PREDICTIVE MODELLING FOR FINANCIAL SUCCESS OF HOLLYWOOD MOVIES DATASET - CASESTUDY

BY- HARSHITA VED

Summary:

Predicting a movie's box office success before release is a challenging task because of the unpredictable nature of the problem. The report summarizes my attempt to perform the same on the dataset of 1196 movies with production year between 2007-2011. On a high level, I have explored Decision Tree method to classify movies on a scale of 1-9 i.e. flop-blockbuster.

Methodology:

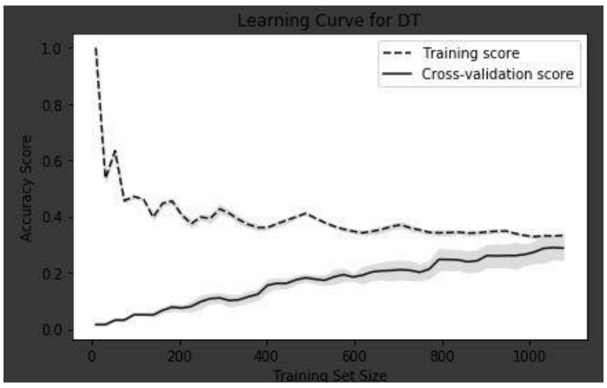
The "Hollywood movie" case study dataset has 1196 data points and is given in a tidy format. After loading the dataset, First, I tried to get intuition on how dependent variables and independent variables are related using Exploratory Data Analysis and by reading Dataset description. Checking the unique values of all features I could eliminate features "name", "id" and "display_name" as they were not informative for the response. Feature selection might not be very helpful as almost all features except "Language" contributed in determining "total"/"Category"(by EDA results). As the features "total" and "Category" are highly correlated and are missing from the Scoring sheet, we could predict "total" first and then the "Category". In other words, "total" would be a proxy variable for "Category". But the choice of arbitrary thresholds for each category might introduce bias. Also, the research paper suggests the "total" feature is industry secret and hence may not be good for generic models (keeping in mind that someone might want to replicate the same in Bollywood). Hence, I decided not to continue with the "total" column. Moving ahead with other remaining features that are all categorical, there was a need to one-hot encode them. I developed a general intuition about the dataset as follows, The "board_rating_reason" is more like the reasoning for the category labeled. Had "board_rating_reason" been a quantitative variable, then I believe it would have been the most correlated feature as it shows basically the reason for that category. The column "board_rating_reason" has textual data and hence I created tangible features out of it. The transformation was to fetch the top 15 keywords appearing in it and giving cumulative weights to each row. This was done using TfidfVectorizer and also manually. As the top word in the list has equal weightage as the top 15th word, I thought of using sentiments the boards might have used for labeling. Thus, analyzing the overall sentiments was also incorporated.

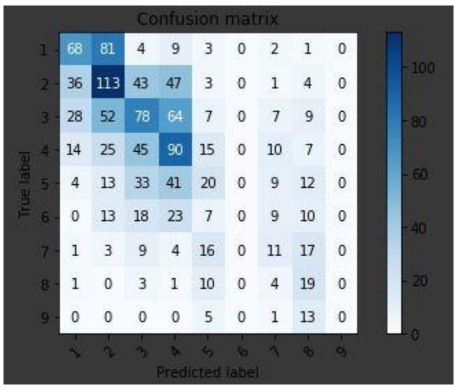
I am using **Decision Tree** as the primary model as it forces the consideration of all possible outcomes of a decision and traces each path to a conclusion and our dataset label is highly imbalanced. The parameters of the decision tree are summarized as below:

max_depth	10	Since there are not too many data points, we choose a small tree.
min_sample_leaf	50	For the samples in each leaf to be statistically significant, i.e. 5% confidence interval we need to have 50 samples per leaf.

Conclusion:

Using Decision Tree, I am able to obtain a 10 fold cross-validation accuracy of 29%. The Learning curve and confusion matrix are as follows.





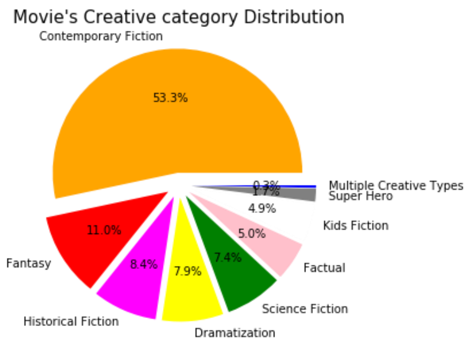
	1	2	3	4	5	6	7	8	9
1	68	81	4	9	3	0	2	1	0
2	36	113	43	47	3	0	1	4	0
3	28	52	78	64	7	0	7	9	0
4	14	25	45	90	15	0	10	7	0
5	4	13	33	41	20	0	9	12	0
6	0	13	18	23	7	0	9	10	0
7	1	3	9	4	16	0	11	17	0
8	1	0	3	1	10	0	4	19	0
9	0	0	0	0	5	0	1	13	0

PREDICTIVE MODELLING FOR FINANCIAL SUCCESS OF HOLLYWOOD MOVIES DATASET - CASESTUDY

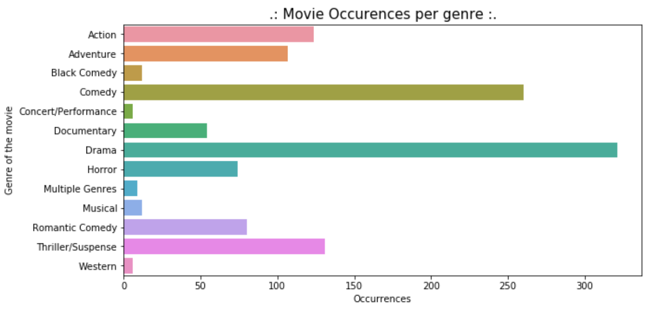
BY- HARSHITA VED

Appendix:

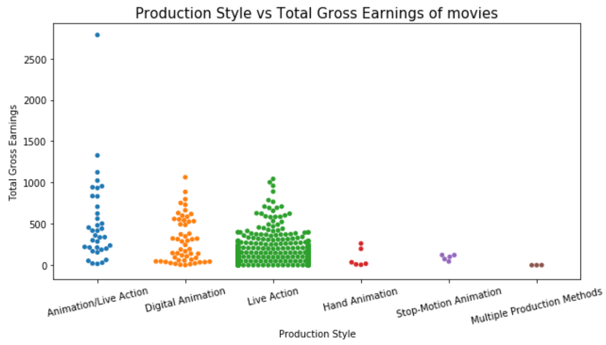
Exploratory Data Analysis



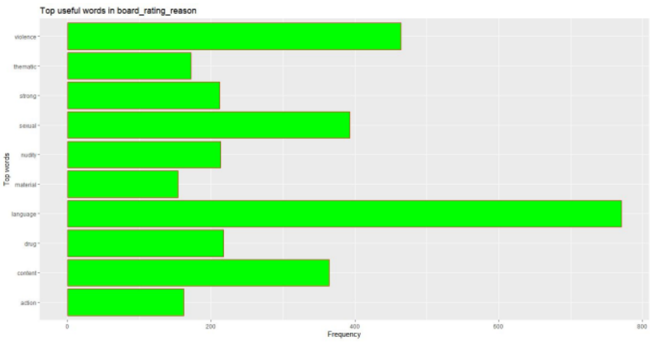
Contemporary Fiction is the top category with more than 50% share, but other categories are also present with considerable share.



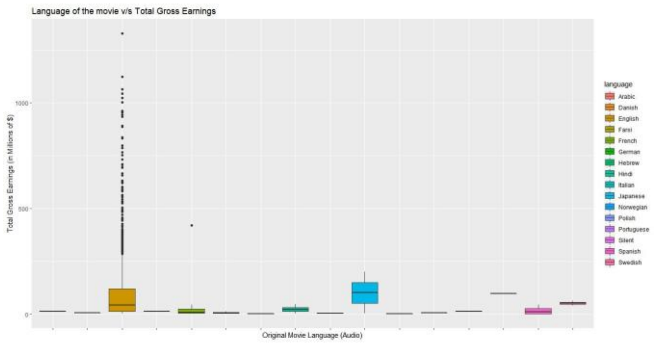
Drama and Comedy being prominently the top genre in the dataset indicate us to analyse the effect of them towards "total"/"category".



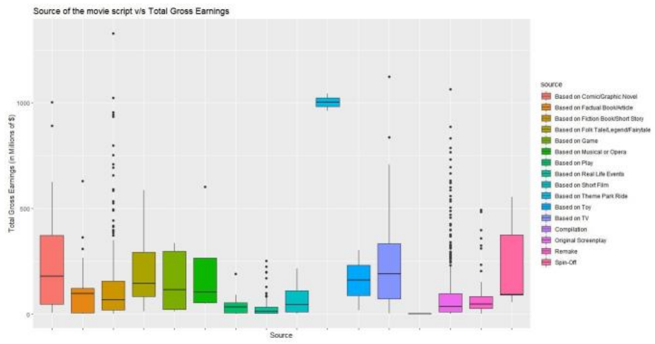
Production Style of "Live Action" contributes most towards Total gross earnings followed by Digital Animation. Impact of multiple category towards "total" suggests we cannot drop this feature.



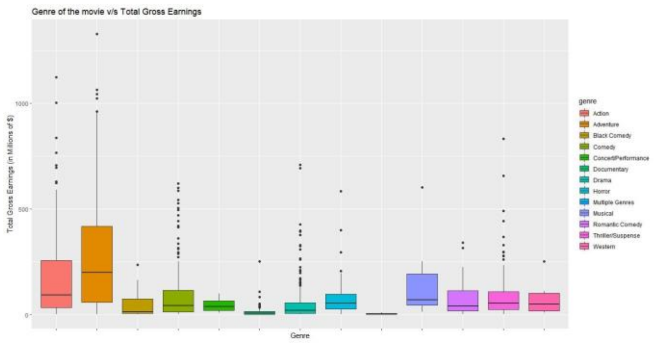
Top 10 useful words in "board_rating_reason" are language, violence, sexual, content, drug, nudity, action etc. suggesting a possibility of sentiment analysis on the feature as well.



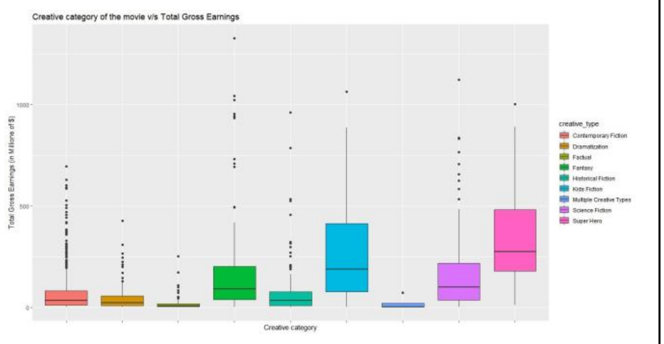
English and Japanese are the most visible original movie language, all others are almost negligible. The plot suggests that this feature can be dropped.



Almost every "source" has a measurable impact on Total gross earnings except "compilation", "Comic/Graphic Novel" being on top. There is no prominent skewness hence we should keep this feature.



Main genres of movies generating earnings are "Adventure" and "Action". However we cannot ignore other genres as the number of categories is high and others have considerable contribution too.



Leading creative categories are "Superhero" and "Kids Fiction". The contribution of the creative category towards "total" is majorly driven by these two and slightly by "Fantasy" and "Science Fiction".