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 consisted of Exploratory Data Analysis and Statistical Modelling of a movies database from Kaggle – "The Movies Dataset". This database consisted of attributes such as Cast, Crew, Plot keyword, Budget, Revenue, Posters, Release date, Languages, Production, Companies, Countries, IMDB votes, IMDB Ratings, and Vote average.

The project goal is achieved by the following 3 steps.

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

Exploratory Data Analysis: Univariate data visualization summaries are used to explore the relationship between the dependent and independent variables. These summaries are helpful to get insights into the correlation between any Movie attribute and how successful a movie is.

Feature Analysis: The influence of each attribute is estimated through Feature Importance Analysis using classification regression. Also, multiple regression models are run to compare performance of each model. The parameters chosen by the model can be viewed as decisive features. Following are the predictor variables: Release Year, Languages of Production, Popularity Score, IMDB Rating, IMDB Vote Count, Adult Rated (Y/N), Genre, Production Countries, Movie Run-time. The target variable is the ratio of Global Revenue and Budget, converted into a 1/0 to categorize a movie as Hit or Not Hit.

Data Sources

The Movie Dataset: This is the primary data set. It contains data of ~45k movies and has following columns:

Adult: A Boolean flag that tells whether a movie was 'Adult' rated or not
Belongs to collection: Tells the collection a movie belongs to (e.g., Lord of the Rings collection)
Budget: The budget of a movie. However, budget will be sourced from a different file as this column contains a lot of missing values
Genres: All the genres of a movie
Homepage: URL Link to the homepage of a movie
ID: Numeric ID of a movie
IMDB ID: Standard 9 Character IMDB ID of the format <ttxxxxxxx>

Original Language: Keys of Original Language of the movie in encoded as a key (e.g., ‘en’ for English)
Original Title: Original title of the movie in original language script
Overview: Summary about the story of the movie
Popularity: Continuous variable with a popularity score
Poster Path: JPG Image file name of the movie poster
Production Companies: Production house of a movie
Production Countries: Production countries of a movie
Release Date: Release date of a movie.
Revenue: Revenue earned. However, revenue will be sourced from a different file as this column contains a lot of missing values
Run-time: Run-time of a movie in minutes
Spoken Languages: All languages of the movie in {key:value} format
Status: Categorical variable showing status of the movie
Tagline: Tagline of a movie
Title: Title of the movie in English
Video: A Boolean Flag
Vote Average: IMDB rating of a movie. However, ratings data will be sourced from a different file as this column contains lot of discrepancies
Vote Count: #Ratings of a movie

Only the columns in bold are used from this file.

IMDB Movies dataset: This data set is a repository of ~85k movies from IMDB. It was used to source Budget, Revenue and Release Year.

Budget: Budget of a movie, not necessarily in USD
World-Wide Gross: Revenue earned by a movie globally
Year: Release Year of a movie

IMDB Ratings dataset: This data set is a repository of ~85k movies from IMDB. It is used to source only ratings of movies. The ratings data is continuous variable with range from 0-10

Methods:

Data Tidying

The data contained a lot of issues and needed to be cleaned before use. The following steps are used to prepare and tidy the data:

1. Merge main data, ‘The Movie Dataset’ with ‘IMDB Movies Data’ and ‘IMDB Ratings Data’ to source Budget, Revenue, Release Year and Ratings of movies
2. Select Relevant features
3. “Genres” and “Production Countries” data was present in a form of {key:value} pair and needed to be split. Further, each cell was populated with multiple {key:value} pairs, indicating that a movie has multiple genres and multiple production countries. The first step to tidy this data was to extract only

the *values* from the {key:value} pairs. This resulted in all genres/production countries being populated in a comma separated format. It was observed that overall, there were 20 unique genres and 111 production countries. Hence, 20 columns for each genre and 111 columns for each country were added to the dataset. Using the comma separated genres and countries in the first step, the 20 genre columns and 111 country columns were populated as 1 or 0. Thus, all genres and production countries a movie were obtained as separate features

4. Revenue and Budget Columns were populated with different currencies. Hence, they were converted to a single currency (USD) to bring the values on the same scale using the 'priceR' package. This package uses an API that sources exchange rates against USD of the current day from the world bank.
5. Y-Variable/Target variable: The Y-variable/Target variable of the model was created from the data. The ratio of the revenue and the budget was used to categorize a movie as 'Hit' or 'Not Hit'. The value of the ratio greater than 1 was categorized as 'Hit' and less than 1 was categorized as 'Not Hit'.

Exploratory Data Analysis:

Data scientists with the help of Exploratory data analysis (EDA) analyze and explore datasets and document important characteristics utilizing data visualization methods. It helps to determine how best to manipulate data sources to get the answers needed, and enables data scientists to discover patterns, identify anomalies, test hypotheses, or test assumptions. EDA is mainly used for studying what information can be inferenced other than formal modeling or hypothesis testing, which gives a better understanding of the variables in a data and the correlations between them. It can also help to determine if a particular statistical analysis method considered is suitable for analyzing the data. First developed by American mathematician John Tukey in the 1970s, the EDA method is still widely used in data retrieval processes.

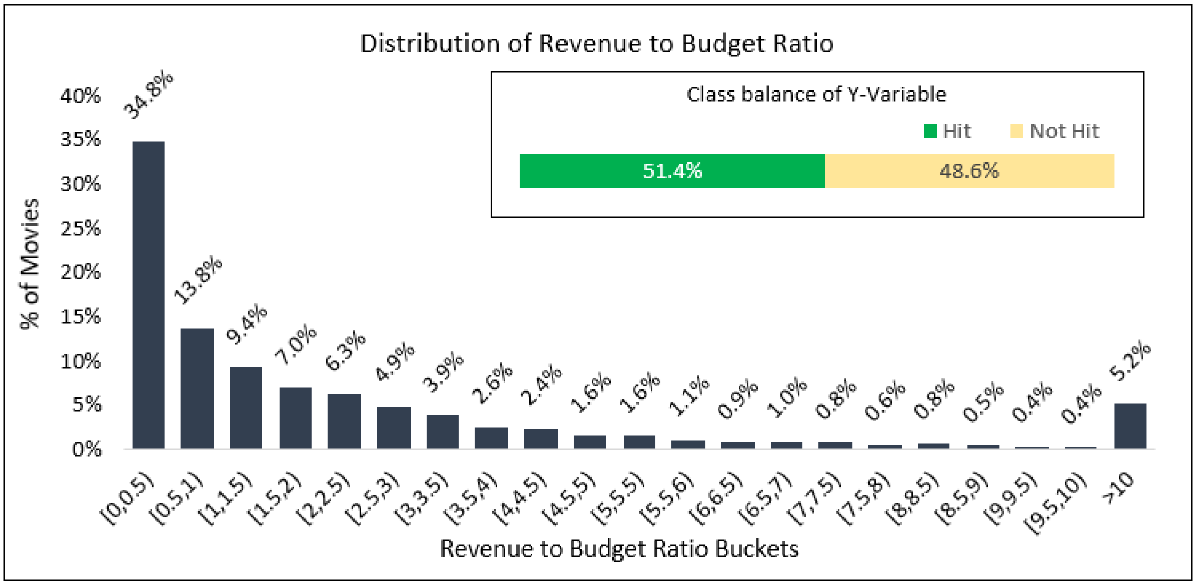
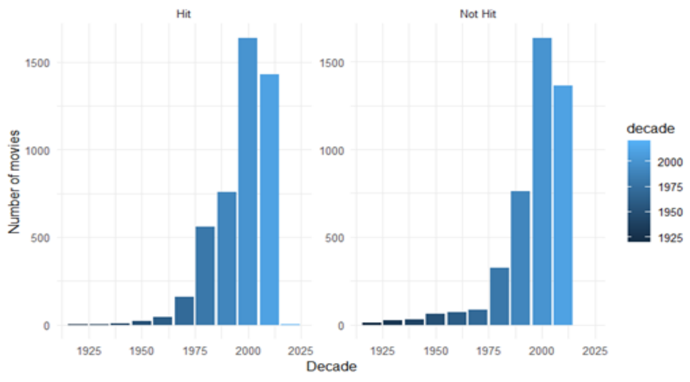
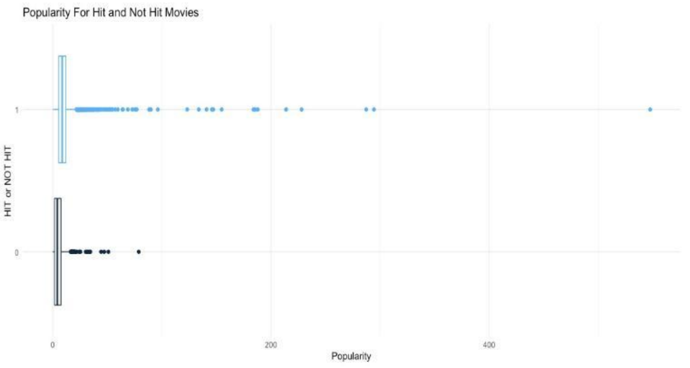
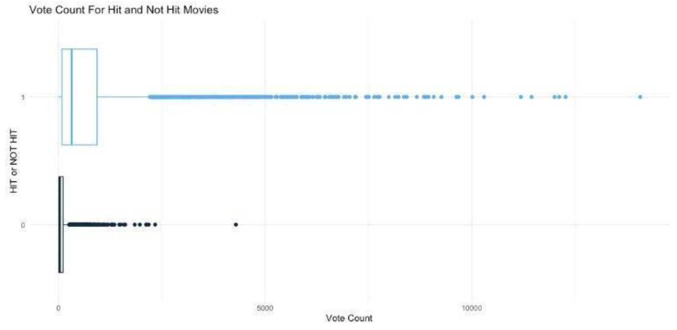


Figure 1: Distribution of Response Variable

Distribution of the Response Variable:

The response variable is calculated based on the ratio of already existing variables named revenue and budget. If the ratio of the revenue to budget comes out to be greater than 1, meaning profits were generated, it is marked as a hit movie, or is not a hit otherwise. The distribution of the response variable can be seen from figure 1. As expected, the number of movies decreases with increase in Revenue to Budget ratio. Also, 51.4% of movies have Revenue to Budget ratio greater than 1. Thus 51.4% of the movies are categorized as 'Hit' and remaining as 'non-Hit'. The Y-variable thus has a balanced class.

Relationship of variables with Y-Variable:

 *Figure 2: Time Series distribution of movies*	Looking at *Figure 2*, we see that there is no strong correlation between the year a movie was released and whether it will be a hit or not. Another inference is that as time passes by, the number of hit and non-hit movies have increased. As expected, this variable does feature in the model
 *Figure 3: Distribution of Popularity Score*	Looking at *Figure 3*, Popularity has a positive relationship with our target variable, Popularity is relatively more right skewed for hit movies. The median popularity is higher for hit movies, and both hit and non-hit movies are right skewed for Popularity. As expected, Popularity score features in the model
 *Figure 4: Distribution of Vote Count*	Looking at *Figure 4*, Vote Count has a positive relationship with our target variable. The median Vote Count is higher for hit movies and both hit and not hit movies are right skewed for Vote Count. As expected, Vote Count features in the model

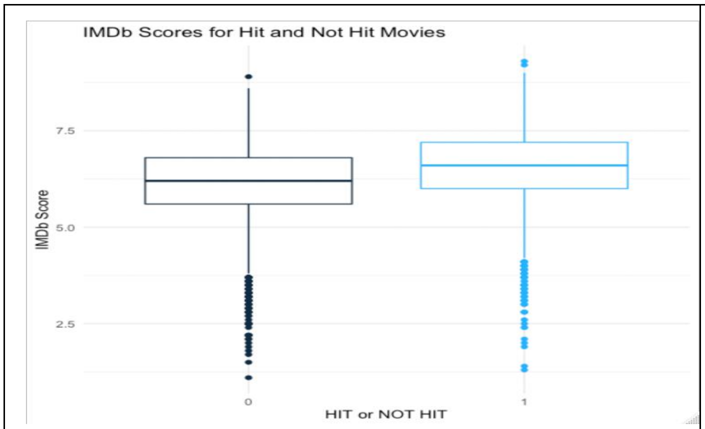


Figure 5: Distribution of IMDb Rating

Looking at *Figure 5*, Hit Movies have higher IMDB scores than non-Hit movies. As expected, IMDB Rating features in the model

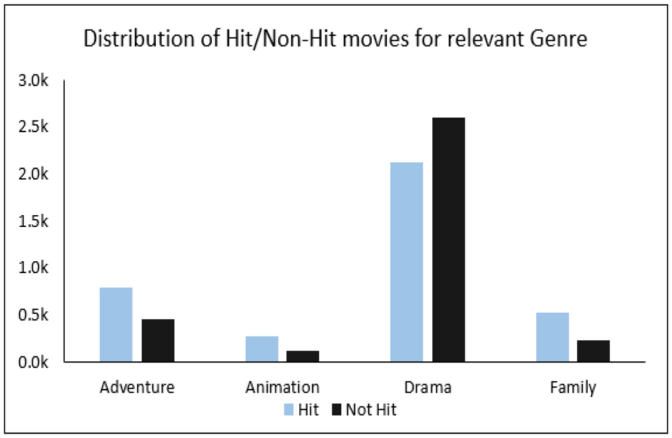


Figure 6: Distribution of Movies by Genre

From *Figure 6*, it can be inferred that Family, Animation and Adventure genres have higher percentage of hit movies which is also reflected in the correlation Matrix. On the other hand, Drama has a negative correlation which can be inferred from the adjacent figure. As expected, these four Genres feature in the model.

MODELLING:

Modelling is the methodology of trying to make a Machine Learning model to take in our dataset and predict a target variable. It also includes minimizing error, optimizing feature weights, and evaluating the model against different metrics such as Accuracy, Precision, Recall, F1-score, Area Under the Curve (AUC) etc. Four supervised machine learning models such as Logistic Regression, Decision Tree, KNN and Random Forest are chosen as models to train and test with our dataset.

LOGISTIC REGRESSION:

Logistic regression is a statistic that uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a form of binary regression. Mathematically, a binary logistic model has a dependent variable with two possible values which is represented by an indicator variable, where the two values are labeled either 0 or 1.

K-NEAREST NEIGHBOURS (kNN):

The nearest neighbor (kNN) algorithm is a nonparametric classification method. Used for classification and regression. In either case, the input consists of the k closest training examples in the data set. Results depend on whether kNNs are used for classification or regression

DECISION TREE:

A decision tree is a flowchart-like structure in which each inner node represents a "test" for an attribute (such as whether a coin toss is a heads or tails), each branch represents a test result, and each leaf node represents a class. brand. (Determined after calculating all properties). The path from root to leaf is a classification rule.

RANDOM FOREST:

Random Forests or Random Decision Forests are ensemble training techniques for classification, regression, and other problems that work by constructing multiple decision trees during training. For classification problems, the output of a random forest is the class chosen from most trees. For regression problems, the mean or mean prediction of the individual trees is returned.

Choosing Model Predictors:

From the *Figure 7* below, there is strong correlation between surprising factors like Drama which has negative correlation with our target variable.

Movies from Countries like Belgium, Luxembourg has positive correlation with our target variable.

Family centered genres like Animation, Comedy, Family have strong correlation with our target feature.

A heat map of correlation coefficients describes the measure of relationship between variables on one axis with the variables on the other axis. According to the palette used here, the lighter the color the greater the relationship. The correlation heat map shown in the figure 7 shows the top features, based on correlation with the Y-variable. These features were use as the predictor variable.

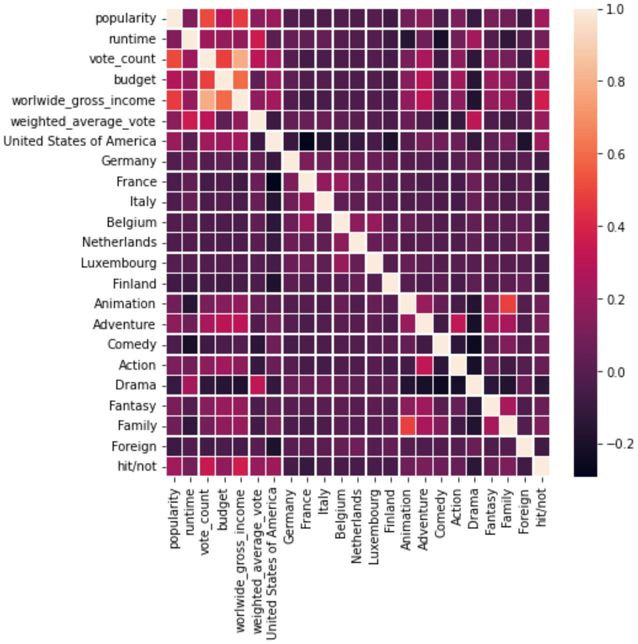


Figure 7: Correlation Heat Map for the Predictor Variables

Results:

Model	Specificity	Accuracy	Precision	F1-Score	Recall/ Sensitivity	AUC
Logistic Regression	1.0	100	1.0	1.0	1.0	1.0
kNN	0.9908	99.27	0.9913	0.9929	0.9945	0.9927
Decision Tree	0.8219	85.38	0.8395	0.8612	0.8840	0.8529
Random Forest	0.9527	95.27	0.9554	0.9538	0.9538	0.9527

Table 1: Model Evaluation Metrics

From Table 1 above:
A basic model such as Logistic regression and kNN performed better with our dataset when compared to complex models like decision Tree.
The reason Decision Tree has the worst, among the four, evaluation metrics might be due to presence of outliers in our training set.

Discussion:

Initial Data Analysis gave some interesting and unexpected relationships. Like how United States movies have more 'non-Hit' in comparison to other countries. This can be associated with the fact that the US has released a greater number of movies. Or how the genre "Drama" has negative correlation with whether a movie becomes a hit or not. It is also observed how Production Country has a strong correlation with our target variable. Features such as Production countries, specific genres, and the budget put into a movie also matters. Even though there are exceptions to each of the above-mentioned factors, in general, it is seen that a hit movie would have to be in one of the 8 genres.

The metrics (mentioned in *Table1*) shows the Evaluation metrics that determine the fit of models to our dataset. The modelling showed that the simplest model, Logistic Regression, with the simplest of error calculation methods give the best model. It is expected that models such as Random Forest and Decision Tree would perform better, but it is seen that decision tree has succumbed to outliers and exceptions and given the worst result out of the 4 models. Random Forest is not as sensitive to outliers, but it still did not out-perform Logistic Regression, even if hyperparameters were tuned.

Hence, movie makers, producers, Film distributors can make use of these above-mentioned factors to decide if the movie will be a hit or not. However, there may be exceptions to each case mentioned. This may be attributed to other important factors that make a movie hit or not, such as Story, screenplay, actors etc. In future, the analysis could include text mining on the summary of a movie. This could make the model even more accurate.

STATEMENT OF CONTRIBUTION:

Madhusudan Malhar Deshpande: Data Exploration and Data Tidying
Reem Ghabayen: Data Visualization and Documentation.
Arif Sarfaraz Waghbakriwala: Data Cleaning and Data Visualization.
Aanand Dhandapani: Data Cleaning and Data Modelling

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.

APPENDIX