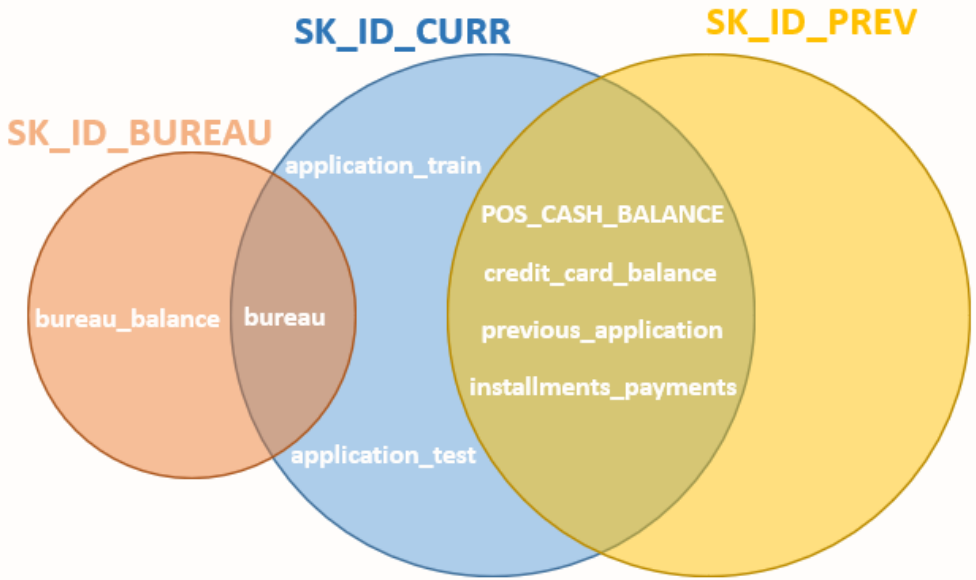
Contexte

Développer un algorithme de scoring pour aider à décider si un prêt peut être accordé à un client

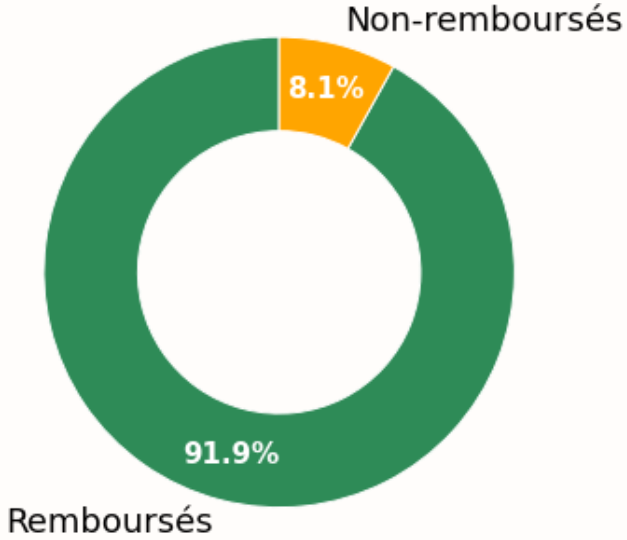
→ Modèle facilement interprétable
→ Mesure de l'importance des variables

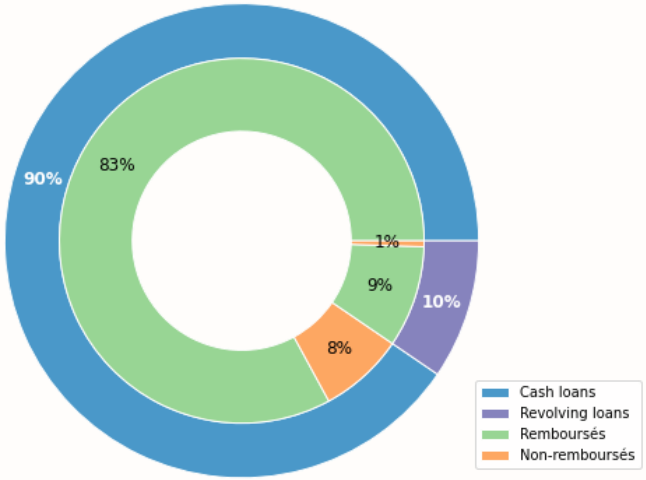
Analyse basée sur
les jeux de données
Home Credit Default Risk

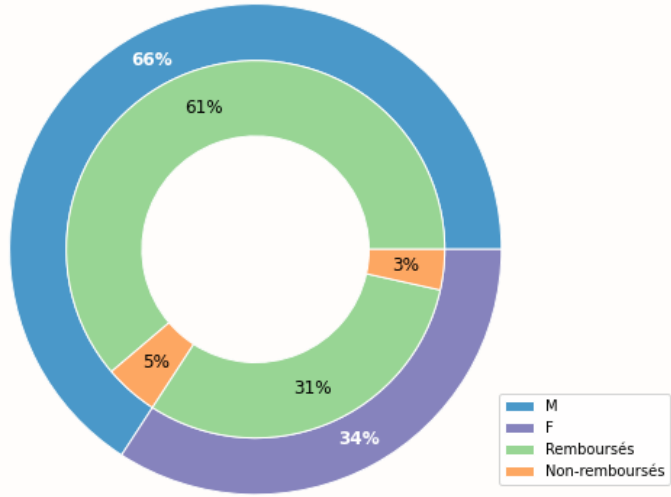
8 bases de données

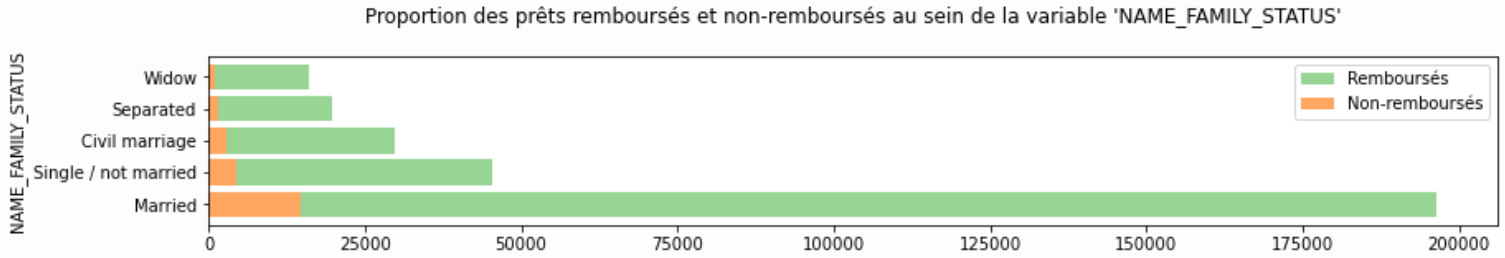


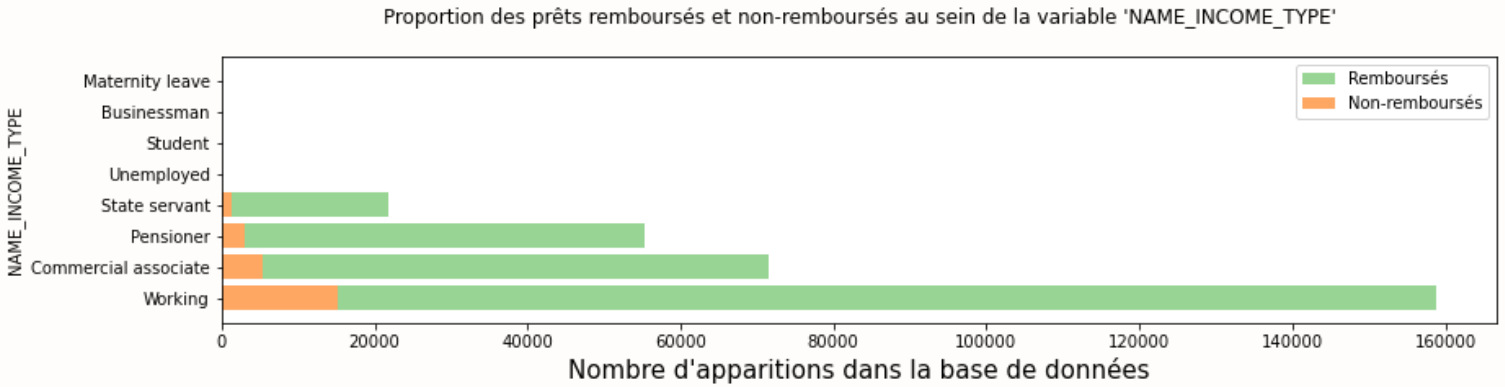
Description du jeu de données

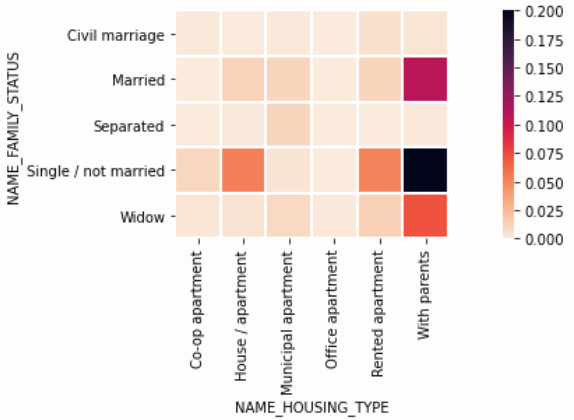






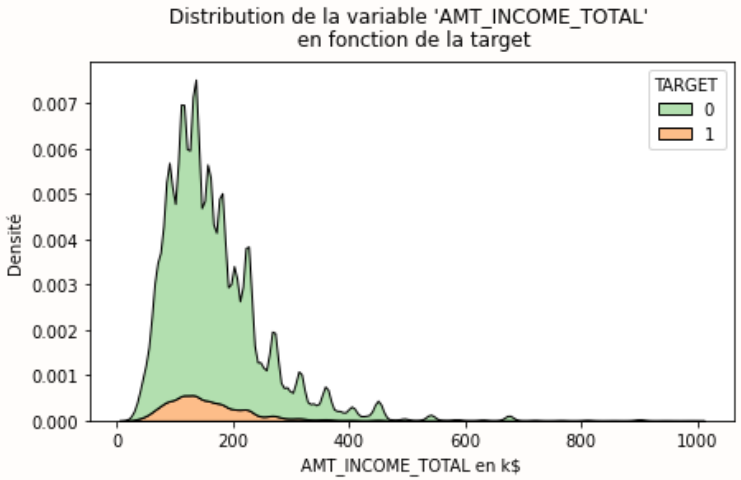


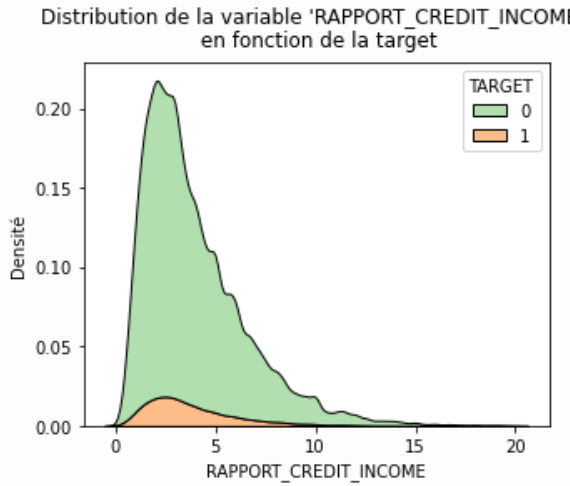




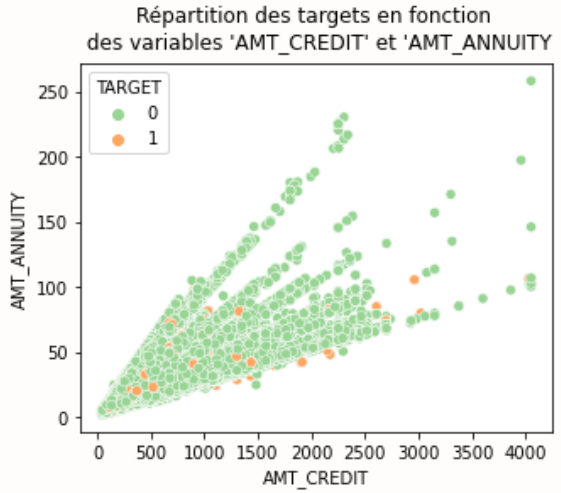
Transformation du jeu de données

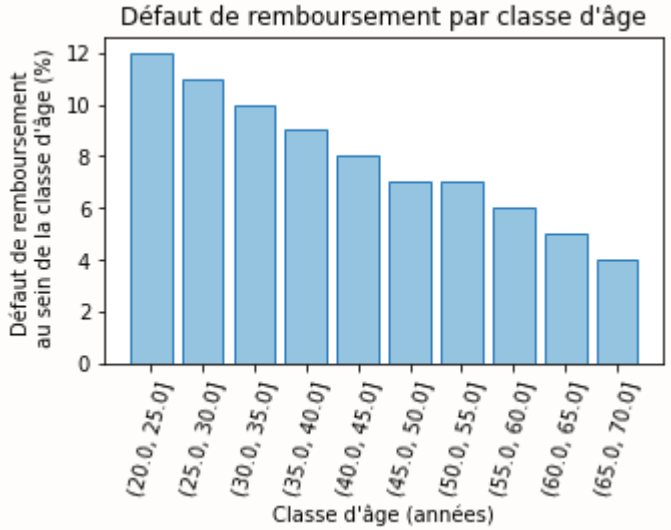
Les variables quantitatives 106 variables initiales

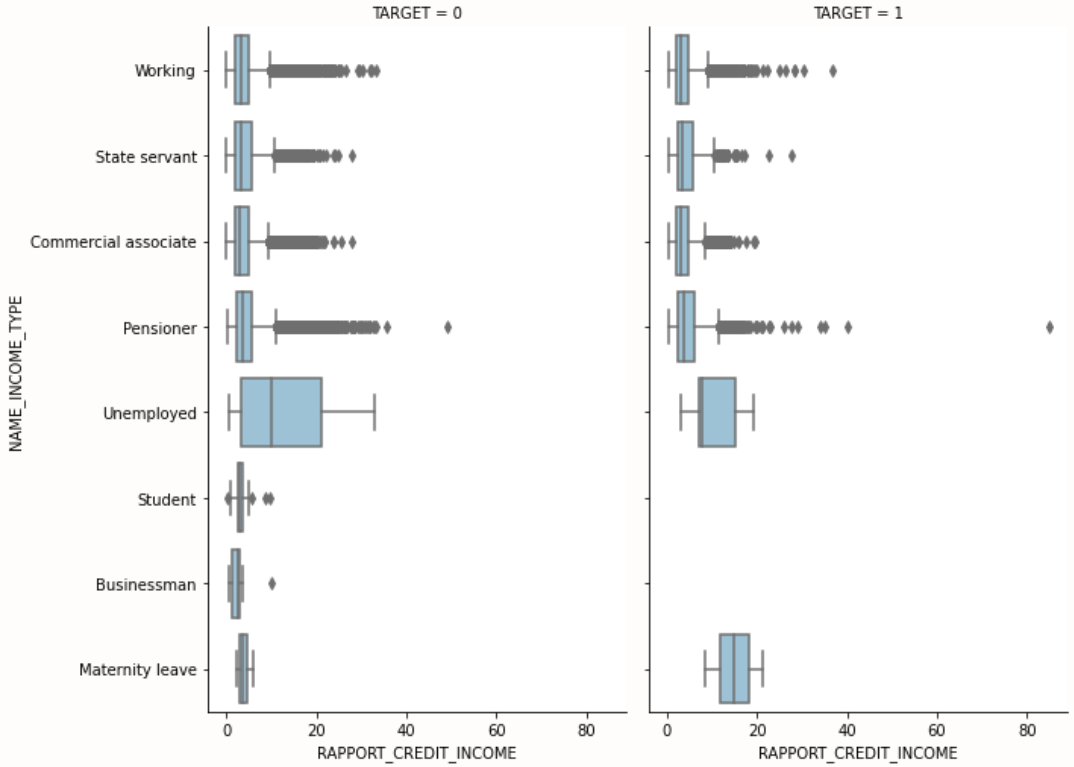




- **Suppression des valeurs impossibles (exemple : temps de travail de 1000 ans)**
- **Création de 15 nouvelles variables**
- **Visualisation des données**

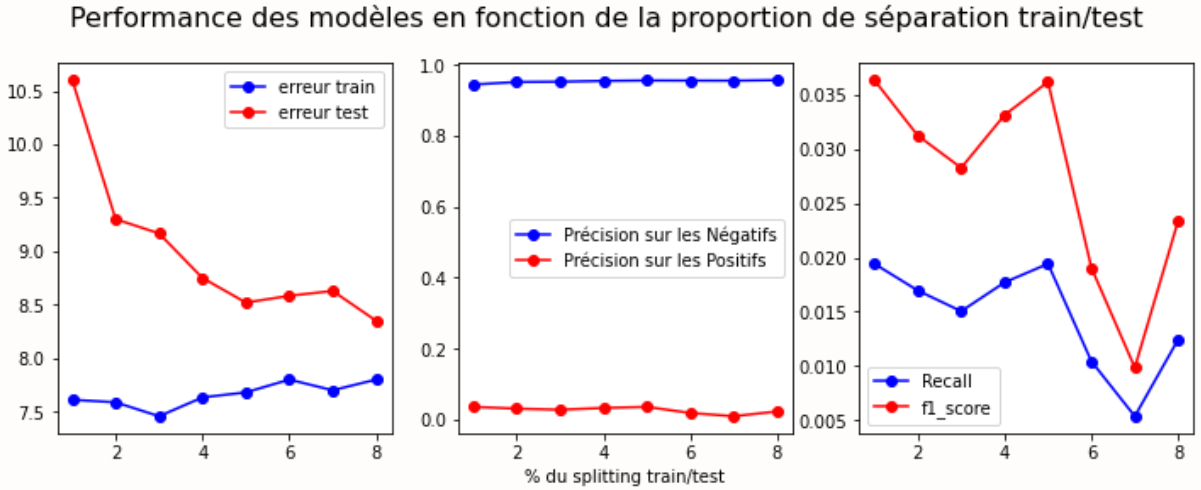




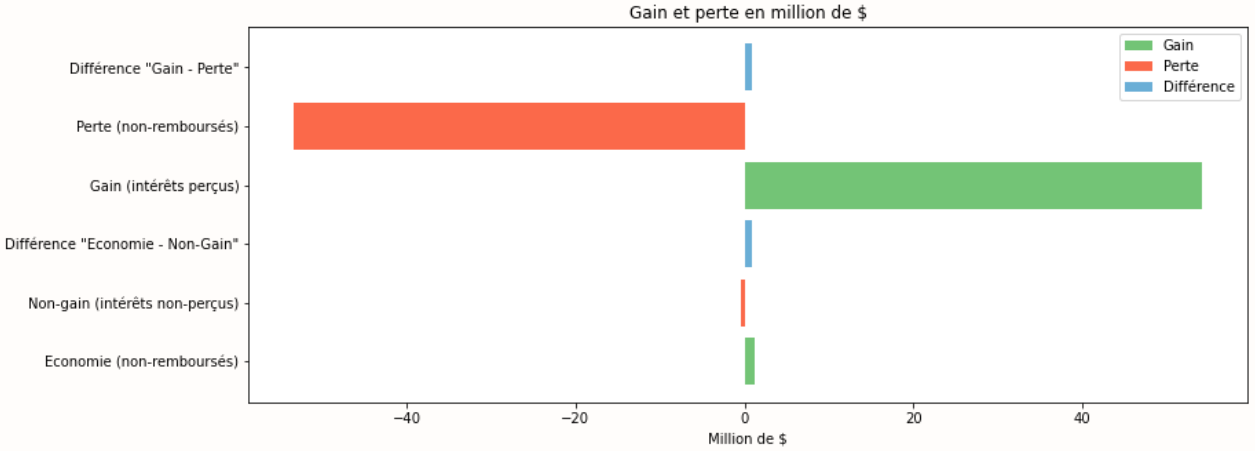


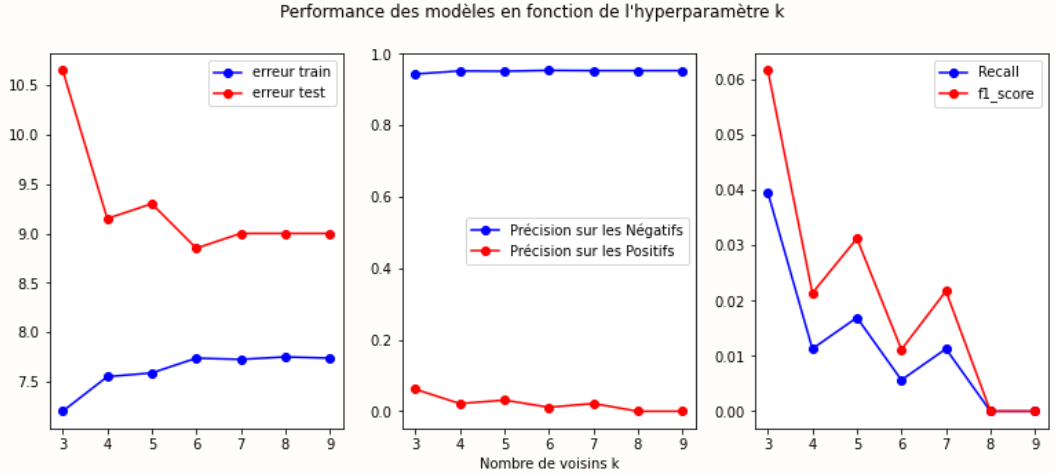
Comparaison et synthèse des résultats

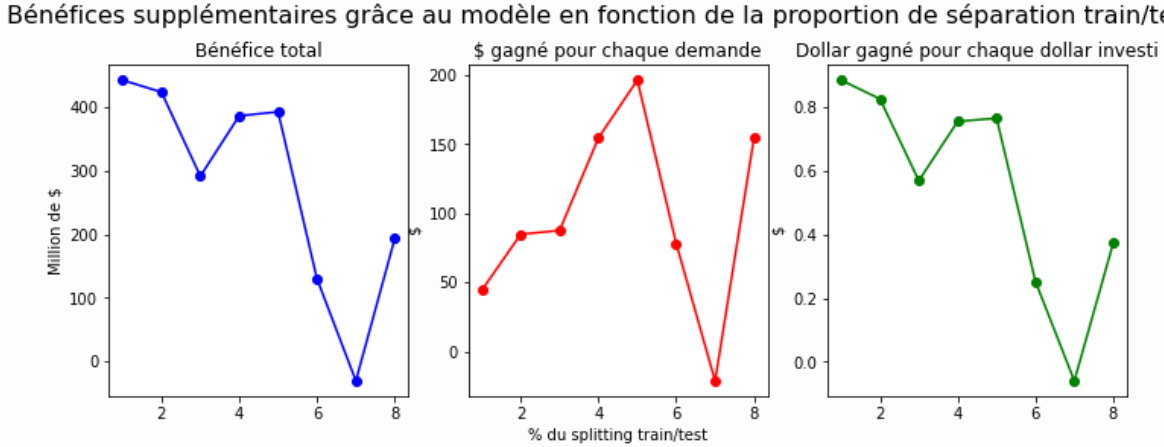
Modèle de comparaison : **KNN**



CALCUL DU BENEFICE SUPPLEMENTAIRE



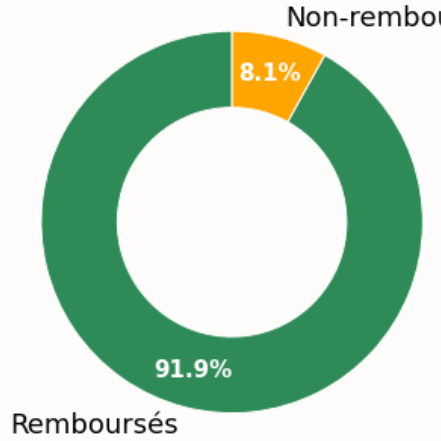




Comparaison et synthèse des résultats

Modèle de comparaison : **KNN**

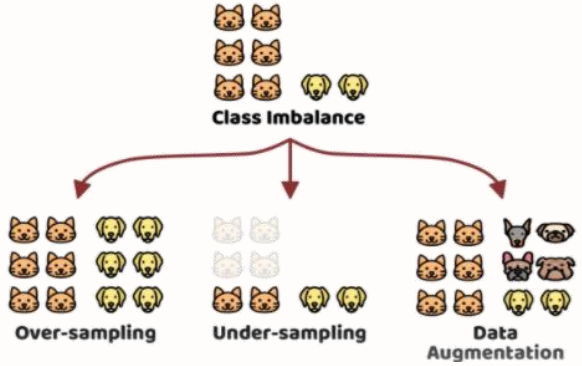
Proportion de prêts remboursés et non-remboursés



Fort déséquilibre

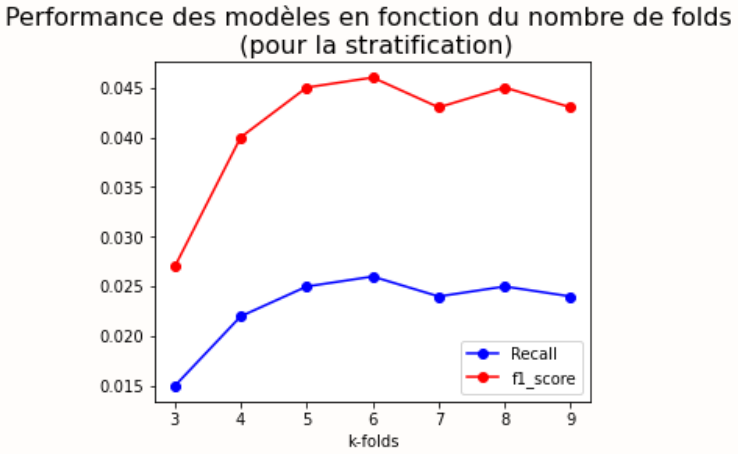
SMOTE / SVSMOTE

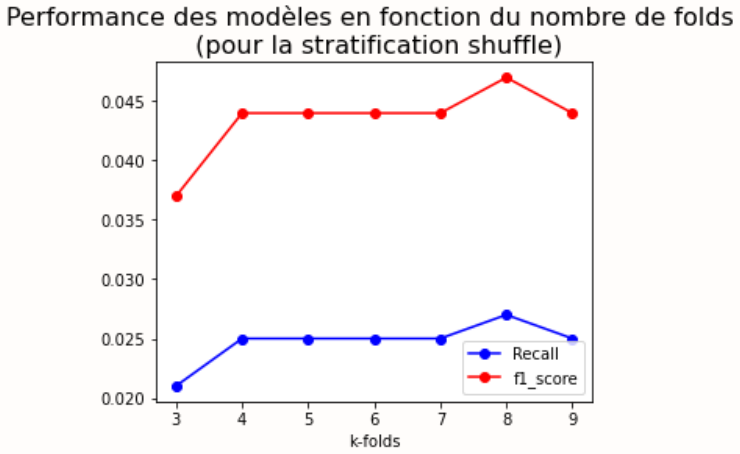
temps de calcul : 0,9 sec
Recall (1) = 0.59
Bénéfice = 11.5 M$



temps de calcul : 6.2 sec
Recall (1) = 0.45
Bénéfice = 11.0 M$

Stratification Kflod / shuffle

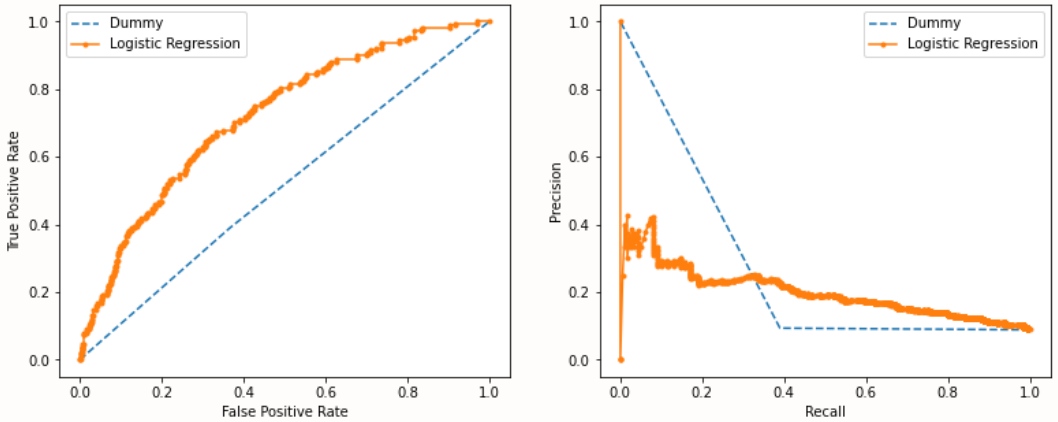




Comparaison et synthèse des résultats

Les modèles linéaires

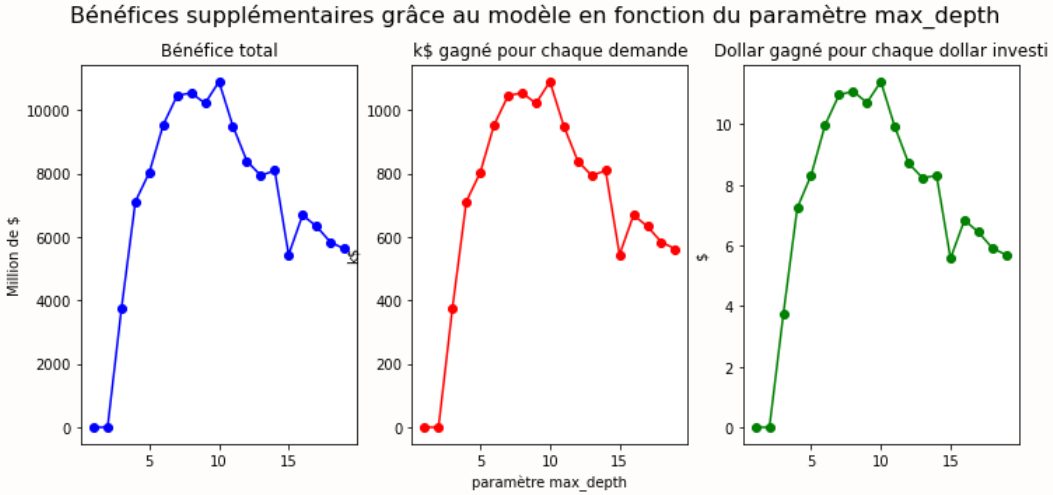
Logistic Regression



Recall (1) = 0.44 Bénéfice = 13.3 M$

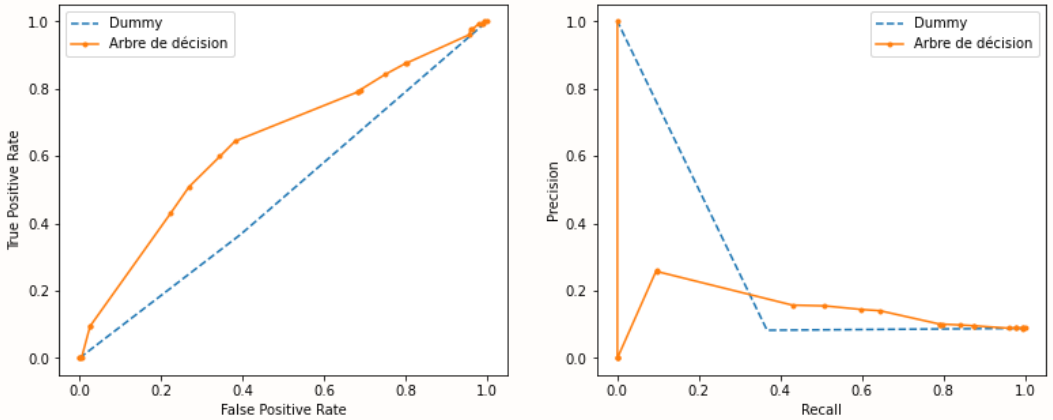
Random Forest Classifier

(réduisent la variance observée avec arbres de décision)

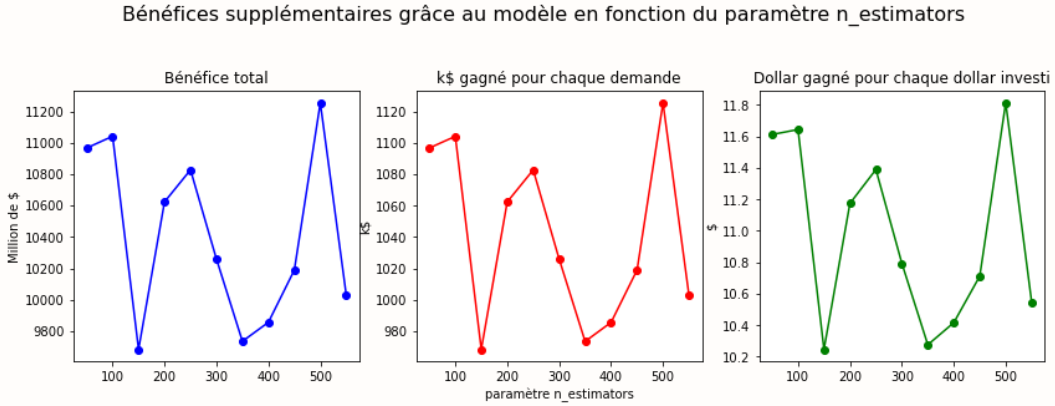


Les modèles non- linéaires

Arbre de décision



Recall (1) = 0.64 Bénéfice = 13.1 M$



Recall (1) = 0.31 Bénéfice = 10.4 M$

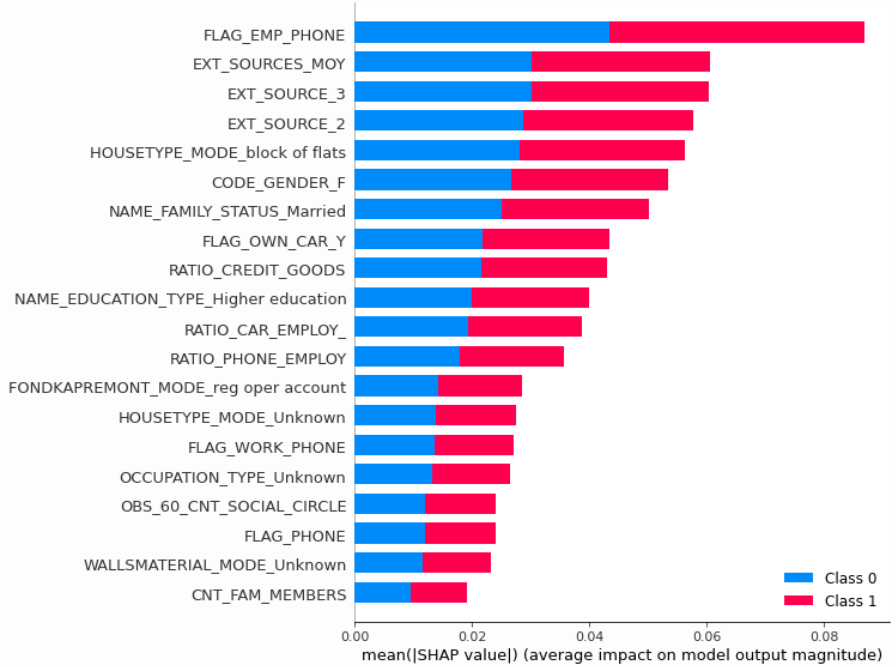
Comparaison et synthèse des résultats

Synthèse des résultats

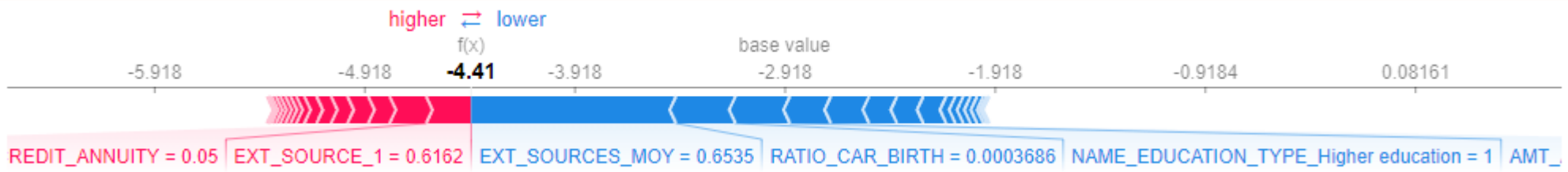
modeles	temps de calcul	recall	mesure f1	roc auc	benefice total (en k$)	benefice par demande (en $)	benefice par dollar (en $)
KNN	0.9	0.44	0.23	0.16	12877.0	1287.72	15.81
Logistic Regression	0.4	0.44	0.27	0.20	13307.0	1330.70	15.10
SVM	37.4	0.14	0.15	0.13	4486.0	448.64	4.64
Arbre de décision	0.3	0.64	0.23	0.15	13116.0	1311.58	19.76
Random Forest	3.3	0.31	0.25	0.18	10429.0	1042.91	10.96
Gradient Boosting Classifier	4.3	0.44	0.27	0.19	15067.0	1506.73	17.07
XGBClassifier	8.4	0.42	0.28	0.19	16358.0	1635.82	17.89
LGBMClassifier	1.3	0.40	0.29	0.18	16083.0	1608.30	17.22
MLPClassifier	2.4	0.45	0.28	0.20	13772.0	1377.19	15.47
Voting	6.8	0.43	0.27	0.19	13912.0	1391.15	15.60

Interprétabilité du modèle

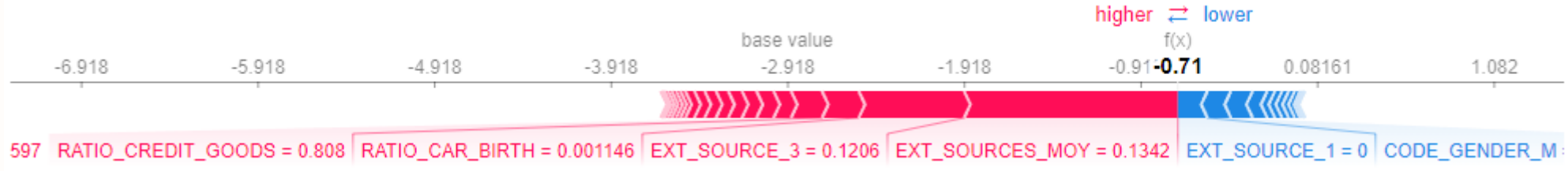
Random Forest Classifier



Target 0 remboursé



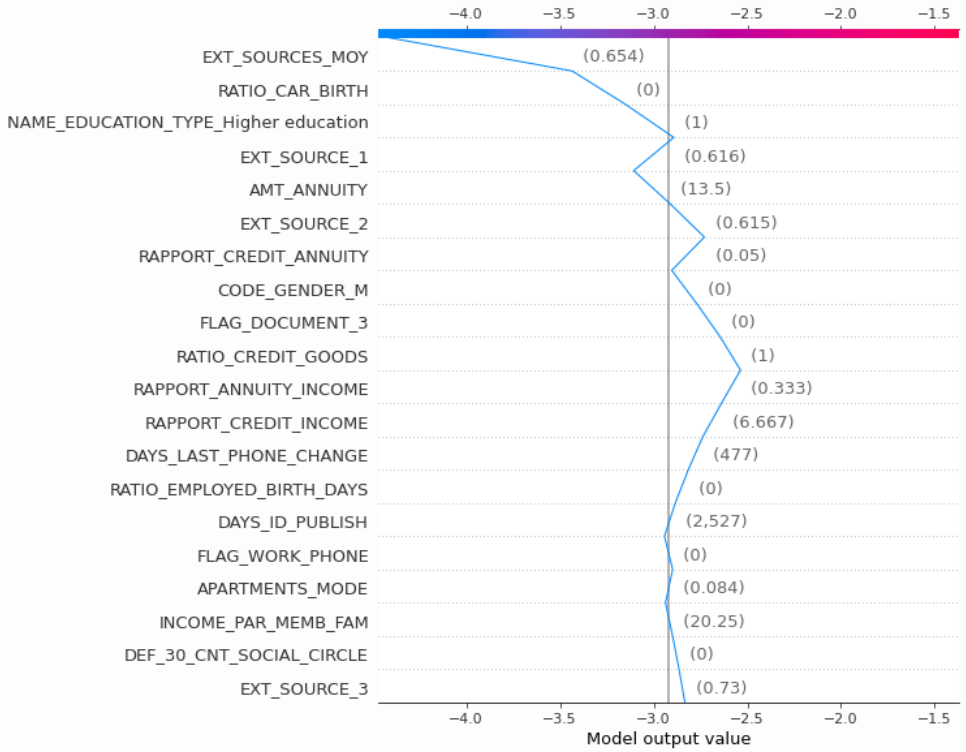
Target 1 non-remboursé



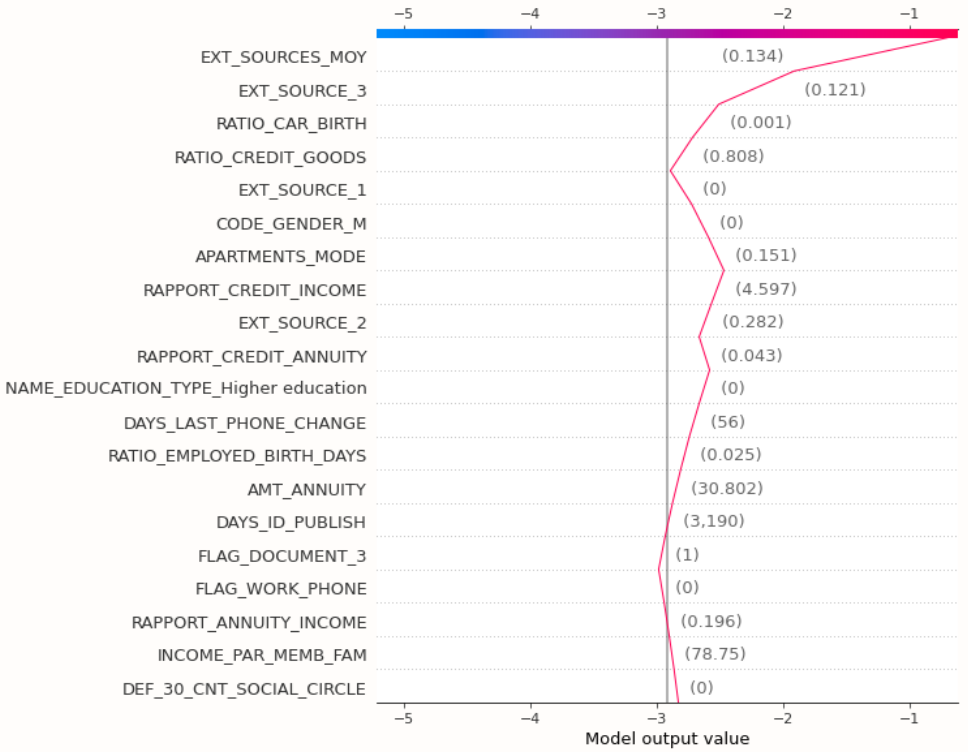
Interprétabilité du modèle

Gradient Boosting Classifier

Target 0 : remboursé



Target 1 : non-remboursé



Attention aux données privées et discriminatoires !! (exemple : CODE_GENDER)

Conclusion

- Les nouvelles variables améliorent le modèle
  (Bénéfice supplémentaire augmenté de 18% avec les nouvelles variables)

- Evaluation du modèle :
  - Score standard : recall, precision, f-measure
  - Calcul du bénéfice supplémentaire engendré par l'utilisation du modèle

- Meilleure combinaison de modèle :
  - Logistic Regression (modèle linéaire)
  - Arbre de Décision
  - Gradient Boosting Classifier (modèle ensembliste)
  - MLP (réseaux de neurones artificiels – deep learning)

- Interprétabilité du modèle :
  Possibilité de mettre en place une méthode globale (pour le modèle) et une méthode locale (pour un client donné) de mesure de l'importance des variables

Réflexion

- Le meilleur modèle est-il le moins éthiquement et juridiquement discriminant ?
- Comment justifier l'accès aux données privées et leur utilisation (genre, origine ethnique et sociale) ?
- Comment rendre compte des aspects humains non quantifiables ? Est-ce éthique ?