[](http://www.cequitysolutions.com/)

Summer Internship Project:

# Movie Prediction Model

## Introduction:

The project involved analyzing the revenue and budget metrics of movies that have released in the last few years (2011 to June 2015) alongside their Social Media presence, in terms of trailer hits and response on YouTube and Facebook likes. The analysis is used to understand the correlation, if any between social media presence and the box office performance of a movie and whether a movie’s result at the box office can be predicted using the available social media parameters.

The first two weeks involved initial data collation and parallel learning of the basics of R programming.

We are working with 272 unique rows of data. There are two aspects to the data:

1. The Revenue metrics- this encompasses the opening day collection, opening weekend collection, opening week collection, lifetime box office collection and the approximate budget.
2. Social Media presence- this comprises the number of Facebook likes, whether the movie has an active Facebook page, the trailer video hits on YouTube and the response recorded through the “thumbs up/down” aspect of the video.

The subsequent two weeks comprised of cleaning and preprocessing the data. This involved Form different functions and their requirements to pull the necessary information from the data in hand.

## Functions:

1. Movie related functions:

ROI - calculates the return on investment in a film in Crores as well as the percentage turnover based on the approx budget of the film.

Verdict – classifies a movie in terms of its success of failure based on the % return on investment.

Expected ROI% - calculated based on the average ROI of the actors in the cast of the movie.

1. Actor functions:

No\_of\_Films - lists out the different actors and the number of films he/she has been a part of within the list.

No\_of\_Hits - lists out the different actors and the number of successful films he/she has been a part of within the list.

Act\_Avg\_ROI – calculated the average return on investment for an actor and the average budget of a movie he / she acts in.

1. Social Media functions:

Tweets.R file contained the function for scraping tweets associated with the movie prior to release. However as the search API has an unstable database that keeps on short term tweets, we could not incorporate the twitter presence as a parameter for the model.

Raw Data:

The raw data is contained in the excel file, Bollywood Data. It comprises of three sheets, Movie Revenues, Social Media and Upcoming Movies

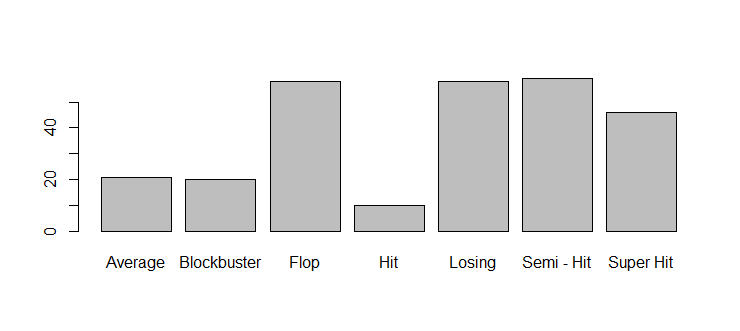
Processed Data:

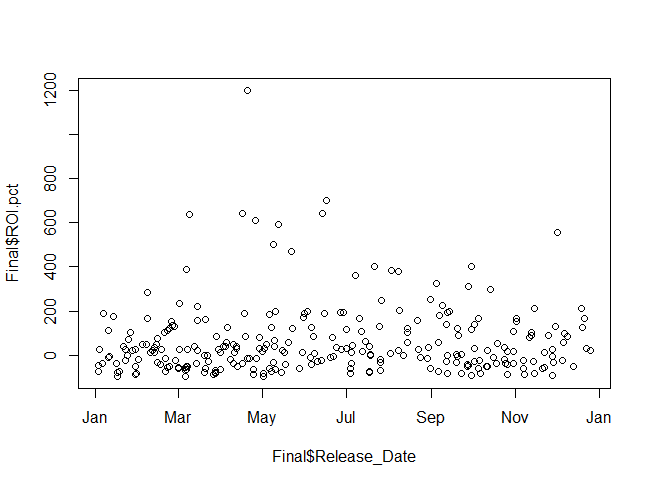
The Movie Revenues and Social Media sheets were merged to give the CSV file, Merged.

The Actors.csv file was created using the functions in the Actor.R script.

Final.csv contains the processed data with the calculated verdict and ROI function results and the social media metrics.

|  |
| --- |
|  |
| Average Blockbuster Flop Hit Losing Semi - Hit Super Hit |
| 21 20 58 10 58 59 46 |



 Observing ROI % against the different months a year

**Correlation between the expected ROI % and actual ROI % values:**

0.724322

**Correlation between the YouTube Hits and actual ROI% values:**

0.07494341

**Correlation between Days between Trailer and Movie release to ROI %:**

-0.01121751

**Correlation between Trailer hits and FB following:**

cor(FB.Following,Youtube.Trailer.hits)

[1] 0.03732381

**Correlation between Trailer hits and trailer thumbs up:**

cor(Trailer.Thumbs.Up,Youtube.Trailer.hits)

[1] 0.8111163

**Correlation between Budget and Trailer hits:**

cor(as.numeric(as.character(Budget)),Youtube.Trailer.hits)

[1] 0.5655812

**Correlation between Trailer and Movie release and Expected ROI %:**

cor(as.numeric(as.character(Expected.ROI.pct)),Youtube.Trailer.hits)

[1] 0.1146286

cor(as.numeric(as.character(Expected.ROI.pct)),as.numeric(as.character(Budget)))

[1] -0.02337642

**Correlation between Budget to ROI %:**

cor(as.numeric(as.character(ROI.pct)),as.numeric(as.character(Budget)))

[1] -0.1089411

## Linear Model:

**Possible Predictors (XVAR):**

YouTube Trailer hits, Days.between.Releases, FB.Following, Expected ROI %, Genre, Budget

**Predicted (YVAR):**

ROI.pct

**Model fit:**

When using Youtube Trailer Hits, FB Following, Trailer Thumbs Up and Days between Releases

lm(formula = ROI.pct ~ Expected.ROI.pct + Budget + FB.Following +

Youtube.Trailer.hits + Trailer.Thumbs.Up + Days.between.Releases +

Genre, data = Final)

Residuals:

Min 1Q Median 3Q Max

-353.02 -46.12 -0.66 37.07 758.47

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.415e+02 1.099e+02 2.197 0.0290 \*

Expected.ROI.pct 1.355e+00 8.569e-02 15.815 <2e-16 \*\*\*

Budget -1.293e+00 3.972e-01 -3.254 0.0013 \*\*

FB.Following 2.551e-05 1.007e-05 2.533 0.0119 \*

Youtube.Trailer.hits 9.183e-07 4.057e-06 0.226 0.8211

Trailer.Thumbs.Up 3.262e-04 1.166e-03 0.280 0.7799

Days.between.Releases -1.979e-01 2.387e-01 -0.829 0.4079

GenreAction -2.143e+02 1.127e+02 -1.902 0.0584 .

GenreAction / Drama -2.422e+02 1.353e+02 -1.790 0.0747 .

GenreAction / Sci -Fi -2.145e+02 1.656e+02 -1.295 0.1966

GenreAction / Sci Fi -1.247e+02 1.621e+02 -0.769 0.4424

GenreAction / Thriller -1.902e+02 1.579e+02 -1.204 0.2297

GenreAdult Comedy -1.654e+02 1.353e+02 -1.222 0.2230

GenreBiopic -2.009e+02 1.203e+02 -1.671 0.0961 .

GenreComedy -2.113e+02 1.110e+02 -1.903 0.0583 .

GenreComedy -2.087e+02 1.550e+02 -1.346 0.1796

GenreComedy / Action -2.437e+02 1.558e+02 -1.564 0.1192

GenreComedy / Drama -1.837e+02 1.156e+02 -1.590 0.1132

GenreCrime -2.351e+02 1.137e+02 -2.067 0.0398 \*

GenreCrime -2.788e+02 1.559e+02 -1.789 0.0749 .

GenreDance -3.103e+02 1.584e+02 -1.959 0.0513 .

GenreDrama -2.482e+02 1.110e+02 -2.235 0.0263 \*

GenreHorror -2.578e+02 1.132e+02 -2.276 0.0237 \*

GenreKids -2.692e+02 1.343e+02 -2.005 0.0460 \*

GenreNeo Noir -2.422e+02 1.552e+02 -1.561 0.1200

GenreReligious -2.367e+02 1.608e+02 -1.471 0.1425

GenreRom-Com -2.664e+02 1.115e+02 -2.389 0.0176 \*

GenreRomance -1.797e+02 1.143e+02 -1.571 0.1174

GenreSports -2.337e+02 1.554e+02 -1.503 0.1341

GenreSports -2.260e+02 1.553e+02 -1.456 0.1468

GenreThriller -2.192e+02 1.117e+02 -1.963 0.0508 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.4 on 241 degrees of freedom

Multiple R-squared: 0.5845, Adjusted R-squared: 0.5328

F-statistic: 11.3 on 30 and 241 DF, p-value: < 2.2e-16

Observation:

It is observed that only Expected ROI pct, Budget, FB Following and Genre hold any significance. Thus we continue working with only these variables on the model.

> train<-Final[1:200,]

> fit<-lm(formula=ROI.pct~ Expected.ROI.pct+Budget+FB.Following+Genre,data=train)

> summary(fit)

Call:

lm(formula = ROI.pct ~ Expected.ROI.pct + Budget + FB.Following +

Genre, data = train)

Residuals:

Min 1Q Median 3Q Max

-316.79 -41.15 -0.31 31.63 469.20

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.571e+02 9.416e+01 2.730 0.006986 \*\*

Expected.ROI.pct 1.191e+00 8.419e-02 14.145 < 2e-16 \*\*\*

Budget -1.106e+00 3.570e-01 -3.098 0.002277 \*\*

FB.Following 3.366e-05 9.438e-06 3.566 0.000469 \*\*\*

GenreAction -2.438e+02 9.722e+01 -2.508 0.013057 \*

GenreAction / Drama -2.811e+02 1.331e+02 -2.112 0.036104 \*

GenreAction / Sci -Fi -2.411e+02 1.377e+02 -1.751 0.081772 .

GenreAction / Sci Fi -1.642e+02 1.390e+02 -1.181 0.239185

GenreAction / Thriller -2.146e+02 1.344e+02 -1.596 0.112221

GenreAdult Comedy -1.892e+02 1.147e+02 -1.649 0.100902

GenreBiopic -1.682e+02 1.049e+02 -1.603 0.110690

GenreComedy -2.356e+02 9.510e+01 -2.477 0.014200 \*

GenreComedy -2.297e+02 1.326e+02 -1.732 0.085054 .

GenreComedy / Action -2.660e+02 1.329e+02 -2.001 0.046928 \*

GenreComedy / Drama -1.862e+02 1.016e+02 -1.833 0.068468 .

GenreCrime -2.610e+02 9.840e+01 -2.653 0.008721 \*\*

GenreCrime -2.875e+02 1.326e+02 -2.169 0.031466 \*

GenreDance -3.282e+02 1.359e+02 -2.415 0.016756 \*

GenreDrama -2.694e+02 9.533e+01 -2.826 0.005266 \*\*

GenreHorror -2.813e+02 9.724e+01 -2.893 0.004308 \*\*

GenreKids -2.855e+02 1.148e+02 -2.488 0.013808 \*

GenreNeo Noir -2.639e+02 1.327e+02 -1.988 0.048354 \*

GenreReligious -2.458e+02 1.326e+02 -1.854 0.065373 .

GenreRom-Com -2.943e+02 9.535e+01 -3.086 0.002360 \*\*

GenreRomance -2.085e+02 9.846e+01 -2.118 0.035619 \*

GenreSports -2.522e+02 1.326e+02 -1.902 0.058868 .

GenreThriller -2.245e+02 9.683e+01 -2.318 0.021602 \*

---

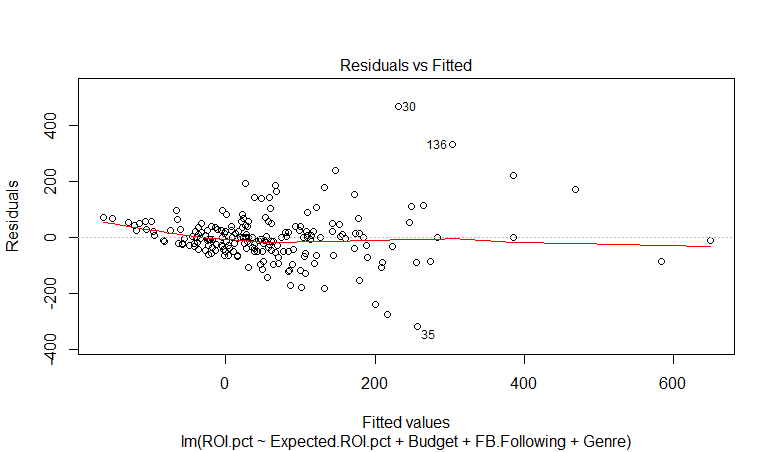
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

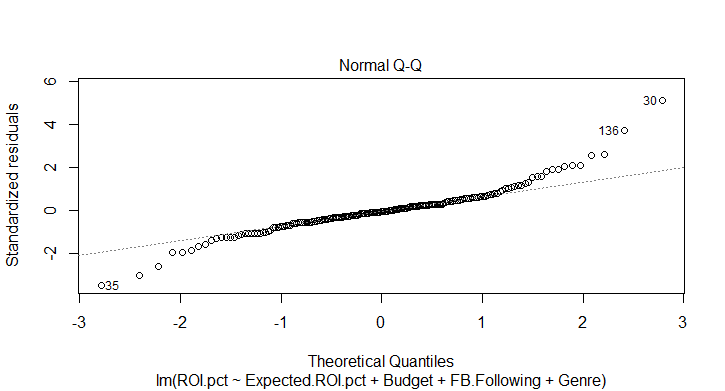
Residual standard error: 93.6 on 173 degrees of freedom

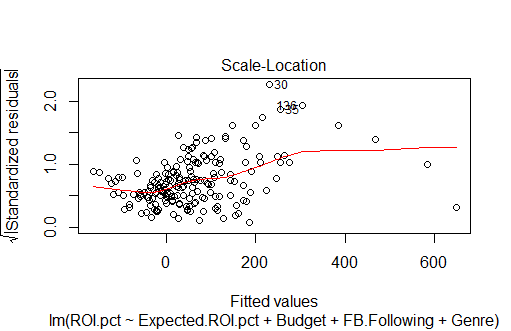
Multiple R-squared: 0.6333, Adjusted R-squared: 0.5782

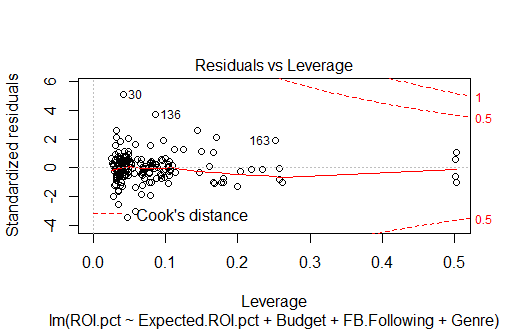
F-statistic: 11.49 on 26 and 173 DF, p-value: < 2.2e-16

On plotting the above model, we get the following graphs:









Testing the model on a sample of 100 records from the dataset:

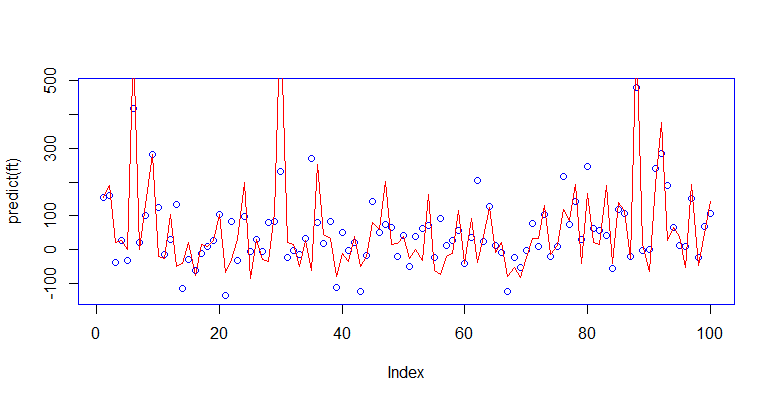
> test<-Final[1:100,]

> ft<-lm(formula=ROI.pct~ Expected.ROI.pct+Budget+FB.Following+Genre,data=test)

> plot(predict(ft),col="blue")

> lines(test$ROI.pct)

> lines(test$ROI.pct,col="Red")



The blue circles represent the model plotted points and the red line demarcates the actual ROI %.