Milestone 3

NETIDs: el3418 ac8480

Methodology: Describe the methodology that you applied.

For our final milestone, we wanted to improve upon the results from our baseline Linear Regression model, as well as compare its performance to other models. As we speculated in Milestone 2, average career VORP did turn out to be too difficult to use as a threshold due to varying career lengths (including the many players who are still active). Therefore, we normalize the career VORP to generate an average per season by dividing each player's career VORP by their number of seasons played to get seasonal VORP. For this milestone, we also wanted to evaluate the R^2 to get a better sense of how much of the variance in our target variable we were capturing.

As mentioned in Milestone 2, we chose a Random Forest algorithm to reduce the variance of our model outputs through averaging. For the purpose of outputting a classifier, we decided to transform our average seasonal VORPs into player caliber classifications of "Good":1 or "Bad":0. This was decided by a threshold value (1.72) close to the median seasonal VORP to split the labels.

To further improve model robustness, we leveraged PCA further. In Milestone 2's additional analysis, we stated that we would explore using PCA to generate multiple models. However, ultimately we decided on a different approach - we would apply dimensionality reduction on subsets of features that were highly correlated. These PCA reduced features would be representative of the various archetypes of basketball players (Ex: guard, center, wing). One possible approach to deciding which features to subset on would have been an extensive correlation analysis between each feature and examining interaction terms. However, this would have required an exponentially increasing amount of work per feature added to the dataset. A faster, more naive approach that we took was to examine the coefficients of a PCA in 3 directions performed on a dataset containing all features. If the absolute difference between a coefficient in one direction was larger than an arbitrary threshold (0.15) in either of the other two directions and the feature had not been included in any prior subsets, then that feature would be included in the set of features that we considered to be correlated in that direction. For example, the coefficients for 3PA were 0.280061, -0.045850, and 0.132491 in their respective directions. Therefore, it was included in the first subset of features to apply dimensionality reduction on. Next, we remove certain covariates iteratively such as turnover % and blocks to see if our baseline metrics would improve. We also applied PCA to more covariates we noticed in our heatmap in Figure 9. Notably, there was a high correlation among FG, FGA, FT, TOV, USG% with PTS, thus we decided to create a PCA reduced "scoring attribute" feature using these covariates.

Finally, we investigated the distribution of true test labels. Upon inspection, a large imbalance was observed in our target variable - there are hundreds of NBA players, but very few reach an all star level of play (Figure 1). Our assumption was that the model was potentially overfitting. To address this issue, we apply upsampling through the SMOTE technique (Figure 2).

In Milestone 2, we discussed looking into player clusters to run separate models. However, we did not have a large enough dataset to allow us to split on position without running into dataset size issues, which may underfit our models, so we decided not to go this route. We also chose to not use draft pick order as a dummy variable because it would introduce even more bias into our models and overfit on the training set by increasing model complexity.

Results: Describe your results in a succinct and effective way.

In Milestone 2 we only used MAE to evaluate our Linear Regression, where the 5-folds cross-validation came out to be 7.96, and the MAE on the test set was 7.53. For this milestone, we added R^2. After switching over to average seasonal VORP, our MAE was still high. However, after using our dimensionality reduced features, our MAE improved significantly to ~0.72 and our R^2 was 0.006. One observation worth pointing out was that the PCA reduced feature using scoring attributes did not improve our models substantially - perhaps because these features were hand selected outside of the method we described above, and thus had more collinearity with the other PCA reduced features. We also used the gini index to evaluate how good our results were relative to each other. A good gini coefficient (close to 0) would suggest that, even if our values were not calibrated well, the positioning of predicted VORPs relative to other player's VORPs was good. The max gini in our dataset was 0.426, the gini of our Linear Regression predictions was 0.066, and the normalized gini was 0.155.

Looking at the Random Forest model, our initial results show that the AUC was not great at 0.56 using the aforementioned threshold of 1.72. We pruned our tree to have a max depth of 2. We also tried using a new binary labeling threshold value at 2.04, which was the mean seasonal VORP, and found that our Random Forest's AUC improved slightly to 0.58.

Utilizing Ridge Regression with a regularization term of 1e-3 seemed to not improve results. As the regularization term increased, the training set errors decreased, but the test set error increased, showing that our model was overfitting. The Ridge Regression model's MAE became 0.7235 and the R^2 reduced to 0.0034.

The most significant improvements we saw were after upsampling our target value. After applying this to our dataset, both our linear regression and random forest models improved significantly. Our Linear Regression MAE improved to ~0.4639, and its R^2 went to 0.119 (Figure 3). Additionally, many of the previously insignificant features now had p-values < 0.05. The max gini in our upsampled dataset was 0.457, the Linear Regression gini was 0.164, and the normalized gini was 0.360. The Random Forest's AUC went up to 0.68 (Figure 6), and its precision was 0.1153. We also added F1-score to the evaluation because teams would most likely want to balance between avoiding draft busts and missing out on great prospects. Thus, we calculated the F1-score and obtained a value of 0.1923.

Analysis: What conclusion can you draw from your results?

The top 5 and worst 5 players according to our Linear Regression model's seasonal VORP projections (Figures 4 & 5) show that the model was correctly assigning a high VORP for Dwight Howard, but less so the other players, such as a bottom 5 seasonal VORP projection for Joakim Noah, and Tim Hardaway Jr. (who is still active as a bench player). The top players and worst players according to our Random Forest model (Figures 7 & 8) shows Damian Lillard labeled as a top player (Probability of 1: ~0.609), which is correct because he is a multiple time all-star. The worst players are the same as our Linear Regression model, and highlights the shortcomings of our models.

In general, our models suggest that higher usage players who take a lot of shots tend to get drafted higher. In conclusion, player projection from college data is a very difficult task that requires an abundance of accurate data and proper feature selection. Sports data is hard to predict on due to the high variance, small sample size, and constant covariate shift among players.

Appendix

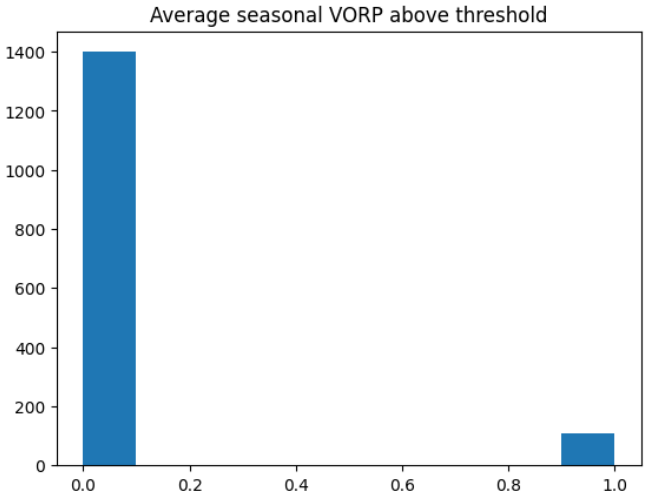

Figure 1

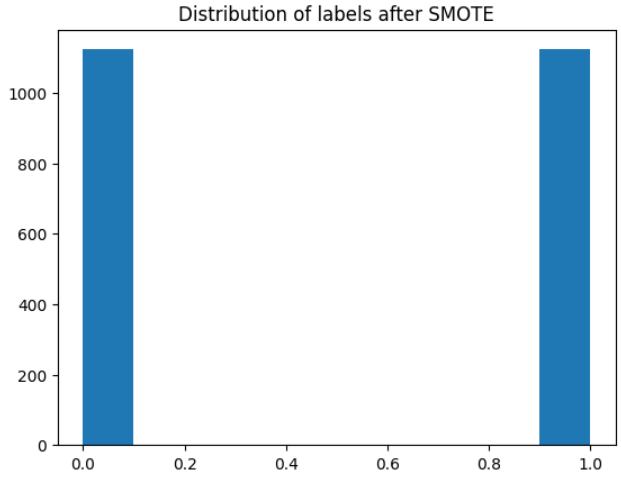

Figure 2

OLS Regression Results

Dep. Variable:	Caliber	**R-squared (uncentered):**	0.119
Model:	OLS	**Adj. R-squared (uncentered):**	0.115
Method:	Least Squares	**F-statistic:**	30.25
Date:	Sat, 06 May 2023	**Prob (F-statistic):**	3.33e-55
Time:	22:46:21	**Log-Likelihood:**	-2266.0
No. Observations:	2246	**AIC:**	4552.
Df Residuals:	2236	**BIC:**	4609.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pca_guard	-0.1069	0.019	-5.634	0.000	-0.144	-0.070
pca_big	0.0055	0.013	0.425	0.671	-0.020	0.031
pca_wing	-0.0300	0.012	-2.605	0.009	-0.053	-0.007
scoring_attr	0.0348	0.015	2.371	0.018	0.006	0.064
FG%	-0.2679	0.061	-4.426	0.000	-0.387	-0.149
2P%	-0.2022	0.039	-5.241	0.000	-0.278	-0.127
3P%	-0.0701	0.018	-3.874	0.000	-0.106	-0.035
BLK	0.0661	0.021	3.160	0.002	0.025	0.107
eFG%	0.1382	0.050	2.771	0.006	0.040	0.236
TOV%	0.0054	0.019	0.293	0.770	-0.031	0.042

Omnibus:	1089.615	**Durbin-Watson:**	0.352
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	133.317
Skew:	-0.191	**Prob(JB):**	1.12e-29
Kurtosis:	1.870	**Cond. No.**	13.9

Figure 3

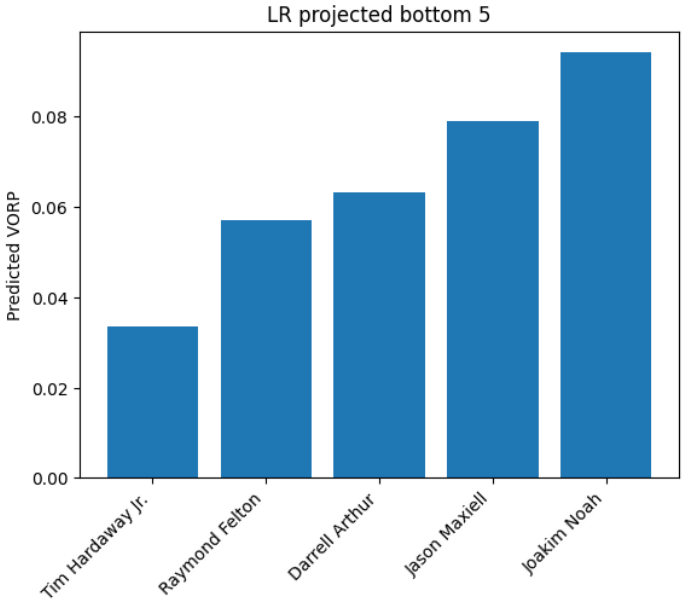

Figure 4

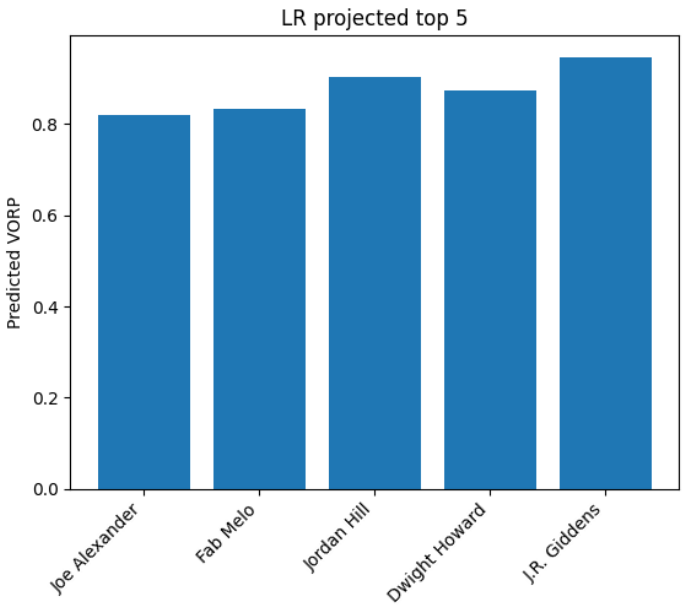

Figure 5

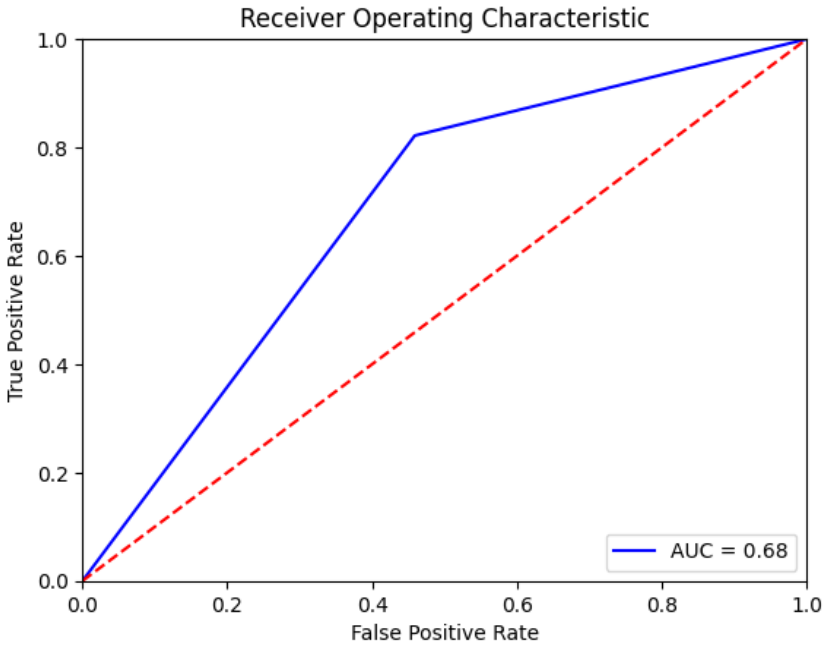

Figure 6

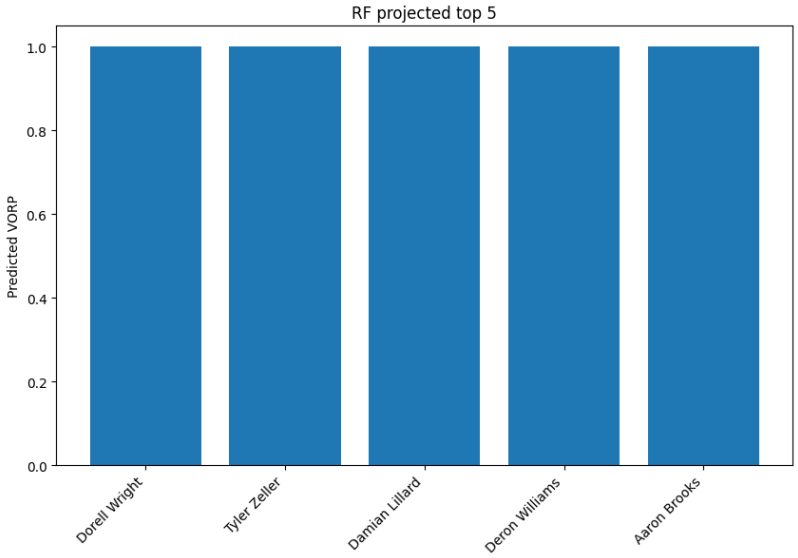

Figure 7

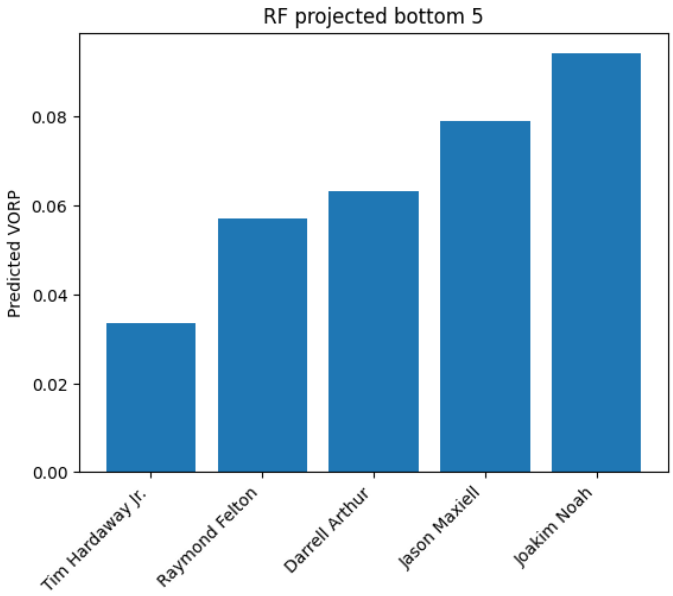

Figure 8

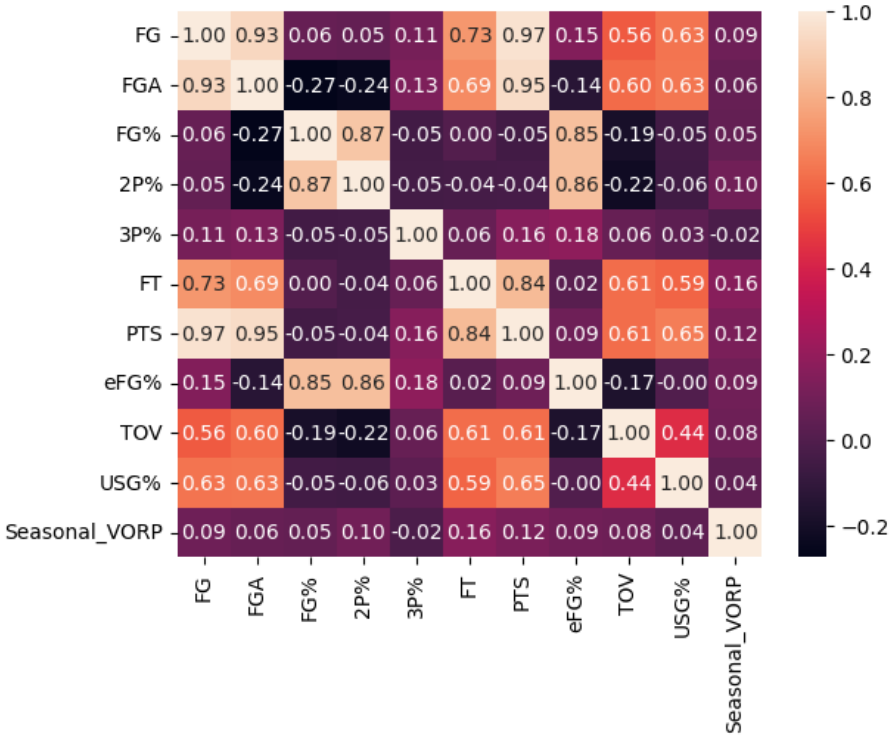
FG 1.00 0.93 0.06 0.05 0.11 0.73 0.97 0.15 0.56 0.63 0.09
FGA 0.93 1.00 -0.27 -0.24 0.13 0.69 0.95 -0.14 0.60 0.63 0.06
FG% 0.06 -0.27 1.00 0.87 -0.05 0.00 -0.05 0.85 -0.19 -0.05 0.05
2P% 0.05 -0.24 0.87 1.00 -0.05 -0.04 -0.04 0.86 -0.22 -0.06 0.10
3P% 0.11 0.13 -0.05 -0.05 1.00 0.06 0.16 0.18 0.06 0.03 -0.02
FT 0.73 0.69 0.00 -0.04 0.06 1.00 0.84 0.02 0.61 0.59 0.16
PTS 0.97 0.95 -0.05 -0.04 0.16 0.84 1.00 0.09 0.61 0.65 0.12
eFG% 0.15 -0.14 0.85 0.86 0.18 0.02 0.09 1.00 -0.17 -0.00 0.09
TOV 0.56 0.60 -0.19 -0.22 0.06 0.61 0.61 -0.17 1.00 0.44 0.08
USG% 0.63 0.63 -0.05 -0.06 0.03 0.59 0.65 -0.00 0.44 1.00 0.04
Seasonal_VORP 0.09 0.06 0.05 0.10 -0.02 0.16 0.12 0.09 0.08 0.04 1.00
FG
FGA
FG%
2P%
3P%
FT
PTS
eFG%
TOV
USG%
Seasonal_VORP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Figure 9