Mason: Real-time NBA Matches Outcome Prediction

by

Rongyu Lin

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved March 2017 by the
Graduate Supervisory Committee:

Hanghang Tong, Chair
Jingrui He
Huan Liu

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

The National Basketball Association (NBA) is the most popular basketball league in the world. The world-wide mighty high popularity to the league leads to large amount of interesting and challenging research problems. Among them, predicting the outcome of an upcoming NBA match between two specific teams according to their historical data is especially attractive. With rapid development of machine learning techniques, it opens the door to examine the correlation between statistical data and outcome of matches. However, existing methods typically make predictions before game starts. In-game prediction, or real-time prediction, has not yet been sufficiently studied. During a match, data are cumulatively generated, and with the accumulation, data become more comprehensive and potentially embrace more predictive power, so that prediction accuracy may dynamically increase with a match goes on. In this study, I design game-level and player-level features based on real-time data of NBA matches and apply a machine learning model to investigate the possibility and characteristics of using real-time prediction in NBA matches.

DEDICATION

To the past and future me

To my beloved family and friends

The next episode awaits

ACKNOWLEDGMENTS

First and foremost, I would like to thank my thesis advisor Dr. Hanghang Tong. Without your patient instruction and dedicated involvement in every step throughout the process, this thesis would have never been accomplished. No words can fully express my thanks for your guidance and understanding over the past two years. It is my great honor to be your student.

I would also like to show gratitude to Dr. Jingrui He and Dr. Huan Liu. Thank you for giving me this precious chance to work with you on my thesis. I am gratefully indebted to you for your valuable comments on this thesis.

I must also give my thanks to some members of DATA Lab, STAR Lab and DMML Lab, Mr. Liangyue Li, Ms. Chen Chen, Mr. Si Zhang, Mr. Boxin Du, Mr. Yao Zhou, Mr. Dawei Zhou, Mr. Xu Liu and Miss Kewei Cheng. Even though I have not had the opportunity to work with all of you, the impact of your academic advice and work on my own study is obvious throughout this thesis.

Getting through this journey required more than academic support, and I have many, many people to thank for listening to and, at times, having to tolerate me over the past three years. I can not hold to express my gratitude and appreciation for their friendship. Mr. Luning Zhao, Mr. Yuchen Shang, Mr. Bojun Sun, Miss Tianyi Wang, Mr. Jian Shang, Mr. Xiaoyu Zhang and Mr. Fan Hong have been unwavering in their personal support during the time I spent at Arizona State University. There is no way I could overcome all difficulties without your persistent help. For many memorable evenings out and in, I must also thank Mr. Pak Lun Kevin Ding, Miss Ting Yan Betrica Fok, Mrs. Haiqin Lin, Mr. Chenyang Li, Mr. Tong Zhou, Mr. Enlai Zhang, Mr. Lin Luo, Mrs. Si Wang, Mrs. Hanshuang Liang who always opened their heart to me whenever I needed someone to talk to. It is my great fortune to have met you along this journey.

Finally, most importantly, none of this could have happened without my family. My parents and grandparents, who offered both their financial support and endless encouragement through these years. Every second, no matter happy or sad, cheerful or depressed, knowing that you will always be there for me, I was able to stand on high mountains and walk on stormy seas. This thesis stands as a testament to your unconditional love and encouragement.

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

The National Basketball Association (NBA) [1] is the most popular basketball league in the world. Each year, 30 teams in the league plays against each other in different kinds of matches, including pre-season, season regular, playoff and the finals. Recent years some teams also played with international teams or clubs in pre-season matches, seeking to popularize both the league and basketball all over the world. The world-wide mighty high popularity to the league leads to large amount of interesting and challenging research problems, such as team tactics, league marketing impact, player trading, draft and charity. Among them, predicting the outcome of an upcoming NBA game between two specific teams according to their historical data is especially attractive. With rapid development of machine learning and data mining techniques nowadays, it opens the door to examine the correlation between statistical data and outcome of matches.

From data mining perspective, match outcome prediction is mainly determined by two factors, feature design and prediction algorithm. Existing studies have made many attempts to both factors. When designing features, previous work have tried taking into account traditional box statistics, home advantage, coach, odds, player injury and so on. Machine learning models such as support vector machine, hidden Markov model were used as prediction algorithms.

Despite satisfying results, to the best of our knowledge, these existing methods typically make predictions before the game starts. In-game prediction, or to say, real-time prediction, has not yet been sufficiently studied. During a match, data are

[1] http://www.nba.com/

cumulatively generated, and with the accumulation, data become more comprehensive and potentially embrace more predictive power, so that prediction accuracy may dynamically increase with a match goes on. Besides, existing studies designed complex features and prediction algorithms to embrace more prediction power. These features and algorithms are sometimes hard to understand.

In this study, we design game-level and player-level features based on real-time data of NBA matches in recent 5 seasons and apply a machine learning model to investigate the possibility and characteristics of utilizing real-time prediction in NBA matches. Meanwhile, we try to find if simple features and algorithms could also gain much prediction power.

The rest of this thesis is organized as follows. Chapter 2 reviews some related work. Chapter 3 introduces our methodology. Chapter 4 shows our experiments. Chapter 5 delivers the experimental results and analysis. Chapter 6 concludes the thesis and discusses future work.

Chapter 2

RELATED WORK

In this chapter, we review the related work in terms of basketball matches (not just NBA) outcome prediction. We also review some match prediction methods for other sports, including esports, as these methods for different sports may inspire research in basketball.

In order to predict basketball matches outcome, previous researchers either designed effective features for the match or invent new prediction algorithms. Some of them also tried both ways to improve prediction result. In early stage, researchers utilized individual statistics, like Melnick (2001), or used statistical analysis of team performance to understand relationship between outcome and features (Sampaio and Janeira (2003)). Zak *et al.* (1979) ranked individual teams by combining defensive and offensive elements. With machine learning and data mining techniques developing so fast in recent years, diverse machine learning models, such as logistic regression by Cox (1958), support vector machine by Cortes and Vapnik (1995) and neural networks by Minsky and Papert (1988), were applied by previous researchers such as Loeffelholz *et al.* (2009) according to their different input data or feature sets. Cao (2012) gave a comprehensive review of data mining techniques used in predicting outcomes of basketball matches. Kvam and Sokol (2006) invented LRMC method (logistic regression and Markov chain) for predicting National Collegiate Athletic Association (NCAA) basketball matches. As follow-up work, Brown *et al.* (2010) improved the method to bring better performance. Also focusing on NCAA, Lopez and Matthews (2015) attempted to quantify the degree of luck played in a game. Making use of homogeneous Markov model, Štrumbelj and Vračar (2012) were able to forcast outcome of a match by simulating the progression. Trawinski (2010) utilized fuzzy classification system to predict the Asociacin de Clubs de Baloncesto (ACB) league matches. In the same year, Miljković *et al.* (2010) used Naive Bayes method in predicting NBA season games, while Hu and Zidek (2004) and Wei (2011) focused more on playoffs exploiting special contextual features and naïve bayes algorithms, respectively. In Vaz de Melo *et al.* (2008), complex network metrics provided decent prediction without using box score statistics. Considering both individual performance and group cohesion, Berri (1999) first measured how individual players contribute to a team's success, and DeLong *et al.* (2013) designed a series of frameworks named TeamSkill and applied them to NBA season games.

Besides basketball, previous researchers also studied and excavated making prediction in other sports, both virtual world (esports) and real-world. Haghighat *et al.* (2013) briefly reviewed and analyzed data mining techniques used in predicting sports results. Although (e)sports like soccer, football, tennis and League of Legends [1] have different data structures and determining factors to basketball, methods used or created for predicting their outcomes may still inspire basketball. Leung and Joseph (2014) explored predicting US college football games with sports data mining approach. DeLong *et al.* (2011), inspired by Elo (1978), Glickman (1993) and Glickman (1999), modeled team chemistry with a series of frameworks named TeamSkill and DeLong and Srivastava (2012) implemented the framework to an on-line multi-player game, Halo 3. Chen and Joachims (2016a) presented a framework for predicting pairwise matchups, in which a model called BLADE-CHEST is utilized to represent one player. They then applied their method to both tennis (real-world) and Starcraft II [2] (virtual world) in Chen and Joachims (2016b). Min *et al.* (2008) proposed

[1] http://leagueoflegends.com/

[2] http://us.battle.net/sc2/en/

a compound framework combining Bayesian inference, rule-based reasoning and in-game time-series approach in predicting soccer matches. Same with Min et al., Rue and Salvesen (2000) and Aslan and Inceoglu (2007) tried to solve the problem with Markov chain Monte Carlo methods and neural network, respectively. Modeling football or soccer matches with multi-layer perceptron, McCabe (2002) and McCabe and Trevathan (2008) covered the prediction of four major league sports, including the Australian National Rugby League [3], the Australian Football League [4], Super Rugby [5] and English Premier League [6]. Also doing research on English Premier League, Langseth (2013) looks at statistical models for prediction of soccer matches.

[3] http://www.nrl.com/

[4] http://www.afl.com.au/

[5] http://www.sanzarrugby.com/superrugby/

[6] https://www.premierleague.com/

Chapter 3

METHODOLOGY

In this chapter, we present our methodology of making real-time NBA matches predictions. Before introducing our methods, we first present definition of real-time used in this study. Real-time prediction means that for a single match, we make one prediction every 2 minutes based on data generated from the beginning of the match (0th minute) to current time point. There are 48 minutes of regular time and 23 *in-game* time points in an NBA match (2nd, 4th, 6th,..., 46th minute), so we make 23 predictions with our methods for each match.

Existing studies only made predictions before a match starts, and thus they did not utilize any real-time data. During a match, data are cumulatively generated, and with the accumulation, data become more comprehensive and potentially embrace more predictive power, so that prediction accuracy may dynamically increase with a match goes on. On the other hand, previous studies with more prediction power typically designed complex features or training models, and we are curious if simple features and models could also bring decent prediction power. Based on the two aspects mentioned above, we propose our hypotheses and verify them with our method.

Our method contains two parts, corresponding to the two main factors that may influence match outcome prediction. In feature design part, we first design game-level and player-level features based on the data set individually, then combine them together to formulate a new feature set. And in training model part, we apply a machine learning model that is easy to understand to our feature sets.

3.1 Hypotheses

Hypothesis 1: Prediction accuracy dynamically increases with a match goes on if predictions are made with same feature set and same training model.

Hypothesis 2: Prediction power can be embraced with feature sets and models that are easily understood.

3.2 Feature Design

Feature design is the most critical part of making predictions. Quality of features may have great influence on final result. In this section, we first design our feature sets based on real-time data of NBA matches from two separate aspects, game-level and player-level. Then combine the two features sets together to formulate a new feature set that contains both game-level and player-level features.

3.2.1 Baseline Methods

We provide two baselines for our feature design, History Difference (H-Diff) and Present Difference (P-Diff). H-Diff is a simple pre-match prediction method, which is similar to Rote Learning in Chen and Joachims (2016b), considering only history records between teams and ignoring any other factors. P-Diff is a simple real-time prediction method, taking the most basic real-time game-level information into consideration.

History Difference (H-Diff)

H-Diff makes prediction before a match starts by comparing history records between two teams, the one with better history records is forcasted to win the upcoming match. Here, history record only contains games belong to previous seasons, which means records of the same season will not be included. For example, suppose we want to predict the outcome of third match between San Antonio Spurs and Houston Rockets in 2014-2015 season with H-Diff of 2 previous seasons, we only compare the winning record between Spurs and Rockets against each other in 2013-2014 and 2012-2013 season, the first and second match between two teams in 2014-2015 season will not be used in prediction, even though they are already history. This method does not make use of any game or player level information besides history records of recent seasons, nor contains any real-time information or involves with any learning models. Also, due to the characteristics of NBA league:

(1). frequent and sharp player changes each season;

(2). each pair of teams only meets each other at most 4 times per season,

results of H-Diff may vary significantly when choosing different number of seasons' history records for predicting.

Present Difference (P-Diff)

P-Diff contains the most basic game-level real-time information. It predicts match outcome according to points difference at current time point between two teams. Like H-Diff, P-Diff also does not utilize any game-level or player-level features and has no relationship with any training model. The only factor that effects prediction result is the points difference between two teams at current time point, and as leading team may change multiple times in a match, result of this method may vary as a game goes on. Take the match between Cleveland Cavaliers and Washington Wizards on Feb 7th, 2014 [1] as an example, the Wizards led by 3 at 8th minute in 1st quarter, so Wizards was predicted to win the game at the time point; however, when it came to 32th minute, Cavaliers took the lead and was forcasted to win this game.

[1] http://www.nba.com/

3.2.2 Game-level Features

To contain more real-time, game-level information, we present this approach, Recent X Differences (RX-Diff). We use a sliding window to include recent X differences of current time point as features of the match. Each points difference is taken as one feature. X stands for the length of sliding window, or the upper limit number of recent points differences we consider as features for the game, and since we have 23 in-game time points per game, X should not be greater than 23. For example, in Table 3.1, when X equals to 5, we have 1 feature for 2nd minute, since we only have one recent points difference (2nd minute); we have 4 features for 8th minute (8th, 6th, 4th and 2nd minute, successively) and 5 for 14th minute (14th, 12th, 10th, 8th and 6th minute, successively) and all time points after 10th minute. When X is greater than 5, we still have 1 and 4 features at 2nd and 8th minute, respectively, but have 7 for 14th minute and number of game-level features will still be increasing until it reaches X in one of the following time points. Figure 3.1 gives intuition of this feature design.

3.2.3 Player-level Features

To design features with real-time, player-level information, we present the approach Top K Stats (TPK). Different to RX-Diff, we have fixed number of features at all time points in TPK. There are 18 traditional box statistics in our data set. For each statistic of each team, we pick the highest K numbers at current time point to formulate K features, so number of features in this method is *18*K*2*. Take the match mentioned above as example, Table 3.2 shows part statistics of Cleveland Cavaliers at 12th minute. With different values of K, we have different feature sets based on the same data. Table 3.3 gives an intuition of this.

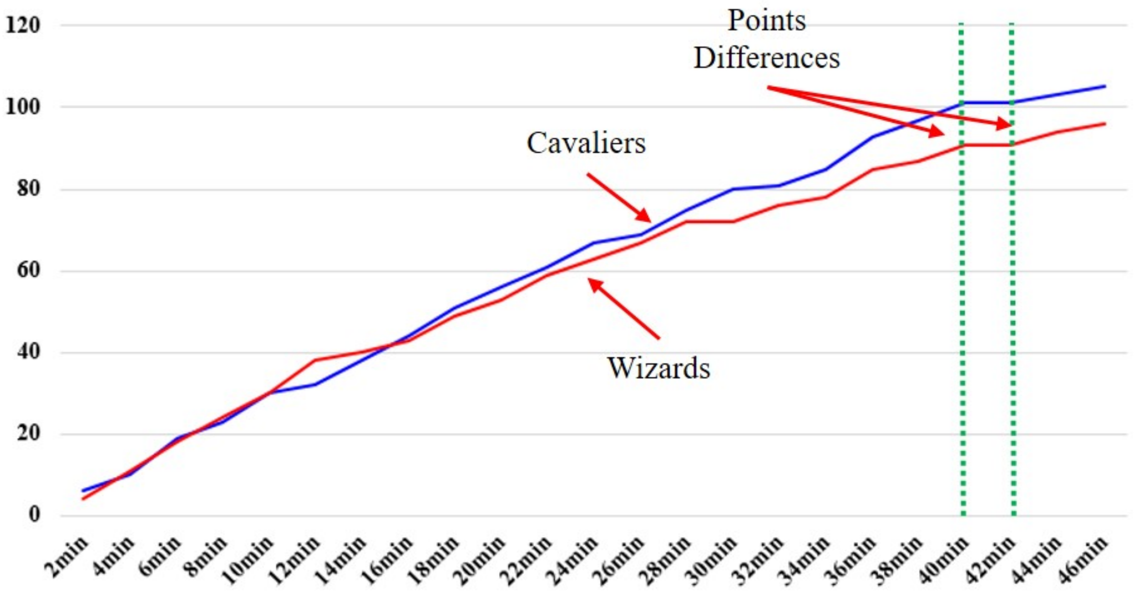


Figure 3.1: Points Change over Time in Match between Cleveland Cavaliers and Washington Wizards on Feb 7th, 2014

Table 3.1: Example of Game-level Features

X	Time Point	Game-level Features
X = 5	2nd minute	PD at 2min
	8th minute	PD at 8min, PD at 6min, PD at 4min, PD at 2min
	14th minute	PD at 14min, PD at 12min, PD at 10min, PD at 8min, PD at 6min
X = 10	2nd minute	PD at 2min
	8th minute	PD at 8min, PD at 6min, PD at 4min, PD at 2min
	14th minute	PD at 14min, PD at 12min, PD at 10min, PD at 8min, PD at 6min, PD at 4min, PD at 2min

In each game, we have player-level features of both teams and combine them together to formulate our feature set with home team on the left and away team on the right. Since there are at least 5 players that have played in one match at all time points (this usually happens at 2nd minute, when 5 starting lineup players are still on the court and substitutions have not appeared), K should be no more than 5.

Normalization

Traditional statistics has different scales of evaluation. For example, a player may score more than 40 points in a game, but can not commit more than 6 personal fouls. To eliminate the effect from different scales, we normalize the player-level features obtained to a 0-1 scale. For each statistic, we retrieve the maximum and minimum achieved in the data set, and map them to 0 and 1, respectively. All other numbers are mapped to the range of 0 and 1.

3.2.4 Game-level and Player-level Features

Since previous two approaches take real-time game-level and player-level information into consideration individually, we are curious if advantages of these two approaches could be complementary and disadvantages could be reduced when combined together. Thus comes this approach, Top K Stats + Recent X Difference (TPK-RX-Diff). In this approach, we combine feature sets of previous two approaches together to formulate a new feature set, with game-level features in the front and player-level features after. This new feature set contains at least *1+18*2*K* features and at most *X+18*2*K* features for each game at each time point, with X representing maximum number of recent points differences as game-level features and K for the highest K players' stats in every traditional statistic as player-level features. Table 3.4 shows comparison of feature sets with different Xs and Ks.

Table 3.2: Part Statistics of Cleveland Cavaliers at 12th minute of match against Washington Wizards on Feb 7, 2014

Player	**Rebounds**	**Assists**	**Points**
CJ Miles	0	0	12
Tristan Thompson	3	1	6
Anderson Varejao	3	1	2
Jarrett Jack	0	0	0
Kyrie Irving	6	1	6
Dion Waiters	1	1	6
Anthony Bennett	1	0	0
Matthew Dellavedova	0	0	0
Alonzo Gee	1	0	0
Tyler Zeller	0	0	0

Table 3.3: Example Player-level Features at 12th minute for Cleveland Cavaliers Based on Data in Table 3.2

K	**Rebounds Features**	**Assists Features**	**Points Features**
2	[3, 3]	[6, 1]	[12, 6]
3	[3, 3, 1]	[6, 1, 1]	[12, 6, 6]
4	[3, 3, 1, 1]	[6, 1, 1, 1]	[12, 6, 6, 6]
5	[3, 3, 1, 1, 0]	[6, 1, 1, 1, 0]	[12, 6, 6, 6, 2]

Table 3.4: Example of Combined Features

X & K	**Time Point**	**Combined Features**
X=5,K=2	2nd minute	PD at 2min, Top 2 stats of 18 traditional statistics *1+18*2*2 = 73 features*
	8th minute	PD at 8min, PD at 6min, PD at 4min, PD at 2min, Top 2 stats of 18 traditional statistics *4+18*2*2 = 76 features*
X=5,K=5	8th minute	PD at 8min, PD at 6min, PD at 4min, PD at 2min, Top 5 stats of 18 traditional statistics *4+18*2*5 = 184 features*
	14th minute	PD at 14min, PD at 12min, PD at 10min, PD at 8min, PD at 6min, Top 5 stats of 18 traditional statistics *5+18*2*5 = 185 features*
X=10,K=5	14th minute	PD at 14min, PD at 12min, PD at 10min, PD at 8min, PD at 6min, PD at 4min, PD at 2min, Top 5 stats of 18 traditional statistics *7+18*2*5 = 187 features*

Table 3.5 summerizes feature designs used in all above approaches.

3.3 Training Model

The training model we use in this study is logistic regression. As outcome of an NBA match is either win or loss, we expect to use a 2-class classifier to train and test our data. Logistic regression is not only a good model for classifying 2 classes, but also

Table 3.5: Summary of Feature Designs

Approach	**Description**	**Number of Features**
H-Diff	Historical Wins and Losses of Home Team against Away Team	2
P-Diff	Present Points Difference of Home Team to Away Team	1
RX-Diff	Recent X Points Differences of Home Team to Away Team	[1,X]
TPK	Top K stats of 18 Traditional Statistics of Each Team	18*2*K
TPK-RX-Diff	Recent X Points Differences of Home Team to Away Team + Top K stats of 18 Traditional Statistics of Each Team	[1+18*2*K, X+18*2*K]

decent simple comparing to other machine learning models, which meets our needs of a simpler model. To ensure the best prediction result, we use N-fold cross-validation with a series of learning rates and shuffle data set for each training to reduce effect from match order. We will classify a match as 0 if home team if predicted to win the upcoming match, and 1 if away team is predicted as the winner.

Chapter 4

EXPERIMENTS

4.1 Data Set

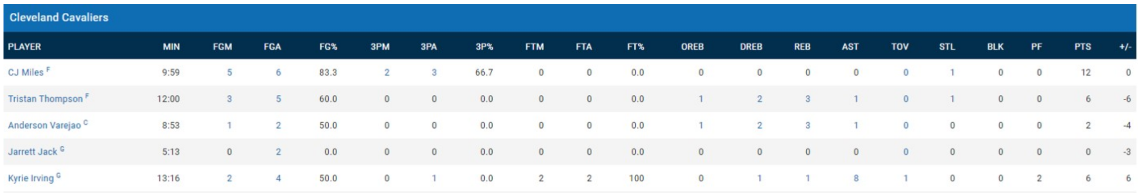
The dataset [1] [2] used in this study was derived from season games played in previous 5 NBA seasons. In total, we have 7140 matches with each season containing 1230 matches except 2011-2012 season has only 990 due to lockout of the league. For each match, we collect real-time data of both teams every 2 minutes.

Figure 4.1 and Figure 4.2 show traditional statistics of part Cavaliers players at 12th and 30th minute in the example match mentioned in previous chapter. Seeing from the figures, with game goes on, statistics of each individual player changes and becomes more comprehensive. For each match in our data set, we will have 23 similar data tables, corresponding to each time point.

Note that we only apply our model to regular time of season games. Pre-season, playoff, all-star, the finals and overtime scenarios are NOT studied in this thesis. Description of 18 traditional basketball box statistics involved in TPK is shown in Table 4.1.

[1] http://www.nba.com/

[2] http://sports.yahoo.com/nba/



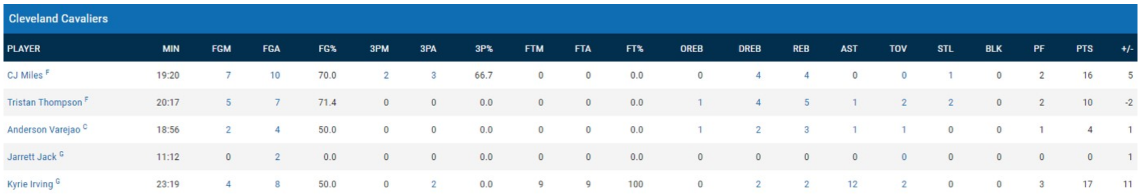
Cleveland Cavaliers

PLAYER	MIN	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	PTS	+/-
CJ Miles ^{F}	9:59	5	6	83.3	2	3	66.7	0	0	0.0	0	0	0	0	0	1	0	0	12	0
Tristan Thompson ^{F}	12:00	3	5	60.0	0	0	0.0	0	0	0.0	1	2	3	1	0	1	0	0	6	-6
Anderson Varejao ^{C}	8:53	1	2	50.0	0	0	0.0	0	0	0.0	1	2	3	1	0	0	0	0	2	-4
Jarrett Jack ^{G}	5:13	0	2	0.0	0	0	0.0	0	0	0.0	0	0	0	0	0	0	0	0	0	-3
Kyrie Irving ^{G}	13:16	2	4	50.0	0	1	0.0	2	2	100	0	1	1	8	1	0	0	2	6	6

Figure 4.1: Traditional Statistics of Part Cavaliers Players at 12th minute

Table 4.1: Description of 18 Traditional Basketball Box Statistics

Feature	**Description**
FGM	Field Goal Made
FGA	Field Goal Attempted
FG%	Field Goal Percentage
3PM	3-Pointers Made
3PA	3-Pointers Attemped
3P%	3-Pointers Percentage
FTM	Free Throws Made
FTA	Free Throws Attemped
FT%	Free Throw Percentage
OREB	Offensive Rebounds
DREB	Defensive Rebounds
REB	Rebounds
AST	Assists
TOV	Turnovers
STL	Steals
BLK	Blocks
PF	Personal Fouls
PTS	Points



Cleveland Cavaliers

PLAYER	MIN	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	PTS	+/-
CJ Miles F	19:20	7	10	70.0	2	3	66.7	0	0	0.0	0	4	4	0	0	1	0	2	16	5
Tristan Thompson F	20:17	5	7	71.4	0	0	0.0	0	0	0.0	1	4	5	1	2	2	0	2	10	-2
Anderson Varejao C	18:56	2	4	50.0	0	0	0.0	0	0	0.0	1	2	3	1	1	0	0	1	4	1
Jarrett Jack G	11:12	0	2	0.0	0	0	0.0	0	0	0.0	0	0	0	0	0	0	0	0	0	1
Kyrie Irving G	23:19	4	8	50.0	0	2	0.0	9	9	100	0	2	2	12	2	0	0	3	17	11

Figure 4.2: Traditional Statistics of Part Cavaliers Players at 30th minute

Table 4.2: Parameter Settings for Training Model

Learning Rates	0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10
Folds	5, 10
Maximum Iterations	500

4.2 Parameter Settings

4.2.1 Feature Design

As stated in Chapter 3, value of X in RX-Diff and value of K in TPK are dynamic. Changes of values of X and K may have influence on prediction accuracy. In order to exacavate the effect on accuracy of X and K, or parameter sensitivity, we set X = 5, 10 and K = 2, 3, 4, 5 for Recent X Differences and Top K Stats, respectively. For TPK-RX-Diff, we implement with different combinations of X and K in RX-Diff and TPK.

4.2.2 Training Model

For logistic regression, we use 5 and 10-fold cross-validation with a series of learning rates and maximum iterations of 500. Detailed settings of parameters for our training model can be found in Table 4.2.

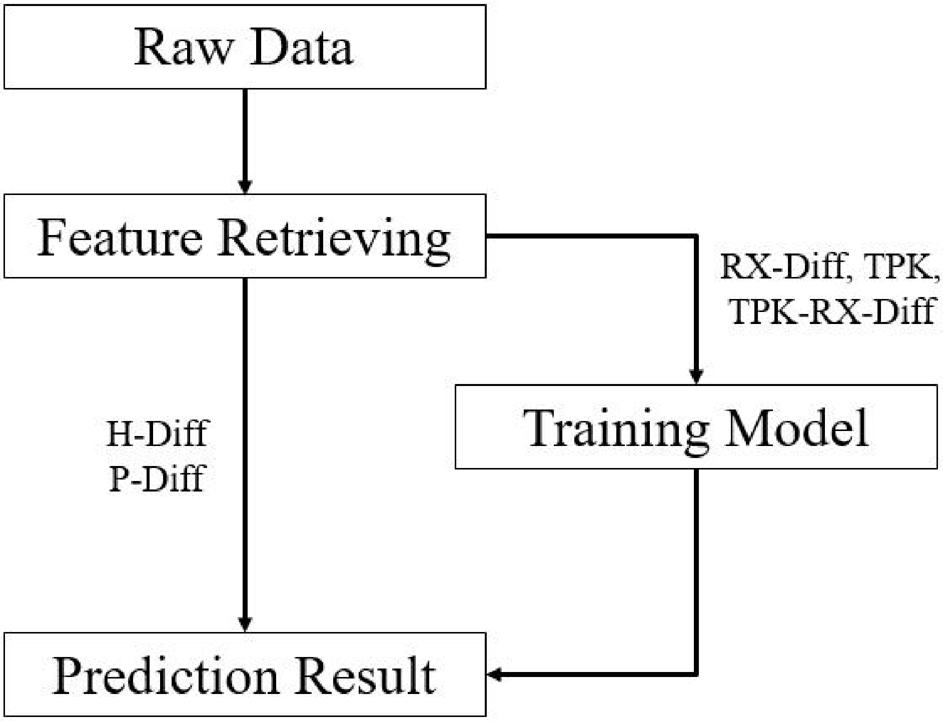


Figure 4.3: Flowchart of Experiment Process

4.3 Experiment Process

We first retrieve features from data set with feature designe methods mentioned in Chapter 3 and different groups of parameters mentioned in above section. Then, input feature sets of RX-Diff, TPK and TPK-RX-Diff to training model, and train the model with different parameters to obtain the prediction accuracies. According to our parameter settings, for each training model, we have 2 sets of results for RX-Diff, 5 sets of results for TPK and 10 for TPK-RX-Diff. For two baseline methods, H-Diff and P-Diff, we can directly get the prediction results by simple comparisons. A more intuitive way to show our process can be found in Figure 4.3.

Chapter 5

RESULTS & ANALYSIS

We present the results based on history records of recent 2 and 3 seasons for baseline H-Diff. Thus, there are in total 3 baselines in our result for comparison, H-Diff(2), H-Diff(3) and P-Diff. For RX-Diff, TPK and TPK-RX-Diff, we record its average, maximum, minumum and variance of testing accuracy in their experiments.

Evalutation Objectives

Objective 1: Average Prediction Accuracy - Early Game and Late Game

Objective 2: Stability in Predicting

Objective 3: Parameter Sensitivity (X, K)

N-fold Cross-Validation

Table 5.1 shows summary of N-fold Cross-validation results different set of parameters in Table 4.2. Comparing two blue columns, we observe that there is little difference between different number of folds. And comparing two columns in red, we find there is also little difference among different learning rates. Therefore, any combination of learning rate and number of folds delivers nearly the same results. For concision, in the following, we only show the result of 5-fold cross-validation and learning rate of 0.1 for our training model.

Baselines

As shown in Figure 5.1, H-Diff(2) (yellow line) and H-Diff(3) (black line) do not change over time. P-Diff, as it contains real-time information, has a dynamically

Table 5.1: Summary of Cross-valition Results

Time	**5-Fold**				**10-Fold**			
	Avg	**Max**	**Min**	**Gap**	**Avg**	**Max**	**Min**	**Gap**
2min	53.118	53.118	53.118	0	53.175	53.175	53.175	0
4min	57.691	57.892	57.374	0.518	57.786	58.02	57.531	0.489
6min	60.279	60.412	60.146	0.266	60.314	60.358	60.176	0.182
8min	62.483	62.694	62.316	0.378	62.628	62.841	62.542	0.299
10min	64.708	64.796	64.572	0.224	64.678	64.752	64.612	0.14
12min	66.094	66.28	65.97	0.31	66.187	66.476	66.063	0.413
14min	66.792	66.852	66.686	0.166	66.808	66.874	66.705	0.169
16min	68.020	68.17	67.904	0.266	68.067	68.109	67.98	0.129
18min	69.160	69.262	69.094	0.168	69.216	69.381	68.884	0.497
20min	70.442	70.548	70.324	0.224	70.448	70.514	70.36	0.154
22min	71.281	71.38	71.122	0.258	71.218	71.368	70.753	0.615
24min	72.765	72.846	72.692	0.154	72.788	73.026	72.67	0.356
26min	72.903	73.01	72.844	0.166	72.914	73.15	72.792	0.358
28min	74.527	74.616	74.426	0.190	74.517	74.692	74.444	0.248
30min	76.003	76.11	75.638	0.472	75.990	76.159	75.854	0.305
32min	76.847	76.968	76.734	0.234	76.837	76.913	76.782	0.131
34min	78.160	78.232	78.086	0.146	78.200	78.336	78.117	0.219
36min	79.281	79.408	79.214	0.194	79.329	79.395	79.248	0.147
38min	80.642	80.728	80.582	0.146	80.654	80.728	80.58	0.148
40min	81.796	81.858	81.71	0.148	81.804	81.86	81.77	0.09
42min	83.295	83.366	83.202	0.164	83.311	83.366	83.249	0.117
44min	84.767	84.924	84.614	0.31	84.785	84.849	84.742	0.107
46min	86.465	86.6	85.852	0.748	86.521	86.579	86.475	0.104

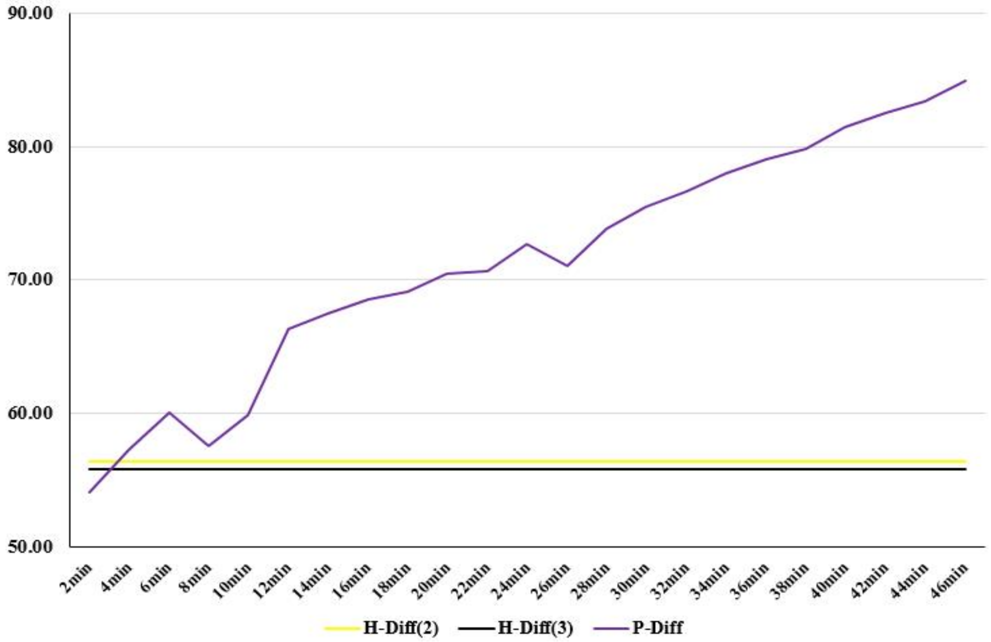


Figure 5.1: Accuracy Comparison of Baselines

increasing accuracy. When it gets close to end of match (46th minute), P-Diff has an accuracy around 85%, which is a reasonable result since some games were ”close” match-ups and we did not include overtime scenarios. Although P-Diff has high average late game accuracy, it performs bad in early game.

RX-Diff

Accuracies of R5-Diff (square dot line) and R10-Diff (diamond dot line) in Figure 5.2 have similar trend to P-Diff. At the beginning of a game, both R5-Diff and R10-Diff have low accuracy, even lower than P-Diff on average, but increase more steadily than P-Diff and become slightly higher than P-Diff when getting close to the end of match. Besides, value of X has little effect to prediction accuracy, as square dot line almost overlaps with diamond dot line. However, when we try to analyze the variances of RX-Diff in Figure 5.3, we find that variace of both R5-Diff and R10-Diff dramatically increases from 36th minute (start of 4th quarter in a match). This

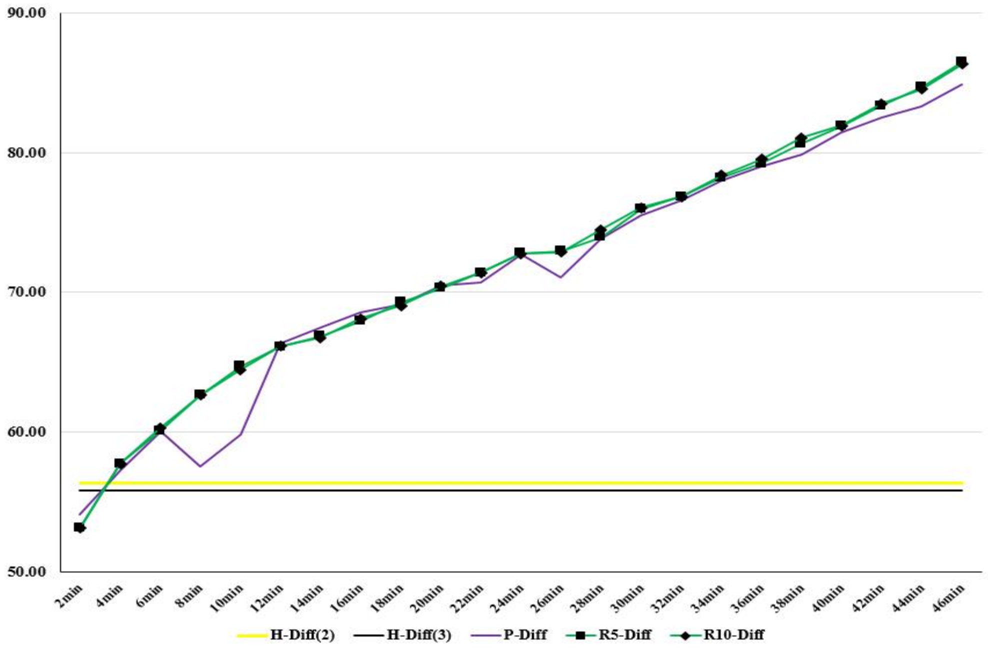


Figure 5.2: Average Accuracy Comparison of RX-Diff (X = 5, 10)

phenomenon leads the range of prediction accuracy to become wider in 4th quarter. Intuitively shown in Figure 5.4, most of the ranges of accuracy are within 5% before 36th minute, which is somehow acceptable, yet ranges after 36th minute are more than 5%, even 10% at 46th minute. This shows that RX-Diff has unstable performance in 4th quarter, which is not a beneficial thing in actual predictions, even though its average performance beats baselines.

Therefore, RX-Diff has low average early game accuracy, high average late game accuracy, low stability, and low parameter sensitivity.

TPK

Figure 5.5 shows the result of TPK feature sets. Different to RX-Diff, TPK has higher accuracies at the beginning of match yet increases much slower over time, which causes accuracies to be lower than P-Diff since the end of 1st quarter (12th minute). When K is greater than 2, accuracies at in 2nd half (after 24th minute)

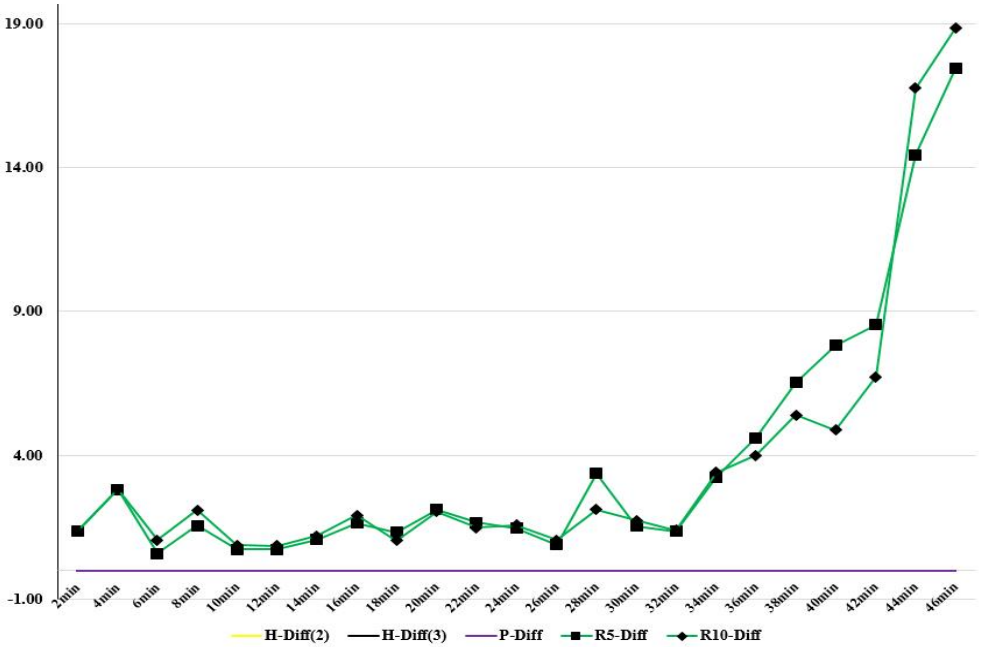


Figure 5.3: Variance Comparison of RX-Diff (X = 5, 10)

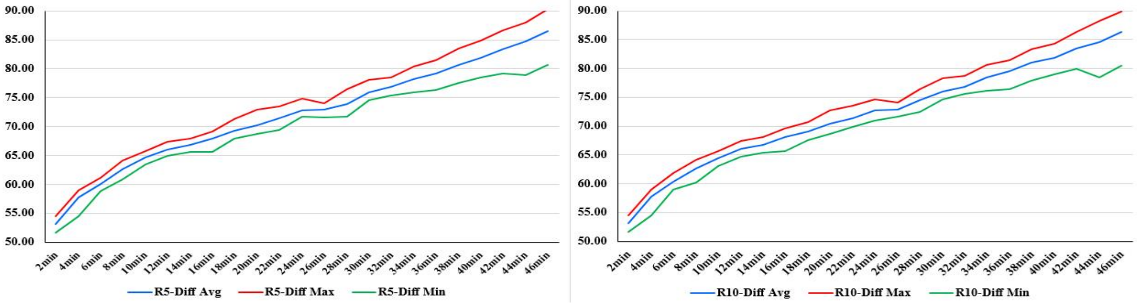


Figure 5.4: Internal Comparison of RX-Diff (X = 5, 10)

first drop dramatically, then goes back to similar increasing trend as TP2 (square dot line), while TP2 does not have accuracy drop between 24th minute and 26th minute. We call this phenomenon *half game drop* in the following and we can see that half game drop in TPK is significant. Potential reasons that may cause this phenomenon could be five starting linup players return on the court together again after half break, or half break (10 minutes) brings more discontinuity than quarter breaks (2 minutes). By comparing TP2 (square dot line) to TP3 (diamond dot line), we find that this may

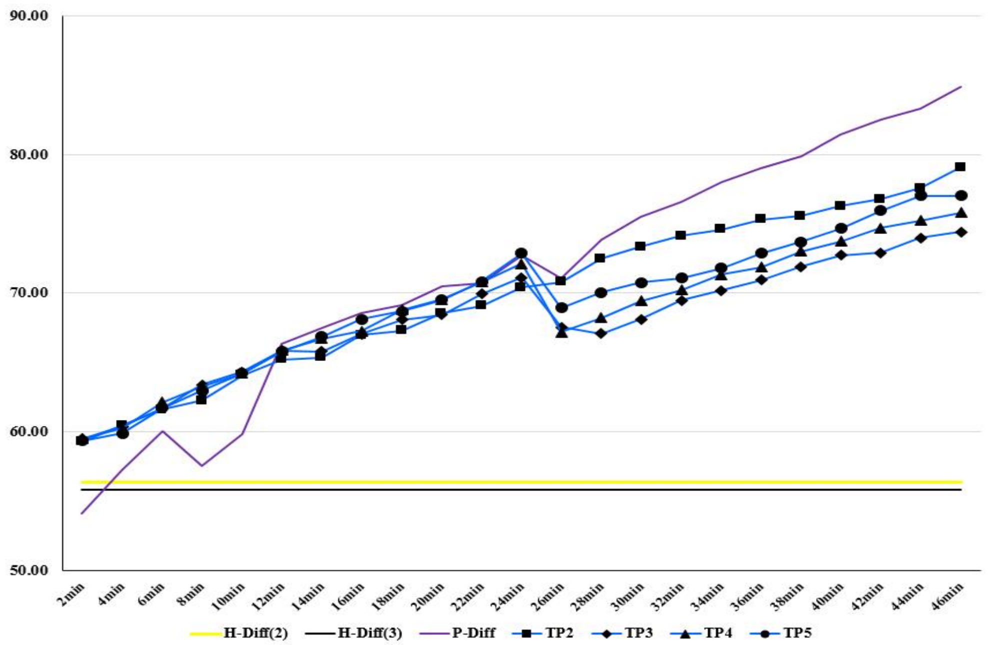


Figure 5.5: Average Accuracy Comparison of TPK

happen when we start to include 3rd highest number in each traditional statistics, and reduce due to continuous growing in number of features (TP4 (triangle dot line) and TP5 (circle dot line)).

In addition, change of K also has large influence on accuracy, especially in 2nd half, which may be another consequence caused by half game drop. Although TPK does not outperform P-Diff and RX-Diff in late game, it has a variance that keeps steady and low throughout the game (Figure 5.6). Internal comparison in Figure 5.7 shows that nearly all ranges are within 5%, which draws that TPK is much more stable than RX-Diff at any time of a match.

Thus, TPK has high average early game accuracy, low average late game accuracy, high stability, high parameter sensitivity and half game drop phenomenon.

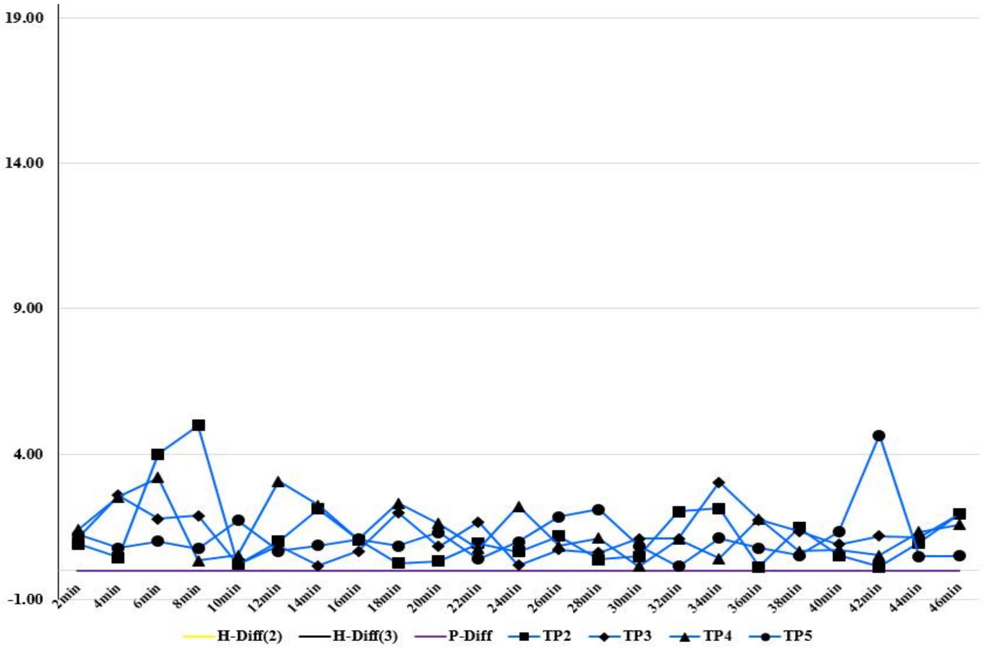


Figure 5.6: Variance Comparison of TPK

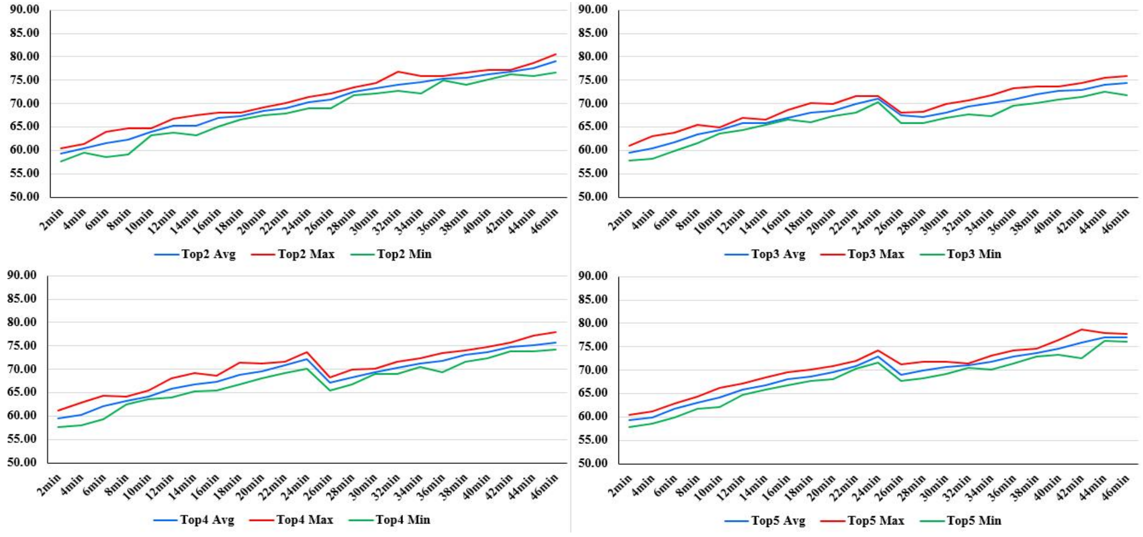


Figure 5.7: Internal Comparison of TPK (K = 2, 3, 4 and 5)

TPK-RX-Diff

Combining feature sets of TPK and RX-Diff, we have results shown in Figure 5.8 for X = 5 . We can see from the figures that although half game drop still exists in TPK-R5-Diff, it is reduced comparing to results of TPK. Besides, TPK-R5-Diff has both high average early game accuracy and high average late game accuracy, and TP2-R5-Diff, one that does not have half game drop, outperforms all other methods in the 2nd half.

Comparing Figure 5.8 and Figure 5.9, we see that accuracies are nearly the same at all time points between X = 5 and X = 10. This indicates that TPK-RX-Diff has low parameter sensitivity of X, which is advantage of RX-Diff. Besides average accuracy, TPK-R10-Diff also performs very close to TPK-R5-Diff in variance and internal comparison. Therefore, for concision, we only show results of TPK-R5-Diff in the following.

Accuracy variance (Figure 5.10) of TPK-RX-Diff keeps steady and low throughout the game, and internal comparison (Figure 5.11 and Figure 5.12) shows (i). half game drop is reduced; (ii). range of accuracy is small at any time point; (iii). parameter sensitivity of K is reduced.

Therefore, TPK-RX-Diff has high average early game accuracy, high average late game accuracy, high stability, reduced half game drop, low parameter sensitivity of X and reduced parameter sensitivity of K.

A summary of all experimental results can be found in Table 5.2. TPK-RX-Diff combines the advantages and reduces the disadvantages of both RX-Diff and TPK.

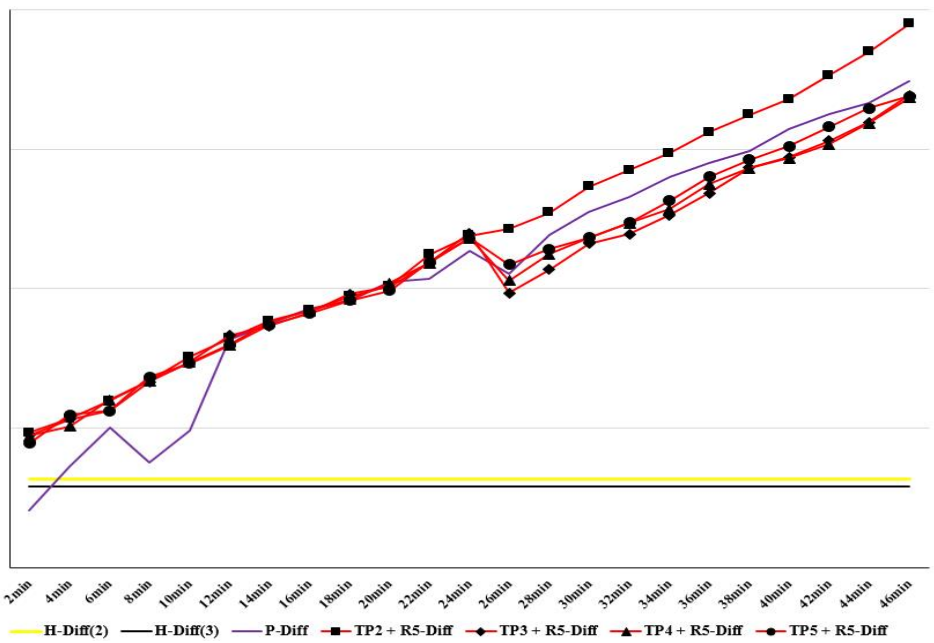


Figure 5.8: Average Accuracy Comparison of TPK-R5-Diff

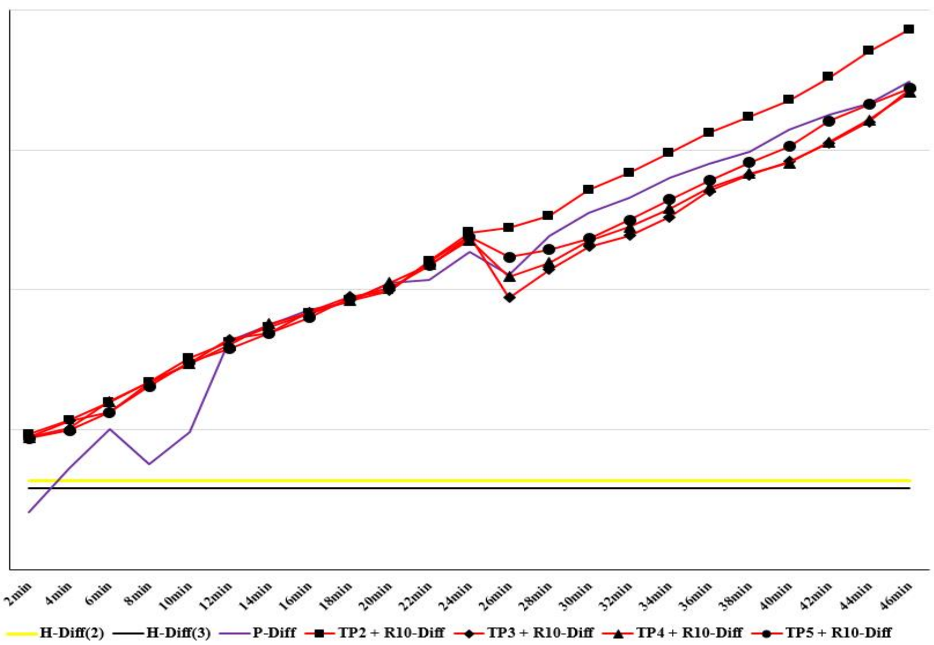


Figure 5.9: Average Accuracy Comparison of TPK-R10-Diff

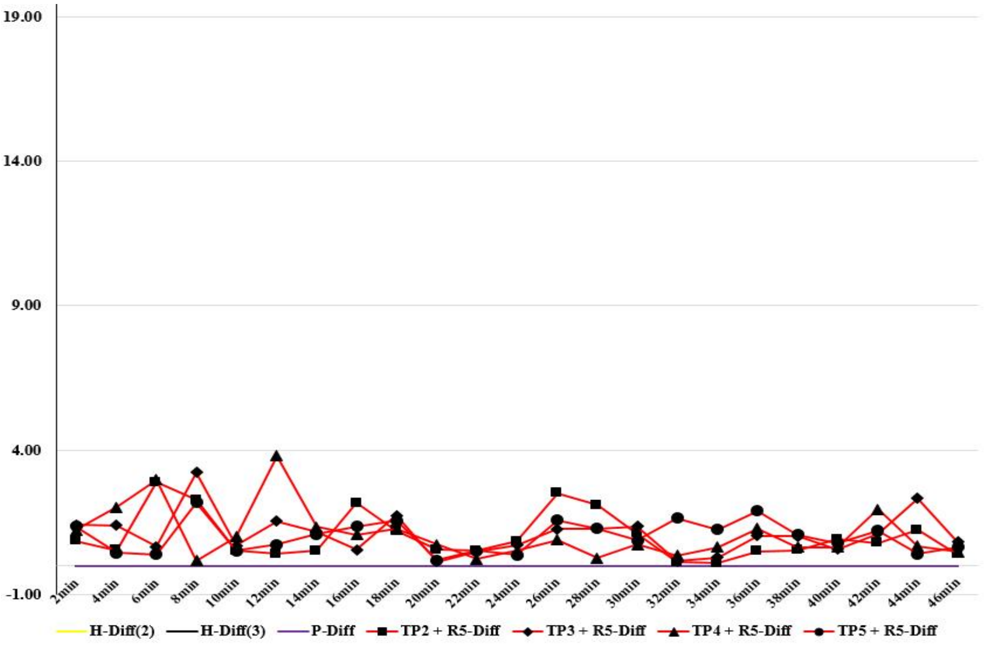


Figure 5.10: Variance Comparison of TPK-R5-Diff

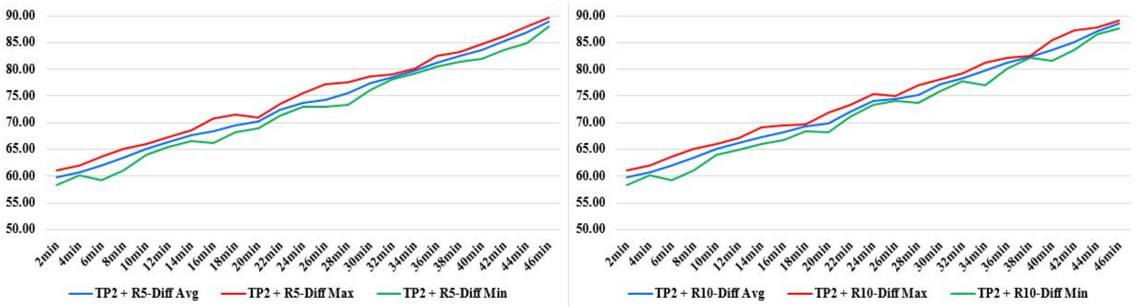


Figure 5.11: Internal Comparison of TPK-RX-Diff, Fixed K, X = 5, 10

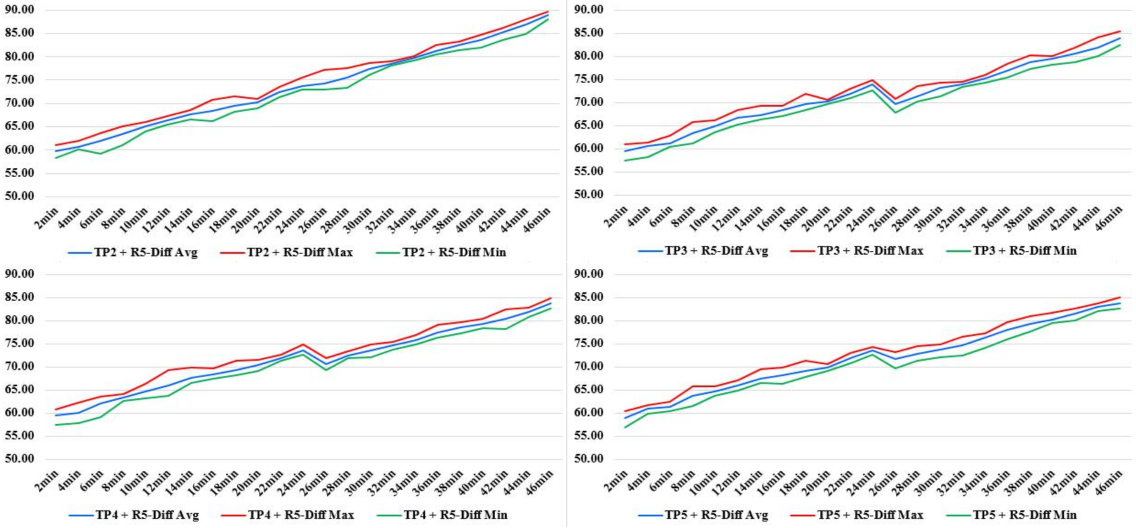


Figure 5.12: Internal Comparison of TPK-RX-Diff, Fixed X, K = 2, 3, 4, 5

Table 5.2: Summary of All Experimantal Results (Bold Italics are Advantages)

Method	**Early Accuracy**	**Late Accuracy**	**Stability**	**Sense X**	**Sense K**	**Half Game Drop**
H-Diff	low	low	N/A	N/A	N/A	N/A
P-Diff	low	high	N/A	N/A	N/A	N/A
RX-Diff	low	***high***	low	***low***	N/A	N/A
TPK	***high***	low	***high***	N/A	high	significant
TPK-RX-Diff	***high***	***high***	***high***	***low***	***reduced***	***reduced***

Chapter 6

CONCLUSION & FUTURE WORK

6.1 Conclusion

In this study, we verified both our hypotheses raised related to real-time prediction with the methods we present. Results of our experiments all support that prediction accuracy increases with match goes on and prediction power can be achieved with feature sets that are easily understood. By introducing RX-Diff, TPK and TPK-RX-Diff, we provide simple feature designs that also embraces much prediction power, especially in real-time match outcome prediction. Besides, to the best of our knowledge, we are the first to investigate the possibility and characteristics of real-time prediction in NBA matches.

6.2 Future Work

(i). As our result shows, prediction accuracy significantly drop between 24th minute and 26th minute when features from TPK are involved in predicting. Influencing factors that cause this problem is worthy to be studied and understood in the future;

(ii). Besides logistic regression, there are other machine learning models that can be used for predicting NBA matches. We would like to validate and verify our finding with alternative machine learning models;

(iii). We used data set of season games in regular time, while pre-season, playoff, the finals and overtime scenarios are not studied in this thesis. Therefore, apply our methods in these scenarios may be another direction for future study;

(iv). Finally, there are other factors that may influence outcome of a match can also be considered as features of both game-level and player-level. We would like to explore and select these feature to enrich our feature set.

REFERENCES

Aslan, B. G. and M. M. Inceoglu, “A comparative study on neural network based soccer result prediction”, in “Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on”, pp. 545–550 (IEEE, 2007).

Berri, D. J., “Who is’ most valuable’? measuring the player’s production of wins in the national basketball association”, Managerial and decision economics pp. 411–427 (1999).

Brown, M., J. Sokol *et al.*, “An improved lrmc method for ncaa basketball prediction”, (2010).

Cao, C., “Sports data mining technology used in basketball outcome prediction”, (2012).

Chen, S. and T. Joachims, “Modeling intransitivity in matchup and comparison data”, in “Proceedings of the Ninth ACM International Conference on Web Search and Data Mining”, pp. 227–236 (ACM, 2016a).

Chen, S. and T. Joachims, “Predicting matchups and preferences in context”, in “Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining”, pp. 775–784 (ACM, 2016b).

Cortes, C. and V. Vapnik, “Support-vector networks”, Machine learning **20**, 3, 273–297 (1995).

Cox, D. R., “The regression analysis of binary sequences”, Journal of the Royal Statistical Society. Series B (Methodological) pp. 215–242 (1958).

DeLong, C., N. Pathak, K. Erickson, E. Perrino, K. Shim and J. Srivastava, “Teamskill: modeling team chemistry in online multi-player games”, in “Pacific-Asia Conference on Knowledge Discovery and Data Mining”, pp. 519–531 (Springer, 2011).

DeLong, C. and J. Srivastava, “Teamskill evolved: Mixed classification schemes for team-based multi-player games”, in “Pacific-Asia Conference on Knowledge Discovery and Data Mining”, pp. 26–37 (Springer, 2012).

DeLong, C., L. Terveen and J. Srivastava, “Teamskill and the nba: applying lessons from virtual worlds to the real-world”, in “Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining”, pp. 156–161 (ACM, 2013).

Elo, A. E., *The rating of chessplayers, past and present* (Arco Pub., 1978).

Glickman, M. E., “Paired comparison models with time-varying parameters”, Tech. rep., DTIC Document (1993).

Glickman, M. E., “Parameter estimation in large dynamic paired comparison experiments”, Applied Statistics pp. 377–394 (1999).

Haghighat, M., H. Rastegari and N. Nourafza, “A review of data mining techniques for result prediction in sports”, Advances in Computer Science: an International Journal **2**, 5, 7–12 (2013).

Hu, F. and J. V. Zidek, “Forecasting nba basketball playoff outcomes using the weighted likelihood”, Lecture Notes-Monograph Series pp. 385–395 (2004).

Kvam, P. and J. S. Sokol, “A logistic regression/markov chain model for ncaa basketball”, Naval Research Logistics (NrL) **53**, 8, 788–803 (2006).

Langseth, H., “Beating the bookie: A look at statistical models for prediction of football matches.”, in “SCAI”, pp. 165–174 (2013).

Leung, C. K. and K. W. Joseph, “Sports data mining: predicting results for the college football games”, Procedia Computer Science **35**, 710–719 (2014).

Loeffelholz, B., E. Bednar, K. W. Bauer *et al.*, “Predicting nba games using neural networks”, Journal of Quantitative Analysis in Sports **5**, 1, 1–15 (2009).

Lopez, M. J. and G. J. Matthews, “Building an ncaa mens basketball predictive model and quantifying its success”, Journal of Quantitative Analysis in Sports **11**, 1, 5–12 (2015).

McCabe, A., “An artificially intelligent sports tipper”, in “Australian Joint Conference on Artificial Intelligence”, pp. 718–718 (Springer, 2002).

McCabe, A. and J. Trevathan, “Artificial intelligence in sports prediction”, in “Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on”, pp. 1194–1197 (IEEE, 2008).

Melnick, M. J., “Relationship between team assists and win-loss record in the national basketball association”, Perceptual and Motor Skills **92**, 2, 595–602 (2001).

Miljković, D., L. Gajić, A. Kovačević and Z. Konjović, “The use of data mining for basketball matches outcomes prediction”, in “Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on”, pp. 309–312 (IEEE, 2010).

Min, B., J. Kim, C. Choe, H. Eom and R. B. McKay, “A compound framework for sports results prediction: A football case study”, Knowledge-Based Systems **21**, 7, 551–562 (2008).

Minsky, M. and S. Papert, “Perceptrons: an introduction to computational geometry (expanded edition)”, (1988).

Rue, H. and O. Salvesen, “Prediction and retrospective analysis of soccer matches in a league”, Journal of the Royal Statistical Society: Series D (The Statistician) **49**, 3, 399–418 (2000).

Sampaio, J. and M. Janeira, “Statistical analyses of basketball team performance: understanding teams’ wins and losses according to a different index of ball possessions”, International Journal of Performance Analysis in Sport **3**, 1, 40–49 (2003).

Štrumbelj, E. and P. Vračar, "Simulating a basketball match with a homogeneous markov model and forecasting the outcome", International Journal of Forecasting **28**, 2, 532–542 (2012).

Trawinski, K., "A fuzzy classification system for prediction of the results of the basketball games", in "Fuzzy Systems (FUZZ), 2010 IEEE International Conference on", pp. 1–7 (IEEE, 2010).

Vaz de Melo, P. O., V. A. Almeida and A. A. Loureiro, "Can complex network metrics predict the behavior of nba teams?", in "Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 695–703 (ACM, 2008).

Wei, N., "Predicting the outcome of nba playoffs using the naïve bayes algorithms", University of South Florida, College of Engineering (2011).

Zak, T. A., C. J. Huang and J. J. Siegfried, "Production efficiency: the case of professional basketball", Journal of Business pp. 379–392 (1979).