Beyond the Boxscore: Applications of the Four Factors of Basketball in U SPORTS

David Awosoga

Honours Thesis

Department of Mathematics and Computer Science
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

Abstract

Following trends pioneered by baseball, the increased use of statistical tools and methodologies in basketball has led to significant optimizations in resource allocation concerning player and team evaluations. This quantitative analysis continues to supply critical insights into producing predictive frameworks and providing teams with comparative advantages over their opponents. Most calculable metrics are readily available in professional leagues. However, they are particularly underutilized at lower levels of sport, even though similar raw data is collected through mediums such as boxscores. This thesis discusses the statistical background of some boxscore-based metrics used in professional basketball leagues and estimates their model coefficients to fit data from university basketball within Canada. Following a brief history of the advancements of analytics in basketball and background on the boxscore and its uses, the paper will focus on metrics that evaluate team performance. Here, the Four Factors of Basketball [27] (four factors) will be used extensively to derive expected win percentage and win probability models. The introduction of these models will be followed by discussions on their statistical background and characteristics, and case studies will illustrate unique findings resulting from their construction.

Acknowledgements

I would like to sincerely thank everyone who has played a role in the completion of this thesis, starting with my supervisor Dr. John Sheriff, whose willingness to take me on and flexibility with my topic of choice made this project possible. I would also like to thank the University of Lethbridge Pronghorns Men's Basketball Head Coach Jermaine Small for giving me the opportunity to volunteer as the Head of Analytics for the team and get a taste of what life as a sports statistician is. Thank you to the rest of the assistant coaches and players for teaching me more about basketball than I could have ever imagined or learned from a boxscore. Thank you to Mr. Martin Timmerman for your willingness to provide me with the historical data used in this project and for your diligence in promoting U SPORTS basketball within Canada. Finally, a special thanks goes to Pronghorn Men's Basketball Assistant Coach Zale Smordin for always being so bullish about my ideas and willing to put them into practice. Thank you for believing in the value of my work - even at times when I didn't.

Table of Contents

List of Tables

List of Figures

Abbreviations

A	Assists
A/G	Assists per Game
A/TO	Assist-to-Turnover Ratio
BLK	Blocks
BLK/G	Blocks per Game
DEF	Defensive Rebound
DQ	Disqualifications
FG	Field Goal
FGA	Field Goal Attempt
fgPCT	Field Goal Percentage
FT	Free Throw
FTA	Free Throw Attempt
ftPCT	Free Throw Percentage
GP	Games Played
GS	Games Started
MIN	Total Minutes Played
minAVG	Minutes Played per Game
OFF	Offensive Rebound
No.	Jersey Number
PF	Personal Fouls
PTS	Total Points
ptsAVG	Points per Game
rebAVG	Rebounds per Game
STL	Steals
STL/G	Steals per Game
TO	Turnover
TO/G	Turnovers per Game
TOT	Total Rebounds
3FG	3-Point Field Goal
3FGA	3-Point Field Goal Attempt
3PCT	3-Point Field Goal Percentage

Extended definitions of these basic statistics are available from many online resources such as the Basketball Reference Glossary [31], with some terms given slightly different abbreviations the those used in this paper.

Chapter 1

Introduction

Since its invention in 1891 by Canadian James Naismith, basketball has been enjoyed by millions of people worldwide. From elementary school gym classes to the National Basketball Association (NBA), players, fans, and managers continue to ask fundamental questions about the biggest contributors to winning. Such questions include how to gauge team performance, what factors dictate a game's outcome, and how to determine the best player on a team, in the league, or even ever. Attempts to address these questions have been historically underwhelming and incredibly subjective, and while strides have been made in professional leagues, insufficient progress has occurred in amateur levels of basketball outside of the National Collegiate Athletic Association (NCAA)[28]. This paper will attempt to provide insight into these questions in Canadian university basketball by applying some statistical methods utilized in the NBA.

Data analytics in sport has a rich history, with countless hours devoted to developing and maintaining robust records for analysis. Pioneered by baseball writer and statistician Bill James in 1977 and popularized by the 2002 Oakland Athletics [16], the numerical approach to measuring player and team performance has rapidly expanded as the intersection between mathematical models and the framework that governs success in sport is better understood [2]. Knowledge of statistical methodologies, machine learning algorithms, and computational optimizations continues to advance the field, and many software packages and libraries have been developed to aid in the statistical analysis of sports, including basketball [41]. Nevertheless, even the best metrics are not universally trusted by NBA executives and decision-makers [12], and the long-term goal to make data visualization and sports analytics more accessible, digestible, and understandable continues to progress.

1.1 Background

Basketball statistics began with tabulating points scored, personal fouls, and free throws and dividing these aggregate totals by the number of games played to produce *per-game* statistics. These statistics are easy to compute, widely understood by the general public, and still used by sports broadcasters and fans alike. However, they hold very little value to analysts when describing player performance because they fail to consider playing time, making them poor indicators of efficiency and productivity. For example, players with different minutes played per game that average identical points per game could be erroneously considered as of similar value under this framework. Introducing basic *percentages* to measure shooting proficiency resolved some differences in usage rates but gave way to untrustworthy sample sizes and did not properly account for the lower-percentage but increased value of the 3-point shot. As new metrics were introduced to try and combat these deficiencies, ambiguity arose about their context and employment, and little headway in basketball analytics was made until the late 1990s [21].

Inspired by a similar movement in baseball, *regression-based* statistics were introduced to compare the predictive power of models. Per-game statistics were succeeded by the more stable and informative *per-minute* statistics, and *possession-based* metrics introduced a new wave of analytics that provided a strong correlation with wins on a team level and could be broken down for individual players as well [5]. Since teams alternate possessions in a game, playing basketball was abstracted to an optimization problem of how to maximize a team's per-possession efficiency [14]. *Tempo-free* statistics such as offensive, defensive, and net ratings became useful because they account for *pace*, how "fast" a team plays [27]. Pace is calculated as the number of possession a team averages per game and is significant because faster teams will have offensive totals significantly different than teams that play at a slower pace, even if their overall efficiencies are similar. Current leaders in advanced NBA analytics involve *plus-minus* statistics [3], such as Box Plus-Minus (BPM) [22], Estimated Plus-Minus (EPM), and Daily Plus-Minus (DPM). Other metrics with more exotic acronyms have experienced increased popularity in recent years, including LEBRON, a player role and luck-adjusted metric from BBall Index [23], and RAPTOR, the Robust Algorithm (using) Player Tracking (and) On/Off Ratings developed by Nate Silver of FiveThirtyEight [33].

In terms of the mediums used to store basketball data, boxscores have been and continue to be the premier source of detailed game analysis at all levels of basketball. First introduced to baseball in 1859 by writer and avid fan Henry Chadwick [25], boxscores are compact tables that contain the summarized raw counting statistics of players, which are then aggregated to comprise team performance. Since photography was not very prevalent at the time, boxscores were printed in newspapers and became the primary way that people could gain both a cursory and in-depth review of a game's happenings with greater detail than just viewing the final score. However, boxscores do not tell the whole story, and they have been succeeded by more informative data sources such as *play-by-play* and *player tracking.* Play-by-play data refers to a game report that gives the sequential listing of events that determine a game's outcome, chronicling the score and players on the court at the time of any event occurrence. They contain useful information for predicting the outcome of a particular matchup and simulating the progression of events that can lead to such an outcome [39]. The introduction of spatio-temporal data via player tracking and wearable technologies in the 2010s has revolutionized basketball analytics, as cameras are now able to determine the locations of all the players on the court, as well as the (x,y,z) coordinates of the ball [37] with incredible accuracy and precision. This has given researchers the ability to determine the relationship between shot selection and opponent location, as well as provide improved analysis of individual defensive performance [7] - something historically undervalued by boxscore and play-by-play-based statistics. Bayesian regression models and Markov Chains have been used to model player career projections and team evaluations, leading to the creation of new metrics [6].

1.2 Data Collection and Structure

The league under consideration is University Sports, or **U SPORTS**. Formerly known as Canadian Interuniversity Sport (CIS), U SPORTS underwent a rebranding in 2016 and is the governing body and leader of university sport in Canada [34]. As of the 2021-2022 season, there are 48 USPORTS basketball teams representing four conferences across Canada - Canada West (CW), Ontario University Athletics (OUA), le Réseau du Sport Étudiant du Quebec (RSEQ), and Atlantic University Sport (AUS). Table 1.1 shows the distribution of U SPORTS basketball teams across Canada. For analysis of U SPORTS Basketball games, player tracking is unavailable and play-by-play is difficult to parse efficiently. Therefore, game boxscores will be the primary data source used in this paper, an appealing choice due to their widespread availability and consistent and structured format.

The first set of data is taken from the Coach's View Season Summary Statistics, provided for each basketball team by PrestoSports [29]. This data set includes all of the traditional game data recorded in a boxscore, as well as per-game averages and additional metrics such as assist-to-turnover ratio. The individual counting statistics are then aggregated to comprise the overall team results, including the overall record, home/away splits, and opponent performances in each statistic. There are $rosterSize + 2$ rows and 31 variables (columns) in each season summary, listed in Table 1.2 and explicated in the Abbreviations section of this paper. Statistics from the 2015-2016 through 2019-2020 seasons

CW	**OUA**	**RSEQ**	**AUS**
Alberta	Algoma	Bishop's	Acadia
Brandon	Brock	Concordia	Cape Breton
Calgary	Carleton	Laval	Dalhousie
Lethbridge	Guelph	McGill	Memorial
MacEwan	Lakehead	UQAM	Saint Mary's
Manitoba	Laurentian		StFX
Mount Royal	McMaster		UNB
Regina	Nipissing		UPEI
Saskatchewan	Ontario Tech[a]		
Thompson Rivers	Ottawa		
Trinity Western	Queen's		
UBC	Ryerson		
UBC Okanagan	Toronto		
UNBC	Waterloo		
UFV	Western		
Victoria	Wilfred Laurier		
Winnipeg	Windsor		
	York		

[a] Ontario Tech joined U SPORTS in the 2019-2020 season.

Table 1.1: Member Schools for Basketball in Each U SPORTS Conference

were used as the training set for created models, and the 2021-2022 season was used as the test set for predictions (the 2020-2021 season was cancelled due to the covid-19 pandemic). There were 236 total entries for both the men's and women's leagues.

No.	Player	GP	GS	MIN	minAVG	FG	FGA
fgPCT	3FG	3FGA	3PCT	FT	FTA	ftPCT	OFF
DEF	TOT	rebAVG	DQ	A	A/G	TO	TO/G
A/TO	BLK	BLK/G	STL	STL/G	PTS	ptsAVG	

Table 1.2: Variables Contained in Coach's View Season Summaries

The second data set contains the in-game boxscores for every regular season game played in U SPORTS during the same five-year period. Kindly provided by Martin Timmerman [38], the information from these boxscores was also transformed and partitioned in the same fashion as the coach's view season summaries. There were 2747 games played in the men's league and 2737 games in the women's league, the discrepancy in games played caused by cancellations in the 2021-2022 season due to covid-19.

Two types of statistical analyses can arise when examining data. *Descriptive statistics* help us understand the events that contributed to a particular outcome, and they will be extensively used when exploring the results from the test data set. *Predictive statistics* can be used to forecast future outcomes, and the inputs used in predictive models for basketball must be carefully considered to maximize the value that they provide to coaches, player development staff, and sports fans.

1.3 The Four Factors of Basketball

Early basketball analysts used primitive statistics for the quantitative measurement of team performance, and more robust methods were not introduced until the 1990s. One of the trailblazers in this field was statistician Dean Oliver, who wrote data-centred basketball articles on his now-defunct website called the Journal of Basketball Studies (JoBS) [26]. A former Division III basketball player who graduated from the California Institute of Technology with an engineering degree, Oliver's basketball research focused primarily on the effect of pace on team performance. He was also integral in introducing of the aforementioned possession framework [14] that has revolutionized contemporary basketball statistics. After the popularity that followed the publication of his 2003 book *Basketball on Paper* [27], Oliver was hired as the first full-time statistical analyst in NBA history [40]. Since then, he has held multiple sports statistician positions and is currently an assistant coach for the Washington Wizards of the NBA [24].

One of Oliver's most outstanding contributions to basketball analytics, and the guiding principle from which stems the majority of the statistical work performed in this paper, was **The Four Factors of Basketball**. The four factors are based on the four main events that affect an offensive possession in basketball - a field goal attempt, a turnover, an offensive rebound, and a foul drawn. Since possessions are not officially tracked in U SPORTS boxscores, estimates with good empirical accuracy are used in their place based on these events. A team's possessions are thus estimated as the sum of a their field goal attempts, turnovers, and free throw attempts (the factor of 0.44 accounts for the fact that approximately 44% of free throws are possession-ending), and subtracts offensive rebounds, which extend possessions: $POSS \approx FGA + 0.44 \times FTA - OFF + TO$ [14]. To define their relative importance, Oliver assigned linear weights to each factor based on its frequency of occurrence and impact on the outcome of a basketball game, a calculation repeated in Section 2.3. Since the factors can be expressed in terms of percentages instead of raw totals, they are comparable across seasons and even eras of basketball with some slight modifications and estimations.

A field goal attempt is the most common way that a possession ends, and it comes as no surprise that Oliver defined a team's **Effective Field Goal Percentage** (eFG%) as the most important factor to their success. eFG% calculates how efficiently a team scores, adjusting for the extra value gained from a 3-point shot versus a 2-point shot.

$$eFG\% = \frac{FG + 0.5(3P)}{FGA}$$

Turnovers are the second most frequent way that a possession ends, and **Turnover Percentage** (TOV%) is simply the percentage of a team's possessions that end in a turnover. Unlike the other three offensive factors, TOV% is one that teams try to minimize. Turnovers can be classified into two main categories - live-ball turnovers that result in steals for the other team and dead-ball turnovers resulting from a fundamental rule violation such as a travel, shot clock violation, or offensive foul.

$$TOV\% = \frac{TO}{POSS}$$

If a team secures an offensive rebound after a missed field goal attempt, the shot clock will be reset (to 14 seconds in U SPORTS basketball) regardless of how much time was on the original shot clock, making them incredibly valuable for generating additional scoring opportunities. **Offensive Rebound Percentage** (ORB%) is the third most important factor, defined as the percentage of offensive rebounds that a team obtains out of total available rebounds from missed field goal attempts and free throws.

$$ORB\% = \frac{OFF}{TOT}$$

The final factor is **Free Throw Factor** (FTF%), the percentage of "free" points a team gets per field goal attempt. In principle, the rate at which a team gets to the free throw line (FTR) can be used in its place, but FTF% provides slightly more accurate model predictions empirically. It also underscores the conventional wisdom that the ability to get to the line does not have much bearing if a team is not proficient at making free throws.

$$FTF\% = \frac{FT}{FGA}$$

It is worth noting that while other more obscure factors can affect or terminate a possession, the four factors described above capture the overwhelming majority of possession-ending occurrences. This makes models that utilize these factors relatively simple and incredibly informative.

Men		Factor	Women	
2015-2020	2021-2022	Name	2015-2020	2021-2022
48.0	47.8	eFG%	41.3	41.4
18.7	19.0	TOV%	22.9	22.9
29.4	28.9	ORB%	32.0	31.3
20.4	20.5	FTF%	19.4	18.6

Table 1.3: U SPORTS Factor Averages for Test and Train Datasets

Table 1.3 displays the average performances in each of the four factors across U SPORTS basketball, split into the 5-year period that comprises the training data set and the 2021-2022 season which acts as the test set. Comparing the training and test averages yields little difference, which gives optimism about the predictive ability of the generated models. There are many underlying dissimilarities in the way men's and women's basketball is played that contribute to the perceived differences in factor averages. eFG% has the largest gap between the two games, and while a component of this may be attributable to shooting skill, a greater portion is better explained by physiological differences. Empirical data reflects the intuitive notion that field goal percentage has an inverse relationship to the proximity of a player to the basket. Since men and women play on a net of the same height (10 feet) and male basketball players are significantly taller on average than their female counterparts, they have an inherently higher field goal probability in every area of the court, regardless of skill. Additionally, the option of slam dunks, the most efficient field goal type, only increases overall eFG% for men. This fundamentally changes how offenses are run in the women's league, as the reduced threat of interior scoring means that teams have less need to collapse in the paint and can be more spread out to defend 3FGA better [15]. This lack of spacing has major implications on eFG% and could partially explain the disparity in average 3PCT in U SPORTS, 32.27% for men and 28.58% for women.

Another interesting observation from the factor differences is that ORB% is a few points higher among women than with U SPORTS men. Again, a physiological explanation can give some preliminary insights, but the answer is ultimately unclear. ORB% is difficult to partition into individual player percentages since it is contingent on a player's location relative to the basket during a field goal attempt, information not available in a box score. However, further details can be garnered by generating a histogram of offensive rebounds per player and examining the spread, such as in Figure 1.1. As expected, the data is heavily right-skewed for both men and women, with most players accruing less than 30 offensive rebounds in a season. However, the data for females is slightly more spread out than the male results, which implies a greater range of female players capable of successfully grabbing an offensive rebound. This could explain the corresponding larger ORB%.

The rest of the paper is organized in the following manner. Chapter 2 will use the

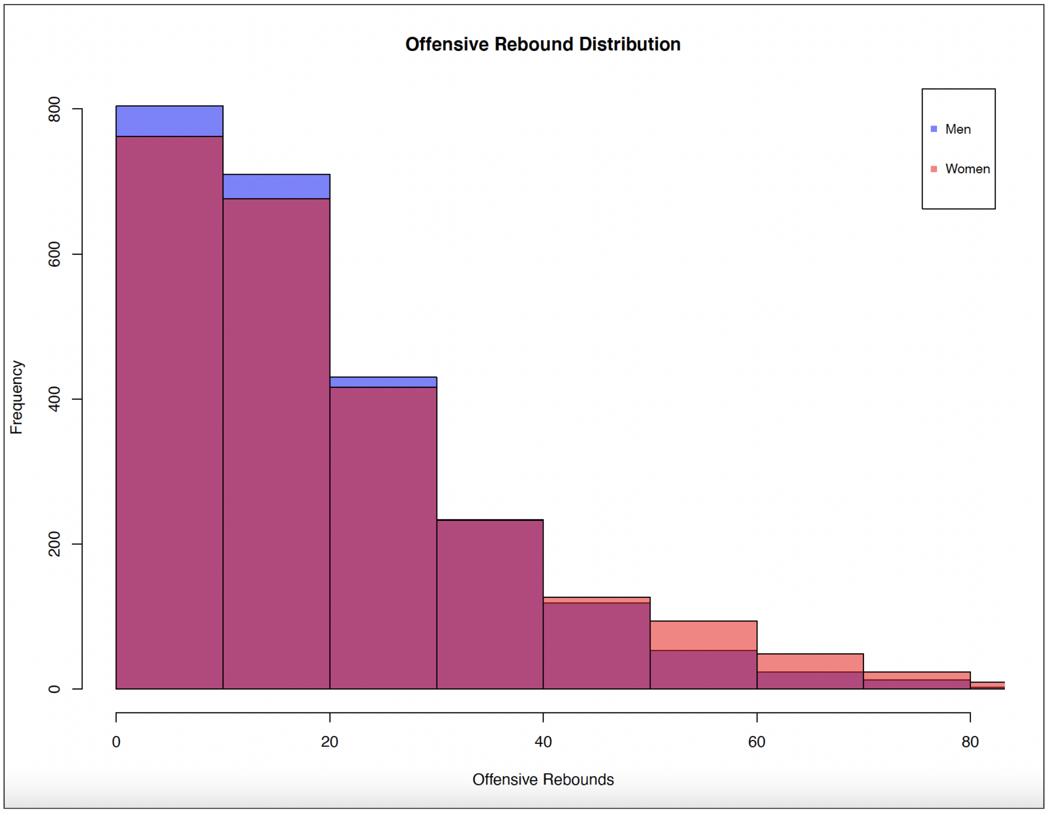


Figure 1.1: Men's and Women's Offensive Rebound Distribution

four factors to model a team's expected win percentage, recalculating the coefficients used in Oliver's original model and analyzing the relative weights to better understand each factor's impact on the men's and women's games. Chapter 3 will use these factors to predict win probability, and the accuracy of predictions on the test set will be computed and discussed for their potential use within a team's strategy and player development operation. Finally, chapter 4 will summarize the findings from the constructed models and introduce techniques that attempt to quantify individual player value by extending the methods used to assess team performance.

Chapter 2

Expected Win Percentage

Using expectations to predict team success is a technique borrowed from Major League Baseball's well-defined "Expected Runs" metric [35], which predicts the number of runs that can be scored within an at-bat based on any of the 24 combinations of outs and occupied bases that can occur at a time. The expected wins model modifies this concept by regressing a team's wins on the four factors, a useful application of which is the ability to determine the "importance" of each factor. Once the factors have been ranked, a strategic framework can be devised within a team's operations to properly allocate time to developing skills that better align with the significance of the factor in question. By adhering to a regimented training model based on the four factors and their relationship to winning, teams can manipulate their factor scores to overcome talent constraints and optimize their expected win percentage [4]. This chapter will adapt Oliver's procedure to create an expected winning percentage model that accounts for some peculiarities that arise within U SPORTS and contrast the norms of the NBA. Following a description of the model construction and analysis of its predictive ability, there will be two case studies that consider some of its unique properties.

2.1 Methodology

A multiple linear regression is a common empirical research tool that attempts to explain a measurable outcome (the dependent variable, Y_i) using well-understood determinants (independent variables), while accounting for random error ϵ. Independent variables are represented as the set $X = \{X_m, m \in 0, 1, 2, \ldots M\}$, where $X_0 \equiv 1$ to provide the traditional intercept (constant) term. The general form of a multiple regression model is then defined as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_M X_M + \epsilon$$

such that

$$E(Y|X_1, \ldots, X_M) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_M X_M$$

The following assumptions must hold in a multiple regression model [9].

1. The model observations must satisfy the population relationship $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_M X_{iM} + \epsilon_i$ for $i = 1, \ldots, N$, where N represents the number of observations.

2. The model must have strict exogeneity, meaning that $E(\epsilon_i|X) = 0$.

3. The variance of the error term must be constant, such that $Var(\epsilon_i|X) = \sigma^2$

4. The covariance between the different error terms ϵ_i and ϵ_j conditional on X is zero.

5. There must not be a linear relationship between the explanatory variables.

The parameters $\beta_0, \beta_1, \ldots, \beta_M$ of the model are then estimated using *Ordinary Least Squares Estimators.* Ordinary least squares estimators minimize the sum of squared errors for each parameter, represented as

$$S(\beta_m) = \sum_{i=1}^{N} (y_i - E(y_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_M x_{iM})^2$$

With these assumptions satisfied, this process will result in the estimators $\hat{\beta_m}$ being the best linear unbiased estimators of the parameters. If the error term of the model is normally distributed, then the least-squares estimators and the dependent variable y_i for each observation will also be normally distributed.

The coefficient of determination, more commonly referred to as a model's R^2 value, shows how much of the variability of a model can be explained by its explanatory variables. It is calculated as

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

and will be used in the following discussions to demonstrate how well the four factors predict team success in a season. *Adjusted* R^2 *(*$\bar{R^2}$*)* is an alternative measure of goodness-of-fit that accounts for some deficiencies present within the basic R^2 construct and penalizes the inclusion of variables that do not add value to a model's predictive ability.

2.2 Model Specification

To perform these calculations with the U SPORTS dataset, modifications to the original expected wins model must be made to account for differences between the NBA and U SPORTS schedule structures. Firstly, U SPORTS teams play exclusively within conference regular season games, with the only non-conference matchups occurring as exhibition games or during the playoffs. On the other hand, teams in the NBA play each other at least twice during the regular season, regardless of the conference. Secondly, due to the much shorter season length of between 16 and 24 games, teams can and frequently achieve perfect and win-less regular seasons in U SPORTS, a feat statistically improbable in a typical 82-game NBA season. Finally, the variation in games played by teams across the U SPORTS conferences has the greatest effect on the model construction. The NBA has an equal number of teams in each conference, and each team plays the same number of games in a season. This is not the case in U SPORTS, with each conference playing a different number of games within their regular season and these totals varying season-to-season for some conferences, such as the OUA. Therefore, using expected wins as the dependent variable results in a biased estimation that does not suitably describe the winningness of a particular team. To combat this, a team's win percentage is used as the response variable since it transfers across conferences and can simply be multiplied by the corresponding games played to give the more familiar expected wins output.

To use the four factors as explanatory variables to model a team's expected win percentage, their offensive and defensive factor performances are split into two categories. The offensive factors are denoted by their familiar name, while the factors of their opposition are prepended with "opp", for eight total variables. Since the actual win percentage is a value ranging from 0 to 100 inclusive and is linear in nature, there is no guard against expected values outside the realm of possibility, such as values over 100% or negative percentages. Due to its proprietary nature, it is unclear whether Oliver's model included an intercept term or not, and the terms' effects are worth considering. The inclusion of a constant term often has an inconsequential interpretation, but excluding it could have significant ramifications on a model's R^2, F-statistic, and normality assumptions. Therefore, models with and without the constant term will be generated in the following manner and compared during analysis:

$$\begin{aligned} WinPecentage = {} & \beta_0 + \beta_1 eFG\% + \beta_2 TOV\% + \beta_3 ORB\% + \beta_4 FTF\% \\ & + \beta_5 oppeFG\% + \beta_6 oppTOV\% + \beta_7 oppORB\% + \beta_8 oppFTF\% \end{aligned} \tag{2.1}$$

$$\begin{aligned} WinPecentage = {} & \beta_1 eFG\% + \beta_2 TOV\% + \beta_3 ORB\% + \beta_4 FTF\% + \beta_5 oppeFG\% \\ & + \beta_6 oppTOV\% + \beta_7 oppORB\% + \beta_8 oppFTF\% \end{aligned} \tag{2.2}$$

Table 2.1: U SPORTS Men's Win Percentage

	Dependent variable:	
	Win Percentage	
	M_1	M_2
eFG	2.333^{***}	2.855^{***}
	(0.214)	(0.161)
TOV	-1.962^{***}	-1.734^{***}
	(0.260)	(0.258)
ORB	0.695^{***}	0.936^{***}
	(0.167)	(0.157)
FTF	0.562^{***}	0.702^{***}
	(0.185)	(0.185)
oppeFG	-2.726^{***}	-2.231^{***}
	(0.224)	(0.181)
oppTOV	2.062^{***}	2.022^{***}
	(0.227)	(0.232)
oppORB	-0.970^{***}	-0.596^{***}
	(0.234)	(0.214)
oppFTF	-0.544^{***}	-0.479^{**}
	(0.185)	(0.189)
Constant	0.747^{***}	
	(0.209)	
Observations	236	236
R^2	0.872	0.975
Adjusted R^2	0.868	0.974
Residual Std. Error	0.087 (df = 227)	0.090 (df = 228)
F Statistic	194.106^{***} (df = 8; 227)	$1,101.860^{***}$ (df = 8; 228)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2.2: U SPORTS Women's Win Percentage

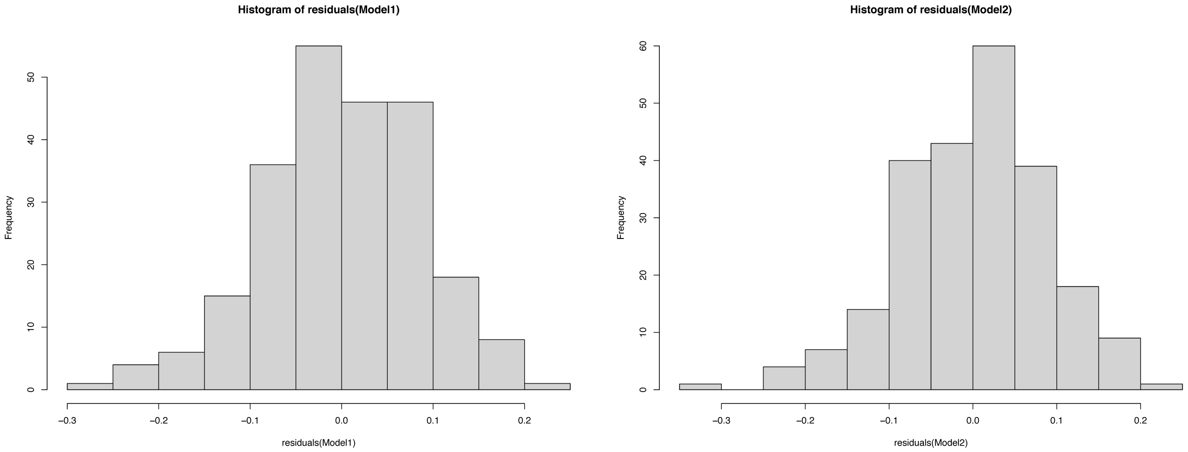
	Dependent variable:	
	Winning Percentage	
	W_1	W_2
eFG	2.242***	2.779***
	(0.201)	(0.164)
TOV	−1.761***	−1.554***
	(0.186)	(0.187)
ORB	0.678***	0.813***
	(0.140)	(0.142)
FTF	0.659***	0.718***
	(0.177)	(0.183)
oppeFG	−2.294***	−1.793***
	(0.219)	(0.193)
oppTOV	1.795***	1.867***
	(0.211)	(0.218)
oppORB	−1.038***	−0.806***
	(0.169)	(0.166)
oppFTF	−0.734***	−0.621***
	(0.184)	(0.189)
Constant	0.643***	
	(0.149)	
Observations	236	236
R^2	0.904	0.977
Adjusted R^2	0.901	0.976
Residual Std. Error	0.084 (df = 227)	0.087 (df = 228)
F Statistic	268.248*** (df = 8; 227)	1,221.205*** (df = 8; 228)
Note:		*p<0.1; **p<0.05; ***p<0.01

2.3 Analysis

Tables 2.1 and 2.2 report the output of the regressions run from equations 2.1 (M_1 and W_1) and 2.2 (M_2 and W_2), courtesy of the stargazer package from R [18]. A cursory glance at the beta value of each factor for both the men's and women's results gives assurance about the model specification. The signs on each coefficient align with a basic understanding of the game of basketball: shooting, rebounding, and free throws will have a positive impact on a team's success, while turnovers will have a negative effect. The opposite intuitively holds for the opposition and is reflected by the coefficients on the "opp" variables. Each coefficient is statistically significant at the 1% level except for oppFTF in M_2, which has a p-value of 0.012.

In M_1, the adjusted R^2 of 0.868 means that the explanatory variables account for nearly 87% of the variation in the response variable (win percentage) of the model. This is extremely high and comparable with results from NBA datasets [1][13], providing strong evidence in defence of the factors' relationship to winning. The minuscule difference between the original and adjusted goodness-of-fit values demonstrates how little penalty each added variable receives. Since the factors and dependent variable can be expressed with percentages, they have a very straightforward interpretation. For example, a 1 percent change in eFG% is expected to change a team's win percentage by 2.333 percent, and a team that reduces its TOV% by 1 percent is expected to enjoy a 1.962 percent increase in win percentage. The F-statistic of the model is also very significant, which increases the confidence that we have in its validity. M_2 omits the constant term and consequently produces some interesting results. To start, the coefficient of determination in M_2 experiences a substantial increase to 0.974, nearly 11 points greater than that of M_1. This value is nearly identical to the un-adjusted metric and the model's F-statistic jumps to over 1100 with the extra degree of freedom provided in the denominator. The interpretation of the coefficients remains the same as in M_1, and there is some discussion to be had about the effects of removing the constant term from M_2.

A general observation from comparing the two models is the defensive emphasis of M_1 compared to the more offense-heavy focus given to the coefficients of M_2. The values of eFG% and oppeFG% are seemingly flipped between the models, and rebounding also experiences a similar switch in coefficients on both sides of the ball, from 0.695 to 0.936 offensively and -0.97 to -0.596 on defense. Some coefficients remain relatively unchanged, such as oppTOV (2.062 vs 2.022) and oppFTF (-0.544 vs -0.479). Recall that residuals are the difference between an observed outcome and its expected result. They form the basis of the ordinary least squares formulation of a linear regression model. The intercept term β_0 is incorporated to ensure that the mean of the residuals will be zero and protect the model against bias, the second assumption referenced in Section 2.1. However, Figure 2.3 shows two adjacent histograms comparing the residuals of the M_1 and M_2, and while they have slightly different shapes, M_2 is still very normal in its distribution, with a mean of 0.000555. Normality tests such as that of Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling all report p-values greater than 0.05 (0.084, 0.85, and 0.33, respectively), so any uncertainties about M_2's legitimacy can be quickly put to rest.



The women's regression output W_1 was similar to the men's model, with intuitively correct coefficients and an excellent R^2 value of 0.904. The F-statistic was also very high and slightly outperformed the men's model, giving reason to believe that the four factors may better model the women's game. This proposition will be examined further in Sections 2.4 and 3.4 by comparing the predictive abilities of the men's and women's models. W_2 exhibits a comparable offensive emphasis to M_2 when estimating the factor coefficients and experiences a large increase in R^2 with almost zero penalization from the dependent variables. The ability to explain nearly 98% of the model's variation is substantial and further proves that the four factors are transferable across multiple levels of basketball. Apart from eFG%, which displays a considerable difference in offense and defense between W_1 and W_2, the other factor coefficients are relatively unchanged. It is interesting to note that even though the average eFG% in U SPORTS basketball is vastly different between men and women (48% to 41.3%), the coefficient on these terms for both models is similar. Another peculiarity in W_2 is that opponent TOV% (1.867) has a greater absolute effect on the expected win percentage than opponent eFG% (-1.793), although this difference is not statistically significant. This is incongruous with the other three models, where eFG% has a greater impact than TOV% on both sides of the ball.

The weight of each factor is determined by first computing the combined absolute value of the coefficient assigned to its offensive and defensive components. This value is then divided by the sum of the factor coefficients in the model, again in absolute terms. The result is a linear weight that describes the factor's importance and impact on how many games a team is expected to win. Since the constant term has an unobservable influence on the regression results, W_2 and M_2 will be used so that the sum of weights is 100. The calculation for the weight of eFG% in women's basketball is illustrated as an example in Equation 2.3.

$$\begin{aligned} Weight_{eFG} &= \frac{|2.779|+|-1.793|}{|2.779|+|-1.554|+|0.813|+|-0.718|+|-1.793|+|1.867|+|-0.806|+|-0.621|} \\ &= 41.75\% \end{aligned} \tag{2.3}$$

Table 2.3 shows the complete factor weights for U SPORTS basketball. For reference, when Dean Oliver initially computed his expected wins model for the NBA, he came up

Factor	Men	Women
eFG%	44	42
TOV%	33	31
ORB%	13	15
FTF%	10	12

Table 2.3: Four Factor Weights in U SPORTS

with factor weights of 40% for eFG%, 25% for TOV%, 20% for ORB%, and 15% for FTF%. A recreated model of expected wins in the NBA from 2017 had weightings of 46%, 35%, 12%, and 7% [11], which has closer alignment with the U SPORTS results. A noteworthy takeaway is that even though their magnitudes have significantly changed since their inception, the order of factor importance has remained the same. Efficient shooting and smart, mistake-free basketball are imperative to success in U SPORTS basketball, while a constant presence on the offensive glass and free throw line round out the necessities and contribute significantly.

2.4 Predictive Ability

The models used the games played during the 2021-22 season to test their predictive abilities and were compared using their R^2, root-mean-square error (RMSE), and residual square error (RSE). RMSE is calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

and measures how spread out the residuals in a model are, showing how well the data is centred around the line of best fit [8]. Residual Standard Error is a related test of prediction accuracy that takes the square root of the residual sum of squares divided by the degrees of freedom, expressed as

$$RSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{df}}$$

Model	R^2	RMSE	RSE
M_1	0.876	0.093	0.174
M_2	0.870	0.096	0.179
W_1	0.875	0.103	0.166
W_2	0.868	0.106	0.172

Table 2.4: Comparing Predictive Abilities of Models

Table 2.4 displays the values that the four models had in each of the three described statistics. Even though the models without intercepts had superior R^2 values to those with them in the regression output with the training set, they were very similar in predicting win percentage on the test data set. In fact, the models that included the constant term were slightly better, but not by a substantial amount. The men's and women's models also produced nearly identical results in explaining the variation of win percentage in the test set, which clouds the initial impression that the four factors better model women's basketball in U SPORTS. The RMSE values for all four models are around 10%, with the men's outputs being marginally better. However, the RSEs of the women's models were lower than M_1 and M_2, with W_1 being the only model to dip under 17%. Figure 2.1 compares the expected win percentage from M_1 and W_1 (horizontal axis) and actual win percentage (vertical axis) for the 48 men's and women's basketball teams during the 2021-2022 season. Within this figure lie various applications of the four factors, two of which are explored in the following case studies.

2.4.1 Case Study: Measuring Statistical Accomplishment

One application of the expected win percentage model is using the predicted values as measures of statistical dominance and inferiority. Here, the "best" and "worst" statistical teams are defined as those with the highest and lowest expected win percentages, the extreme values from the x-axis shown in Figure 2.1.

Coming as little surprise to anyone familiar with U SPORTS men's basketball over the past decade, the best statistical team during the 2021-2022 season were the eventual

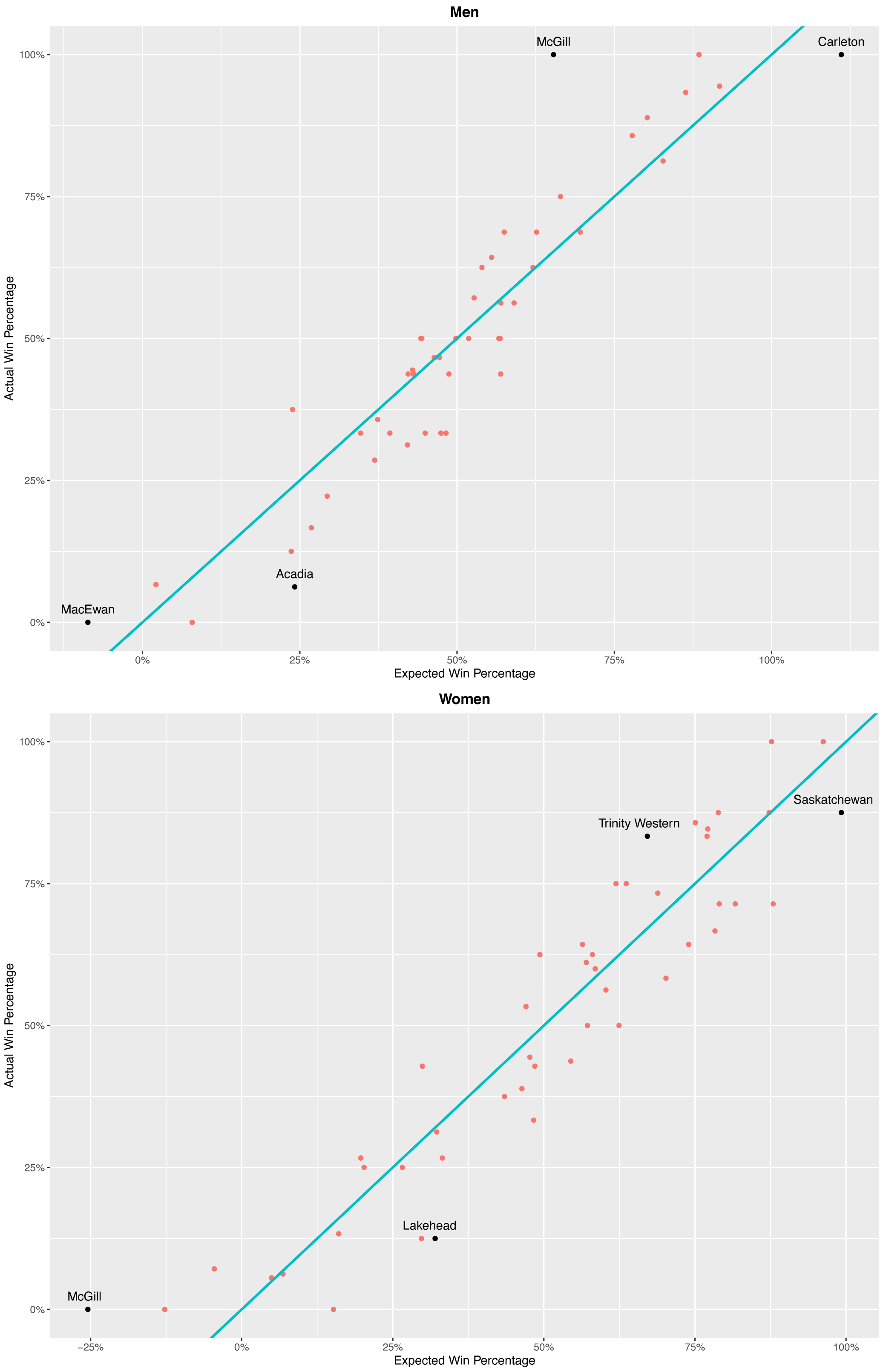


Figure 2.1: Comparing Expected and Actual Win Percentage for the 2021 - 2022 Season

national champion *Carleton Ravens*. They were the cream of the crop amongst a deep pool of talented teams, with an expected win percentage nearly 20 points greater than the next best school. The Ravens cruised to a 14-0 regular season before losing in surprising fashion to Queen's University in the OUA Semifinals. They regained their composure in the U SPORTS Final 8, narrowly defeating the host Alberta Golden Bears in the semifinal and overpowering the runner-up Saskatchewan Huskies in the final to win the national title. Statistically, Carleton boasted the 2^{nd} best eFG%, 2^{nd} best ORB%, and the 2^{nd} best opponent ORB% in U SPORTS, at 53.7%, 38.2%, and 20.9%, respectively. However, their defensive performance is what warranted their infeasible expected winning percentage of 111%. Their 37.8% opponent eFG was the lowest recorded in U SPORTS men's basketball, not only in the season of interest but across the entire 6-year data set. Only four other teams, including the 2017-2018 Ravens, had recorded a season allowing an opponent eFG under 40% since the 2015-2016 season.

The Saskatchewan Huskies have spent nearly 15 years as a powerhouse women's basketball program under the leadership of longtime head coach Lisa Thomaidis, and the 2021-2022 season proved no different. The Huskies had the largest expected win percentage in U SPORTS, at a shade over 99%. They ended up going 14-2 during the regular season, with both losses coming at the hands of the Winnipeg Wesmen, the team that the Huskies ultimately defeated to win the Canada West Final. The top-ranked team going into the U SPORTS Final 8, Saskatchewan was upset in their opening-round matchup against Queen's University and then cruised past Laval and UPEI to win the consolation final, finishing 5^{th} overall. The Huskies led U SPORTS in eFG% (48.3%) by over 1 point, and were top 10 in 4 other categories, FTF% (21.5%), opponent eFG% (36.8%), opponent TOV% (29.4%) and opponent FTF% (13.4%). Compared historically, this edition of the Huskies had the 8^{th}-best eFG%, following a trend that has them occupying four of the top five spots in this category.

When examining statistical inferiority among men's U SPORTS teams, one does not

eFG	TOV	ORB	FTF	oppeFG	oppTOV	oppORB	oppFTF
40.4	23.0	22.3	16.6	56.8	19.2	34.8	14.9

Table 2.5: Macewan Griffins 2021-2022 Regular Season Statline

have to look any further than the Grant Macewan University Griffins. The Griffins failed to win a game during the regular season and their prohibitively poor performance resulted in them achieving an expected win percentage of -8%, the only predicted value below zero among men's results. Shown in Table 2.5, their 40.4% eFG% was the lowest in U SPORTS in the 2021-2022 season, and the second-worst output across the entire 6-year data set. Defensively, Macewan's opponent eFG% of 56.9% was the highest surrendered in the six years under consideration, nearly 1% worse than the next closest team, the winless 2017-2018 Trinity Western Spartans. The Griffins were among the bottom-five teams in nearly every other statistical category during 2021-2022, including the 3^{rd}-worst TOV%, 4^{th}-worst ORB% and FTF%, and 2^{nd}-worst opponent ORB%. An interesting thing to note is that Macewan had the 3^{rd}-best opponent FTF% in the country, and this detail demonstrates why FTF% is the least valuable of the four factors - excellence in this category cannot overcome poor results among the other factors.

On the women's side, three teams had negative expected win percentages. None of these teams managed a win during the 2021-2022 regular season, and the McGill Martlets resided in the statistical basement. The Martlets ranked at or near the bottom in each of the four factors, including having the worst eFG% (33.7%) and second-worst ORB% and opponent eFG% (20.7% and 48.3%, respectively). Although they did not produce any historically bad performances, their combination of poor offense, rebounding deficiency, and inability to get to the free throw line leaves little to wonder unto why McGill finished the season with an expected win percentage of -25%, well outside of the interpretable realm.

2.4.2 Case Study: Assessing Achievement Using Residuals

Another byproduct of the predicted model construction in Figure 2.1 is that the difference between expected and actual win percentage gives the residuals, visually described as the vertical distance between the 45° line and the point of interest. Taking the maximum and minimum values of these residuals yields a unique interpretation - teams with the greatest under/over-achievement.

Across the men's teams, Acadia University had the dubious distinction of U SPORTS' most underachieving team. The lowly Axemen finished dead last in the AUS conference with a 6.25 win percentage (1 win, 15 losses) but had an expected win percentage nearly 18 points higher, at 24% (4 wins and 12 losses). Acadia had solid offensive production with a 49.6% eFG% that was good for 16^{th} in the country, and coupled this shooting proficiency with an above-average 18.1% TOV%, 20^{th} in U SPORTS. This set the precedence for perceived success but is where the positive reviews come to a crashing halt. The Axemen were ranked 41^{st} in the country in opponent eFG%, 44^{th} in opponent FTF%, and 46^{th} in ORB% at a paltry 21.4%. However, there is more than meets the eye about Acadia's undisciplined defense, a detail that provides clarity to Acadia's win percentage discrepancy. The Axemen received 13 disqualifications during the regular season, the most among AUS teams and third highest total in U SPORTS. This statistic is not shown by the four factors and occurs when a player is ejected from a game for committing five personal fouls. The early exits of players who were key offensive contributors cost Acadia on multiple occasions late in games, with five of their losses coming by 6 points or less.

On the opposite end of the achievement spectrum, the McGill Redbirds were an expected 8-win 4-loss team that took advantage of a weak conference schedule en route to a perfect 12-0 record, winning the RSEQ division. The Redbirds' 50% difference between their win percentage and that of the second place team (Concordia, 6-6) was the largest such gap in any division in U SPORTS. Additionally, RSEQ had the smallest divisional parity index as per win percentage standard deviation ratios [32], an indicator of the overall closeness of talent within that division. McGill achieved success with an elite defensive scheme, ranking 5th overall in opponent FTF%, 6th in opponent TOV%, and 11th in opponent eFG%. This staunch defense won the Redbirds multiple "close" regular season games - 9 of their 12 victories came by 6 points or less. However, their modest 49.5% eFG% and below-average 21.9% TOV% proved to be insurmountable against the stronger out-of-conference opponents that McGill faced in the USPORTS Final 8. In the quarterfinal, they were overwhelmed by the statistically superior and eventual U SPORTS bronze medallists Alberta Golden Bears and then fell to the Canada West Champion Victoria Vikes in the consolation semifinal. This marked a disappointing end to McGill's once hopeful season, summarized by a lack of offensive prowess that was hidden behind the relative weaknesses of their regular season opponents.

Women's teams in U SPORTS had less drastic achievement differences, with a few noteworthy exceptions. During the 2021-2022 season, the most underachieving team was the Lakehead University Thunderwolves. The Thunderwolves finished the year with 2 wins and 14 losses, well below their predicted 5-11 record. Although most of their losses did not result in particularly close final scores, diving into Lakehead's factors provides clarity on this matter. Statistically, their most glaring weaknesses were their 23.1 ORB%, which was the fifth-lowest mark in U SPORTS, and their opponent TOV% of 18.4%, which was ranked sixth lowest. Their losses resulted from an inability to eliminate opponent possessions by forcing turnovers while at the same time not generating extra possessions for themselves via offensive rebounds. Although the Thunderwolves defended relatively well and were not *terrible* offensively, this combination produced lopsided overall FGA totals (854 for, 950 against), extra opportunities to score that their opponents did not relinquish.

Among overachieving teams, the team with the largest legitimate gap (the aforementioned McGill Martlets had the largest difference, from -25% to 0%) was the Trinity Western University Spartans. The Spartans finished the regular season with an 83% win per-

eFG	TOV	ORB	FTF	oppeFG	oppTOV	oppORB	oppFTF
44.7	23.1	34.7	19.8	38.5	22.3	30.0	19.0

Table 2.6: Trinity Western Spartans 2021-2022 Regular Season Statline

centage, 16 points (3 wins) more than their expected total. Their season statline is shown in Table 2.6, where Trinity Western averaged solid four factor totals across the board, particularly in offensive and defensive eFG%. However, their 28^{th}-ranked TOV% and 27^{th}-ranked opponent TOV% were not reflective of the success that they ended up having, and the plausible culprit of their inflated win percentage. Recall that TOV% carries nearly 1/3 of the four factor weightings in women's basketball, and even with positive differences in the other three categories, it was expected to have a greater detriment to the predicted outcomes of Trinity Western's games. So, while the Spartans were able to avoid this reality during the regular season, turnovers are what ultimately cost them a chance to contend for a berth in the national championship tournament. In the Canada West playoffs, Trinity Western won their first-round matchup against the Macewan Griffins but were bounced by the lower-seeded University of Lethbridge Pronghorns in the Canada West quarterfinals, largely due to committing 21 turnovers to Lethbridge's 13.

Chapter 3

Win Probability

Now that the predictive capabilities of the four factors with a team's expected win percentage over the course of a season have been established, the next logical progression is to explore how well the four factors perform on a game-by-game basis. With so many possible factor combinations that can result in a win, a large data set is required to account for the increased variability of inputs, maximizing the model's predictive success. A well-defined model that produces accurate experimental results is appealing and could boast more concrete functionality than the original win percentage construct. For example, mini-applications can easily be created to update four factor scores in real-time throughout a game and relay important information to coaches and players about their performance. These totals can then be input into a win probability model to determine whether the current game state will result in a victory for the team in question. If not, the team will know exactly how much improvement is required, and in which factor areas. Target factor performances can also be prepared in advanced by fitting the model with sample inputs to simulate both a most likely outcome and one that maximizes win probability. After examining the attributes of the devised probability model, its results will be considered on the 2021-2022 data set, including a case study on some particularly noteworthy games that occurred during the regular season.

3.1 Methodology

When considering discrete events that take a binary outcome, probability models provide important insights into the relationships between explanatory variables and the result. Sporting contests such as basketball games are ideal candidates for such models, with an outcome of 1 signifying a win and 0 representing a loss.

The probability model of choice is logistic regression due to its classification capability. Given **X**, the set of explanatory variables X_1, X_2, $\ldots$, X_m, logistic regression predicts the probability of response variable Y = 1, Pr(Y), in the cumulative standard logistic distribution, which is evaluated at $\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$: [9]

$$Pr(Y = 1|\mathbf{X}) = F(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m)$$

where F is the cumulative logistic distribution function:

$$F(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m)}}$$

Since a logistic regression model is confined to probability values between 0 and 1, the interpretation of its coefficients is different than in a linear regression. This is facilitated by the logit function, which gives the log odds of an event occurring as:

$$logit(Pr(Y)) = log\left(\frac{Pr(Y)}{1 - Pr(Y)}\right)$$

This means that a 1 unit change in explanatory variable X_i will change the log odds of Y (winning a game) by its coefficient β_i, and these β_i's can be transformed into odds and probabilities for a more familiar interpretation.

Since logistic regression models are fit differently than their linear counterparts, maximum likelihood estimates are used instead of least squares. Therefore, to compare logistic regression models, we use measures connected to these estimates such as the Akaike Information Criterion (AIC), $AIC = 2k - 2ln(\hat{L})$, where k is the number of parameters in the model and $\hat{L}$ is the maximum value of the likelihood function. In general, the higher the Log-Likelihood, the better the model. The converse is true with AIC, where models with a smaller AIC indicate a better fit model.

McFadden's Pseudo R-squared [20], which can be interpreted as the overall effect size, is another effective tool for measuring model effectiveness and is calculated as

$$R^2_{McFadden} = 1 - \frac{ln(L_{full})}{ln(L_{null})}$$

It is the ratio of the log-likelihood of the complete model (L_{full}) over the log-likelihood of the model with no explanatory variables, (L_{full}).

3.2 Model Description

Basketball games are symmetric in definition such that for two teams T_1 and T_2 that play each other, T_1's offensive factor scores $eFG\%_{T_1}$, $TOV\%_{T_1}$, $ORB\%_{T_1}$, $FTF\%_{T_1}$ will correspond to T_2's defensive factor scores, and T_2's offensive factor scores $eFG\%_{T_2}$, $TOV\%_{T_2}$, $ORB\%_{T_2}$, $FTF\%_{T_2}$ will also be T_1's defensive factor scores. Therefore, for the 2376 games in the training set, there will be $2376 \times 2 = 4752$ statlines, and a "team of interest" for each game outcome must be defined to ensure that these statlines are not double-counted. This did not arise in the expected win percentage model because unlike an individual game, no two teams will have symmetric offensive and defensive factor scores over a season.

Ambiguity arises when deciding which team to define as T_1 and which team to denote as T_2. If the winning team is always chosen as T_1, then the response variable would be 1 for each observation which would not properly train the model to predict outcomes based on the factor magnitudes. To bypass this potential for error and ensure consistent representation, T_1 was specified as the **home team** (no prefix) and T_2 as the **away team** (prepended with "opp"). The binary response variable then determines the probability of the home team winning the game ($WinProbability_h$)given the statline with the home team's offensive and defensive factors. The results can also be interpreted symmetrically since the win probability p for the home team implies a win probability of 1 - p for the away team. With the response variable appropriately specified, the construction of the win probability model becomes very straightforward. Analogous to the expected win percentage variant, versions of the model with and without constant terms will be considered.

$$\begin{aligned} z = \beta_1 eFG\% + \beta_2 oppeFG\% \beta_3 TOV\% + \beta_4 oppTOV\% \\ + \beta_5 ORB\% + \beta_6 oppORB\% + \beta_7 FTF\% + \beta_8 oppFTF\% \end{aligned} \tag{3.1}$$

$$WinProbability_h = \frac{1}{1 + e^{-(\beta_0 + z)}} \tag{3.2}$$

$$WinProbability_h = \frac{1}{1 + e^{-z}} \tag{3.3}$$

3.3 Analysis

Table 3.1 reports the stargazer output of the regressions run from equations 3.2 and 3.3, with the same nomenclature as the expected win percentage models from Chapter 2. Each coefficient across the four models has a p-value of less than 1%, and the signs on each term are identical to those from the output of Tables 2.1 and 2.2, strengthening the notion of correct model specification. Unlike the linear regression models however, the variable with the largest impact on the log odds of winning in all four logistic regression models is defense ($aeFG$), affecting win probability by more than an equivalent change in any other factor. From model M_1 we see that a 1 percent change in the home team's ORB% is predicted to change their log odds of winning a game by 0.245, and the rest of the coefficients in each

Table 3.1: Logistic Regression Output for Men's and Women's Games

	Dependent variable:			
	Win Probability			
	M_1	M_2	W_1	W_2
eFG	0.578***	0.580***	0.868***	0.868***
	(0.035)	(0.034)	(0.068)	(0.068)
oppeFG	−0.609***	−0.607***	−0.912***	−0.915***
	(0.037)	(0.036)	(0.073)	(0.072)
TOV	−0.540***	−0.539***	−0.660***	−0.662***
	(0.037)	(0.036)	(0.056)	(0.055)
oppTOV	0.538***	0.538***	0.744***	0.743***
	(0.038)	(0.038)	(0.063)	(0.062)
ORB	0.245***	0.245***	0.314***	0.313***
	(0.019)	(0.018)	(0.029)	(0.029)
oppORB	−0.202***	−0.201***	−0.313***	−0.314***
	(0.016)	(0.016)	(0.029)	(0.028)
FTF	0.146***	0.146***	0.239***	0.239***
	(0.015)	(0.015)	(0.024)	(0.024)
oppFTF	−0.146***	−0.146***	−0.227***	−0.227***
	(0.014)	(0.014)	(0.024)	(0.023)
Constant	0.264		−0.243	
	(1.134)		(1.427)	
Observations	2,376	2,376	2,376	2,376
Log Likelihood	−287.613	−287.640	−155.191	−155.206
Akaike Inf. Crit.	593.227	591.281	328.382	326.411
McFadden Pseudo R^2	0.823	0.825	0.905	0.906
Note:				*p<0.1; **p<0.05; ***p<0.01

model can be interpreted similarly. Another noticeable detail about the models is that they are extremely sensitive to small changes in factor scores. For a numerical example, consider two women's teams playing a game where at halftime they report identical factor scores of league average values, such as those taken from Table 1.3. With a 41.3% eFG%, 22.9% TOV%, 32.0% ORB%, and 19.4% FTF%, each teams' win probability would be approximately 50% (53%/47% in favour of the home team using W_2). If the home team then, whether by choosing more efficient play types or by improving their shot selection, raised their overall eFG% by just 1 percent in the second half of the game to 42.3%, their expected win probability would increase to nearly 73% ($z = 0.9826$, please refer to Equations 3.1 and 3.3 for full calculation details). These seemingly small adjustments, taking into account which factor the improvement targets and its relative importance, can substantially alter the trajectory of a game, with execution of gameplanning acting as a key component of their implementation.

The constant term has a minimal effect on model performance and the models that do and do not include it have nearly identical values in their coefficients, Log-Likelihood, AIC, and Pseudo R-squared. Unlike W_2 and M_2 from the linear regression models in Tables 2.2 and 2.1, however, there is a noticeable difference between the men's and women's results. The four factors account for over 90% of the variability in the women's game, a much better fit than the 82% covered in the men's game, which is still remarkably high. The predictive results on the test set will inform how well the models can be extended into the future, but it is clear that the women's results over this time frame were more conclusively described than those of men. Reviewing other data sources such as game film to analyze differences in playing styles could prove an interesting exercise to determine why such discrepancies in variation exist.

To begin comparing and contrasting offensive and defensive contributions with the models from Table 3.1, we note that turnover percentage and free throw factor are equal on both sides of the ball for M_2. The largest difference belongs to rebounding, as own offensive

Factor	Men	Women
eFG%	39	41
TOV%	36	33
ORB%	15	15
FTF%	10	11

Table 3.2: Four Factor Weights for Win Probability

rebounding is higher rated by almost 4.5 points over the opponent's score. However, there is no such difference within the women's game, presumably because of the more even offensive rebound distribution discussed in Chapter 1. Instead, the most substantial difference in W_2 is the 8-point swing that forcing opponent turnovers instead of committing them has on win probability. The magnitude of the coefficients also provides context for relative factor importance, and Table 3.2 shows the recalculated linear weights of each factor in determining win probability. It is interesting to note that although most of the weights are similar to their expected win percentage counterparts, within-game TOV% see an increased level of importance in both basketball leagues, from 33% to 36% among men and from 31% to 33% in the women's game. The TOV% increase in men's basketball mirrors the surprising drop in eFG%, from 44% across a season to 39% in a game. By developing a strategy built around an aggressive defense that focuses on winning the turnover battle, teams can potentially circumvent poor shooting performances and exploit the diminished eFG% importance to ”steal” games. Of course, the expected win percentage model demonstrates that such emphasis is not sustainable across the course of a season since the number of total possessions will reveal superior teams. Nevertheless, these gameplans can make a big difference for isolated games and winner-take-all matchups.

3.4 Predictive Ability

Similar to the expected win percentage models discussed in Chapter 2, the regular season games from the 2021-2022 season were used to test the performance of the win probability

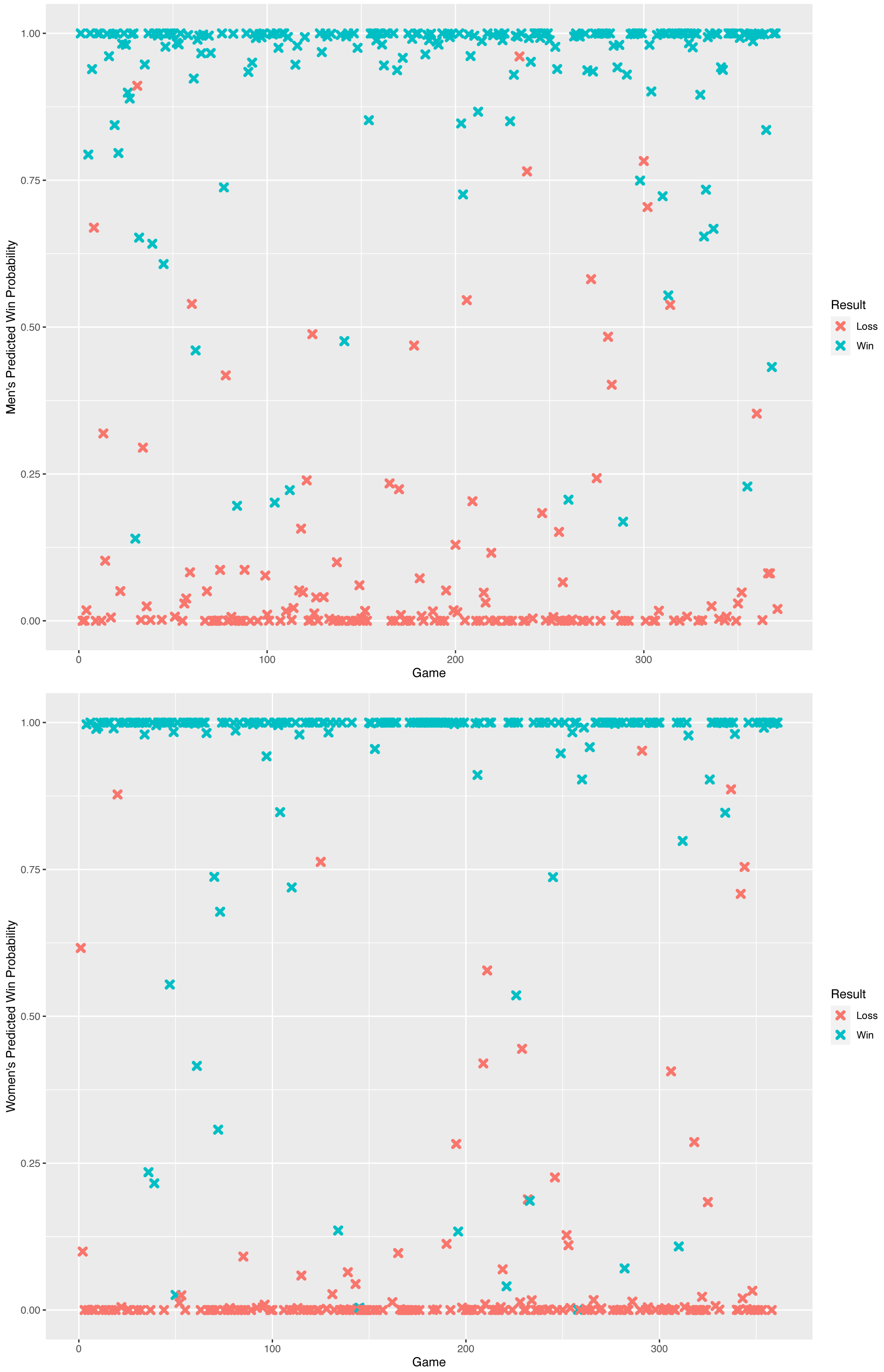


Figure 3.1: Men's and Women's Win Probability Predictions

models. Four factor scores from the results of these games were used as inputs, and a predicted win probability was computed. This probability was then classified as a win if it was ≥ 0.5 and a loss otherwise. After classifying the predicted values from the data set of nearly 400 games, they were compared to the actual outcomes, and the model accuracies were determined. Although the models that did not include the intercept term produced slightly different individual win probabilities than those that included the constant, the predicted outcomes were identical in terms of wins and losses. The men's model correctly predicted 351/371 games, good for a 94.61% accuracy, and the women's model experienced similar success by correctly predicting 94.18% (340/361) of games. These results are shown in Figure 3.1 and come as a bit of a surprise since the women's model was hypothesized to have greater predictive power due to its performance on the training set.

What immediately stands out when viewing the figure is how densely packed the data points are at the extreme ends of each image. Looking at games with indeterminate or less certain outcomes (arbitrarily defined as win probabilities between 25% and 75%), the men's graphic contains more variability in the predicted win probabilities, whereas the number of games with uncertain outcomes can easily be counted in the women's plot. Contrary to the natural inclination that games would experience more variation in the expected outcome, the win probability model demonstrates how conclusive the four factors' assessment of a game can be. To illustrate, consider that 72% of women's games had win probabilities greater than 99%, even though the final score of 35 (over a tenth) of these games was determined by a single-digit margin. Such games would be deemed "close" to the casual observer when they instead had decisive outcomes. While this clustering of probabilities demonstrates the effectiveness of the model in confidently predicting the outcomes of games, it unintentionally restricts its interpretation to simply a win or a loss since most factor combinations will result in probabilities with extreme values.

3.4.1 Case Study: Investigating Statistically Improbable Wins

Motivated by the remarkable accuracy of the win probability model, one cannot help but fixate on the games that were mispredicted and attempt to reconcile the events that led to such an upset. In this context, the word “upset” does not describe a lower-seeded team taking out a higher ranked-opponent but rather a team that was statistically outperformed but still came away with a victory. For the 2021-2022 season, the set of such results is the union of teal **X**’s in the bottom half of a plot in Figure 3.1 and red **X**’s in the top half of that plot. Two of the 20 total upsets in men’s basketball fulfilled the criteria of a “major” upset, arbitrarily defined as a game in which the victor had a predicted win probability of less than 10%. Of these two games, the most statistically improbable victory belonged to the Nipissing Lakers, who escaped with a 66-65 win over the Toronto Varsity Blues on February 12, 2022. Figure 3.3 details the four factors from that game, a statline that gave the Lakers a win probability of a measly **3.93%**.

The Varsity Blues had a 50.0% eFG%, an above-average output, while the Lakers committed a turnover on over a quarter of their possessions. The second-chance opportunities granted to Nipissing by their excellent offensive rebounding were essentially neutralized by their substantial TOV%. In fact, the Varsity Blues finished the game with more field goal attempts, making free throws the culprit of this game’s result. Toronto’s minuscule 10.2% FTF% in this game did not come as a result of missed attempts, as their 66.7% ftPCT was only slightly below their season average of 67.3% and better than Nipissing’s 63.0% performance. Therefore, the x-factor in this game was not the percentage of made free throws but rather the sheer volume of attempts. Toronto managed to get to the free line a mere nine times, compared to Nipissing’s 27 trips to the charity stripe. This cost the Varsity Blues an otherwise statistically sure win and demonstrates a unique situation where counting totals proved more influential than percentages.

There were *six* major upsets in women’s basketball during the 2021-2022 season, the

Factor	Nipissing	Toronto
eFG%	42.2	50.0
TOV%	26.1	15.6
ORB%	46.3	25.7
FTF%	29.3	10.2

Table 3.3: Four Factor Scores for Nipissing vs Toronto from February 12, 2022

Factor	Waterloo	Laurier
eFG%	20.5	30.9
TOV%	28.6	14.9
ORB%	44.7	18.4
FTF%	24.2	14.5

Table 3.4: Four Factor Scores for Waterloo vs Wilfred Laurier from February 18, 2022

most improbable of which came less than a week after Nipissing's victory over Toronto, down Highway 401 in nearby Waterloo. On February 18, 2022, the host Warriors squeezed past their close rival Wilfred Laurier Golden Hawks by a final score of 43-42. A cursory look at the four factor game scores shown in Table 3.4 quickly justify the Warriors **0.046%** win probability, as Waterloo was dominated in both eFG% and TOV%, the factors that dictate a combined 74% of a team's expected win probability. Waterloo's 20.5% eFG% was the eighth-lowest recorded in the 361 games played that season, but a strong rebounding performance on both sides of the glass kept the game close. Similar to the men's result, further investigation into the FTF% recorded by both teams reveals the cause of this unthinkable outcome. The Warriors went an impressive 16/17 from the free throw line, the *third-best* performance in U SPORTS (minimum 15 attempts), and a stark contrast to Laurier's 8/21 effort, the *fourth-worst* mark that season. Although it has been clearly demonstrated that proficient FTF% will not sufficiently compensate for poor performance in the other factors, Waterloo's wide differential in this game was enough overcome all of the odds and snatch a clutch victory from the jaws of near-certain defeat.

Chapter 4

Conclusion

Advanced statistical methods utilizing regressions and Bayesian-based constructs have been used for over 20 years in the NBA to enhance the quantitative understanding of the game of basketball. This paper investigated an application of such frameworks, the Four Factors of Basketball, within U SPORTS to assess their efficacy when addressing questions of how to evaluate team performance. These factors were used in multiple linear regression models to determine how well they represent and explain team success. They produced excellent results on the training data set, with the models that did not include the intercept term explaining nearly 98% of the variation in win percentage. Each factor was assigned a linear weight based on its coefficient and compared to past and present NBA benchmarks. These weights defined the importance of each factor and its impact on a game, suggesting areas for players and coaches to focus on optimizing to maximize their team's success. The final test compared predicted win percentages with those reported during the 2021-2022 regular season, and the models fared incredibly well, with low error rates and R^2 values near or above 87%. This predictive success led to two case studies that illustrated some unique consequences of the model construction. The first analyzed extreme values to determine the best and worst statistical teams in U SPORTS, and the second used residuals to identify teams with the greatest perceived overachievement and underachievement.

The paper concluded by using the 2376 boxscores that comprised the 5-year training set to create logistic regression win probability models. Like the expected win percentage tests, models with and without constant terms were developed and compared. The men's games demonstrated more volatility than those on the women's side, but the models still proved to be excellent descriptors of the factors that affected the game outcomes, with McFadden's Pseudo R-Squared values greater than 0.82 and 0.91, respectively. The performance of the models on the test set further cemented their dependability, as they accurately predicted over 94% of game outcomes. Such reliability initiated curiosity about the incorrect predictions, leading to an investigation of the most statistically improbable victories during the regular season. It has been demonstrated that advanced analytical instruments used within the NBA maintain excellent functionality when applied to data from U SPORTS basketball. However, the introduction of tools such as the four factors has only scratched the surface of resources available to increase the quantitative understanding of the game. Advances in these areas will ensure that U SPORTS continues to produce high-quality basketball for both participants and spectators, generating much excitement about the game's future.

4.1 Future Works

A question that arises following the effective performance of advanced NBA analytical tools in evaluating team performance within U SPORTS basketball is whether these methods can be extended to serve as proxies for quantifying a player's talent. Individual player metrics are a hotly contested subject area in NBA circles due to the relative secrecy of their formulation, but they generally fall into two categories. The first are statistics that measure a player's value in terms of their contributions to their team's success in a specific area. For example, possession-ending events can be classified per player so that their statline can be used to calculate their individual factors totals, a technique used in applications such as Synergy Sports Technology [36]. Additionally, play-by-play and spatial data make it possible to determine a player's ORB%, which cannot be computed using only boxscore numbers.

The other branch of individual player metrics are called "catch-all" statistics, which aim to summarize the observable contributions of a player and aggregate them into one number to determine their overall value. Preliminaries on U SPORTS data have been run for metrics such as regularized adjusted plus-minus (RAPM) and expected box plus-minus [17], and another metric worthy of exploration within U SPORTS is Player Efficiency Rating (PER) [10]. PER was created by former ESPN columnist John Hollinger and was the first all-in-one statistic to attempt to quantify a player's value. It takes a players' boxscore contributions: field goals, free throws, 3's, assists, rebounds, blocks and steals, turnovers, missed shots, and fouls, adding the positive contributions and penalizing the negative ones using detailed formulas [19]. It adjusts for per-minute productivity and team pace, as well as the league averages in all the statistical categories, and is normalized such that a PER of 15 is considered league average. It is biased towards offensive output since these totals are better recorded in boxscsores so players with an unseen defensive impact will be considered underrated overall, but as the first of its kind PER revolutionized player evaluations during the 2010s. The top 10 male and female U SPORTS players ranked by PER during the 2021-2022 season are displayed in Tables 4.1 and 4.2.

Rank	**Player**	**Team**	**PER**
1	Tajinder Lall	Brock	34.6
2	Sukhman Sandhu	UBC	32.6
3	Biniam Ghebrekidan	Carleton	32.1
4	Michael Okafor	Lakehead	31.0
5	Lloyd Pandi	Carleton	30.7
6	Thomas Kennedy	Windsor	30.4
7	Prince Kamunga	York	29.0
8	Olivier Simon	Concordia	28.8
9	Grant Shephard	Carleton	28.7
10	Keevan Veinot	Dalhousie	28.2

Table 4.1: 2021-2022 Top 10 U SPORTS Players by PER - Men

Rank	Player	Team	PER
1*	Kiyara Letlow	Cape Breton	38.7
2	Brigitte Lefebvre-Okankwu	Ottawa	36.5
3	Sarah Gates	McMaster	35.9
4	Julia Chadwick	Queens	34.9
5	Burke Bechard	Guelph	33.6
6	Amaiquen Siciliano	Bishops	33.4
7	Jade Belmore	Regina	33.0
8	Nicole Fransson	Trinity Western	32.6
9	Summer Masikewich	Saskatchewan	31.7
10	Carly Ahlstrom	Saskatchewan	31.5

*Jenna Mae Ellsworth of UPEI had a PER of 44.9
but only played in 7 of the team's 13 games.

Table 4.2: 2021-2022 Top 10 U SPORTS Players by PER - Women

These tables give insight into Carleton's statistical dominance on the men's side during the 2021-2022 season, as they produced three of the top 10 PER accumulators, including U SPORTS Player of the Year Lloyd Pandi. The women's results stand out with Cape Breton's Kiyara Letlow, who took home U SPORTS rookie of the year honours, and two Saskatchewan Huskies cracking the top 10 as well. There are limitations to this U SPORTS-specific application of PER however, due to lack of information surrounding its original regression and plus-minus methods. A low-level description of the coefficients that compute the league value of an offensive possession and the importance of each player's contribution is not publicly available. Therefore, while PER is effective in general for intra-league player comparisons, its values become increasingly approximate the farther away the style of play of the league is from that of NBA. For example, the discussion from Section 1.3 about differing offensive rebounding distributions and the large eFG% discrepancy between U SPORTS men and women should result in a different PER weight given to each contribution, which isn't the case in the actual formulation. Therefore, attaining the additional information necessary to recreate PER and similar catch-all metrics from the ground up will improve our notion of player value within U SPORTS.

References

[1] Tarek Al Baghal. Are the four factors indicators of one factor? an application of structural equation modeling methodology to nba data in prediction of winning percentage. *Journal of Quantitative Analysis in Sports*, 8, 01 2012.

[2] J. Albert, M.E. Glickman, T.B. Swartz, and R.H. Koning. *Handbook of Statistical Methods and Analyses in Sports*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2017.

[3] Austin Clemens. Nylon calculus 101: Plus-minus and adjusted plus-minus. `https://fansided.com/2014/09/25/glossary-plus-minus-adjusted-plus-minus/`. Accessed: 2022-02-26.

[4] eCoach. How to use analytics to help improve your basketball team with coach dean oliver (ecoachcast #2). `https://www.youtube.com/watch?v=yVlBBkPVF-M`. Accessed: 2022-08-19.

[5] Jeremias Engelmann. Possession-based player performance analysis in basketball (adjusted +/- and related concepts). 2016.

[6] Edgar Santos Fernandez, Paul Wu, and Kerrie Mengersen. Bayesian statistics meets sports: a comprehensive review. *Journal of Quantitative Analysis in Sports*, 15(4):289–312, 2019.

[7] Alexander Franks, Andrew Miller, Luke Bornn, and Kirk Goldsberry. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9, 05 2014.

[8] Stephanie Glenn. Rmse: Root mean square error. https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/. Accessed: 2022-04-14.

[9] R. Carter Hill, William E Griffiths, and G. C. (Guay C.) Lim. *Principles of econometrics / R. Carter Hill, William E. Griffiths, Guay C. Lim.* Wiley, Hoboken, fifth edition. edition, 2018.

[10] J. Hollinger. *Pro Basketball Forecast*. PRO BASKETBALL PROSPECTUS. Potomac Books, Incorporated, 2005.

[11] Justin Jacobs. Introduction to oliver's four factors. https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/. Accessed: 2022-04-16.

[12] Bryan Kalbrosky. What is the best advanced statistic for basketball? nba executives weigh in. https://hoopshype.com/lists/advanced-stats-nba-real-plus-minus-rapm-win-shares-analytics/. Accessed: 2022-03-01.

[13] Konstantinos Kotzias. The four factors of basketball as a measure of success. https://statathlon.com/four-factors-basketball-success/. Accessed: 2022-05-09.

[14] Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan Rosenbaum. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3:1–1, 02 2007.

[15] B.G. Lemmon. What should the rim height be in the wnba? https://medium.com/@bglemmon/what-should-the-rim-height-be-in-the-wnba-3ffb339822be. Accessed: 2022-04-15.

[16] M.M. Lewis. *Moneyball: The Art of Winning an Unfair Game.* Norton paperback. W.W. Norton, 2003.

[17] Peter L'Oiseau. Process-oriented player evaluation metrics for usport basketball. 2021.

[18] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2.3, 2022.

[19] Eric Maroun. Understanding advanced statistics: Player efficiency rating. https://web.archive.org/web/20170910105350/https://hardwoodparoxysm.com/2012/03/07/understanding-advanced-statistics-player-efficiency-rating/. Accessed: 2022-04-18.

[20] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1972.

[21] Milos Mudric. How the nba data and analytics revolution has changed the game. https://www.smartdatacollective.com/how-nba-data-analytics-revolution-has-changed-game/. Accessed: 2022-08-19.

[22] Daniel Myers. About box plus/minus (bpm). https://www.basketball-reference.com/about/bpm2.html. Accessed: 2022-02-21.

[23] Krishna Narsu and Tim/Cranjis McBasketball. Lebron introduction. https://www.bball-index.com/lebron-introduction/. Accessed: 2022-03-01.

[24] NBAstuffer. Dean oliver. https://www.nbastuffer.com/analytics101/dean-oliver/. Accessed: 2022-03-25.

[25] National Baseball Hall of Fame. Henry chadwick. https://baseballhall.org/hall-of-famers/chadwick-henry. Accessed: 2022-02-21.

[26] Dean Oliver. Journal of basketball studies. http://www.rawbw.com/~deano/. Accessed: 2022-03-21.

[27] Dean Oliver. *Basketball on paper : rules and tools for performance analysis / Dean Oliver.* Brassey's, Inc., Washington, D.C, 1st ed. edition, 2004.

[28] Ken Pomeroy. Advanced analysis of college basketball. https://kenpom.com. Accessed: 2022-04-27.

[29] PrestoSports. Prestosports — all-in-one sports technology platform. https://www.prestosports.com/landing/index. Accessed: 2022-02-21.

[30] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022.

[31] Basketball Reference. Glossary. https://www.basketball-reference.com/about/glossary.html. Accessed: 2022-02-23.

[32] Duane Rockerbie. *The Economics of Professional Sports 2017.* 11 2017.

[33] Nate Silver. Introducting raptor, our new metric for the modern nba. https://fivethirtyeight.com/features/introducing-raptor-our-new-metric-for-the-modern-nba/. Accessed: 2022-03-01.

[34] U SPORTS. U sports. https://usports.ca/en. Accessed: 2022-02-21.

[35] T.M. Tango, M.G. Lichtman, and A.E. Dolphin. *The Book: Playing the Percentages in Baseball.* Potomac Books, 2007.

[36] Synergy Sports Technology. Synergy sports. https://synergysports.com. Accessed: 2022-04-18.

[37] Zachary Terner and Alexander Franks. Modeling player and team performance in basketball, 2020.

[38] Martin Timmerman. U sports hoops, university baskbetball in canada. https://usportshoops.ca. Accessed: 2022-02-21.

[39] Petar Vračar, Erik Štrumbelj, and Igor Kononenko. Modeling basketball play-by-play data. *Expert Systems with Applications*, 44:58–66, 2016.

[40] Sports Management Worldwide. Dean oliver. https://www.sportsmanagementworldwide.com/users/dean-oliver. Accessed: 2022-03-25.

[41] P. Zuccolotto and M. Manisera. *Basketball Data Science: With Applications in R*. Chapman & Hall/CRC Data Science Series. CRC Press, 2020.

Appendix A

Code

The following R [30], Libraries were used:

- dplyr
- rvest
- stringr
- tibble
- tidyr

The files in this section concern the main precomputational task of this project - scraping the information for every team and game, and calculating the advanced metrics discussed in the paper.

The following scripts are included in this appendix:

- TeamSummaries.R - Scrape Coach's View Team Summaries
- GameBoxscores.R - Scrape Individual Game Boxscores
- FourFactors.R - Compute Four Factors for either a game or a season

A.1 TeamSummaries.R

This script scrapes the Coach's View Team Summaries for each team. The required inputs are the team name, the season, and the link of the website, separated into the link prefix and suffix.

```r
## Specifying the url
teamUrl <- paste(linkPrefix, season, teamName, linkSuffix)

## Reading the HTML code from the website
webpage1 <- read_html(teamUrl)

team_html <- html_nodes(webpage1, "td:nth-child(1)")
team_data <- html_text(team_html)

# Remove Unnecessary Characters
for (x in 1:length(team_data)) {
  team_data[x] <-  str_replace_all(team_data[x], "[\\s]", "")
}
team_data <- team_data[team_data != ""]

## Scrape Roster Data
roster <- length(team_data)-1
team <- ""
for ( i in 2:(roster+1)) {
  cssSection <- paste0("tr:nth-child(", (2*i))
  if (i == roster+1) {
  cssSection <- paste0(cssSection, ")␣td␣")
  } else
    cssSection <- paste0(cssSection, ")␣td,␣")
  team <- paste(team, cssSection)
}

## Use CSS selectors to scrape the team section
playerStats_html <- html_nodes(webpage1, team)
cssSection <- paste0("tr:nth-child(", (2*roster)+5)
cssSection <- paste0(cssSection, ")␣.align-center,␣tr:nth-child(")
cssSection <- paste0(cssSection, (2*roster)+7)
cssSection <- paste0(cssSection, ")␣.align-center")
cssSection
teamStats_html <- html_nodes(webpage1, cssSection)
playerStats_data <- html_text(playerStats_html)
teamStats_data <- html_text(teamStats_html)
```

After *playerstats_data* and *teamStats_data* have been formatted, they are saved as data frame called *team*.

A.2 GameBoxscores.R

This mostly functional script scrapes in-game boxscores off of the usportshoops website, requiring the link prefix, roster sizes of both teams, and a specially structured gameID.

```r
gameUrl <- paste0(linkPrefix, gameId)

#Reading the HTML code from the website
webpage1 <- read_html(gameUrl)
homeTeam <- substr(gameId[1], 10, 12)
awayTeam <- substr(gameId, 13, 15)

#Using CSS selectors to scrape the boxscore data
winningTeam_html <- html_nodes(webpage1, "tr:nth-child(1) b")
winningTeam_data <- html_text(winningTeam_html)

## Get home and away boxscores
offset <- 4
firstCharacter <- ""
while (firstCharacter != "*") {
  cssSection <- paste0("tr:nth-child(", (homeRoster)+offset)
  cssSection <- paste0(cssSection, ") td")
  homeStats_html <- html_nodes(webpage1, cssSection)
  homeStats_data <- html_text(homeStats_html)
  offset <- offset - 1
  firstCharacter <- substr(homeStats_data[1], 1, 1)
}
offset <- 0
firstCharacter <- ""
while (firstCharacter != "*") {
  cssSection <- paste0("tr:nth-child(", (awayRoster+homeRoster+offset))
  cssSection <- paste0(cssSection, ") td")
  awayStats_html <- html_nodes(webpage1, cssSection)
  awayStats_data <- html_text(awayStats_html)
  offset <- offset + 1
  firstCharacter <- substr(awayStats_data[1], 1, 1)
}
if (length(homeTeamIndex)==0) {  # The away team won and is listed first
  homeTeamWon <- 0
  gameStats <- c(awayStats_data, homeStats_data)
} else {  #The home team won and is listed first.
  homeTeamWon <- 1
  gameStats <- c(homeStats_data, awayStats_data)
}
```

After *gameStats* has been parsed, the data is formatted as a data frame called *team*.

A.3 FourFactors.R

This script computes a team's Four Factors from either the team summary or the game boxscore using the formulas discussed in Section 1.3. It also calculates a team's winning percentage (for season summaries) and pace, another useful metric for calculating team success. The input for this function is the roster size, called rs.

```
eFG <- ((team$FG[rs+1]) + (0.5*team$`X3FG`[rs+1]))/team$FGA[rs+1]
oppeFG <- ((team$FG[rs+2]) + (0.5*team$`X3FG`[rs+2]))/team$FGA[rs+2]

POSS <- ((team$FGA[rs+1]) + (0.44*(team$FTA[rs+1]))
    - (team$OFF[rs+1]) + (team$TO[rs+1]))
oppPOSS <- ((team$FGA[rs+2]) + (0.44*(team$FTA[rs+2]))
    - (team$OFF[rs+2]) + (team$TO[rs+2]))

TOV <- (team$TO[rs+1]) / POSS
oppTOV <-(team$TO[rs+2]) / oppPOSS

ORB <- (team$OFF[rs+1]) / ((team$OFF[rs+1]) + (team$DEF[rs+2]))
DRB <- (team$DEF[rs+1]) / ((team$OFF[rs+2]) + (team$DEF[rs+1]))

oppORB <- (1 - DRB)
oppDRB <- (1 - ORB)

FTF <- ((team$FT[rs+1]))/((team$FGA[rs+1]))
oppFTF <- ((team$FT[rs+2]))/((team$FGA[rs+2]))

NAME <- teamName
WinPercentage <- Wins / (Wins + Losses)
Pace <- (5*((POSS + oppPOSS)/2) / team$MIN[rs+1])
```

These calculations are then saved in a data frame and used for the different regression models.