COMP4211 Project – TMDB Box Office Prediction

Shayan Imran

20316792

2019

Contents

[Introduction 2](#_Toc8238700)

[Dataset and Preprocessing 3](#_Toc8238701)

[Machine Learning Task 4](#_Toc8238702)

[Machine Learning Methods 5](#_Toc8238703)

[Experiments and Results 6](#_Toc8238704)

[Computing Environment 7](#_Toc8238705)

# Introduction

This project presents experimentation using multiple types of machine learning techniques on the [TMDB Box Office Prediction Kaggle Competition](https://www.kaggle.com/c/tmdb-box-office-prediction/overview) (<https://www.kaggle.com/c/tmdb-box-office-prediction/overview>). The aim of the competition is to predict the revenue for movies given metadata for over 7000 past movies.

# Dataset and Preprocessing

The dataset provided in this competition consists of metadata for 7398 movies from the TMDB movie database. Only 3000 of the movies in this list contain a column with the revenue earned by the movie, predicting which is the target of this competition. These 3000 rows are therefore used to train the machine learning models that are used.

The metadata consists of the following 22 columns:

* id
* belongs\_to\_collection  
  The name of the series this movie belongs to, if it belongs to one eg.   
  [{'id': 256377, 'name': 'The Muppet Collection', …}]
* budget
* genres  
  A list of genres this movie belongs to eg.  
  [{'id': 28, 'name': 'Action'}, {'id': 35, 'name': 'Comedy'}, …]
* homepage  
  A link to the website of the movie, if it has one
* imdb\_id  
  The id of the movie on the Internet Movie Database
* original\_language  
  The language of the movie
* original\_title  
  The original title of the movie
* overview  
  A few sentences introducing the movie
* popularity  
  A real number which seems to be higher the more popular the movie is, but it is not explained how this is obtained
* poster\_path  
  A web location for the poster in the TMDB website
* production\_companies  
  A list of the companies involved in the production of the movie eg.  
  [{'name': 'Paramount Pictures', 'id': 4}, {'name': 'United Artists', 'id': 60}, …]
* production\_countries
* release\_date
* runtime
* spoken\_languages
* status
* tagline
* title
* Keywords
* cast
* crew

In addition to this, the training data consists of an extra column for the output of the task:

* revenue

# Machine Learning Task

# Machine Learning Methods

# Experiments and Results

# Computing Environment