Naima Asif​, Faizan Bhutto​, Shravanthi Polu​, Emma Stefanovich​, Sivaramakrishna Yarra​

DSC 423 Project Report

**Understanding the Revenue and Consumer Ratings of Movies**

**Introduction**

A movie production company’s main goal is to produce movies that are successful—that is, movies that turn a profit and rate well with viewers. Our project focuses on how we can use data on movies to find what features of a movie contribute most to its revenue and viewer score. After preparing our data, we explored each variable of interest and found two models: one to predict revenue, the other to predict viewer score. After splitting our data into training and testing sets, we made predictions on the testing set and assessed each model’s accuracy. We found that each of our models provided accurate predictions, but that the model on a movie’s score was far more accurate than the one on a movie’s revenue.

**Problem**

If we were a movie production company, we would want to find out what features of a movie influence its success. A movie’s revenue and reviews are largely indicative of a its success, both monetarily and terms of public opinion. If we could find out what features of a movie contribute to higher revenues and reviews, then we could focus on making movies that are more likely to succeed. With models on revenue and review score, we could also make predictions on how much revenue a movie might make and how viewers might score it.

**Dataset Description**

The dataset we used is called [*TMDB 5000 Movie Dataset*](https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata). It contains information on movies from [The Movie Database](https://www.themoviedb.org/?language=en-US). The dataset has 4,803 observations on movies released from 1916 to 2016. It contains the following variables:

* *budget*: The budget of the movie's production, in USD.
* *genres*: A list of genres applicable to the movie.
* *homepage*: The movie's website, if it has one.
* *id*: The movie's ID number in The Movie Database.
* *keywords*: The list of keywords associated with the movie in The Movie Database.
* *original\_language*: The original language the movie was produced in.
* *original\_title*: The original name the movie was released with.
* *overview*: The description of the move in The Movie Database.
* *popularity*: The current popularity score of the movie in The Movie Database, determined by the average number of users searching for the movie per day.
* *production\_companies*: A list of production companies that created the movie.
* *production\_countries*: A list of countries the movie was produced in.
* *release\_date*: The date on which the movie was first released.
* *revenue*: The total earnings of the movie, in USD.
* *runtime*: The length of the movie in minutes.
* *spoken\_languages*: The language(s) that the movie is currently available in.
* *status*: If the movie is released, rumored, or in post-production.
* *tagline*: The tagline of the movie in The Movie Database.
* *title*: The current title of the movie.
* *vote\_average*: The average user review score of the movie in The Movie Database.
* *vote\_count*: The number of users who reviewed and scored the movie.

Not all of these variables are useful for analysis. We will be excluding the following variables from any analyses:

* *homepage* (website url, unnecessary text)
* *id* (unnecessary id column)
* *keywords* (unnecessary text)
* *original\_language* (unnecessary information, will use spoken\_languages instead)
* *original\_title* (unnecessary text)
* *overview* (unnecessary text)
* *status* (unnecessary information, all movies are released)
* *tagline* (unnecessary text)
* *title* (unnecessary text)

The rest of the variables will be useful for analysis. The main variables we want to predict using various models are *revenue* and *vote\_average*.

**Procedure**

Before we could do any data exploration or analysis, we first had to clean the data. A full documentation of our data cleaning is available in Appendix A. In short, we ended up creating some new variables and excluding others based on their usability. A full list of the final variables we used is present in the section below.

1. **Data Exploration**

Before creating any models, we first explored our cleaned data to look for patterns or issues. Below are the variables in our cleaned data that we will explore: