Board Game Geek Blog Analysis

By William Dew

Introduction

Board Game Geek (BGG) is an online forum for board gaming hobbyists. It contains a database that holds reviews, images, videos, blogs and information about thousands of different board games. These board games range from European-style board games to card games and everything in between. BBG also lists game specifications provided by the manufacturer, and allows users to rate the games which are then used to generate an overall user score.

Board game companies can use BBG to their advantage. Hundreds of new games come out each year, and it is hard to not have a game get lost in all the noise and fanfare. A board game company can use BGG to find out what types of board games, including their themes and mechanisms, are trending popular at the moment. A predictive model could be built that can predict the score of a board game based on the features of the board game.

Data

The data to be analyzed comes from a public database called Kaggle collected from BGG in March of 2017. (https://www.kaggle.com/mrpantherson/board-game-data).

The board game data has 20 columns and 5000 rows. Each column is a characteristic of a board game and each row is a unique board game. The 20 characteristics can be divided into three types:

1. Qualitative:

    - *mechanic*: list of board game mechanics
    - *category*: list of themes or categories of the board game
    - *designer*: list of designers

2. Quantitative:

    - *min_players*: minimum number of players
    - *max_player*: maximum number of players
    - *avg_time*: average time to play the game
    - *min_time*: minimum time needed to play game
    - *max_time*: maximum time needed to play game
    - *year*: year board game was released
    - *avg_rating*: mean of all board game user ratings
    - *geek_rating*: BGG rating based on avg_rating but altered with BGG algorithm
    - *num_votes*: number of users that have rated the game
    - *age*: recommended minimum age of players
    - *owned*: number of users who stated they own game
    - *weight*: average rating of how hard a game is to understand given by users

3. Unique-each value in column is unique, no duplicates

    - *rank*: each game is ranked by popularity from 1 to 5000
    - *bgg_url*: url of game on BBG
    - *game_id*: each game is given a unique ID
    - *names*: name of board game
    - *image_url*: url to picture of board game cover

One problem with the data is some categories are provided by the user, such as weight and rating. Other categories' information should come from the board game box, but may be entered wrong by the users or board game companies. These factors must be taken into account when exploring the BGG's data.

Data Wrangling

It is important to "clean" the data to help build a better predictive model. This will be done by removing unique values, null values, and outliers.

The first thing to do is to remove the columns with unique features because they won't help the model because unique features do not have a pattern that the model can analyze. These columns will be removed during the cleaning process: *bgg_url*, *image_url*, and *rank*.

For columns *min_players, max_players, avg_time, min_time, max_time* having a 0 as a value does not make sense. 127 games had 0 in one of the previously mentioned columns. A game has to have at least one player and time has to take longer than 0 minutes to be played, so these values were probably entered in error and so will be removed.

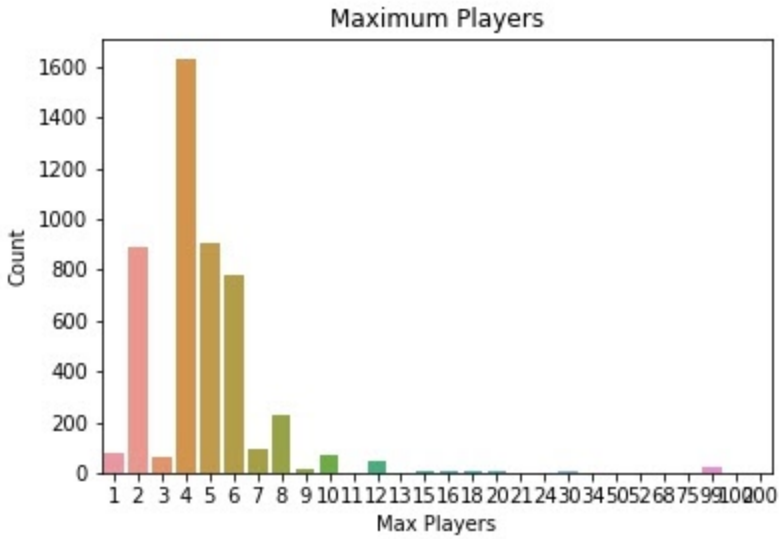

Figure 1

Next outliers will be removed so as to not skew the model. Outliers were found in 3 different columns: *max_players, max_time*, and *year*. Most games have a maximum player count of 4 with 2, 5, and 6 being the next most common. In figure 1 there seems to be a significant number of games at 99 players. There are only 120 games that have a max of 11 players or more and most times you will not have more than 10 people at a board game party. Therefore any game with more than 10 players will be removed. Some users report that their games last longer than 1440 minutes which is longer than a 24 hour period. These games are probably campaign style games that can be left set up and come back to. Most people can not or will not devote that much time to playing a board game. Only 0.3% of games are reported as taking longer than 480 minutes, and so those games will be removed. Looking at the data in the year column 97.5% of board games were released after 1974. All games before 1974 will be removed for the data set.

Exploratory Analysis

Exploratory Analysis is an approach to summarize the main characteristics of the data. Exploring the data can help data scientists better understand and select useful features that will impact the model's performance. Looking at both quantitative and qualitative features in more depth will bring better insights about BGG data.

Quantitative Data

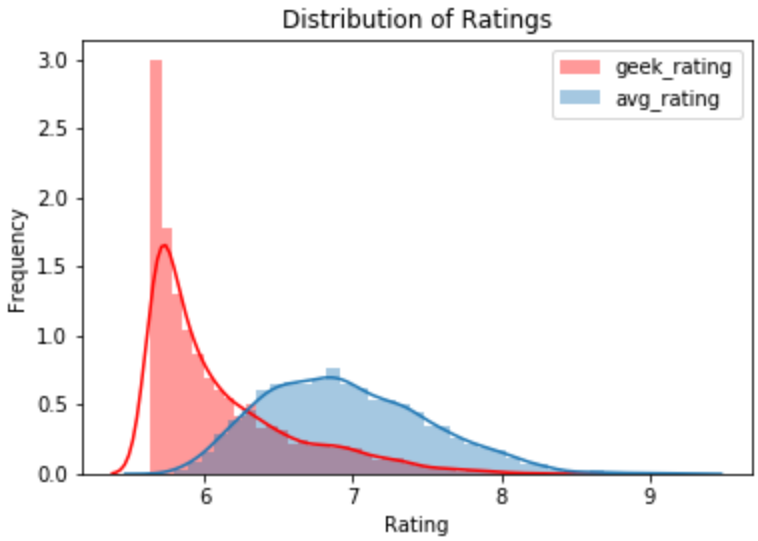

Figure 2

The BGG data has two different types of rating features. One is *avg_rating*, the average of all ratings from registered BGG users that the game has received, calculated by adding up all individual ratings and dividing by the number of ratings. The second is *geek_rating* that the rank column is based on. The feature *geek_rating* is based on the *avg_rating* but the number is altered by BGG using an algorithm and bayesian statistics to prevent games with relatively few votes climbing up to the top of the BGG ranking. The distribution of the two ratings we see in figure 2 that *geek_rating* is skewed

towards the left. This makes sense since the *geek_rating* is a ranking system of all board games 1 to 5000. The *avg_rating* has much more of a normalized distribution centered around a rating of 7.

Looking at the correlation between the ratings and other features will help in understanding the data. Instead of looking at the rating vs min and max players individually, a new column was added by combining the min and max players to form a player range. The most popular player ranges for board games is 2-4 players with 2 player games popularity coming as a surprise second. Looking at the average ratings of the player ranges, 1 player games are the best rated but also have the fewest board games with at least 50 or more games. 2-4 player games which are the most popular player count have a wide range of ratings.

The *weight* column has an average rating by users on how complex a game is. It is a scale from 0-5. We need to keep in mind that BGG is user driven data and that the data might be skewed towards people who are involved in the hobby and they may rate games differently than the average board gamer. The hobbyist might not think a game is complex and the rate is lower. But we will take the weight as a good measure. The *weight* distribution seems to be a normal distribution skewed a little to the left. *Weigh*t has a high correlation with *avg_rating* (0.545) and *geek_rating* (0.632). *Weight* will be a good metric for predicting ratings. Looking at the relationship between *age* and *weight* we see that there are a lot of games that have an age of 0 but have a wide range of weight. This is weird. Babies will not understand a game with complexity over 4. I think these are missing values. We will replace all 0 in the *age* column with the mean of the *age* column.

Playing time is divided into three features: *min_time, max_time*, and *avg_time*. The correlation between *avg_rating* and these three features is slightly positive and *geek_rating* doesn't have a correlation positive or negative with the three features. Looking at average time vs. max time in figure

3 below we see that they are very similar. If they are it would be redundant to have both average time and max time. We will remove the average time feature from the data set.

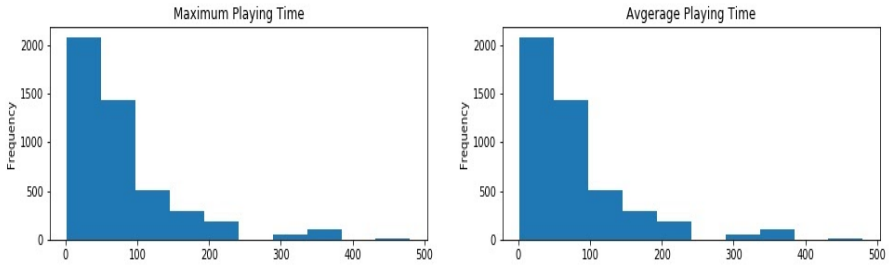

Figure 3

Qualitative Data

Trying to compare *mechanics, designers*, or *categories* with *avg_rating* and *geek_rating* will take some cleaning. Each of the qualitative features includes a list of values. In order to compare the individual values with the ratings the values must be separated out. There are 52 different *mechanics*, 84 *categories*, and 2582 *designers* . This is way too many features to include into the predictive model. We combined all *mechanics* of board games of 250 or less into a column of *other_mechanic*. The *categories* column has 10 of different values with categories about war. We combined these categories into one column called *combined_war*. We then combined all *categories* with less than 250 into *other_categoreis*. The *designers* column shows that 96% designers haven't made more than 10 games. It seems that the *designer* category will have very little influence on predicting the rating of a game and will be dropped from the data set.

Looking at two popular board game themes, Science Fiction and Fantasy, is there a statistical difference between the two themes average score or geek score. The *avg_rating* and *geek_rating* between Fantasy and SciFi board games are very close:

Avg rating for Fantasy game: 7.044

Avg rating for Science Fiction game: 7.103

Geek Avg rating for Fantasy game: 6.182

Geek Avg rating for Science Fiction game: 6.198

Using bootstrap replicates statistical analysis with ratings for both categories stating that the null hypothesis states that there is no difference between the two themes of board games while hypothesis claims there is a difference.

H0 = fantasy ratings = science fiction ratings OR fantasy ratings - science fiction ratings = 0

Ha = fantasy ratings != science fiction ratings

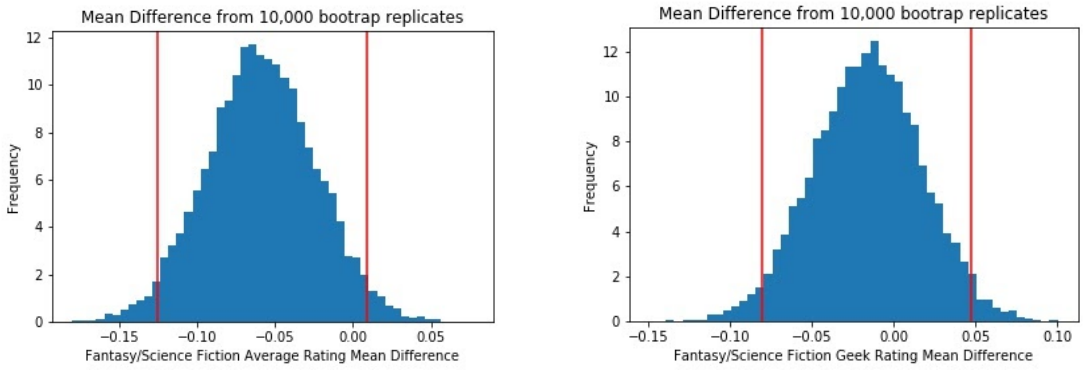

Figure 4

The above histograms (figure 4) is the mean difference in bootstrap deviation of 10,000 samples. The red lines mark the 95% confidence interval. The 95% confidence interval does contain 0 on both *avg_rating* and *geek_rating*. This means the null hypothesis can be accepted as true and that the *avg_rating* and *geek_rating* of the themes Fantasy and SciFi have no statistical difference.

Model

Now that the data is cleaned and organized, a predictive model may be built with sklearn algorithms. The model will be trying to predict the avg_score of each game using machine learning regression algorithms . The *avg_score* of a game is the average of all ratings given by users on a particular game. It hasn't been modified by BGG like geek_score has.

The data will be scaled using the normalization function. Scaling creates non-dimensional features so that those features with larger units do not have an undue influence on the classifier as would be the case if the classifier uses some sort of distance measurement as a similarity metric.

After normalization, algorithms need to be chosen to test with. Regressor algorithms will be used for the model to predict the average rating. Different regression algorithms will be tested in Linear, Decision Tree, Random Forest, and KNN. The algorithms will also be tuned with cross validation to find the best algorithm parameters for the model.

The scoring parameter the model will use is Mean Absolute Error (MAE). MAE is a mean of error (actual value minus the absolute value of the predicted value). A smaller MAE suggests the model is great at prediction. Each model during the RandomizedSearchCV and GridSearchCV will give a MAE and a mean and standard deviation will be taken of each of MAE of all models.

The data was split into 4 different feature sets to see which one is the best for the predictive model. Train and test data sets will be created and a kfold object to have the data split into 10 folds. After running all the different models with Random Grid Search or Grid Search cross validation we see the results in the following table:

Algorithm	Feature Set	training_score	validation_score
Linear Regression	1	0.300 (0.001)	0.305 (0.012)
Linear Regression	2	0.317 (0.002)	0.318 (0.016)
Linear Regression	3	0.310 (0.002)	0.311 (0.015)
Linear Regression	4	0.307 (0.002)	0.309 (0.015)
Decision Tree Regressor	1	0.252 (0.001)	0.298 (0.011)
Decision Tree Regressor	2	0.247 (0.001)	0.296 (0.009)
Decision Tree Regressor	3	0.250 (0.001)	0.295 (0.010)
Decision Tree Regressor	4	0.250 (0.001)	0.297 (0.008)
Random Forest Regressor	1	0.154 (0.000)	0.260 (0.006)
Random Forest Regressor	2	0.178 (0.001)	0.267 (0.008)
Random Forest Regressor	3	0.163 (0.000)	0.262 (0.007)
KNN	1	-0.000 (0.000)	0.342 (0.009)
KNN	2	-0.000 (0.000)	0.289 (0.009)
KNN	3	-0.000 (0.000)	0.293 (0.011)
KNN	4	-0.000 (0.000)	0.315 (0.006)

From the table above Random Forest Regressor with features set 1 had the best MAE validation score of 0.260 and a standard deviation of (0.006). Another interesting thing the table shows is that the KNN had a score of 0.000 which means that the KNN model was overpredicting.

Now the model is built and is ready to predict the average rating for games using the best estimator found. Using Random Forest Regressor with Features 1 we will test the test data that was set aside at the beginning. The model predicts a MSE of 0.498 with the test data This is higher than any of the training scores but still low so the model was a good predictor of average score.

Conclusion

Board Game Companies can learn a lot from BGG. Analyzing the data games that are 2-4 players and play within 2 hours with a higher weight are the most popular on BGG. This data could be used to determine what type of games that will appeal to the widest audience.

A predictive model was built with Random Forest Regression and a scoring metric of MAE. Using a custom feature list and a best estimator found with Random Search cross validation. Being able to predict an average score from users with different features would be useful for a company to determine what type of board game to make.

Some areas to explore further would be to see the popularity of different categories and if that affects the scores at all. Would the number of categories or mechanics affect the score at all? Could you have too many or too few? These are questions to look into.