Stage data scientist

Objectif du projet :

Le but est ici de tester tes réflexes dans le cadre d'un projet de ML. Il n'y a pas d'objectif de résultat (dans la classification) et il n'est donc pas nécessaire d'utiliser des algorithmes compliqués. En revanche, il est important d'être rigoureux.se dans ton projet, d'être capable d'expliquer et justifier tes choix et de pouvoir correctement évaluer tes résultats finaux (éventuellement pour dire qu'ils ne sont pas bons). **Aucune connaissance métier ne t'es demandée seule la démarche ML est évaluée.**

Description du projet :

Nous essayons ici de mettre en place un moteur de recommandations personnalisées pour nos utilisateurs. Nous avons déjà construit un dataset en situation réelle en proposant, par mail, des projets à nos utilisateurs (selon les caractéristiques décrites ci dessous) et nous voulons aujourd'hui améliorer nos résultats. Pour ce faire, nous cherchons ici à construire un modèle capable de prédire si l'utilisateur va, ou pas, contribuer à un projet qu'on lui propose (problème de classification). Nous utiliserons ensuite le modèle trouvé pour recommander les projets les plus pertinents pour chaque utilisateur.

Rendu attendu :

Ton rendu doit comporter le code source du projet, une explication rapide de ta démarche et tes résultats (on en parlera ensuite ensemble) et un README qui me permette de savoir comment installer d'éventuelles librairies et lancer ton projet. Je dois pouvoir être capable de lancer ton modèle sur le dataset de test et de récupérer le score total (à toi de choisir le score qui te semble le plus pertinent pour ce projet) de ton modèle sur ce dataset de test.

Description du dataset :

Le dataset est constitué de 18489 data points séparés en un dataset d'entraînement (recos_training.csv, environ ¾ des données) et un dataset de test (recos_test.csv, environ ¼ des donnés). Le dataset est constitué de 25 colonnes :

- id : juste un id de la recommandation
- contrib : variable cible, 1 veut dire que l'utilisateur a participé au projet recommandé, 0 veut dire que ce n'est pas (encore) le cas.

Et 23 variables qui constituent l'ensemble des variables d'entrée du problème :

- same_owner : booléen qui exprime le fait que l'utilisateur a déjà soutenu un projet du même porteur de projet. 1 signifie que oui, 0 que non.

- owner_friend : booléen qui exprime le fait que l'utilisateur est un ami facebook du porteur de projet. 1 signifie que oui, 0 que non.
- nb_friends : entier représentant le nombre d'amis facebook de l'utilisateur ayant soutenu le projet.
- amount_friends : nombre réel représentant le montant engagé par les amis facebook de l'utilisateur sur le projet.
- nb_copledgers : nombre réel représentant le nombre de personnes qui avaient contribué au même projet que l'utilisateur par le passé et ont participé à ce projet (peut ne pas être un entier).
- amount_copledgers : nombre réel représentant le montant engagé par les copledgers (terme défini précédemment) sur le projet.
- desc_score_mean : nombre entre 0 et 1 exprimant la similarité entre la description du projet et celles des projets auxquels l'utilisateur a déjà participé. Plus le nombre est proche de 1, plus les descriptions sont jugées similaires.
- dist : nombre entre 0 et 1 représentant la distance physique entre l'utilisateur et le projet. 1 veut dire que le projet est dans la ville de l'utilisateur, 0 veut dire que le projet est à plus de 100 km de l'utilisateur.

Les autres variables concernent les différentes catégories des projets de notre plateforme et symbolisent le fait que l'utilisateur a, ou pas, déjà soutenu des projets de la même catégorie que le projet présenté. Cet attrait pour une catégorie est représenté par un nombre entre 0 et 1, 1 signifie un attrait total pour la catégorie, 0 veut dire que l'utilisateur n'a jamais participé à un projet de cette catégorie ou que le projet proposé n'est pas dans cette catégorie. La liste des catégories est la suivante :

- music
- live-performance
- journalism
- book-and-publishing
- design-and-innovation
- event
- film-and-video
- style
- photography
- social
- web-and-tech
- education
- art
- adventure-and-sport
- food
- ecology