Predicting Box Office Revenue

Ronnen Nagal
Ben-Gurion University

March 25, 2019

Abstract

The film industry is in a constant growth trend. The global box office was worth 41.7 billion in 2018. Hollywood has the world's most massive box office revenue with 2.6 billion tickets sold and around 2000 films produced annually.

One of the main interests of the film studios and related stakeholders is a prediction of revenue that a new movie can generate based on a few given input attributes.

Introduction

Starting in 1929, during the Great Depression and the Golden Age of Hollywood, an insight began to evolve related to the consumption of movie tickets. It appeared that even in that bad economic period, the film industry kept growing. The phenomenon repeated in the 2008 recession.

The primary goal is to build a machine-learning-based model that will predict the revenue of a new movie given such features as cast, crew, keywords, budget, release dates, languages, production companies, and countries.

EDA was the first step followed by introducing an initial linear model and comparing it to other models at the end of the process. 7398 movies data collected from The Movie Database (TMDB) as part of a kaggle.com Box Office Prediction Competition. A train/test division is also given to build and evaluate the developed model.

1 Challenges

Consumer behaviours have changed over the years: the MeToo movement, as well as other social developments, have surfaced in our society, and that reflected in movie scripts. However, some of the preferences that were relevant 50 years ago are still relevant today; hence, an analysis based on the last few decades of movies production is always appropriate and will be able to serve any stakeholders that have an interest in predicting a new movie revenue.

2 The System

The machine learning model is based on a supervised learning system, as it's the most common state of many real-world problems.

The categorical columns welcomed a challenge. Starting with converting the 'collection' column yield that only 20% of the films do (600 out of 3000), and from a simple boxplot, we can observe that movies that belong to a collection tend to have higher revenue.

The categorical column 'overview' welcomes a semantic analysis; as we read an overview of a movie, we decide whether to see that movie or not; hence, semantic analysis has a place here, although I found no correlation between revenue and the 'overview' and 'tagline' of a movie using NLP with the VADER sentiment package.

Other categorical columns, such as cast, crew, genre, and more were the main focus in analyzing the data and creating the features engineering process.

3 The Models

Different models implemented and the score, as assessed by kaggle.com competition, was observed and evaluated. The models started with a linear regression model, moving on to a random forest regression model and finishing with an LGBMRegressor model.

4 The Outcome

The LGBMRegressor model is the winner with the best result; the model was sent to the competition on kaggle.com twice (last time at March 20, 2019), and successfully overcame 70% of the participating teams. The contest will end in two months, so there is time to make more incremental steps to improve the score.

Regards.
Ronnen Nagal