Capstone Project 1: NBA Player Analysis (Springboard - Christos Magganas)

Objective: [Hypothetically] As the new General Manager for the Seattle Supersonics, an expansion NBA franchise, I want to be able to evaluate players appropriately, without bias. As the owner of an expansion team in a competitive league with less money to offer potential free agents, our approach to building a competitive team has to be to acquire undervalued players that we predict will have success.

Data Wrangling

- Source of the Data and Method of Acquirement

The data was scraped and collected from Basketball-Reference and ESPN. The data includes statistics about all the NBA players from 2014-2018.

Links: https://www.basketball-reference.com/, http://www.espn.com/nba/statistics/

The goal of this project is to create an accurate predictor for the metric RPM. Real-Plus-Minus is an unbiased comparative metric to evaluate player value. Specifically, a player's estimated on-court impact on team performance, measured in net point differential per 100 offensive and defensive possessions. The data collected from these two sites, will help achieve the goal of creating this predictor.

- Cleaning Steps

Once the data was collected via python requests. The html format needed to be converted (BeautifulSoup), and then needed to be sorted through such that the contents of the tables were stored into some combination of lists and dictionaries. The data was then arranged into a DataFrame, where it would be easily analyzed. This process was done for both sites.

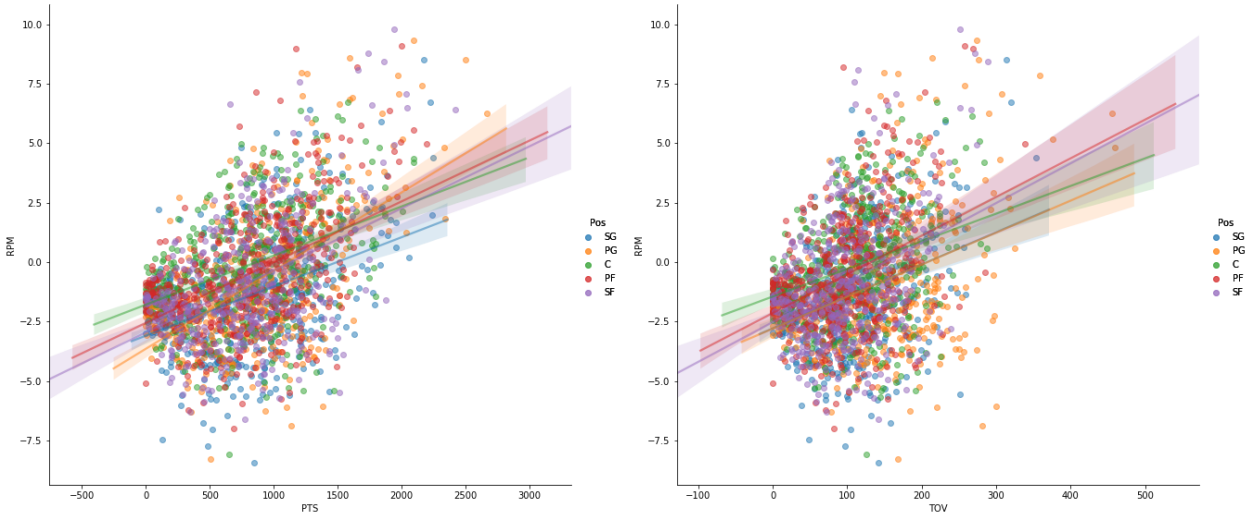
- Merging and Sorting

The two DataFrames would have to be merged together such that no data would be lost or changed. The data also had to be arranged such that all information for a player that year was on a single row, and then the two could be merged into one. Some players were missing from the data because the spelling of their names was different in each DataFrame. Those names had to be changed to match the others. Additionally, some positions were different, so the same had to be done for them. No outliers were removed, due to the high number of variables and proportionally lower number of datapoints. At this point, the data had been cleaned and sorted enough to be saved into a csv for EDA.

Data Storytelling

From the data collected, containing information of the NBA players from 2014-2018, analysis on variables such as assists, points and rebounds are conducted to find their correlation to RPM (player rating). Here is some idea of what that data looks like now that it has been put together.

- **Embarking on our Data Journey**
    - Finding trends

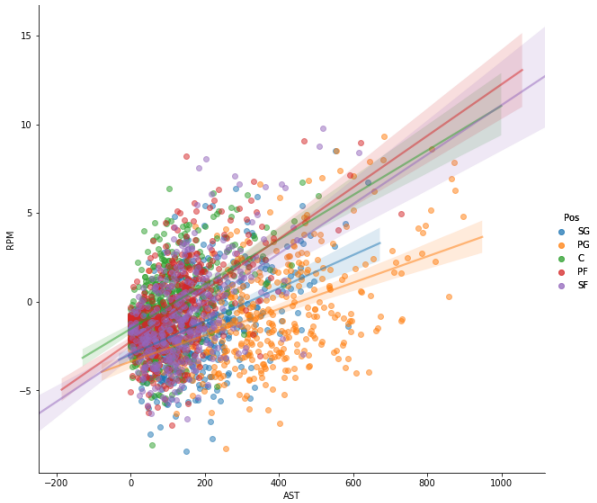


scatter plots comparing PTS (points) and TOV (turnovers) against RPM

- Insights / Further Questions

If we look at the graph on the left, we see that PTS and RPM have a positive relationship for all positions. This of course is a very intuitive relationship. If you score points, you directly impact your team in a positive way. However, we see in the right graph that the relationship between TOV and RPM is also positive. This means that players who lose the ball more often also have a higher positive impact on their team. This is initially counterintuitive, but once you consider that players who lose the ball are more often in possession of the ball ie. the “ball-carriers” are more often the talented players, it fits the data. This leads us to ask the question of which other variables have trends counter to belief.

Another interesting note, is that when comparing assists (AST) to RPM, we get very different slopes for each position. Above, we can see that the regression lines for each position generally are close to one another and have similar slope. We must consider which variables when evaluated produce better results in a general model with more data, and which ones ought to be compared separately.

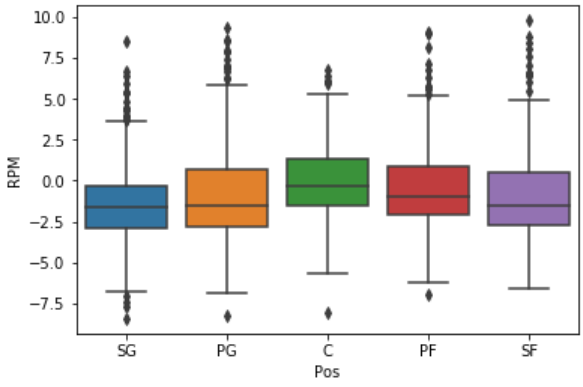


Exploratory Data Analysis - Inferential Statistics

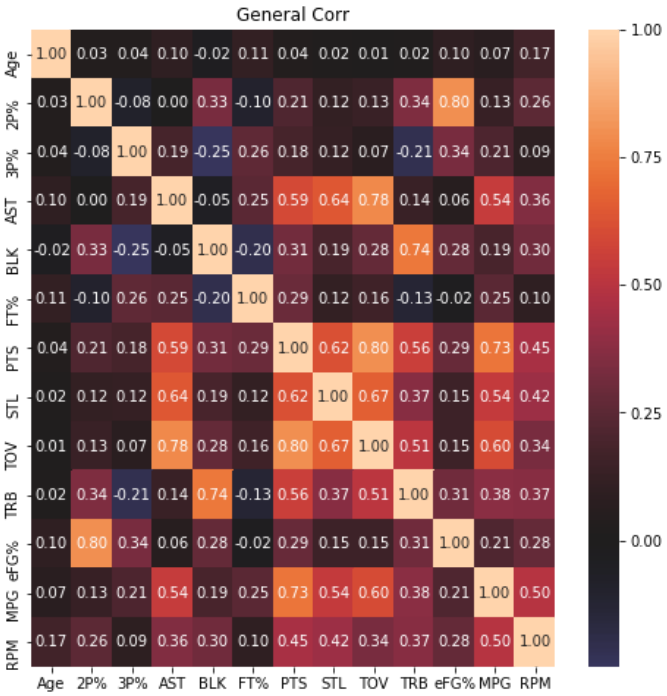
- Initial Analysis:

We see from our analysis of the data that indeed each variable has some value of correlation to the compared variable RPM. Most values that are expected to have a strong correlation to RPM such as points and assists do in fact have a high correlation coefficient. However, some values counterintuitively have high correlation coefficients such as turnovers (which decreases offensive impact and likely increases defensive impact negatively). These trends are most likely the cause of many variables acting simultaneously.

Plot (below-left): Box-Pot of RPM by Position (all years)



Plot (right): Correlation Matrix of all variables against RPM (all years)



- Primary Plot Analysis:

RPM is a good unbiased tool to evaluate players comparatively versus one another, but the process of evaluation occurs first within the team. A player's RPM should be not be directly compared to a player from another team until the quality of the team as a whole is taken into consideration first. Additionally, in the first plot, we can see that RPM also differs between positions. To understand this skewness, we must analyze our data by separating players by position and to seeing which variables are more or less impactful than the general model with accounts for all the positions.

- Insights of Primary Plots / Further Analysis / Additional Visuals:

Position	C	PF	SF	PG	SG	General
GS	0.2166	0.269	0.285	0.3788	0.2622	0.2769
MP	0.2336	0.2981	0.2597	0.3334	0.2587	0.2571
MPG	0.2379	0.2904	0.2804	0.3379	0.2529	0.2478
FT	0.1868	0.2369	0.212	0.3231	0.1556	0.2227
FTA	0.182	0.1995	0.1973	0.2955	0.1523	0.2141
PTS	0.1592	0.2333	0.1987	0.3161	0.1723	0.2066
FG	0.1326	0.2044	0.1753	0.2638	0.1567	0.1851
STL	0.1804	0.1895	0.2233	0.2681	0.2284	0.1793
DRB	0.175	0.1524	0.1862	0.2259	0.1196	0.1673
2P	0.1128	0.1503	0.1414	0.1722	0.1268	0.1533
TRB	0.1405	0.1147	0.163	0.2212	0.1305	0.1388
FGA	0.1038	0.1717	0.1313	0.2132	0.1184	0.1321
AST	0.218	0.3101	0.2678	0.24	0.2006	0.1284
2PA	0.0846	0.1218	0.106	0.1305	0.1014	0.1191
TOV	0.0879	0.1436	0.1326	0.1298	0.1201	0.1125
BLK	0.0843	0.1203	0.108	0.0815	0.1259	0.093
eFG%	0.063	0.0515	0.0741	0.1149	0.0528	0.0759
FG%	0.0465	0.0319	0.0796	0.078	0.0695	0.0752
2P%	0.0493	0.0338	0.0532	0.0624	0.0385	0.0604
ORB	0.0536	0.0252	0.0437	0.1317	0.0914	0.0603
3P	0.0356	0.0738	0.0825	0.2502	0.0886	0.0432
PF	0	0.0216	0	0.0439	0.0286	0.03
3PA	0.0298	0.0626	0.0664	0.2138	0.063	0.0294
Age	0.0315	0.0638	0	0	0.0487	0.0288
FT%	0	0.0458	0.0203	0.0373	0	0.0143
3P%	0.0402	0.0425	0.0325	0.0622	0.0395	0.0082

The end goal is to create a best-fit model that accurately predicts RPM. To do that and be able to see clear connections, we must strip away unrelated information and adjust the related information in a way that can be understood. From our analysis above of the primary plots, we know that we must first strip away the unrelated variables. By sorting these variables by their r-squared values (the variable's correlation to RPM) and using their p-value (in this case whether the relationship is statistically significant), we can determine which variables we should use. In the chart to the left, the variables are sorted by their r-squared value for all positions and given value zero (0) if insignificant (p-value greater than 0.01).

From this chart we now know which variables have relevance. To create a good model, we must use the right combination of variables to produce the most accurate prediction of what RPM is based on those

variables. For most of the variables, the general model has an r-squared value averaging somewhere close to and between the values for each position. A general model would in those cases be sufficient to describe them. But if we look at assists (AST), we see that all the r-squared values are higher for each position than the general model. If we look back at the scatter plot of 'AST' and 'RPM', we can see the data is much different for each position. Therefore, we need to keep in mind, during model testing, which variables must be evaluated separately and which ones can follow a general analysis. During the model testing process, we see that the more variables we have in our model, the higher the correlation is to the RPM data. But that is merely because the model is from one dataset, and having more variables just means the model can describe itself better. To be able to predict future RPM accurately, we must determine which variables have a direct impact.

To determine which variables to use, we initially include all the variables in our model. The model will initially contain variables that are redundant (Multicollinearity). To get the right combination of variables for our model, we have to determine which of these variables are redundant of one another. Then one group of redundant variables at a time, only use the variable with highest correlation. Using the Variance Inflation Factor (VIF), we can determine which variables are likely correlated to one another. Variables with higher VIFs are likely related in this way. When you systematically remove one variable, every other variable's VIF value decreases, significantly more so for the variable related to the removed variable. For example, the variables "PTS" (points), "FG" (field-goals), "2P" (two-pointers) and "3P" (three-pointers) all describe scoring two/three points for a made basket for that individual. These are essentially combinations of the same thing. We then leave only one variable in the model, that which has the highest correlation, in this case "PTS". After repeating this process, we get a model without insignificant or redundant independent variables (highest r-squared first).

'RPM ~ GS + MPG + FT + PTS + STL + DRB + AST + TOV + BLK + eFG%'

In-depth Analysis (Machine Learning)

Objective Recap and Re-evaluation

Our 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
```

Honing and Tuning

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).

PCA & Dimension Reduction

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) + (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)'

Interpreting Outcome

It must be stated that we are assuming the model is as good as it can be given the data. We can assume it is the best it can be because of the thorough statistical processes we took throughout the model creation, as seen above.

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.