Board Game Prediction

Final Presentation (Model and Results)

Michael McCann | April 2022

Project Goals – Business Use Case

I had three goals for this project at the outset: use regression to predict board game rank, use categorization to determine if a game would make it in the top 100, and develop a game recommendation engine.

- **Board Game Rank:** Use a regression model to predict ranking on Board Game Geek.
  - Use Case: Assist a game publisher make games that will succeed.

- **Board Game Top 100:** Use categorization to determine if a game will break into the top 100 on Board Game Geek (a metric used in the industry).
  - Use Case: Assist local store determine if a new game will be 'popular' prior to placing orders or otherwise taking up inventory space.

- **Recommendation Engine:** Develop a tool using user ratings to recommend games to people looking for new experiences. Ended up falling outside the scope of this project and is on hold.
  - Use Case: Help gamers find new games they will love and enjoy!

Data

Data is derived from two separate Kaggle datasets merged into one larger dataframe for machine learning and a few smaller dataframes for EDA and visualizations.

Sources

- Board Game Database from BoardGameGeek: Comprised of nine CSVs looking at various facets of 22k board games (Ratings, Designer, Artist, Publisher, Theme, Mechanics, etc). Accessed via BGG's API.
- Board Game Geek Rankings: Top 5000 games as ranked on Board Game Geek going back to October 2018. Data is scraped weekly from BGG and published to Kaggle.

Combined these datasets based on the games that ranked in the top 5,000 games over the past 2.5 years. This amounted to approximately 6,000 games during that period. Eliminated inconsistencies in the Board Game Database.

Caveats and Considerations

- Board Game Database included a number of obscure games with little data or ratings information.
- Formatting for the ranking data is not ideal and required some work to get into a useable format.

Best Model

LightGBM outperformed all other regression models across all measured metrics for both our classification and regression problem.

<u>Regression</u>

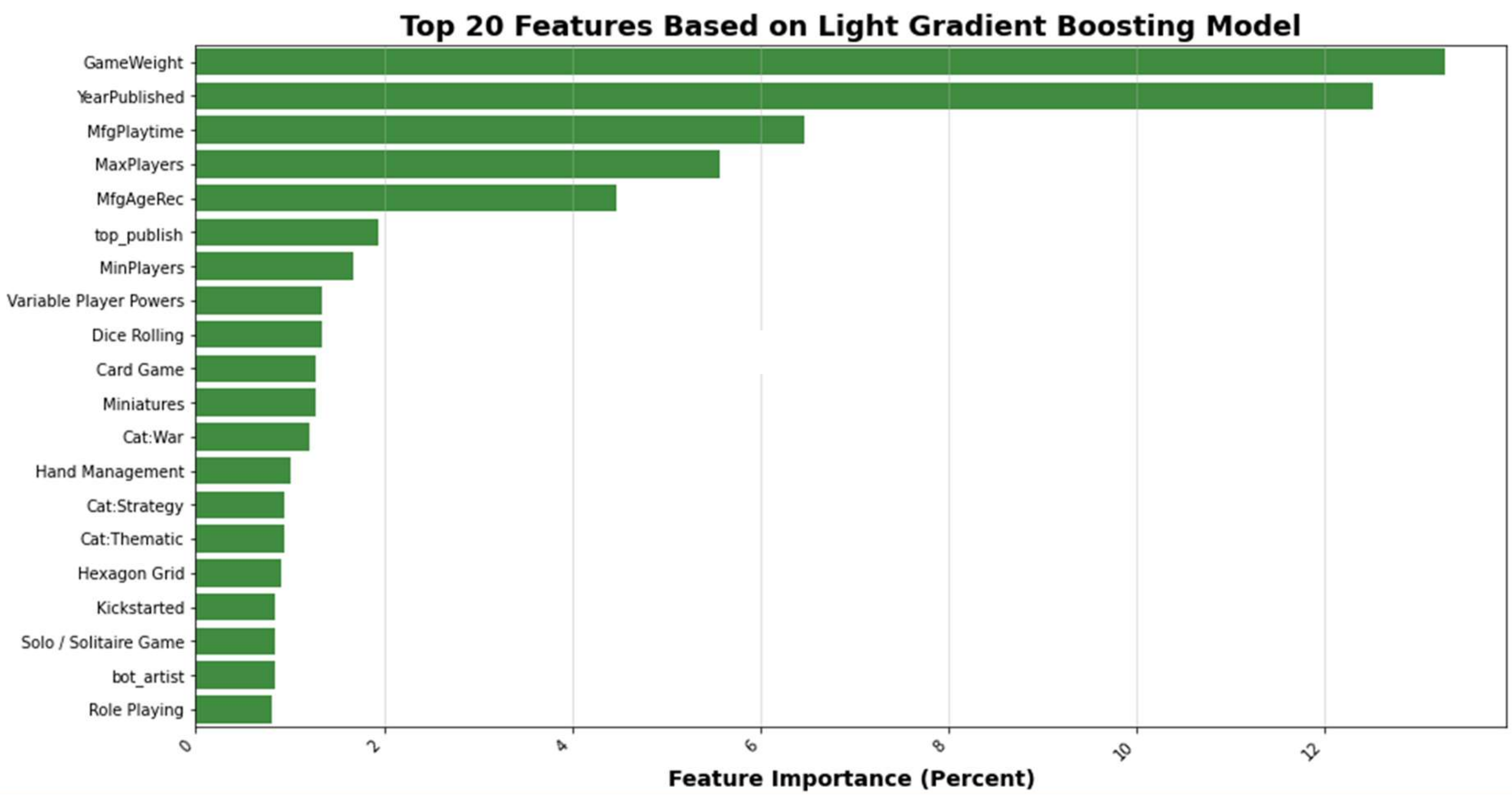
- Dummy Model explained 0% of variance within the dataset.
- Majority of other models explained 50 to 55% of variance.
- LightGBM explained 61% of variance.

<u>Classification</u>

- Majority of models predicted 0-30% of the True cases (games that were in the Top 100).
- LightGBM was able to predict about half of the True cases (games that were in the Top 100).
  - Very small number of false positives (games that were not in the Top 100 being categorized as such).

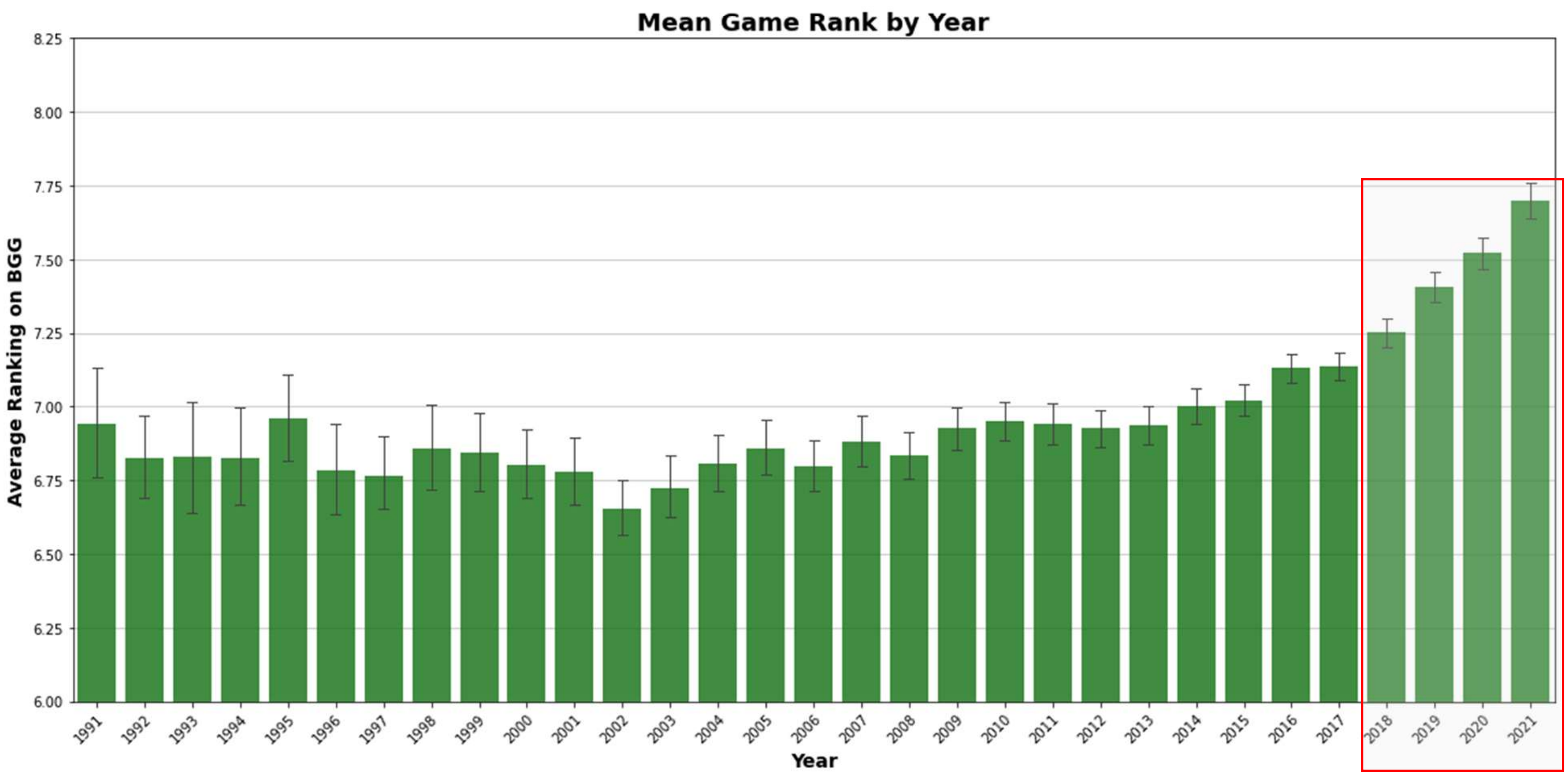
Feature Importance

What columns or features are most important to the model? Higher value means it is more important to the model. Does not necessarily mean it is a positive/negative correlation.



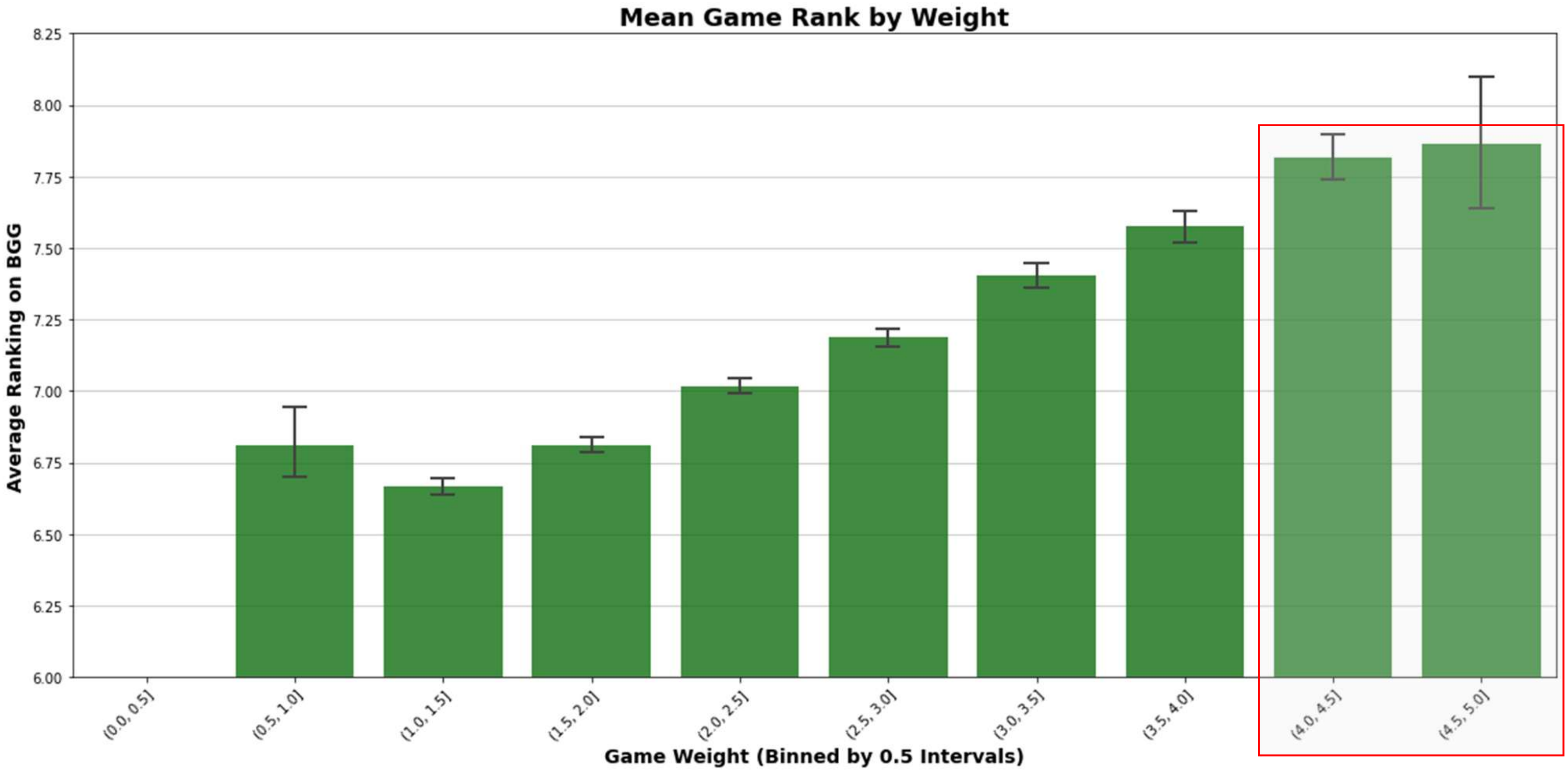
Year Published – Recency Bias

It appears that newer games get ranked better. This could be that older games are not as attractive anymore and get ranked lower compared to their newer counterparts.



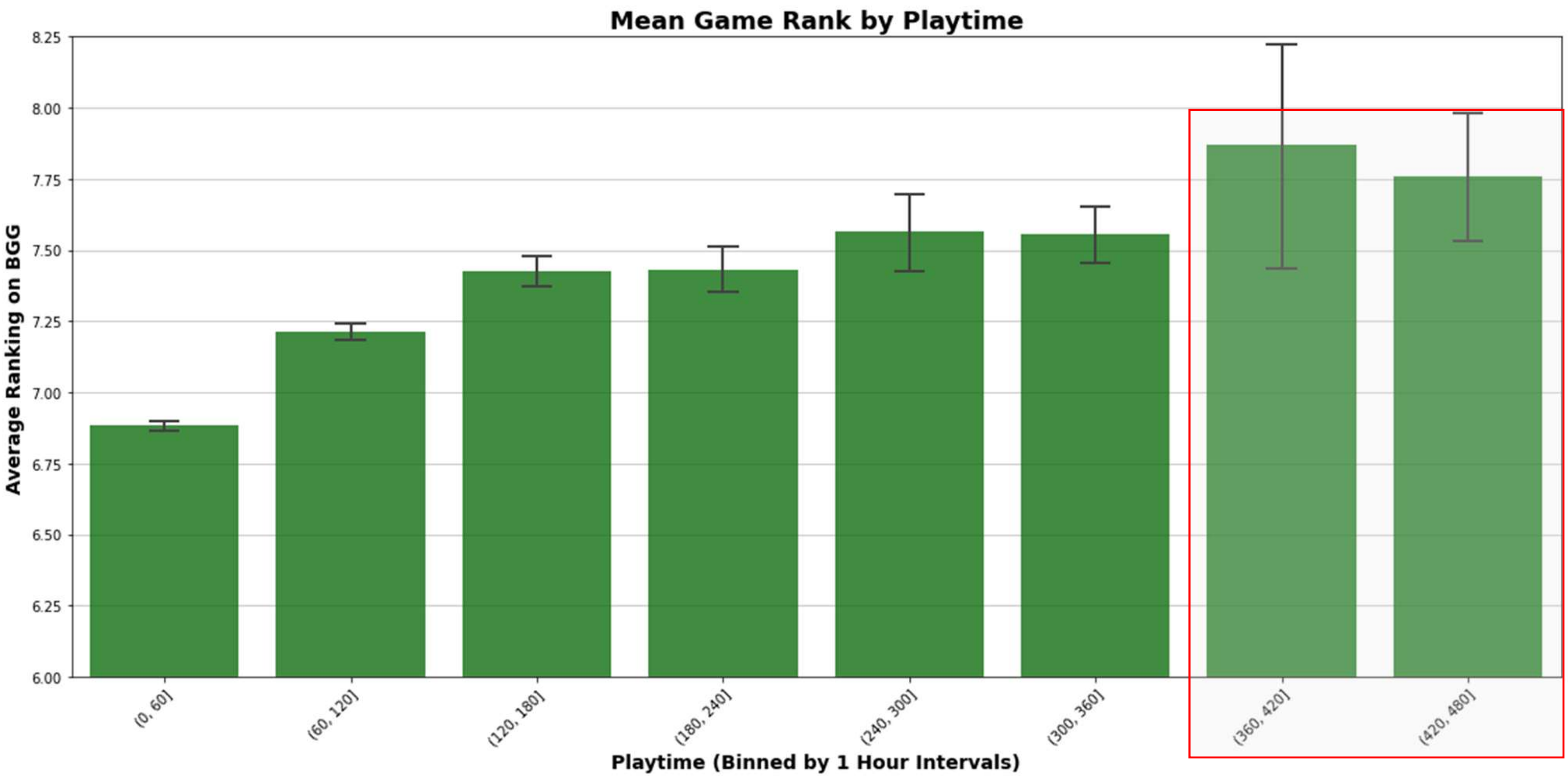
Game Weight – More Complex? No Problem.

The more complex a game the higher on average it was rated. This could be pointing towards the trend towards ‘euro’ games.



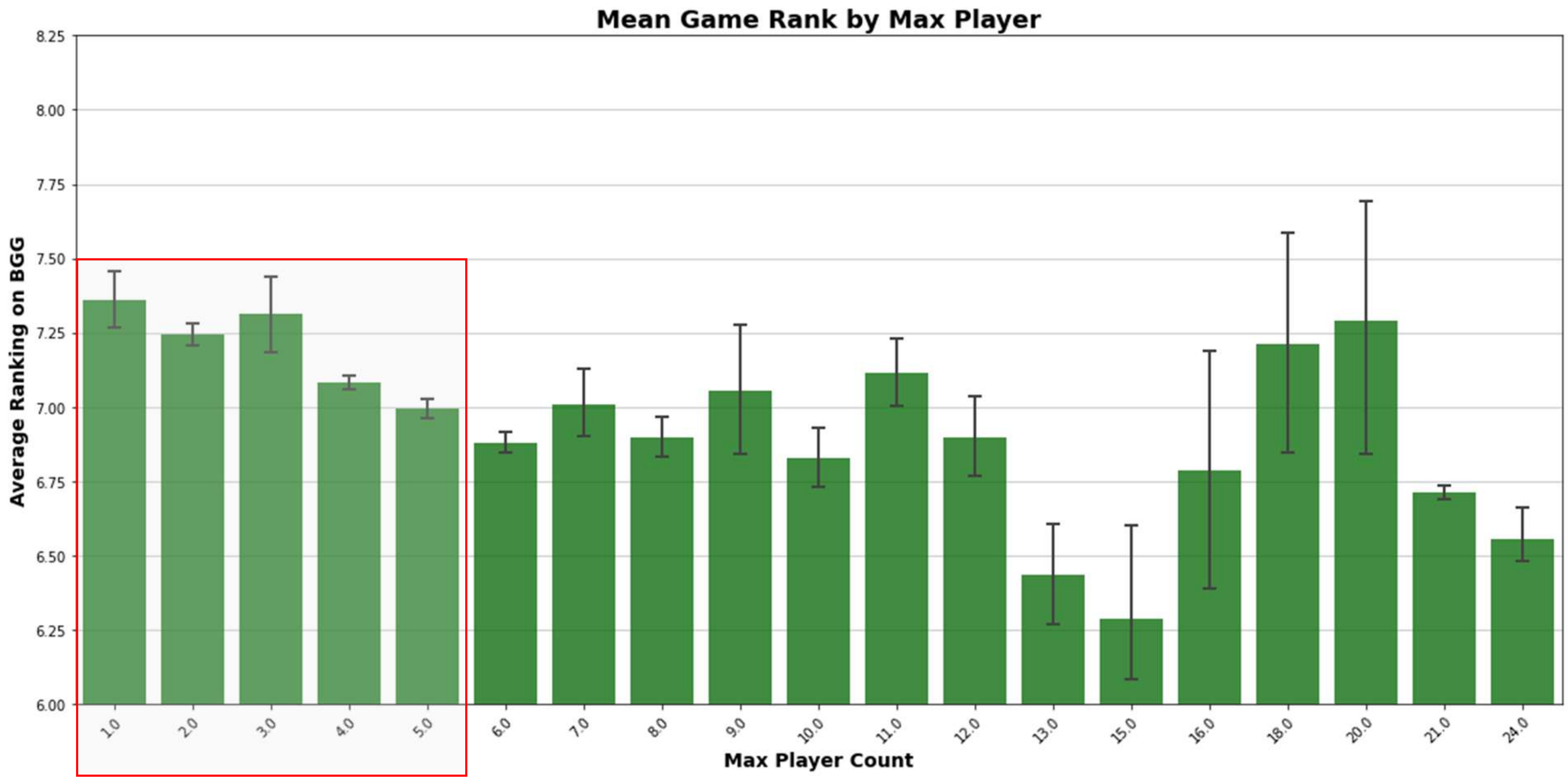
Playtime – We want to Play More!

Longer games proved to be more popular than shorter games. This included legacy games or games with ongoing sessions/campaigns.



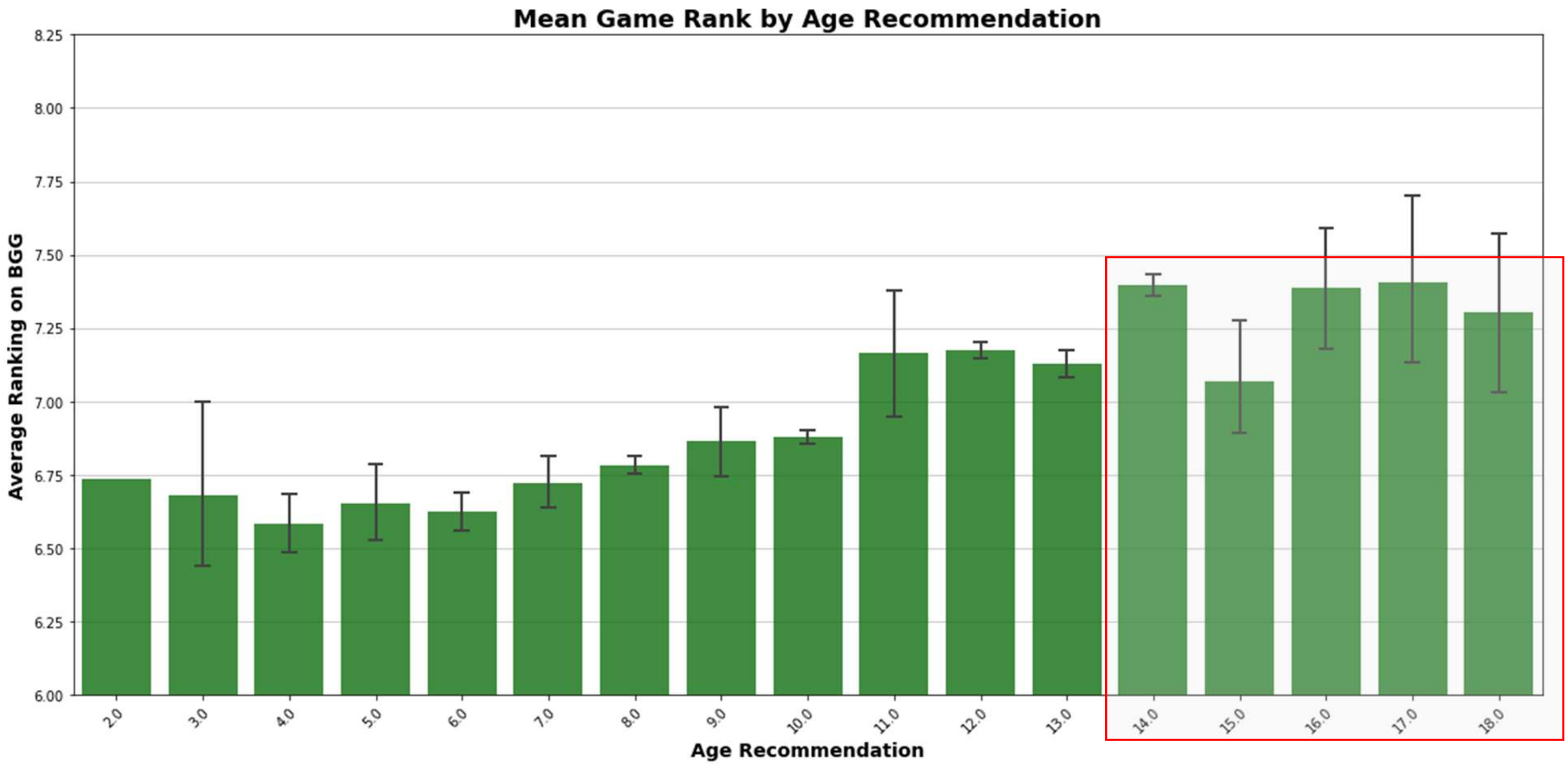
Number of Players (Max) – Sometimes Less is More

Games with a lower player count did marginally better than games with more players. Games with 18 and 20 players appear to be outliers with higher rankings on board game geek.



Recommended Age – Games Not Just for Children

Games targeting players age 14 and up do better than games aimed at children. This could be because children are less likely to rate games online and/or a move towards this being a legitimate hobby.



Important Categories, Themes, and Mechanics

The below lists were not heavily weighted based on feature importance. However, they were the highest weighted respectively and might be worth considering when developing a new game.

Category	Theme	Mechanics
War	Fighting	Variable Player Powers
Thematic	Sports	Dice Rolling
Strategy	Fantasy	Hand Management
Family	Science Fiction	Hexagon Grid
Card Game	Civilization	Solo / Solitaire Game
Party Game	Theme_Colonial	Role Playing
Children's Game	Medieval	Scenario / Mission / Campaign Game
	Video Game Theme	Deck / Bag / Pool Building
	City Building	Drafting
	Humor	Cooperative Game

Outlook:

Recommendation: Make games that are more complex, that are geared towards older players and have a longer playtime. Most importantly… make fun games!

Results:

- Regression – Our final model is statistically significant but not overwhelmingly so. It is possible that this is as good as the model gets which could be good for the gaming industry on the whole (if not for the specific company).
- Classification – The 50 / 50 positive result is promising and will help narrow what games shops should pursue to avoid having excess inventory or purchasing ‘bad’ games.

Next Seps:

- Regression – Continue to tinker with the model, perhaps trying a neural network. Additional, feature selection may help tighten the model.
- Categorization – Testing with sampling to avoid some of the downfalls of this dataset may help increase the accuracy of the model.

Stretch Goal – Recommendation Engine:

- Will be pursuing this in the future.

Questions?

Prepared by: Michael McCann

Contact: msmccann10@gmail.com