

# Outil d'interprétation de score

## B. Thieurmel

Datastorm

ENSAE | ENSAI | CREST | CEPE | CASD

60 rue Etienne Dolet - 92240 Malakoff

benoit.thieurmel@datastorm.fr

**Mots clefs** : Statistique, Scoring, Interprétation, Visualisation

Les méthodes usuelles de scoring telles que les forêts aléatoires ou les réseaux de neurones mènent à des résultats difficilement interprétables.

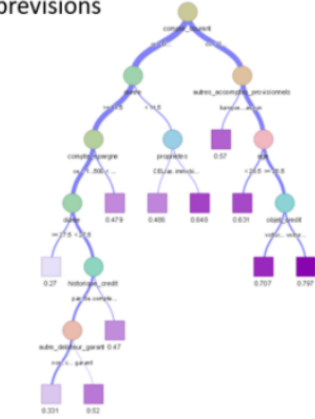
En effet, le fait de ne pas obtenir de coefficients associés aux variables du modèle ou dont l'interprétation n'est pas immédiate ne rend pas aisée la compréhension des résultats. Par ailleurs, la détermination de la valeur seuil du score déterminant l'appartenance ou non à une classe n'est généralement pas simple et dépend de la finalité du modèle.

Nous avons donc développé un outil d'aide à l'interprétation d'un score, se traduisant en une application **shiny**[1]. A partir d'un jeu de données contenant la variable réponse, des variables explicatives (prises en compte ou non dans la modélisation initiale) ainsi que le score obtenu, cet outil permet de répondre notamment aux questions suivantes :

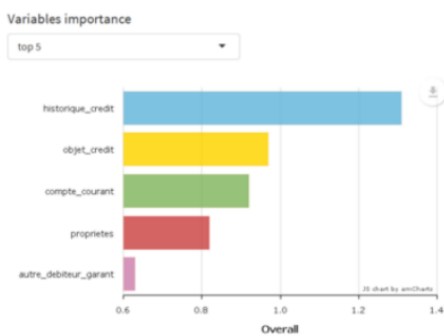
- Quel choix faire pour la valeur du seuil de score ?
- Quelles sont les variables qui ont le plus contribué à l'élaboration du score ? Et pouvons-nous expliquer ce score ?
- Peut-on expliquer ou catégoriser les bonnes ou les mauvaises prédictions ?
- Quelle est la répartition du score et sa performance selon les valeurs d'une des variables explicatives ?

Les visualisations dynamiques proposées sont réalisées avec les packages **rAmCharts**[2] et **visNetwork**[3], développés par les équipes de Datastorm. Des nouvelles fonctionnalités ayant été rajoutées dans le cadre de ce projet.

Arbre CART expliquant la variable réponse, le score ou la qualité des prévisions



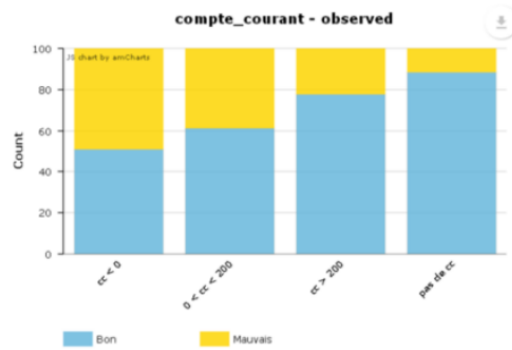
Importance des variables



Matrice de confusion partielle en fonction des modalités ou des intervalles d'une variable explicative

Modality	N	Accuracy	Real Pos. (%)	Of which pred. Pos. (%)	Real Neg. (%)	Of which pred. Neg. (%)
1	cc < 0	61.31%	50.73%	27.34%	49.27%	96.30%
2	0 < cc < 200	57.62%	60.97%	40.24%	39.03%	84.76%
3	cc > 200	52.38%	77.78%	65.31%	22.22%	7.14%
4	pas de cc	88.32%	88.32%	93.10%	11.68%	23.91%

Répartition de la réponse en fonction d'une variable explicative



## Références

- [1] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.5. <https://CRAN.R-project.org/package=shiny>
- [2] Benoit Thieurmél, Antanas Marcelionis, Jeffery Petit, Elena Salette and Titouan Robert (2018). rAmCharts: JavaScript Charts Tool. R package version 2.1.6. [http://datastorm-open.github.io/introduction\\_ramcharts/](http://datastorm-open.github.io/introduction_ramcharts/)
- [3] Almende B.V., Benoit Thieurmél and Titouan Robert (2018). visNetwork: Network Visualization using 'vis.js' Library. R package version 2.0.4. <http://datastorm-open.github.io/visNetwork/>