NBA team analysis using Hidden-Markov-Model

Hongye Li, Xue Sheng, Tianyuan Gu
hyeli@ucdavis.edu, xuesheng@ucdavis.edu, ttgu@ucdavis.edu

Abstract

In this paper, we analyzed the features that will influence an NBA team's game results W/L (i.e. win or lose). We implement Hidden-Markov-Model to achieve our goal of game prediction, transition between win and lose estimation, and W/L features analysis. We mainly try to answer these questions by 04-05 season NBA game dataset. We found that it's hard to predict well during that time due to the even transition probability distribution and game style. We conclude that 04-05 season emphasized more on defense which corresponds to the truth. In the end, we predict today's second NBA Finals game between GSW and BOS, and get the result that GSW will win today's game. We will verify this later. We also propose the features each team should perform well for winning in *Section 5*. It suggests that GSW need to pay attention to rebounds, while BOS need to strive on 3 points shooting and defense. Finally, we discuss the potential study direction in the future.

1 Introduction

As data analysis becomes more and more popular today, data driven evaluation starts to show its advantages to team analysis. Compared with classic personal experience analysis, data driven evaluation can provide more solid and logical supports to the results of evaluation. The search results of 'NBA data analysis project' returns a bunch of related projects, which shows how popular it is in the NBA industry. For instance, Houston Rockets is the pioneer in data analysis to its players and team in the NBA League. Data analysis in sports can chase back to a long time ago. The famous movie 'Moneyball' which is based on the 2003 nonfiction book by Michael Lewis, is an account of the Oakland Athletics baseball team's 2002 season and their general manager Billy Beane's attempts to assemble a competitive team under limited budget . The assistant general manager helped Beane to search for latent 'golden' players in the free market by data analysis and finally got a big win. These are all the reasons why we prefer to spend effort on data analysis to the NBA teams in this paper.

As the technology develops, the NBA League and team managers focus more on data than ever, such as creating new measuring features to evaluate players' behavior during the game, hiring data scientists for prediction, evaluation and analysis for the game. The data analysis in the NBA can fall into many different detailed sections of the players and games. For this project, based on the great work of Vashisht Madhavan (2017) and the exciting NBA Finals, it gains our attention to find the features that can influence W/L (i.e win or lose) of a game. By these features, we may evaluate which team has a better chance to win. Another topic we are interested in is about the streak. There are many latent features we cannot quantify for a team, such as morale, emotion, chemical reaction, etc. These features may influence the team to win a streak, which is pretty important in a series games during the Finals. Although, we cannot get those features, we can estimate the transition probability of W/L instead. We are also interested in predicting the games with no labels. Without labels, we can regard the predicting procedure as an unsupervised classification problem, so it can have different explanations. For illustration, we will regard the prediction is about W/L. Finally, we will implement HMM to this year's playoffs to compare the game performance trend in the NBA. As a big fan of NBA, we decide to focus on these topics in this paper.

2 Problem and Related Work

As mentioned above, we are interested in the things related to W/L in this paper. In summary, our attentions mainly falls on the following questions:

- Which are the features that decide W/L of a game?
- What are the probabilities of transition between win and lose?
- How is the prediction accuracy of the model? How to explain it?
- How do basketball games change in the NBA during different eras?

Our idea for this project comes from the work of Vashisht Madhavan (2017), who compared supervised and unsupervised methods for predicting NBA games, such as HMM, GMM, SVM, etc. Bernard Loeffelholz, et al. (2009) examined the use of neural networks for predicting the win of games in the NBA, and got an accuracy of winning games 74.33 percent, which is higher than the experts' prediction. Cheng Ge, et al. (2016) proposed a method based on maximum entropy principle to construct an NBA Maximum Entropy (NBAME) model, and got a 74.4 percent accuracy for winning teams. In Daniel Jurafsky and James H. Martin's book (2021), it presents a detailed and clear explanation of Hidden-Markov-Model and the algorithms to solve this problem. In view of the above works, we focus more on data science related topics, and try to propose reasonable explanations of our findings.

The data we use in this paper is from BASKETBALL REFERENCE, a website provides tons of sports data and a great place for preparation for data analysis projects, since it is always the most difficult step for data analysis. In this paper, we mainly use 04-05 regular season data for analysis. Later, we will use 21-22 season playoffs data for further understanding of game performance changes in the NBA. Our data contains a **features dataset** and a **ground truth dataset**.

- **features dataset**: It is a combined team-based features data of all NBA teams. It has 27 features (8 home team advanced features, 4 home team offensive features, 4 home team defensive features, 1 home court feature, 1 home team value feature, 8 opponent advanced features, and 1 opponent value feature, explanation of the features are in the **Appendix**), which captures the performance measurements of both the home team and the opponent team. The home team value and opponent value feature are calculated by the BPM factor of each player in the team. In addition, the **features dataset** is standardized.
- **ground truth dataset**: It contains the W/L data of different teams in the League with encoding $win = 1,\ lose = 0$.

The features **home team value** and **opponent value** are calculated by a factor called Box Plus-Minus (BPM). It is developed by Daniel Myers (2020), who defines it as a basketball box score-based metric that estimates a basketball player's contribution to the team when that player is on the court. It is based only on the information in the traditional basketball box score–no play-by-play data or non-traditional box score data (like dunks or deflections) are included. Figure 1 shows the **home team value** of each team in the NBA during 04-05 regular season.

The result in Figure 1 corresponds with the true team value during the same time, such as the **Spurs** who won the championship has the highest team value.

Before we get deeper of the project, we wants to have a preliminary insight of the data. By extracting the winning teams' features data, we calculated the mean of each feature of 04-05 regular season and got Figure 2. In this way, we can have a glance of feasible influential features to wins.

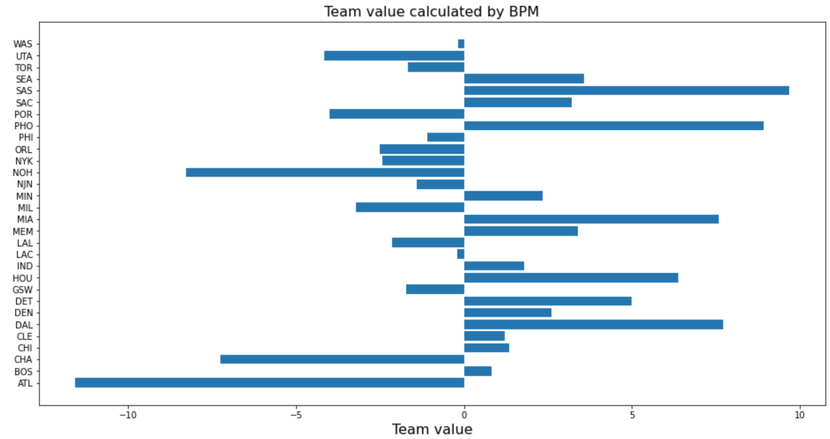


Figure 1: Team values

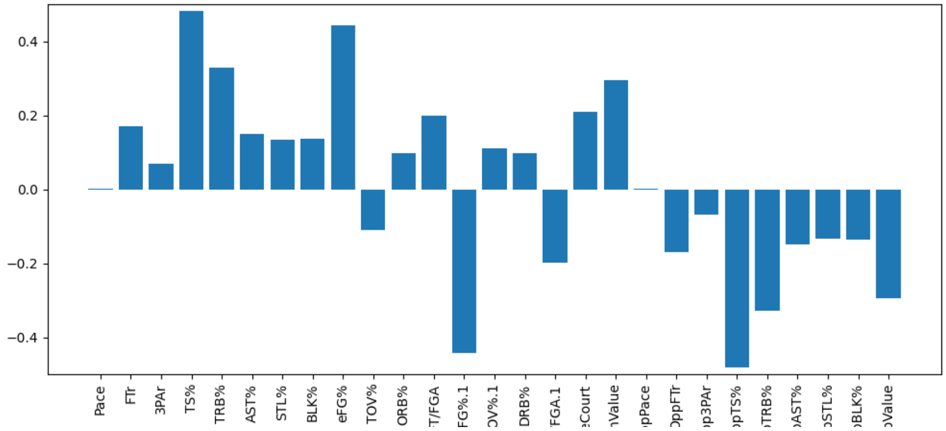


Figure 2: An image of a galaxy

Based on Figure 2, we notice that the features with relatively higher positive values are *TS%*, *TRB%*, *eFG%*, *HomeCourt*, and *TeamValue*, and those with negative values are *FG%.1*, *OppTS%*, *OppTRB%*, and *OppValue*. Where *TS%*, *TRB%*, *eFG%*, *HomeCourt* represents true shooting percentage, total rebound percentage, effective field goal percentage, and home team value respectively, and *FG%.1*, *OppTS%*, *OppTRB%*, and *OppValue* represents opponent effective field goal percentage, opponent true shooting percentage, opponent total rebound percentage, and opponent value respectively.

The result seems reasonable, as we know that back to 10's of 21st century, teams focus more on defense which can explain why we have higher value for rebound and lower value for opponent rebound and true shooting percentage. The winning teams seems have higher average team value, which is also reasonable. We also notice the winning teams have negative turnover percentage while forcing its opponent to have positive turnover rate. All above is about defense. The figure implies that winning team tends to perform better at offense too. Winning teams have high positive effective field goal percentage and true shooting percentage while limiting opponents effective field goal percentage and true shooting percentage much lower. From Figure 2, we got a result that winning teams have better performance at both offense and defense.

3 Hidden-Markov-Model for analyzing NBA teams

Based on the research by Vashisht Madhavan (2017), we think that Hidden-Markov-Model (HMM) fits our expectation of the project best. The main reason is that the model can do prediction and find the relations between the game measurements and the game results together. It may not be the most accurate model, since it is an unsupervised learning. However, it can provide us the answers for our questions mentioned above. In addition, the idea of the model also fits parts of our comprehension to the NBA games. Although there are many factors that will influence a game's result, the previous game may have a big influence to team's behavior in the next game, especially for a series like the playoffs. The team morale, mental state and even physical condition could be better for next game after a win before. That's why we concern about the probability of transition between win and lose. It can not only help us explain the predictions, but also give us a preview of the forthcoming game. Another reason is that we want to find the most influential features to the game results. The idea of emission matrix of HMM which represents the probability of observations given the "hidden states" can provide us which features are important to a game. It can help us understand the game with reasonable quantity support.

The HMM is a Markov model in which the system is being modeled is assumed to be a Markov process (called "hidden" states). As what mentioned from an influential tutorial by Rabiner (1989), our job is to solve the three fundamental problems of HMM, which corresponds to find the likelihood, parameters and the best "hidden" state sequence. In the model, we denote the components as follows.

Table 1: Model components

π_0	initial probability, with the constraint $\sum \pi_0 = 1$
$S = 1, 0$	a set of N states, where 1 represents win and 0 represents lose
$O = o_t,\ t = 1, \ldots, T$	a sequence of T observations
$A = a_{ij},\ i = 1, 2,\ j = 1, 2$	**transition matrix**, where a_{ij} represents the probability of moving from state i to j , with the constraint $\sum_{j=1}^{N} a_{ij} = 1$
$E = e_j(o_t),\ j = 1, 2,\ o_t = t - th\ observation$	**emission matrix**, where $e_j(o_t)$ represents the probability of observation o_t being viewed given state j

Before implementing the model, we must make some assumptions which fit our prior comprehension of the NBA games. Like the Markov chain assumption, we assume that the upcoming game result is only due to the present game result, which can be written as $p(S_t|S_{t-1}) = p(S_t|S_{t-1}, \ldots, S_0)$, where S_t represents the state at time t. We also need to acknowledge that the observations are identical and independent with each other. Therefore, we can deduce that $p(o_t|S_1, \ldots, S_T, o_1, \ldots, o_T) = p(o_t|S_t)$, which means that the observation only depends on the state at time t. Furthermore, we assume the **emission matrix** B of our problem follows multivariate normal distribution for easy understanding and calculation. Therefore, our goal is to predict the future states, and estimate parameters (A, E). For estimating parameters (A, E), we use **forward-backward algorithm** proposed by Baum (1972), which is a special case of EM algorithm. The algorithm can be implemented as follows.

We use the package *hmmlearn* to estimate the parameters A, E, and predict the game results in this paper. The default algorithm of the package is as above.

4 Data analysis

4.1 Train and prediction results

To utilize the Hidden-Markov-Model, we fit the model with selected features from three combinations of teams during 04-05 regular season. First we try twenty-five teams data out of all 30 teams and treat

```
By the **forward-backward algorithm**, we start the iteration given the initialization and
terminate the iteration until convergence.
**Given initialization:** $\pi_0,\ A,\ B$
**Iteration:** until convergence
    E-step:
        A$_{ij} = \frac{\alpha_t(i)a_{ij}e_j(o_{t+1})\beta_{t+1}(j)}{\sum \alpha_t(j)\beta_t(j)}$
        E$_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\sum \alpha_t(j)\beta_t(j)}$
        where $\alpha_t(j) = p(o_1, \ldots, o_t, S_t = j|A, E),\ \beta_t(i) = p(o_{t+1}, \ldots, o_T|S_t = i, A, E)$
    M-step:
        a$_{ij} = \frac{A_{ij}}{\sum_k A_{ik}}$
        e$_j(o_t) = \frac{E_t(j)}{\sum_t E_t(j)}$
**return:** A, E
```

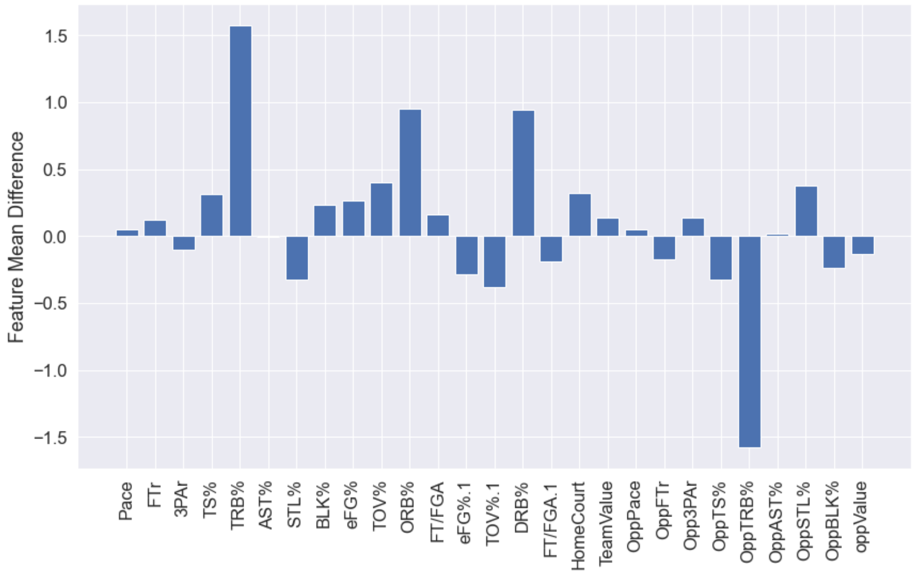
each team's games as a sequence. Then we predict all thirty teams' Win/Lose results and report the accuracy rate of predicting the result for each team and each game. For each team, we also explore the true positive(win) rate and true negative(lose) rate of the prediction. To compare different train sets, we involve other two combinations, three teams with most won games and three teams with least won games. Their accuracy rates of prediction are tested on all 30 teams' sequential games.

Table 2: Accuracy of HMM Prediction

Train Set	Accuracy on all teams
25 teams out of 30	63.29%
3 teams with most won games	54.80%
3 teams with least won games	61.26%

Feature means, or known as a projection of emission matrix from fitting models can present the influence of each feature on the Win/Lose result. In Figure 3, the differences between features means of win and lose prediction give us an immediate understanding of their influences on the game results, involving whether a certain feature is strengthening the probability of winning and how important the effect of a certain feature is on the probability of winning or losing.

Figure 3: Feature Mean Difference of Predicted Win over Lose



Within the win games of each team, the true positive and false negative games are reported in Figure 4. The aim of this result is to explore how the prediction performs on each team's won games. From the result, teams' won games are split into count of predicted winning and predicted losing in clear visualization. Most teams have true positive rates over 50%. In three teams with most won games, two of them have rather high true positive rates.

Figure 4: True Positive and False Negative in Win Games of Each Team

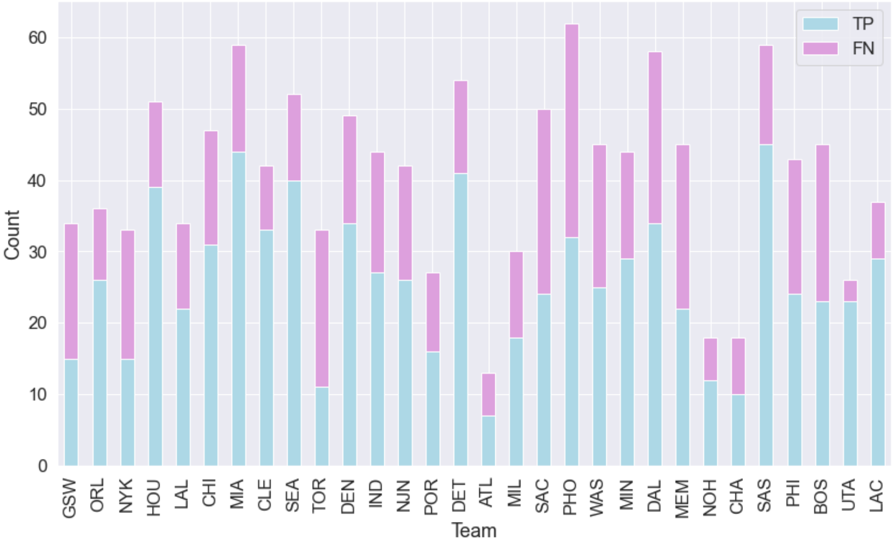


Table 3: Transition Matrix of 25 teams out of 30

	Lose	Win
Lose	0.53	0.47
Win	0.49	0.51

Table 4: Transition Matrix of best 3 teams

	Lose	Win
Lose	0.49	0.51
Win	0.51	0.49

Table 5: Transition Matrix of worst 3 teams

	Lose	Win
Lose	0.46	0.54
Win	0.44	0.56

Moreover, we want to study the transition matrix between W/L situation, in order to explore the effect of consecutive win or lose. Conveniently, the transition matrix of HMM using different combinations of train sets are effective in answering this question. The effect of consecutive winning or losing seems not so significant in 2004-2005 games. The results we got are shown in *Table 3*, *Table 4* , and *Table 5*.

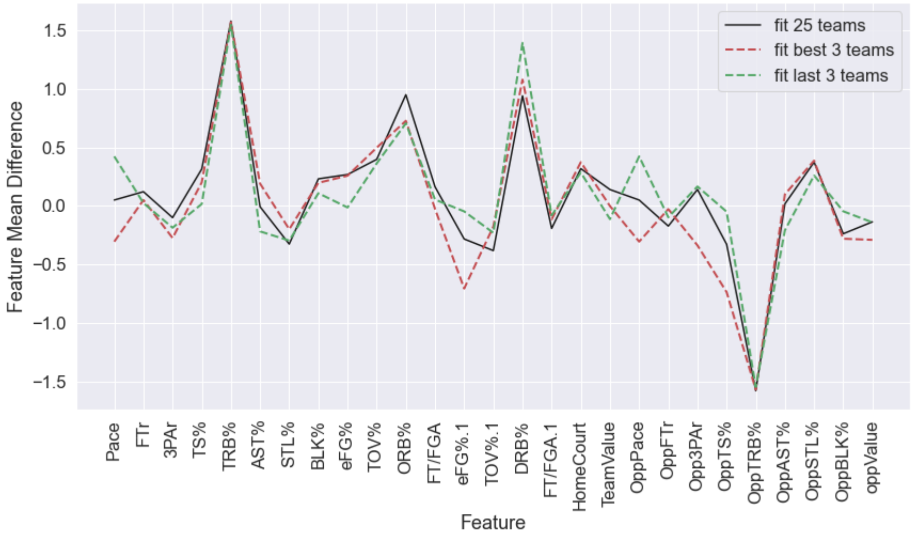
The model is fitted onto different combinations of train teams produce three sets of feature mean difference. Figure 5 presents where these fitting sets differ from each other. From this figure, all three fitting sets in year 2004-2005 show similar distributions among these features. However, there are some patterns of selected features hidden in best/last teams, where the difference of best/last jump up and down around the twenty-five teams fitting.

4.2 Analysis on the results

4.2.1 Prediction analysis

The prediction results in *Table 2* show that our prediction accuracy is obvious not good, which is reasonable. As we mentioned above, the model is an unsupervised learning that cannot guarantee for a high prediction accuracy. Although we do so bad at accuracy, we want to get deeper of the prediction section to see what may influence the accuracy. The first thing comes to our mind is that it may caused by the "streak phenomenon", which means that due to HMM assumption, the results are highly correlated with the previous information. Therefore, the model may be good at predicting winning streak or losing streak, while performs bad at predicting win or lose changes.

Figure 5: Feature Mean Difference(Win - Lose) from Fitting Different Train Set



For further study, we dive to see each teams prediction results to get a deeper understanding. According to Figure 4, the model generates a pretty good result. For the 2004-2005 season, the top 6 regular season ranking was: PHO(Phoenix Suns), SAS(San Antonio Spurs), MIA(Miami Heat), DAL(Dallas Mavericks), DET(Detroit Pistons), SEA(Seattle Supersonics). The Top 6 teams generated by our prediction is SAS, MIA, DET, SEA, HOU(Houston Rockets), DAL. HOU actually ranked 7th in that season. The team that finished last in that season is CHA(Charlotte Bobcats),NOH(New Orleans Hornets),ATL(Atlanta Hawks) which is consistent with the prediction outcome. Therefore, it can be seen that for very high ranking and very low ranking teams, the model has a good predictive power at the total W/L games.

4.2.2 Transition analysis

From *Table 3*, *Table 4* , and *Table 5*, we notice that during 04-05 regular season, the transition probabilities fluctuate slightly around 0.5 for teams from all levels. It implies that winning a streak may not be easy during that time. This may help to explain what we found in the *Prediction analysis* section. Because we suppose HMM is good at predicting streaks, while the transition probabilities tell us it's difficult to win a streak during the season. For our intuition, it may be caused by the emphasis of defense during that time. This suggestion can be verified in the next section.

4.2.3 Features analysis

According to the result in Figure 5, TRB is the most important factor affecting the game in 2004-2005 season. Correspondingly, ORB and DRB become big factors. TOV also affects the prediction of outcome because turnovers are more likely to be converted into points by opponents. The fact that 3PAr has absolutely no effect on the game indicates that in 2004-05, long-range shooting was not a key factor and players preferred physical plays in the post(paint). The influence of FT/FGA seems to confirm this conclusion since players are more likely to get free throws from collisions in the paint than outside . Surprisingly, TS has less impact on the game than TOV or even FT/FGA. This is perhaps further evidence that the scale and intensity of refereeing was different in 2004-05 from nowadays because in more intense rivalry, the TS between teams may not be high. At the same time, the team with a higher TRB can ensure more offensive opportunities and thus score more points. In addition, an interesting find is that the *best 3 teams* have lower **Pace** feature than the *last 3 teams*, which indicates that low pace game is mainstream and encouraged during that time.

5 21-22 NBA Finals analysis

As we are writing the report, the second game of the Finals starts. After analyzing teams in 2004-05 season , we want to analyze the two teams in this year's(2021-2022) Finals which are GSW(Golden State Warriors) and BOS(Boston Celtics). Due to the prediction results from our model, we predict that Golden State Warriors will win today's game and we can verify this 3 hours later. The result seems meaningful, because GSW has home court advantage and they cannot lose another home court game anymore. GSW will pour out all their strengh to win this game, or they will have high probability to lose the whole series.

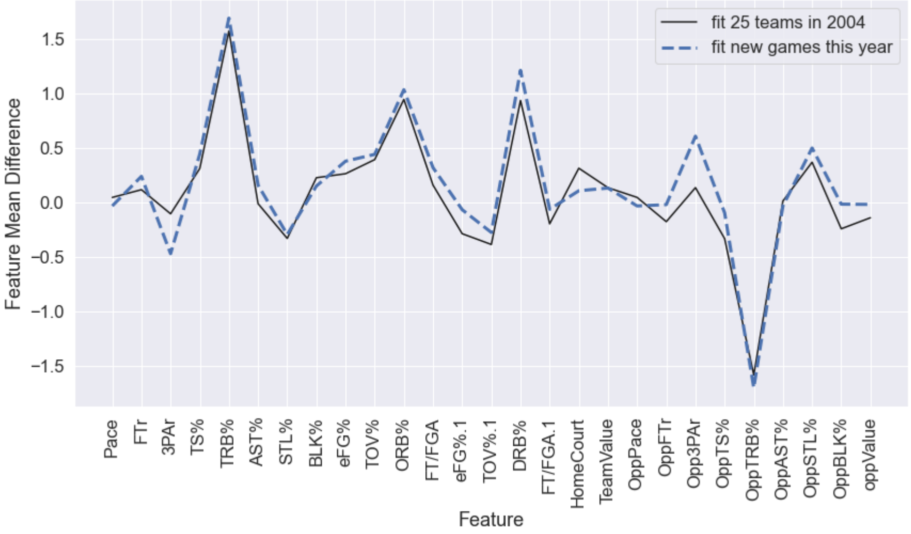
In the process of predicting today's game, we found something interesting. The first is the transition matrix in *Table 6*, which tells us that it tends to win or lose a streak easier nowadays than 04-05 season. This phenomenon can be explained by the different game styles between these two eras. The old days emphasize more on defense. As a consequence, every game seems very intense and every team seems to have a chance to win during the game. However, today's teams focus more on offense, which improve the viewing experience of the game, but the game may lose intensity.

Table 6: Transition Matrix of new games this year

	Lose	Win
Lose	0.60	0.40
Win	0.29	0.71

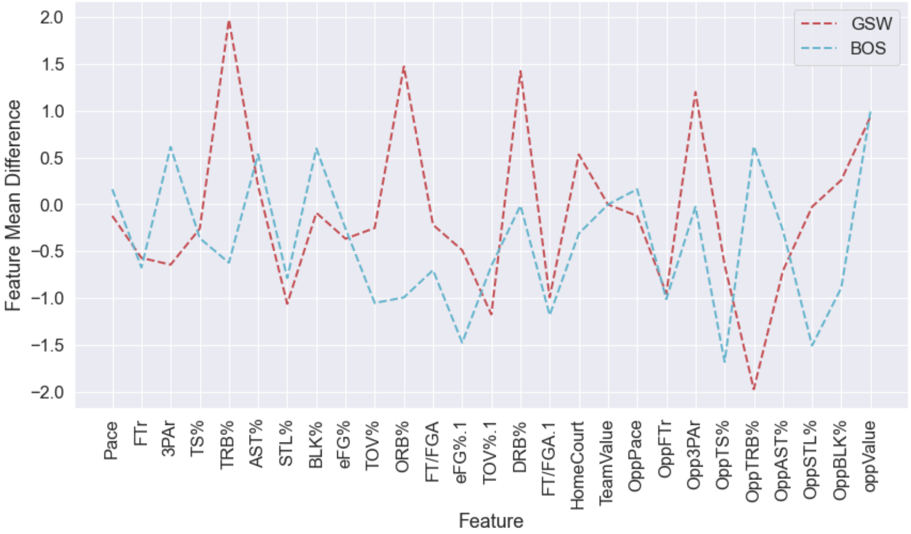
Comparing with 04-05 season, it's not hard to see that TRB is always a deciding factor in any periods from Figure 6, even it's 21-22 playoffs. In addition, the biggest change among all features is 3PAr. As the increase number of offensive rounds, Teams need more 3PAr and less Opp3PAr to win which indicates the style changing of NBA nowadays. Compared with before, whether the game happened at homecourt becomes less important. This is an interesting finding. One possible explanation is that the overall style of shooting makes the players 'calm' both home and away.

Figure 6: Feature Mean Difference(Win - Lose) from Fitting Different Period Set



We compare the influential features to W/L of both GSW and BOS in Figure 7. According to the stats, the Warriors need to control TRB better, so as ORB and DRB, which can be explained that the Warriors is a shooting preferred team. To our surprise, the 3 points shooting seems to have a higher weight for BOS to win than GSW. To recap, for GSW to win, they need to protect and strive for rebounds, together with better using their home court advantage. On the other hand, for BOS to win, they need to try more 3 points shooting, keep passing the ball, hold their defense, and lower their turnovers. The one who performs better at these features may have a higher probability to win.

Figure 7: Feature Mean Difference(Win - Lose) from Fitting Two Teams This Year



6 Conclusion and Discussion

From **Section 4**, we found that although our prediction for every game is not accurate enough, the prediction of number of win games is pretty good. This may be caused by HMM's core assumption that it's better for predicting a "chain" like games, and the time during 04-05 season had more intense game style that makes the transitions between win and lose have nearly the same probability. The features analysis implies that it was a defense emphasized era, which also corresponds to what we found from the prediction and transition probabilities. Then we predict this year's playoffs as a comparison to the previous season in **Section 5**. The results indicates that teams are easier to win or lose a streak in nowadays, because the diagonal values in transition matrix is much higher than before. By comparing the features plot, we found that game style changes from defense to emphasis of offense today, which corresponds to people's commonsense. Finally, we compare the main characters of this year's NBA Finals participants, GSW and BOS. We conclude that GSW should pay more attention on rebounds, and use their home court advantage for today's game. While BOS should try more 3 points shooting, keep passing the ball, hold their defense, and lower turnovers for victory. Our prediction for today is that GSW will win the game. We will know this 2 hours later. Let's verify it.

For further study, we want to pay more attention on the features learning. In this paper, we made very strong assumptions to the problem which is not robust enough, and may cause many mistakes in the process. We may need loose the assumptions to see what will happen. Since the HMM is an unsupervised learning method, we need to implement some supervised learning method to see if we can improve the prediction accuracy. Another point we need to pay attention to is that the HMM is being modeled as a Markov process, which only depends on the previous information. However, the basketball game is a long time battle from time to time. We may need to include more previous information and the latest information to get a clearer view of the essence of the game in the future.

References

Madhavan, V. (2017). *Predicting NBA Game Outcomes with Hidden Markov Models.*

Loeffelholz, B., Bednar, E., Bauer, K. (2009) *Predicting NBA Games Using Neural Networks*, Journal of Quantitative Analysis in Sports, 5(1). https://doi.org/10.2202/1559-0410.1156

Cheng Ge, Zhenyu Zhang, Moses N. Kyebambe, and Nasser Kimbugwe. (2016). *Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle*, Entropy 18(12): 450. https://doi.org/10.3390/e18120450

Daniel Jurafsky and James H. Martin. *Appendix Chapters A: Hidden Markov Models*, Speech and Language Processing, 2021. https://web.stanford.edu/ jurafsky/slp3/A.pdf

Daniel Myers, developer of Box Plus/Minus. (2020). *About Box Plus/Minus (BPM)*, BASKETBALL REFERENCE. https://www.basketball-reference.com/about/bpm2.html

Rabiner, L. R. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, 77(2):257–286.

Baum, L. E. (1972). *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*. Inequalities III: Proceedings of the 3rd Symposium on Inequalities. Academic Press.

Appendix

Advanced Factors:

- **Pace** – Pace Factor: An estimate of possessions per 48 minutes
- **FTr** – Free Throw Attempt Rate
- **3PAr** – 3-Point Attempt Rate
- **TS%** – True Shooting Percentage
- **TRB%** – Total Rebound Percentage
- **AST%** – Assist Percentage
- **STL%** – Steal Percentage
- **BLK%** – Block Percentage

Offensive Four Factors:

- **eFG%** – Effective Field Goal Percentage
- **TOV%** – Turnover Percentage
- **ORB%** – Offensive Rebound Percentage
- **FT/FGA** – Free Throws Per Field Goal Attempt

Defensive Four Factors:

- **eFG%.1** – Opponent Effective Field Goal Percentage
- **TOV%.1** – Opponent Turnover Percentage
- **DRB%** – Defensive Rebound Percentage
- **FT/FGA.1** – Opponent Free Throws Per Field Goal Attempt

Team Factors:

- **HomeCourt** – Home Court vector
- **TeamValue** – Team value calculated by players' BPM

Opponent Factors:

- **OppPace** – Opponent Pace Factor: An estimate of possessions per 48 minutes
- **OppFTr** – Opponent Free Throw Attempt Rate
- **Opp3PAr** – Opponent 3-Point Attempt Rate
- **OppTS%** – Opponent True Shooting Percentage
- **OppTRB%** – Opponent Total Rebound Percentage
- **OppAST%** – Opponent Assist Percentage
- **OppSTL%** – Opponent Steal Percentage
- **OppBLK%** – Opponent Block Percentage
- **TeamValue** – Opponent team value calculated by players' BPM