PROJET 8

PARTICIPEZ A UNE COMPETITION KAGGLE !

Rapport

AMERICAN EXPRESS - DEFAULT PREDICTION

Predict if a customer will default in the future

Étudiant : Zeruk Viktoriya

Mentor : Louis Willems

SOMMAIRE

CONTEXTE

Kaggle est une plateforme qui organise des compétitions en Data Science et qui récompense les meilleurs analystes internationaux.

La mission est donc de rechercher et participer à une compétition réelle et en cours sur la plateforme et de partager les résultats obtenus avec la communauté.

Plan du travail :

Première étape (Notebook 1) : choix du modèle, 10 % des données ont été analysées, y compris toutes les variables avec des valeurs manquantes relatives.

Dans la deuxième étape (Notebook 2), le modèle LGBM a été entraîné sur 100 % des données.

La troisième étape était prévue pour entraîner le modèle LGBM sur des données contenant uniquement des variables sélectionnées avec un poids d'information plus important.

Il serait vraiment intéressant de comparer les résultats des différentes phases. Cette partie du travail n'a pas été réalisée, par manque de temps.

1. PRESENTATION DE LA COMPETITION

1.1 Vue d'ensemble de la compétition

Que ce soit au restaurant ou pour acheter des billets de concert, la vie moderne compte sur la commodité d'une carte de crédit pour effectuer les achats quotidiens. Elle nous évite de transporter de grandes quantités d'argent liquide et peut également avancer un achat complet qui peut être payé dans le temps. Comment les émetteurs de cartes savent-ils que nous rembourserons ce que nous facturons ? Il s'agit d'un problème complexe pour lequel il existe de nombreuses solutions, et encore plus d'améliorations potentielles.

La prévision de défaut de crédit est essentielle à la gestion du risque dans une entreprise de prêt à la consommation. Elle permet aux prêteurs d'optimiser leurs décisions de prêt, ce qui se traduit par une meilleure expérience client et une économie d'entreprise saine. Il existe des modèles pour aider à gérer le risque, mais il est possible d'en créer de meilleurs qui peuvent surpasser ceux qui sont actuellement utilisés.

American Express est une société de payement qui travaille à l'échelle mondiale. Premier émetteur de cartes de payement au monde, elle offre à ses clients l'accès à des produits, des connaissances et des expériences qui enrichissent leur vie et contribuent à leur succès commercial.

Dans cette compétition, il nous faut appliquer nos compétences de Machine Learning pour prédire les défauts de payement. Plus précisément, il faut exploiter un ensemble de données à l'échelle industrielle pour construire un modèle d'apprentissage automatique qui défie le modèle actuel en production. Les ensembles de données de formation, de validation et de test comprennent des données comportementales chronologiques et des informations anonymes sur le profil des clients.

1.2 Aperçue des données

L'ensemble de données contient des caractéristiques de profil agrégées pour chaque client à chaque date de relevé. Les caractéristiques sont rendues anonymes et normalisées, et sont classées dans les catégories générales suivantes :

- **D_* = variables de délinquance**
- **S_* = variables de dépenses**
- **P_* = variables de payement**
- **B_* = variables de solde**
- **R_* = variables de risque**

Les caractéristiques suivantes étant catégorielles :

['B_30', 'B_38', 'D_114', 'D_116', 'D_117', 'D_120', 'D_126', 'D_63', 'D_64', 'D_66', 'D_68', 'S_2'].

1.3 Objectif

L'objectif est de prédire la probabilité qu'un client ne rembourse pas le montant du solde de sa carte de crédit à l'avenir, sur la base de son profil client mensuel. La variable binaire target est calculée en observant la fenêtre de performance des 18 mois avant le dernier relevé de carte de crédit, et si le client ne paye pas le montant dû dans les 120 jours suivant la date de son dernier relevé, il est considéré comme étant en défaut de payement.

1.4 Métrique d'évaluation

La métrique d'évaluation pour cette compétition, notée M, est la moyenne de deux mesures de classement : le coefficient de Gini normalisé, G, et le taux de défaut capturé à 4%, D.

$$M = 0,5(G + D)$$

Le taux de défaut capturé à 4% est le pourcentage d'étiquettes positives (défauts) capturées

dans les 4% des prédictions les mieux classées, et représente une statistique de sensibilité/rappel. Pour les deux sous-métriques G et D , un poids de 20 est attribué aux étiquettes négatives pour tenir compte du sous-échantillonnage. Cette métrique a une valeur maximale de 1.

2. EXPLORATION DES DONNÉES

Les DataFrames mis à disposition par American Express étant lourds, des utilisateurs Kaggle en ont proposé des versions compressées plus légères. Nous avons ainsi utilisé une de ces versions afin de ne pas saturer la mémoire RAM de notre machine.

2.1 Valeurs manquantes

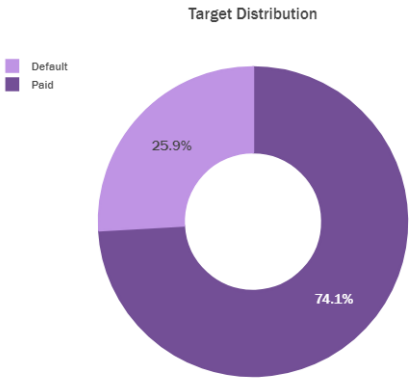
Une fois les données chargées et utilisables, une première étape a été de rechercher les éventuelles valeurs manquantes.

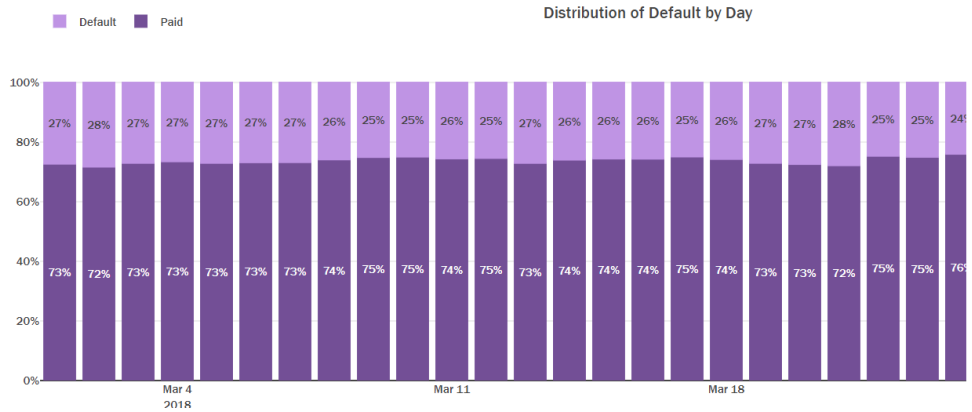
Il s'est avéré que le jeu de données contenait un très grand nombre de valeurs manquantes. Toutefois, en gardant en tête que certaines variables sont liées à des évènements de non payement ou de fraude, évènements relativement rares à l'échelle d'une banque, on peut considérer certaines valeurs manquantes comme des indicateurs d'évaluation des clients. De plus, les données étant agrégées, en regroupant les données après chaque période de relevés bancaires, certaines informations peuvent évoluer d'une période à l'autre, créant ainsi de nouvelles valeurs manquantes.

	Number of missing values	% of missing values
Delinquency 87	458268	99.860000
Delinquency 88	458086	99.820000
Delinquency 108	456286	99.430000
Delinquency 110	455235	99.200000
Delinquency 111	455235	99.200000
Balance 39	454808	99.110000
Delinquency 73	454674	99.080000
Balance 42	452771	98.660000
Delinquency 137	442518	96.430000
Delinquency 134	442518	96.430000
Delinquency 138	442518	96.430000
Delinquency 135	442518	96.430000
Delinquency 136	442518	96.430000
Risk 9	431960	94.130000
Balance 29	431589	94.050000
Delinquency 76	409597	89.250000
Risk 26	407770	88.860000
Delinquency 106	407265	88.750000
Delinquency 132	407153	88.720000
Delinquency 49	407150	88.720000

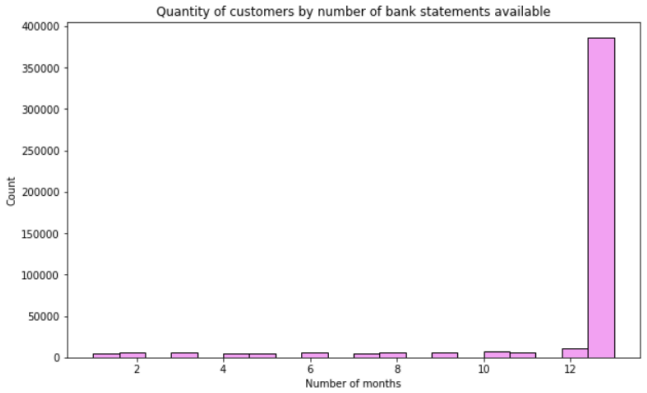
Il semble que pour certaines variables, beaucoup de données manquent. Mais il faut garder à l'esprit que les cas de fraude ou de non-paiement sont assez rares, d'où un nombre élevé de valeurs manquantes. Nous considérerons que le jeu de données ne contient pas d'erreurs et conserverons pour le moment les valeurs manquantes.

2.2 Target Distribution | Clients





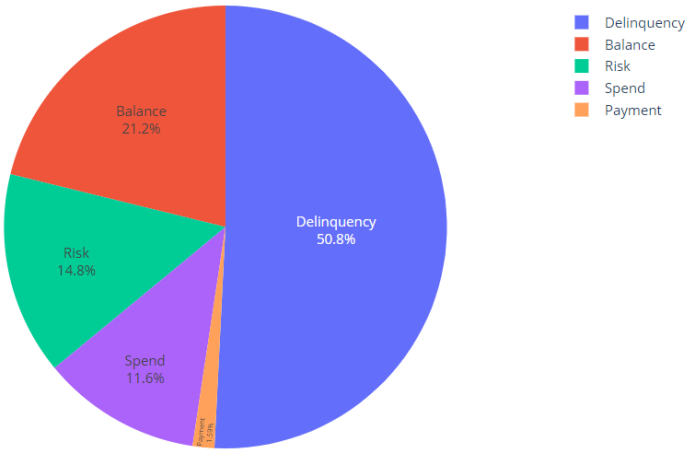
Les données concernent 458913 clients distincts, pour la quasi-totalité desquels on dispose des données bancaires des 13 derniers mois. En liant la valeur de la variable target, indiquant la présence ou non de défaut de payement, on se rend compte que plus de 25% des clients ont eu des défaut de payement, bien plus que ce que l'on supposait plus haut.



Cette proportion est constante pour chaque jour de l'ensemble de formation, avec une tendance saisonnière hebdomadaire le jour du mois où les clients reçoivent leurs relevés.

Cependant, cette proportion n'est peut-être pas révélatrice de la situation générale. On peut imaginer qu'une proportion plus importante de clients en défaut de payement a été incorporée dans le jeu de données dans le but d'obtenir de meilleurs résultats lors de l'entraînement des modèles de classification.

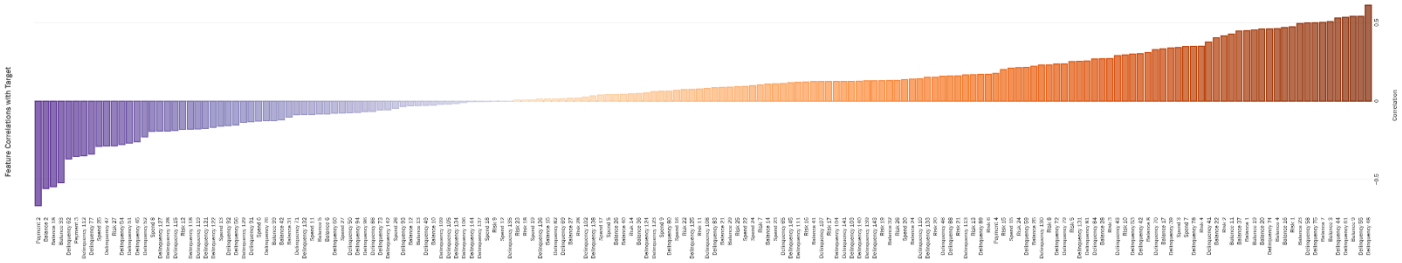
2.3 Les variables



Les variables présentes dans le jeu de données étant déjà normalisées, il était difficile de tirer de l'information à partir de leurs distributions. En revanche, il était toujours possible de rechercher d'éventuelles corrélations, entre les variables d'une même catégorie d'une part, et entre les variables et la target d'autre part. La conclusion générale de cette étude est qu'il ne semble pas exister de tendance particulière au niveau des corrélations, et qu'il serait donc préférable de garder un maximum de variables.

2.4 Analyse de la valeur de l'information

2.4.1 Corrélations



Il existe plusieurs fortes corrélations avec la variable cible. Le paiement 2 est le plus corrélé négativement avec la probabilité de défaut avec une corrélation de -0,67, tandis que la délinquance 48 est globalement la plus corrélée positivement à 0,61. La délinquance 87 est également absente des corrélations ci-dessus en raison de la proportion de valeurs nulles. En fait, 24 des 30 principales caractéristiques avec des valeurs manquantes se trouvent dans les variables de délinquance.

2.4.2 Information Value (IV)

La valeur de l'information est l'une des techniques les plus utiles pour sélectionner des variables importantes dans un modèle prédictif. Il permet de classer les variables en fonction de leur importance.

$$IV = \sum (\% \text{ of non-events} - \% \text{ of events}) * WOE$$

Si la statistique IV est :

- Moins de 0,02, alors le prédicteur n'est pas utile pour la modélisation (séparant les Biens des Mals)
- 0,02 à 0,1, alors le prédicteur n'a qu'une faible relation avec le rapport de cotes Biens/Mauvais
- 0,1 à 0,3, alors le prédicteur a une relation de force moyenne avec le rapport de cotes Biens/Mauvais
- 0,3 à 0,5, alors le prédicteur a une forte relation avec le rapport de cotes Biens/Mauvais.
- 0,5, relation suspecte (vérifier une fois)

Référence : https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

Des valeurs IV pour chaque *features*

```
P_2     3.732496
D_48    2.406650
B_18    2.212995
B_7     2.106675
D_75    2.070369
D_61    2.070309
B_23    2.059052
B_10    1.998735
B_9     1.994288
D_44    1.980849
B_2     1.905190
B_6     1.900564
B_3     1.858991
B_1     1.829305
B_37    1.801300
dtype: float64
```

2.4.3 Poids de la prevue | Weight of Evidence (WOE)

Le poids de la preuve indique le pouvoir prédictif d'une variable indépendante par rapport à la variable dépendante. Depuis qu'il a évolué à partir du monde de la notation de crédit, il est généralement décrit comme une mesure de la séparation des bons et des mauvais clients. Les « mauvais clients » désignent les clients qui ont fait défaut sur un prêt. et "Bons clients" fait référence aux clients qui ont remboursé le prêt.

$$WOE = \ln(\%\ \text{of non-events} \div \%\ \text{of events})$$

- Distribution des bons clients (%)dans un groupe particulier
- Distribution des mauvais clients (%) dans un groupe particulier
- ln - Log naturel

Étapes du calcul du WOE

- Pour une variable continue, diviser les données en 10 parties (ou moins selon la distribution).
- Calculer le nombre d'événements et de non-événements dans chaque groupe (bin)
- Calculer le % d'événements et le % de non-événements dans chaque groupe.
- Calculer WOE en prenant le logarithme naturel de la division du % de non-événements et du % d'événements

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 0.18)	29733	0.064790	2631	27102	0.911512	-3.383762	0.745582	0.064705
1	[0.18, 0.32)	34240	0.074611	8586	25654	0.749241	-2.146085	0.409141	0.043147
2	[0.32, 0.43)	40102	0.087385	15556	24546	0.612089	-1.507621	0.242465	0.027729
3	[0.43, 0.52)	41474	0.090374	22460	19014	0.458456	-0.884958	0.083160	0.010069
4	[0.52, 0.58)	28361	0.061800	19500	8861	0.312436	-0.262764	0.004528	0.000564
5	[0.58, 0.64)	32780	0.071430	26363	6417	0.195760	0.361492	0.008501	0.001057
6	[0.64, 0.69)	27969	0.060951	20396	7573	0.270771	0.018748	0.000039	0.000005
7	[0.69, 0.73)	23797	0.051855	22171	1626	0.068328	1.561144	0.080413	0.009141
8	[0.73, 0.77)	24881	0.054222	23930	951	0.038219	2.117939	0.118583	0.014247
9	[0.77, 0.82)	33368	0.072711	32634	734	0.021997	2.743082	0.246278	0.023740
10	[0.82, 0.85)	21436	0.046709	21187	249	0.011620	3.487376	0.230131	0.019643
11	[0.85, 0.88)	26085	0.056841	25872	213	0.008166	3.748106	0.278419	0.022721
12	[0.88, 0.93)	40208	0.087616	39988	220	0.005472	4.151189	0.480421	0.036617
13	[0.93, 0.97)	29920	0.065198	29808	112	0.003743	4.532515	0.392997	0.028094
14	[0.97, inf)	24588	0.053579	24527	61	0.002481	4.945137	0.354106	0.023646
15	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
16	Missing	2969	0.006470	2474	495	0.166723	0.557515	0.001733	0.000214
Totals		458913	1.000000	340085	118828	0.258934		3.732496	0.330007

P_2 est une fonctionnalité continue, car nous l'avons divisé en 15 groupes

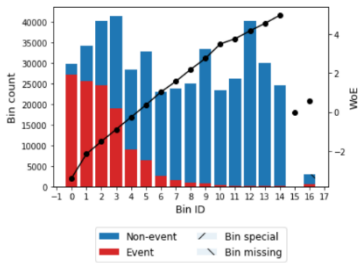
Chaque casier a des nombres et des taux de non-événements et d'événements

Chaque bin a des valeurs WOE et IV

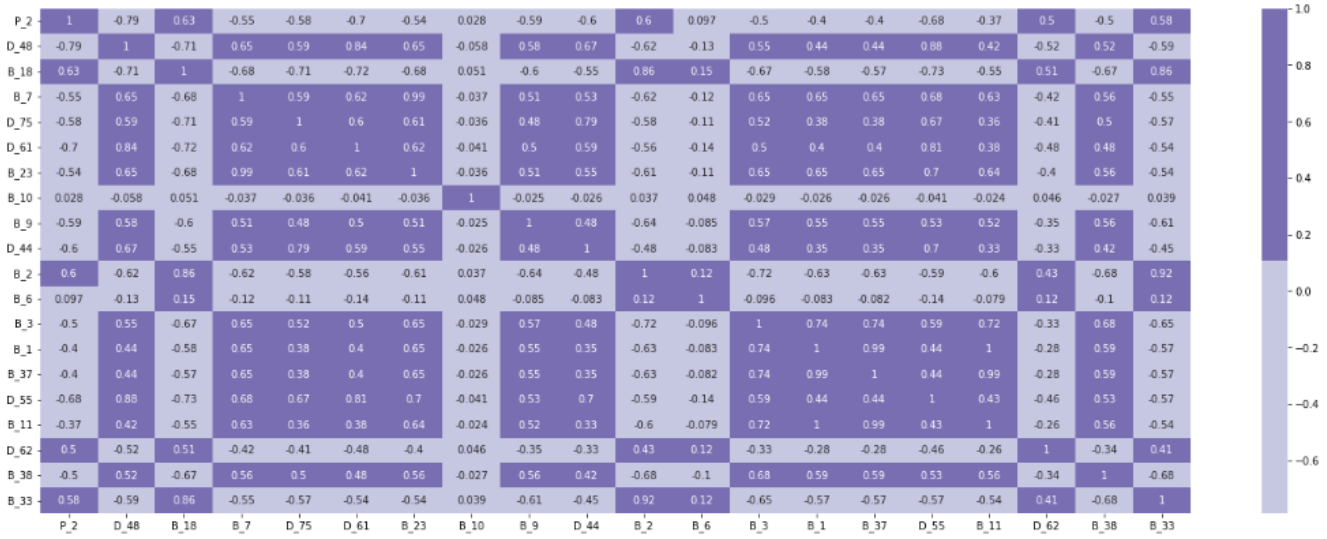
Pour les valeurs manquantes, il est créé 16e bin

Nous pouvons observer que tout en augmentant les bins, le taux d'événements diminue.

La ligne pointillée noire qui est positivement corrélée avec la cible.



Heatmap des features sélectionnées



Remarque :

L'analyse effectuée est importante.

En effet, dans la **première étape** du travail (Notebook 1) sur la choix du modèle, 10 % des données ont été analysées, y compris toutes les variables avec des valeurs manquantes relatives.

Dans la **deuxième étape** (Notebook 2), le modèle LGBM a été entraîné sur 100 % des données.

La troisième étape était prévue pour entraîner le modèle LGBM sur des données contenant uniquement des variables sélectionnées avec un poids d'information plus important.

Il serait vraiment intéressant de comparer les résultats des différentes phases. Cette partie du travail n'a pas été réalisée, par manque de temps.

Cette l'analyse a été utile pour comparer l'importance des variables identifiées par le modèle LGBM.

3. CHOIX DE MODEL DE PREDICTION

Les données mises à disposition sont déjà séparées en jeu d'entraînement et jeu de test. Cependant, les données étant conséquentes, nous nous sommes limités à 10% des données pour tester et comparer les différents modèles dans le but d'accélérer les calculs et ne pas saturer la mémoire. En plus du temps de calcul et du score Accuracy, nous avons utilisé la métrique fournie par la compétition pour comparer les différents modèles.

Sélection des hyperparamètres éte effectué avec GridSearchCV.

3.1 Préparation des données

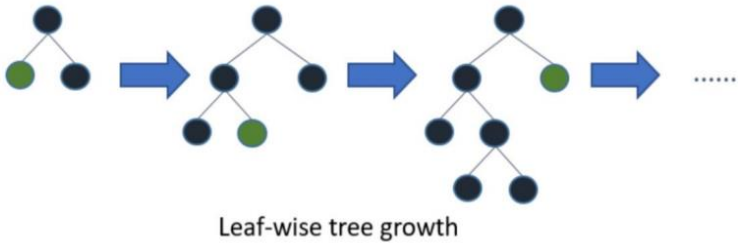
Après encoder les variables catégorielles, nous avons procédé à traitement des valeurs manquantes : conserver toutes les valeurs manquantes et laisser les algorithmes effectuer les traitements. Nous restons, que les variables de délinquance, qui contiennent de nombreuses valeurs manquantes, peuvent révéler de petits détails qui sont très importants pour indiquer un défaut de payement. Même si cela n'améliore pas la qualité générale du modèle, la prise en compte de toutes les variables permettrait peut-être d'éviter certains non-remboursements ou fraudes et donc d'éviter des conséquences dramatiques compte tenu du rôle central des banques dans nos sociétés.

3.2 Modèles

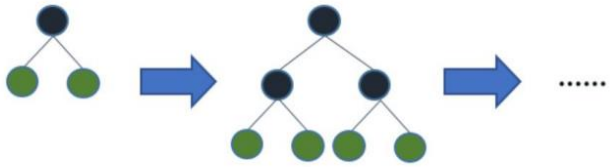
Nous avons cherché des algorithmes de classification binaire capables de traiter automatiquement les valeurs manquantes. Nous avons retenu les deux modèles suivants :

1. **LightGBM (LGBM)** est un framework de gradient boosting basé sur des arbres de décision. Il est conçu pour être distribué et efficace avec les avantages suivants :
    - Vitesse d'entraînement plus rapide et efficacité accrue
    - Utilisation réduite de la mémoire
    - Meilleure précision
    - Prise en charge de l'apprentissage parallèle et par le GPU
    - Capacité à traiter des données à grande échelle

Le LightGBM fait croître l'arbre de décision verticalement tandis que les autres algorithmes d'apprentissage basés sur les arbres le font généralement horizontalement. Cela signifie que le LGBM fait croître l'arbre en fonction des feuilles (Leaf-wise tree growth), alors que les autres algorithmes le font croître en fonction des niveaux (Level-wise tree growth).



Leaf-wise tree growth



Level-wise tree growth

2. CatBoost a été développé par l'entreprise russe Yandex. CatBoost signifie Categorical Boosting parce qu'il est conçu pour fonctionner parfaitement sur des données catégorielles. 7 Voici quelques caractéristiques de CatBoost, qui le distinguent de tous les autres algorithmes de boosting :
    - Haute qualité sans réglage des paramètres
    - Prise en charge des caractéristiques catégorielles
    - Version GPU rapide et évolutive
    - Amélioration de la précision en réduisant l'overfitting
    - Prédictions rapides
    - Fonctionne bien avec moins de données

    L'algorithme CatBoost est intéressant à utiliser lorsque les données sont très hétérogènes, c'est-à-dire avec beaucoup de variabilité, des types différents ou encore une quantité importante de valeurs manquantes.

3.3 Résultats

Les deux modèles donnent des résultats similaires :

	LGBM	CatBoost
computation time	201.125230	265.051841
metric score	0.510733	0.511213
accuracy score	0.880854	0.880547

Le modèle final retenu pour la mise en production est le LGBM entraîné sur le jeu de données original dans lequel on a conservé toutes les variables et toutes les valeurs manquantes.

4. DEFAULT PREDICTION

Certains paramètres ont été identifiés précédemment.

Deux configurations ont été testées.

Configuration 1

```
params = {'boosting_type': 'gbdt',
          'n_estimators': 1000,
          'num_leaves': 50,
          'learning_rate': 0.05,
          'colsample_bytree': 0.9,
          'min_child_samples': 2000,
          'max_bins': 500,
          'reg_alpha': 2,
          'objective': 'binary',
          'random_state': 21}
```

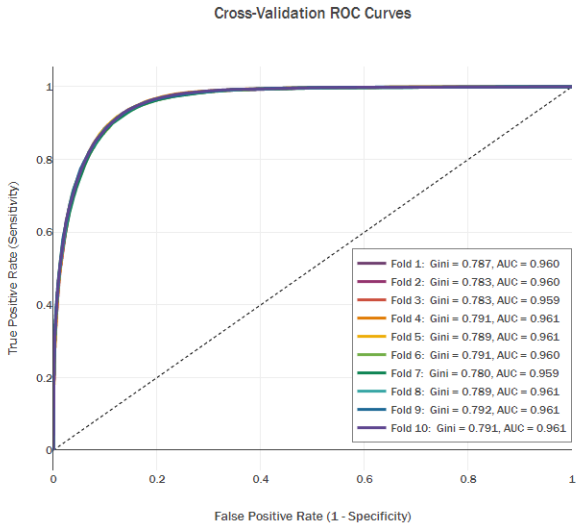
Configuration 2

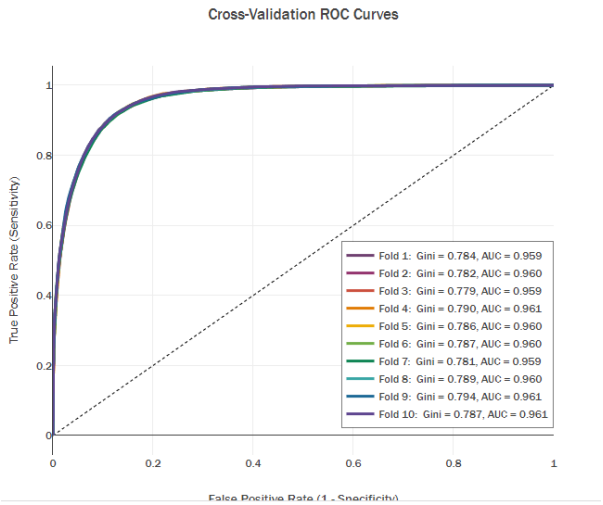
```
params = {'boosting_type': 'dart',
      'n_estimators': 1000,
      'num_leaves': 100,
      'learning_rate': 0.1,
      'colsample_bytree': 0.9,
      'min_child_samples': 2000,
      'max_bins': 500,
      'reg_alpha': 2,
      'objective': 'binary',
      'random_state': 21,
      'bagging_freq': 10,
      'bagging_fraction': 0.50,
      'n_jobs': -1,
      'lambda_l2': 2}
```

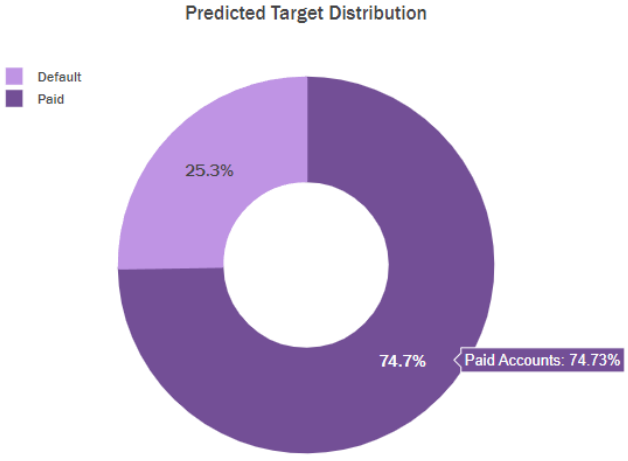
Résultats Configuration 1

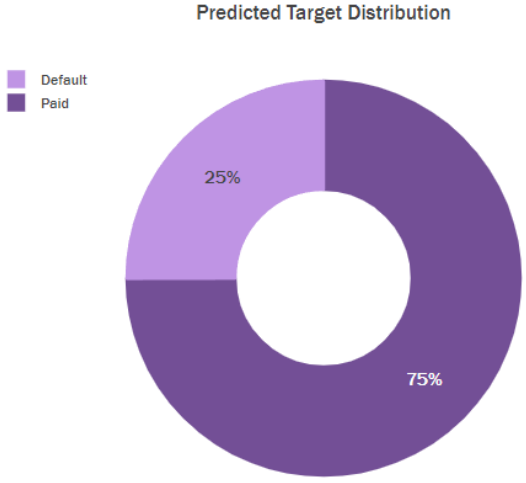
Résultats Configuration 2

https://www.kaggle.com/code/victoriazeruk/amex-default-prediction-test

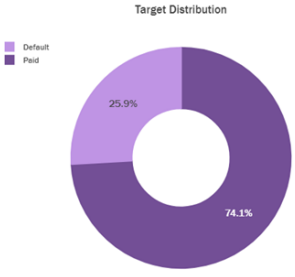








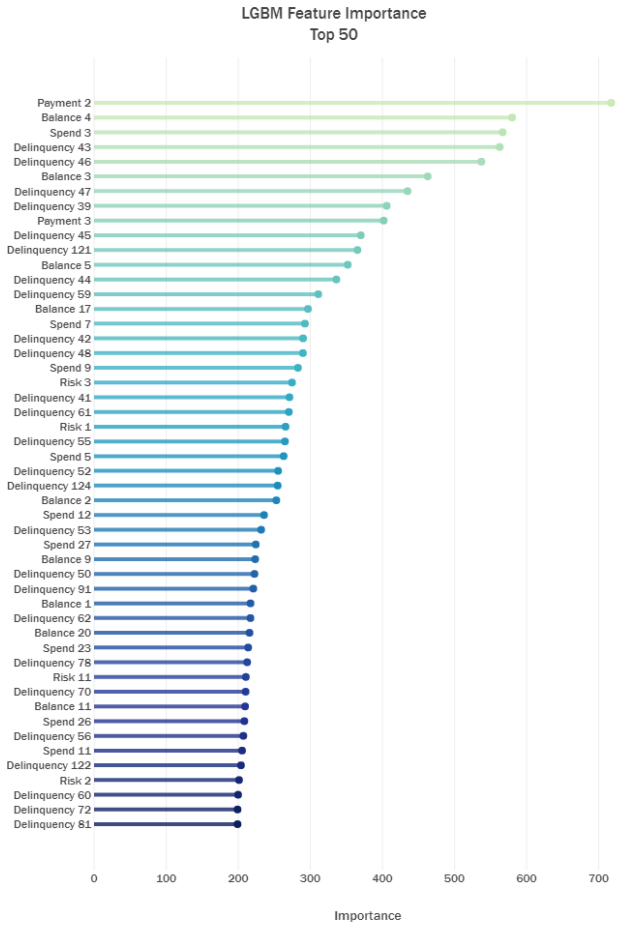
Distribution initiale :



Configuration choisie : 1

5. IMPORTANCE DES *FEATURE*

Le modèle final a identifié les *feature* importantes suivantes :



Des valeurs IV pour chaque *features*

```
P_2     3.732496
D_48    2.406650
B_18    2.212995
B_7     2.106675
D_75    2.070369
D_61    2.070309
B_23    2.059052
B_10    1.998735
B_9     1.994288
D_44    1.980849
B_2     1.905190
B_6     1.900564
B_3     1.858991
B_1     1.829305
B_37    1.801300
dtype: float64
```

Les correspondances importantes : Payment 2 (P_2), Delinquency 48 (D_48), Balance 3 (B_3), Balance 3 (B_2), Balance 3 (B_3).

Il serait intéressant d'approfondir le sujet de l'importance des variables.

6. CONCLUSION

Le modèle final retenu pour la mise en production est le LGBM entraîné sur le jeu de données original dans lequel on a conservé toutes les variables et toutes les valeurs manquantes.

Les caractéristiques identifiées par le modèle LGBM correspondent en partie à l'analyse de la valeur de l'information IV.

Ce concours est mon premier concours Kaggle et personnellement je le trouve très utile pour échanger des idées et étudier de nouvelles techniques d'apprentissage automatique.

Par exemple, il était intéressant de réaliser le test du modèle GRU basé sur les travaux de

CHRIS DEOTTE

(TensorFlow GRU Starter - [0.790] : https://www.kaggle.com/code/cdeotte/tensorflow-gru-starter-0-790/) - Notebook 3.

L'approche proposée est très prometteuse.

Grâce à Openclassrooms j'ai découvert beaucoup de choses sur l'analyse de données et machine learning, j'ai acquis de nouvelles compétences et j'ai amélioré mes compétences en programmation python.