Making Better Board Games: An Analysis of Board Game Geek's Dataset

Introduction

What makes a board game popular? What makes a game critically successful? How does the thematic category of a board game influence its success? Does mechanic, complexity, play time or number of players play a role? How might a game designer best optimize these various factors and should they publish on their own or do so through a major publishing company? These and similar questions constitute the core of this project, which concerns a dataset of board game data, mined from the gaming website Board Game Geek. Board Game Geek is a community repository of information about board games, including such significant details as publisher, year of publication, and mechanic, as well as subjective details about the game such as its rating average and its recommended number of players.

This project is aimed at board game publishers and developers. It creates predictive models that describe the main characteristics of critically successful games, as well as building a classifier that shows the broader historical changes that have taken place in the industry. I use the Board Game Geek site metrics average and geekscore as response variables. These metrics are based on ratings from users and a weighted average of those ratings, respectively. These metrics have some direct importance for producers of games, as each unit of average or geekscore is correlated with a significant increase in users of the site reporting that they own the game. Thus, any enterprising board game publisher or developer should pay close attention to these metrics, as high performance on them translates into better performance in the gaming market, to the degree that self-reported ownership on Board Game Geek is a useful proxy for the market in board games more broadly.

A secondary purpose of this project is to test a hypothesis about the historical changes taking place in the gaming industry. Conventionally, games are divided into Euro games and War games, labels that reflect two main styles: a shorter, more interactive and mechanic focused game (i.e. Settlers of Catan) and a historical simulation game, which often represents military campaigns, and takes exceedingly long amounts of playtime (i.e. Squad Leader, Axis and Allies). These were neve the only games out there, but the labels serve as the archetypes of two distinct eras in gaming. The transition to Euro games began in the 1990s (Settlers of Catan was published in 1995), but strong evidence of this transition emerges in 2000. Using 2000 as an inflection point, I test the hypothesis that these two eras in gaming have distinct features by means of a classifier.

Brief Description of the Data Set

This SQL database includes about 23,930 individual observations, as well as 80 different types of data. I found the dataset on Kaggle, at this link https://www.kaggle.com/gabrio/board-games-dataset, and it was scraped from the Board Game Geek website early in 2017. The key measures in this dataset are average and geekscore (as mentioned above). Year of publication has also turned out to be a crucial category too, because of the Euro game/War game split. Other parts of the dataset categorical, which includes data types such as category of game, game mechanic, and publisher. The dataset also includes the categories wanting, owned, and trading, which provides a sense of the ownership and relative diffusion of the game. Having reduced the dataset down to manageable components, I thought that I would provide the following key to the data in the dataset:

Id: the id of the board game, according to Board Game Geek

Name: the name of the board game

Description: A paragraph long description of the game by the publisher

Average Players The sum of the minimum and maximum number of players divided by two

AveragePlayingTime: Average length of the game in minutes

Category: The genre or genres of the game (some have no category)

Mechanic: The mechanic(s) used in the game

Publisher: the publisher(s) of the game

Average: The average rating of the game given to it by raters, a scale of 1-10

Weight: A measurement of the complexity of the game, on a scale of 1-5, provided by users of the site

Geekscore: A Bayesian average used by raters, on a scale of 1-10 (though there are no scores higher than 9)

Median: The median average rating of a game

Numcomments: The number of comments made on a game

Numweights: The number of people who have provided a weight rating (complexity) for the game

Owned: The number of people who currently report owning the game.

Stddev: The standard deviation of average ratings of the game

Trading: The number of copies of the game currently being traded on the Board Game Geek marketplace (in 2017)

Usersrated: The number of users that have provided an average rating of the game

Wanting: The number of users who reported wanting the game

The main problem with this dataset is that observations such as copies of a game owned are reported by a self-selecting group of individuals, this data is not the actual earnings data of a particular game (although intriguingly, not wholly inaccurate, according to the records, the game with the most owners is Scrabble). On the other hand, Settlers of Catan, which has sold 15 million copies, has only been registered by 4000 users on Board Game Geek. As others have noted, total sales figures of publishing companies are proprietary information, and are not released to the public.[1] And therefore, the category owned is the best publicly available proxy for a game's success. Arguments based on limited data are suggestive, not portrayals of the actual reality of the gaming industry.

Data Wrangling Steps

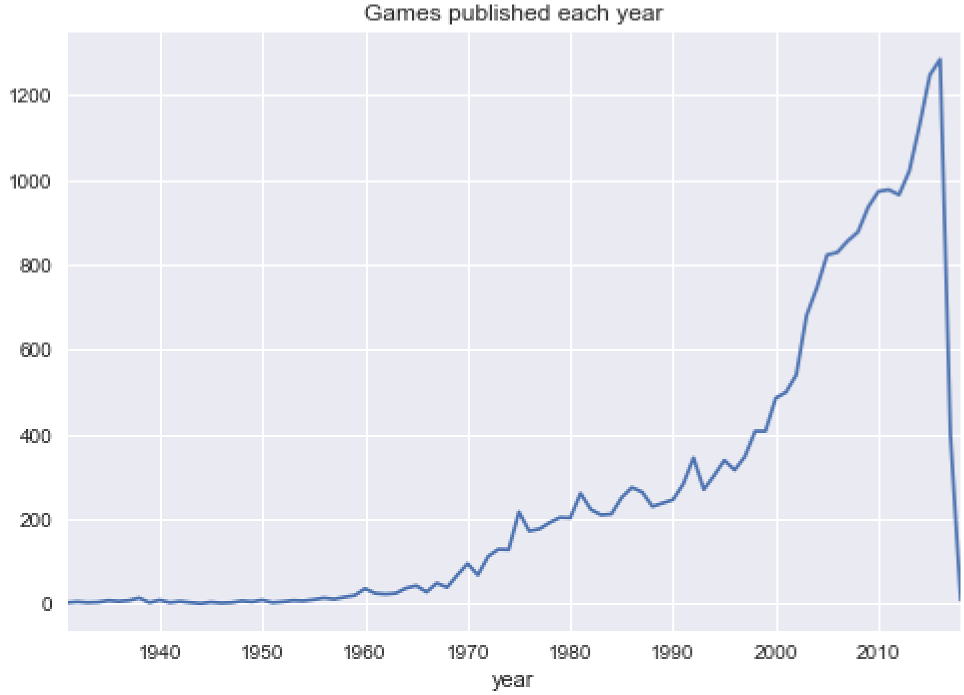
[1] This forum has a fairly limited list of sales data for a small set of games (about 352), which is perhaps the most comprehensive list available https://Board Game Geek.com/geeklist/22976/board-game-sales-stats-now-sources, but as far as I can tell, it just derives from promotional materials. It also does not shed much light on the problem, as this is a mere fraction of the games available in my dataset.

I undertook the following steps to clean the data:

- Filtered and deleted extraneous data, such as data that concerned videogames
- Deleted games with less than five ratings as a means of taking out any unpublished or poorly circulating games
- Filtered the database to only contain games after 1930
- Replaced subjective measurements in some categories such as "No" or "Best" with numeric values
- turned the SQL database into a csv file
- Turned maxplayer and minplayer, as well as max and min playing time into averageplayers and average playing time.

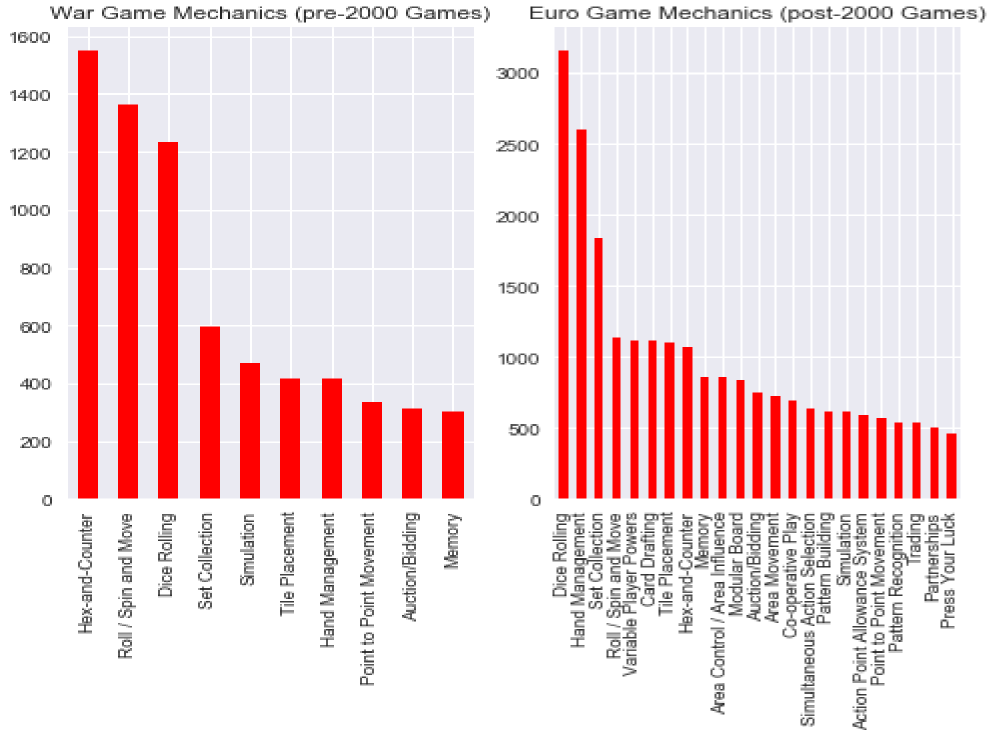
Exploratory Data Analysis

In the EDA, it became immediately clear that year was a crucial variable, as more games were produced after 2000 than in all the years before. Clearly, some sort of change has occurred, and the big jump in game production occurred around 2000 or so. I posited that this change represented a distinction between Euro games and War games, based on the popular hypothesis of change in the industry. The following chart illustrates this dramatically:



Using 2000 as a boundary, I found that in areas such as categories, mechanics, and publishers, the distribution of games had changed dramatically. To just give one example, compare the two charts, listing pre and post

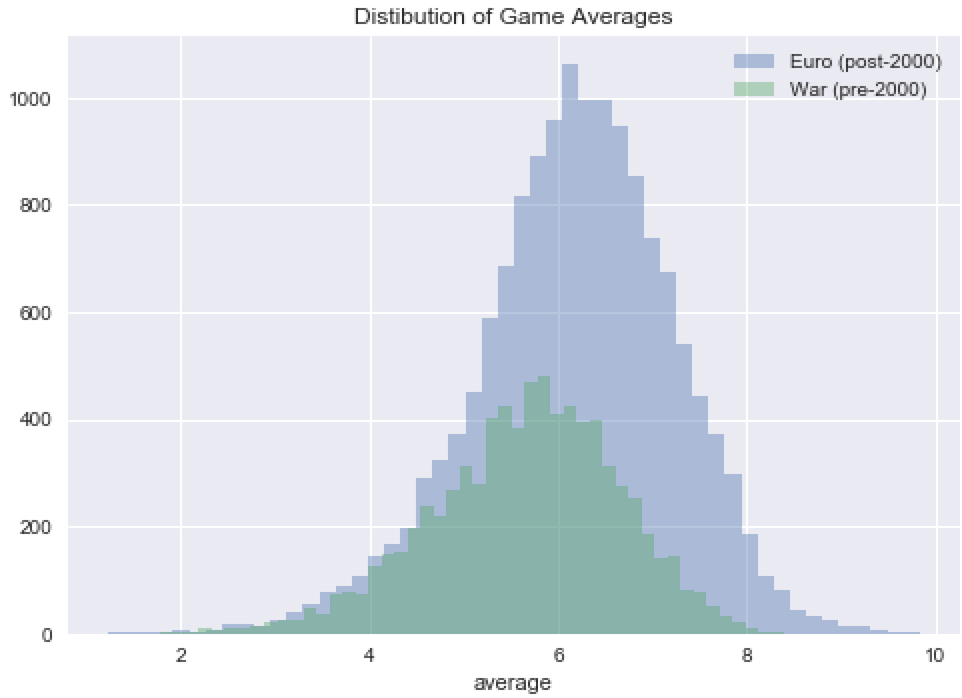
2000 games by mechanic.



War Game Mechanics (pre-2000 Games)

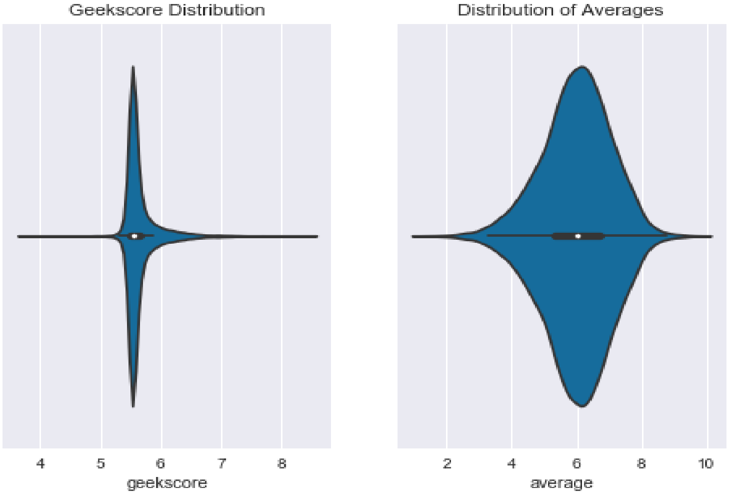
Euro Game Mechanics (post-2000 Games)

This is an aggregation of game mechanics with more than 400 instances, from pre and post 2000. As is fairly clear, the mechanics are quite a bit different. Hex and Counter and Roll/Spin and Move play very important roles in War games, but decline somewhat significantly in importance. Euro games are typified by a surge of new mechanics, which had little prominence before, such as Cooperative Play or Action Point Allowance Systems. I found similar patterns when looking at category, publishers and average ratings of the game, as demonstrated in this graph below:



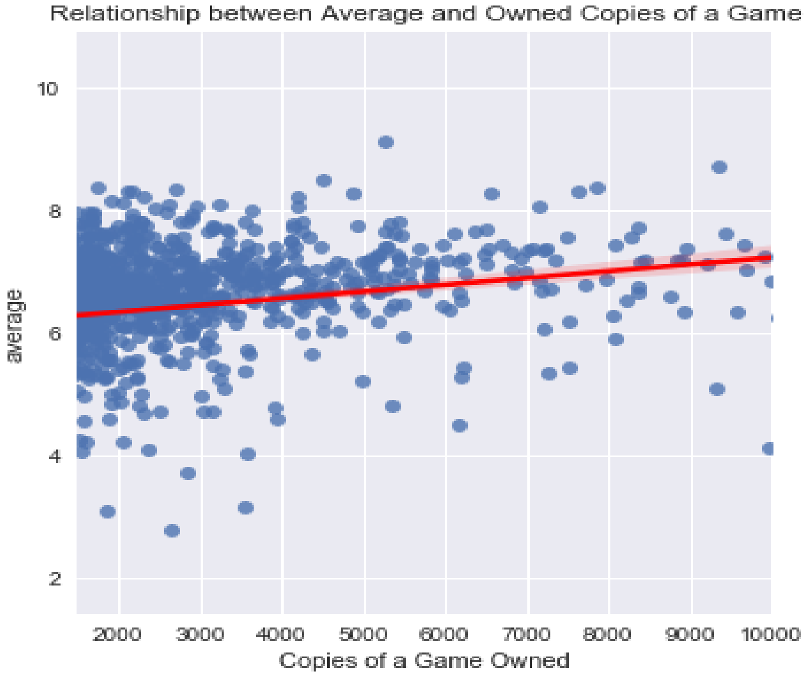
Average ratings of games post-2000 tended to be significantly higher than pre-2000 games, and there were quite a few more of these post-2000 games. This exploratory analysis suggested to me that there were significant differences between Euro and War games, and that it would be worthwhile to design a classifier to help distinguish between them.

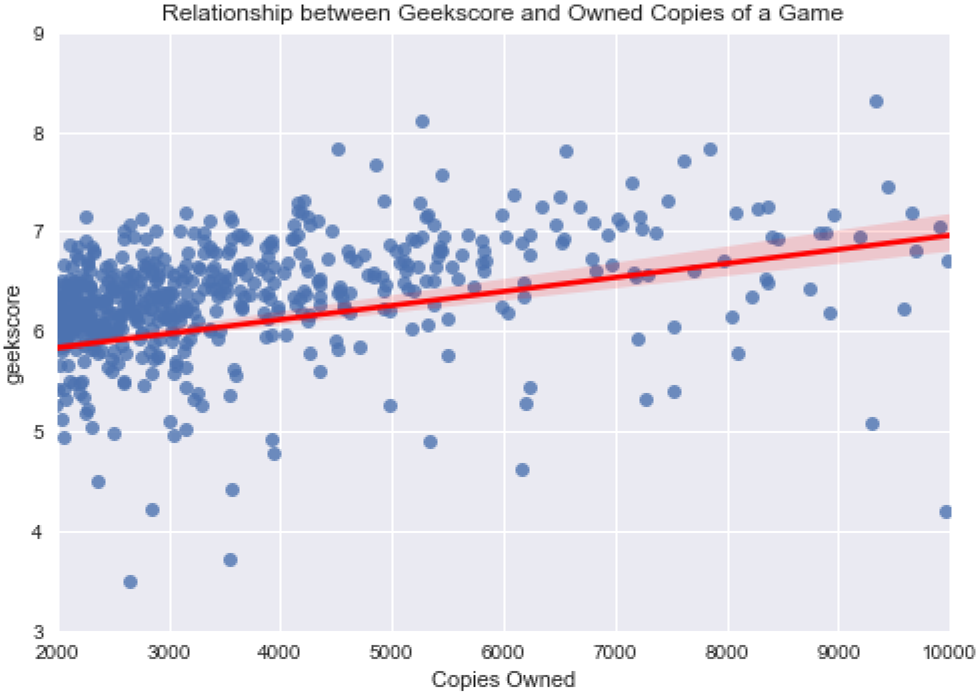
In my exploratory data analysis, I also looked at the distribution of the two response variables, average and geekscore.



These charts show the difference between these metrics of critical success. First, there are far fewer games with geekscores than averages (about 10,000 versus 30,000), thus, the geekscore graph is narrower. Second, the scale of the geekscore and the average differ. Geekscores run from 3 to about 8.5, whereas averages run from 1.5 to 10. Finally, the weight of the score is distributed differently. Most geekscores are between 5 and 6, whereas most averages are between 4 and 8, with the average average being 6. So as measures of a game's quality, these average and geekscore differ markedly, and need to be addressed with different models.

The core relationship between geekscore and average and the commercial success of a game is illustrated in this set of regression charts.





These lines show a direct relationship between average/geekscore and the number of copies owned. Based on a fairly unsophisticated regression analysis, an increase of .002 in average correlates with 50 more copies of a game being owned. For geekscore, the results are even more dramatic, each .002 increase correlates with 111 more copies being owned. To put it simply, better ratings lead to more people purchasing a game. These models also probably understate the effect to a certain degree, as they are probably held down by several massively successful games like Yahtzee or Monopoly Deal with low averages and geekscores. Such games are legacy institutions which serve a mass market public, rather than new publications, which cannot count on similar levels of distribution and saturation in the market place. Hence, the very extreme outlier cases are exceptions that prove the rule, suggesting that average/geekscore are crucial to the commercial success of any game. Similar linear correlations exist between average/geekscore and other categories that are proxies for a game's commercial success, such as copies of a game being traded or copies of a game on someone's wishlist (wanting). This was the core discovery of this exploratory analysis, finding a linear relationship between the commercial categories and the geekscore/average. The next section looks at the factors related to the formation of these scores, as they have such clear significance for the success of a game to a designer or a publisher.

Building models, answering questions

There are two main questions to investigate here. First, what factors provides a game with good ratings (including both geekscore and average)? This question is crucial, as these factors are directly tied to a game's commercial viability. Second, my exploratory data analysis has found that there are two main categories of games, whose primary feature is temporal, and I used a classifier to distinguish between these two main categories of games.

For the purpose of model building, I changed the formulation of the data from Board Game Geek, grouping some data together under labels, so as to more easily replace it with dummy variables. Using the different playing time columns, I created an average playing time column, which was reported in minutes, and I designated games that were less than 60 minutes as short, between 60 and 120 minutes as average, and over 120 minutes as long. For the category average players, I categorized 1-2 person games as fewplayers, 3-5 player games as average number of players, and six or more players as many players. I did something similar with publishers, denoting publishers with over 90 games as big publishers, between 90 and 24 games as average sized publishers (I noticed that a larger cutoff for average sized publishers ended up categorizing several significant publishers as small), and < 24 games were identified as small publishers. I also turned the columns category and mechanic into lists, and then transformed them into dummy variables, which ended up providing me with a fairly large (if somewhat sparse) dataset to work with. Finally, I designated all pre-2000 games as War games, and all post-2000 games as Euro games, creating a column gametype, which could serve as a dependent variable for my classifiers.

Methodology

I built models for three response variables: geekscore, average, and gametype. The first two response variables were analyzed with regression models and the third one with a classifier and a Naïve Bayes text classifier. As I only reported on the models that proved the best at predicting these response variables, I will briefly describe the process by which I arrived at these results.

For geekscore and average, I first split the data into training data and test data. I then split the training data into training and validation data, using a for loop for tuning parameters (although in the case of Random Forest Regression, I used GridSearchCV). For these two response variables, I built a linear regression model, a version using the lasso technique, and another version with the ridges technique. I also tried Decision Tree Regression and SVM Regression, tuning the parameters of each. I ultimately settled on Random Forest Regression for both geekscore and average, as it had a far lower root mean squared error than any of these other models.

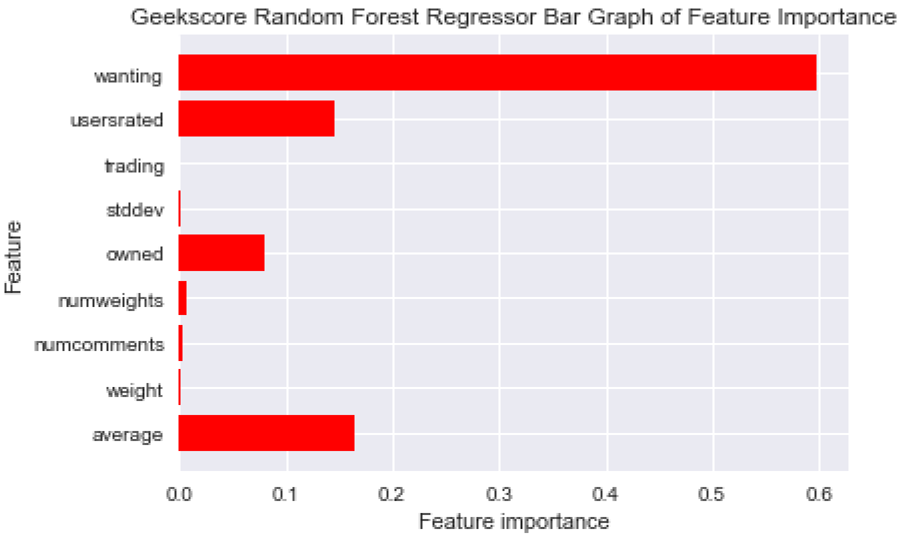
For the classifier of gametype, I also split the data into training and test data, and then split the training data again into training and validation data. I used for loops to tune parameters on the following models: SVM, Decision Tree, Logistic Regression, and a Random Forest Classifier. Based on the accuracy of the Random Forest Classifier, I ultimately chose to report and employ this as my model. Finally, the text classifier went through a process of development and experimentation, as I originally included the game's description along with the text columns category, publisher, and mechanic. Before implementing the Naïve Bayes model, I also built a Latent Dirichet Allocation for this text data, an unsupervised method for clustering topics in text data. To do this, I processed the data by eliminating stopwords, stemming the words, and eliminating any non-text elements from the description. My LDA analysis ultimately proved of little worth reporting, and I instead built a Naïve Bayes model, which used text data to predict whether a game was a Euro game or a War game, which I will detail below.

Analysis

My first model uses Random Forest Regression to look at what factors influence the formulation of the geekscore. I used a random forest regressor, because I had a variety of different features in the dataset, many of which (I assumed) were somewhat extraneous to the project. So this seemed like a good way of building a lot of different decision trees and seeing which features performed best. I used sklearn.model_selection train_test_split to split the data (.75 for train, .25 for test), and used GridSearchCV with 5 fold crossvalidation to optimize for parameters (100 trees, max_depth=9, min_samples_leaf=7, max_leaf_nodes=83). This model provided me with an extremely low RMSE:

Root Mean Squared Error on Training Data: 0.023022710400284016

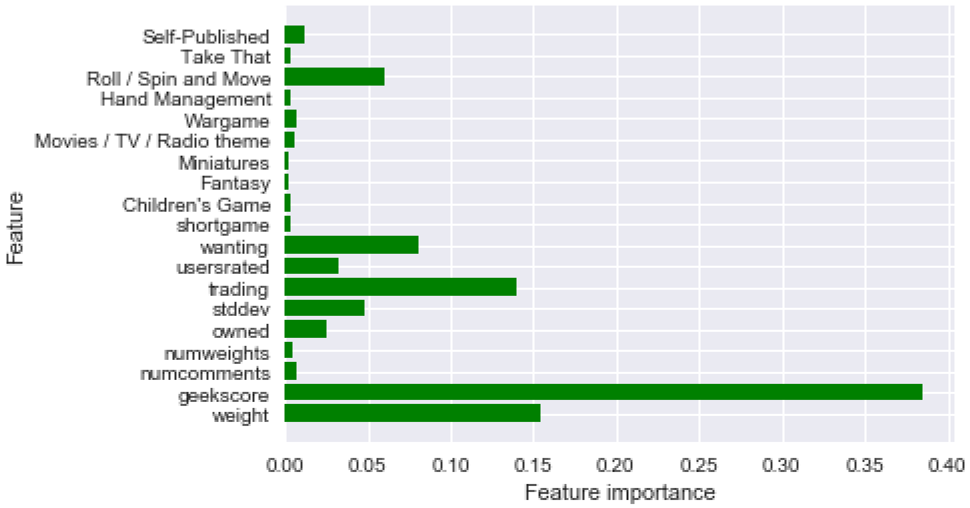
Root Mean Squared Error on Test Data: 0.03631883462324047



I filtered the feature importance at an importance of 0.0001, and it identified the "wanting" feature as highly, highly significant in the geekscore accounting for .77 of the feature importance. Also significant were usersrated, stdev, average, trading, numweights, numcomments, and weight. Wanting is a numeric count of how many users have expressed a desire for a particular game. I took out this variable to see how the model faired without it (as it was so predictive that I was worried that there might be some error), and it had an equal rmse, but shifted more importance to the other variables mentioned above. The game length and some categories and mechanics had a relatively small feature importance (below .00001). Based on this model, the features that are the most predictive of a geekscore concern user engagement, rather than category, mechanic, or length of game. This makes sense, as geekscore is fundamentally a measure of those who rated it, and a higher number of people rating and engaging with it would seem to drive geekscore up, in one sense. However, it is a bit surprising in a way too, as one might think of geekscore as elite or critical opinion (as it shores up user ratings with weights), but it is instead highly responsive to user ratings and opinions. Geekscore seems to be shaped primarily by user interactions with the games.

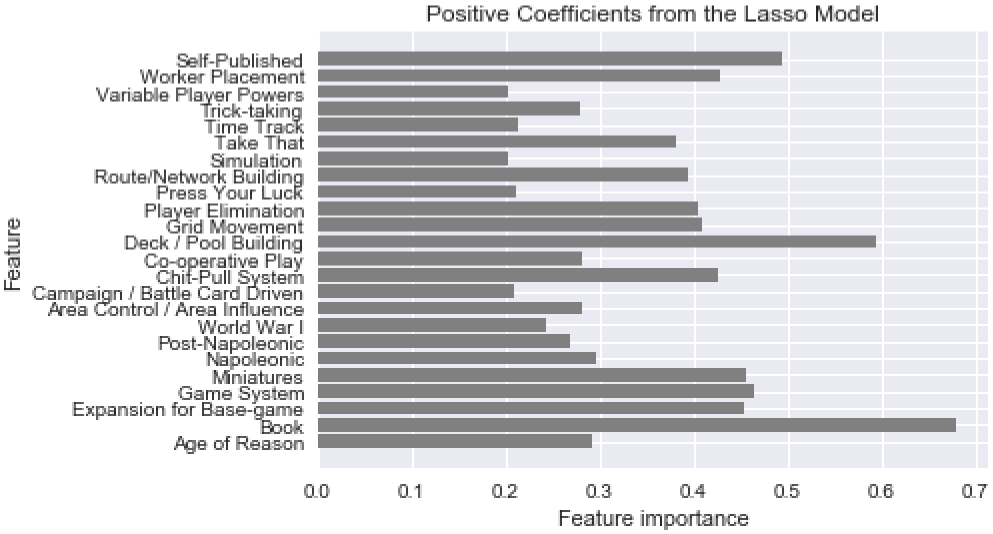
I also trained a model to predict the average of a game. Averages are purely the input of users, and are thus, much more vulnerable to skewing than a category like geekscore. Averages tend to be higher, if a bit more variable, than geekscores. I again used a random forest regressor to account for the many categories in the dataset. I used train_test_split with a random state of 1 and a training size of .7. I cross validated the model using GridSearchCV, and arrived at the best parameters, which were n_jobs, min_samples_leaf, max_depth, and max_leaf_nodes.

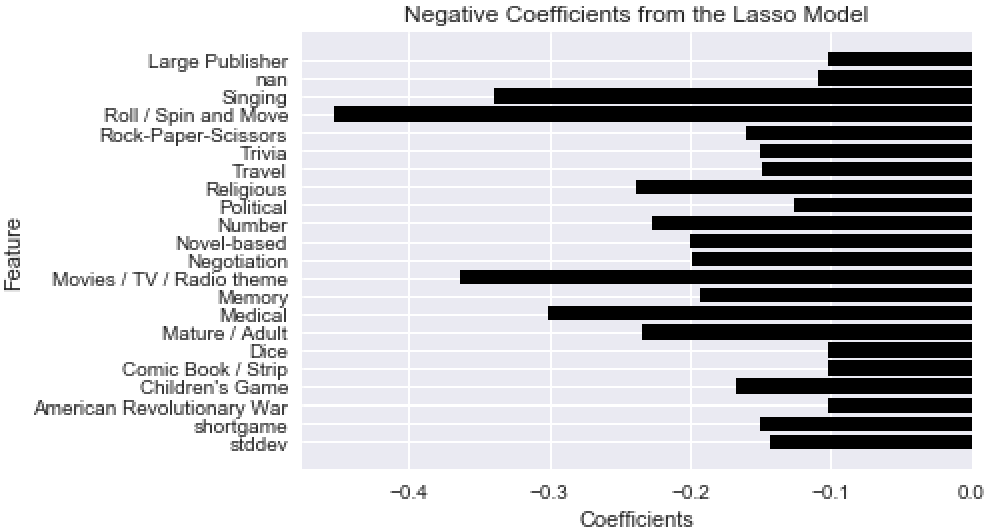
Root Mean Squared Error on Training Data: 0.7132977515882472
Root Mean Squared Error on Test Data: 0.7493511953604849



This model provided me with a fairly low RMSE, but one which is quite a bit higher than geekscore, but that is to be expected because of the specific features of averages. Most predictive was geekscore, which is interesting, as there seems to be less division between these two response variables than the exploratory data analysis suggested. I filtered the importance of features at 0.002 in the graph above, and found that wanting, weight, trading, and stdev were highly important in forming the trees for the random forest. This resembles what was mentioned above with geekscore, which also used similar variables extensively. Average also was influenced by some other categorical variables, such as the mechanic Roll/Spin and Move, which seems oddly important here. Self-published is also an important variable, which seems surprising, but makes sense inasmuch as many games are first self-published before being picked up by major publishers. Other games that have this level of feature importance might have dedicated fanbases that raise the average (Fantasy 6.26, Miniatures 7.47) or be loathed (Movies/TV/Radio Themes 4.98, Children's Game 5.23). I also built a Linear Regression version of the model, which uses the lasso technique to reduce dimensionality, and while not as accurate, it is highly informative about the consequences of particular mechanic/category/design choices, and thus, I include the two charts here. Despite the somewhat higher MSE, it provides a more informative picture of the effects of category and mechanic, as feature importance is neutral, and the lasso model provided positive and negative coefficients. As I filtered it positively (> 0.2) and negatively (< -0.1), it provides a very useful visual of how a specific category or mechanic can influence the average of a game, based on coefficients (NB—nan in this case is the category None).

```
Number of features used: 146
Root Mean Squared Error on train data: 0.8782871750708969
Root Mean Squared Error on test data: 0.8892818363281227
```





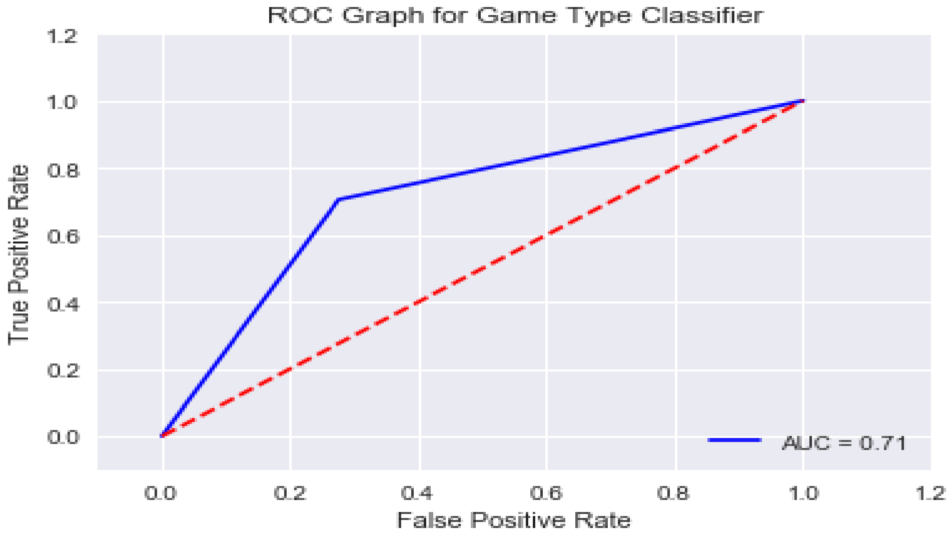
While the RMSE here is quite a bit higher, this model provides a much more nuanced account of the penalties for particular categories and mechanics. For instance, there are major penalties in the average for games that use Singing, Roll/Spin and Move or Negotiation. Certain categories are also poorly rated, such as Religious, Political, Number, Memory, Trivia, and Children's Game; these might seem to reflect more traditional board games, which have fairly unsophisticated mechanics and often obvious goals. Certain types of thematic games also do poorly, such as Medical, Mature/Adult, Comic Book/Strip, and Movie/TV/Radio Theme, all of which reflect the difficulty of making games that work with this material, or more accurately, the profusion of bad games in these categories. For instance, many movie series have poorly implemented board games as part of their marketing (I count seven Twilight board games alone in this dataset). Positive coefficients tended to be focused on mechanics, and tended to include somewhat newer mechanics like worker placement, cooperative play, and deck pool building. Expansions and Miniatures have higher ratings here, I think because the audiences for these games are self-selecting, and only those who really love the original game or spending a lot of money on miniatures will actually rate these games.

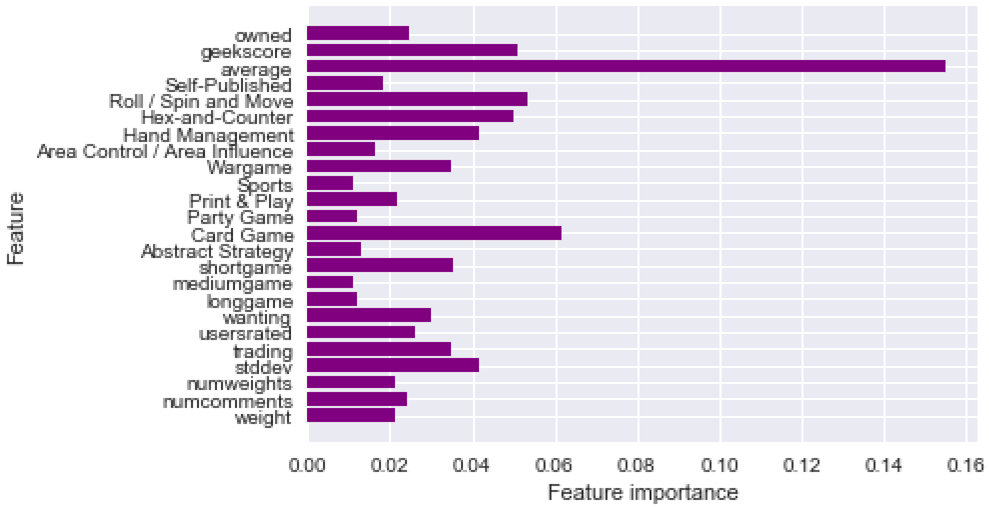
Classification Models

I wanted to test my hypothesis that the distinctiveness of Euro and War games was a meaningful way of organizing the games in this dataset, drawing on the popular narrative of dramatic change in the gaming industry. To do so, I used a classifier. I (somewhat arbitrarily as the underlying data is more complex) assigned a label to all games after 2000 as Euro games, and pre-2000 as War games. I tried a few models, and the most successful was a Random Forest Classifier with the following parameters, n_estimators =100, max_depth=27, min_samples_leaf=11, max_leaf_nodes=90. Random Forest Classifiers also are good for dealing with datasets with extensive features, and weeding out the more or less extraneous ones. There were also more Euro than War games, so I weighted the classes in such a way that they would be balanced, which undersamples the larger class.

Accuracy on Training Set: 0.7323144886872426

Accuracy on Test Set: 0.712216186098324

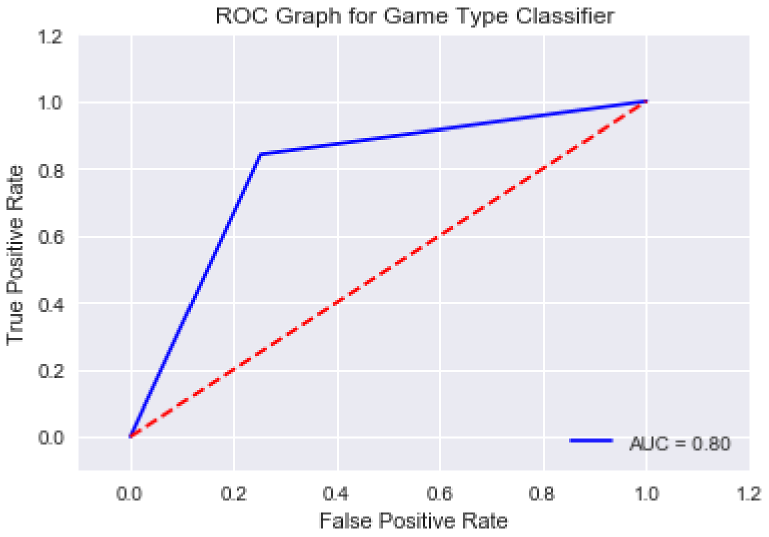




This is a significant accuracy, if not an especially high one. It is rather significant, given that the labels are arbitrary (based on years, rather than intrinsic features of the game). The important features here are all relatively low numbers; the largest is average (and only at 0.155), suggesting that there is a decisive division in terms of averages between Euro and War games (other statistical analysis suggests that Euro games have higher averages). Other areas include some general metrics of engagement, such as numweights, numcomments, and usersrated, as well as specific differences in mechanics, genre, and runtime. Hex and Counter and War game, which seem to be more typical of War games, whereas mechanics like Hand Management and Card Drafting, as well as the category Card Games are likely to be more typical of Euro games.

I wanted to keep testing my assumption that Euro/War games are significantly different through other means, so I built a Multinomial naïve Bayes text classifier to classify them based on a series of categorical text data, which included things like category, mechanic, and publisher (I had originally included description as well—a paragraph long text category, but it made the model the worse). I combined these text categories and passed them as a dataframe column into a vectorizer, and then trained a model with an alpha of 0.001, asking the model to classify games as either Euro games or War games. This dataset proved to be significantly more accurate than my random forest classifier, with the following accuracies:

Accuracy on training data: 0.873799
Accuracy on test data: 0.808549



The final keyword step of the argument looks at the probability of a particular game being associated with Euro or War games, and classifies many publishing companies with Euro or War games. Indeed, the model suffers acutely if publisher data is removed, suggesting that a core difference between the type of games is not necessarily just mechanic or category, but the publisher involved in the process. Indeed, a transition between Euro and War games may be based on the rise of new publishing houses.

Conclusions and Recommendations

For board game designers, it is important to pay attention to the factors that influence consumer adoption of and use of board games. In a more segmented marketplace, games like Clue, Risk, and Scrabble offer trajectories that are not able to be duplicated (and indeed, similar such games would probably inspire hostility from most well informed gamer groups). On the other hand, critical success in the Board Game Geek community offers an alternate path to commercial success in a more mainstream sense or just among the community of hardcore gamers. Measures of geekscore and average suggest that particular mechanics, genres, and categories can influence the success or failure of a game, and designers should be thoughtful about making a game in generally loathed categories such as Movie/TV/ Radio Theme and Religious or with a mechanic that is problematic, such as Negotiation.

More important than the specific insights that can be derived from this model is the more general portrait of how average and geekscore are created, namely by doing well on a series of metric internal to the Board Game Geek website. For geekscore, these metrics are primarily wanting, usersrated, owned, and average (along with the much smaller numweights and numcomments). For average, these metrics are wanting, trading, stdev, owned, and weight (with numweights and numcomments being a lot smaller). This modeling suggests that publishers ought to spend as much time promoting and spreading their game among hardcore devotees as playtesting it. Average and geekscore are dictated primarily by how the users of Board Game Geek interact with and think about a game, as well as their willingness to rate, acquire, and discuss the game. A more active outreach and publicity campaign could provide high returns for a game publisher or designer.

The classifier model generally shows that there is a decently strong distinction between Euro and War games, and that there is some truth to the popular narrative of a changing board game market. These distinctions are fairly significant, and include mechanics, categories, averages, and publishers. However, the relatively low RMSE scores (.75-.8) suggest that the story could be made a bit more complex.

Future Directions

My operating theory is that more engagement with a game would produce higher geekscores and averages, and thus, more copies sold. And Board Game Geek does collect data on number of views and other measurements of engagement (just not in an easily scrapable form), so I could use this data to confirm my hypothesis. Another route might be to look at the role of awards in shaping public tastemaking and acquisition in the board game industry—thus, for instance, what exactly does winning the Spiel des Jahrens mean? Does that translate directly into more games owned? Finally, and a much broader question, is what exactly helps a game transition from the gaming world to the mainstream? For instance, a game like Pandemic or Ticket to Ride is now sold in retailers across America, and how do games move from specialty play to showing up on the shelves alongside Monopoly? My theory is that this is a huge driver of returns on investments for board games companies, and that the process is probably fairly random, as the data is limited. Nevertheless, it would be possible to mine catalogs of retailers to see what board games they stock, and then perhaps have a sufficient amount of data to see how the jump might have been made.

As for the classifier, it works well enough for what it is, but clearly the underlying data is a bit more complex and the Euro/War dichotomy is not wholly sufficient to account for the dataset. A more complex version might try a series of different labels—perhaps by eliminating some categories or grouping others together—or might also use unsupervised learning to organize the dataset in a more efficient manner.