Classification des crédits bancaires avec l'algorithme Gradient Boosting Machine

Élaboré par : Safa MEDIOUNI
Encadré par : Nabil BELGASMI
École Supérieure de la Statistique et de l'Analyse de l'Information de Tunis

1. Introduction

La gestion des risques bancaires est une composante indispensable du secteur financier à l'échelle mondiale. Depuis des années, la solvabilité des clients et l'identification des emprunteurs douteux restent toujours parmi les préoccupations majeures des banquiers. Plusieurs méthodes simplistes ont été implémentées par ces banques afin de contourner ces problématiques, et avec l'avancement des sciences statistiques, de nouveau horizons s'ouvrent pour ce domaine.L'émergence du "Machine Learning" a fourni aux banquiers des outils innovants, qui ont permit d'avoir plus de précisions et de minimiser les erreurs d'estimation.

Ce papier explore l'implémention et la mise en place d'un modèle de scoring crédit au sien de la "Banque de Tunisie" en tirant profit d'un algorithme innovant de Machine Learning appelé : GBM ou "Gradient Boosting Machine".

2. Objectif

L'objectif de notre projet est d'accorder à chaque client de la banque un score de risque. Ce score va fournir une information sur la solvabilité du client en cas d'empreint d'un crédit.
Pour ce faire, l'algorithme GBM a été implémenté sur des données historiques, et une optimisation des paramètres du modèle a été faite afin d'avoir une précision maximale du modèle.

3. Données

Nous avons travaillé sur un ensemble de données composé d'environ 84.000 clients uniques.La durée d'observation de ces clients s'étendait sur une période de 3 ans (entre 2013 et 2016). Nos variables d'intérêt couvraient les aspects qualitatifs des clients comme l'âge,le genre et la profession et aussi les aspects qualitatifs tels que le salaire mensuel, les montants des crédits octroyés auparavant et les montants remboursés.
Pour la phase d'apprentissage, nous avons utilisé la variable catégorique Bon/Mauvais qui donne le classement affecté aux clients par la banque suite à leur historique d'empreint de crédits.

4.Techniques et méthodes

La méthode implémentée se base sur l'algorithme 'Gradient Boosting Machine' qui est une combinaison de la descente de gradient et du Boosting.
Le modèle GBM apprend automatiquement des données disponibles et prédit les classes d'appartenance de chaque client selon les profils semblables de ces derniers. Les techniques utilisées sont :

- **Validation croisée :** Nous avons utilisé la technique de « validation croisée » qui consiste à diviser, de façon aléatoire et de manière à être indépendants et identiquement distribués, les données en un ensemble d'apprentissage constituant 80% et un ensemble test constituant 20% des données.

- **Grille de recherche :** Nous avons à chaque fois varié les paramètres du modèle, qui ont été croisé pour donner une combinaison pour une performance donnée.Par la suite, nous avons mesuré les performances avec la « Cross-validation ». Nous avons utilisé cette méthode pour déterminer les meilleurs paramètres maximisant la performance du modèle.

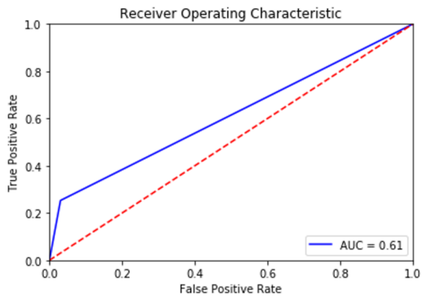
- **Sélection des variables :** Cette étape est d'une grande utilité dans la recherche de modèles parcimonieux. Elle consiste à enlever les variables à faible apport et ne garder que les variables qui ont le plus grand pouvoir prédictif.Une démarche générale est de commencer par l'ensemble global des variables disponibles et faire des éliminations individuelles. Plusieurs combinaisons sont possibles. Le choix du meilleur arrangement sera décider en prenant en compte la plus grande précision du modèle correspondant.

5.Résultats du modèle

La sortie du modèle donne une précision égale à 0.9382, avec une valeur du ROC de l'ensemble d'apprentissage égale à 0.986411. La valeur de la validation croisée est d'une moyenne 0.82436, d'un écart-type 0.00408, et variant entre 0.8175 et 0.8298.

- Validation du modèle:[1 cm]

ROC result:



Metrics for GBM model's validation :

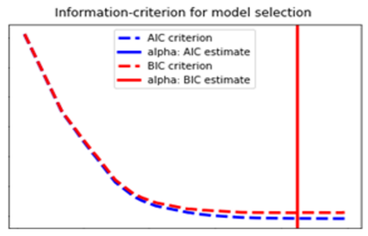
1. The Cohen kappa's score is : 0.29198894548
2. The Jaccard similarity is : 0.854549811292
3. The matthews correlation coefficient is: 0.32865156162
4. The balanced F-score: 0.358350951374

Classification Report:

	precision	recall	f1-score	support
class 0	0.87	0.97	0.92	14014
class 1	0.61	0.25	0.36	2679
avg / total	0.83	0.85	0.83	16693

Confusion Matrix:

```
[[13564   396]
 [ 2036   697]]
```



6.Démarche d'analyse

La capacité prédictive du modèle de scoring a été testée par la « courbe ROC ». Nous remarquons à partir des résultats que la courbe ROC est au-dessus de la première bissectrice ce qui signifie que le modèle est bon pour prédire.

Divers métriques ont été utilisées afin d'évaluer et de valider notre modèle telles que la matrices de confusion, le coefficient kappa...

7. Conclusion

A partir des données historiques des clients de la banque, nous avons pu construire un algorithme de Machine Learning qui accorde un score pour les nouveaux demandeurs d'emprunts et les classe comme étant bons ou mauvais.
Les résultats des tests effectués sur le modèle implémenté ont donné une précisions égale à 93%, Ainsi ce modèle nous garanti une meilleure performance que les modèles traditionnels.

Ceci dit, l'adoption de cet algorithme par la banque porte une véritable opportunité pour cette dernière dans la prise de décision d'accords des emprunts ainsi que dans la gestion du risque crédit. Cette précision garanti à la banque une meilleure politique d'évaluation de sa clientèle et optimise le processus d'octroi de crédits.

8. Références

- "The Elements of Statistical Learning,Data Mining,Inference,and Prediction"-Trevor Hastie.Second Edition.
- "Greedy Function Approximation : A Gradient Boosting Machine" - Jerome H.Friedman IMS 1999 Reitz lecture
- "PROJECTION PURSUIT REGRESSION" -Jerome H. Friedman