UNIVERSITY of DENVER

Daniel Felix Ritchie School of Engineering and Computer Science

MOVIE RATING PREDICTION - BAYESIAN ANALYSIS

Pooja Radhakrishnan

ABSTRACT

Movie rating prediction sounds fascinating and is pretty useful for predicting the popularity of a movie so that the theatre coordinators can schedule movie timings accordingly. Also, for the users, it is very useful to determine whether to watch the movie or not. In this paper, the methods to predict a movie's rating is implemented using Bayesian Analysis. It includes various perspectives of analyzing the data and implementation of the method to achieve the result. The results obtained were close to the actual values.

RESEARCH QUESTION

Can a movie's rating be predicted having other determining attributes such as genre, title type, season of release, MPAA rating, runtime, IMDB and Tomatometer ratings, Critics score and Oscar win or lose data using Bayesian Analysis?

DATA SET

The dataset used for the analysis is obtained from www.imdb.com and www.rottentomatoes.com. It includes 32 attributes with about 651 films in the dataset having information about the movie title, title type, genre, MPAA rating, runtime, month of release, Oscar win/lose etc. A sample chunk of the dataset is shown in the figure below:

	title	title_type	genre	runtime	mpaa_rating	studio	thtr_rel_year	thtr_rel_month	thtr_rel_day
1	Filly Brown	Feature Film	Drama	80	R	Indomina Media Inc.	2013	4	19
2	The Dish	Feature Film	Drama	101	PG-13	Warner Bros. Pictures	2001	3	14
3	Waiting for Guffman	Feature Film	Comedy	84	R	Sony Pictures Classics	1996	8	21
4	The Age of Innocence	Feature Film	Drama	139	PG	Columbia Pictures	1993	10	1
5	Malevolence	Feature Film	Horror	90	R	Anchor Bay Entertainment	2004	9	10
6	Old Partner	Documentary	Documentary	78	Unrated	Shcalo Media Group	2009	1	15
7	Lady Jane	Feature Film	Drama	142	PG-13	Paramount Home Video	1986	1	1
8	Mad Dog Time	Feature Film	Drama	93	R	MGM/United Artists	1996	11	8
9	Beauty Is Embarrassing	Documentary	Documentary	88	Unrated	Independent Pictures	2012	9	7
10	The Snowtown Murders	Feature Film	Drama	119	Unrated	IFC Films	2012	3	2
11	Superman II	Feature Film	Action & Adventure	127	PG	Warner Bros. Pictures	1981	6	19
12	Leap of Faith	Feature Film	Drama	108	PG-13	Paramount Home Video	1992	12	18
13	The Royal Tenenbaums	Feature Film	Comedy	110	R	Buena Vista Distribution Compa	2002	1	4
14	School for Scoundrels	Feature Film	Comedy	100	PG-13	MGM	2006	9	23
15	Rhinestone	Feature Film	Comedy	111	PG	20th Century Fox	1984	6	20
16	Burn After Reading	Feature Film	Drama	96	R	Focus Features	2008	8	27
17	The Doors	Feature Film	Drama	140	R	Sony Pictures Home Entertainment	1991	3	1
18	The Wood	Feature Film	Drama	106	R	Paramount Pictures	1999	7	16
19	Jason X	Feature Film	Horror	91	R	New Line Cinema	2002	4	26
20	Dragon Wars	Feature Film	Drama	90	PG-13	Sony Pictures Home Entertainment	2007	9	13
21	Fallen	Feature Film	Drama	124	R	Warner Home Video	1997	6	1
22	The Gleaners and I	Documentary	Documentary	82	Unrated	Zeitgeist Films	2001	4	6
23	Tamara Drewe	Feature Film	Drama	107	R	Sony Pictures Classics	2010	10	8
24	Not Without My Daughter	Feature Film	Drama	116	PG-13	MGM Home Entertainment	1991	1	11
25	The Yes Men Fix the World	Documentary	Documentary	87	Unrated	Cinetic Media	2009	1	18
26	Oh, God! You Devil	Feature Film	Comedy	97	PG	WARNER BROTHERS PICTURES	1984	11	9
27	Meet Joe Black	Feature Film	Drama	178	PG-13	Universal Pictures	1998	11	13
28	The Bay	Feature Film	Horror	84	R	Roadside Attraction	2012	11	2
29	Scandal	Feature Film	Art House & International	115	R	HBO Video	1989	5	28
30	Imagine: John Lennon	Documentary	Musical & Performing Arts	100	R	WARNER BROTHERS PICTURES	1988	10	7

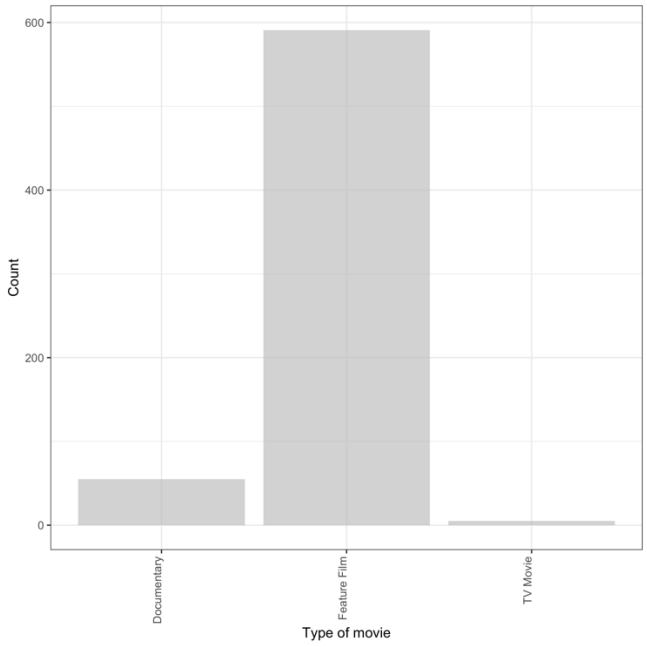
Showing 1 to 32 of 651 entries, 39 total columns

Console

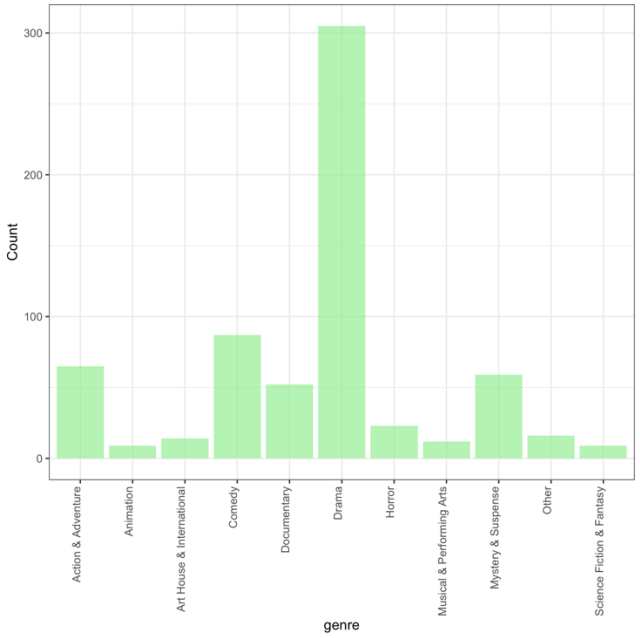
EXPLORATORY DATA ANALYSIS

Having the dataset, let's try to visually analyze the data to have more insight about the variables. Using *ggplot()* functionality of R, we plot the characteristics of the dataset as follows:

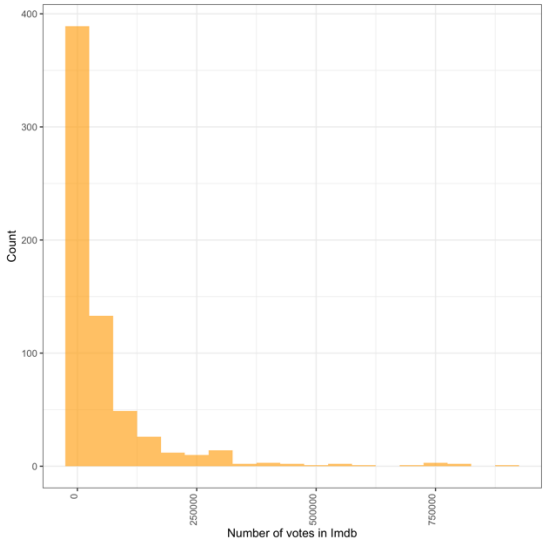
- Type of movie:



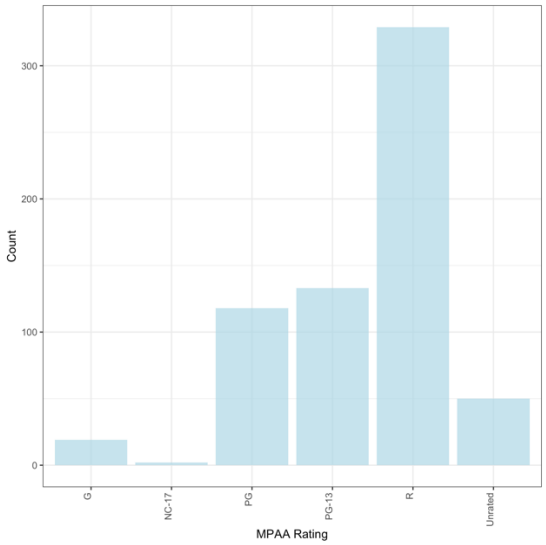
- Genre:



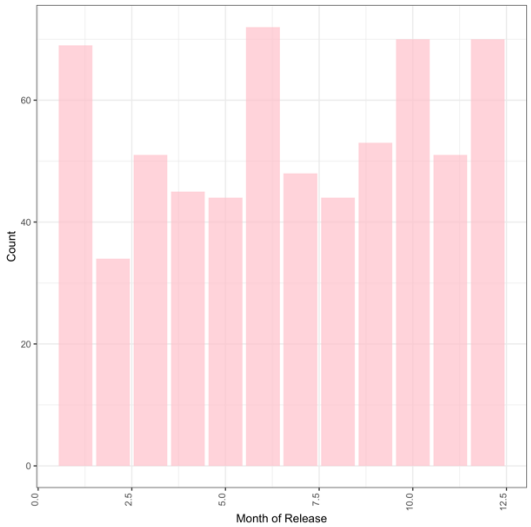
- Number of votes in IMDB.com



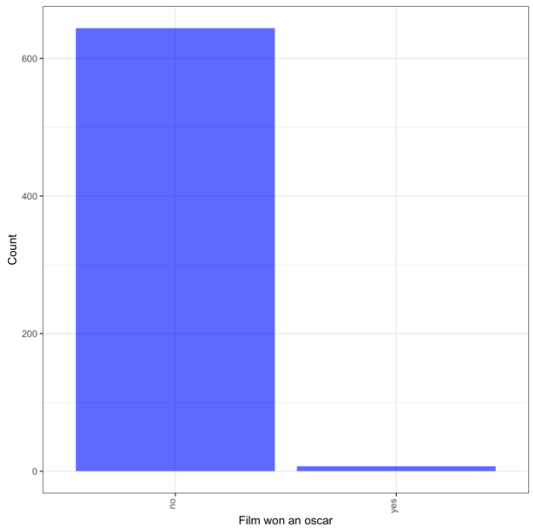
- MPAA rating



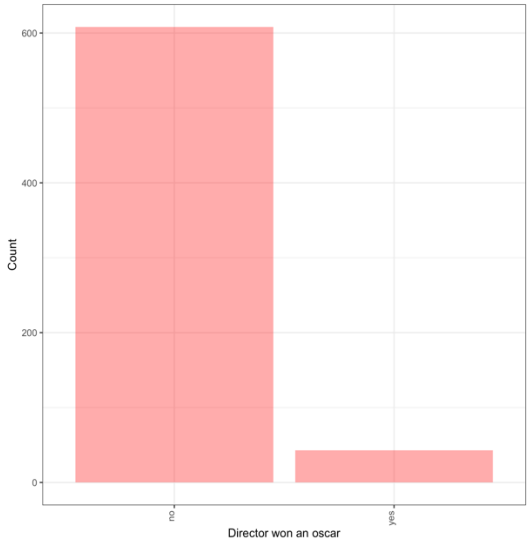
- Month of Release



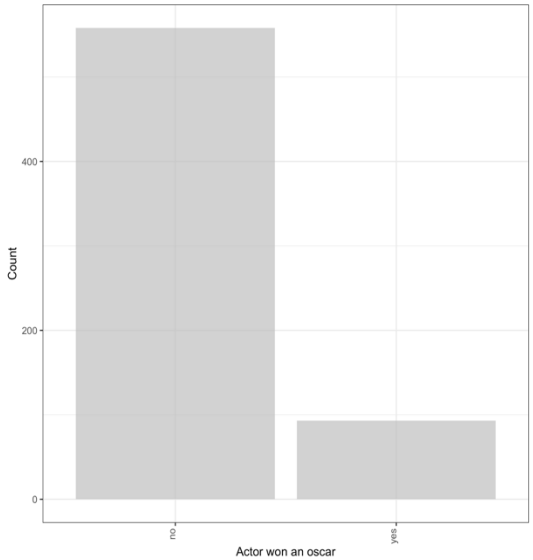
- Film won an Oscar?



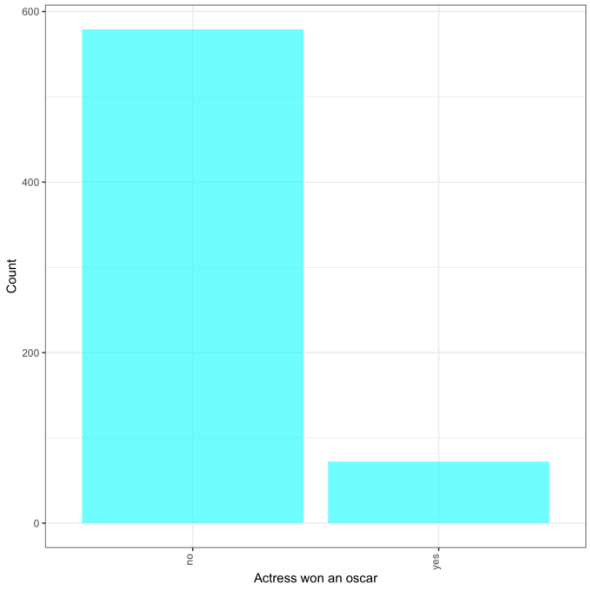
- Director won an Oscar?



- Actor won an Oscar?

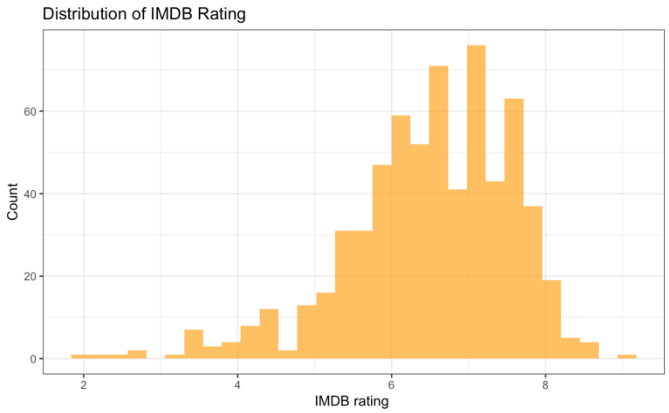


- Actress won an Oscar?

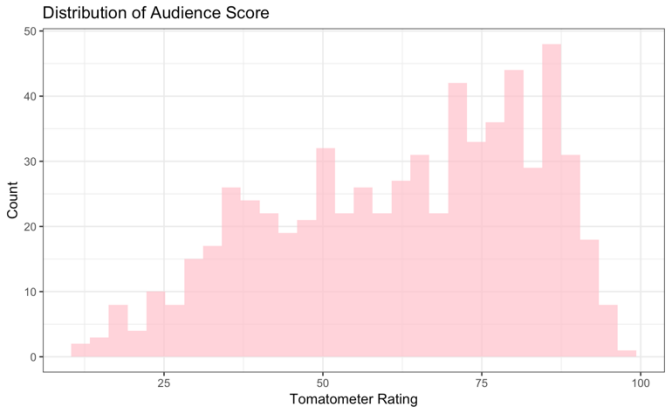


From the summary of the dataset above, let's try extracting the most dominant features and try to build a model. So extracting then *Feature Film* factor from *title_type* as there are 591 films under this factor out of 651, *'Drama'* from genre as there are 305 films under this factor, *R* factor from *mpaa_rating* as there are 329 films under this factor. As we have have Oscar nominations and winners data as well, we will consider that as a factor by creating a new field *oscar_season*. We'll also consider creating a new attribute *summer_season* as generally most of the movies get released during this time.

Now, let's try to choose our response variable: *audience_score (Tomatometer – Rottentomatoes* rating) or *imdb_rating.* Let's visually analyze it:

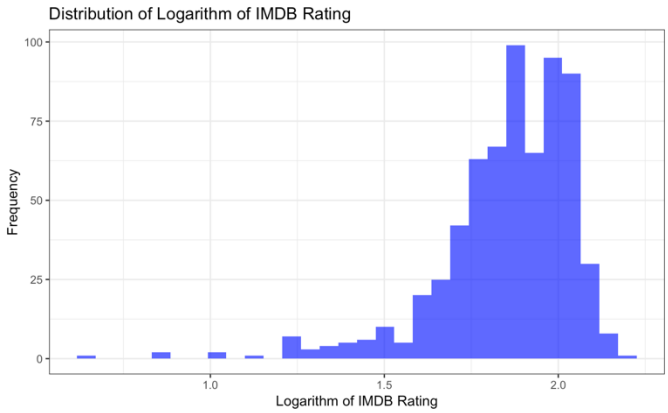


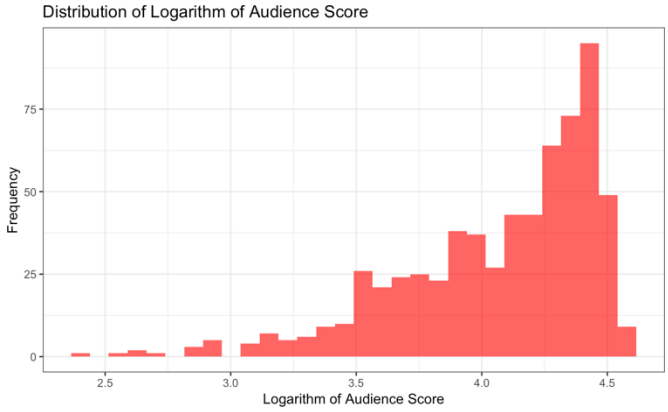
	mean	sd	median	IQR	min	max
1	6.493	1.085	6.6	1.4	1.9	9



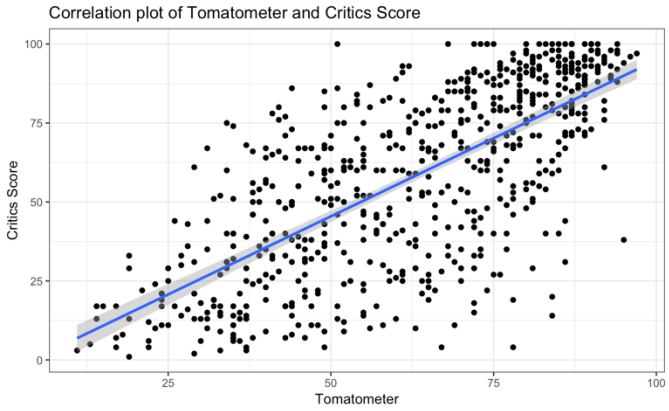
	mean	sd	median	IQR	min	max
1	62.363	20.223	65	34	11	97

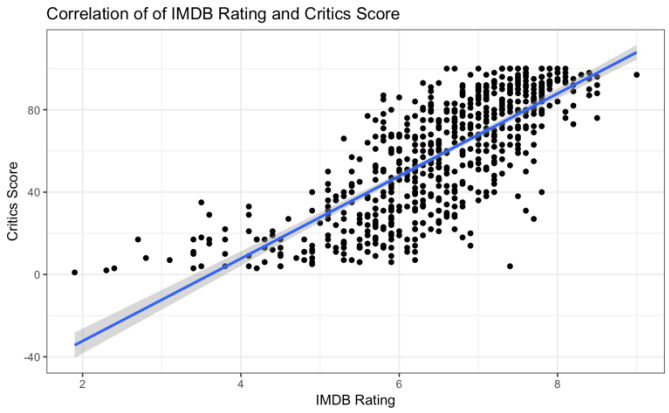
From the plots, it is evident that the distribution of IMDB ratings closely follows a normal distribution, whereas the Rottentomatoes rating is more uniform. So, let's consider *imdb_rating* to be our response variable. But to confirm this, let's consider the logarithmic graphs of the two variables to confirm the same.

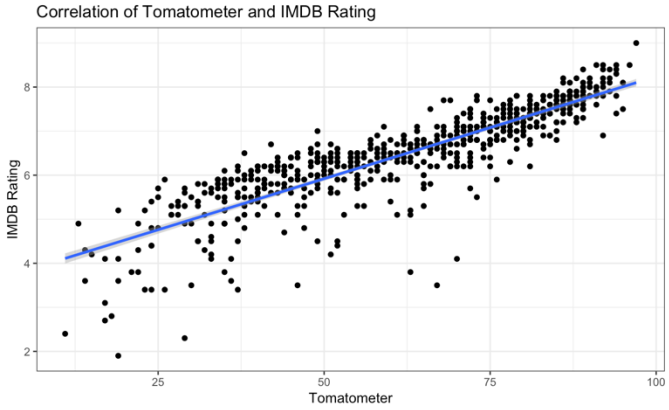


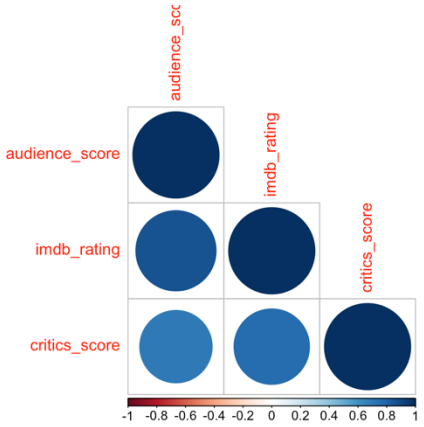


We can clearly see that the logarithmic graph of *audience_score* is highly right skewed and the IMDB rating graph appears to resemble normal distribution. Hence, we can confirm our assumption of *imdb_rating* to be our response variable.
The next goal to find out the predictor variables can be achieved by checking correlation between the variables. This is accomplished as follows:









The following interpretations are made:
- The first scatter plot between *Tomatometer* rating and the critics score is a bit spread out from the line, which conveys that they're not very highly correlated.
- The second scatter plot between *imdb_rating* and critics score seems to a bit better than the previous graph which conveys that they're not very highly correlated.
- The third scatter plot between *Tomatometer* rating and the *imdb_rating* seems to be highly correlated as the points are very close by.
- The fourth plot - Correlation plot also proves visually that *imdb_rating* and *Tomatometer* rating are highly correlated as it's in a darker shade of blue than the other combinations.
In addition to these plots, the correlation matrix also proves the same numerically:

	audience_score	imdb_rating	critics_score
audience_score	1.0000000	0.8648652	0.7042762
imdb_rating	0.8648652	1.0000000	0.7650355
critics_score	0.7042762	0.7650355	1.0000000

We can infer that, when we use *imdb_rating* as the response variable, we shouldn't use *audience_score* as a predictor variable as they're very highly correlated.

BAYESIAN MODELING

Bayes rule helps us to infer from the hypotheses of data. It relies mainly on two concepts: *Bayes Theorem* and *Conditional Probability.*

Conditional probability is given by:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

given that P(B) not equal to 0.

The conditional probability of an event B is the probability that the event will occur given the information that an event A has already occurred.

Bayes theorem is close to the Conditional probability. It states that the probability of an event B is nothing but the sum of the conditional probabilities of B given the information that event A has already occurred. When A and B are independent, P(B) is given as:

$$P(B \cap A_i) = P(B \mid A_i) * P(A_i)$$

Thus, substituting P(B) in the equation of conditional probability, we get:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{\text{Sum}(P(B \mid A_i) * P(A_i))}$$

This can be generalized as:

P(hypothesis | data) is given as:

$$\frac{P(data \mid hypothesis) * P(hypothesis)}{P(data)}$$

where,
P(hypothesis | data) – posterior probability
P(data | hypothesis) – likelihood
P(hypothesis) – prior probability
P(data) – Evidence

So, we can see that *Bayesian statistics* is a mathematical procedure that applies probabilities to statistical problems.

The key to build the model under the Bayesian approach is: *Bayes Factor.* The Bayes factor is the ratio of the likelihood probability of two competing hypotheses and it helps us to quantify the support of a model over another one. In Bayesian modelling, the choice of prior distribution is a key component of the analysis and can modify the result. But, the prior starts to lose weight when we add more data.

METHODOLOGY

The approach to address the research question involved the following steps:

- Exploratory Data Analysis
- Determination of the predictor variables
- Filtration of data to build the final dataset
- Try to determine the deterministic features
- Build the Bayesian Model
- Analyze the model
- Infer from the result of the model graphically
- Prediction

The preceding steps were applied to the dataset by analyzing the initial dataset, extracting the informative attributes needed, rebuilding the dataset for the model. The model was built upon this dataset and the graphical results of the model was observed. This led to determine the most determining attributes of the model and upon prediction, the model gave good results.

DATA SATISFACTION OF METHOD

The methodology was applied to predict the movie rating and popularity. Upon Exploratory Data Analysis being performed, the interesting predictor variables were determined by plotting them visually. New variables were also added to the dataset for analysis and the final dataset was built for the model after filtration.

The final dataset was analyzed to have an initial guess of what the highly influencing variables may be. A sample of that is shown below:

feature_film	n	Mean	Sd
no	60	7.53	0.7808
yes	591	6.39	1.056

drama	n	Mean	Sd
no	346	6.33	1.217
yes	305	6.67	0.8798

mpaa_rating_R	n	Mean	Sd
no	322	6.46	1.159
yes	329	6.52	1.008

oscar_season	n	Mean	Sd
no	460	6.43	1.093
yes	191	6.64	1.054

summer_season	n	Mean	Sd
no	443	6.54	1.076
yes	208	6.4	1.101

From the table summaries and the plots above, we can infer the following:
- A relationship between *feature_film* and *imdb_rating* is present in this dataset.

- A clear proof of there exists a relationship between *drama* and *imdb_rating* doesn't exist as the mean and variance are almost similar.
- There's no relationship between *mpaa_rating_R* and *imdb_rating* doesn't exist as the mean and variance is almost the same.
- There's no relationship between *oscar_season* and *`mdb_rating* doesn't exist as the mean and variance is almost the same.
- There's no relationship between *summer_season* and *imdb_rating* doesn't exist as the mean and variance is almost the same.

So, we can infer that *feature-film* has a stronger relationship with *imdb_rating* than the other variables. The process of building the model and the prediction is reviewed in the following sections.

Now, the dataset is good for the model to be built upon. This can be implemented by the inbuilt library **BAS** in R. The function *bas.lm()* has four parameters *: data, method, prior, modelprior.*

Method – A character variable indicating which sampling method to use

Prior – Prior distribution for regression coefficients
Modelprior – Family of prior distribution on the models

The dataset is the final dataset that we have with the variables that we are interested in. Markov Chain Monte Carlo (MCMC) is used as the method because it improves the search capability of the model.
Other alternative methods are: *'deterministic', 'BAS', 'MCMC + BAS'.*

Zellner-Siow Cauchy (ZS-null) and uniform methods are used as prior to assign equal probabilities to all the models.
Other alternative priors are*: "AIC", "BIC", "g-prior", "JZS", "ZS-full", "hyper-g", "hyper-g-laplace", "hyper-g-n", "EB-local", "EB-global".*
Other alternative modelpriors are: *"Bernoulli", "beta.binomial", "tr.beta.binoamial", "tr.poisson", "tr.prior.power".*

APPLICATION OF METHOD

The model is implemented as follows:

```
> library(BAS)
> bayes_dat_lm <- bas.lm(imdb_rating ~ ., data = bayes_dat,  method = "MCMC", prior = "ZS-null, modelprior = uniform())
```

After the model is built, the summary of the model is analyzed:

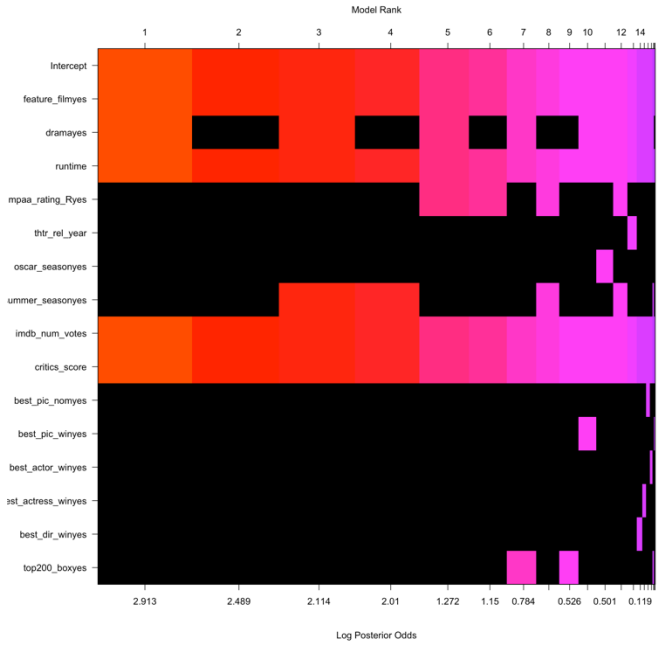
```
> bayes_dat_lm
> summary(bayes_dat_lm)
```

```
Call:
bas.lm(formula = imdb_rating ~ ., data = bayes_dat, prior = "ZS-null",
    modelprior = uniform(), method = "MCMC")


 Marginal Posterior Inclusion Probabilities:
         Intercept      feature_filmyes             dramayes              runtime     mpaa_rating_Ryes
           1.00000              0.99997              0.57483              0.98173              0.18113
      thtr_rel_year      oscar_seasonyes     summer_seasonyes       imdb_num_votes       critics_score
           0.06702              0.07672              0.33966              1.00000              0.99999
    best_pic_nomyes      best_pic_winyes    best_actor_winyes  best_actress_winyes      best_dir_winyes
           0.06185              0.08534              0.05710              0.06006              0.05841
     top200_boxyes
           0.11405

                  P(B != 0 | Y)  model 1     model 2     model 3     model 4     model 5
Intercept            1.00000000  1.0000   1.0000000   1.0000000   1.0000000   1.0000000
feature_filmyes      0.99996643  1.0000   1.0000000   1.0000000   1.0000000   1.0000000
dramayes             0.57483063  1.0000   0.0000000   1.0000000   0.0000000   1.0000000
runtime              0.98172913  1.0000   1.0000000   1.0000000   1.0000000   1.0000000
mpaa_rating_Ryes     0.18112640  0.0000   0.0000000   0.0000000   0.0000000   1.0000000
thtr_rel_year        0.06702271  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
oscar_seasonyes      0.07672424  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
summer_seasonyes     0.33966064  0.0000   0.0000000   1.0000000   1.0000000   0.0000000
imdb_num_votes       1.00000000  1.0000   1.0000000   1.0000000   1.0000000   1.0000000
critics_score        0.99998779  1.0000   1.0000000   1.0000000   1.0000000   1.0000000
best_pic_nomyes      0.06184540  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
best_pic_winyes      0.08534241  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
best_actor_winyes    0.05710297  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
best_actress_winyes  0.06006165  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
best_dir_winyes      0.05840912  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
top200_boxyes        0.11404877  0.0000   0.0000000   0.0000000   0.0000000   0.0000000
BF                           NA  1.0000   0.6524129   0.4482058   0.4023594   0.1831721
PostProbs                    NA  0.1771   0.1159000   0.0796000   0.0717000   0.0343000
R2                           NA  0.6408   0.6371000   0.6431000   0.6398000   0.6421000
dim                          NA  6.0000   5.0000000   7.0000000   6.0000000   7.0000000
logmarg                      NA 314.5823 314.1552445 313.7798194 313.6719125 312.8849932
```

The model yields the posterior inclusion probabilities of all the attributes. Since there are 16 attributes, 2^{16} models will be generated, and the best 5 appear in the summary in accordance with the Bayesian Factor (*BF*) value. The values 0 and 1 for all the models indicate the inclusion and exclusion of the attributes in the respective models. The Posterior Probabilities(*PostProbs*) of all the models, R value of the models(*R2*), dimensions and the logarithm of marginal likelihood(*logmarg*) of the models are also listed in the summary.

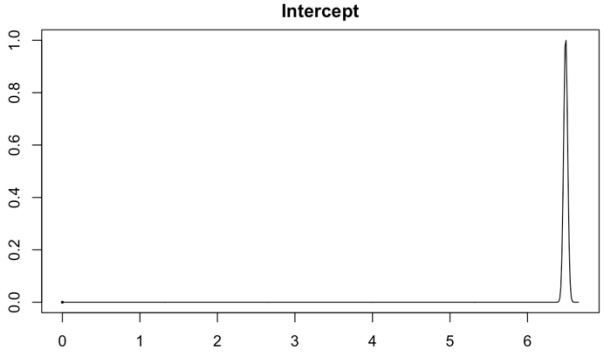
The following plot is used to visually interpret the Log Posterior Odds and the Model Rank.

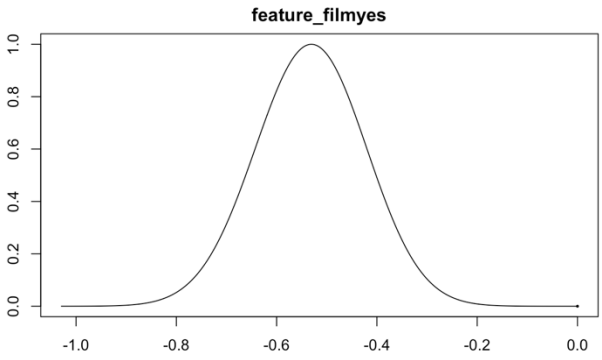
Visually interpreting the model:

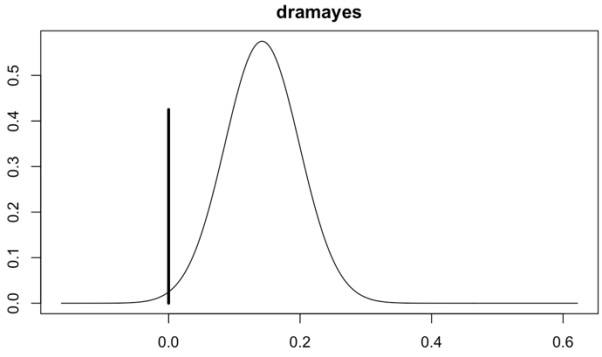


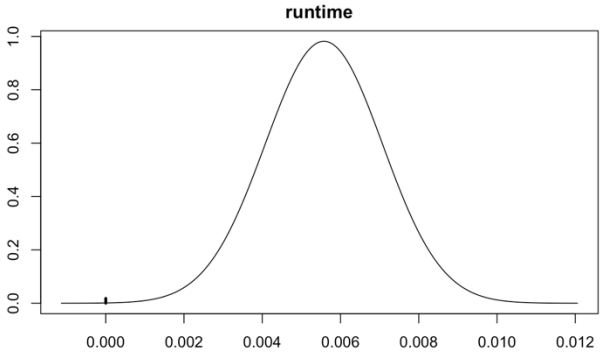
From the plot and the summary, we can infer that:
- *feature_film* appears in all the top 5 models with a mariginal probability of 0.99997864
- *runtime* appears in all the top 5 models with a mariginal probability of 0.98245544
- *critics_score* appears in all the top 5 models with a mariginal probability of 0.99995880
- *imdb_num_votes* appears in all the top 5 models with a mariginal probability of 0.99997711
- *drama* appears in 3 of 5 top models with a mariginal probability of 0.57723083

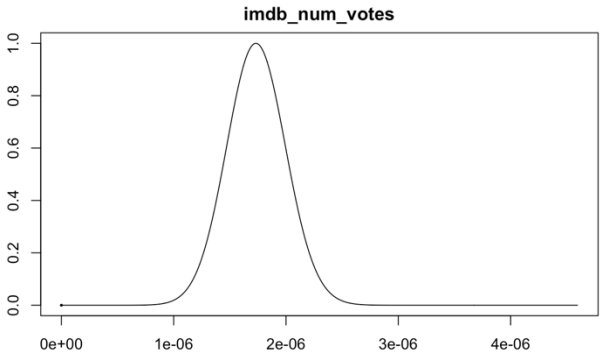
Hence, the best model includes the five attributes above along with the intercept.
The coefficients of the attributes of the model is derived to visually analyze the distribution:

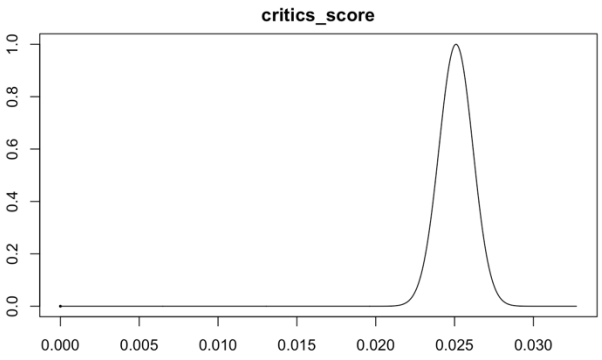






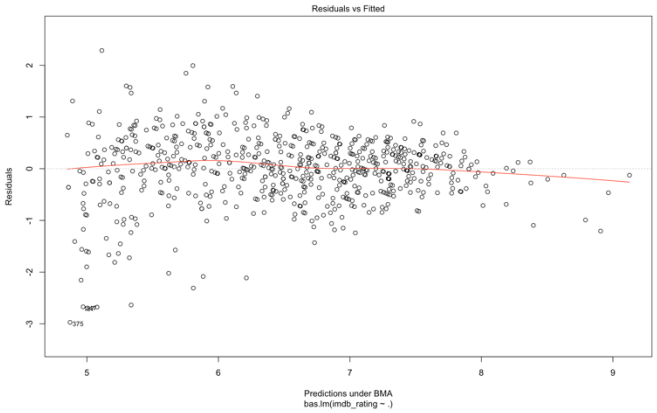




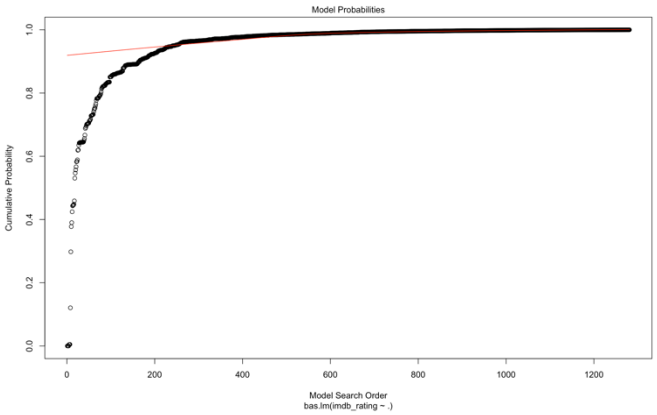


The plots above show the possible values of probability that each variable can take when the coefficient is non-zero.
The vertical line denotes the probability when the coefficients are 0.

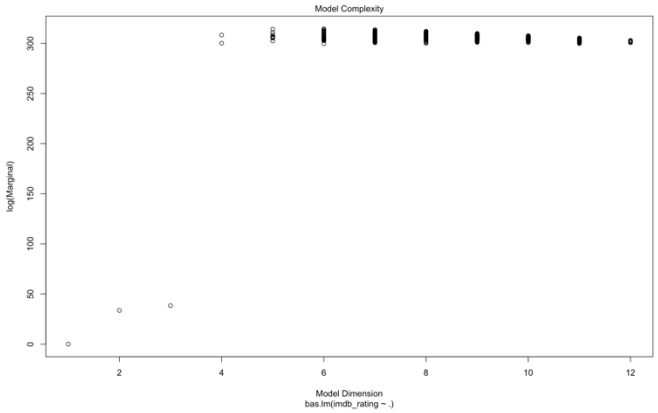
Now, let's analyze significance of the model graphically using the *plot* function with *which* parameter:



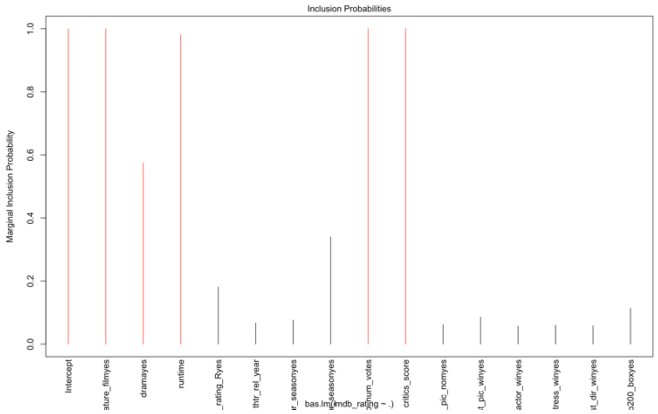
The 'Residuals vs Fitted' plot shows that there aren't many outliers except 2 with the variance being constant and the spread also being constant.



The 'Model Probabilities' curve shows a bend nearly at 300 and becomes flat after that. This plot shows the cumulative probabilities of all the models.



The 'Model Complexity' graph shows the dimension of each model, which is the number of regression coefficients with the intercept vs the log of the model's marginal likelihood. The highest log marginal is be reached between 5 to 12 dimensions.



The 'Inclusion Probabilities' graph shows the marginal posterior inclusion probabilities for all the variables. The red lines denote the marginal posterior inclusion probabilities that are greater than 0.5. This graph shows the final variables to be included for the model.

PREDICTION

Five movies were given as the input to test the Bayesian model's prediction capacity. The following chunk of code is used for the prediction. The model is given as the object, the new movie 'Manchester' is the new test data, indicating that BMA should be used as the estimator for the model, the type of interval calculation is given to be

predict and we include the standard errors of the model by setting it to TRUE.

> predict_1 <- predict(bayes_dat_lm, manchester, estimator="BMA", interval = "predict", se.fit=TRUE)

The following results were achieved for the five movies:

Movie <fctr>	Estimated.IMDB.rating <dbl>	Real.IMDB.rating <dbl>
Doctor Strange	7.568577	7.5

1 row

Movie <fctr>	Estimated.IMDB.rating <dbl>	Real.IMDB.rating <dbl>
Jungle Book	7.268279	7.4

1 row

Movie <fctr>	Estimated.IMDB.rating <dbl>	Real.IMDB.rating <dbl>
Lights Out	6.41153	6.3

1 row

Movie <fctr>	Estimated.IMDB.rating <dbl>	Real.IMDB.rating <dbl>
Manchester	7.915975	7.8

1 row

Movie <fctr>	Estimated.IMDB.rating <dbl>	Real.IMDB.rating <dbl>
Zootropolis	7.886496	8

1 row

RESULT

Bayesian modeling has achieved good result in terms of prediction as seen from the results above. There predicted value lies in the interval of -1 < real value < 1. We can see that the model's prediction is close to the actual IMDB rating. Thus, we have implemented the Bayesian regression model to successfully predict the movie rating using the IMDB dataset.

REFERENCES

- https://stackoverflow.com/questions/13279213/compare-bayesian-linear-regression-vs-linear-regression
- https://stat.ethz.ch/R-manual/R-devel/library/stats/html/predict.lm.html
- https://www.rdocumentation.org/packages/BAS/versions/1.5.3/topics/bas.lm
- https://www.analyticsvidhya.com/blog/2016/06/bayesian-statistics-beginners-simple-english/
- https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7
- https://rdrr.io/cran/BMA/man/plot.bic.html