Predicting On-Base Percentage in Major League Baseball: A Regression Analysis Approach

Ben Klassen

September 8, 2024

Abstract

This report presents a comprehensive analysis of predicting on-base percentage (OBP) for Major League Baseball (MLB) players using various regression models. The analysis covers data preprocessing, feature engineering, model training, evaluation, and hyperparameter tuning to identify the most effective models. Ridge Regression emerged as the top-performing model. Future work may involve more advanced models and additional features to enhance predictive capabilities.

1 Introduction

The objective of this analysis is to predict the OBP of players for the 2021 MLB season. The dataset, `obp.csv`, contains historical player statistics spanning from 2016 to 2021. OBP is a key metric in baseball, reflecting a player's ability to reach base and is widely used in player evaluations, game strategies, and contract negotiations. By leveraging machine learning techniques, this analysis aims to identify key indicators that influence OBP and develop models that provide high prediction accuracy.

2 Analysis

2.1 Understanding the Data

The dataset contains player statistics from 2016 to 2021, including columns for plate appearances (PA), OBP, and birth date. Preliminary exploration revealed many missing values, particularly in the historical PA and OBP columns. These missing values were significant as they could potentially bias the model if not properly handled.

Name	Number of Missing Values
Name	0
playerid	0
birth_date	0
PA_21	0
OBP_21	0
PA_20	106
OBP_20	106
PA_19	135
OBP_19	135
PA_18	213
OBP_18	213
PA_17	274
OBP_17	274
PA_16	325
OBP_16	325

Table 1: Summary of Missing Values

2.2 Data Visualization

A scatter plot of PA versus OBP revealed no discernible relationship between these variables. For players with fewer than 50 PAs, there is significant variation in OBP, likely due to the small sample size.

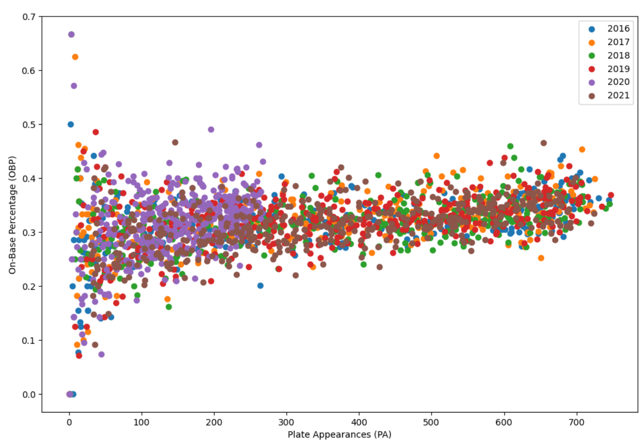

Figure 1: Scatter Plot of PA vs. OBP

2.3 Data Preprocessing

Data entries with insufficient PAs (less than 100 combined across the years 2016-2020) were removed. Missing values were handled by imputing the median, which, compared to the mean, reduces the impact of outliers. Additionally, birth dates were converted to player age in 2021 to provide a more interpretable variable for regression modeling.

2.4 Feature Analysis and Engineering

Important features were identified based on their correlations with OBP in 2021. OBP from recent years (e.g. OBP_20, OBP_19) showed stronger correlations, which is a logical result as this is a more accurate reflection of their current skill level.

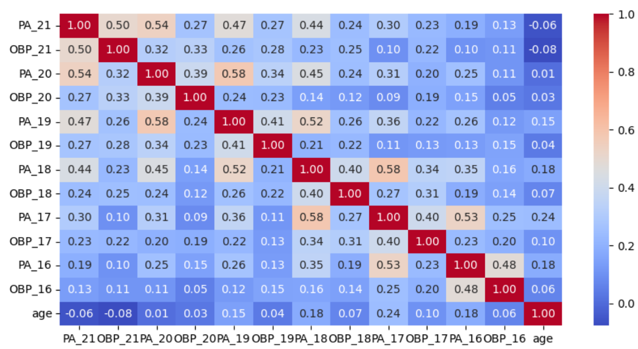

Figure 2: Correlation Heatmap of Features

Three new features were created to reflect an enhanced weighting on OBP from more recent years as well as trends in OBP year to year. A positive or negative trend in a player's performance could indicate a change in trajectory in skill that is not captured in static OBP values.

```
# Create a weighted OBP feature from other columns
# giving more weight to OBP from recent years
df['OBP_weighted'] = (df['OBP_16'] * 0.1 + df['OBP_17'] * 0.15
                      + df['OBP_18'] * 0.2 + df['OBP_19'] * 0.25
                      + df['OBP_20'] * 0.3)

# Create OBP trend features from other columns to
# account for how the OBP is moving year-year
df['OBP_trend_1920'] = df['OBP_20'] - df['OBP_19']
df['OBP_trend_1819'] = df['OBP_19'] - df['OBP_18']
```

2.5 Model Training

The dataset was split into training and testing subsets, with 70% of the data used for training and 30% of the data used for testing. Since OBP is a continuous numeric variable, regression is an appropriate modeling technique. Several regression models were implemented and trained, including Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, Decision Tree Regression, Gradient Boosting, AdaBoost, and ElasticNet Regression.

2.6 Model Evaluation

The models were evaluated on the test data, and their performances were assessed using the Mean Squared Error (MSE) and the Coefficient of Determination (R-squared). Ridge Regression emerged as the best model, exhibiting the lowest MSE and highest R-squared values. The performance of other models, such as Decision Tree Regression, could have been hindered by overfitting or weak generalization to unseen data.

Model	**MSE**	**R-squared**
Linear Regression	0.0018	0.1411
Ridge Regression	0.0018	0.1556
Lasso Regression	0.0019	0.0716
Random Forest Regression	0.0020	0.0678
Decision Tree Regression	0.0037	-0.7820
Gradient Boosting	0.0019	0.0937
AdaBoost	0.0020	0.0289
ElasticNet Regression	0.0019	0.1063

Table 2: Model Performance Metrics

2.7 Feature Selection and Hyperparameter Tuning

To improve the Ridge Regression model, two strategies were employed: feature selection and hyperparameter tuning. Feature selection was performed using recursive feature elimination based on feature importance derived from the coefficients of the Ridge Regression model. It was discovered that PA, particularly from earlier years (e.g. PA_16 and PA_17), had a small impact on OBP in 2021.

Feature	**Importance**
OBP_20	0.086292
OBP_weighted	0.057276
OBP_19	0.057047
OBP_18	0.046819
OBP_17	0.043048
OBP_trend_1920	0.029245
OBP_16	0.013060
OBP_trend_1819	0.010228
age	0.001556
PA_20	0.000109
PA_18	0.000034
PA_17	0.000025
PA_19	0.000011
PA_16	0.000001

Table 3: Feature Importance

Hyperparameter tuning was performed using grid search cross-validation to determine the optimal value for alpha, which controls the amount of regularization in ridge regression models (essentially the amount of underfitting or overfitting). An optimal alpha value of 0.39 was determined.

The combination of feature selection and hyperparameter tuning yielded a slight increase in R-squared and no improvement in MSE. This could suggest that the model was already close to optimal with the default parameters or that the data provided little room for further improvement.

Ridge Regression Model	**MSE**	**R-squared**
Before improvements	0.0018	0.1556
After improvements	0.0018	0.1568

Table 4: Model Performance Before and After Improvements

2.8 Limitations and Future Work

Given the simplicity of the dataset, current models may not generalize well to new players, especially those with significant gaps in their historical data. Additionally, using regression techniques assumes certain relationships between variables that might not capture all the complexities of OBP prediction, such as team strategy and ballpark factors. Future work could involve exploring more advanced models and additional features that provide a more granular view of a player's hitting quality and consistency, such as exit velocity and launch angle.

3 Conclusion

This analysis identified several models for predicting OBP, with Ridge Regression emerging as the best-performing model. Although feature selection and hyperparameter tuning only marginally improved model performance, further refinement could enhance predictive capabilities.

Appendix

A Predictions using Best Performing Models

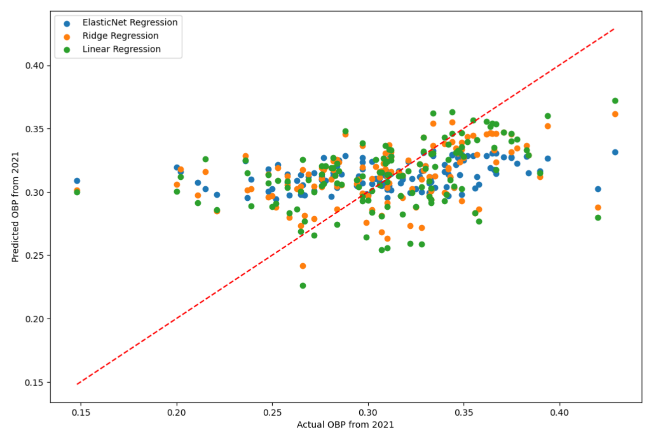

Figure 3: Predicted vs. Actual OBP in 2021

B Players with Largest Prediction Errors

Analysis of the optimized Ridge Regression model identified the players with the largest prediction errors where the predicted OBP was farthest from the actual OBP. These discrepancies could be due to unique playing styles or situational factors not captured by the model.

Player	**Actual**	**Predicted**	**Error**
Albert Almora Jr.	0.148	0.292	0.144
Chris Owings	0.420	0.284	0.136
Evan White	0.202	0.317	0.115
Todd Frazier	0.200	0.306	0.106
Andrew Knapp	0.215	0.321	0.106

Table 5: Top 5 Players with the Largest Prediction Errors