A.I Technical Report

Contents

[Introduction 3](#_Toc189728807)

[The Problem 4](#_Toc189728808)

[The Data 5](#_Toc189728809)

[Methodology 6](#_Toc189728810)

[Code Overview 7](#_Toc189728811)

[Results/Findings 8](#_Toc189728812)

[Conclusion 9](#_Toc189728813)

[References 10](#_Toc189728814)

# Introduction

To do.

# The Problem

When going through the process of conceptualising, writing, filming and producing a movie, it can be hard to say if the movie will do well. Merely throwing money at a movie’s production will not make it successful. For example, both “The 13th Warrior" and “Mortal Engines” were given budgets of around $150 million, but the former only grossed $62 million and the latter grossed a slightly better $84 million (Saha, 2023).

To assist this problem, this document will go over the creation of a simple machine learning model that will attempt to find a new movie’s potential success from the movie’s information (such as genre and age ratings). While a movie’s rating does not show the monetary success of a movie, it is hoped that this tool will help to show if a new movie would be popular or not to help possibly remove some of the ambiguity that comes with movie creation.

# The Data

The data for this project is sourced from the metadata of the online movie database website known as IMDb (Verma, n.d.). IMDb’s database contains the information of thousands of movies and allows users to see a movie’s rating, cast and the people who produced said movie. It also allows users to find other movies similar to ones they already like (IMDb, n.d.). But most importantly, IMDb gives each movie a rating based on user ratings rather than ratings from professional reviews. This helps to give a better idea on what movies are popular rather than just being considered “good”.

The data itself contains the following headers:

* name: The name of each movie.
* year: The year each movie was released.
* movie\_rated: The age rating of each movie.
* run\_length: Runtime of each movie.
* genres: The genres of each movie separated by a semicolon.
* release\_date: The date each movie was released.
* rating: The rating of each movie from IMDb.
* num\_raters: Number of people who rated each movie.
* num\_reviews: Number of people who reviewed each movie.

(Verma, n.d.)

# Methodology

For this project, the data will be loaded into the Python programming language and, through the use of the numpy library, will be loaded into tables. Once the data is loaded, it will be analysed, cleaned and normalised before then being used to create four models using K-Nearest Neighbours and linear regression, with those four models being:

1. Age Rating (KNN)
2. Runtime Length (KNN)
3. Genre (KNN)
4. Release Month (Linear Regression)

Each model will give its predicted user ratings. Then, the average of each of these user ratings will be finally given as the predicted user rating for the new movie. After this, graphs will be created using another Python library known as matplotlib in order to show how the models came to the predicted user rating. After the models have been created, a website will be created to allow the inputting of movie information and the prediction of the potential user rating.

# Code Overview

# Results/Findings

# Conclusion

# References

* IMDb. (n.d.). About IMDb. IMDb. <https://www.imdb.com/pressroom/about/?ref_=fea_eds_center-29_pr_related_1>
* Saha, P. (2023, November 10th). Hollywood: These Are The Most Expensive Flops In Movie History. CEOWORLD Magazine. <https://ceoworld.biz/2023/11/10/hollywood-these-are-the-most-expensive-flops-in-movie-history/>
* Verma, K. (n.d.). IMDb Movies Dataset. Kaggle. <https://www.kaggle.com/datasets/krishnanshverma/imdb-movies-dataset>