Projet : Prédire la popularité d'un animé

Contexte et présentation du dataset

Le dataset utilisé dans ce projet provient de la plateforme Kaggle, sous le nom Anime Dataset. Il rassemble des données issues de MyAnimeList (MAL) et de l'API Jikan, et couvre plus d'un siècle de production d'animés. Cette base de données, composé de treize fichiers, nous renseigne sur :

- les animés eux-mêmes (informations générales, genres, studios, synopsis, scores),
- leurs personnages et les relations personnage–anime,
- les membres du staff (réalisateurs, doubleurs, musiciens),
- les utilisateurs de MyAnimeList et leurs évaluations,
- les relations entre animes (système de recommandations proposé par MAL).

Nature du problème

Ce projet vise à répondre à deux questions :

1. Prédire la popularité et la note globale d'un nouvel anime.

Il s'agit d'un problème de régression, où l'on cherche à expliquer les variables cibles issues de MyAnimeList à partir des caractéristiques intrinsèques de l'œuvre.

2. Identifier les utilisateurs les plus susceptibles d'apprécier un nouvel anime.

Même si la recommandation individuelle peut se formuler comme un problème de classification (prédire si un utilisateur aimera un anime ou non), nous adoptons ici une approche en régression, plus naturelle compte tenu des notes continues disponibles.

Présentation du dataset

Pour réaliser ce projet, deux tableaux ont été construits.

1) Dataset anime_level

Ce tableau rassemble uniquement des caractéristiques intrinsèques aux animés :

- les méta-données descriptives (year, episodes, synopsis_length) ;
- des variables catégorielles transformées en indicatrices (type d'anime, rating, saison, source d'adaptation) ;
- des variables multi-labels dérivées (genres, thèmes, studios, producteurs) ;
- des caractéristiques liées au contenu : nombre de personnages, qualité du cast, richesse musicale, recommandations sortantes.

Ont été exclues toutes les variables qui reflètent directement la réaction du public (score, rang, popularité, nombre de votes ou de membres). Ces informations apparaissent uniquement après la sortie d'un anime, elles ne permettraient donc pas de prédire un nouvel anime. La variable cible Y correspond à la colonne 'score' de *details.csv*, qui correspond à la note moyenne globale donnée par les utilisateurs.

2) Dataset user_anime_level

Ce tableau vise à modéliser l'intérêt potentiel d'un utilisateur pour un anime. Nous ne conservons que les variables nécessaires pour décrire :
- l'interaction (note passée, statut de visionnage, ré-visionnage, épisodes vus) ;
- le profil utilisateur, via des résumés statistiques (nombre d'animes vus, score moyen, taux de rewatch...) ;
- les préférences explicites, via ses favoris par catégorie et un profil de genres (moyenne des indicateurs de genres des animes qu'il apprécie) ;
- le contenu minimal de l'anime, réduit à quelques variables compactes (genres principaux, type, rating, année, synopsis_length) afin de limiter la dimensionnalité et la consommation mémoire.

Nous avons mis de côté les caractéristiques les plus lourdes (studios, producteurs, ou l'ensemble des multi-hot très détaillés), car elles génèrent un nombre énorme de colonnes et compliquent fortement l'analyse. Dans user_anime_level, la variable cible est la colonne 'score' de *ratings* qui correspond à la note réellement donnée par un utilisateur à un anime.

Le fichier *ratings.csv* contient plus de 120 millions d'observations, ce qui dépasse les capacités mémoire disponibles sur Rstudio. Nous avons décidé d'utiliser un sous-échantillon aléatoire de 2 millions de lignes

La *Figure 1* compare la distribution des notes du fichier complet *ratings.csv* à celle d'un échantillon aléatoire de 2 millions de lignes. Les proportions observées pour chaque note (1 à 10) se superposent presque parfaitement.

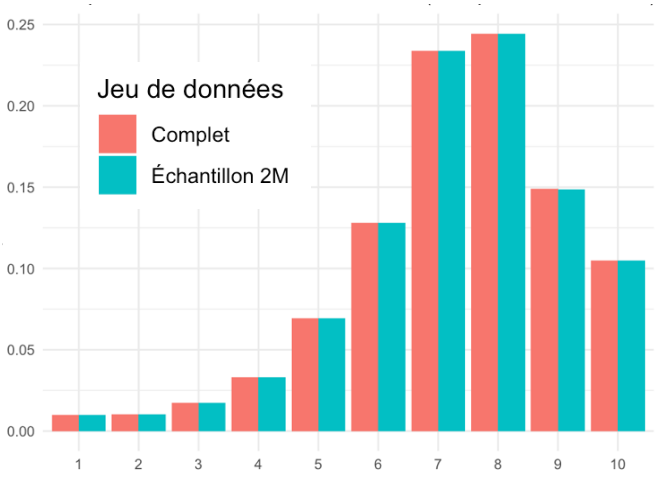

Figure 1 : Distribution des notes (Complet vs Échantillon) $\in [1,10]$

Cette coïncidence confirme que l'échantillonnage est représentatif, et qu'il peut être utilisé de manière fiable pour entraîner les modèles sans perte d'information significative, tout en respectant les contraintes mémoire.

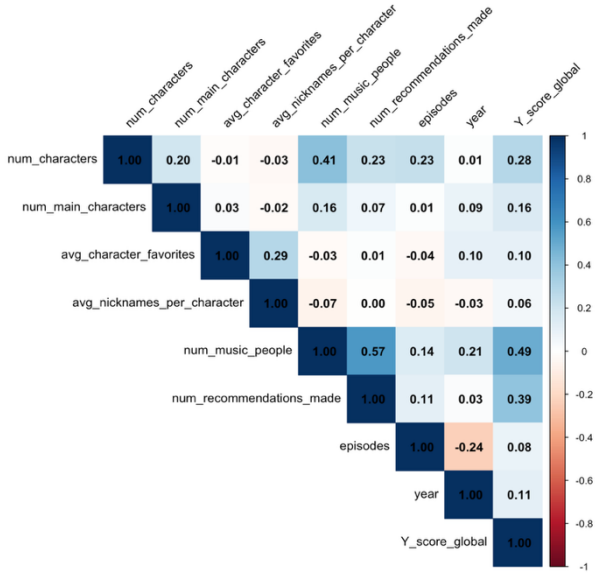

Figure 2 : Matrice de corrélation anime_level

Nous avons observé la matrice de corrélation en *Figure 2* entre certaines variables du tableau anime_level, et celle-ci révèle que seules quelques relations ressortent nettement — notamment les liens positifs entre participation musicale, recommandations et note globale — tandis que la majorité des autres variables sont faiblement corrélées avec la note.

En définitive, ces observations nous aident à repérer les relations les plus utiles et orientent directement le choix des variables à intégrer dans les futurs modèles de régression.