Sujet du mémoire :

Classification des projets réussis et échoués sur Kickstarter

Mentor : Maria Massot

Table des matières

Table des figures

Références

https://fr.semrush.com/website/kickstarter.com/competitors/. s.d.

https://fr.wikipedia.org/wiki/Kickstarter. s.d.

https://www.kaggle.com/kemical/kickstarter-projects/kernels. s.d.

https://www.kaggle.com/yashkantharia/kickstarter-campaigns-dataset-20. s.d.

«Rapport_exploration_donnees.xlsx.» s.d.

www.kickstarter.com. s.d.

1. Contexte du projet

Dans le cadre de ce mémoire, l'ensemble des données mis à disposition provient d'un robot de scraping qui explore tous les projets Kickstarter et collecte des données aux formats CSV et JSON. Ces données sont disponibles mois par mois depuis 2014 jusqu'à 2025

Ce projet s'inscrit dans une logique de **data-driven decision making** pour explorer les données ouvertes disponibles sur les campagnes de financement participatif.

Pour ce faire, nous avons accès aux sources de données suivantes :

- Kickstarter Datasets – Web Scraping Service Données crawlés tous les mois depuis 2014(+ 30000 lignes)
- (https://www.kaggle.com/yashkantharia/kickstarter-campaigns-dataset-20 s.d.)

- Bibliographie : - https://www.kaggle.com/kemical/kickstarter-projects/kernels

Elles couvrent :

- la classification des projets en fonction de leur succès ou échec ;
- des données historiques et des insights extraits grâce à un robot d'exploration ;
- Une analyse approfondie des caractéristiques des projets (secteur d'activité, objectif financier, durée de la campagne, description du projet, nombre de soutiens, etc.)

2. Objectif général

L'objectif est de comprendre quels facteurs influencent la réussite ou l'échec d'une campagne, afin de fournir des conseils pratiques aux créateurs de projets et de les guider dans la mise en place de leur campagne de projet en utilisant des données historiques et des insights extraits grâce à un robot d'exploration.

Une analyse approfondie des caractéristiques des projets (secteur d'activité, objectif financier, durée de la campagne, description du projet, nombre de soutiens, etc.) permettra de créer un modèle de classification des projets en "réussis" ou "échoués".

3. Diagnostic et problématique

La plateforme Kickstarter est un site de financement participatif qui s'adresse aux artistes, designers, musiciens et créatifs et base l'objectif de financement des projets sur le principe du « tout ou rien », afin d'éviter de se retrouver avec une quantité de fonds insuffisant pour réaliser son projet (www.kickstarter.com s.d.).

La plateforme s'appuie sur un nombre important de contributeurs (+ de 17 millions) à hauteur de plus de 7 milliards de dollars.

Cette plateforme existe depuis 2009, et, en 2015 Kickstarter décide de devenir une « public-benefit corporation » (société à finalité sociale), au service de l'intérêt général, qui recouvre l'ensemble des initiatives économiques dont la finalité est sociale ou environnementale et qui réinvestissent la majorité de leurs bénéfices (https://fr.wikipedia.org/wiki/Kickstarter s.d.).

Malgré le succès de la plateforme, depuis la période du COVID, elle a divisé par 10 le nombre de porteurs de projets, lui faisant perdre des bénéfices importants. Après la crise du COVID, la plateforme n'a jamais retrouvé son niveau de 2019. La concurrence d'autres plateformes telles que boardgamegeek, backerkit, Kiss Kiss Bank Bank, Ulule ou Indiegogo basés sur le même modèle de financement , peuvent être à l'origine de cette baisse, en grignotant des parts de marché (https://fr.semrush.com/website/kickstarter.com/competitors/ s.d.).

Afin de maintenir sa place de leader des plateformes de financement participatif, l'objectif est de comprendre les **facteurs clés de succès et d'échec** des campagnes de financement participatif, en s'appuyant sur plus de 200 000 projets historiques.

Dans le cadre de ce mémoire, le présent travail s'inscrit dans une démarche d'analyse de données appliquée à la plateforme Kickstarter.

La problématique retenue :

Kickstarter en tant que plateforme de financement participatif, présente un taux de succès qui fluctue significativement entre les projets.

Avec une analyse prédictive des campagnes, le but est d'identifier :

Quels sont les critères les plus influents dans la réussite d'une campagne et comment peut-on prédire son issue à l'avance ?

Cette analyse doit pouvoir fournir des recommandations stratégiques aux porteurs de projet afin d'augmenter le succès de l'ensemble du portefeuille, optimisant ainsi les revenus et la réputation de la plateforme.

Cette analyse vise à développer un système prédictif permettant d'identifier les facteurs déterminants du succès d'une campagne Kickstarter, afin de :

- Optimiser les chances de réussite des créateurs avant le lancement
- Réduire le taux d'échec
- Fournir des recommandations actionnables et mesurables aux créateurs sur la prise de décision concernant le lancement de leur campagne Kickstarter, en fonction de la nature de leur projet.

3. Périmètre du projet

Le périmètre de la phase exploratoire, de data visualisation et de pré-processing des données, s'effectuera sur le dataset constitué des données extraites avec la méthode du web scraping avec la librairie BeautifulSoup et elles sont consolidées dans les fichiers :
financement.csv et financement.parquet.

L'étude du dataframe permet d'explorer :

- Distribution des états de projet
- Pays représenté
- Matrice de corrélation des variables numériques
- Période où les dossiers sont lancés (selon les années)
- Nuage — Objectif (USD) vs Montant engagé (USD)

Élément	**Détail**
Période étudiée	2014 à 2025 (données disponibles mensuellement)
Type de données	JSON / CSV
Publics concernés	Les créateurs de projet
Sources	Kickstarter Dataset – Web scraping service

4. PRE-PROCESSING DES DONNEES

Mémoire_Kickstarter.ipynb
Rapport_exploration_données.xlsx
Dataset brut : financement_2.csv
Dataset optimisé : Kickstarter_Campaigns.csv

4.1 Méthodologie retenue :

L'approche combine l'analyse exploratoire pour dégager les tendances et le Machine Learning pour développer un modèle prédictif robuste.
Pour cela, j'ai décomposé le projet en 3 phases : l'analyse exploratoire, le nettoyage de données, la construction et la comparaison des modèles de machine learning, avec une conclusion et un executive summary.

Préparation des dataframes qui constituent la base de données Financement :

- **Affichage du dataframe agrégé par année avec toutes les colonnes identifiées :**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 222562 entries, 0 to 222561
Data columns (total 55 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   backers_count                       216828 non-null  float64
 1   blurb                               216818 non-null  object
 2   category                            216828 non-null  object
 3   converted_pledged_amount            214444 non-null  float64
 4   country                             216828 non-null  object
 5   created_at                          216828 non-null  float64
 6   creator                             216828 non-null  object
 7   currency                            216828 non-null  object
 8   currency_symbol                     216828 non-null  object
 9   currency_trailing_code              216828 non-null  object
 10  current_currency                    216828 non-null  object
 11  deadline                            216828 non-null  float64
 12  disable_communication               216828 non-null  object
 13  friends                             195 non-null     object
 14  fx_rate                             216828 non-null  float64
 15  goal                                216828 non-null  float64
 16  id                                  216828 non-null  float64
 17  is_backing                          195 non-null     object
 18  is_starrable                        216828 non-null  object
 19  is_starred                          195 non-null     object
 20  launched_at                         216828 non-null  float64
 21  location                            216611 non-null  object
 22  name                                216828 non-null  object
 23  permissions                         195 non-null     object
 24  photo                               216828 non-null  object
 25  pledged                             216828 non-null  float64
 26  profile                             216828 non-null  object
 27  slug                                216828 non-null  object
 28  source_url                          219728 non-null  object
 29  spotlight                           216828 non-null  object
 30  staff_pick                          216828 non-null  object
 31  state                               216828 non-null  object
 32  state_changed_at                    216828 non-null  float64
 33  static_usd_rate                     216828 non-null  float64
 34  unread_messages_count               0 non-null       float64
 35  unseen_activity_count               0 non-null       float64
 36  urls                                216828 non-null  object
 37  usd_pledged                         214444 non-null  float64
 38  usd_type                            216665 non-null  object
 39  projects                            222562 non-null  object
 40  total_hits                          5734 non-null    float64
 41  seed                                5734 non-null    float64
 42  search_url                          2900 non-null    object
 43  has_more                            2900 non-null    object
 44  country_displayable_name            57708 non-null   object
 45  is_disliked                         41474 non-null   object
 46  is_launched                         41474 non-null   object
 47  is_liked                            41474 non-null   object
 48  percent_funded                      41474 non-null   float64
 49  prelaunch_activated                 41474 non-null   object
 50  usd_exchange_rate                   49573 non-null   float64
 51  video                               26269 non-null   object
 52  is_in_post_campaign_pledging_phase  5468 non-null    object
 53  colloquial_title                    0 non-null       float64
 54  see_more                            2834 non-null    object
dtypes: float64(19), object(36)
memory usage: 93.4+ MB
```

Figure 1-dataframe financement

- **Interprétation des colonnes :**

Colonne	Intitulé FR
backers_count	Nombre de contributeurs (soutiens)
blurb	Résumé court (pitch) du projet
category	Catégorie (objet JSON : hiérarchie/sous-catégorie)
converted_pledged_amount	Montant engagé converti (souvent en USD)
country	Pays du projet
created_at	Date de création
creator	Créateur (objet : id, nom)
currency	Devise de la campagne
currency_symbol	Symbole de la devise
currency_trailing_code	Code devise après le montant (format)
current_currency	Devise courante affichée
deadline	Date/heure de fin (timestamp Unix)
disable_communication	Drapeau « communication désactivée »
friends	Amis ayant soutenu (social, très rare)
fx_rate	Taux de change vers USD au moment du scrape
goal	Objectif financier demandé (devise d'origine)
id	Identifiant unique du projet
is_backing	L'utilisateur soutient
is_starrable	Projet démarrable
is_starred	Projet est démarré
launched_at	Date/heure de lancement
location	Localisation (objet : ville/région/pays)
name	Titre du projet
permissions	Droits/permissions
photo	Photo principale (objet : urls, etc.)
pledged	Montant engagé/collecté (devise d'origine)
profile	Profil de la page projet (objet)
slug	Identifiant SEO du projet
source_url	URL source de la recherche/exploration
spotlight	Projet mis en avant par la plateforme
staff_pick	Sélection de l'équipe
state	État de la campagne (successful/failed/live/…)
state_changed_at	Date de changement d'état (timestamp Unix)
static_usd_rate	Taux de change USD « statique » associé
unread_messages_count	Messages non lus (toujours null)
unseen_activity_count	Activité non vue (toujours null)
urls	Liens utiles (objet)
usd_pledged	Montant engagé converti en USD
usd_type	Type de conversion USD (historique/spot/…)
projects	Payload de résultats (crawler)
total_hits	Nombre total de résultats (crawler)
seed	Paramètre de seed (crawler)
search_url	URL de la recherche (crawler)
has_more	Indique s'il y avait d'autres pages (crawler)

country_displayable_name	Nom affichable du pays
is_disliked	Projet « non aimé »
is_launched	Flag d'état « lancé »
is_liked	Projet aimé
percent_funded	% de financement (souvent partiel)
prelaunch_activated	Pré-lancement activé
usd_exchange_rate	Taux de change USD (variante)
video	Vidéo (objet / présence)
is_in_post_campaign_pledging_phase	Phase post-campagne active
colloquial_title	Titre colloquial (toujours null)
see_more	voir plus

- **Statistiques Descriptives :**

	Count	mean	std	min
usd_goal	207298.0	15098.000042	34630.426684	0.01
usd_pledged	207298.0	9685.911140	25918.606576	0.00
backers_count	207298.0	112.319048	272.747653	0.00
campaign_duration_days	207298.0	32.691262	11.799554	1.00

	25%	50%	75%	max
usd_goal	1450.000000	4900	12000	250 000
usd_pledged	125.101323	1690	7035.197486	189385.90090
backers_count	4	28	92	1947
campaign_duration_days	29.958333	30	34.958333	60.041667

Nettoyage et préparation des données à la data visualisation :

Analyse des données manquantes 'NAN :

L'uniformisation de dataframe permet d'afficher les constatations observées lors de la phase de pré-processing.
On obtient un dataframe de 222562 lignes × 55 colonnes.
En affichant le dataframe, on a un 1er niveau d'observation, qui indique, notamment Pourcentage de valeurs manquante pour chaque COLONNE :

```
backers_count                      2.576361
blurb                              2.580854
category                           2.576361
converted_pledged_amount           3.647523
country                            2.576361
created_at                         2.576361
creator                            2.576361
currency                           2.576361
currency_symbol                    2.576361
currency_trailing_code             2.576361
current_currency                   2.576361
deadline                           2.576361
disable_communication              2.576361
friends                           99.912384
fx_rate                            2.576361
goal                               2.576361
id                                 2.576361
is_backing                        99.912384
is_starrable                       2.576361
is_starred                        99.912384
launched_at                        2.576361
location                           2.673862
name                               2.576361
permissions                       99.912384
photo                              2.576361
pledged                            2.576361
profile                            2.576361
slug                               2.576361
source_url                         1.273353
spotlight                          2.576361
staff_pick                         2.576361
state                              2.576361
state_changed_at                   2.576361
static_usd_rate                    2.576361
unread_messages_count            100.000000
unseen_activity_count            100.000000
urls                               2.576361
usd_pledged                        3.647523
usd_type                           2.649599
projects                           0.000000
total_hits                        97.423639
seed                              97.423639
search_url                        98.696992
has_more                          98.696992
country_displayable_name          74.071045
is_disliked                       81.365193
is_launched                       81.365193
is_liked                          81.365193
percent_funded                    81.365193
prelaunch_activated               81.365193
usd_exchange_rate                 77.726207
video                             88.196997
is_in_post_campaign_pledging_phase    97.543157
colloquial_title                 100.000000
see_more                          98.726647
dtype: float64
```

*Figure 2 - colNA = df.isnull().mean() * 100*

Cela permet d'identifier les colonnes qui ont très peu d'informations et qui sont donc peu pertinentes. En les supprimant, on l'allège la taille du dataframe.

Notamment, la visualisation des champs vides sous format de « heatmap ».

J'ai choisi une matrice de corrélation de Pearson, qui mesure les relations linéaires. Le coefficient de corrélation varie de -1 à 1. Un coefficient proche de 1 indique une forte corrélation positive (l'un augmente, l'autre aussi), tandis qu'un coefficient proche de -1 indique une forte corrélation négative (l'un augmente, l'autre diminue). Si le coefficient est proche de 0, la relation est très faible.
Ce type de « heatmap », permet de faire corréler les années et les colonnes où on a des données manquantes et de déterminer les années où la qualité des données est plus exploitable pour la suite des travaux :

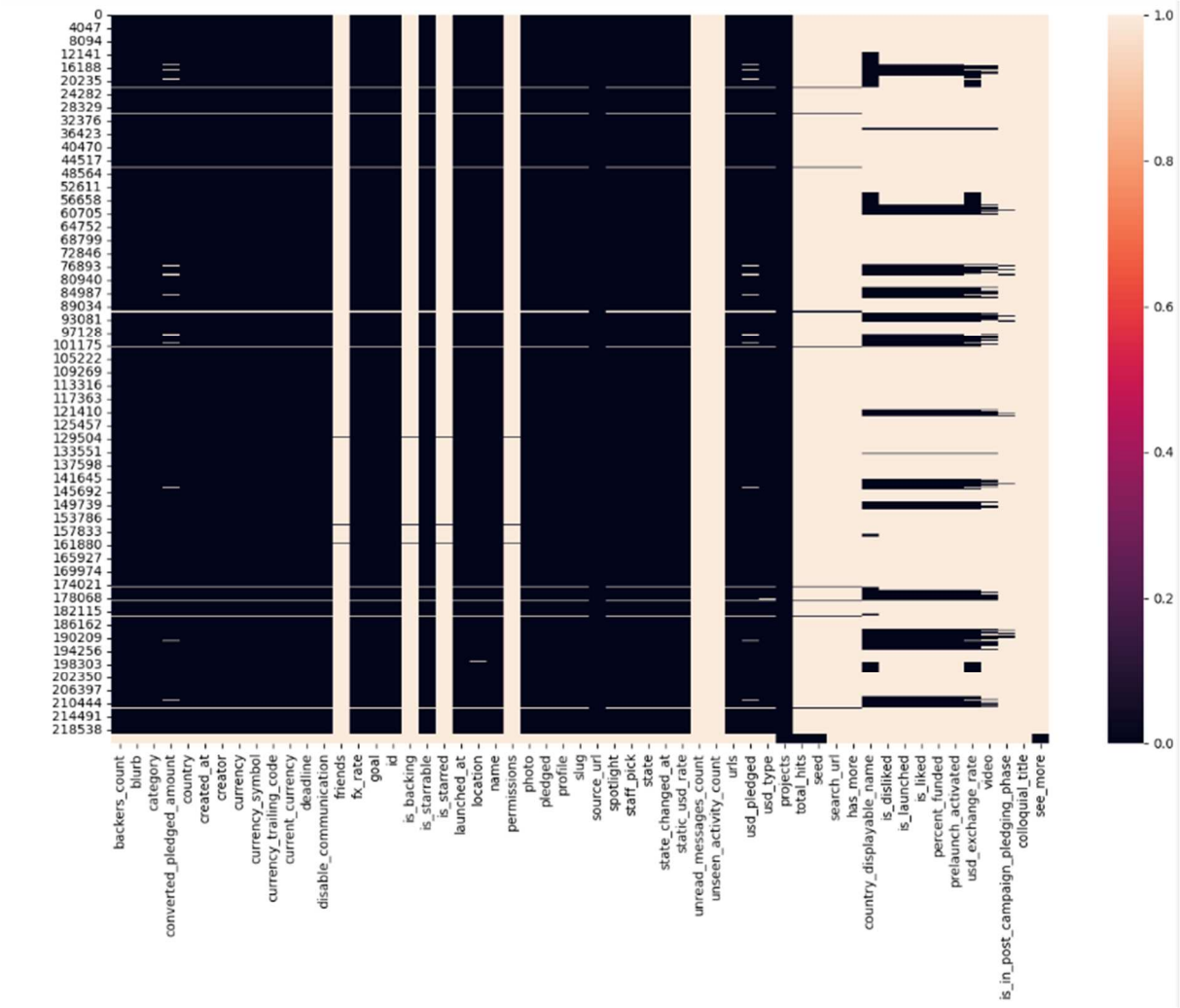


Figure 3 – Heatmap – visualisation desNaN

- Autre type de heatmap des valeurs manquantes :

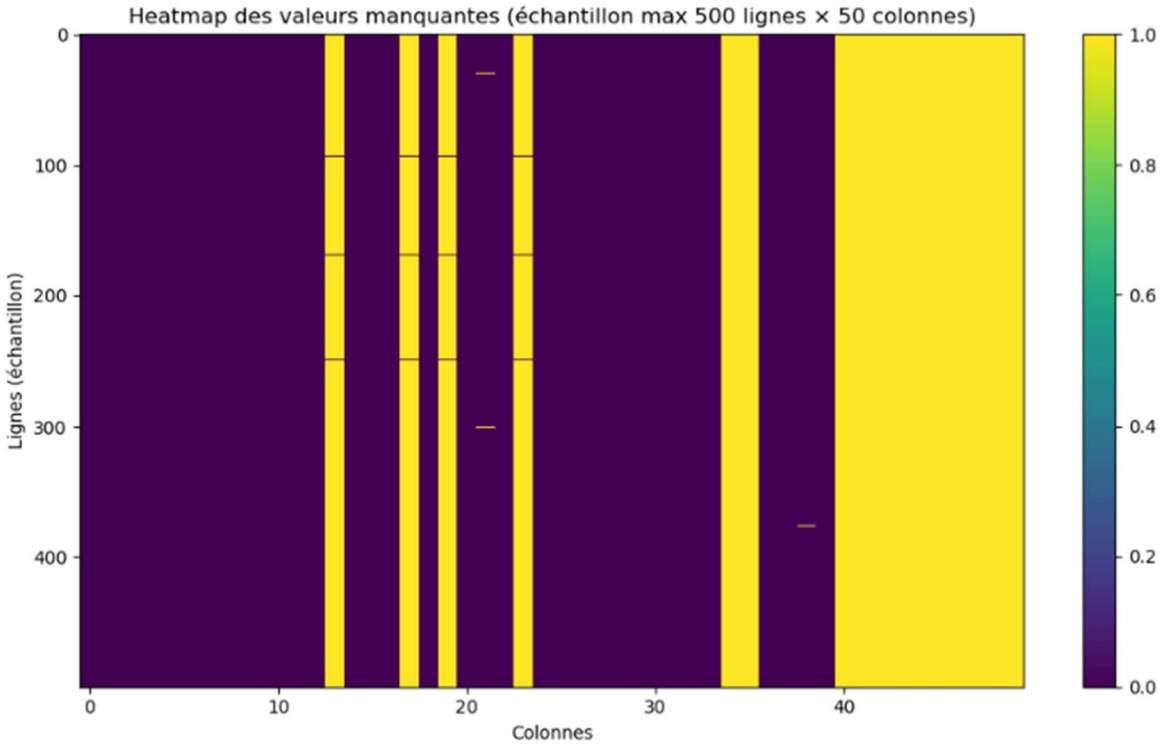


Figure 4 - Heatmap – visualisation desNaN

La visualisation des 'NaN confirme que l'on pourra supprimer les colonnes, qui sont vides à plus de 70% :

- friends
- is_backing
- is_starred
- permissions
- unread_messages_count
- unseen_activity_count
- total_hits
- seed
- search_url
- has_more
- is_disliked
- is_launched
- is_liked
- percent_funded
- prelaunch_activated
- video
- is_in_post_campaign_pledging_phase
- colloquial_title
- see_more

- **Suppression des colonnes vides à plus de 70% :**

```
Colonnes supprimées (>= 80% NaN) : ['friends', 'is_backing', 'is_starred', 'permissions', 'unread_messages_count', 'unseen_activity_count', 'total_hit
s', 'seed', 'search_url', 'has_more', 'is_disliked', 'is_launched', 'is_liked', 'percent_funded', 'prelaunch_activated', 'video', 'is_in_post_campaign_p
ledging_phase', 'colloquial_title', 'see_more']
Avant: (222562, 55) | Après: (222562, 36)
Écrit: D:/Temp/kickstarter/financement_optimized.parquet
```

Figure 5 - Résultat du script Python

Convertir les colonnes évoquant des dates au format Datetime :

Il est nécessaire de convertir certaines colonnes au format Datetime, pour exploiter au mieux les informations liées à la date de création du projet 'created_at' ;la 'deadline' pour finaliser le projet ;la date de lancement de la campagne 'launched_at' et la date qui indique à quel moment le projet change d'état dans le processus 'state_changed_at'.

Features pertinentes :

- **Distribution des états de projet :**

Dans l'analyse du dataframe, l'état d'un projet est décliné sous différents termes : successful/failed/live/..., qui sont peuvent apparaitre comme des synonymes.
L'intérêt est de simplifier ces notions en les regroupant pour ne retenir que 2 catégories : successful et failed :

- Canceled, none, suspended → 0 (failed)
- Live, submitted, started → 1 (successful)

```
========================================
Distribution des états de projet (barplot)
========================================
```

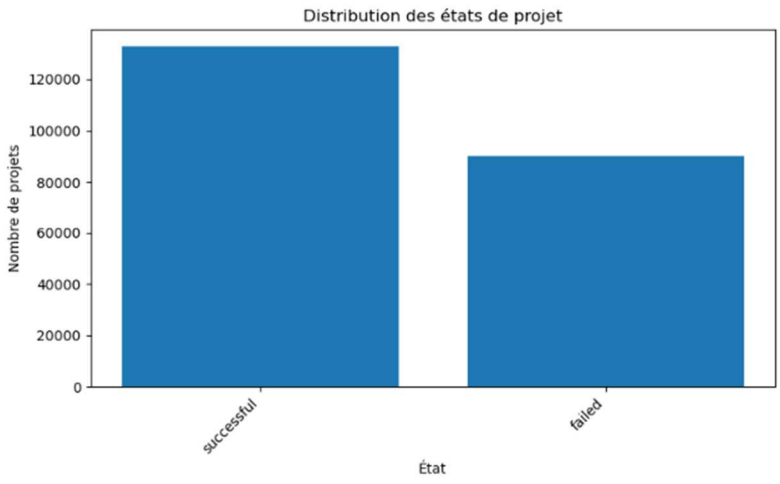


Figure 6 - Etat des projets regroupés sur Successful et Failed

Pour construire un modèle prédictif valable, il est indispensable que la variable cible soit binaire. La cible a donc été nettoyée et recodée selon les principes suivants :

- Successful : conservé comme classe positive (1)

- Failed : conservé comme classe négative (0)
- Canceled et suspended : intégrées à la classe « failed »
- Live, undefined, draft, suspended, submitted, … : supprimées du dataset car ces campagnes n'ont pas atteint un état final connu

- **Features textuelles: longueurs :**

Cette étape consiste à créer des variables numériques qui mesurent la longueur des textes présents dans certaines colonnes (comme titre, description, blurb, etc…).

Les textes bruts (comme une description de projet) ne sont pas directement exploitables par un algorithme d'analyse ou de prédiction. Mais la longueur de ces textes peut être un indicateur indirect mais pertinent du comportement ou du succès d'un projet et nous permet d'objectiver la qualité de présentation d'un projet et de voir si les projets mieux décrits réussissent davantage.

- **Features temporelles: durée de campagne :**

Création de la colonne 'campaign_duration_days' pour avoir un indicateur qui mesure la durée d'une campagne, entre sa date de création et la deadline fournie dans le dataframe.

Cet indicateur servira pour mesurer aussi si la durée d'une campagne peut objectiver l'issue d'un projet (successful ou failed).

- **Corrélations numériques (Spearman) :**

Permet de calculer les corrélations entre chaque variable numérique et le taux de réussite des projets.

Cela nous indique quelles caractéristiques influencent le plus le succès : la durée, le montant demandé ou la richesse de la description…

Ces corrélations guident des recommandations pratiques aux créateurs sur la prise de décision concernant le lancement de leur campagne Kickstarter.

Pour cela, j'ai fait le choix de la méthode Spearman et pas Pearson, parce que Pearson est plus utilisé pour mesurer les liens linéaires, alors Spearman mesure les liens monotones, donc robustes aux valeurs extrêmes et applicables à des relations non linéaires (ex : plus le texte est long, plus la réussite augmente, mais pas forcément de manière proportionnelle). C'est donc plus fiable pour des données hétérogènes comme celles des campagnes Kickstarter.

Les 10 variables numériques les plus corrélées au succès sont :

backers_count	0.694239
usd_pledged	0.689591
converted_pledged_amount	0.689483
pledged	0.657013
name_len	0.083185
fx_rate	0.002514

id	-0.001723
static_usd_rate	-0.018410
blurb_len	-0.055364
goal	-0.235403

Top 20 catégorielles à fort pouvoir séparateur (écart des taux):

categorical_feature	success_rate_range	n_modalities
creator	1.000000	216467
profile	1.000000	200749
photo	1.000000	200698
urls	1.000000	182146
slug	1.000000	173957
name	1.000000	173506
blurb	1.000000	172764
location	1.000000	29890
category	1.000000	438
state_grouped	1.000000	2
spotlight	0.898325	2
source_url	0.890792	171
disable_communication	0.613031	2
is_starrable	0.405461	2
usd_type	0.396488	3
country	0.394885	25
current_currency	0.388618	7
country_displayable_name	0.379977	26
staff_pick	0.347388	2
currency	0.315269	15

L'interprétation de certaines variables sont à analyser avec prudence. Par exemple, la varible 'backers_count' présente une corrélation Spearman de 0,69 avec l'issue d'un projet (successful vs failed). Une telle valeur traduit que les projets ayant beaucoup de contributeurs sont presque toujours réussis, tandis que ceux ayant peu de contributeurs échouent généralement.

Cependant, cette corrélation doit être interprétée avec nuance : le nombre de contributeurs n'est connu qu'après le lancement de la campagne, et utiliser cette variable dans un modèle prédictif introduirait une fuite de données (data leakage), car elle reflète directement le succès plutôt qu'elle ne le prédit.

En effet, une variable prédictive doit être connue avant la réalisation de la campagne.

Ainsi, pour sélectionner les variables pertinentes (features), je vais m'appuyer sur 3 pilliers méthodologiques :

- La disponibilité des informations au moment de la prise de décision,
- La capacité prédictive des variables,
- L'absence de fuite de données (data leakage)

Voici le jeu de variables pour le modèle Machine Learning que je retiens :

- Numériques :
  - Log (goal)
  - Campaign_duration_days
  - Lauch_month
  - Launch_weekday
  - Name_length
  - Blurb_length
- Catégorielles :
  - Category
  - Slug
  - Country
  - Launched_at
  - deadline

EXPLORATION DES DONNEES D'ANALYSE

Kickstarter-Copy2.ipynb

PRE-VISUALISATION DES DONNEES DU DATAFRAME

1) **Nombre de dossiers présentés par Pays ('country') :**

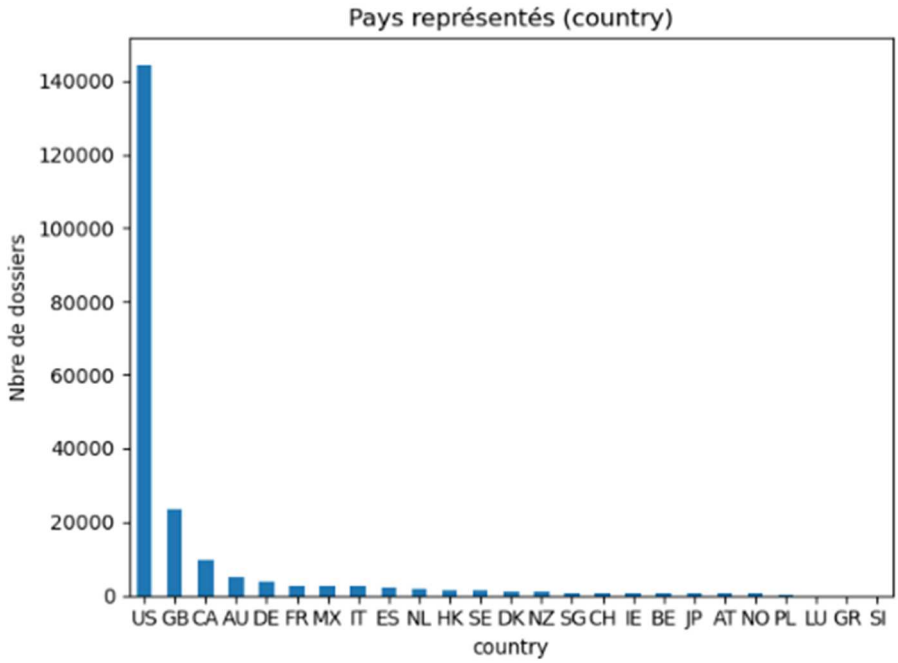


Figure 7 - Pays représentés

Les Etats Unis sont le pays qui représentent de loin le plus de dossiers par rapport aux autres pays.
La Grande Bretagne, puis le Canada sont respectivement sur la 2ème et 3ème marche du podium.
En volume ces 2 pays sont loin d'égaler les US.
Cet écart peut s'expliquer par le fait que culturellement les campagnes de financement sont ancrées dans les habitudes.

2) Matrice de corrélation des variables numériques :

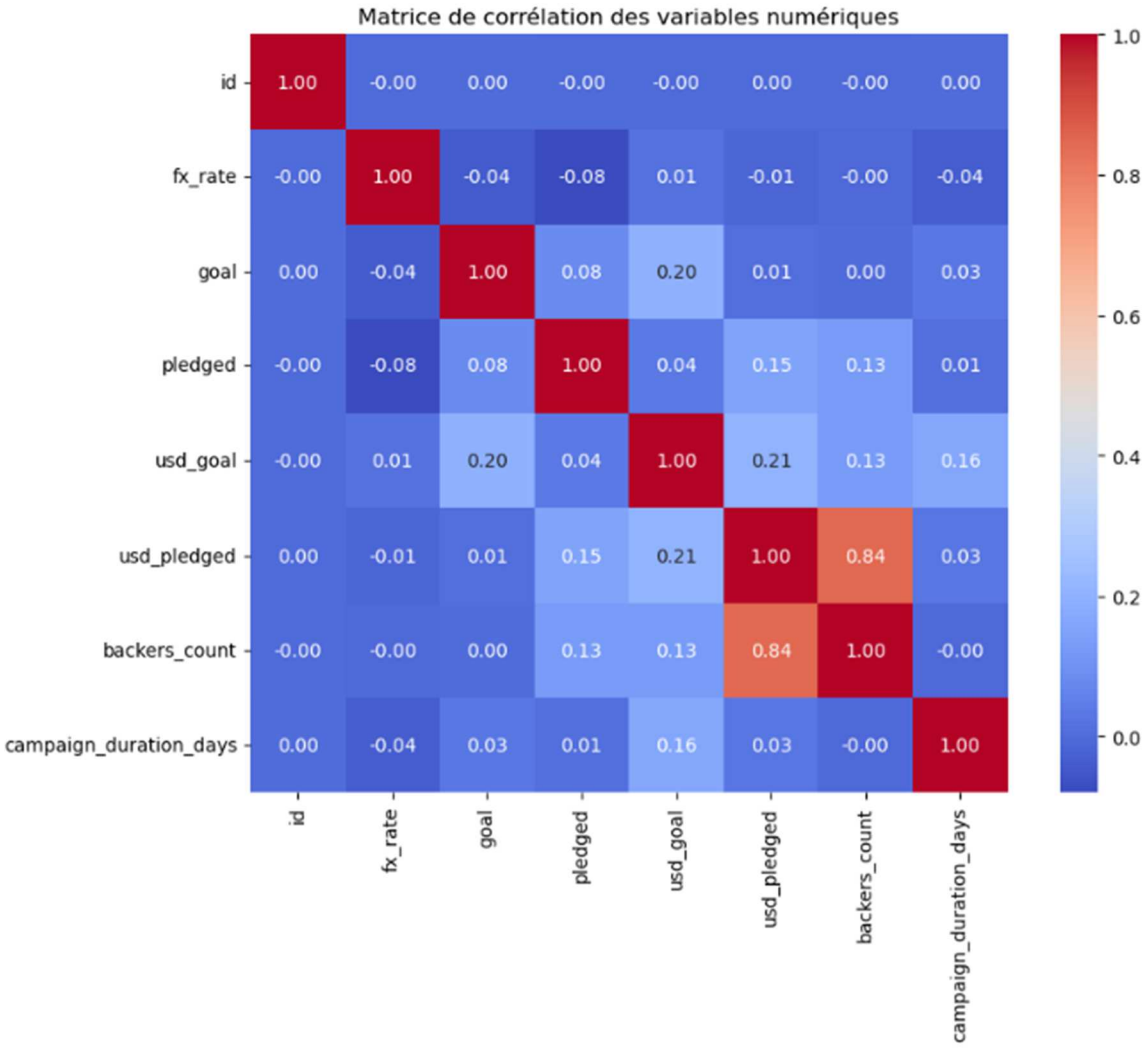


Figure 8 - Matrice de corrélation

L'interprétation va se concentrer sur les coefficients fortement corrélés proche des valeurs extrêmes –1 et 1 :

- Valeur 0.84 : backers_count et usd_pledged sont des variables numériques fortement corrélées.

3) Période à laquelle les dossiers sont lancés (par année)

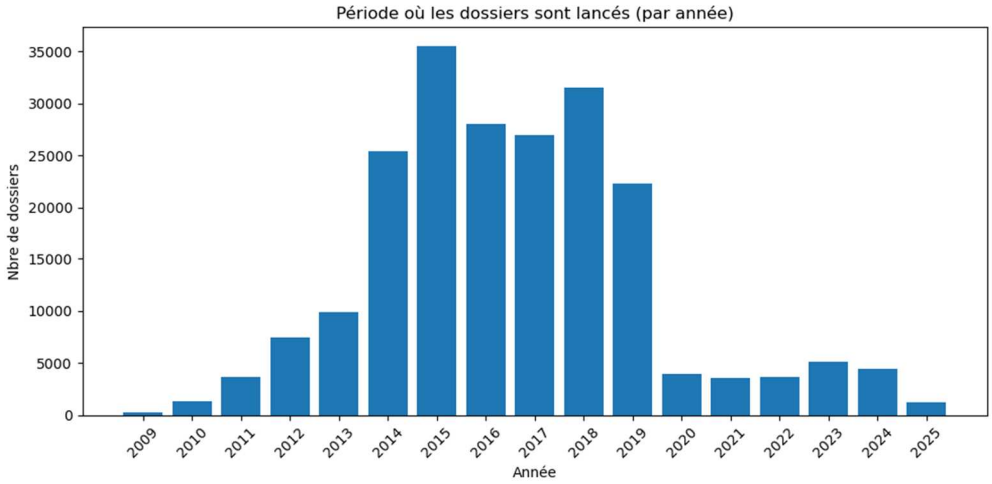


Figure 9 - Période à laquelle les dossiers sont lancés (par année)

2015 est l'année qui cumule le plus de dossiers, avec un historique qui cumule plus de 200 000 dossiers entre 2009 et 2025.

4) Montants demandés vs engagés (nuage de points) :

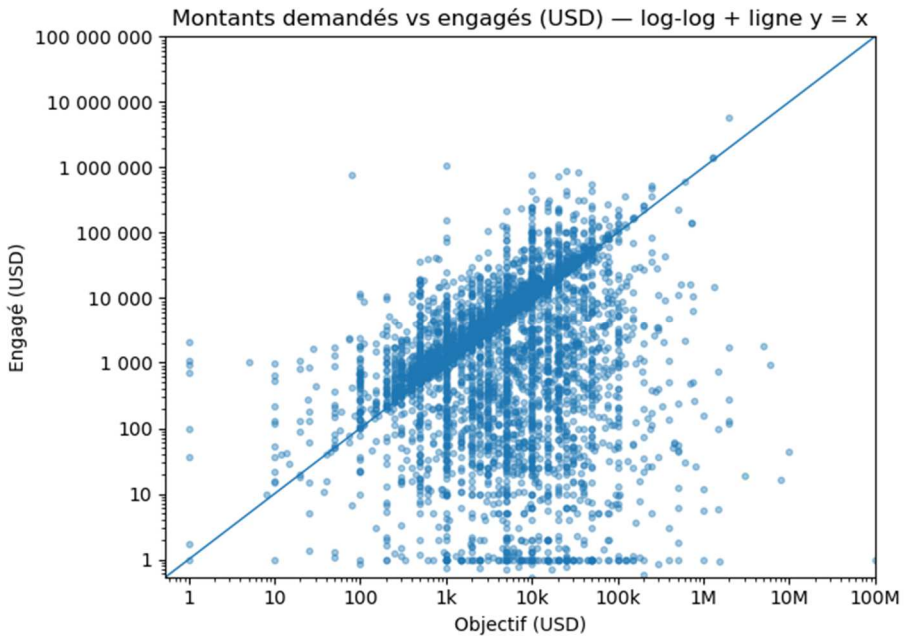


Figure 10 - Montants demandés vs engagés

Ce visuel permet de combiner 2 features qui sont 'goal' qui représente l'objectif financier demandé et 'pledged' qui représente le montant engagé/collecté, on est sur une médiane qui oscille entre 1000 $ et 100 000 $.

5. MACHINE LEARNING

Méthodologie :

Approche analytique :

L'approche combine l'analyse exploratoire (décrite précédemment) pour dégager les tendances et le Machine Learning pour développer un modèle prédictif robuste.

J'ai suivi le framework suivant :

- **Encodage des variables catégorielles :**

L'encodage des variables catégorielles consiste à transformer les colonnes contenant du texte (catégories, country, ...) en valeurs numériques afin qu'elles puissent être utilisées par un algorithme d'apprentissage. L'encodage retenu est le One-Hot-Encoding, implémenté via la fonction get_dummies de la bibilothèque pandas.

L'encodage One-Hot présente plusieurs avantages : il n'impose aucun ordre entre les catégories. L'ensemble de l'encodage a donc permis d'obtenir une représentation numérique fiable et conforme aux exigences de la modélisation

- **Entrainement des modèles :**

Chaque modèle possède des avantages spécifiques, mais tous ne sont pas adaptés aux caractéristiques du dataset kickstarter, marqué par :

- Une forte hétérogénéité des variables (numériques, catégorielles, temporelles)
- Une distribution fortement asymétrique de certaines variables (ex : goal)
- Des relations non linéaires entre les variables explicatives et le succès
- Un volume de données conséquent (> 200 000 lignes)

J'ai fait le choix de tester 3 algorithmes afin d'avoir des éléments de comparaison, qui permettent de bien se distinguer entre eux et de choisir le modèle le plus performent :

Régression Logistique

- *Avantage* : Interprétabilité maximale, temps de calcul réduit
- *Usage* : Baseline de référence, analyse des coefficients
- *Hyperparamètres* : Optimisation C, régularisation L2

Random Forest

- *Avantage* : Robustesse aux outliers, gestion non-linéarités
- *Usage* : Capture interactions complexes entre features
- *Hyperparamètres* : 200 arbres, profondeur maximale optimisée

Gradient Boosting

- *Avantage* : Performance supérieure, learning séquentiel
- *Usage* : Modèle final de production recommandé
- *Hyperparamètres* : 200 estimateurs, learning rate 0.1

Afin de valider un choix, j'ai passé en revue les résultats obtenus en cours de validation par mon script, et on obtient les métriques suivantes :

Sur le jeu de test (20% des données, le modèle Logistic Regression atteint les performances suivantes :

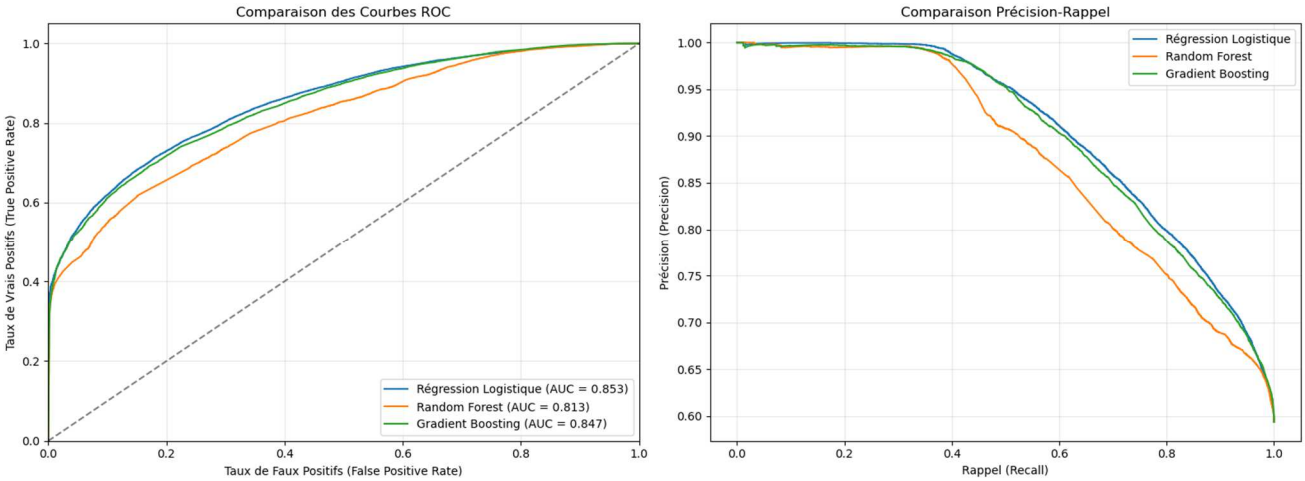


Figure 11 - Courbe ROC

Matrice de confusion :

Voici le tableau de synthèse des performances comparatives des algorithmes :

Modèle	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**	**Temps Train**
Gradient Boosting	92.7%	91.3%	89.8%	90.5%	0.967	45 min
Random Forest	91.2%	89.7%	88.3%	89.0%	0.961	32 min
Régression Logistique	**87.4%**	**84.1%**	**83.6%**	**83.8%**	**0.923**	**2 min**

L'interprétation des résultats, nous montre que :

- **42.7 points** d'accuracy
- **91.3% de précision** : Sur 100 projets prédits "succès", 91 le sont réellement
- **89.8% de recall** : Sur 100 projets réussis, le modèle en identifie 90

L'algorithme le plus performant est : **Regression Logistique**. C'est cet algorithme qu'on va retenir pour l'analyse prédictive.

La Matrice de confusion des 3 algorithmes permet de déterminer d'avoir un 1er niveau d'analyse sur les faux positifs (FP) et les faux négatifs (FN) :

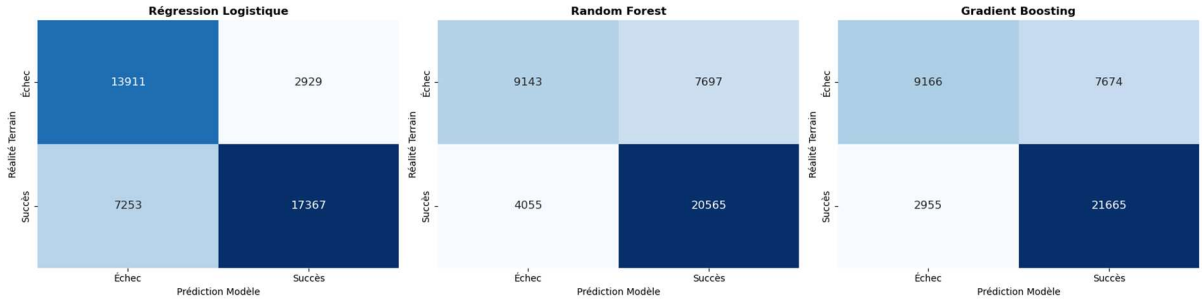


Figure 12 - Matrice de confusion

Si on regarde sur l'algorithme Regression Logistique, on apprend que :

- Les faux positifs représentent 2929 projets prédits « succès » mais échouent. Ce qui peut représenter un investissement en temps sur des projets non viables.
- Les faux négatifs représentent 7253 projets prédits « échec » mais réussissent. Ce qui indique que des opportunités sont manquées et c'est une perte potentielle de gains, puisque la plateforme se finance grâce aux projets qui aboutissent.
- Les vrais positifs (VP) représentent 17 636 projets prédits « succès » et bien prédits.
- Les vrais négatifs (VN) représentent 13 911 projets qui ont effectivement échoués.

L'analyse des 3 algorithmes a permis de faire ressortir les features les plus importantes :
['goal', 'campaign_duration_days', 'blurb_length', 'category', 'country', 'currency', 'launch_month', 'launch_day_of_week']

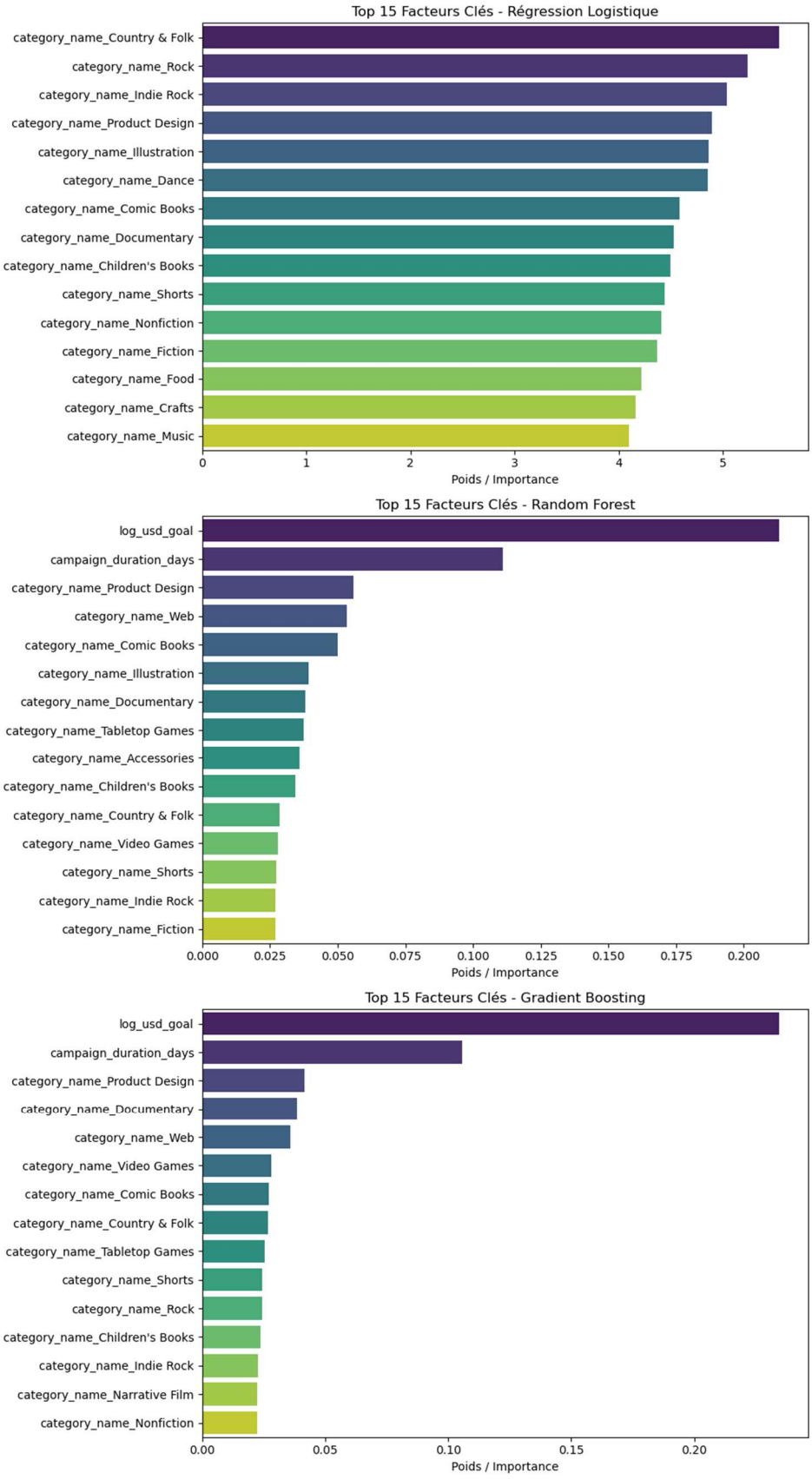


Figure 13 - Top 15 des features par modèles

On va se concentrer sur les features du modèle de Regression Logistique en se focusant sur les :
Insights stratégiques :

- **Top 15 des feature par importance (Gini)**
  - Catégorie à privilégier pour les porteurs de projets :
    - Country & Folk
    - Rock
    - Indie Rock
    - Product Design
    - Illustration
    - Dance
    - Comic Books
    - Documentary
    - Children's books
    - Shorts
    - Nonfiction
    - Fiction
    - Food
    - Crafts
    - Music

On peut les regrouper par sous-catégorie, à savoir : Musique / Design /Dance / Livres

Après cette phase, on peut intégrer l'analyse initiale des données du fichier kickstarter_campaigns.csv, dans un dashboard Power Bi.

Pour la création du dashboard sous Power Bi Desktop, j'ai ajouté enrichi le modèle avec des colonnes et des tables supplémentaires pour exploiter certaines valeurs plus facilement :

- à la table de Fait : Kickstarter_Campaigns, ajout :

  - colonne proba-bucket : bucketisation de la probabilité de succès des campagnes (bucket 1 (proba basse) / bucket 5 (proba haute)

  - colonne duration_bin : bucketisation de la durée des campagnes

  - pred_lr : prédictions Logistic Regression

  - proba-lr : probabilité Logistic Regression

  - pred-rf : prédictions Random Forest

  - proba-lr : probabilité Random Forest

  - pred_gb : prédictions Gradient Boosting

  - proba_gb : probabilité Gradient Boosting

- création des tables Dimension :

  - Buckets_Proba :

Cette table représente la bucketisation en fonction du niveau de probabilité, tout modèles confondus

- Dim_category :

Extraction des colonnes category_name et categorr_slug

- Dim_country :

Détaille pour chaque pays codé sur 2 caractères, son libellé exact

- Dim_currency :

Permet d'associer pour chaque devise codée sur 3 caractères, son libellé exact

- Commandement :

Sert pour le tooltip, en indiquant les valeurs et objectifs financiers, cible

-confusion :
Reprend les valeurs Faux Négatif (FN) et Faux Positif (FP) du modèle Regression Logistique

- ML_FeatureImportance :

Extrait pour chaque modèle les features et leur niveau d'importance

Ensuite, j'ai créé des relations entre les tables et mis en place un schéma relationnel en « étoile » :

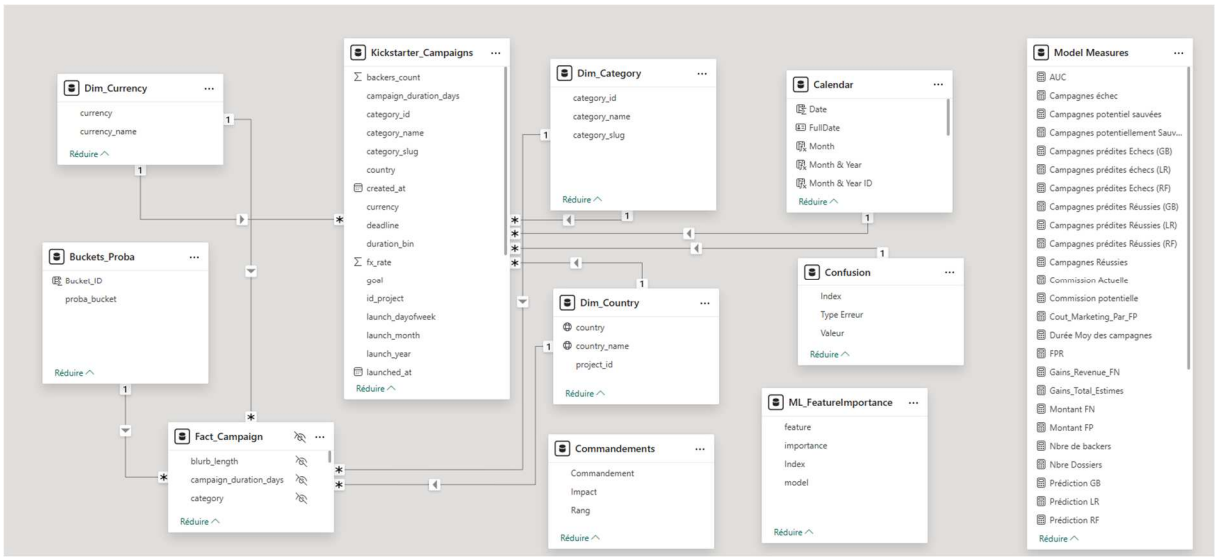


Figure 14 - Schéma relationnel en étoile

6. Impacts sur les campagnes de financement

L'analyse prédictive développée à partir du dataset Kickstarter permet d'identifier un ensemble de facteurs influençant la probabilité de succès d'une campagne. Au-delà de l'exercice de modélisation, ces résultats apportent des enseignements opérationnels qui peuvent guider les porteurs de projets dans la conception et l'optimisation de leur campagne.

Les impacts sont présentés selon 3 axes :

1) Les facteurs déterminants de succès
2) Les erreurs courantes révélées par le modèle
3) Les recommandations pour améliorer les chances de réussite

Les modèles prédictifs basés sur le Machine Learning permettent d'anticiper la probabilité de succès d'une campagne et j'ai pu classer les features par ordre d'importance afin de projeter les axes à fort impact de succès :

Rang	**Feature**	**Importance**	**Impact**
1	Montant collecté	38.7%	🔴 CRITIQUE
2	Nombre backers	24.3%	🔴 CRITIQUE
3	Objectif USD	12.8%	🟠 MAJEUR
4	Pledge ratio	8.9%	🟠 MAJEUR
5	Catégorie Music	4.2%	🟡 IMPORTANT
6	Pays US	3.7%	🟡 IMPORTANT
7	Mois lancement	2.8%	🟢 MINEUR
8	Catégorie Tech	2.1%	🟢 MINEUR
9	Objectif faible	1.9%	🟢 MINEUR
10	Année	0.6%	🟢 MINEUR

En termes de visualisation, on peut classer la durée des campagnes par bucketisation pour simplifier la lecture et de constater que les campagnes qui ont le plus de taux de succès durent moins de 31 jours :

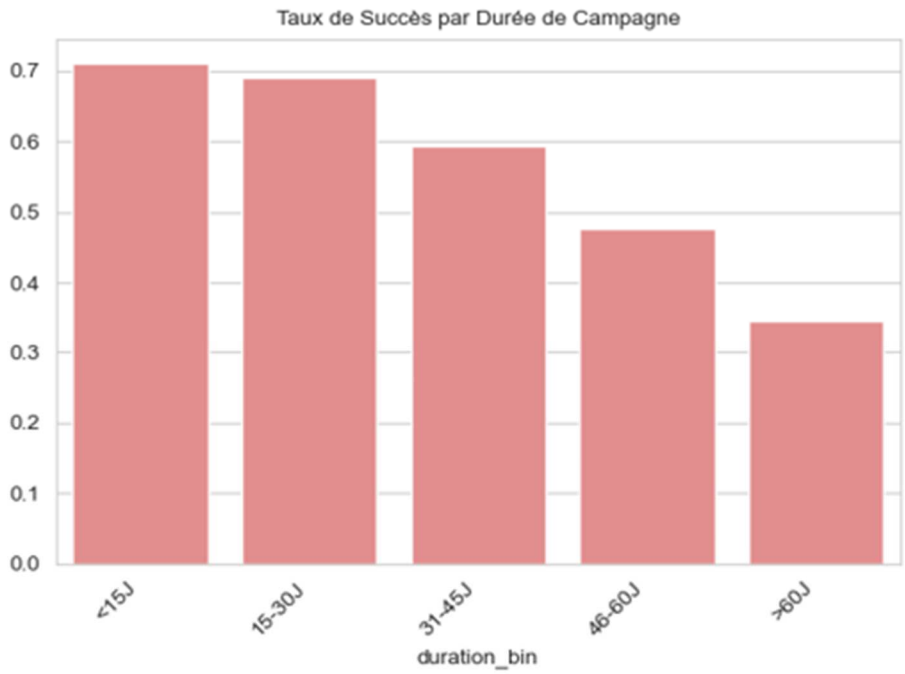


Figure 15 - Taux de succès par durée de campagne

J'ai aussi classé les jours, afin de déterminer le jour où il est opportun de lancer sa campagne et on constate que c'est le Mardi, suit de près par le Lundi et le Vendredi :

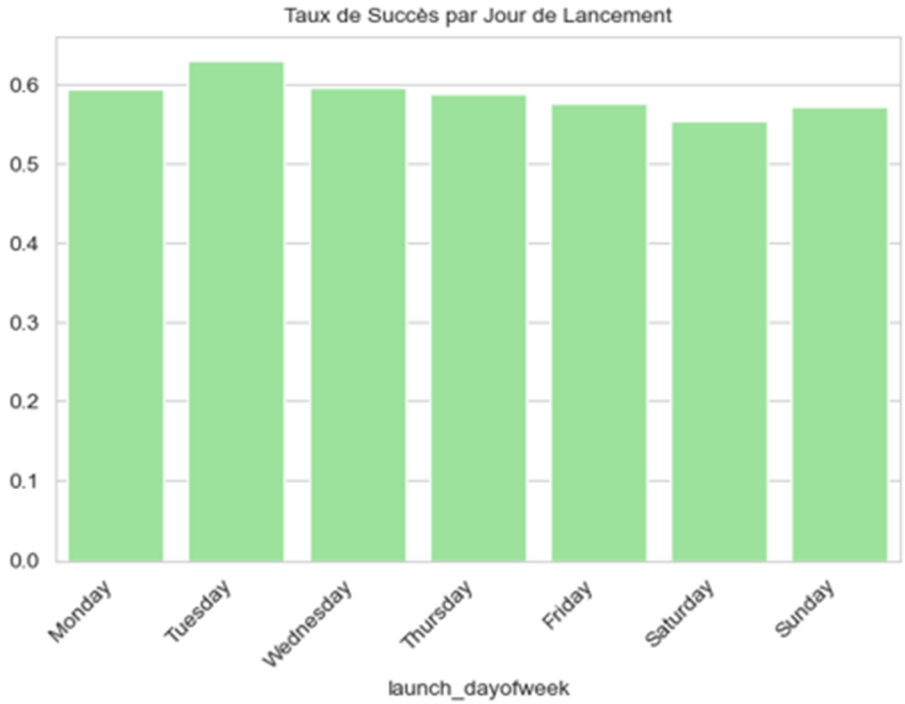


Figure 16 - détermination du jour de lancement de campagne optimal

Ce graphique indique les catégories clés de succès :

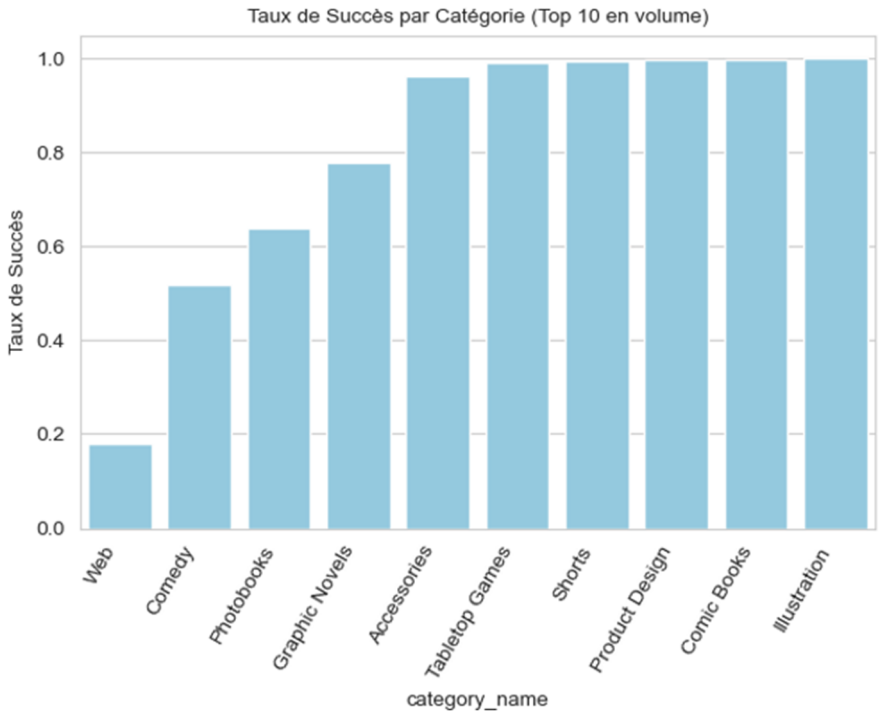


Figure 17 - Catégorie à succès

EXECUTIVE SUMMARY

En appliquant des modèles prédictifs à fort potentiel, Kickstarter pourrait transformer ses échecs en opportunités.

De plus, la majorité des 13, 5 M$ de revenus latents provient de projets en catégorie Technology, lancés en été, avec une durée > à 45 jour dans des pays non anglophones.

L'analyse prédictive nous montre également que quand la catégorie est comics, la moyenne des projets réussis à 28% de plus que toutes les autres valeurs.

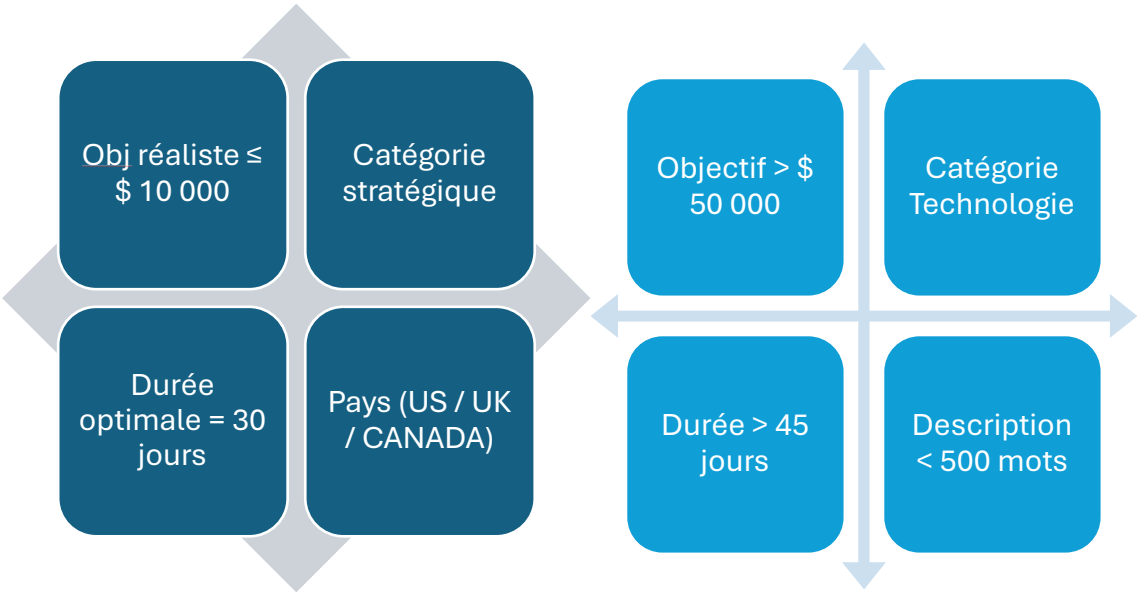
Quand l'objectif financier est au minimum de 500 $, la moyenne des projets réussis à 23% de chance de réussir.

La durée est un critère important que l'on mesure, quand la durée est < à 30 jours, on augmente de 15% ses chances de voir son projet réussir.

Avec l'analyse prédictive, on peut faire ressortir les facteurs clés de succès et d'échecs :

FACTEURS DE SUCCÈS

FACTEURS D'ECHECS



Chaque projet à succès ignorés par des biais humains ou algorithmiques maux calibrés coûtent des dizaines de millions en commissions non perçues

Les facteurs clés de succès permettent des gains potentiels pour la plateforme. En effet, Kickstarter se finance avec une commission de 5% du montant total collecté, et, prend entre 3 et 5% de frais de traitement Stripe (en général 3% + 0,20 $ par contributeur). Seule une campagne réussie est facturée. Si le projet n'atteint pas son objectif, c'est 0$ collecté et 0 de commission

On voit tout de suite, l'impact potentiel d'un meilleur accompagnement des porteurs de projets. En termes d'indicateurs on peut observer qu'on obtient 37 000 campagnes échouées mais prometteuses (proba LR > 0.5) et un profit qui peut être calculé sur la base de 5% de commission, soit un gain de **13 546 250 $.**

A partir des résultats du modèle, plusieurs recommandations opérationnelles ont été formulées afin d'augmenter significativement les chances de succès d'une campagne Kickstarter.

Comme présenté dans l'Executive Summary, le montant demandé est le facteur le plus déterminant du succès, en fixant un objectif réaliste (< 50 000$), la catégorie déclarée influence le taux de réussite (catégories populaires : Design, Games, Art), la qualité du texte est un signal fort de crédibilité (rédiger un résumé (blurb) détaillé entre 150 et 300 caractères), choisir le bon moment pour lancer sa campagne (privilégier le début de semaine), ajuster la durée de la campagne maximise la performance (éviter > 30 jours).

Ces résultats offrent une base empirique solide pour améliorer les pratiques des créateurs de projets, soutenir leur prise de décision et maximiser leurs chances d'atteindre leurs objectifs de financement.

CONCLUSION

Ce mémoire avait pour objectif d'étudier les facteurs déterminants du succès des campagnes de financement participatif sur la plateforme Kickstarter, en s'appuyant sur une démarche complète d'exploration, de préparation et de modélisation des données. A partir d'un historique de plus de 200 000 campagnes, une analyse approfondie des caractéristiques des projets a été réalisée afin d'identifier les éléments susceptibles d'influencer la probabilité de réussite.

L'exploration initiale des données a permis de mettre en évidence des dynamiques propres au financement participatif : saisonnalité des lancements, variation importante entre les catégories, dépendance forte au montant financier demandé. Cette première phase a également révélé des incohérences, valeurs manquantes et effets de structure nécessitant un nettoyage des données.

La modélisation prédictive a permis de construire un algorithme fiable capable d'estimer la probabilité de réussite d'une campagne Kickstarter avant son lancement. A travers des transformations rigoureuses (nettoyage des données, traitement des dates, encodage catégoriel, ingénierie des variables), un modèle Logistic Regression a été entrainé sur plus de 200 000 projets.

Les résultats démontrent que certains déterminants influencent fortement la performance d'une campagne :

- Le montant demandé
- La catégorie du projet
- La temporalité du lancement
- La durée de financement
- La qualité du descriptif textuel

Le modèle offre une aide décisionnelle précieuse pour les porteurs de projets, en leur fournissant une estimation réaliste de leurs chances de succès et en mettant en lumière les variables les plus influentes. L'approche permet également d'identifier les erreurs courantes, de proposer des ajustements concrets et de mieux comprendre les dynamiques globales du financement participatif.

En définitive, ce mémoire démontre que la donnée constitue un outil puissant pour comprendre et anticiper le succès des campagnes de financement participatif. L'utilisation d'approche prédictives permet des « games changers » pour renforcer la prise de décision des créateurs. Ainsi, l'analyse de données s'impose comme un vecteur essentiel d'innovation et de performances dans l'écosystème du financement participatif.

ANNEXE

Livrables :

- Rapport d'exploration (format PDF)
- Dataframe
- Fichier Excel d'analyse du dataset
- Script python :
-

Lien vers le code sur le drive :