Moov AI Test technique

Fatima-Zahra Banani

Rapport

Table des figures

Table des matières

CHAPITRE 1

Problèmes de qualité des données

La table de données consiste en :
— 323750 entrées.
— 17 colonnes.
Il existe 13 colonnes nommées (*[ID ,name ,category ,main_ category ,currency ,deadline, goal, launched, pledged, state, backers, country, usd_ pledged]*) et 4 colonnes sans noms.

1.1 Colonnes pas informatives

Dans cette étape, j'ai commencé par supprimer les colonnes avec des valeurs manquantes représentant plus que 85%.

Les colonnes en question étaient les colonnes non nommées de la table.

1.2 Inconsistence des données

Le but de cette étape est d'éviter d'avoir des valeurs incohérentes au sein d'une même colonne.

Pour les colonnes supposées être d'un certain type ("numérique","date"), les valeurs n'appartenant pas au type en question étaient remplacées par des "nan".

Pour les valeurs de colonnes représentant une option parmi une liste fixe d'options offertes par le site "kickstarter", comme "currency", "category" par exemple,

les valeurs n'appartenant pas à la liste possible des valeurs de la colonne étaient remplacées par des "nan".

1.3 Valeurs manquantes

Après le traitement des données précédemment expliqué, J'ai remarqué qu'il y avait des lignes avec majoritairement des valeurs manquantes, puisque ces lignes sont devenues pas informatives, elle étaient supprimées.

Pour le "state" du projet, il existait des projets avec comme état : "canceled","live","suspended". Ces projets étaient supprimés de la table de données.

Pour les projets avec état "canceled", l'annulation du projet est un choix du promoteur du projet pour une raison inconnue. Les projets "suspended", sont à cause du non respect d'une loi de la plateforme "Kickstarter".

Par contre, pour les projets "live", ce sont des projets qui n'ont pas encore atteint la date limite, une modification possible est de garder les projets qui ont déjà dépassé leur but puisque pour ces projets, l'état final est normalement connu.

1.4 Relations entre colonnes

Dans cette étape, la cohérence des valeurs appartenant au même projet, est remise en question.

Puisque la plateforme "Kickstarter" oblige les promoteurs de projets de ne pas dépasser 60 jours pour la promotion, j'ai créé une nouvelle colonne "campaign_period" ("deadline"-"launched"), cette colonne représente la période en jours. Les projets avec une période qui ne figurent pas dans l'intervale [1,60], étaient supprimés.

Aussi chaque catégorie ("category") doit appartenir à une catégorie principale ("main_category") correspondante, les valeurs de catégorie ne respectant pas cette propriété étaient remplacées par la catégorie principale.

1.5 Déduction des valeurs manquantes

Pour la colonne "usd_pledged", ses valeurs manquantes étaient déduites des colonnes "currency","pledged" et "deadline", en faisant une conversion de devise en "usd" en question dans la journée "deadline".

Pour la colonne "state", la différence entre les colonnes "goal" et "pledged" permettait de déduire l'état du projet.

1.6 Réponses

Pour résoudre les problèmes de qualité des données, les étapes précédentes étaient suivies.

Après le traitement des données, 13,97% des entrées étaient supprimées.

CHAPITRE 2

Identification des "insights"

Dans ce chapitre, des outils de visualisations étaient utilisées pour indentifier des "insights".

2.1 Analyse univariée

FIGURE 2.1 – Distribution de la colonne "campaign_period"

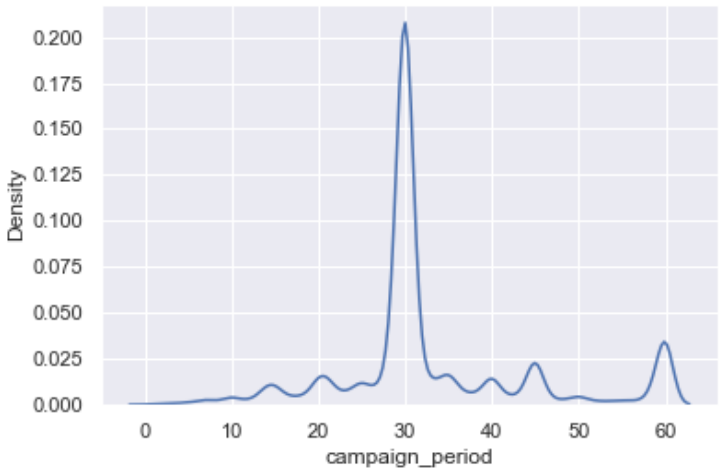


FIGURE 2.2 – Histogramme du nombre de projets par pays

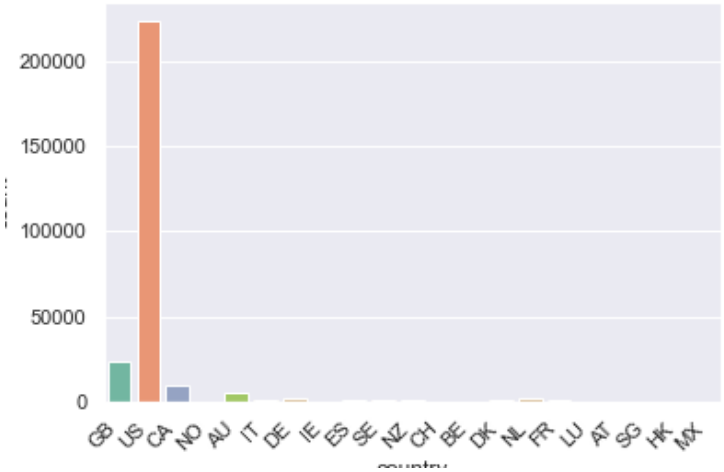
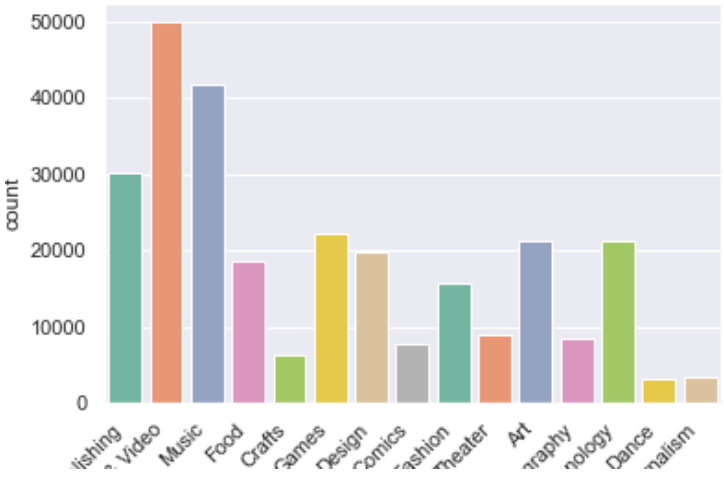


FIGURE 2.3 – Histogramme du nombre de projets par catégorie principale



— La majorité des promoteurs de projets choisissent comme période une période d'environ 30 jours.
— Le pays le plus présent dans la table de données est "Etats Unis".
— Il existe une variance de nombre de projets par catégorie principale.

2.2 Analyse statistique

Les figures suivantes représentent des visualisations de la moyenne et de l'écart-type de quelque colonnes dépendamment de l'état du projet.

FIGURE 2.4 – Diagramme en boîte de la période de campagnes

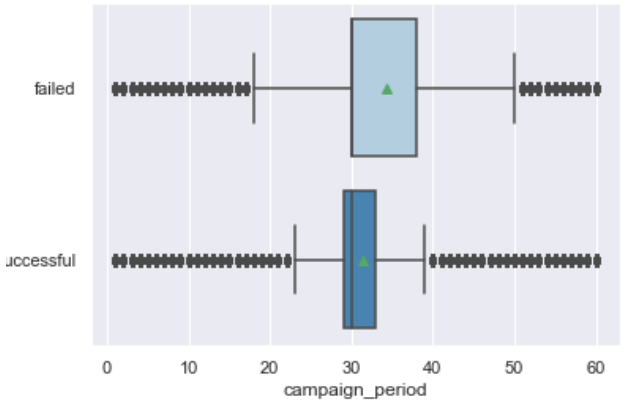


FIGURE 2.5 – Diagramme en boîte des "backers"

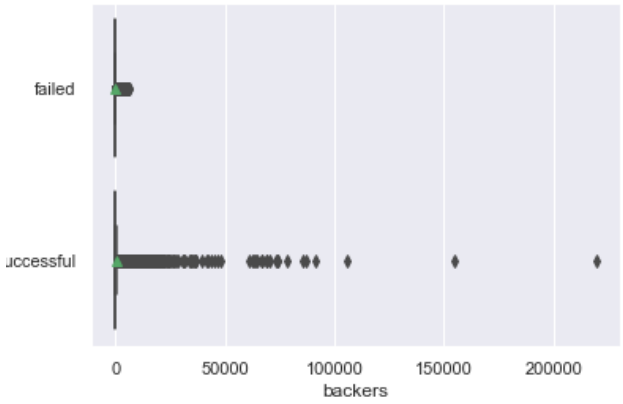
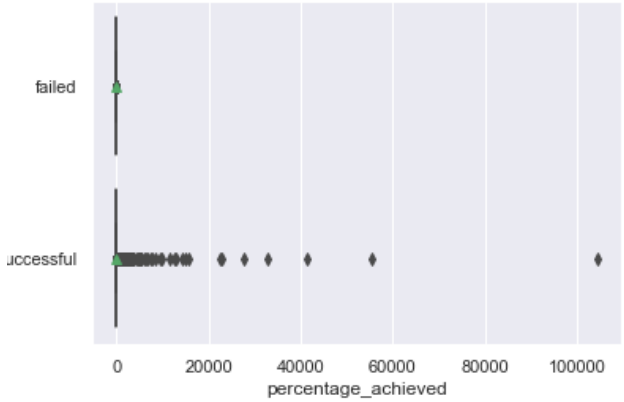


FIGURE 2.6 – Diagramme en boîte du pourcentage du but atteint ("pledged"/"goal")



— Les projets qui ont atteint le succès ont une période de campagne plus concentrée dans les alentours de 30 jours. Il y a plus de variance de périodes pour les projets qui ont échoués.
— Pour le pourcentage du but atteint et le nombre de backers, on remarque beaucoup plus de variance chez les projets qui ont réussi.

2.3 Analyse bivariée

Les figures suivantes représentent des visualisations de pourcentage d'échec et de succès des projets dépendamment de quelque colonnes.

FIGURE 2.7 – Pourcentage de succès/échec du projet dépendamment de la période de la campagne

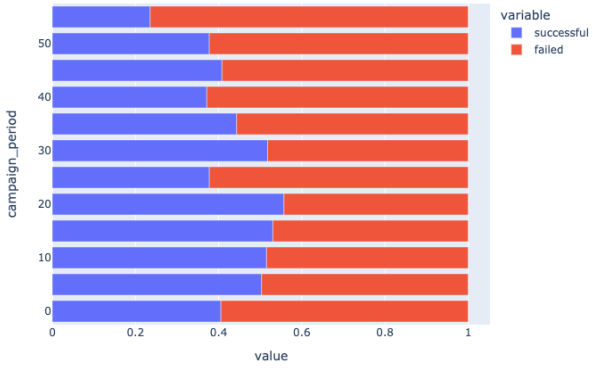


FIGURE 2.8 – Pourcentage de succès/échec du projet dépendamment du pays

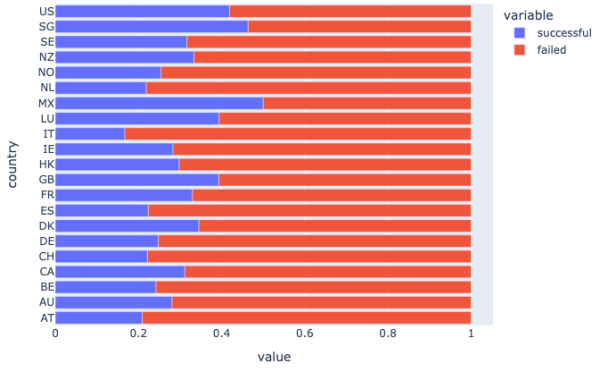
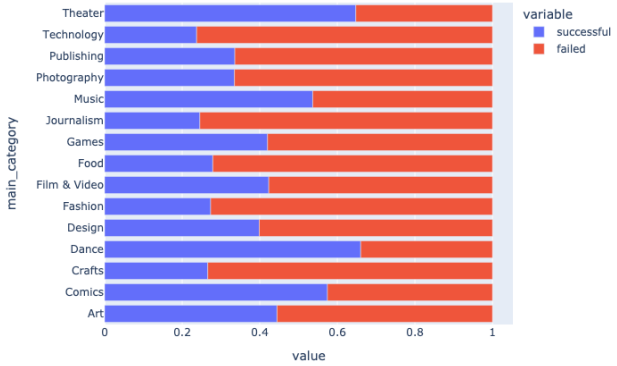


FIGURE 2.9 – Pourcentage de succès/échec du projet dépendamment de la catégorie principale



— On remarque une variance du pourcentage de succès/échec du projet dépendamment des trois colonnes.
— Pour les pays, puisque les pays autre que les états-unis sont faiblement représentés, je pense que ajouter cette caractéristique pour l'entraînement du modèle, va introduire un biais.
— Un colonne qui sera très probablement intéressant d'avoir et de visualiser et "goal_usd", puisque cette caractéristique aidera à estimer la grandeur du projet et ainsi le "goal" de la campagne.

2.4 Réponses

— Pour s'assurer que les "insights" sont statistiquement significatifs, on doit utiliser un test statistique. L'hypothèses "null" est que les "insights" identifiés sont un résultat de chance (Ils ne reflètent pas les vrais relations entre les caractéristiques de données mais dépendent uniquement de l'échantillon de données étudié).
— D'après les visualisations précédentes, je pense que les caractéristiques qui peuvent être utilisées pour entraîner le modèle ML sont : "campaign_period", "main_category", "goal", "currency". "goal_usd" sera une meilleure alternative des deux caractéristiques "goal" et "currency".

CHAPITRE 3

Solution ML

3.1 Détection des valeurs aberrantes

Selon les visualisations de la section précédente, il y avait une grande variance des deux colonnes "backers" et "usd_pledged", ce qui signifiait que les valeurs sont aberrantes.

3.2 Balance des données

Puisque la variance était plus remarquable pour les projects avec l'état "successful", la majorité des entrées de données supprimées, appartenaient à la classe "successful" ce qui a créé un déséquilibre de classes.

3.3 Réponses

— Le choix de l'algorithme de classification était motivé du fait que les arbres de décision sont facilement interprétables.
— Pour les conditions de décision dans l'arbre, chaque nœud avait un seuil pour les caractéristiques, donc je pense qu'un meilleur encodage du caractéristique "catégorie principale" serait un encodage de fréquence du caractéristique dans les données et non un encodage de "labels". Je pense cette encodage aura une meilleure interprétabilité et contribuera à un meilleur classificateur.

— Le caractéristique "goal" a le poids le plus élevé sur la classification des projets, suivi par "campaign_period".
— En utilisant un modèle capable de classifier si un projet peut atteindre le succès ou pas, les promoteurs de projets peuvent utiliser la prédiction du modèle, pour avoir une estimation de risque de leur projet. Donc si le projet va probablement échouer selon le modèle, ils peuvent travailler plus sur le "pre-campaign" marketing pour s'assurer d'avoir préalablement un public qui peut soutenir le projet, ou remettre en question l'argent visée par la campagne ("goal") , en diminuant l'échelle du projet.

CHAPITRE 4

Risque après déploiement :

4.1 Réponses

— Le risque à considérer après le déploiement du modèle est la dérive des données "data drift". Si les données changent avec le temps par rapport aux données utilisées pour l'entraînement du modèle, par exemple une catégorie devient plus populaire ou en demande qu'une autre, la prédiction de l'état d'un projet du modèle sera mauvaise. On peut ajouter un module d'évaluation des dérives qui réentraîne le modèle en utilisant les nouveaux données reçues si le module détecte une dérive.

— Une des méthodes pour évaluer si une dérive est fort probable est d'utilser des tests statistiques qui comparent les distributions estimées des données. En utilisant "kl-divergence" par exemple, si la valeur de la divergence est haute, on peut conclure qu'une dérive des données est fort possible, et on réentraîne le modèle dans ce cas.