TMDB Movie Revenue Prediction

Team Otter: Hyoin An, Hyeran Cho, Jonghoo Lee

Github: https://github.com/hyoin-an/erdosbootcamp

Overview

This project is part of the Data Science Bootcamp at the Erdos Institute. Our objective is to predict the revenue of movies by leveraging data sourced from TMDB (The Movie Database), inspired by a Kaggle competition. (https://www.kaggle.com/competitions/tmdb-box-office-prediction/overview).

Dataset

Description

We first considered histograms of six major quantitative variables to understand their distributions. The variables were budget, revenue, runtime, popularity, vote_count, and dates. Since the distributions of budget, revenue, popularity and vote_count were extremely right skewed, we took log of base 10 to transform their distribution more symmetric. The budget and revenue of a large number of movies were around a hundred million dollars. Most of the movies had a runtime between 1 hour and 2 hours. To see associations among the quantitative variables, we used scatter plots. All pairwise correlations were at least positive, and particularly budget and revenue showed the strongest positive correlation coefficient, over 0.7. We also investigated the number of movies by year and associations between revenue and year or month. In addition, we explore proportions of particular genres by month. For thriller and horror, the proportion was the highest in October. We guessed that it might result from Halloween. On the other hand, the proportion of Drama and Family movies was the highest in November and December. We interpreted that it was affected by Thanksgiving and Christmas.

Data Cleaning & Feature Engineering

We gathered the dataset from TMDB by utilizing TMDB API. There are about 6500 movies available with valid revenue data. Features include genres, budget, production countries/companies, release date, cast, popularity, etc.

For genres, the raw data contains a column of lists of genres. We added a column for each genre whose row stores a boolean indicating the genre of the corresponding movie. For the cast, there are a total of 130,005 actors/actresses. We added a column for each actor/actress so that each column indicates whether the actor/actress appears in the movie using boolean. A similar transformation is applied to the directors data.
For the release date, the raw data is in datetime format. We have normalized them by mapping them linearly onto [-1,1] so the oldest movie has the value -1 and the most recent movie has the value 1. We have also added the month column indicating the month of the release date of the movie. For the production countries, we have added a column indicating whether the movie is produced in the US or not. As a result, the preprocessed dataset contains ten key features:

budget, popularity, runtime, video, vote average, vote count, number of genres, collection, homepage, release date, 59 dummy variables: genre (15), language (32), and month (12). Finally, since there are so many casting members and directors, we have applied PCA to cast and directors using 10 principal components.

Method

We partitioned the dataset into training (n = 4800, movies from 1902-04-17 to 2017-09-22) and test (n = 747, movies 2017-09-22 to 2023-10-25) set. We conducted model training and parameter-tuning exclusively on the training set. Utilizing the XGBoost model framework, we refined hyperparameters through 5-fold cross-validation. As a benchmark, we chose a naive model estimating log revenue using log(budget + 1). The key performance indicator for the model is the root mean squared error (RMSE).

Results

- RMSE (Our model): 1.3118
- RMSE (Benchmark): 1.6552

The variable importance plot highlights key factors crucial to model performance, with the top variables identified as follows: budget, vote count in TMDB, movie language, genre (Horror), and popularity measured by TMDB.

Discussion

In future work, movie title and overview information could prove beneficial for enhancing predictive accuracy. Additionally, exploring alternative methods of feature engineering for cast and director information may offer valuable insights.