TMDB Movies Dataset

Adham Howaidy

April 15, 2024

**Abstract:**

The TMDB Films dataset offers an all-inclusive collection of data on different motion pictures delivered throughout the years. This dataset contains a large number of properties including movie titles, release dates, budgets, revenue, genres, cast, and production companies. In this analysis, I investigated the attributes and trends of movies based on the provided data. Our analysis starts with data preprocessing steps including taking care of missing values and data cleaning to guarantee the integrity of the dataset. Then, at that point, I performed exploratory data analysis to gain understandings into the distribution and relationships among various variables. EDA includes visualizations, for example, histograms, scatter plots, heatmaps, and such to show correlations between variables in the data. Moreover, I performed statistical analyses to look at different aspects of the dataset, for example, budget and revenue trends over the long run, genre distributions, and the effect of specific factors on movie success. Therefore, I built predictive models to predict movie revenue with the chosen features (variables).

**Methods:**

The TMDB movies dataset was obtained from Kaggle which is a well-known website that has a lot of datasets regarding many different topics. Data preprocessing was used on this dataset. I inspected it well and as expected, there were missing values so I used appropriate strategies to handle the missing data such as removal of the missing values. Additionally, data cleaning procedures were used to address inconsistencies, errors, and outliers in the dataset to ensure data integrity for further data analysis. Exploratory data analysis techniques were used for example descriptive statistics, visualizations, and genre analysis. First for descriptive statistics, I calculated summary statistics, standard deviation, quartiles to better understand the distribution of numerical columns which were budget, revenue, and runtime. For visualization, I used different visualization strategies, for example, scatter plots, and heatmaps to explore the distribution and correlations among various features (variables). This allowed me to distinguish patterns, and correlations inside the dataset. As for genre analysis, I made a summary statistic to see which genres had the most budget, which one had the most revenue to see if there’s some trend. Next thing I used is some further statistical analysis for example I performed further statistical analyses to examine the relationship between movie budget and revenue. The next method I used was predictive modeling to attempt to predict the revenue of movies based on various features such as budget, popularity, and runtime. The goal for the model is to estimate the revenue. Revenue prediction can be very useful for various stakeholders in the movie industry, including production companies, investors, and distributors. I used various models such as random forest, gradient boosting, decision trees, but they didn’t work as well as the classification tree model that I ended up sticking with in the end. I also used the ARIMA model and a time failure analysis model as well to see if there’s any bit of improvement.

**Conclusions:**

My analysis revealed a positive correlation between movie budget and revenue which means that higher-budget movies tend to generate higher revenues. The top 3 genres that were the most popular genres were Adventure, Action, and Science Fiction. The classification tree was able to predict movie revenue based on features (variables) such as budget, popularity, and runtime with revenue being the target variable. The one limitation that I would say is the accuracy of the model. At first, it was 16%, then after some tweaks and further expanding the parameters, I managed to get it to 26% and couldn’t get it any higher than that. Maybe if I explore additional factors like critical reception and audience demographics then it might enhance the model and that might increase the accuracy. Also, the results from the ARIMA model and the time to failure analysis weren’t any better with the ARIMA model giving me a RMSE of over a billion which is of course too high of an error. Predicting movie revenue is intrinsically difficult because of the large number of elements impacting a film's prosperity, for example, marketing endeavors, release timing, competition, and audiences’ preferences. The ARIMA model, which depends entirely on historic revenue information, might have struggled to capture the intricacy and inconstancy of these elements. Perhaps it needed other features (variables) to be used such as marketing budgets, genre, cast, and critical reviews but not all those features were available in this dataset which is another limitation to why I failed on getting on a higher accuracy percentage in my predictive model.

**Statement of Business Value:**

My predictive model and analysis offer huge business value to partners in the film business by giving significant experiences to improve direction and enhance results. The potential business worth of my solution incorporates:

* Audience Engagement: Understanding people's preferences and patterns is very essential for expanding audience engagement and fulfillment. My investigation gives bits of knowledge into genre preferences, and demographics which will make stakeholders to fit their contributions to meet crowd expectations and improve the viewer's experience.
* Strategic Planning: My examination offers key experiences for long haul arranging and dynamic in the film business. By recognizing market patterns, competitive dynamics, and upcoming opportunities, stakeholders can develop informed strategies to maintain an upper hand and benefit from market trends.
* User Interaction: Users can interact with my model through a user-friendly interface, where they can enter parameters like for example budget, genre, and release date to get revenue predictions for their movie projects.

Investing in my project will enable companies to use information driven insights to drive business development, improve efficiency, and maintain a competitive edge in the marketplace.