CARLETON UNIVERSITY

Process Oriented Player Evaluation Metrics for USports Basketball

by

Peter L'Oiseau

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN STATISTICS

OTTAWA, ONTARIO

April, 2021

Abstract

This thesis focuses on quantifying the true talent of each individual player on the court of a Canadian USports basketball game. In this pursuit, the ultimate measure of success is whether or not the quantification of player value can be used to project future outcomes of players and the teams which they comprise. The paper explores both bottom-up and top-down approaches to assigning value to individual players. These assignments of value are then tested and contrasted to observe what measures should to be used to understand properly what a player can contribute to a team. The name of the game in basketball is to score more points than the other team and thus a player who can contribute to this end is valuable. However, due to the smaller number of games played by individual players at the Canadian University level as compared to professional levels of basketball, more noise is present in measures that focus on team outcomes, which obscures player's true talent. To combat this, a deeper examination is undertaken to understand the processes which predict better outcomes and tell a more salient picture of a player's true talent.

Preface

The data used to inform this thesis comes from Synergy Sports[1], which does an incredible service to basketball at all levels across the world, making data and video available to coaches, players and analysts. Their hosting, along with the database maintained by the Ontario University Athletics, makes research and analysis possible for so many levels of competitive basketball. It is also crucial to understand the well built foundation laid by forerunners in the field of basketball analytics. Authors like Dean Oliver, John Hollinger, Jeff Sagarin, Dan Rosenbaum and Nate Silver pioneered the concepts upon which analysts can build today. The top-down approach to assigning value to individual players through regression models was a monumental shift in the way basketball teams, fans, players, and analysts have come to understand the game. A passage from Alan Schwarz's *The Numbers Game*[2] articulates these people's work in the space perfectly. Schwarz is writing about the role of Elias Sports Bureau in the early analytics movement in baseball in 1980's:

> The Elias books... *Analysts* became the primary cause of the statistics epidemic of the 1980's and beyond where fans were

[1] Synergy Sports, Synergy Sports^{TM} , www.synergysportstech.com/synergy/.
[2] Schwarz, Alan. The Numbers Game: Baseballs Lifelong Fascination with Statistics.

> deluged with incessant statistical gobbledygook. They put millions of statistics in the hands of people who didn't know how to use them, like putting a chainsaw in the hands of a hyperactive teenager with similarly grisly results. Announcers would cite Benny Distefano's slugging percentage with the bases empty in late innings as if it were meaningful.

Having access to well-kept granular data is one incredibly important piece to *understanding* a problem, but using the rigorous statistical methods developed by academicians is crucial to *knowing*. With these strategies and data, we can develop hypotheses and tests in order to make accurate predictions about the way the world will behave. These people and many others introduced rigor into the world of basketball analysis and turned the granular data we have from an unwieldy chainsaw into a precision craftsman's knife.

Acknowledgments

I would like to sincerely thank all the people who made the research of sport statistics possible for me and that starts with my supervisor Dr. Shirley Mills. This thesis is born of work I did as a full-time intern of the Carleton University Ravens Women's and Men's basketball teams and if it were not for Dr.Mills I would have never have been in this position. The position too would never have existed without the generosity and vision of an incredible donor. Additionally, the person who made the partnership work between the teams and the School of Mathematics and Statistics, and the Director of Recreation & Athletics at Carleton University, Jennifer Brenning. I also add my thanks to all the people on these teams who I worked with everyday and put up with my incessant questions in trying to learn the intricacies of the game, truly experts in their field and absolute honour to work with: Director of Basketball Operations Dave Smart, Men's Head Coach Taffe Charles, Women's Head Coach Brian Cheng, Men's Assistant Coach Rob Smart, Men's Assistant Coach Jamie Campbell and Women's Assistant Coach Michelle Abella. And lastly, I thank all the players, who are the reason why we all do this. Your incredible talents and work as conscientious student-athletes is an incredible asset to the community. It does not go unnoticed.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The aim of this thesis is to quantify the talent of individual Canadian USports basketball players in one number metrics. The novel approach taken will be using play-by-play data, which describes the gameplay in text, as the basis for quantifying a player's decision making. By examining what goes into the process of making plays on the basketball court, we gain new information about a player's talent which is not ordinarily captured in other one number metrics. This information is then fed into the industry-leading metrics to improve their ability to tell us about how talented a player is.

Fundamental to the quantification and projection of individual basketball players is the idea of true talent. When we measure any random variable, there is variance associated with it and that applies to basketball players' performance as well. When we measure any statistics with respect to an individual player, there is presumed to be some element of signal and some element of noise. In this model, noise is the variability and the signal is a player's true talent. Some random variables are nearly all noise and actually provide no insight into how a player is likely to perform in the future. For

any metric introduced in this paper, it will be crucial to measure how well it predicts future performance, i.e. does this metric tell us something about a durable skill that a player can possess or is it merely a reflection of noise. Additional to the stability of a metric, it has to predict some other variable which we care about. While a metric which can be used to predict itself into the future may be an indication of a durable skill, if it is completely isolated from any other variables, then the skill is not very interesting. In the case of player performance metrics, we also care that the summation of players' performance in a metric for a team predicts the success of that team.

Two structural ways to approach assigning statistics to individual players are top-down and bottom-up. Traditionally, sports statistics have relied on the bottom-up model where we look at individuals and track how many times he/she commits an act on the floor. Typical examples of these statistics are points scored, assists, rebounds, steals, etc. These measures are captured in the box score but often lack context, so many in the statistical community have presented excellent ways of improving the utility of these numbers by using a relevant denominator. This is in contrast to the top-down method which attempts to quantify team performance and subsequently attribute the results to the component parts of the team. Top-down metrics capture actions and potential skills a player has which are not easy to define and count over the course of a game.

At our disposal is the last five seasons of Canadian USports play in the Synergy Sports Technology database[1]. Play-by-play data in this database is extremely detailed. There is substitution data which allows us to know who is on the court at all times. There is sufficient information to determine

[1] Synergy Sports, Synergy Sports^{TM} , www.synergysportstech.com/synergy/

possession which is the fundamental building block in analytics that allows us to determine efficiency of a team or a player's offense and defense. Most importantly, with respect to the task of creating a more granular metric for player decision making, we have detailed commentary about what shots are taken and what series of events lead to it. This data is what will allow us to build models to predict the outcome of shots which follow specific play-by-play patterns in the Synergy Sports database.

The predictions from these models will be the expected points from the series of events described. Expected points can then be used to make the metrics which describe quality shot creation and shot-making (the difference between actual points and expected points). These are useful individual metrics, but combining the two is what will allow us to account for more information. Conceptually, we are trying to isolate a player's decision-making and abilities to get high quality shots offensively and prevent such defensively.

From here we can begin to build on many of the best metrics for player performance with the shot-quality and shot-making information. Regularized Adjusted Plus-Minus (RAPM), which relies on the number of points scored per possession, can be adjusted to account for the shot-making and quality of the players on the court. While a well-built expected points model will provide useful bottom-up data about shot-quality and shot-making, it is as a supplementation in regression models which will be its biggest contribution of value. Box Plus-Minus (BPM) uses a group of variables available in the basketball box score for individual players and uses them in a regression for RAPM to determine how much these variables matter. These coefficients are then used to determine an estimate for a player's true talent.

Adding these shot-quality and shot-making variables for the offensive and defensive sides of the ball serves as novel and valuable information in this regression.

Chapter 2

Literature Review

The bottom-up approach has wonderful potential since the number of events that can be tracked during a basketball game is limitless. Number of passes, sprint speed, number of steps taken and more continue the flow of new and useful bottom-up data, which improves our understanding of what a basketball player does and how it contributes to winning[1]. Dave Heeren's Tendex[2] was the first introduction of linear weights into basketball statistics. The goal was to summarize in one number a player's contribution to their team by adding and subtracting box score statistics with simple weights. This introduced the idea of *efficiency* to basketball statistics since it penalized players for missed shots. John Hollinger [3], utilised these bottom-up measures in his Player Efficiency Rating (PER) and built on the idea of linear weights in basketball statistics. Hollinger tried more or less to add up the good events a player does on the court, assign each event a weight and subtract out the bad events, again assigning each a weight. Subsequently,

[1] Silver, Nate. "How Our RAPTOR Metric Works."
[2] Oliver, Dean. Basketball on Paper
[3] Hollinger, John. "Calculating PER."

the score was adjusted relative to the context of the league. This was a huge step in the field since its more thoughtfully reasoned linear weights made an attempt to combine multiple aspects of a player's game and look to the league context to determine how great his impact actually was. Another crucial innovation this statistic made was the acknowledgement of pace and its variance. Pace is a measure that quantifies the number of possessions a team plays in a game which affects the numbers in the box score. When emphasis was put on possessions rather than totals in a game, a corner was turned since we could now put players who were on different teams, and playing in different eras on a more equal footing with respect to the bottom-up numbers we care about. This was truly laudable and an incredible innovation but the bottom-up approach of analyzing the box score fails to recognize the symmetry that exists between offense and defense in the game of basketball.

A point prevented is just as good as a point scored in basketball is somewhere very close to the truth. There is a dogmatic belief in sports culture that "defense wins championships" and there certainly is a case to be made that, in basketball, great defense makes offense easier by a larger factor than a great offense makes defense easy. This being said, the box score in basketball comes nowhere close to capturing how impactful a player is on defense. Steals, blocks, defensive rebounds, and personal fouls are the extent of what a standard box score tells us about defense. Personal fouls are not even an exclusively defensive stat and defensive rebounds tell us more about how close to the rim a player stands on defense, who they guard and how tall they are, than they do about a player's defensive value to a team. This then leaves the box score relatively useless for being the arbiter of defensive

talent of an individual player. There are new bottom-up metrics[4] being added to the record of data at the highest levels of basketball, with new technologies that are allowing for insights into the individual actions that make a player an effective defender. That said, the most effective way to currently quantify defense is via top-down approaches apportioning team defensive performance to the players on the court.

The top-down approach is the idea that we can apportion some fraction of credit for each event on the court to the 10 players that occupy it at a given time. Initially, this approach was quite naive and did a very poor job at apportioning value to individual players. The forefather of this method of player evaluation is plus-minus, a statistic where a player is given a point for every point their team scores while they are on the floor and deducted one for each point scored against while on the floor. The way we know this is not a good way to evaluate individual players is that it cannot predict itself consistently through time, it is coated irredeemably with noise. However, this is the scaffolding the pioneering researchers in basketball analytics used to build top-down statistics that do hold up to the test of capturing signal. Fundamentally, plus-minus is interesting but its fatal flaw is its lack of acknowledgement that some players deserve more or less credit for the events happening on the court. There are more distinctions to be drawn and issues to be dealt with but, by addressing this fatal flaw, we have entered into the realm of useful top-down player evaluation metrics.

ON/OFF rating builds on this top-down approach by measuring the scoring rate for and against a player's team when a player is on the court and compares it to when they are off the court. This is interesting because

[4]Dowsett, Ben. "Nylon Calculus: How Second Spectrum Is Redesigning the NBA."

there is an attempt to adjust plus-minus for the opponents and teammates however it does not go all the way. It can be used to compare teammates against each other and would be perfect at doing just that if teams played their players at random with no correlations, but of course this is not the case. Teams use reserves in tandem and certain teammates very rarely play together, like the starting and backup point guards. Additionally, some players get more playing time against bench units, which will then overvalue their production.

Adjusted Plus-Minus (APM) came into the public domain from the work of Jeff Sagarin, Wayne Winston[5] and Dan T. Rosenbaum[6]. Adjusted Plus-Minus looks at all the continuous segments of time with the same ten players on the court, and uses the difference between the two team's scores divided by how long the segment was (in time or possessions) as the dependant variable and runs a regression where the players are the independent variables. This takes on the fatal flaw of plus-minus by apportioning distinct values to each player on the court but goes further by actually parsing both offensive and defensive impact of a player. However, it is by no means the end of the story, since there are still issues to be addressed. First and foremost, is the issue of multicollinearity, players that play with each other a high percentage of the time cannot be properly apportioned value, due to a lack of controlled comparison. Regularized Adjusted Plus-Minus (RAPM)[7] takes on this problem by using a ridge regression which increases the bias in the model to a tolerable amount in the bias-variance trade off and reduce the

[5]Hruby, Patrick. "Numbers Game."

[6]Rosenbaum, Dan T. " Measuring How NBA Players Help Their Teams Win."

[7]Zou, Simon. Open Source Data Science Pipeline for Developing "Moneyball" Statistics in NBA Basketball.

multicollinearity problem.

In the past decade a big push has been made to blend the bottom-up and top-down approaches. Box Plus-Minus (BPM)[8] uses a player's box score statistics and bottom-up numbers and applies them in a regression for RAPM, to determine weights for each statistic. RAPTOR (Robust Algorithm using Player Tracking and On-Off Ratings)[9] similarly regresses RAPM, in this case using many more independent variables including state-of-the-art bottom-up metrics like distance travelled and player match-up data on defense, as well as other top-down metrics like a player's ON/OFF rating. RAPTOR, as any of the plus-minus metrics, suffers from the limitation of assuming that performance is linear and additive. This assumption is not fatal, as demonstrated by the fact that these metrics do predict out-of-sample well when the training covers multiple years, but factors like system, team and teammate interactions undeniably affect performance. For example, when analyzing the game of Chris Bosh with the Toronto Raptors in 2009-10 and comparing it to the following year when he was in Miami, plus-minus statistics say he performed worse in Miami. However, the context around him greatly changed. Usage rate is an estimate of the number of possessions which a player ends for their team on offense. Bosh went from being the best player on his team with a usage rate of 28.7, expending most of his energy on the offensive end, to being a clear third option with a usage rate of 23.5. While Bosh had not lost his ability to be a great basketball player, the actions he needed to take on the court were different because of his context and that is something these plus-minus metrics do not account for.

[8] Myers, Daniel. "About Box Plus/Minus (BPM)."
[9] Silver, Nate. "How Our RAPTOR Metric Works."

All these metrics in their first iterations suffered from problems which were easily identifiable by statisticians and basketball lifers alike. Sample sizes and stability of the metrics are always a consideration. Answering the question of how many games, minutes, or possessions does a player need to accumulate in order for the metric to give an accurate picture of the player's true talent is crucial for the quality of the metric. In most cases, the answer appears to be that several NBA seasons' worth of data is needed to predict out-of-sample metrics reliably and that poses a major problem for using similar metrics for college players who have seasons with between 25 and 35 games. Additionally, the idea of "garbage time" proves to be a statistically significant factor when calculating these metrics. Garbage time is the idea that, when the margin between two teams' scores in a game is high and there is little time remaining, the performance of players changes. This is an additional adjustment that needs to be made along with quality of competition because the effort being put forth by the players on the court is assumed to diminish in these situations.

Since the strict separation between the top-down and bottom-up measures has been eliminated by metrics like RAPTOR and Steve Ilardi and Jeremias Engelmann's Real Plus-Minus (RPM) [10], the focus of the analytics community appears to be obtaining better bottom-up statistics for individual players so that the role of the top-down statistics can be reduced. The purpose of the top-down statistic is to try to quantify the value that players are bringing that we are failing to capture. However, if there truly is talent there that we are missing and it is not merely noise, it must be manifesting itself in some action the player is taking on the court. Therefore, the role

[10] Ilardi, Steve. "The next Big Thing: Real plus-Minus."

of researchers in this space is to hypothesize what those hidden acts may be, collect data on them and test to see if those are indeed durable talents that we should value in a basketball player. To develop a hypothesis on this topic, researchers are trying to determine the chain of causality that leads players to be successful. Success is easy to define points scored vs. points allowed by a team and it is reflected as the dependent variable in the top-down regression models like APM and RAPM. Thinking about how to use these measures in ways beyond the box score requires us to neatly pick at the process that teams, and the players which comprise them, take to have success.

First and foremost, scoring points in basketball requires a team to take a shot that is likely to return points, thus *shot-quality* is the measure for which we search. The tact for defining shot-quality in the analytics community has been to collect as much data as possible about every shot and to determine what is significant in predicting points per shot attempt. Second Spectrum is a technology company which has been tracking National Basketball Association (NBA) player movement on the court since 2013 and produced a paper[11] in 2014 to try to attack this problem. Variables which they used to model the expected point value of a shot are distance from the basket when the shot is taken, the distance from the nearest defender and whether or not the shooter was dribbling the ball before they shot it. They divided their findings into two metrics the shot-quality a player was able to attempt and the ability for the player to convert those shots. These are two aspects of the process of scoring points that are not entirely independent but both tell us about an individual player's abilities when it comes to shooting and

[11] Chang, Yu-Han, et al. "Quantifying shot-quality in the NBA."

shot creation. While we do not have the same variables available to use as Second Spectrum, this model serves as the inspiration for the shot-quality and shot-making metrics developed by using the play-by-play data.

This by no means is the end of the shot-quality exploration because it leaves many more questions to be answered. When it comes to shot-making, what data can be collected about this process that will lead to a better understanding of what makes a player successful? Nearest defender and dribbling status are undeniably good ones but data on the mechanics of the body and how a basketball player should shoot the ball is on the forefront of data collection and analysis today[12]. Gathering data about the moment of the shot is undeniably a good thing but, diving deeper into the process which leads to the shot ought to tell us more about what makes a good offensive player and conversely what a good defensive player prevents. Narrowing in on exactly what these process attributes are ought to give us a metric that predicts outcomes better with less variability. More granular shot-quality metrics may help combat the sample size issue which is so prevalent in the case of college and university basketball analytics.

[12] Ewing, Lori. "Nick Nurse Learned Importance of Shooting Repetition at Early Age."

Chapter 3

Data Structure

Data used for these expected points models, separated by gender, are all the games where at least one men's or women's Canadian USports team is competing and is captured by Synergy from the 2015-16 season through to the 2019-20 season. The play-by-play which describes each possession is what we are interested in. There are approximately half a million observations for both men and women. Below is a snippet of data used to model the expected number of points on a possession:

Table 3.1: Women Data Summary

team	q	time	$play_A$	$play_B$	$play_C$	$play_D$	$play_E$	$play_F$	$play_G$	$play_H$	away score	home score	year	margin	preamble	pts
BRA	2	0:01:11	hailey mass	Spot-Up	No Dribble Jumper	Open	Long/3pt	Make 3 Pts	0	0	50	40	2020	-10	1	3
WIN	3	0:09:23	antoinette miller	ISO	Top	Drives Left	Dribble Jumper	Short to < 17'	Miss 2 Pts	0	33	39	2018	6	0	0
CQU	4	0:04:27	ashley moss	No Play Type	Turnover	0	0	0	0	0	50	42	2018	-8	0	0
RYU	2	0:00:23	cara tiemens	PnR Ball Handler	High PnR	Dribble Off Pick	Turnover	0	0	0	27	41	2018	14	0	0
TRW	4	0:02:08	nicole fransson	No Play Type	Foul	0	0	0	0	0	73	74	2019	1	0	2
CAR	4	0:03:48	tatyanna burke	No Play Type	Basket	Make 2 Pts Foul	0	0	0	0	40	70	2020	30	0	3
CMT	1	0:03:32	hannah dean	PnR Roll Man	Pick and Pops	Drives Left	Basket	Make 2 Pts	0	0	6	15	2016	9	1	2
MAN	1	0:07:06	laurern bartlett	PnR Ball Handler	Right PnR	Side	Dribble Off Pick	Basket	Miss 2 Pts	0	4	1	2020	-3	0	0
WLA	4	0:09:13	brianna iannazzo	Transition	Ballhandler	Basket	Foul	0	0	0	50	56	2017	6	0	1
BRC	2	0:06:12	kara spotton	Transition	First Middle	Dribble Jumper	Short to < 17'	Miss 2 Pts	0	0	16	30	2017	14	0	0

By variable names the meanings are as follows: *team* is the acronym of the team with possession of the ball, q represents the quarter, $play_A$ is the name of the player who finishes the possession, $play_B$ is the type of play being run by the offense. Proceeding $play_B$, the remainder of the play variables listed alphabetically from C to H follow the rest of the play sequentially through until the second last non-zero play. The second last non-zero play of an observation describes the type of shot in the case of a shot or foul and the last non-zero play describes the outcome of the possession. *Away score*, *home score* and *margin* represent the score of the game at the time of the possession. *Year* represents the season in which the game took place, where games occurring in the second half of the calendar year are grouped with the preceding year to correspond to the USports season cycle. *pts* describes how many points were gained from the possession. In the case where a player is fouled, the number of points made in free throws is assigned to the observation under *pts*. The play variables do have a meaningful effect on free throw shooting, which is reflected in the data. For example, Post-Up plays which result in a foul have a lower percentage of free throws made than a foul occurring on a PnR Ball Handler play. *Preamble* is a binary variable which is 1 when there was a stretch of play-by-play eliminated from the observation describing who delivered the player the ball and 0 if there is no information removed.

Specifically, there is a play type labelled from a list of: Spot Up, Post Up, Transition, Cut, Pick and Roll: Roll Man, Pick and Roll: Ball Handler, Isolation, Offensive Rebound, Hand Off, Off Screen and No Play Type. This along with shot categories based on distance from the basket: 3-Point Shot, 17' to Three Point Line, Inside 17' and Basket, provide some granular positioning metrics. In the Synergy Sports database there is more novel data which is specific to the play type. For example, a Pick and Roll: Roll Man play will also have information about whether the Roll Man rolled to the basket or if they popped, or a Transition play will state where on the floor the ball was passed to (left, right, trailer) or if the ball handler shot the ball themselves.

Another conceptual question to ponder in creating this expected points metric is whether or not to include fouls and turnovers. Not doing so, neglects shot attempts where a player is fouled. Based on the independent variables used, it is likely we can predict with some measure of accuracy, the likelihood of being fouled. A failure to model this will not account for differences in the foul gathering process of different players. Players who go to the line more than expected will be undervalued and players who go to the line less than expected will be overvalued when building an expected model. Variables accessible in this analysis will also do a fine job at modelling the likelihood of a foul, thus the models will not be binary miss-or-make models.

Synergy Sports play-by-play data, does tell us a lot about what happens before a shot. This can inform us about the outcome of the shot attempt, but it can also tell us about whether or not a shot is likely to happen. Filtering out the noise in the turnover data would be a great benefit to any metric built on top of this model. Therefore, turnovers are considered in-scope

for the expected points model and now changes the scope of the model from shot-quality to possession quality based on the player who finishes the possession.

Possessions where non-shooting fouls result in free throws, by bonus implications or intentional fouls, are excluded from the modelling processes. These possessions tell us little about the talent of the players on the court in that possession and more about what proceeded the situation. Information which can be gained from these possessions is already addressed by modelling the previous possessions.

Complications exist when creating this same expected points model for individual defenders. There is a clear endpoint created by a decision made by an offensive player on most possessions. The instinct to mimic this on the defensive side would be to assign the person who is guarding the player who ended the possession the credit. This is for the most part what we will do when creating a bottom-up statistic because there are players who are able to suppress field goal percentages of the players they are guarding and not foul them. However, since the play-by-play is capturing what is happening on the offensive side of the ball, it likely will not capture the tactics individual defenders take to refrain from fouling and thus will incorrectly regress them to the mean. Additionally, there is the challenge posed by the fact that defending a player well off the ball will *prevent* them from even recording a shot, which goes unnoticed by the bottom-up approach. Lastly, the problem of help defense and team rotations affects the way "nearest" defender is tracked in Synergy's database. So, if a defender rotates onto a player who shoots or drives the ball, they are credited with the outcome despite the fact the process which led to it was largely out of their control.

It should be noted, not all play types are assigned a primary defender in Synergy, namely, No Play Type, Cut, Offensive Rebounds and Transition. This is justified since these situations rely on the team to work as a unit defensively more so than other plays where an individual defender can be apportioned more of the credit, like an isolation or post-up. However, this leaves a severe hole in the analysis of defenders, since transition is such an important and efficient way for teams to score and conversely, prevent points.

This said, there should be less emphasis put on the bottom-up portion of the defensive analysis and much more on the top-down. Good off-ball defense and timely rotations are designed to suppress the other teams points and ultimately result in lower quality shot attempts and more turnovers for the opponent. These skills are not captured in any bottom up metric but likely could be captured over large samples of the top-down metrics. Therefore, the top-down process oriented defensive metrics will serve as excellent tools to fill in the gaps left by bottom-up numbers for evaluating individual defensive true talent.

The crucial note about the structure of the data is that when the data are separated by the play type, $play_B$, the plays recorded in the possession follow a standard format. This means that each observation can be grouped by play type and the following play values can be used as categorical variables to predict the number of points. Note for example, when $play_C$ is Ball Handler, that means something materially different about the play being run when $play_B$ is Transition or PnR Ball Handler. This is because the person who has the ball has different options available to them in the subsequent moments if the preceding play type is different, therefore using them in the

same model is misguided. Moreover, each possession does not necessarily have the same number of plays recorded in the play-by- play. This means if we wish to use the data, not all observations can be used to train the same model since there are missing variables.

Below in Tables 3.2 and 3.3 are exploring the relationship between play type ($play_B$), point scoring, shot attempts plus turnovers and the length of the play-by-play entry. Note, the work is done for the women and subsequently the men's data. Table 3.2 details the play type sorted by frequency, n, and the percent column describing the proportion of total plays run. pts, is the average number of points scored when the play is run and finally, the variance describes the variance in the number of points scored for the play type. Over 90 percent of the variance in the variance column can be explained by the turnover rate, and percentage of 3-point shots for that play type. Higher turnover rates means lower variance, since those are automatic zero point possessions. Higher percentage of 3 point shot attempts increases variance since it offers the widest spread of possible points on a possession, 0 or 3.

Table 3.2: Women's Play Type Summary

Play Type	n	pts	variance	percent
Spot-Up	125079	0.742	1.388	25.26
Transition	92355	0.83	1.157	18.65
PnR Ball Handler	50902	0.622	0.996	10.28
Cut	48866	0.914	1.004	9.87
No Play Type	44115	0.272	0.449	8.91
Post-Up	36748	0.736	0.935	7.42
Offensive Rebound	24999	1.04	1.003	5.05
Off Screen	23784	0.731	1.314	4.8
ISO	19939	0.651	0.952	4.03
Hand Off	15012	0.684	1.173	3.03
PnR Roll Man	13358	0.752	1.063	2.7

The more detailed Table 3.3 offers the same statistics for play types given the length of the play-by-play line. Specifically, the subscript is the number of plays after the play type ($play_B$) until the zeros begin. Each of these play types and play lengths will be used to create an individual model for expected points since each observation fits into one and only one of these categories. Since these are the data sets which will be individually modelled, the descriptive statistics are useful in understanding which models will have an outsized influence on the overall expected points model and deserve more attention.

Table 3.3: Women's Play Type Detailed Summary

Play Type	n	pts	variance	percent
Spot-Up $_4$	87303	0.817	1.616	19.12
Cut $_2$	48841	0.914	1.004	10.7
Transition $_3$	47624	1.081	1.016	10.43
Post-Up $_5$	29768	0.814	0.977	6.52
Spot-Up $_3$	27090	0.773	0.959	5.93
Offensive Rebound $_4$	23967	1.06	1.009	5.25
PnR Ball Handler $_5$	23660	0.63	1.033	5.18
Transition $_5$	19085	0.909	1.753	4.18
PnR Ball Handler $_4$	17809	0.614	0.888	3.9
Transition $_2$	16911	0.09	0.146	3.7
ISO $_4$	15814	0.688	0.986	3.46
Off Screen $_6$	12842	0.826	1.652	2.81
Spot-Up $_2$	9386	0.06	0.1	2.06
PnR Ball Handler $_6$	8491	0.662	1.172	1.86
Transition $_4$	8363	0.752	1.187	1.83
PnR Roll Man $_2$	6130	0.828	0.971	1.34
Off Screen $_4$	5666	0.81	0.961	1.24
Hand Off $_4$	4992	0.804	0.972	1.09
PnR Roll Man $_5$	4763	0.712	1.267	1.04
Hand Off $_6$	4410	0.838	1.696	0.97
No Play Type $_2$	4246	1.004	1.009	0.93
Post-Up $_4$	4144	0.532	0.783	0.91
Hand Off $_5$	3488	0.684	1.159	0.76

Off Screen $_5$	3097	0.663	1.042	0.68
ISO $_5$	3038	0.622	0.923	0.67
No Play Type $_3$	2157	0.486	0.905	0.47
Off Screen $_3$	2149	0.06	0.101	0.47
Hand Off $_3$	2108	0.081	0.138	0.46
PnR Roll Man $_4$	1576	0.723	0.921	0.35
Post-Up $_2$	1616	0.303	0.419	0.35
Post-Up $_3$	1218	0.091	0.15	0.27
ISO $_3$	1040	0.202	0.311	0.23
No Play Type $_4$	1063	0.673	1.252	0.23
Offensive Rebound $_2$	1010	0.563	0.637	0.22
PnR Ball Handler $_3$	858	0.232	0.344	0.19
PnR Roll Man $_3$	886	0.505	0.752	0.19

Lastly, a frequency diagram of the play types with colour to describe the distribution by the shot type: 2-point field goal attempt, 3-point field goal attempt, turnover, free throws, 2-point field goal make and foul, and 3-point field goal make and foul.

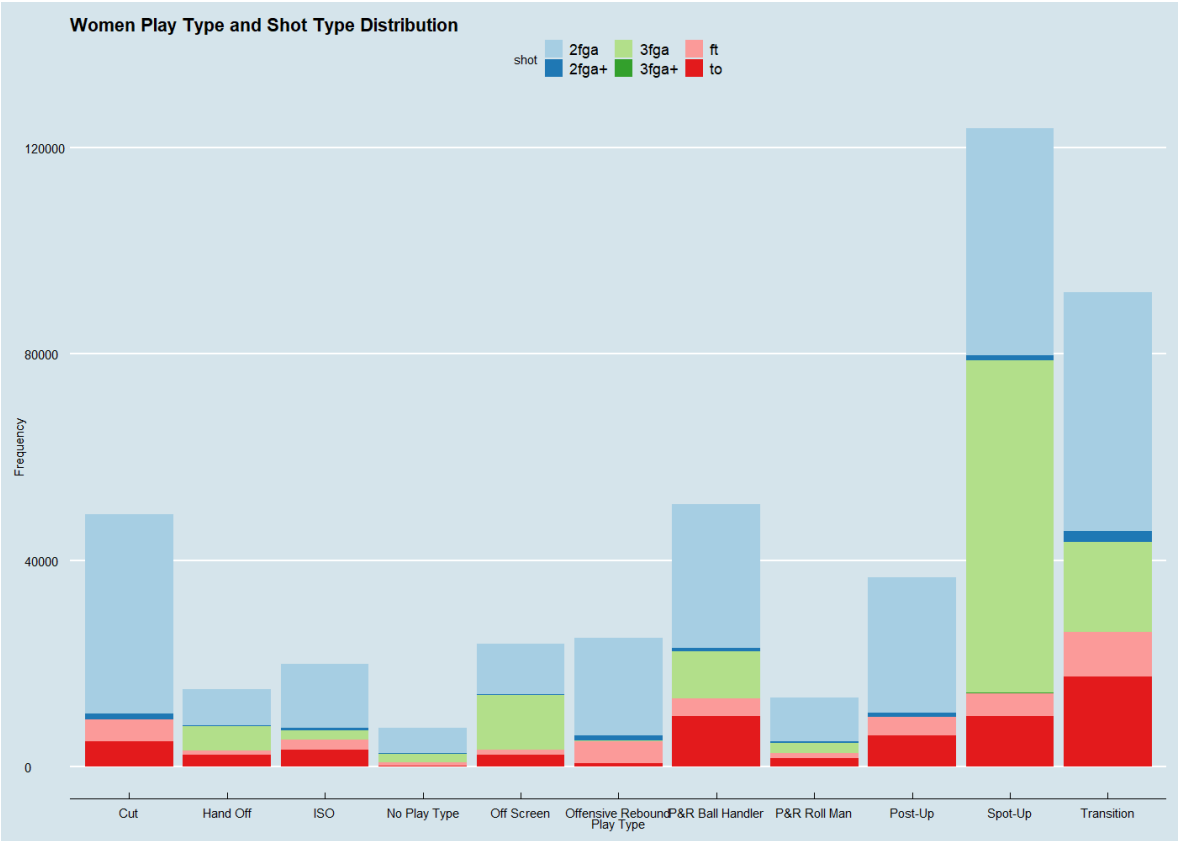

Figure 3.1: Women's Play Type and Shot Distribution

The same tables and figures are displayed for the men's data and illustrates some of the differences between the stats in the two leagues. Notably, the average points and variances in outcome for a given play type are consistently higher for the men because shooting percentages in the men's game are higher (and specifically closer to 50 percent). The distribution of play types is also slightly different, notably more Pick and Rolls are run and a higher percentage of them end with the roll man in the men's game.

Table 3.4: Men's Play Type Summary

Play Type	n	pts	variance	percent
Spot-Up	127880	0.856	1.579	25.93
Transition	95841	0.965	1.296	19.43
PnR Ball Handler	49256	0.73	1.152	9.99
Cut	42822	1.074	1.03	8.68
No Play Type	34324	0.373	0.596	6.96
ISO	31355	0.731	1.085	6.36
Post-Up	31338	0.762	0.962	6.35
Offensive Rebound	25528	1.176	1.019	5.18
Off Screen	22362	0.828	1.469	4.53
PnR Roll Man	16628	0.912	1.267	3.37
Hand Off	15848	0.776	1.344	3.21

Table 3.5: Men's Play Type Detailed Summary

Play Type/ Length	n	pts	variance	percent
Spot-Up $_4$	92580	0.939	1.816	19.93
Transition $_3$	52397	1.18	1.031	11.28
Cut $_2$	42803	1.074	1.03	9.21
Post-Up $_5$	25071	0.834	1.001	5.4
Spot-Up $_3$	24708	0.881	1.029	5.32
Offensive Rebound $_4$	24681	1.198	1.018	5.31
ISO $_4$	23075	0.772	1.127	4.97
PnR Ball Handler $_5$	21773	0.753	1.238	4.69
Transition $_5$	20086	1.071	2.03	4.32
PnR Ball Handler $_4$	19199	0.69	0.969	4.13
Transition $_2$	14689	0.116	0.19	3.16
Off Screen $_6$	11876	0.968	1.849	2.56
Spot-Up $_2$	9179	0.082	0.136	1.98
Transition $_4$	8347	0.888	1.562	1.8
PnR Ball Handler $_6$	7293	0.827	1.433	1.57
PnR Roll Man $_2$	7067	0.982	1.016	1.52
ISO $_5$	6585	0.705	1.053	1.42
PnR Roll Man $_5$	6300	0.901	1.677	1.36
Hand Off $_6$	5215	0.921	1.861	1.12
Off Screen $_4$	4740	0.899	1.039	1.02
Hand Off $_4$	4447	0.918	1.056	0.96
Post-Up $_4$	4369	0.578	0.827	0.94
No Play Type $_2$	4313	1.072	1.02	0.93

Hand Off $_5$	3976	0.812	1.339	0.86
Off Screen $_5$	3519	0.748	1.177	0.76
No Play Type $_3$	2306	0.544	1.091	0.5
Hand Off $_3$	2196	0.083	0.138	0.47
Off Screen $_3$	2191	0.058	0.102	0.47
PnR Roll Man $_4$	1975	0.89	1.049	0.43
ISO $_3$	1619	0.282	0.416	0.35
PnR Roll Man $_3$	1280	0.614	0.856	0.28
No Play Type $_4$	1074	0.76	1.487	0.23
Post-Up $_2$	1030	0.371	0.477	0.22
PnR Ball Handler $_3$	907	0.283	0.417	0.2
Post-Up $_3$	867	0.101	0.151	0.19
Offensive Rebound $_2$	834	0.543	0.613	0.18

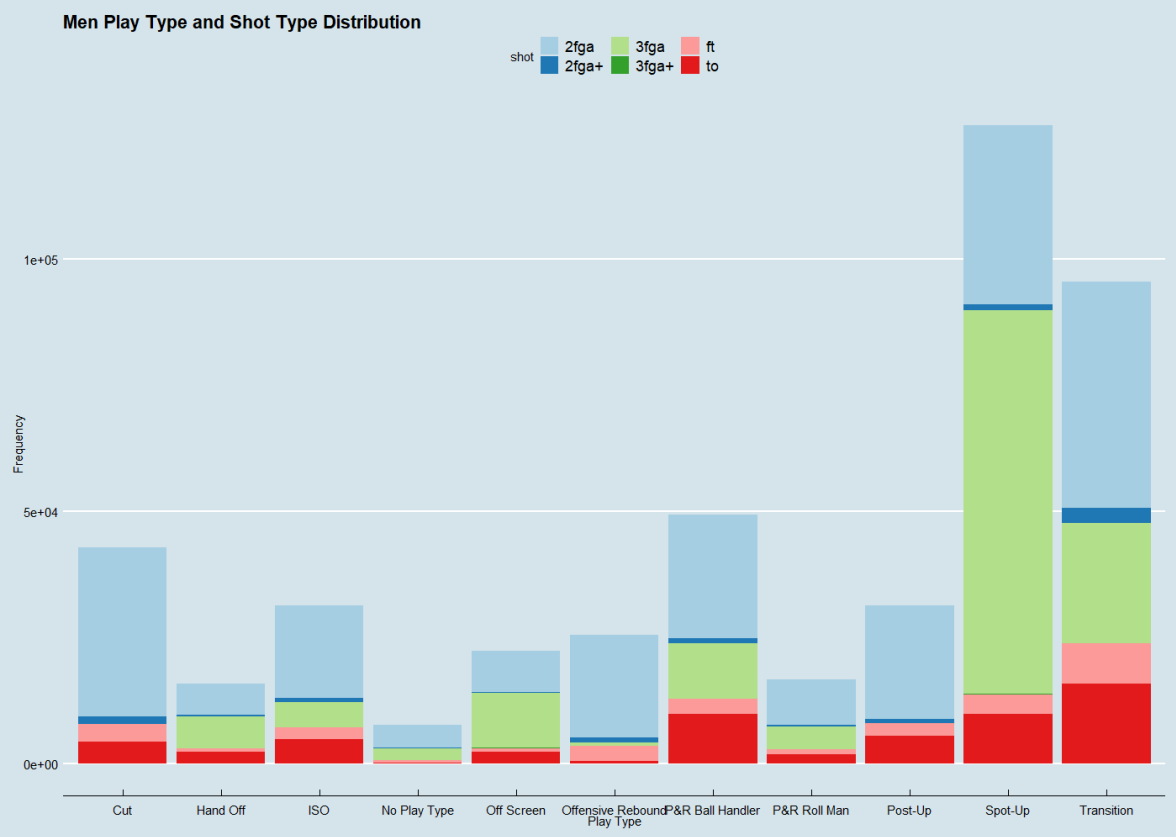

Figure 3.2: Men's Play Type and Shot Distribution

Chapter 4

Methods

4.1 Separation of Leagues by Gender

As currently constituted USports separates basketball along the binary of men's and women's teams. One could combine the two sets of data and create an indicator variable for the league from which the game play is sourced. However, the intent of the project is to create models of player performance by accounting for how players relate to one another as opponents and teammates. Since the leagues are currently segregated, men and women don't relate to each other as teammates or opponents, so each model described is built once for each of the two genders represented in USports. Building a model which contained data from both leagues may be able to tell us more about the difference between the two leagues but that is considered out-of-scope for this thesis

4.2 The Problem of Try

One the biggest critiques I received in my time as a statistical analyst for the Carleton Ravens basketball teams from coaches was: statistics compiled when the effort level of the players on the court is low are meaningless. This is commonly referred to as, garbage time[1], or the problem of try. The critique requires a couple of things to be true: first, players actually give more or less effort in certain situations and second, their level of effort changes how the game is played. While the quantification of each player's effort level is not yet possible, the latter contention is possible via proxy, namely, how the game is played changes based on the score and time remaining in the game. When effort level is discussed with people making the critique, the main contributing factor mentioned is the amount of doubt one can have about the final outcome of the game. Therefore, the level of uncertainty of outcome in a game, based on time remaining and score, can be used to estimate the level of try expected for the players on the court. *Win probability* and *leverage* are ideas imported from early baseball analytics work by Tom Tango[2] which will serve us well here.

Win probability is exactly as it sounds, an attempt to model likelihood of each team winning the game at any moment in the game. The basic stripped down model here accounts for scoring margin between the teams and the time remaining in the game. Adding the effect of home court in basketball is low cost and benefits the model. More variables could be added such as "quality of team measures" like scoring margin on the season and quality of opponents or player performance measures, to improve the performance of

[1] Oliver, Dean. Basketball on Paper

[2] Tango, Tom. "Crucial Situations."

the win probability model[3]. *Leverage* builds off of this model and purports to reflect how important a moment in the game is based on how much the win probability can possibly change into the next moment.

$$Leverage_m = Var(WP_{m+1}|m) \tag{4.1}$$

Where m is defined as a moment and WP the win probability. In baseball, a moment and its subsequent can be defined quite precisely since there is a finite number of outs each team gets, arranged in innings and scoring is limited by the finite number of base runners a team has on during a plate appearance. The moment is the plate appearance and the win probability can be estimated by analyzing the score, outs, inning, and position of base runners (and home team, team/ player quality etc.). There is a finite number of new situations that can occur based on a single plate appearance and the variance of those potential outcomes in terms of win probability, defines leverage. In a plate appearance, the number of outs possible is known: 0, 1, or 2, and the number of outs for a team is clearly defined within an inning and a game. If a plate appearance occurs with the bases loaded and 2 outs in the bottom of the ninth inning with the home team down by 1, the leverage is at its peak since the three outcomes of that plate appearance that are possible in the next moment are win probability of 0, 1 or slightly greater than 50 percent. This contrasts a situation where one team is up 10 runs with 1 out remaining and no base runners on, since no matter what occurs in the plate appearance the win probability will change almost not at all.

[3] Paine, Neil. "How Our NBA Predictions Work."

Basketball poses a challenge since from one moment to the next is not so easy to define. The relationship between outs and plate appearances is very clean in baseball and makes separating one moment from the next possible. Time and possessions in basketball are not so simple. There is a maximum number of seconds in a possession, but measures of time are continuous, not categorical. Further, a team can end a possession and immediately begin a new one with an offensive rebound, which puts the modelling of potential WP_{m+1} states in an endless recursive loop since time of possession can only approach 0 time and be as long as the entire game. With this in mind, a moment is defined in terms of time and disregards the number of possessions which can happen in these intervals. This allows us to clearly define how many moments will occur in a game and how close we are to the ultimate decision of a winner. Specifically, define a moment, m, by the time into the game, the scoring margin and home court advantage. Then the set of $WP_{m+1}|m$ is defined as all the win probabilities in the subsequent moment given m, which can be deduced from the observed training data. This conditional distribution is then smoothed by using the conditions which define m to predict $Leverage_m$.

4.3 Sample Size

Sample size is a problem which can plague any statistical investigation and is top of mind in this case. In some portions of the investigation there is no problem at all. For example, the possessions used to model the expected number of points for both men and women is approximately half a million. Therefore, creating training, validation and testing sets for this data will not

be a concern. However, using the outputs of those models to make inference on the talent of players will be. Some players play a limited number of possessions, which makes the task of determining their true talent an issue. Moreover, a typical USports team will play somewhere around 25 games a season which are logged in Synergy's database. When compared to an NBA season, where metrics like BPM and RAPM were designed, 25 games is less than a third of the length of the season. There is some research which shows the NBA season captures information about team talent very quickly[4] relative to other sports leagues, likely because of the large number of possessions, each with a scoring outcome, which happen in every game. On balance though, there is some evidence that shows multiple seasons of NBA player data are necessary to find a good stable estimate of player talent[5]. This is undeniably a challenge especially considering there is a maximum number of seasons, five, a player is allowed to play in USports.

This fact will reduce expectations around accuracy of estimates, so for each given metric, focus will be on how well it predicts future performance. Techniques to mitigate this problem will include using multi-year samples of player data and incorporating other non-playing data which predicts future performance. For example, position, height, and weight (for men only, unavailable for women) have predictive value on individual statistics over and above the historical statistics and will be used to supplement the player evaluation statistics.

[4]Tango, Tom M., et al. The Book: Playing the Percentages in Baseball.
[5]Myers, Daniel. " About Box Plus/Minus (BPM)."

4.4 Changing True Talent

Expanding on the commentary about the limited number of seasons a player can play in the USports system, a lot of growth happens from year to year. When a player plays a certain season of their eligibility (first, second, third, etc.) there is a variance in the age, experience and maturity of the player. This makes pinning down a player's true talent at a given moment very difficult. There is not a typical aging curve which one might see in professional sports (which looks like a parabola)[6], but rather something that looks like an exponential increasing curve in the best of cases and a logarithmic curve in others. Furthermore, there is a survivorship bias effect which artificially increases the amount of increase in player growth in upper years. This bias occurs since players who perform worse in early years in their career are less likely to continue playing into their senior seasons as compared to higher performing players.

One way to combat the issue of changing true talent is to narrow in at specific pivotal points in a player's career and evaluate them at these points, say the end of a season. However, this increases the problem of small sample sizes by shrinking them even further. Therefore, the larger concern will win out and all previous seasons of data collected on an individual player will be used to evaluate their current true talent. Having the metric be stable for a player over time is the concern being addressed. Therefore the predictability of the metric will be used as the performance measure in a cross-validation which test different weights on previous seasons.

[6] Weinberg, Neil. "The Beginner's Guide To Aging Curves."

4.5 Coaching and Teammate Effects

Teammate effects are a massive consideration in a sport as interconnected as basketball. Who a player plays with can have a massive effect on their top-down numbers like ON/OFF rating, especially if there are large correlations between when a set of players on a team are either on or off the court. Given a perfectly random rotation of players on the court at a given time, creating top-down statistics would be a simple task, but that is not the reality. Coaches choose to use players as substitutes for one another or pair players together who work in some aspect of the game. This introduces the problem for a regression model of multicollinearity. Multicollinearity occurs when 2 or more of the independent variables in a multiple regression model are highly linearly related. So while the estimates of the regression coefficients for the collinear predictors work well in sample to predict the dependant variable, if the collinearity does not continue apace in the future, there exists an out-of-sample prediction problem. Since the task at hand is creating *individual* player evaluation metrics, we want to isolate the effect of an *individual* player. Commonly the solution to this problem, in statistical literature (and utilised by the architects of RAPM), is ridge regression[7]. The details of this process are discussed in depth in Chapter 4.7.

Something more which is not being accounted for is the synergies between teammates which either make them better or worse together. These are teammate specific effects which will manifest for some teammates and not others, which appear as additional value assigned to the players which will not appear in other contexts. Modelling these anomalous behaviors

[7] Gruber, Marvin H. J. Improving Efficiency by Shrinkage.

is an extraordinarily interesting topic of research which requires granular specificity while being cautious of overfitting the data. Moreover, there are coaching specific effects which go largely unmodelled. This is the idea that a coach's decision around plays, spacing, rotations, transition and philosophical focus will affect how a player performs. Luckily, player movement from team to team and therefore coach to coach is less a part of the system of USports basketball than professional basketball, but its role is not zero. To combat this effect, which of the 48 USports team the player is on will be considered in the development of metrics. Implicit in this is the stylistic differences in play between conferences (AUS, RSEQ, OUA, Can West) which may occur due to traveling and schedule discrepancies will also be accounted for. Note however, teammate synergistic effects and coach game plan and philosophy effects largely fall outside of the scope of this project and should be considered fertile areas for future research.

Specifically with respect to team effects, Box Plus-Minus (BPM) accounts for the pace the team plays relative to the league and the shooting metrics of the team. A player is compared against the shooting abilities of the team, as a method to counteract system effects. The role a player plays in a team's offense is also adjusted for in the BPM statistic. Notice we assign value to a player who assists a basket. It is natural then to reduce the value we assign to the player *scoring* that basket. The key here is players who can generate more of their baskets unassisted, all things being equal, are more valuable to a team than a player who generates more of their baskets through assists. This quality allows a player's BPM to be less dependant on circumstances of which they cannot control and thus tells us more about the true talent of the player. Position plays a large part in this equation, where point guards do

much of the work of creating and centers do more of the finishing. Therefore, the listed position is used as a proxy for role of a player on the team and is adjusted based on his or her demonstrated abilities with respect to creating, finishing, rebounds and personal fouls committed.

4.6 Neural Networks and Random Forests

With the data structure as described, we model the expected points for each of the unique play types and lengths, 37 for each gender. Each data set is trained and tested using a series of different neural network and random forest models to find which configuration produces the lowest test error. Neural network and random forest models were chosen as they are leading machine learning algorithms which can produce reliably accurate and precise models. The Classification and Regression Tree (CART) model provided by a random forest seemed especially attractive to fit the structure of the data. The play-by-play follows a step-by-step logic between variables where the possible values of a variable are indeed contingent upon what comes before it. A decision tree structure which can create specific branches of the categorical input models some of the data sets very well. A little more background on random forests is warranted to justify the decision.

A random forest is built on a foundation of many decision trees, which can be modified to regress a numeric outcome variable or classify a categorical one. A decision tree is a supervised learning algorithm and is composed of internal decision nodes which recursively split the data[8]. In a multivariate situation each decision node searches specific demarcation points along

[8]Alpaydin, Ethem. Introduction to Machine Learning.

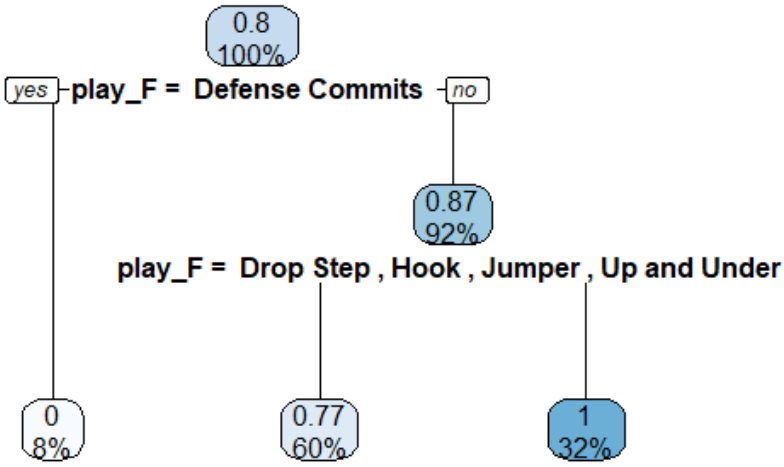

Figure 4.1: Decision Tree

each of the the independent variables and finds which point creates the two groups, called leafs, which are the most homogeneous within and heterogeneous between. Each node in the tree uses a decision function which attempts to measure the impurity of each node produced by a potential split. Impurity is meant to capture how similar the contents of a leaf are to each other. In the case of a classification tree, there are many impurity measures which can be chosen such as Gini index or entropy, but mean squared error is often used in the case of regression. These trees follow an If-Then format as you move down the tree and can be visualised as seen in Figure 4.1.

The data for men's post-up$_5$ is used and in this tree. The root node uses a

the average number of points in the data of 0.8 on the first decision node and tests if the value in the variable $play_F$ is *Defense Commits* (it is 8 percent of the time). In that case, the expected points are 0 and conversely if $play_F$ is not *Defense Commits* the next decision node tests whether or not $play_F$ is equal to *Drop Step*, *Hook*, *Jumper* or *Up and Under*. It is one of those four options 60 percent of the time and the expected points in that case is .77 while, if it is not, the expected points moves up to 1. This tree structure can explain the mechanics of a basketball play and thus made it a natural choice for this exercise.

Players are running plays but this does not happen in a vacuum. The skills of the players involved in the plays influences the decisions players makes. Therefore, the player who concluded the play is considered as a factor in the modelling process since it affects the decision making of player's which is one thing we are trying to evaluate. This results in a very large set of binary indicator variables to be considered by the algorithm and can lead to very long and wide trees. This sort of depth in a tree can result in overfitting when final leaves have only a handful of observations or less from which they are drawing. Random forests offer a few countervailing methods to protect from this type of overfitting.

Firstly, based on the principle from which the technique draws the "forest" moniker, it is a collection of many trees, the number of which is tuned by cross-validation. Random forests use a technique called bootstrap aggregation which is a type of sampling which repeatedly takes different samples of the training data with replacement. Note however this bootstrapping is not necessary and is a hyperparameter which is tuned and when it is not done, the trees are trained on samples of the training data without replacement.

These samples are used to create a decision tree on some subset of the given features available in a process called feature bagging. Feature bagging is the process of sampling some number of features (tuned in cross-validation) with replacement into a series of many bags. The feature bags are then used to create a decision tree with the many samples of data and are validated against the remaining data. This validation will dictate the out-of-bag (OOB) error which is used to choose the optimal construction of the tree given that feature bag. Ultimately the predictions of the forest become an average of all the trees of which the forest is composed. This results in a loss of intrepretability but is a strong bulwark against over-fitting.

Another technique to combat against overfitting is pruning. Pruning is the technique of cutting off the branch of a tree at some point, so that all node decisions subsequent to the cut are disregarded. There are a few ways to prune the tree that are derived from explicit rules prior to the tree's construction. One is to limit the depth of the tree, i.e. that one observation in the training set can only pass through so many decision nodes. Another way is to specify an impurity limit, i.e. if a decision produces nodes which has a high enough mean squared error, the decision should be forgotten and the node ought to be a final leaf. Finally, a specific number can be set as the minimum number of observations acceptable to be on a node and if a rule produces a number lower than that specified, it shall be disregarded and the node will become a final leaf. All three of the pruning techniques were used in this modelling process since the number of independent variables is so large and the data is quite sparse.

The other supervised machine learning technique used to model the data is a neural network. The mechanics of the algorithm begin with the

inputs of some training set along n dimensions. The values of these n dimensions for each the k observations in the training set are sent through a chosen N number of layers each with a number m_i of neurons where $i = 1, ..., N$, each of which has a bias neuron associated with it. At the end of the N layers of the network, in the case of regression, a final prediction for the observation is offered as a function of all the composite neurons. On each neuron there are incoming values from the previous layer's neurons which are multiplied by some weights in a linear fashion added to the bias neuron associated with that hidden layer and applied to an activation function to create a new value which will then be fed to the next layer. Resulting in the j^{th} neuron in the $(i+1)^{th}$ hidden layer, y_j, following the equation,

$$y_j = F(\underline{X}^T \underline{W} + b_j) \tag{4.2}$$

where F is the activation function, $X = [x_1, ...x_{m_i}]^T$ the incoming neurons from the i^{th} layer, $W = [w_1, ..., w_{m_i}]^T$ the weights associated with connecting X to y_j and b_j as the bias introduced between the i and $(i+1)^{th}$ layer associated specifically to y_j. The activation function F and number of neurons in each layer $m_i : i = 1, ..., N$ are hyperparameters which are cross-validated in the training loop while the number of hidden layers N is fixed. The beauty of the algorithm comes from its ability to hone in on what the best set of W and b_j are necessary for each y_j to ultimately produce the best estimate at the end of the network. There will be a total of $n * m_1 + m_N * 1 + \sum_{i=1}^{N-1} m_i * m_{i+1}$ weights in a fully connected network and $\sum_{i=1}^{N} m_i$ biases that must be tuned to create an output. The initial values for weights and biases must be chosen and often times are selected at random. But after

the first observation of the training set passes through the initially chosen weights and biases, a chosen loss function is used to measure performance of the network.

A loss function calculates the error between the produced output and the true value, in our case the mean squared error is the chosen loss function. The loss is then computed for all the observations in the training set and an optimization function is used to modify the weights and biases in order to reduce the loss of the network in the next iteration. Optimization functions rely on and vary mildly from stochastic gradient descent. Gradient Descent (GD) is the process of taking all the weights and biases as an input and attempting to adjust them in the direction and at the magnitude that the loss function and learning rate dictate. However, this process takes an enormous amount of computational power and time to complete. In order to ease that concern Stochastic Gradient Descent (SGD) is used instead. SGD works by taking batches of the data to move quickly calculate the GD on a smaller sample of data. This may not be the optimal descent, but the process is iterated and does eventually move into a local minimum more quickly then if one attempted to calculate the GD on the entire data set. The learning rate is the amount of movement the weights and biases can have during the propagation process. Since the function we are trying to minimize exists in an extraordinarily large dimensional space, namely $n*m_1+m_N*1+\sum_{i=1}^{N-1} m_i*m_{i+1}+\sum_{i=1}^{N} m_i$ dimensional, it is very susceptible to descending into a local minimum which is not the global minimum. To counteract this, the learning rate can be changed to either allow more movement when trying to find the global minimum or less. Optimization functions like Adaptive Moment Estimation (Adam) and Root Mean Squared

Propagation (RMSprop) vary the learning based on the values of the most recent movement[9].

The key to SGD is backpropagation which allows the network to calculate the effect a change in a neuron has on the loss of the network[10]. Let J be the loss function, let T_j be the activation function applied to the j^{th} neuron and let $\underline{s_j}$ be the inputs into the j^{th} neuron in a hidden layer. If we wish to find the relative change in J with respect to x_j we can use the chain rule of derivatives and find,

$$\frac{\partial J}{\partial x_j} = \frac{\partial J}{\partial T_j} * \frac{\partial T_j}{\partial x_j} \tag{4.3}$$

$\partial T_j / \partial x_j$ is the derivative of the activation function since it is a function of a single variable and $\partial J / \partial T_j$ is the gradient of the loss function with respect to the activation. Note $T_j = T(\underline{w_j} * \underline{x_j^T} + b_j)$ where $\underline{w_j}$ and b_j are the weights and the bias. This means we have connected the activated neuron T_j to $\underline{x_j}$, the input into the j^{th} neuron. This formula allows us to calculate the current gradient of J with respect to the weights and biases, which in turn allows us to manipulate the weights and biases in order to lower that gradient. Therefore what makes backpropagation so powerful chain rule applied to the composition of readily differentiable operators.

An additional hyperparameter considered in this exercise is a hidden layer in the middle of network, which randomly drops some proportion of the neurons that connect two layers. This means the dropped neurons' contributions to the activation will not be counted on the forward pass and

[9]Arcos-García, Álvarez-García. "Deep Neural Network for Traffic Sign Recognition Systems: An Analysis of Spatial Transformers and Stochastic Optimisation Methods."

[10]Rebala, Gopinath., et al. An Introduction to Machine Learning.

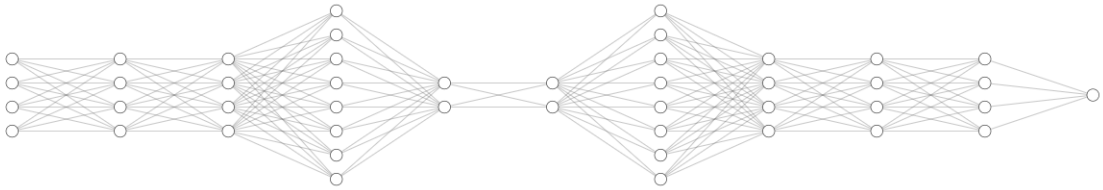

Figure 4.2: Neural Network Architecture

their weights and biases will not be updated on the backpropagation[11]. This is done as a regularization technique which pushes back on the tendency of supervised learning algorithms to overfit the training data.

The design of the neural networks is also important since the number of hidden layers and the number of neurons in the layers are hyperparameters. After some experimentation a network with 10 hidden layers centered around a dropout layer was chosen. The first 3 and last 3 layers contained 64 neurons, the 4^{th} and 7^{th} layers contained 128 neurons and the middle layer which is subject to dropping out was tested with 32, 64 and 128 neurons, while the dropout rate was tested at between a quarter and a half of neurons. Figure 4.4 visualizes this network where each node represents 16 for the sake of compactness. Note that the number of nodes in the 5^{th} and 6^{th} layers varies based on the model due to hyperparameter tuning and the connections between the nodes in the layer between the two are subject to random dropout on the forward pass.

[11] Srivastava, Nitish et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting."

The Rectified Linear Unit (ReLU) and sigmoid activation functions were tested while the optimization functions Adam, RMSprop and the standard SGD were tested. The sigmoid function is $f(x) = 1/(1+e^{-x})$, while the ReLU function is $g(x) = max(0, x)$, both of which attempt to scale the input coming from a neuron into a number which better represents the neurons relative activity or contribution to the network. We also test the utility of a Exponential Linear Unit (ELU) for our models where the function

$$h(x) = \begin{cases} x & \text{x>=0} \\ \alpha(e^x - 1) & \text{x<0} \end{cases}$$

allows negative value to take on a non zero value output in accordance to some exponential function dictated by a chosen alpha.

The rationale behind choosing this technique is its ability to produce useful approximations of any functions. In fact a single hidden layer neural network given an arbitrary bounded and non-constant activation function can approximate any function[12].

4.7 Regularized Adjusted Plus-Minus and Ridge Regression

Regularized Adjusted Plus-Minus (RAPM) uses ridge regression as a method to combat the multicollinearity problem due to the fact that teammates are the only ones who can play with each other despite all players being used as variables in the regression. This section is devoted to explaining the differences between ridge regression and ordinary least squares regression and why it is useful in our case.

[12] Hornik, Kurt. "Approximation Capabilities of Multilayer Feedforward Networks."

First we will consider ordinary least squares regression of the form:

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon} \tag{4.4}$$

Such that $\underline{Y} = (y_1, ..., y_n)^T$, $n \epsilon \mathbb{Z}$ is the vector of n observations of the response variable. Next, $\beta = (\beta_1, ..., \beta_p)^T$, $p \epsilon \mathbb{Z}$ is the p-dimensional vector of regression coefficients. X is then the $n \times p$ design matrix $X = (x_1, ..., x_n)^T$ where each x is observed at each of the p independent variables, $x_i \epsilon \mathbb{R}^p$ $\forall i = 1, ..., n$. Finally, $\epsilon = (\epsilon_1, ..., \epsilon_n)$ are n independent and identically distributed random errors. In the context of the RAPM equation, each y_i is the difference in the two teams' scores (scoring margin) for a continuous time period in a game with no changes in the 10 players on the court. x_i contains an observation for each of the p players who could be on the court, which is 0 if a player is not on the court and is a value weighed by which side of the margin the player is on, the leverage of the situation and how recently the game took place (weights subject to cross-validation). Therefore, we estimate the $\underline{\beta}$ coefficients to be a reflection of the true talent of each of the players they represent. The least squares estimates (LSE) $\underline{\hat{\beta}}$, are then estimated by minimizing the residual sum of squares (RSS) equations given by:

$$\min_{\underline{\beta}}\{(\underline{Y} - X\underline{\beta})^T(\underline{Y} - X\underline{\beta})\} = \min_{\underline{\beta}}\{L(\underline{\beta})\} \tag{4.5}$$

such that $L(\underline{\beta}) = \underline{Y}^T\underline{Y} - 2\underline{\beta}^T X^T Y + \underline{\beta}^T X^T X\underline{\beta}$. To do this, we solve for $\underline{\beta}$ in the equation

$$\frac{\partial L(\underline{\beta})}{\partial \underline{\beta}} = -2X^T\underline{Y} + 2X^T X\underline{\beta} = 0 \tag{4.6}$$

and conclude

$$\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{Y} \tag{4.7}$$

Moving back to the original regression equation, (4.4), we suppose the error $\underline{\epsilon}$ has $E(\underline{\epsilon}) = \underline{0}$ and $E(\underline{\epsilon}\underline{\epsilon}^T) = \sigma^2 I_n$ for some $\sigma^2 > 0$ and I_n the $n \times n$ identity matrix. Under this constraint it can be shown the variance of the LSE is

$$Var(\underline{\hat{\beta}}) = \sigma^2 (X^T X)^{-1} \tag{4.8}$$

Note referring back to the problem of multicollinearity, where one or more of the independent variables of X are highly correlated, one consequence of this is that information is being captured multiple times. Thus one of the associated eigenvalues for highly correlated variables will be close to 0 and equal in the case of perfect correlation[13]. With this in mind, we can analyze the effect eigenvalues approaching 0 have on the efficiency of estimation of β and how to combat those effects. Suppose the matrix $X^T X$ is positive definite. Then the spectral decomposition of the matrix is $X^T X = \Gamma \Lambda \Gamma^T$ where $\Gamma \epsilon \mathbf{M}_{p \times p}$ is a column orthogonal matrix and $\Gamma = Diag(\lambda_1, ..., \lambda_p)$ where $\lambda_i > 0$ for $i = 1, ...p$ is the order eigenvalue matrix corresponding to $X^T X$. Let $T = X\Gamma$ and $\xi = \Gamma^T \beta$, then using the product of these two formulas, we reconstruct the regression equation (4.4) as

$$\underline{Y} = T\xi + \underline{\epsilon} \tag{4.9}$$

The LSE of ξ, using the same process as above, is then

$$\begin{aligned} \hat{\xi} &= (T^T T)^{-1} T^T \underline{Y} \\ &= \Gamma^{-1} T^T \underline{Y} \end{aligned} \tag{4.10}$$

[13] Gruber, Marvin H. J. Matrix Algebra for Linear Models

The covariance matrix of $\hat{\xi}$ is then given by [14]

$$Var(\hat{\xi}) = \sigma^2 \Lambda^{-1} = \sigma^2 Diag(1/\lambda_1, ..., 1/\lambda_p) \tag{4.11}$$

Looking at the trace of the covariance matrix yields $tr(Var(\hat{\xi})) = \sigma^2 \sum_{j=1}^{p} 1/\lambda_j$. We can now see that when we have small λ_j, the total variance of $X^T X$ explodes. Therefore, the problem becomes how can we prevent the variance from exploding, given the fact of small eigenvalues. In a basic sense, we have to add a term in the denominator of the equation which is is exploding the variance of our estimates, explicitly

$$Var(\hat{\xi}) = \sigma^2 (\Lambda + k I_p)^{-1}, k > 0 \tag{4.12}$$

Then by replacing ξ in (4.9) with the estimated matrix in (4.10), we have $\hat{\xi} = (\Lambda + k I_P)^{-1} T^T Y$ with covariance matrix

$$\begin{aligned} Var(\hat{\xi}) &= (\Lambda + k I_p)^{-1} T^T Var(Y) T (\Lambda + k I_p)^{-1} \\ &= \sigma^2 (\Lambda + k I_p)^{-1} \Lambda (\Lambda + k I_p)^{-1} \end{aligned} \tag{4.13}$$

which for well-chosen k, essentially alleviates our problem of explosion of the variances of the estimates in the case of multicollinearity, since the trace of the covariance matrix becomes $tr(Var(\hat{\xi})) = \sigma^2 \sum_{j=1}^{p} \lambda_j/(\lambda_j + k)^2$. This gives the variance of the estimates.

Introducing this well-chosen k, would then increase the efficiency of the estimation of $\underline{\beta}$. Namely, there would be a constraint on each β_j, such that $\sum_{j=1}^{p} \beta^2 < t$ for some positive t, however this constraint changes the

[14] Arashi. Theory of Ridge Regression Estimators with Applications.

minimization problem to be the Penalized Residual Sum of Squares (PRSS)

$$\min_{\underline{\beta}}\{(\underline{Y}-X\underline{\beta})^T(\underline{Y}-X\underline{\beta})+k||\underline{\beta}||^2\}, \tag{4.14}$$

$$||\underline{\beta}||^2=\sum_{j=1}^{p}\beta_j^2 \tag{4.15}$$

We then dub the internal equation $PL(\underline{\beta})$ and take its derivative with respect to $\underline{\beta}$, set it to zero and solve for $\underline{\beta}$ and discover the ridge regression estimator $\hat{\underline{\beta}}^{RR}(k)$.

$$\begin{aligned}\partial PL(\underline{\beta})/\partial\underline{\beta}&=-2X^T\underline{Y}+2X^TX\underline{\beta}+2k\underline{\beta}=0\\ \Rightarrow\hat{\underline{\beta}}^{RR}(k)&=(X^TX+kI_p)^{-1}X^T\underline{Y}\end{aligned} \tag{4.16}$$

These are the new estimators we use, in our context to estimate the true talent of the each individual player. However it becomes apparent the quality of the estimates hinges on the choice of k. When $k=0$, we actually have returned to the LSE and when $k\to\infty$ each estimator approaches 0. There are many papers [including Hoerl et al. (1970) and Kibria (2003)] attempting to estimate k. Multiple hypothesized k were tested in cross-validation and methods outlined in Friedman et al. (2009), returned the best results for our specific situation.

Note there is also a variant of this regression which attempts to quantify the offensive and defensive values of each player separately. In this case, the regression has two observations for each continuous time period with no player substitutions where the regressands are the sum of points scored, one for the home team, one for the away team. Additionally, there are two independent variables for each player, the offense variable is positive when the player is on the court and the defense variable is negative when the player is on the court and zero otherwise. This regression has four times the

amount of data but provides a more in-depth look into which side of the ball a player excels on.

4.8 Box Plus-Minus and Team-Specific Adjustments

Box Plus-Minus (BPM) relies on a series of linear regression models which attempt to capture the bottom-up skills a player possesses and adjusts for team context. The fitted regression has RAPM as the regressand and variables available in the box score per 100 possessions on the court as the independent variables, namely, a player's assists, turnovers, steals, offensive rebounds, blocks, 3-point field goals, 3-point field goal attempts, 2-point field goals, 2-point field goals attempts, free throws and three other variables which are derived based on team context-adjusted points, position, and offensive role. Additionally, there is an offensive variant Offensive Box Plus-Minus (oBPM) which regresses on Offensive Regularized Adjusted Plus-Minus (oRAPM).

Adjusted points ($apts$) begins with the box score statistic, points and normalizes a player's total for the quality of a team's offense. Specifically,

$$apts = ((pts/tsa - Tpts/tsa) + Lpts/tsa) * tsa \tag{4.17}$$

where tsa is true shooting attempts, field goal attempts $+0.44*$ free throw attempts. The .44 constant applied to free throws represents the estimate for how much of a possession a free throw takes up where a field goal attempt is exactly 1. $Tpts$ represents the team's points and the $Lpts$ represents the

league's points. This means a player is judged against the quality of their offense and subsequently the league.

Generally, a player does not always play the same position throughout the course of a season or even a game based on the lineup around them. Player position is scored on a continuous scale from 1 to 5 where 1 represents a point guard and 5 is a center. The initial prior for the position is taken from an aggregating website[15] which has player positions from each team's roster. Most players are listed as guard or forward but the positions are adjusted based on the team-context and number of minutes the players play. A model is used to create a greater spread of positions across the 1, 5 spectrum rather than the guard (1.5) and forward (3.5) that is gained from the source. Additionally, it captures the fact that most player's do not play just one position 100% of the time they are on the court. This model uses the position as the regressand and the independent variables are the proportion of the team's box statistics a player accumulates while the player is on the court. The box score statistics in this case are points, assists, turnovers, blocks, total rebounds, and personal fouls. Statistics like assists and offensive rebounds allow us to infer which players play more point guard than shooting guard or more small forward than power forward and so on. This estimate is then put into the context of all their teammates estimates under the suppositions that a team will play most of their minutes with players at different position and that no estimate will go below 1 or above 5. The histogram of positions for players between 2015 and 2020 shown in Figures 4.3 and 4.4 for women and men respectively. The bimodal distribution reflects the bias which teams have to merely list their player as

[15]Timmerman, Martin. U Sports Hoops usportshoops.ca/mbb2020/index.php.

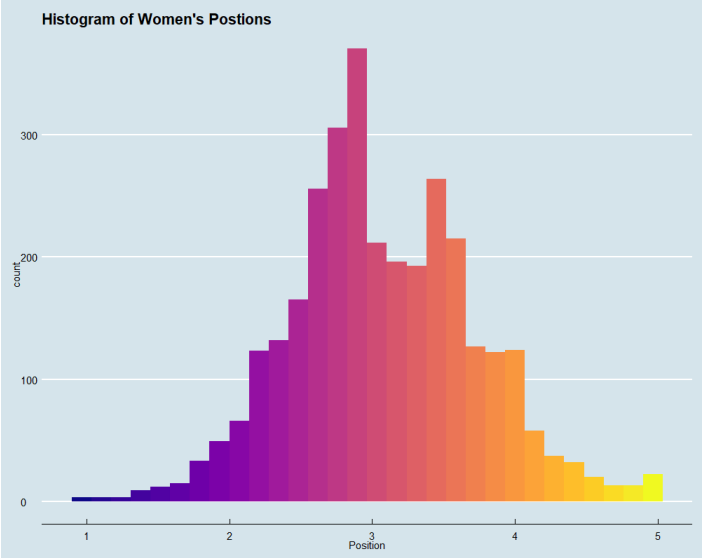

Figure 4.3: Women's Positions

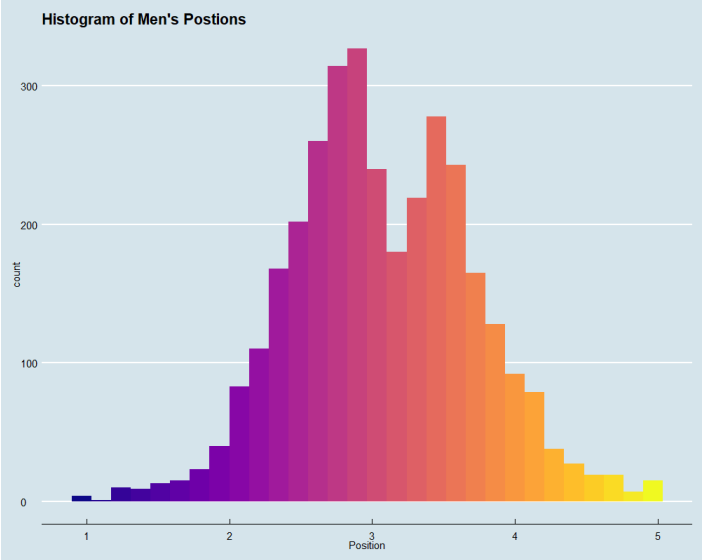

Figure 4.4: Men's Positions

either forwards or guards.

The final team adjusted statistic which is used in the BPM model is one which attempts to quantify which players *create* shots on offense for them and their teammates and which ones are primarily *beneficiaries* of this

creation. The prior of this metric is based on the percentage of baskets which a player scores that are assisted since player's who have a higher percentage of their baskets assisted are benefiting heavily from other players in the offense. This metric can be modelled by a number of box score statistics like assists per 100 possessions, 2-point field goal attempts, 3-point field goal attempts as well the proportion of shots a player takes: at the basket, short range, mid range and 3 point range. This estimate is then scaled to be between 1 and 5 to mimic the position scale and again put through the same team adjustment process as positions. The distributions are again provided in Figures 4.5 and 4.6. These show a left skewed distribution which demonstrates that there are only a handful of players which create most of a team's offense while most players fall into a creator sometimes but beneficiary most of time role.

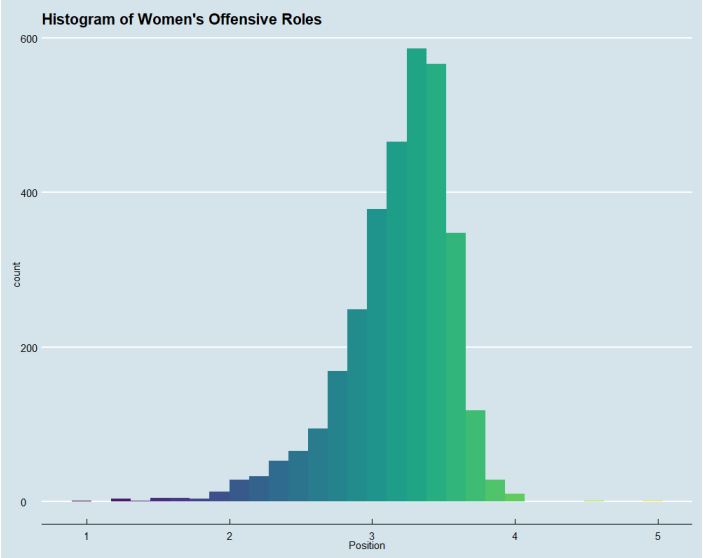

Figure 4.5: Women's Offensive Role

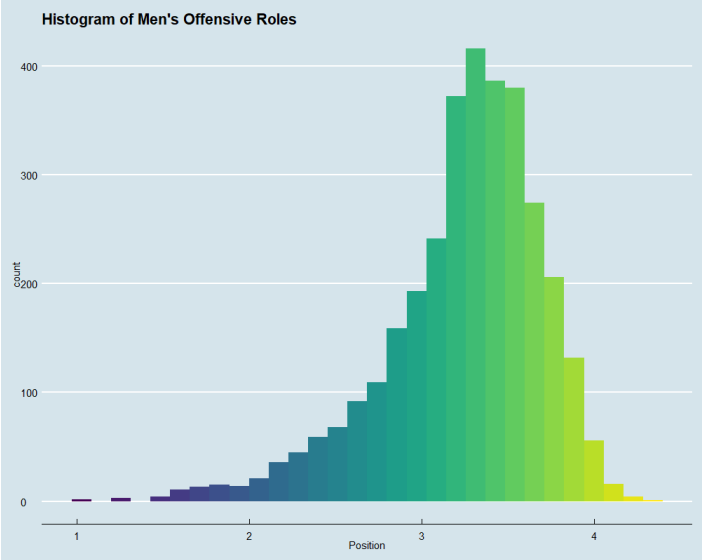

Figure 4.6: Men's Offensive Role

With these statistics gathered, the models fitted and estimates calculated, the base for BPM and oBPM can be created, but the raw output of the model is once more contextualized for the players. There is still an aspect of BPM which is top-down, since the model assumes that the adjusted net rating

of a team is a sum weighted by playing time of the BPMs of the player which compose that team. Note that adjusted net rating for a team is a point differential per possession that accounts for strength of schedule and league environment. In the case of oBPM, adjusted offensive rating is used instead of adjusted net rating (which just doesn't subtract point per possession allowed). Any discrepancies between the team adjusted rating and the weighted sums of players' raw BPMs is then divided by playing time and added back to the players' BPMs on that team. Note that usually the summed raw BPMs of a team is greater than the adjusted net rating, which means a tax will be paid on the raw BPM in proportion to playing time.

Adjusted ratings are scaled and normalized to reflect how a team compares to the rest of the country where the average is zero. However, there proved to be a crucial difference in the calculations of these ratings as compared to calculations in an National Basketball Association (NBA) context. The high and low ends of the scores were significantly greater in magnitude for the USports distribution. For context, the 2017 Golden State Warriors, one of the greatest teams in basketball history, had an adjusted offensive rating of 7 and adjusted net rating of 12.4. On the women's side Saskatchewan consistently has an adjusted offensive rating near 20 and adjusted net rating around 35. On the men's side Carleton is around 30 and 60 for the same statistics. This speaks to the distribution of talent across the country and the fact that there is a much greater range of talent on USports Basketball rosters across Canada than there is in the NBA. Another thing to consider is that there is a clear system advantage which can be cultivated by a university retaining players for 4-6 years and coaches and staff for even longer.

When these final team adjustments are made to the raw BPM produced

by the model, the metric is over-determined by the team of a player, since the team ratings on either side are extreme. An easy way to see this is when a player transfers from a team that is very different in quality, their BPM from season to season shifts erratically in a way that does not comport with the actual development of the player. Additionally, a ranking of the best players is more or less just a ranking of the best teams and makes comparing players across teams impossible if the team performances are not similar. This is not a good method for calculating BPM in our context, but there is something to be said for combining the top-down and bottom-up philosophies, so a penalty term to normalize the adjusted ratings is put in place to make the distribution more like that of the NBA. After this, BPM and oBPM statistics are calculated, and the Defensive Box Plus-Minus (dBPM) is calculated as the difference between BPM and oBPM under the assumption the sum of offensive and defensive value of a player is the equivalent of the value of that player. This makes dBPM an inherently more conservative statistic than its offensive counterpart. The box score has always struggled to capture the true value of a defender and dBPM, while valiantly attempts to do so, still fails. Defensive RAPM over a large sample is a more reliable indicator of a player's true talent but the metrics do generally have similar assessment of players, as can be seen in Figures 4.7 and 4.8.

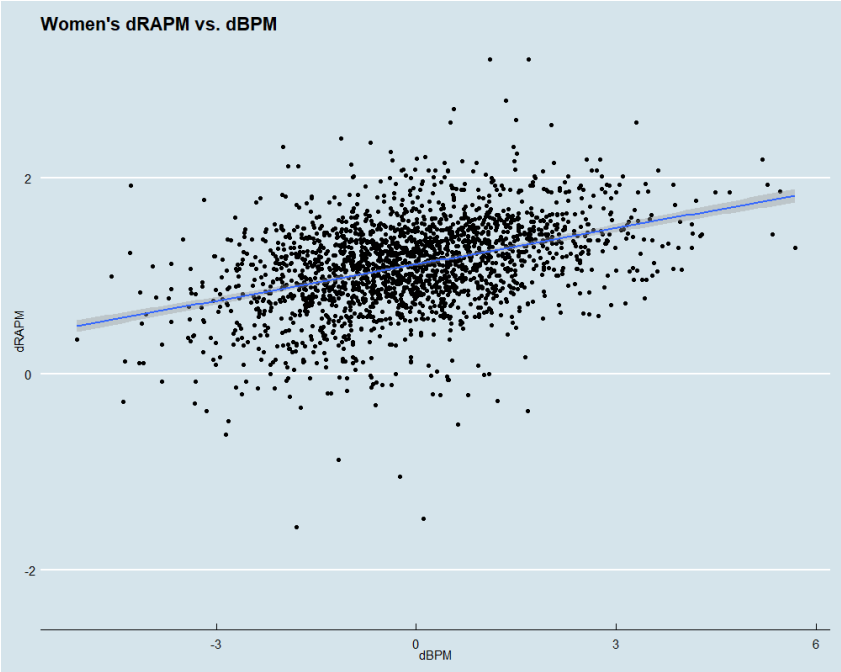

Figure 4.7: Women's DRAPM vs. DBPM

Figure 4.8: Men's DRAPM vs. DBPM

Chapter 5

Results

5.1 Expected Points

After running many training loops and tuning the listed hyperparameters, the effectiveness of the random forests and neural networks was approximately equal in the case of the women's data. Twenty-one of the data sets were best modelled by the neural network while eighteen of them were best modelled by the random forest. The test mean squared error of the data set ranges from approximately 0 to 1.7 $points^2$.

*Note the extremely low MSE is due to very little variation in the dependant variable.

Table 5.1: Women's Neural Network Best Models

Data	MSE	Number of 5th Layer Neurons	Number of 6th Layer Neurons	Drop Out Rate	Optimization Function	Activation Function
cut_2	0.973	32	128	0.25	adam	relu
npt_2	1.007	32	64	0.5	RMSprop	relu
$pnrman_2$	0.936	32	128	0.5	adam	relu
$postup_2$	0.409	128	32	0.25	sgd	relu
$spotup_2$	0.101	32	64	0.25	sgd	relu
iso_2*	1.46E-06	128	64	0.25	adam	relu
$oscreen_2$*	1.84E-04	128	128	0.25	adam	relu
$trans_2$	0.141	32	64	0.25	sgd	relu
$postup_3$	0.153	32	64	0.5	sgd	relu
$oscreen_3$	0.051	128	64	0.5	sgd	relu
$handoff_3$	0.132	64	128	0.5	sgd	relu
$trans_3$	1.001	32	128	0.5	sgd	relu
$spotup_4$	1.588	32	32	0.5	RMSprop	relu
$trans_4$	1.176	64	128	0.5	RMSprop	sigmoid
$pnrball_5$	0.954	64	64	0.5	sgd	relu
$postup_5$	0.926	64	128	0.5	sgd	relu
iso_5	0.856	32	64	0.25	RMSprop	sigmoid
$oscreen_5$	1.077	32	128	0.5	RMSprop	sigmoid
$trans_5$	1.732	64	128	0.5	adam	sigmoid
$pnrball_6$	1.136	32	32	0.5	adam	sigmoid

$oscreen_6$	1.643	64	64	0.5	adam	sigmoid

Table 5.2: Women's Random Forest Best Models

Data	MSE	Bootstrap	Max Depth	Min Samples Leaf	no Estimators
PnR Ball Handler $_2$	0.264	TRUE	10	1000	1000
Offensive Rebound $_2$	0.613	TRUE	5	2	5000
No Play Type $_3$	0.87	TRUE	10	100	500
PnR Roll Man $_3$	0.525	TRUE	500	100	500
PnR Ball Handler $_3$	0.34	TRUE	500	100	1000
Spot-Up $_3$	0.953	TRUE	10	100	5000
ISO $_3$	0.245	TRUE	10	1000	500
No Play Type $_4$	1.182	FALSE	5	100	500
PnR Roll Man $_4$	0.961	TRUE	100	1000	1000
PnR Ball Handler $_4$	0.75	TRUE	5	100	1000
Post-Up $_4$	0.687	TRUE	5	5	500
ISO $_4$	0.925	TRUE	100	100	500
Off Screen $_4$	0.932	FALSE	5	100	500
Hand Off $_4$	0.953	FALSE	5	100	5000
Offensive Rebound $_4$	0.985	TRUE	10	100	500
PnR Roll Man $_5$	1.244	FALSE	5	1000	1000
Hand Off $_5$	1.165	TRUE	10	100	1000
Hand Off $_6$	1.631	TRUE	5	5	500

On the men's side, the data was best modelled by neural networks twenty-three times and random forests fifteen times with the relevant hyperparameters listed below.

Table 5.3 Men's Neural Network Best Models

Data	MSE	Number of 5th Layer Neurons	Number of 6th Layer Neurons	Drop Out Rate	Optimization Function	Activation Function
$pnrman_2$	0.971	32	128	0.5	sgd	relu
$pnrball_2$	0.11	64	32	0.5	RMSprop	relu
$postup_2$	0.428	64	32	0.25	sgd	relu
$spotup_2$	0.125	32	128	0.5	sgd	relu
orb_2	0.559	32	32	0.5	adam	relu
npt_3	0.904	64	32	0.25	sgd	elu
$postup_3$	0.103	64	64	0.25	sgd	relu
$spotup_3$	1.031	32	128	0.5	sgd	relu
$oscreen_3$	0.061	128	32	0.25	RMSprop	elu
$handoff_3$	0.107	64	128	0.5	adam	elu
$pnrman_4$	1.029	64	32	0.25	sgd	relu
$pnrball_4$	0.778	128	32	0.5	sgd	elu
$postup_4$	0.746	64	64	0.25	sgd	elu
$spotup_4$	1.798	64	128	0.5	sgd	elu
iso_4	1.058	64	64	0.5	sgd	elu
$oscreen_4$	1.051	64	128	0.25	adam	elu
$pnrball_5$	1.133	32	128	0.25	adam	elu
$postup_5$	0.923	128	32	0.5	sgd	elu
iso_5	1.07	128	128	0.5	adam	elu
$handoff_5$	1.333	32	64	0.5	sgd	elu

$trans_5$	2.004	128	128	0.25	RMSprop	elu
$oscreen_6$	1.826	128	128	0.5	sgd	elu
$handoff_6$	1.836	128	32	0.5	sgd	elu

Table 5.4: Men's Random Forest Best Models

Data	MSE	Bootstrap	Max Depth	Min Samples Leaf	no Estimators
Cut $_2$	0.989	TRUE	10	5	500
No Play Type $_2$	1.007	TRUE	100	1000	1000
ISO $_2$	0.251	TRUE	100	100	500
Transition $_2$	0.193	TRUE	10	5	500
PnR Roll Man $_3$	0.547	TRUE	500	100	500
PnR Ball Handler $_3$	0.364	TRUE	5	5	500
ISO $_3$	0.315	FALSE	5	100	5000
Transition $_3$	1.014	TRUE	10	100	1000
No Play Type $_4$	1.404	TRUE	5	2	1000
Hand Off $_4$	1.046	TRUE	500	1000	1000
Offensive Rebound $_4$	0.995	TRUE	5	100	5000
Transition $_4$	1.554	TRUE	10	1000	1000
PnR Roll Man $_5$	1.707	TRUE	10	100	1000
Off Screen $_5$	1.197	FALSE	5	100	500
PnR Ball Handler $_6$	1.404	TRUE	500	1000	1000

With these models in place, every possession is given an expected point total and the player who ended the play (either by shot or turnover) is credited. The expected points are then summarised by possession and player and scaled against the league. This is the Shot-Quality (squ) metric for a player. Subsequently this is compared against actual points per possession. The difference between a player's squ and their actual points per possession is called their Shot-Making (sma). These metrics can also be applied to player who was deemed the primary defender of a possession which gives us Defensive Shot-Quality (dsqu) and Defensive Shot-Making (dsma). As mentioned, not all possessions are given a primary defender and defense has many aspects which are not about playing on-ball in the half court. Regardless, these metrics do end up having some predicative power on team performance.

When it comes to stability of these metrics, all are significant predictors of themselves year-over-year into the future, but there is a range of results which span from certainly capturing an important skill to possibly just capturing noise. First, the shot-making metric, sma, correlates year over year for individual women at .28 and for men at .22. So, there is some amount of true talent which is being captured over the course of a season but likely would require a multi-year sample to give a more accurate picture of a player's shot-making abilities. When it comes to the defensive side of the same statistic, the year-over-year correlation for both men and women is .02. So there is a relationship year-over-year but it explains almost none of the variation. Intuitively, it makes sense because as a defender, you have really only have an ability to affect a player up until the moment they shoot. The shot quality metric conceptually would seem to capture something more

like a durable skill which a player could possess on defense. Indeed the year-over-year correlation of dsqu is higher, .06 for women and .04 for men. However, this is not a particularly strong relationship and if it is capturing anything important, it would need a larger sample than just one season of defensive plays to find it. Lastly, we consider the offensive shot quality which most directly comes from the expected points model. In the case of the women's data, squ correlates year-over-year at a rate of .53 and .4 for the men's data. This lends credence to the idea that the squ metric really is capturing the decision-making of offensive players and the skill which they possess for getting quality shots.

5.2 Effects of Leverage and Player Development

We will investigate the effect that low leverage moments have on the play of USports basketball as a best attempt to deal with the problem of try described in Chapter 4.2. Under this model, two categories will be established based on leverage: low effort for low leverage situations and regular effort which will account for all other situations. Note that for measures like RAPM, all the players on the court are accounted for, so when substitute players are placed in the game, this information is already accounted for. The leverage component will merely be used to dock importance from minutes played under significantly low value circumstances.

To test this, a cross-validation was done where predictability of the metric was optimized. Specifically, a regression is done using the 2015-2016 through 2018-19 seasons to calculate each player's RAPM and then the regression with the same weights was done for the 2019-2020 season. Pre-

dictability is defined as the R-squared value in a regression, for players who appeared in both samples, where the regressand is the 2019-2020 RAPM and the independent variable is the 2015-16 to 2018-19 RAPM. The weights which produced the best predictability were selected as the optimal weights. First, a regression which included the average leverage of the game play was included as an independent variable. This made no difference in the case of the men but was useful for the women's data. What did affect the men's metric was choosing a leverage threshold for which any game play under that barrier would be discounted by some amount. This thresholding worked best when any leverage situation scored 0 was discounted by 50 percent.

Important to note, in the men's game, these low-value circumstances defined as blowouts occur for between 2 and 4.5 percent of the minutes played consistently across the last 5 seasons of play. However, the distribution of blowouts is not randomly distributed across teams and has a lot to do with talent. Ontario Tech which has only played the most recent season and struggled quite a lot, played 11.7 percent of their games under blowout conditions. Carleton, on the other end of the spectrum, has won the national championship in four of the seasons in the data and played 12.5 percent of their minutes under these same blowout conditions. So, players on these teams have significantly more of their minutes severely discounted due to the lack of effort which tends to occur in these low leverage moments. Ontario Tech suffered a similar fate on the women's side but there is not an outlier which exists in a similar way to Carleton for the women. Since the same thresholding did not benefit the women's regression, it is not as important to find blowouts as it is to understand which teams face differ-

ent leverage distributions across their minutes and how that will affect the accounting of their statistics.

We also consider the manner in which a player's development, or changing true talent affects our ability to predict their true talent into the future. In the case of the women's game, decreasing the value of past seasons' performance decreased the effectiveness of predicting future performance. Along with a training loop which applied a series of different weights to leverage and number of possessions played, on average there was 8.5 percent less variation in season-to-season RAPM when including any sort of weights on past seasons. Additionally, the more heavily discounted past seasons were, the more predictability was lost. Therefore, the conclusion was drawn that no discounts were applied to performance further in the past and all seasons of data were treated equally in the calculations. This was not the case for the men's game, where decreasing the weight of past seasons actually increased the predictability of the metric. Additionally, the R-squared in all iterations of the cross-validation of this model was higher for the women's game than the men's game. This is another point which suggests there is less season-to-season variation in individual player performance for women as compared to men in USports basketball.

On the men's side, in an equivalent training loop, the weights applied on average increased the predictability of the season to season RAPM by 9.7 percent as compared to no weights on past seasons. The optimal weight on past performance was an exponential decay function which values a season at 75 percent of the prior. The net effect of these adjustments on the men's RAPM calculations is that a regression over four seasons of data will explain 28 percent of the variation in RAPM for an individual in the next

season. More of the variation is explained for players who play more seasons within the four season sample. The women's data had no adjustment for possessions or, as previously mentioned, past seasons but it does have an adjustment for the leverage of the situation. With this adjustment, the four seasons regression predicts 41 percent of the following season's variation in RAPM.

5.3 Women's Expected RAPM

To calculate stability of the metric, RAPM, it is calculated on the first four seasons of the data and then singularly on the last season. For players who appear in the 30^{th} percentile or higher in terms of possession played in both calculations, a linear regression was run using the first four season to predict the last season. The R-squared for this regression is used as the measure of stability for RAPM. Note, that since there is only one independent variable used in the regression, R-squared is used instead of the adjust R-squared. A cross-validation is run to find the optimal hyperparameters to apply to the data with the stability metric as the deciding factor. Equation 5.1 describes the regression equation for RAPM where the coefficient estimates are the scores for each player.

$$margin = \beta_0 + \beta_1 * x_1 + ... + \beta_p * x_p + \beta_{p+1} * lev; \tag{5.1}$$

$$lev = \frac{\sum_{\forall t} lev_t}{|t|} \tag{5.2}$$

Where $margin$ is equal to the difference in the two teams' scores in the observed continuous game section. x_1,...,x_p represent all players who could

possibly be on the court during the period

$$x_i = \begin{cases} 1 & \text{if player i is a home player on the court} \\ -1 & \text{if player i is an away player on the court} \\ 0 & \text{otherwise} \end{cases}$$

$\forall i = 1,...p$. Additionally, t is the list of times in the period where leverage has been calculated. To test whether or not length of the continuous period should be considered, the x_i value is multiplied by the number of possessions in that period. When it comes to past seasons, a variety of weighting schemes were attempted by multiplying the x_i value by a constants which made production in the past less influential on the estimate of β_i than more recent seasons.

Expected RAPM (XRAPM) is calculated with two additional terms which utilize the information gained in the expected points modelling process. Specifically, it uses two more explanatory variables which are the average difference in shot-making and shot-quality between the players shooting the ball for the two teams in the continuous period.

$$margin = \beta_0 + \beta_1 * x_1 + ... + \beta_p * x_p + \beta_{p+1} * lev + \beta_{p+2} * sma + \beta_{p+3} * squ; \tag{5.3}$$

$$sma = (\frac{\sum_{\forall s_h} psma_{s_h}}{|s_h|} - \frac{\sum_{\forall s_a} psma_{s_a}}{|s_a|}), \tag{5.4}$$

$$squ = (\frac{\sum_{\forall s_h} psqu_{s_h}}{|s_h|} - \frac{\sum_{\forall s_a} psqu_{s_a}}{|s_a|}) \tag{5.5}$$

s_h and s_a are the list of shots taken during the period by the home and away teams respectively in that time period. $psma$ and $psqu$ is the shot-making

statistic and shot-quality statistic of the player taking the shot at the denoted subscript.

In the case of the women's RAPM, the tuning of the hyperparameters which delivered the highest stability produced and R-squared of 40.8%. In this case, the leverage variable remained in the equation but x_i was not adjusted based on when the season occurred or number of possessions. The lack of adjustments, was not the case for the calculation of XRAPM. The leverage continued to be held as the explanatory variable, while each year in the past was weighted at 85% of the year prior and situations below the 10^{th} percentile in leverage were discounted entirely. This resulted in XRAPM being slightly less stable than RAPM, coming in at an R-squared value of 38.2%. While giving up some stability, Table 5.5 compares the amount of variation the two metrics can explain with respect to measures of team success. Each observation represents a model where the team success measure is listed under Outcome. The sum of the team's players' RAPM and XRAPM scores weighted by playing time are used as predictors. R-squared is used as the measure of stability for the single predictor models, but adjusted R-squared is used for the aggregate offense/defense models.

Table 5.5 Women's RAPM XRAPM comparison

Outcome	Predictors	(Adj.) R Squared
aNrtg	RAPM	0.7094
aNrtg	XRAPM	0.6884
Win%	RAPM	0.7978
Win%	XRAPM	0.7911
aNrtg	oRAPM+dRAPM	0.6943
aNrtg	XoRAPM+XdRAPM	0.5781
Win%	oRAPM+dRAPM	0.7488
Win%	XoRAPM+XdRAPM	0.6438
aOrtg	oRAPM	0.6411
aOrtg	XoRAPM	0.4213
aDrtg	dRAPM	0.5959
aDrtg	XdRAPM	0.437

The results do not show much of a difference between the overall statistics (RAPM and XRAPM) but XRAPM does perform a little worse at predicting the quality of a team. However, the individual break downs of offense and defense (oRAPM vs. XoRAPM and dRAPM vs. XdRAPM) are significantly worse for the XRAPMs than the RAPMs. The correlation between XRAPM and RAPM in total, offensive and defensive value are .97, .83 and .79 respectively. This implies that XRAPM is an inferior version of RAPM and ought not be valued for the women's set of data.

5.4 Men's Expected RAPM

The process for determining the utility of XRAPM for the men's data is the same as that of the women's. It should be noted that adjustments were made to the data for calculating RAPM, as outlined in Section 5.2. Specifically, previous seasons are weighted at three quarters of the previous season, there is no weighting for length of continuous period and plays which fall below the first percentile in leverage are weighed half as much. This configuration resulted in a 27.7% R-squared value which is notably lower than the RAPM stability for the women's game. When the differential in average shot-quality and shot-making of the players shooting during a continuous period is accounted for we see enormous increases in stability. The hyperparameters were tuned such that length of continuous period is accounted for, first percentile and lower leverage minutes were weighed a little bit heavier at 70% of standard and finally past years were weighed at 80% their prior. This resulted in a 36.0% R-squared, a 30% increase in stability compared to standard RAPM. It is interesting to note,that the women's data did not show radically different results in stability but the men's certainly did. There may be more value in these shot-quality metrics for men's data than there is for women's but there may also be other contributing factors not examined in this analysis. Table 5.6 is the comparison of XRAPM and RAPM in predicting team success.

Table 5.6 Men's RAPM XRAPM comparison

Outcome	Predictors	(Adj.) R Squared
aNrtg	RAPM	0.7259
aNrtg	XRAPM	0.6971
Win%	RAPM	0.6784
Win%	XRAPM	0.6099
aNrtg	oRAPM+dRAPM	0.6382
aNrtg	XoRAPM+XdRAPM	0.6565
Win%	oRAPM+dRAPM	0.5893
Win%	XoRAPM+XdRAPM	0.591
aOrtg	oRAPM	0.4929
aOrtg	XoRAPM	0.5049
aDrtg	dRAPM	0.5728
aDrtg	XdRAPM	0.6025

We see the same slightly worse results for the total value version of the expected metric but in this case the expected offensive and defensive metrics are actually better predictors of a team's offensive and defensive outcome than their standard version. This is an encouraging sign which suggests there is indeed information in the Synergy Sports play-by-play which can tell us more about a player's true talent. Having metrics which can stabilize more quickly is important for any practical use of these metrics since the number of games in a USports season is so small compared to that of the NBA.

5.5 Women's Expected BPM

First, we will examine the BPM and oBPM models which use the following variables to predict an individual's RAPM:

Table 5.7: Women's BPM Model

term	estimate	std.error	statistic	p.value
Intercept	-0.81	0.62	-1.307	0.191
apts	-0.695	0.021	-33.459	0
ast	0.164	0.024	6.899	0
to	-0.213	0.02	-10.749	0
stl	0.198	0.03	6.664	0
orb	0.218	0.022	10.034	0
blk	0.177	0.041	4.29	0
3fg	2.08	0.079	26.384	0
3fga	0.076	0.023	3.371	0.001
2fg	1.442	0.043	33.283	0
ft	0.724	0.026	27.496	0
position	0.746	0.278	2.686	0.007
position^2	-0.14	0.044	-3.175	0.002
orole	0.179	0.087	2.054	0.04

Table 5.8 Women's oBPM Model

term	estimate	std.error	statistic	p.value
Intercept	1.309	0.088	14.896	0
apts	-0.298	0.014	-20.806	0
ast	0.135	0.014	9.805	0
to	-0.121	0.014	-8.718	0
stl	0.059	0.02	2.968	0.003
orb	0.079	0.014	5.522	0
3fg	1.039	0.045	23.025	0
2fg	0.677	0.029	23.435	0
fta	-0.092	0.027	-3.459	0.001
ft	0.478	0.037	13.031	0

Interestingly, the BPM model has a quadratic relationship with position and Figure 5.1 shows the relative value position has on BPM.

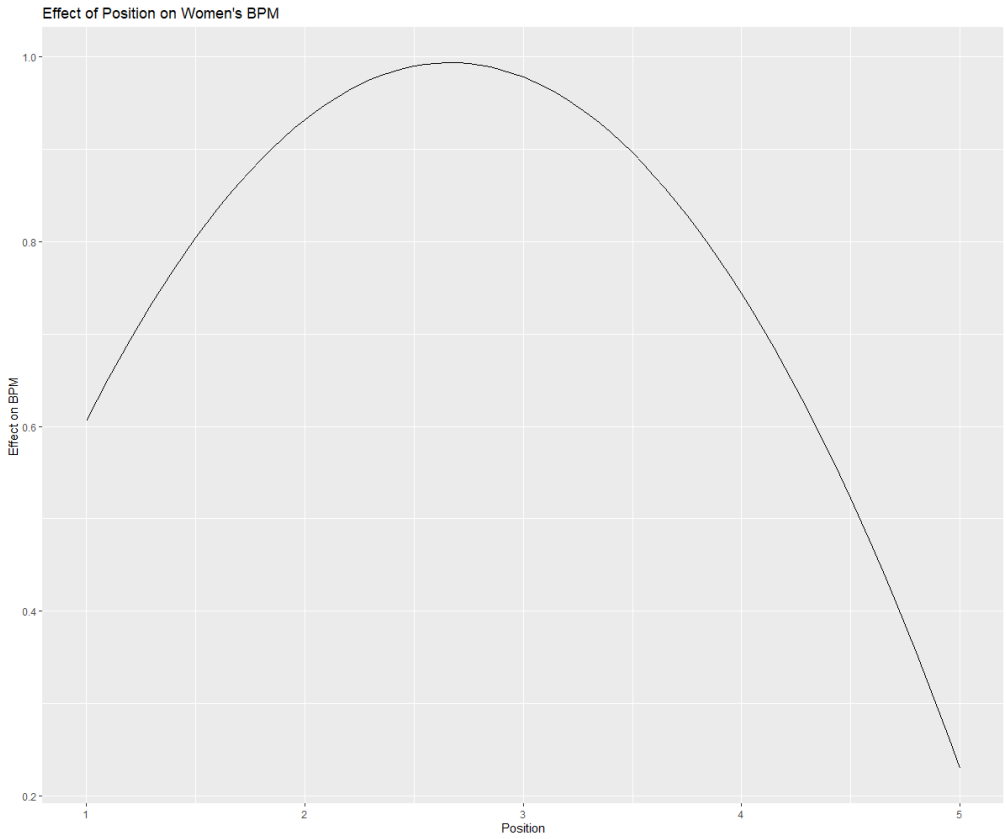

Figure 5.1: Women's BPM Position Relationship

The score peaks at 2.66 and ends up disfavouring players who play near the larger end of the spectrum and really likes wings and guards. This aligns with many of the trends at the NBA level and makes a fair bit of sense intuitively to me. Looking at the coefficient estimates for these models can tell us more about what value the box score has in understanding a player's true talent. When it comes to the overall model, the coefficient estimates for 3-point field goals (3fg), 2-point field goal attempts (2fg) and free throws (ft) are approximately in proportion to the value in points each are worth. Notably, 3-point field goal attempts (3fga) even when they don't go in are still deemed to be somewhat of positive value. The adjusted points (apts)

statistic penalizes players for shooting inefficiently and more so when the player is on a good offense. A player is also taxed for turning the ball over to the other team (to). Offensive Role (orole) is a positive indicator for a player, meaning that initiating offense and having the ball is a positive for estimating a player's true talent. The remaining coefficient estimates are positive and each statistic is measured per 100 possessions. Here we can say a steal and a block are worth approximately the same, with a slight edge to steals. Lastly, note the statistics which do no appear in the model but were tested: defensive rebounds (drb), personal fouls (pf), free throw attempts (fta), 2-point field goal attempts (2fga). A lot of the information in the 2fga and fta is already captured in apts. However, it's interesting that drb is not included despite it's widespread usage as an indicator of value for a player

The offensive model which regresses on oRAPM tells a similar story but does not consider position or orole. Interestingly, there is a tax applied to missed free throws which does not explicitly occur in the overall model. There is also a defensive statistic which makes it into the model, steals (stl). This points to the fact that the two sides of the ball are connected in basketball and transition offensive created by steals is extremely beneficial to a team's offense.

Now we can examine the utility of the shot quality and shot making metrics we created by leveraging the process information contained in the play-by-play data. The task of incorporating the expected points model into the BPM formulation leaves one with a couple of options. The obvious first move is to use, in the ultimate BPM regression, a player's shot-quality and shot-making both offensively and defensively as independent variables. However, the question arises, is it more useful to regress on the RAPM of

player or their XRAPM from the 5 year sample?

Table 5.9 Women's XBPM Model on RAPM

term	estimate	std.error	statistic	p.value
Intercept	2.21	0.91	2.43	0.02
apts	-0.7	0.02	-31.36	0
sma	0.26	0.05	5.6	0
squ	0.26	0.08	3.28	0
dsma	-0.17	0.02	-7.58	0
dsqu	-0.59	0.09	-6.51	0
dposs	0.01	0	2.23	0.03
2fga	0.1	0.02	4.61	0
ast	0.13	0.02	5.74	0
to	-0.14	0.03	-5.28	0
stl	0.18	0.03	6.54	0
orb	0.2	0.02	9.09	0
blk	0.11	0.04	3.17	0
3fg	1.75	0.1	17.99	0
3fga	0.15	0.03	5.87	0
2fg	1.18	0.06	21.19	0
ft	0.67	0.03	24.08	0
orole	0.17	0.09	1.9	0.06

Table 5.10 Women's XoBPM Model on oRAPM

term	estimate	std.error	statistic	p.value
Intercept	0.29	0.43	0.68	0.5
apts	-0.28	0.02	-18.64	0
squ	0.13	0.06	2.33	0.02
sma	0.09	0.02	3.86	0
ast	0.14	0.01	9.63	0
to	-0.08	0.02	-4.41	0
stl	0.06	0.02	3.23	0
orb	0.07	0.02	4.38	0
3fg	0.94	0.05	18.03	0
2fg	0.61	0.03	18.48	0
fta	-0.08	0.03	-3.15	0
ft	0.44	0.04	11.35	0

Table 5.11 Women's XBPM Model on XRAPM

term	estimate	std.error	statistic	p.value
Intercept	3.03	0.77	3.95	0
apts	-0.6	0.02	-30.52	0
dsma	-0.13	0.02	-6.66	0
dposs	0.01	0	2.1	0.04
dsqu	-0.48	0.08	-5.92	0
sma	0.19	0.04	4.73	0
squ	0.16	0.07	2.22	0.03
2fga	0.07	0.02	3.77	0
ast	0.13	0.02	7.45	0
to	-0.14	0.02	-6	0
stl	0.13	0.02	5.32	0
orb	0.19	0.02	9.45	0
3fg	1.53	0.09	17.82	0
3fga	0.12	0.02	5.23	0
2fg	1.05	0.05	21.38	0
ft	0.58	0.02	23.69	0

Table 5.12 Women's XoBPM Model on XoRAPM

term	estimate	std.error	statistic	p.value
Intercept	0.71	0.09	7.6	0
apts	-0.03	0.01	-2.89	0
sma	0.17	0.02	8.14	0
2fga	0.02	0.01	2.41	0.02
3fga	0.03	0.01	3.26	0
pf	0.02	0.01	2.19	0.03
ast	0.05	0.01	4.18	0
to	-0.04	0.01	-3.66	0
orb	0.04	0.01	3.12	0

The models which regress on RAPM and oRAPM can be compared pretty seamlessly to the actual BPM and oBPM models since the dependant variable is the same as well as many of the independent variables. In the overall model, we can see the addition of the squ and sma variables as well as their defensive counterparts plus the percentage of possession a player defends while on the court (dposs) are significant. All these variables are significant and the adjusted R-squared is improved from the original BPM. This fact holds true when just the offensive information is used to supplement the oBPM regression. This a good sign for XBPM because it is able to leverage process information which we would not otherwise have, in a meaningful way. Specifically, it tells us that players who score high in shot making and shot quality and suppress these metrics for the player's they are guarding, are doing something positive for their team.

However, the model which regresses on XRAPM cannot be compared

evenly along adjusted R-squared to the model which regress on RAPM. Therefore, in order to conclude which model is superior, stability was tested. BPM relies on assigning a team's overall success to a sum of its players which means using the metric to predict team success is circular logic. Thus, we use metric stability as the ultimate test of which version of the metric is more useful. In the case of BPM, we have measurements for each player in each season dating back to 2015-16. With the available data, Ontario University Athletics (OUA) teams could have BPM measurements back to 2009-10 however this is considered out-of-scope.

To measure stability, any two subsequent seasons for a player are put in a regression with the prior season used to predict the latter. The XBPM model which regresses on RAPM produces R-squared values for stability of .58, .65 and .26 for the total value, offensive value and defensive value statistics respectively. Compared to the XBPM model which regresses on XRAPM, the R-squared values are .56, .66 and .25. So narrowly the XBPM which regresses on RAPM is more stable, which allows us to make an apples-to-apples comparison with BPM.

BPM does end up being slightly more stable in total value and offense and a big improvement in defense, giving R-squared values of .61, .67 and .38 respectively. However, I am convinced of the merit of XBPM in the case of the women's data, based on the construction of the fundamental regression. Both BPM and XBPM regress on the same RAPM statistic and use the same independent variables except for the additions of a player's offensive and defensive shot-quality and shot-making. The BPM model without these predictors explains 55.1% of the variation in RAPM but by adding these variables derived from the expected points model, the model explains 57.8%

of the variation in RAPM. As mentioned in Chapter 4.8, the defensive version of this statistic is much less reliable than the RAPM estimate and that holds for the XBPM estimates as well. However, this improvement in the BPM statistic is another reason to feel confident that the play-by-play process oriented expected points model can help to bolster the utility of the best metrics available to basketball analysts.

5.6 Men's Expected BPM

Again, we will examine the BPM and oBPM models which use the following variables to predict an individual's RAPM but this time on the men's side:

Table 5.13: Men's BPM Model

term	estimate	std.error	statistic	p.value
Intercept	1.219	0.146	8.331	0
apts	-0.739	0.027	-27.74	0
ast	0.08	0.019	4.171	0
to	-0.14	0.026	-5.311	0
stl	0.326	0.035	9.307	0
orb	0.133	0.025	5.392	0
blk	0.104	0.039	2.679	0.007
3fg	2.205	0.086	25.65	0
3fga	0.052	0.024	2.211	0.027
2fg	1.616	0.051	31.459	0
2fga	-0.053	0.021	-2.498	0.013
ft	0.774	0.032	24.48	0

Table 5.14 Men's oBPM Model

term	estimate	std.error	statistic	p.value
Intercept	1.603	0.084	19.179	0
apts	-0.241	0.016	-15.44	0
ast	0.105	0.011	9.737	0
to	-0.113	0.016	-7.04	0
orb	0.07	0.014	4.886	0
3fg	0.857	0.047	18.066	0
2fg	0.635	0.03	21.247	0
2fga	-0.059	0.013	-4.58	0
ft	0.282	0.019	15.004	0

The coefficient estimates take on a similar values to that of the women's model. Notable differences are the lack of account for player's offensive role or position. The men's overall model deducts more points for inefficient 2-point shooting than is the case for the women's overall model. In fact the men's overall model values a missed 2 point field in direct negative proportion to how it values a missed 3-point field. I would assume that this difference between values comes from the fact that missed 3-point field goals are more likely to rebounded by one's own team than a generic 2-point field goal. A more nuanced parsing of the value of 2-point field goal attempts likely would provide a different result. Fortunately some of that parsing is done by the squ and sma metrics which we calculated. The following tables outline the XBPM and XoBPM models which attempt to leverage this information.

Table 5.15 Men's XBPM Model on RAPM

term	estimate	std.error	statistic	p.value
Intercept	0.47	0.99	0.47	0.64
apts	-0.71	0.03	-28.21	0
sma	0.21	0.04	5.96	0
squ	0.36	0.07	5.41	0
dsma	-0.18	0.03	-6.85	0
dsqu	-0.3	0.1	-3.06	0
ast	0.08	0.02	4.07	0
stl	0.29	0.03	8.31	0
orb	0.11	0.03	4.2	0
blk	0.08	0.04	2.17	0.03
3fg	1.85	0.1	18.15	0
3fga	0.11	0.03	4.42	0
2fg	1.4	0.05	25.84	0
ft	0.71	0.03	22.16	0

Table 5.16 Men's XoBPM Model on oRAPM

term	estimate	std.error	statistic	p.value
Intercept	-0.19	0.48	-0.41	0.68
apts	-0.24	0.02	-14.86	0
squ	0.18	0.05	3.51	0
sma	0.08	0.03	2.78	0.01
2fga	-0.03	0.02	-1.77	0.08
3fga	0.04	0.02	2.19	0.03
ast	0.1	0.01	8.2	0
to	-0.07	0.02	-3.37	0
stl	0.07	0.02	3.31	0
orb	0.06	0.02	3.9	0
3fg	0.71	0.06	11.3	0
2fg	0.55	0.04	14.44	0
ft	0.27	0.02	13.45	0

Table 5.17 Men's XBPM Model on XRAPM

term	estimate	std.error	statistic	p.value
Intercept	2.18	0.37	5.95	0
apts	-0.27	0.01	-29.32	0
sma	0.09	0.01	7.4	0
squ	0.14	0.02	5.75	0
dsma	-0.08	0.01	-7.99	0
dsqu	-0.21	0.04	-5.94	0
blk	0.03	0.01	2.09	0.04
pf	0.01	0.01	1.83	0.07
ast	0.04	0.01	5.57	0
stl	0.1	0.01	8.21	0
orb	0.06	0.01	6.86	0
3fg	0.71	0.04	19.27	0
3fga	0.04	0.01	4.63	0
2fg	0.52	0.02	26.38	0
ft	0.26	0.01	22.05	0

Table 5.18 Men's XoBPM Model on XoRAPM

term	estimate	std.error	statistic	p.value
Intercept	-0.15	0.25	-0.62	0.54
apts	-0.03	0	-6.39	0
squ	0.08	0.03	2.94	0
2fg	0.11	0.01	9.29	0
3fg	0.18	0.02	10.74	0
drb	0.01	0	2.29	0.02
ast	0.07	0.01	10.71	0
to	-0.06	0.01	-5.49	0
stl	0.05	0.01	4.8	0
orb	0.02	0.01	2.73	0.01

Notably, as compared to the women's XBPM models, the men's does not use the dposs variable. However, many of the significant variables are the same and valued in similar proportion to the women's models. Most importantly, the sma, squ, dsma and dsqu variables again are deemed to be significant predictors for RAPM, XRAPM and their offensive variants.

Considering the increased stability in the XRAPM metric for men as compared to RAPM, one may expect the XBPM metric which regresses upon XRAPM to once again show more stability than its RAPM alternative. Indeed this is the case, as the stability of XBPM regressed on XRAPM is .56, .62 and .32 for the total, offensive and defensive value statistics respectively. This compares to the one which regresses on the RAPM statistic which produces .45, .61 and .19 R-squared values in the same order. However, this does not allow for a clean comparison of the fundamental regression which underlies

the BPM and XBPM metrics for the men since the regressand is different. We can say the BPM is slightly more stable than XBPM with R-squared values of .6, .66 and .37 for total, offensive and defensive values. We can also say that including the offensive and defensive shot-quality and shot-making metrics increases the percentage of the variation explained by the model and all have significant p values. This again ultimately lends credence to the idea that the expected point model adds additional information about the quality of a player and their ability to contribute to a winning team.

Chapter 6

Discussion

6.1 Conclusion

To review, on a conceptual level we investigated if the semi-standardized information in the Synergy Sports play-by-play with regards to offensive possessions can improve the best player evaluation metrics in basketball. This thesis is an attempt to evaluate the process which underlies the outcome of a possession: made shot, missed shot, turnover, etc. Using machine learning algorithms such as random forests and neural networks, we were able to effectively model the number of points a possession would produce based on this play-by-play. Due to the structure of the play-by-play, a combination of models were used to best model, the data each with varying degrees of accuracy but ultimately combining to create a useful set of predictions. Using these predictions and following the process of the work of Chang, et al. (2014), we were able to create metrics for individual players which quantified their shot-quality as well as their shot-making by comparing their actual performance and their expected performance.

These shot-making and shot-quality metrics utilise the information contained within the play-by-play in a new way. This information then served as a supplement to the information already contained in the highest quality basketball metrics for individual player performance. To test whether this information was actually adding value, two basic principles were set forth to prove the worth of a metric. The sum of the metric across a team should predict that team's performance. This means, if a metric deems that many players on a team are of high quality that team should win games and outscore opponents and precisely the opposite should occur when a team is comprised of players a metric deems to be low quality. Additionally, a player's performance on the metric in the past should predict in some measure their performance on that metric in the future. If a metric is to have any use to an analyst of basketball, the metric must be capturing some repeatable skill which a player can possess. Possession of a skill implies that a player has and will continue to have this ability, hence the ability for a metric to predict itself into the future is considered an asset.

To test if the information added value to RAPM, each continuous time period used in the regression, which regressed on margin, accounted for each player on the court, leverage, how far in the past it happened, and length of the continuous time period. We then added the average differential of shot-making and shot-quality of players who shot the ball during the continuous time period. To test the two versions of the metric along the outlined principles, each team season for the past 5 years (total of 236) had the sum of their player's RAPM and XRAPM weighed by possessions played compared against their team performance. The total value statistics were compared to the team's winning percentage and adjusted net rating.

Offensive and Defensive statistics were compared against adjusted offensive and defensive rating. Additionally, the stability principle was tested by calculating the metrics in a 4 season sample and a subsequent 1 year sample and analyzing its ability to predict itself for players who played a minimum number of possessions in both samples.

In the case of BPM, the shot-quality and shot-making abilities on both offense and defense were used in the underlying regression to increase its predictive value. Additionally, using XRAPM instead of RAPM as the regressand of this equation was investigated. When it comes to testing, the measure actually uses adjusted net rating for each team and apportions value to the player which makes the logic of testing the first principle circular. Therefore to compare BPM against its expected extension, any two consecutive individual player seasons above a certain minimum possessions played threshold is used to test the stability of the metrics.

Since this project utilises Canadian USports basketball data from the 2015-16 season to the 2019-20 season, there is a symmetry between the men's and women's data. This allows us to run the experiment twice and actually leads to bifurcation of results between the two data sets. In the case of the women's data, the process information within the play-by-play did not add value. In fact in some cases it degraded the utility of the metrics. With respect to women's RAPM, the stability was slightly diminished and its ability to predict team success was as well. With respect to BPM, using XRAPM as a regressand clearly diminished stability. Using RAPM as the regressand and merely adding the shot making and shot quality metrics on offense and defense as independent variables did increase the adjusted R-squared of that regression but slightly decreased stability. Ultimately,

adding this type of information was useful in a very minimal way to the calculation of BPM but by no means was an intervention worth the hardship of modelling. The men's data, however, produced a much more compelling reason to utilize this information.

Using the same process as with the women's data, RAPM's stability metric was increased by 30 percent. This is an immediately compelling reason to use this information but the result is supplemented by the fact the offensive and defensive version of XRAPM were more predictive of a team's offensive and defensive performance, respectively. Additionally, when performing the underlying BPM regression with the supplementary independent variables, using XRAPM as the regressand produced a more stable metric than with RAPM as the regressand. It did not ultimately produce a more stable metric than the standard BPM but its increased predictiveness in the underlying regression is reason enough to value the metric.

In conclusion, the process-oriented approach of harvesting play-by-play information in the lead-up to a shot can add value to the state-of-the-art player evaluation metrics in basketball. The bifurcation of results with respect to the genders means this approach is by no means guaranteed to improve RAPM and BPM metrics. However, with the significant increases in performance seen in the men's RAPM and BPM metrics, it can be speculated that accounting for a player's ability to create quality shots and make them at a higher rate than expected is important for assessing their true talent. Further this can tell us that getting high quality shots and making shots are distinct but not uncorrelated skills which a player can possess. Lastly, it tells us that these skills are important for a team that wishes to optimize offensive performance and suppress that of their opponent.

6.2 Further Research

This research was conducted on five seasons' worth of Synergy Sports data across Canada, which is not an insignificant amount. Given the resumption of this league and continued use of the Synergy Sports play-by-play system, the models which predict expected points will only become more accurate. Additionally, more computational resources are required to tune the adjustments in the offense-defense regression for RAPM and XRAPM. There are twice as many observations and even more problematically, approximately twice as many dimensions, which stunted the research in determining the optimal parameters. With these two adjustments, a clearer picture can be painted of the effect which including process-oriented information can have on the best available player evaluation metrics.

Most importantly, a reason as to why the bifurcation of results occurred between the genders ought to be studied. Speculating on differences between gender is an especially fraught subject, but the fact of the matter is that they are currently segregated in Canadian University athletics. After studying the data for an extended period of time there are some consistent and durable statistical differences which exist between the two leagues. For example, field goal and three point percentages have been consistently 12-14% higher in the men's game than the women's; whereas the free throw percentage show no appreciable difference. I am not speculating on what that means about gender but rather just noting the two leagues have differences which affected the outcome of this study. Delving deeper into what differences in the two leagues caused the process-oriented information to add value to RAPM and BPM in the men's game will be important to under-

standing the utility of this adjustment. Setting a list of parameters which must hold true for this task to produce a valuable adjustment could be a research project. This sort of list would help reduce the time spent testing and coding for any analyst looking to reproduce these metrics for a different use case.

Appendix

Basketball Terms

Assist: refers to when a pass from a teammate leads directly to a made field goal.

Block: refers to when a defensive player touches a field goal attempt and prevents it from being successful. It can also refer to when a defensive player commits a foul by trying to stand in front of an offensive player without setting their feet.

Drive: refers to when an offensive player with the ball makes and attempt to dribble the ball toward the basketball in spite of defensive pressure from the opponent.

Field Goal Attempt: refers to when a shot is attempted in the course of regular game play. This does not include free throws or when a player is fouled and does not make the shot.

Free Throw Attempt: refers to when an offensive player is allowed to shoot unimpeded by defenders, a shot from the free throw line. This occurs mainly from the defensive team fouling a player while shooting the ball. It can also occur from technical fouls for unsportsmanlike conduct, violations, bonus implications among others.

Jump Shot (Jumper): refers to when a player shoots a shot while facing the hoop, jumping, bring the ball above their head with two hands and releasing it. This is the most common shooting form especially as a player moves further from the hoop.

Layup: refers to a shot that comes from near the hoop usually within 3 feet but can be expanded if there are no players between the player with the ball and the hoop.

Personal Foul: is a penalty assessed to a player when they violate the rules by acting on another player in an improper way.

Pick and Pop: refers to when an offensive player sets a screen for the ball handler and after having impeded the progress of the ball handler's defender, they do not go toward the hoop but rather stands in place to shoot a jump shot when passed to.

Points: are the scoring system of basketball and are awarded when the ball goes through the hoop. Some shots are worth more than others but when made, they all result in points scored which cannot be reversed.

Possession: refers to a continuous time period where one team possess the basketball. Possessions are extended when a team records an offensive rebound, so there is no limit on how long a possession can take.

Rebound: refers to when a player grabs the ball after a shot has been missed.

Screen: refers to when an offensive player moves to a teammate's defender and plants their feet in an attempt to impede the progress of that defender. This is then followed by that teammate moving in way which forces the defender to make a decision about how to navigate the player standing in their way. Outside of dribbling and moving with the basketball,

this is the most frequent tool used to create open shots on offense.

Steal: refers to when a defender takes the ball from the offensive team.

Three Point Field Goal Attempt: refers to when an offensive player shots the ball from behind the three point line. All other field goals are worth two points if made.

Turnover: refers to when a team loses possession of the ball to the other team. This is commonly done by passing the ball to the wrong team, dropping the ball or being the last player to touch the ball before it goes out of bounds.

Violation: is a penalty assessed to a player when they violate the rules by failing to do an action in the allotted time. Violations include standing in the painted area continuously for too long, failing to inbound the ball in 5 seconds among others.

Wing: refer to a side of the court on the offensive or defensive end. It can also be used to refer to a player which often times stands in that position on the court on offense.

Play Types

Cut: refers to when an offensive player without the ball runs toward to the hoop receives a pass and makes an attempt to shoot the ball.

Hand Off: refers to when an offensive player without the ball runs toward their teammate with the ball and takes the ball and from there makes an attempt to shoot the ball.

ISO: refers to when an offensive player with the ball is isolated from the rest of their teammates and make an attempt to shoot the ball.

No Play Type: is a miscellaneous category where plays are categorized when they have no obvious structure.

Off Screen: refers to when an offensive player without the ball has their defender screened and moves to a place where they receive the ball and make an attempt to shoot the ball.

Offensive Rebound: refers to when an offensive player rebounds a missed shot from their own team and makes an attempt to shoot the ball.

PnR Ball Handler: refers to when an offensive player with the ball has their defender screened and make an attempt to shoot the ball.

PnR Roll Man: refers to when an offensive player with the ball has their defender screened and then passes to the player who screened for them and they make an attempt to shoot the ball.

Post-Up: refers to when an offensive player with the ball has their back to the hoop in an area near the hoop an makes an attempt to shoot the ball.

Spot-Up: refers to when an offensive player without the ball receives a pass and without dribbling makes an attempt to shoot the ball.

Transition: refers to when a team gets a steal or a rebound on defense and runs to the offensive end and makes an attempt to shoot the ball.