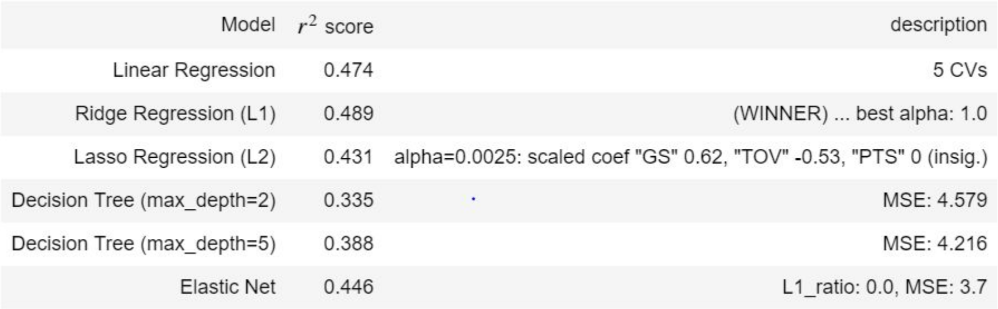
CapProj1: In-depth Analysis (Machine Learning)

Hypothesis Recap

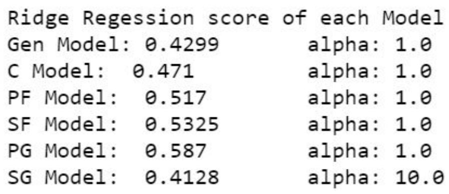
The goal is to build a model from the NBA stats dataset that can accurately predict RPM (unbiased but comparative to teammates only). Consider that RPM is a volatile singular score that can be skewed by the performance of your teammates. Now consider our predictive model, with numerous variables, that estimates RPM. The RPM estimator is a prediction of an unbiased score that ranks an individual NBA player's value, but it is based on a multitude of other variables making it much less prone to variability from the actions of other players.



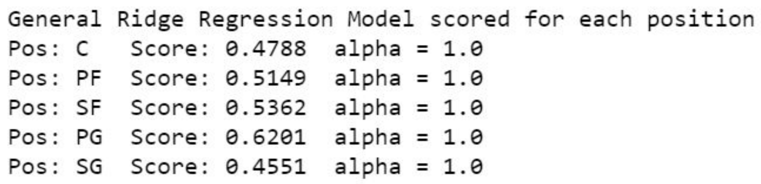
Model	r^2 score	description
Linear Regression	0.474	5 CVs
Ridge Regression (L1)	0.489	(WINNER) ... best alpha: 1.0
Lasso Regression (L2)	0.431	alpha=0.0025: scaled coef "GS" 0.62, "TOV" -0.53, "PTS" 0 (insig.)
Decision Tree (max_depth=2)	0.335	MSE: 4.579
Decision Tree (max_depth=5)	0.388	MSE: 4.216
Elastic Net	0.446	L1_ratio: 0.0, MSE: 3.7

Determining Technique

During Exploratory Data Analysis, it was clear that since all the variables are continuous numbers, a regression needed to be done on the feature set to determine the prediction variable. It was then time to begin testing the different types of regression techniques that would best understand the feature data and would also give the best prediction. For regression prediction, the best techniques to use are linear regression, ridge and lasso regression, and decision tree regression. Linear regression is the simplest technique to build models with higher dimensionality. This technique minimizes the sum of residuals and creates a linear model based on that principle. Ridge and Lasso regression work in the same way as linear regression but have an additional penalty factor which weights the coefficients of the regression with a given hyperparameter (alpha = 0.1, 1, 10, etc. for example). Different values of alpha are tested to determine which one produces the best score. This method worked very well, and ultimately the technique with the highest score was Ridge regression. The last method I used, was a Decision Tree regression, which works for classification as well. This technique scored the lowest.



```
Ridge Regession score of each Model
Gen Model: 0.4299       alpha: 1.0
C Model:  0.471         alpha: 1.0
PF Model:  0.517        alpha: 1.0
SF Model:  0.5325       alpha: 1.0
PG Model:  0.587        alpha: 1.0
SG Model:  0.4128       alpha: 10.0
```



```
General Ridge Regression Model scored for each position
Pos: C   Score: 0.4788  alpha = 1.0
Pos: PF  Score: 0.5149  alpha = 1.0
Pos: SF  Score: 0.5362  alpha = 1.0
Pos: PG  Score: 0.6201  alpha = 1.0
Pos: SG  Score: 0.4551  alpha = 1.0
```

<u>Honing and Tuning</u>

Now that we have determined which method of regression to use, we want to make sure that our prediction is as accurate as possible. I used an influence plot which takes the data points and measures their influence on the model's prediction. The data points that have high influences are usually outlier or contain incorrect information, and are then removed. Only a couple points were removed but it did help improve the scores. Before running regressions of the data, it is important to make sure that all features are scaled the same way. In this case, it did not affect our score but it is important to still do it so that coefficients are not inflated by their unit difference. Each feature was scaled to the normal distribution (mean: 0, variance: 1).

<u>PCA & Dimension Reduction</u>

Partial Component Analysis, done on features that are correlated (>~0.6 in our correlation matrix), groups together features that are correlated and finds the partial components that make them related, such that the number of partial components we want to keep is less or equal to the number of grouped features, but simultaneously explains 99% of the variance that the grouped features explained. By removing the redundant partial components through this process, it reduces the number of features without reducing the score.

Features grouped (in parentheses) by their collinearity:

'RPM ~ (GS+MPG+FT+PTS+STL+twP+AST+TOV+thP) + (DRB+BLK+ORB) + (eFG%+FG%+2P%)'

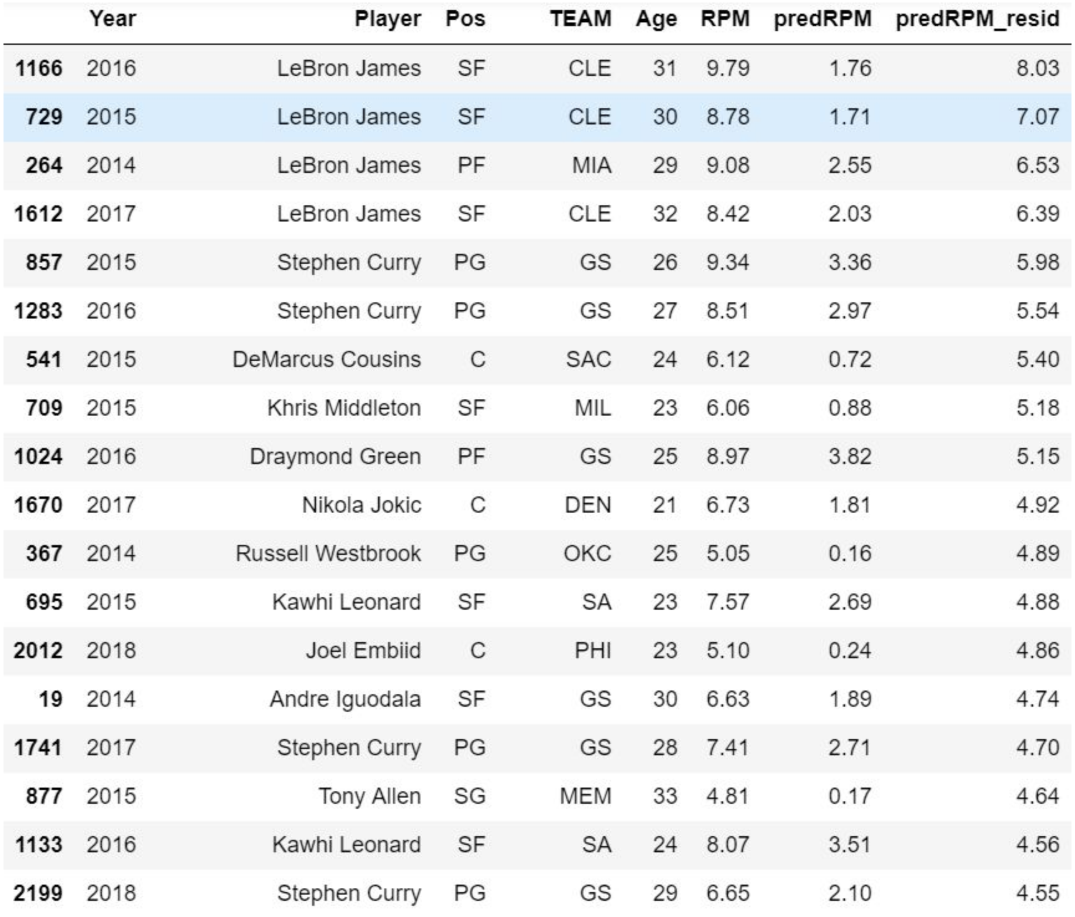
Essential Partial Components of each group remain:

'RPM ~ (PC_1+PC_2+PC_3+PC_4+PC_5+PC_6+PC_8) + (PC_9+PC_11) + (PC_12)'

<u>Interpreting Outcome</u>

It must be stated that we are assuming the model is as good as it can be given the data. After completing the prediction model, we are able to get a predicted RPM for each observation (player's stats for that year). The predicted RPM ("predRPM") is a predicted value based on all the individual's stats, and the actual RPM ("RPM") is based on team performance while the player is on or off the court. Because the predicted RPM accounts for the portion of RPM attributed to the individual's stats, we can infer that the difference between the actual and predicted RPM (ie. prediction residual: "predRPM_resid") is then the portion of the player's effect that is not based on their individual stats but on the intangible effect that they contribute to the team. We then infer that the value of the residual is actually a measure of the intangible effect. Assuming the model is as good as it can be given the data, if a player has a higher actual RPM than predicted (“RPM”>”predRPM”), their RPM is being underrated and that unaccounted portion can be attributed to their intangible input. For players with a lower RPM than predicted (“RPM”<”predRPM”), their RPM is being overrated because their intangible impact on the team is detrimental.

[only RPM and predRPM greater than zero shown, sorted by predRPM_resid]



	Year	Player	Pos	TEAM	Age	RPM	predRPM	predRPM_resid
1166	2016	LeBron James	SF	CLE	31	9.79	1.76	8.03
729	2015	LeBron James	SF	CLE	30	8.78	1.71	7.07
264	2014	LeBron James	PF	MIA	29	9.08	2.55	6.53
1612	2017	LeBron James	SF	CLE	32	8.42	2.03	6.39
857	2015	Stephen Curry	PG	GS	26	9.34	3.36	5.98
1283	2016	Stephen Curry	PG	GS	27	8.51	2.97	5.54
541	2015	DeMarcus Cousins	C	SAC	24	6.12	0.72	5.40
709	2015	Khris Middleton	SF	MIL	23	6.06	0.88	5.18
1024	2016	Draymond Green	PF	GS	25	8.97	3.82	5.15
1670	2017	Nikola Jokic	C	DEN	21	6.73	1.81	4.92
367	2014	Russell Westbrook	PG	OKC	25	5.05	0.16	4.89
695	2015	Kawhi Leonard	SF	SA	23	7.57	2.69	4.88
2012	2018	Joel Embiid	C	PHI	23	5.10	0.24	4.86
19	2014	Andre Iguodala	SF	GS	30	6.63	1.89	4.74
1741	2017	Stephen Curry	PG	GS	28	7.41	2.71	4.70
877	2015	Tony Allen	SG	MEM	33	4.81	0.17	4.64
1133	2016	Kawhi Leonard	SF	SA	24	8.07	3.51	4.56
2199	2018	Stephen Curry	PG	GS	29	6.65	2.10	4.55

It must be pointed out that when we plot players who scored a high number of points against the prediction RPM residuals, we see that most residuals are positive. However, it cannot be determined whether this trend is simply an anomaly, if there is more that needs to be explained and the model ought to be more complex, or if our theory of positive residuals as indication of an undervalued player is in line. But based on the dataset size and the logically appropriate approaches taken to create the model, it is sufficient to assume that the relationship between actual and predicted RPM has been adequately represented.