SOUTENANCE

CREDIT SCORING

Cédric Dietzi

Sommaire

- Contexte
- Jeu de données
- Transformation des données
- Evaluation des modèles
- Analyse métier et conclusion

Contexte

La société « Prêt à dépenser » souhaite disposer d'un outil d'aide à la décision d'octroi de prêts

Utilisé par les chargés de clientèle

Simple à interpréter

Le jeu de données

Prêt à dépenser

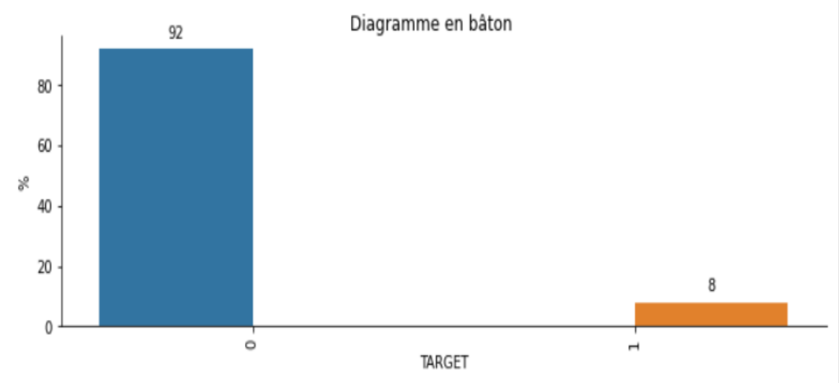
307 000 CLIENTS

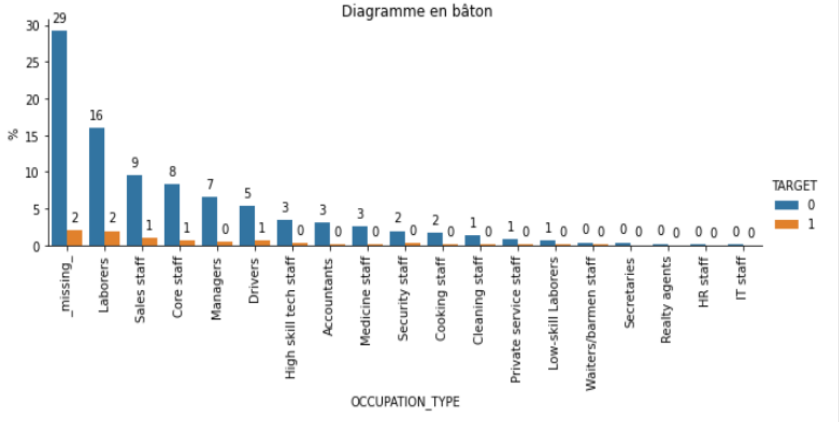
7 FICHIERS

220 VARIABLES

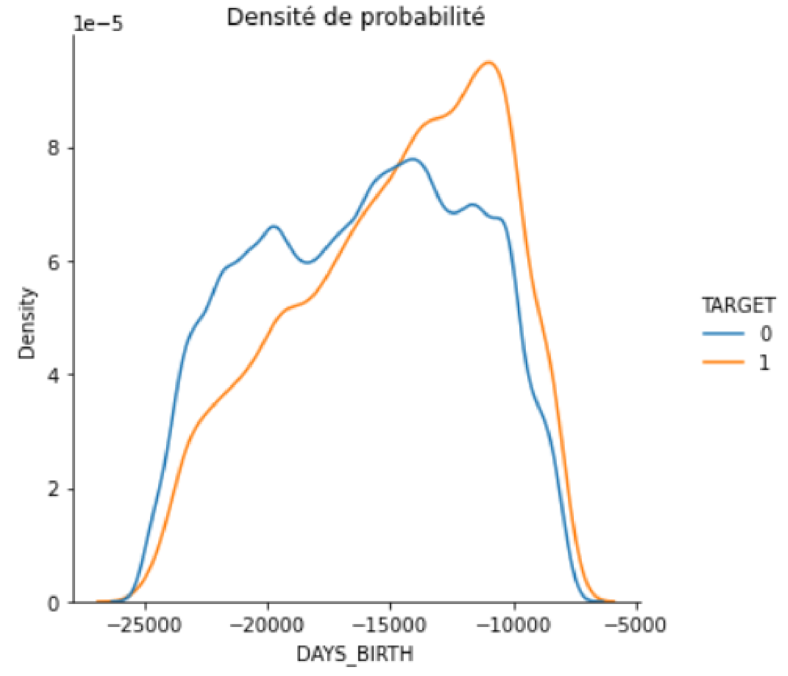
CLIENTS, PRÊTS ACTUELS,
PRÊTS ANTÉRIEURS

Analyse exploratoire de données

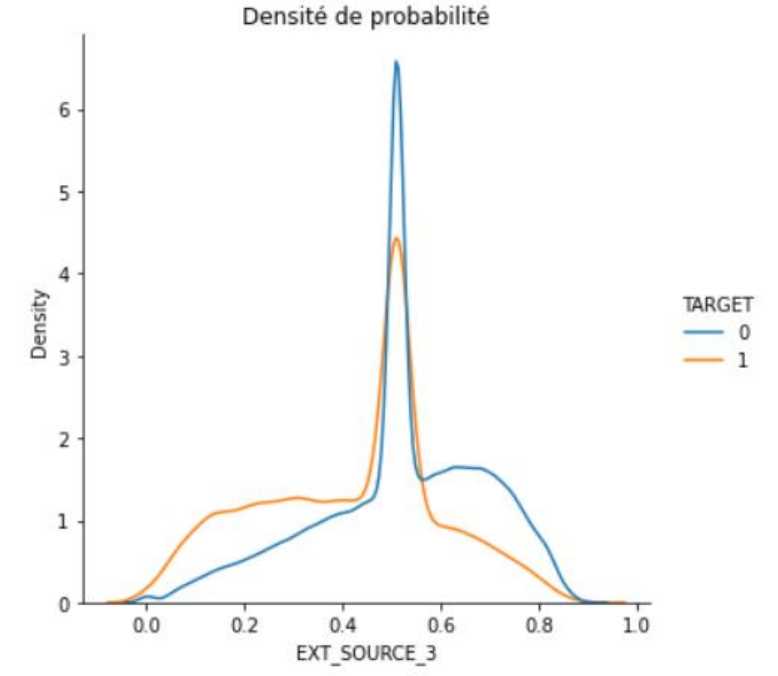


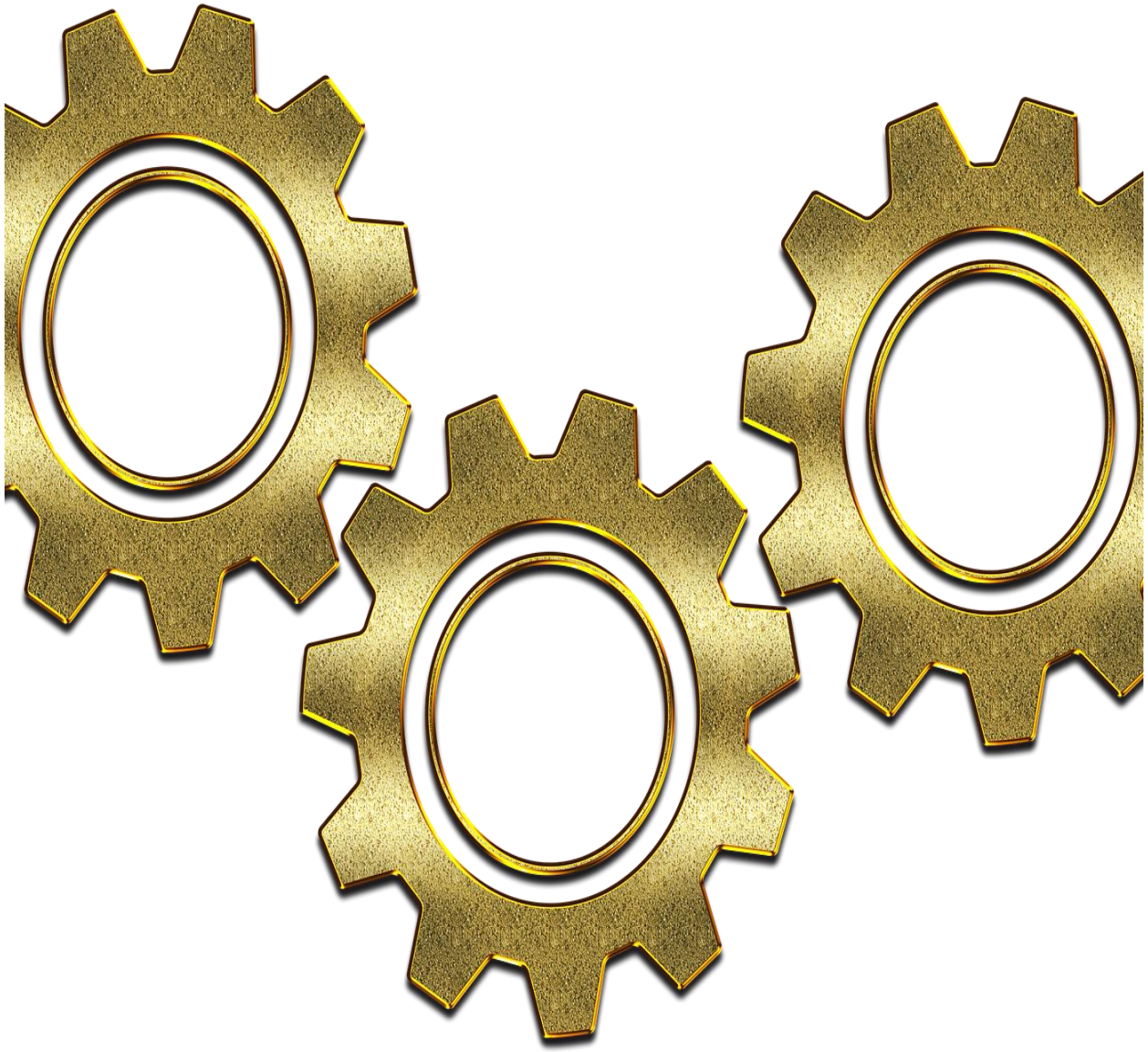


	Moyenne	Ecart type
0	-16136.593	4360.669
1	-14878.819	4189.156



	Moyenne	Ecart type
0	0.519	0.171
1	0.418	0.187





Transformation des données

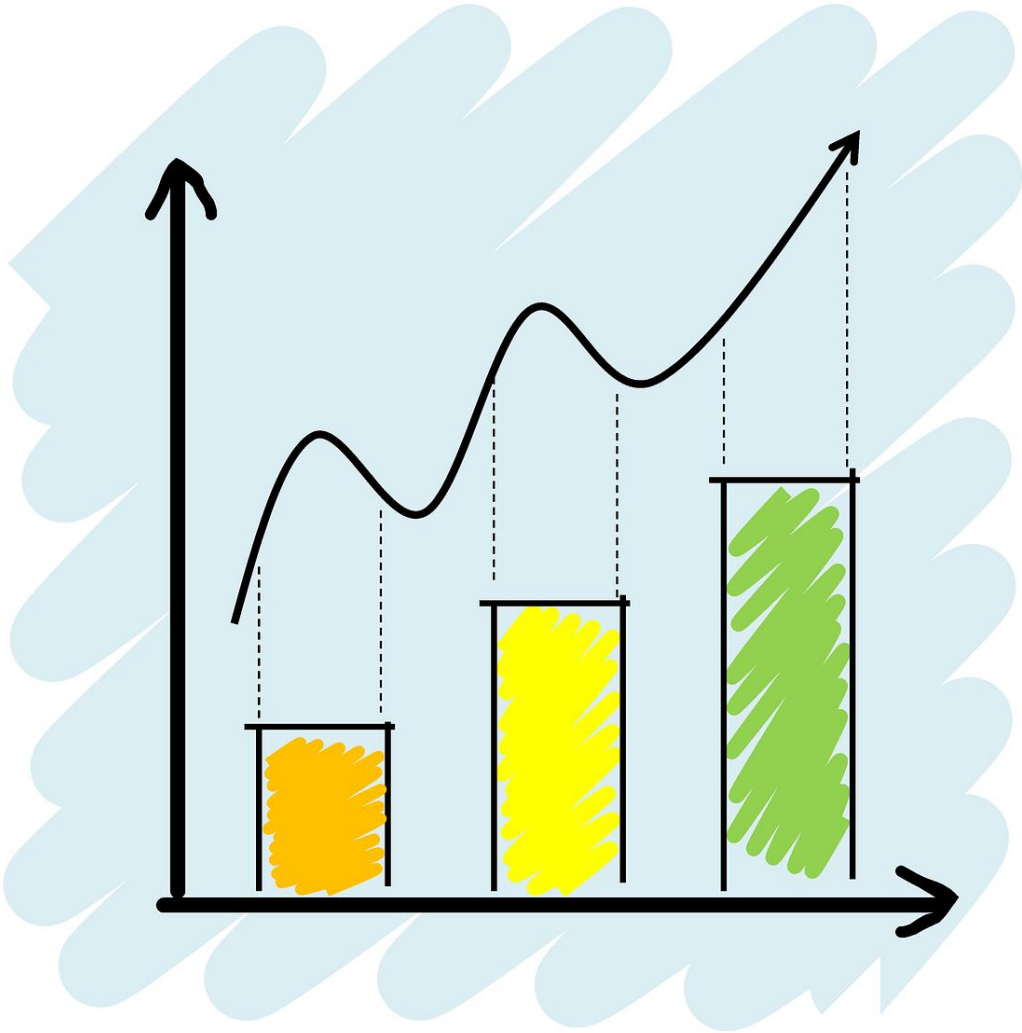
Features engineering

- Ratios
- Statistiques

Nettoyage

- Valeurs aberrantes remplacées par nan
- Imputation des valeurs manquantes qualitatives par « _missing_ »
- Imputation des valeurs manquantes quantitatives par la moyenne
- Encodage des variables qualitatives binaires
- Encodage des variables qualitatives non binaires
- Normalisation et Centrage

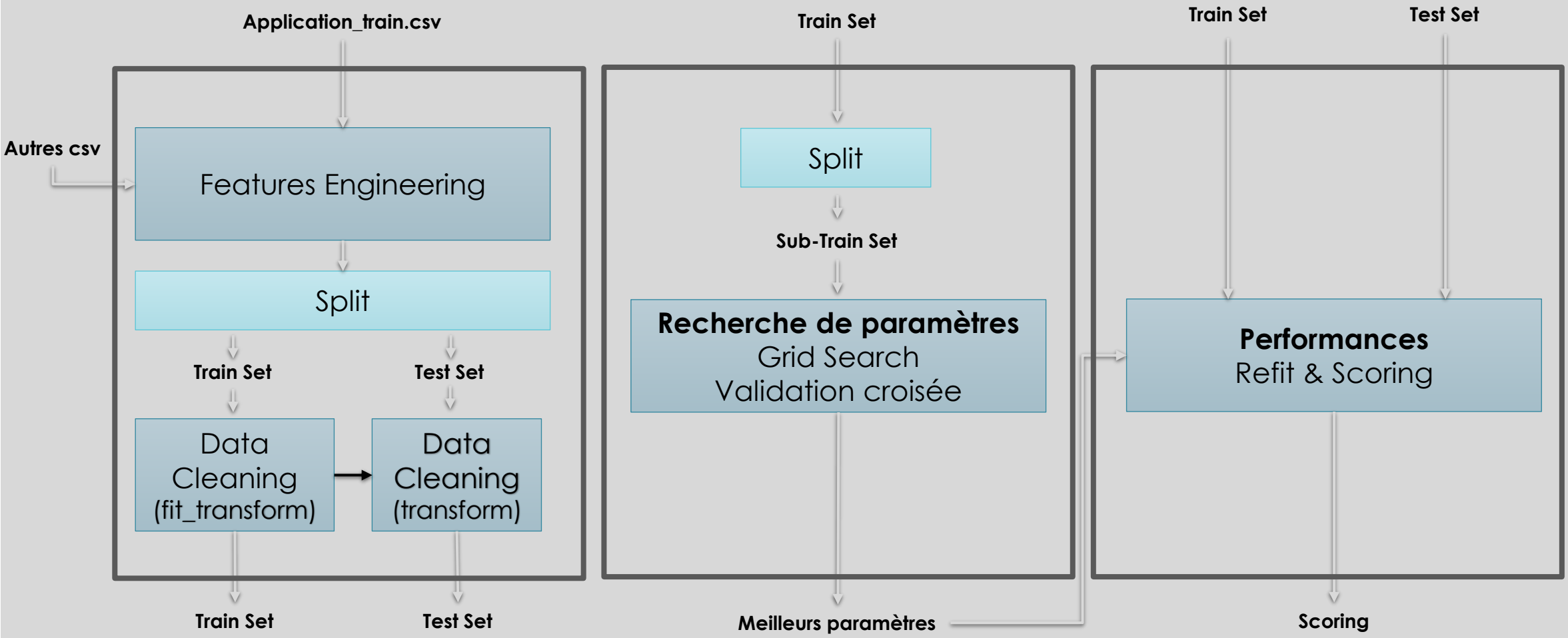
On passe de 121 à 877 variables

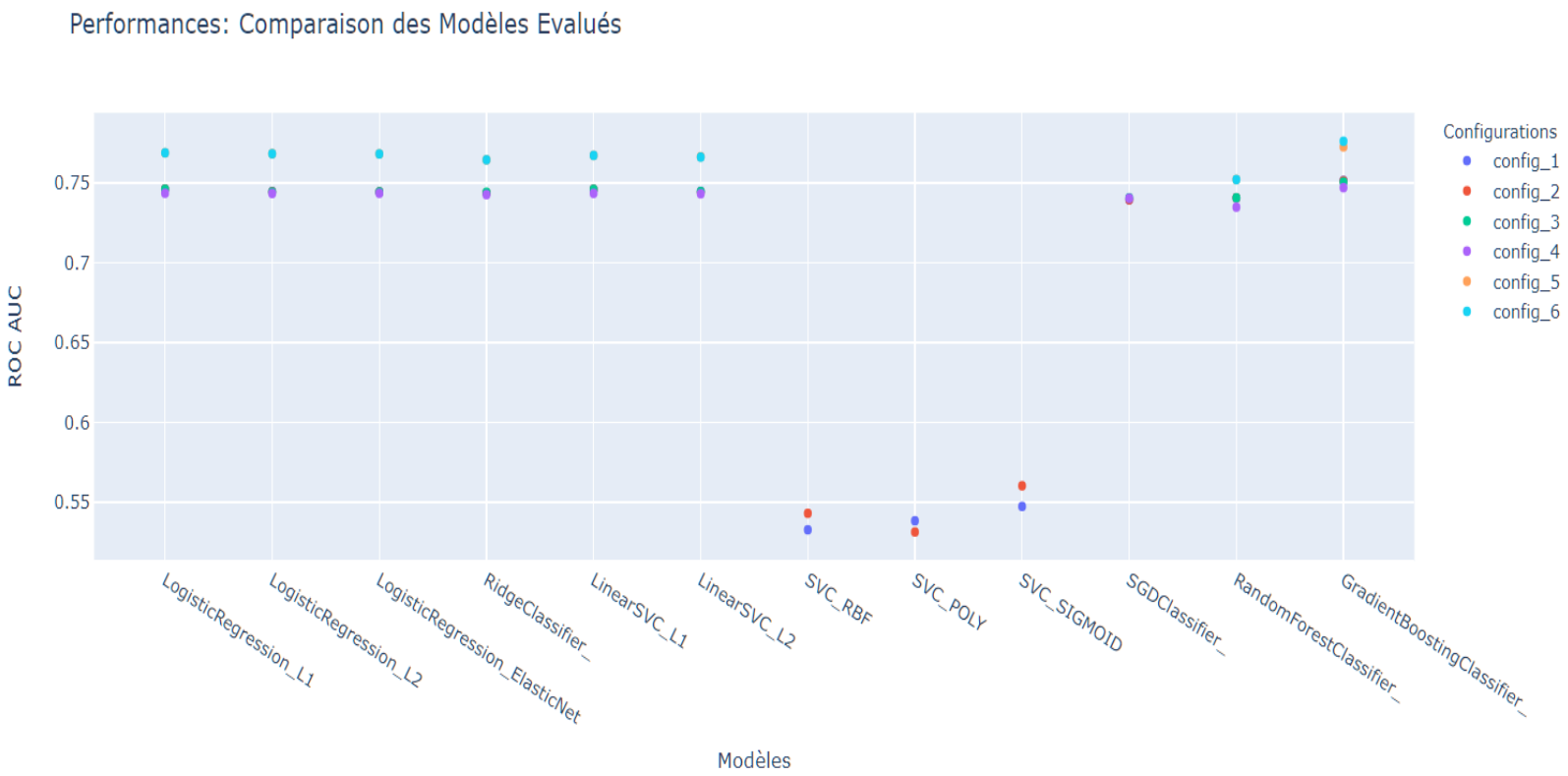


Evaluation de modèles

& Résultats

Processus d'évaluation des modèles





Modèle	Score
GradientBoostingClassifier_	0.776124
LogisticRegression_L1	0.768871
LogisticRegression_L2	0.768285
LogisticRegression_ElasticNet	0.768188
LinearSVC_L1	0.767243
LinearSVC_L2	0.766188
RidgeClassifier_	0.764495
RandomForestClassifier_	0.752121

Meilleur candidat technique
Gradient Boosting

Performances sur jeu de test

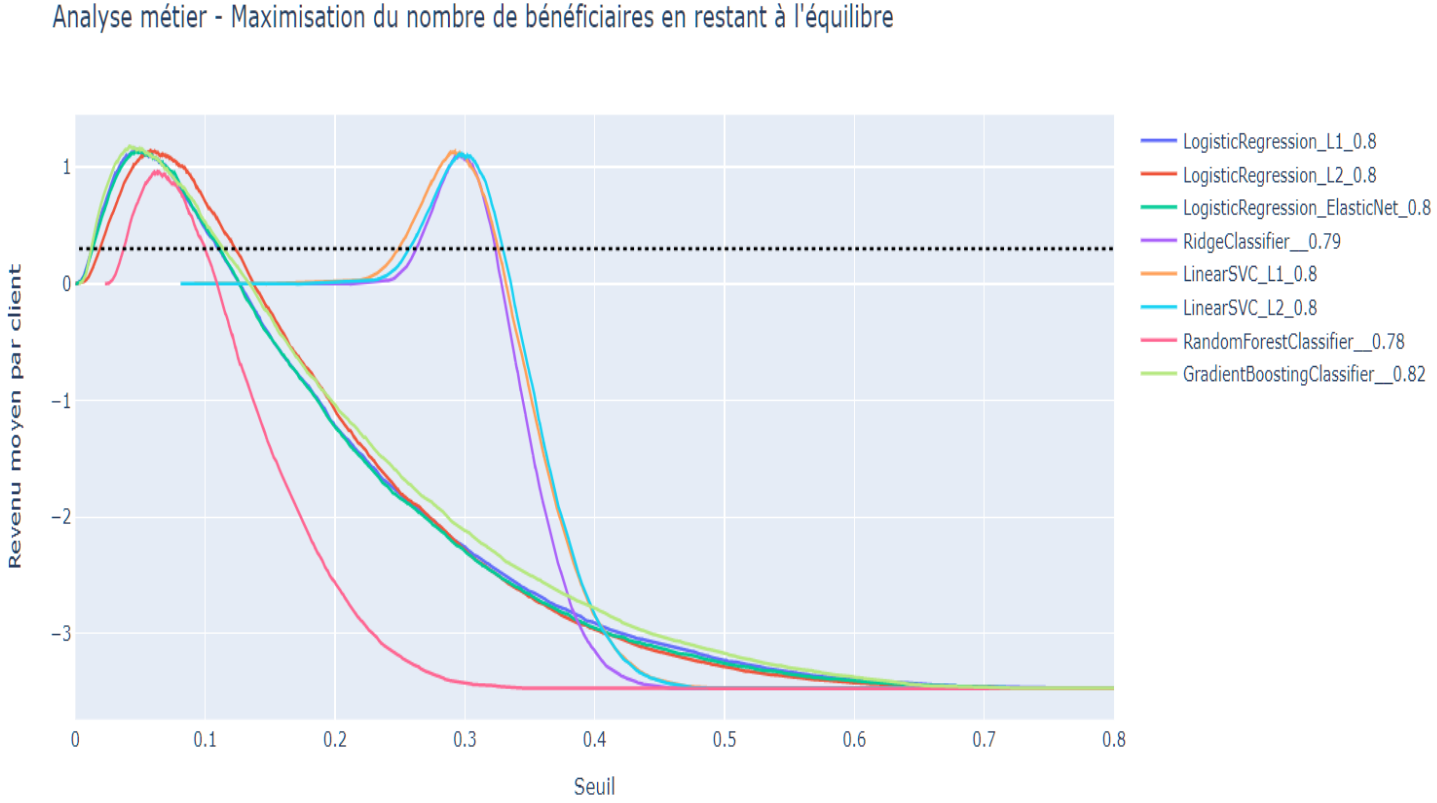
Mesure: ROC AUC. Aire sous la courbe d'efficacité.

Score d'une decision aléatoire: 0,5.

Config_1 et 2: paramètres estimés sur jeu réduit

Config_3 et 4: paramètres confirmés sur un jeu plus large

Config_5 et 6: impact de l'ajout de variables construites



Modèle	Score bénéficiaire
GradientBoostingClassifier_	0,82
LogisticRegression_L1	0,8028
LogisticRegression_ElasticNet	0,8016
LogisticRegression_L2	0,8008
LinearSVC_L1	0,7979
LinearSVC_L2	0,7974
RidgeClassifier_	0,7935
RandomForestClassifier_	0,778

Meilleur candidat métier
Gradient Boosting

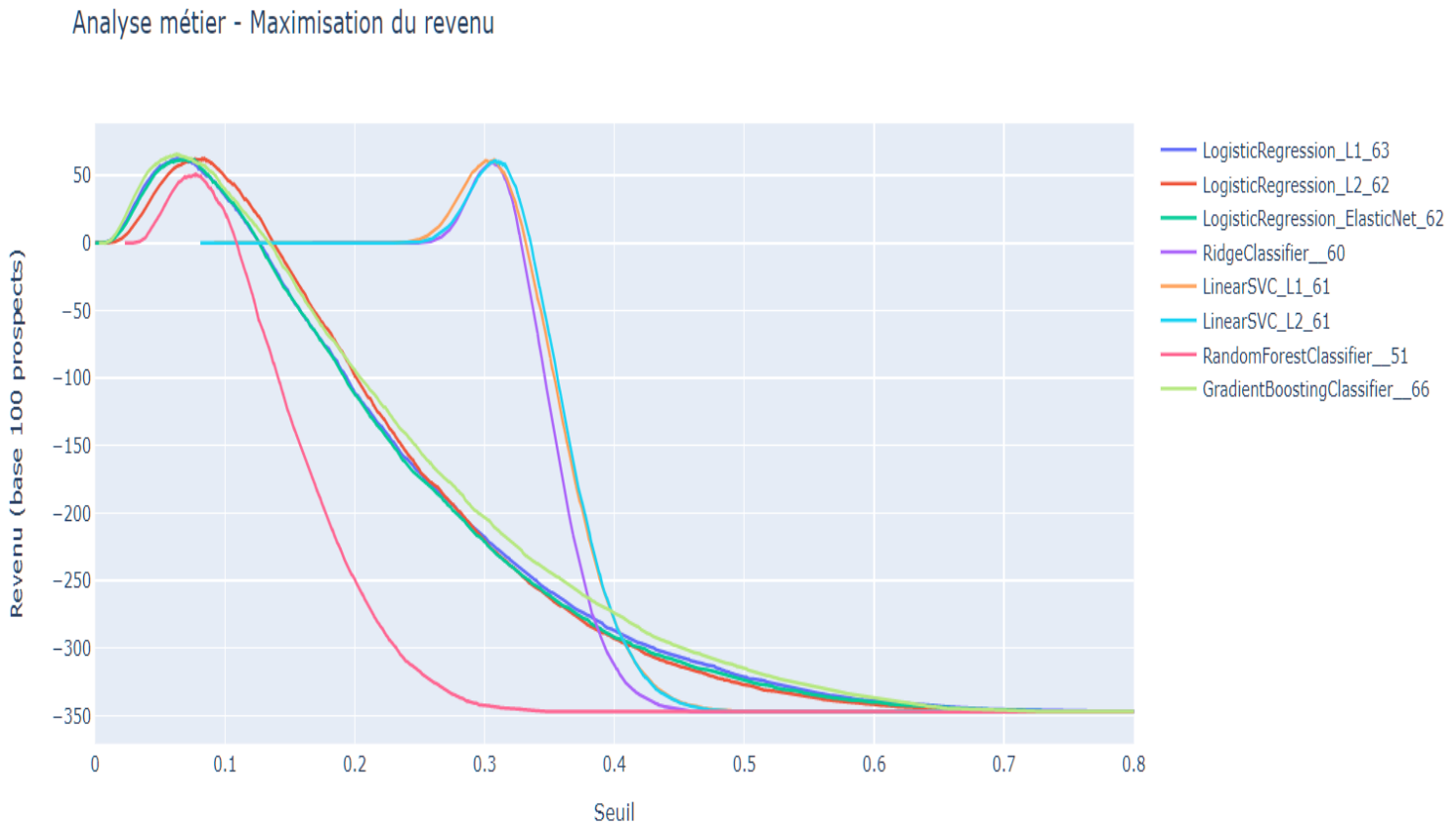
Analyse métier

Un bon client rapporte 5

Un mauvais client fait perdre 100

Coûts fixes par client: 0,3

=> Comment maximiser le **nombre de bénéficiaires** en restant à l'équilibre ?



Modèle	Score revenu
GradientBoostingClassifier_	65,5277
LogisticRegression_L1	62,6343
LogisticRegression_L2	62,4049
LogisticRegression_ElasticNet	62,2272
LinearSVC_L1	61,0036
LinearSVC_L2	60,7896
RidgeClassifier_	60,1688
RandomForestClassifier_	50,87

Meilleur candidat métier
Gradient Boosting

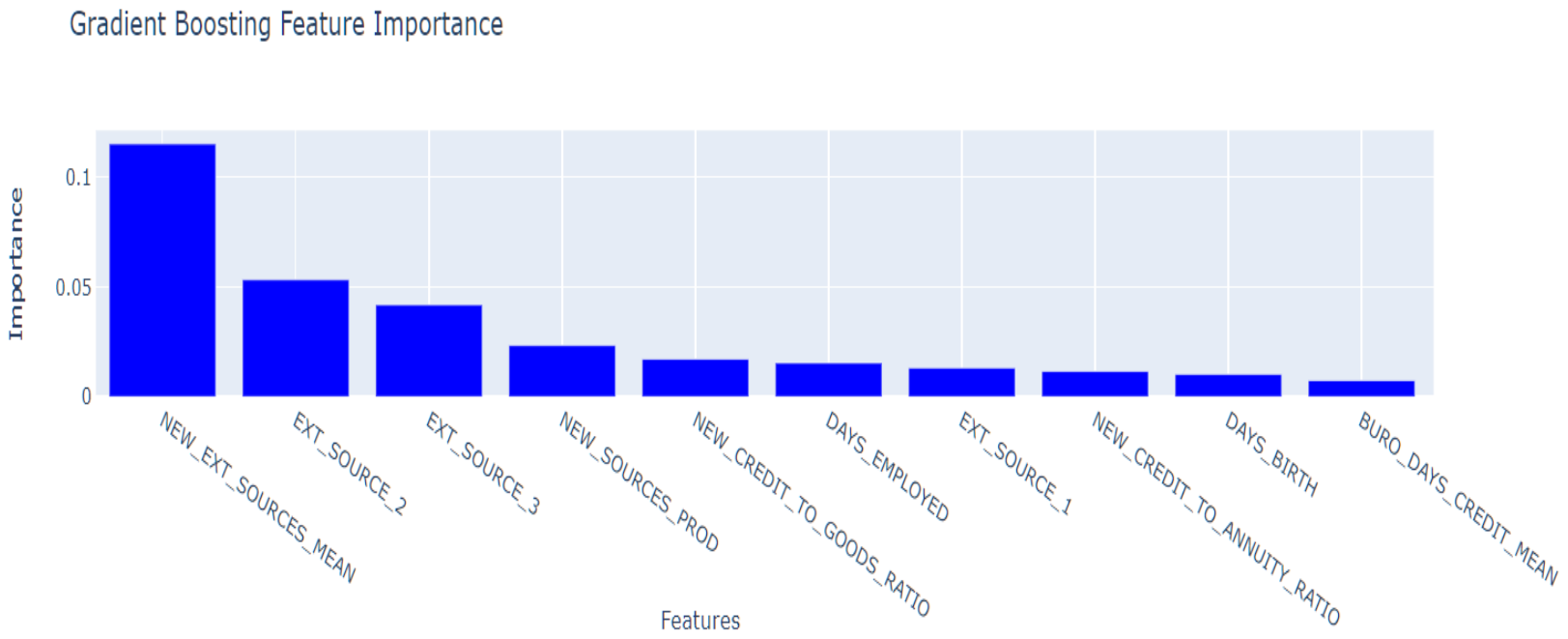
Analyse métier

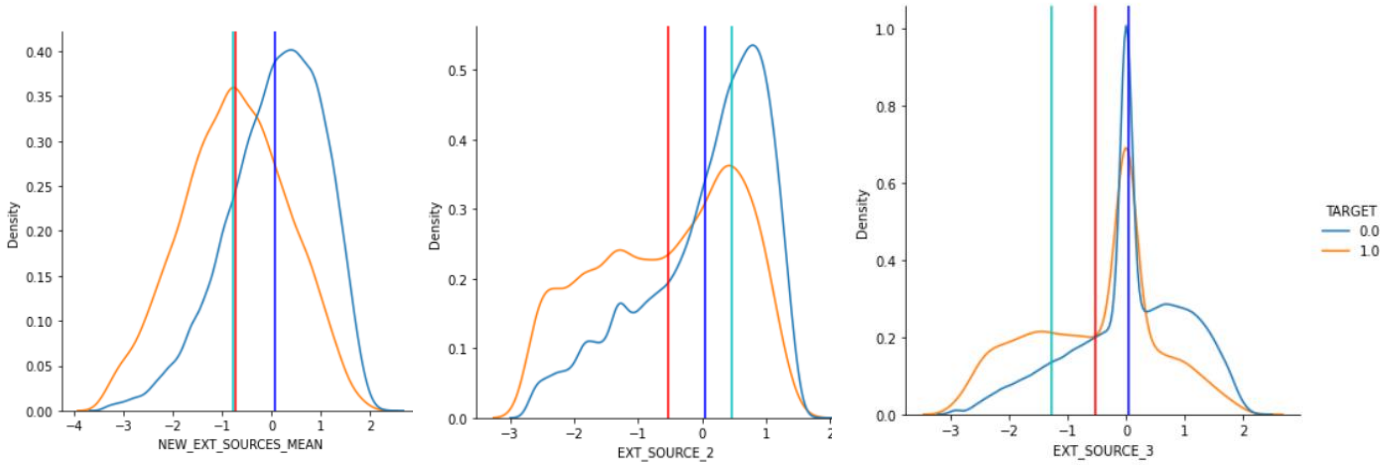
Un bon client rapporte 5

Un mauvais client fait perdre 100

Coûts fixes par client: 0,3

=> Comment maximiser le **revenu** ?





Exemple de prediction et interprétation

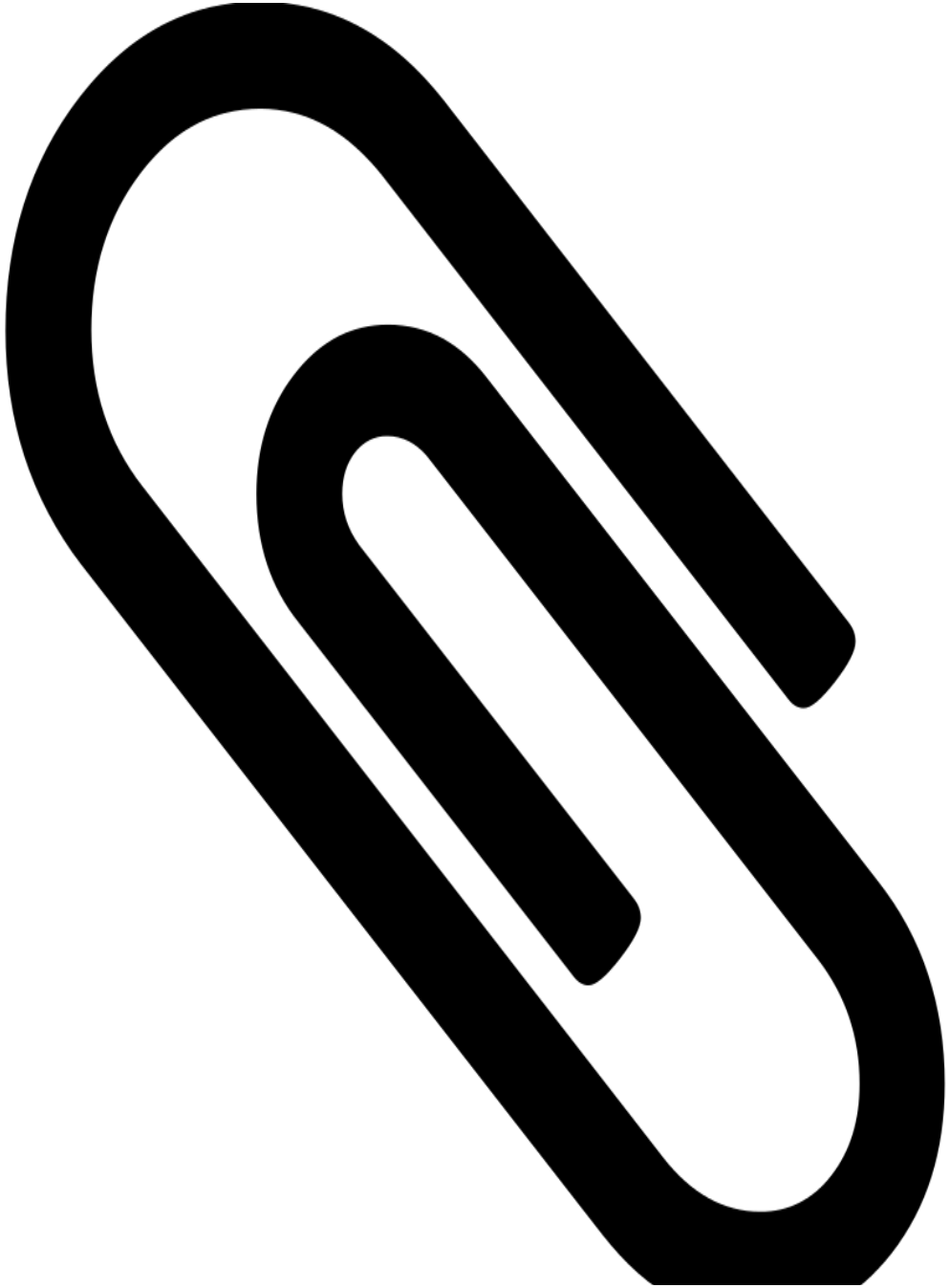
Un client est rejeté. On interprète la décision en fonction de sa position sur les distribution des variables les plus importantes.

Graphs:

- Rouge - Moyenne des mauvais clients
- Bleu – Moyenne des bons clients
- Cyan – Client évalué

Conclusion

- Le modèle Gradient Boosting est le plus performant d'un point de vue métier pour les deux cas d'usage envisagés.
- Les variables les plus déterminantes dans la décision sont liées aux sources d'information externe.
- Une décision particulière peut-être interprétée en visualisant la position du demandeur par rapport aux distributions de ces variables les plus déterminantes.
- Dans les deux cas d'usage envisagés, le modèle permet de faire passer la société d'une position de perte à l'atteinte de son objectif de maximisation des bénéficiaires ou du revenu.



Annexe Guide d'utilisation

Recherche de paramètres

Exemple sur un modèle de regression logistique avec régularisation L2.