{'}$;

iii) in the random case, the C curve overlaps the 45-degree line;

iv) in the generic case, the C curve lies in the area between the Y response variable Lorenz curve, L_Y and its dual, $L_Y^{'}$. The distance between C and the 45-degree line measures how a model improves over random predictions.

Co-financed by the Connecting Europe Facility of the European Union

The C and ROC curves

In the case of a binary response variable, the C curve and the ROC curve have the following behavior:

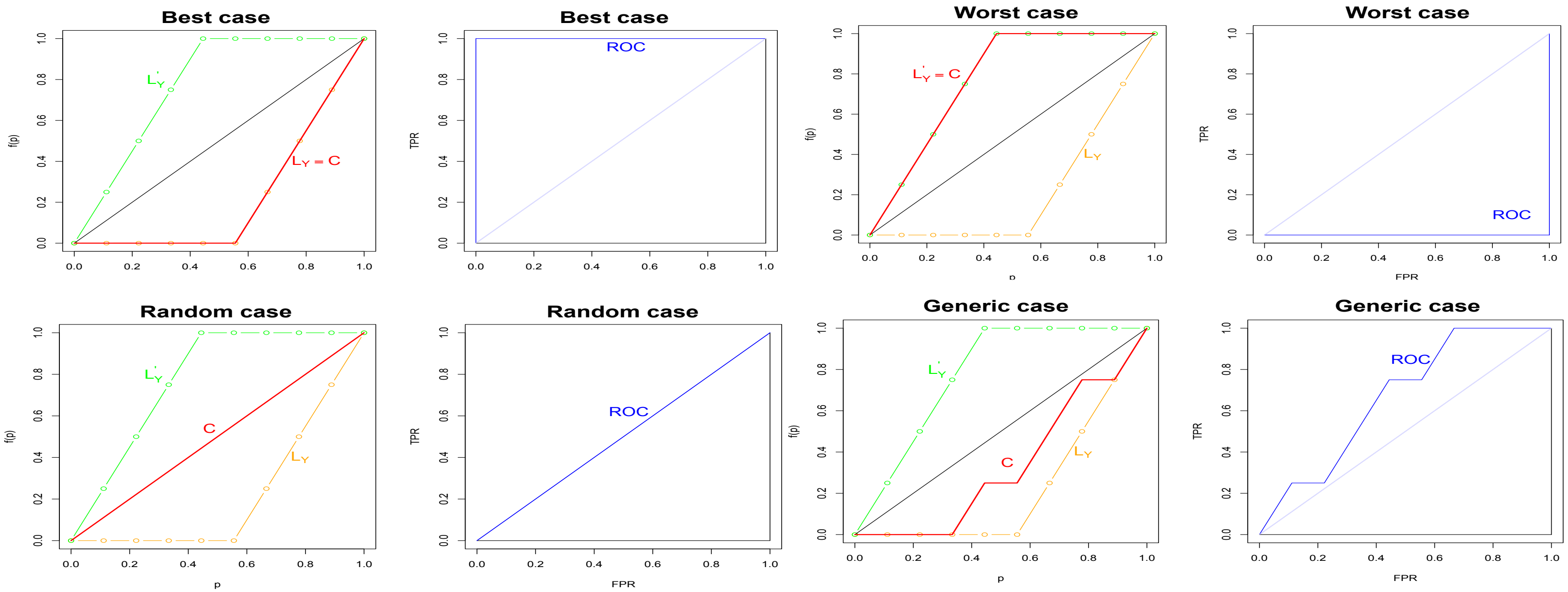


Figure: The concordance C curve and the ROC curve in the best, worst, random and generic cases

Co-financed by the Connecting Europe Facility of the European Union

Definition of the Rank Graduation Accuracy (RGA) measure

On the analogy between the ROC and the C curve, a summary measure for the C curve of a model can be derived. The resulting

measure, named Rank Graduation Accuracy (RGA) measure, is defined by the following expression:

$$RGA = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j} \right) \right\}}.$$

Remark

When tied predictions occur, it may be unclear how to order the observed values in the expression of RGA. In this case, we suggest to replace the observed response values corresponding to the predictions with their mean values.

Co-financed by the Connecting Europe Facility of the European Union

The RGA properties

Property 1 - Simplification

The RGA measure can be simplified as follows:

$$RGA = \frac{\sum_{i=1}^{n} iy_{\hat{r}_i} - \sum_{i=1}^{n} iy_{r_{n+1-i}}}{\sum_{i=1}^{n} iy_{r_i} - \sum_{i=1}^{n} iy_{r_{n+1-i}}}.$$

Property 2 - Normalisation

i) $0 < RGA < 1$ for an intermediate model; ii) $RGA = 1$ for the best model; iii) $RGA = 0$ for the worst model; iv) $RGA = 0.5$ for a random model.

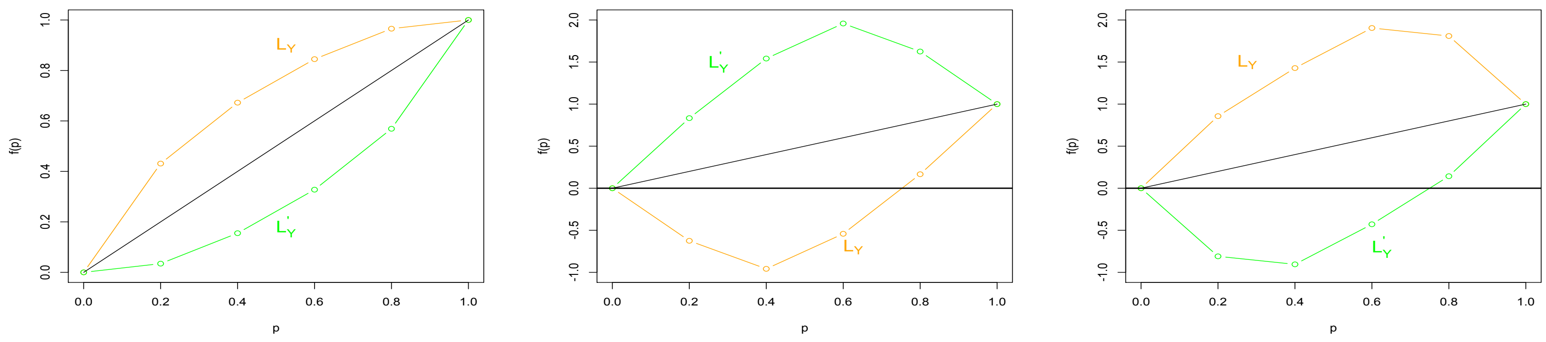
Property 3 - Invariance

RGA is invariant with respect to translations of Y, meaning that $RGA = RGA^k$, where RGA^k denotes the RGA measure computed on the transformed variable $Y^k = Y + k$, where k is a constant such that $k \in \mathbb{R}$.

Property 4 - Equivalence between RGA and $AUROC$

Property 5 - Equivalence between RGA and the Wilcoxon-Mann-Whitney statistic

The case of a target variable with negative values



(a) Negative values and $\bar{y} < 0$ (b) Mixed values and $\bar{y} > 0$ (c) Mixed values and $\bar{y} < 0$

Note that, in case a), the Lorenz and dual Lorenz curves are reversed, but the Lorenz curves remain inside the unit square, satisfying Property 2.

Differently, in cases b) and c), the Lorenz curve extends below $y = 0$ and the dual Lorenz curve extends above $y = 1$. In these cases, to fulfill Property 2, we can subtract from the Y variable its minimum negative value. This translation leaves the measure invariant (according to Property 3) and can thus be exploited to satisfy Property 2.

A test for the RGA measure - I

Proposition

In the case of a continuous response variable, the RGA index can be translated in terms of covariance operators. It can be shown that:

$$RGA = \frac{cov(Y_{r(\hat{Y})}, F(Y)) - cov(Y, 1 - F(Y))}{cov(Y, F(Y)) - cov(Y, 1 - F(Y))}, \tag{3}$$

where F is the cumulative continuous distribution functions of Y and $1 - F$ is the retro-cumulative distribution function of Y.

Given a more complex model Mod_1 and a simpler model Mod_2, their predictive accuracy can be compared by setting the following hypotheses:

$$H_0 : \psi(Y, \hat{Y}_{Mod_1}) = \psi(Y, \hat{Y}_{Mod_2}) \quad \text{vs} \quad H_1 : \psi(Y, \hat{Y}_{Mod_1}) \neq \psi(Y, \hat{Y}_{Mod_2})$$

where $\psi(Y, \hat{Y}_{Mod_1}) = \frac{cov(Y_{r(\hat{Y}_{Mod_1})}, F(Y))}{cov(Y, F(Y))}$ and
$\psi(Y, \hat{Y}_{Mod_2}) = \frac{cov(Y_{r(\hat{Y}_{Mod_2})}, F(Y))}{cov(Y, F(Y))}$.

A test for the RGA measure - II

By denoting with $\hat{\delta} = \hat{\psi}(Y, \hat{Y}_{Mod_1}) - \hat{\psi}(Y, \hat{Y}_{Mod_2})$, the test statistics for testing the null hypothesis

$$H_0 : \psi(Y, \hat{Y}_{Mod_1}) = \psi(Y, \hat{Y}_{Mod_2})$$

becomes:

$$Z = \frac{\hat{\delta}}{\sqrt{\widehat{Var(\hat{\delta})}}} \to N(0, 1),$$

where the estimated variance $\widehat{Var(\hat{\delta})} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\delta}_{(-i)} - \bar{\delta})^2$, $\hat{\delta}_{(-i)}$ are the values of $\hat{\delta}$ by omitting one pair $(Y, \hat{Y})$ at a time and $\bar{\delta}$ is the average of the values $\hat{\delta}_{(-i)}$, for $i = 1, \ldots, n$.

For a fixed significance level α, the rejection region corresponds to the values of $|Z| \geq z_{\alpha/2}$.

Robustness of the RGA - I

It is important that the measurement of predictive accuracy is not affected by outlying observations, which may bias model comparison.

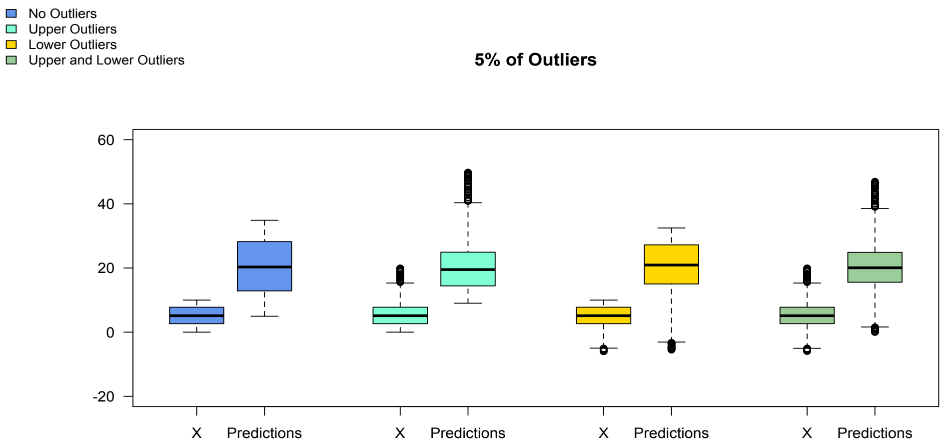
Without loss of generality, let X and Z be two independent continuous random variables with $X \sim U(0,10)$ and $Z \sim N(0,1)$ and let $Y = 5 + 3X + Z$, from which we can simulate a set of observations.

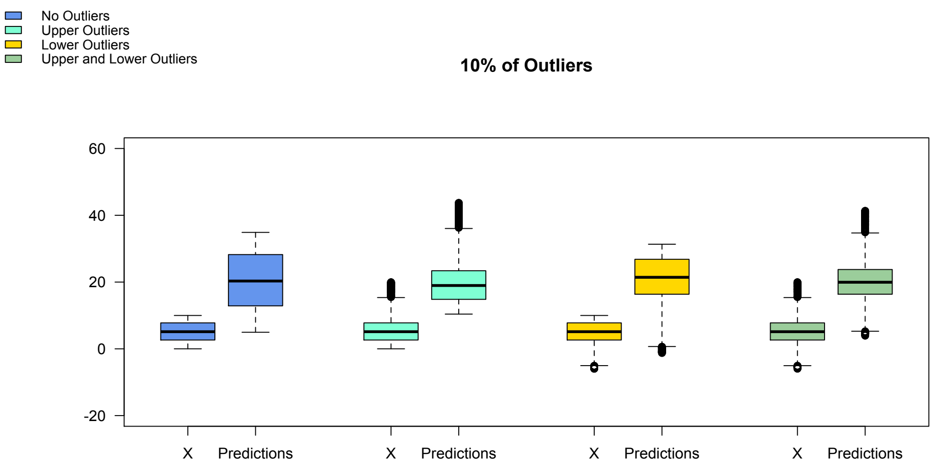
To assess robustness, we replace the obtained left and right tail observations of the X distribution with outliers in the tails of the distribution. Without loss of generality, we consider six alternative replacements, as follows:

a) the observations greater than the 95% percentile are replaced by observations sampled from a $U(15,20)$ distribution;
b) the observations lower than the 5% percentile are replaced by observations sampled from a $U(-10,-5)$ distribution;
c) the observations greater than the 95% percentile are replaced by observations sampled from a $U(15,20)$ distribution and the observations lower than the 5% percentile are replaced by observations sampled from a $U(-10,-5)$ distribution;
d) the observations greater than the 90% percentile are replaced by observations sampled from a $U(15,20)$ distribution;
e) the observations lower than the 10% percentile are replaced by observations sampled from a $U(-10,-5)$ distribution;
f) the observations greater than the 90% percentile are replaced by observations sampled from a $U(15,20)$ distribution and the observations lower than the 10% percentile are replaced by observations sampled from a $U(-10,-5)$ distribution.

Co-financed by the Connecting Europe Facility of the European Union

Robustness of the RGA - II





Predictive accuracy measures	***RMSE***	***RGA***
Without outliers	***0.981***	***0.997***
Scenario a) (upper 5% outliers)	3.769	0.997
Scenario b) (lower 5% outliers)	3.695	0.997
Scenario c) (upper and lower 5% outliers)	4.119	0.997
Scenario d) (upper 10% outliers)	4.163	0.996
Scenario e) (lower 10% outliers)	4.018	0.996
Scenario f) (upper and lower 10% outliers)	4.027	0.996

Co-financed by the Connecting Europe Facility of the European Union

Application to "Employee" data

"Employee" dataset

Data report information on: gender, age, educational degree, employment category, job time in months since hire, total work experience, minority classification, starting salary and current salary (in dollars).

Aim

Understanding whether salary growth is affected by personal characteristics.

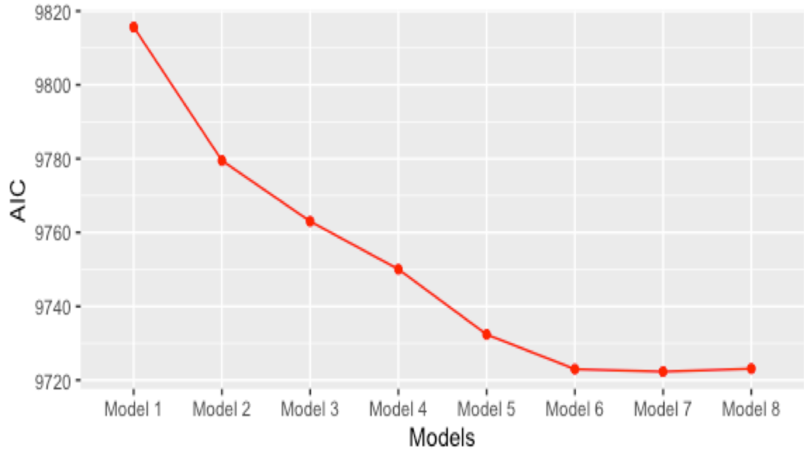
Procedure

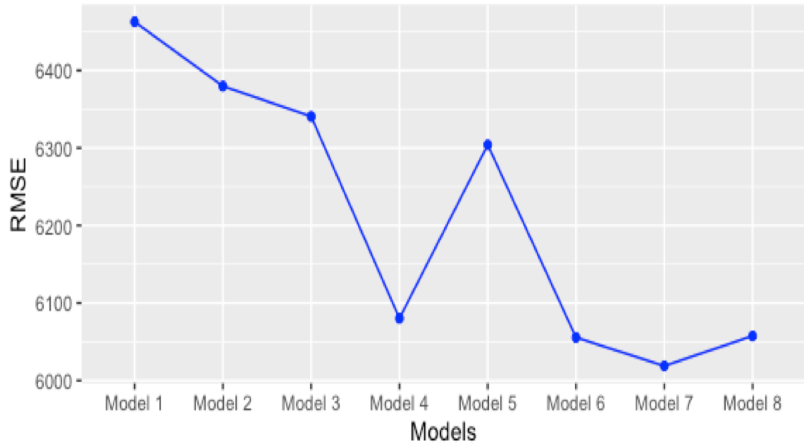
- Salary growth is considered as the response variable (measured either on a continuous scale and on a binary scale).
- Both linear and logistic regression models are considered.
- Stepwise model selection is applied to the data.
- For each possible model size (from 1 to 8), we compare all possible models by means of the AIC criterion.
- Dataset is split into a train dataset (including the 80% of all the observations) and a test dataset (including the remaining 20% of the observations).
- The *RMSE* and the *RGA* of each of the best 8 linear regression models, and the *BS* (Brier score) and the *RGA* of each of the best 8 logistic regression models are computed.

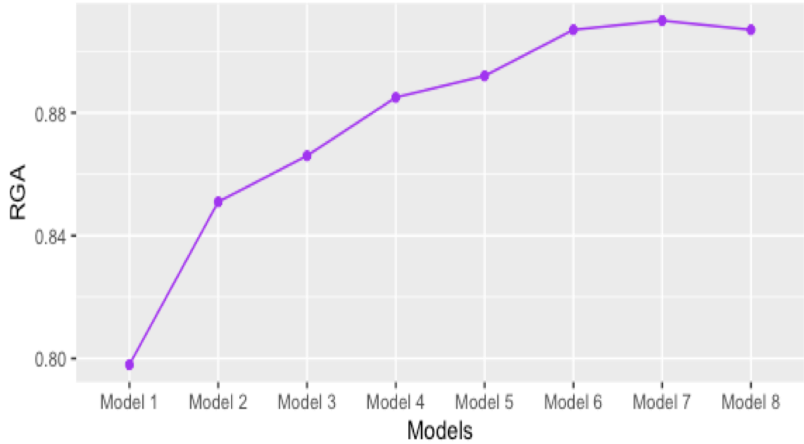
$$\boxed{AIC = \log \frac{1}{N} \sum_{i=1}^{N} e_i^2 + \frac{2p}{N}}$$

Co-financed by the Connecting Europe Facility of the European Union

Results from the linear regression model





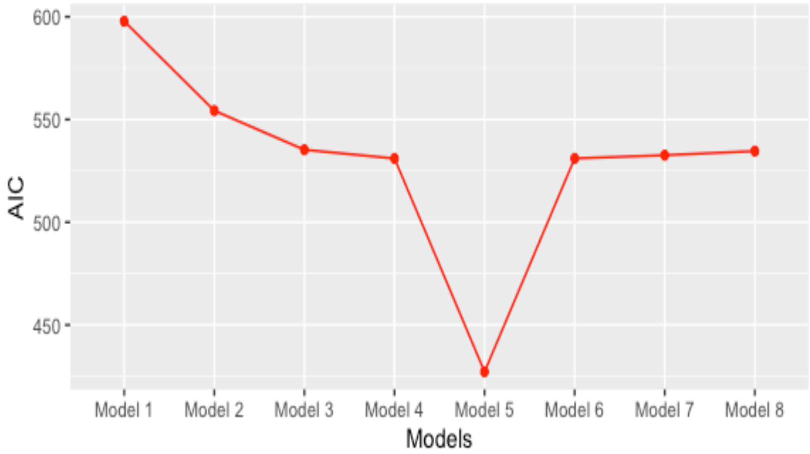


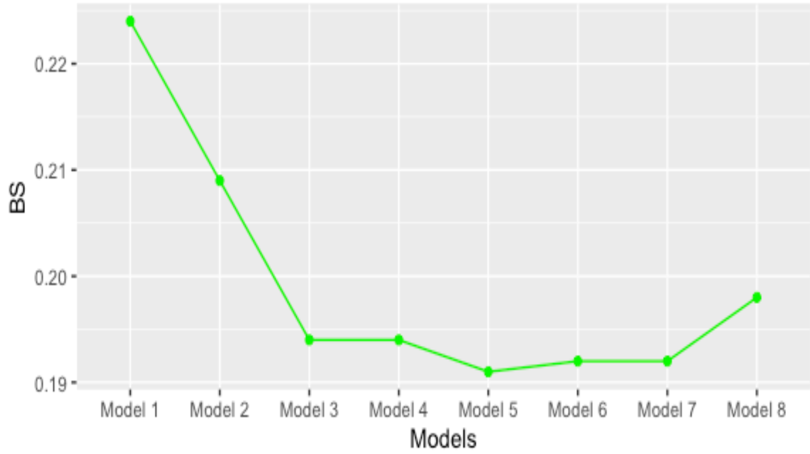
Target variable: current salary - starting salary

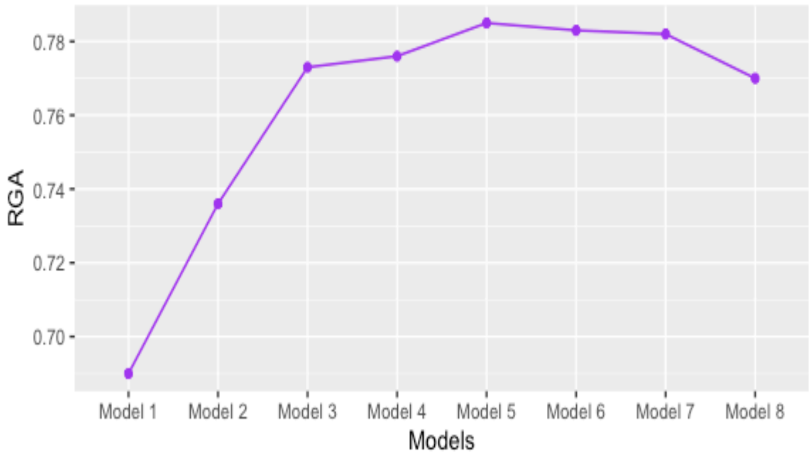
Model	Variables	*RMSE*	*RGA*	*AIC*
Model 1	Manager	6426.728	0.798	9815.657
Model 2	Manager, ed. degree	6379.797	0.851	9779.493
Model 3	Manager, ed. degree, job time	6340.604	0.866	9763.017
Model 4	Manager, job time, age, male	6080.111	0.885	9750.053
Model 5	Manager, ed. degree, job time, age, custodial	6304.019	0.892	9732.383
Model 6	Manager, ed. degree, job time, male, custodial, tot. job time	6055.528	0.907	9722.994
Model 7	Manager, ed. degree, job time, male, custodial, tot. job time, no minority	6018.835	0.910	9722.324
Model 8	Manager, ed. degree, job time, male, custodial, tot. job time, no minority, age	6057.536	0.907	9723.128

Co-financed by the Connecting Europe Facility of the European Union

Results from the logistic regression model







Target variable: "doubling" of the starting salary

Model	Variables	*BS*	*RGA*	*AIC*
Model 1	Age	0.224	0.690	597.853
Model 2	Age, job time	0.209	0.736	554.320
Model 3	Age, job time, custodial, manager	0.194	0.773	535.267
Model 4	Age, job time, custodial, manager	0.194	0.776	531.037
Model 5	Age, job time, custodial, manager, gender	0.191	0.785	427.186
Model 6	Age, job time, custodial, manager, gender, tot. job time	0.192	0,783	531.047
Model 7	Age, job time, custodial, manager, gender, tot. job time, no minority	0.192	0.782	532.616
Model 8	Age, job time, custodial, manager, gender, tot. job time, no minority, ed. degree	0.198	0.770	534.615

Co-financed by the Connecting Europe Facility of the European Union

Application to Bitcoin data

Bitcoin price data

Data report information on several time series of financial prices.

The daily bitcoin prices in the Coinbase exchange, from 18 May 2016 to 30 April 2018, is used as our response variable.

The daily prices of classical assets, such as oil, gold and SP500, together with the exchange rates (dollar/yuan and dollar/euro), are considered as candidate predictors.

Aim

Comparing the model selection performance of the *RGA* against that of the RMSE.

Procedure

- Linear regression model is applied.
- Stepwise model selection is applied to the data.
- For each possible model size (from 1 to 5), we compare all possible models by means of the AIC criterion.
- To predict bitcoin prices, we follow a rolling windows procedure.

Co-financed by the Connecting Europe Facility of the European Union

The rolling procedure

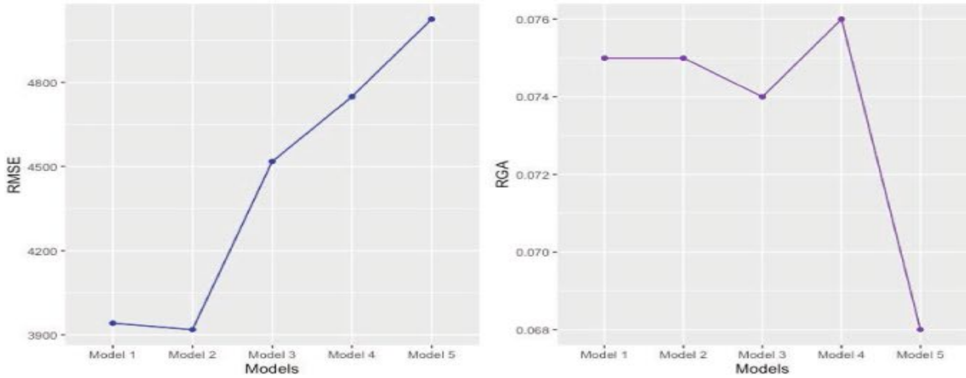
The rolling procedure can be summarised as follows

- models are trained using only data between 1st January 2017 and 31st December 2017. Forecasts are derived for a time window (first window) that starts on January 1st, 2018 and ends at January 31st, 2018;
- models are trained using only data between 1st February 2017 and 31st January 2018. Forecasts are derived for a time window (second window) that starts on February 1st, 2018 and ends at February 28th, 2018;
- models are trained using only data between 1st March 2017 and 28th February 2018. Forecasts are derived for a time window (third window) that starts on March 1st, 2018 and ends at March 31st, 2018;
- models are trained using only data between 1st April 2017 and 31st March 2018. Forecasts are derived for a time window (fourth window) that starts on April 1st, 2018 and ends at April 30th, 2018.

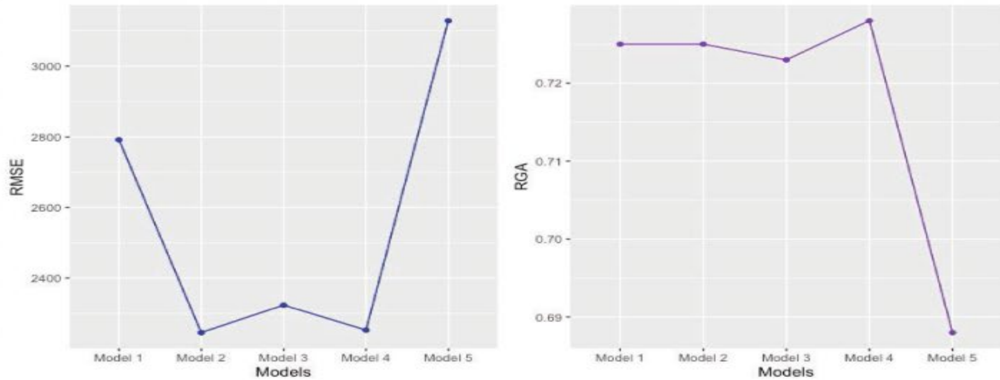
Results - continuous target variable

Model	**Variables**
Model 1	sp500
Model 2	sp500, exchange rate dollar/yuan
Model 3	sp500, gold, oil
Model 4	sp500, gold, oil, exchange rate dollar/euro
Model 5	sp500, gold, oil, exchange rate dollar/ euro, exchange rate dollar/yuan

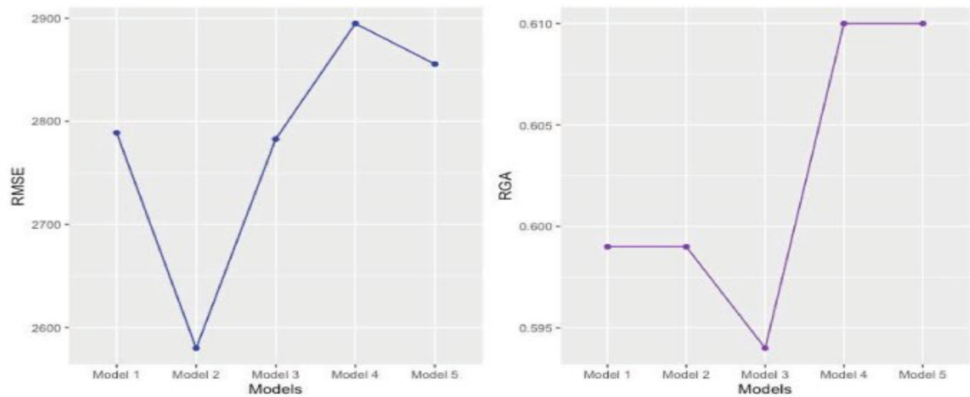
FIRST WINDOW



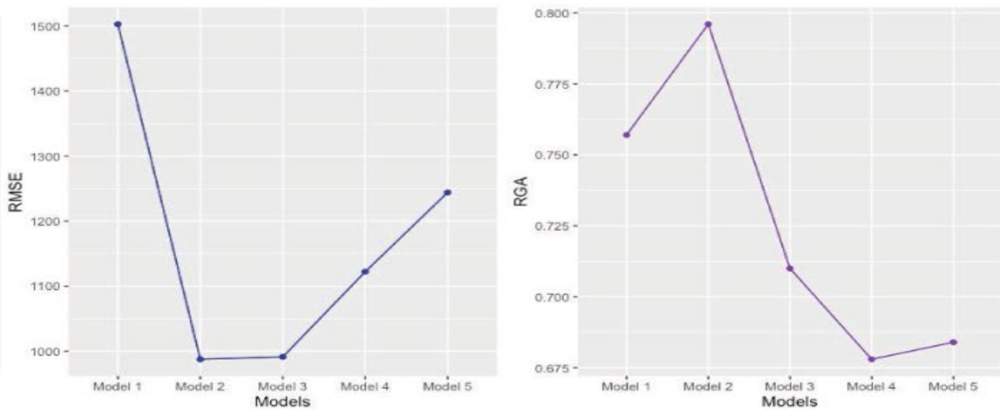
SECOND WINDOW



THIRD WINDOW



FOURTH WINDOW



Co-financed by the Connecting Europe Facility of the European Union

Application to ordered Bitcoin data

To better strength the *RGA* role to the evaluation of predictive accuracy, we suppose that bitcoin prices are only available on an ordinal scale based on five categories encoded by 1, 2, 3, 4, and 5.

Aim

Comparing the model selection performance of the *RGA* against that of the MSE.

Procedure

- Rank regression model is applied.
- Stepwise model selection is applied to the data.
- For each possible model size (from 1 to 5), we compare all possible models by means of the AIC criterion.
- The model is trained on data referred to year 2017 and tested on data referred to year 2018.

Let Y be a response variable, expressed through h ordered categories

Procedure:

- assign a rank $r_1 = 1$ to the smallest ordered category of Y;
- assign rank $(r_{j-1} + n_{j-1})$ to the following ordered categories, where n_{j-1} is the absolute frequency associated with the $(j-1)$-*th* category with $j = 2, \ldots, h$;
- the phenomenon described by the Y variable can be re-formulated in terms of its ranks R, where:

$$R = \left\{ \underbrace{r_1, \ldots, r_1}_{n_1}, \underbrace{r_2, \ldots, r_2}_{n_2}, \ldots, \underbrace{r_h, \ldots, r_h}_{n_h} \right\},$$

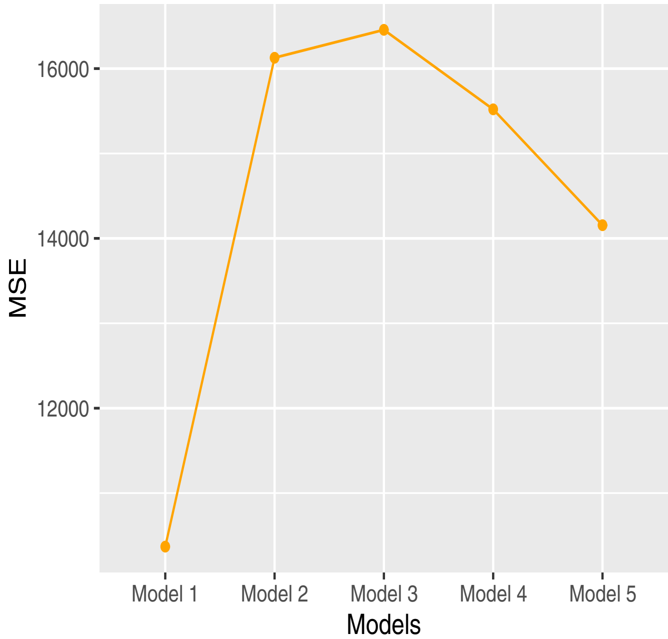
  with $r_1 = 1$, $r_2 = r_1 + n_1$ and $r_h = r_{h-1} + n_{h-1}$.
- Given K explanatory variables, a regression model for R can be specified as:

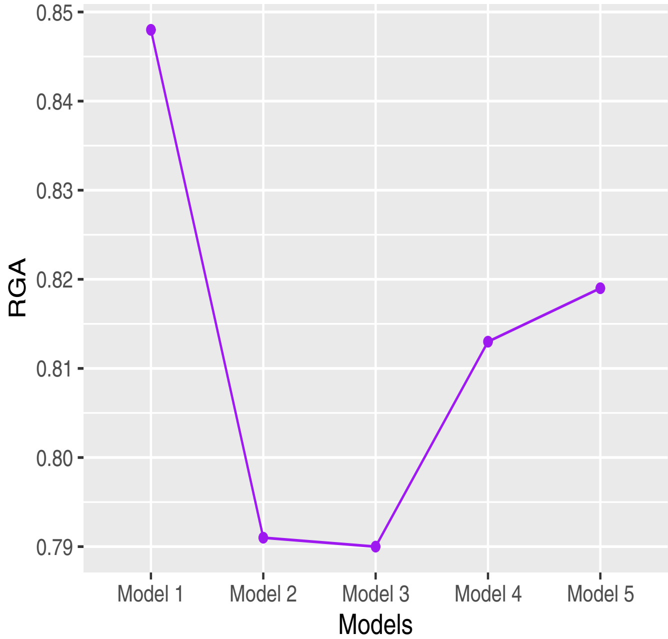
$$\hat{R} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_K X_K,$$

  whose unknown parameters can be estimated by the OLS method.

Results - ordinal target variable

Model	Variables
Model 1	sp500
Model 2	sp500, exchange rate dollar/yuan
Model 3	sp500, exchange rate dollar/yuan, oil
Model 4	sp500, exchange rate dollar/yuan, oil, gol(
Model 5	sp500, exchange rate dollar/yuan, oil, gold, exchange rate dollar/euro





Co-financed by the Connecting Europe Facility of the European Union

Application to binarised Bitcoin data

Beside the issue of the bitcoin price prediction, also the forecast of the returns derived from cryptocurrencies becomes a crucial topic, especially for those investors who are interested in measuring the potential gains or losses. In order to cover this perspective, we first transform the bitcoin prices into returns and then we proceed to their binarisation by assigning value equal to 1 to the negative returns (losses) and value equal to 0 to the positive returns (gains).

Aim

Comparing the model selection performance of the *RGA* against that of the AUROC.

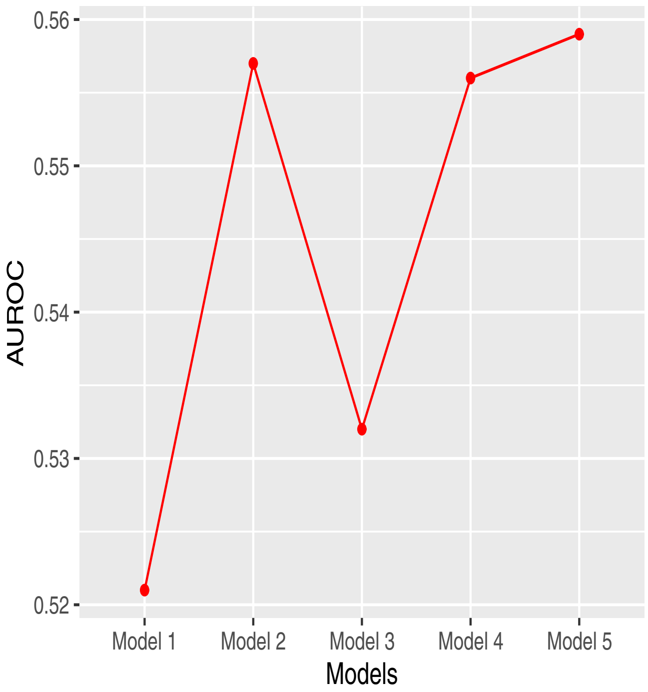
Procedure

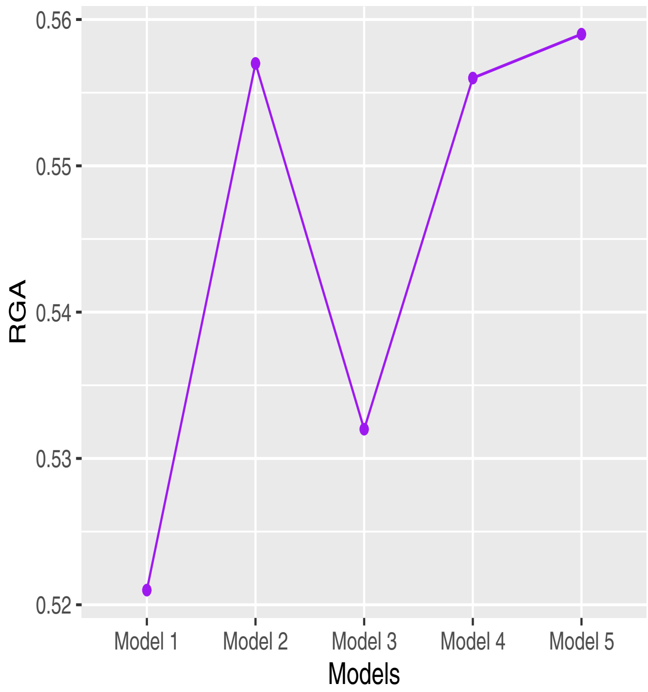
- Logistic regression model is applied.
- Stepwise model selection is applied to the data.
- For each possible model size (from 1 to 5), we compare all possible models by means of the AIC criterion.
- The model is trained on data referred to year 2017 and tested on data referred to year 2018.

Co-financed by the Connecting Europe Facility of the European Union

Results - binarised target variable

Model	Variables
Model 1	Gold
Model 2	Gold, exchange rate dollar/euro
Model 3	Gold, exchange rate dollar/euro, oil
Model 4	Gold, exchange rate dollar/euro, oil, sp500
Model 5	Gold, exchange rate dollar/euro, oil, sp500, exchange rate dollar/yuan





Co-financed by the Connecting Europe Facility of the European Union

Reference

- Raffinetti E.: A Rank Graduation Accuracy measure to mitigate Artificial Intelligence risks, Quality & Quantity (2023) and the references therein (available at https://link.springer.com/article/10.1007/s11135-023-01613-y)