Productivity Spillovers with Assist Network

Johnny Cheung Chun Yu (1155194973)[1*†],

Classic economic model predicts that workers will be paid the value of their marginal product (*1*). In the sport industry, from a pure offensive side, the marginal product of a player is the contribution on scoring. Such contribution could be examined by his countable scoring and, in a law ,uncountable, how much could he help others score. I use a play-by-play data from NBA to develop a assist matrix to better understand how. This matrix describe an assist network in a fixed environment. I then apply deep learning estimation methods to examine Team Dynamics and Synergy Analysis.

1 Introduction

The paper makes two notable contributions to the analysis of teamwork using basketball lineups and player dynamics. The first part of my paper introduces an ”assist matrix” to meticulously record play-by-play data, enabling a detailed examination of individual and team performance. Most dat scientist, the use intelligent machine learning framework is for predicting the results of games played at the NBA. They wish to figure the influential features set as the input that affects the outcomes of NBA games. The effect of star players to current lineup could vary by different team (*1*). The play-by-play data is imperative for rigorous basketball analysis as it furnishes a detailed, moment-by-moment record of game events, offering contextual insights that aggregate statistics often overlook. Such data preserves the sequential nature of the game, facilitating a more profound understanding of player interactions, team dynamics, and the ramifications of individual actions on overall game progression. The models help us better understanding the impact of star players

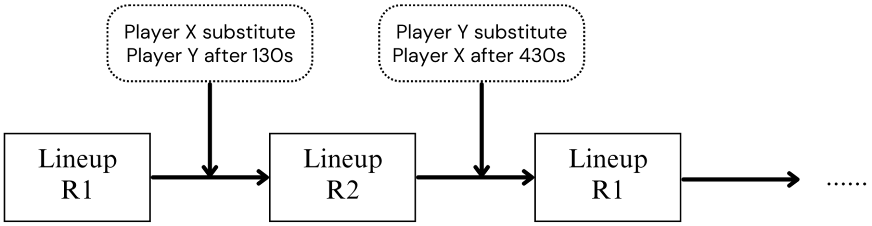

Figure 1: A demonstration of Labeling

like LeBron James and Chris Paul on team dynamics, considering the dynamic environment and varying playtime. By applying economic theory to relate players' wages to their marginal product, it quantifies players' contributions to scoring (*2*). Additionally, the later part of this paper employs deep learning methods to simulate optimal player lineups, demonstrating how effective assistants enhance overall team performance. This comprehensive approach provides valuable information on optimizing the lineups of the basketball team and understanding the interplay between players.

In NBA games, the dynamics of player 'sets' is influenced by a combination of team strategies (*3*), in-game situations, and individual development. Coaches frequently adjust roles based on offensive and defensive schemes, exploiting match-ups and adapting to the game's context, such as score and time remaining. Figure 1 demonstrate the time line of an NBA match up. Although detailed play-by-play data is available on the NBA website, there are still some elements missing: One example would be the players. The 'player' environment in the game is dynamic as the coach could be substituted anytime. In my work, I first scrape the (i) the starting-line up and (ii) the salary data from NBA.reference.com. The machine will then record the player environment and players' contribution. We recorded 4 contributions, the scorer, assist, duration and score-differences. For simplicity, I used a token, for example R1000 to represent a set of players. The machine will automatically detect who is currently in the court and so to memorize. And then the columns for assist and scorers will detect the description on who scores solely or who get assisted, and by who. USing this method, we can easily construct a matrix of play-by-play data under a dynamic set of players. The website https://www.basketball-reference.com/ hass provided the sufficient personal data and most importantly the starting lineup.

2 Data Modification

Table 2 describes some general statics for the duration. As different sets of player will have different playtime, the longer the duration of that matrix will lead to an inflation of statics. After scrape all the necessary data. The a very important thing is to make sure the the matrix is not inlfated by the duration. Although coaches had tried different combinations of players, some of the combinations induration is too short to examine. Of previous analysis on the matrix had already excluded those matrix that were less than 24 minutes. Also, I throw away all the free throws. The environment means different set of players. Let the pairing matrix be A in the environment i, the effective Ball Engagement Matrix E_i would be

$$B_i = A_i^T + S_i$$

For example in R15432, the five players are arranged LAL_Davis, LAL_Prince, LAL_Russell, LAL_Reaves and LAL_LeBron.

$$A_1 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad S_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

I would Compare the both B_1 and E_1

$$B_1 = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 3 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 & 0 \\ 0 & 1 & 0 & 5 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{2}$$

Which S is the matrix of solo handling ball moment. Subject to the duration, I fixed the environment into 12 minutes that is the modified matrix would be:

$$B_i = 720(A_i^T + S_i)/Duration_i$$

As an example table 2 presents a matrix that describe in environment R1, The first row for (CLE) Allen (from 0.14293 to 0.061256) repersents the expected effective shoot per 12 minutes, assisted by the column player Allen, Garland ... respectively. 'Allen assist Allen' implies the player handle the ball by himself without any assistant. The sum of all values in the the matrix computes the frequency of shooting. Afterwards, adjusting the inflation by duration, table 3 recorded the summary statistics of effective shoring frequnect and table 4 recorded the top 5 teams with the highest 'efficiency' in effective shooting frequency, noted that the shooting frequency does not necessarily implies better win rate due to two reason, one is the qlality of shoot, different location (for example 3-pt line) will generate different scores. Second is the impact with 0, the matrix may underestimate in short period time that stable assist that greater than zero is not constructed.

3 Experiment

In my first experiment, I build and train a neural network to predict matrices based on different combinations of player tokens, while accounting for various durations. The process starts by loading matrix CSV files and duration data from specified directories. Each matrix is normalized by its corresponding duration to handle variations in duration. All matrices are padded or truncated to ensure they have the same size. The unique player tokens are then extracted and converted into one-hot encoded feature vectors. The data is split into training and validation sets, and the one-hot encoded features are used to train a simple feedforward neural network. The model is trained using the training set and evaluated on the validation set. Finally, the training history is plotted to visualize the model's performance over epochs. The overall goal is to create a predictive model that can generate matrices for different player combinations, taking into account the duration of each combination.

Figure 2 illustrate the relation validation Loss and the training Loss over epoches under the NLP model. Different models may have different line-shape due to its model design. The graph shows the training and validation loss over epochs while training a neural network. The training loss (blue line) decreases rapidly and stabilizes near zero, indicating that the model is fitting the training data. The validation loss (orange line) stabilizes at a higher level, suggesting that the model is over-fitting. Figure 3 demonstrates more model examples of over-fitting.

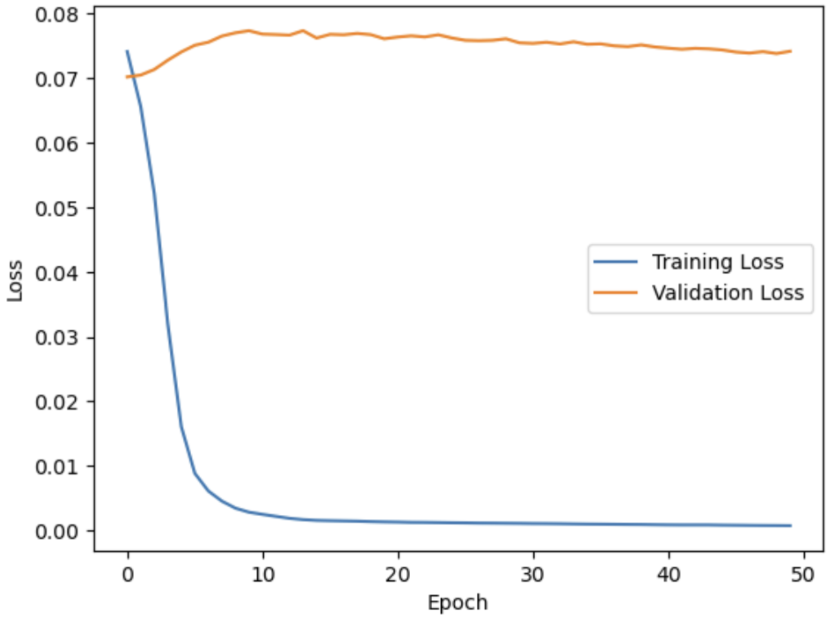

Figure 2: A demonstration of MLP using tensorflow

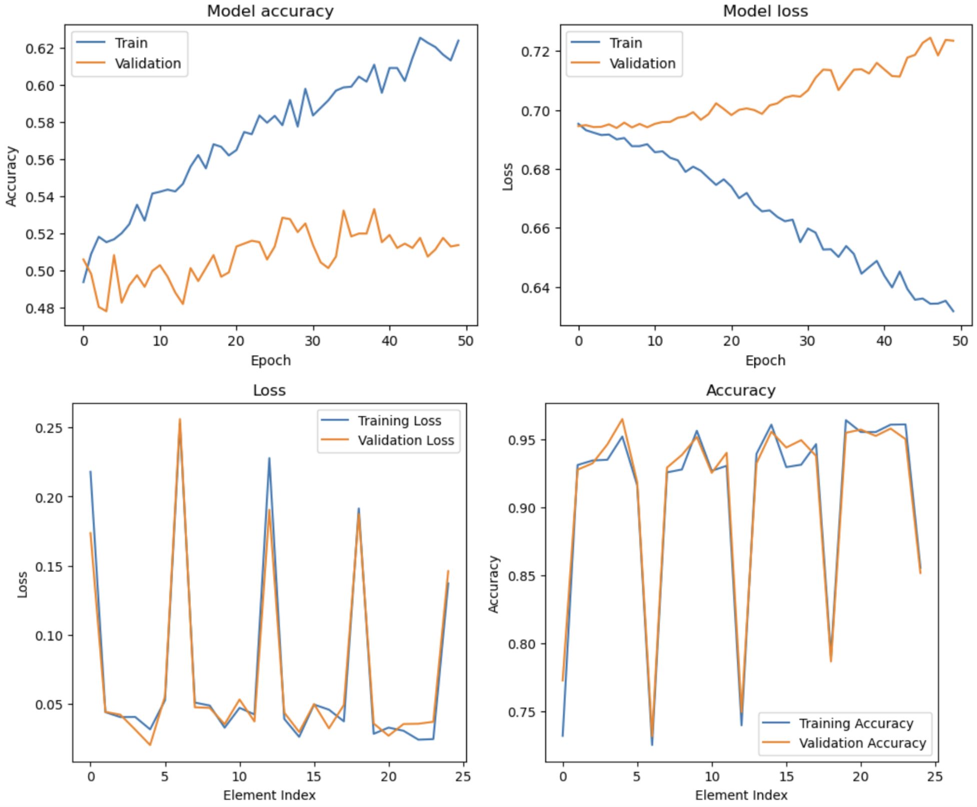

Figure 3: A demonstration of FNN (up) and CNN model (down)

Table 5 presents the performance metrics of four different machine learning models—MLP, CNN, LightGBM, and FNN—evaluating their training and validation accuracies and losses. Each row corresponds to a model, displaying its training accuracy and loss alongside its validation accuracy and loss. Overall, the table highlights that LightGBM significantly outperforms the other models, while MLP and CNN have similar performance levels, and FNN shows particularly weak results. Both the MLP and CNN models show relatively low training and validation accuracies (0.5364 and 0.5205 for training; 0.4926 and 0.5136 for validation), suggesting they may struggle with the dataset.

The second experiment is to use neural network model to predict the pattern of the matrix. In our example, we use Anthony Davis from Los Angeles Lakers to simulate the situation off player trade in substitution. When Anthony Davis is traded to Cleveland CLE and the coach change the lineup R1. Consider the team chemistry of the remaining four unaffected players remind the same and therefore the value of the corresponding column engagement matrix should remain unaffected.The classical economic model suggest that under the diminishing marginal return over concentration over one player will cause decrease in the marginal product.

$$[0.0101, 0.04332, 0.0111, 0.0619, 0.06795], RMSE : 0.4547$$

In our case ”LAL_Davis,” to a basketball lineup R1, (CLE_Allen CLE_Garland CLE_Mitchell CLE_Mobley) using a neural network. The Root Mean Square Error (RMSE) suggests a moderate level of prediction error. It begins by loading and prepossessing historical performance matrices and duration data for various player combinations. Matrices are normalized based on duration and padded to a uniform size to ensure consistency. The player combinations are then encoded into feature vectors using one-hot encoding. The neural network model is defined, compiled, and trained on the preprocessed data, with the goal of learning to predict performance matrices for given player combinations. And because of that, such matrix help fans better understanding dynamics helps coaches make strategic decisions during games, such as matchup against opposing players or adjusting lineups based on performance. However, this data only simulate the ball received by Davis, Davis may have also affected other existing player by giving assist.

4 Conclusion

This paper introduce only very abstract ideas for a future research proposal. I have discussed the some technical and statistical limitation. To extend the porposal. One thing potentially do in later is to access detail ball movment data and compute the shooting accuracy instead of frequency. Do passing necessarily lead to better shooting accuracy? In NBA, peer effects can manifest in several ways (*2*). One the most important. Performance Enhancement: Players may perform better when surrounded by high-performing teammates. This can be due to increased motivation, improved chemistry, and collaborative strategies that maximize individual strengths. Our framework provides a theoretical approach that allows labor economists to examine the value of teamwork using deep learning (*4*). Such methods could be alternative used for other sports that are teamwork intensive. For economic studies, an example could be using the matrix to examine the value of partnership in research, for example writing a good quality paper together may not be economically optimal (compared to solo publishing). In the future course on ECON4140, I would like to apply such estimation with field experiment data. For example, the result of the signing group project and co-authorship, or the partnership in companies as a potential topic financial economics.

Moreover, such analysis do not consider the defensive factors. Player behavior in assist mainly comes from the pressure of the oppose side. Historically, teams with robust defensive stats tend to have a higher impact in game. Meanwhile the another method to compute marginal product of players is also to compute how effective can that player reduce their opposing team scoring. Ultimately, a comprehensive analysis of defensive factors enhances game strategies, player evaluations, and overall fan engagement.

References and Notes

1. P. Arcidiacono, J. Kinsler, J. Price, Productivity Spillovers in Team Production: Evidence from Professional Basketball. *Journal of Political Economy* **124** (4), 1142–1170 (2016), doi:10.1086/687529, https://www.journals.uchicago.edu/doi/full/10.1086/687529.

2. F. Thabtah, L. Zhang, N. Abdelhamid, NBA Game Result Prediction Using Feature Analysis and Machine Learning. *Annals of Data Science* **6**, 103–116 (2019), doi:10.1007/s40745-018-00189-x, https://link.springer.com/article/10.1007/s40745-018-00189-x.

3. X. Zhu, *et al.*, The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity. *International Journal of Sports Science & Coaching* **16** (3), 485–500 (2021), doi:10.1080/24751839.2021.1977066, https://www.tandfonline.com/doi/full/10.1080/24751839.2021.1977066.

4. P. Arcidiacono, J. Foster, N. Goodpaster, J. Kinsler, Estimating spillovers using panel data, with an application to the classroom. *Quantitative Economics* **3** (3), 421–470 (2012), doi:10.3982/QE145.

5 Appendix

*

Token	Duration	Team	Players
R257	68656.0	OKC	Dort, Giddey, Gilgeous-Alexander, Holmgren, Williams
R1628	57695.0	DEN	Caldwell-Pope, Gordon, Jokić, Murray, Porter
R143	56238.0	WAS	Avdija, Gafford, Jones, Kuzma, Poole
R213	55281.0	MIN	Conley, Edwards, Gobert, McDaniels, Towns
R179	55061.0	HOU	Brooks, Green, Sengun, Smith, VanVleet

Table 1: Top 5 Tokens with the Longest Durations

	CLE_Allen	CLE_Garland	CLE_Mitchell	CLE_Mobley	CLE_Strus
CLE_Allen	0.142930	0.081674	0.081674	0.102093	0.061256
CLE_Garland	0.367534	0.306279	0.102093	0.326697	0.061256
CLE_Mitchell	0.224604	0.183767	0.530883	0.163349	0.122511
CLE_Mobley	0.122511	0.061256	0.000000	0.285860	0.102093
CLE_Strus	0.204186	0.102093	0.102093	0.102093	0.040837

Table 2: Matrix Representation

Statistic	**Value**
Mean	0.114340
Median	0.000000
SD	0.298511

Table 3: Summary Statistics of Effective scoring frequency

Token	**Effective Shoot**	**Player Set**
R4263	15.09933775	Oubre, Embiid, Harris, Beverley, Melton
R6347	14.82758621	George, Westbrook, Leonard, Mann, Zubac
R157	13.97210431	Bridges, Dinwiddie, Finney-Smith, Johnson, Sharpe
R489	13.74823197	Banchero, Bitadze, Anthony, Harris, Ingles
R9350	13.58490566	Bagley, Jones, Kispert, Kuzma, Poole

Table 4: Effective Shooting with per 12 mins

Model	Train Accuracy	Train Loss	Val Accuracy	Val Loss
MLP	0.5364	0.6854	0.4926	0.7028
CNN	0.5205	0.7237	0.5136	0.7235
LightGBM	0.9074	0.0720	0.9105	0.0699
FNN	0.1745	0.0704	0.1808	0.0741

Table 5: Train and Accuracy result from different Machine learning Model