Abe Woycke

Final Report:
Korean Baseball Organization (KBO) Projections

Problem Statement

I set out to project the full-season performance of a baseball player based on early partial-season data. The original hypothetical use case was a marketing firm attempting to select an athlete spokesperson. They could want to select someone early in the season with limited information, while making sure the player they selected remains high-performing deep into the season.

Other potential applications of this forecasting are sports gambling, informing a team's internal player valuation, or adding information to an assessment of the future financial value of a KBO franchise.

Using my ridge regression models projections of players' overall performance based off of a small sample of early season data (~25 games), taken June 10th, 2020, I formulated a top-ten list of players recommended to be spokespeople. All ten recommendations remained in the top 28 for overall performance as of September 10th, 2020 (the league year is still in progress at the time of this writing), and the top three overall performers are all included in the model's Top Ten list.

Data Wrangling

KBO historical statistics were not as readily available as I had hoped, so I had to iteratively scrape them from 313 individual sub-pages of the site Baseball Reference (BR). The eventual resultant dataset consisted of 28 features and 8,025 rows. As with almost all real-world data, a lot of the columns had messy formatting issues, so there were lots of little cleaning steps required. I also used IterativeImputer to fill in missing values for Stolen Bases, Caught Stealing, and International Walks, all of which weren't recorded in our dataset before 2001.

Our data also included a lot of features that are simple counting stats that accrue over the season. These stats are higher for players based on the number of opportunities they have. So, to make them more informative, I converted many of the dataset's counting stat features to rates per plate appearance. I also dropped some features that, although informative, are really just transformations of other features in the dataset.

Abe Woycke

Exploratory Data Analysis

The project's goal was to project a player's full-season performance based off of their in-season performance to date. Subsequently, I selected the old-school triple crown (HR, BA, RBI) as three different response variables to measure performance. I measured overall performance by the average ordinal rankings of players in these three categories. The main things I looked for in the EDA were strong feature correlations with the response variables. Some of those initial findings are shown below.

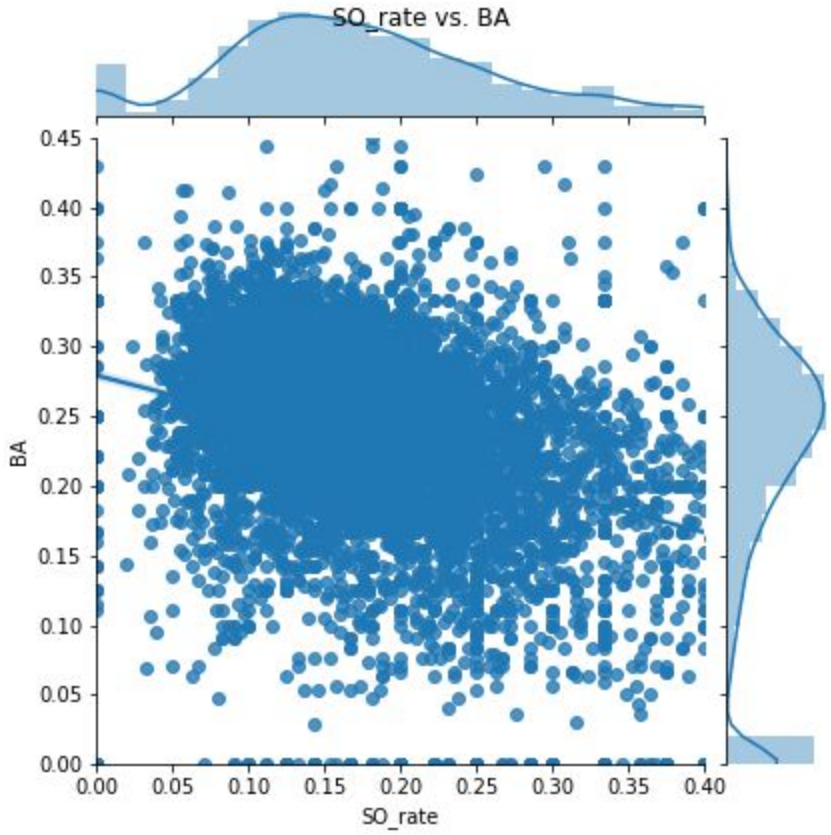

Figure 1: a joint plot of Strike Out Rate vs. Batting Average

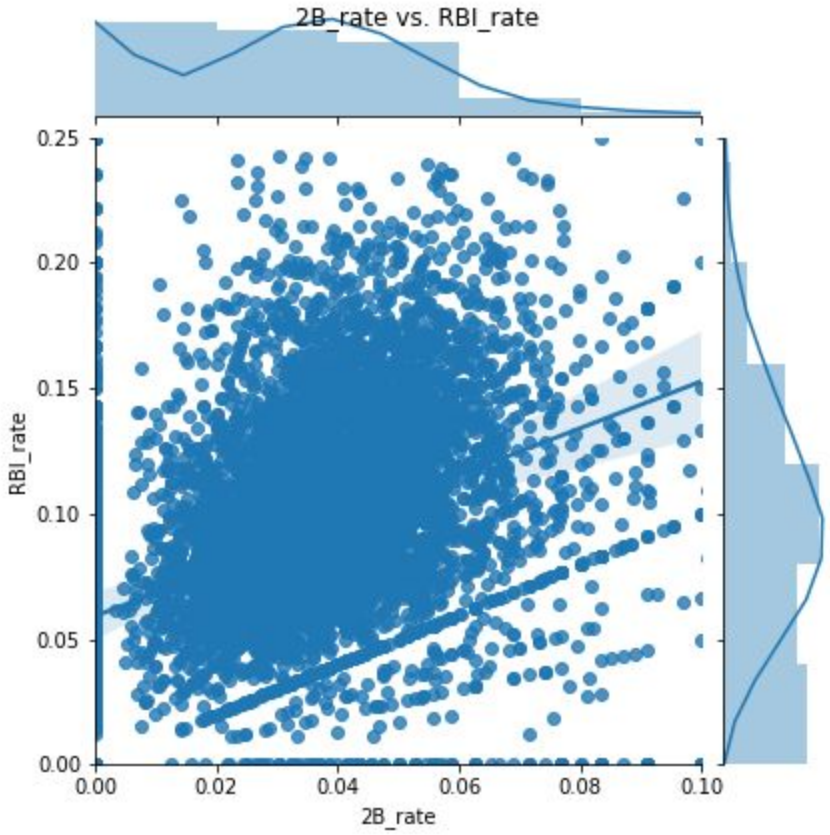

Figure 2: a joint plot of Double Rate vs. Runs Batted In Rate

Both charts show data relationships you would expect from baseball, but it was informative to visualize the strength of these relationships.

Preprocessing

I made a few more minor transformations to my data in this step, including dropping the categorical variable Batting Stance, since a large portion of the dataset did not have a value assigned, as visible in the chart below. The only other thing to do as part of Preprocessing was to execute a Train Test Split. In this case, I split off the partial current season (2020) as a holdout set for testing our recommendations, but also split the larger remaining dataset into train and test subsets. I used StandardScaler() to transform the data as well.

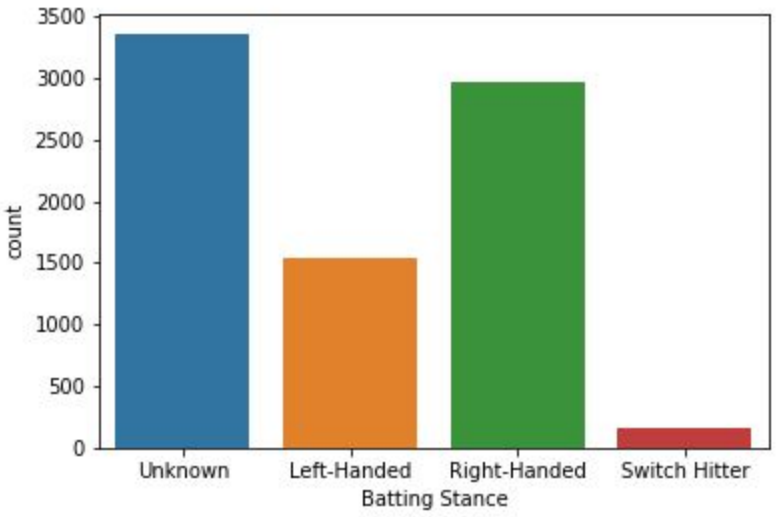

Figure 3: Distribution of Batting Stances for Entire Dataset

Model Selection

I originally tried six regression models from scikit-learn out-of-the-box (untuned), for each of the three response variables. They were Lasso, ElasticNet, Logistic Regression, Ridge Regression, SVR(kernel-'linear'), and SVR(kernel-'linear'). I used RMSE as my leading indicator for model performance. For all three response variables, Ridge Regression performed the best. It had a low RMSE and high r2 for both HR_rate and BA. And it had better scores than the others for RBI_rate (which appears comparatively harder to predict than our other two variables).

Having selected our model type for all three response variables, I proceeded to tune the Ridge Regressor. Because Ridge is not intensely computationally demanding, I used a fairly wide GridSearchCV in lieu of Bayesian hyperparameter tuning or RandomSearchCV.

I also plotted the RMSE by different alpha values (the primary tuning parameter for the RidgeRegressor), to visualize the subsequent changes based on the tuning. A couple of those charts are included below. Each of the models for each response variable is tuned slightly differently.

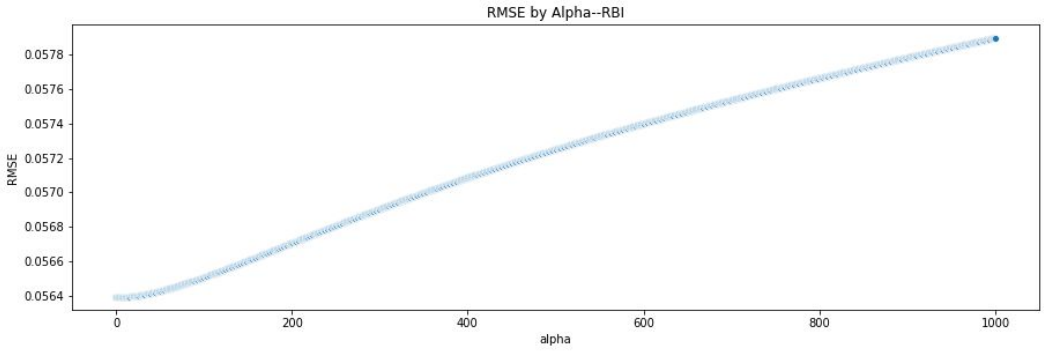

Figure 4: RMSE for Different Alphas (RBI_rate)

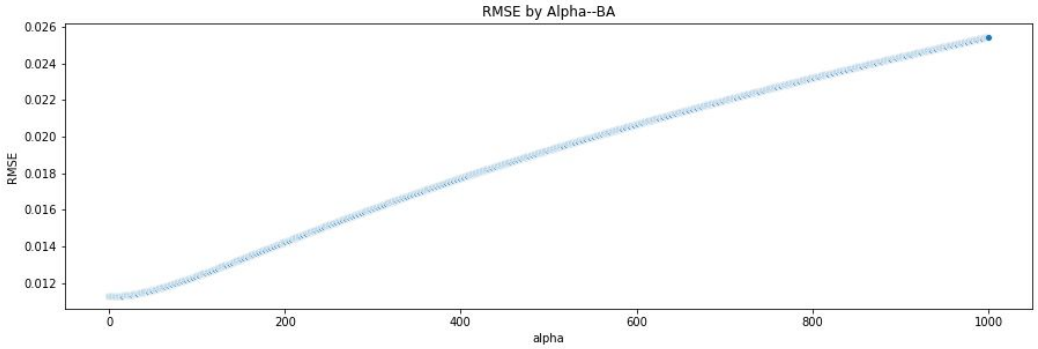

Figure 5: RMSE for Different Alphas (Batting Average)

Takeaways

The regressors performed fairly well. There were a number of players whose performance varied quite a bit from the projections, but not by any earth shattering margins. I have included visualizations of the projections vs. the actual rates below. Additionally, here are the Top Ten Recommended Players from the projection system, with their ranking three months later in parentheses.

1. Roberto Ramos (Tied 10th)
2. Mel Rojas (1)
3. Dong-won Park (28)
4. Sung-Bum Na (2)
5. Jose Miguel Fernandez (15)

6. Jung-hoo Lee (17)
7. Preston Tucker (Tied 10th)
8. Ui-ji Yang (3)
9. Dae-ho Lee (Tied 20th)
10. Weun-Sung Chae (Tied 20th)

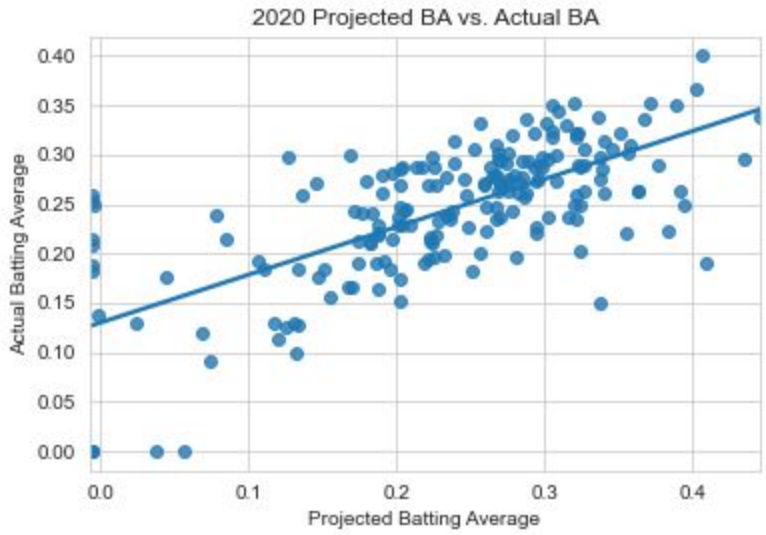

Figure 6: Model Projected 2020 Batting Average vs. Actual

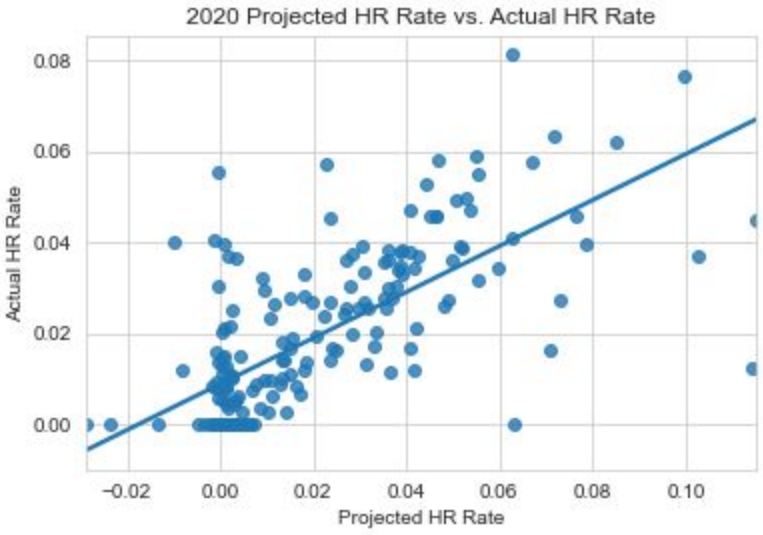

Figure 7: Model Projected 2020 Home Run Rate vs. Actual

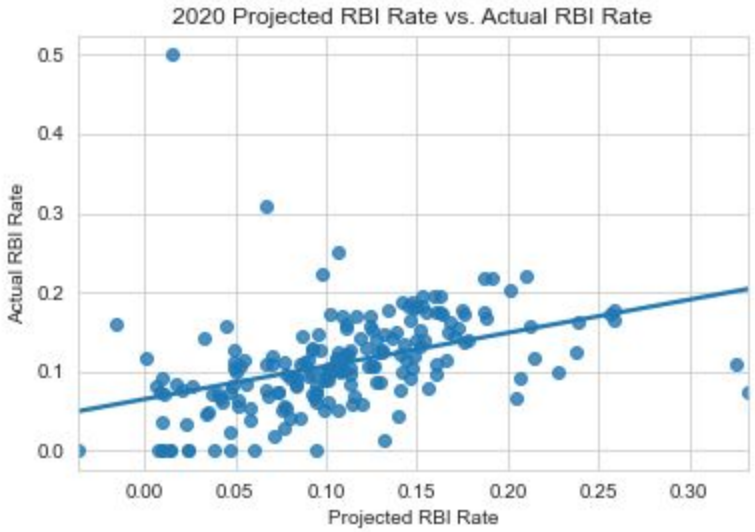

Figure 8: Model Projected 2020 RBI Rate vs. Actual

Future Research

If I were to revisit this problem or problems like it, I would seek more detailed game-by-game or at-bat data, which may exist behind a paywall somewhere, to form a more sophisticated picture of the ebbs and flows of a player's performance over the course of a full KBO season. I would also reform the dataset to include individual player performance indicators based on their performances from previous seasons.

I would be particularly interested in building a forecasting system that takes into account pitcher performance as well. A lot of the existing sabermetric analysis that is applied to Major League Baseball data just hasn't been applied to the KBO, and I'm sure there's some fun insights into the differences in styles of play that could be uncovered.