COMPTE RENDU DU PROJET

Roussel Desmond NZOYEM NGUEGUIN

N° étudiant : 21911823

Courriel : roussel-desmond.nzoyem-ngueguin@etu.unistra.fr

Université de Strasbourg

UFR de Mathématique et d'informatique

Master 1 CSMI

Traitement et fouilles de données

Sous la supervision de M. **Vincent VIGON**

20 juin 2020

Contenu

I. INTRODUCTION

Nous allons nous intéresser à la prédiction des genres d'un film, et de son succès financier. Tout d'abord, nous retraiterons les données et nous les exprimerons sous une forme appropriée en vu de l'analyse des variables qualitatives et quantitatives. Enfin nous effectuerons les apprentissages en utilisant plusieurs algorithmes différents.

II. RETRAITEMENT

1. Description des données originales

Le jeu de données original a la forme suivante :

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	status	tagline	title	vote_average	vote_count
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577	[{"name": "Ingenious Film Partners", "id": 289...	[{"iso_3166_1": "US", "name": "United States o...	2009-12-10	2787965087	162.0	[{"iso_639_1": "en", "name": "English"}, {"iso...	Released	Enter the World of Pandora.	Avatar	7.2	11800
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	139.082615	[{"name": "Walt Disney Pictures", "id": 2}, {"...	[{"iso_3166_1": "US", "name": "United States o...	2007-05-19	961000000	169.0	[{"iso_639_1": "en", "name": "English"}]	Released	At the end of the world, the adventure begins.	Pirates of the Caribbean: At World's End	6.9	4500
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	en	Spectre	A cryptic message from Bond's past sends him o...	107.376788	[{"name": "Columbia Pictures", "id": 5}, {"nam...	[{"iso_3166_1": "GB", "name": "United Kingdom"...	2015-10-26	880674609	148.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...	Released	A Plan No One Escapes	Spectre	6.3	4466
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	en	The Dark Knight Rises	Following the death of District Attorney Harve...	112.312950	[{"name": "Legendary Pictures", "id": 923}, {"...	[{"iso_3166_1": "US", "name": "United States o...	2012-07-16	1084939099	165.0	[{"iso_639_1": "en", "name": "English"}]	Released	The Legend Ends	The Dark Knight Rises	7.6	9106
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	en	John Carter	John Carter is a war-weary, former military ca...	43.926995	[{"name": "Walt Disney Pictures", "id": 2}]	[{"iso_3166_1": "US", "name": "United States o...	2012-03-07	284139100	132.0	[{"iso_639_1": "en", "name": "English"}]	Released	Lost in our world, found in another.	John Carter	6.1	2124

Figure 1: Données originales

Il contient 4803 lignes et quelques-unes de ses 20 colonnes sont décrites ci-dessous :

- **budget** : le budget du film en dollar
- **revenue**: les box-office international
- **genres** : les genres du film (par exemple action, comédie, etc..) ; un film peut bien sûr appartenir à plusieurs genres différents
- **keywords** : les mots clés associés à ce film. Il s'agit d'une particularité de ce film (adaptation d'un roman, violence, etc.)
- **overview** : l'intrigue du film
- **tagline** : le "slogan" du film (mentionné par exemple sur l'affiche)
- **original_langage** : l'unique langue originale du film ; très souvent la langue du pays de production
- **sopken_language** : les langues parlées dans le film ; les langues inconnue sont indiqués par '???????'.
- **vote_count** : le nombre de votes enregistrés par le site www.themoviedb.org
- **vote_average** : la moyenne des votes
- **popularity** : une mesure de la popularité

A vu d'œil, quelques problèmes sont détectables dans la data frame.

- L'identifiant des films "id" n'est pas significatif. Nous le réexprimons immédiatement sous forme d'un entier compris entre 0 et la taille de la data frame.
- Les budgets sont exprimés en dollars ; on préfère le million de dollars.
- Certaines colonnes indispensables sont exprimées au format JSON ; nous préférons avoir des types python qui sont plus facilement manipulables.

D'autres problèmes sont plus difficiles à remarquer, par exemple :

- Des intrigues manquantes
- Des langues inconnues
- Des budgets, revenus ou durées aberrants

2. Retraitement des données

a. Suppression du format JSON

Pendant la suppression du format JSON sur les colonnes "genres", "keywords", "production_companies", "production_companies", "spoken_languages", on crée des dictionnaires qui seront utilisés pour chacune de ces catégories plus tard.

	budget	genres	id	keywords	original_language	overview	popularity	production_companies	production_countries
0	237.0	[Action, Adventure, Fantasy, Science Fiction]	0	[culture clash, future, space war, space colon...	en	In the 22nd century, a paraplegic Marine is di...	150.437577	[Ingenious Film Partners, Twentieth Century Fo...	[United States of America, United Kingdom]
1	300.0	[Adventure, Fantasy, Action]	1	[ocean, drug abuse, exotic island, east india ...	en	Captain Barbossa, long believed to be dead, ha...	139.082615	[Walt Disney Pictures, Jerry Bruckheimer Films...	[United States of America]
2	245.0	[Action, Adventure, Crime]	2	[spy, based on novel, secret agent, sequel, mi...	en	A cryptic message from Bond's past sends him o...	107.376788	[Columbia Pictures, Danjaq, B24]	[United Kingdom, United States of America]
3	250.0	[Action, Crime, Drama, Thriller]	3	[dc comics, crime fighter, terrorist, secret i...	en	Following the death of District Attorney Harve...	112.312950	[Legendary Pictures, Warner Bros., DC Entertai...	[United States of America]
4	260.0	[Action, Adventure, Science Fiction]	4	[based on novel, mars, medallion, space travel...	en	John Carter is a war-weary, former military ca...	43.926995	[Walt Disney Pictures]	[United States of America]
5	258.0	[Fantasy, Action, Adventure]	5	[dual identity, amnesia, sandstorm, love of on...	en	The seemingly invincible Spider-Man goes up ag...	115.699814	[Columbia Pictures, Laura Ziskin Productions, ...	[United States of America]
6	260.0	[Animation, Family]	6	[hostage, magic, horse, fairy tale, musical, p...	en	When the kingdom's most wanted-and most charmi...	48.681969	[Walt Disney Pictures, Walt Disney Animation S...	[United States of America]
7	280.0	[Action, Adventure, Science Fiction]	7	[marvel comic, sequel, superhero, based on com...	en	When Tony Stark tries to jumpstart a dormant p...	134.279229	[Marvel Studios, Prime Focus, Revolution Sun S...	[United States of America]
8	250.0	[Adventure, Fantasy, Family]	8	[witch, magic, broom, school of witchcraft, wi...	en	As Harry begins his sixth year at Hogwarts, he...	98.885637	[Warner Bros., Heyday Films]	[United Kingdom, United States of America]
9	250.0	[Action, Adventure, Fantasy]	9	[dc comics, vigilante, superhero, based on com...	en	Fearing the actions of a god-like Super Hero l...	155.790452	[DC Comics, Atlas Entertainment, Warner Bros.,...	[United States of America]

Figure 2: Résultat de la suppression du format JSON

Observons le nombre de données manquantes dans notre data frame :

```
Valeurs manquantes:
budget                     0
genres                     0
id                         0
keywords                   0
original_language          0
overview                   3
popularity                 0
production_companies       0
production_countries       0
release_date               1
revenue                    0
runtime                    2
spoken_languages           0
tagline                  844
title                      0
vote_average               0
vote_count                 0
```

Figure 3: Nombre de données manquantes par colonnes

Les 3 films à intrigues manquants seront supprimés. Le film sans date de sortie est maintenu, tout comme les films avec durée et slogan indéterminés.

b. Traitement des budgets, revenus et durées aberrants

Il n'y a pas de film à budget, revenu, ou durée anormalement élevée. En revanche, il y a en a qui sont anormalement faibles (environ 1500 films sur 4800). Ça serait trop douloureux de tous les supprimer. On a deux options :

- On peut remplacer ces valeurs aberrantes par les moyennes de chaque colonne. Ceci conduit à une data frame nommée "*df_1*", et enregistrée sous le nom "*tmdb_5000_movies_imputed.csv*". Elle contient **4800 lignes** et sera utilisée pour la prédiction du genre à partir des intrigues, vu que cet apprentissage n'est pas du out affecté par les budgets, revenus ou durées.
- On peut tout simplement supprimer ces valeurs. Ceci conduit à une data frame nommée "*df_2*", et enregistrée sous le nom "*tmdb_5000_movies_ommited.csv*". Cette data frame est moins riche (**3211 lignes**) mais plus fiable. Elle sera donc utilisée pour le second apprentissage. Il s'agira là de la prédiction du succès d'un film à partir de son budget, de sa durée, de ses genres, de ses mots clés, des compagnies qui l'on produit et des langues qui y sont parlées.

c. Créations de nouvelles variables

i. Des dummy variables pour le genre

Dans la suite, il nous sera nécessaire d'analyser les corrélations entre différents genres de films. Mais les genres sont qualitatifs. Nous allons donc créer des variables muettes correspondant à chacun des genres de films qu'on rencontre dans la data frame. Ces dummies seront la cible du premier apprentissage que nous allons effecteur (classification multi-label).

	title	genres	Action	Adventure	Fantasy	Science Fiction	Crime	Drama	Thriller	Animation	Family	Western	Comedy	Romance	Horror	Mystery	History	War	Music	Documentary	Foreign	TV Movie
0	Avatar	[Action, Adventure, Fantasy, Science Fiction]	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Pirates of the Caribbean: At World's End	[Adventure, Fantasy, Action]	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Spectre	[Action, Adventure, Crime]	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	The Dark Knight Rises	[Action, Crime, Drama, Thriller]	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	John Carter	[Action, Adventure, Science Fiction]	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4: Des variables muettes pour les genres des films

ii. Un meilleur indicateur du succès ou de l'échec

Il nous faut une métrique pour définir le succès ou l'échec d'un film. On décrète qu'il s'agit d'un :

- **échec** s'il rapporte moins que son budget ; ou s'il rapporte moins d' 1.25 fois son budget alors que ce budget était très grand (supérieur à 100 millions de dollars).
- **succès massif (**ou **retentissant)** s'il rapporte plus de 10 fois son budget ; ou s'il rapporte plus de 350 millions de dollars, ayant nécessité au plus le tiers de cela.
- **succès** dans tous les autres cas.

Ces catégories seront la cible du deuxième apprentissage que nous allons effectuer (classification multi-classe).

Observons la distribution des 3 catégories crées dans les deux data frames définies précédemment.

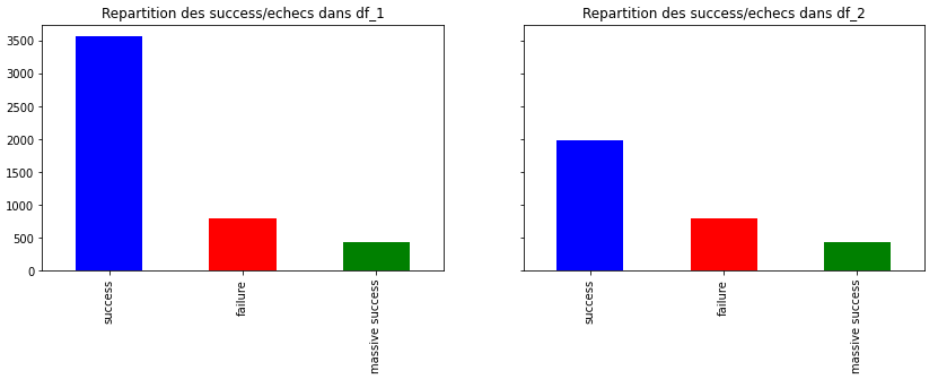

Figure 5: Répartition des succès/échecs

La df_2 est mieux équilibrée que la df_1 (même si elle n'est pas aussi riche). Ceci nous donne une motivation supplémentaire pour utiliser df_2 lors de la prédiction du succès des films.

On peut des à présent observer les 3 plus grands films à succès dans notre jeu de données.

	id	title	release_date	tagline	overview	keywords	genres	runtime	budget	revenue	return	return_type
4577	4577	Paranormal Activity	2007-09-14	What Happens When You Sleep?	After a young, middle class couple moves into ...	[haunting, psychic, entity, demonic possession...	[Horror, Mystery]	86.0	0.015	193.356	12890.400000	massive success
4496	4496	The Blair Witch Project	1999-07-14	The scariest movie of all time is a true story.	In October of 1994 three student filmmakers di...	[witch, voodoo, legend, sorcery, maryland, for...	[Horror, Mystery]	81.0	0.060	248.000	4133.333333	massive success
4724	4724	Eraserhead	1977-03-19	Where your nightmares end...	Henry Spencer tries to survive his industrial ...	[baby, mutant, claustrophobia, nightmare, pare...	[Drama, Fantasy, Horror, Science Fiction]	89.0	0.010	7.000	700.000000	massive success

Figure 6: Les trois plus grands films à succès

Bizarrement ce sont tous des films d'horreur à très faible budget. Les gens aiment-ils tant les films d'horreur que ça ? Mais plus important : est-ce là la solution pour percer dans l'industrie du cinéma ????

Procédons à présent à l'analyse des données. En général, nous utiliserons df_1 pour l'analyse des variables qualitatives et df_2 pour l'analyse des variables quantitatives, même si cela n'influence que très peu les conclusions.

III. ANALYSE

1. Analyse des données qualitatives

a. Analyse du genre des films

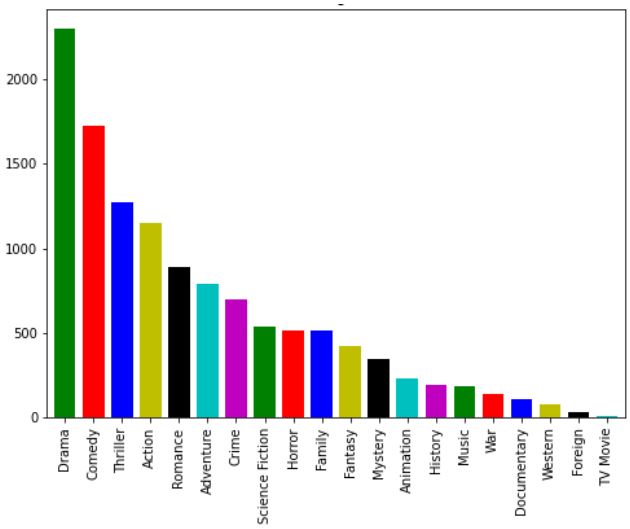

Figure 7: Classement des genres des films

Beaucoup de films se considèrent comme des drames. C'est normal car la définition d'un drame n'est pas du tout rigoureuse. À l'opposé personne ne veut faire des films destinés au petit écran. C'est aussi normal car la principale source de revenu du film (son box-office) est alors supprimée.

b. Analyse des compagnies de production

On connait tous les titans de l'industrie du cinéma (le "Big Five" : Universal, Paramount, Warner Bros., Walt Disney, et Columbia). Mais quelle sont ces compagnies qui investissent le plus sagement, faisant ainsi des bons films de façon consistante. Je soupçonne Pixar et Marvel ; il parait qu'ils ne font jamais de mauvais films.

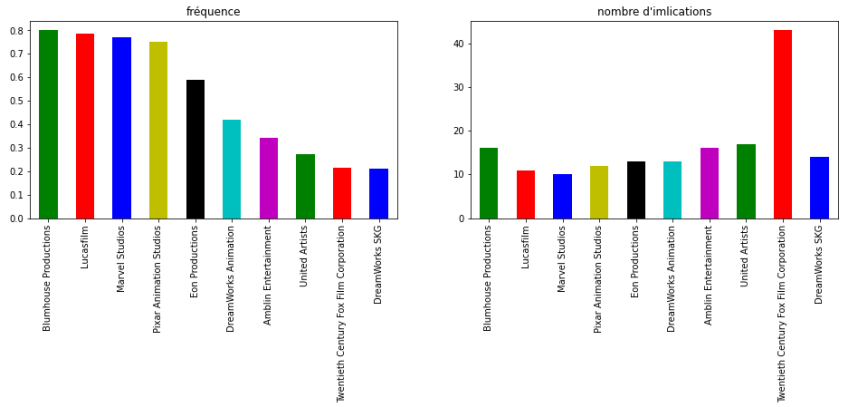

Figure 8: Fréquence et nombre d'implication dans les succès massif (dans le top 20)

Dans le top 20 des compagnies qui font le plus de films à succès massif :

- Pixar et ses films d'animation est 4ème en termes de fréquence (consistance) de production de ces magnifiques films ; L'univers Cinématique de Marvel est aussi bien classé en 3ème position. Mes suspicions (et celles du public) sont bien correctes sur ce point.
- Ça ne me surprend pas que Lucasfilm soit 2ème, vu que les films Star Wars "rapporteront toujours de l'argent", qu'ils restent bons ou pas.
- Je suis cependant très surpris par Blumhouse Productions. Je n'avais aucune idée qu'ils étaient aussi chirurgicaux. D'après leur portfolio, Blumhouse semble spécialisé dans les films d'horreur, ce qui confirme l'intuition que j'ai eu lorsque j'observais les 3 plus grand films à succès (figure 6). Est-ce que mon algorithme d'apprentissage réussira à capter leur recette sécrète ?

A présent, analysons les compagnies qui échouent très souvent.

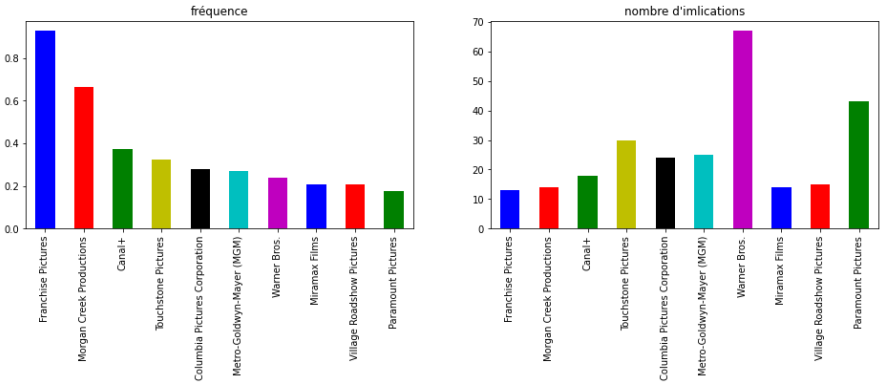

Figure 9: Fréquence et nombre d'implication dans les échecs (dans le top 15)

Je suis assez surpris de retrouver Warner Bros. et Canal+ dans cette ce top 15. C'est trop triste ! DC (et son parent Warner Bros.) a intérêt à redoubler d'efforts pour rattraper Marvel (et son parent Disney).

c. Analyse des langues des films

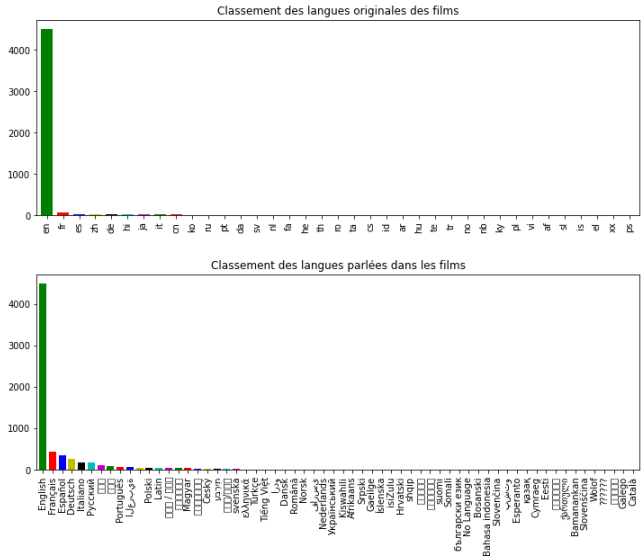

Figure 10: Classement des langues des films

La majorité des films dans notre data frame étant Hollywoodien, naturellement l'anglais domine. On y trouve quand même quelques films français (langue originale = fr). On rencontre encore plus de films qui emploient le français (langue parlée = Français), qu'ils soient made in France ou non. Ça confirme bien le fait que le français est une langue bien aimée des Américains.

d. Analyse des mots clés

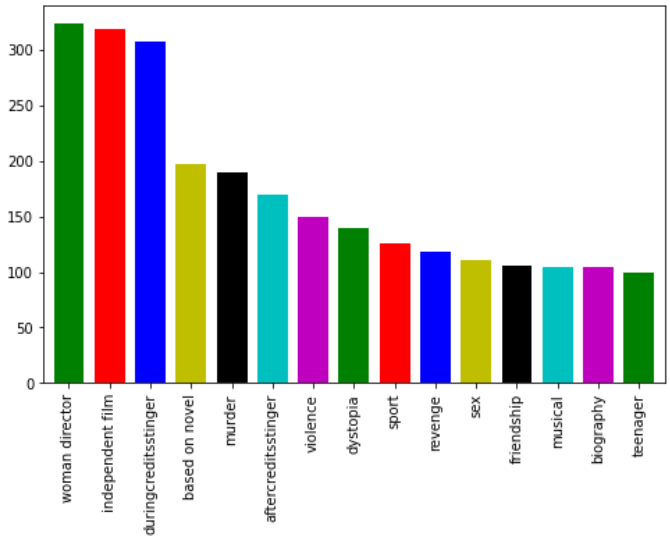

Figure 11: Classement des mots clés associés aux films

- Certains de ces mots clés étaient prévisibles. Par exemple "*during credits stinger*" et "*after credits stinger*". S'il y a une scène pendant ou après le générique de fin, il vaudrait mieux être averti avant d'aller au cinéma. Aussi, c'est normal d'avoir "*independent film*" dans cette liste. Les réalisateurs ont généralement plus de liberté sur ces films. On s'attend alors à une "vrai" expérience artistique.
- Par contre je suis un peu choqué par la mention "*woman director*" si fréquente. Il faut tout simplement croire que nos données ont un petit problème de sexisme. Après tout, certains de ces films remontent à la création du cinéma.

2. Analyse des données quantitatives

a. Matrice de corrélations

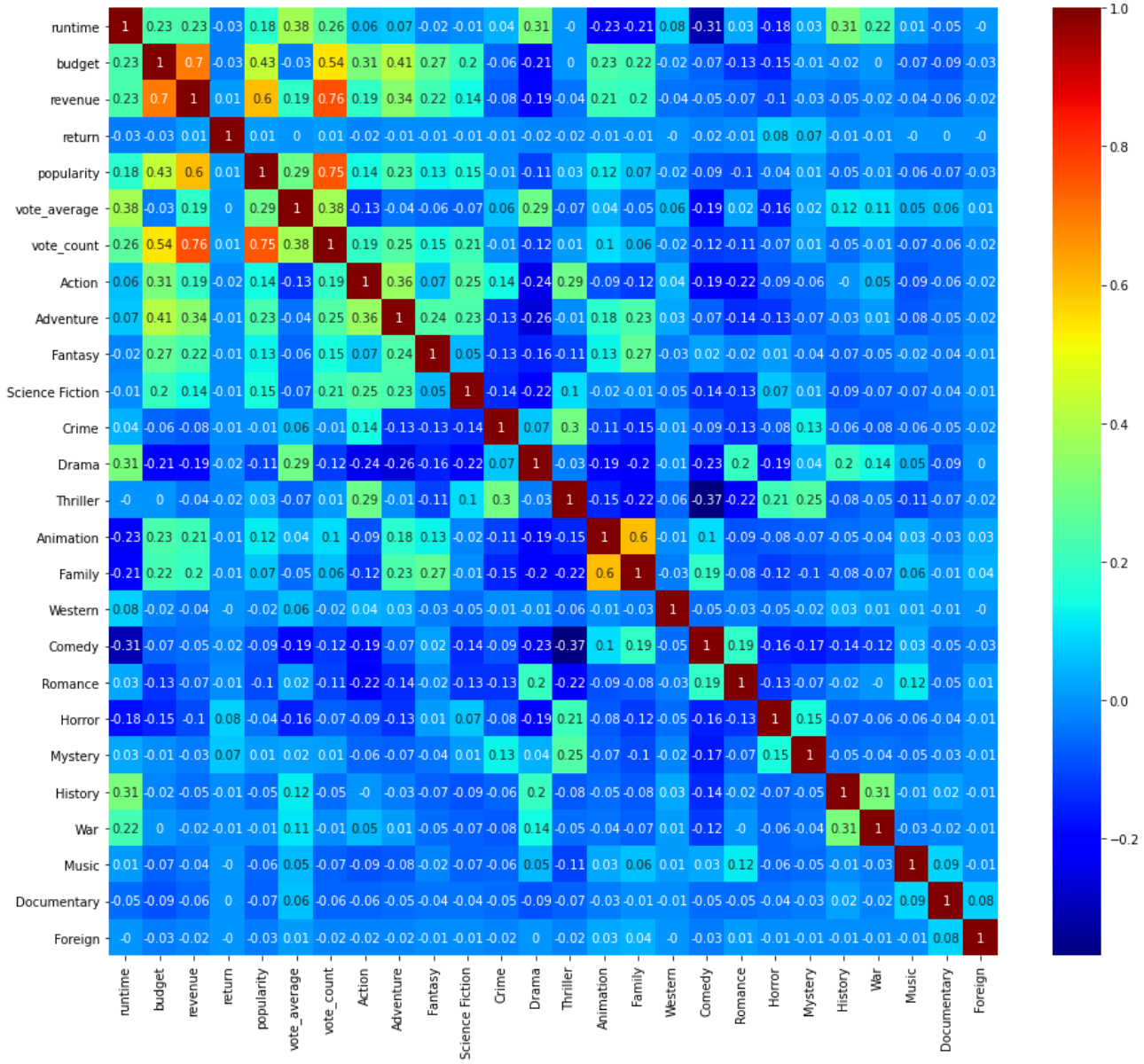

Figure 12: Matrice de corrélation

Concernant les corrélations les plus importantes. On voit sans surprise que :

- Les films à gros revenu ont tendance à avoir un gros budget et sont les plus votés.
- Ces films à grand nombre de votes deviennent naturellement les plus populaires. Ceci indique que, les votes (quand ils effectués) ont tendance à être favorables.
- Les films d'aventure ont tendance à contenir de l'action
- Les films d'histoire sont très souvent des films de guerre. C'est la nature belliqueuse de l'homme.

Mais à ma grande surprise :

- Les films d'animation sont très souvent les films de famille. De nos jours, les animés (surtout japonais) viennent sous toutes les formes. Je me disais que cela équilibrerait un peu les choses.

- Le budget est un meilleur indicateur du genre Adventure que du genre Action (tous deux ayant tendance à couter cher). C'est probablement dû au fait que les films d'action ne sont pas très bien définis, alors que les films d'aventure le sont.

Quant aux faibles corrélations, on remarque sans surprise que :

- Un film à suspense (Thriller) est très difficilement amusant !
- Les films d'animation et de comédie ont un temps réduit. C'est normal car les enfants ont une durée d'attention faible.

Cependant, l'information la plus importante confirmée par le tableau de corrélation est que les **return = revenue/budget** ne sont pas du tout corrélés avec le **budget**. L'industrie du cinéma présente des risques. Il ne suffit pas d'investir une grosse somme pour espérer en gagner par la suite.

b. Relation budget - revenu - popularité - votes - succès

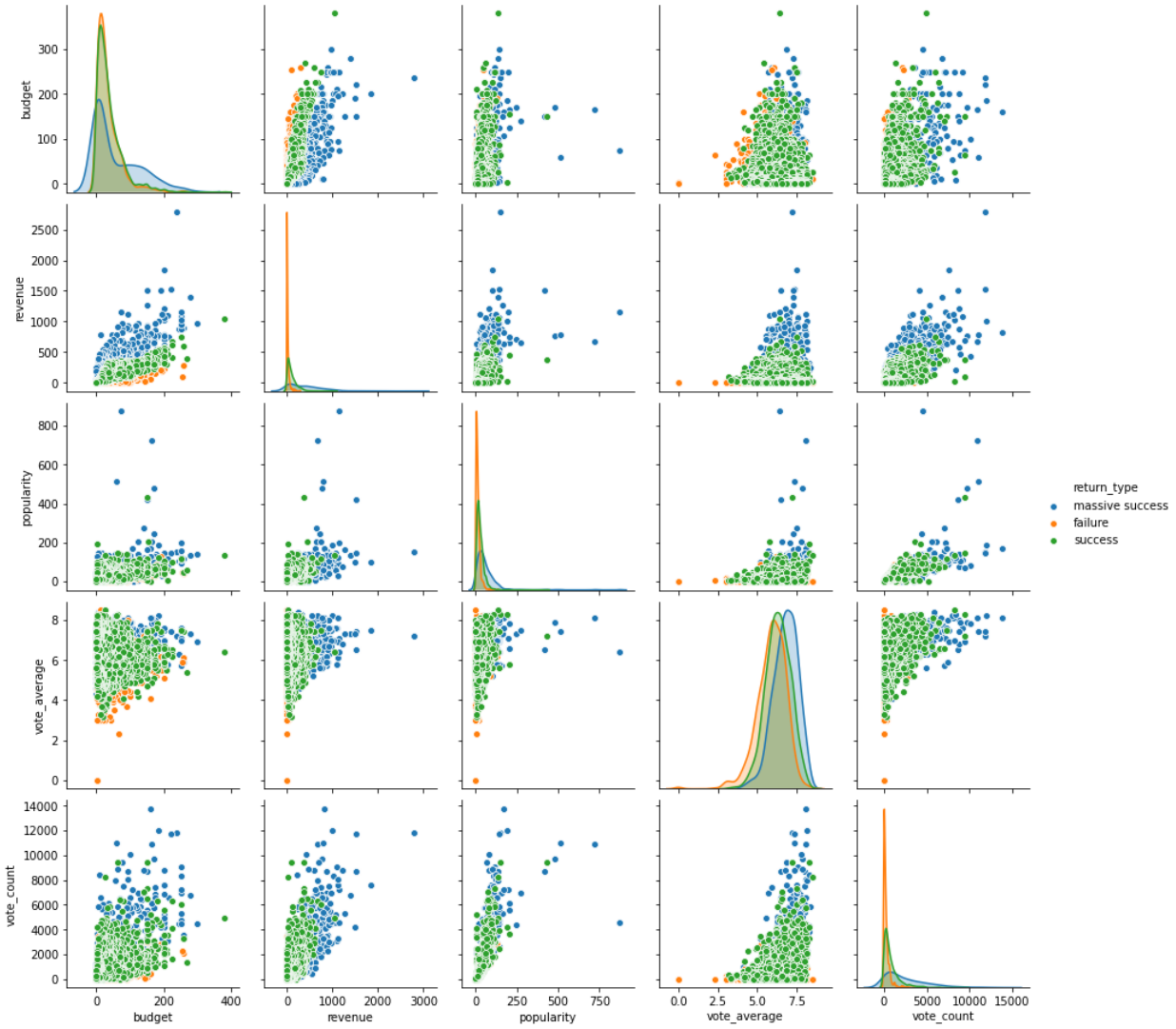

Figure 13: Pairplot pour les variables budget - revenu - popularité - moyenne des votes - nombre de votes, et colorié par les types de retour

Cette figure confirme bien que :

- Le revenu, la popularité, et le nombre de votes évoluent avec les budgets des films. (colonne 1).
- En ce qui concerne la dépendance par rapport au revenu (colonne 2), on voit que la popularité et le nombre de vote croient avec le revenu. Tandis que la moyenne des votes a tendance à rester autour de 7. Autrement dit, même les films qui ont perdu de l'argent sont bien votés.
- La popularité est un bon prédicteur du nombre de votes (colonne 3).
- Il n'y a de dépendance concrète (d'aucune des autres variables) en fonction de la moyenne des votes (colonne 4).
- Quand le nombre de votes est fort, le budget, le revenu, la popularité, et surtout la moyenne des votes ne peuvent être qu'élevés (colonne 5). Ça traduit le fait qu'on risque toujours d'avoir des films nuls qui sont mis en avant par un groupe de personnes très (très) motivé (en votant plusieurs fois par exemple).

On observe les couleurs sur la figure 13. Elles indiquent le type du retour : succès massif (bleu), échec (orange), ou succès (vert).

- Les films à succès massifs n'ont pas forcément un gros budget.
- Un indicateur pour les films à succès massifs une **moyenne des votes relativement élevée** (courbe "vote_average" - "vote_average"). Ça se confirme dans la réalité : on paye les critiques pour bien noter nos films ; et puis dénigrer les films de la compétition (Marvel vs. DC). Malheureusement pour nous, la moyenne des votes est connue après la sortie du film. Nous ne pouvons donc pas l'utiliser pour la prédiction du succès.

Une fois cette analyse faite, on sauvegarde nos data frames (au format CSV) et les dictionnaires (sous forme de module Python 3). Nous passons à présent à l'apprentissage (voir notebook numéro 2).

IV. APPRENTISSAGE

1. Prédiction du genre

On désire prédire le(s) genre(s) d'un film à partir de son intrigue.

a. Préparation des données

Pour obtenir les donnes nécessaires à l'apprentissage, transformons les intrigues, slogans, et mots clés (chaines de caractères) en liste d'entiers. Utilisons le dictionnaire imdb de Keras pour les mots simples, et nos propres dictionnaires pour les groupes de mots et les noms propres.

Version originale:

	overview	genres
0	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]
1	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]
2	A cryptic message from Bond's past sends him o...	[Action, Adventure, Crime]
3	Following the death of District Attorney Harve...	[Action, Crime, Drama, Thriller]
4	John Carter is a war-weary, former military ca...	[Action, Adventure, Science Fiction]

Version numerisee:

	overview	genres
0	[1, 10, 3, 50823, 1116, 5, 2, 7734, 8, 15876, ...	[1, 2, 3, 4]
1	[1, 1704, 2, 195, 2416, 7, 29, 350, 46, 215, 1...	[2, 3, 1]
2	[1, 5, 14257, 748, 38, 2, 500, 3291, 89, 22, 5...	[1, 2, 5]
3	[1, 1044, 3, 340, 6, 7794, 4816, 4354, 23376, ...	[1, 5, 6, 7]
4	[1, 307, 3513, 8, 5, 2, 1137, 1247, 1704, 870,...	[1, 2, 4]

Figure 14: Version originale et version numérisée des 5 premières entrées et sorties

On utilise ensuite la méthode d'encodage **one_hot_encoding** pour vectoriser nos données. Un exemple est donné ci-dessous (il s'agit du chef-d'œuvre Avatar de James Cameron).

```
input : [0 1 1 ... 0 0 0]                          In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission,
output: [1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  ['Action', 'Adventure', 'Fantasy', 'Science Fiction']
```

Figure 15: Encodage d'un input et son output associé

On divise ensuite nos données en train, val et test. On obtient des tenseurs de formes ci-dessous :

```
x shapes: (3000, 88586) (1000, 88586) (800, 88586)
y shapes: (3000, 20) (1000, 20) (800, 20)
```

Figure 16: Forme des tenseurs train, val, et test

b. Réseau de neurones

On utilise un réseau de neurones (de la librairie Keras) à deux couches complètement connectées avec une fonction d'activation *relu*.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 256)               22678272
_________________________________________________________________
dense_2 (Dense)              (None, 64)                16448
_________________________________________________________________
dense_3 (Dense)              (None, 20)                1300
=================================================================
Total params: 22,696,020
Trainable params: 22,696,020
Non-trainable params: 0
_________________________________________________________________

Infos supplémentaires
---------------------------
layer: 0
weights shape: (88586, 256)
bias shape: (256,)
---------------------------
layer: 1
weights shape: (256, 64)
bias shape: (64,)
---------------------------
layer: 2
weights shape: (64, 20)
bias shape: (20,)
```

Figure 17: Réseau de neurones utilisé

Pour la compilation :

- On utilise l'optimiseur Adam
- On utilise la "binary_crossentropy" pour fonction loss
- On observe l'accuracy

On lance ensuite l'apprentissage sur 15 époques par paquets de 512. On opte pour la méthode d'**early stopping** pour lutter contre le surapprentissage.

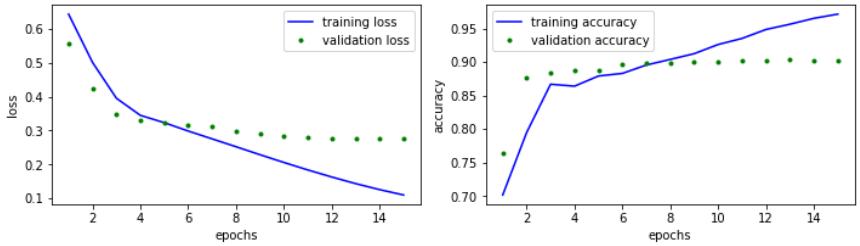

Figure 18: Loss et accuracy sur les données train et val

Observons quelques prédictions :

```
Beer League (2006)
original:   [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']
prediction: [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']

The Lives of Others (2006)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Drama', 'Thriller']
prediction: [0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Drama']

Buried (2010)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0]  -  ['Drama', 'Thriller', 'Mystery']
prediction: [1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Thriller']

Road Hard (2015)
original:   [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']
prediction: [0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Drama', 'Comedy']

Sex With Strangers (2002)
original:   [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]  -  ['Documentary']
prediction: [0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Comedy', 'Romance']

In Her Line of Fire (2006)
original:   [1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Drama', 'Thriller']
prediction: [1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action']

The Sisterhood of Night (2015)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0]  -  ['Drama', 'Thriller', 'Mystery']
prediction: [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  []

The Toxic Avenger (1984)
original:   [1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0]  -  ['Action', 'Science Fiction', 'Comedy', 'Horror']
prediction: [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']

Like Crazy (2011)
original:   [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']
prediction: [0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Drama']

The Gallows (2015)
original:   [0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0]  -  ['Thriller', 'Horror']
prediction: [0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Drama']
```

Figure 19: Prédictions avec le réseau de neurones

On constate que les prédictions sont assez bonnes. Mais le model a l'air d'avoir peur de se tromper, et ne donne très souvent aucune prédiction. La figure 20 confirme bien cela. On résoudra ce problème en ajustant le seuillage dans la suite.

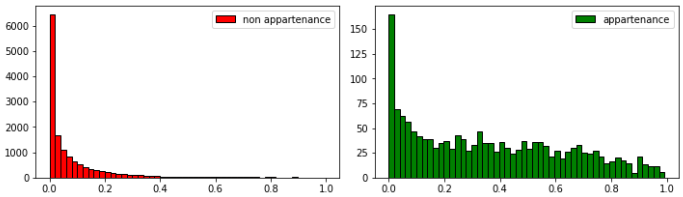

Figure 20: Fiabilité des prédictions

Le score obtenu est de 90% ce qui a priori n'est pas mal. Néanmoins, il faut prendre en compte le fait qu'on est intéressé par l'appartenance d'un film a un genre de film (label = 1), et non le contraire (label=0).

On veut à présent trouver un seuil meilleur que le 0.5 utilisé jusqu'à présent. Pour cela, calculons la précision et le rappel. On décrète que la classe positive c'est la classe des 1 (pour chacun des 20 genres possibles). C'est de toute évidence la classe minoritaire.

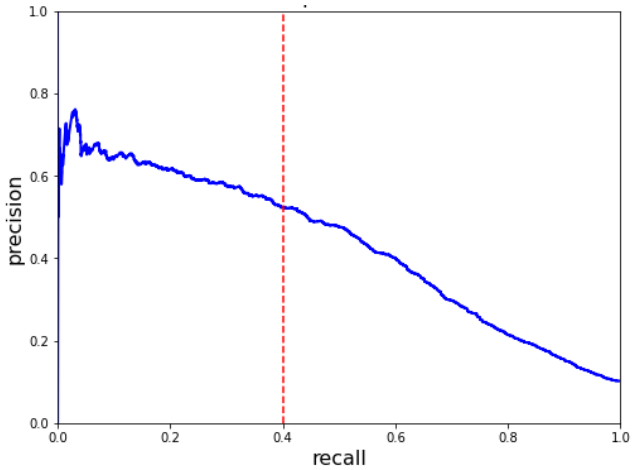

Figure 21: Courbe précision/rappel pour effecteur un seuillage tel que rappel=0.4

Nous voulons obtenir un fort rappel. Car ce n'est pas grave si notre réseau donne quelques mauvaises prédictions sur le genre, du moment qu'on a au moins une catégorie dans laquelle classer le film. Cependant, la figure 21 montre que la dépendance précision/rappel est loin d'être idéale. Un compromis acceptable entre bonne précision et bon rappel peut être pris tel que rappel=0.4.

Ceci fait, observons les prédictions obtenues avec le nouveau seuil calculé (égale à 0.39) :

```
Beer League (2006)
original:   [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']
prediction: [0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Drama', 'Comedy']

The Lives of Others (2006)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Drama', 'Thriller']
prediction: [1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Drama']

Buried (2010)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0]  -  ['Drama', 'Thriller', 'Mystery']
prediction: [1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Thriller']

Road Hard (2015)
original:   [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']
prediction: [0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Drama', 'Comedy']

Sex With Strangers (2002)
original:   [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]  -  ['Documentary']
prediction: [0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Comedy', 'Romance']

In Her Line of Fire (2006)
original:   [1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Drama', 'Thriller']
prediction: [1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action']

The Sisterhood of Night (2015)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0]  -  ['Drama', 'Thriller', 'Mystery']
prediction: [0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Thriller']

The Toxic Avenger (1984)
original:   [1 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0]  -  ['Action', 'Science Fiction', 'Comedy', 'Horror']
prediction: [0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0]  -  ['Adventure', 'Family', 'Comedy']

Like Crazy (2011)
original:   [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']
prediction: [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']

The Gallows (2015)
original:   [0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0]  -  ['Thriller', 'Horror']
prediction: [0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Drama']
```

Figure 22: Prédictions obtenue après seuillage pour le réseau de neurones

Cette fois, les prédictions sont bien meilleures que dernièrement (figure 19). Même quand le model se trompe, la catégorie qu'il indique n'est pas très éloignée de la vraie catégorie, comme nous pouvons l'observer sur la matrice de corrélations (figure 12).

c. Arbre de décision

Vu qu'il s'agit d'une classification multi-label, nous devons utiliser un estimateur qui supporte l'approche **one_vs_all**. On se tourne naturellement vers les arbres de décisions.

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=42, splitter='best')
```

Figure 23: Arbre de décisions utilisé pour la prédiction des genres

On obtient les prédictions suivantes :

```
Beer League (2006)
original:   [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']
prediction: [0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Comedy', 'Romance']

The Lives of Others (2006)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Drama', 'Thriller']
prediction: [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']

Buried (2010)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0]  -  ['Drama', 'Thriller', 'Mystery']
prediction: [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']

Road Hard (2015)
original:   [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]  -  ['Comedy']
prediction: [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']

Sex With Strangers (2002)
original:   [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]  -  ['Documentary']
prediction: [0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Adventure', 'Animation', 'Romance']

In Her Line of Fire (2006)
original:   [1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Drama', 'Thriller']
prediction: [1 0 0 0 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0]  -  ['Action', 'Crime', 'Drama', 'Thriller', 'Mystery']

The Sisterhood of Night (2015)
original:   [0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0]  -  ['Drama', 'Thriller', 'Mystery']
prediction: [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']

The Toxic Avenger (1984)
original:   [1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0]  -  ['Action', 'Science Fiction', 'Comedy', 'Horror']
prediction: [1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Fantasy', 'Thriller']

Like Crazy (2011)
original:   [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']
prediction: [0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]  -  ['Drama', 'Romance']

The Gallows (2015)
original:   [0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0]  -  ['Thriller', 'Horror']
prediction: [1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]  -  ['Action', 'Science Fiction', 'Thriller']
```

Figure 24: Prédictions obtenues pour l'arbre de décisions

Les prédictions ne sont pas vraiment meilleures que les précédentes (figure 22). En l'occurrence, le model semble fixé sur la dualité 'Drama' - 'Romance'. C'est quand même vrai que les romances demandent beaucoup de drames, assez pour en décourager certains ? :)

Le calcul du score F1 (égal à 0.32) confirme bien que ce modèle est moins performant que le réseau de neurones, qui avait un F1 score de 0.45.

2. Prédiction du succès

On désire prédire le succès/échec d'un film ("return_type") en fonction du **budget** qui est investi, de la **durée** du film, des **genres** de ce film, des **langues** qui y figurent, des **compagnies de production** qu'on embauche pour le produire, et des **mots clés** qui caractérisent le film.

a. Préparation des données

On commence par numériser nos données.

	budget	runtime	genres	spoken_languages	production_companies	keywords	return_type
0	237.0	162.0	[Action, Adventure, Fantasy, Science Fiction]	[English, Español]	[Ingenious Film Partners, Twentieth Century Fo...	[culture clash, future, space war, space colon...	massive success
1	300.0	169.0	[Adventure, Fantasy, Action]	[English]	[Walt Disney Pictures, Jerry Bruckheimer Films...	[ocean, drug abuse, exotic island, east india ...	massive success
2	245.0	148.0	[Action, Adventure, Crime]	[Français, English, Español, Italiano, Deutsch]	[Columbia Pictures, Danjaq, B24]	[spy, based on novel, secret agent, sequel, mi...	massive success
3	250.0	165.0	[Action, Crime, Drama, Thriller]	[English]	[Legendary Pictures, Warner Bros., DC Entertai...	[dc comics, crime fighter, terrorist, secret i...	massive success
4	260.0	132.0	[Action, Adventure, Science Fiction]	[English]	[Walt Disney Pictures]	[based on novel, mars, medallion, space travel...	failure

	budget	runtime	genres	spoken_languages	production_companies	keywords	return_type
0	237.0	162.0	[1, 2, 3, 4]	[1, 2]	[1, 2, 3, 4]	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	massive success
1	300.0	169.0	[2, 3, 1]	[1]	[5, 6, 7]	[22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 3...	massive success
2	245.0	148.0	[1, 2, 5]	[3, 1, 2, 4, 5]	[8, 9, 10]	[38, 39, 40, 41, 42, 43, 44]	massive success
3	250.0	165.0	[1, 5, 6, 7]	[1]	[11, 12, 13, 14]	[45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 5...	massive success
4	260.0	132.0	[1, 2, 4]	[1]	[5]	[39, 66, 67, 6, 68, 10, 69, 70, 71, 72, 73, 74...	failure

Figure 25: Versions originales et numérisée des entrées et sorties pour la prediction du succes

On génère ensuite nos données train et test. Ici, nous n'aurons pas besoins de données de validation, vu qu'on effectuera (implicitement) une cross-validation.

Un vecteur d'entrée $\boldsymbol{x_i}$ est juste la concaténation de :

- La forme normalisée du budget $\frac{\boldsymbol{budget\ du\ film\ i}}{\boldsymbol{maximum\ des\ budgets}}$ (de longueur 1)
- La forme normalisée de la durée (de longueur 1)
- La forme vectorisée des genres (de longueur 20)
- La forme vectorisée des langues (de longueur 62)
- La forme vectorisée des compagnies de production (de longueur 5017)
- La forme vectorisée des mots clés (de longueur 9813)

Ce qui nous donne une longueur totale de 14914.

Un scalaire $\boldsymbol{y_i}$ vaut :

- 0 pour la classe "failure"
- 1 pour la classe "success"
- 2 pour la classe "massive success"

Visualisons un exemple (il s'agit du film Avatar (2009)) :

```
input : [0.62 0.48 0.   ... 0.   0.   0.  ]

output: 2
```

Figure 26: Visualisation d'un vecteur d'entrée et son label associé

b. Réseau de neurones

Le modèle est décrit ci-dessous :

```
MLPClassifier(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
              beta_2=0.999, early_stopping=True, epsilon=1e-08,
              hidden_layer_sizes=(100, 10), learning_rate='constant',
              learning_rate_init=0.001, max_fun=15000, max_iter=200,
              momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
              power_t=0.5, random_state=None, shuffle=True, solver='adam',
              tol=0.0001, validation_fraction=0.1, verbose=False,
              warm_start=False)
```

Figure 27: Réseau de neurones pour la prédiction du succès

On obtient les prédictions suivantes :

	title	release_date	return_type	prediction
2800	The Visit	2015-09-10	massive success	success
2841	The Last Exorcism Part II	2013-02-28	success	success
2882	The Nun's Story	1959-06-18	success	success
2923	The Apartment	1960-06-15	success	success
2964	Timecrimes	2007-09-20	failure	success
3005	Thirteen	2003-08-20	success	success
3046	From Here to Eternity	1953-08-04	massive success	failure
3087	From a Whisper to a Scream	1987-09-25	success	success
3128	Ruby in Paradise	1993-10-08	success	failure
3169	Tupac: Resurrection	2003-01-23	massive success	success
3210	El Mariachi	1992-09-04	success	success

Figure 28: Prédiction pour le réseau de neurones

On obtient un score d'accuracy de 0.4842, et sa matrice de confusion est données ci-dessous :

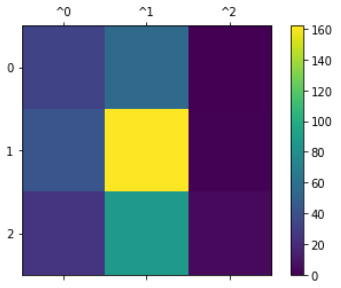

Figure 29: Matrice de confusion pour le réseau de neurones

On voit que le modèle prédit bien les succès, probablement parce qu'il y a beaucoup plus de films labélisées "success" dans nos données.

c. Régression logistique

Le modèle est décrit ci-dessous :

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=1000,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=42, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

Figure 30: Régression logistique pour la prédiction du succès

On obtient les prédictions suivantes :

	title	release_date	return_type	prediction
2800	The Visit	2015-09-10	massive success	success
2841	The Last Exorcism Part II	2013-02-28	success	success
2882	The Nun's Story	1959-06-18	success	success
2923	The Apartment	1960-06-15	success	success
2964	Timecrimes	2007-09-20	failure	failure
3005	Thirteen	2003-08-20	success	success
3046	From Here to Eternity	1953-08-04	massive success	failure
3087	From a Whisper to a Scream	1987-09-25	success	success
3128	Ruby in Paradise	1993-10-08	success	failure
3169	Tupac: Resurrection	2003-01-23	massive success	success
3210	El Mariachi	1992-09-04	success	success

Figure 31: Prédiction pour la régression logistique

On obtient un score d'accuracy de 0.4720, et sa matrice de confusion est donnée ci-dessous :

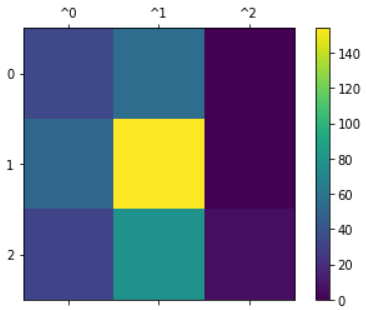

Figure 32: Matrice de confusion pour la régression logistique

d. Forêt aléatoire

Le modèle est décrit ci-dessous :

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=500,
                       n_jobs=-1, oob_score=False, random_state=42, verbose=0,
                       warm_start=False)
```

Figure 33: Forêt aléatoire pour la prédiction du succès

On obtient les prédictions suivantes :

	title	release_date	return_type	prediction
2800	The Visit	2015-09-10	massive success	success
2841	The Last Exorcism Part II	2013-02-28	success	success
2882	The Nun's Story	1959-06-18	success	success
2923	The Apartment	1960-06-15	success	success
2964	Timecrimes	2007-09-20	failure	failure
3005	Thirteen	2003-08-20	success	success
3046	From Here to Eternity	1953-08-04	massive success	success
3087	From a Whisper to a Scream	1987-09-25	success	success
3128	Ruby in Paradise	1993-10-08	success	failure
3169	Tupac: Resurrection	2003-01-23	massive success	success
3210	El Mariachi	1992-09-04	success	success

Figure 34: Prédiction pour la forêt aléatoire

On obtient un score d'accuracy de 0.5231, et sa matrice de confusion est donnée ci-dessous :

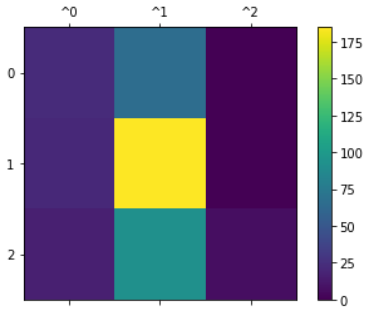

Figure 35: Matrice de confusion pour la forêt aléatoire

Il s'agit-là de notre meilleur score jusqu'à présent. Est-ce qu'on peut faire mieux ?

e. Ensemble learning

On combine les trois modèles précédents dans l'espoir de former un qui soit meilleur.

```
VotingClassifier(estimators=[('nn',
                              MLPClassifier(activation='relu', alpha=0.0001,
                                            batch_size='auto', beta_1=0.9,
                                            beta_2=0.999, early_stopping=True,
                                            epsilon=1e-08,
                                            hidden_layer_sizes=(100, 10),
                                            learning_rate='constant',
                                            learning_rate_init=0.001,
                                            max_fun=15000, max_iter=200,
                                            momentum=0.9, n_iter_no_change=10,
                                            nesterovs_momentum=True,
                                            power_t=0.5, random_state=None,
                                            sh...
                                                    criterion='gini',
                                                    max_depth=None,
                                                    max_features='auto',
                                                    max_leaf_nodes=None,
                                                    max_samples=None,
                                                    min_impurity_decrease=0.0,
                                                    min_impurity_split=None,
                                                    min_samples_leaf=1,
                                                    min_samples_split=2,
                                                    min_weight_fraction_leaf=0.0,
                                                    n_estimators=500,
                                                    n_jobs=-1, oob_score=False,
                                                    random_state=42, verbose=0,
                                                    warm_start=False))],
                  flatten_transform=True, n_jobs=None, voting='soft',
                  weights=None)
```

Figure 36: Soft-voting pour la prédiction du succès

On obtient les prédictions suivantes :

	title	release_date	return_type	prediction
2800	The Visit	2015-09-10	massive success	success
2841	The Last Exorcism Part II	2013-02-28	success	success
2882	The Nun's Story	1959-06-18	success	success
2923	The Apartment	1960-06-15	success	success
2964	Timecrimes	2007-09-20	failure	failure
3005	Thirteen	2003-08-20	success	success
3046	From Here to Eternity	1953-08-04	massive success	failure
3087	From a Whisper to a Scream	1987-09-25	success	success
3128	Ruby in Paradise	1993-10-08	success	failure
3169	Tupac: Resurrection	2003-01-23	massive success	success
3210	El Mariachi	1992-09-04	success	success

Figure 37: Prédiction pour l'ensemble method

On obtient un score d'accuracy de 0.5036, et sa matrice de confusion est donnée ci-dessous :

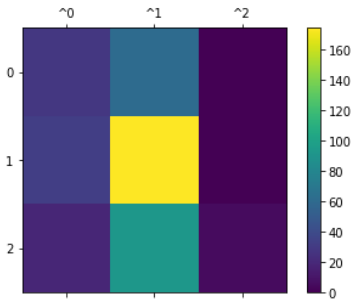

Figure 38: matrice de confusion pour l'ensemble method

Pas de chance, le modèle est moins performant que le meilleur de ses composants (la foret aléatoire).

f. Bagging

Nous avons à notre disposition très peu de données (3200 éléments environ), on fait du Bagging pour remédier à cela. La forêt aléatoire nous a donné le meilleur résultat jusqu'à présent ; utilisons-la ici :

```
BaggingClassifier(base_estimator=RandomForestClassifier(bootstrap=True,
                                                        ccp_alpha=0.0,
                                                        class_weight=None,
                                                        criterion='gini',
                                                        max_depth=None,
                                                        max_features='auto',
                                                        max_leaf_nodes=None,
                                                        max_samples=None,
                                                        min_impurity_decrease=0.0,
                                                        min_impurity_split=None,
                                                        min_samples_leaf=1,
                                                        min_samples_split=2,
                                                        min_weight_fraction_leaf=0.0,
                                                        n_estimators=500,
                                                        n_jobs=-1,
                                                        oob_score=False,
                                                        random_state=42,
                                                        verbose=0,
                                                        warm_start=False),
                  bootstrap=True, bootstrap_features=False, max_features=1.0,
                  max_samples=10, n_estimators=100, n_jobs=-1, oob_score=False,
                  random_state=42, verbose=0, warm_start=False)
```

Figure 39: Bagging pour la prédiction du succès

On obtient les prédictions suivantes :

	title	release_date	return_type	prediction
2800	The Visit	2015-09-10	massive success	success
2841	The Last Exorcism Part II	2013-02-28	success	success
2882	The Nun's Story	1959-06-18	success	success
2923	The Apartment	1960-06-15	success	success
2964	Timecrimes	2007-09-20	failure	success
3005	Thirteen	2003-08-20	success	success
3046	From Here to Eternity	1953-08-04	massive success	success
3087	From a Whisper to a Scream	1987-09-25	success	success
3128	Ruby in Paradise	1993-10-08	success	success
3169	Tupac: Resurrection	2003-01-23	massive success	success
3210	El Mariachi	1992-09-04	success	success

Figure 40: Prédiction pour le bagging classifier

On obtient un score d'accuracy de 0.5012, et sa matrice de confusion est donnée ci-dessous :

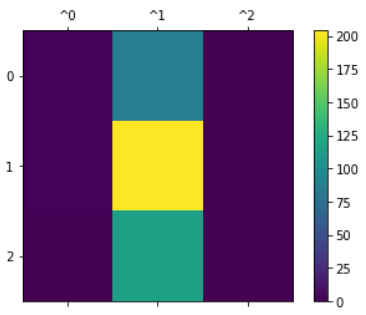

Figure 41: Matrice de confusion pour le bagging classifier

C'est avec tristesse qu'on remarque qu'il reste moins performant que la forêt qui le compose.

g. Conclusions sur l'étude du succès

Les scores obtenus sont recapitulés ci-dessous.

Modèle	Réseau de neurones	Régression Logistique	Foret Aléatoire	Ensemble Learning	Bagging Classifier
Accuracy	0.4842	0.4720	0.5231	0.5036	0.5012
F1 score pondéré	0.4117	0.4119	0.4362	0.4228	0.3347

Table 1: Comparaison des scores

En conclusion, on a du mal à prédire le succès avec plus de 50% d'exactitude (ce qui représente à mon avis le minimum de crédibilité). D'une part je suis fautif car je suis un peu stricte dans la catégorisation des succès/échecs. Par exemple, un succès massif prédit comme un simple succès est tout à fait acceptable. J'encouragerais avec enthousiasme la production d'un tel film.

D'autre part mon jeu de données n'est pas parfait :

- On a très peu de données : 3200 films, c'est assez petit pour se faire une idée des goûts cinématographiques de l'humanité.
- Ces données ne sont pas assez diversifiées. Il y a beaucoup trop de succès pour très peu d'échecs. Peut-être que, dans quelques années, une métrique universelle de définition du succès sera créée. En plus on aura plus de films à étudier. Ça sera potentiellement plus facile de conduire une telle étude.
- Le succès d'un film dépend énormément des stars à l'affiche. Ça aurait donc été intéressant d'étudier le jeu de données contenant l'équipe de tournage.

Une autre explication est que ces faibles scores sont juste naturels. La nature contrôle tout et je n'y peux rien. Si la prédiction du succès d'un film était facile (voire possible) alors tout le monde rentrerait dans l'industrie du cinéma pour se faire riche ; et il n'y aurait plus de data scientist. Je ne veux pas ça :) !!

V. PERSPECTIVES

Nous avons prédit les genres des films, et leur succès grâce à des données toutes connues avant la sortie du film. Les résultats obtenus pour la prédiction des genres sont encourageants. Le réseau de neurones construit peut être adapté pour la création d'un système de recommandation de films.

En ce qui concerne la prédiction du succès, nous avons obtenu moins de réussite. Nous aurions pu faire une analyse en composante principale pour déterminer quelles variables contribuent le plus au succès. Cela aurait probablement améliorer nos résultats. Nous aurions pu ajouter à tout cela la prédiction de la note moyenne du film, et/ou la somme d'argent que celui-ci rapportera. Cependant, nous avons déjà une bonne idée de la relation entre nos prédictions et ces dernières grâce à la matrice de corrélations de la figure 12.