What Do Consumers Want In A Boardgame?

GameStat Visionaries
Andrew Kulp, Talisa Manker, and Garrett Zimmerman

Recap of Project Objectives

- **Business Problem**
  - Build a better boardgame
    - Design a boardgame that is highly rated, and that people will want to own.
    - Look at features of existing highly rated board games to determine how to design the new game.

- **Project Goals**
  - Ratings and Ownership Relationship
    - Use Linear Regression to relate highly rated games with the number of copies owned
  - Variable Selection
    - Narrow down the numerous board game features to those that are correlated with highly rated board games
  - Ratings Prediction
    - Use Logistic Regression to train a classification model that predicts if a board game will be highly rated based on its features

Georgia Tech.

Data Collection Update

- **Data Sources**
  - Preapproved data from Board Game Geek
    - Combines game manufacturer data with player community data
  - Information Contained in Data Sources
    - Reviews
    - Number Owned, Wanted, or Wished
    - Categories , Themes and Mechanics of Games
    - Artists and Publishers of Games
  - Data Relevance
    - Predict which features of a board game will lead to higher ownership
    - Guide the design of a board game which tend to have higher user ratings

- **Data Collection Status**
  - Data on minimum sales of boardgames needed for profit

- **Data Challenges**
  - There are over 5,000 variables in the 9 data sources
 - Variable selection methods are being used to narrow down variables

Georgia Tech.

Exploratory Data Analysis

Initial Insights:

- Relationship between Numowned and Bayes Average Rating
    - Positive Correlation between NumOwned and Bayes Average Rating (See Heat Map)
    - Relationship becomes more linear for Bayes Average Ratings at 6 and above (See Graph 1)
    - Conclusions: Designing a boardgame to obtain a higher Bayes Average Rating leads to more owned

- Average number of variables
    - Partition the data into Train, Validate and Test Data Sets
    - Eliminated Variables which consumer decides on
        - EX: Rank.Strategy
    - Will Use Variable Selection methods to reduce the number of variables in the regression models using train data set
    - Will further reduce variables by considering the average number of categorical variables to create more models and test models using validation data set (See Table 1)

Data Quality:

- All Data Sets are Related by BGGID, the boardgame ID
- To perform Natual Log transform all numerical data values received a +1 to avoid ln(0)
- Created Categorical Variable if entry had an omitted data
    - EDA was performed on non-omitted data, followed by analysis on data with omissions.
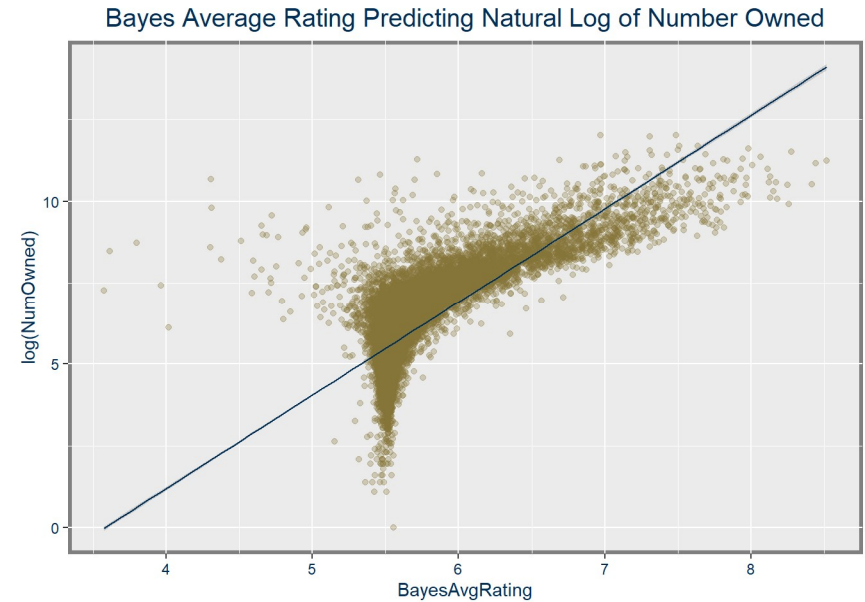
Graph 1
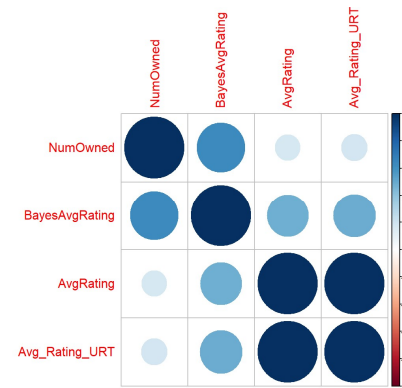
Heat Map

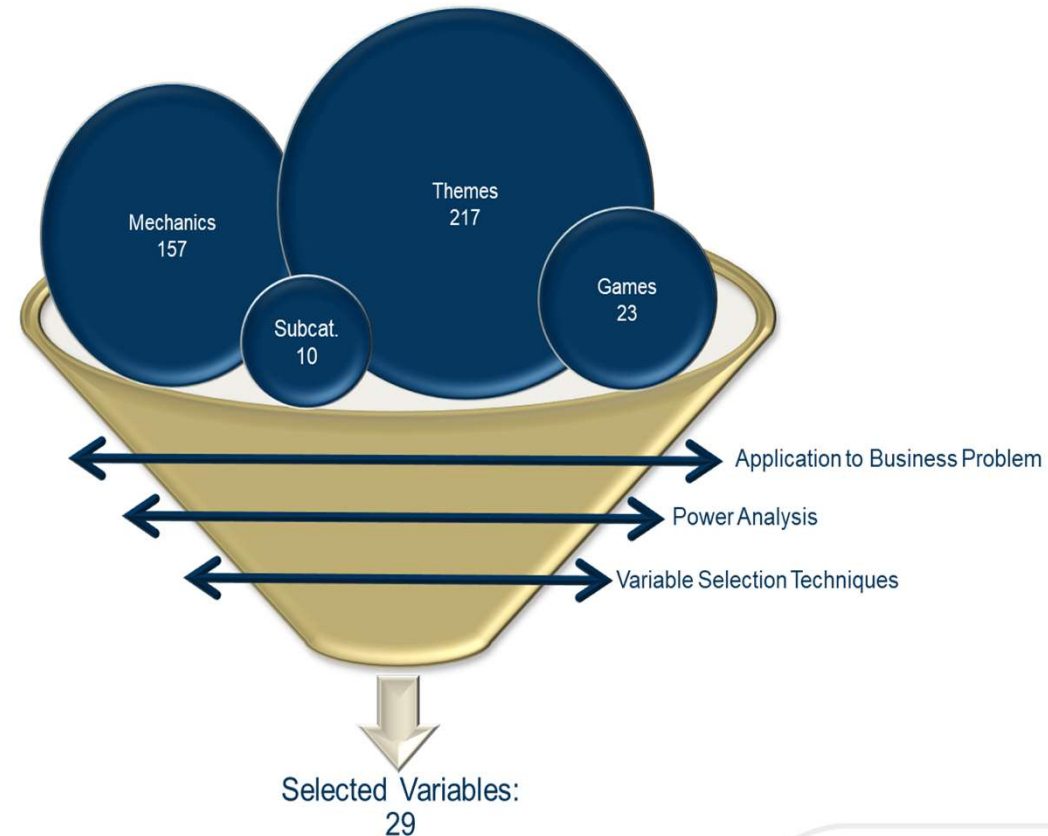| Category	Averages
Artists	1.1935234
Designers	1.2197491
Mechanics	3.1051311
Publishers	2.5819384
Subcategories	0.5386545
Themes	1.4768073
Categories	1.0000000

Table 1

Georgia Tech

Analytical Approach and Methodology

- **Techniques Applied**
  - Linear Regression
    - **Rationale:** find relationship between rating and number owned
  - Logistic Regression
    - **Rationale:** predict games that will have high ratings based on features
  - Power Analysis
    - **Rationale:** remove features with sample sizes too small for reliable hypothesis testing
  - Variable Selection
    - Forward step-wise, Elastic Net, Lasso
    - **Rationale:** further limit number of variables

- **Preliminary Results**
  - There is a linear relationship between rating and log of number owned
  - Power analysis greatly reduced the number of variables
  - Preliminary linear model obtained with only 29 variables using forward step-wise regression

Mechanics
157

Themes
217

Games
23

Subcat.
10

Application to Business Problem

Power Analysis

Variable Selection Techniques

Selected Variables:
29

Georgia Tech.

Progress Towards Objectives

Milestone Achieved	Goal Alignment
Completed exploratory analysis	All 3 project goals
Decided on 6.5 as rating threshold for logistic regression	Ratings Prediction
Completed variable selection for linear model	Variable Selection Ratings and Ownership Relationship
Used power analysis to reduce number of variables	Variable Selection
Turned in deliverables: Project Proposal Progress update	All 3 project goals

Project Goals

- Ratings and Ownership Relationship
- Variable Selection
- Ratings Prediction

Adjustments Made

- Lowered rating threshold for logistic regression from 7 to 6.5
- Changed from anova to t-test in power analysis
- Enhanced variable selection based on average number of categories, sub-categories, mechanic, & themes in games

Georgia Tech.

Plan for Completion

Next Steps

Final Variable Selection	Final Linear Model	Final Logistic Model	Final Presentation

Key

Key	
Andrew	(tan)
Talisa	(navy)
Garrett	(gray)
Team	(light gray)

Task	Nov 10 - Nov 16	Nov 17 - Nov 23	Nov 24 - Nov 30	Dec 1 - Dec 7	8-Dec
EDA					
Potential Threshold	▓ (Andrew)				
Adress Challenges	▓ (Team)	▓ (Team)			
Linear Regression Model	▓ (Garrett)				
Log Model Threshold	▓ (Andrew)				
Variable Selection	▓ (Talisa)	▓ (Talisa)	▓ (Talisa)		
Choose Rating Thresholds	▓ (Team)	▓ (Team)			
Mid Semester Update	▓ (Team)	**Due Nov 25**			
Final Variable Selection	▓ (Talisa)	▓ (Talisa)			
Final Log Model Threshold		▓ (Andrew)	▓ (Andrew)		
Final Linear Model		▓ (Garrett)	▓ (Garrett)		
Work on Slides for Final Pres		▓ (Team)	▓ (Team)	▓ (Team)	
Final Pres Submission					▓ (Garrett)

Georgia Tech.

Expected Challenges and Mitigation

Potential Obstacles

1. Selecting binary variables to use
2. Choosing a probability threshold cut off for highly rated games
3. Selecting Artist, Publishers, and Designers

Mitigation Strategies

1. Fit logistic regression models against multiple combinations of variables and choose the one with the best performance
2. Test the logistic regression model with different probability thresholds and choose the threshold with the highest performance
3. Pick artist, publishers, and designers that have worked on highly rated games with similar attributes to designed boardgame chosen by logistic regression model

GT Georgia Tech.

Questions and Feedback

Please email us any feedback you have. We would love to hear from you.

Open Questions

- What is the final combination of variables that will be used with the logistic regression?
- What publishers, designers and artists will we recommend?

Discussion Points

- What does testing probability thresholds add to our presentation? Is testing different combinations of variables enough to address the business problem?
- Should we use the validation data to choose multiple logistic models and choose one using the test data? Or should we just pick 1 model using the validation data?

Email Us Feedback
Andrew Kulp akulp3@gatech.edu
Talisa Manker tmanker3@gatech.edu
Garrett Zimmerman gzimmerman8@gatech.edu

GT Georgia Tech.

Summary and Thank You

Thank you for helping us to build a better board game.

- We are trying to build a boardgame with the best chances of being rated at 6.5/10 or above.

- We have made progress on selecting variables and creating our linear and logistic models.

- Our next steps are to finalize our models and use them to design a board game consumers will want to own.

Acknowledgements

- Scott Kuhn
  - Great guidance at our initial meeting
  - Good suggestions in proposal feedback

- Dr. Fan
  - Great project tips during office hours

- All the wonderful TAs
  - Answering project questions on Piazza

Email Us Feedback
Andrew Kulp akulp3@gatech.edu
Talisa Manker tmanker3@gatech.edu
Garrett Zimmerman gzimmerman8@gatech.edu

Georgia Tech