Feature Analysis of Movie Attributes

Madhusudan Malhar Deshpande, Arif Sarfaraz Waghbakriwala, Reem Ghabayen, Aanand Dhandapani

Summary:

A Movie can be classified as a successful movie or a 'hit' based on the revenue it earns for the budget put into its making. Of course, the story of the Movie plays a vital role in determining that. However, there could be hidden patterns in the movie attributes that have more influence over others in deciding whether a movie will be a hit or not. By analyzing various attributes of a movie, this project aims to achieve that.

The project will consist of Exploratory Data Analysis and Statistical Modelling of a movies database from Kaggle – "The Movies Dataset." This database consists of data of ~45,000 movies released before July 2017. The data consists of the Movie attributes such as Cast, Crew, Plot keyword, Budget, Revenue, Posters, Release date, Languages, Production, Companies, Countries, IMDB votes, and Vote average.

Project Plan:

Importing and Tidying Data: The primary data table consists of about 45,000 movies released before July 2017. Data tables, specifically 'IMDB Ratings' and 'IMDB Movies' will be used to source information of a few attributes into the primary data table to construct a database that will be used for performing further analysis. The data tables are not in tidy format and will be parsed to create a tidy and usable database.

Exploratory Data Analysis: Univariate data visualization summaries will be used to explore the relationship between the dependent and independent variables. These summaries will be helpful to get insights into the correlation between any Movie attribute and how successful a movie is. This will also be used to detect any data anomalies and remove them.

Feature Analysis: The influence of each attribute will be estimated through Feature Importance Analysis using regression, where the parameters of predictor variables can be viewed as 'Feature Importance.' Following are the predictor variables: Cast, Director, Release Year, Languages of Production, Popularity Score, IMDB Rating, Adult Rated (Y/N), Genre, Production Countries, IMDB Vote Averages. The ratio of Global Revenue and Budget will serve as the dependent variable.

Preliminary results:

The primary data table consists of data ~45,000 Movies. The dependent variable, which is being calculated as ratio of Revenue and Budget, and Cast, Director and Ratings data is being sourced from a different IMDB database, consisting of data of ~85,000 movies. However, post merging this data, only ~ 20,000 movies have valid data and only these records will be used for analysis. The Genre Data is in key-value format and will be extracted before it can be used. Categorical Variables will be trimmed down to have only top 5 categories, and the remaining values will be clubbed under the category 'Others'. (e.g., Variables like Language, Director, Country etc.)

References:

- Banik, R. (2017, November 09). *The movies dataset*. Kaggle. Retrieved November 16, 2021, from https://www.kaggle.com/rounakbanik/the-movies-dataset/code.
- Leone, S. (2020, September 14). *IMDB movies extensive dataset*. Kaggle. Retrieved November 16, 2021, from https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset.

The Movie Data set

This is our primary data table with ~45,000 rows of data.

```
movie_metadata <- read.csv("D:\\NEU\\IDMP\\Project\\Movi.csv")

movie_metadata <- subset(movie_metadata,
                         select=-c(budget,ratings,
                                   revenue,Hit.Not.Hit,Gross.Income,
                                   budget.1,asd,X,imdb_title_id,
                                   worlwide_gross_income,budget.2
                                   ))

head(as_tibble(movie_metadata))
```

```
## # A tibble: 6 x 22
##   adult belongs_to_collection genres    homepage  id    imdb_id original_langua~
##   <chr> <chr>                 <chr>     <chr>     <chr> <chr>   <chr>
## 1 FALSE ""                    [{'id': ~ ""        3418~ tt4679~ ml
## 2 FALSE ""                    [{'id': ~ ""        3209~ tt5128~ ta
## 3 FALSE ""                    [{'id': ~ ""        49029 tt0376~ ta
## 4 FALSE ""                    [{'id': ~ "http://~ 25121 tt1417~ ta
## 5 FALSE ""                    [{'id': ~ ""        66526 tt0220~ ta
## 6 FALSE ""                    [{'id': ~ ""        47940 tt0213~ ta
## # ... with 15 more variables: original_title <chr>, overview <chr>,
## #   popularity <chr>, poster_path <chr>, production_companies <chr>,
## #   production_countries <chr>, release_date <chr>, runtime <int>,
## #   spoken_languages <chr>, status <chr>, tagline <chr>, title <chr>,
## #   video <lgl>, vote_average <dbl>, vote_count <int>
```

IMDB Ratings File

This file is used for sourcing ratings attribute for our primary data table. It contains about 85855 rows of data.

```
ratings <- read_csv("D:\\NEU\\IDMP\\Project\\IMDb ratings.csv\\IMDb ratings.csv")
head(ratings)
```

```
## # A tibble: 6 x 49
##   imdb_title_id weighted_average_vote total_votes mean_vote median_vote votes_10
##   <chr>                         <dbl>       <dbl>     <dbl>       <dbl>    <dbl>
## 1 tt0000009                       5.9         154       5.9           6       12
## 2 tt0000574                       6.1         589       6.3           6       57
## 3 tt0001892                       5.8         188       6             6        6
## 4 tt0002101                       5.2         446       5.3           5       15
## 5 tt0002130                       7          2237       6.9           7      210
## 6 tt0002199                       5.7         484       5.8           6       33
## # ... with 43 more variables: votes_9 <dbl>, votes_8 <dbl>, votes_7 <dbl>,
## #   votes_6 <dbl>, votes_5 <dbl>, votes_4 <dbl>, votes_3 <dbl>, votes_2 <dbl>,
## #   votes_1 <dbl>, allgenders_0age_avg_vote <dbl>, allgenders_0age_votes <dbl>,
## #   allgenders_18age_avg_vote <dbl>, allgenders_18age_votes <dbl>,
## #   allgenders_30age_avg_vote <dbl>, allgenders_30age_votes <dbl>,
## #   allgenders_45age_avg_vote <dbl>, allgenders_45age_votes <dbl>,
## #   males_allages_avg_vote <dbl>, males_allages_votes <dbl>, ...
```

IMDB Movie File

This file is used for sourcing Budget, Gross Revenue attributes for our primary data table. It contains about 85855 rows of data.

```
movie <- read_csv("D:\\NEU\\IDMP\\Project\\IMDb movies.csv\\IMDb movies.csv")
head(movie)
```

```
## # A tibble: 6 x 22
##   imdb_title_id title original_title  year date_published genre duration country
##   <chr>         <chr> <chr>          <dbl> <chr>          <chr>    <dbl> <chr>
## 1 tt0000009     Miss~ Miss Jerry      1894 1894-10-09     Roma~       45 USA
## 2 tt0000574     The ~ The Story of ~  1906 1906-12-26     Biog~       70 Austra~
## 3 tt0001892     Den ~ Den sorte drøm  1911 1911-08-19     Drama       53 German~
## 4 tt0002101     Cleo~ Cleopatra       1912 1912-11-13     Dram~      100 USA
## 5 tt0002130     L'In~ L'Inferno       1911 1911-03-06     Adve~       68 Italy
## 6 tt0002199     From~ From the Mang~  1912 1913           Biog~       60 USA
## # ... with 14 more variables: language <chr>, director <chr>, writer <chr>,
## #   production_company <chr>, actors <chr>, description <chr>, avg_vote <dbl>,
## #   votes <dbl>, budget <chr>, usa_gross_income <chr>,
## #   worlwide_gross_income <chr>, metascore <dbl>, reviews_from_users <dbl>,
## #   reviews_from_critics <dbl>
```