Udacity Machine Learning Nanodegree

Capstone Proposal

Movie Box-Office Predictive Analysis

Joel Vilanilam Zachariah

May 18th, 2019

Domain Background

In a world where movies made an estimated $41.7 billion in 2018, the film industry is more popular than ever. The multi-billion dollar industry involves a lot of risk in terms of how well the public will perceive the result that the team creates. Ideally, the revenue gained should be greater than the production budget to keep the work sustainable.

According to wikipedia[1], profit is a key force in the industry, due to the costly and risky nature of filmmaking. For example, *Avengers Endgame* by Marvel Studios garnered over $1.2 billion dollars[2] in the first weekend after release[3], while the budget for the movie was $356 million dollars. The production had to make risk assessments and decide the best conditions to arrive at this successful end result.

By drawing relevant conclusions the production team can optimize their resources to produce the movie with better success probability. I am personally interested to work on this domain because it will be interesting to not only read about the past movies but also to understand the factors that affect the audiences satisfaction in the content.

Problem Statement

A successful movie requires proper planning to minimize production cost and maximize revenue, that is, to maximize profit. Though it is true that experience plays a vital role in deciding the course of action to take for the creation of the movie, the audience acceptance rate can be enhanced if research is done. It can be done from past movies and studying the inter-relations in the system. This risk assessment[4] can improve the production teams success rate.

We need to predict the overall worldwide box office revenue by studying the correlations in the system and estimating which parameters affect the result the most (in terms of revenue). Following that, we need to train the model with the training data set and measure its accuracy with the testing data set.

The purpose is this project is to identify the main characteristics that affect the revenue of the movie based on a handful of machine learning algorithms that we shall test on the remaining features. The experimental results will be assessed by statistical metrics, As we get higher scores against these metrics, we will be able to better judge the success of a movie.

Datasets and Inputs

The dataset is from *The Movie Database* (TMDB)[5], which is a database with metadata of over 3,000 movies from the past. I have used the dataset provided for a competition on Kaggle by TMDB[6].

The "tmdb-data.csv" data file has 3000 rows of data across 23 features.

Input Variables:

1. Id [integer] (Serial number in the database)
2. Belongs_to_collection [object file] (Collection it belongs to)
3. Budget [integer] (Cost of production)
4. Genres [object file] (Theme of the movie)
5. Homepage [object file] (Link to the internet page)
6. Imdb_id [object file] (reference id at IMDB)
7. Original_language [object file] (primary movie language)
8. Original_title [object file] (Title of the movie)
9. Overview [object file] (Short description of the movie)
10. Popularity [float] (relative measure of popularity)
11. Poster_path [object file] (Directory pathway to movie poster)
12. Production_companies [object file] (Name of production companies)
13. Production_countries [object file] (Country of production)
14. Release_date [object file] (Date when the movie was released)
15. Runtime [integer] (duration of the movie in minutes)
16. Spoken_languages [object file] (Languages used by actors)
17. Status [object file] (Released/ Not yet released/ etc)
18. Tagline [object file] (The one-line caption of the movie)
19. Title [object file] (Movie title)
20. Keywords [object file] (Related keywords of the movie)
21. Cast [object file] (List of actors)
22. Crew [object file] (List of crew members)

Let us examine a few attribute in more detail:

- Belongs_to_collection: [{'id', 'name', 'poster_path', 'backdrop_path'}]
- Genre: [{'id', 'name'}]
- Production_companies: [{'name', 'id'}]
- Production_countries: [{code, 'name'}]
- Keywords: [{'id', 'name'}]
- Cast: ['cast_id', 'character', credit_id', 'gender', 'id', 'name', 'order', 'profile_path']
- Crew: [{'credit_id', 'department', 'gender', 'id', 'job', 'name', 'profile_path'}]

Output Variable:

1. Revenue [integer] (Revenue of the movie)

Missing Attribute Values:
There are several missing values in the dataset. These can be treated as possible using imputation techniques.

Solutions Statement

We will prepare the data by splitting feature and target/label columns and also for quality of given data after performing data cleaning. To check for the quality of the model, we will split the data into training and validation sets to check the accuracy of the model. We will split the given data in two - 70% of which will be used for training our model while remaining 30% to utilize as validation set.

Since there are a lot of non-numeric columns, we need to convert them suitably. We can perform label encoding to do the necessary conversion - that is, we create dedicated columns for each kind of input and mark a '1' against the corresponding row. Which contain that value.

I plan to create a neural network as a model and use mean absolute error as the loss function. We are using 3-fold cross validation to estimate accuracy. This will split our dataset into 3 parts - train on 2 and test on 1. We repeat for all combinations of train-test splits.

We are using the metric of mean absolute error to evaluate models. This is the mean of the deviation between predicted value and target value. We will be using the scoring variable when we build and evaluate each model next.

Benchmark Model

As a benchmark, I plan to compare the result of my model against the model that uses only the mean of the training dataset revenue. This is a naive solution but works well to set the baseline for the system.

Evaluation Metrics

For the training phase, we will utilize all the 23 parameters while for testing phase, the first 22 parameters are used to predict the revenue. Mean absolute error will be the measure of quality for the model. Over several epoch, we shall try to minimize the error and then predict the revenue for a given movie characteristics.

Project Design

The workflow of solving this problem will be in the following order:

1. Loading the libraries and dataset
2. Understanding the dataset
3. Dealing with missing values
4. Determine Correlations
5. Feature Engineering
6. Creating the model
7. Model Evaluation

Visualizations will be provided as needed.

References

[1] https://en.wikipedia.org/wiki/Film
[2] https://www.vox.com/2019/4/29/18521581/avengers-endgame-box-office-1-2-billion
[3] https://www.the-numbers.com/movie/Avengers-Endgame-(2019)
[4] https://www.filmsourcing.com/film-production-risk-assessment/
[5] https://www.themoviedb.org/en
[6] https://www.kaggle.com/c/tmdb-box-office-prediction