Réponses aux questions du test technique de Newforma

Mise en situation

Certains promoteurs de projet tentent de comprendre comment augmenter le taux de réussite de leurs futures campagnes. Ils ont à leur disposition des données historiques de campagnes annoncées sur la plateforme Kickstarter.

Objectif:

Développe en Python une approche ML (supervisée et/ou non supervisée) pour aider les promoteurs de projet à lancer des campagnes à fort potentiel de réussite.

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?

Réponse Q1:

Pendant la manipulation de la base de données de Kickstarter, j'ai rencontré plusieurs problèmes, notamment :
1- Les trois dernières colonnes sont vides et doivent être supprimées.
2- Les valeurs dans les colonnes sont de type objet. J'ai analysé chaque colonne individuellement pour déterminer le type de données approprié à leur assigner. Par exemple, j'ai identifié deux colonnes de type datetime ('deadline' et 'launched'), une colonne de type entier ('backers'), etc.
3- Traitement des variables catégoriques. Il existe deux types de variables catégoriques :

- Les variables catégoriques ordinales (par exemple : 'state') qui présentent un ordre ou une hiérarchie spécifiques entre leurs modalités. Pour les convertir en valeurs numériques, il est essentiel de prendre en compte cet ordre et de l'interpréter correctement.
- Les variables catégoriques nominales (par exemple : 'category', 'main category', 'country', etc.).

4-Conversion de la colonne 'goal' en USD en tenant compte de la colonne 'currency'.
5-Traitement des données manquantes. Certaines colonnes présentent des valeurs manquantes, notamment 'usd pledged' (3790 valeurs manquantes), 'category' (5 valeurs

manquantes), 'name' (4 valeurs manquantes). Plusieurs techniques peuvent être utilisées, telles que la suppression des observations avec des valeurs manquantes ou le remplissage des valeurs manquantes en utilisant des techniques telles que la moyenne, la médiane, le mode, la dernière valeur valide, etc. Pour la colonne 'usd pledged', il est possible d'estimer la valeur manquante en utilisant la colonne 'pledged' et 'currency'.
6- Correction des valeurs incorrectes dans différentes colonnes, notamment dans la colonne 'state'. J'ai filtré les données pour ne conserver que les observations valides.
7- Suppression des colonnes non nécessaires pour le modèle d'apprentissage machine.
8- J'ai vérifié la qualité des données afin de les préparer pour l'apprentissage machine.
9- J'ai établi des règles pour vérifier la cohérence des données, telles que la concordance entre la colonne 'currency' et 'usd pledged' lorsque 'currency' est égal à USD, ou la relation entre 'status' et 'usd pledged' lorsque 'status' est égal à 'successful'.
10- J'ai filtré les données pour ne conserver que les exemples ayant le statut 'successful' ou 'failed', car la tâche consiste à concevoir un outil d'aide à la décision pour prédire le succès des campagnes avant leur lancement.
11- La base de données n'est pas équilibrée, c'est-à-dire que le nombre d'exemples avec le statut 'successful' n'est pas égal au nombre d'exemples avec le statut 'failed'. J'ai utilisé la méthode de sur-échantillonnage pour équilibrer la base de données.

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

- Limite-toi aux trois observations les plus pertinentes selon-toi (appuies ces observations avec un visuel).
- 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 ?
- Est-ce que les « insights » trouvés peuvent être transformés en « features » qui faciliteront l'apprentissage du modèle ML ?

Réponse Q2:

Après avoir préparé les données, c'est-à-dire les avoir extraites et transformées, la prochaine étape consiste à effectuer une analyse complète des données. L'objectif est de comprendre, d'identifier les tendances, d'analyser les relations entre les colonnes, de tirer des conclusions et de générer des idées sur la distribution de chaque variable. Cette étape nous permettra de déterminer quelles sont les colonnes les plus pertinentes pour notre problématique et celles ayant le plus grand potentiel pour prédire la variable dépendante.
Dans le notebook "**2-exploratory_data_analysis_EDA**", j'ai créé différents graphiques, tels que des histogrammes et des distributions, pour repérer d'éventuelles données aberrantes et comprendre les relations et corrélations entre les colonnes.
Dans le notebook "**3-advanced_analysis**", j'ai cherché à répondre à plusieurs questions pertinentes qui ont contribué à la génération d'idées et à la création de nouvelles variables représentant des "features". Ces questions incluent :

- Identifier la catégorie principale (main category) la plus fréquente dans la base de données.
- Calculer les taux de succès et d'échec pour chaque catégorie principale.
- Déterminer la valeur moyenne de 'goal' pour les exemples ayant réussi et échoué.
- Calculer la valeur moyenne de 'pledged' pour les exemples ayant réussi et échoué.
- Examiner le rapport entre 'goal' et 'pledged' pour les exemples réussis et échoués.
- Étudier l'influence de la valeur de 'goal' sur le succès du projet.
- Analyser l'impact du nombre de backers sur le succès du projet.
- Évaluer l'influence de la localisation du projet sur son succès.

Ces analyses approfondies des données sont essentielles pour mieux comprendre les tendances et les relations au sein de la base de données, ainsi que pour identifier les variables importantes pour notre modèle d'apprentissage machine.
En ce qui concerne les facteurs de confusion, il est effectivement possible que certaines variables puissent biaiser nos analyses. Ces variables, s'ils ne sont pas correctement

contrôlés, peuvent fausser nos résultats. Il est donc essentiel de les prendre en compte dans nos analyses et d'utiliser des méthodes statistiques appropriées pour contrôler leur impact. Enfin, vous avez tout à fait raison de considérer la création de "features" à partir des insights obtenus. Ces nouvelles caractéristiques peuvent améliorer la performance de votre modèle d'apprentissage machine en capturant des informations pertinentes découvertes au cours de l'analyse des données.

Question#3 : Au niveau de la solution ML:

- 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 ?
- 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?

Réponse Q3:

- Mon interprétation des résultats de ma solution ML : La solution que je propose permet d'estimer la probabilité de succès ou d'échec d'un projet sur Kickstarter en utilisant des données d'entrée telles que la catégorie, la catégorie principale, la devise, le pays, le montant demandé (goal), etc.
- En ce qui concerne les parties prenantes, nous pouvons identifier trois principaux acteurs : les porteurs de projet, les investisseurs et la plateforme Kickstarter. Du point de vue des porteurs de projet, cette solution ML leur offre un outil précieux pour évaluer leurs chances de succès avant de lancer leur campagne sur la plateforme Kickstarter. Cela leur permet d'ajuster leur stratégie en conséquence et d'identifier des objectifs de financement réalistes pour optimiser leurs chances de réussite.
- Pour les investisseurs, cette solution ML représente un moyen de prendre des décisions éclairées en sélectionnant des projets ayant une probabilité plus élevée de réussite. Cela réduit le risque d'investir dans des projets non rentables.
- Enfin, du point de vue de la plateforme Kickstarter, cette solution ML renforce sa crédibilité en offrant un outil d'aide à la décision. Cela inspire confiance aux investisseurs et les incite à participer davantage aux campagnes hébergées par la plateforme.
- À mon avis, pour gagner la confiance de toutes les parties prenantes, il est essentiel de garantir une présentation claire et transparente des résultats de ma solution ML. Cela pourrait inclure des résultats de probabilité de succès accompagnés d'un rapport explicatif professionnel, permettant à toutes les parties de répondre à la question "Pourquoi". De plus, pour soutenir la solution ML et améliorer sa performance, il serait judicieux de maintenir le modèle en le réentraînant avec des données fraîchement collectées sur la plateforme.

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 ?

Réponse Q4:

Dans le cycle de vie d'un modèle d'apprentissage machine, on commence par la définition du problème et on termine par le déploiement de la solution. Cependant, il est essentiel de suivre l'évolution de la performance du modèle au fil du temps afin de contrôler et d'améliorer sa précision.
Le scénario décrit dans cette question est souvent appelé "data drifting" en anglais, ce qui évoque la notion de dérive des données. La dégradation de la performance de la solution ML est généralement due à des évolutions des variables au fil du temps. Pour faire face à cette situation, il est nécessaire de mettre en place une surveillance continue de la performance du modèle et de planifier des réentraînements réguliers avec des données fraîchement collectées. Cette approche permet de maintenir le modèle à jour et de refléter les changements dans l'environnement, assurant ainsi sa pertinence continue.