REJECT INFERENCE TECHNIQUES AND SEMI SUPERVISED METHODS IMPLEMENTED IN THE PROCESS FOR GRANTING CREDIT

Asma GUIZANI
Computational Mathematics Laboratory-Tunisia
Besma SOUISSI
Computational Mathematics Laboratory-Tunisia
Salwa BEN AMMOU
Faculty of Economic Sciences and Management of Sousse
Gilbert SAPORTA
Laboratory Cédric – CNAM, Paris

Outline

1- Introduction for the problem of reject inference
2- Seven methods for reintegration of reject applicants
   a- Simple augmentation
   b- Re-weighting
   c- Iterative Reclassification
   d- Parceling
   e- Mixed Classification
   f- AdaBoost
   g- Gentle AdaBoost
3- Application
4- Results
5- Conclusion

Introduction for the problem of reject inference

- Credit scoring is a fundamental tool of risk prediction based on the characteristics of the loan applicant.
- The use of different statistical techniques to build a score model.
- Assign for each applicant a score.

Introduction for the problem of reject inference

- The data set used is based only on accepted applicant whose the predicted variable is known.
- The probability of default for refused applicants is not estimeted.
- The results of the score model are biased because estimations are done on a non-representative data set(selection bias).
- Solution : consider the refused applicants in the initial sample.

Simple augmentation

Step 1:

- build a score model for only the accepted applicants (labeled on good or bad payers).

Step 2 :

- The score model established is applied on the refused applicants to determine their probability of default
- Assigning refused applicants to their corresponding class (good or bad) depending on the probability of default.

Step 3 :

- Add the inferred goods and bads to the known good and bad to build a new score model using the new data set.

Re-weighting

Step 1:

- An accept/reject model is build to get the probability of acceptance for each applicant.

Step 2 :

- A good/bad model is build with the accepted applicants and adjusted using for each case a weight that is inversely proportional to the probability of acceptance.

Iterative Reclassification

Step 1:

- Build a known good/bad model to get the probability of default.
- The rejected applicants are assigned to classes (good or bad) based on the default probability established.

Step 2 :

- Combine inferred rejects and accepted applicants , and a new score model, based on this "augmented data set", is determined.
- Rescore reject applicants and reassign them to corresponding classes. Rebuild score models based on the new "augmented data set". The process will be repeated until stabilization of scores.

Parceling

Step 1:

- Build a model using known good and bad

Step 2 :

- The population (accepted and refused) is splitted into classes defined by score intervals and the default rate is determined within each score interval.
- The score model is applied to the rejects to assign them to each score interval respecting the assumption that the default rate is the same as accepted applicants.

Step 3 :

- The rejected of each interval are classified, randomly, into good and bad classes respecting the same proportion of good and bad for accepted applicants of each score interval.

Step 4 :

- The inferred rejects are combined with the known good and bad to rebuild a new known good/bad model.

Mixed classification

Step 1:

- A first classification with k-means to cluster the entire population (accepted and rejected) into "k" homogeneous groups.

Step 2 :

- A second clustering is established on the "k" previous clusters by a Hierarchical Classification applied to the centroids of the "k" groups in order to be reduce to “q” groups ($q<k$).

Step 3 :

- the rejects belonging to each of these classes will be assigned to the category of “good” or “bad” according to the most frequent category in their class.
- The inferred rejects are combined with the known good and bad to rebuild a new known good/bad model.

Preliminary step for Boosting algorithm

- learning data set of labeled data $S = \{(x_1, y_1), \cdots, (x_l, y_l)\};\ y_l \in \{-1,1\}$ and unlabeld data $\{x_u\}_{1 \leq u \leq U}$ with $N = U + L$.
- Assigning unlabeled data to pseudo-classes determined by Factorial Discriminant Analysis (FDA).
- Initial weight :

$$p_0 = \begin{cases} l/N & i \in L \\ u/N & i \in U \end{cases}$$

- Normalization of p_0 to obtain weights w_0 which $(\sum_{i=1}^{N} w_0 = 1)$

AdaBoost

- Training data:

$D = \{(x_1, y_1), \cdots, (x_N, y_N)\};\ y_N \in \{-1,1\}$

- For t = 1 …T do :

1. Fit the classifier $f_t(x)$ using weight on the training data
2. Compute the weight error :

$$\varepsilon_t = \sum_i w_t [y_i \neq \hat{y}_i],\ i = 1, \ldots, L+U$$

3. If $\varepsilon_t > 0{,}5$ stop the process else :
4. Compute : $\alpha_t = \frac{1}{2}\log\left({}^{(1-\varepsilon_t)}/_{\varepsilon_t}\right)$
5. Update the weight :

$$w_{t+1}(i) = \frac{w_t(i)\exp(-\alpha_t\, y_i f_t(x_i))}{z_t}$$

with z_t is a normalisation factor

- Output the classifier

$$sign[F(x)] = sign\left(\sum_{t=1}^{T} \alpha_t f_t(x)\right) \begin{cases} if\ F(x) > 0\ so\ y = 1 \\ if\ F(x) < 0\ so\ y = -1 \end{cases}$$

Gentle AdaBoost

- Training data:

$D = \{(x_1, y_1), \cdots, (x_N, y_N)\}$; $y_N \in \{-1,1\}$ and $F(x) = 0$

- If t = 1 …T so :
  1. Fit the regression function $f_t(x)$ by weighted least squares of y_i to x_i with weights w_i
  2. Updates: $F(x) \leftarrow F(x) + f_t(x)$
  3. Updates $w_i \leftarrow w_i \exp(-y_i f_t(x))$ and normalize

- Output the classifier

$$sign[F(x)] = sign\left(\sum_{t=1}^{T} f_t(x)\right) \begin{cases} if\ F(x) > 0\ so\ y = 1 \\ if\ F(x) < 0\ so\ y = -1 \end{cases}$$

Data

- A data bank of 9892 applicants of credit with 15 independent variables measured for each unit.
  - 7986 accepted applicants known reponse variable.
  - 1906 refused applicants unknown reponse variable.
- Source : external rating agency « Experian ».
- Applicants credit from « Financo » for the two years 2000 et 2001
- The reponse variable indicates whether or not an applicant is a good payer.

Process simulation

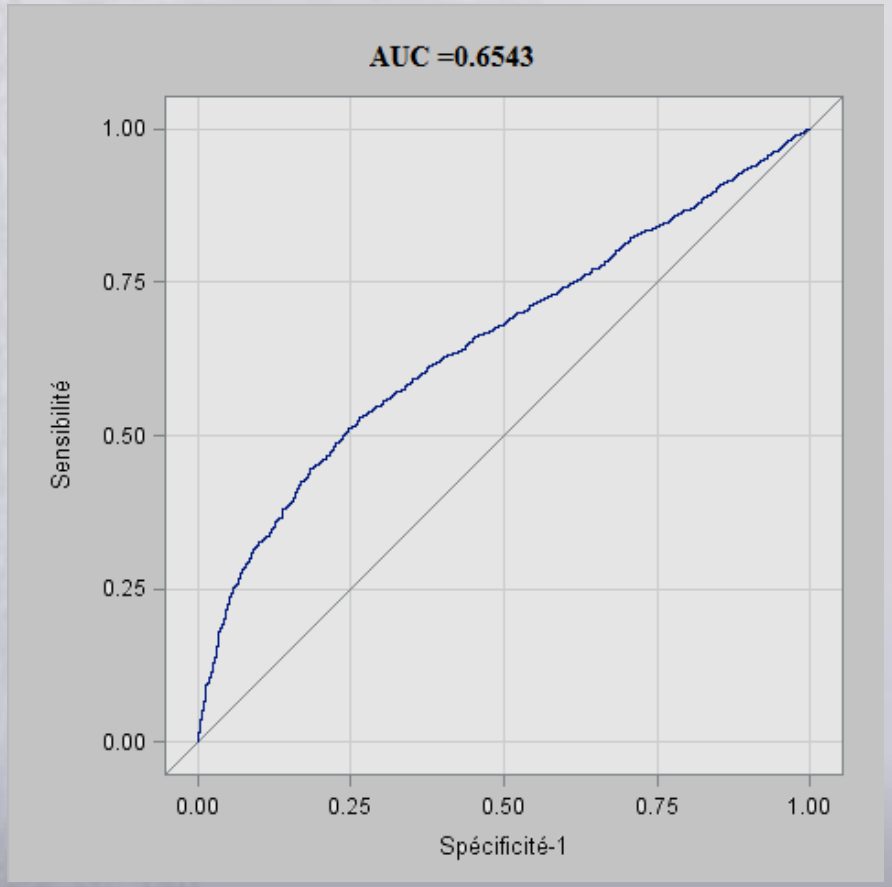
- Simulation of the rejection process on the 7986 accepted applicants.
- Create a uniform variable U_i for each observation .
- Compare U_i to the probability of default $Pr(_i)$ established by the discrimination between accepted and rejected.
  - If $U_i < Pr(_i)$: refused applicant ⟹ 1300
  - If $U_i > Pr(_i)$: accepted applicant ⟹ 6686
- Repeat the random process simulation 50 times : Stability comparaison between the different AUC index.

Application

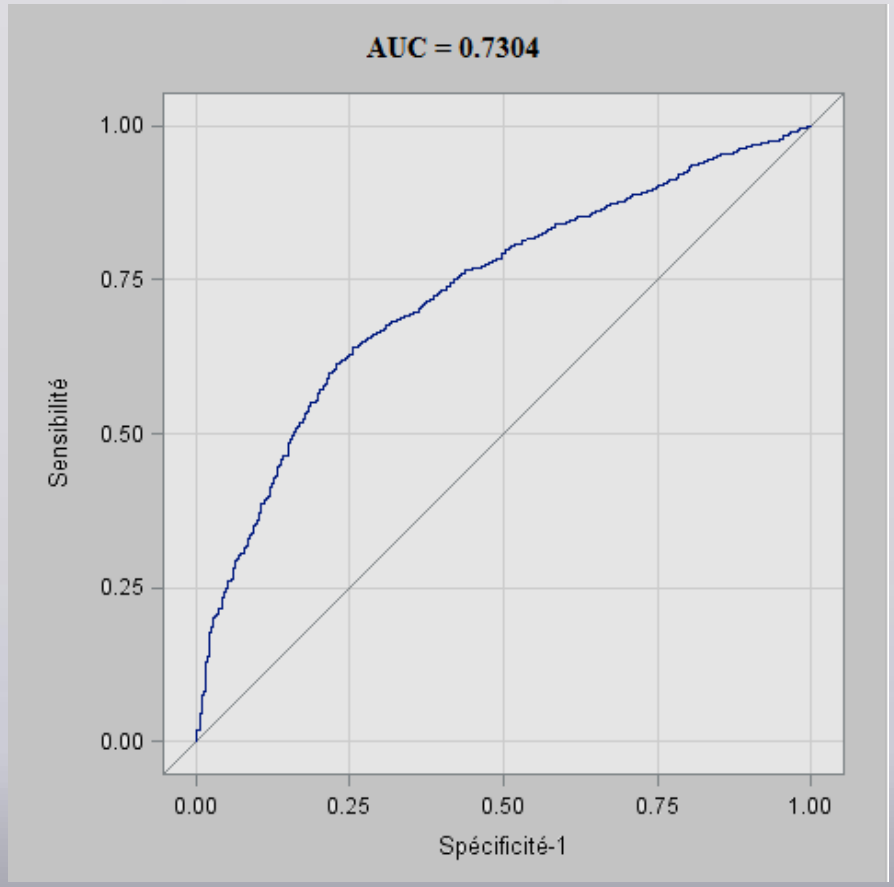
- Performance comparaison between score models with the ROC curve
- A synthesis of score performance for any threshold s
- Using s as a parameter, the ROC curve links the true positive fraction (good applicants classified as good) to the false positive fraction (bad applicants classified as good).
- AUC index :Widely used mesure of score performance.

Performance comparaisons

Parceling

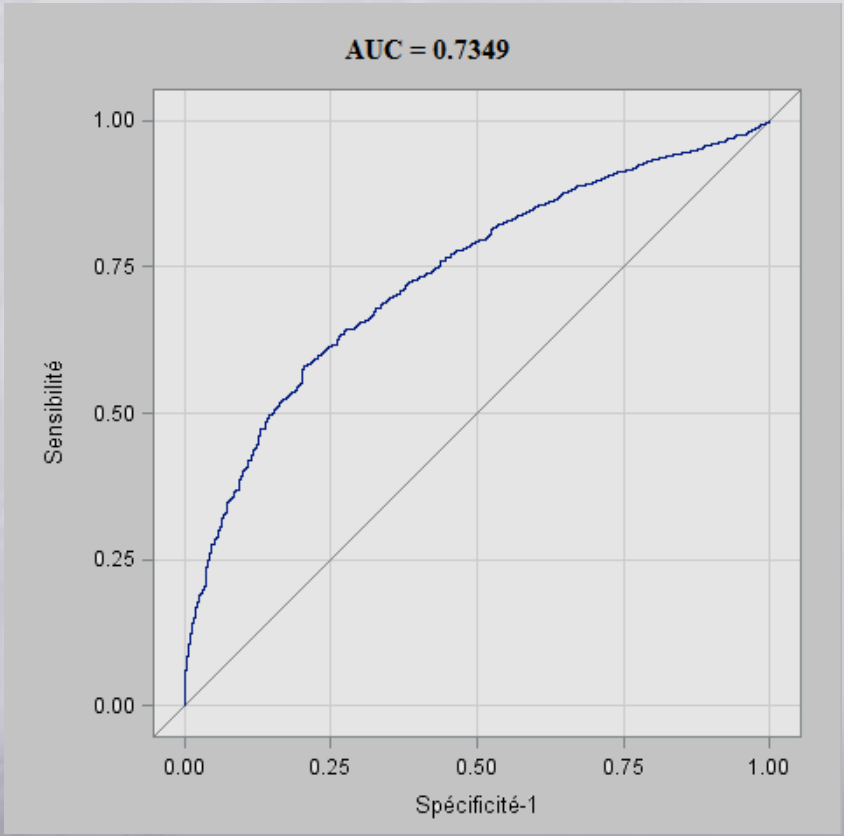


Re-weighting

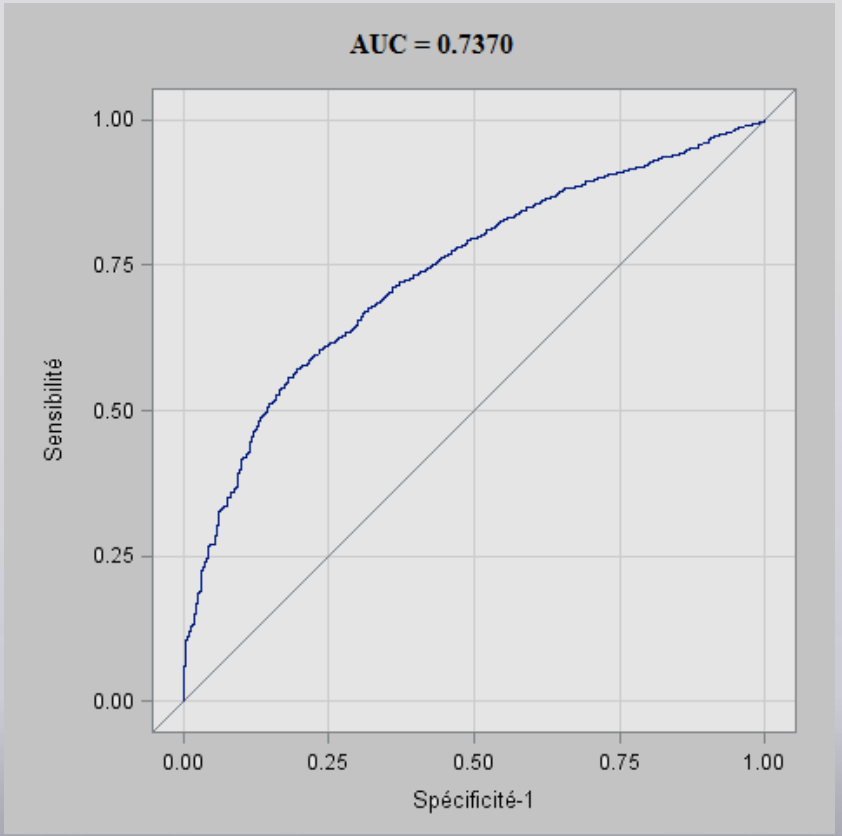


Performance comparaisons

Iterative reclassification

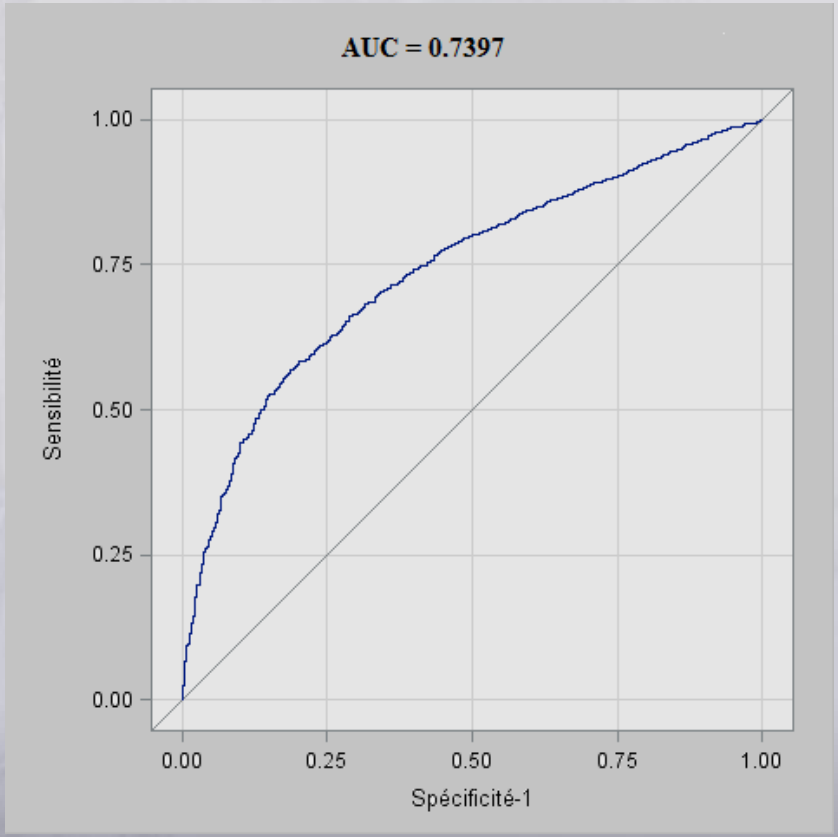


Simple augmentation

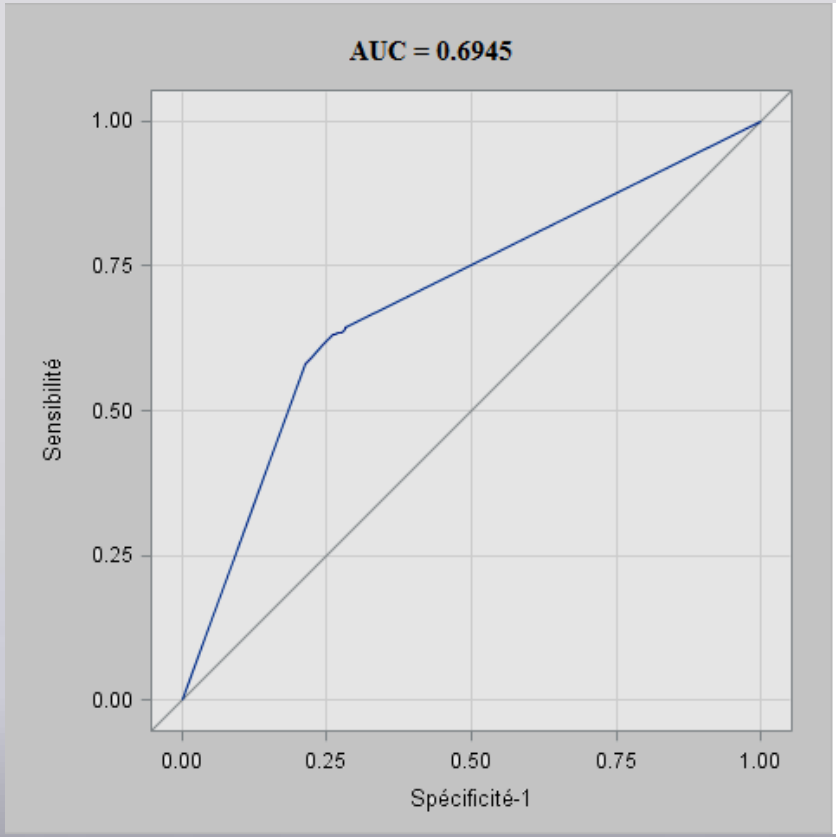


Performance comparaisons

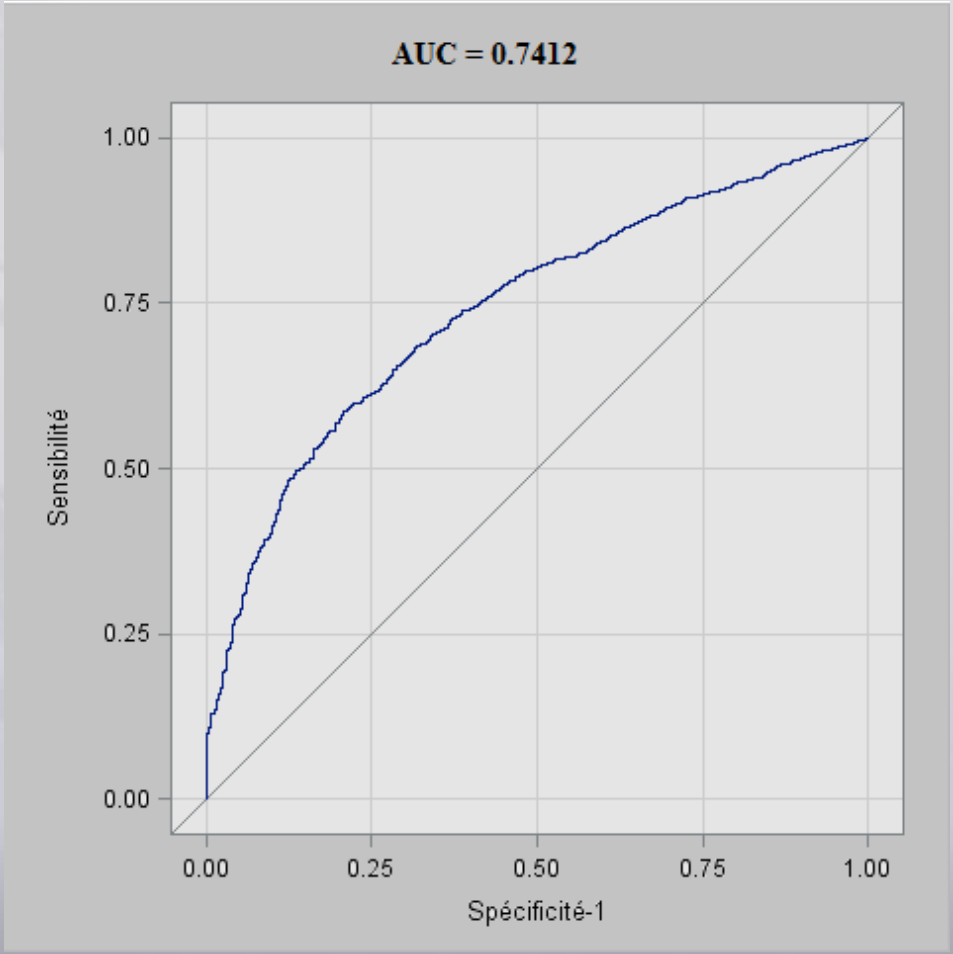
Mixed classification



AdaBoost



Gentle AdaBoost



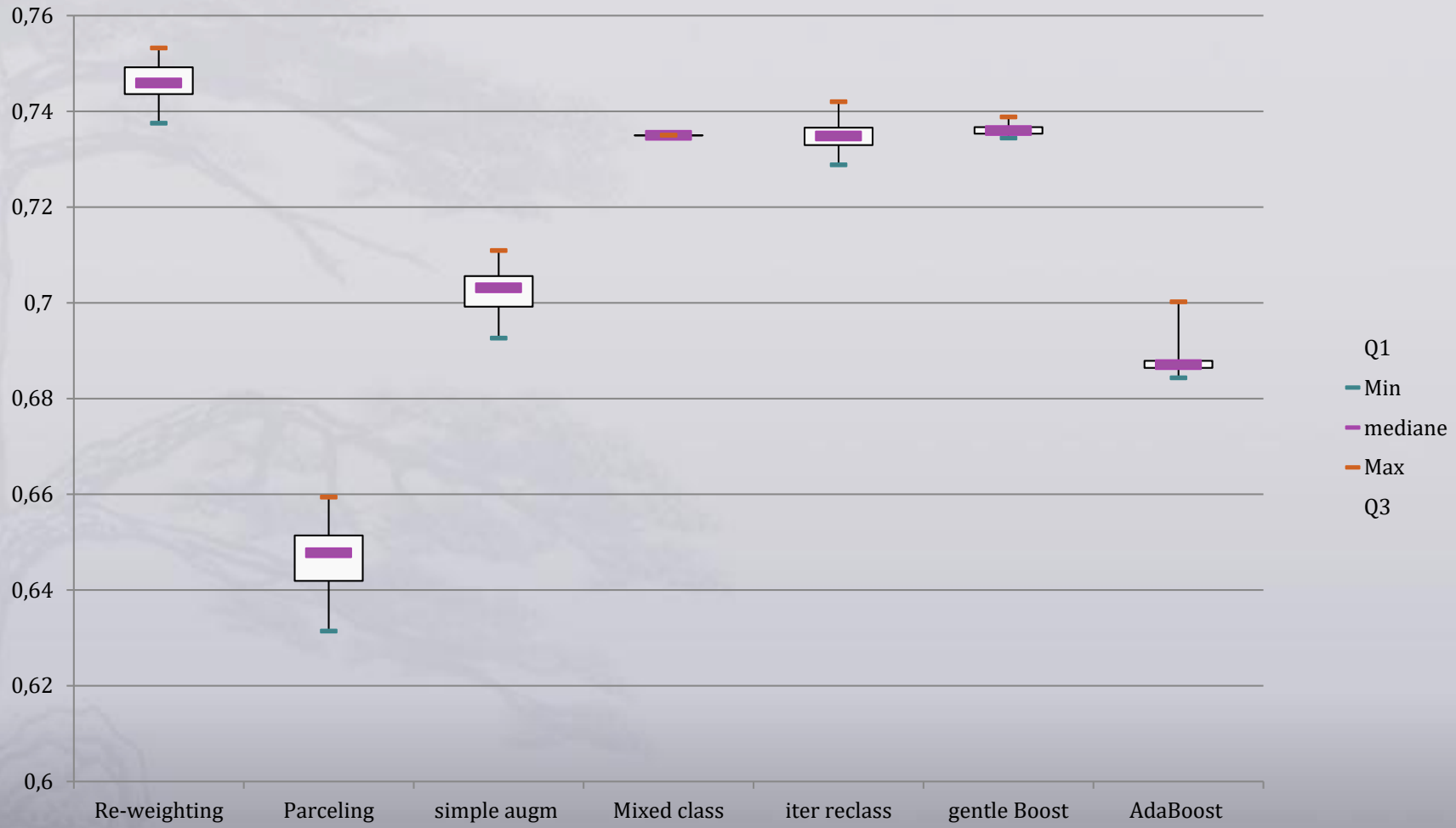
Comments

- Performance of the 7 methods :

Gentle AdaBoost > Mixed classification > Simple augmentation > Iterative reclassification > Re-weighting > AdaBoost > Parceling

- The 7 methods have a good predictive performance

Variability



Comments

- Expect for re-weighting and simple augmentation, the 7 methods keep the same performance.

- AUC has a small variability for the 50 samples.

Conclusion and future work

- The results of seven methods are promising.
- Simulate other rejection process
- Compare other methods applied on reject inference
- More comparaisons needed with Confusion matrix.

References

- Banasik, J., Crook, J., 2007. Reject inference, augmentation, and sample selection. European Journal of Operational Research, 183, 1582-1594.
- Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive Logistic Regression: A Statistical View of Boosting. The Annals of Statistics, 28, 2, 337-407.
- Siddiq, N., 2006. Credit risk scorecards developing and implementing intelligent credit scoring. John Wiley & Sons, Inc., New Jersey.
- Tuffery, S., 2011. Data Mining and Statistics for Decision Making. Wiley, New York.