IMDB Rating Prediction Proposal

Cody Kesler, Brandon Marshall, Jordan Andersen

April 5, 2020

1 Project Description

The rating of movies is often very subjective from person to person but averaged ratings, such as those found on IMDB, can be useful for assessing the general quality of a movie. Ratings that the critics give usually do not line up with average crowd-sourced review because the variance in critics scores is much higher as there are fewer critics to average the score. As more people rate a movie, the crowd-sourced score approaches what may be considered the 'real' more true to life score. Because of this, this project proposes to regress on the information of a film to determine its crowd-sourced average rating. This regression will give a predicted 'crowd-sourced' rating without the need for hundreds or thousands of people needing to see the movie and rate it.

2 Models

Our approach will be to regress on the crowd-sourced rating label given our input features. Our input feature space consists of variable length text, categorical, and numerical data with a label between 1 and 10. To achieve the highest generalization accuracy we will test two different methods, standard statistical machine learning regression models and a deep learning model.

2.1 Kernel/Word-2-Vec with Regression

As mentioned above, part of the difficulty with the data is the combination of both text features and numerical features. The second method we intend to try and use is to encode these words into a numerical value using either a kernel or some other method. Doing so will allow us to run a regression algorithm (such as linear regression, knn, etc). As such, we intend to research into commonly used kernels in Natural Language Processing, along with other potential methods to encode these text values into a number. Once these texts are encoded, we intend to run Linear Regression, KNN Regression, and RF Regression on these features.

2.2 Deep Learning

The first architecture that we intend to test is a neural network. We have three types of data we need to vectorize the data to regress. For this, we will need 2 different deep learning networks working in tandem. One to encode the text data and one to handle the regression using the numerical data and the encoding from the text encoder. Numerical data can simply be normalized, and fed into the main regression network. We will vectorize the categorical data using one-hot encoding.

The most difficult of these data-types will be the variable-length text data from the movie title and description. We will embed and vectorize the variable-length text data using a GRU, which will output and embedded vector in a latent space which will be fed into the main regression network. The benefits of this method are that it can be pre-trained by creating an auto-encoder to boost the regression training. We can then backpropagate through both the main regression network and the GRU, the GRU will 'learn' what sequences of text are most important, by getting error signal from the main model.

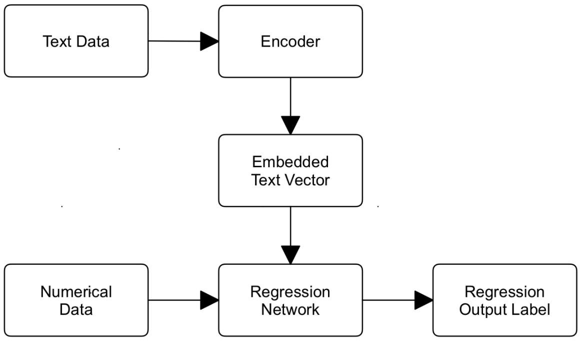


Figure 1: Deep Leaning Model

Thus, as shown above in Figure 1, with the two networks, we can create a model that takes numerical and text data and regresses on the given labels.

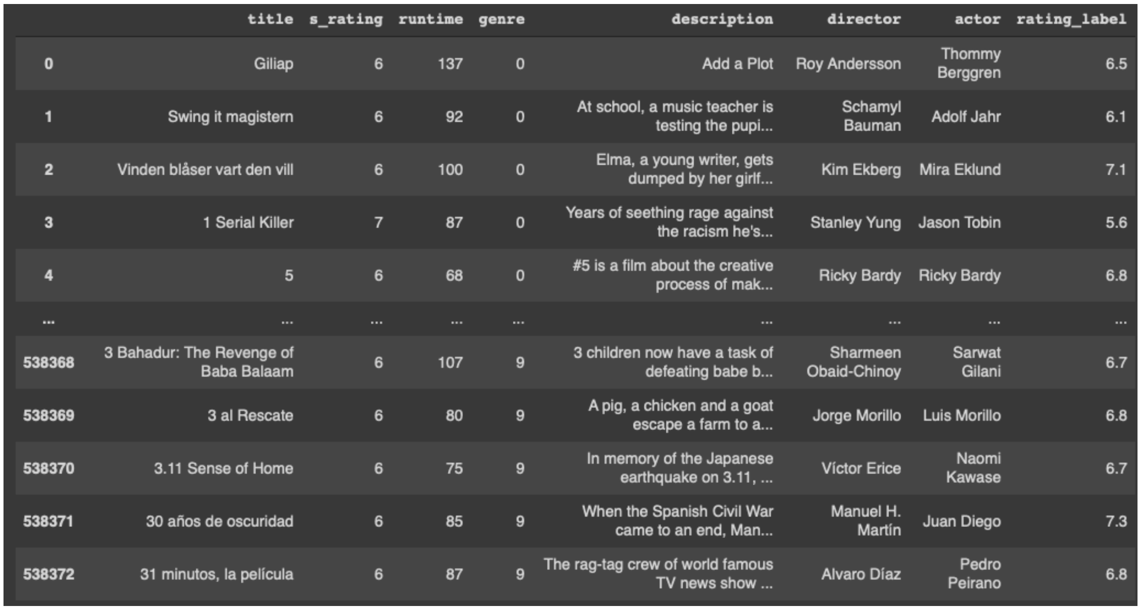
3 Data

3.1 Scraping

The data is being scraped from IMDB from the url $https : //www.imdb.com/search/title/?title\_type = movie\&view = advanced\&sort = alpha, asc\&genres = g$ where g is the genre to access.

3.2 Description

There are 15 genre categories on IMDB: ”comedy”, ”sci-fi”, ”horror”, ”romance”, ”action”, ”thriller”, ”drama”, ”mystery”,”crime”, ”animation”, ”adventure”,”fantasy”, ”comedy-romance”, ”action-comedy”, ”superhero”. Each genre is denoted with enumeration. The ratings: ””, ”Unrated”, ”Not Rated”, ”R”, ”PG-13”, ”PG”, are also all enumerated in the data. We plan to have 100000 instances per genre which will give us approximately 1.5 million instances of the data, which should be enough to avoid over fitting and allow for good generalization. Below in Figure 2 instances of the data are shown.



	title	s_rating	runtime	genre	description	director	actor	rating_label
0	Giliap	6	137	0	Add a Plot	Roy Andersson	Thommy Berggren	6.5
1	Swing it magistern	6	92	0	At school, a music teacher is testing the pupi...	Schamyl Bauman	Adolf Jahr	6.1
2	Vinden blåser vart den vill	6	100	0	Elma, a young writer, gets dumped by her girlf...	Kim Ekberg	Mira Eklund	7.1
3	1 Serial Killer	7	87	0	Years of seething rage against the racism he's...	Stanley Yung	Jason Tobin	5.6
4	5	6	68	0	#5 is a film about the creative process of mak...	Ricky Bardy	Ricky Bardy	6.8
...	...	...	...	...	...	...	...	...
538368	3 Bahadur: The Revenge of Baba Balaam	6	107	9	3 children now have a task of defeating babe b...	Sharmeen Obaid-Chinoy	Sarwat Gilani	6.7
538369	3 al Rescate	6	80	9	A pig, a chicken and a goat escape a farm to a...	Jorge Morillo	Luis Morillo	6.8
538370	3.11 Sense of Home	6	75	9	In memory of the Japanese earthquake on 3.11, ...	Víctor Erice	Naomi Kawase	6.7
538371	30 años de oscuridad	6	85	9	When the Spanish Civil War came to an end, Man...	Manuel H. Martín	Juan Diego	7.3
538372	31 minutos, la película	6	87	9	The rag-tag crew of world famous TV news show ...	Alvaro Díaz	Pedro Peirano	6.8

Figure 2: Data Instances