Predicting Attendance by Game

Joshua Amunrud
Oakland Athletics Practicum

December 1, 2016

1 Introduction

1.1 Goal

Develop a model to predict Oakland Athletics home game attendance. The model will be most useful if it only includes features that would be available at the beginning of a season. The ultimate goal is to predict attendance for future seasons in order to determine demand for each game. This information could then be used to help determine which games should be classified as white, green, and gold for the Fielder's Choice plan.

1.2 Methodology

Over the course of this project, I'll be exploring several features in order to reduce the mean-squared error of the difference between predicted attendance and actual attendance. I will be using cross-validation with data from 2009 to 2016 in order to develop an optimal model.

2 Model Development

This section will be added to as new features are added, tweaked, or subtracted in order to improve performance. Unless otherwise noted, the metric used for model performance is root-mean-square error (RMSE). The units for this measurement are tickets sold.

2.1 First model

The initial features developed for this model were:

- Day of the Week
- Day or Night game
- Month
- Holiday

Note: March and October were treated as April and September respectively.

Performance

Using Linear Regression, this basic model has an RMSE of 7,215.

2.2 Promotion Categories

Promotions were added as a feature to the model. As there are many possible promotions, they were broken up into a few categories:

- Fireworks
- Bobble Heads

- Jersey
- Opening Day
- Exhibition
- Others

Performance

Adding promotional categories improved the model to an RMSE of 6,413. The R^2 of the new model is 0.181, meaning 18.1% of the variation in tickets sold can be explained by these three factors. The remaining 81.9% is due to factors not yet considered, or inherent variability.

2.3 Opponent and Last Season Wins

The next iteration of the model included:

- MLB opponent as a categorical variable
- Opponent's wins last season
- A's wins last season
- Whether or not the game took place on a holiday

Performance

This resulted in a big improvement of performance with an RMSE of 4,507. The R^2 is 0.594, meaning almost 60% of the variation is explained by factors included in the model.

3 Next Steps

The next action steps in the development of the model are:

- **Additional Features:** Currently, the next feature to develop is average road attendance per opponent. This may be used in addition to or instead of using the opponent as a categorical variable

- **Other Algorithms:** So far, linear regression is the only algorithm used, but other could be tested to see if there is a bump in performance
- **Verify Assumptions:** After engineering features and landing on an algorithm, the assumptions will need to be verified.