Modeling MLB Attendance

By: Ethan Liu and Ryan Oscherwitz

Popularity of baseball, specifically the MLB, has been decreasing for a significant amount of time. Jinuk Park and Sanghyun Park are the authors of an article in which they model the attendance of Korean Baseball Organization (KBO) games using different types of models. They found that their artificial neural network model more accurately predicted the attendance of KBO games than their multiple linear regression model. We wish to extend their research to the MLB by exploring what factors fans consider when choosing to attend a MLB game. The methods we will implement include decision trees, random forests, gradient boosting, and artificial neural networks. We have data containing the attendance, time, weather, promotions, and a number of other facts about each MLB game from 2015 until 2018.

We started by modeling per game attendance by using decision trees. Doing so yielded an MAE of 5382.89 and an RMSE of 6784.99. From the resultant variable importance plot, we found that promotions were not amongst the more important variables in predicting attendance. Despite attempting to prune and optimize our decision tree, we were not able to develop a very productive model and instead, shifted gears towards random forests, a modeling technique that combines the output of a number of decision trees in order to give a final output. Doing this yielded our best results, an MAE of 3813.25 and an RMSE of 5087.54, though with variable importance values that were quite close to the ones found from our singular tree.

Gradient boosting builds a sequence of “shallow” decision trees where each successive tree learns and improves from the previous models. Each new tree corrects the error from the previous collection of models. We choose to compare three different types of gradient boosting: Generalized Boosting (GBM), XGBoost, and LightGBM. Starting with GBM, we first consider a model that uses all of our numerical predictors. From there, we select only the most important predictors from that first model to give us the best model using GBM. The best model has a RMSE of 5,747. Moving on to XGBoost, our XGBoost model yielded a RMSE of 5275.17 and a MAE of 3978.91. These results are slightly better than the ones given by just using GBM, though still not as good as those given by the random forests. In order to be able to take our categorical variables into account, we also created a model using LightGBM. However, this model yielded a RMSE of 7701.76 and a MAE of 6533, the worst results yet.

Artificial neural networks (ANNs) are another machine learning algorithm. ANNs efficiently recognize patterns in raw data. Their structure is based on the neurons in the brain. There are input nodes that are connected to hidden layers which are then connected to output nodes. One downside of ANNs is that the results can be difficult to interpret, as the hidden layers between the input and output nodes do their calculations behind the scenes. ANN models require the data to be scaled. As such, only numeric data can be used when tuning the model. We choose to use all of our numeric predictors when fitting this model. We find the RMSE to be 7269.247.

Multivariate adaptive regression splines can be used to model nonlinear relationships, which is what we seem to have here. Using MARS, we end up with an R^2 value of 0.655 and and RMSE of 5893.5. These values aren’t as good as the values achieved with other models, but the ability to see the weights of each variable does give insight as to their relationship with per game attendance.

Despite how long we spent getting promotion data, we found that promotions were not influential in predicting attendance. We did find that attendance was higher with stadiums that were very new or very old, and in stadiums that were more heavily invested in. Interleague and weekend games seemed to increase attendance, while intradivision games did not. Our best model was the random forests, as gradient boosting did not markedly improve our model. In fact, LightGBM produced the worst predictions of any of our models.