Determinants of Successful and Unsuccessful Movies Based on Worldwide Box Office Revenue

Francesca De Simone, Lucia Quintero, Jaron Fay, Varun Gauba, Matthew A. Lanham

Purdue University Krannert School of Management

fdesimon@purdue.edu; lquinter@purdue.edu; vgauba@purdue.edu; jfay@purdue.edu; lanhamm@purdue.edu

Abstract

This study develops predictive models to identify the drivers of success and failure for movies based on worldwide box office revenue. Movies are big business and data analytics can help decision-makers design and create movies that are most likely to be successful. In 2018, the film industry made an estimated $41.7B worldwide, but if more resources went into creating the movies most people want to see, revenues are predicted to increase. Using movie features from 7,398 movies from The Movie Database (TMDB) we developed predictive models that were accurate at predicting revenue, but also interpretable to explain the drivers of successful outcomes, as well as negative outcomes. We discuss our recommendations to movie makers from a business perspective.

Introduction

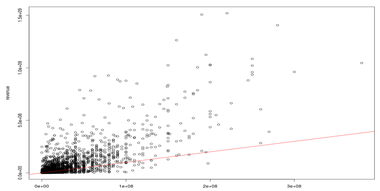

Figure 1. Revenue ~ Budget

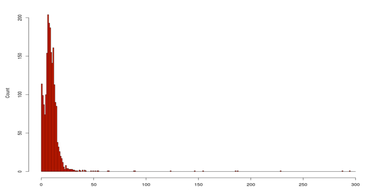

Figure 2. Skew of Popularity

These figures are just an example of the insight the original data provided us. We found there was a relationship between the different variables and the revenue.

Our original train dataset contained 3000 observations with 24 variables. But not all the variables were important in predicting the revenue.

➢ **What factors are better at predicting revenue?**

Our plan: create a model that highlights the main factors in what makes or breaks a movie will not only lead to an increase in box office revenue but will lead to the creation of high-quality movies.

Literature Review

Table 1. Main Models Used in Prediction of Success

Authors	Year	Method
Rijul Dhir, Anand Raj	2018	Random Forest
Márton Mestyán, János Kertész. Taha Yasseri	2013	Cross-Validation
Jeffrey S. Simonoff, Ilana R. Sparrow	2012	Linear Regression
Shyam Gopinath, Pradeep K. Chintagunta, Sriram Venkataraman	2013	Econometric Two-Stage Model
Ajay Siva Santosh Reddy, Pratik Kasat, Abhiyash Jain	2012	Text Mining

Through the review, we discovered that Random Forest would be the best method to build our predictive model. In order to ensure the most accurate results, we tried many different variations of this method. This will save millions of dollars in investment in a potentially unsuccessful movie.

Methodology

Data

We used the Movie Database (TMDB) to create our model. The raw data that we were given was mostly categorical variables; revenue, budget, popularity, and runtime being the only numerical variables.

Data Cleaning & Pre-Processing

For the variables that we did keep, we had to determine which ones had missing data that meant something or not. With most of our data being categorical we had to establish which ones would be useful to us, like genre or collection.

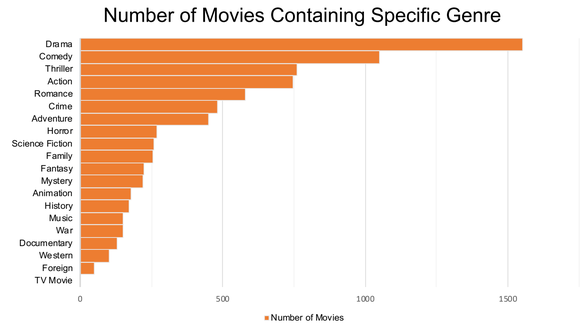

Figure 3. Example of Data Cleaning with Genre

Model Design

We used Random Forest to create our predictive model

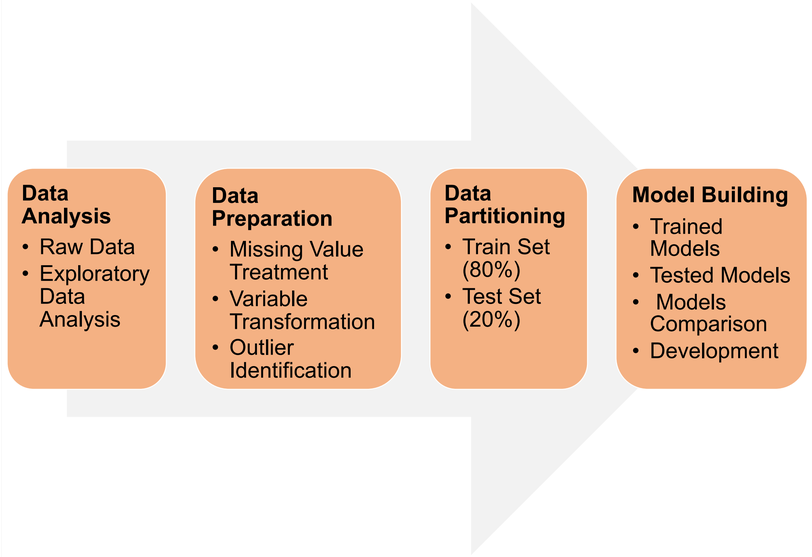

Figure 4. Study Design

Results

This figure represents the different parameters that we tried in our model to determine what the effects were on our RMSE. Each different parameter gave us different results in what was an important feature or not.

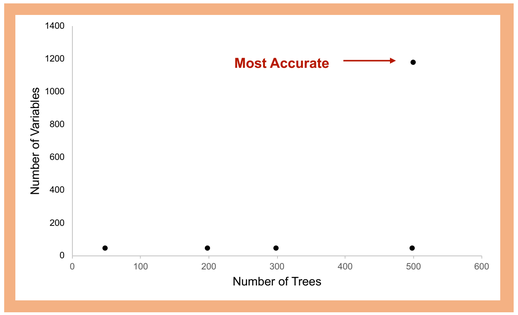

Figure 5. Comparison of Model Performance

The table below shows which variables hold the most importance in predicting the successfulness of a movie. Budget and release date on a Sunday and whether it is part of a collection or not are the three highest predictors of movie success in terms of revenue.

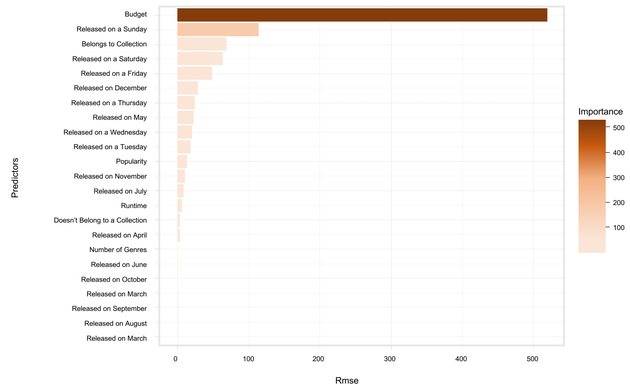

Figure 6. Importance of Predictors

Conclusions

In conclusion,

- Budget, releasing a movie on a Sunday, and whether it is part of a collection or not are the most important predictors of movie success
- This will help save thousand of dollars for companies and producers
- In the future, we will like to delete budget and part of a collection to see if a new movie will be successful regardless of movies before it, or the amount of money they put in
  - We will also like to determine success base on profit

Acknowledgements

We thank Professor Matthew Lanham, BAIM master students, and CRAN repository for constant guidance on this project.