Final Project Description Fall 2021

Background and Introduction

The final project for the course will require you to complete some tasks based on a hypothetical request from an independent film company. They are trying to decide how to allocate their resources in order to get more views on Netflix. There is a website, FlixGem https://flixgem.com (https://flixgem.com), which collects data from different sources and produces what they call a “Hidden Gem” score, which users then use to choose new movies and television series to watch that they might not come across. The company has several questions to ask of this data, all of which will help them going forward. The data for this project comes from a Kaggle project, the details of which can be found at https://www.kaggle.com/syedmubarak/netflix-dataset-latest-2021 (https://www.kaggle.com/syedmubarak/netflix-dataset-latest-2021).

I have attached the dataset to the project description on Crowdmark so that you do not have to register for an account on Kaggle. I have also reduced the number of variables, to help to limit the scope of the project. The data set you can download from Crowdmark contains the following data from the original dataset:

Variable	**Description**
Title	Movie or series title
Languages	Languages in the film
Series or Movie	Series or standalone movie
Hidden Gem Score	Hidden Gem Score from FlixGem
Runtime	Runtime Category
Director	Director
IMDb Score	IMDb Score
Rotten Tomatoes Score	Rotten Tomatoes Score
Metacritic Score	Metacritic Score
Release Date	Release Date
Summary	Movie Summary

Note that IMDb, Rotten Tomatoes, and Metacritic are all differnt websites which specialize in rating movies based on general public and movie critic reviews.

Objectives and evaluation

The project requires you to complete three tasks, detailed below. You should prepare a single report containing your answers to all tasks. Include the code for each task in your report, for reproducibility purposes. You may include the code as code chunks where the analyses are taking place or, if you prefer, you may include it at the end (although the code should be clearly commented so that it is clear which task each block of code corresponds to).

The completion of each task is worth 25 points. The quality of presentation will also be worth 25 points, i.e. clarity of explanation, plots, tables, and code.

The length of the projects will vary, depending on the number and formatting of figures and tables and the conciseness of the writing. Rather than focusing on the number of pages, I encourage students to focus on completing each task (and subtask) below to the best of their ability in the clearest and most efficient manner.

Tasks to complete

Task 1: Data wrangling and exploratory data analyses

The first task is to do some data wrangling (i.e. cleaning and manipulation) and conduct some exploratory data analyses. The film company **DOES NOT** want results for Series, only for Movies, since they only produce movies. Second, they know that there is missingness in some of the variables, but they are content to allow you to drop any records containing any missing values for the purposes of this analysis (so you should). **Include any plots and summary statistics that you think will aid in supporting your assessments.**

Based on the subsetted and cleaned data, please answer the following questions:

a. Does the Hidden Gem Score seems to be associated to the Runtime Category or the languages used in the film? Explain briefly the reasons behind your assessment. **Hint:** You may need to do some re-coding of one or both of these variables. Any reasonable re-coding is fine, just be sure to be clear what you've done.

b. Do any of the three review site scores (IMDb, Rotten Tomatoes, Metacritic) seem to be strongly or weakly correlated with the Hidden Gem Scores? Explain briefly the reasons behind your assessment and the nature of those associations.

c. The company has a theory that people are becoming more acceptable of longer movies because they can watch them at home on Netflix and other content-collecting sites. Do you notice any trend over time in the Hidden Gem Scores by category of RunTime Length? Explain briefly the reasons behind your assessment.

Task 2: Factors of the Hidden Gem Score

Recall that the goal of the company is to make decisions about what the most important factors are that contribute to the Hidden Gem Score. The company has suggested that a Regression Tree could be used to maybe identify those factors. Regression trees would work particularly well for this problem due to the categorical nature of the data. A description of regression trees can be found here: https://uc-r.github.io/regression_trees (https://uc-r.github.io/regression_trees) with example code. Apply the `rpart` function to the data using the Hidden Gem Score as the outcome and Languages, Runtime, IMDb Score, Rotten Tomatoes Score and Metacritic Score as predictors. Summarize what you think are the most important features for predicting the hiddden Gem Score based on the fitted tree and summarize how well your predictions perform. **NOTE: You DO NOT have to implement any Bagging or Split Optimization from the article beyond what the `rpart` function already provides.** (but of course you can if you're excited to do so).

Task 3: An H-index for directors

The last task they would like you to complete is to find a way to identify **directors** who produce films that have high Hidden Gem Scores. The problem is how to use the Hidden Gem Scores for directors, given that all of the directors have directed different numbers of films in the dataset. If you use the maximum Hidden Gem

Score for each director as the measure of how good they are, then directors with more movies are likely to look better because they have more chances to have a high score. If you use the average Hidden Gem Score, then directors with fewer movies can look spuriously good because they produce a single hgih score movie.

This is similar to the problem that we see with trying to rank researchers based on their citations. Researchers who publish lots of papers will have lots of citations to their work, even if none of their work is not cited often. Researchers who publish a small number of highly cited papers have a much smaller of body of work to be judged upon. What has been proposed is a measure to balance quantity and quality, the **H-index**. The H-index in reseach for a researcher is equal to the number, H, of publications for that researcher which have all been cited AT LEAST H times. For example, a researcher who has published three papers which have been cited 1 time, 4 times and 100 times respectively has an H-index of 2, because they have 2 papers that have been cited at least 2 times. A researcher who has published 5 papers that have been cited 3, 6, 7, 8, and 9 times has an H-index of 4 because they have 4 papers that have been cited at least 4 times.

For this task, find the top 10 directors in the dataset according to an Hidden Gem H-index (an HG-H index?) defined as the the number of films, H, in the dataset that they have directed which have Hiddden Gem Scores that are greater than or equal to H and produce them in a table with their associated $HG-H$ index.