Implémentez un modèle de scoring (version 2022) - Note méthodologique

Dans ce projet, nous réalisons la mission de définir et exploiter un modèle de machine learning qui soit utile à une société financière proposant des crédits à la consommation. Aussi, nous sommes chargés de rendre disponible et interprétable la décision du modèle sur l'octroi du prêt dans un tableau de bord accessible à des personnes non expertes.

Nous partons d'une série de caractéristiques de demandes de prêt déjà clôturées, dont un indicateur binaire d'existence de difficultés de remboursement (notre variable cible). Ce support sera le point de départ pour l'entraînement de notre classificateur.

Méthodologie d'entraînement du modèle

Nous résumons à la suite les opérations réalisées dans la phase d'entraînement. Puis nous procéderons à les détailler.

1. Prétraitement générique du jeu de données complet.
2. Séparation du dataset en jeu d'entraînement et jeu de test.
3. Traitement du jeu d'entraînement. Ce traitement consiste à:
    a. Sélectionner les 100 features les plus significatives et,
    b. Dans la plupart des cas, adresser le déséquilibre de la variable cible.

    Le résultat est un dataset d'entraînement dérivé du premier, celui qui sera effectivement utilisé pour l'entraînement.
4. Entraînement du modèle.
5. Évaluation de la performance du modèle.

La phase 1 aborde les traitements communs de nettoyage du dataset. Il comprend aussi l'imputation des valeurs manquantes par la médiane de la variable en question.
L'opération à la phase 2 est cruciale et elle est exécutée parmi les premières afin d'éviter des phénomènes de data leakage, soit de contaminer l'information passée à l'entraînement des estimateurs qui en améliorerait la performance d'une manière non réaliste. Nous séparons donc les données entre un set que nous allons élaborer afin d'entraîner les estimateurs au mieux (jeu d'entraînement) et un autre qui nous servira de référence pour le test de la performance du modèle (jeu de test).

On observe que les deux valeurs de la variable cible (0, correspondantes à un crédit remboursé sans difficultés et 1, associés à des difficultés de remboursement) sont représentées dans les proportions respectivement, de 82% et 8%.

Nous avons donc à faire avec un problème de classification sur des données déséquilibrées.

En général, ce type de déséquilibre doit être traité avant que le jeu soit utilisé pour entraîner un modèle. La raison étant qu'un modèle ainsi entrainé favoriserait la classe la plus représentée, à laquelle on se réfère comme la classe majoritaire.
Nous sommes alors confrontés avec trois possibilités:

1. Utiliser une méthode qui tient compte du déséquilibre lors de l'entraînement. Ces modèles sont rares et un tel choix exclurait la majorité des méthodes, qui requièrent un dataset équilibré.
2. L'autre possibilité est de traiter ce déséquilibre en amont et générer un dataset équilibré à partir du dataset original. Dans ce cas, on modifie l'une des deux classes pour qu'elle atteigne le même nombre d'instances que l'autre. On parle alors de:
    a. Sur-échantillonnage, quand la classe minoritaire est peuplée de nouveaux éléments (nous avons utilisé l'algorithme SMOTE).
    b. Sous-échantillonnage, quand la classe majoritaire est réduite. Nous avons utilisé l'algorithme Cluster Centroids et le random undersampling. Cette approche semble avoir l'avantage de ne pas modifier la classe la plus délicate dans notre scénario, celle des cas non solvables.

C'est au moment de choisir comment traiter le déséquilibre qu'on commence à définir un protocole de classification, qui se compose d'une combinaison de choix à niveau 1) du traitement du déséquilibre du dataset, 2) du classificateur, et 3) des paramètres du classificateur.

Nous allons donc définir le protocole à travers des actions suivantes:
- traiter le déséquilibre du dataset et générer des dataset d'entraînement équilibré.
- réaliser une première feature selection sur le dataset. On obtient un set réduit des colonnes les plus déterminantes pour la prévision avec kbest.
- répéter le traitement du déséquilibre sur le dataset réduit aux colonnes issues de la feature selection. Un tel dataset régénéré améliorera la qualité de l'entraînement.
- réaliser l'entraînement du modèle de notre choix sur le dataset réduit et retraité.
- évaluation de la performance d'un modèle sur le jeu de train par validation croisée sur 5 segments et sur le jeu de test.

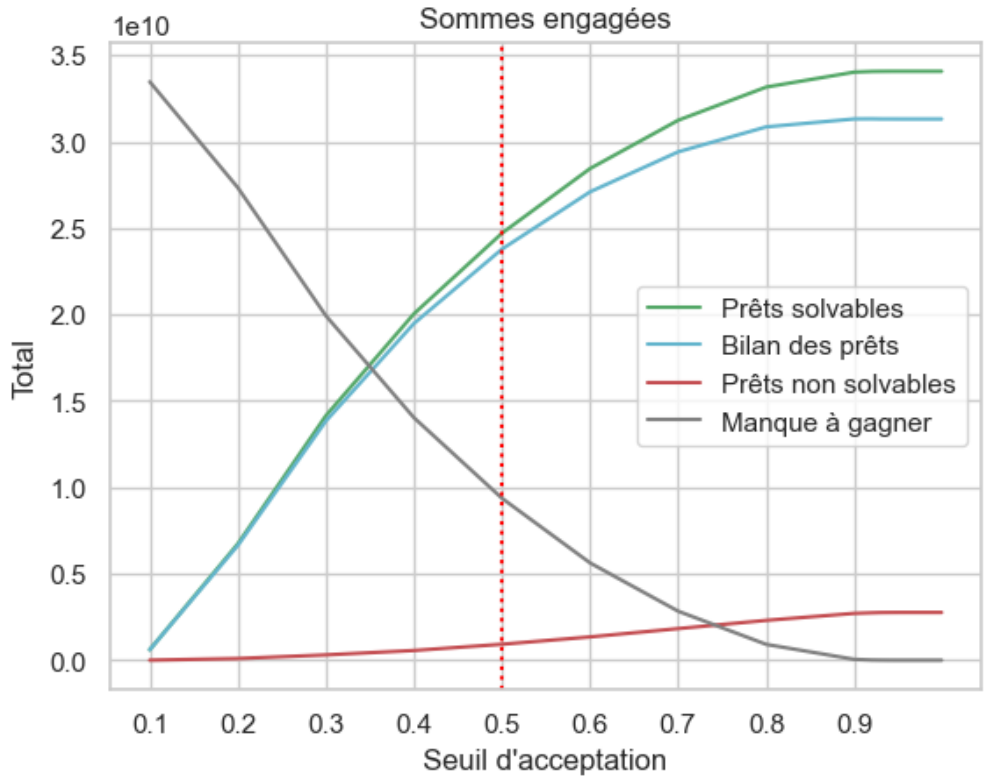
Les mesures de performances qui ont guidé le choix du classificateur, mesurées à travers de 5 métriques, sont montrées dans le tableau annexe. Suite à leur analyse, l'estimateur HistGradientBoostingClassifier avec les paramètres {'max_iter': 200, max_leaf_nodes': 40} issus de GridSearch a été retenu et entraîné sur un jeu équilibré par Random Undersampling.

Comparaison des modèles

Classificateur	Dataset	Performance sur le jeu d'entraînement					Performance sur le jeu de test					Performance calculée par validation croisée sur le jeu d'entraînement					Performance calculée par validation croisée sur le jeu de test équilibré par random undersampling				
		Accuracy	Recall	F_β ($\beta = 3$)	Roc auc	R^2	Accuracy	Recall	F_β ($\beta = 3$)	Roc auc	R^2	Accuracy	Recall	F_β ($\beta = 3$)	Roc auc	R^2	Accuracy	Recall	F_β ($\beta = 3$)	Roc auc	R^2
'Dummy '	'rus'	0,500	**0,000**	**0,000**	0,500	-1,000	0,919	**0,000**	**0,000**	0,500	-0,088	0,500	**0,000**	**0,000**	0,500	-1,000	0,500	**0,000**	**0,000**	0,500	-1,000
'LogReg '	'rus'	0,690	0,685	0,686	0,690	-0,242	0,688	0,691	0,521	0,689	-3,206	0,688	0,682	0,683	0,688	-0,250	0,686	0,680	0,681	0,686	-0,255
'GauNB '	'rus'	0,631	0,561	0,569	0,631	-0,475	0,687	0,564	0,434	0,631	-3,219	0,624	0,588	0,591	0,624	-0,503	0,630	0,610	0,611	0,630	-0,481
'SVM '	'rus'	0,690	0,685	0,686	0,690	-0,239	0,689	0,691	0,522	0,690	-3,187	0,689	0,683	0,684	0,689	-0,245	0,689	0,684	0,684	0,689	-0,244
'XGBClass '	'rus'	0,800	0,796	0,796	0,800	0,200	0,690	0,690	0,522	0,690	-3,174	0,686	0,681	0,682	0,686	-0,257	0,672	0,676	0,675	0,672	-0,313
'RandFC '	'rus'	1,000	1,000	1,000	1,000	1,000	0,682	0,683	0,513	0,683	-3,282	0,678	0,678	0,681	0,683	-0,277	0,672	0,683	0,685	0,678	-0,293
'RandFC '	'rus'	1,000	1,000	1,000	1,000	1,000	0,691	0,680	0,515	0,686	-3,168	0,686	0,672	0,675	0,685	-0,274	0,683	0,672	0,671	0,684	-0,280
'KNC '	'rus'	0,714	0,704	0,705	0,714	-0,144	0,616	0,597	0,428	0,607	-4,169	0,604	0,590	0,591	0,604	-0,583	0,597	0,575	0,578	0,597	-0,613
'HistGBC '	'rus'	0,758	0,755	0,756	0,758	0,034	0,696	0,704	0,535	0,700	-3,098	0,694	0,689	0,686	0,693	-0,221	0,680	0,683	0,683	0,680	-0,280
'Dummy '	'ccus'	0,500	**0,000**	**0,000**	0,500	-1,000	0,919	**0,000**	**0,000**	0,500	-0,088	0,500	**0,000**	**0,000**	0,500	-1,000	0,500	**0,000**	**0,000**	0,500	-1,000
'LogReg '	'ccus'	0,826	0,838	0,836	0,826	0,303	0,369	0,832	0,476	0,580	-7,497	0,822	0,836	0,834	0,822	0,289	0,666	0,662	0,663	0,666	-0,337
'GauNB '	'ccus'	0,824	0,862	0,856	0,824	0,297	**0,221**	0,853	0,441	0,509	-9,498	0,819	0,860	0,853	0,819	0,277	0,526	**0,087**	**0,094**	0,526	-0,894
'SVM '	'ccus'	0,832	0,847	0,845	0,832	0,327	0,353	0,839	0,474	0,574	-7,724	0,827	0,844	0,841	0,827	0,309	0,666	0,665	0,665	0,666	-0,335
'XGBClass '	'ccus'	1,000	1,000	1,000	1,000	1,000	**0,150**	0,938	0,460	0,509	-10,452	0,961	0,945	0,948	0,961	0,843	0,642	0,643	0,643	0,642	-0,431
'RandFC '	'ccus'	1,000	1,000	1,000	1,000	1,000	**0,163**	0,921	0,456	0,509	-10,275	0,952	0,929	0,933	0,952	0,805	0,646	0,651	0,651	0,651	-0,382
'RandFC '	'ccus'	1,000	1,000	1,000	1,000	1,000	**0,163**	0,923	0,457	0,510	-10,277	0,952	0,930	0,934	0,951	0,805	0,646	0,663	0,662	0,652	-0,409
'KNC '	'ccus'	0,921	0,913	0,915	0,921	0,684	**0,289**	0,823	0,446	0,532	-8,586	0,848	0,823	0,827	0,848	0,391	0,568	0,550	0,552	0,568	-0,728
'HistGBC '	'ccus'	0,980	0,965	0,968	0,980	0,919	**0,150**	0,940	0,461	0,511	-10,451	0,961	0,944	0,947	0,961	0,847	0,660	0,657	0,658	0,660	-0,362
'Dummy '	'smote'	0,500	**0,000**	**0,000**	0,500	-1,000	0,919	**0,000**	**0,000**	0,500	-0,088	0,500	**0,000**	**0,000**	0,500	-1,000	0,500	**0,000**	**0,000**	0,500	-1,000
'LogReg '	'smote'	0,678	0,688	0,687	0,678	-0,287	0,667	0,659	0,491	0,664	-3,484	0,678	0,688	0,686	0,678	-0,288	0,664	0,662	0,663	0,664	-0,345
'GauNB '	'smote'	0,664	0,661	0,661	0,664	-0,344	0,648	0,471	0,354	0,567	-3,741	0,663	0,659	0,660	0,663	-0,347	0,615	0,471	0,485	0,615	-0,540
'SVM '	'smote'	0,680	0,714	0,709	0,680	-0,278	0,651	0,682	0,498	0,665	-3,706	0,680	0,713	0,709	0,680	-0,280	0,666	0,666	0,666	0,666	-0,338
'XGBClass '	'smote'	0,962	0,925	0,932	0,962	0,847	0,916	**0,042**	**0,047**	0,518	-0,128	0,952	0,912	0,916	0,952	0,807	0,642	0,646	0,645	0,642	-0,430
'RandFC '	'smote'	1,000	1,000	1,000	1,000	1,000	0,913	**0,041**	**0,044**	0,515	-0,171	0,949	0,911	0,915	0,949	0,794	0,655	0,658	0,653	0,651	-0,388
'KNC '	'smote'	0,876	0,999	0,975	0,876	0,505	0,585	0,535	0,377	0,562	-4,593	0,781	0,996	0,955	0,781	0,124	0,567	0,581	0,580	0,567	-0,731
'HistGBC '	'smote'	0,958	0,918	0,925	0,958	0,833	0,918	**0,027**	**0,030**	0,512	-0,104	0,953	0,911	0,916	0,953	0,814	0,652	0,650	0,651	0,652	-0,391
'Dummy '	'orig'	0,919	**0,000**	**0,000**	0,500	-0,088	0,919	**0,000**	**0,000**	0,500	-0,088	0,919	**0,000**	**0,000**	0,500	-0,088	0,500	**0,000**	**0,000**	0,500	-1,000
'LogReg '	'orig'	0,919	**0,016**	**0,017**	0,507	-0,086	0,920	**0,014**	**0,016**	0,507	-0,084	0,919	**0,015**	**0,017**	0,507	-0,086	0,690	0,684	0,685	0,690	-0,239
'GauNB '	'orig'	0,574	0,695	0,474	0,629	-4,742	0,576	0,707	0,482	0,635	-4,720	0,596	0,672	0,468	0,631	-4,450	0,636	0,615	0,616	0,636	-0,458
'SVM '	'orig'	0,919	**0,000**	**0,000**	0,500	-0,088	0,919	**0,000**	**0,000**	0,500	-0,088	0,919	**0,000**	**0,000**	0,500	-0,088	0,691	0,686	0,687	0,691	-0,236
'XGBClass '	'orig'	0,934	**0,196**	**0,213**	0,597	0,109	0,918	**0,056**	**0,062**	0,525	-0,102	0,918	**0,055**	**0,060**	0,524	-0,110	0,674	0,675	0,675	0,674	-0,306
'RandFC '	'orig'	0,986	0,823	0,838	0,911	0,807	0,919	**0,015**	**0,017**	0,507	-0,098	0,918	**0,015**	**0,016**	0,507	-0,101	0,626	0,553	0,560	0,630	-0,484
'KNC '	'orig'	0,930	**0,217**	**0,233**	0,605	0,057	0,904	**0,047**	**0,051**	0,513	-0,288	0,905	**0,046**	**0,050**	0,513	-0,283	0,576	0,561	0,563	0,576	-0,696
'HistGBC '	'orig'	0,921	**0,043**	**0,047**	0,521	-0,058	0,920	**0,031**	**0,034**	0,514	-0,078	0,919	**0,028**	**0,031**	0,513	-0,084	0,682	0,684	0,684	0,682	-0,274

Fonction coût métier, algorithme d'optimisation et métrique d'évaluation

Même le classificateur le plus performant présente des marges d'erreurs. Notre approche à la fonction associée aux erreurs de classification s'est concentrée sur les sommes impliquées dans ces erreurs. Nous avons alors déterminé le nombre d'erreurs sur le jeu de test et extrait la valeur financière correspondante. Les faux négatifs correspondent aux erreurs les plus graves, car le classificateur n'identifie pas un cas de difficultés à rembourser, donc potentiellement une perte de capital ou, à tout le moins, un retard dans son recouvrement. Les faux positifs sont moins problématiques car ils entraînent des manques à gagner plutôt que des pertes réelles.
Bien que le classificateur restitue un résultat binaire, sa classification se fonde sur une valeur de probabilité qu'il peut restituer. Par défaut, le classificateur applique 0.5 comme seuil minimum pour classifier comme insolvable un profil. Cette probabilité est un levier ultérieur qui nous permet d'exercer un contrôle plus fin sur la classification. À partir de cette probabilité et des sommes engagées, le data scientist peut accompagner la direction de l'entreprise dans la calibration de sa stratégie financière. Il est donc intéressant de voir comment les sommes engagées varient en fonction du seuil de probabilité pour la prédiction d'un cas non solvable. D'après nos résultats, fixer un seuil souple comme 0.9 mène, raisonnablement, les prêts non solvables à augmenter, mais en même temps réduit à zéro le manque à gagner. Il en résulte que le bilan atteint son maximum avec une stratégie d'octroi très permissive. Ce résultat peut refléter une attitude prudente des demandeurs vis-à-vis des conséquences d'un prêt non remboursé.

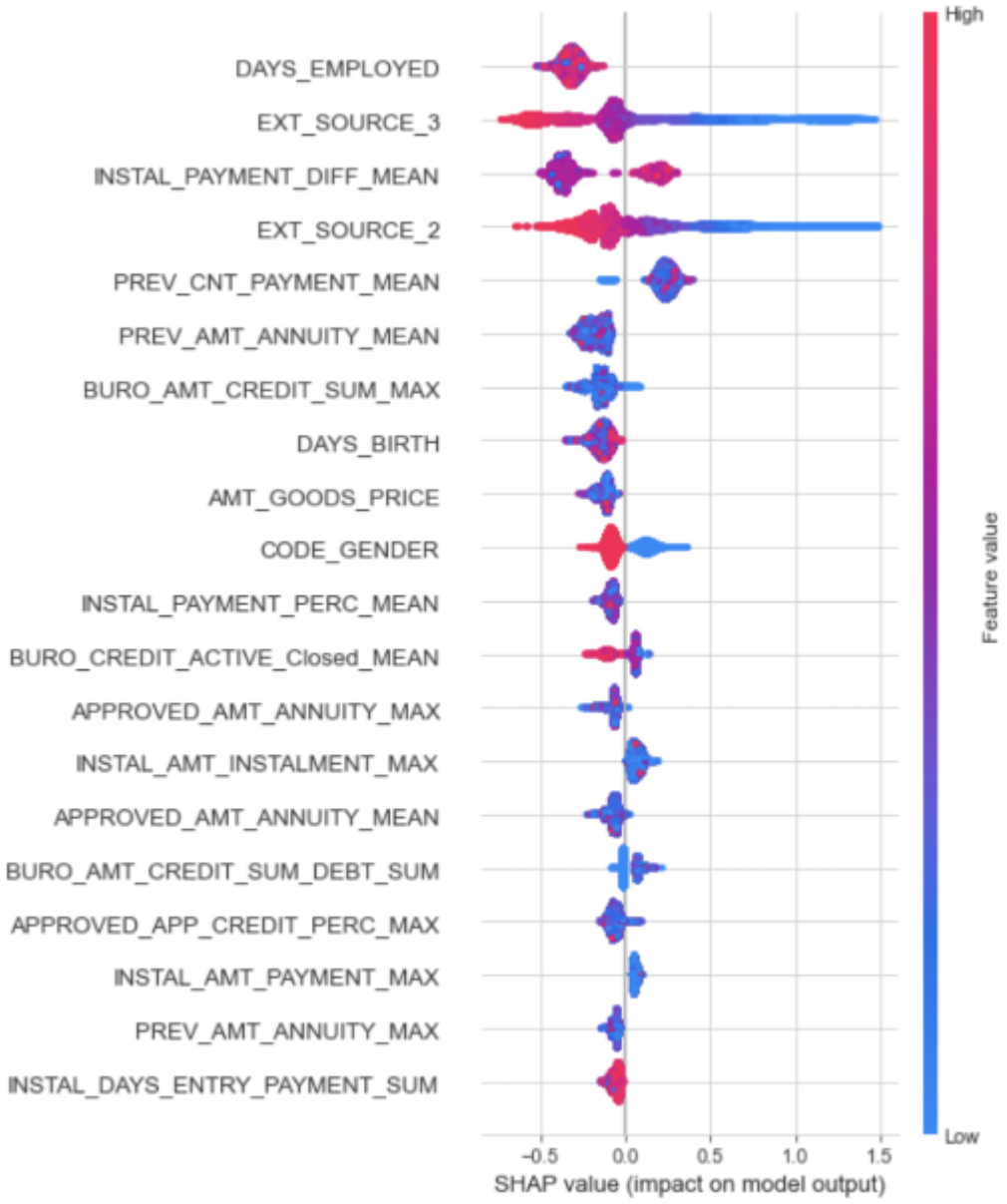


Pour cette raison, nous avons choisi un seuil de 0.9 pour notre tableau de bord.

La fonction de scoring choisie dans gridsearch pour l'optimisation était le recall. Cette fonction est une mesure adaptée à notre cas, car elle maximise le nombre de résultats positifs exacts, étant définie comme la fraction des vrais positifs sur les total des positifs.

Interprétabilité globale et locale du modèle

Nous avons utilisé le valeurs Shapley pour interpréter les prédictions du modèle. Cette approche permet de quantifier l'impact de chaque variable sur l'ensemble des prédictions (interprétabilité globale), mais aussi de faire de même pour un profil pris individuellement (interprétabilité locale). Cette approche dérivée de la théorie des jeux s'est répandue pour expliquer les prévisions d'un modèle complexe, comme il est typique en machine learning. Dans ce domaine, l'idée derrière son usage est que chaque variable contribue comme un joueur dont on peut estimer l'impact dans un contexte coopératif.
Pour notre modèle, à niveau global, les variables les plus signifiantes sont montrées à la suite:



Les variables identifiées dans le cadre de l'interprétabilité locale occupent une place prééminente dans le tableau de bord parce qu'elles indiquent les facteurs déterminants dans l'estimation de la probabilité d'insolvance.

Limites et améliorations possibles

À la suite, je présente une liste non exhaustive de points d'améliorations de notre approche.

- À niveau du modèle, on voit facilement des marges d'amélioration du point de vue de l'exposition financière de la société. Comme solutions, on pourrait implémenter une différenciation de la politique des prêts selon leur montant, et des niveaux de garde à ne pas dépasser sur une période donnée, mais cela dépasse la portée de ce projet.

- Le tableau de bord indique les variables à travers des noms codifiés. Elles nécessitent d'être accompagnées de leur définition pour être comprises.

- Du point de vue de l'architecture, le tableau de bord est fonctionnel seulement comme prototype. Il rencontrerait de grandes limites de mémoire et de sécurité dans un contexte réel.