Analyse des données de MovieLens

Rebecca Leygonie & Walid Laouer

Dans le cadre de l'UE 3i026 : Intelligence Artificielle et Science des Données

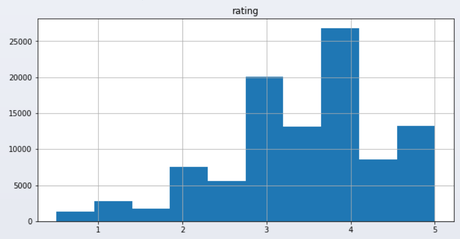
Visualisation des données

Deux bases de données complémentaires :

❖ MovieLens :
- Movies
- Links
- Ratings
- Tags

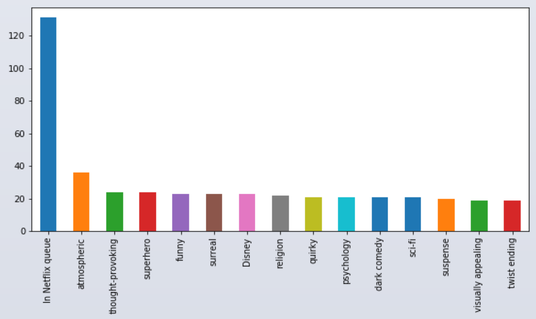
❖ TMDb :
- Crew
- Films
- Acteurs

Fréquence de distribution des notes



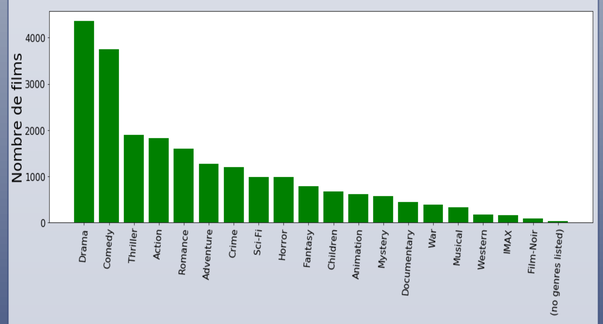
➢ En abscisse : les notes allant de 0 à 5
➢ En ordonné : le nombre de fois ou la note à été donné

Les quinze mots les plus utilisés dans les commentaires écrits par les internautes



➢ En abscisse : les mots
➢ En ordonné : leurs fréquence d'apparition

Popularité des catégories de films

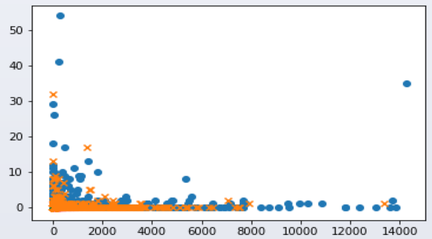


Pour la suite du projet, nous avons mis les données sous forme de DataFrame en utilisant par exemple le one-hot encoding pour les catégories.

Classification supervisée

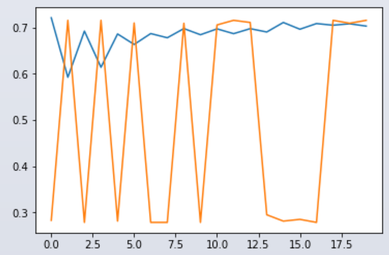
❖ Problématique : On cherche à savoir si lorsque les gens s'expriment sur un film, c'est pour le dénigrer ou le louer.
❖ Méthode : on définit comme une bonne toute note supérieure ou égale à 3/5 et on essaye de prédire si un film obtient une bonne note en fonction du nombre de votes ainsi que du nombre de commentaires écrits par les internautes.
❖ Algorithmes : KNN et Perceptron

Visualisation graphique des données sur lesquelles vont être appliquées les modèles



➢ En abscisse : le nombre de votes
➢ En ordonné : le nombre de commentaires

Évolution de l'accuracy des modèles de classifications.



➢ En abscisse : La valeur de k (nombre de voisins dans KNN)
➢ En ordonné : l'accuracy
➢ En bleu : KNN
➢ En rouge : Perceptron

On remarque qu'à la fin, l'accuracy des deux modèles est la même et vaut : 0,71
Bien que l'évolution de l'accuracy de l'algorithme du Perceptron ait l'air d'agir comme un algorithme aléatoire, ce n'est pas le cas. Il s'agit juste d'un échange de label qu'on peut ne pas prendre en compte.

Visuellement parlant, il n'est pas évident que les modèle, utilisés soient fonctionnels pour la résolution de notre problématique.
De plus l'accuracy qui est de 0,71 montre que le nombre de votes et le nombre de commentaires ne sont pas suffisants pour prédire la note.

On peut tout de même voir graphiquement que les points bleus (note>=3) sont plus présents que les oranges (note<3) lorsque le nombre de commentaires et de votes augmentent.

C'est pourquoi, bien que peu encourageants, ces résultats ne nous arrêteront pas dans l'exploration de l'hypothèse que les utilisateurs s'expriment plus lorsque leurs avis sur un film sont positifs.

Regression supervisé

❖ Problématique : On cherche toujours à savoir si lorsque les gens s'expriment sur un film, c'est pour le dénigrer ou le louer.
❖ Méthode : prédire la note d'un film donnée par les internautes, en fonction du nombre total de mots des commentaires de ce film.
❖ Algorithmes : Linear Regression

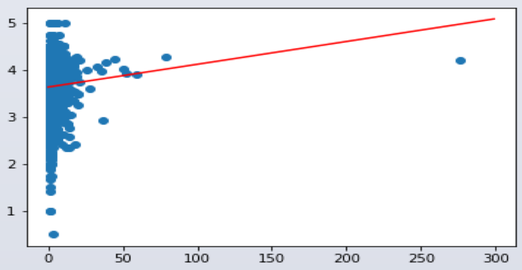
Résultats de la régression linéaire :

R square: 0.0062217227096326155
Corrélation: 0.07887789747218554

On remarque que le score de régression est très mauvais.
Ainsi, notre modèle de régression n'a pas réussis a prédire les notes en fonction du nombre total des mots contenus dans les commentaires.

On remarque aussi que le coefficient de corrélation est très faible, ce qui signifie que le nombre de mots des commentaires n'est pas un descriptif assez important pour prédire la note.

Visualisation des données après application de la régression linéaire



➢ En abscisse : la note allant de 0 à 5
➢ En ordonné : le nombre de mots

On remarque sur ce graphe que les points ne sont pas clairement alignés de part et d'autre de la droite, ce qui explique le faible coefficient de corrélation.
Pour pouvoir prédire la note, il faudrait rajouter davantage de variables explicatives, c'est-à-dire des variables de contrôle.

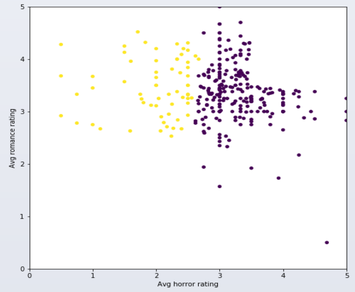
Malgré le faible pouvoir explicatif de cette régression, on note que le coefficient de corrélation associé au nombre de mots est de signe positif, ce qui suggère que les gens auraient plutôt tendances a laisser des commentaires pour exprimer une opinion positive sur les films plutôt que négative.

Pour conclure, cette première analyse des notes attribuées aux films en fonction des commentaires est un peu décevante.
Nous aimerions, en disposant de plus de temps, la compléter afin de prendre en compte, parmi les variables explicatives, non seulement le nombre de mots mais aussi mais le nombre de commentaires (donc le nombre de mots par commentaire). Jouent également certainement le mois de sortie du film, la réputation du réalisateur et des acteurs, etc.

Catégorisation non supervisé

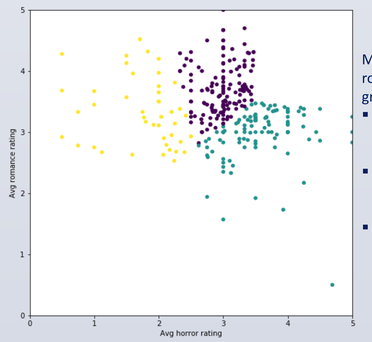
❖ Problématique : Dans cette partie, nous allons essayer d'étudier les préférences cinématographique des utilisateurs en terme de catégorie de films. Nous allons prendre deux catégories : Romance et Horror, qui nous semblent opposés
❖ Méthode : calcul de la note moyenne de chaque utilisateur pour tous les films d'amour et tous les films d'horreur. Pour partitionner notre ensemble de données nous supprimons les personnes qui aiment à la fois l'horreur et la romance, afin que nos clusters aient tendance à les définir comme aimant un genre plus que l'autre.
❖ Algorithme : K-moyennes

Visualisation des données après catégorisation (nombre clusters=2)



Nous pouvons voir que les groupes sont principalement basés sur la façon dont chaque personne a évalué les films d'horreur. Si leur cote moyenne de films d'horreur est supérieure à 2,5 étoiles, ils appartiennent à un groupe. Sinon, ils appartiennent à l'autre groupe. Que se passerait-il si nous les divisions en trois groupes?

Visualisation des données après catégorisation (nombre clusters=3)



Maintenant, la note moyenne en romance commence à jouer. Les groupes sont:
- les gens qui aiment la romance mais pas les films d'horreur (jaune)
- les gens qui aiment les films d'horreur mais pas la romance (vert)
- les gens qui aiment les films d'horreur et la romance (violet)

Nous avons constaté que plus nous divisions notre ensemble de données en un grand nombre de groupes, moins nous parvenions à obtenir des groupes d'individus aux goûts homogènes au sein des groupes, et différents entre groupes.

Conclusion

Nous avons tenté d'analyser une base de données de films. Nous avons exploré des hypothèses qui découlaient de nos observations de la vie quotidienne, à savoir la pertinence de deux catégories de films, et la valeur des commentaires laissés sur internet. Les données n'ont pas validé ces hypothèses. Loin de nous décourager, ce résultat démontre à nos yeux l'importance du travail empirique, notamment statistique.