The Data

The data were collected from each of the 10 schools in the Liberty League from the 2012 season through the 2019 season. Overall, this amounted to hitting data from 698 players. The statistics gathered from the schools websites were batting average, on base plus slugging, at bats, runs, hits, total bases, slugging percentage, and on base percentage. I then created calculated columns for runs, hits, and total bases per at bat. In addition to the statistics previously mentioned, I created columns for the difference between the statistics over two years (for example, avg_trend1 would denote the difference in batting average from the first two seasons).

The data were split into three data frames to create three different models -- these were for players who played for at least two, three, and four seasons. There were 411 players who played at least two seasons, 264 who played at least three seasons, and 143 players who played four years.

The Models

For each of the three data frames, I ran a random forest model with 25 estimators to predict batting average, OPS, OB%, and SLG%. A table of the RMSE for each model is below:

	AVG	OPS	OB%	SLG%
Two Seasons	.1413	.3016	.1467	.1828
Three Seasons	.0904	.2303	.1294	.1359
Four Seasons	.0973	.2370	.1097	.1452

I anticipated that the first model would have the greatest errors, as there is only one season of data available. I assumed that the models would sequentially get better, which didn't seem to be the case, as the model with four seasons performed slightly worse than the model with three seasons. This could be because there were only 143 players in that dataset, compared to the 264 players in the dataset with three seasons.

Predicting Rochester 2020 Data

In this analysis, I wanted to predict the player statistics for the 2020 season and compare those to the predictions of Coach Reina. I used both a linear regression model and a random forest model to create the predictions and compare the performance of the two models. Below is a table of the RMSE for the models:

	AVG	OPS	OB%	SLG%
Two Seasons	.0678	.1754	.0750	.1116
Three Seasons	.0509	.0753	.0719	.0185
Four Seasons	.0783	.1629	.0695	.1085
Overall	.0715	.1556	.0775	.0972

The predictions for the random forest and the resulting errors are below. The error is calculated as Coach Reina's prediction - the model's prediction, and the percentage error is that value divided by the model's prediction.

	Name	AVG	OPS	SLG	OB
1	Matzat,Jacob	0.31492	0.77088	0.32692	0.396960
8	Sy,Harper	0.31808	0.81420	0.39992	0.373880
10	Rende,Joseph	0.25292	0.58436	0.23072	0.334440
13	McKinsey,Brian	0.26076	0.66804	0.32816	0.397720
16	Piontek,Luke	0.26796	0.80072	0.40384	0.326973
4	Pickering,Steve	0.30516	0.80872	0.39152	0.406720
7	Bankovich,Drew	0.26684	0.71080	0.30816	0.330680
1	Rieth,David	0.35084	0.94468	0.51128	0.389400
2	Hertz,Jake	0.33120	0.83368	0.42580	0.353920
4	Trombley,Kyle	0.29724	0.77680	0.41244	0.362520
5	Miraz,Zach	0.32368	0.74136	0.34172	0.451800
6	McNabb,Ryland	0.27864	0.72096	0.39184	0.322760
8	Craig,Aaron	0.31628	0.77016	0.35760	0.388840

	Name	AVG	OPS	SLG	OB
0	Matzat,Jacob	0.005 (1.61%)	-0.001 (-0.11%)	0.093 (28.47%)	-0.047 (-11.83%)
1	Sy,Harper	-0.008 (-2.54%)	0.036 (4.4%)	0.0 (0.02%)	0.076 (20.36%)
2	Rende,Joseph	0.097 (38.38%)	0.326 (55.73%)	0.219 (95.04%)	0.126 (37.54%)
3	McKinsey,Brian	0.059 (22.72%)	0.142 (21.25%)	0.082 (24.94%)	0.002 (0.57%)
4	Piontek,Luke	0.012 (4.49%)	-0.068 (-8.46%)	-0.004 (-0.95%)	0.006 (1.84%)
5	Pickering,Steve	-0.04 (-13.16%)	-0.054 (-6.64%)	-0.012 (-2.94%)	-0.032 (-7.8%)
6	Bankovich,Drew	-0.078 (-29.17%)	-0.161 (-22.62%)	0.042 (13.58%)	-0.131 (-39.52%)
7	Rieth,David	0.059 (16.86%)	0.075 (7.97%)	0.089 (17.35%)	0.031 (7.86%)
8	Hertz,Jake	0.004 (1.15%)	-0.034 (-4.04%)	-0.026 (-6.06%)	0.046 (13.02%)
9	Trombley,Kyle	-0.047 (-15.89%)	-0.202 (-25.98%)	-0.112 (-27.26%)	-0.088 (-24.14%)
10	Miraz,Zach	-0.014 (-4.23%)	0.109 (14.65%)	0.108 (31.69%)	-0.052 (-11.47%)
11	McNabb,Ryland	-0.029 (-10.28%)	-0.138 (-19.14%)	-0.142 (-36.2%)	0.01 (3.17%)
12	Craig,Aaron	-0.026 (-8.31%)	-0.06 (-7.81%)	-0.008 (-2.13%)	-0.029 (-7.42%)

The same results from the linear model are below.

	Name	AVG	OPS	SLG	OB
1	Matzat,Jacob	0.31492	0.77088	0.32692	0.396960
8	Sy,Harper	0.31808	0.81420	0.39992	0.373880
10	Rende,Joseph	0.25292	0.58436	0.23072	0.334440
13	McKinsey,Brian	0.26076	0.66804	0.32816	0.397720
16	Piontek,Luke	0.26796	0.80072	0.40384	0.326973
4	Pickering,Steve	0.30516	0.80872	0.39152	0.406720
7	Bankovich,Drew	0.26684	0.71080	0.30816	0.330680
1	Rieth,David	0.35084	0.94468	0.51128	0.389400
2	Hertz,Jake	0.33120	0.83368	0.42580	0.353920
4	Trombley,Kyle	0.29724	0.77680	0.41244	0.362520
5	Miraz,Zach	0.32368	0.74136	0.34172	0.451800
6	McNabb,Ryland	0.27864	0.72096	0.39184	0.322760
8	Craig,Aaron	0.31628	0.77016	0.35760	0.388840

	Name	AVG	OPS	SLG	OB
0	Matzat,Jacob	0.033 (11.49%)	0.053 (7.41%)	0.082 (24.31%)	-0.029 (-7.66%)
1	Sy,Harper	0.024 (8.29%)	0.121 (16.64%)	0.046 (12.99%)	0.075 (20.09%)
2	Rende,Joseph	0.094 (36.81%)	0.267 (41.56%)	0.146 (48.21%)	0.121 (35.61%)
3	McKinsey,Brian	0.086 (36.56%)	0.19 (30.75%)	0.112 (37.68%)	0.078 (24.33%)
4	Piontek,Luke	0.072 (34.35%)	0.169 (30.0%)	0.139 (53.39%)	0.03 (9.87%)
5	Pickering,Steve	-0.078 (-22.7%)	-0.081 (-9.66%)	-0.041 (-9.64%)	-0.04 (-9.68%)
6	Bankovich,Drew	-0.083 (-30.5%)	-0.148 (-21.18%)	0.008 (2.29%)	-0.156 (-43.76%)
7	Rieth,David	0.072 (21.29%)	0.093 (10.05%)	0.105 (21.25%)	-0.012 (-2.79%)
8	Hertz,Jake	0.061 (22.37%)	0.084 (11.73%)	0.017 (4.5%)	0.067 (20.05%)
9	Trombley,Kyle	0.043 (20.7%)	-0.028 (-4.69%)	0.004 (1.42%)	-0.033 (-10.57%)
10	Miraz,Zach	0.0 (0.13%)	0.074 (9.55%)	0.083 (22.45%)	-0.008 (-2.06%)
11	McNabb,Ryland	-0.125 (-33.26%)	-0.328 (-36.0%)	-0.192 (-43.47%)	-0.136 (-28.94%)
12	Craig,Aaron	0.066 (29.41%)	0.056 (8.51%)	0.06 (20.9%)	-0.005 (-1.35%)