Logistic Regression on IMDb Score Classification

Shuhui Guo
Group 5

1 Introduction

1.1 Background

IMDb is the initial of the Internet Movie Database, which is an online database of information related to films, television programs reviews and ratings. A successful movie should not only have the ability to gain tremendous profit, but also get a high rating. For the IMDb score, the score of 7 is a cutoff. The movies with the score greater than 7 are regarded as a good movie, while the movies with the score lower than 7 cannot be regarded as a good movie.

1.2 Question Statement

Based on the background, it is essential to predict whether a movie can be scored greater than 7 before it is released. The results can help film companies to understand the secrets of generating a successful movie.

1.3 Data Description

The dataset is from Kaggle website. The original dataset contains 28 variables for 5043 movies, spanning from the year of 1916 to 2016 in 66 countries. This dataset contains a variety of information, including the box office performance metrics, IMDb ratings, information on the director and the cast, and the social media metrics. After removing observations with missing values, the number of observations after data cleaning is approximately 3800. The variables chosen to do prediction in this question are as below:

Variable Name	Property	Description
duration	numerical	Duration in minutes
color	categorical	Film colorization. 'Black and White' or 'Color'
genres	categorical	Film categorization like 'Animation', 'Comedy', 'Romance', 'Horror', 'Sci-Fi', 'Action', 'Family'
title_year	numerical	The year in which the movie is released (1916-2016)
language	categorical	English, Arabic, Chinese, French, German, Danish, Italian, Japanese, etc
country	categorical	Country where the movie is produced
content_rating	categorical	Content rating of the movie
aspect_ratio	numerical	Aspect ratio the movie was made in
facenumber_in_poster	numerical	Number of the actor who featured in the movie poster
budget	numerical	Budget of the movie in Dollars
imdb_score	numerical	IMDB Score of the movie on IMDB

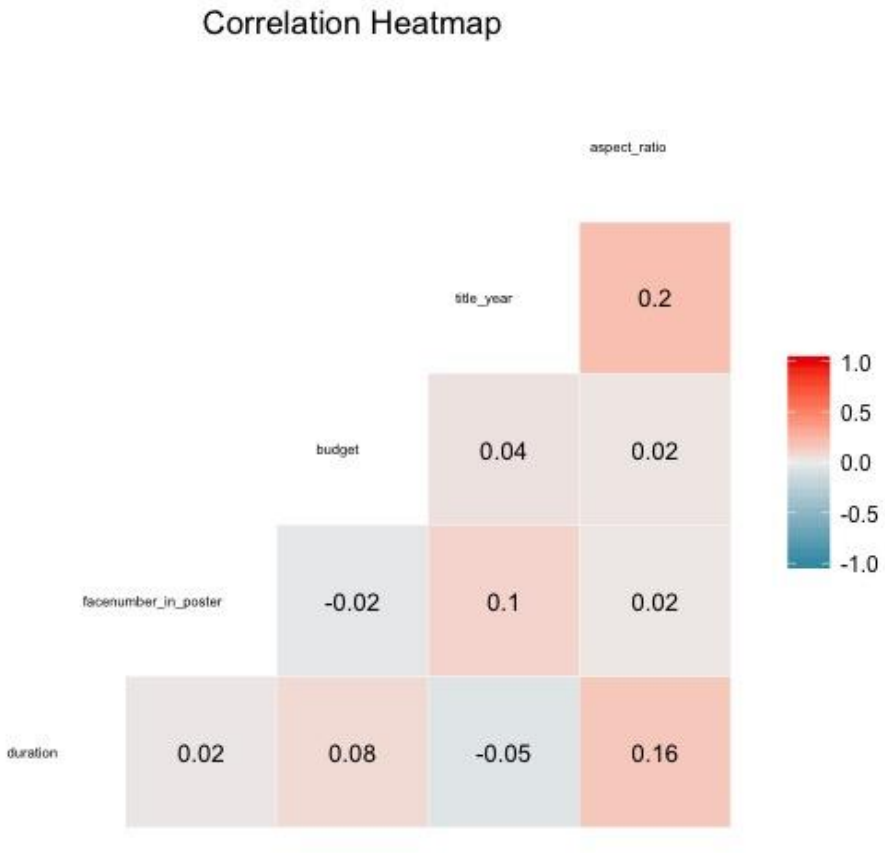
Among the above variables, the response variable is imdb_score. Level 1 refers to imdb_score greater than 7, and level 0 refers to imdb_score no greater than 7. Other 11 variables are possible predictors, which contain the information before the release of movies.

The descriptive statistics of the numerical variables are as below:

variable	Min	1st Qu	Median	Mean	3rd Qu	Max
duration	37.0	96.0	106.0	110.2	120.0	330.0
facenumber_in_poster	0.000	0.000	1.000	1.379	2.000	43.000

budget	2180	10000000	25000000	45850000	50000000	12220000000
title_year	1920	1999	2005	2003	2010	2016
aspect_ratio	1.18	1.85	2.35	2.11	2.35	16.00
imdb_score	1.600	5.900	6.600	6.466	7.200	9.300

The correlation heatmap of the numerical predictors is plotted as below:



Based on the heatmap, there is no high correlation (greater than 0.7) between predictors. Therefore, the predictors can be used for prediction.

The descriptive statistics of the categorical variables are as below:

imdb_score	greater than 7	no greater than 7
count	1165	2636

color	black and white	color
count	129	3671

genres	Action	Adventure	Animation	Biography	Comedy	Crime	Documentary	Drama	Family
count	962	371	45	207	1000	259	30	684	3
genres	Horror	Fantasy	Musical	Mystery	Thriller	Sci-Fi	Romance	Western	
count	164	37	2	23	1	8	2	3	

language	Aboriginal	Arabic	Aramaic	Bosnian	Cantonese	Czech	Danish
count	2	1	1	1	8	1	3
language	Dari	Dutch	Dzongkha	English	Filipino	French	German
count	2	3	1	3626	1	36	13
language	Hebrew	Hindi	Hungarian	Icelandic	Indonesian	Italian	Japanese
count	1	8	1	1	2	7	12
language	Kazakh	Korean	Mandarin	Maya	Mongolian	None	Norwegian
count	1	5	15	1	1	1	4
language	Persian	Portuguese	Romanian	Russian	Spanish	Swedish	Telugu
count	3	5	1	1	23	1	1
language	Thai	Vietnamese	Zulu				
count	3	1	1				

country	Afghanistan	Argentina	Aruba	Australia	Belgium	Brazil	Canada
count	1	3	1	40	2	5	62
country	Chile	China	Colombia	Czech Republic	Denmark	Finland	France
count	1	17	1	3	8	1	103
country	Georgia	Germany	Greece	Hong Kong	Hungary	Iceland	India
count	1	82	1	13	2	2	10
country	Indonesia	Iran	Ireland	Israel	Italy	Japan	Mexico
count	1	4	7	1	11	17	8
country	Netherlands	New Line	New Zealand	Norway	Official site	Peru	Poland
count	3	1	11	4	1	1	1
country	Romania	Russia	South Africa	South Korea	Spain	Sweden	Thailand
count	2	3	3	8	22	1	4
country	UK	USA	West Germany				
count	322	3005	1				

content_rating	Approved	G	GP	M	NC-17	Not Rated
count	17	87	1	2	6	34
content_rating	Passed	PG	PG-13	R	Unrated	X
count	3	567	1313	1709	23	10

After acknowledging and exploring the data, analysis will be conducted in the next parts.

2 Methods and Results

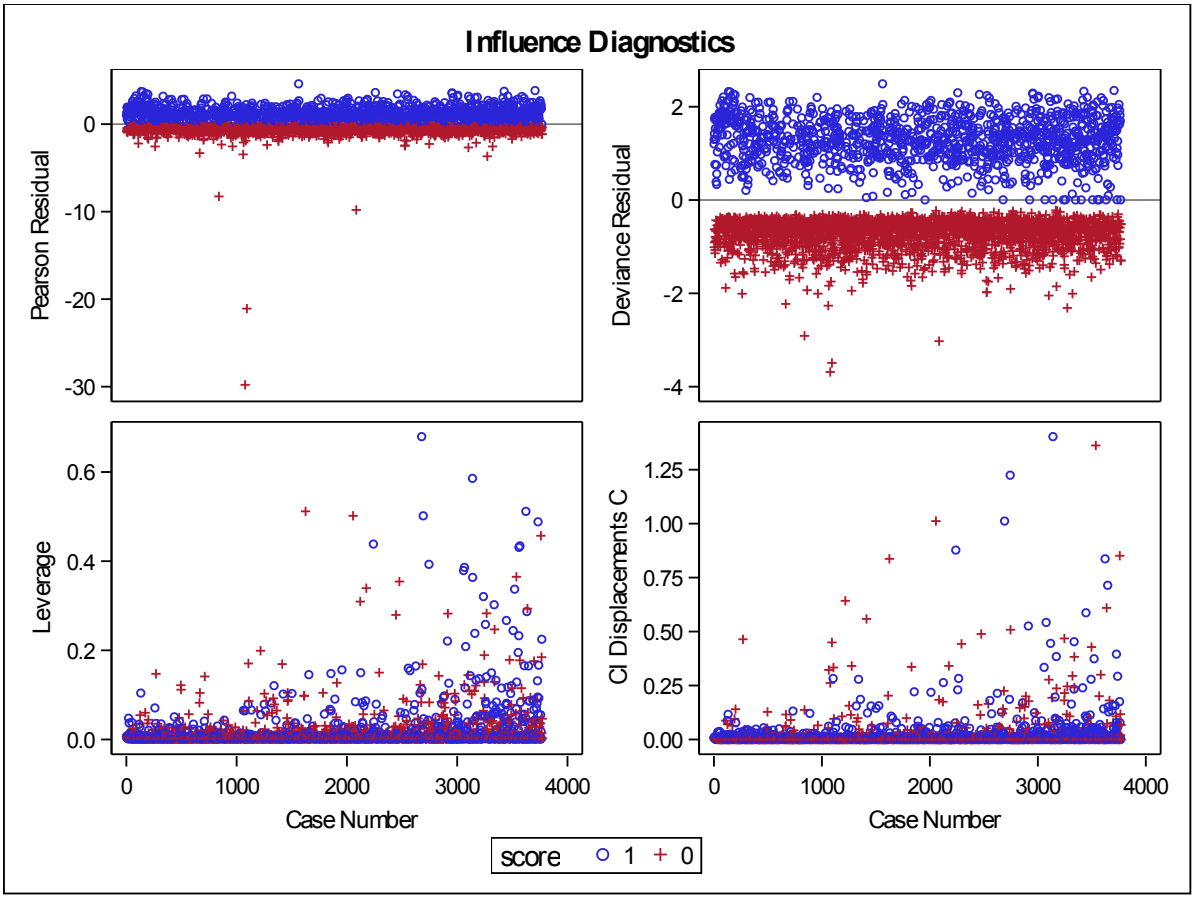
2.1 Model Selection

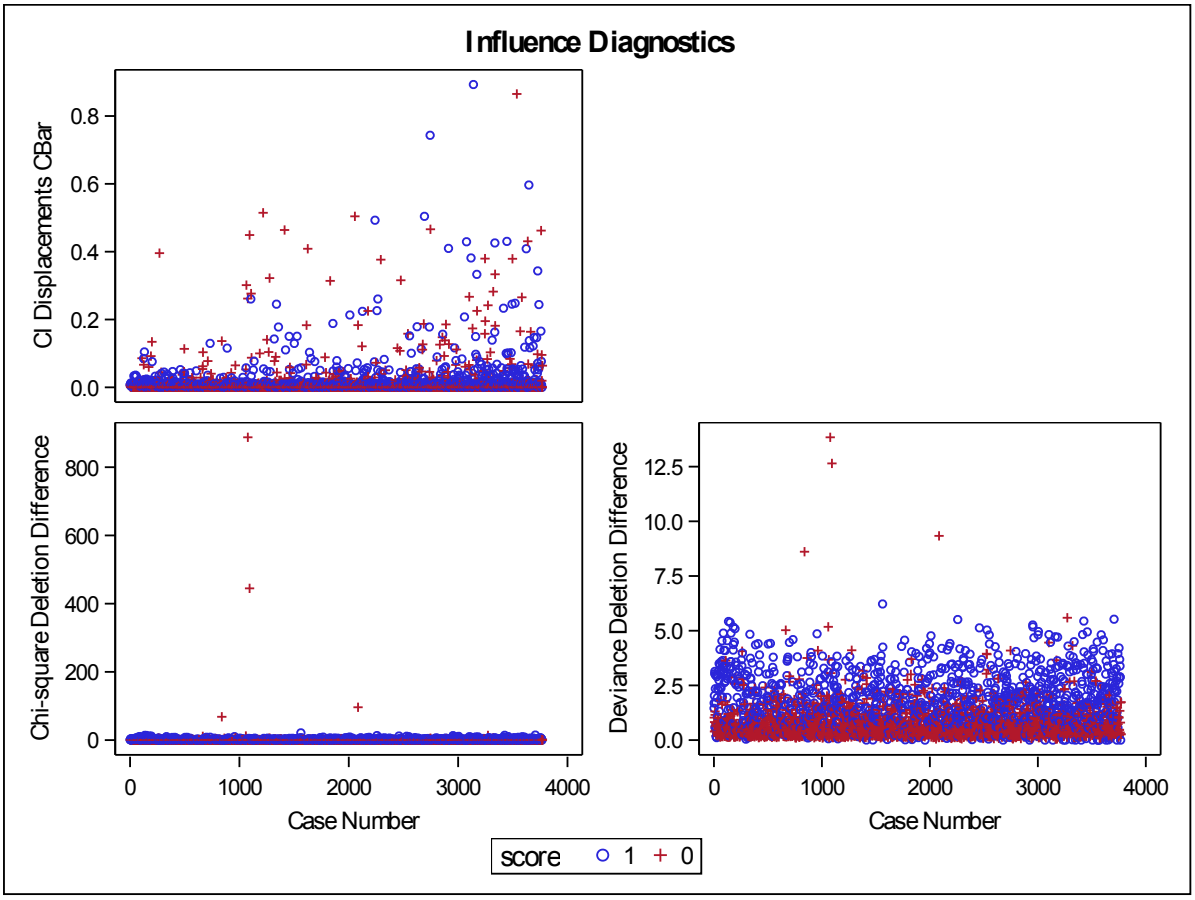
In this part, 10 iterations of the model selection are conducted (details in appendices). In each iteration, the stepwise selection is conducted and extremely unduly influential points are removed. Totally 30 points are removed and the selection results in the final iteration are as below:

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	duration		1	1	338.9972		<.0001
2	genres		13	2	258.9168		<.0001
3	language		32	3	144.4476		<.0001
4	title_year		1	4	52.1417		<.0001
5	content_rating		8	5	49.2163		<.0001
6	country		21	6	51.9603		0.0002
7		language	32	5		38.1408	0.2102
8	language		28	6	65.7658		<.0001
9		language	32	5		38.1408	0.2102

According to the above table, the variables in the final model are **duration**, **genres**, **title_year**, **content_rating**, and **country**.

The influence diagnostics are as below:





According to the above plots, the highest Cbar value is about 0.8, which is in a relatively lower level. But there is a remaining issue with diagnostics. There are certain observations with absolute value of Pearson residual greater than 2, which might be high influential points. However, there is no obvious trend in residuals and the Cbar values do not appear that these points are extremely unduly influential. Therefore, there is no need to refit this model.

2.2 Quasi-complete separation problem

Although the above diagnostics are fine, the quasi-complete separation problem is detected in the final model. The quasi-complete separation happens when a logistic model perfectly predicts the response. Under this condition, unique maximum likelihood estimates do not exist. Some estimates of parameters are abnormal. One solution is to remove the variables that cause this problem. It is often difficult to know exactly which variables cause the separation, but variables that exhibit large parameter estimates or standard errors are likely candidates. To check the cause of the quasi-complete separation problem, the estimates of predictors are shown as below:

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	50.0590	17.8916	7.8283	0.0051
duration		1	0.0303	0.00225	182.3506	<.0001
genres	**Action**	1	-0.9769	0.1828	28.5604	<.0001
genres	**Adventure**	1	0.0634	0.2012	0.0993	0.7527
genres	**Animation**	1	0.9999	0.3437	8.4653	0.0036
genres	**Biography**	1	0.5249	0.2190	5.7463	0.0165
genres	**Comedy**	1	-0.7956	0.1810	19.3086	<.0001
genres	**Crime**	1	0.0274	0.2084	0.0172	0.8955
genres	**Documenta**	1	1.7413	0.4315	16.2824	<.0001
genres	**Drama**	1	0.0555	0.1823	0.0927	0.7607
genres	**Family**	1	-0.1303	1.1759	0.0123	0.9118
genres	**Fantasy**	1	-1.3608	0.4930	7.6173	0.0058
genres	**Horror**	1	-1.4936	0.2882	26.8554	<.0001
genres	**Mystery**	1	0.1928	0.4547	0.1798	0.6715
genres	**Sci-Fi**	1	0.3146	0.7612	0.1708	0.6794
title_year		1	-0.0254	0.00466	29.6077	<.0001
content_rating	**Appro**	1	-0.0166	0.5645	0.0009	0.9766
content_rating	**G**	1	0.00847	0.2886	0.0009	0.9766
content_rating	**NC-17**	1	-0.3774	0.8381	0.2028	0.6525
content_rating	**Not R**	1	0.3912	0.4391	0.7937	0.3730
content_rating	**PG**	1	-0.4316	0.1980	4.7517	0.0293
content_rating	**PG-13**	1	-0.5756	0.1926	8.9367	0.0028
content_rating	**R**	1	0.0447	0.1836	0.0592	0.8078
content_rating	**Unrat**	1	-0.1583	0.4936	0.1028	0.7485
country	**Arg**	1	-2.1480	15.3175	0.0197	0.8885
country	**Aus**	1	-2.7271	15.2756	0.0319	0.8583
country	**Bra**	1	9.2940	147.9	0.0040	0.9499
country	**Can**	1	-3.0495	15.2747	0.0399	0.8418
country	**Chi**	1	-1.4956	15.2826	0.0096	0.9220
country	**Cze**	1	10.1397	220.8	0.0021	0.9634
country	**Den**	1	-1.3239	15.2952	0.0075	0.9310

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
country	**Fra**	1	-2.3923	15.2730	0.0245	0.8755
country	**Ger**	1	-2.9601	15.2738	0.0376	0.8463
country	**Hon**	1	-1.0625	15.2819	0.0048	0.9446
country	**Hun**	1	-2.6890	15.3396	0.0307	0.8608
country	**Ind**	1	9.7680	155.1	0.0040	0.9498
country	**Ira**	1	9.4376	180.5	0.0027	0.9583
country	**Ire**	1	-2.0455	15.2917	0.0179	0.8936
country	**Ita**	1	-3.1382	15.2888	0.0421	0.8374
country	**Jap**	1	-2.2473	15.2822	0.0216	0.8831
country	**Mex**	1	-1.8992	15.2897	0.0154	0.9011
country	**Net**	1	8.5913	176.9	0.0024	0.9613
country	**New**	1	-0.8896	15.2923	0.0034	0.9536
country	**Nor**	1	-1.4315	15.3039	0.0087	0.9255
country	**Rom**	1	-2.0977	15.3477	0.0187	0.8913
country	**Rus**	1	-2.6330	15.3272	0.0295	0.8636
country	**Sou**	1	-1.6930	15.2839	0.0123	0.9118
country	**Spa**	1	-1.9044	15.2779	0.0155	0.9008
country	**Tha**	1	-2.0945	15.3179	0.0187	0.8912
country	**UK**	1	-2.2963	15.2719	0.0226	0.8805

Based on the above table, the standard errors of variable **Country** are abnormal, which indicates that **Country** might cause the quasi-complete separation problem. Therefore, **Country** is removed and the parameter estimates of the refitted model are as below:

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	41.3565	9.1706	20.3371	<.0001
duration		1	0.0301	0.00220	187.5428	<.0001
genres	**Action**	1	-0.8856	0.1806	24.0514	<.0001
genres	**Adventure**	1	0.0971	0.1993	0.2375	0.6260
genres	**Animation**	1	0.9795	0.3428	8.1625	0.0043
genres	**Biography**	1	0.6380	0.2168	8.6574	0.0033

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
genres	**Comedy**	1	-0.7952	0.1799	19.5375	<.0001
genres	**Crime**	1	0.0373	0.2066	0.0326	0.8566
genres	**Documenta**	1	1.6823	0.4289	15.3835	<.0001
genres	**Drama**	1	0.1491	0.1805	0.6825	0.4087
genres	**Family**	1	-0.2021	1.1724	0.0297	0.8631
genres	**Fantasy**	1	-1.2304	0.4585	7.2024	0.0073
genres	**Horror**	1	-1.5200	0.2861	28.2218	<.0001
genres	**Mystery**	1	0.1720	0.4456	0.1490	0.6995
genres	**Sci-Fi**	1	0.2054	0.7582	0.0734	0.7864
title_year		1	-0.0224	0.00459	23.8199	<.0001
content_rating	**Appro**	1	0.2232	0.5725	0.1520	0.6966
content_rating	**G**	1	-0.0904	0.2830	0.1021	0.7494
content_rating	**NC-17**	1	-0.3655	0.8147	0.2013	0.6537
content_rating	**Not R**	1	0.6924	0.4162	2.7672	0.0962
content_rating	**PG**	1	-0.5272	0.1936	7.4147	0.0065
content_rating	**PG-13**	1	-0.7007	0.1881	13.8824	0.0002
content_rating	**R**	1	-0.0311	0.1796	0.0300	0.8625
content_rating	**Unrat**	1	-0.1462	0.4868	0.0901	0.7640

The quasi-complete separation problem is solved. Therefore, the best set of predictors includes **duration**, **genres**, **title_year**, and **content_rating**. Take the variables duration and genres-Action as example. For each one unit change in duration, the log odds of getting imdb score greater than 7 increases by 0.0301. This might because the movies with longer duration are more likely to enrich their plots. For each one unit change in genres-Action, the log odds of getting imdb score greater than 7 decreases by 0.8856. This might because action movies are more likely to being outdated and not attractive to audiences.

2.3 Model Assessment

After fitting the logistic regression model using **duration**, **genres**, **title_year**, and **content_rating** as predictors, the results are shown as below:

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	700.9080	23	<.0001
Score	666.0828	23	<.0001
Wald	521.0927	23	<.0001

Based on the above table, the p-values of Likelihood ratio test, score test and Wald test are all less than 0.05, which rejects the null hypothesis that the global parameter estimate is 0. It is reasonable to say that the model is significant.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
8.3252	8	0.4024

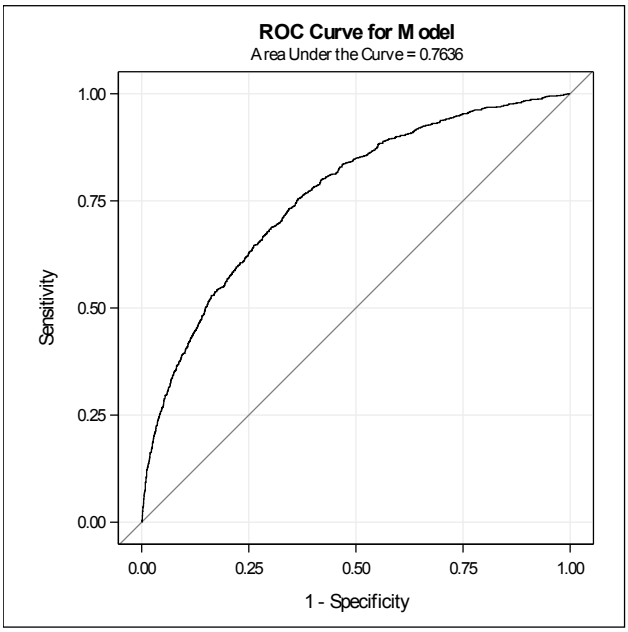
In Hosmer-Lemeshow Goodness of fit test, the p-value is 0.4024>0.05, which does not reject the null hypothesis so we can conclude that there is no lack of fit in this model.

Based on the above analysis, the logistic regression model using **duration**, **genres**, **title_year**, and **content_rating** as predictors can be used for prediction.

2.4 Prediction Performance

Set the threshold as 0.5. The prediction probabilities higher than 0.5 should be classified to level 1, and the probabilities lower than 0.5 should be classified to level 0.

The ROC curve is plotted as below:



The area under the curve (AUC) is used to summarize the performance of the model, and it is currently 0.7636. This value is pretty high, which indicates that the model has a good performance.

The confusion matrix is constructed as below:

Table of score by test			
score	**test**		
Frequency	**0**	**1**	**Total**
0	2405	212	2617
1	738	416	1154
Total	3143	628	3771

Based on the above table, out of 3771 observations in total, the model classifies 2821 correctly, which is 0.75. This result seems to be good. Nevertheless, the classification accuracy of each outcome is quite different. Out of the 2617 observations which get scores no greater than 7, the model classifies 2405 correctly, which is 0.92. On the other hand, out of the 1154 observations which get scores greater than 7, the model classifies 416 correctly, which is only 0.36. Therefore, there exists an unbalanced issue in the prediction results.

2.5 Model Adjustment

In this part, the unbalanced issue will be addressed by finding the best cutoff to do classification. To make the classification results better, the false positive rate (fpr) and false negative rate (fnr) should be as small as possible. To ensure this, there should be a threshold in which the distance between (fpr, fnr) and (0, 0) is the smallest. The Euclidean distance is used to measure this distance:

$$d = \sqrt{fpr^2 + fnr^2}$$

Finally the smallest distance is selected to be 0.4468. The corresponding threshold is 0.26. The prediction probabilities higher than 0.26 should be classified to level 1, and the probabilities lower than 0.26 should be classified to level 0.

The confusion matrix is constructed as below:

Table of score by test			
score	**test**		
Frequency	**0**	**1**	**Total**
0	1687	930	2617
1	311	843	1154
Total	1998	1773	3771

Based on the above table, out of 3771 observations in total, the model classifies 2530 correctly, which is 0.67. This result is lower than before, but the classification accuracy of each outcome should also be checked. Out of the 2617 observations which get scores no greater than 7, the model classifies 1687

correctly, which is 0.64. On the other hand, out of the 1154 observations which get scores greater than 7, the model classifies 843 correctly, which is only 0.73. The results are balanced and the overall model is much more accurate than before.

3 Conclusions

According to the above analysis of the movies' IMDb scores, certain conclusions could be made. Before a movie's release date, whether this movie will get an IMDb score greater than 7 can be predicted by logistic regression model. The selected information used to do prediction includes duration, genres, title year, and content rating of movies. The cutoff should be set as 0.26 to get prediction results. The prediction probabilities higher than 0.26 should be classified to the level greater than 7, and the probabilities lower than 0.26 should be classified to the level no greater than 7. The prediction accuracy rate is 0.67, which is good but not very well.

To get the prediction results which are not only balanced but also with high prediction accuracy, further analysis like support vector machine and random forest could be tried in the future. Also, the counts of certain categories in the variable **language** and **country** are too small. The analysis would be better if such categories were merged.

4 Appendices

4.1 Model Selection Process

The highly influential points in iteration 1:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
989	1047	Color	137	Crime	3	English	Cze	R
2365	2606	Color	120	Action	3	Mandari	Tai	PG-13
2483	2738	Color	105	Action	1	Mandari	Tai	Not R
2534	2811	Color	127	Action	0	English	Ira	PG
2898	3277	Color	176	Drama	4	Hindi	Ind	Not R
2922	3312	Color	300	Action	3	Thai	Tha	R
2960	3376	Color	142	Action	2	English	UK	M
3019	3466	Color	120	Action	2	English	UK	GP
3106	3582	Color	110	Musical	2	English	USA	PG-13
3109	3585	Color	110	Biography	1	English	USA	M
3159	3660	Color	111	Action	0	Thai	Tha	R
3338	3954	Color	106	Drama	2	Portugu	Bra	Unrat
3432	4158	Black	102	Adventure	3	English	USA	Passe
3727	4813	Black	100	Musical	8	English	USA	Passe

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1
989	50000000	2015	6.4	2.35	0	1.638

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
2365	15000000	2000	7.9	2.35	1	1.317		
2483	15000000	2015	6.4	1.37	0	1.317		
2534	14000000	1978	6.5	2.35	0	1.041		
2898	7217600	2012	6.9	2.35	0	1.019		
2922	400000000	2001	6.6	1.85	0	10.871		
2960	7000000	1969	6.8	2.35	0	1.134		
3019	7200000	1971	6.7	2.35	0	157.374		
3106	6000000	1978	7.2	2.35	1	1.280		
3109	6000000	1969	8.1	2.35	1	1.134		
3159	200000000	2005	7.1	1.85	1	8.090		
3338	4000000	2014	6.1	2.35	0	1.134		
3432	2800000	1939	8.1	1.37	1	1.587		
3727	379000	1929	6.3	1.37	0	1.280		

The point with Cbar1 greater than 150 is removed.

The highly influential points in iteration 2:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
989	1047	Color	137	Crime	3	English	Cze	R
1036	1098	Color	110	Action	2	English	Per	R
1407	1492	Color	91	Action	0	English	Aru	R
2172	2370	Color	109	Biography	1	English	Bel	R
2249	2457	Color	112	Adventure	0	English	Gre	PG-13
2365	2606	Color	120	Action	3	Mandari	Tai	PG-13
2429	2676	Color	117	Drama	4	English	Off	PG-13
2483	2738	Color	105	Action	1	Mandari	Tai	Not R
2534	2811	Color	127	Action	0	English	Ira	PG
2670	2971	Color	293	Adventure	0	German	Wes	R
2707	3014	Color	113	Action	9	English	Geo	R
2898	3277	Color	176	Drama	4	Hindi	Ind	Not R
2922	3312	Color	300	Action	3	Thai	Tha	R
2960	3376	Color	142	Action	2	English	UK	M
3105	3582	Color	110	Musical	2	English	USA	PG-13

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
3108	3585	Color	110	Biography	1	English	USA	M
3158	3660	Color	111	Action	0	Thai	Tha	R
3175	3683	Color	93	Comedy	1	French	Fin	Not R
3337	3954	Color	106	Drama	2	Portugu	Bra	Unrat
3345	3980	Color	99	Drama	2	English	Ice	R
3375	4044	Color	101	Crime	0	Spanish	Col	R
3431	4158	Black	102	Adventure	3	English	USA	Passe
3464	4214	Color	101	Drama	0	English	Pol	PG-13
3496	4285	Color	90	Animation	0	Hebrew	Isr	R
3690	4736	Color	83	Drama	1	Dari	Afg	PG-13
3726	4813	Black	100	Musical	8	English	USA	Passe

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2
989	50000000	2015	6.4	2.35	0	1.6382	1.64
1036	45000000	1994	5.4	1.85	0	0.0000	2025.90
1407	35000000	1998	4.8	2.35	0	0.0000	2025.89
2172	15000000	2011	7.1	2.35	1	0.0000	1459.60
2249	18000000	2016	6.7	2.35	0	0.0000	2025.95
2365	15000000	2000	7.9	2.35	1	1.3175	1.32
2429	15000000	2010	6.3	2.35	0	0.0000	2025.90
2483	15000000	2015	6.4	1.37	0	1.3175	1.32
2534	14000000	1978	6.5	2.35	0	1.0411	1.04
2670	14000000	1981	8.4	1.85	1	0.0000	1459.61
2707	20000000	2011	5.6	2.35	0	0.0000	2025.95
2898	7217600	2012	6.9	2.35	0	1.0194	1.02
2922	400000000	2001	6.6	1.85	0	10.8710	10.87
2960	7000000	1969	6.8	2.35	0	1.1344	1.13
3105	6000000	1978	7.2	2.35	1	1.2802	1.28
3108	6000000	1969	8.1	2.35	1	1.1344	1.13
3158	200000000	2005	7.1	1.85	1	8.0901	8.09
3175	3850000	2011	7.2	1.85	1	0.0000	1459.58
3337	4000000	2014	6.1	2.35	0	1.1339	1.13
3345	3800000	2009	6.9	2.35	0	0.0000	2025.89

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2
3375	3000000	2004	7.5	1.85	1	0.0000	1459.60
3431	2800000	1939	8.1	1.37	1	1.5872	1.59
3464	2400000	2006	5.1	2.35	0	0.0000	2025.94
3496	1500000	2008	8	1.85	1	0.0000	1459.58
3690	46000	2003	7.4	1.85	1	0.0000	1459.56
3726	379000	1929	6.3	1.37	0	1.2802	1.28

The points with Cbar2 greater than 1000 are removed.

The highly influential points in iteration 3:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
881	929	Color	106	Romance	2	English	USA	PG-13
989	1047	Color	137	Crime	3	English	Cze	R
2529	2811	Color	127	Action	0	English	Ira	PG
2891	3277	Color	176	Drama	4	Hindi	Ind	Not R
2915	3312	Color	300	Action	3	Thai	Tha	R
2953	3376	Color	142	Action	2	English	UK	M
3098	3582	Color	110	Musical	2	English	USA	PG-13
3101	3585	Color	110	Biography	1	English	USA	M
3151	3660	Color	111	Action	0	Thai	Tha	R
3329	3954	Color	106	Drama	2	Portugu	Bra	Unrat
3421	4158	Black	102	Adventure	3	English	USA	Passe
3713	4813	Black	100	Musical	8	English	USA	Passe
3726	4846	Color	95	Thriller	3	English	USA	R

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3
881	50200000	2011	7.1	1.85	1	0.0000	0.0000	9789.25
989	50000000	2015	6.4	2.35	0	1.6382	1.6382	1.63
2529	14000000	1978	6.5	2.35	0	1.0411	1.0411	1.10
2891	7217600	2012	6.9	2.35	0	1.0194	1.0194	1.07
2915	400000000	2001	6.6	1.85	0	10.8710	10.8710	9.06
2953	7000000	1969	6.8	2.35	0	1.1344	1.1344	1.08
3098	6000000	1978	7.2	2.35	1	1.2802	1.2802	1.41
3101	6000000	1969	8.1	2.35	1	1.1344	1.1344	1.08

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3
3151	200000000	2005	7.1	1.85	1	8.0901	8.0901	6.63
3329	4000000	2014	6.1	2.35	0	1.1339	1.1339	1.16
3421	2800000	1939	8.1	1.37	1	1.5872	1.5872	1.60
3713	379000	1929	6.3	1.37	0	1.2802	1.2802	1.41
3726	300000	2014	4.8	1.85	0	0.0000	0.0000	13689.51

The points with Cbar3 greater than 9000 are removed.

The highly influential points in iteration 4:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
988	1047	Color	137	Crime	3	English	Cze	R
2528	2811	Color	127	Action	0	English	Ira	PG
2890	3277	Color	176	Drama	4	Hindi	Ind	Not R
2914	3312	Color	300	Action	3	Thai	Tha	R
2952	3376	Color	142	Action	2	English	UK	M
3097	3582	Color	110	Musical	2	English	USA	PG-13
3100	3585	Color	110	Biography	1	English	USA	M
3150	3660	Color	111	Action	0	Thai	Tha	R
3328	3954	Color	106	Drama	2	Portugu	Bra	Unrat
3420	4158	Black	102	Adventure	3	English	USA	Passe
3712	4813	Black	100	Musical	8	English	USA	Passe

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3	cbar4
988	50000000	2015	6.4	2.35	0	1.6382	1.6382	1.62556	1.62556
2528	14000000	1978	6.5	2.35	0	1.0411	1.0411	1.09652	1.09652
2890	7217600	2012	6.9	2.35	0	1.0194	1.0194	1.06590	1.06590
2914	400000000	2001	6.6	1.85	0	10.8710	10.8710	9.05786	9.05786
2952	7000000	1969	6.8	2.35	0	1.1344	1.1344	1.08448	1.08448
3097	6000000	1978	7.2	2.35	1	1.2802	1.2802	1.40681	1.40681
3100	6000000	1969	8.1	2.35	1	1.1344	1.1344	1.08448	1.08448
3150	200000000	2005	7.1	1.85	1	8.0901	8.0901	6.63014	6.63014
3328	4000000	2014	6.1	2.35	0	1.1339	1.1339	1.16286	1.16286
3420	2800000	1939	8.1	1.37	1	1.5872	1.5872	1.60135	1.60135
3712	379000	1929	6.3	1.37	0	1.2802	1.2802	1.40681	1.40681

The points with Cbar4 greater than 9 are removed.

The highly influential points in iteration 5:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating	budget
988	1047	Color	137	Crime	3	English	Cze	R	50000000
2528	2811	Color	127	Action	0	English	Ira	PG	14000000
2890	3277	Color	176	Drama	4	Hindi	Ind	Not R	7217600
2951	3376	Color	142	Action	2	English	UK	M	7000000
3096	3582	Color	110	Musical	2	English	USA	PG-13	6000000
3099	3585	Color	110	Biography	1	English	USA	M	6000000
3327	3954	Color	106	Drama	2	Portugu	Bra	Unrat	4000000
3419	4158	Black	102	Adventure	3	English	USA	Passe	2800000
3711	4813	Black	100	Musical	8	English	USA	Passe	379000

Obs	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3	cbar4	cbar5
988	2015	6.4	2.35	0	1.63821	1.63821	1.62556	1.62556	1.63264
2528	1978	6.5	2.35	0	1.04110	1.04110	1.09652	1.09652	1.10340
2890	2012	6.9	2.35	0	1.01943	1.01943	1.06590	1.06590	1.07832
2951	1969	6.8	2.35	0	1.13438	1.13438	1.08448	1.08448	1.10115
3096	1978	7.2	2.35	1	1.28018	1.28018	1.40681	1.40681	1.40878
3099	1969	8.1	2.35	1	1.13438	1.13438	1.08448	1.08448	1.10115
3327	2014	6.1	2.35	0	1.13393	1.13393	1.16286	1.16286	1.16207
3419	1939	8.1	1.37	1	1.58723	1.58723	1.60135	1.60135	1.60183
3711	1929	6.3	1.37	0	1.28018	1.28018	1.40681	1.40681	1.40878

The points with Cbar5 greater than 1.6 are removed.

The highly influential points in iteration 6:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating	budget
2527	2811	Color	127	Action	0	English	Ira	PG	14000000
2889	3277	Color	176	Drama	4	Hindi	Ind	Not R	7217600
2950	3376	Color	142	Action	2	English	UK	M	7000000
3095	3582	Color	110	Musical	2	English	USA	PG-13	6000000

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating	budget
3098	3585	Color	110	Biography	1	English	USA	M	6000000
3326	3954	Color	106	Drama	2	Portugu	Bra	Unrat	4000000
3709	4813	Black	100	Musical	8	English	USA	Passe	379000

Obs	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3	cbar4	cbar5	cbar6
2527	1978	6.5	2.35	0	1.04110	1.04110	1.09652	1.09652	1.10340	1.10171
2889	2012	6.9	2.35	0	1.01943	1.01943	1.06590	1.06590	1.07832	1.09430
2950	1969	6.8	2.35	0	1.13438	1.13438	1.08448	1.08448	1.10115	1.07800
3095	1978	7.2	2.35	1	1.28018	1.28018	1.40681	1.40681	1.40878	1.03105
3098	1969	8.1	2.35	1	1.13438	1.13438	1.08448	1.08448	1.10115	1.07800
3326	2014	6.1	2.35	0	1.13393	1.13393	1.16286	1.16286	1.16207	1.16515
3709	1929	6.3	1.37	0	1.28018	1.28018	1.40681	1.40681	1.40878	1.23306

The points with Cbar6 greater than 1.2 are removed.

The highly influential points in iteration 7:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating	budget
2527	2811	Color	127	Action	0	English	Ira	PG	14000000
2889	3277	Color	176	Drama	4	Hindi	Ind	Not R	7217600
2950	3376	Color	142	Action	2	English	UK	M	7000000
3098	3585	Color	110	Biography	1	English	USA	M	6000000
3326	3954	Color	106	Drama	2	Portugu	Bra	Unrat	4000000

Obs	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3	cbar4	cbar5	cbar6	cbar7
2527	1978	6.5	2.35	0	1.04110	1.04110	1.09652	1.09652	1.10340	1.10171	1.10171
2889	2012	6.9	2.35	0	1.01943	1.01943	1.06590	1.06590	1.07832	1.09430	1.09430
2950	1969	6.8	2.35	0	1.13438	1.13438	1.08448	1.08448	1.10115	1.07800	1.07800
3098	1969	8.1	2.35	1	1.13438	1.13438	1.08448	1.08448	1.10115	1.07800	1.07800
3326	2014	6.1	2.35	0	1.13393	1.13393	1.16286	1.16286	1.16207	1.16515	1.16515

The points with Cbar7 greater than 1.1 are removed.

The highly influential points in iteration 8:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating	budget
2888	3277	Color	176	Drama	4	Hindi	Ind	Not R	7217600
2949	3376	Color	142	Action	2	English	UK	M	7000000
3097	3585	Color	110	Biography	1	English	USA	M	6000000

Obs	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3	cbar4	cbar5	cbar6	cbar7	cbar8
2888	2012	6.9	2.35	0	1.01943	1.01943	1.06590	1.06590	1.07832	1.09430	1.09430	1.09429
2949	1969	6.8	2.35	0	1.13438	1.13438	1.08448	1.08448	1.10115	1.07800	1.07800	1.08207
3097	1969	8.1	2.35	1	1.13438	1.13438	1.08448	1.08448	1.10115	1.07800	1.07800	1.08207

The points with Cbar8 greater than 1 are removed.

The highly influential points in iteration 9:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
2359	2606	Color	120	Action	3	Mandari	Tai	PG-13
2476	2738	Color	105	Action	1	Mandari	Tai	Not R
2742	3076	Color	193	Drama	2	Hindi	Ind	R
3092	3582	Color	110	Musical	2	English	USA	PG-13
3326	3975	Color	107	Biography	0	English	USA	Passe

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3	cbar4
2359	15000000	2000	7.9	2.35	1	1.31749	1.31749	0.93664	0.93664
2476	15000000	2015	6.4	1.37	0	1.31749	1.31749	0.93664	0.93664
2742	700000000	2006	6	2.35	0	0.90824	0.90824	0.93688	0.93688
3092	6000000	1978	7.2	2.35	1	1.28018	1.28018	1.40681	1.40681
3326	3768785	1950	7	1.37	0	0.21580	0.21580	0.15620	0.15620

Obs	cbar5	cbar6	cbar7	cbar8	cbar9
2359	0.93476	0.98623	0.98623	0.96644	1.07
2476	0.93476	0.98623	0.98623	0.96644	1.07
2742	0.94581	0.95288	0.95288	0.95134	1.40
3092	1.40878	1.03105	0.00000	0.00000	35802.94
3326	0.15550	0.00000	0.00000	0.00000	22429.36

The points with Cbar9 greater than 20000 are removed.

The highly influential points in iteration 10:

Obs	index	color	duration	genres	facenumber_in_poster	language	country	content_rating
2359	2606	Color	120	Action	3	Mandari	Tai	PG-13
2476	2738	Color	105	Action	1	Mandari	Tai	Not R
2742	3076	Color	193	Drama	2	Hindi	Ind	R

Obs	budget	title_year	imdb_score	aspect_ratio	score	cbar1	cbar2	cbar3	cbar4
2359	15000000	2000	7.9	2.35	1	1.31749	1.31749	0.93664	0.93664
2476	15000000	2015	6.4	1.37	0	1.31749	1.31749	0.93664	0.93664
2742	700000000	2006	6	2.35	0	0.90824	0.90824	0.93688	0.93688

Obs	cbar5	cbar6	cbar7	cbar8	cbar9	cbar10
2359	0.93476	0.98623	0.98623	0.96644	1.06649	1.06649
2476	0.93476	0.98623	0.98623	0.96644	1.06649	1.06649
2742	0.94581	0.95288	0.95288	0.95134	1.39885	1.39885

The points with Cbar10 greater than 1 are removed. And finally there are no points with Cbar greater than 1 in the next iteration.