# REPORT

GROUP PROJECT
IMDB DATA SET ANALYSIS AND MOVIE RATING PREDICTION

PROGRAMMING FOR DATA SCIENCE

*MASTERS DATA ANALYTICS FOR BUSINESS*

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO – UNIVERSIDADE DE LISBOA

André Viana – l54543
*andreviana@aln.iseg.ulisboa.pt*

Gonçalo Duarte – l48505
*goncaloduarte@aln.iseg.ulisboa.pt*

José Cabral – l54997
*l54997@aln.iseg.ulisboa.pt*


*May 2021*

**Table of Contents**

## Abstract

The entertainment industry is one of the most powerful and impactful industries in the world. We all consume movies, tv shows, docuseries… all of us are contributors to the global success of the industry.

As big as it is, it also creates different feelings and opinions on the public. Some people love a certain movie, while others may dislike it. It is what it is, we all have different tastes, we all have our preferences and that is why the industry has to keep reinvent itself. Actors and actresses need to constantly push themselves to be more versatile, writers and producers have to study ways of improving the quality of their movies… it is a process in constant development. Nowadays, people are more demanding than ever. With the insane amount of movie production on the market, it is also very easy for us, consumers, to jump from a movie to another, from a series to another. So, the decision factors for any of us to choose a certain movie to watch are many times in the details. Details maybe how exciting a trailer is, if it was shot in a certain country, the sound effects on the trailer, which company produced it, which actors and actresses are in… (Computer vision can be a great tool to help us understand the importance of this features)

In fact, actors and actresses are significant elements of the success – or lack of it – of a movie. The presence of a certain actor in a movie might – positively or negatively – affect its final rating. As mentioned above, the public is demanding, and so having a high rating is a good sign for a movie to be profitable. Receiving good feedback from who buys a product is a goal for any industry, anywhere in the world. When it comes to entertainment, ratings are a way of "reading" that feedback.

Through the IMDb data base, it is possible to develop analysis and understand what might influence the ratings. Regression models, convolutional neural network and keras models are examples of possible paths in order to explain the dependent variable (movie rating).

## Key – words

Entertainment industry, Movie rating, Computer vision, IMDb data base, Regression, Convolutional neural network, Keras models.

## Introduction

This project aims at analyzing movie data – available at the IMDb database – explore its evolution over the years concerning different features, develop a model to predict ratings and increase our knowledge in computer vision through a small example of its use.

Living through a pandemic is an unprecedented situation (in the most recent decades) that surely took a tool on the entertainment industry. Further on the project, it will be possible to visualize the evolution of the number of movies produced by year. An intense decrease is expected when comparing 2020 to previous years.

When focusing on movie ratings, one big question that arises is: *which features are determinant to the rating?* In fact, many characteristics can lead to a movie having a higher or lower evaluation. For example, is the movie genre relevant to the overall rating? Can certain actors and actresses increase (or decrease) the movie rating? To be able to come up with answers to these questions, data will be explored and a sample will be used to develop prediction models – having as regressors the previous features.

Lastly, we want to push our boundaries when it comes to computer vision and develop a small model that will get the movie name from its front cover and automatically search it on the IMDb data base.

***Disclaimer:*** The full code written to develop this project (and the corresponding jupyter notebooks) is available on the **github repository**: https://github.com/ISEG-MDAB/PDS-imdb-dataset

When going through the repository, please start by reading the file *READEME.md* to get all information on both data (also explained on the *Methodology* of this report) and on what is the utility of each notebook (as well as in which order they should be analyzed).

## Literature Review

Being so popular and powerful, movie industry is also a target for many different researches. To develop this project, it was important to explore some papers and analyze some already tested models. Through this literature review, three research reports on movie rating prediction will be mention (each of them considering different techniques).

On table 1, the main aspects of each paper are referenced:

| Authors | Objectives | Method | Results |
|---|---|---|---|
| Augustine, A., & Pathak, M. (2020)[1] | Build a model to predict movie ratings based on its crew members' data. | Neural networks | Model with 9.83% of accuracy. Main errors due to missing information on the training set |
| Abarja, R. A., & Wibowo, A. (2020)[2] | Develop a predicting model for IMDb movie ratings based on movie features. | Convolutional neural network (CNN) | Best results using one dimensional CNN. May be used as a model for small sized samples. |
| Dixit, P., Hussain, S., & Singh, G (2020)[3] | Produce a model to predict IMDb rating. | Regression models and Classification models | Best model is the classification model Gradient Boosting with 83% accuracy |

Table 1 – Literature review: Papers

The first paper (Augustine, A., & Pathak, M. (2020)) develops a prediction model based on Neural Networks mainly based on crew members information. It was important to go through this research since the accuracy obtained is fairly low – which also shows us the importance of testing different models and try different paths to achieve the best possible results.

Using one dimensional convolutional neural networks from (Abarja, R. A., & Wibowo, A. (2020)) proved to be an accurate approach to predict movie ratings when working with somehow small samples.

Finally, on the last paper (Dixit, P., Hussain, S., & Singh, G (2020)), the authors explore regression and classification models, trying different approaches and obtaining the best accuracy when using the Gradient Boosting classification model.

Note that Gradient Boosting classification model was not a topic covered on the class for which this report is intended. For further information on the topic:

- https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html *(sklearn Gradient Boosting documentation)*

## Methodology

The methodology fallowed is the CRIST-DM/POST-DS[4]: business understanding, data understanding, data preparation, modelling and model validation.

*Business understanding:*

In IMDb website, we can find several datasets, regarding information about all kinds of tv shows and movies and its crew. Analyzing this type of information can be a powerful tool to understand what the consumer is looking for a movie or a tv show.

When writers or directors plan their movies, they can have a better understanding of which actors and genres work better to achieve the best rating possible.

*Data understanding:*

To develop this project, we worked with **5 IMDb data sets**, regarding movies from 2010 to 2021:

- ***title_basics***

  Data set containing a row per movie/series, movie/series code as index and columns:

  - *titleType* : type of title (ex: for movies, 'titleType': short; for episodes of tv shows, 'titleType': tvEpisode)
  - *primaryTitle* : movie title;
  - *originalTitle* : original movie title;
  - *isAdult* : 0 if movie is not Adult; 1 if it is;
  - *startYear* : release year;
  - *endYear* : year in which the series stopped being streamed. '\N' for all movies;
  - *runtimeMinutes* : movie duration, in minutes;
  - *genres* : which genres are associated with the movie (does not have to have only one genre, can store multiple separated by commas (ex: Action,Drama)

Note that this data set also includes data regarding series, not only movies, but only movies' data were considered for this project.

- ***title_ratings***

  Data set regarding movies' ratings. For each movie, stores by column:

  - *averageRating* : stores the average rating attributed to the movie (on a scale from 1 to 10);
  - *numVotes* : number of votes on the corresponding movie;

- ***title_crew***

  Data set with crew information (directors and writers) per movie. Each individual is given a code, so there are no names on this data set. Columns:

  - *directors* : codes of the directors of the movie;
  - *numVotes* : codes of the writers of the movie;

  On both columns, each cell can store more than one code, doing so separating codes by commas. *(ex: code1,code2)*

- ***title_principals***

  Data set storing information regarding all professionals who took part on the movie

  Columns:

  - *tconst* : movie code;
  - *ordering* : works as an individual index for each movie. Integers starting at 1, incrementing 1 by each row while the movie code does not change. When the entire movie crew is characterized, 'tconst' column receives a new code and 'ordering' starts from 1 again;
  - *nconst* : crew member code;
  - *category* : stores the in which category the person works on this particular movie;
  - *job* : stores the actual rolle of the person on the movie; (ex: 'category': cinematographer, 'job': director of photography);
  - *characters* : For actors/actresses, stores the character(s) played on the movie. For all other crew members, '\N'.

- ***name_basics***

  Data set storing information for each professional related to movies. Has as columns:

  - *primaryName* : person's name;
  - *birthYear* : Year in which he/she was born;
  - *deathYear* : Year in which he/she died. If the corresponding professional is alive, stores '\N';
  - *primaryProfession* : profession(s) related to movies;
  - *knownForTitles* : Movie codes in which he/she worked on. Can store multiple codes, separated by commas (*ex: code1,code2,code3*).

*Data preparation:*

- **Data preparation for visualization**
    - Filter the dataset on movies from 2010 to 2021;
    - Delete all the rows with NA values on startYear and genre;
    - Merge the rating data frame to the title_basics data frame;
    - Filter all the other data frames on the movies that are on the title_basics data frame;
    - Create dummies on the title_basics data frame to all genres;
    - Export all this new table to csv files;
- **Data preparation for rating prediction**
    - Filter the visualization dataset on movies from 2020 to 2021;
    - Remove all movies with NA values on runtimeMinutes;
    - Create dummies for all actors;
    - Scale all numeric values;

*Modelling and validation*

To model this dataset, we selected a set of models: OLS, Ridge (alpha = 0.5), Lasso (alpha = 0.5), Bayesian Ridge, Neural Network and a Sequential Model from keras. We then measured the quality of the models using 3 indicators: $R^2$, Mean Absolute Error and Mean Squared Error.

After developing the model and choosing the best one, each movie can have a predicted rate. Having this prediction writers and directors can:

- Chose the optimal crew and movie genre;
- Maximize the movie rating;
- Understand what the audience prefers.

## Results/Discussion

When presenting and analyzing the results obtained, we will work in three different stages. Each stage corresponds to one jupyter notebook available on the **github repository** (link displayed on report introduction):

1. Data Exploration: notebook*: movie_ratings_visualization.ipynb*
2. Movie Rating Prediction: notebook: *Rating_prediction.ipynb*
3. Computer Vision: notebook: *Computer_vision.ipnyb*

*Data Exploration*

Total number of movies produced by year (impact of COVID 19)

| Year | Number of movies | % Change |
|---|---|---|
| 2010 | 6925 | - |
| 2011 | 7474 | 7.35 |
| 2012 | 7869 | 5.02 |
| 2013 | 8199 | 4.02 |
| 2014 | 8750 | 6.30 |
| 2015 | 8903 | 1.72 |
| 2016 | 9248 | 3.73 |
| 2017 | 9558 | 3.24 |
| 2018 | 9392 | -1.77 |
| 2019 | 8989 | -4.48 |
| 2020 | 6011 | -49.54 |
| 2021 | 1044 | - |

Table 2 – Total number of movies produced by year (% change)

As expected, 2020 was a rough year for movie production. When compared with previous years (since 2010), it registered the lowest number. COVID 19 pandemic forcing all of us to live 2020 in lockdown and social distancing to be the “new normal”, really translates into these numbers. In fact, from 2019 to 2020, movie industry suffered an approximate 49.54% decrease (in bright red on the table). This is a totally different number from what has happened in recent years: from 2010 onwards, the level of movie production has been growing, with the exceptions of 2018 and 2019 in which it saw decreases of 1.77% and 4.48%, respectively. Note that 2021 having a low number does not yet translate into real conclusions, since by the time this project was developed, more than half of the year was still to come.

*Notebook Visualization: Scatter plot – “Total number of movies produced by year”*

Number of movies produced by year (genre comparison)

| Genre | Number of movies |
|---|---|
| Drama | 2247 |
| Documentary | 1445 |
| Comedy | 1179 |
| Thriller | 736 |

Table 3 – Top 4 genres of most produced movies 2020

| Genre | Number of movies |
|---|---|
| Drama | 3704 |
| Documentary | 2118 |
| Comedy | 1910 |
| Thriller | 1002 |

Table 4 – Top 4 genres of most produced movies 2019

The genre is also a significant characteristic of a movie. We all have certain expectations once we know the genre – for example, we do not expect the same from animation and horror movies.

On the jupyter notebook, you will find an interactive scatter plot, from which it is possible to compare the number of movies produced through the years (2010 to 2021) for any two genres. On tables 3 and 4, there are displayed the top 4 movie genres (of most produced movies) in 2020 and 2019 respectively. As you can see, the four-place podium did not change – all the numbers are, as expected, lower in 2020, due to the overall decrease registered (analyzed on the previous topic) – these being Drama, Documentary, Comedy and Thriller.

What insight can we take? Taking back to the beginning of the report, we mentioned the increase in public's demand and the need for variety. These results show exactly that: the top four genres of movies produced are distinct (documentary movies have different characteristics from thriller, same with drama and comedy). So, the industry is heavily producing a lot of different content, due to many different preferences from the consumers.

*Notebook Visualization: Interactive scatter plot – "Number of movies produced by year (Genre comparison)"*

Average rating per movie genre, per year

| Genre | Average rating |
|---|---|
| Documentary | 7.295 |
| Talk Show | 7.233 |
| Music | 7.058 |
| News | 7.04 |

Table 5 – Top 4 genres of highest average movie rating in 2020

| Genre | Average rating |
|---|---|
| Talk Show | 10 |
| Game Show | 8.5 |
| News | 7.233 |
| Documentary | 7.213 |

Table 6 – Top 4 genres of highest average movie rating in 2019

When analyzing the jupyter notebook, an interactive bar chart is displayed. By choosing the year, the user will get the average rating for each movie genre (for movies produced on the corresponding year). For simplification, on this report, we present the top four genres from 2019 and 2020.

Contrary to what happened regarding the number of movies produced by genre, the top 4 genres of highest average movie rating change from 2019 to 2020. Note that the ratings attributed to the movies are in a scale from 1 to 10 (10 being the best possible evaluation and 1 the worse). One thing to note is the consistency of Documentary movies: not only it is one of the preferred genres in both 2019 and 2020, but also one of the most type of movies produced. This tells us that Documentaries, in general, are both intensively produced and of very high quality.

Finally, also note that Talk Shows reached the first position in 2019 with a 10.0 average rating, and only fell down to second place in 2020, only 0.062 behind the highest rated genre (documentaries).

*Notebook Visualization: Bar chart – "Average ranking per movie genre, per year"*

Movie rating evolution per actor

Another possible analysis to be develop is regarding the average ratings per actor/actress (i.e., for a certain actor, the average rating of the movies in which he/she was a part of). Once again, this is an interactive analysis on our notebook, allowing the user to choose the name of the actor/actress. Then, there will be displayed both a bar chart with the average rating of the movies for that particular actor, and a scatter plot with the evolution of the ratings per year.

Note that, when the average is zero for a particular year, it means the actor was not a part of any movies in that particular year.

Once again to simplify our report, we will focus on a single case: Tom Holland (well known as Peter Parker on *Spider Man*).

Case of Tom Holland

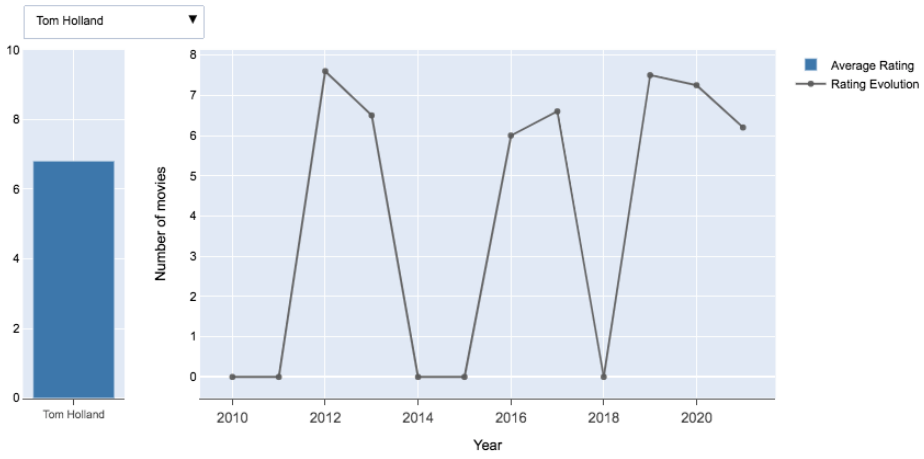


Figure 1 – Average rating and movie rating evolution (Tom Holland)

After selecting “Tom Holland” on the menu, we get the visuals displayed above (figure 1). The bar chart on the left indicates us that the average movie rating for movies in which Tom Holland was casted is around 6.8. Going into more detail on that number, we note that in 2010, 2011, 2014, 2015 and 2018, the actor was not in any movies (meaning the values on the scatter plots equal to zero for those years). On the remaining years, he remained very consistent, with averages always between 6 and 8.

*Notebook Visualization: “Movie rating evolution per actor/actress – Average Rating”*

Versatility indicator

In order to measure actors' versatility, we developed a measure regarding the number of different genres a certain actor/actress was in.

To create the versatility indicator, two steps were taken:

1. Assign a value to each actor/actress, according to how many different movie genres he/she had worked on.

*For example: if an actor only worked on action movies between 2010 and 2021, value is 1; if an actor worked on thriller, drama and action movies between 2010 and 2021, value is 3.*

2. Then standardize the indicator to a range from 0 to 10 (0 being the lowest value and 10 the highest).

| Actor | Versatility indicator |
|---|---|
| Eric Roberts | 10.0 |
| Forest Whitaker | 10.0 |
| Ewan McGregor | 9.474 |
| James Franco | 8.947 |
| Sean Patrick Flanery | 8.947 |
| Sean Astin | 8.947 |
| Michael Peña | 8.947 |
| Malcom McDowell | 8.947 |
| Danny Glover | 8.947 |

Table 7 – Top 9 actors (Versatility indicator)

Table 7 displays the top 9 actors on the developed versatility indicator. We thought it was interesting to find a way of measuring versatility because – as in any job – the more versatile an actor is, the more value he can bring. With the maximum possible value (10.0 score) there are Eric Roberts (three times golden globes nominee) and Forest Whitaker (golden globe for best actor with the movie "The Last King of Scotland"). The average value for the indicator is 2.333 (meaning these top 9 actors are way ahead of the average on these matter).

Just like the previous visuals, this indicator is presented on the notebook through an interactive chart, allowing the user to select the actor/actress and get the correspondent value for the measure.

*Movie Rating Prediction*

As described on the methodology, to predict movie ratings, 6 different models were trained and tested. On tables 8 and 9, values for the indicators used to measure models' performance are displayed (on train data and test data, respectively).

| Model | $R^2$ | MAE | MSE |
|---|---|---|---|
| OLS | 1.000000 | 0.000000 | 0.000000 |
| Ridge | 1.000000 | 0.000000 | 0.000000 |
| Lasso | 0.967598 | 0.239817 | 0.090002 |
| BaeysianRidge | 1.000000 | 0.000000 | 0.000000 |
| Neural Network | 0.999180 | 0.022509 | 0.002277 |
| Keras Model | 0.997455 | 0.062004 | 0.007069 |

Table 8 – Model performance indicators (Train Data)

| Model | $R^2$ | MAE | MSE |
|---|---|---|---|
| OLS | 0.999997 | 0.000614 | 0.000009 |
| Ridge | 1.000000 | 0.000218 | 0.000000 |
| Lasso | 0.967568 | 0.237683 | 0.087596 |
| BaeysianRidge | 1.000000 | 0.000000 | 0.000000 |
| Neural Network | 0.994833 | 0.092314 | 0.013957 |
| Keras Model | 0.989359 | 0.124104 | 0.028739 |

Table 9 – Model performance indicators (Test Data)

All models used produce extremely good results for the evaluation measures used. Note that we want a model with:

- $R^2$ as close to 1 as possible;
- MAE and MSE as close to 0 as possible.

To decide which model to use, we first focus our attention on Test data (Train data was used to train the model). The model that produces the best results is the **BayesianRidge Model**.

More than being extremely good, the numbers obtained for the models' measurement indicators allow us to note that the actors and genre are directly related to the final rating.

## Computer vision

On this course, we had an introduction on computer vision, and so we decided to explore this field in this project to try to get a better understanding of this area of the Data Science field. For this, we used a pre-trained model called EASY-OCR. Through this model, we were able to extract the name of the movie from the movie covers. After that, it was possible to search for the movie on the IMDb dataset, using the API imdbpy, that enables us to access the IMDb database directly from a Python Notebook.

This pre-trained model detects all text on a given image. In our case, we only needed to get a movie title, so to solve this problem (reading all information, not just the title), we decided to only save the text box with the greatest area (H x W). This solution brought some limitations on our implementation. For our implementation to work smoothly, we need to make sure that the text from the movie tittle is the biggest one on the cover, otherwise the model will consider other text and it will not return the desired value.


Figure 2 – Movie name recognition from front cover (Avatar and Doorman)

We didn't go has deep we wanted in this part of the project, because we focused more on the rating prediction. Nonetheless, with this implementation, we got an understanding of the main challenges in this field.

## Conclusion

Reaching the end of our project, we are now able to answer some questions raised in the introduction of this report.

One of the main goals was to understand which features were determinant to the rating of a movie. After developing all the analysis, it is possible to state that actors and actresses are a significant part of the rating. With this being said, let's consider a scenario with two movies: same script, same producer, directors, writers, cinematographers… the only aspect separating them are the actors. The movies would probably end up with different ratings depending on how dear (or not) the main actors and actresses were to the public eye. As the genres were also regressors of the models estimated, the same is possible to conclude: movie types will also affect the final rating.

Despite the prediction models developed, two different analysis were conducted: data exploration and computer vision. Through data exploration, we were able to visualize data from different perspectives, which gave us the ability to characterize it. As crucial as it is to develop great prediction models, it is also important to understand the data, to "get to know it". A variety of charts were displayed on this project in order for us to get more insight into the entertainment world, and in particular, movie features. On this section, it was also possible to measure the impact of COVID 19 on this industry. 2020 surely was not an easy year all across the world. When plotting the evolution of the overall number of movies produced per year (from 2010 to 2021) it was impactful to get almost a 50% decrease from 2019 to 2020. As we keep on riding this unstable wave, we hope what is coming is better than what has gone.

Lastly, we wanted to challenge ourselves and get a little bit into the wide world of computer vision. This was more of a complementary work, which we think was great to show some of the amazing possibilities through these tools and also to captivate and motivate us to learn more.

## References

[1] Augustine, A., & Pathak, M. (2008). User rating prediction for movies. *Technical Report. University of Texas at Austin.*

[2] Abarja, R. A., & Wibowo, A. (2020). Movie Rating Prediction using Convolutional Neural Network based on Historical Values. *International Journal*, *8*(5).

[3] Dixit, P., Hussain, S., & Singh, G. (2020). Predicting the IMDB rating by using EDA and machine learning Algorithms.

[4] C. J. Costa and J. T. Aparicio, "POST-DS: A Methodology to Boost Data Science," 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), Seville, Spain, 2020, pp. 1-6, doi: 10.23919/CISTI49556.2020.9140932.