Worldwide Box-office revenue prediction for movies

Problem Statement

With the advancement in technology and cinematography, the film industry is growing rapidly and made huge business in last few years. According to reports, the American film industry generated $43.4 billion in revenue last year, increasing in each of the past five years at an annualized rate of just 2.2%. Thus, it is important to understand the factors which are behind the success of each movie; whether a huge star cast can make money at the box-office, a good director, the huge budget or there is any other factor behind the successful box-office story. In order to understand various factors behind the successful movie and predict the revenue of the movie, analysis has been carried out to build a prediction model.

Data Description - Test and Train Data Set

1. Purpose of Data set – Record the information of the upcoming or released movies.
2. Source of Data set – The data set is available on Kaggle, downloaded from https://www.kaggle.com/c/tmdb-box-office-prediction
3. Time window – Year 1930 – Year 2029
4. Cost of Data – Downloadable for free.
5. Collection technique – Update the database as the new products gets available on the store and even update the database as soon as any attribute of the product gets change.
6. Collection tools – TMDB API and downloadable from Kaggle.
7. Quality – The data is genuine since it has been provided by the internal source and thus there is a transparency in the data been available.
8. Completeness – The data isn't complete, there are some missing fields in the Train and Test data set.

Variables that are used in the Movie Dataset

Name	Definition	Variable Classification
id	Unique movie ID	int
belongs_to_collection	The collections to which the movie belongs.	object
budget	The budget for the movie.	float
genres	The movie's genre.	object
homepage	The movie's homepage.	object
imdb_id	Movie’s IMDB ID	object
original_language	The original language of the movie	object
original_title	The original title of the movie	object
overview	A brief plot	object
popularity	An index of movie’s popularity	float
poster_path	The poster of the movie	object
production_companies	The producer companies	object
production_countries	Countries involved in the production	object
release_date	Release date of the movie	object
runtime	The length of the movie	float
spoken_languages	The language spoken in the movie	object
status	Released or rumored	object
tagline	Tagline of the movie	object
title	Title of the movie	object
Keywords	Keyword used	object
cast	Details of the cast involved	object
crew	Details of the crew involved	object

Techniques and methodologies involved in Analysis and building model

1. Hypothesis Generation

This step involved understanding factors that may influence the revenue of the movie and hypothesizing different variables in order to understand their impact on the target variable (Revenue). This helped in performing EDA and focusing on important variables. Following are some of the hypothesis been made:

- **Budget:** The high budget movie may be successful at the box-office and earn high revenue.
- **Release date:** Releasing the movie during holidays or during festive season may earn high revenue.
- **Cast:** Popular actors or a multi-starrer movie may earn huge revenue.
- **Production company:** Big or popular production company may attract the audience and earn huge revenue.

2. Exploratory Data Analysis

Based on the hypothesis, analyzed the data statistically and visually to find the key insights in the data as well as any anomalous behavior present which can be rectified during Data Processing. Following are the key findings obtained upon analyzing the dataset:

- The train and test dataset have 3000 and 4398 number of entries respectively and have 24 and 23 attributes (columns) in the dataset.
- Upon analysis, it was observed that columns ‘belongs_to_collection’ and ‘homepage’ have more than 50% of missing value, thus they have been dropped from the dataset

- The other columns which have missing values are keywords (9%), cast (0.35%), crew(0.51%), genres (0.31%), overview(0.29%), poster_path (0.02%), production_companies (5.5%), production_countries (2.12%), release_date (0.013%), runtime (0.08%), spoken_languages (0.83%), status (0.02%), tagline (19.73%) and title (0.04%).
- Other columns, 'id', 'imdb_id', 'original_title', 'overview', 'popularity', 'status', 'tagline', 'title' have been dropped since it has been assumed that these parameters would not have significance impact on the sales.

3. Data Pre-Processing

Both the training as well as test dataset were combined before applying any modifications. Once, the datasets were combined, statistical summary of the combined dataset was obtained.

- It was observed that Budget and runtime have missing values. Thus, a **mean** for both the quantities was determined to impute the missing values.
- New columns were introduced – genre_number (number of genres of the movie), production_company_number (number of production companies involved in the movie), production_countries_number (number of countries involved in production), spoken_languages_number (number of languages spoken in the movie), cast_number (number of actors in the movie), crew_number (number of crew member involved). For all these columns, if there exists more than one values, then first three values have been recorded in separate columns (ex. Cast1, cast2, cast3 for cast column). In case of crew, only ‘Director’ has been used.

- Release date has been transformed into month and weekday. For any missing values in weekday, ‘Friday’ has been used to impute missing cell and in case of ‘month’, month with the highest movies has been used to impute missing value.

4. Feature Engineering

a) Determine the Budget – cast ratio

There may be a possibility that a movie with high budget and a multi start cast earns huge revenue or a movie with high budget and lower number of star-cast also gets successful and even low budget movie with a good star cast breaks record at the box-office. Thus, determining the budget-cast ratio may have an impact on the predictive model.

b) Determine the Budget – runtime ratio

The budget of the movie also varies depending on the length of the movie. Thus, it would be interesting to know if there is any significant relation between the budget-runtime ration and the revenue.

c) Determine mean-budget-by-year

Each year, with the advancement in the technology, the cost of making the movie is also changing. Thus, the average budget of the movie in each year can be a deciding factor – how much money should be invested to earn a good revenue.

5. Model Building

In order to build a predictive model, two Machine Learning algorithms Linear Regression and Random Forest were used. Firstly, the model was built using Linear Regression where some predictor variables were used to fit the model on the training dataset. After fitting the variables into the model, only those variables were retained which were significant i.e. having low p-value (<0.05). Later, these variables were fitted into the model using ‘backward-stepwise selection’ to obtain a reduced best-fit model.

This model iteratively deletes the insignificant predictors until no further reduction is possible. The only significant predictor which were used to fit the model were budget + runtime + production_company_number + production_countries_number + keywords_number + cast_number + crew_number + Budget_cast_ratio + Budget_runtime_ratio + weekday_0 + weekday_2 + weekday_3 + weekday_4 + month_0 + month_3 + month_5 + month_6 + month_8. The R-squared value of the model is 0.5508917 which is quite low and the RMSE value is 99162913 which is quite high.

Table 1: R^2, RMSE and Adjusted R-square value of a model

Model	**R^2**	**RMSE**	**Adjusted R-square**
Linear Regression – Reduced Model (Backward stepwise selection)	0.5508917	99162913	0.6084

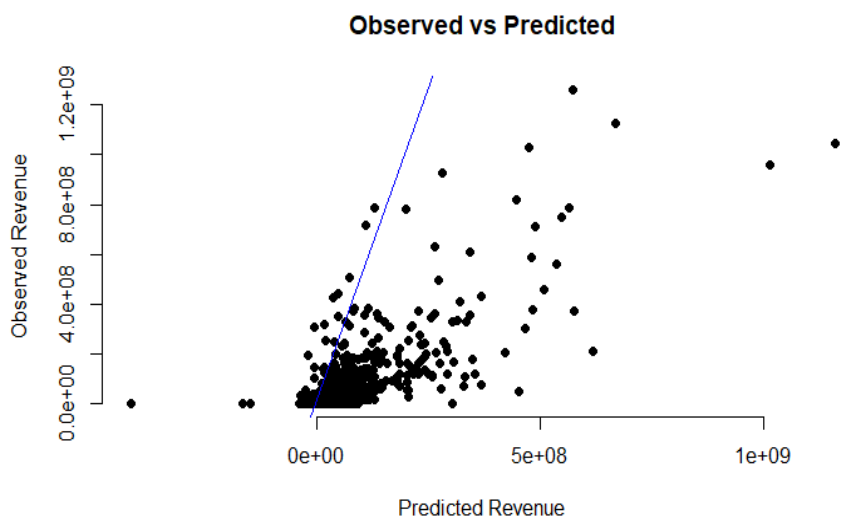

Figure 1: Relationship between observed and predicted values across the regression line for Linear Regression – Reduced Model.

The reason behind the low variability can be the existence of the Non-Linear relationship between the dependent as well as independent variable, thus fitting a model using Random Forest regression method would be a good choice. On fitting the Random Forest regression model, the R-squared value obtained was 0.5927156 and the RMSE was 94432748 which is comparatively better than the Linear Regression model. The model also explained approximately 62% variability. This model helped us understand the feature importance based on the IncNodePurity value (higher the IncNodePurity value, higher the importance of the feature) of the predictor variables. It was found that Budget and Budget-runtime-ratio are the most important variable and has_poster (poster of the movie is available or not) is the least important once.

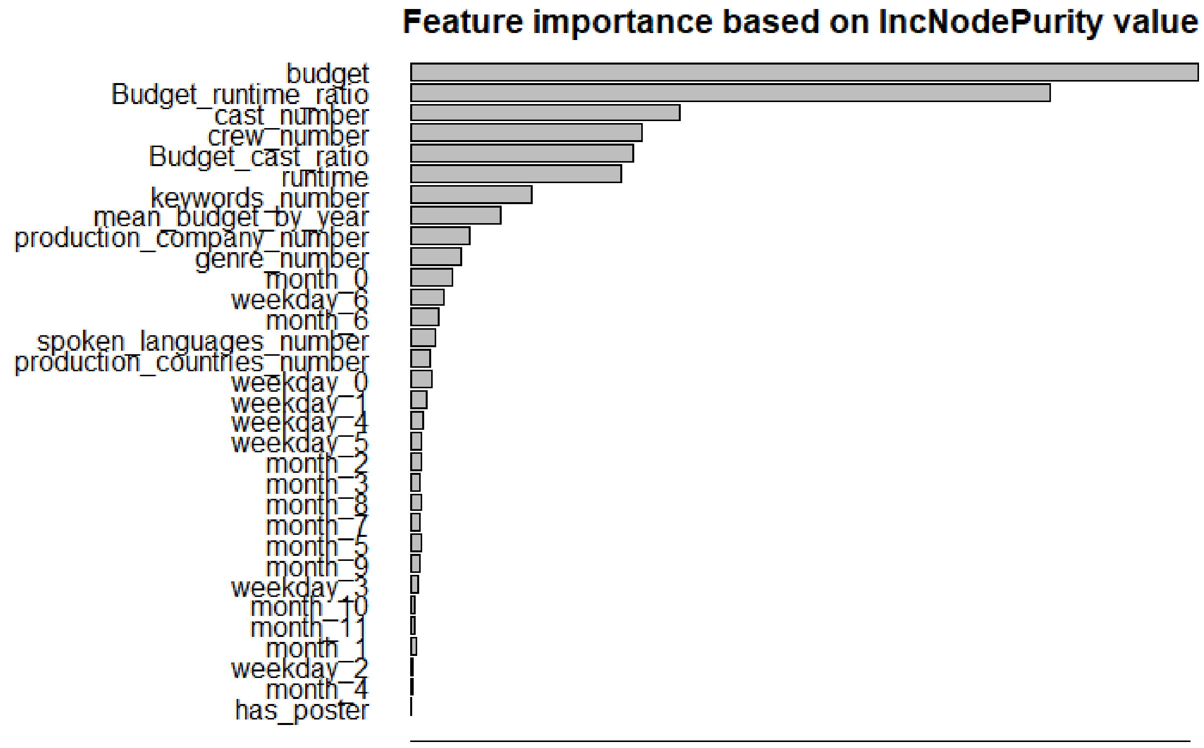

Fig 1: Bar graph representing the feature importance of the predictors upon implementing Random Forest model.

Conclusion

The Random Forest regression model would be beneficial in estimating the revenue of each movie worldwide in comparison to the Linear regression model. This would also help the film maker improve their box-office revenue by taking measures analyzing Feature importance of different variables from the dataset that could greatly impact. The film makers should also focus upon the least important feature and come up with some strategies to use those variables in improving the revenue. For instance, has_poster which has low feature importance can be strategically prepared and marketed to attract more audience. Further analysis of the actors' and director’s rating can be incorporated in the dataset to analyze their impact on the revenue. The accuracy of this model can also be enhanced by introducing more variables that may have direct impact on the revenue, this may further reduce the RMSE score and increase the R-squared value. Moreover, other Machine Learning models can also be used to predict the revenue with higher accuracy.

References

Kaggle. TMDB Box office Prediction. Retrieved from https://www.kaggle.com/c/tmdb-box-office-prediction

Donges, N. (2018, Feb 22). The Random Forest Algorithm. Retrieved from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Jha, V. (2017, June 18). Decision Tree Algorithm for a predictive model. Retrieved from https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-predictive-model/

Microsoft Azure. (2017, Nov 03). The Team Data Science Process Lifecycle. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Statistics Solution. (2013). What is Linear Regression. Retrieved from https://www.statisticssolutions.com/what-is-linear-regression/