UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

MASTER 2 TIDE

JUILLET 2022

Projet Final d'Apprentissage Statistique Avancée

*L*andy CLEMENT
Berthony SULLY
Agui TCHABOU
Jerrold NEMBA

Contents

1 Résumé

Le credit scoring est généralement considéré comme une méthode qui évalue le niveau de risque associé à un dossier de crédit potentiel et prédit la solvabilité du demandeur de crédit. Cette méthode implique l'utilisation de différentes techniques statistiques pour aboutir à un modèle de scoring basé sur les caractéristiques du client. Les institutions financières utilisent ce modèle pour estimer la probabilité de défaut qui va être utilisée pour affecter chaque client à la catégorie qui lui correspond le mieux: bon payeur ou mauvais payeur.

Aujourd'hui, les techniques de Machine Learning permettent d'exploiter les données rendues disponibles par la digitalisation de la relation clientèle et les réseaux sociaux. Ainsi, la conjonction de l'émergence de ces technologies et des données a ainsi modifié de façon structurelle l'industrie du crédit et favorisé l'émergence de nouveaux acteurs. Au niveau microéconomique, ces nouvelles approches favorisent l'inclusion financière et l'accès au crédit des emprunteurs les plus fragiles. Toutefois, il faut signaler que l'application du Machine Learning sur des données peut aussi conduire à des biais et à des phénomènes de discrimination.

Dans la première partie de ce rapport, nous utilisons et comparons différents modèles d'apprentissage automatique (classifieurs uniques) et les méthodes ensemblistes ou d'agrégation (en anglais: Ensemble Learning) pour la modélisation du risque de crédit. Plus précisément, les méthodes Régression Logistique, Arbres de Décision, Random Forest, XGBoost et Stacking. D'abord, on a prédit les défauts de paiement sans faire de réduction de dimension et on a vu que le modèle d'Arbre de Décision est le meilleur pour la modélisation de notre problème avec un score d'AUC de 93 % sur le test. Ensuite, on a fait une réduction de dimension en utilisant les techniques AFDM et ACM. Cette dernière nous a permis d'améliorer les prédictions des modèles Régression Logistique et l'Arbre de Décision. Comparativement aux autres modèles, sur le test, le score de la Régression Logistique varie de 82% à 83% et celui de l'Arbre de Décision de 93% à 96%. Donc, le modèle **Arbre de Décision** est plus pertinent pour détecter les défauts de paiement que celui de **la Régression Logistique**. De plus, la réduction de dimension permet d'améliorer fortement la prédiction de l'Arbre de Décision. Il faut souligner qu'on a utilisé deux types de métriques statistiques pour évaluer la qualité de ces modèles, à savoir : les métriques qui mesurent la qualité générale du modèle, telles que l'AUC et la Logloss et celles basées sur la matrice de confusion qui dépendent du choix du seuil (cut-off) telles que le rappel et la précision.

Pour la deuxième partie du rapport, l'objectif est de connaitre les caractéristiques des entreprises sur le marché européen et de constituer des groupes homogènes basées sur les données dont on dispose. On a alors constaté qu'elles peuvent être regroupées en trois (3) grands groupes que sont les industries spécialisées, les industries traditionnelles et celles d'économies d'échelle et les industries dans le domaine de la technologie et celles non renseignées.

2 Introduction

De nos jours, les décisions basées sur des algorithmes deviennent prédominantes dans la quasi-totalité de domaines tels que le diagnostic médical, le traitement du langage naturel, la reconnaissance faciale, la reconnaissance vocale, la détection de fraudes, la recherche d'emploi ou la recherche de texte. Ainsi, le monde financier n'échappe pas à cette révolution de la science des données. L'intelligence artificielle (IA) et les techniques d'apprentissage automatique (Machine Learning) sont particulièrement utiles en matière de connaissance client, d'allocation d'actifs, de détection de blanchiment et de transactions illégales, de gestion des risques de crédit, ou d'amélioration des processus internes.

Dans cette étude, notre objectif est d'analyser le risque de crédit qu'est une forme d'analyse effectuée par un analyste de crédit sur des emprunteurs potentiels afin de déterminer leur capacité à honorer leurs obligations en matière de dette. Si l'emprunteur présente un niveau acceptable de défaut de paiement, l'analyste peut recommander l'approbation de la demande de crédit aux conditions convenues. Le résultat de l'analyse de défaut de paiement détermine la note de risque qui sera attribuée à l'emprunteur et sa capacité à accéder au crédit. Lorsqu'ils calculent le risque de crédit d'un emprunteur particulier, les prêteurs tiennent compte de divers facteurs communément appelés les 5 C du crédit qui comprennent : la capacité de l'emprunteur à rembourser le crédit, son caractère, son capital, ses conditions et ses garanties.

Par ailleurs, nous cherchons à répondre à la problématique suivante : La réduction de dimension par les techniques de l'ACM et de l'AFDM permet-elle d'améliorer les prédictions du risque de dédaut de paiement? Pour répondre à cette question, nous allons appliquer les modèles Régression Logistique, Arbre de Décision, Random Forest, XGBOOST et Stacking sur des données de crédit bancaire. Ensuite, nous allons analyser les principales caractéristiques des entreprises sur le marché européen.

3 Choix et description de notre Dataset

En vue de la réalisation de notre travail, on a disposé de données de défaut de paiement. Pour des questions de confidentialité, l'identité de la banque en question est restée secrète. La dimension de notre base de données est de 5000 observations et de 23 variables dont 7 numériques et 13 catégorielles plus la variable cible. Dans cet ensemble de données, chaque entrée représente une personne (un client) qui prend un crédit bancaire. Chaque personne est classée comme bon ou mauvais payeur selon l'ensemble des attributs.

4 Matériels et méthodes

La littérature montre que les méthodes ensemblistes permettent d'améliorer de façon spectaculaire des classifieurs standards en apprentissage supervisé et en particulier le problème de données déséquilibrées. C'est pourquoi, nous avons testé sur ces données les méthodes d'ensembles et individuels cités ci-dessus en utilisant le langage de programmation Python. Dans un premier temps, on a prédit les défauts de paiement sans faire de réduction de dimension et on a vu que le modèle d'Arbre de Décision est le meilleur pour la modélisation de notre problème. Dans un second temps, on a fait une réduction de dimension en utilisant les techniques AFDM et l'ACM. Cette dernière technique nous a permis d'améliorer les modèles Régression Logistique et l'Arbre de Décision qu'est encore le meilleur selon les résultats.

5 Compréhension de notre Dataset

De ce résumé des données, nous pouvons voir l'étendue de chaque variable. La valeur la plus élevée est 921 pour le montant du crédit (Credit Amount) et la plus faible est une forme binaire pour les variables Dependents et Default_On_Payment. Nous pouvons voir également qu'il n'y a pas de valeurs manquantes dans les données. Les sorties ci-dessous en témoignent.

[120]:
```
        # Variables quantitatives
        print(info_df_num_2)
```

```
          variables    Nb value distinct  Nb value missing  %Modalité_missing

0   Duration_in_Months                 33                 0                0.0
1        Credit_Amount                921                 0                0.0
2       Inst_Rt_Income                  4                 0                0.0
3  Current_Address_Yrs                  4                 0                0.0
```

4	Age	53	0	0.0
5	Num_CC	4	0	0.0
6	Dependents	2	0	0.0
7	Default_On_Payment	2	0	0.0

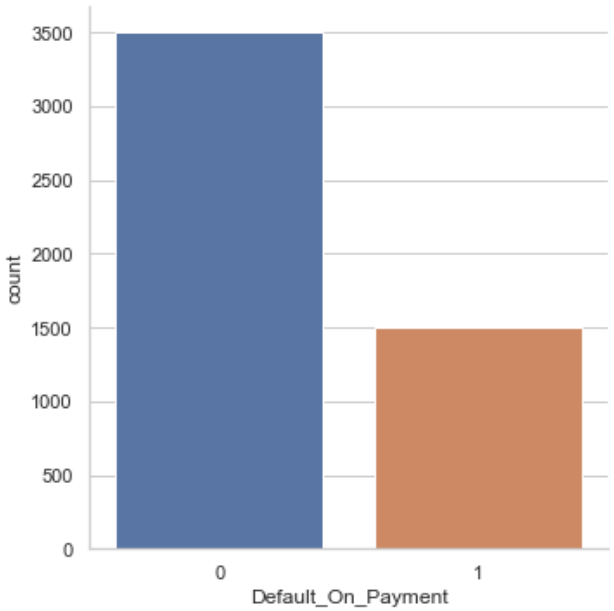
[121]:
```
# Varibales catégorielles :

print(info_df_cat)
```

	variables	Nb value distinct	Nb value missing
0	Status_Checking_Acc	4	0
1	Credit_History	5	0
2	Purposre_Credit_Taken	10	0
3	Savings_Acc	5	0
4	Years_At_Present_Employment	5	0
5	Marital_Status_Gender	4	0
6	Other_Debtors_Guarantors	3	0
7	Property	4	0
8	Other_Inst_Plans	3	0
9	Housing	3	0
10	Job	4	0
11	Telephone	2	0
12	Foreign_Worker	2	0

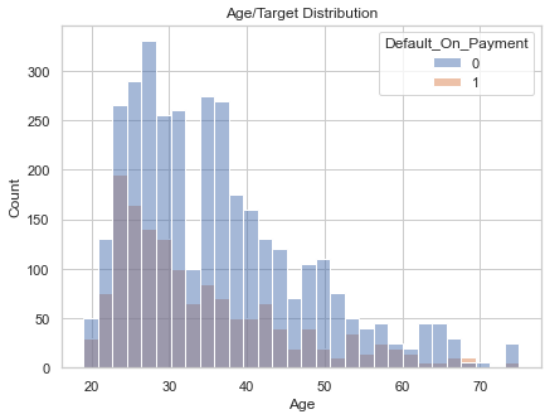
6 Analyse Exploratoire des Données

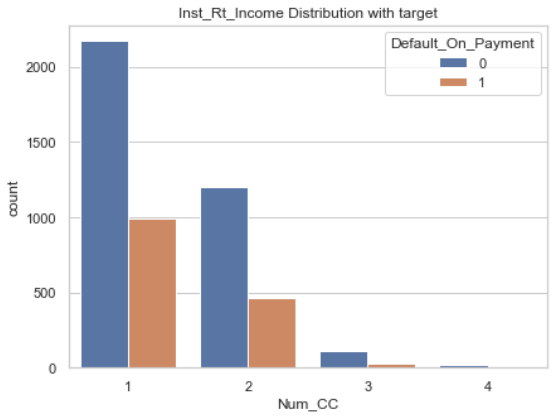
On a fait une analyse exploratoire des données en vue de répondre à deux objectifs spécifiques, à savoir : trouver la distribution de chaque variable individuelle pour évaluer sa qualité, c'est ce qu'on appelle l'analyse univariée et trouver les relations entre les variables et la variable cible, qu'est l'analyse bivariée.

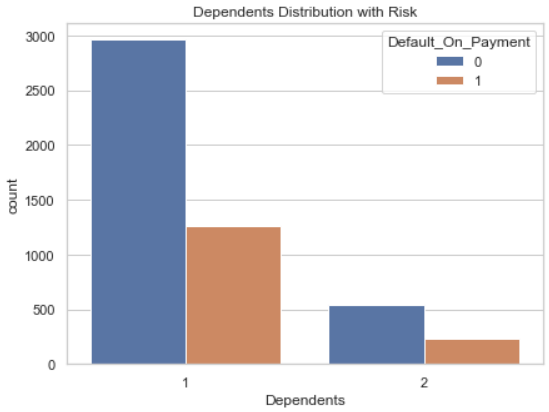
L'analyse univariée des variables numériques montre que les variables Age et Credit_Amount sont distribuées de manière inégale et sont décalées vers la gauche. Cela peut affecter la prédiction ultérieure. Quand à notre variable cible Default_On_Payment : 1 = défaut de paiement / 0 = Remboursement, elle est déséquilibrée. Le rapport est de 3500 / 1500 entre 0 et 1. Nous devons respecter cette répartition dans le jeu d'apprentissage et de test. On va traiter tous ces cas avant de modéliser pour ne pas biaiser prédictions.

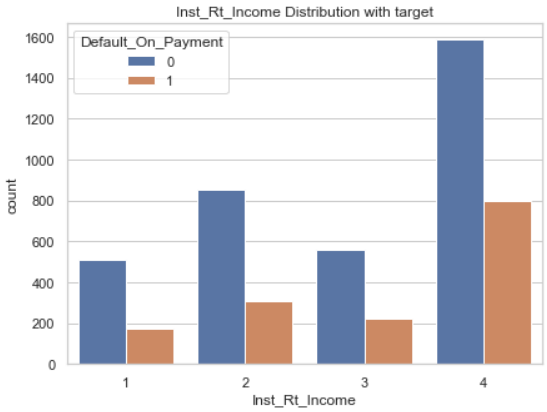


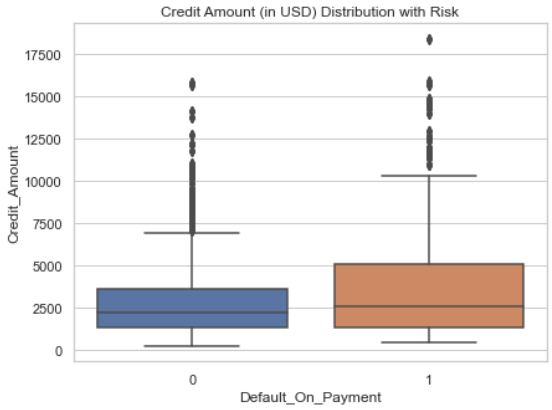
Ensuite, on a fait une analyse bivariée en croisant d'abord les variables numériques avec notre target et on a vu que la variable age n'est pas discriminante. C'est-à-dire, on ne peut pas tenir compte de l'âge d'une personne pour décider si elle a droit à un prêt ou pas puisque l'âge n'affecte pas beaucoup l'évaluation du risque. Le fait qu'on est jeune ne signifie pas qu'on est plus susceptible à faire des défauts bancaires que les adultes et vice versa. En ce qui concerne la durée et le montant du crédit, on voit bien qu'un montant de crédit plus élevé et une durée plus longue signifient un risque plus élevé pour la banque. Donc, les financiers doivent mettre beaucoup d'accent sur le montant et la durée des prêts au moment de l'évaluation des risques de défaut bancaire. Les graphiques ci-dessous en témoignent.

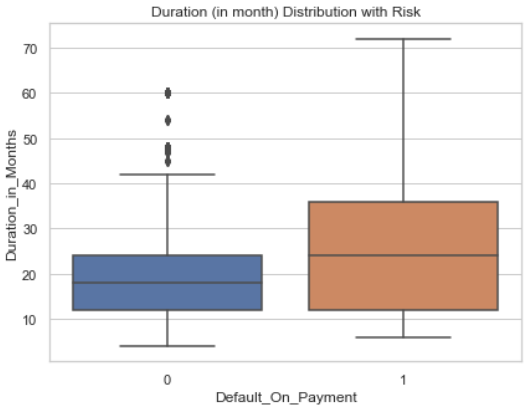




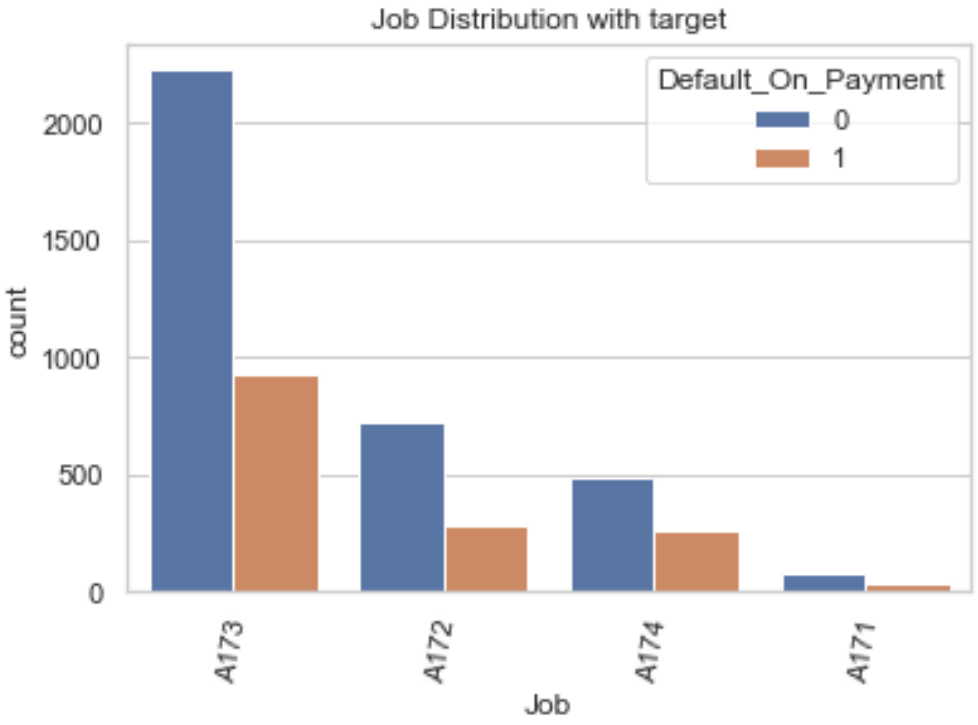








Plus loin, en croisant les variables catégorielles avec notre target, on voit que la plupart des personnes enregistrées ont un niveau de compétence professionnelle de 2 (Skilled). Cependant, le niveau de compétence professionnelle n'affecte pas beaucoup l'évaluation du risque. Donc, le niveau académique de quelqu'un n'est pas discriminant pour l'évaluation de risque de crédit bancaire. Les personnes qui possèdent une maison représentent un faible risque et un bon classement pour la banque.



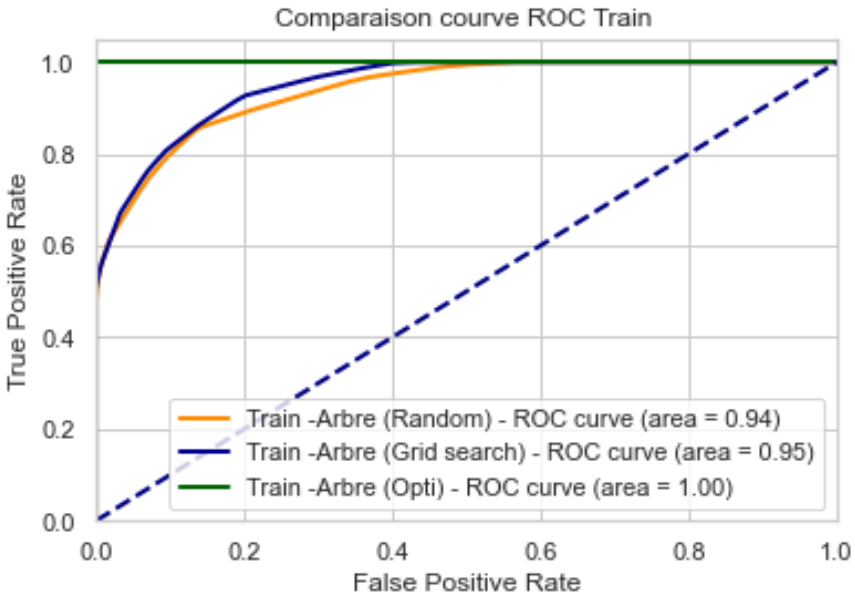
7 Modélisation

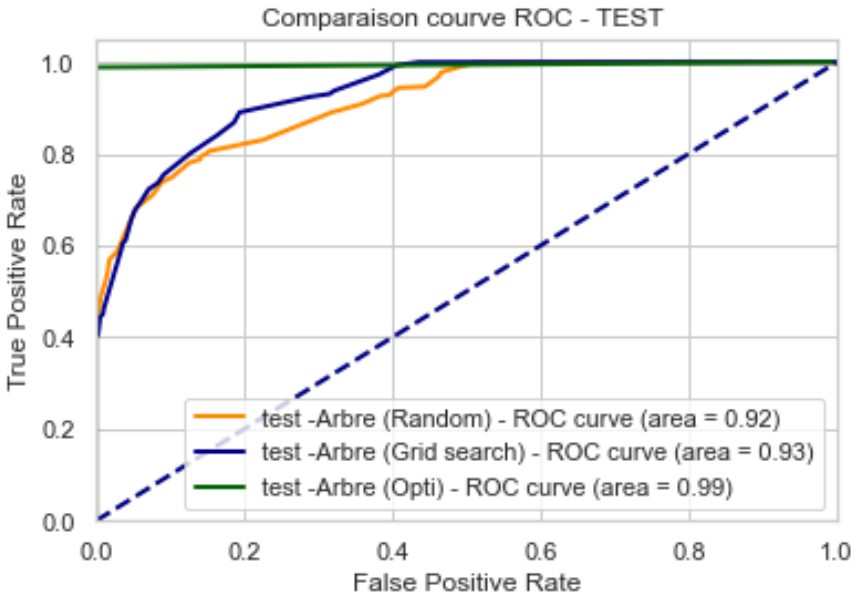
On a modélisé notre arbre de décision en fixant les paramètres criterion, splitter, max_depth. Après avoir appliqué la prédiction sur le train et calculé les probabilités prédites, on a eu une prédiction égale à 0 lorsque la probabilité est supérieure ou égale à 0.5 et 1 sinon. Etant donné que notre objectif est de prédire la probabilité de défaut bancaire d'un client, on a conservé en mémoire uniquement la probabilité de l'événement cible. C'est-à-dire, toutes les probabilités pour lesquelles la prédiction est égale à 1. L'arbre a l'avantage d'être visuel, et la fonction "feature importance" nous a permis d'afficher les variables qui ont le plus de pouvoir prédictif.

Pour évaluer la qualité générale de notre modèle, on a utilisé les méthodes de validation qui sont destinées à mesurer la capacité du modèle pour la prise de décision quant à son utilisation ou à son rejet. Ces méthodes de validation s'appuient sur des tests de robustesse appliqués sur un échantillon témoin non utilisé pour la construction du modèle. La validation du modèle est une étape décisive qui permet de vérifier la conformité des coefficients du modèle de score et évaluer ses performances et sa capacité prédictive. En effet, la robustesse de notre modèle est vérifiée à travers la matrice de confusion, l'aire sous la courbe de ROC et la logloss. Ladite matrice permet de calculer certains paramètres pour évaluer la capacité prédictive du modèle. Celle-ci est d'autant meilleure que l'Accuracy, la Precision et Recall sont élevés. Ensuite, on a ciblé 11% de notre population pour une précision de 92%. Donc, notre seuil tel que si Score est supérieur ou égal

au seuil et qui implique que Ypred =1 est de 0.68.

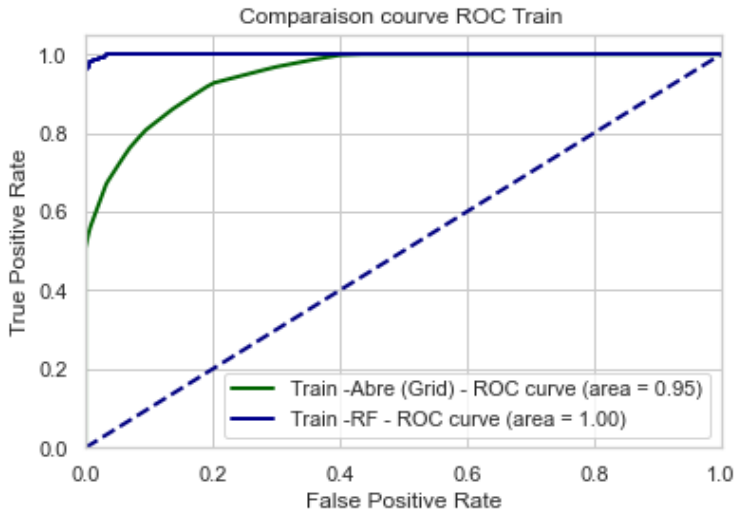
Pour optimiser les résultats d'un modèle, le choix des paramètres est très important car, un mauvais choix de paramètre peut impliquer du sur-apprentissage ou du sous-apprentissage. En effet, on a utilisé trois (3) grandes méthodes pour choisir nos hyper-paramètres, à savoir : les méthodes aléatoire, grid Search et optimisation bayésienne. Parmi les trois (3) méthodes utilisées pour l'optimisation de notre modèle, on a retenu la Méthode Grid Search puisqu'elle est plus stable sur le train et sur le test. Ci-dessous les résultats de notre modèle Arbre de Décision.

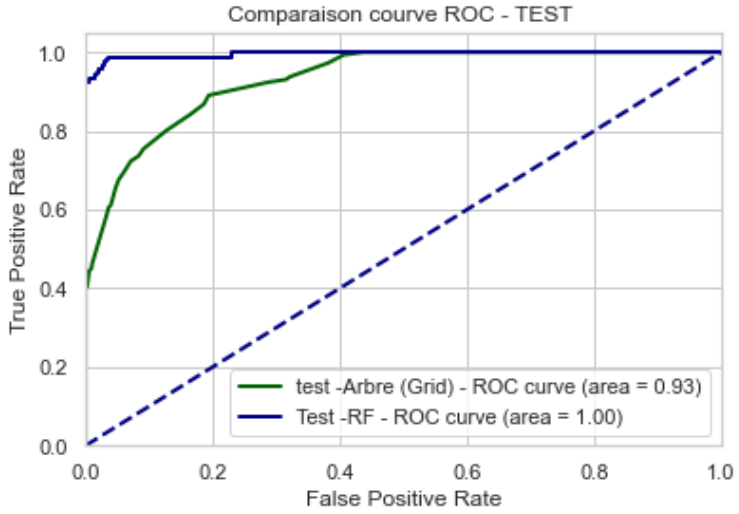




7.1 Random Forest

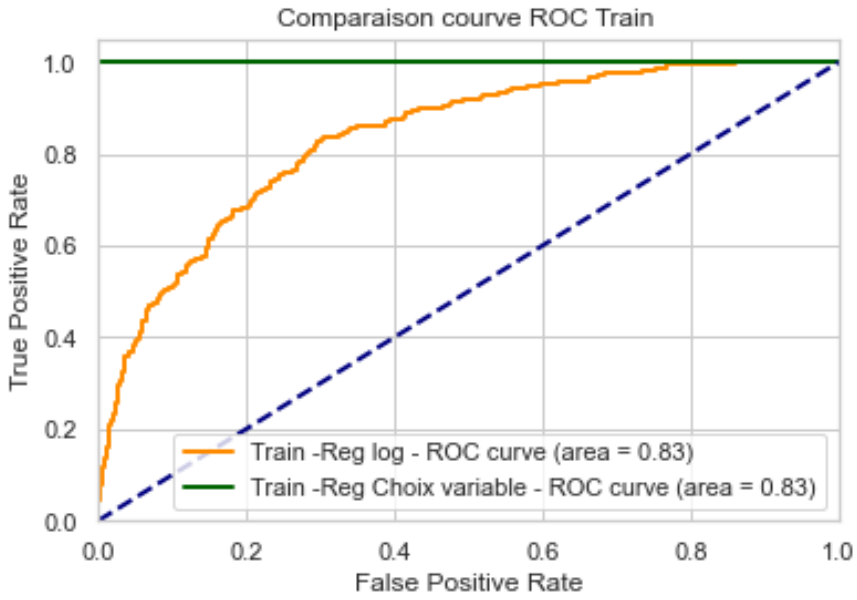
Après avoir mis en place notre modèle Random Forest qu'est un bagging basé sur des arbres de décisions puis, l'optimisé par la méthode grid Search pour sélectionner les hyperparamètres, on a trouvé une AUC égale à 1 sur le train et sur le train. Ce modèle a été ensuite challengé avec le modèle Arbre de décision pour voir lequel est le plus pertinent à détecter les défauts bancaires. On a vu que le modèle Arbre de décision est plus pertinent pour détecter les défauts de paiement que celui de Random Forest puisqu'il est plus stable. Ci-dessous, les graphiques illustrant la comparaison des modèles Arbre de Décision et Random Forest.

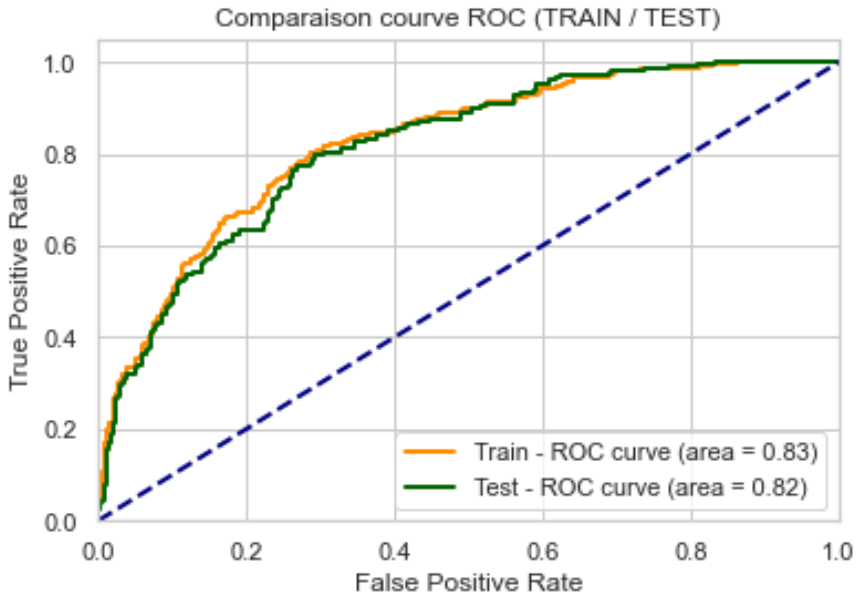




7.2 Régression logistique

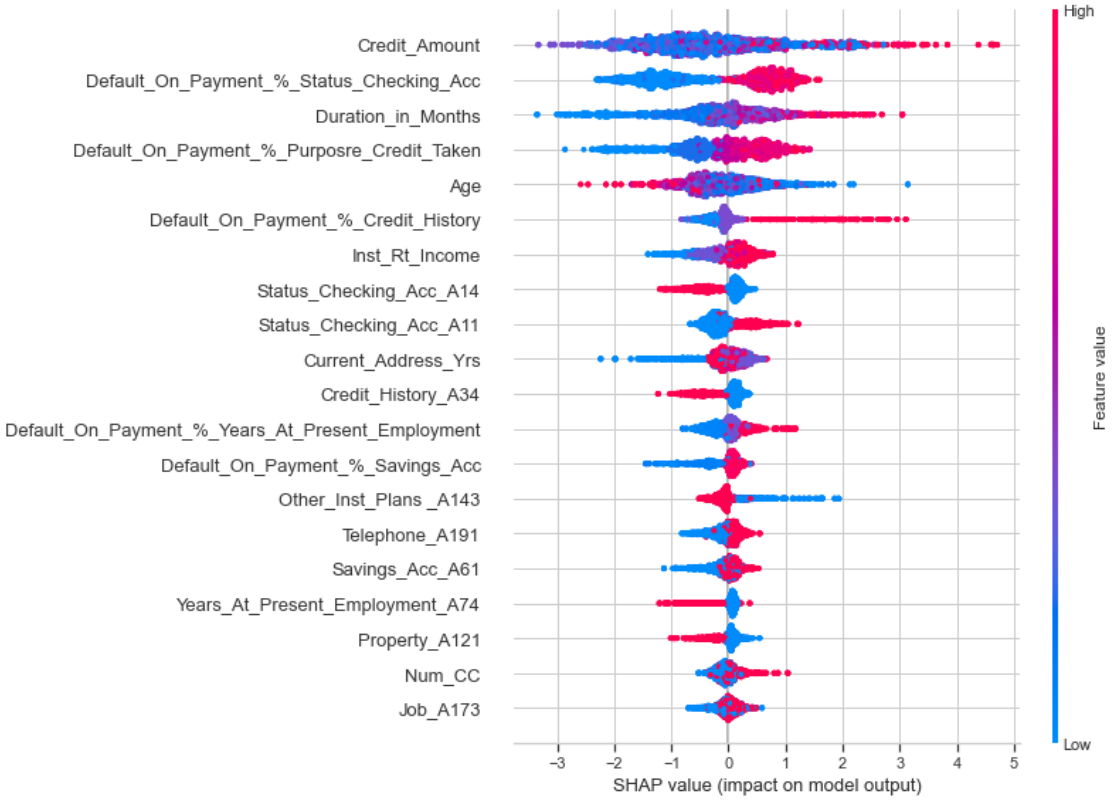
La régression logistique nécessite une plus grande préparation des données. Ainsi, pour pouvoir interpréter les résultats de notre régression, la normalisation des données s'avérait être importante. En vue de construire un modèle plus robuste, nous avons utilisé les variables les plus importantes de Random Forest. Après avoir utilisé les variables les plus importantes pour notre modèle, nous obtenons un score d'AUC de 82% sur le test. Voir les graphiques ci-dessous :



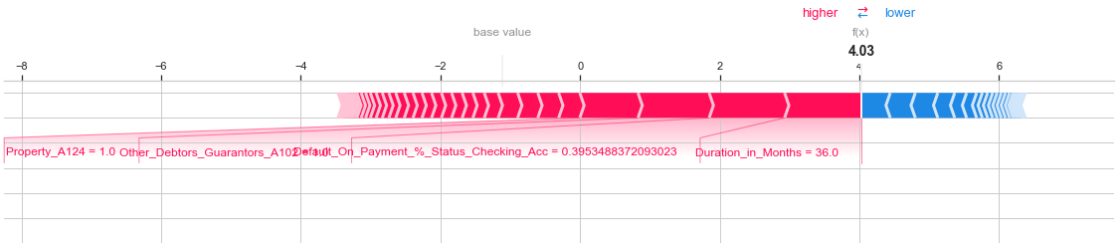


7.3 XGBOOST

L'algorithme XGBoost fait partie des algorithmes de Boosting qui se basent sur le même principe que ceux de Bagging. La seule différence apparaît lors de la création des apprenants faibles. Pour le Boosting, les algorithmes ne sont plus indépendants. Au contraire, chaque apprenant faible est entraîné pour corriger les erreurs des apprenants faibles précédents. En un mot, le Boosting de modèle est un modèle itératif où le modèle à l'étape k essaie de prédire l'erreur de prédiction des k-1 modèles précédents. Ainsi, le score d'AUC de notre modèle XGBOOST est de 1. Pour aller plus loin et de comprendre les prédictions de Xgboost, on a utilisé le package shap qui nous a permis de comprendre quelles variables qui impactent les résultats pour chaque observation. On a vu que plus les variables Credit_Amount, Status_Checking_Acc, Duration_In_Months, Credit_History sont positives, plus le risque de défaut de paiement sera élevé.



Plus loin, on a vu que pour la prédiction de la 4ème observation sur X_test, les variables qui poussent vers une augmentation du score de risque sont Property_A124, Other_Debtors_Guarantors_A102 et Duration_In_Months.

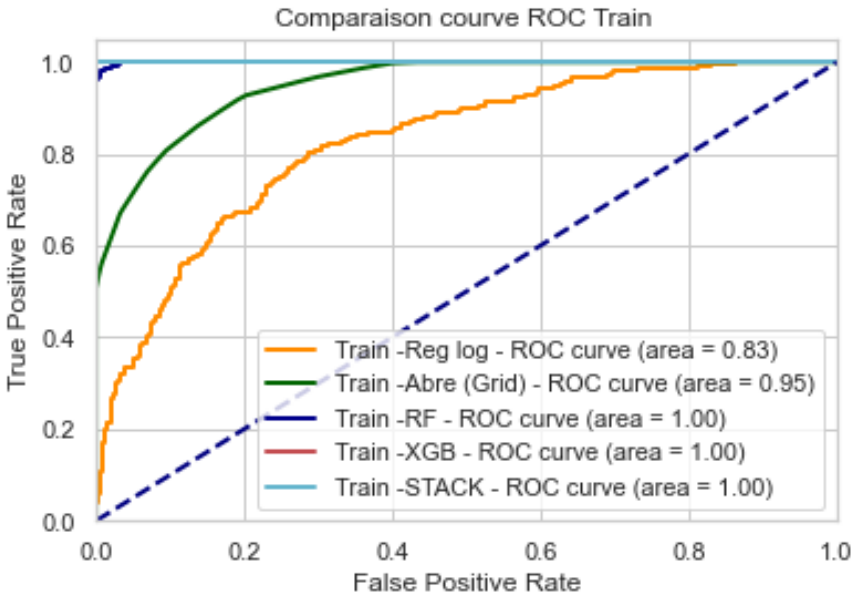


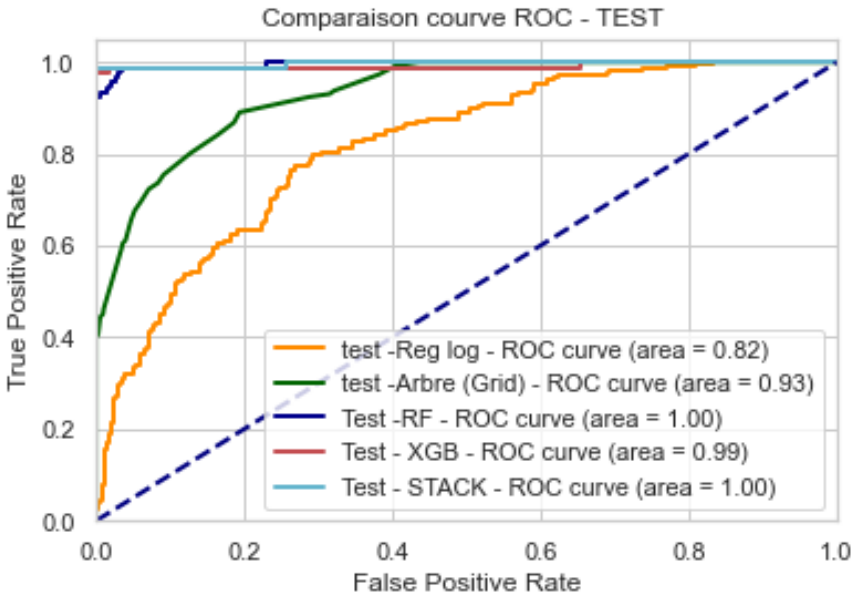
7.4 Stacking

Le stacking consiste à combiner les prédictions de plusieurs modèles d'apprentissage automatique sur le même ensemble de données, comme le bagging et le boosting. Donc, ce modèle est construit à partir des scores des modèles Arbres de Décision, Random Forest, Regression Logistique et XGBOOST. Le stacking a été conçu pour améliorer la performance de la modélisation, mais il n'est pas garanti que cela se traduise par une amélioration dans tous les cas. Ainsi, notre modèle ensembliste a un score d'AUC de 1.

7.5 COMPARAISON DE TOUS LES MODELES

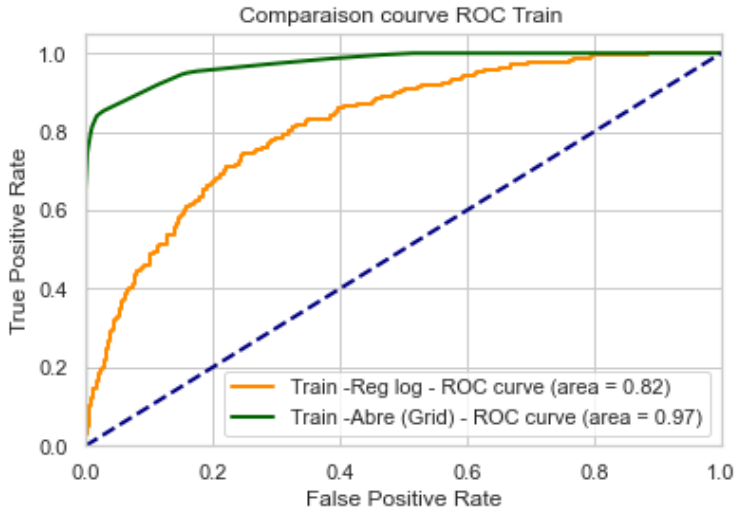
Afin de selectionner un modèle parmi tous les modèles qu'on a mis en place dans le cadre de notre étude, nous avons décider de comparer tous les modèles. On a vu que tous les modèles ensemblistes ont une AUC égale à 1 sur le train et sur le test. Par contre, on a vu que les modèles Arbre de Décision et Régression Logistique ont une meilleure performance avec une AUC successivement de 93% et de 82% sur le test. Donc, l'Arbre de Décision est le meilleur modèle pour détecter les défauts de paiement.

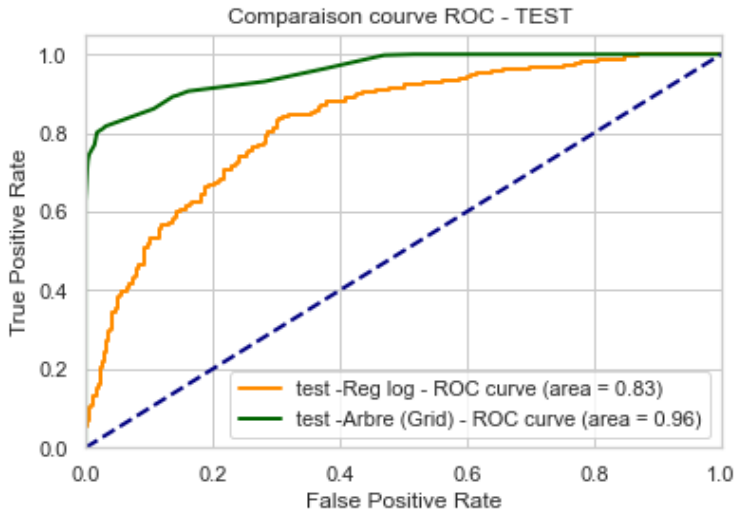




8 Réduction de dimension : ACM

Pour aller plus loin, on a fait une réduction de dimension en utilisant la technique de l'ACM. En effet, on a appliqué l'ACM sur les variables catégorielles et on s'est retrouvé à la fin avec 53 dimensions. Cependant, nous avons vu qu'on a atteint le maximum d'informations à partir de la 40e dimension. Ainsi, on a décide de prendre 95% d'informations en ne gardant que 36 dimensions. Ensuite, on a fusionné ces 36 dimensions avec les variables numériques pour refaire la prédiction avec les memes modèles utilisés précédemment. Par conséquent, cette technique nous a permis d'améliorer les modèles Régression Logistique et l'Arbre de Décision. Ainsi, sur le test, le score desdits modèles varient successivement de 82% à 83% et de 93% à 96% comparativement aux prédictions faites sans réduction de dimensions.





9 Réduction de dimension : AFDM

Tout comme l'ACM, AFDM est une Méthode factorielle de réduction de dimension pour l'exploration statistique de données qualitatives complexes. Cette méthode est une généralisation de l'Analyse Factorielle des Correspondances, permettant de décrire les relations entre p (p > 2 variables) variables qualitatives simultanément observées sur n individus. L'AFDM résulte de la combinaison entre l'ACP et l'ACM. Lorsqu'on possède un jeu de données qui comporte à la fois des données quantitatives et qualitatives, l'utilisation de l'AFDM est préconisée. Elle consiste dans un premier temps soit à convertir les données qualitatives en quantitatif et d'appliquer l'ACP et dans un second temps ou soit à convertir les données quantitatives en qualitatives , dans ce cas on applique

l'ACM.

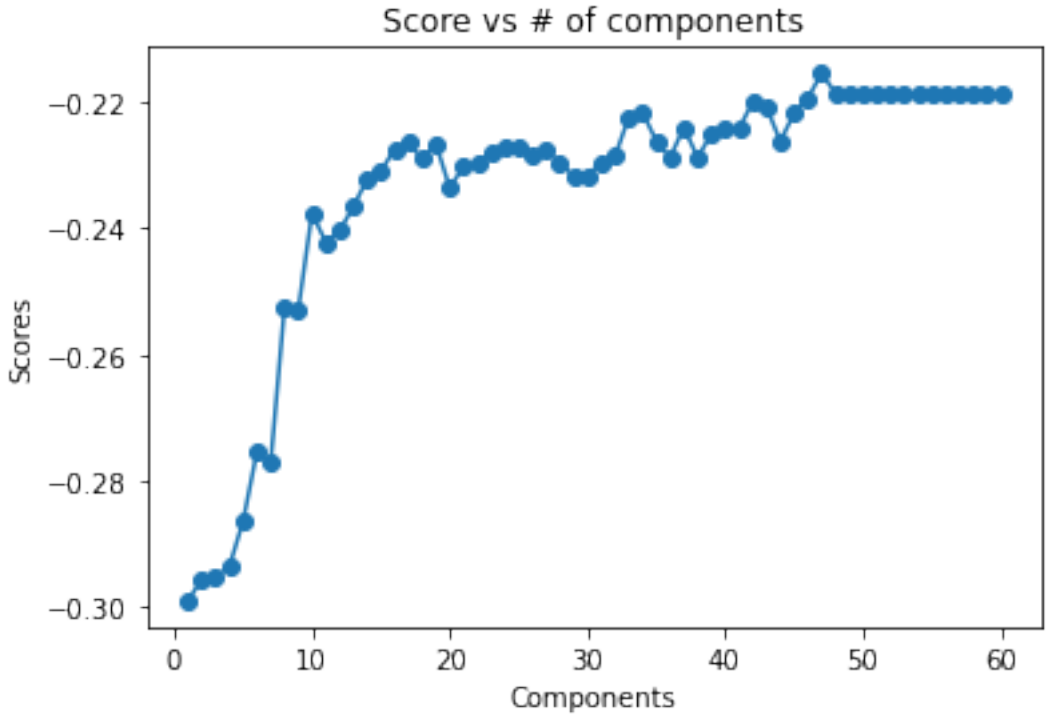
Dans notre cas, nous avons converti les treize (13) variables qualitatives de la base en quantitatives et d'appliquer l'ACP. Pour ce faire, on a d'abord fait une one hot encoding sur les données qualitatives puis nous avons utilisé le code suivant pour normaliser ces données en gardant le poids de la varible (le nombre de modalités).

[18]:
```
        table_rouge[["Telephone__A191", 'Telephone__A192']] =␣
↪table_rouge[["Telephone__A191", 'Telephone__A192']].apply(lambda x: x /␣
↪np.sqrt(1/2))
```

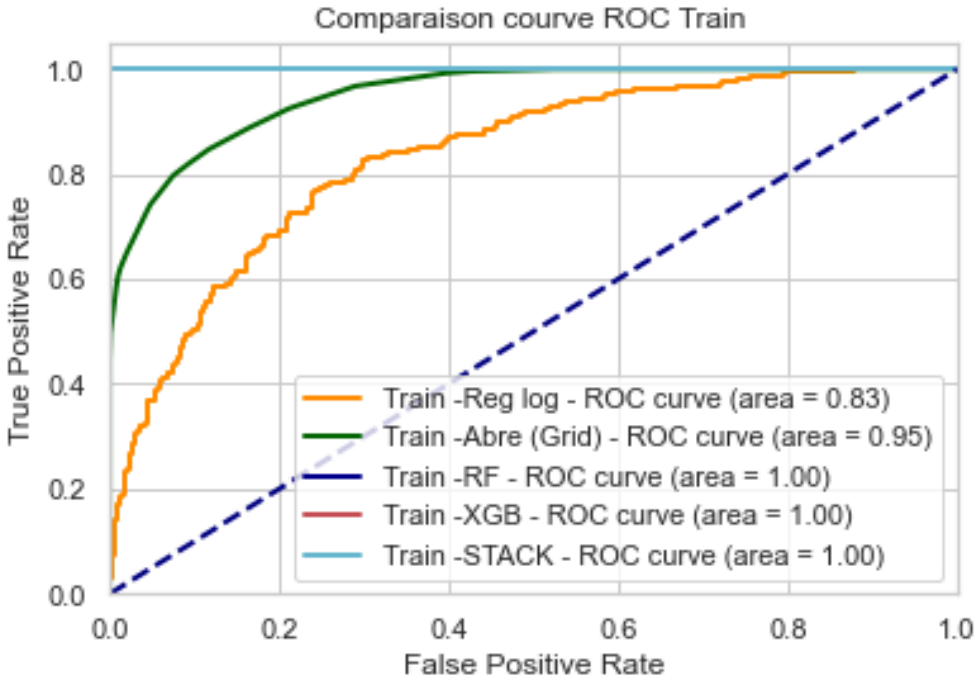
Pour ce qui est des variables quantitatives, on les a normalisées en soustrayant leur moyenne et en divisant par l'écart type. Voir l'exemple de code ci-dessous :

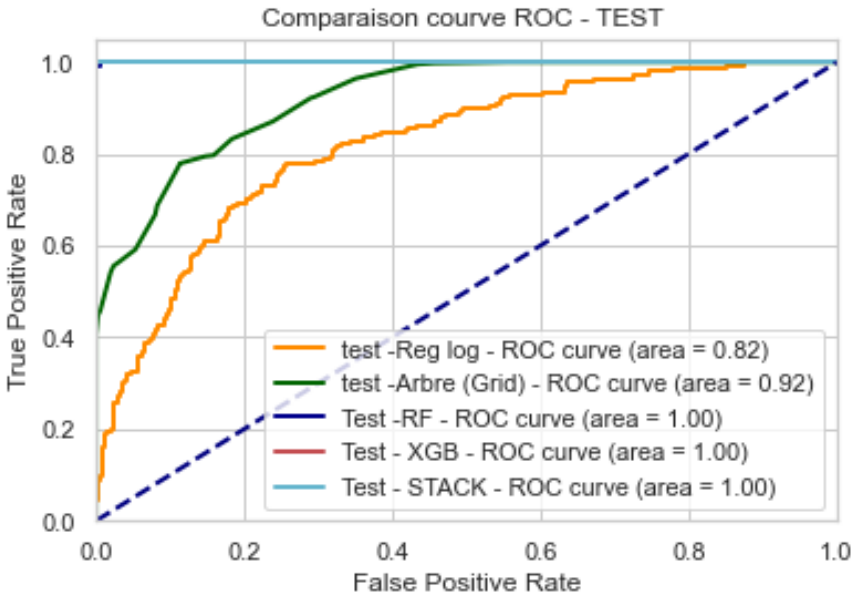
[17]:
```
        table_rouge[['Duration_in_Months',␣
↪'Credit_Amount','Inst_Rt_Income','Current_Address_Yrs','Age','Num_CC','Dependents']]␣
↪= table_rouge[['Duration_in_Months',␣
↪'Credit_Amount','Inst_Rt_Income','Current_Address_Yrs','Age','Num_CC','Dependents']].
↪apply(lambda x: (x-np.mean(x)) / np.std(x))
```

Après avoir réalisé l'ACP, on a déterminé le nombre optimal de composantes à retenir pour la modélisation et le graphique ci-dessous de sélectionner que les 47 premières composantes.



On a vu que l'AFDM n'a pas permis d'améliorer grandement les résultats car, les modèles ensemblistes comme XGBOOST, Stacking et Random Forest tombent encore à 1 sur le train et sur le test. Pour ce qui est des modèles Régresssion Logistique et Arbre de Décision, on a obtenu successivement un scofre d'AUC de 82% et 92% sur le test. Voir les graphiques ci-dessous comparant les résultats de tous les modèles.





DEUXIÈME PARTIE : CONNAITRE LES PRINCIPALES CARACTÉRISTIQUES DES ENTREPRISES EUROPÉENNES

10 Introduction

Les crises financières et économiques ne constituent pas un nouveau phénomène. En effet, les crises, abus de confiance ou détournements d'actifs ont toujours accompagné l'activité économique. Cependant, l'environnement dans lequel l'entreprise évolue, le fonctionnement structurel, les moyens techniques et de communication actuels (messageries et réseaux sociaux), accessibles à un public large, ouvrent des nouvelles opportunités aux crises systémiques dépendamment de l'entreprise, tant en termes de son domaine d'activité.

11 Méthodologie

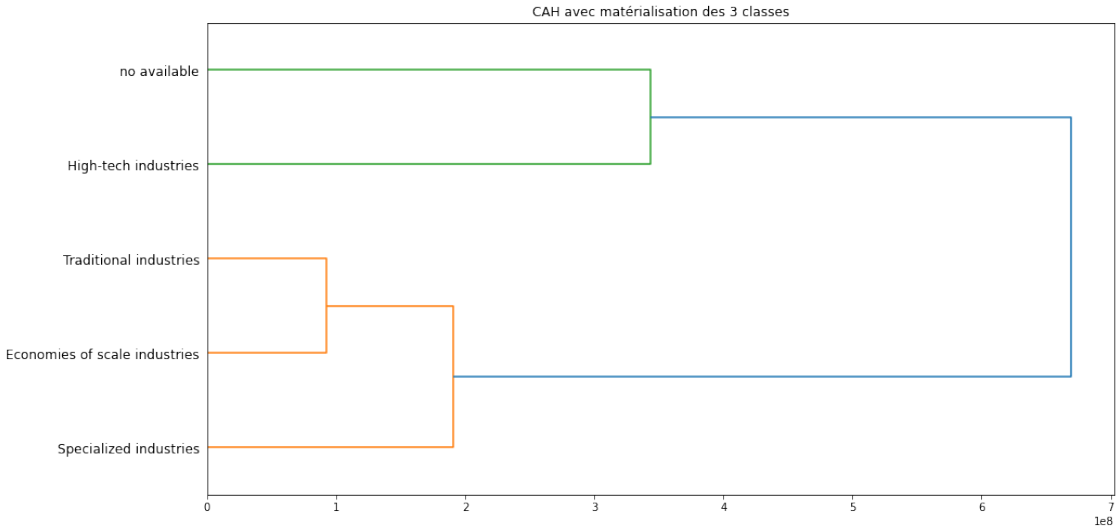
Pour cette partie, on a décidé de faire une classification ascendante hiérarchique (CAH) pour essayer de rapprocher les observations. La CAH est un algorithme de machine learning de la catégorie non supervisée. Comme les KMeans, elle permet d'identifier des groupes homogènes dans une population, on parle aussi de clustering. Ses champs d'application sont divers : segmentation client, analyse de donnée, segmenter une image, apprentissage semi-supervisé....

11.1 Traitement des données

- La base de données contient 14759 observations et 489 variables, dont beaucoup d'entre elles, comprennent une grande partie de valeurs manquantes.
- On a ensuite supprimé celles qui continent plus de 10% de valeurs manquantes. Cette stratégie a laissé 137 des 489 variables de départ.
- On a constaté que la plupart des variables restantes étaient catégorielles et on a exécuté une imputation des valeurs manquantes par le mode.
- On a décidé d'éliminer les variables catégorielles restantes et de garder la variable "pavitt" qui se désigne comme la spécialisation des entreprises car on veut que notre analyse se porte sur cet angle. La finalité est de créer des groupes homogènes pouvant distinguer ces entreprises dont on dispose les une des autres.

Un groupement des données a été faite selon la variable "pavitt" avec une agrégation par moyenne et le traçage du dendrogramme a été réalisé.

[266]:
```
# importation des outils
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster

#CAH
Z = linkage(group, method='ward',metric='euclidean')

#affichage du dendogramme
plt.figure(figsize=(15, 8))
plt.title('CAH avec matérialisation des 3 classes')
dendrogram(Z,labels=group.index,orientation='right')
plt.show()
```



Le dendrogramme présente une classification qui peut se faire en trois groupes. Les groupes sont très cohérents au point de vue métier car les industries traditionnelles et nouvelles se distinguent assez bien sur le dendrogramme. On a finalement :

- Groupe 1 : les industries spécialisées
- Groupe 2 : les industries traditionnelles et celles d'économies d'échelle
- Groupe 3 : les industries dans le domaine de la technologie et celles non renseignées

Ce qui serait intéressant d'explorer comme piste pour continuer serait de construire des clusters à partir de ces données pour pouvoir extraire le plus d'informations que possible. Les méthodes comme KMeans sont généralement utilisées avec ce genre de classification pour approfondissement de l'analyse.

12 Conclusion

Après avoir modélisé notre problème de détection de défaut de paiement sans réduire les dimensions, on a voulu les challengé avec les mêmes modèles de prédiction mis en place suite à une réduction de dimensions en utilisant les techniques ACM et AFDM. Cela nous a permis de voir que la réduction de dimension permet d'améliorer fortement la prédiction du modèle Arbre de Décision. Ainsi, la réduction de dimension par la méthode d'ACM a rendu le modèle **Arbre de décision** plus pertinent pour détecter les défauts de paiement que celui de **la Régression Logistique** et les modèles ensemblistes. Car, les scores obtenus sur le train et sur le test sont successivement 97% et 96%.

Pour ce qui sont des caractéristiques des entreprises sur le marché européen, on a constaté qu'elles peuvent etre regroupées en trois (3) grands groupes à savoir les industries spécialisées, les industries traditionnelles et celles d'économies d'échelle et les industries dans le domaine de la technologie et celles non renseignées.