Question #1 : Comme c'est souvent le cas dans les projets, le jeu de données peut nécessiter quelques manipulations pour être utilisable par une approche ML. Si tu rencontres des problèmes de qualité des données durant ta manipulation des données de Kickstarter, comment les as-tu résolus?
Pour bien répondre a cette question, je vais diviser ma réponse en deux cadres. Le premier est un cadre général et le deuxième est un cadre plus spécifique au problème traité dans cet exercice.

Cadre général:

Avant de construire un modèle ou faire un analyse de donnée, il est important d'examiner et de pré-traiter l'ensemble de données pour s'assurer qu'il est adapté à l'entraînement et que l'on peut obtenir des résultats significatifs du modèle. Ainsi, lorsque que je fais face a un problème, je commence par regarder deux facteurs importants qui permettent de juger la qualité ou l'aspect de l'ensemble de données avec lesquelles on va travailler. Premièrement, je regarde l'aspect général des données, la grandeur de l'ensemble, le nombre de caractéristiques, le type de données (texte, numérique, catégorique, temporelle , etc..) . Ceci permet de se familiariser avec l'ensemble de donnée qu'on travail et permettra de déterminer le type d'inspection global à réaliser (résumé de statistiques / compréhension des valeur catégoriques). Grace à cette familiarisation et compréhension, je peux venir vérifier les "impuretés" qui peuvent exister dans les données (informations fautives ou qui n'on aucun sens avec la catégorie).
Par la suite, je regarde le manque de données qu'il y existe dans l'ensemble de données. Cette analyse permet non seulement de connaitre les caractéristiques qui ont le plus de manque d'information mais aussi cela nous permets d'avoir de l'information sur le contenu des données. À titre d'exemple, si on travail avec une base de donnée financière et il existe une caractéristique de date de fermeture de compte où il a un grand manque de données, ceci nous montre non seulement que cette caractéristique devra subir des manipulations pour combler a son manque d'information mais aussi, le manque de cette information nous fais savoir que la majorité des comptes sont toujours ouvert et que la population de l'ensemble de données sera majoritairement des comptes toujours ouverts. Mis a part les manipulations possibles, il y a aussi des cas ou le manque d'information est tellement considérable qu'il vaut mieux se défaire de ces caractéristiques pour pouvoir travailler avec un ensemble de données non biaisé et de qualité.

Cadre spécifique à l'exercice:

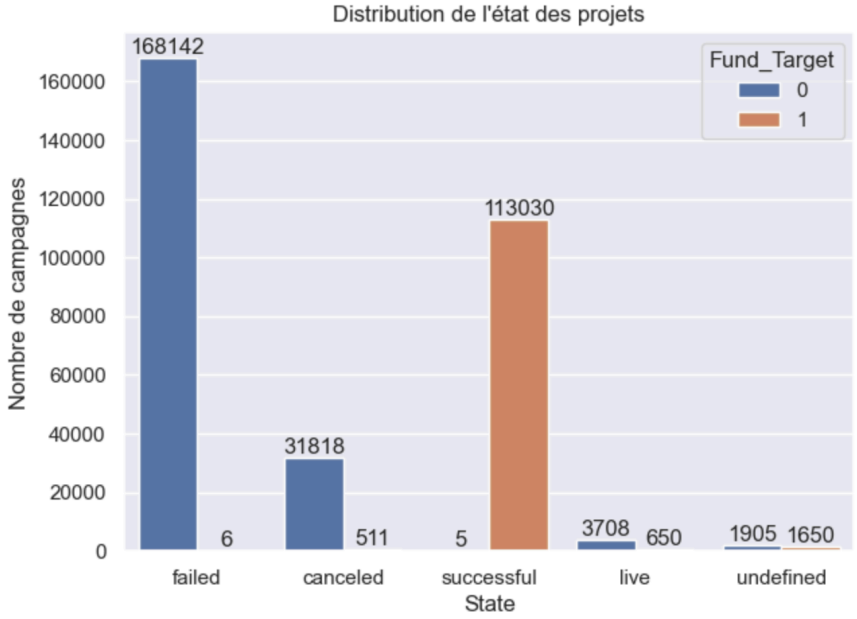
Pour les données obtenus de Kickstarter, en employant la méthodologie mentionnée plus haut, j'ai réussi à trouver plusieurs problème de qualité. Tout d'abord en regardant l'ensemble de données je me suis aperçu qu'il existait plusieurs catégories non définies tel que (Unnamed :13) une catégorie non descriptive n'est usuellement pas utile.
En analysant le manque de donnée de l'ensemble, je remarque que non seulement il s'agit de caractéristique non informative mais aussi elles ne contiennent aucune (our presque) information. Il devient donc évident que des catégories doivent êtres enlevée.
Aussi, un manque d'information été tres présent dans la colonne "usd pledged" ceci a été résolu en utilisant une librairie qui permet de faire la conversion de monnaie a une date précise à l'aide des colonnes "Currency" et "pledged". La date utilisée a été celle de lancement. Pour le reste des caractéristiques, il y avait un grand nombre de données "parasites" celles si on été traité en regroupant toutes les valeurs incohérentes en forme de "other". Le nettoyage pour chacune des caractéristique est différente et montré et commenté dans le code envoyé.

Question #2 : Identifie des « insights » qui, selon toi, peuvent contribuer à comprendre le succès ou non des campagnes.

- o Limite-toi aux trois observations les plus pertinentes selon-toi (appuies ces observations avec un visuel).
- o Basé sur les insights mentionnés plus haut, y a-t-il un risque que des « confounding variables » (i.e facteur de confusion) viennent affecter l'interprétation de tes observations ?
- o Est-ce que les «insights» trouvés peuvent être transformés en «features» qui faciliteront l'apprentissage du modèle ML ?

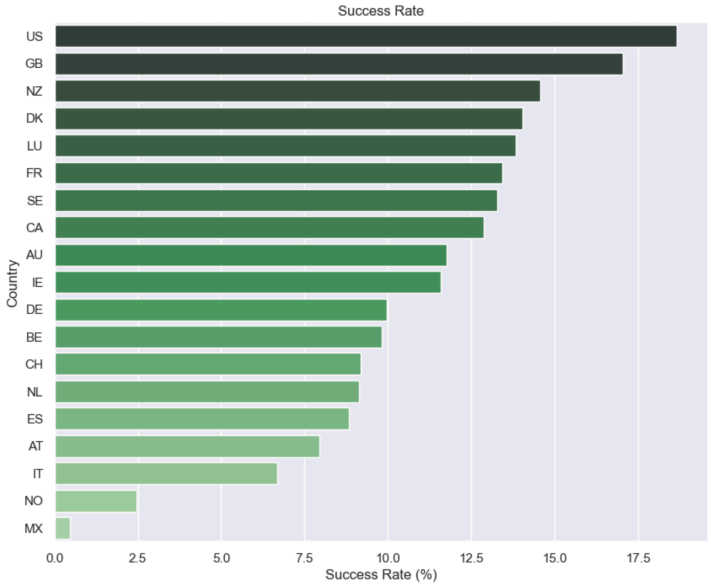
Trois insight pertinents pour le succès des campagnes sont les suivants:

I) Une condition pour qu'une campagne soit un succès, il faut que l'objectif monétaire (goal) soit atteint ou dépassé.

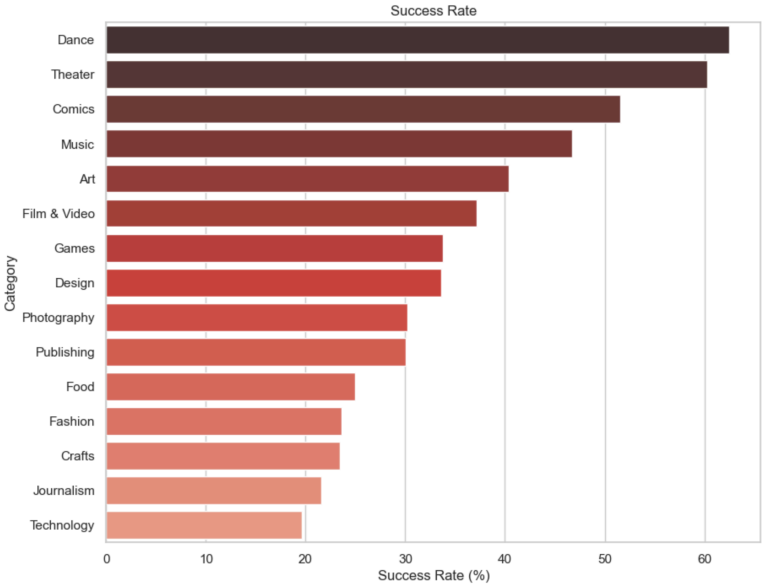


*Fund_Target: 0 si l'objectif n'a pas été atteint. 1 si il a été atteint et/ou dépassé

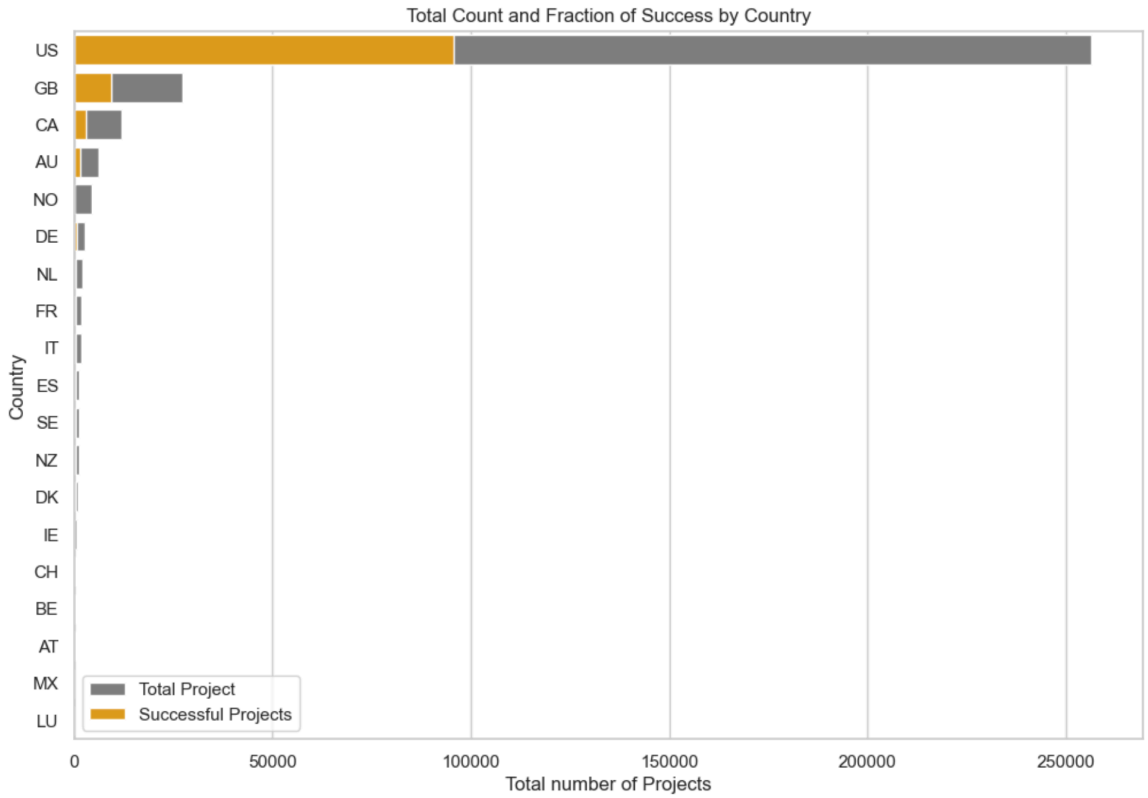
II) Une campagne lancé au États-Unis (US) ou Grande Bretagne (GB) auras plus de chance de s'avérer en un succès que si elle est lancé dans le reste des pays présents dans l'ensemble des données.



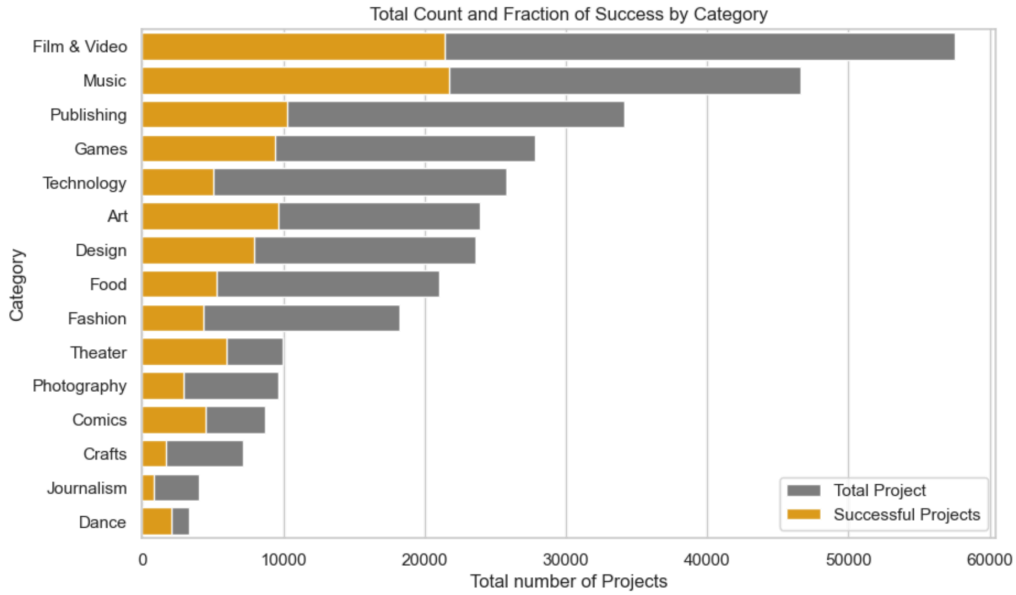
III) Une campagne dans les domaine de Dance, Théâtre ou Comics aura plus de 50% de chance d'être une campagne réussie.



En effet, il existe un risque de faire une conclusion où des confounding variables viennent affecter ou complètement changer l'interpretation des données il faut ainsi faire attention aux informations tirées des graphes de distribution. Une source d'erreur ici serait d'assumer que les campagnes au états unis sont généralement réussies et qu'il serait préférable de lancer una campagnes aux US sans préalablement verifier et comparer le nombre total de campagnes lancées et le nombre total de campagnes réussies (voir graphique plus bas). En effet, le plus de campagnes sont lancées le plus qu'il peut y avoir de campagnes réussies. C'est pourquoi nous nous sommes intéresses au taux de réussite et non pas seulement ou nombre total de réussite (il s'avère qu'en effet les US sont un bon pays pour la réussite d'une campagne)



Un autre exemple est celui des catégories, en regardant le graphique plus bas, on pourrait assumer qu'une campagne lancée en Technologie aura plus de succès qu'une lancée en Danse puisqu'il existe une plus grande quantité de campagnes réussies en Tech qu'en Danse. Cependant, en regardant le taux de réussite montré plus haut, on s'aperçoit que les campagnes avec le taux de réussite le plus bas sont celles en Tech.



En effet, les insights sont d'importants "features" a utilisé pour simplifier et ainsi aider un modèle d'apprentissage. Un exemple concret, est la création du feature "Fund_Target" qui semble être colinéaire avec le succès ou l'échec d'une campagne. En utilisant le feature Fund_Target plutôt que "State" on peut faciliter l'apprentissage du modèle ML puisqu'il devient plus facile de prédire si un campagne sera réussie ou pas en sachant si l'objectif monétaire a été atteint (ou pas). Ceci est d'autant plus vrai qu'il existe de features a tenir en compte. Il faut aussi savoir qu'en faisant ainsi il existe aussi une perte d'information liée au fait que l'on simplifie l'information en utilisant cet "insight". (ex: nous ne tenons plus compte des campagnes annulées, en cours et indéfinies).

Question#3 : Au niveau de la solution ML:

o En tenant compte des parties prenantes visées par ta solution, comment interprètes-tu les résultats produits par ta solution ML ? Comment cette solution ajoute-t-elle de la valeur pour ces parties prenantes ?

o Selon toi, comment envisage-tu que les parties prenantes vont utiliser ta solution pour tenter de comprendre comment lancer des campagnes à haut taux de succès?

Avant tout, il faut comprendre quel est l'intérêt de la partie prenante lors de l'utilisation de cette solution. Il peut s'agit de promouvoir une image de succès de campagne ou bien de diminuer les risques (coûts) lors de lancement de campagne afin de garantir une campagne réussie. Une fois cette étape a été réalisée, on peut s'intéresser aux résultats produits par cette solution. Dans mon cas, j'ai essayer de diminuer au plus possible les faux négatifs (opportunité ratée) ainsi que les faux positifs (utilisateurs mécontents). Ces métriques tirées de la matrice de confusion permettent d'évaluer les résultats afin de pouvoir se prononcer sur la qualité du modèle.

Pour ce qui est de les résultats produits par ma solution, plutôt que prendre une prédiction comme un assurance de réussite, ils permettent aux parties prenantes de prédire si une campagne s'avérera en échec. Ainsi, si ma

solution développée prédis que la campagne sera un échec, elle permettra aux parties prenantes d'ajuster les paramètres nécessaires afin d'augmenter fortement les chances à cette même campagne d'avoir du succès et c'est là où sa valeur ajouté réside. Il s'agit plutôt d'une solution (ou un outil) d'amélioration de lancement de campagne et non pas une solution restrictive de campagnes.

Question#4 : Imaginons que ta solution est déployée et roule maintenant en production. Tu remarques que la performance de ton modèle se dégrade progressivement depuis les derniers mois. De plus, tu identifies également certaines variables dont les valeurs semblent avoir évoluées durant la même période. Selon toi, quel serait une raison qui explique cette situation et comment la ressouderais-tu ?

Il existe plusieurs raisons qui peuvent expliquer cette situation/comportement du modèle déployé, il peut même s'agir de la combinaison de ces raisons et ceci met en évidence l'importance d'avoir un suivi continu du modèle même après déploiement.

Lorsqu'on constate que certaines variables ont évoluées durant la période ceci suggère que les propriétés (d'un point de vue statistique) des variables ont changé. Cependant, la relation existante entre ces variables et la variable que l'on veut prédire reste la même que lorsqu'on le modèle a été entrainé et ne suit pas ces changements. Cette relation n'est donc plus valide due aux changements de la distribution générale des données. Ceci peut être régler en effectuant un suivi du modèle en évaluant sa performance et si jugé nécessaire l'entrainer avec des changement qui sont en harmonie avec la dynamique changeante des données avec laquelle le modèle travaille.

Un autre concept inhérent aux modèles de prédiction en ML est la détérioration du modèle issu du fait que les patterns appris par le modèle deviennent de moins en moins cohérent avec les changements des donnée, témoignant aussi d'un manque d'adaptabilité du modèle. Si le modèle ne peut pas s'ajuster aux nouveaux patterns sa performance sera rétrograde.

Finalement, le modèle aura aussi un impact sur l'environnement dans lequel il fonctionne, ceci entraine des boucles de rétroaction qui viennent influencer directement les données et le comportement où il opère, ceci est un aspect que le modèle ne peut pas prendre pas en compte.

En général, pour remédier à la détérioration des performances d'un modèle, il est essentiel de surveiller en permanence ces performances, de comprendre les causes de la dégradation, de mettre à jour le modèle, de le réapprendre à partir de données récentes et de mettre en œuvre des stratégies qui permettent au modèle de s'adapter à l'évolution des conditions. Une maintenance régulière et une approche proactive de la gestion du modèle sont essentielles pour atténuer la détérioration des performances qui sont incontournables.