Rapport de projet

DATA731 : Modélisation Stochastique

Prédiction de films

Enseignant :

Atto Abdourrahmane

décembre 2020

Déborah Guilly
Simon Guilbert

Introduction - Choix des données

Ce projet consistait à répondre à un problème posé en se servant de données analysées grâce à des méthodes statistiques. Nous avions le choix entre 3 sujets. Le premier était basé sur l'analyse d'évènements naturels (coulées de lave), le deuxième se concentrait sur l'évolution du PIB dans un certain nombre de pays, et le troisième sujet était un sujet "libre" qui nous permettait de choisir les données de notre choix, dans n'importe quel domaine. Nous avons choisi de traiter ce dernier sujet.

Nous avons donc cherché sur plusieurs sites internet qui fournissent gratuitement des données et nous nous sommes arrêtés sur un ensemble de fichiers qui répertorient un certain nombre d'informations en rapport avec des films. Nous avons trouvé ces données sur le site tensorflow.org.

A l'intérieur de ces fichiers, on avait accès :

- aux titres et aux genres de 59000 films
- aux notes attribuées pour ces films par 283000 spectateurs
- aux tags (mots clés) attribués aux films par les spectateurs

Problématique

Le problème auquel nous allions essayer de répondre nous est rapidement venu à l'esprit : Quels films pourrions-nous recommander aux spectateurs en fonction de leurs goûts cinématographiques ?

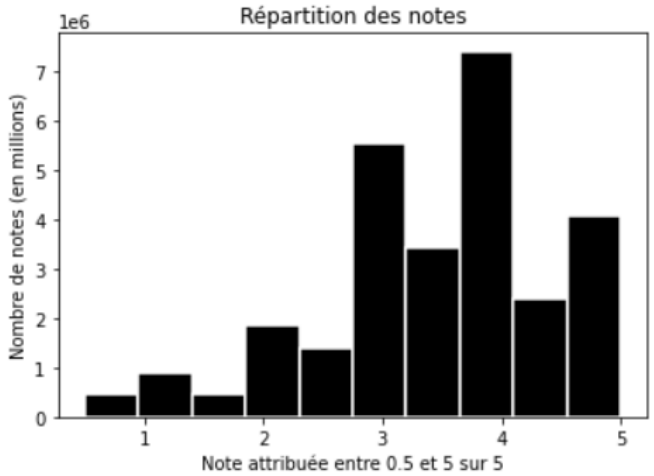
Pour répondre à cette problématique, nous nous sommes fixés pour objectif de prédire le film le plus intéressant à voir pour chaque utilisateur et de fournir le résultat sous forme d'un tableau semblables à celui ci-dessous :

Utilisateur	Titre du film conseillé
001	Titanic
002	Toy Story
003	Pulp Fiction
004	Finding Nemo

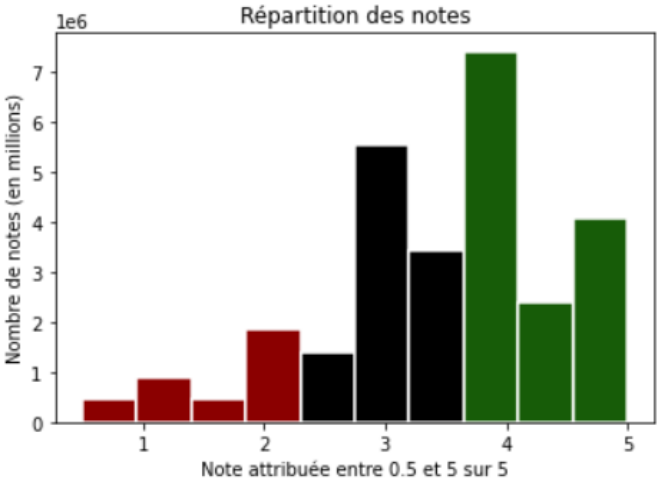
Recherche des informations utiles pour répondre au problème

1. Les notes

Parmi les informations que nous avions à disposition, nous avons commencé par regarder comment les spectateurs évaluent les films qu'ils voient, c'est-à-dire s'ils ont tendance à les surévaluer ou au contraire à les sous-évaluer. Pour cela, nous avons tracé l'histogramme de toutes les notes comprises entre 0.5 et 5/5 :



Sur l'histogramme ci-dessus, on peut remarquer qu'il y a plus de notes à droite du graphique qu'à gauche, ce qui veut dire que les spectateurs ont tendance à surévaluer les films. De plus, la médiane est égale à 3,50 donc la moitié des notes sont supérieures à 3.5/5. On peut alors considérer que si un utilisateur a vraiment apprécié un film, c'est qu'il lui a mis une note d'au moins ⅘ et à l'inverse s'il ne l'a pas aimé c'est qu'il l'a noté moins de ⅖.



Dans la suite de l'analyse, on va essentiellement se concentrer sur les notes représentées en rouge et en vert sur l'histogramme ci-dessus puisque ce sont ces notes qui déterminent le mieux le ressenti d'un film sur un spectateur. Par exemple, si un utilisateur attribue la note de 1,5/5 au film Matrix et 5/5 au film Inception, on se souviendra que cet utilisateur en particulier n'aime pas le premier film mais adore le second. En revanche, s'il note le film Forrest Gump

3.5/5, ça signifie qu'il l'aime moyennement et donc ce film ne sera pas pris en compte dans la suite de l'analyse.

Utilisateur	Film	Note
007	Matrix	1.5
007	Forrest Gump	3.5
007	Jurassic Park	0.5
007	Apollo 13	3
007	Inception	5

2. Les genres

Pour chaque film sont attribués des genres qui permettent de classer les films par catégorie. Parmi ces genre on peut retrouver :

- science fiction
- drame
- thriller
- action
- aventure
- …

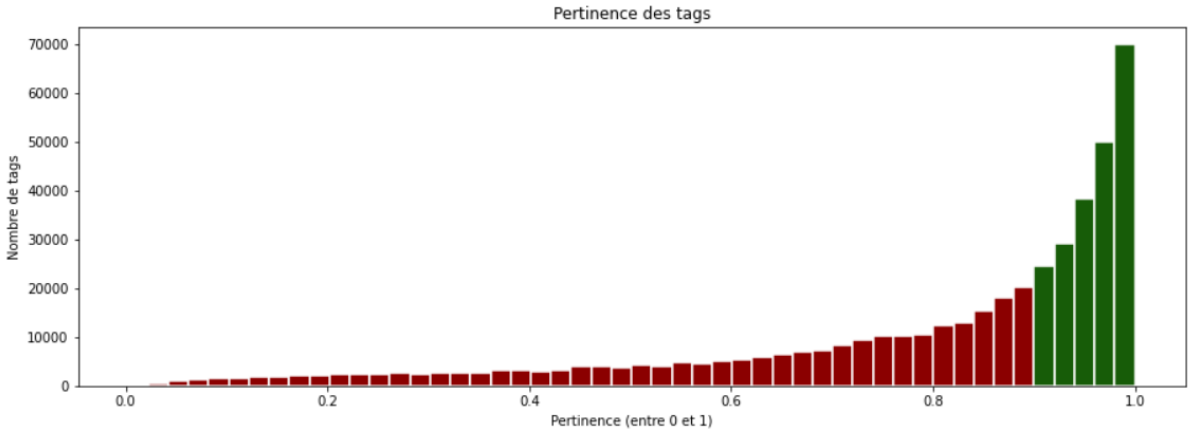
Il peut être intéressant de créer des liens entre les films du même genre pour améliorer notre système de prédiction. C'est ce que nous ferons par la suite.

3. Les tags

En plus de donner une note à un film, les utilisateurs ont la possibilité de donner un mot clé (appelé tag) en rapport avec le film. Pour chaque tag, une intelligence artificielle détermine la probabilité entre 0 et 1 que le tag soit cohérent avec son film associé. Par exemple, dans le tableau ci-dessous on peut voir que le tag "Robot" a un niveau de pertinence très faible (1.58%) avec le film "The Lion King" alors que le mot clé "Disney" est beaucoup plus pertinent (94.63%).

Film	Tag	Pertinence
The Lion King	Robot	0.0158
The Lion King	Simba	0.9024
The Lion King	Funny	0.6859
The Lion King	Disney	0.9463
The Lion King	Family	0.9178

Nous avons donc voulu savoir si les tags attribués par les spectateurs en rapport avec les films étaient plutôt pertinents ou pas. On a donc tracé l'histogramme suivant :



En s'appuyant sur cet histogramme et en sachant que la médiane est égale à 0.88, on en conclut que les tags sont en général pertinents puisque la moitié d'entre eux sont pertinents à plus de 88%. Dans la suite de l'analyse on va garder uniquement les tags les plus pertinents, c'est-à-dire ceux qui sont représentés en vert dans l'histogramme (pertinents à plus de 90%).

Film	Tag	Pertinence
The Lion King	Robot	**0.0158**
The Lion King	Simba	**0.9024**
The Lion King	Funny	**0.6859**
The Lion King	Disney	**0.9463**
The Lion King	Family	**0.9178**

Ainsi, les tags "Simba", "Disney" et "Family" seront conservés tandis que "Robot" et "Funny" ne seront pas pris en compte dans l'analyse puisque leur niveau de pertinence est inférieur à 90%.

Le système de points

Pour pouvoir conseiller au mieux l'utilisateur sur un film qu'il pourrait aimer, on commence par lister tous les films qu'il n'aime pas (notés moins de ⅖) et tous ceux qu'il apprécie (notés plus de ⅘).

Ensuite, on calcule pour chacun des 59000 films un nombre de points de la façon suivante :

Même genre que ses films préférés	→	+ 5 points
Même genre que ses films les moins appréciés	→	- 3 points
Même tag que ses films préférés	→	+ 2 points
Même tag que ses films les moins appréciés	→	- 1 points

Pour finir, il ne reste plus qu'à faire ressortir le film qui a le plus de points et ce sera celui qui sera conseillé à l'utilisateur.

Résultats obtenus

Voici un extrait des résultats obtenus. On a bien ce qu'on voulait au départ, c'est-à-dire une colonne pour identifier l'utilisateur, une deuxième colonne avec le titre du film conseillé et une dernière avec le nombre de points.

Utilisateur	Film Conseillé	Points
82	Fight Club (1999)	61
83	Stargate (1994)	26
84	Pulp Fiction (1994)	70
85	Lady and the Tramp (1955)	20
86	Thing, The (1982)	24
87	Strange Days (1995)	11
88	Pulp Fiction (1994)	77
89	Twister (1996)	17
90	Matrix, The (1999)	48
91	Bottle Rocket (1996)	17
92	Green Mile, The (1999)	32
93	Superman (1978)	30
94	Footloose (1984)	21
95	Pulp Fiction (1994)	79
96	Inception (2010)	57
97	Waiting for Guffman (1996)	19
98	American Beauty (1999)	35
99	Lord of the Rings: The Fellowship of the Ring, The (2001)	61
100	Pulp Fiction (1994)	79

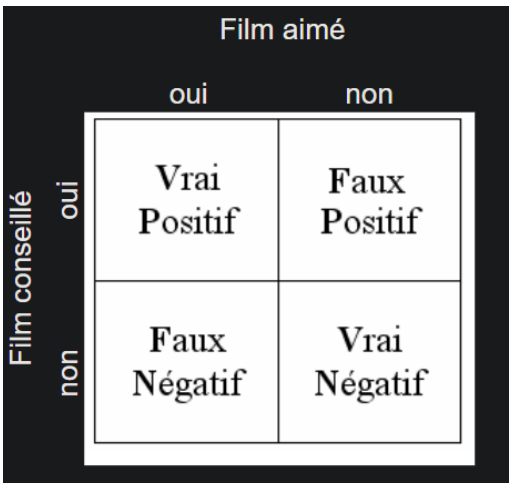
Problèmes rencontrés - Améliorations possibles

La première remarque que l'on peut faire sur ce résultat est que certains films comme par exemple "Pulp Fiction" reviennent très souvent. C'est peut-être un hasard que plusieurs personnes aient exactement les mêmes goûts, mais c'est quand même étrange que parmi les 59000 films à disposition certains reviennent fréquemment. Pour corriger ce problème, il faudrait modifier le système de points pour éviter de trop avantager certains films.

Un autre élément qui nous a beaucoup freiné dans notre analyse est que parmi tous les fichiers de données à disposition il n'y avait aucune information démographique sur les spectateurs. On ne connaissait par exemple pas leur sexe, leur age, ou encore leur nationalité. Ce manque d'informations nous a empêché de classer les utilisateurs afin de leur conseiller des films plus appropriés.

Ensuite, comme expliqué précédemment, ce sont les utilisateurs qui créent les tags : plusieurs personnes peuvent vouloir dire la même chose sans l'orthographier de la même façon. Par exemple deux personnes différentes peuvent attribuer les tags "science fiction" et "sci-fi" mais ces 2 tags ne seront pas considérés comme identiques par notre algorithme bien qu'ils veuillent dire la même chose. On a alors une perte d'informations.

Nous ne pouvons pas dire si les résultats obtenus sont cohérents puisque pour le savoir il faudrait que les utilisateurs à qui on conseille des films voient, et qu'ils nous donnent leur avis. Si on avait pu avoir ces avis, on aurait créé une matrice de confusion pour savoir si notre système fonctionne.



Enfin le plus gros problème que nous avons rencontré est que le calcul du nombre de points pour chaque film est très long. Il faut compter à peu près 18 secondes pour trouver le meilleur film pour 1 utilisateur, et donc pour en traiter 283000 cela prendrait un temps considérable.

Conclusion

Pour conclure, nous avons atteint l'objectif et à répondre au problème initial en conseillant un film à voir pour chaque utilisateur. Cependant notre système est perfectible notamment sur la durée de traitement beaucoup trop longue pour pouvoir satisfaire tous les utilisateurs.

Nous avons bien conscience que le travail réalisé au cours de ce projet par notre groupe a été très différent de celui des autres équipes de la promotion. Cette différence peut s'expliquer par notre choix des données à traiter puisqu'elles ne contenaient pas de séries temporelles à analyser contrairement au sujet sur le PIB par exemple.