1. Background and business goals

2. Data collection and preparation

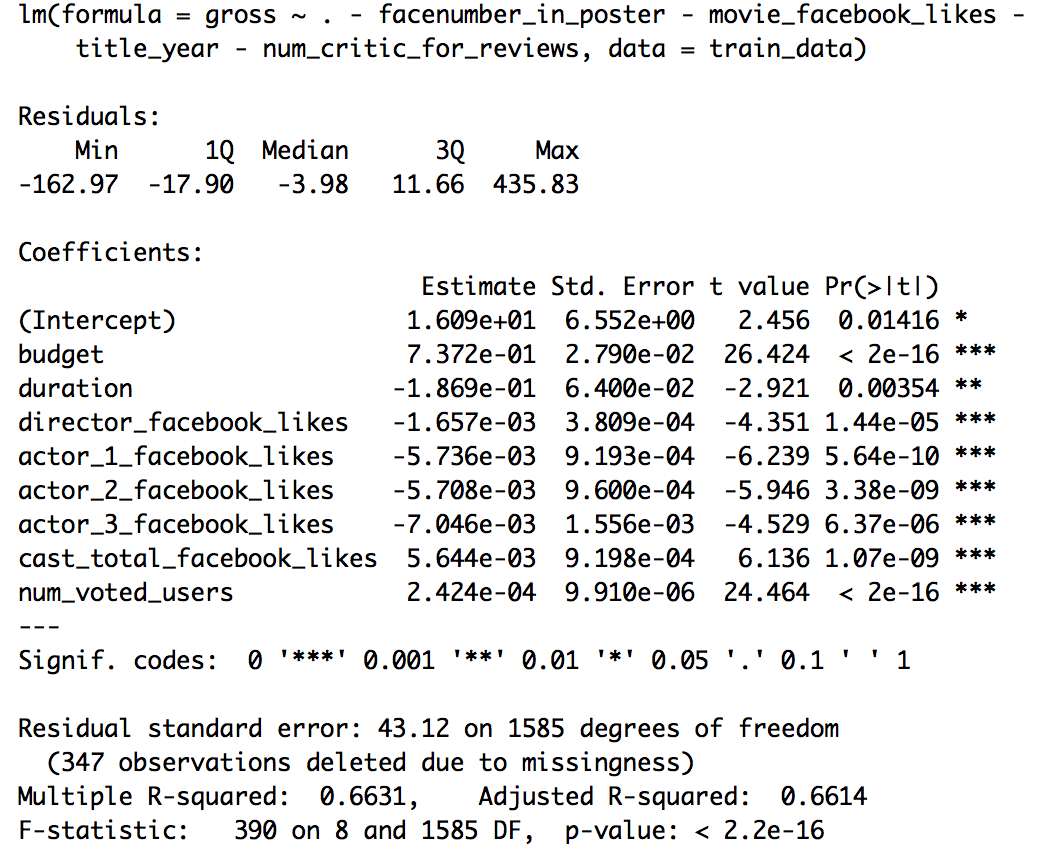
3. Modeling

3.1 Analysis into IMDb score

3.2 Analysis into gross

3.2.1 Linear regression

In building linear regression model, we involve only numerical variables. Numerical variables in the dataset include: The numerical variables include: number critic for reviews, duration, director Facebook likes, actor 3 Facebook likes, actor 1 Facebook likes, number voted users, cast total Facebook likes, face number in poster, budget, actor 2 Facebook likes, IMDb score, aspect ratio, movie Facebook likes. We assume that gross has relationship with both production and marketing process. Thus, gross is probably affected by factors including budget, duration and popularity of director and cast (which is reflected by Facebook likes). Additionally, gross is also affected by word-of-mouth marketing online, which could be reflected through IMDb score (which may be a reference whether to watch the movie), number of voted users and number of critics for reviews.

We remove variables that has highest P-value one by one and finally identify those have obvious relationship with gross.

From this model, budget has a strong positive relationship with gross. Duration has an obvious negative relationship with gross. It is interesting that although cast total Facebook likes has a slight positive relationship with gross, coefficients for director and actor 1,2,3 are all negative. Consequently, conclusion could be made that gross does not show obvious relationship with Facebook likes. Number of voted users shows a really slight positive impact on gross.

3.2.2 Ridge model

We run ridge model in order to obtain insights into categorical variables in the dataset including: genres, keywords, directors and actors. We scale gross (scaled grosses are one digit numbers) and use it as dependent variables to calculate coefficients of each genre, keyword, director and actor to find their impact on gross.

For ridge model on genres, we get a satisfying mean square of errors at 0.823. And the related coefficients for top 20 genres are as exhibited. For keywords, we get a satisfying mean square of errors at 0.926. And the related coefficients for top 20 keywords are as exhibited.

|  |  |  |  |
| --- | --- | --- | --- |
| Genre | Coefficient | Keyword | Coefficient |
| Adventure | 0.54039629 | iron man | 1.7325129 |
| Animation | 0.33085388 | velociraptor | 1.19511312 |
| Action | 0.30981386 | experiment gone wrong | 1.19503439 |
| Fantasy | 0.22553026 | disaster film | 1.19489973 |
| Sci-Fi | 0.20380205 | physical appearance | 1.14607346 |
| History | 0.15351692 | dc comics | 1.09140279 |
| War | 0.13753685 | teddy bear | 0.95738073 |
| Biography | 0.08883533 | day care | 0.95737811 |
| Sport | 0.03640468 | tentacle | 0.94949297 |
| Family | 0.03503043 | goblin | 0.92559591 |
| News | 0.03409123 | snowman | 0.91979423 |
| Mystery | 0.01213348 | sister love | 0.91976416 |
| Short | 0 | self survival | 0.90420161 |
| Romance | -0.00106688 | fight to the death | 0.90417362 |
| Music | -0.04436885 | terrorist plot | 0.82381366 |
| Musical | -0.04822326 | lawlessness | 0.82378445 |
| Thriller | -0.05213305 | imprisonment | 0.82369967 |
| Crime | -0.05813061 | death of child | 0.80385281 |
| Comedy | -0.0647272 | missile attack | 0.79208552 |
| Western | -0.14187135 | human bomb | 0.7920397 |

For ridge model on directors, we get a relatively satisfying mean square of errors at 1.055. And the related coefficients for top 20 directors are as exhibited. For actors, we get a relatively satisfying mean square of errors at 0.974. And the related coefficients for top 20 directors are as exhibited.

|  |  |  |  |
| --- | --- | --- | --- |
| Director | Coefficient | Actors | Coefficient |
| Colin Trevorrow | 7.307043342 | Anthony Reynolds | 1.400930145 |
| Lee Unkrich | 4.420763514 | Conrad Vernon | 1.378292484 |
| Chris Buck | 4.24737759 | Don Rickles | 1.335078496 |
| Sam Raimi | 4.099274173 | Omar Sy | 1.27030978 |
| George Lucas | 3.998223279 | Maurice LaMarche | 1.201861892 |
| Joss Whedon | 3.864177499 | Livvy Stubenrauch | 1.201848519 |
| Tim Miller | 3.788397717 | Alan D. Purwin | 1.04722418 |
| Kyle Balda | 3.459938788 | Lester Speight | 1.042884567 |
| Pete Docter | 3.181569619 | Ed Skrein | 1.037281294 |
| Francis Lawrence | 2.9211836 | Stefan Kapicic | 1.037272367 |
| Andrew Adamson | 2.888433709 | Keir O'Donnell | 1.032398895 |
| Peter Jackson | 2.473987788 | Leonard Roberts | 1.032380053 |
| Mark Andrews | 2.258396103 | Billy Boyd | 1.006461543 |
| Andrew Stanton | 2.132797309 | Kevin Dunn | 0.942740474 |
| Zack Snyder | 2.130138218 | Orlando Bloom | 0.902153161 |
| Don Hall | 2.07836302 | Zack Ward | 0.841528596 |
| Anthony Russo | 2.077894329 | John Ratzenberger | 0.839564831 |
| Betty Thomas | 2.04332642 | Jess Harnell | 0.835581957 |
| Christopher Nolan | 2.039271466 | Willow Smith | 0.829381564 |
| John Lasseter | 2.020873264 | James Coburn | 0.822328922 |

3.2.3 Lasso model

For lasso model on genres, we get a relatively satisfying mean square of errors at 0.974. And the related coefficients for top 20 directors are as exhibited.

For lasso model on keywords, we get a relatively satisfying mean square of errors at 0.974. And the related coefficients for top 20 directors are as exhibited.

For lasso model on actors, we get a relatively satisfying mean square of errors at 0.974. And the related coefficients for top 20 directors are as exhibited.

For lasso model on actors, we get a relatively satisfying mean square of errors at 0.974. And the related coefficients for top 20 directors are as exhibited.

4. Evaluation

Cross-analysis between IMDb score and gross

5. Conclusions and limitations

6. Program improvements