Malis Project Definition : Success Movie

Clément Bernard and Maxence Brugeres

January 19, 2020

Context

It is hard for movie industry to know whether a movie will have success or not. Nevertheless, it seems that some companies have found a way to attract spectators. But there are still cases where high budget films are flopping and vice versa. The public has a lot of different interests, and some films succeeded whereas no one would have predicted that. Thus, some actors or companies seem to be enjoyed a lot, more than others. In this case, are there rules that can be made to say if a movie will be accepted to its public ?

Purpose

The project is aiming several goals. The first one is to determine what are the parameters that influence the most the popularity of a movie. How the gender, duration, cost or cast of a movie are important ? We are aiming to give conclusion about the efficiency of these parameters.

The second one is to know how successful a movie will be, given the most relevant parameters we have found previously.

If time allows us, we can refine our predictions by :

1. Splitting the dataset according to the gender of the film. Then we can create several models for each gender of film and try to compare them. So that it could improve the accuracy of our model if the relevant parameters which impact the film popularity differ from a gender to another.
2. Adding new data such as the number of oscars the cast of a film have already won. It will require to get another dataset (from Kaggle for example). Have they already been nominated to win an oscar ? Has the realisator already won an oscar ? We can also try to determine the cast's popularity and how it affect the popularity of the film. To do so we may add data link to social networks, for instance the number of facebook likes of the actors of a given movie.

Methodology

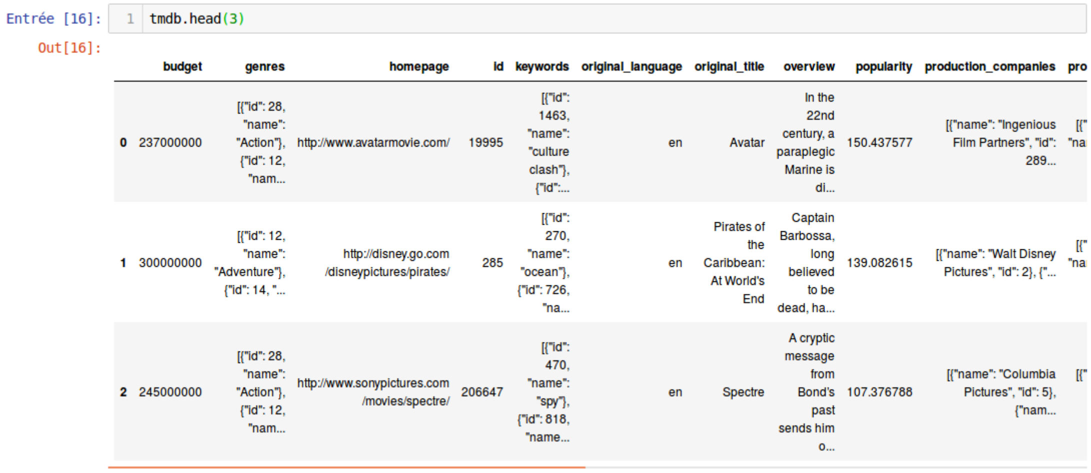
First of all, we need to pre-process our data. There are a lot of literal or incomplete values. Some data need to be modified (like 1-of-K method in order to highlight more the impact of a feature). We also need to convert some data in a JSON format into classic columns in order to uniformize every feature.

After that, we should split our dataset into a training and a test set. Our final accuracy will be done with the test set.
Then, we need to choose a good classifier. We will test several classifiers until we get the best accuracy.
Finally, we'll be able to draw conclusions on our features and the classifiers used.

Dataset

We will use a dataset from Kaggle. This dataset comes from TMDb which is one of the most popular source for movie contents. It contains around 5000 different movies and 30 features. These features are divided into 2 csv files. One for the general information about the movie (budget, gender, language, title, . . .) and the other one for more details (like the casting).

Some of the features contain JSON format.



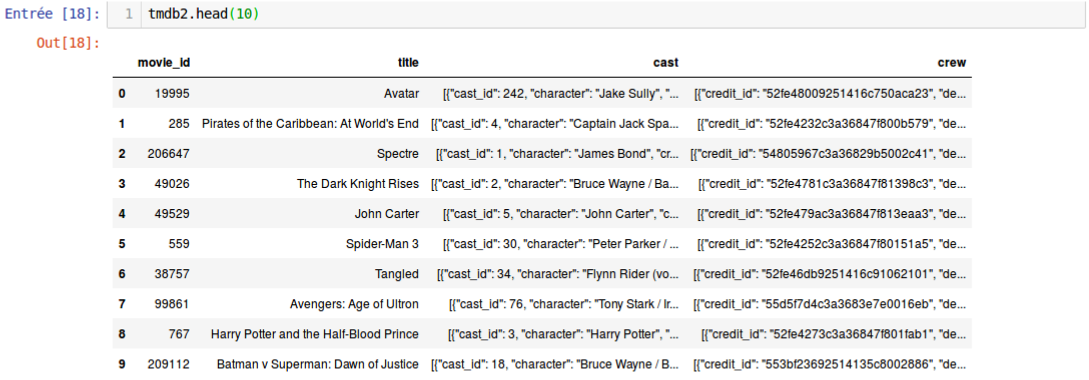
Entrée [16]:

```
tmdb.head(3)
```

Out[16]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	pro
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...	[{" "na
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}, {"...	[{" "na
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...	[{"

Figure 1: Head of the dataset from the first .csv file



Entrée [18]:

```
tmdb2.head(10)
```

Out[18]:

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...
5	559	Spider-Man 3	[{"cast_id": 30, "character": "Peter Parker / ...	[{"credit_id": "52fe4252c3a36847f80151a5", "de...
6	38757	Tangled	[{"cast_id": 34, "character": "Flynn Rider (vo...	[{"credit_id": "52fe46db9251416c91062101", "de...
7	99861	Avengers: Age of Ultron	[{"cast_id": 76, "character": "Tony Stark / Ir...	[{"credit_id": "55d5f7d4c3a3683e7e0016eb", "de...
8	767	Harry Potter and the Half-Blood Prince	[{"cast_id": 3, "character": "Harry Potter", "...	[{"credit_id": "52fe4273c3a36847f801fab1", "de...
9	209112	Batman v Superman: Dawn of Justice	[{"cast_id": 18, "character": "Bruce Wayne / B...	[{"credit_id": "553bf23692514135c8002886", "de...

Figure 2: Head of the dataset from the second .csv file

Movies	Real values	LR	Classification (20)	Classification (100)
Transformers: Age of Extinction	5.620250	5.655568	(5.5, 6)	6.3
The Hobbit: The Desolation of Smaug	7.451325	7.452443	(7, 7.5)	7.2
Apollo 13	7.049922	6.933888	(6.5, 7)	6.7
Interstellar	8.004072	7.415334	(7, 7.5)	7.3

Movies	Real values	Drama regression	Drama classification (20)	Drama classification (10)
Apollo 13	7.049922	6.576370	(6, 6.5)	6.1
Interstellar	8.004072	7.110654	(6, 6.5)	6.2