Daniel Brilliant

May 23, 2023

DATA604

**Using Machine Learning to Project Individual OPS for MLB Players- Executive Summary**

In Major League Baseball (MLB), advanced data analytics and predictive statistics were once considered to be irrelevant to building rosters and developing players. However, methods first began developing in the 1970s when the Society for American Baseball Research (SABR) formed the Statistical Analysis Committee under Dick Cramer, Bill James, and Pete Palmer (Mizels et al., 2022). As this committee developed and understanding of baseball statistics grew, several new statistical metrics for pitching, hitting, and defense were developed over time (Mizels et al., 2022). These were further expanded when advanced Statcast tracking of motion for pitching and hitting was introduced (Mizels et al., 2022). Additionally, AI and Machine Learning models have begun being developed for injury prediction based on risk in specific anatomical locations (Mizels et al., 2022). The growth of these statistics and prediction models has led to the proliferation of the “sabermetric revolution” (named for SABR and the metrics they developed) within baseball front offices.

The “sabermetric revolution” really began with the book *Moneyball* by Michael Lewis, which told the story of how the 2002 Oakland Athletics cut spending but built an efficient baseball team through gained edges in statistical analysis and focus on statistics like on-base percentage (OBP) over a traditional statistic like batting average (Baumer & Zimbalist, 2014, pp. 7–42). The success of the book led to the massive proliferation of analytics departments in baseball front offices, to the point that over 75% of MLB teams had staff members hired in this field by 2014 (Baumer & Zimbalist, 2014, pp. 7–42). The following study attempts to follow these ideals and the methods developed by sabermetric analysis in combination with AI and machine learning.

One statistic that became very popular over time in sabermetric circles is OPS. OPS combines on-base percentage (aka OBP, or how often a player reaches base) and slugging percentage (aka SLG, or how many bases per at bat a player obtains) to define a catch-all offensive statistic (Melling, 2018). A previous Machine Learning project developed a method for predicting this statistic using the previous five years’ offensive statistics and the previous season’s postseason offensive statistics, team, primary position, and awards using Lahman baseball database data from 2000-2016 (Melling, 2018). This model was a lasso regression with an RMSE of 0.170 (Melling, 2018). The following project aimed to develop a model to improve upon these results.

Using a CRISP-DM approach, the project was developed as follows. The business problem, as defined by the original model, was understood and developed based on the above information. The data was culled from the Lahman Baseball Database, which had further information available from the original model. Thus, the new model was made to predict OPS for the next season from 2001-2022 using data from 1996-2021. The data was prepared in multiple stages using a Jupyter Notebook containing Python code. First, an OPS target variable (stored in a one-dimensional NumPy array) from Lahman database statistics from 2001-2022 was produced using Pandas dataframe building and filtering. Then, the target variable dimensions were used to build the template for the feature dataframe in Pandas, where the features of interest were added. For the previous 5 years of offensive statistics, data from the previous 5 years each player played in (starting as early as 1996 for the 2001 predictions) were culled and filtered into the dataframe, with any missing values replaced by zeroes. For the previous year’s team and position, the team and position columns were one-hot encoded and any player who played for multiple teams in a season had their statistics combined into one row to ensure full-season offensive statistics were used for all analysis, while the position with the most games and innings played was chosen as the primary position. For the awards, the different awards themselves were again one-hot encoded and combined for each player in a specific year in case they had won multiple, while any missing values were replaced by 0s. Finally, for playoff offensive statistics, the playoff rounds for each year were one-hot encoded and combined row-wise for each player by year so that all of their postseason offensive statistics from that season could be accounted for, and any missing values were accounted for by replacing with zeroes.

Once the feature matrix was preprocessed and placed into a NumPy array, the data was split into training and test sets (80% in training data, 20% in test data), and the features were scaled using RobustScaler and reduced in dimensionality to include 95% of original variance using PCA. This modified data was used to test several ensemble regression models on the training data to see if they could improve upon the accuracy of the original model. The models tested included GradientBoosting, AdaBoost, RandomForest, ExtraTrees, XGB, Bagging, HistGradientBoosting, and Voting and Stacking models using the top 3 best scoring of the previous models by R2 values. The top 3 highest scores came from ExtraTrees, RandomForest, and Bagging. Hyperparameter tuning was attempted to increase the accuracy of these 3 models but was unsuccessful. Thus, the Voting and Stacking models were made using the default parameters for the top 3. In the end, the two best models by R2 were ExtraTrees and Voting, and the choice between the two was made by evaluating the root-mean squared error (RMSE), as the original project used RMSE as their model selection score. ExtraTrees had the lowest RMSE, so it was chosen as the final model.

A picture containing screenshot, sketch, diagram, plot

Description automatically generatedA black and white image of a city

Description automatically generated with low confidence

Figure 1a: Feature data boxplots showing distribution prior to scaling, Figure 1b: Feature data boxplots showing distributions following scaling of data with RobustScaler

The model evaluation was then performed on the test dataset, which revealed that the model, unfortunately, did not have predictive validity. The R2 value of 0.206 and RMSE of 0.306 were values that were not sufficient for accurate prediction. As seen in the following visuals, this disparity was evidenced in scatterplots. The scatterplot for the training data predictions showed a strong linear relationship, but evidence of uncaptured exploratory information in the residuals made it clear that the model would not necessarily be the best for predicting the test dataset. The resulting scatterplot for the test data predictions showed no clear relationship whatsoever.

A picture containing screenshot, diagram, line, text

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, diagram, design

Description automatically generated

Figure 2a: Predicted values vs actual values for training data, Figure 2b: Predicted values vs residuals for training data, Figure 2c: Predicted values vs actual values for test data

Due to these results, there are clearly more tests that could be performed to determine if a valid predictive OPS model can be produced from the data kept in the Lahman baseball database. Other scaling methods, other regression models, or more historic data could be examined to see if accuracy is improved or decreased. Additionally, further filtering that only includes players who had 5 previous years of experience could be examined to see if accuracy improves. Regardless, the results showed that modifications must be made in order to achieve a better predictive regression model for OPS.

References

Baumer, B., & Zimbalist, A. (2014). *The Sabermetric Revolution Assessing the Growth of Analytics in Baseball* (1st ed., pp. 7–42). University of Pennsylvania Press.

Melling, M. (2018, February 28). *OPS Projections with Machine Learning – Baseball Data Science*. Baseball Data Science. https://www.baseballdatascience.com/ops-projections-with-machine-learning/

Mizels, J., Erickson, B., & Chalmers, P. (2022). Current State of Data and Analytics Research in Baseball. *Current Reviews in Musculoskeletal Medicine*, *15*(4). https://doi.org/10.1007/s12178-022-09763-6