Brooke Ann Coco
Project Luther Proposal
Predicting Tabletop Game Success
April 10, 2019

Overview:

The goal of this project is to use a simple linear regression model to predict the success of tabletop games. Success will be measured by rating (as reported by boardgamegeek.com). Given the growing success of tabletop games within recent years, this project would be of particular interest to tabletop game designers and publishers alike, who can use this information to produce games that are optimized towards consumer desires.

Data:

Much of my data will be scraped from boardgamegeek.com (BGG), a comprehensive games database as well as a popular online forum for tabletop game enthusiasts. The database contains a plethora of information (including videos, images, playtime, number of players, price, community reviews, community ratings, etc.) of over 81,000 tabletop games.

While boardgamegeek.com will be the primary source of information, I will need to find other sources to obtain information such as total yearly sales per game, the country in which each game was released, and the number of games previously released by the creator of each game.

Features:

Throughout this project, I plan to focus on the following features:

Feature Name	Variable Type	Description
rating[*]	Continuous	Average rating of game as scored from 0.0 (awful) – 10.0 (outstanding) Metrics based on BGG community ratings of difficulty: https://boardgamegeek.com/wiki/page/Ratings&redirectedfrom=rating#
release_year	Discrete	Years since the game has been released
min_players	Discrete	Minimum number of people required to play
max_players	Discrete	Maximum number of people able to play
player_diff	Discrete	Difference between maximum number of players and minimum number of players
min_playtime	Discrete	Minimum length of time to complete game (by increments of 15 minutes)
max_playtime	Discrete	Maximum length of time to complete game (by increments of 15 minutes)
playtime_diff	Discrete	Difference between maximum playtime and minimum playtime

[*] Target variable

complexity	Continuous	Difficulty of game as scored from 0.00 (easiest) – 5.00 (most difficult). Metrics based on BBG community ratings of difficulty: https://boardgamegeek.com/wiki/page/Weight
genre	Nominal	Strategy – Children's – Family – War – Abstract – Party – Thematic – Customizable
type	Nominal	Type of game: Card – Dice – Board
price	Continuous	List price of game
expansions	Discrete	Number of expansions offered
*reviews***	Discrete	Number of ratings received on BBG Used as a proxy for sales
*owned***	Discrete	Number of BBG reported owning game
awards	Discrete	Number of awards received by game
country_of_origin	Nominal	Country in which game was released
release_order	Discrete	Number of games previously released by creator (including observation game)

** I would utilize either one variable or the other, not both.

Project Design Considerations:

The potential problem with using average ratings as my target, is that it is a subjective measure that varies from source to source. Thus, if utilizing average ratings, I would need to be mindful that the numbers are not objective, then indicating such in any subsequent reports or presentations of the findings.

Originally, I had planned to utilize yearly game sales as a feature within my dataset, however, upon further inspection, there appears to be no relatively comprehensive source of tabletop game sales. I have yet to locate a reliable source from which to obtain game sales. Thus, I would like to use either the number of ratings or the number of users who own the game (based on the information submitted by BGG users) as a proxy for sales. Though the sample is skewed, it is likely correlated with sales.

Another concern I needed to consider is how to account for several different editions of the same game. Consider Monopoly, which has over 1,100 specialty versions of the game; such as Monopoly: Star Wars, Monopoly: Stock Exchange, Monopoly: London etc. Though Monopoly is an extreme example, it is not uncommon for board games to have several versions of the same game. Though these variants tend to contain the same (if not, similar) game mechanics as the original game, I am inclined to treat these variations as separate games. Often, these variations are thematic, which can lead to fans collecting several versions of the same game. Thus, treating this data as a subset of the original game has the potential to skew the data towards games that have more versions. The decision to treat game variants as distinct from the original version was further re-enforced by the fact that BBG gives these variants their own page, rather than nesting them under the original version.

Finally, I needed to consider how I would represent the number of players and playtime in the dataset, as both represent a range of values. One option was to set each feature as categorical variables. Another option was to separate each feature into two separate columns, one representing the minimum values and the other representing the maximum values. A third option was to calculate the average of each variable. Instinctively, I am leaning towards the second option. Though after further inspection of the data, I may opt for the latter of the three.