

# RAPPORT D'ANALYSE DES DONNÉES

**Projet Kickstarter  
Cursus Data Analyst**

Réalisé par:  
M. Kevin CARO

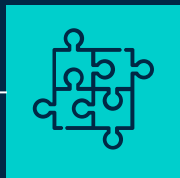


DataScientest • com

The background is a dark blue field populated with various geometric elements. There are numerous small squares in shades of teal, light blue, and orange. Some of these squares are connected to the top edge of the frame by thin, vertical white lines, creating a sense of depth or suspension. The overall aesthetic is modern and minimalist.

# MISE EN CONTEXTE

# DEFINITIONS



01

## CROWDFUNDING

Le crowdfunding est un nouveau mécanisme qui a pour objectif de collecter des apports financiers de particuliers (sous forme de dons, de prêts, d'investissement...) via une plateforme internet comme Kickstarter, ulule...



02

## KICKSTARTER

Créé en 2009, Kickstarter est une plateforme américaine de crowdfunding, qui a été précurseur dans ce domaine. Elle donne la possibilité aux internautes (porteurs de projets) de financer des projets encore au stade d'idée (en réduisant les lourdeurs associées aux modes traditionnels d'investissement, comme les prêts bancaire)

# FONCTIONNEMENT DE LA PLATEFORME

DEPOT  
PROJET

DESCRIPTIF DU PROJET

SECTEUR D'ACTIVITÉ

PAYS/ VILLE

OBJECTIF DE FINANCEMENT

DATE DE DÉBUT DE CAMPAGNE

DATE DE FIN DE CAMPAGNE



VISIBILITÉ  
DU PROJET SUR LA  
PLATEFORME



Projet en cours (Live)



PROJET

NOMBRE DE CONTRIBUTEURS

FIN  
PROCESSUS

MONTANT COLLECTÉ



STATUT DU PROJET



*Transfert des fonds*



*Remboursement des fonds*



## OPTION 1

Dossier succès  
(l'objectif est atteint)

## OPTION 2

Dossier échoué / annulé  
(l'objectif n'est pas atteint)



# NOTRE OBJECTIF

- Comprendre et analyser les tenants qui contribuent au succès ou à l'échec d'un projet
- Créer un modèle capable de prédire les statuts (successful ou failed) des projets dans le but d'aider les créateurs potentiels à concevoir une bonne campagne de financement



The background is a dark blue field populated with various geometric elements. There are numerous small squares in shades of orange, teal, and pink. Thin white vertical lines of varying lengths are scattered across the composition. The text 'JEU DE DONNEES .' is centered in a large, orange, sans-serif font.

# JEU DE DONNEES .

# LE JEU DE DONNÉES

## KAGGLE

- Source: site KAGGLE
- Périodicité : d'avril 2009, jusqu'à la fin 2020.
- Nombre de colonnes: 19 colonnes.
- Nombre de projets: 217 252 projets avant traitement (Chaque ligne correspond à un projet).
- Colonne cible: est la colonne "status" qui détermine si une campagne a fonctionné, échoué, annulé ou est en cours.
- Type informatique : int, float, catégorielle ou numérique.

Nom de la colonne	Description
Unnamed0	Indexation croissante 1,2,3
ID	Identifiant unique du projet
Name	Le nom du projet
Currency	La devise (USD, CAD...)
Launched_at	La date de lancement d'un projet
Backers_count	Le nombre de contributeurs par projet
Blurb	Texte de présentation du projet
Country	Le pays dans lequel la campagne a été lancé
Deadline	La date à laquelle la campagne se termine
Slug	La partie qui identifie chaque page de votre site
Status	Elle détermine si une campagne a fonctionné : successful ou a échoué : failed ou est encore en cours
Used_pledged	Est le montant engagé dans la campagne, c'est-à-dire la somme des dons
Sub_category	La colonne des sous catégories des projets
Main_category	Les catégories principales
Creator_id	L'identifiant du créateur
Blurb_length	La longueur du texte de présentation du projet
Goal_usd	Objectif fixé
City	La ville du projet
Duration	Durée du projet



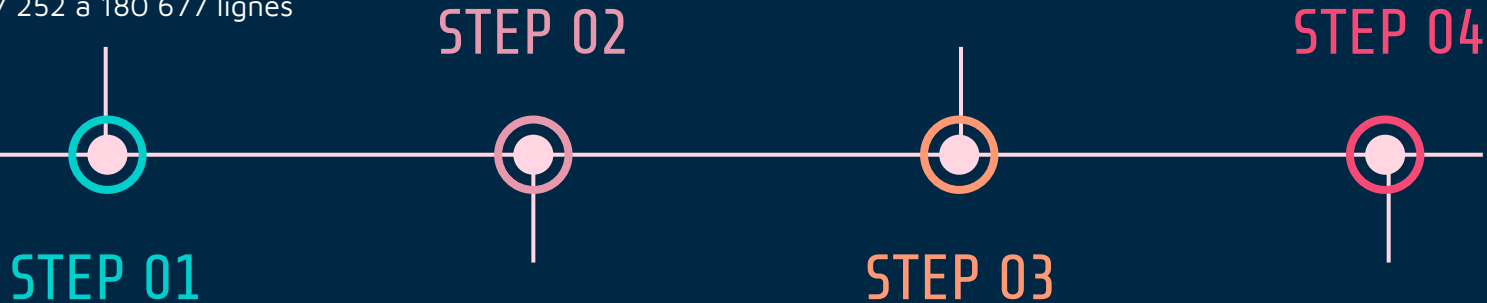
# ■ NETTOYAGE ET PRÉTRAITEMENT



# NOTRE PROCÉDURE DE NETTOYAGE

- Vérification des données manquantes, des NAN (Le Dataset Kaggle ne comportait pas de NAN)
- Suppression des doublons passant de 217 252 à 180 677 lignes

- Supprimer les variables d'indexation ;
- Supprimer certaines colonnes fortement corrélées (Nous avons conservé la colonne Country et supprimé colonne City par exemple) : Matrice de corrélation, V\_Cramer, p-valeur ;
- Splitter la colonne « launched\_at » : jour, mois, année



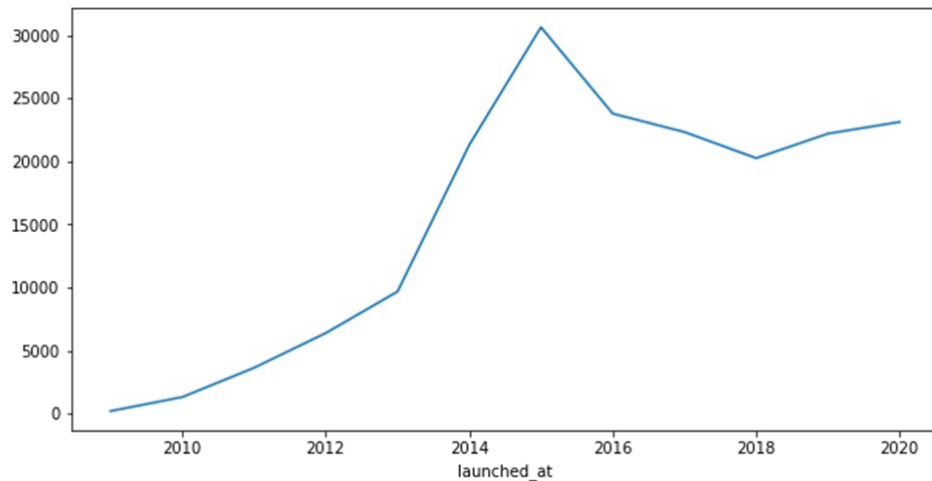
- Corriger les intitulés des colonnes "sub\_category" et "main\_category"

- La variable « status » a été transformée afin de ne garder que les projets échoués (failed) et ceux à succès (successful)
- On a mis de côté les projets "lives" (les projets en cours) pour les utiliser avec notre modèle

# DATA VIZUALISATION

# Nombre de projets par année

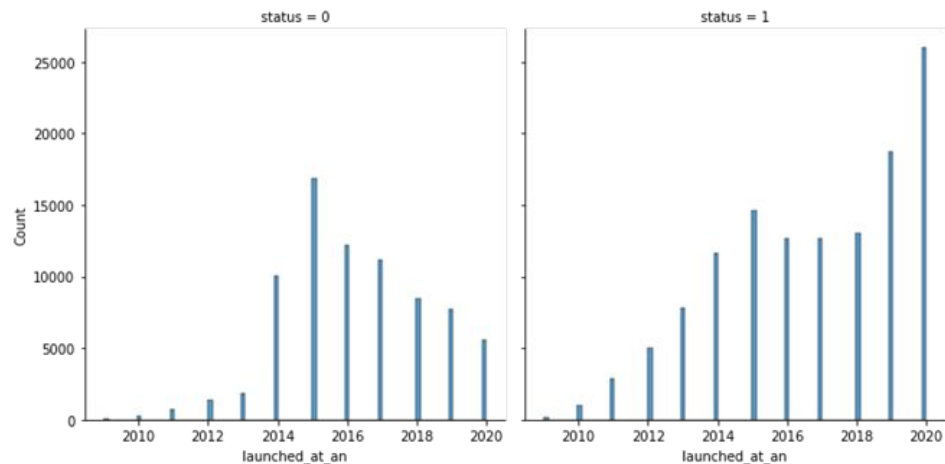
- En 2015, la plateforme a enregistré le plus de projets depuis sa création.
- Ce pic est suivi d'une faible décroissance jusqu'en 2018, où l'on observe une évolution positive.



## Le nombre de projets suivant le « Status » (successful ou failed) par année

- Pour les projets échoués (status=0), à partir de 2014, le nombre de projets échoués a connu un réel pic en 2015. L'année suivante, l'enregistrement des projets échoués n'a cessé de diminuer.
- Pour les projets réussis (status = 1), l'évolution en nombre s'est faite progressivement, et d'une façon continue, de 2009 à 2015. Le nombre de projets réussis a enregistré un pic en 2015, suivi d'une stagnation de 2016 à 2018, puis d'une croissance accrue en 2019 qui s'est confirmée durant l'année 2020.

**Tout porte à croire que la plateforme a entrepris des actions, à partir de l'année 2015 qui ont eu un bénéfice conséquent sur la réussite des projets.**



# Attractivité des projets par secteurs et objectifs de collecte

- L'objectif de collecte **Goal\_usd** est un levier important dans la réussite ou non d'une campagne de financement.
- Il existe donc un **seuil psychologique** à prendre en considération:

□ l'**objectif de collecte moyen** des projets **réussis** est de :

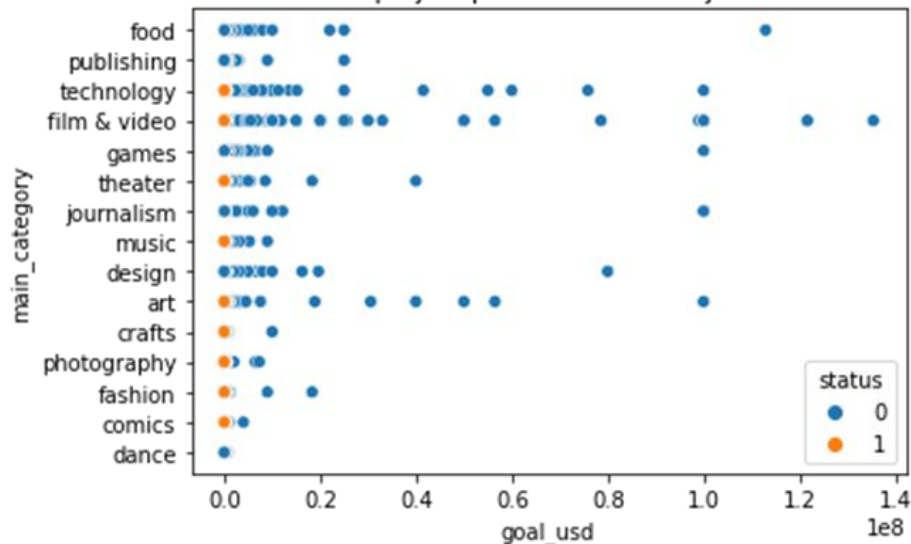
**8 857.67 USD,**

□ Contre un **objectif moyen** des projets **non réussis** est de :

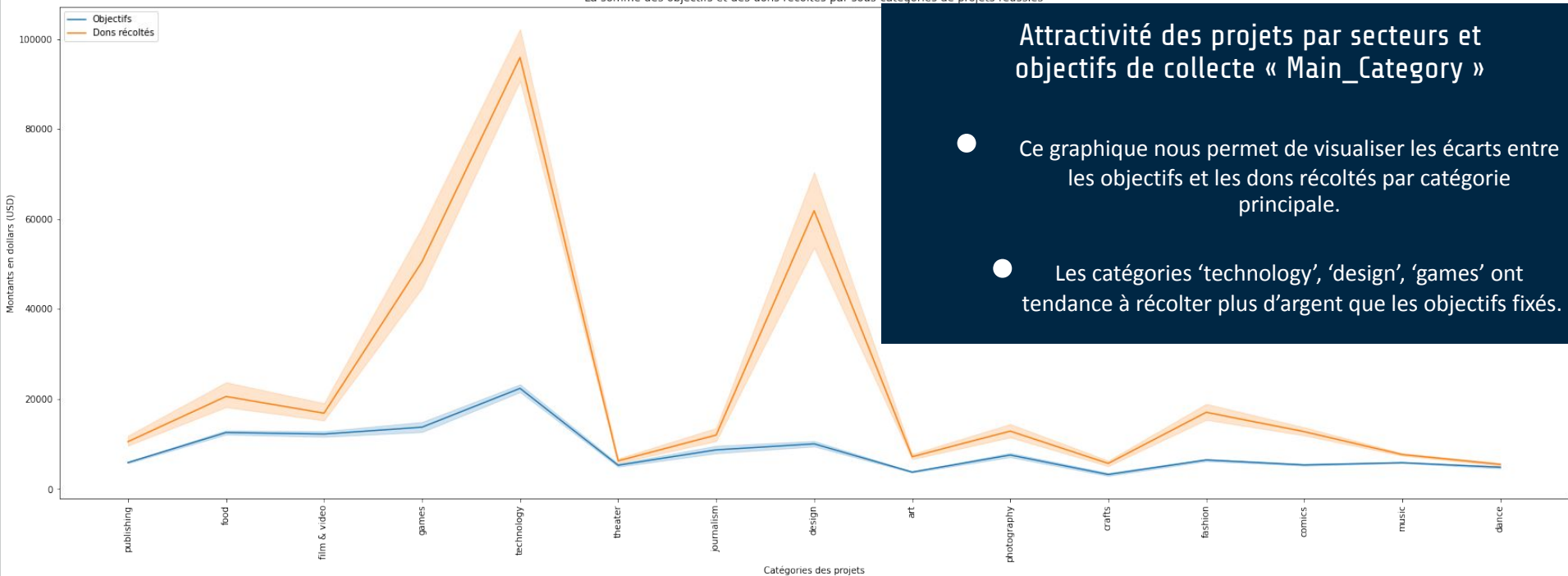
**83 216.44 USD.**



Attractivité des projets par secteurs et objectifs de collecte



La somme des objectifs et des dons récoltés par sous-catégories de projets réussies



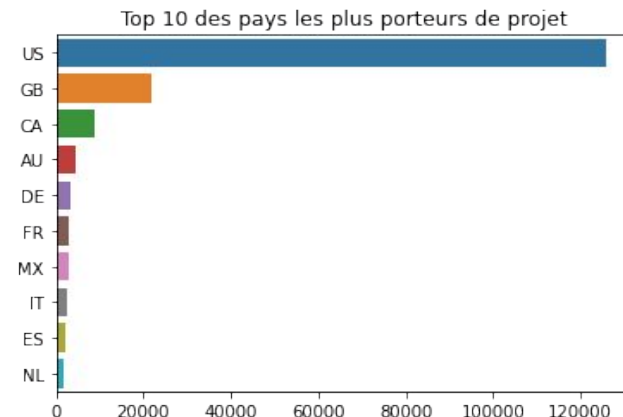
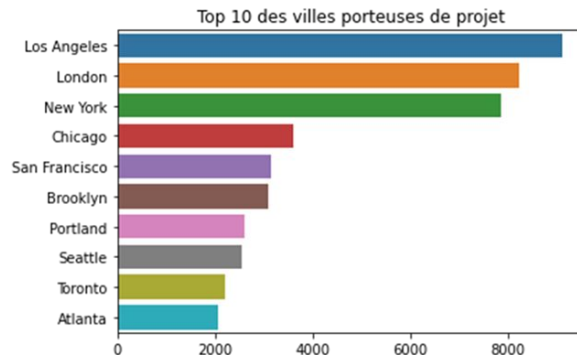
## Attractivité des projets par secteurs et objectifs de collecte « Main\_Category »

- Ce graphique nous permet de visualiser les écarts entre les objectifs et les dons récoltés par catégorie principale.
- Les catégories 'technology', 'design', 'games' ont tendance à récolter plus d'argent que les objectifs fixés.

# L'emplacement géographique

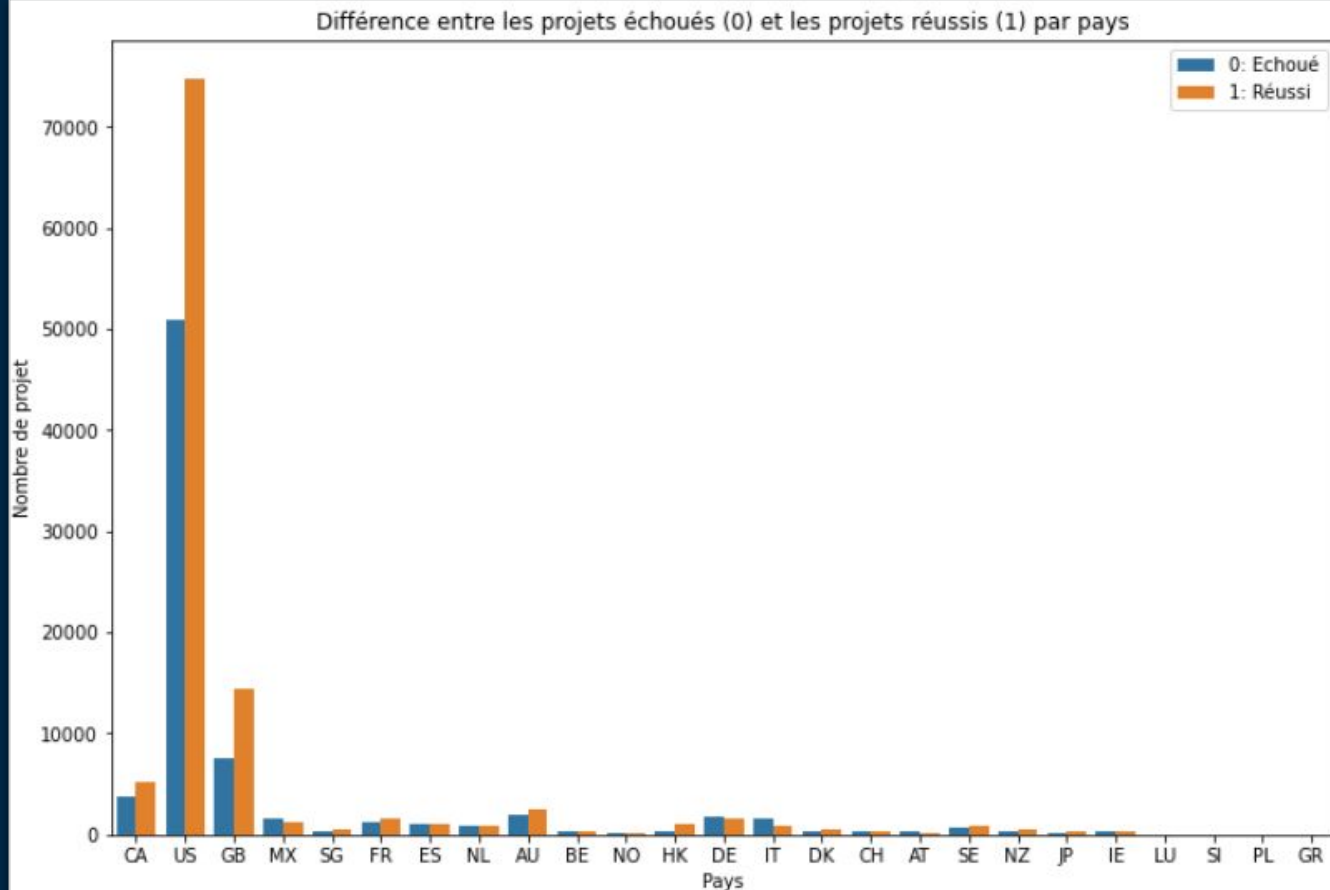
- On constate que les villes se situent majoritairement dans les pays anglosaxons (*Etats-Unis, Angleterre, Canada*). **Probablement, du fait de la généralisation de ce mode de financement, de la culture de l'entrepreneuriat dans ces pays...**
- Le graphique du classement Pays le confirme, à savoir que les pays anglosaxons sont majoritaires (*US, Grande Bretagne, Canada*).

**Les Etats-Unis sont et demeurent largement dominants en termes de projets**



## L'emplacement géographique

- On constate que les Etats-Unis sont le pays qui comptabilise le plus de projets réussis et de projets échoués.





# ITERATION I

The background is a dark blue field decorated with a pattern of small squares and thin vertical lines. The squares are in three colors: orange, teal, and light blue. Some squares are solid, while others are hollow. The vertical lines are thin and white, extending from the top or bottom of the frame. The overall aesthetic is modern and minimalist.

# PRÉPARATION DU JEU DE DONNÉES POUR LES MODÈLES

- Enlever les variables qui sont déterminées qu'après la fin de la campagne, il s'agit de: «`usd_pledged` » et «`backers_count` »

- Utilisation, *pour la mise à l'échelle des valeurs*, de la librairie `StandardScaler` de Scikit-learn.
- Affectation de la variable «`status` » comme étant la variable cible «`target` ».

STEP 01

- Séparation des variables explicatives.
- Création d'un dataframe "feats" dans lequel les variables explicatives (features) dichotomisées ont été stockées.

STEP 02

STEP 03



80% TRAIN  
20% TEST

# NOTRE JEU DE DONNÉES

## FEATS

COUNTRY  
MAIN CATEGORY  
BLURB\_LENGTH  
GOAL\_USD  
DURATION  
LAUNCHED\_AT\_MOIS  
LAUNCHED\_AT\_JOUR  
LAUNCHED\_AT\_ANNE

## TARGET

STATUS

X\_TRAIN

X\_TEST

y\_TRAIN

y\_TEST



# RESULTATS DES MODELES

## Random Forest

67,91 %



>Décèle 60% des projets échoués

>Prédit 74% des projets à succès

>Précision pour les projets à succès 73%

## Regression logistique

68 %



>Décèle 45% des projets échoués

>Prédits 84% des projets à succès.

Ce modèle a une forte tendance à classer les projets comme étant des succès

## KNN

68, 87%



>Décèle 56% des projets échoués

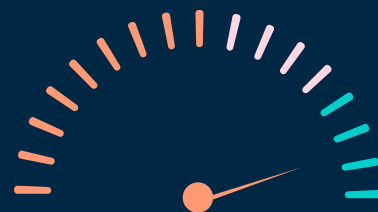
>Prédit 79% des projets réussis

>Précision pour les projets réussis est de 72%

Le KNN est le modèle le plus lent avec 2.8 min. de temps d'exécution

## XGBOOST

71,56 %



>Décèle 51 % des projets échoués

>Prédit 86% des projets à succès.

>Précision pour les projets à succès est estimée à 72%

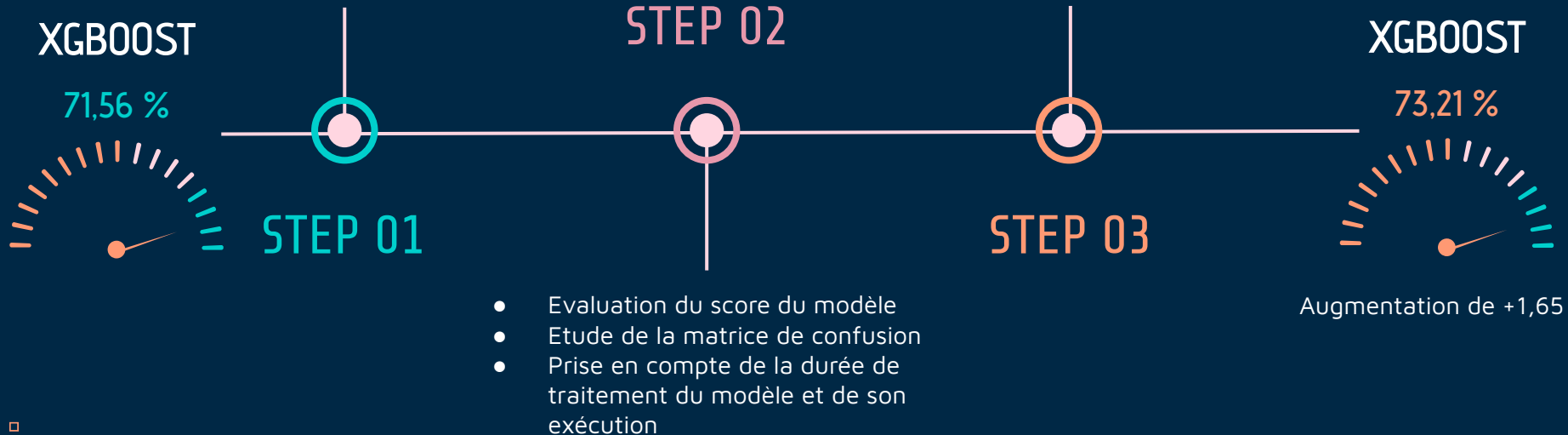
# ITERATION II

The background is a dark blue field decorated with a pattern of small squares and thin vertical lines. The squares are in three colors: orange, teal, and light blue. Some squares are solid, while others are hollow. The vertical lines are thin and white, extending from the top of the frame towards the bottom. The overall aesthetic is modern and minimalist.

# AMELIORATION DU MODELE XGBOOST

Changer les hyper paramètres par défaut

- D'après les résultats précédents, le XGBOOST a obtenu le meilleur score.
- Nous avons réussi à améliorer le score obtenu grâce aux paramètres (max\_depth=10)



# PREDICTIONS DES PROJETS LIVE

Dataset des projets en cours (Live)

	id	Status_pred
0	490067152	1
1	754191545	1
2	56023157	1
3	232907989	1
4	576278674	1

Dataset Webrobots avec les derniers statuts des projets

	id	state
0	1166044523	failed
1	986370978	successful
2	2059825834	failed
3	1462846972	successful
4	1635933643	successful

Merge

- Nous avons conservé les projets en cours (live) pour évaluer notre modèle, grâce aux données de Dataframes plus récents récupérés sur la plateforme Webrobot
- Le but étant de comparer les prédictions

(2459, 3)  
dimension est : (2459, 3)

	id	Status_pred	state
0	490067152	1	0
1	754191545	1	1
2	232907989	1	1
3	232907989	1	1
5	2044095310	1	0

matrice de confusion pred vs résultat réel

Status_pred	0	1
state		
0	244	597
1	90	1528

Score du Model sur des données réelles  
0.725091500610004

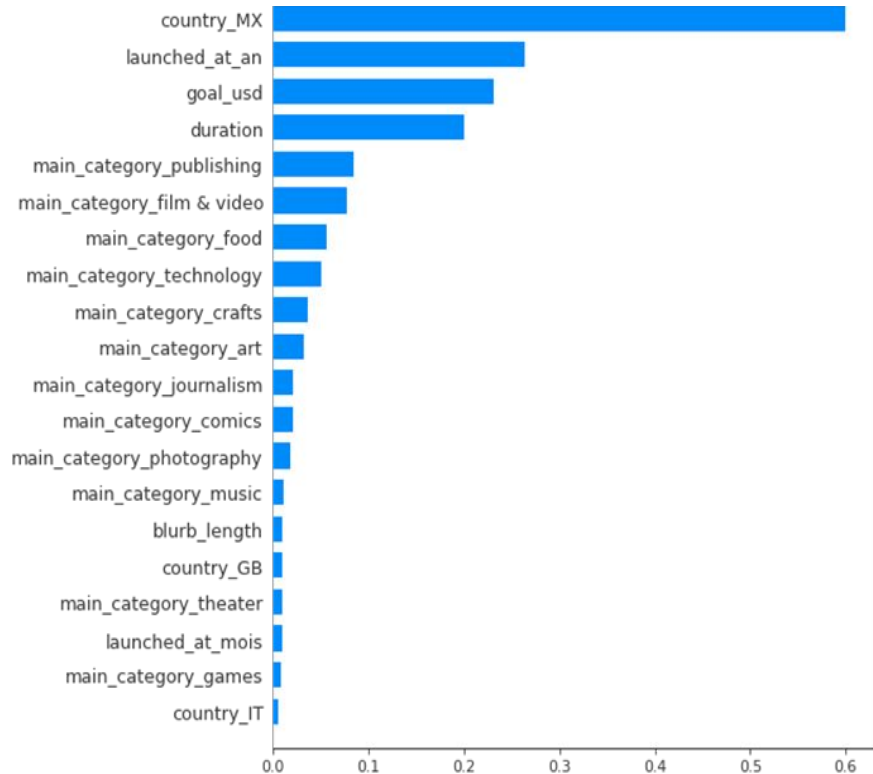
Matrice diagonale  
Score 72,25% vs 73,21%

# INTERPRETABILITE



# INTERPRÉTABILITÉ DU MODÈLE XGBOOST

- Pour analyser notre modèle nous avons utilisé la librairie Shap
- D'après le graphique obtenu, nous remarquons que notre modèle accorde une grande importance aux variables goal\_usd, launched\_at\_annee, duration et country\_MX (Mexique)





- Les valeurs des caractéristiques qui augmentent les prédictions sont en rouge et leur taille visuelle montre leur ampleur dans notre modèle de prédictions.
  - launched\_at\_anne,
  - goal\_usd,
  - main\_category\_publishing...
- Les valeurs qui ont un effet significatif sur la diminution de la prédiction.
  - country\_MX (Mexique)
  - duration



# CONCLUSION

The background is a dark blue gradient. It features several thin, light blue vertical lines of varying lengths. Scattered throughout are small squares in three colors: light blue, pink, and orange. Some squares are solid, while others are outlined. The word 'CONCLUSION' is centered in a large, bold, orange, sans-serif font.

# Ces techniques nous ont permis de:


- Mettre en place une méthodologie visant à comprendre et expliquer notre problématique, à savoir la compréhension au travers une analyse approfondie, **des tenants qui influencent et contribuent au succès, ou à l'échec d'un projet sur la plateforme Kickstarter.**
- Notre **but étant d'aider les créateurs, porteurs de projets potentiels à concevoir la meilleure campagne de financement.**
- Grâce à nos études nous avons pu faire ressortir des tendances qui vont permettre aux futurs porteurs de projets d'optimiser, d'augmenter la possibilité de réussite d'un projet, grâce à notre modèle pleinement fonctionnel.



# Pistes d'amélioration:

- Faire des regroupements par continent afin d'essayer d'améliorer les scores des modèles.
- Tester le comportement des modèles sans la variable Duration, qui a un effet significatif sur la diminution de la prédiction.



The background is a dark blue gradient. It is decorated with several vertical white lines of varying lengths. Scattered throughout are small squares in three colors: light blue, light orange, and light pink. Some squares are solid, while others are outlined.

Place aux questions après une  
courte présentation de notre  
application Streamlit

MERCI