



Predicting MLB Game Attendance

Phillip Richardson



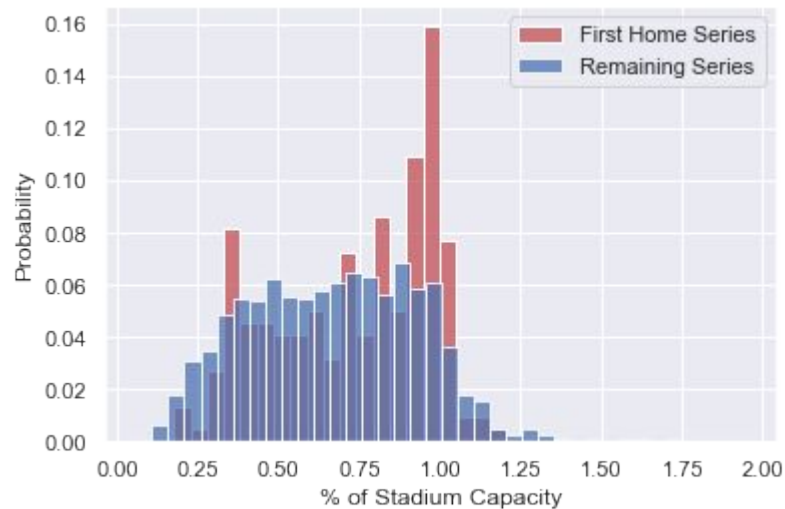
**Goal: To predict % of stadium capacity
for MLB games**

Why?

- Increased ability to be proactive in improving ticket sales
- More informed decisions around ticket sales such as concession stocking.

Data

- 2019 MLB regular season
 - Baseball Reference
 - Beautifulsoup
- Weather in city corresponding to the stadium
 - Weather Underground
 - Beautifulsoup + Selenium
- All teams' home games
- Filtered out first home series





Initial Model Features: Primarily Performance Metrics

Night game (1/0)	Weekend game (1/0)
Max temp	Precipitation (1/0)
Opponents W/L	W/L %
5 game rolling avg Runs Scored	5 game rolling avg Runs Allowed
Win/Loss Streak	Opponent in Division (1/0)
Prior game was rescheduled (1/0)	Prior game was at home and extra innings (1/0)
Games back from division leader	

Removed

Night game (1/0)	Weekend game (1/0)
Max temp	Precipitation (1/0)
Opponents W/L	W/L %
5 game rolling avg Runs Scored	5 game rolling avg Runs Allowed
Win/Loss Streak	Opponent in Division (1/0)
Prior game was rescheduled (1/0)	Prior game was at home and extra innings (1/0)
Games back from division leader	



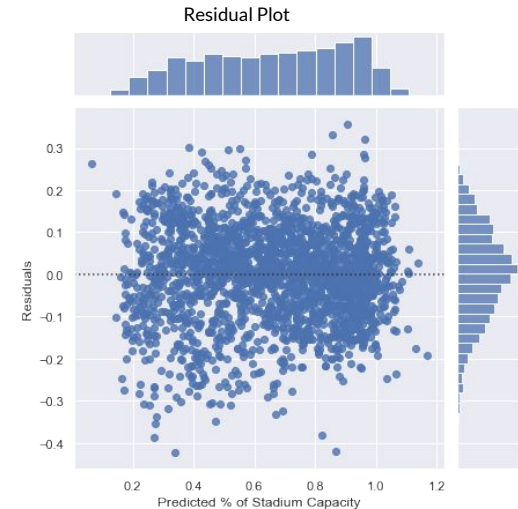
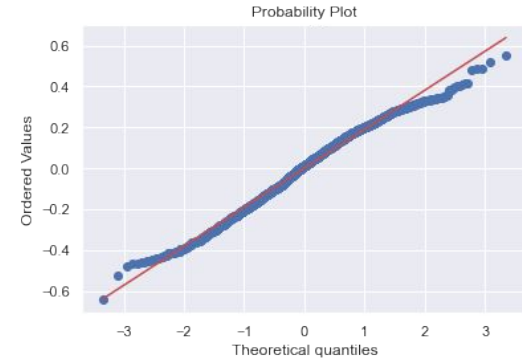
Model Selection: Lasso Polynomial Regression

- Multiple iterations of models
 - adding /removing features
- 5-fold cross validation for each model
- Lasso and polynomial models had features standardized
- Attempted transforming % of stadium capacity

Model Type	Avg R^2	R^2 STD
Base	0.309	0.044
Interactions	0.314	0.043
Lasso (poly)	0.350	0.031

Model Validation

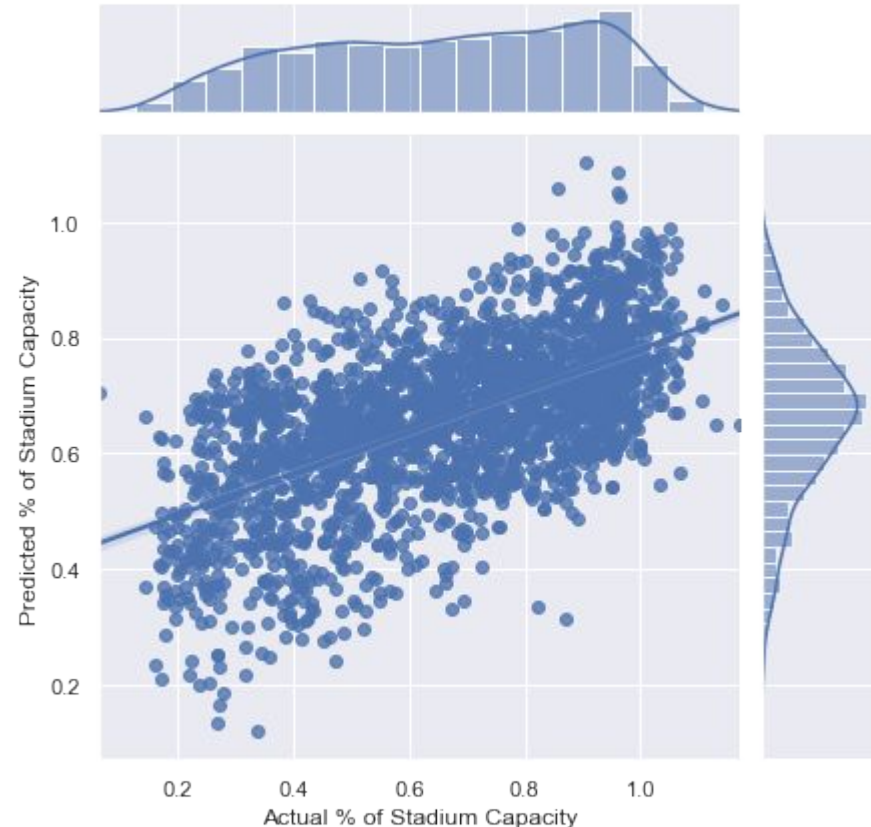
- Mean Absolute Error : 0.158
 - Test MAE: 0.158
 - Training MAE: 0.158
- Model mostly abides by the 5 main assumptions
 - QQ Plot shows a light left skew
 - Errors appear uncorrelated and relatively constant variance



Model Results

- On average, this model has a difference between the actuals and predicted of 15.8 percentage points.
 - Useful context for interpreting results
- Room for improvement, incorporating social features

Predicted vs. Actuals



Model Applications

- **Setting % of capacity thresholds to address low attendance**
 - **Implement targeted social media ads for teams predicted to have low attendance**
 - **Implementing ticket discounts or other promotions such as local business partnerships**
- **Create better revenue forecasts by anticipating concession purchases from predicted attendance**