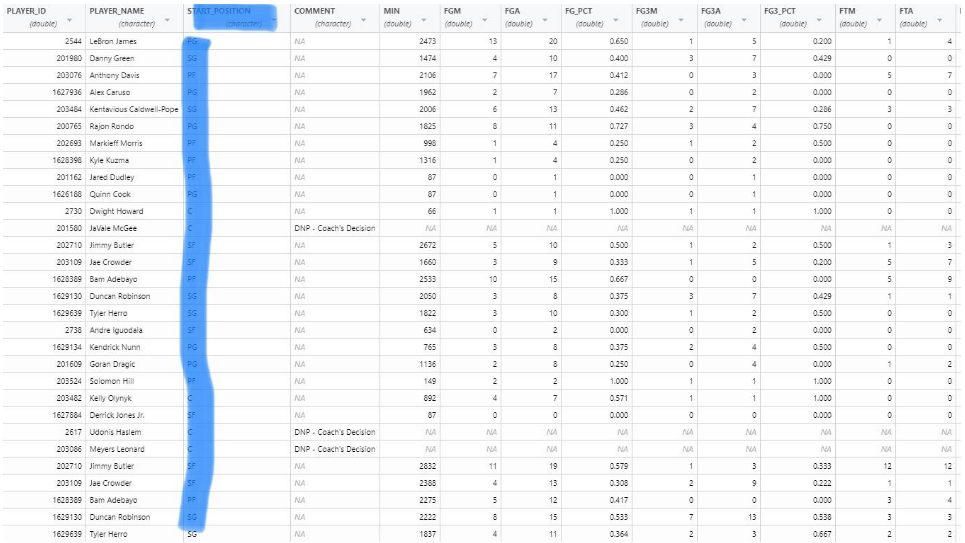
Augustine Xu, Yunzhe Qi, Abhi Vadali
IS 517
Final Project Report

The NBA today uses data analytics to make nearly every decision. More specifically, teams use data to track and predict injuries, to scout new talent, and to make game strategy decisions. One of these decisions is which players to give the most playing time, and more specifically, assessing whether one position is more valuable than another. Generally, teams are biased towards the starters, since they are known to have more talent, but there hasn't been a large-scale exploration of the strategy they should use with regard to bench players. It is useful to employ a statistical learning solution here because the best way to measure each position's impact is quantitatively, rather than subjectively. An ML solution helps us assess and truly gain a quantitative understanding of whether one position is more predictive of winning than another. The primary research question is, for each of the thirty teams' benches, which positions (PG, SG, SF, PF, C) are most predictive of winning? A secondary research question is, for each of the thirty teams' starters, which positions (PG, SG, SF, PF, C) are most predictive of winning? One final secondary research question is, out of all the players in the league, which positions (PG, SG, SF, PF, C) are most predictive of winning? In order to assess this, we gathered statistical data for every player from every game in the 2019-2020 season. Then, we split the dataset into starters and bench players and grouped together each of the five positions. Finally, we used the stats of each position player to predict the win/loss outcome of the game through Logistic Regression, Random Forest, and SVM classifiers and registered the accuracy of the model for each position for the starters, bench players, and the whole league.

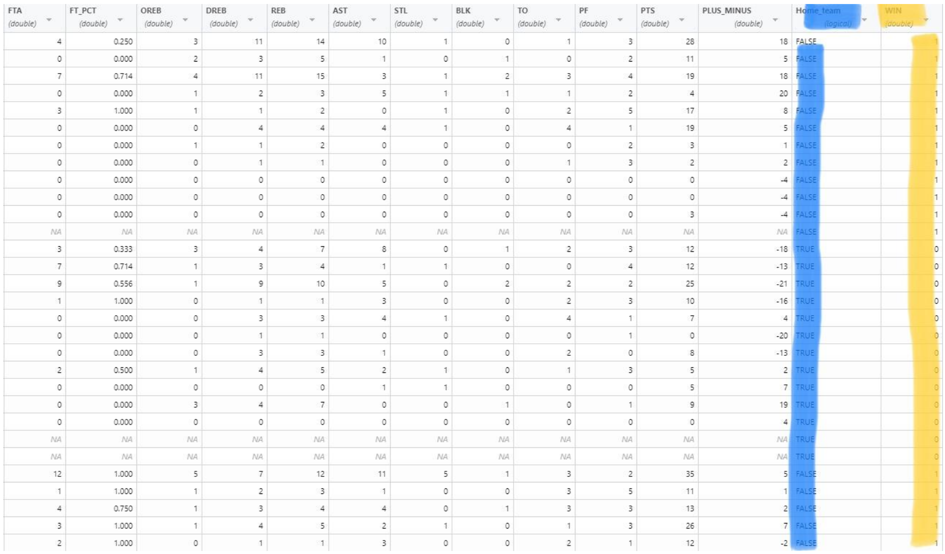
Our datasets come from the Kaggle Competition named "NBA games data". There are 5 CSV files in total, and three of them were used which are "games.csv", "games_details.csv" and "players.csv". Aiming to analyze for our research questions, we did some data cleaning and preprocessing on the datasets. The Fig1 and Fig2 are the screenshots of our annotated dataset. The marked columns are added columns which are necessary for our analysis. The START_POSTION is the significant factor in our research questions and WIN is the only response variable through the whole project.

There are no previous-published solutions to this problem. There are several published articles on making predictions on NBA games. However, these articles are using different predictors in different aspects to predict win-lose of NBA games. For example, the performance of the whole team is considered in one article while in our project, the performance of the single player is the factor we are focusing on. The predictions on NBA games are also common in the gambling industry while different factors are analyzed.



PLAYER_ID (double)	PLAYER_NAME (character)	START_POSITION (character)	COMMENT (character)	MIN (double)	FGM (double)	FGA (double)	FG_PCT (double)	FG3M (double)	FG3A (double)	FG3_PCT (double)	FTM (double)	FTA (double)
2544	LeBron James	PG	NA	2473	13	20	0.650	1	5	0.200	1	4
201980	Danny Green	SG	NA	1474	4	10	0.400	3	7	0.429	0	0
203076	Anthony Davis	PF	NA	2106	7	17	0.412	0	3	0.000	5	7
1627936	Alex Caruso	PG	NA	1962	2	7	0.286	0	2	0.000	0	0
203484	Kentavious Caldwell-Pope	SG	NA	2006	6	13	0.462	2	7	0.286	3	3
200765	Rajon Rondo	PG	NA	1825	8	11	0.727	3	4	0.750	0	0
202693	Markieff Morris	PF	NA	998	1	4	0.250	1	2	0.500	0	0
1628398	Kyle Kuzma	PF	NA	1316	1	4	0.250	0	2	0.000	0	0
201162	Jared Dudley	PF	NA	87	0	1	0.000	0	1	0.000	0	0
1626188	Quinn Cook	PG	NA	87	0	1	0.000	0	1	0.000	0	0
2730	Dwight Howard	C	NA	66	1	1	1.000	1	1	1.000	0	0
201580	JaVale McGee	C	DNP - Coach's Decision	NA	NA	NA	NA	NA	NA	NA	NA	NA
202710	Jimmy Butler	SF	NA	2672	5	10	0.500	1	2	0.500	1	3
203109	Jae Crowder	SF	NA	1660	3	9	0.333	1	5	0.200	5	7
1628389	Bam Adebayo	PF	NA	2533	10	15	0.667	0	0	0.000	5	9
1629130	Duncan Robinson	SG	NA	2050	3	8	0.375	3	7	0.429	1	1
1629639	Tyler Herro	SG	NA	1822	3	10	0.300	1	2	0.500	0	0
2738	Andre Iguodala	SF	NA	634	0	2	0.000	0	2	0.000	0	0
1629134	Kendrick Nunn	PG	NA	765	3	8	0.375	2	4	0.500	0	0
201609	Goran Dragic	PG	NA	1136	2	8	0.250	0	4	0.000	1	2
203524	Solomon Hill	PF	NA	149	2	2	1.000	1	1	1.000	0	0
203482	Kelly Olynyk	C	NA	892	4	7	0.571	1	1	1.000	0	0
1627884	Derrick Jones Jr.	SF	NA	87	0	0	0.000	0	0	0.000	0	0
2617	Udonis Haslem	C	DNP - Coach's Decision	NA	NA	NA	NA	NA	NA	NA	NA	NA
203086	Meyers Leonard	C	DNP - Coach's Decision	NA	NA	NA	NA	NA	NA	NA	NA	NA
202710	Jimmy Butler	SF	NA	2832	11	19	0.579	1	3	0.333	12	12
203109	Jae Crowder	SF	NA	2388	4	13	0.308	2	9	0.222	1	1
1628389	Bam Adebayo	PF	NA	2275	5	12	0.417	0	0	0.000	3	4
1629130	Duncan Robinson	SG	NA	2222	8	15	0.533	7	13	0.538	3	3
1629639	Tyler Herro	SG	NA	1837	4	11	0.364	2	3	0.667	2	2

Figure 1. Screenshot of dataset with annotated column.



FTA (double)	FT_PCT (double)	OREB (double)	DREB (double)	REB (double)	AST (double)	STL (double)	BLK (double)	TO (double)	PF (double)	PTS (double)	PLUS_MINUS (double)	Home_team (logical)	WIN (double)
4	0.250	3	11	14	10	1	0	1	3	28	18	FALSE	1
0	0.000	2	3	5	1	0	1	0	2	11	5	FALSE	1
7	0.714	4	11	15	3	1	2	3	4	19	18	FALSE	1
0	0.000	1	2	3	5	1	1	1	2	4	20	FALSE	1
3	1.000	1	1	2	0	1	0	2	5	17	8	FALSE	1
0	0.000	0	4	4	4	1	0	4	1	19	5	FALSE	1
0	0.000	1	1	2	0	0	0	0	2	3	1	FALSE	1
0	0.000	0	1	1	0	0	0	1	3	2	2	FALSE	1
0	0.000	0	0	0	0	0	0	0	0	0	-4	FALSE	1
0	0.000	0	0	0	0	0	0	0	0	0	-4	FALSE	1
0	0.000	0	0	0	0	0	0	0	0	3	-4	FALSE	1
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	1
3	0.333	3	4	7	8	0	1	2	3	12	-18	TRUE	0
7	0.714	1	3	4	1	1	0	0	4	12	-13	TRUE	0
9	0.556	1	9	10	5	0	2	2	2	25	-21	TRUE	0
1	1.000	0	1	1	3	0	0	2	3	10	-16	TRUE	0
0	0.000	0	3	3	4	1	0	4	1	7	4	TRUE	0
0	0.000	0	1	1	0	0	0	0	1	0	-20	TRUE	0
0	0.000	0	3	3	1	0	0	2	0	8	-13	TRUE	0
2	0.500	1	4	5	2	1	0	1	3	5	2	TRUE	0
0	0.000	0	0	0	1	1	0	0	0	5	7	TRUE	0
0	0.000	3	4	7	0	0	1	0	1	9	19	TRUE	0
0	0.000	0	0	0	0	0	0	0	0	0	4	TRUE	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	TRUE	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	TRUE	0
12	1.000	5	7	12	11	5	1	3	2	35	5	FALSE	1
1	1.000	1	2	3	1	0	0	3	5	11	1	FALSE	1
4	0.750	1	3	4	4	0	1	3	3	13	2	FALSE	1
3	1.000	1	4	5	2	1	0	1	3	26	7	FALSE	1
2	1.000	0	1	1	3	0	0	2	1	12	-2	FALSE	1

Figure 2. Screenshot of dataset with response variable.

We did not have certain expectations going into the project since the results of sports games could be considered as one of the “most unpredictive” things. The unexpected winners were not rare to see in the history of sports. The predictors we used in our project are just a small part of all the factors on the results of group sports games, especially we are focusing on the benches which might have less impacts on the other factors like starters.

In our initial proposal, we planned to analyze the model accuracies for the top bench players?. At that time, our primary goal was to use the model accuracy as a proxy to measure which bench player contributed most to his team. Then, other than the starters and the ‘sixth man’ from each team, we also managed to find the bench player whose individual statistics had the highest correlation with his team’s outcome in the game, specifically winning. Meantime, we also proposed to leverage the registered statistics from different players, such as PPG, RPG, +/-

(etc.) aiming to assess which of these features were most important for the model in the prediction of whether the team wins or not.

During our experiments, we have discovered that the models with pure bench player statistics tend to perform poorly, and their performances are especially close to 'random guessing'. Thus, we instead introduced the statistics from the starters and trained two additional kinds of models, namely the models with pure starters and models with all the players. By comparing the performances of models with different datasets, we would then have more comprehensive insights on why our 'bench player models' failed, and the impacts of statistics from different kinds of players. Experiment details regarding different models and datasets are presented in Figure 3, 4, and 5.

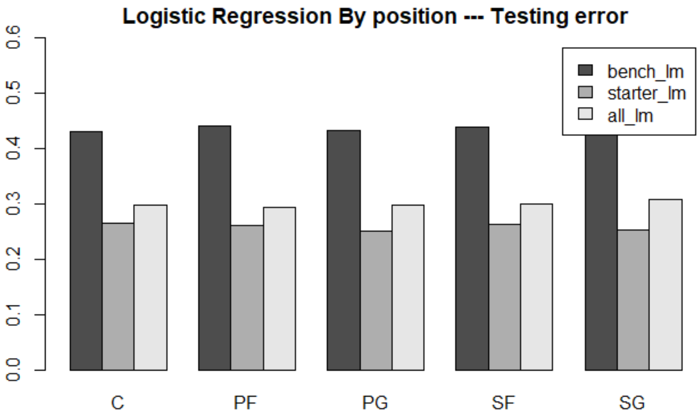


Figure 3. Models with statistics of different kinds of players and positions (Logistic Regression)

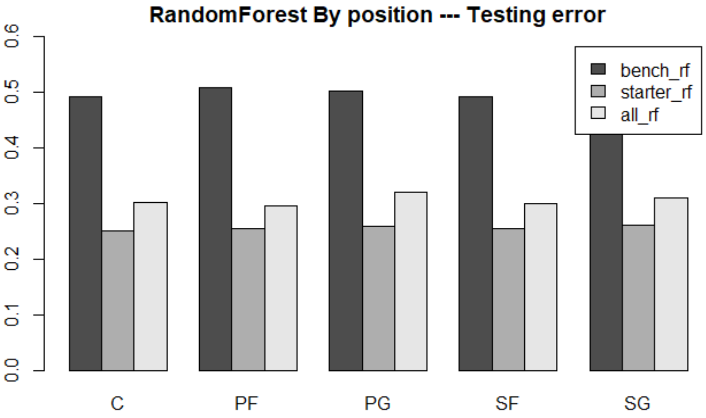


Figure 4. Models with statistics of different kinds of players and positions (Random Forest)

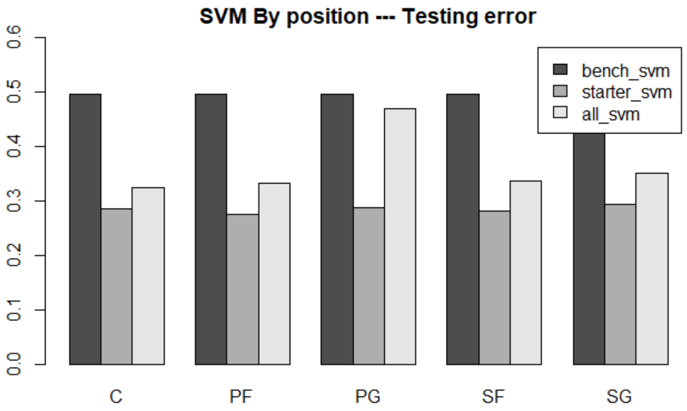


Figure 5. Models with statistics of different kinds of players and positions (SVM + radial kernel)

Secondly, in order to investigate the consistency of bench players and starter players within different teams, we have also trained separate models for each team. For a specific team, we argue that if the model with its bench player statistics only and the model with its bench player statistics only tend to perform similarly, the performances of these two kinds of players within this team would have more consistency. On the other hand, for example, if the model with starters' statistics tends to perform much better than the model with bench player statistics only, the difference between this team's starters and bench players would be significant. Details of the experiments are shown in Figure 6.

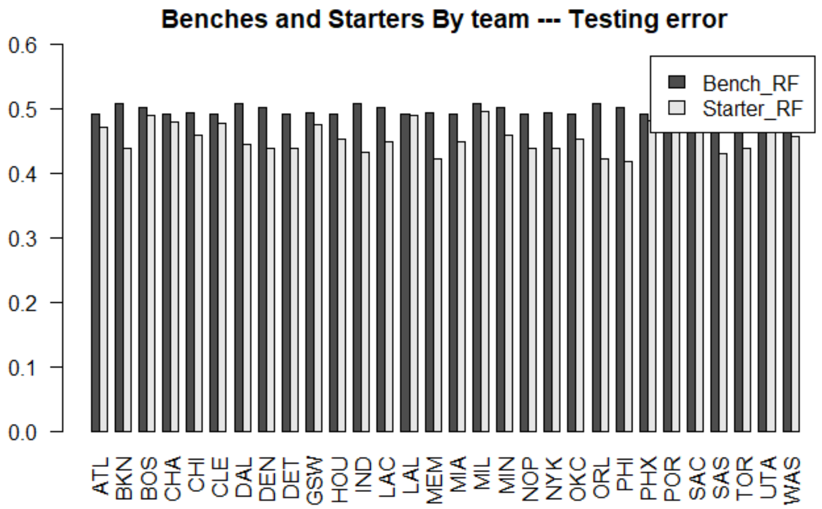


Figure 6. Models with statistics of different teams (Random Forest)

Our team tends to work together on both ideas and the model realization. During our virtual meeting, we would fix the ideas as well as the actual workloads. Then, we would equally split the work into different topics and finish them collaboratively. Summing up the results we have, we then conduct our analysis and reach our conclusions.

Our team split the pieces of the project amongst ourselves and regularly met to make sure we were on track with each of our tasks. Some of the different parts of the project were, preprocessing the data, building the classifiers for all the data, building the classifiers for the starters, building the classifiers for the bench players, and writing the presentation and this report. If we had more time, we would like to explore more the comparison between bench players and starters, rather than focusing on the comparison between positions. More specifically, we would like to normalize the bench players' and starters' statistics and explore whether the starters would still be more predictive of winning, or whether there are some bench players or bench positions that are worthy of more playing time due to outperforming the starters based on their minutes.