**Overview :**

Le but de cette compétition est de prédire quels clients sont plus susceptibles de faire défaut sur leurs prêts. L'évaluation favorisera les solutions qui sont stables dans le temps.

Votre participation pourrait offrir aux fournisseurs de services financiers aux consommateurs un moyen plus fiable et durable d'évaluer le risque de défaut d'un client potentiel.

**Description :**

L'absence d'historique de crédit peut signifier beaucoup de choses, y compris un jeune âge ou une préférence pour l'argent liquide. Sans données traditionnelles, quelqu'un avec peu ou pas d'historique de crédit est susceptible de se voir refuser un prêt. Les fournisseurs de services financiers aux consommateurs doivent déterminer avec précision quels clients peuvent rembourser un prêt et lesquels ne le peuvent pas, et les données sont essentielles. Si la science des données pouvait aider à mieux prédire les capacités de remboursement d'une personne, les prêts pourraient devenir plus accessibles à ceux qui en bénéficieraient le plus.

Actuellement, les fournisseurs de services financiers aux consommateurs utilisent diverses méthodes statistiques et d'apprentissage automatique pour prédire le risque de prêt. Ces modèles sont généralement appelés scorecards. Dans le monde réel, les comportements des clients changent constamment, donc chaque scorecard doit être régulièrement mis à jour, ce qui prend du temps. La stabilité du scorecard à l'avenir est cruciale, car une baisse soudaine de ses performances signifie que les prêts seront accordés à des clients de moins bonne qualité en moyenne. Le cœur du problème est que les fournisseurs de prêts ne sont pas en mesure de repérer les problèmes potentiels avant que les premières échéances de ces prêts ne soient observables. Étant donné le temps nécessaire pour re-développer, valider et mettre en œuvre le scorecard, la stabilité est très souhaitable. Il y a un compromis entre la stabilité du modèle et sa performance, et un équilibre doit être trouvé avant le déploiement.

Fondé en 1997, l'hôte de la compétition, Home Credit, est un fournisseur international de services financiers aux consommateurs se concentrant sur le prêt responsable principalement aux personnes ayant peu ou pas d'historique de crédit. Home Credit élargit l'inclusion financière pour la population non bancarisée en créant une expérience de prêt positive et sûre. Nous avons précédemment organisé une compétition avec Kaggle que vous pouvez voir ici.

Votre travail pour évaluer les risques de défaut potentiels des clients aidera les fournisseurs de services financiers aux consommateurs à accepter plus de demandes de prêt. Cela pourrait améliorer la vie des personnes qui ont été historiquement refusées en raison du manque d'historique de crédit.

**Evaluation :**

Les soumissions sont évaluées à l'aide d'une métrique de stabilité Gini. Un score Gini est calculé pour les prédictions correspondant à chaque WEEK\_NUM.

gini=2∗AUC−1

Une régression linéaire, a⋅x+b, est ajustée aux scores Gini hebdomadaires, et un taux de chute (falling\_rate) est calculé comme min(0,a). Cela est utilisé pour pénaliser les modèles dont la capacité prédictive diminue.

Enfin, la variabilité des prédictions est calculée en prenant l'écart type des résidus de la régression linéaire ci-dessus, en appliquant une pénalité à la variabilité du modèle.

La métrique finale est calculée comme suit :

stability metric=*mean*(*gini*)+88.0⋅*min*(0,*a*)−0.5⋅*std*(residuals)

**Submission File:**

Pour chaque case\_id dans l'ensemble de test, vous devez prédire une probabilité pour le score cible. Le fichier doit contenir un en-tête et avoir le format suivant :

case\_id,score

57543,0.1

57544,0.9

57545,0.5

etc.

**Dataset Description:**

Dans cette compétition, vous prévoyez le défaut des clients en fonction des informations internes et externes disponibles pour chaque client. Le score est calculé à l'aide d'une métrique personnalisée qui évalue non seulement l'AUC des prédictions, mais prend également en compte la stabilité du modèle de prédictions sur l'ensemble des données de l'ensemble de test. Pour mieux comprendre cette métrique, veuillez-vous référer à l'onglet Évaluation.

**Table Description :**

Ce jeu de données contient un grand nombre de tables résultant de l'utilisation de sources de données diverses et des différents niveaux d'agrégation des données utilisés lors de la préparation du jeu de données. Remarque : Tous les fichiers énumérés ci-dessous sont disponibles aux formats .csv et .parquet.

**Base tables :**

Les tables de base stockent les informations de base sur l'observation et le case\_id. Il s'agit d'une identification unique de chaque observation et vous devez l'utiliser pour joindre les autres tables aux tables de base.

Fichiers d'entraînement :

* train\_base.csv

Fichiers de test :

* test\_base.csv (Remarque : le fichier test\_base.csv caché contient environ 90 % du nombre de valeurs de case\_id du fichier train\_base.csv)

static\_0 Propriétés : profondeur=0, source de données interne Fichiers d'entraînement :

* train\_static\_0\_0.csv
* train\_static\_0\_1.csv

Fichiers de test :

* test\_static\_0\_0.csv
* test\_static\_0\_1.csv
* test\_static\_0\_2.csv

static\_cb\_0 Propriétés : profondeur=0, source de données externe Fichiers d'entraînement :

* train\_static\_cb\_0.csv

Fichiers de test :

* test\_static\_cb\_0.csv

applprev\_1 Propriétés : profondeur=1, source de données interne Fichiers d'entraînement :

* train\_applprev\_1\_0.csv
* train\_applprev\_1\_1.csv

Fichiers de test :

* test\_applprev\_1\_0.csv
* test\_applprev\_1\_1.csv
* test\_applprev\_1\_2.csv

other\_1 Propriétés : profondeur=1, source de données interne Fichiers d'entraînement :

* train\_other\_1.csv

Fichiers de test :

* test\_other\_1.csv

tax\_registry\_a\_1 Propriétés : profondeur=1, source de données externe, fournisseur de registre fiscal A Fichiers d'entraînement :

* train\_tax\_registry\_a\_1.csv

Fichiers de test :

* test\_tax\_registry\_a\_1.csv

... (et ainsi de suite pour les autres fournisseurs de registres fiscaux et bureaux de crédit)

Veuillez noter que les mêmes conventions de nommage s'appliquent aux fichiers de test. Il convient de noter que certains fournisseurs de données externes peuvent ne pas être disponibles pour les évaluations futures (test), ce qui est anticipé. Chaque groupe de tables peut comprendre une ou plusieurs tables individuelles. Si un groupe contient plus d'une table, elles sont divisées en fonction de SEMAINE\_NUM. Cette division a été mise en place pour limiter la taille maximale des tables.

Valeurs de profondeur :

* profondeur=0 - Il s'agit de caractéristiques statiques directement liées à un case\_id spécifique.
* profondeur=1 - Chaque case\_id a un enregistrement historique associé, indexé par num\_group1.
* profondeur=2 - Chaque case\_id a un enregistrement historique associé, indexé à la fois par num\_group1 et num\_group2.

Vous pouvez en savoir plus sur les bureaux de crédit (CB) ici : <https://fr.wikipedia.org/wiki/Bureau_de_cr%C3%A9dit>

**Columns :**

Colonnes spéciales :

* case\_id : Il s'agit de l'identifiant unique de chaque dossier de crédit. Vous aurez besoin de cet ID pour joindre les tables pertinentes à la table de base.
* date\_decision : Cela fait référence à la date à laquelle une décision a été prise concernant l'approbation du prêt.
* WEEK\_NUM : Il s'agit du numéro de semaine utilisé pour l'agrégation. Dans l'échantillon de test, WEEK\_NUM continue séquentiellement à partir de la dernière valeur d'entraînement de WEEK\_NUM.
* MONTH : Cette colonne représente le mois et est destinée à des fins d'agrégation.
* target : Il s'agit de la valeur cible, déterminée après une certaine période en fonction du défaut ou non du client dans le cas de crédit spécifique (prêt).
* num\_group1 : Il s'agit d'une colonne d'indexation utilisée pour les enregistrements historiques de case\_id dans les tables de profondeur=1 et profondeur=2.
* num\_group2 : Il s'agit de la deuxième colonne d'indexation pour les enregistrements historiques de case\_id des tables de profondeur=2. L'ordre de num\_group1 et num\_group2 est important et sera clarifié dans les définitions des caractéristiques.

Toutes les autres colonnes brutes dans les tables servent de prédicteurs. Leurs définitions peuvent être trouvées dans le fichier feature\_definitions.csv. Pour les tables avec profondeur=0, les prédicteurs peuvent être directement utilisés comme caractéristiques. Cependant, pour les tables avec profondeur>0, vous devrez peut-être utiliser des fonctions d'agrégation qui condenseront les enregistrements historiques associés à chaque case\_id en une seule caractéristique. Au cas où num\_group1 ou num\_group2 représenterait l'index de personne (ce qui est clair avec les définitions des prédicteurs), le zéro a une signification spéciale. Lorsque num\_groupN=0, il s'agit du demandeur (la personne qui a demandé un prêt).

Divers prédicteurs ont été transformés, donc nous avons la notation suivante pour des groupes similaires de transformations :

* P : Transformer DPD (jours de retard)
* M : Masquer les catégories
* A : Transformer le montant
* D : Transformer la date
* T : Transformation non spécifiée
* L : Transformation non spécifiée

Veuillez noter que les transformations au sein d'un groupe sont indiquées par une lettre majuscule à la fin du nom du prédicteur. Nous espérons que cela simplifiera la manipulation des prédicteurs.

Éditions :

* pmts\_month\_158T est pour un contrat actif
* pmts\_month\_706T est pour un contrat fermé
* dateofcredstart\_181D : Date de début d'un contrat de crédit.