le cnam
Nouvelle-Aquitaine

USSI4X

Entreposage et Fouilles de Données

Adrien Moreau

Raphaël Simon

Auriane Sagard

Jordan Brassan

Mathéo Rouché—Chevaillier

I. Présentation du sujet

Le jeu de données "financement.csv" offre l'opportunité d'explorer les dynamiques du financement participatif à travers une multitude de projets. Ces données riches et variées nous permettent d'entrer dans le monde des campagnes de financement en ligne, en examinant les caractéristiques des projets et les tendances de financement.

Dans le paysage actuel du financement participatif, de nombreux projets cherchent à trouver le soutien nécessaire pour atteindre leurs objectifs. Cependant, prédire le succès d'un projet et estimer le montant potentiel du financement reste un défi complexe. Notre objectif est de comprendre si un nouveau projet, une fois inscrit sur la plateforme, sera financé, et le cas échéant, à quel niveau. Cette problématique essentielle nécessite une analyse approfondie des données disponibles pour identifier les facteurs qui influencent la réussite d'une campagne de financement participatif.

A travers cette étude, nous explorons les différentes variables présentes dans le jeu de données, telles que la catégorie du projet, la devise utilisée, les dates, l'objectif de financement, le nombre de supporters potentiels, et bien d'autres. Nous cherchons à déterminer comment ces variables interagissent et contribuent à la décision de financement d'un projet. En outre, nous nous intéressons également à l'impact potentiel de l'analyse de texte dans le processus de création de nouvelles fonctionnalités pour améliorer la prédiction du succès d'un projet.

En résumé, notre objectif est d'apporter des insights précieux sur les tendances et les dynamiques des campagnes de financement participatif, tout en développant des modèles prédictifs aussi robustes que possible pour évaluer le potentiel de financement des nouveaux projets.

Le dataset "financement.csv" est composé de 15 variables. Ce jeu de données offre un aperçu plutôt complet des projets inscrits sur la plateforme, de leurs objectifs de financement aux montants promis, en passant par l'état de chaque projet.

Voici un aperçu des principales variables présentes dans le dataset :

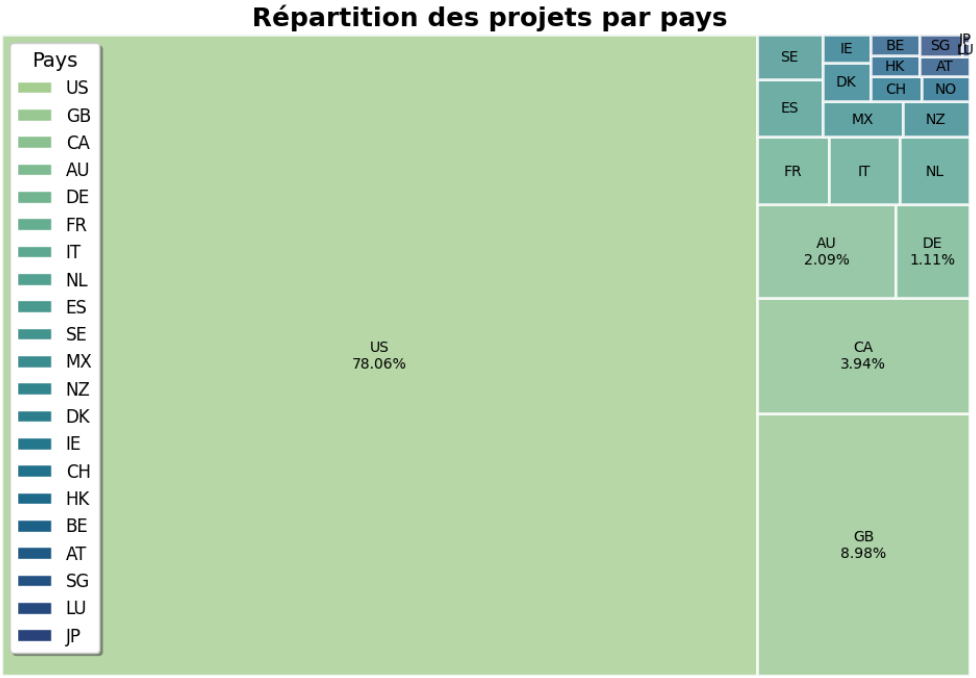
ID	Identifiant unique attribué à chaque projet, permettant de le différencier de manière unique dans le dataset.
Nom	Nom du projet, fournissant une indication sur la nature ou le sujet de la campagne de financement.
Catégorie	Catégorie à laquelle le projet est associé, offrant une classification thématique des différents types de projets.
Catégorie Principale	Catégorie principale de la campagne, permettant une classification plus large des projets.
Monnaie	Devise utilisée pour définir les objectifs de financement et les montants promis pour chaque projet.
Date Butoire	Date limite fixée pour la réalisation du financement du projet. ***Format**: 'dd-mm-aaaa'*
Objectif	Montant financier visé par la campagne de financement participatif pour atteindre ses objectifs.
Lancement	Date de lancement de la campagne de financement, marquant le début de la collecte de fonds. ***Format**: 'dd-mm-aaaa hh:mm:ss'*
Promesse	Montant total promis par les contributeurs potentiels pour soutenir le projet.
État	État actuel du projet, indiquant s'il a été financé avec succès ou s'il est toujours en cours de financement.
Supporters	Nombre de personnes ou d'entités ayant exprimé leur intention de soutenir financièrement le projet.
Pays	Pays d'origine des promesses de financement associées à chaque projet.
Promesse USD1	Montant promis converti en dollars américains par la plateforme de financement.
Promesse USD2	Montant promis converti en dollars américains par l'outil de conversion fixer.io.
Objectif USD	Objectif de financement converti en dollars américains par l'outil de conversion fixer.io.

II. Préparation du dataset

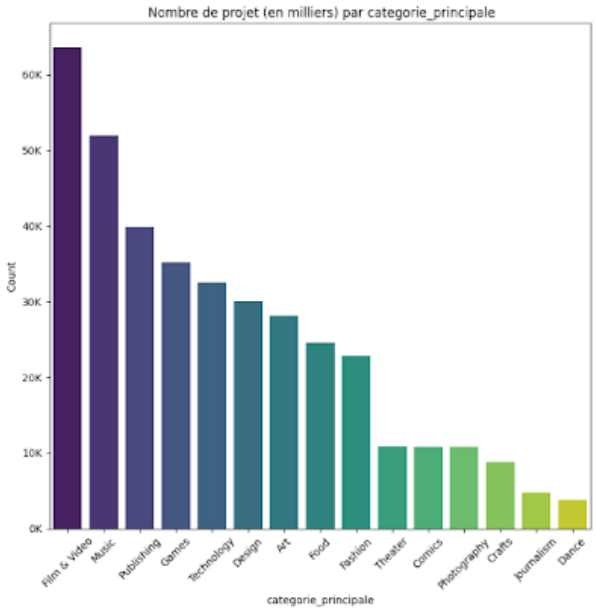
- **Remplacement des valeurs NaN par 0 dans la colonne "promesse_usd1" :** Afin d'assurer la cohérence des données et éviter les problèmes de calcul, les valeurs manquantes dans la colonne "promesse_usd1" sont remplacées par 0. Cette variable est cependant bien moins fiable que promesse_usd2, elle sera ignorée par la suite.
- **Traitement des dates :**
  Les dates dans le dataset sont extraites et décomposées en leurs composants (jour, mois, année, heure, minute, seconde). Ceci est réalisé en définissant des fonctions qui séparent chaque composant de la date ou de l'heure selon le format de la date (lancement ou butoire).
- **Ajout de l'écart entre la date de début de lancement de recherche de financement et la date butoire de financement du projet :**
  En plus de la séparation des différentes parties de la date, nous avons ajouté une nouvelle variable correspondant à l'écart entre ces deux dates.
- **Ajout de variables sur le nom (taille + nombre de mots) :**
  Cette partie consiste à enrichir les données en ajoutant des informations sur les noms des projets. Deux nouvelles variables sont créées : le nombre de mots dans chaque nom et le nombre total de caractères.
- **Feature engineering sur le nom de projet :**
  Dans le but d'ajouter des variables explicatives, nous avons décidé d'incorporer le traitement de texte. Grâce à cette approche, nous pourrons détecter si un mot ou groupe de mots influence l'acceptation d'un projet ou non. Pour ce faire, nous avons nettoyé les noms en retirant toute ponctuation et les mots de liaison. Ensuite, nous avons tokenisé les mots pour les convertir en format numérique avant d'encoder le tout pour les modèles.
- **Ajout de la variable 'continent' et dummification des variables catégorielles :** L'ajout de la variable 'continent' permet de regrouper les pays par continent, ce qui peut faciliter l'analyse géographique des données. Ensuite, les variables catégorielles telles que la catégorie du projet, la catégorie principale, la monnaie et le continent sont transformées en variables binaires (dummy variables) pour être utilisées dans des modèles d'apprentissage automatique.

III. Analyse du jeu de données

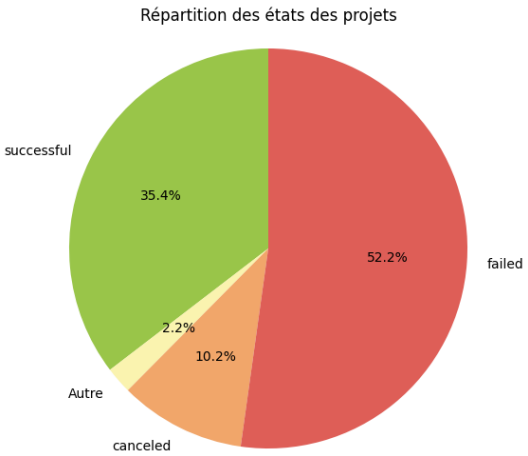
Nous avons commencé par analyser chaque variable du jeu de données afin de bien comprendre les caractéristiques des données qu'il contient.



Tout d'abord, on constate que 78% des projets proviennent des Etats-Unis, et par écho, la monnaie la plus présente est le dollar américain.

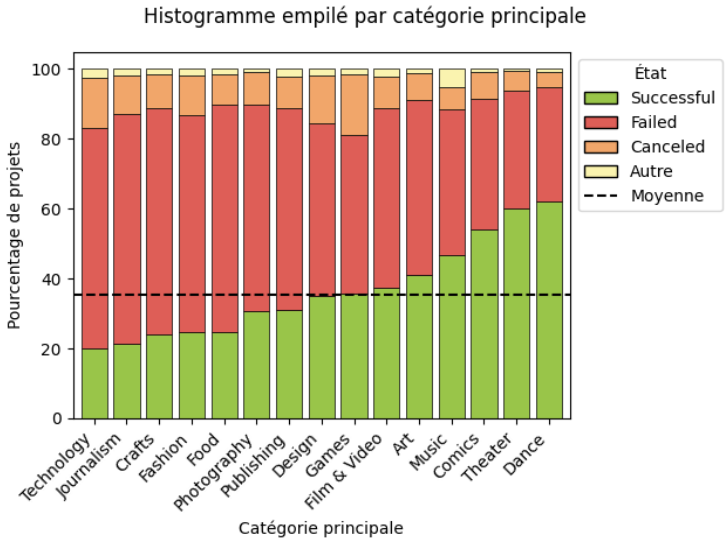


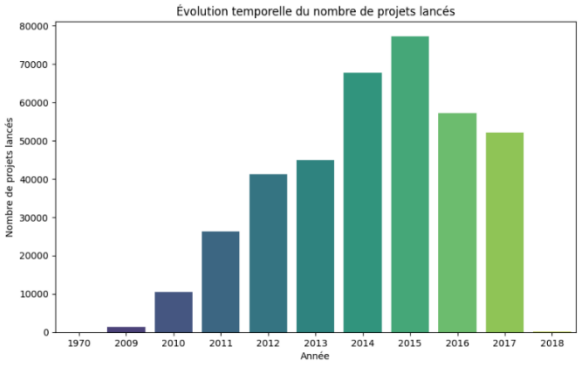
Parmi les catégories principales, les projets de Film & Vidéo sont les plus représentés, tandis que la Danse est en 15ème et dernière place.

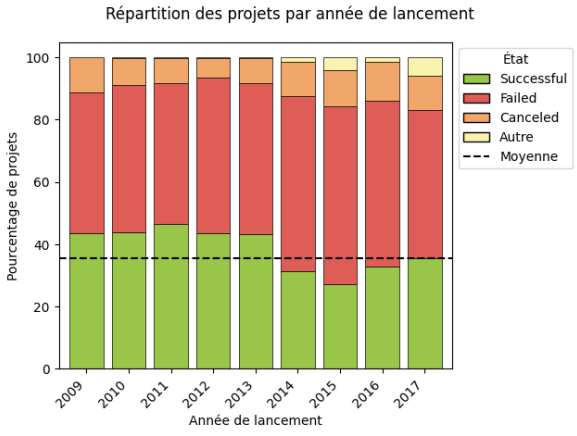


Les échecs représentent plus de la moitié des projets (52,2%), tandis que les succès n'en représentent qu'un tier (35,4%). Cette répartition souligne l'importance pour les porteurs de projet de connaître si leur projet sera financé avec succès, ces derniers étant moins fréquents.

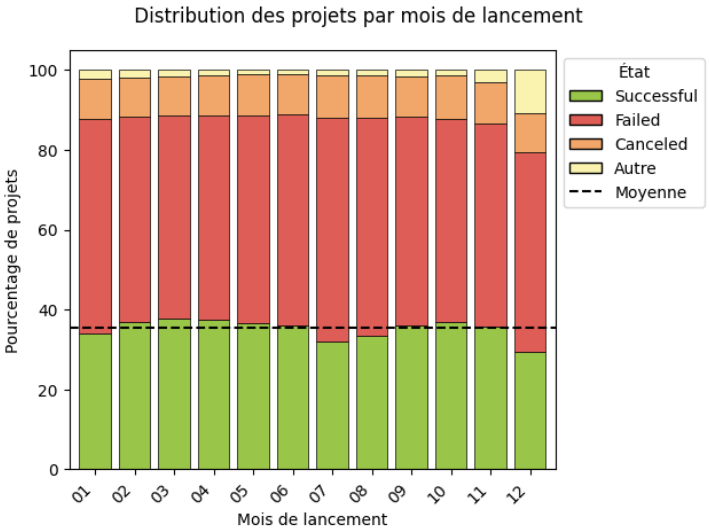
Les principaux projets ayant plus d'échec que la moyenne sont dans le domaine des technologies, du journalisme, de l'artisanat, la mode, l'alimentation, la photographie et l'édition. La danse a environ 60% de succès, c'est bien supérieur à la moyenne de 35,4%.





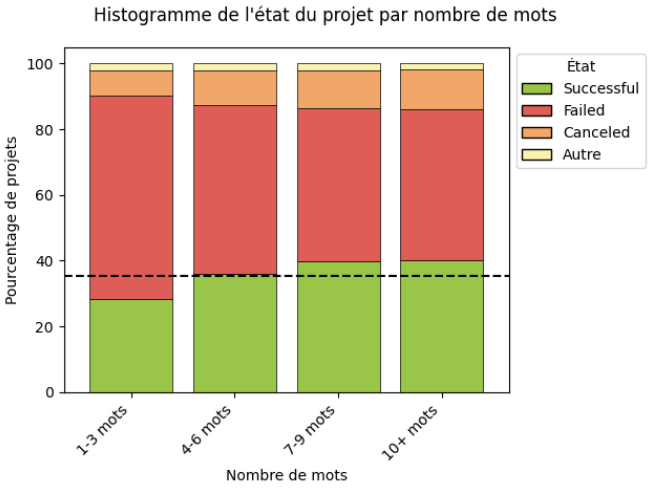


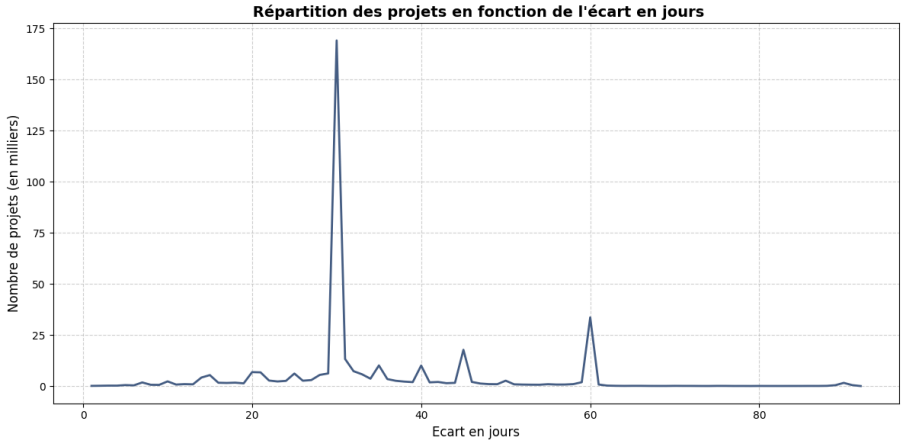
Le nombre de projets a considérablement augmenté en 2014 et 2015. On peut noter que les succès de ces projets sont inferieur à la moyenne, cela suppose que plus le nombre de projet est important, plus il y a d'échecs.



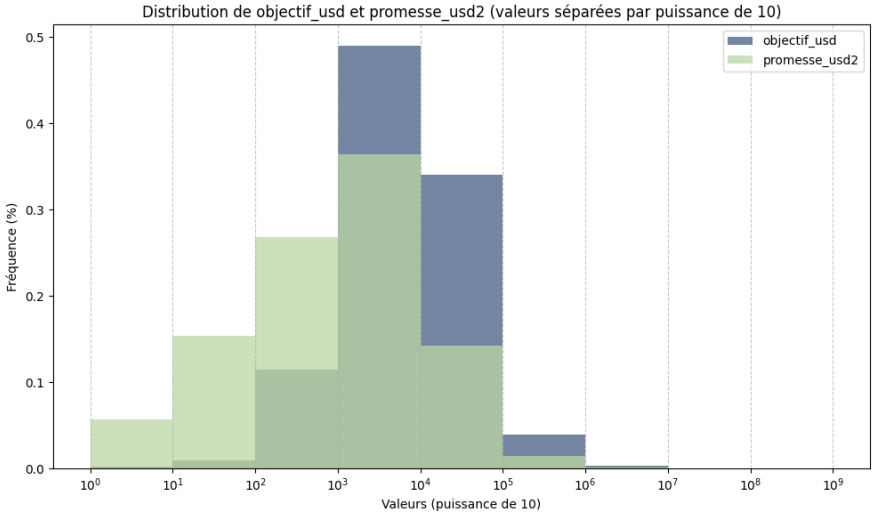
On ne constate pas de différence significative sur le nombre de projets en succès par mois, on remarque une légère baisse pendant juillet, août et décembre, ce qui peut probablement correspondre des périodes de vacances.

D'après les données, plus le nom d'un projet est long, plus le projet a de succès. Cela peut en effet aider les potentiels mécènes à mieux visualiser les enjeux du projet, ses tenants et ses aboutissants.

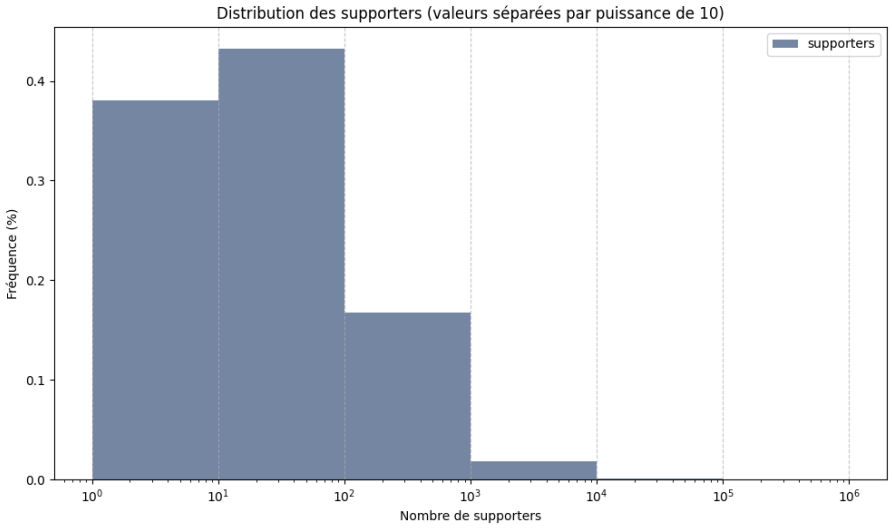




En moyenne, l'écart en nombre de jours d'un projet est de 35 jours. Cela est parfaitement visible et compréhensible étant donné la répartition de ces écarts sur le graphique ci-dessus. En effet, une énorme majorité des projets dispose d'un mois pour atteindre leurs objectifs et certains se laisse jusqu'à 2 mois.

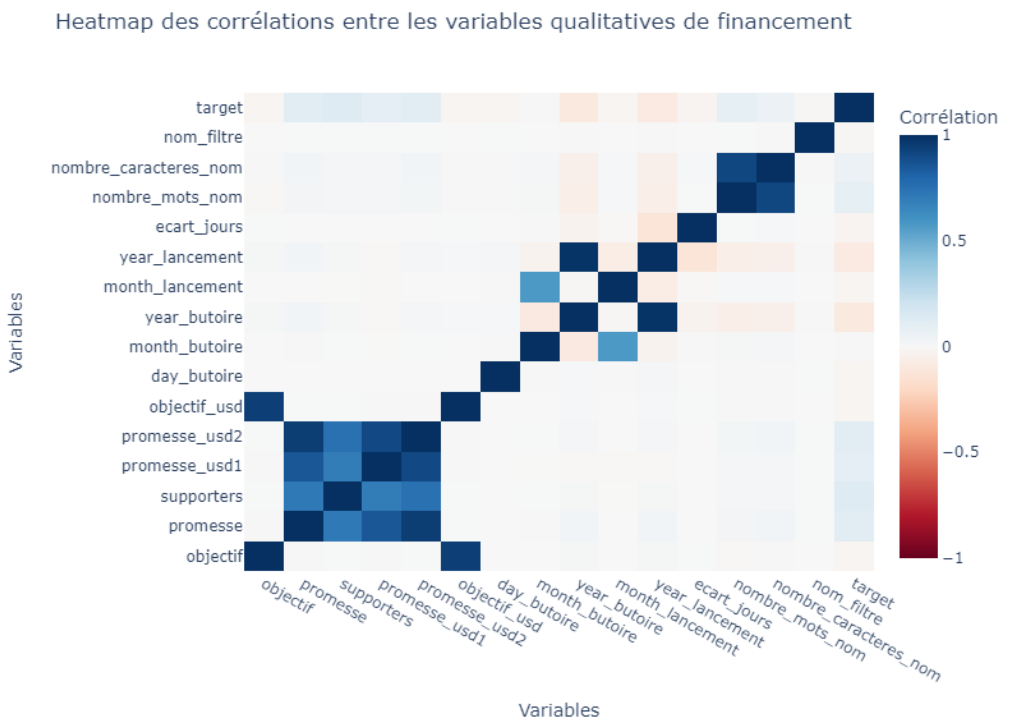


Nous distinguons clairement sur ce graphique une même pyramide plutôt centrée pour les objectifs comme pour les promesses. En revanche, la pyramide des promesses n'est pas en phase avec celle des objectifs. Elle est légèrement décalée sur la gauche, ce qui explique que le taux du succès du financement des projets n'est pas plus haut. Par ailleurs, gardons en tête que chaque barre est une puissance de 10. Un décalage d'une barre peut donc être multiplié jusqu'à 10 fois dans les valeurs.



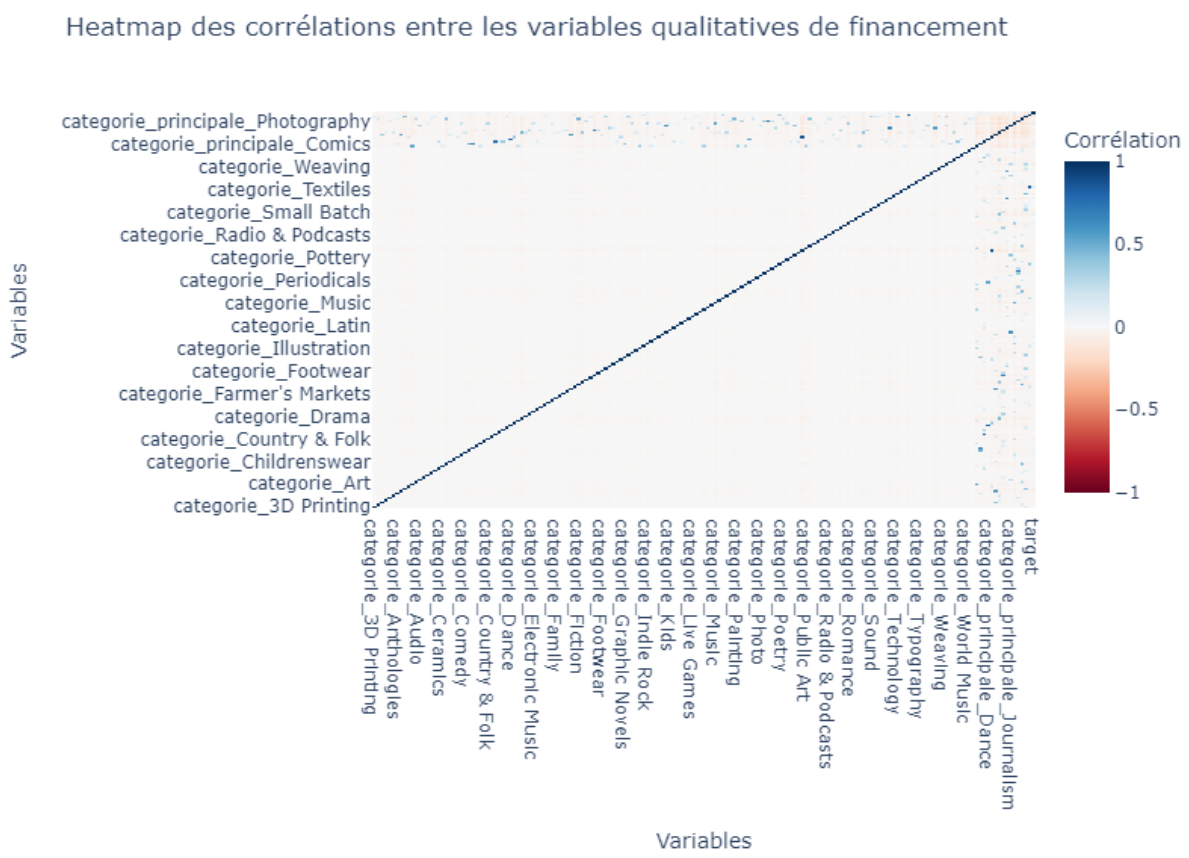
La distribution des supporters ci-dessus montre que la plupart des projets n'en nécessitent que peu. En effet, la masse est concentrée en deçà de 100 supporters au maximum avec seulement 20% des projets au-dessus de ce seuil.

Heatmap des corrélations entre les variables qualitatives



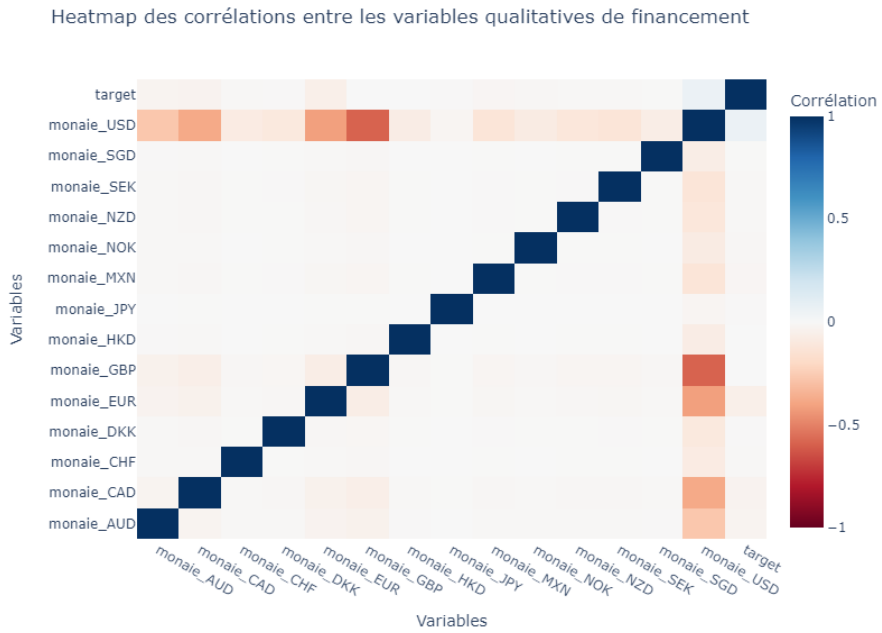
- Objectif et Objectif_usd sont liés ! Jusqu'ici, rien d'étonnant !
- La promesse est corrélée à promesse_usd1 et 2, pour les mêmes raisons que le sont celles du dessus.
- On trouve aussi une forte corrélation entre l'année de lancement de la recherche de fond et l'année butoire, indiquant une récolte de fond sur une période assez courte (35 jours en moyenne).
- On a une corrélation entre la taille du nom du projet et le nombre de mots qui le compose, là encore rien de surprenant.
- Enfin, une corrélation fortement négative relie l'appartenance du projet aux EUA et à l'Europe, là encore c'est parfaitement normal d'après les statistiques observées sur cette variable initialement.

Heatmap des corrélations avec les catégories dummifiées



- On observe quelques points de couleurs entre les corrélations des catégories principales. Cela est encore dû à l'étape de dummification des variables.
- En revanche, aucune corrélation n'est à relever entre la cible (état final = succès) et une quelconque catégorie (Cela est observable en zoomant à l'aide de la version plotly de cette heatmap directement dans le notebook).

Heatmap des corrélations avec les monaies dummifiées



- Nous n'observons encore une fois aucune corrélation avec la cible.
- Seule une corrélation est présente sur la monnaie utilisée puisque le dollar américain représente la majeure partie du jeu de données.

Tous ces graphiques de corrélations nous confortent dans la sélection des variables à utiliser pour entrainer nos différents modèles par la suite !

IV. Nos prédictions

a. Modélisation du financement

Dans cette section, nous abordons la phase de modélisation visant à prédire si un nouveau projet inscrit sur la plateforme sera financé ou non. Pour ce faire, nous avons adopté une approche axée sur l'apprentissage automatique en utilisant uniquement les données disponibles lors de l'inscription d'un projet sur le site.

Sélection des Variables

Dans le cadre de notre problématique, nous avons pris soin de sélectionner uniquement les variables pertinentes, c'est-à-dire celles qui sont susceptibles d'être fournies par un porteur de projet lors de son inscription sur la plateforme. Ainsi, nous avons éliminé les variables telles que le nombre de financeurs potentiels (supporters), le pays de la promesse de financement et le montant promis (promesse). Nous avons également supprimé la variable "objectif", car elle était redondante avec la variable “objectif_usd” qui contient la même information mais dans une seule et même unité.

Méthodologie dans le choix des Modèles

Nous avons choisi d'explorer plusieurs modèles d'apprentissage automatique pour résoudre notre problème de prédiction de financement de projet. Parmi les modèles considérés, nous avons utilisé :

- Régression Linéaire
- Régression Logistique
- Random Forest
- Gradient Boosting Machine (GBM)
- AdaBoost

Cette sélection découle d'une réflexion méthodique. Nous avons commencé par évaluer l'efficacité des modèles linéaires pour déterminer s'il était possible de séparer les deux populations avec une simple droite. Par la suite, nous avons élargi notre exploration à d'autres modèles tels que la forêt aléatoire et le gradient boosting machine (GBM). Bien que ces derniers aient présenté des performances encourageantes, elles ne se sont pas avérées significativement meilleures que celles des modèles linéaires. Dans l'optique d'optimiser la séparation des données, nous avons alors expérimenté le SVM avec un noyau polynomial, espérant ainsi obtenir une amélioration potentielle des performances du modèle.

Evaluation des Modèles

Chaque modèle a été **évalué** en utilisant des métriques appropriées telles que **l'accuracy**, la **précision**, le **rappel** et le **f1-score**. Nous avons également effectué une validation croisée pour sélectionner les hyper **paramètres optimaux** sur la **Forêt Aléatoire**. Ces paramètres ont été attribués au modèle GBM, puisqu'effectuer un **cycle d'optimisation** sur ce modèle requiert beaucoup de ressources machines.

Voici les **performances obtenues** selon les différents modèles évalués :

- Nous avons débuté notre exploration en testant un modèle de **régression linéaire** pour évaluer la possible corrélation linéaire entre la variable cible et les autres champs du dataset. Ce premier modèle a produit une **accuracy** de prédiction de **0.68** pour la faisabilité de financement d'un projet. De plus, le modèle de régression linéaire a obtenu une **précision** de **0.66**, ce qui représente la capacité du modèle à prédire correctement les projets financés parmi toutes les prédictions positives (les projets prédits comme financés). Le **rappel**, quant à lui, était de **0.68**, indiquant la proportion de projets financés que le modèle a réussi à identifier parmi tous les projets réellement financés. Le **f1-score**, qui est une moyenne harmonique de la précision et du rappel, était de **0.68**. Il permet de mesurer l'équilibre entre la précision et le rappel, fournissant ainsi une évaluation globale de la performance du modèle en termes de capacité à identifier les projets financés tout en minimisant les faux positifs. Ces premiers résultats sont **encourageants** étant donné la faible complexité de ce modèle.

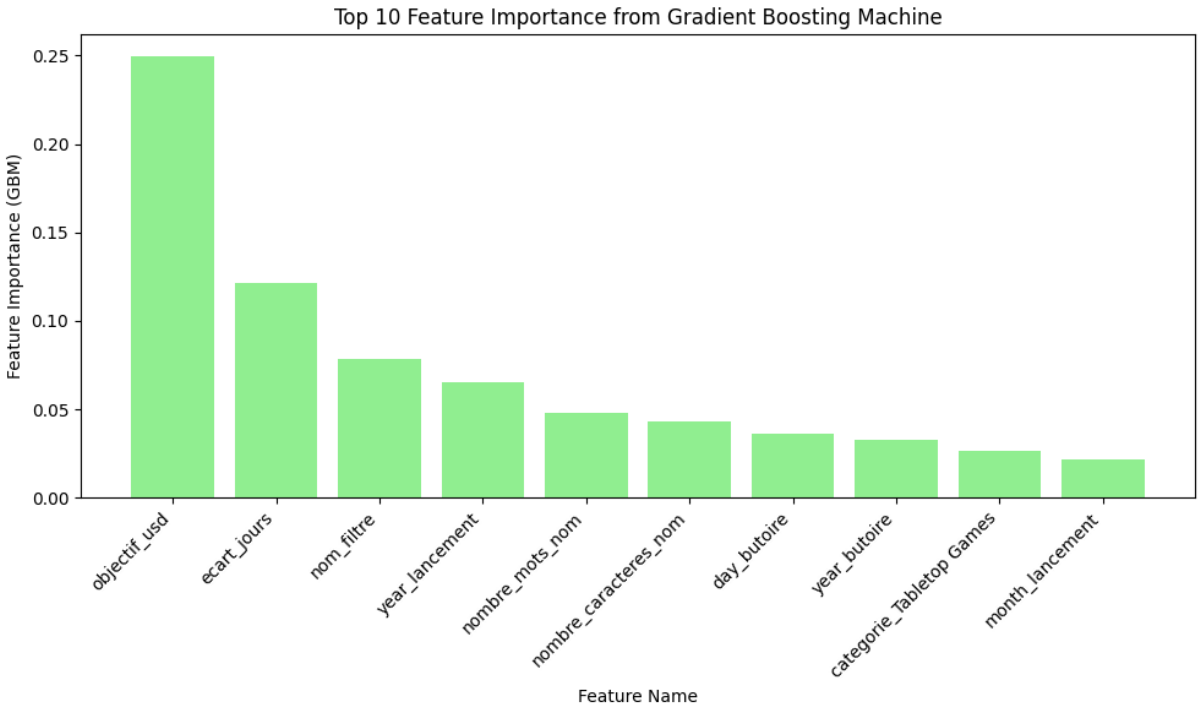
- Le modèle de **régression logistique** a produit une accuracy de **0.65** dans la prédiction de la faisabilité de financement des projets. Sa **précision** était de **0.62**, son **rappel** de **0.65**, et son **f1-score** de **0.65**. Ces résultats sont légèrement en dessous d'une simple régression linéaire, nous allons donc essayer d'explorer d'autre types de modèles.
- Nous avons mis en place un **grid search** pour **optimiser** les **paramètres** d'un modèle de **random forest**. Les paramètres optimaux trouvés par la **cross-validation** sur plusieurs combinaisons possibles de paramètres par le grid search sont :

  n_estimators (nombre d'arbres) = **300**

  max_depth (profondeur max des arbres) = **9**

  Le modèle de **random forest** a produit une **accuracy** de **0.67** dans la prédiction de la faisabilité de financement des projets. Sa **précision** était de **0.68**, son **rappel** de **0.67**, et son **f1-score** de **0.67**. Bien que **légèrement meilleur** que la régression logistique en termes d'accuracy, ses performances dans la prédiction des projets financés sont encore **modestes**.
- Le modèle de **Gradient Boosting Machine** (**GBM**) a obtenu une **précision** de **0.68**, un **rappel** de **0.67**, et un **f1-score** de **0.67** dans la prédiction de la faisabilité de financement des projets. Avec une **accuracy** de **0.72**, il surpasse légèrement les performances de la régression linéaire et de la random forest. Cependant, le **rappel** pour la **classe 1** (0.13) reste **faible**, indiquant une **difficulté** à **détecter** correctement les projets financés parmi tous ceux **réellement financés**.
- Le modèle **AdaBoost** a obtenu une **précision** de **0.67**, un **rappel** de **0.67**, et un **f1-score** de **0.71** dans la prédiction de la faisabilité de financement des projets. Son **accuracy** est de **0.70**. Bien que cette **performance** soit **similaire** à celle du **GBM**, le modèle pourrait sûrement encore être amélioré pour mieux détecter les projets financés.

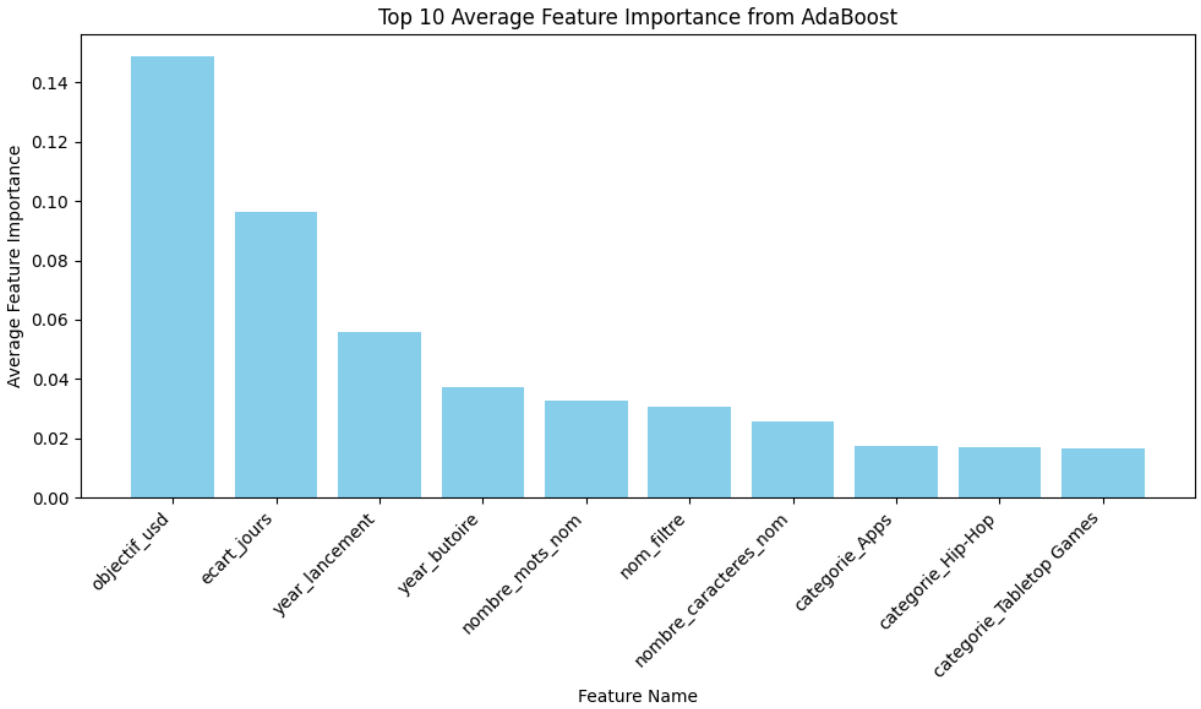
Importance des features du jeu de données d'entraînement



Les **caractéristiques** les plus **importantes** incluent **l'objectif** USD, **l'écart de jours**, le **nom du filtre** et **l'année de lancement.**

D'autres **caractéristiques**, autour du **nom du projet**, telles que le nombre de mots et de caractères dans le nom, ainsi que des **catégories spécifiques** comme "Tabletop games", sont également **notables**.

Certaines variables ont une importance inférieure à 0.05.



Les **caractéristiques clés** sont **similaires** au GBM, avec une **priorité** accordée à **l'objectif** USD, l'écart de jours et l'année de lancement.

En outre, des **catégories spécifiques** telles que "Apps", "Hip-hop" et "Tabletop games" sont considérées comme **importantes**.

Certaines variables ont une importance inférieure à 0.04.

Dans le modèle **GBM**, **l'objectif** USD se démarque comme la caractéristique la plus influente, tandis qu'**AdaBoost** partage une attention **similaire** envers cette variable, bien que légèrement **moins importante**. **GBM** présente une **diversité** plus **large** dans les caractéristiques importantes, avec un accent sur les délais et les catégories spécifiques, tandis qu'**AdaBoost** offre une **priorisation** plus **claire**. Cependant, la complexité d'interprétation est un défi potentiel avec GBM en raison de la dispersion des importances des variables. Ces différences soulignent l'importance de choisir le modèle le mieux adapté aux besoins spécifiques de l'application, en fonction de la manière dont il pondère ces caractéristiques dans ses prédictions.

b. Modélisation du niveau de financement

Dans cette section, nous abordons la phase de modélisation visant à prédire le montant de financement d'un projet. Nous adoptons une approche axée sur l'apprentissage automatique en utilisant uniquement les données disponibles lors de l'inscription des projets sur le site.

Sélection des Variables

Nous avons sélectionné uniquement les variables pertinentes pour notre problématique, en éliminant les variables telles que le nombre de financeurs potentiels, le pays de la promesse de financement, et le montant promis (qui devient ici la cible). Nous avons également ciblé les projets qui ont été financés pour rester cohérents avec notre problématique.

Méthodologie dans le choix des Modèles

Nous avons exploré plusieurs modèles d'apprentissage automatique pour résoudre notre problème de prédiction du montant de financement des projets. Nous avons utilisé une régression linéaire, un random forest, un gradient boosting machine et un SVM. Cette sélection découle d'une réflexion méthodique, commençant par l'évaluation des modèles linéaires avant d'élargir notre exploration à des modèles plus complexes.

Les performances des modèles linéaires et non linéaires ont été encourageantes, mais n'ont pas significativement surpassé celles des modèles linéaires. Dans l'optique d'optimiser davantage la séparation des données, nous avons envisagé d'utiliser un SVM avec un noyau polynomial. Cependant, en raison du temps de calcul excessivement long requis par ce modèle, nous avons décidé de ne pas analyser les résultats obtenus

Evaluation des Modèles

Chaque modèle a été évalué en utilisant des métriques telles que Mean Squared Error R-squared Mean et Percentage Error. Nous avons effectué différents tests sur ces modèles au niveau des hyper paramètres tels que n_estimators ou max_depth.

Régression Linéaire

Le modèle de régression linéaire affiche un MSE de 11,295,008,358.37, indiquant une certaine dispersion des prédictions par rapport aux valeurs réelles.

Le coefficient de détermination R^2 est de 0.371, ce qui suggère que le modèle explique 37.1% de la variance dans les données, démontrant une adéquation modérée entre les prédictions du modèle et les valeurs réelles.

Le MPE est de 884.99, ce qui soulève des préoccupations quant à la précision relative des prédictions, indiquant que les prédictions du modèle peuvent avoir une tendance à être largement éloignées des valeurs réelles en termes de pourcentage.

Random Forest

Le modèle du Random Forest présente un MSE de 14,291,882,719.92, indiquant une plus grande dispersion des prédictions par rapport à la régression linéaire.

Le coefficient de détermination R^2 est de 0.204, montrant une capacité limitée du modèle à expliquer la variance dans les données.

Le MPE est de 136.41, inférieur à celui de la régression linéaire, ce qui suggère que les prédictions du modèle sont plus proches des valeurs réelles en termes de pourcentage.

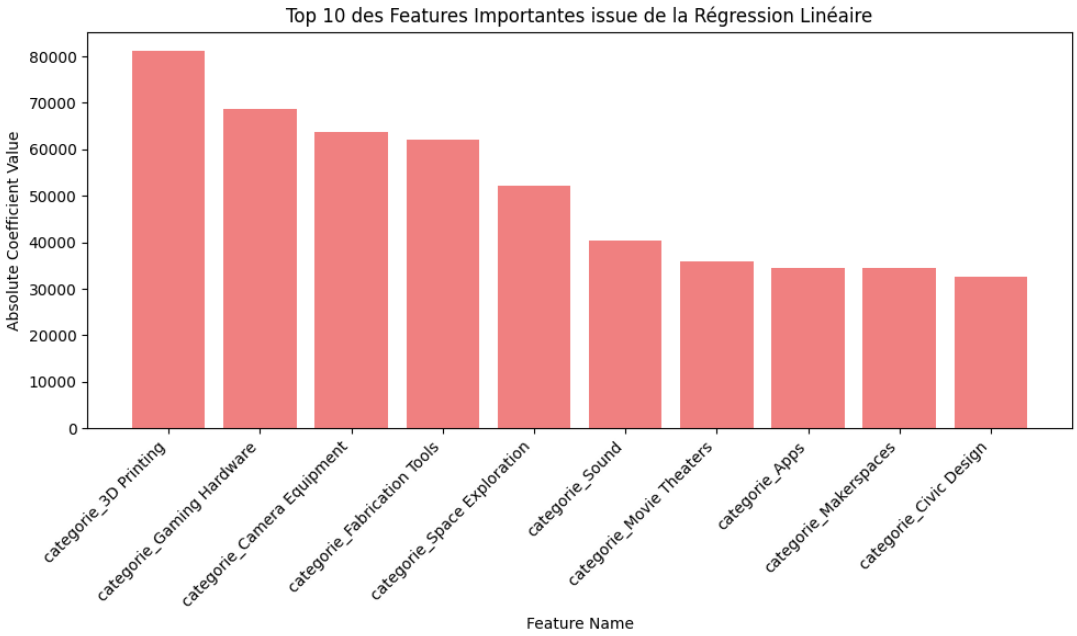
Gradient Boosting Machine (GBM)

Le GBM affiche un MSE de 11,473,466,572.60, légèrement plus élevé que celui de la régression linéaire.

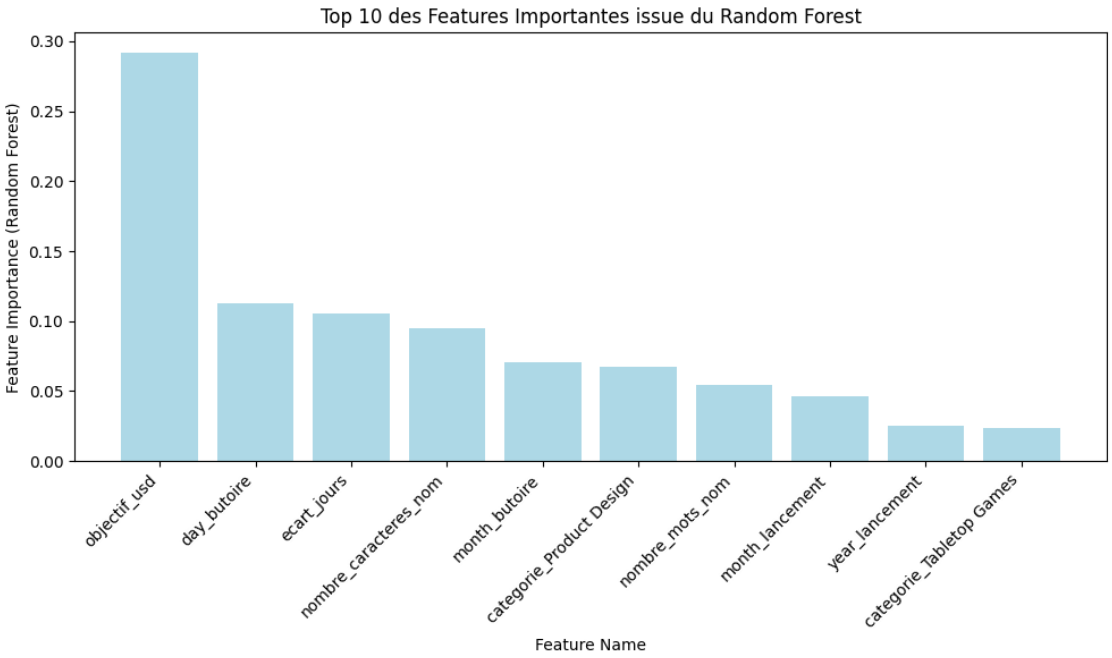
Le coefficient de détermination R^2 est de 0.361, montrant une capacité similaire à la régression linéaire à expliquer la variance dans les données.

Le MPE est de 477.61, plus élevé que celui de la régression linéaire mais inférieur à celui du Random Forest, suggérant une précision relative modérée des prédictions.

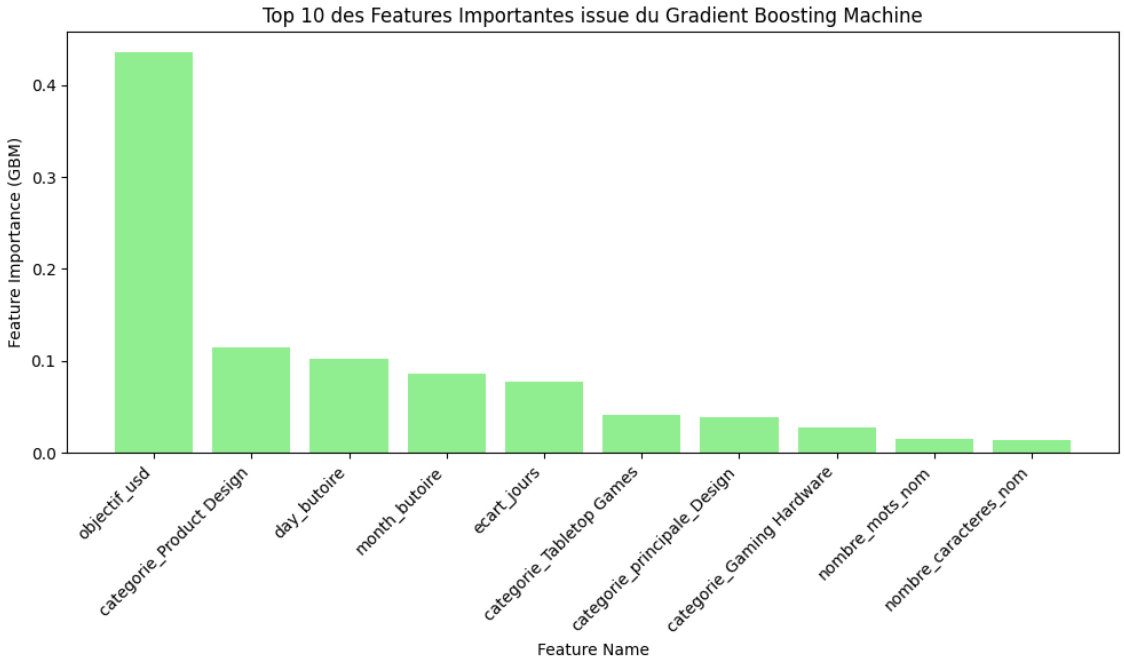
En examinant les performances des trois modèles, nous constatons que la régression linéaire a produit les résultats les plus favorables en termes de MSE et de R^2, bien que le MPE reste élevé. Le Random Forest a montré des performances légèrement inférieures, tandis que le GBM se situe entre les deux autres modèles en termes de précision et d'explication de la variance.



Notre premier modèle de **Régression Linéaire utilise** principalement les variables de notre variable binarisée **catégorie**.



Dans notre modèle **Random Forest**, nous voyons une **importance** très **forte** sur **objectif_usd**. Ce qui est **normal** car tous nos **projets financés dépassent l'objectif**. C'est donc une information incontournable.



Pour ce dernier modèle, le **GBM utilise** aussi beaucoup **l'objectif usd** comme le random forest. On peut noter une certaine **similitude** dans l'utilisation des **champs importants** entre le **random forest et le GBM** alors que notre régression linéaire utilise des variables totalement différentes.

V. Conclusion

En conclusion, cette étude approfondie sur le financement a mis en évidence l'importance cruciale de variables telles que l'objectif financier, la durée de la campagne, et les caractéristiques du projet pour prédire son succès. Les modèles d'apprentissage automatique ont montré une capacité prometteuse mais pas encore parfaite à prédire la faisabilité de financement des projets, bien que la prédiction du montant de financement reste un grand défi. L'analyse textuelle s'avère également être une piste intéressante pour améliorer les prédictions. Nos modèles n'ont, certes, pas des résultats très satisfaisants, mais se rapprochent au maximum de nos attendus. Nous avons aussi fait des tests en mettant les données de supporter et de pays et nos modélisations étaient bien plus performantes, voire trop performantes …