Predicting Board Game Success

Dan Feldman - 9/4/2018

Goals

- *Identify trends to gain insight into the current board game market.*
- *Build a predictor for board game success.*

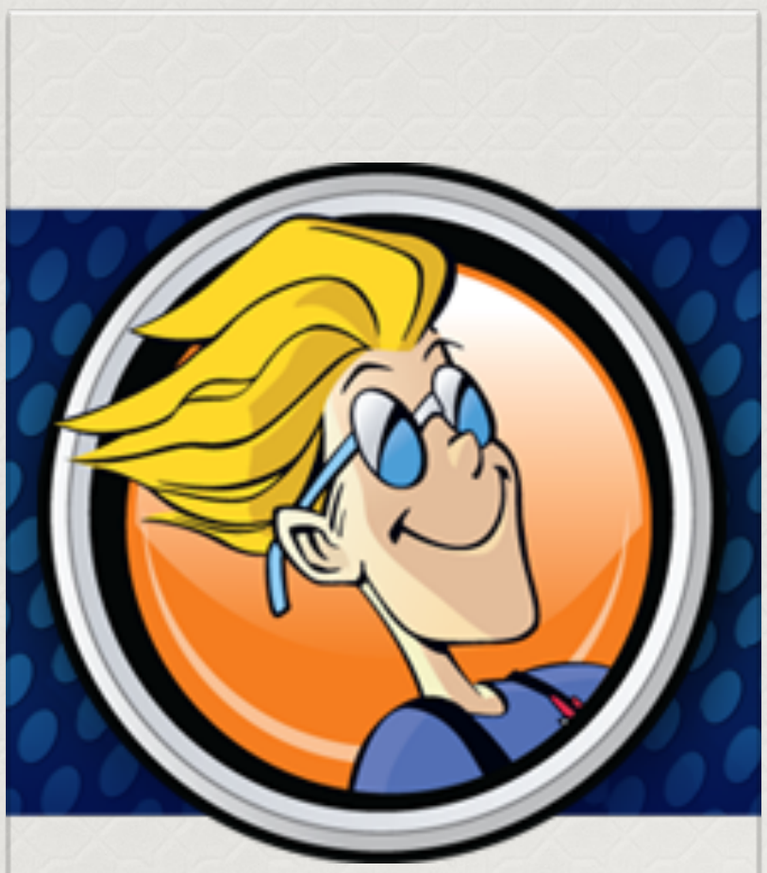
Potential Clients

Table Toys shop in Indonesia (https://www.tokopedia.com/tabletoys)

- *Board game creators: how can they create better games with more effective market strategy?*
- *Board game sellers/stores: which games should they sell and emphasize to their buyers?*

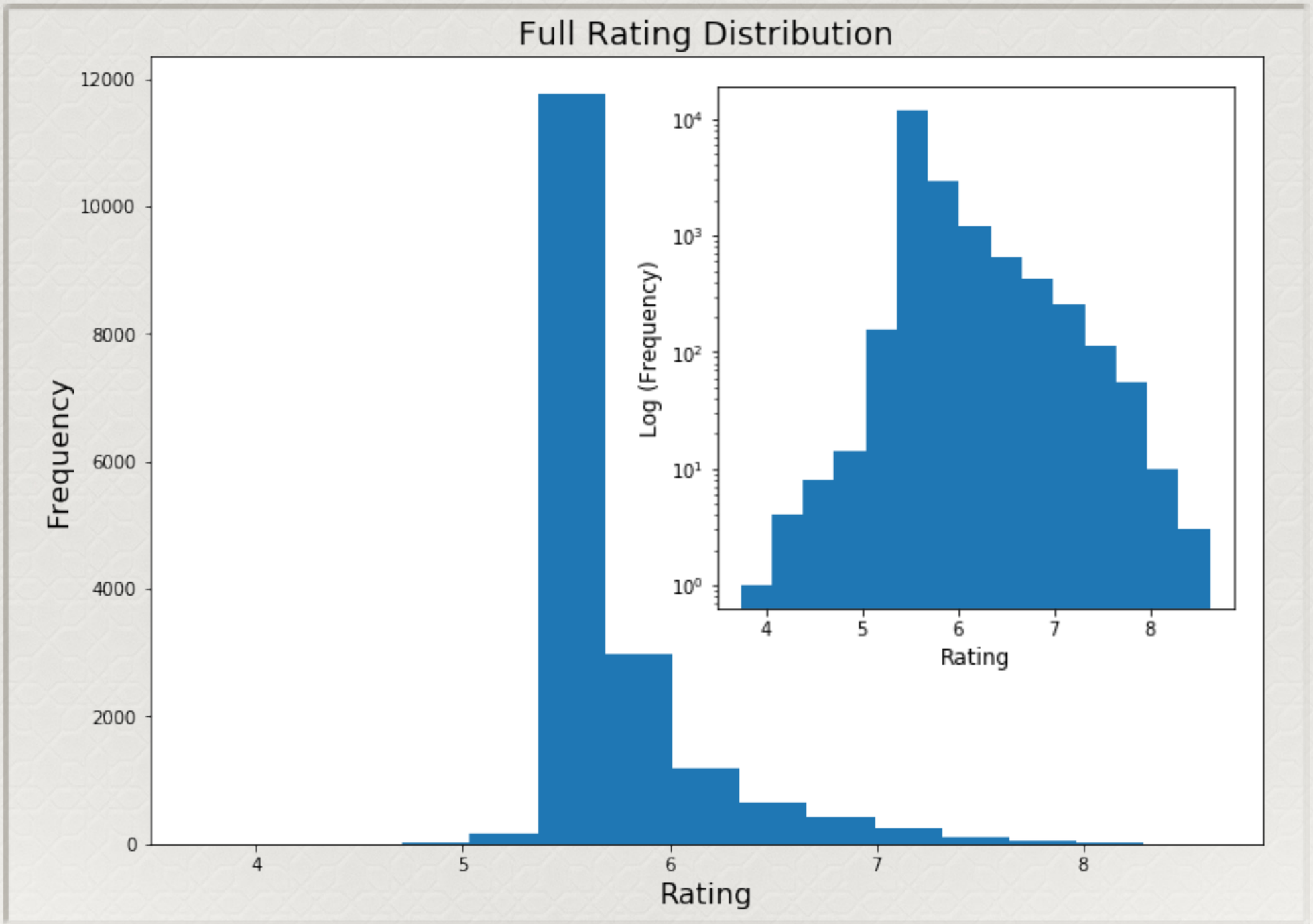
The Dataset: Board Game Geek

- *boardgamegeek.com (BGG)*
- *17,625 total board games*
- *Likely selection bias against casual gamers*
- *Data wrangled from XML API*

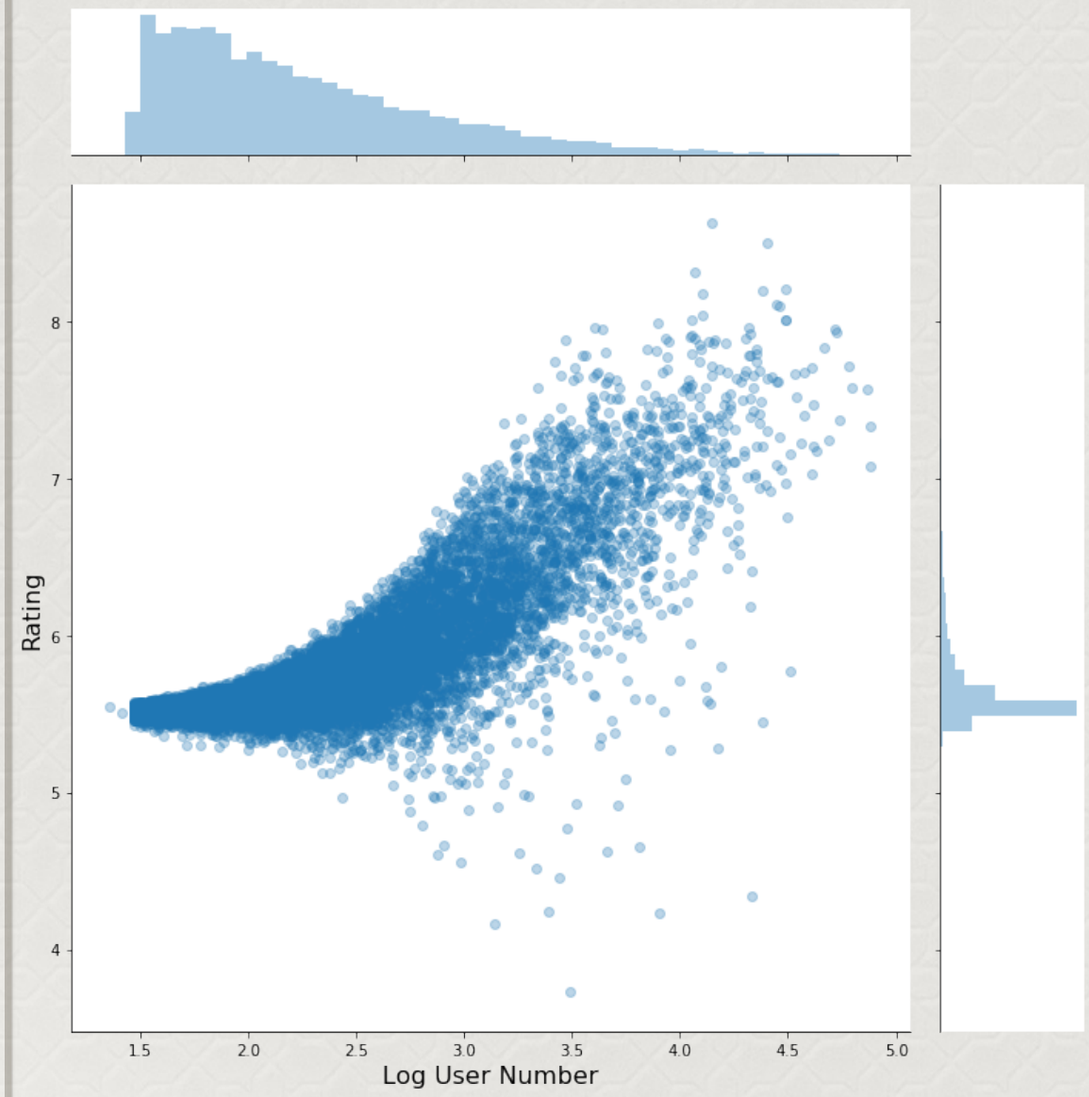


BGG Logo

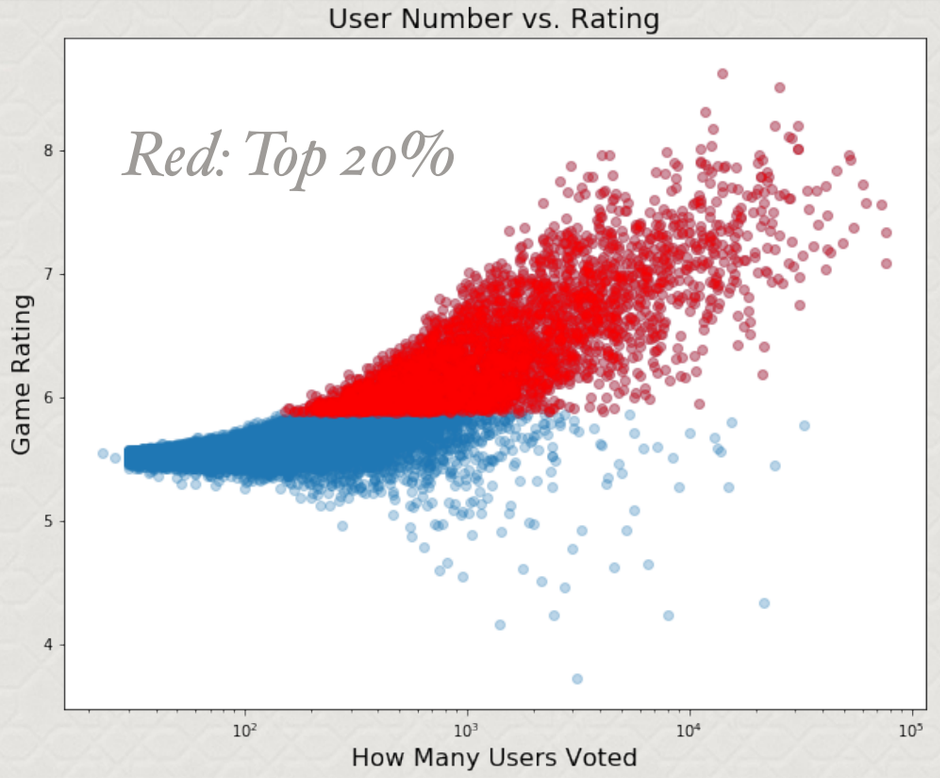
The Target Variable: Ratings

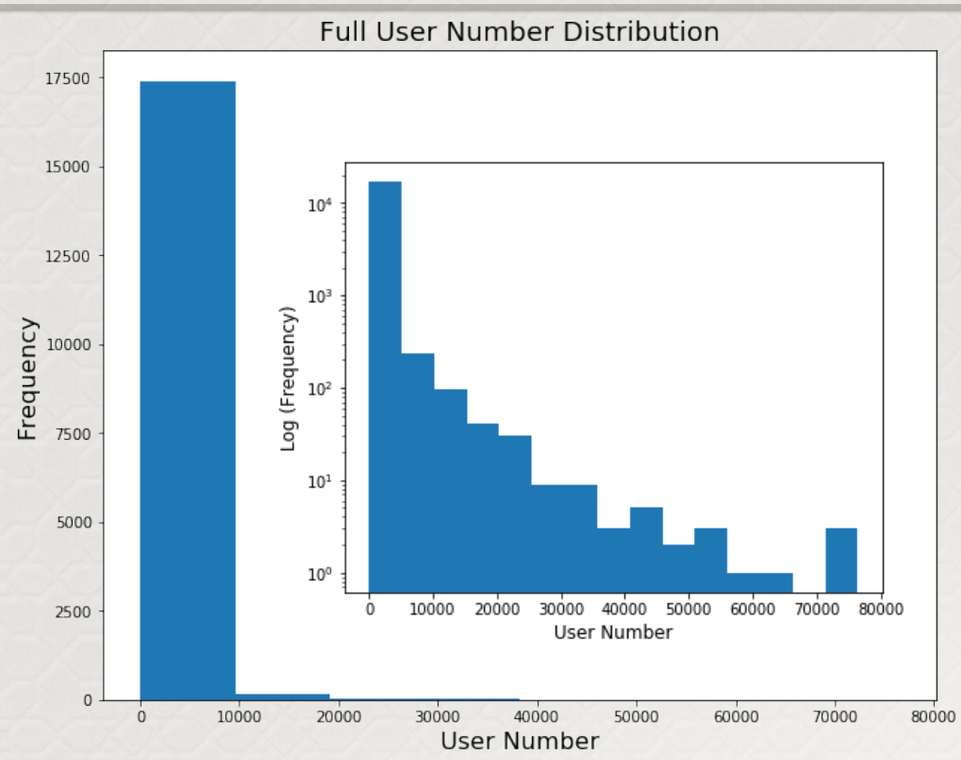


User Number vs. Rating



Highly peaked but strongly correlated

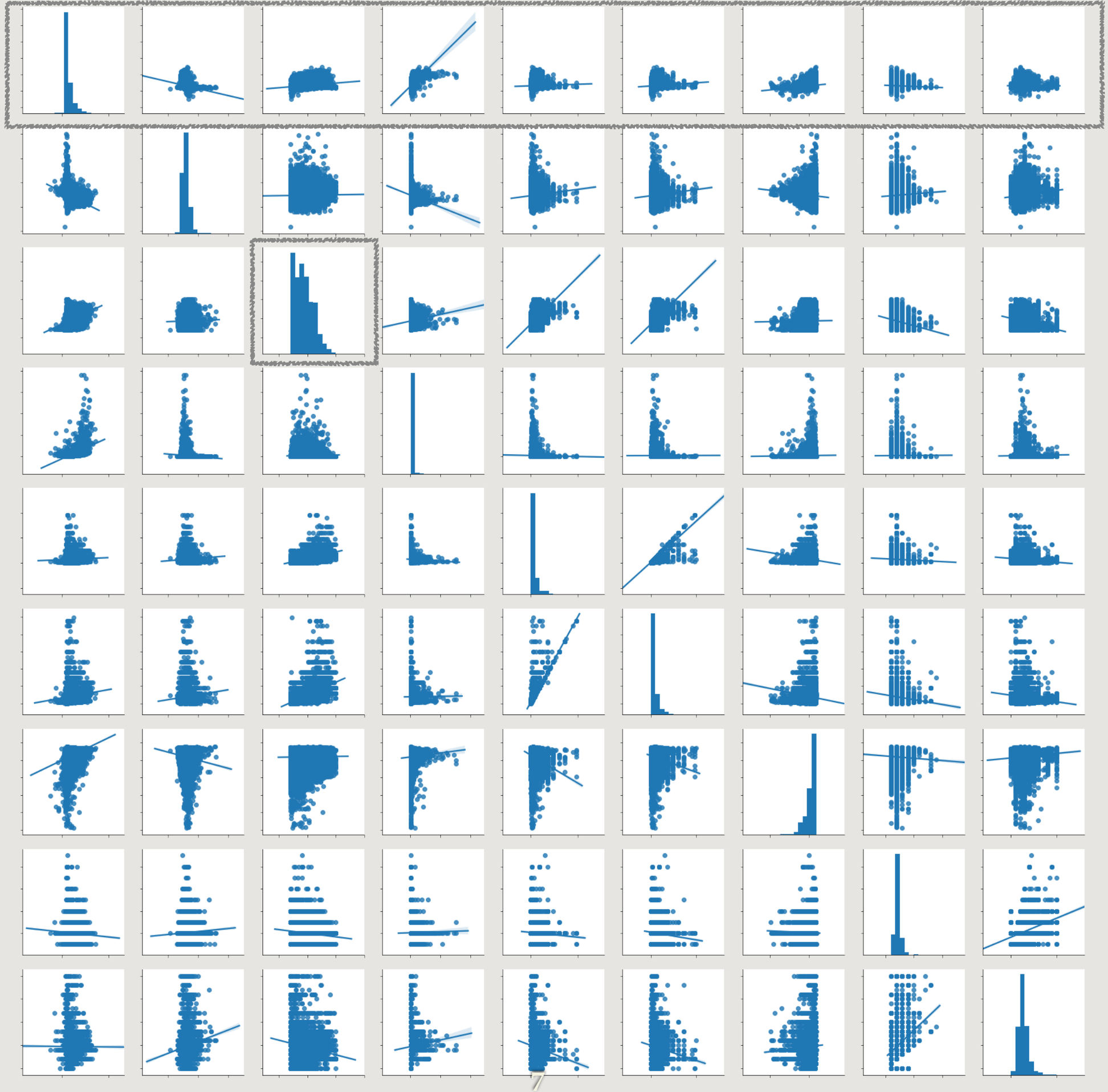




Scatter plots at a glance

Rating

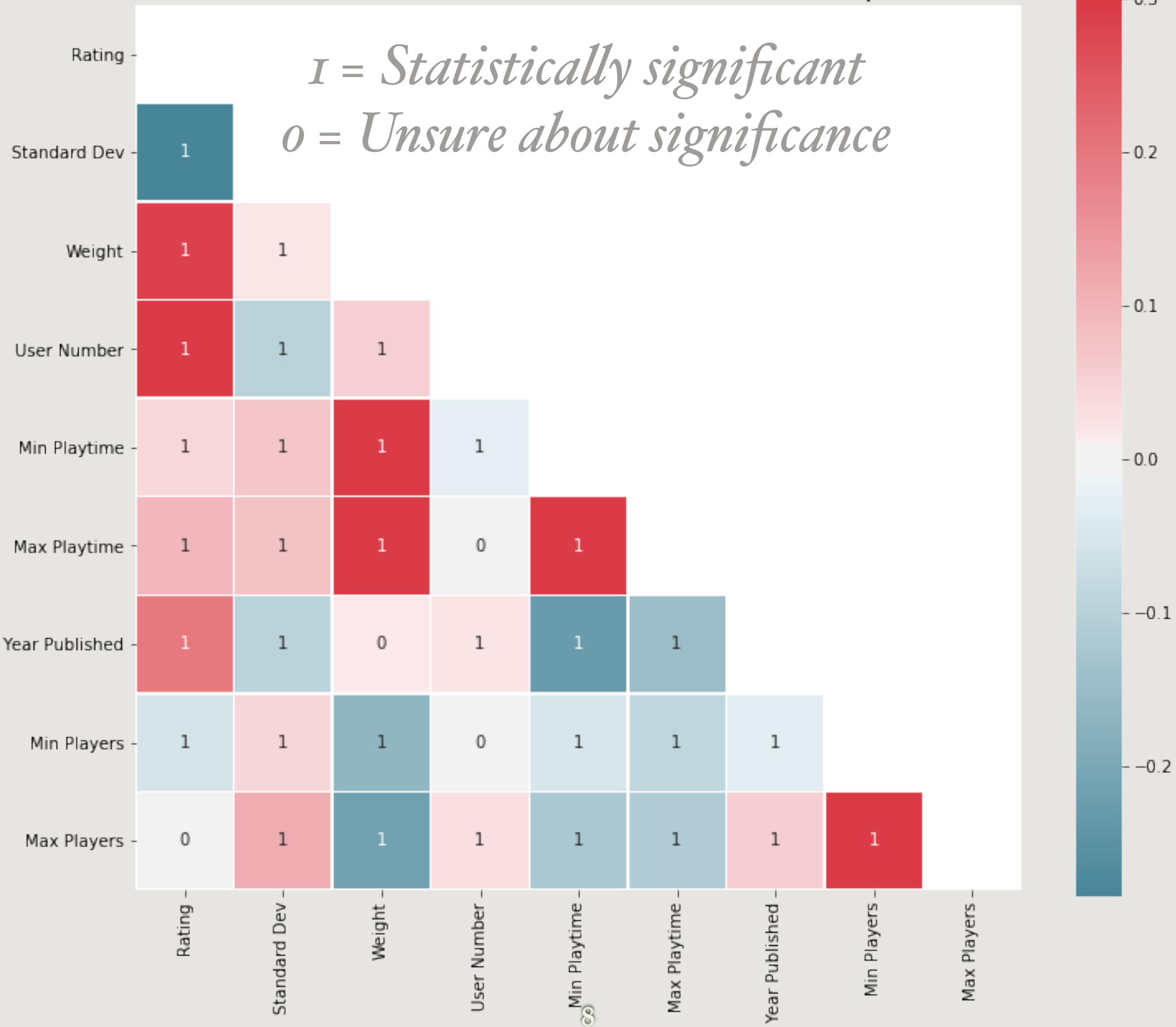
Weight



Pearson Correlation Coefficient Heatmap

1 = Statistically significant

0 = Unsure about significance



Modeling: Binary Classification

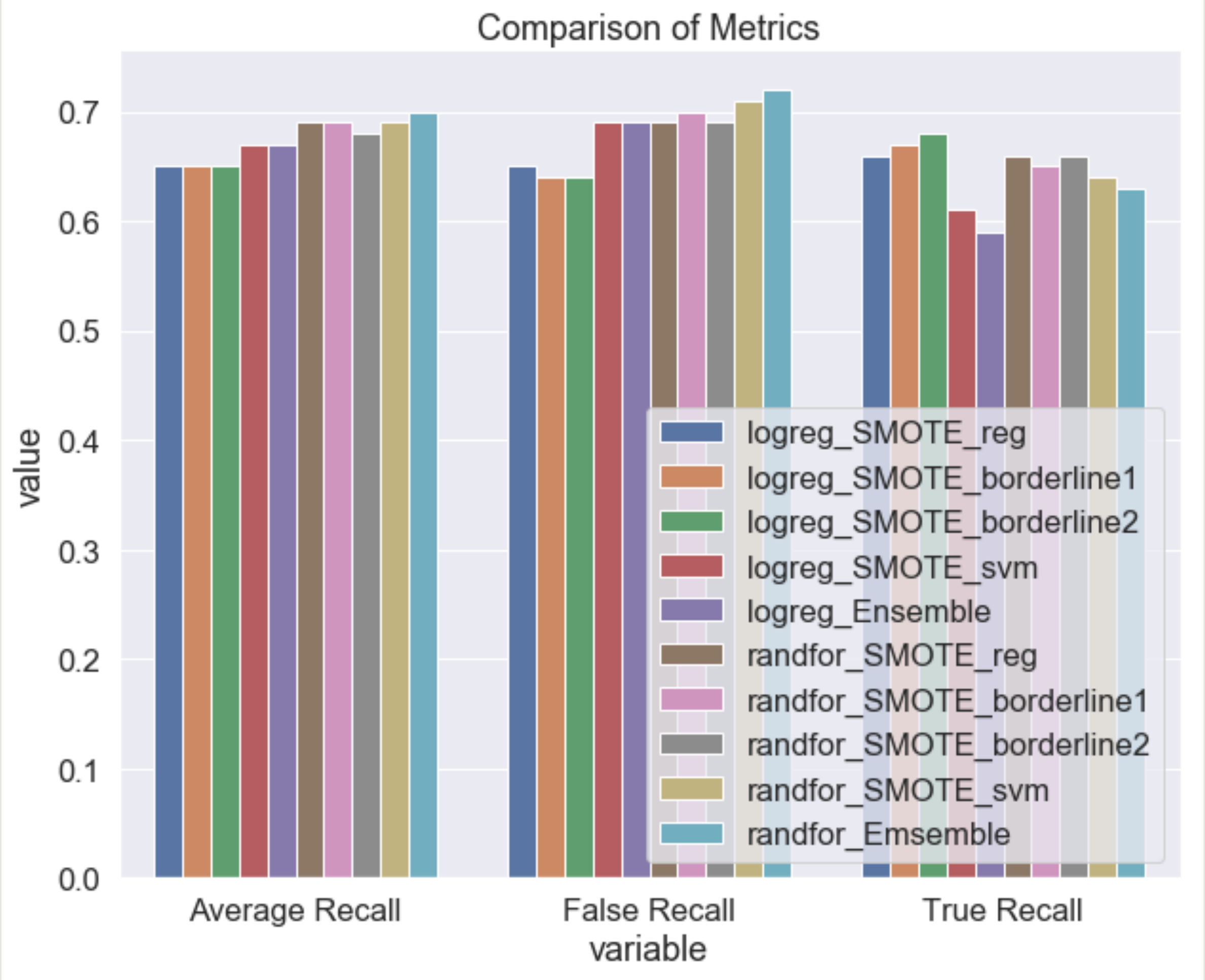
- *To make prediction easier, changed from continuous target variable (ratings) to binary classification*
- *If true: successful game, if false: unsuccessful game*
- *Chose a cutoff of the 80th percentile (rating ~ 5.88 or above considered successful)*
- *Began with Logistic Regression classifier*

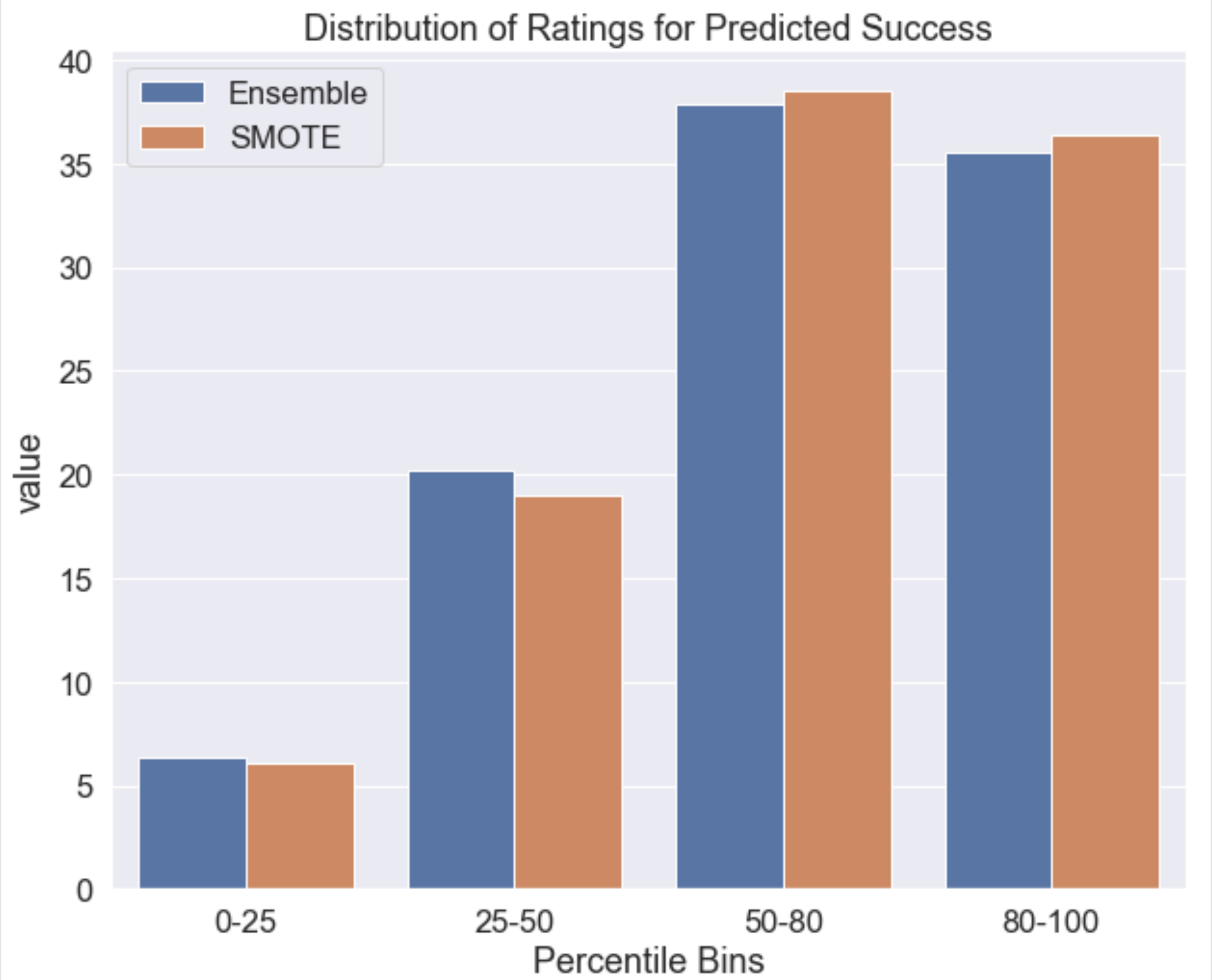
Initial Modeling: Imbalanced Dataset

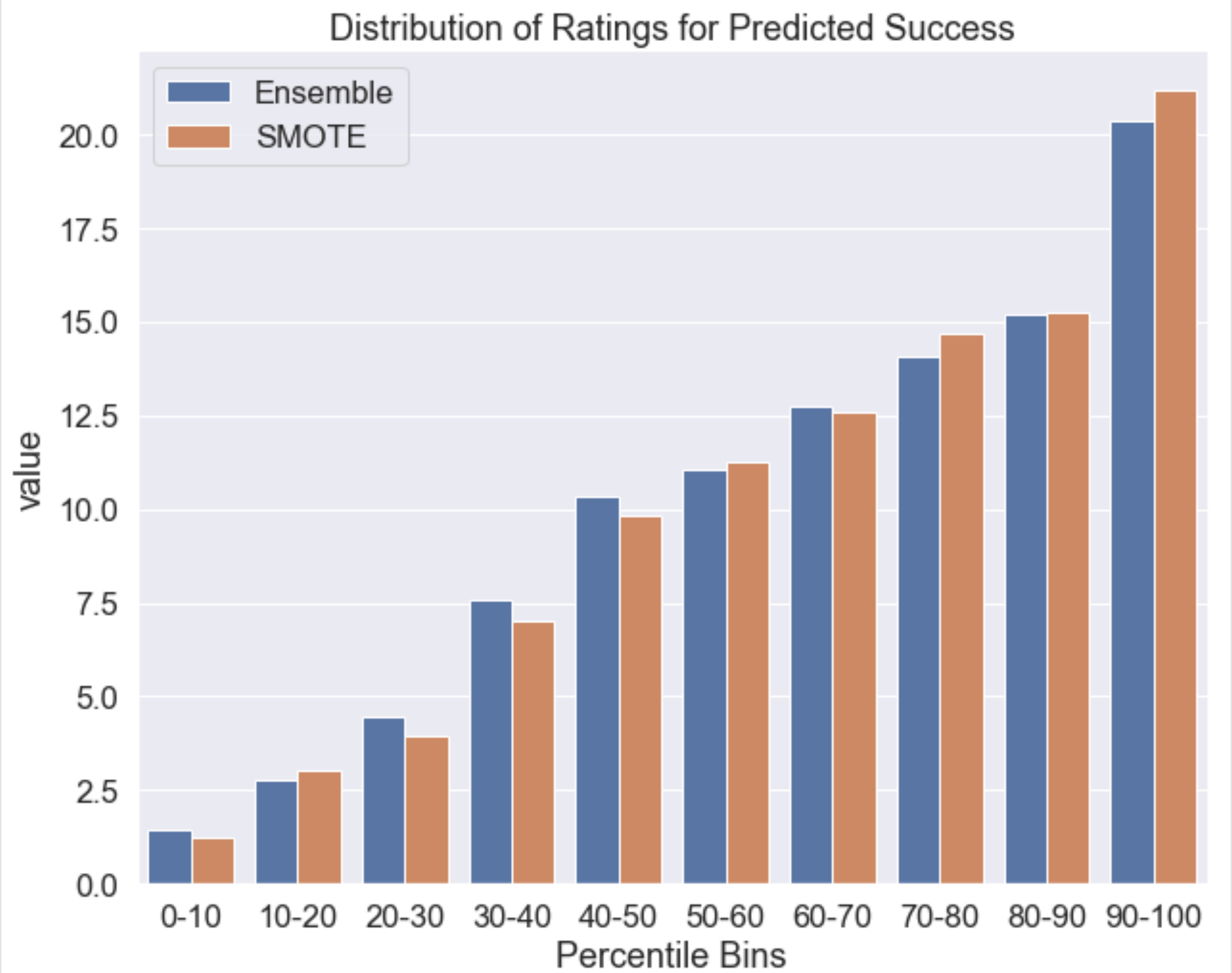
[Test Classification Report:]

	precision	recall	f1-score	support
False	0.80	1.00	0.89	4238
True	0.35	0.01	0.02	1050
avg / total	0.71	0.80	0.72	5288

- *Because the top 20% of games are successful by definition, the ratio of false to true is 4:1*
- *Classifier maximizes accuracy, classifying all games as unsuccessful (false)*
- *Fix Imbalance using over-sampling with SMOTE, and under-sampling with random ensemble*


Comparison of Metrics
value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Average Recall
False Recall
True Recall
variable
logreg_SMOTE_reg
logreg_SMOTE_borderline1
logreg_SMOTE_borderline2
logreg_SMOTE_svm
logreg_Ensemble
randfor_SMOTE_reg
randfor_SMOTE_borderline1
randfor_SMOTE_borderline2
randfor_SMOTE_svm
randfor_Emsemble


Distribution of Ratings for Predicted Success
Ensemble
SMOTE
40
35
30
25
20
15
10
5
0
value
0-25
25-50
50-80
80-100
Percentile Bins


Distribution of Ratings for Predicted Success
Ensemble
SMOTE
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
value
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
Percentile Bins

Findings/Results Summary

- *Ratings are correlated with board game complexity, number of users voting (popularity proxy?), year published; anti-correlated with standard deviation (controversial proxy?)*
- *Random forest performed better than logistic regression, logistic regression coefficients reinforce EDA correlations*
- *Over-sampling with SMOTE and under-sampling ensemble techniques roughly same effectiveness, SMOTE slightly better for both computation time and profitability/risk aversion*
- *Modeling can predict strong board game success approximately 69% of the time*