Expected Possession Value: An Evaluation Framework for Decision-Making, Strategy, and Execution in Basketball

by

Ivan C. Jutamulia

B.S. Computer Science and Engineering,
Massachusetts Institute of Technology, 2020

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 14, 2021

Certified by .....
Anette "Peko" Hosoi
Associate Dean of Engineering; Neil and Jane Pappalardo Professor,
Mechanical Engineering
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Expected Possession Value: An Evaluation Framework for Decision-Making, Strategy, and Execution in Basketball

by

Ivan C. Jutamulia

Submitted to the Department of Electrical Engineering and Computer Science
on May 14, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Quantifying decision-making in professional basketball has been an extremely challenging area of research in the past decade, with potentially very fruitful and powerful insights to be drawn as NBA organizations want to understand cognitive aspects of athlete performance. This work seeks to develop an objective framework for evaluating decision-making, while simultaneously making inferences around strategy and execution efficacy.

I construct a metric called Expected Possession Value (EPV) computed through tracking data that is then leveraged to identify scoring opportunities throughout a game. I then analyze these opportunities as instances of decision-making, quantifying how often those opportunities are missed and how good those opportunities were. Looking at team opportunities as a whole and relying on the notion of expectation, I am then also able to make judgements on how much of a team's performance can be attributed to their strategy versus their execution. Through this analysis, I show that using EPV is an effective framework for extracting quantitative measures to aid in decision-making evaluation through tracking data.

Thesis Supervisor: Anette "Peko" Hosoi
Title: Associate Dean of Engineering; Neil and Jane Pappalardo Professor, Mechanical Engineering

Acknowledgments

I would first like to express my sincere gratitude to my supervisor Professor Peko Hosoi for allowing me the opportunity to conduct this research, and for mentoring me in my Master's studies. Thank you for exposing me to the intersection of science, technology, and sports, and for all the encouragement throughout this project.

My work on this project would not be possible without the support of the San Antonio Spurs organization and the Google Cloud platform. In particular, I would like to acknowledge Nick Repole from the Spurs for the valuable guidance and feedback from the perspective of an NBA organization. I would also like to thank Ramzi BenSaid from Google Cloud for his tremendous help in optimizing my system.

In addition, I would like to express my appreciation to my colleagues in the MIT Sports Lab Consortium for providing insightful thoughts, perspectives, and feedback throughout my project. Being a part of this research group was an extremely constructive and valuable experience for me.

Lastly, I would like to thank my family and friends for their continued support as I complete my studies and this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the sports industry is starting to embrace data-driven analytics more and more, there has been a surge of avenues to explore in the sports research domain. The potential applications range far and wide, with teams and organizations looking to use mathematically based methods to improve athlete performance, team performance, business operations, and more. Numerous work has already been done in many of these fields, resulting in innovative analytical tools having already been adopted into general practice by many sports leagues.

The National Basketball Association in particular is a league that has welcomed this data / analytics revolution with open arms. Over the past decade, the NBA has taken initiative to partner with tracking providers to give teams tracking data for all league games. As of 2020, all 30 teams in the NBA have analytics departments, compared to in 2008 when there were only have 5 such organizations. In fact, NBA Commissioner Adam Silver even said at a 2017 analytics conference at the Wharton School of Business that "analytics are part and parcel of virtually everything we do now", highlighting just how much of a focal point it has become [1].

One area of sports research where there has yet to be substantial progress made from an analytics point of view is that of decision-making. For a team sport such as basketball that incorporates extremely complex strategies and tactical game-plans, player decision-making is integral to a team's success. Being aware of this, scouts and coaches in the NBA have made it a point of emphasis to take into consideration

and evaluate a player's decision-making ability when assessing that player as a whole. However, this evaluation process for decision-making is entirely subjective, as there currently does not exist any kind of objective framework for doing so. Moreover, there is not much work that has been done to quantify decision-making abilities. While simple statistics such as turnovers and assists may offer some intuition, they are far from painting a full picture.

The problem with these simple statistics is that they only capture the outcomes of events as they happen on the court. They ignore the hypothetical events and the counterfactuals that humans reason through during the decision-making process. For example, if a ballhandler makes a pass to a wide open teammate under the basket, their decision to do so is grounded in their belief that the outcome of it will be an open layup and two points for their team. But what if their teammate misses that easy shot? The decision made by the ballhandler to pass is not captured as an assist or any other existing statistic. This, of course, reveals how important counterfactual reasoning is when it comes to evaluating decisions. Regardless of whether the teammate made the shot, the decision made by the ballhandler to pass to their wide open teammate under the basket should be considered "good".

In essence, a key component of cognitive evaluation is that decisions should not be characterized as good or bad based on the actual outcome, but rather the *expectation* of the outcome. More specifically, characterizing all the possible resulting events of a decision and their likelihoods of happening is a more proper way to make a judgement about the decision, agnostic to the actual outcome.

The idea that decision-making evaluation should be rooted in expectation is the foundation to this research, and motivates my work in constructing a valid and accurate evaluation framework. I develop a metric called Expected Possession Value (EPV) which seeks to characterize the expected number of points for every player during an offensive possession at any point in time. EPV captures the hypothetical scenario where the ball is passed to a particular player and they take a shot, quantifying how much value that particular event would contribute to the team's point total. Equipped with this simple form of counterfactual reasoning, we can start to

compare the expected values of different hypotheticals, ultimately comparing them to what actually happened in order to say quantitatively if there were better decisions to be made.

This kind of analysis is then extended to the team scale, where we can evaluate team strategy in terms of the expected point value that their tactics can generate, regardless of how many points they are actually scoring. By leveraging EPV, we can identify opportunities that are being missed to potentially get more points out of a possession, highlighting instances where game-plan and strategy has opened up good opportunities that are not being taken advantage of. This powerful framework would then serve as a unique tool for teams to differentiate between strategy and execution issues. A team may have good strategy to create good opportunities but not be converting their shots, or perhaps they just have a bad strategy and game-plan overall. On the stats sheet however, both of these situations would look the same, simply a lack of points. Looking at this through an expectation lens, it becomes much more clear how to evaluate team strategy while being agnostic to execution. Table 1.1 summarizes the kinds of questions we can start to answer on both the strategy and execution side.

Strategy	**Execution**
Are we creating good shots?	How many good shots are we not taking?
How good are the opportunities we generate?	Are we making the right passes?
Are we leaving points on the table?	Are we shooting above or below expectation?

Table 1.1: Differentiating between strategy vs. execution.

In this paper, I will detail the machine learning approach used to develop the EPV metric, and show its applications to conduct the analysis alluded to above. Our goal with this project is to not only make strides in an underdeveloped area of sports research, but also to provide teams like the San Antonio Spurs a valuable tool for their evaluation systems.

Chapter 2

Overview

I began work on this project in September 2019 during my senior year as an undergraduate at MIT in collaboration with the San Antonio Spurs, and have continued this research throughout the entirety of my Master program. In this chapter, I will summarize each of the components of my work, giving a rationale for my approach and explaining how the system is integrated from a high level. I will also summarize the layout of this paper to prepare the reader.

2.1 Project Components

My work can be segmented into two distinct components. The first component was the formulation and development of the Expected Possession Value metric. The goal was to be able to build a module that could take as input tracking data and game data, and output a series of calculated EPV values for all time frames contained in that data. These EPV values can then be pipelined into the next major component of my work, which leverages EPV to conduct interesting analysis surrounding decision-making and team performance.

2.1.1 Developing EPV

The motivation for creating a metric such as EPV is centered around the most fundamental decisions that an NBA player makes on the offensive end of the court: whether to pass or shoot, and who to pass to. The focus then turns to answering two particular queries given player tracking data:

1. If the ballhandler shot right now, would that be a good shot?

2. Should the ballhandler pass to a teammate who could take a better shot?

Notice that both queries ask a question about the quality of a hypothetical shot. To assess the value of any particular shot, it makes sense to quantify the probability of that shot going in, which we can think of this evaluating the *shot difficulty*. Thus, we now have a much more concrete question to answer in order to quantify shot difficulty:

Shot Difficulty. What is the probability of making this shot?

The second query also incorporates an element of passing. Similarly to our approach with shots, to account for a hypothetical pass being made, we can quantify the probability of completing a particular pass, evaluating the *pass difficulty*. This gives us an analogous question for quantifying pass difficulty:

Pass Difficulty. What is the probability of completing this pass?

These queries can now be reduced to simple prediction problems which we can abstract into machine learning models. Specifically, I formulate each of these difficulty models as taking a particular shot or pass as input, and outputting the probability of converting the shot or completing the pass. The input shot or pass can be featurized entirely by the tracking data, as from it we can extract distances between players, distances to the basket, configuration of the defensive players, etc. The difficulty models are the most important modules for calculating EPV, and I will discuss them in more detail in Chapter 5.

With probability predictions coming from the difficulty models, the actual computation of EPV is very simple and intuitive. Up to this point, I have only described the EPV metric from a relatively high level. A more formal definition is as follows.

Def. **Expected Possession Value (EPV)**

EPV is the expected number of points to be scored by each offensive player if they receive the ball and shoot at that particular instance in time. It is a function of both time (t) and player (p_i), such that each offensive player has an EPV value at every moment during a possession. Given that we know where the ball is (and correspondingly who the ballhandler is), the expected points for player i is the probability of getting the ball to them via a pass and then subsequently shooting and scoring, weighted by the point value of that hypothetical shot.

Understanding intuitively what EPV measures is critical to seeing its relationship with decision-making. Since EPV is calculated for every offensive player, it captures five hypothetical shooting outcomes at every time frame during a possession dependent on the ballhandler's decision. These five outcomes result from the ballhandler either deciding to shoot themselves or to pass to any one of their four teammates to shoot, in addition to the outcome where the ballhandler decides to keep the ball and not do anything. With a concrete number to quantify the value of each of these hypothetical outcomes, we can start to evaluate which passes and which shots would be considered good, comparing to what the ballhandler actually does to make judgements on their decision-making. The system for computing EPV is summarized in Figure 2-1.

2.1.2 Analysis with EPV

A lot of interesting analysis can now be done with respect to decision-making, strategy, and execution. EPV gives us the power to reason counterfactually about hypothetical passes that could have been made and hypothetical shots that could've been taken. In

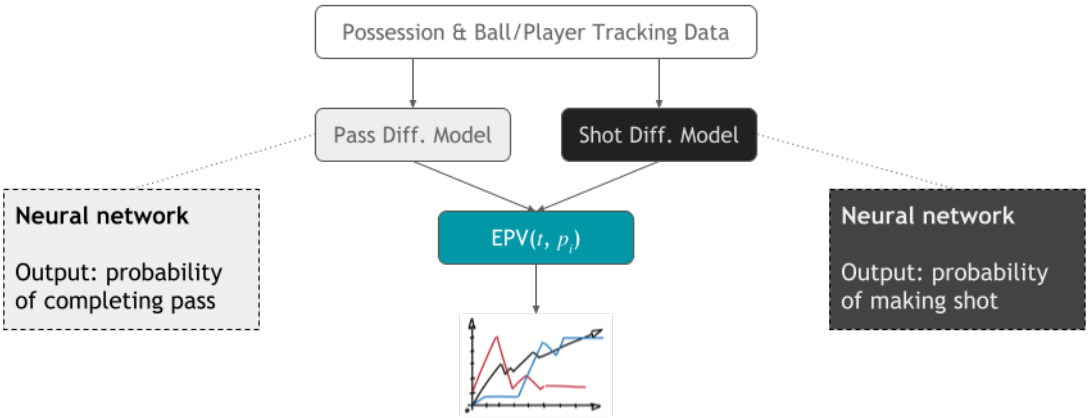

Figure 2-1: EPV Calculation Diagram. EPV for each offensive player at each time frame is calculated by passing tracking data through difficulty models. EPV evolution for each player throughout a possession can then be visualized for interpretability.

addition, quantifying these opportunities gives us a way of comparing what actually happened versus what could've happened, uncovering insights about how effective players and teams are at capitalizing on opportunities as they arise.

Identifying opportunities with high EPV and analyzing them then became the focus my work. To identify good opportunities, I looked at instances in which a player's EPV exceeded a certain threshold for a significant enough duration of time, indicating that if they had gotten the ball and shot, the offensive team would have attained a high expected points from that possession.

While we use EPV to identify these opportunities, in order to make evaluations of the players and teams, we need to consider what actually happened. Now we can classify opportunities as being or "missed" or "converted", dependent on if the ball was actually passed to the player with high EPV or not. With a large corpus of opportunities and classifications of whether they were missed or not, we can see on a player level who is good at identifying and taking advantage of these opportunities. We can also see on a team level which teams are good at creating opportunities, indicative of good strategy and game-plan, and which teams are good at taking advantage of their opportunities, indicative of efficiency and execution ability. By highlighting these missed opportunities, we can also quantify the points that are being left on the table

throughout a game. Identifying these inefficiencies are extremely valuable for NBA teams as they look for every competitive edge. While my work thus far with identifying opportunities has been extremely fruitful, I believe EPV to be a valuable tool that can be applied in many other ways that I have not had the chance to explore.

2.2 Paper Structure

In Chapter 1, I motivated the need for an expectation based framework for evaluating decision-making while also highlighting the importance of this kind of analysis in today's NBA.

In Chapter 2, I have given an high level overview of the different aspects of the project, summarizing the process of conducting this research.

Chapter 3 will be a discussion about previous work done in this area of research and how my work is unique or relates.

The data I use throughout this project is an integral component, one that warrants a chapter of its own. In Chapter 4, I will describe what this data looks like, where it comes from, and why it is useful.

Chapter 5 will cover the development of my difficulty models. I will detail the process of training each of the models in depth, showing their important features and accuracies, while also discussing difficulties and limitations in training.

In Chapter 6, we will see how the difficulty models are integrated to calculate Expected Possession Value, with accompanying visualizations to show how EPV can be interpreted.

The evaluation of my methods will be contained in Chapter 7. This includes a discussion around the accuracy of EPV and how we can evaluate it with no ground truth, as well as some limitations with the current version of my difficulty models.

Chapter 8 will focus on an application of EPV, identifying missed opportunities within a possession and throughout a game. This analysis will reveal fascinating insights both on a player and team level when it comes to decision-making.

In Chapter 9, I will discuss what all this work means on the decision-making

frontier. I will make a case for why this analysis is valid and important, and talk about ways it can be incorporated into the NBA.

Lastly, I conclude in Chapter 10 with a summary of my work and considerations for future extensions.

Chapter 3

Related Work

3.1 Literature Review

As I described earlier, little research has been conducted when it comes to decision-making in the NBA and sports in general. However, the idea of quantifying shot difficulty and leveraging the notion of expectation is a well explored concept in research and the industry. While these works don't relate their models to a discussion around decision-making and strategy, they seek to quantify very similar things that I do, serving as good proof of concepts and evaluation benchmarks for this project.

3.1.1 Quantifying Shot Difficulty

Quantifying shot difficulty or shot quality looking beyond just standard field goal percentage was an idea first proposed Chang et. al at Second Spectrum [2]. In their paper, the authors introduce the notion of *shot quality* and *effective shot quality* as "analogues to field goal percentage and effective field goal percentage". Similar to my approach, they use tracking data to consider numerous properties of a particular shot, then leverage machine learning methods to predict the likelihood of that shot going in. Their work was motivated by the idea that the difficulty of shots is highly contextual and depends on more than just pure location, which previous researchers like Kirk Goldsberry had explored [3]. Using a variety of models, including decision

trees, logistic regression, and Gaussian process regression, the authors were able to show that they could predict shot conversion with about 63% accuracy. In 2016, Second Spectrum expanded on this work to disentangle the shot quality itself and the identify/impact of the shooter [4]. This allowed the models to answer two distinct questions: how hard was the shot and how hard was the finish with respect to the difficulty of the shot. With this more advanced view on shooting quality, Second Spectrum started to answer similar research questions that I posed. In particular, by accounting for the difficulty in shots, they were able to analyze how players were shooting *with respect to expectation*, a key idea that I leverage as well, to make an evaluation on their shot-making ability. Analogously through my work, I can analyze how players are deciding to pass or shoot with respect to the expected points that those actions would result in to make an evaluation on their decision-making ability.

Another example of a piece of work quantifying shot difficulty worth noting was conducted by Krishna Narsu, following a similar approach that both Second Spectrum and I took [5]. Narsu uses a logistic regression model much like Second Spectrum to input features of a shot and output a probability of conversion as a metric to quantify shot difficulty. However, Narsu restricts his focus to high stress, high leverage situations, particularly when game-winning shots are attempted in the playoffs. As such, Narsu attempts to improve his model by incorporating features that help characterize the high stress level and urgency of the shot. Time left in the game, time left on the shot clock, being on the home team or away team, all now make much more of a difference than a typical NBA shot. In fact, when analyzing these variables for high pressure shots, Narsu found them all to be statistically significant in predicting the probability. Once more, Narsu was able to achieve about 63% accuracy in his model, consistent with Second Spectrum's standard for performance.

As a last example to show how shot difficulty has been addressed in research already, Bocskocsky et al. at Harvard University utilized a different approach motivated by an entirely different application [6]. In their work, the authors attempt to support the "hot hand" phenomenon in basketball, where a player who has hit a streak of shots has a higher perceived chance at hitting their next shot. Widely considered

a fallacy of local representativeness in academia at the time, the authors claim that that statistical studies disproving the "hot hand" theory fail to take into account that as players successively make consecutive shots, the shot selection and difficulty gets progressively more difficult. Previously, it was assumed that the next shot in a streak of shots is selected randomly from a generic ensemble of shots, an assumption that the authors relax in this work. Of course as a result, it became necessary to show that shot difficulty was correlated with streaky shooting, another example of needing to quantify shot difficulty. However, rather than computing probabilities, the authors just needed to show positive correlation. They developed an Ordinary Least Squares (OLS) model to relate shot difficulty to a statistical definition of how "hot" a player is. As before, the model used inputs related to tracking data that could characterize the context of a shot. By observing correlations with players on a shooting streak such as shot distance increasing, closest defender distance shrinking, the authors show that as player makes each additional shot in a streak, their next shot becomes 2.5% more difficult. As a conclusion, they regress a measure of how hot a player is with the probability of hitting a shot, importantly controlling for shot difficulty, to demonstrate that there is a small positive correlation, indicating the existence of the "hot hand". While not directly related to my project, this work serves as an example of how shot difficulty can be viewed from many perspectives, and can have a wide number of applications for analysis.

3.1.2 Expected Threat in Soccer

Perhaps the most closely related piece of work to my project was done in the soccer domain by Karun Singh [7]. Singh's work is motivated by the fact that simple statistics such as assists in soccer don't necessarily capture the value of passes as they pertain to contributing to goals. Building on the already developed idea of expected goals (xG) in soccer, being able to look multiple actions / passes in advance would be extremely valuable.

From a high level, Singh's strategy is very similar to my own. He seeks to quantify how many expected goals can be attained when the ball is at a particular point in the

field. This is done by first partitioning the field into a grid in order to discretize (x, y) coordinates. Singh then utilizes the reward-action concept central to reinforcement learning to quantify the value of the ball being relocated to one of the 192 distinct (x, y) coordinates. In particular, the value at each coordinate is computed iteratively as the weighted sum of expected goals of that coordinate and value from all other coordinates, weighted by probability of shooting and probability of receiving the ball from those other coordinates, respectively. More specifically, Singh constructs a transition matrix T that captures probabilities of the ball moving from coordinate pair $(x, y) \rightarrow (z, w)$ and a vector s that specifies probabilities of shooting at every coordinate, where the probabilities are derived empirically from Premier League data. He uses a function $g(x, y)$ that represents the known expected goals for every coordinate, and calculates the values ("expected threat") for each coordinate with value iteration using Equation 3.1.

$$\mathbf{xT}_{x,y} = s_{x,y} \cdot g(x, y) + (1 - s_{x,y}) \sum_{x,y} \sum_{z,w} T_{(x,y)\rightarrow(z,w)} \cdot \mathbf{xT}_{z,w} \tag{3.1}$$

Singh's interpretation of expected threat as developed in this way is very similar to how I interpret expected possession value. It explains how much value a particular pass contributes to scoring a goal. This interpretation is the exact same characterization I seek to find in the basketball domain in order to compare hypothetical outcomes. In this vein, I draw on this work for motivation in an analogous domain of sports.

3.2 Extension of Previous Research Project

A previous student in my lab, Santhosh Narayan, also took on the task of predicting shot probabilities. However, working with Professor Hosoi, Narayan used deep learning approaches to see if he could attain better results, different than all the other models I have described so far. My work is actually a direct extension of the preliminary exploration Narayan and Professor Hosoi did previously in this project, as they

set a foundation for what the deep learning methods could look like for quantifying the difficulty of shots. While that work wasn't necessarily rooted in the motivation of decision-making analysis or strategy evaluation, it was another valuable proof of concept and first iteration that I found useful to work off of.

Chapter 4

Data

This project was significantly dependent on the data that was available to us as well as the tools and resources for processing that data. As such, it is useful to understand exactly what the data looked like, where it came from, and how it was utilized. In this chapter, I will give a deeper view at the data we were working with, as well as highlight the powerful resources we had access to for processing the data and how I was able to leverage Google Cloud tools to address data challenges.

4.1 Second Spectrum Markings

The entirety of data I had access to was provided by the San Antonio Spurs. Every NBA team (including the Spurs) is given the same set of basketball data, which teams are then free to use how they want. All aspects of all NBA games played are captured, including possession data, events, defensive matchups, and most importantly tracking data. This data is collected entirely by Second Spectrum, the official tracking provider for the NBA. I had access to all of the Second Spectrum data accumulated from the 2013-2014 NBA season to the 2018-2019 NBA season, amounting to 8,210 games, 1.5 million unique possessions, and nearly 80,000 hours of tracked game-time.

Through their own methods of machine learning and data processing, Second Spectrum has already on their end attached semantics and basketball events to the tracking data they collect. Thus, the data is provided to us separated in specific

"markings". For example, there is a unique table that captures all shots, one for all passes, and one for all possessions, making it extremely easy to work with. While these three tables are merely a fraction of the total data provided to us, these were primarily the ones that I worked with for the purposes of this project, in addition to the raw player and ball tracking data.

As I said before, in this entire dataset there were approximately 1.5 million possessions that were captured, all tagged with a unique Second Spectrum ID, and accounting for a table of size 1.04 GB. For each of these possessions, the identities of the players, context of the game during that possession, and outcome of the possession are all recorded. A detailed breakdown of the table schema can be found in Table A.1 in Appendix A.

From the 2013-2018 seasons, nearly 4.8 million passes were recorded in this dataset, accounting for a table of size 2.23 GB. Attached to each one of these passes is a full meta-description of the context of the pass as well as tracking information about the pass itself. For instance, details about the game environment are captured for each pass, including at what possession in the game the pass occurred and how much time was left in the game and in the shot clock. Also documented is who the passer is, who the intended target is, and their respective locations. This data was very easily manipulable in order to generate features for my pass difficulty model, and was extremely comprehensive in capturing the characteristics of each pass. The table schema for passes can also be found in Appendix A in Table A.2.

The shots dataset provided by Second Spectrum consisted of 1.4 million shots, 900 MB in size. As with the passes table, included with each shot is similar meta-data describing the context of the game at the time of the shot, as well as information about the defenders and other offensive players. Again, the details provided for each shot were extremely comprehensive and allowed me to easily derive features for my difficulty model development. The table schema for shots can also be found in Appendix A in Table A.3.

The last table that I will highlight is the all-important player tracking table. This table was by far the largest and most burdensome to work with, with over 7.1 billion

rows and taking up nearly 1.7 TB of storage. The tracking captured by Second Spectrum is recorded at a 25 Hertz rate, meaning that every second there are 25 frames of measurements. While having this high frame rate by default was good in the sense that we had more data to work with, it was also challenging in the sense that we had more data to process in our computation. A large part of the reason the tracking data table was so large was due to this high sampling rate, which contributed to lots of inefficiencies. The tracking data table captured each player's location at a particular frame, the ball's location, and game context. Thus, at each frame, there are 10 unique rows for the five offensive players and five defensive players, resulting in $10 \times 25 = 250$ rows per second. With almost 80,000 hours of game-time recorded in the entire dataset, it is easy to see how the size of this table ballooned very quickly. The schema for the tracking table can once again be found in Appendix A in Table A.4.

4.2 Google Cloud

The Google Cloud Platform (GCP) was absolutely critical for all of the software development and data operations throughout this project. The most important aspect of GCP was that it was able to store the immense amount of data we had in a readily accessible way through BigQuery. All the data tables were stored in the BigQuery engine, an integrated data warehouse compatible with extremely efficient SQL querying. Having the data exist in this infrastructure allowed for easy access, quick processing, and simple integration.

My code was also all developed on a virtual machine hosted by Google Cloud, through its Compute Engine functionality. By utilizing the GCP connected VM, I was also able to take full advantage of numerous tools to aid in my computing, including the seamless connection to data through BigQuery and cloud-hosted storage through Google Cloud Storage (GCS). For larger files such as the neural network models that I trained, being able to save these to a remote server in GCS was extremely helpful.

As a last example of how I leveraged Google Cloud, I was able to take advantage

of the parallel computing powers it offers through the Dataflow tool. One of the challenges I faced when training my neural networks was the lengthy time it was taking to process thousands of examples, and scanning the massive tracking table numerous times. The Google Cloud Dataflow platform combined with the Apache Beam SDK were extremely valuable in highly parallelizing these computations and cutting down the runtime of training significantly.

Chapter 5

Difficulty Models

The foundation of EPV is grounded in the pass difficulty and shot difficulty models that I developed. These machine learning models are used to characterize the difficulty or quality of a particular pass or shot, predicting a probability of completing the pass or converting the shot. In this chapter, I will describe the process that I used to build these models, showing the features that were used in the models, and discussing their accuracies.

5.1 Pass Difficulty Model

I used a couple different approaches to develop my pass difficulty model in an attempt to achieve the best accuracy I could. While two distinct classes of models were explored, the features that I used were the same for both. The feature engineering process was an extremely iterative one, with the challenging task of incrementally adding more explanatory components to a feature set that would qualitatively characterize a given pass well. Also importantly of note, these features were all to be generated from the tracking data only. A full description of the finalized features and how they are computed from the tracking data is contained in the following subsection.

For both of these models, I used all 786,208 passes from the 2018-2019 NBA season as data points, splitting them into training and evaluation sets 70% to 30%

respectively. I followed the standard machine learning workflow of training models on the training set, testing for accuracy on the evaluation set, and iterating on features and hyperparameters as necessary. The machine learning algorithms are supervised, with the *complete* column of the pass data serving as the labels.

5.1.1 Pass Features

The simplest characteristics of a hypothetical pass that can be extracted from the tracking data correspond to relative distances between entities involved in the pass. This includes the distance from the passer to the intended target, and the distances from both the passer and the target to the basket.

We can also leverage the defender location data to characterize how much defensive pressure there was on the pass. To do this, I incorporate defensive configuration features as well, including the closest defender distance to the passer, closest defender distance to the target, and closest defender distance to the trajectory of the hypothetical pass, which can be thought of as the most "obstructive" defender relative to the pass.

The relative angles are also good features to incorporate, as they give a better sense of direction with respect to the basket and pass trajectory. I include the angle created from the line segment from passer to basket to target, the angles of the closest defenders to both the passer and target with respect to the pass trajectory, and the angle created from the line segment from passer to target to the most "obstructive" defender. In conjunction with distance features, the angle features give a lot of information about where defenders are located, and as such are great predictors for how likely it might be for the defense to disrupt a pass. These geometric features for a hypothetical pass can be visualized in Figure 5-1.

Lastly, I include a couple features that are more based in game context. I include a binary feature that flags whether or not the passer and target are both located in the backcourt. Passes that occur in the backcourt are usually undefended, and so should be considered outliers by the model. By incorporating this explicitly as a feature, the trained model will be able to pick this up very easily. When the shot

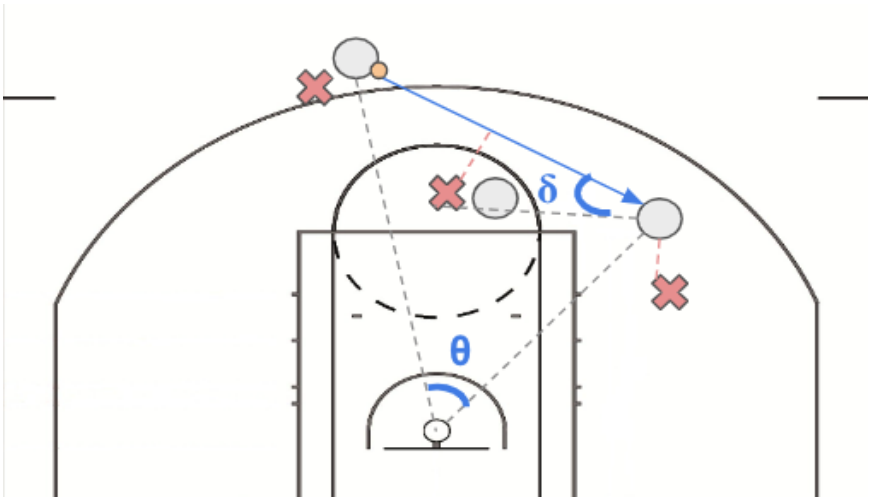

Figure 5-1: Diagram of geometric pass features. Blue arrow indicates trajectory of hypothetical pass. Red X's indicate defenders. Middle defender is most "obstructive" defender, with smallest perpendicular distance to the trajectory of the pass. θ is the angle of the pass with respect to the basket, and δ is the angle of the most obstructive defender.

clock is low, meaning that there is little time remaining in the possession, there is often more pressure on the offense to execute a shot, which in turn makes passes more difficult to complete. To account for this phenomenon, I also include the shot clock as its own feature as well.

A comprehensive list of all pass features used and their descriptions is found in Table 5.1.

5.1.2 Logistic Regression Approach

The desire to predict probabilities from a set of features is a natural problem for machine learning, and is a particularly suitable task for the classical logistic regression model. As such, I used a preliminary logistic regression approach due to its ease of implementation and mathematical interpretability. A logistic regression is an extension of a standard linear regression fitting, but passing the predicted value through a sigmoid function to get a probability output. Instead of fitting $y = \theta^T x$ for

Feature name	Description
dist	Distance of the pass (from passer to target)
passer_basket_dist	Distance from passer to basket
ball_end_basket_dist	Distance from target to basket
basket_angle	Angle of passer → basket → target
closest_def_dist_passer	Closest defender distance to passer
closest_def_dist_ball_end	Closest defender distance to target
closest_def_trajectory	Closest perpendicular defender distance to pass trajectory
closest_def_angle_passer	Closest defender angle to passer w.r.t. pass trajectory
closest_def_angle_ball_end	Closest defender angle to target w.r.t. pass trajectory
closest_def_trajectory_angle	Angle of passer → target → most obstructive defender
backcourt	Whether the pass was made entirely in the backcourt
shot_clock	Time remaining on the shot clock

Table 5.1: Full list of pass features.

parameters θ as per standard linear regression, we instead are fitting:

$$y = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{5.1}$$

where $\sigma(\cdot)$ is the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$. The sigmoid function effectively collapses the real-valued $\theta^T x$ to a (0, 1] range, interpretable as a probability.

Using the features described in Section 5.1.1, I was able to construct the feature vectors x taking special care to standardize each feature such that they were zero-mean and variance of the same order. For a logistic regression, it was crucial to standardize the features in this way to ensure that no one feature was dominating the loss, consequently rendering the other features as non-predictors. This comes from the property that a logistic regression is an extension of a *linear* model.

One more thing I considered when feeding in this training data to the logistic regression model is that the data is biased towards positive examples. More specifically, there is a disproportionately high number of completed passes in the dataset due to players opting not to make difficult passes. To correct for this selection bias, I weight positive (completed passes) and negative (non-completed passes) differently, weighted inversely by the number of positive/negative examples there are in the training set. Letting N_p be the number of positive examples and N_n be the number of negative

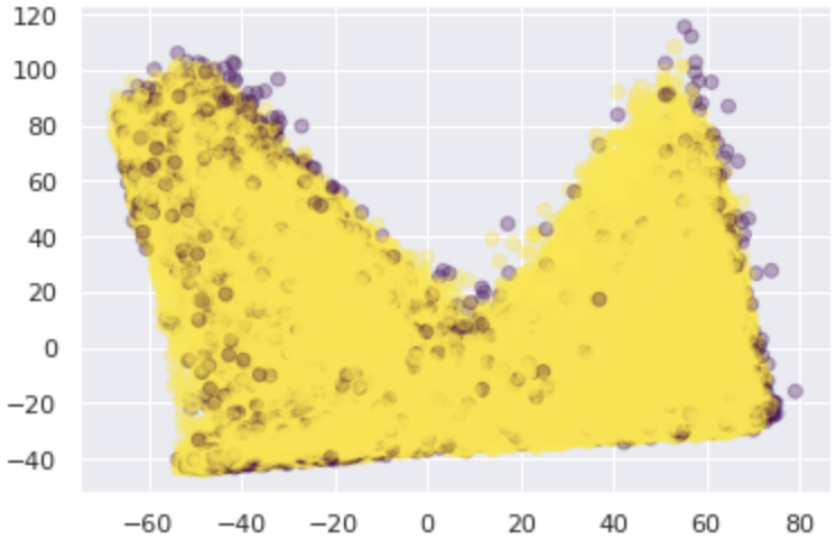

Figure 5-2: 2-dimensional PCA for logistic regression passes. Yellow points are positive examples (completed passes), purple points are negative examples (non-completed passes).

examples, the weights are as follows:

$$w_{pos} = \frac{1}{N_p}, \qquad w_{neg} = \frac{1}{N_n} \tag{5.2}$$

With this weighting scheme, negative examples are weighted much heavier due to there being so many more positive examples in the training dataset. Thus, the logistic regression model will not learn this selection bias and instead be able to balance out this discrepancy in making objective predictions.

A useful visualization we can look at to ensure that the model has a good representation for these passes in order to learn is by plotting out the positive versus negative examples in a 2-D plane. To do this, I ran a 2-dimensional principal component analysis (PCA) to reduce the logistic regression input features into the two most significant components. I then plotted all the passes in both the training dataset and evaluation dataset colored by whether or not the pass was actually completed, shown in Figure 5-2.

What we see in this plot is that negative examples mostly lie at the boundary of the general shape depicted. This is exactly what we want out of a representation for our

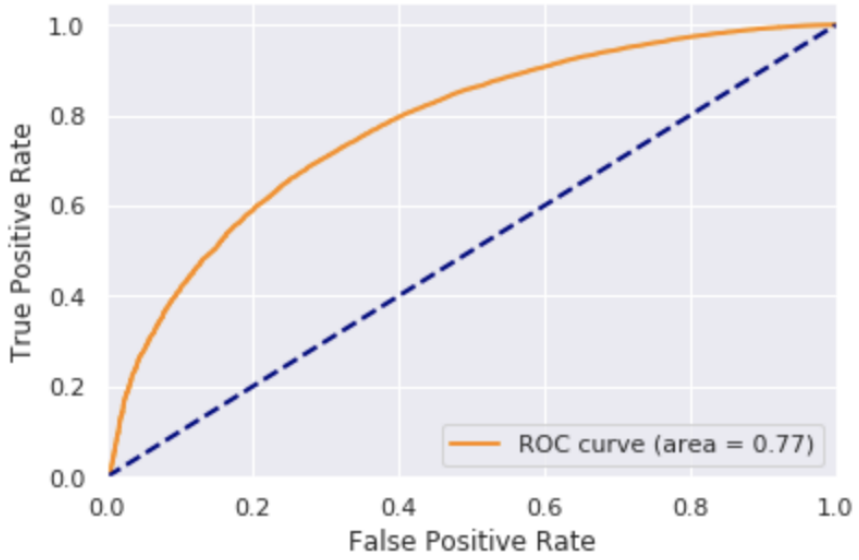

Figure 5-3: Receiving operating characteristic curve for trained logistic regression model.

passes, as we want the data to be linearly separable in high dimensions for the logistic regression model to perform well. We can imagine the learned hyperplane projected into two dimensions being exactly the outline of this shape, such that everything inside is classified as positive, and everything outside is classified as negative. This plot reinforces our confidence that we have selected a good feature set and data representation for our passes to be correctly classified.

After propertly configuring the feature vectors and model training scheme, the trained logistic regression model achieves an evaluation ROC-AUC score of 0.774, which I use as the measure of accuracy for these probabilistic classification models [8]. The receiving operating characteristic is depicted in Figure 5-3.

Already we can see that the model is performing significantly better than chance, which would have an ROC-AUC score of 0.5 (depicted in Figure 5-3 as the dotted blue line). To look deeper into what the model is actually learning, I analyzed at the learned weights in the model corresponding to how important each feature is in predicting a probability of completion. These relative feature importances for the logistic regression model are shown in Figure 5-4.

Based on this chart, the logistic regression uses the backcourt and distance to

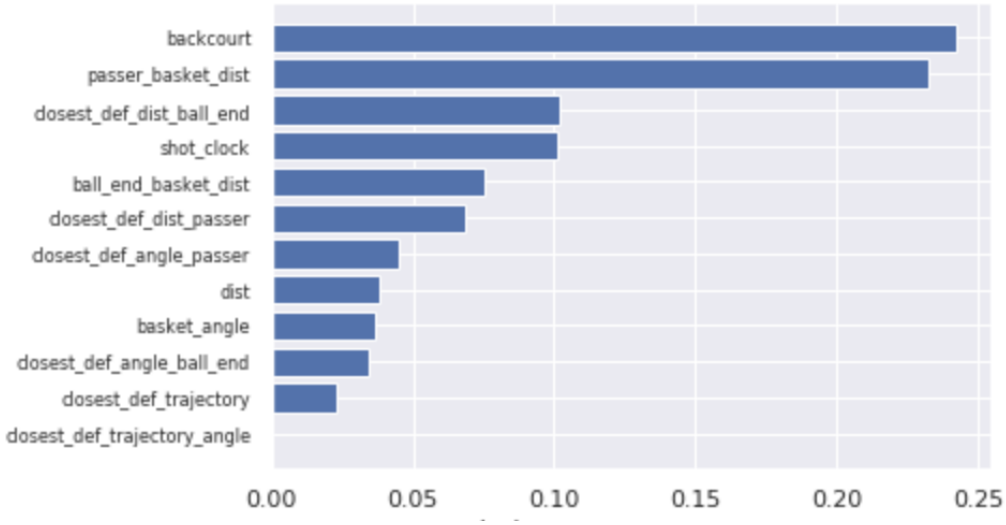

Figure 5-4: Relative feature importances for logistic regression pass difficulty model.

basket features the most when predicting probabilities of completing passes. It also takes significantly into account the defender distances to each of the involved offensive players. This is a valuable sanity check for us to see that the model is indeed using intuitively helpful information to make its predictions. Furthermore, the ability for us to visualize the importance of each feature is a massive benefit to using a simple linear model such as a logistic regression, giving us a lot of interpretability and uncovering any potential black box. As we will see with a deep learning approach, we would have to sacrifice this interpretability in order to achieve better results.

5.1.3 Neural Network Approach

Despite the logistic regression working very well, a research question that persists is whether deep learning approaches could improve accuracy for these classification models. To address this, I also constructed a simple neural network to do the same job: given an input feature vector characterizing a pass, output the probability of completion.

The neural network I built was very shallow and simple, and its basic architecture is shown in Figure 5-5. I used only one fully-connected hidden layer with 128 units, each activated with a ReLU function. As a form of regularization, I also included a

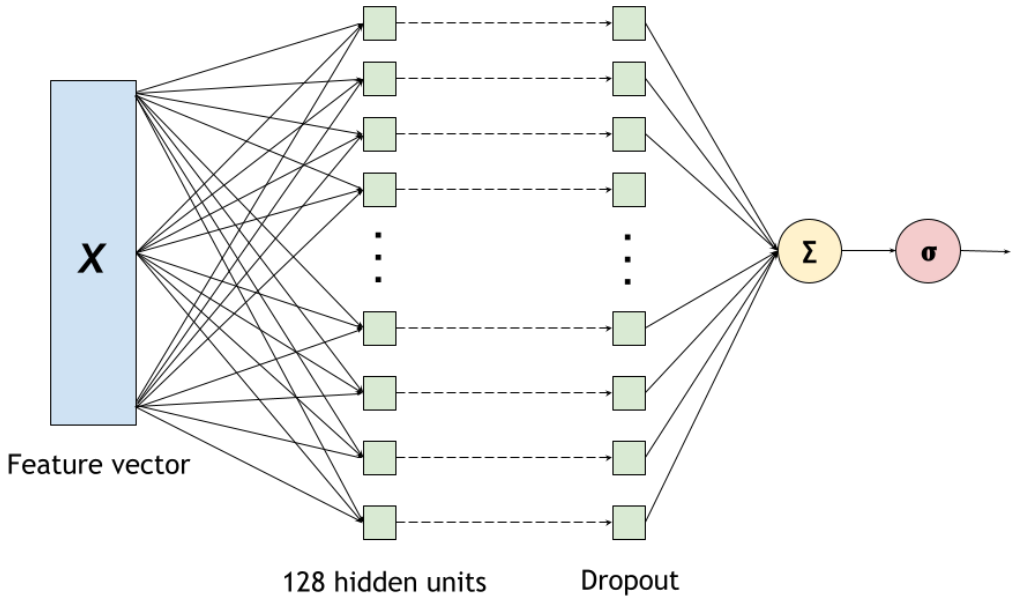

Figure 5-5: Pass difficulty neural network architecture.

dropout layer with dropout parameter 0.2. I included this layer to prevent overfitting, and effectively zeroed out certain random units during training such that the network could not rely too heavily on any particular unit. Finally, I have an output layer that is fully-connected, and uses a sigmoid as a final output activation function. The sigmoid as the final activation converts the prediction into a probability, as desired.

Using the exact same training and evaluation dataset and the same feature set, I trained the model with a binary cross-entropy objective function, such that the neural network is minimizing:

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \Big(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big) \tag{5.3}$$

where y_i is the label (0 or 1) of pass sample i and $\hat{y}_i$ is the predicted probability of completion. With such a simple model, the training was very quick, requiring no more than 10 epochs to converge. The resulting trained network had an evaluation ROC-AUC score of 0.885, significantly higher than the logistic regression model. The ROC curve is shown for reference in Figure 5-6.

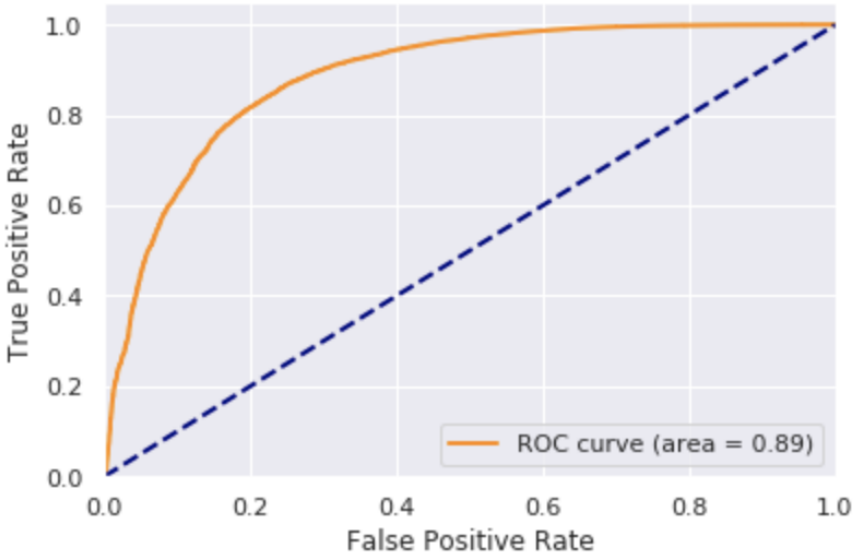

Figure 5-6: Receiving operating characteristic curve for trained pass difficulty neural network.

5.1.4 Comparison of Pass Difficulty Models

Despite the neural network far outperforming the logistic regression model, there is something to be said about the simplicity and interpretability of the logistic regression. Due to the fully-connected nature of the neural network, it is extremely difficult to extract the importances of each feature individually the way we did with the logistic regression model. As such, it's not easy to understand exactly what the neural network has learned semantically, despite it performing so well. This is an open research question within the machine learning community, as deep learning lacks the interpretability that simple linear models offer. With that being said, for the purposes of this project, accuracy is the most important thing for these difficulty models, so I opted to move forward with the neural network model.

As a heuristic method of evaluation, we can visualize the pass difficulty model's performance by looking at the pass probability predictions throughout a real possession. An example possession with pass completion probabilities as calculated from the model is depicted in Figure 5-7.

We can see that the probability predictions align with what we would expect qualitatively. In particular, when defenders are in the way of long range passes,

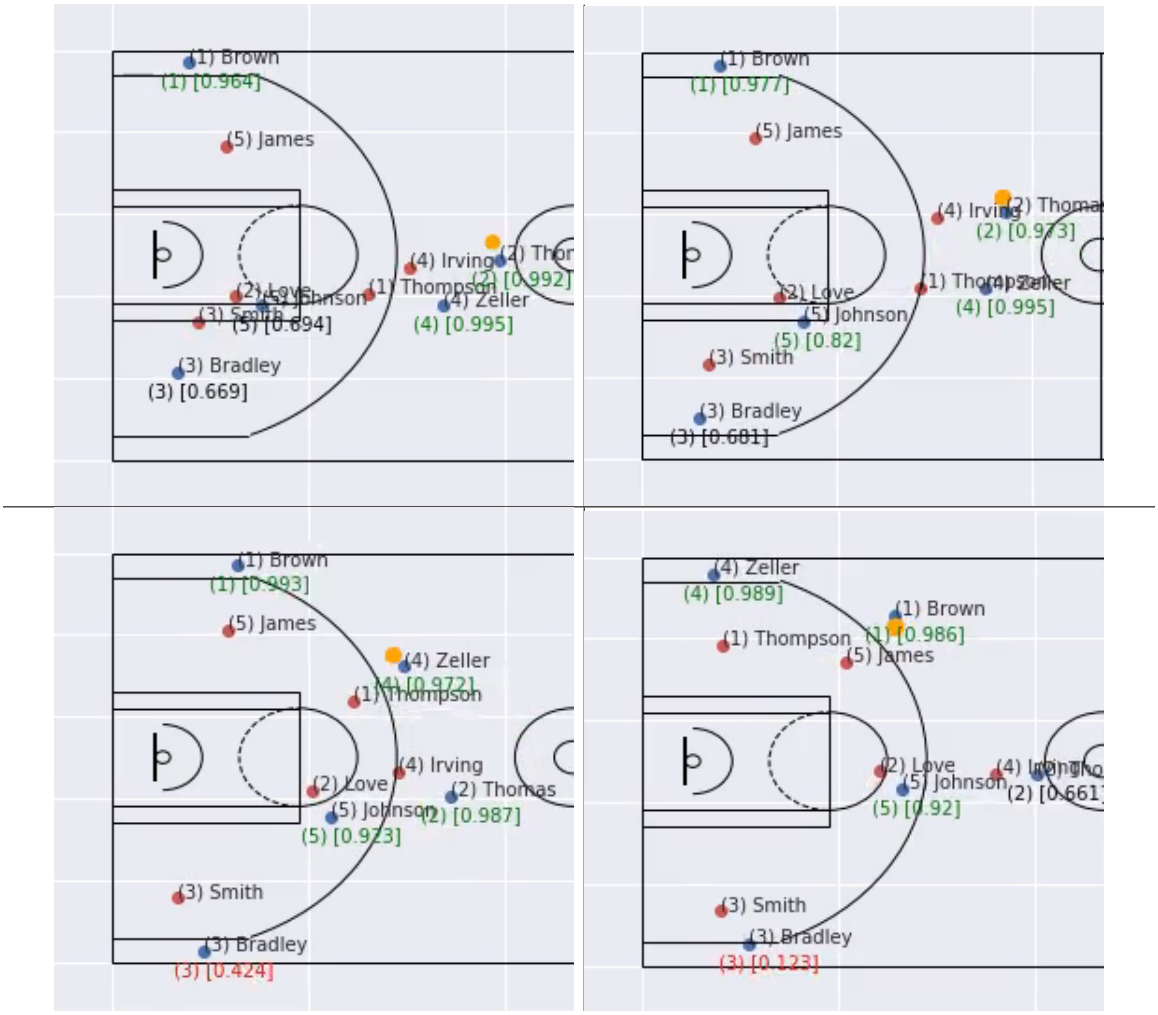

Figure 5-7: Pass probabilities in a possession. Blue dots are offensive players, red dots are defensive players, orange dot is the ball. *Left to right, top to bottom.* 1) Early in the possession, high probabilities for player next to ballhandler and open pass lane to the corner. 2) As defenders spread out in the post, pass completion probability to player 5 increases. 3) Play moves away from player 3, so pass completion probability starts to decrease. 4) Ball is far away from player 3 with many defenders in the way, so pass completion probability is minimal.

we see significant drops in probability. Similarly, when players come closer to the ballhandler, we see the probabilities increase. Observing these patterns is a good indication to us that the pass difficulty model is working as intended, and is accurate beyond just the ROC-AUC metrics.

5.2 Shot Difficulty Model

Developing the shot difficulty model was unsurprisingly a little more challenging and nuanced than the pass difficulty model. Inherently, there is a lot more variability in shot making as compared to pass completion, as making a basket is a much more refined task than completing a pass. As such, being able to create an accurate model for predicting shot conversion probabilities becomes more of a challenge.

In Chapter 3, I described numerous approaches already taken to try and tackle this problem. Seeing how much of an improvement using a neural network provided in the pass difficulty model, and in light of exploring deep learning approaches for quantifying shot difficulty, I directly attempted to develop a neural network for shot difficulty as well, without going through a linear model first. As mentioned in Section 3.2, this idea was first analyzed by Santhosh Narayan, a previous student working with Professor Hosoi. In working on this model, I included some of the features originally proposed, while incorporating new features to characterize the movement of the players. In this section, I will describe details of the model including features used and architecture, while also discussing some challenges that arose in training which weren't prevalent before when working with the pass difficulty model.

5.2.1 Shot Features

Much like with the pass features, a lot of the characteristics of a particular shot can be captured from the geometric configuration of players on the court. In particular, we really only need to consider the shooter and close defenders in their general vicinity.

We can again use distance based features to help gauge how difficult a shot might be. We definitely expect an obvious feature like how far the shooter is away from

the basket to negatively correlate with the likelihood of making the shot. The closest defender distance is also an important feature, which when combined with the number of close defenders, represents how much pressure the shooter is under when they take their shot. I consider close defenders to be any defender within a 4-foot radius of the shooter, as that is a small enough distance to contest the shot in theory, and potentially disrupt the play.

Including relevant angles are also useful for painting the whole geometric picture. To fully characterize the effect that the closest defender has on the shooter, it is wise to also include the angle of the closest defender with respect to the hypothetical shot trajectory path. This in essence describes how "obstructive" this defender is of the shot, which can significantly impact the difficulty of a particular shot. The angle of the shot with respect to court orientation can also be a significant predictor of how hard a shot might be, as it determines the shooters ability to use the backboard. The shot angle is defined as the angle created by the line segment from shooter to basket to center court, which characterizes how off center the shot is. It is typically understood that "straighaway" shots (where the shot angle is 0) are easier to hit, which is why this feature could be important.

A component of shots which make them more difficult to predict when compared to passes is the effect of the players' movements leading up to the shot. If the shooter is having to suddenly decelerate before taking the shot, or if their velocity is very high while taking the shot, it becomes much more difficult to make that shot compared to the same shot if they were standing still. Similarly, if a defender is jumping very quickly at the shooter or is accelerating towards them as they shoot, that becomes much more disruptive than if they were just standing still. To capture this intuition, I also deemed it necessary to include relevant velocity and acceleration information.

Recall that we only have access to the raw positional data, which doesn't explicitly provide velocities and accelerations. To obtain these values, we need to compute them through differentiation of the positional data with respect to time. Given that we have a discrete set of measurements of positions (25 measurements per second), to differentiate this data we need to smooth and interpolate the points such that

we can treat it as a differentiable function. My method of doing this is with a Savitsky-Golay filter, a technique in signal processing to smooth a set of potentially noisy data points [9]. The idea behind the Savitsky-Golay filter, which takes in two parameters window size w and order o, is very simple and intuitive. The operation consists of sliding a window of size w through all the data points, fitting a polynomial of degree o to each window. The points are then re-positioned to fall within the polynomial fits within each window, effectively removing erratic jumps and noise. Applying the Savitsky-Golay filter to the positional data, then applying a standard Cubic Spline to interpolate the points, gives us a smooth differentiable function to use now for calculating velocity. Taking the derivative of this function gives us a velocity function, for which if we apply the same technique of smoothing, allows us to take the derivative once more to obtain an acceleration function.

For the purposes of quantifying shot difficulty, I use the past second as the time frame to compute these derivatives for, giving 25 positional data points to work off of for each player. The Savitsky-Golay filter I use has a window size of 5 and polynomial order of degree 3, which is the simplest mathematical formulation for this kind of smoothing. Using this technique, I can obtain velocity and acceleration functions of the player within the last one second before the shot, which allows me to extract instantaneous velocity and acceleration values at the time of the shot. Being able to incorporate these movement features was extremely beneficial to the performance of the model.

All the geometric features that I used can be visualized in Figure 5-8. The only other feature that I used that wasn't geometric was the shot clock feature, which I incorporated to help characterize how much pressure the shooter was under, similar to the pass difficulty model. This was the only game context based feature I felt was important for the predictability of making a certain shot. A comprehensive list of the shot features used for this model is detailed in Table 5.2.

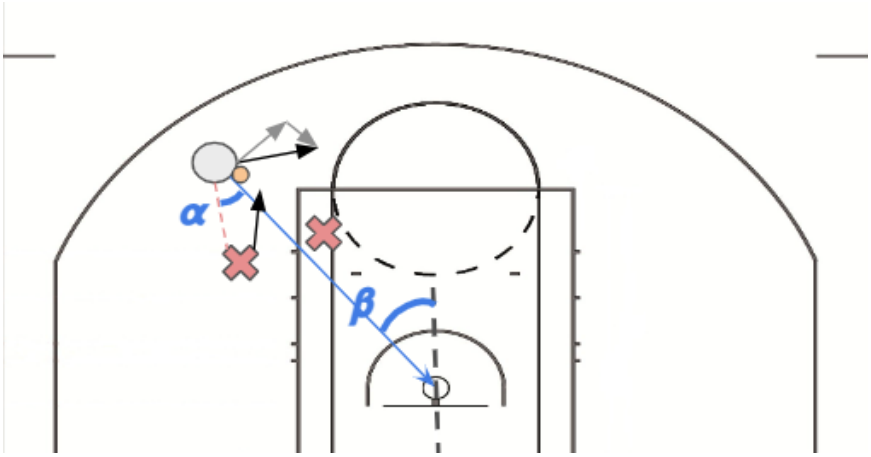

Figure 5-8: Diagram of geometric shot features. Blue arrow indicates trajectory of hypothetical shot. Black arrows represent movement vectors for players, decomposed into gray parallel and perpendicular components. In this example, the shooter is fading to their left while shooting, while the defender is moving towards them to contest. α is the angle of the closest defender with respect to the shot trajectory, and β is the angle of the shot with respect to center court.

Feature name	**Description**
dist	Distance of the shot (from shooter to basket)
x	x coordinate of shooter
y	y coordinate of shooter
shot_angle	Angle of the shot w.r.t. court center
closest_def_dist	Closest defender distance to shooter
closest_def_angle	Closest defender angle to shooter w.r.t. shot trajectory
num_close_defs	Number of defenders within 4 feet of shooter
shot_clock	Time remaining on the shot clock
shooter_par_vel	Parallel velocity of shooter w.r.t. shot trajectory
shooter_perp_vel	Perpendicular velocity of shooter w.r.t. shot trajectory
closest_def_par_vel	Parallel velocity of closest defender w.r.t. shot trajectory
shooter_perp_vel	Perpendicular velocity of closest defender w.r.t. shot trajectory
shooter_par_acc	Parallel acceleration of shooter w.r.t. shot trajectory
shooter_perp_acc	Perpendicular acceleration of shooter w.r.t. shot trajectory
closest_def_par_acc	Parallel acceleration of closest defender w.r.t. shot trajectory
shooter_perp_acc	Perpendicular acceleration of closest defender w.r.t. shot trajectory

Table 5.2: Full list of shot features.

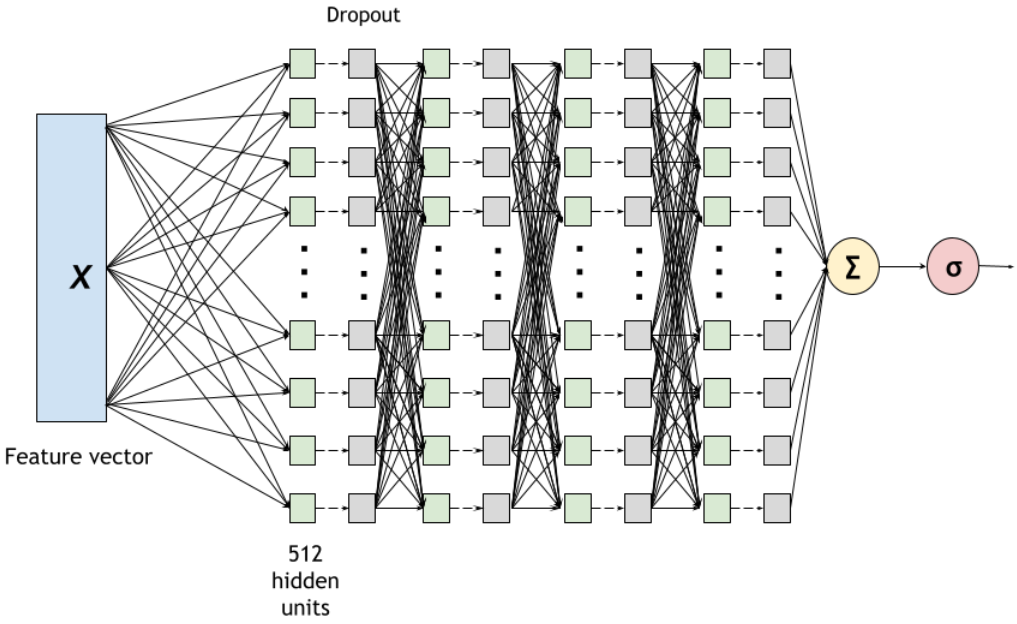

Figure 5-9: Shot difficulty neural network architecture.

5.2.2 Model Architecture

I wanted to replicate as much as possible the architecture of the pass difficulty model given how successful it was at predicting pass completion probabilities. However, I found that the simple one-layer model was nowhere near rich enough to get a good accuracy for the much harder shot difficulty problem.

I used a deeper neural network with 4 hidden fully-connected layers, each with now 512 units. All of these units are activated with a leaky ReLU, which Santhosh found to be slightly more performant than the standard ReLU activation function. Again, I incorporate dropout layers with dropout parameter 0.2 after each fully-connected layer to prevent overfitting. Lastly, I again use a sigmoid final activation to convert the real-valued outputs into probability predictions. The model architecture is depicted in Figure 5-9.

5.2.3 Training and Evaluation

The added computation of velocity and acceleration features from the tracking data created a significant bottleneck for the overall training process for this model. This

extra component in the feature generation step would theoretically run for over 100 hours on the virtual machine on just a small test dataset of 10,000 shots, needing to repeatedly query sets of 25 data points from the 1.7 TB tracking table and apply smoothing. Needless to say, this was not a sustainable solution with the current computing infrastructure that I had if I wanted to scale the training up to a larger dataset.

Aided by the resources in Google Cloud, I was able to parallelize the computations that went into the feature generation. I leveraged Google Dataflow and Apache Beam to build a pipeline object that would spin up multiple virtual machines to perform these calculations. I grouped the tracking and shots dataset into months, and passed each month worth of shots as an input for each machines. By doing so, I reduced the computation needed to be done by each worker, while simultaneously shrinking the tracking table to a fraction of the original 1.7 TB size to be repeatedly queried. Using Dataflow to do this, I was able to speed up the runtime nearly 300 times, such that I was able to get a full set of features for 10,000 shots in a little over 20 minutes. This was immensely beneficial to the development of my model, as it removed the bottleneck of not being able to use more data points without sacrificing efficiency.

In the end, I was able to generate features and train on the entire shots dataset, consisting of 1.4 million shots from 2013-2019. This time, using a 80% to 20% train-validation split, I trained on over 1 million shots, again using the binary cross-entropy objective function in Equation 5.3, and achieved the best ROC-AUC score of about 0.62, comparable to other shot difficulty models that have been developed. The resulting ROC curve is shown in Figure 5-10.

Clearly, in comparison to the pass difficulty model, the shot difficulty model falls well short in terms of accuracy. This does match our intuition that shot success is inherently harder to predict, but we can ask the question: if we had more data, could our accuracy get better? To show that this isn't the case, I plotted the accuracy of the model and the mean squared error of the model as a function of the number of training examples the model is seeing, which is shown in Figure 5-11.

Evidently, we see that beyond ten to fifty thousand shots, we don't really observe

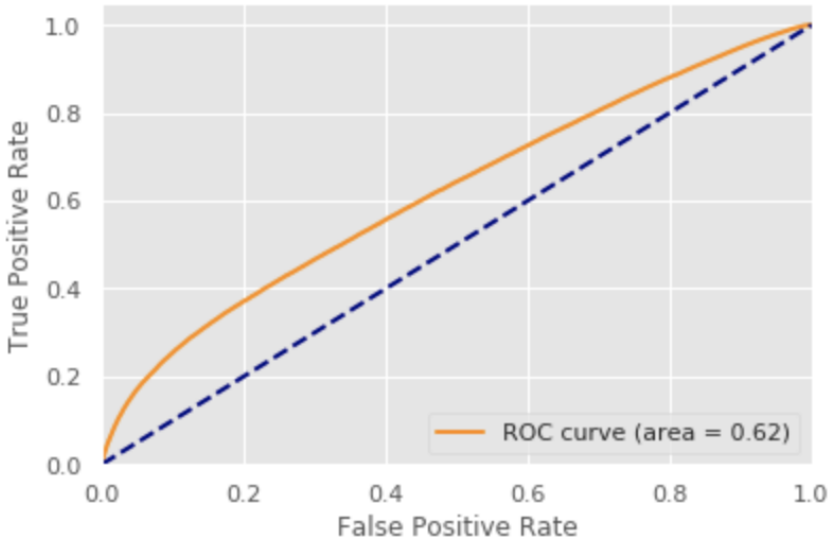

Figure 5-10: Receiving operating characteristic curve for trained shot difficulty neural network.

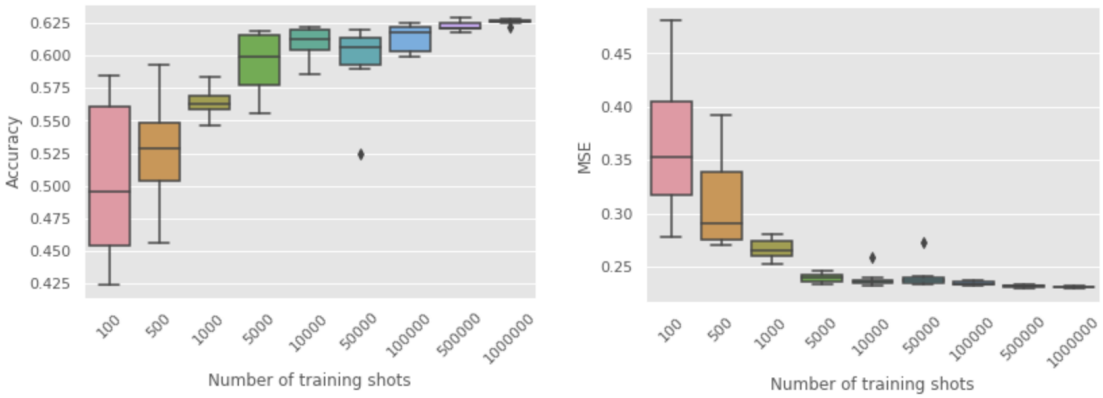

Figure 5-11: Shot difficulty neural network performance as the number of shots used during training increases.

a significant increase in accuracy. With both the accuracy and the mean squared error, we see an asymptotic effect in achievable performance, indicating that adding significantly more, and even infinite training data, won't help the model achieve much higher accuracy. Seeing how consistent the performance of this model is with the models described in Section 3.1.1, it stands to reason that this might be a theoretical bound when it comes to quantifying shot probabilities. Even with the neural network approach that I used, which in theory has much more expressive power than simple linear models and can capture more complex and dynamic relationships between features and outcome, I wasn't able to get any improvement. While this is an ongoing research question specifically for quantifying shot difficulty, perhaps it is just too random of a process by nature to accurately model consistently.

With that being said, I believe there are still improvements that could be made to my model that could help the performance, even if minimally, for which I was not able to implement. I will discuss some limitations of my current approach and areas where more information can be incorporated later in Chapter 7.

Chapter 6

Expected Possession Value

Equipped with methods for quantifying shot difficulty and pass difficulty as probabilities of success, the computation of expected possession value becomes a very simple integration. Recall that the goal of expected possession value as a metric is to capture the hypothetical scenarios created by the ballhandler deciding to either shoot or pass to one of their teammates to shoot, quantifying how many points the team can expect from any one those decisions. Calculating this expectation is nothing more than weighting the probability of success with the corresponding point value, an intuitive idea that I formalize now.

6.1 EPV Computation

As defined in the project overview in Chapter 2, EPV can be described as the expected number of points to be scored by each offensive player if they receive the ball and shoot at that particular instance in time, meaning that each offensive player has an EPV value at every moment during a possession. Leveraging both the shot and pass difficulty models, we now have the ability to characterize how likely it would be to first get the ball to each offensive player, then subsequently how likely it would be for them to score. We can run this computation for each frame in the possession.

To see this explicitly, consider a time t during the possession. From the tracking data, we are able to extract who the ballhandler is, which we will denote as $bh^{(t)}$, all

offensive players locations, all defensive players locations, the ball location, and other game context details. This is all the input we need to feed into our difficulty models to obtain shot conversion and pass completion probabilities. If we want to calculate the EPV for a specific offensive player $p_i^{(t)}$, we need to consider the situation where the ballhandler passes to them, and they immediately shoot. Thus, we can run the pass difficulty model using the location of $p_i^{(t)}$ as the target location to output the probability of the ballhandler completing a pass to player i. We can also run the shot difficulty model using the location of $p_i^{(t)}$ as the shooter location to output the probability of them scoring if they had the ball and shot. Note that multiplying these probabilities results in the probability that both events occur, meaning that player i successfully receives the ball, shoots and scores at time t. Finally, we just need to take this probability of success and weight by the point value of the hypothetical shot, determined by whether or not $p_i^{(t)}$ is standing behind the three point line.

Formally, denoting the ballhandler at time t as $bh^{(t)}$, player i at time t as $p_i^{(t)}$, a completed pass as the arrow $\rightarrow$, and the point value of the shot as a variable v which can be either 2 or 3, EPV is calculated according to the equation:

$$EPV(t, p_i) = \mathbb{P}(bh^{(t)} \rightarrow p_i^{(t)}) \cdot \mathbb{P}(p_i^{(t)} \text{ scores}) \cdot v \tag{6.1}$$

where $\mathbb{P}(bh^{(t)} \rightarrow p_i^{(t)})$ is an output of the pass difficulty model, and $\mathbb{P}(p_i^{(t)}$ scores) is an output of the shot difficulty model.

For the purposes of computational efficiency, EPV by default is calculated at 5 frames per second, such that at each second there are 5 frames $\times$ 5 players $=$ 25 EPV values output. I found that computing EPV at 5 frames per second resulted in qualitatively the same average values as computing at 25 frames per second, with a 5 times speedup in runtime.

6.2 EPV Visualization

Since EPV can be computed for each player at every frame during the possession, it is extremely easy to convert into a visualization plot such that we can see how EPV values evolve over a possession. Shown below in Figure 6-1 is an example of what an EPV visualization for a possession looks like.

There are a few things to dissect with what the visualization is showing, as it is split into two distinct components. For each frame, on top is a dot diagram representing the locations of the offensive and defensive players, with offensive players denoted as circles and defenders denoted as X's, as well as the ball. The EPV of each offensive player is also incorporated into this dot diagram. The offensive player's circle is not only color-coded so it is easy to match color to player, but is also scaled in terms of size to match their EPV at that point in time. More specifically, a large circle would indicate that player to have a high EPV, while a small circle would indicate that player to have a low EPV.

For each frame, the bottom plot is the actual depiction of EPV values for each player, with the x-axis being time (represented as time left on the shot clock), and the y-axis being EPV value. There are five lines in the plot corresponding to each offensive player, color-coded to match the corresponding dot diagram above. In addition, an orange circle is also plotted at each timestep indicating which player has the ball. The lines are filled in as we progress through the possession, and are aligned temporally with the positional dot diagram. The black vertical line seen in the last frame indicates when a shot was taken, at which point the EPV computation stops.

Lastly, a horizontal red line is plotted at EPV value 1.5, which was arbitrarily chosen to indicate a "good shot". Given that teams average just about 100 offensive possessions per game, having a target of achieving a shot with 1.5 EPV every possession would result in 150 points for a game, which is slightly above what NBA teams realistically achieve right now. By plotting this red line here, it serves as a reference for what would be great opportunities in a possession. The selection of 1.5 as a threshold will be expanded on in Chapter 8 where we discuss how opportunities

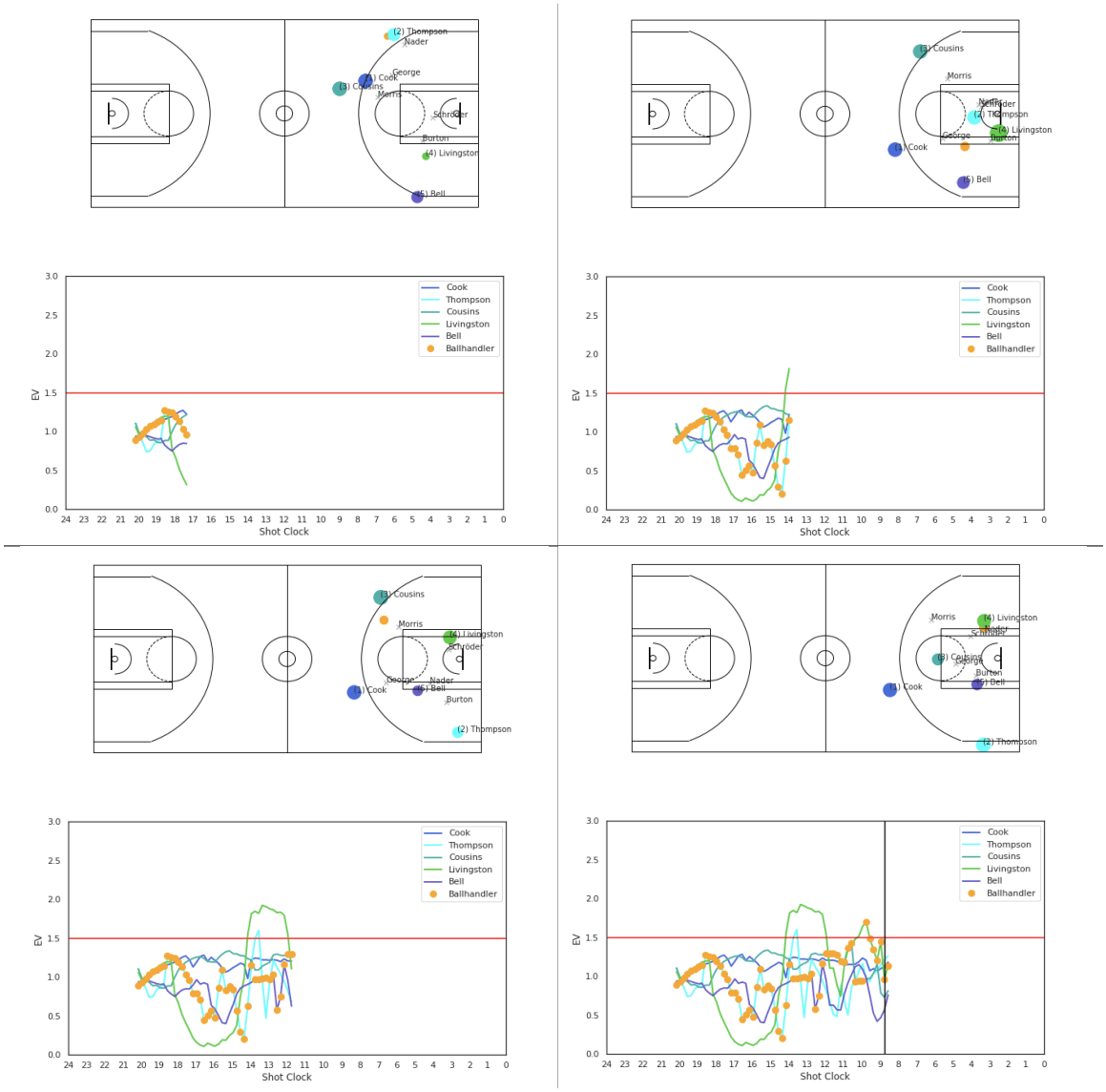

Figure 6-1: EPV visualization plots. Evolution of possession is read from left to right, top to bottom. Shown here are four example still frames taken from the animated visualization.

are characterized.

Evidently, this visualization tool is extremely valuable for our analysis with EPV. Seeing anomalies and spikes in the EPV evolution plots start to motivate questions we may ask about what caused EPV values to go high or low. By observing where the ball was, we can also start to see which opportunities were taken advantage of, and which ones were missed. Viewing a possession through this lens is something that NBA teams haven't been able to do up to this point, and can paint a much clearer story when it comes to evaluating decision-making. Through these visualizations, I was able to explore some anomalies in order to deduce decision-making ability, specifically motivated by the opportunities that are either taken or missed. This interesting analysis is the primary application of EPV that I have examined so far, which will be discussed in Chapter 8.

Before doing so, I think it is crucial to discuss the validity of EPV as it is calculated, while acknowledging its limitations. I will spend the next chapter doing this evaluation.

Chapter 7

Evaluation

7.1 Evaluating EPV Accuracy

While we have already shown evaluation metrics of the pass and shot difficulty model as independent machine learning models, it is not so straightforward to do so with EPV as a whole. When developing the pass and shot difficulty models, we had access to labels in the form of whether a pass was completed or not, and whether a shot went in or not. However, when we integrate it all into an EPV metric, there is no such ground truth that we can test against, which makes the evaluation process for EPV a little more convoluted.

Since EPV can be interpreted as points expected to be obtained from a possession, one strategy we can use is to aggregate computed EPV values from a large sample size of shots, comparing the expected total points and the actual total points scored from that collection of shots. Naturally, doing this on a game-by-game basis is the most intuitive grouping of shots, for which we know how many points are scored. We can essentially use points scored in games as our ground truth data, and make predictions for the points scored in games with EPV, ultimately comparing the two to judge accuracy of EPV. If the expected total points predicted from EPV matches the actual total points scored, then we can more confidently assert the usefulness of the metric, rather than just rely on the eye test to claim its validity.

In the testing that I ran, I computed EPV values for all shots that occurred in 42

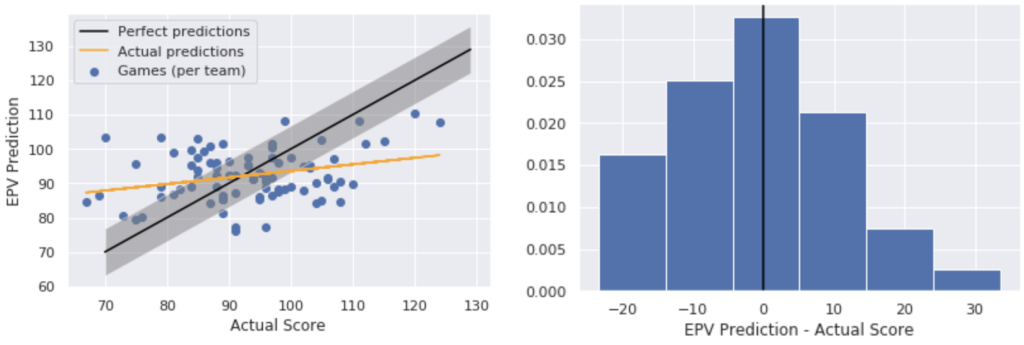

Figure 7-1: Comparison of EPV predicted game scores and actual game scores. *Left.* Scatter plot showing each game's predicted score vs. actual score, separated by team. Black line represents perfect predictions, meaning that the predicted EPV score always equals the actual score, with the gray area being one standard deviation in each direction. Orange line represents actual line of predictions we get from the data. *Right.* Histogram plotting (EPV Prediction - Actual Score) on a per-game basis. The centering around 0 means that on average EPV predicts the final score correctly.

randomly selected games in January 2019. Importantly, I only included the last shot per possession and didn't include missed shots which resulted in a foul. I only include the last shot per possession because that is the only shot from that possession that has the possibility of being reflected in the final score of the game. If a team shoots more than once in a possession, then it is a given that they missed their first shots in the possession, which won't affect the final score of the game. If we want to use the final score as a comparison with the aggregate EPV predictions, we can only include shots that are accounted for in the final score. Similarly, a missed shot in which the shooter gets fouled isn't a reflection of the expected points from that shot. Because it is known to us that the shot doesn't go in, we also ignore these shots as they don't affect the final score.

I aggregated the computed EPV values by game and by team, such that I obtained predictions of the final score, discounting free throws, for both teams in each of the 42 games. The comparison of these predicted final scores and the actual final scores is shown in Figure 7-1.

There are two important takeaways from these figures with respect to how we can evaluate EPV. The first is that on average, it works very well when it comes

to predicting final score. We can see this through the histogram plotted on the right depicting the difference in EPV prediction and actual score. The fact that the histogram is centered around 0 means that on average, the EPV prediction and actual score are very close, indicating that the EPV predictions for shots generally speaking are pretty accurate. However, when we look at the left scatter plot, we see that there's another important insight we can draw about EPV. This plot simply shows for each team in a game what the actual points scored was compared to the predicted EPV point total. If the EPV predictions were 100% accurate, meaning they always predicted the correct score total, we would observe the points falling on the black line, which depicts `actual score` = `EPV prediction`.

What we actually observe is shown by the orange line. In particular, the EPV predictions tend to cluster around the mean, tending to stay within the range 80-110 points while the actual score varies from 65-125. As such, we can say that although EPV generally speaking is very accurate, it does exhibit a bias towards the mean. One possible explanation why this could be the case is simply the efficiency of shooting for a particular team. We can reasonably expect the quality of shots taken and the quantity of shots taken not to vary much game-to-game, leading to EPV values that are generally close together. However, teams can generally be better or worse shooters than league average, and consistently have games where they shoot above or below expectation throughout, such that the actual points scored is significantly under or over expectation. Since EPV doesn't account for these kinds of discrepancies between teams and relies solely on expectation, it is very possible that it is unable to capture these inconsistencies.

To test for this, I looked closely at two opposite teams when it comes to shooting ability. While both the Golden State Warriors and Houston Rockets were heavy volume shooters in the 2018-2019 season, the Warriors led the league in field goal percentage (% of shots made) at 49.0%, while the Rockets were 7th worst at 44.9%. I ran the same EPV game score predictions for 36 games for both the Warriors and Rockets, which when compared with the actual game scores gave the plots shown in Figure 7-2.

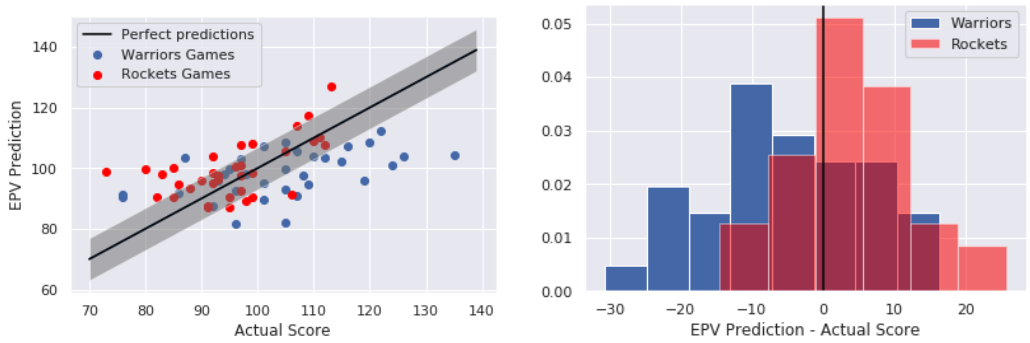

Figure 7-2: Comparison of EPV predicted game scores and actual game scores for the Golden State Warriors and Houston Rockets.

These plots seem to confirm our intuition that the better shooting teams (i.e. the Warriors) typically out-shoot expectation, meaning that their actual game scores exceed their EPV predictions. This results in a majority of the blue dots in the scatter plot falling below the black line. Conversely, the Rockets consistently shooting below average is reflected by their dots falling above the black line. This is more directly highlighted in the histogram on the right, depicting for each game the difference in actual score and EPV prediction. The Warriors distribution is shifted to the left with mean -5.44, while the Rockets distribution is shifted to the right with mean 3.87. Seeing this large discrepancy between teams that shoot at significantly different levels of efficiency is a sign that perhaps if we weight teams according to their field goal percentage, we can reduce the bias to the mean that we observed in Figure 7-1.

Correcting the EPV predictions with team field goal percentages is a relatively straightforward calculation. We can see how each team's field goal percentage compares to the average field goal percentage, then add or subtract a corresponding amount from the EPV predictions. To do this, we first convert both field goal percentage and EPV predictions into standardized units, then correct accordingly. Denoting the corrected EPV prediction as $\hat{E}$, original EPV prediction as E, the team's field goal percentage as $FG\%$, the average field goal percentage as μ_{FG} and standard deviation as σ_{FG}, the average EPV prediction as $\bar{E}$ and standard deviation as σ_E,

the corrected EPV prediction is computed according to Equation 7.1 as follows.

$$\begin{aligned}\hat{E} &= \Big(\frac{E-\bar{E}}{\sigma_E} + \frac{FG\% - \mu_{FG}}{\sigma_{FG}}\Big)\sigma_E + \bar{E} \\ &= E + \Big(\frac{FG\% - \mu_{FG}}{\sigma_{FG}}\Big)\sigma_E\end{aligned} \tag{7.1}$$

To derive this equation, we standardized both the raw EPV prediction and field goal percentage relative to their means, offsetting appropriately in standard units before converting back into points. As we see in the equation, this is equivalent to the intuitive formulation that we simply add to the raw EPV prediction a term proportional to the field goal percentage adjustment.

With the 72 games between the Warriors and Rockets combined with the original 42 sampled games, the raw EPV predictions compared with the corrected EPV predictions are shown in Figure 7-3. We see that extreme values with respect to the x-axis get pulled closer to the black line, and as a result the new line of best fit for the points more closely aligns with the perfect prediction line. In fact, the original slope of the orange line was 0.327, while the slope of the FG% adjusted orange line is now 0.515 (perfect prediction line is slope 1).

Indeed we see that the efficiency of shooting a team-by-team basis does contribute to the underestimation in variance when it comes to EPV game score predictions, while not entirely accounting for it. With that being said, seeing this "bias towards the mean" phenomenon is not overly concerning. Since we are motivated by the problem of evaluating decision-making, we only care about expectation and not so much about the anomalies. However, this does demonstrate that EPV isn't perfect, not developed to readily adjust to specific skill levels or efficiency metrics. In the next section I elaborate on such limitations.

7.2 Limitations

Seeing EPV work empirically well is essential to proceeding to use it for credible analysis around decision-making. However, it is also important to note areas in the

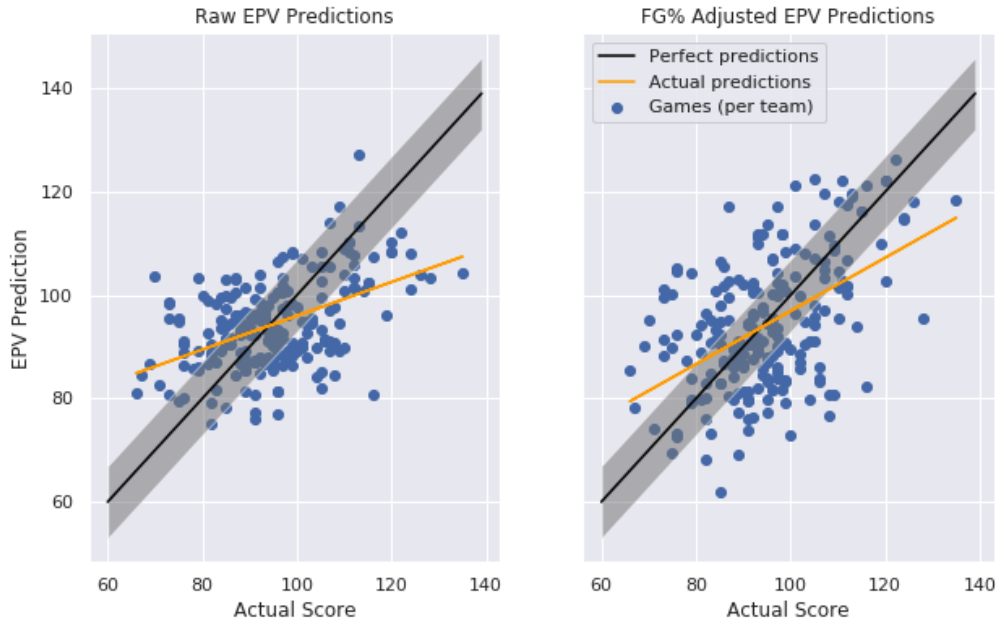

Figure 7-3: Comparison of EPV predicted game scores and actual game scores while correcting for team field goal percentage.

development of EPV which have room for improvement, or are currently limited in their capability.

7.2.1 Improvements for Difficulty Models

While deliberate steps were taken to simplify the features for the shot difficulty model, there are a couple additional details that could significantly improve the accuracy.

An obvious improvement that could be made would be to incorporate the shooter's identity into the prediction of whether they will make the shot. Needless to say, there is a wide range of shooting ability in the NBA, and two highly different players won't have the same shot conversion probability despite the hypothetical configuration of players being the same for both. The reason this was not included in my current iteration of the model was due to its lack of generalizability. If we wanted to incorporate the shooter's identity, we would have to manually assign some kind of metric for all 450 players in the NBA and update that data as new players come in and out of the league. While this is theoretically possible to do, it is a manually burdensome task to do, and falls outside the scope of my research goal.

Similarly, incorporating the identity of the defenders can also significantly boost the performance of both the models. If we can take into account how good of a defender the closest defender is, or how tall they are, we might get some added predictability for how much they can disrupt a shot. Again, while this would be great to incorporate in theory, practically speaking it is difficult to implement and would require a lot of work to maintain this framework.

With that being said, a potential solution would be take an approach that isn't as granular as specifying ability for each player. Instead, we could bin players into discrete types, perhaps grouped by a collection of stats that would characterizing shooting or defending ability. This would be an interesting avenue to explore as an extension of my work.

7.2.2 Oversimplification of Offensive Scenario

It would be an understatement to say that this approach to estimating expected points is a vast oversimplification of what actually happens in an offensive possession in basketball, for a variety of reasons. For one, we have limited the offensive player's decisions to simply passing and shooting. Nothing in this framework accounts for the decision to drive, the decision to dribble backwards and reset, or the decision to simply hold the ball and wait for the play to develop. As such, using EPV in the way that we have developed it is limited to only evaluating the decisions to pass and shoot.

Another slight simplification to the scenario that we have constructed where the ballhandler passes to a teammate and then they shoot is the fact that we have held the defense static. In reality, if a pass were to be made from the ballhandler to a teammate, the defense would likely shift and adjust to the new player having the ball. Currently in my implementation, we assume the defensive configuration to remain the same from hypothetical pass to hypothetical shot. While there is some research work done to simulate defensive rotations as the ball moves around, namely by Hsieh et al. [10], the drawback to incorporating defensive movement is that it introduces another source of uncertainty. Since the time between pass to hypothetical shot is limited

to the duration of the pass, which is typically less than one second, for the sake of simplicity I have considered the lack of defensive simulation negligible.

Lastly, my current approach does not take into account fouls while shooting. In particular, the expected points calculation does not consider the possibility of drawing a foul on any shot, meaning that this isn't incorporated when evaluating a decision to shoot. This would be a somewhat challenging feature to incorporate, as fouls can occur for a variety of reasons. With that being said, incorporating fouls would definitely give a more accurate depiction of what goes into an offensive possession and the outcome possibilities.

Despite these limitations, we have already shown both the pass difficulty and shot difficulty models to be very accurate, and that EPV works well in practice. While evaluating EPV without ground truth is challenging, we are able to still provide a good argument for how accurate it is, and thus can use it convincingly for analysis.

Chapter 8

Application: Missed Opportunities

8.1 Motivation

Though we now have a way of accurately quantifying the expected values of hypothetical actions of passing and shooting, the bridge to decision-making hasn't quite been explicitly articulated. The question of interest now becomes how we can leverage and apply EPV to identify instances where good decisions and bad decisions were made. The novel concept that EPV brings to us is that we can identify these instances based on not just what the outcome of a particular action was, but rather the *expectation* of that action. Thus, we can expand our analysis of bad decisions beyond just concrete outcomes such as turnovers, missed shots, and stolen passes, instead trying to identify *opportunities* without necessarily an attached outcome.

It is easiest to understand this motivation through an EPV possession visualization, as shown in Figure 8-1. This figure depicts approximately a three second window in the possession, in which the ballhandler, DeRozan, dribbles to the right, then looks back to the left and passes to his teammate, Belinelli, at the three point line. With a relatively uneventful short time span that simply consisted of a single non-meaningful pass, one would think naively that there could be nothing to be said about decision-making and execution. Admittedly, this would likely be true, until we take a closer look at what the EPV diagram shows us.

In this short window of time, our eyes are drawn to a spike in EPV for another

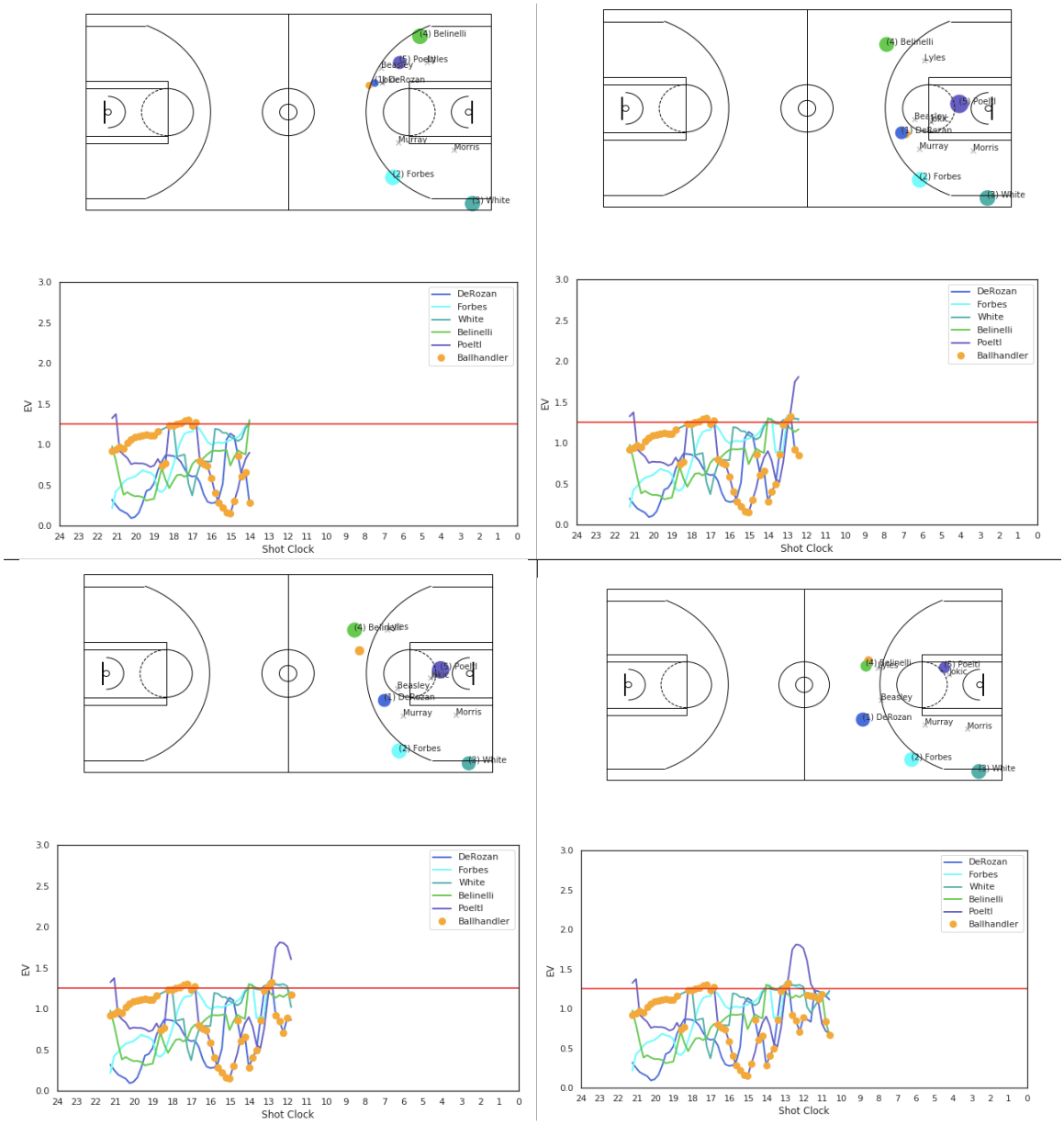

Figure 8-1: Example missed opportunity EPV visualization. Evolution of possession is read from left to right, top to bottom.

player on the offensive team, Poeltl. In particular, we notice that his EPV shoots up to nearly 2.0 for just a second, before falling back down. In observing the dot diagram, we can see why his EPV went up. After setting a screen for the ballhandler DeRozan, both Poeltl's defender (Jokic) and DeRozan's defender (Beasley) follow DeRozan, leaving Poeltl free to cut to the basket. At the instant where Poeltl's EPV is highest (shown in the second frame of Figure 8-1), he is standing wide open right under the basket with a passing lane to receive the ball from DeRozan. Of course, this pass isn't made and instead the ball is circulated back out to another player, at which point the defender Jokic rotates back onto Poeltl and the EPV goes back down.

Obviously, if we paint this EPV-integrated picture for any NBA team's personnel, they would be quick to make judgement on DeRozan's decision not to pass to Poeltl. It is exactly these kinds of instances, extremely hard to pick out to the human eye, that we seek to identify with EPV as moments where decision-making can be evaluated. Heuristically, this means searching for occurrences resembling this example where we see spikes in EPV, and analyzing what the ballhandler actually does to either take advantage of the "opportunity", or to overlook it. By analyzing these opportunities on a large scale, we can then start to derive insight as to how often opportunities are being missed and how bad they are both on a player and team level, giving us some quantitative measures to evaluate decision-making and execution ability.

8.2 Defining Opportunities

To establish a solid framework for analyzing opportunities in this way, we need a careful and formal formulation for what classifies as an "opportunity". As stated before, heuristically an opportunity can be interpreted as a spike in EPV, meaning that there exists a pass or shot at that moment in time that would be highly valuable to the expected points for the possession. To formalize this, we define an opportunity as any offensive player attaining a high EPV value for a sufficient amount of time, such that a decision can be made and consequently an appropriate action can be taken

realistically in that time frame. We also want to differentiate between opportunities in which a shot should be taken, and opportunities in which a pass should be made, as they correspond to different decision choices.

Def. **opportunity**: An opportunity is classified as an instance in which an offensive player p_i has their EPV cross above a threshold T for 2 seconds or longer.

- *Shot opportunity*: p_i is the ballhandler, meaning that their high EPV comes from a chance to take a good shot.
- *Pass opportunity*: p_i is not the ballhandler, meaning that they should get the ball passed to them.

For the sake of evaluating decision-making, we can also now classify opportunities as being either "missed" or "converted". This distinction is very simple in our formalism. If it is a shot opportunity and the player does not take the shot within 2 seconds, the shot opportunities is considered "missed". If it is a pass opportunity and the ballhandler either passes to someone with lower EPV or does not pass at all, the pass opportunity is considered "missed". With this distinction, we can now calculate the rate that certain players or teams miss opportunities, by comparing how many opportunities they miss and how many they convert. We define the *missed opportunity rate* as the percentage of total opportunities that are missed on either a player or team level.

$$\text{Missed opportunity rate } = \frac{\#\text{ missed opportunities}}{\#\text{ total opportunities}} \tag{8.1}$$

In addition to just classifying opportunities discretely as missed or converted, we can also use a continuous measure to quantify how much value an opportunity creates. Intuitively, the higher the EPV spikes, the better the chance. In a similar vein, the longer the duration of the EPV spike, the better the chance. To capture both of these aspects, we can look at the area under the EPV curve (with respect to the threshold

T) as a proxy measure for how good an opportunity is. This area will be large if either the height is large or the width is large, incorporating both the high value of EPV and long duration of an opportunity. Formally, we quantify the *opportunity value* according to Equation 8.2.

$$OV = \int_{t=s}^{t=s+\Delta t} \Big(EPV(t, p_i) - T \Big) dt \tag{8.2}$$

where s represents the start of the opportunity, Δt represents the duration of the opportunity, and we integrate over the time (t) dimension.

As a final metric to define with missed opportunities, it is useful when evaluating decision-making to compare these opportunities with what the ballhandler actually does. As such, for each opportunity we want to compare the hypothetical max EPV with the resulting EPV dependent on the ballhandler's decision. We call this the *missed opportunity delta*, which is simply the max EPV from the actual ballhandler(s) during the opportunity subtracted from the max EPV for p_i in that duration. A high difference would indicate that the opportunity could have resulted in a much better chance for the team with respect to what actually happened, whereas a small difference would indicate that the opportunity may not have been a drastic improvement over what the ballhandler actually decided to do.

Having defined these metrics that we want to use to aid in our analysis, we are nearly ready to run these calculations on our dataset. However, there is an important parameter that we have to choose before doing so, namely, the used a threshold of 1.5 arbitrarily based on intuition, observing that scoring an expected 150 points in an NBA game would be exceptional. To confirm this as an appropriate choice, I ran 200 possessions through this opportunity identification process, varying the threshold from 1.0 to 2.0 to see how it would affect opportunities being classified. Calculating the total opportunities and missed opportunity rate on a player basis, the results of varying the threshold are shown in Figure 8-2.

As expected, we see that as we increase the EPV threshold, the number of total opportunities per player goes down, as it becomes less and less likely for spikes to be

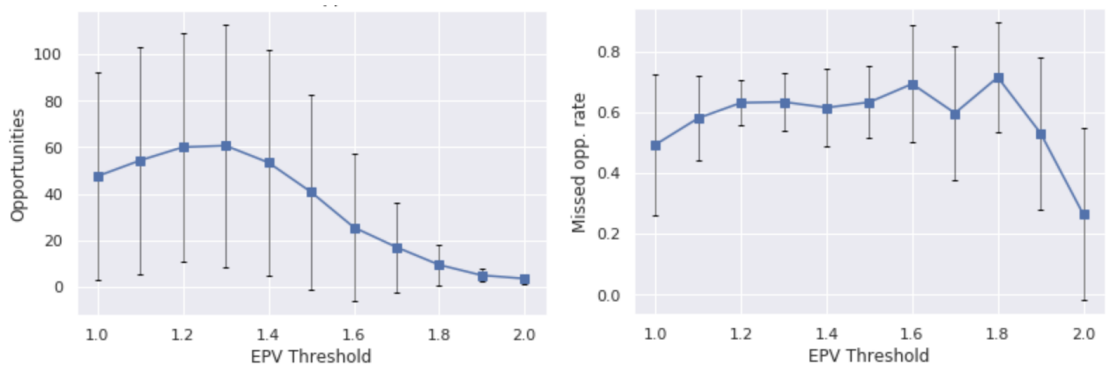

Figure 8-2: Average opportunity metrics by player as a function of threshold. *Left.* Total opportunities vs. threshold. *Right.* Missed opportunity rate vs. threshold.

classified as opportunities. Interestingly though, we don't observe that same effect in the missed opportunity rate until the EPV threshold is above 1.8. We don't want to select the inflection point 1.8 as our threshold however, because that would result in too few opportunities to analyze, while also being way too high of a bar to set, expecting teams to score 180+ points in a game. Instead by choosing 1.5, we still get to analyze a significant amount of opportunities, while not changing the missed opportunity rate. Thus, in our remaining analysis, we set the opportunity threshold $T = 1.5$, though it is worth noting that this is a tunable parameter in the framework.

8.3 Opportunities Calculations

While we wanted to use as large a sample size as possible to look at opportunities, we were limited computationally in how many possessions we could analyze. Because EPV relies on tracking data for 10 players at once for computing values 5 times a second, the runtime for computing EPV over the course of 100 possessions can extend beyond an hour in our current infrastructure. With this limitation in mind, it was important to determine how many possessions we would have to generate EPV values for, such that the sample size of opportunities was large enough to draw real conclusions.

Translating this into a testing strategy, I calculated a few opportunity metrics as defined in Section 8.2 by running EPV calculations for 1500 possessions for the San

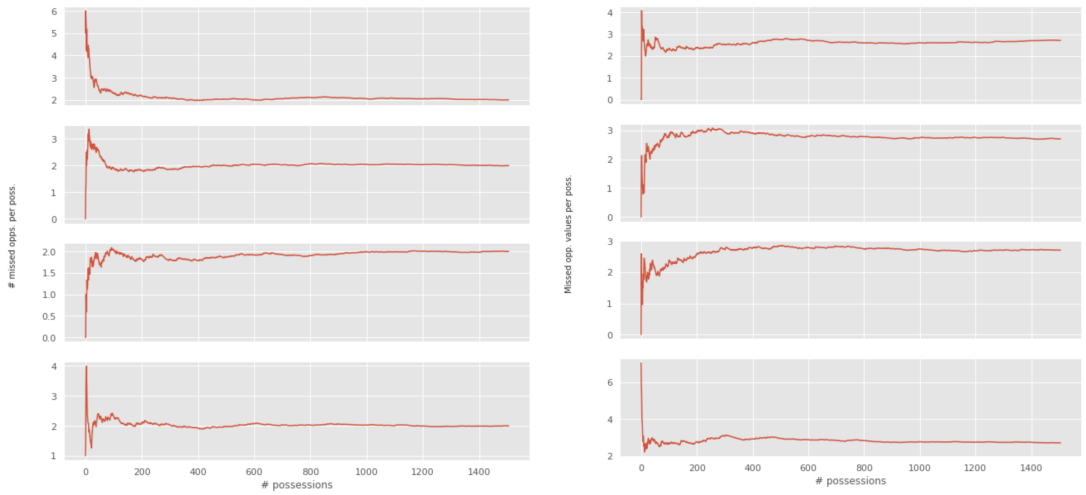

Figure 8-3: Four trials of running averages of opportunity metrics for Spurs as a function of number of possessions. *Left.* Average number of missed opportunities. *Right.* Average missed opportunity value.

Antonio Spurs. I computed the running averages for these metrics as a function of how many possessions had been run already, with the idea that we should see the running average converge beyond a certain number of possessions. In particular, I calculated a running average of the missed opportunities per possession, as well as the average missed opportunity values. Figuring these would be important metrics to look at when we considered decision-making, it was worth seeing how they changed as a function of the number of possessions included.

Shown in Figure 8-3 are the running average of opportunity metrics (for the San Antonio Spurs as a team) plotted against the number of possessions. I ran the EPV calculations for all 1500 possessions four separate times, each time randomizing the order of possessions, such that we could observe convergence regardless of the order in which the possessions are processed. Each row in the figure depicts one of these four run-through calculations.

We observe the exact type of convergence that we are looking for surprisingly early on, at about 200 possessions. Specifically this tells us that possessions after the 200 mark would just be excess data that will not necessarily affect any insights we draw in our analysis. This is immensely useful to know, because it gives us a high level of confidence that running a finite set of possessions is enough to derive accurate metrics

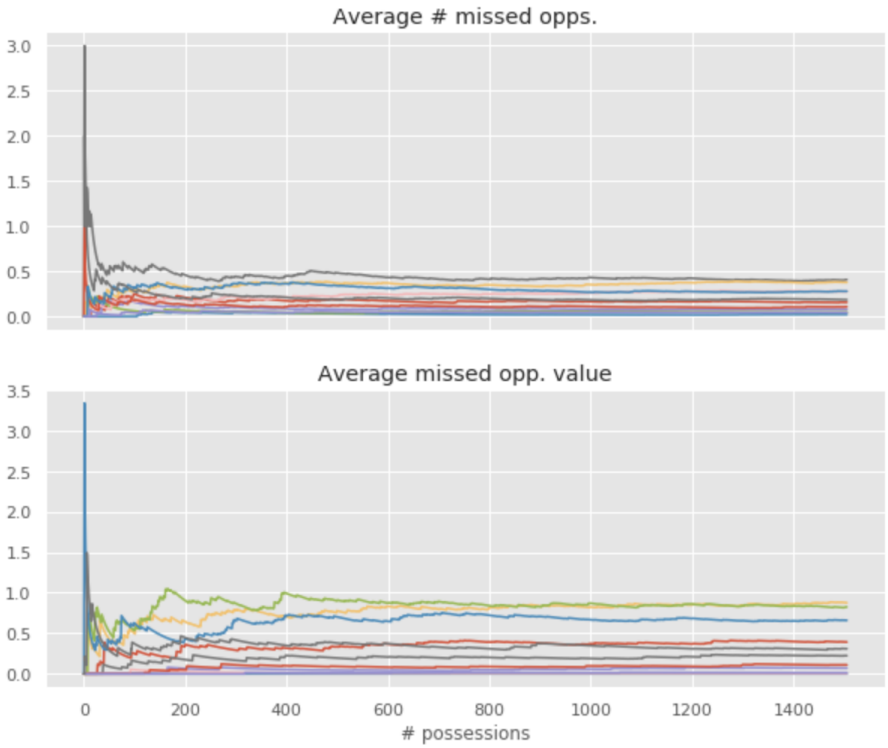

Figure 8-4: Running average of opportunity metrics for players as a function of number of possessions. Each color represents a different player.

for our evaluation and analysis. In addition, we can expect to run this smaller set of EPV calculations in a much more reasonable amount of time computationally, which helps us tremendously for practical purposes.

I ran a similar test calculating the same opportunity metrics but this time on player-by-player basis to see if we would observe a similar convergence. The results are shown in Figure 8-4. Again, while we see some fluctuations with a small number of possessions, as we get beyond 200 possessions each player's metrics flatten out, indicating that is a sufficient number of possessions to consider to also evaluate players' decision-making with respect to opportunity analysis.

In my actual analysis, I decided to use 400 possessions for each team as a way of being conservative with a larger sample size while still being able to run them in a reasonable amount of time. I focused my attention on the month of January 2019, extracting the first 400 possessions for each team in this period to run my analysis on. The reason I did this instead of randomly selecting possessions was to minimize

the variability within each team with regards to strategy, play-style, personnel, and overall team ability. The results of this analysis will be discussed next.

8.4 Player Opportunities

In light of evaluating decision-making, I started by looking at how individual players took advantage of opportunities presented to them as ballhandlers. In addition to tracking opportunities, I also track how much time each player spends as a ballhandler, such that we can appropriately compare players who rarely have the ball in their hands versus ball dominant players. The ballhandler percentage is computed as the number of frames spent as the ballhandler divided by the total number of frames the player is involved in. Semantically, this translates to how ball dominant this player is when they are on the court.

Of course, the more ball dominant a player is, the more opportunities will arise for them compared to off-ball players. When using missed opportunity metrics as proxies for aiding in our decision-making evaluation, we don't want this correlation to bias our analysis. In particular, we have to treat ballhandler time as a confounding variable that we must control for before making direct comparisons with missed opportunity metrics. These comparisons can be made as long as we are careful to normalize by ballhandler time, meaning that we compute opportunity metrics per unit of time spent as the ballhandler.

8.4.1 Player Missed Opportunity Rates

After collecting opportunity metrics for the set of NBA players throughout January 2019, the first thing I wanted to look at was a breakdown of the missed opportunity rates. Shown in Figure 8-5 are histograms showing the missed opportunity rates for all players during this time span.

Interestingly, we see that the overall average missed opportunity rate is pretty high at about 65%. If we think about how we have defined opportunities, at each point in time when an opportunity arises, there are 6 decisions the ballhandler can

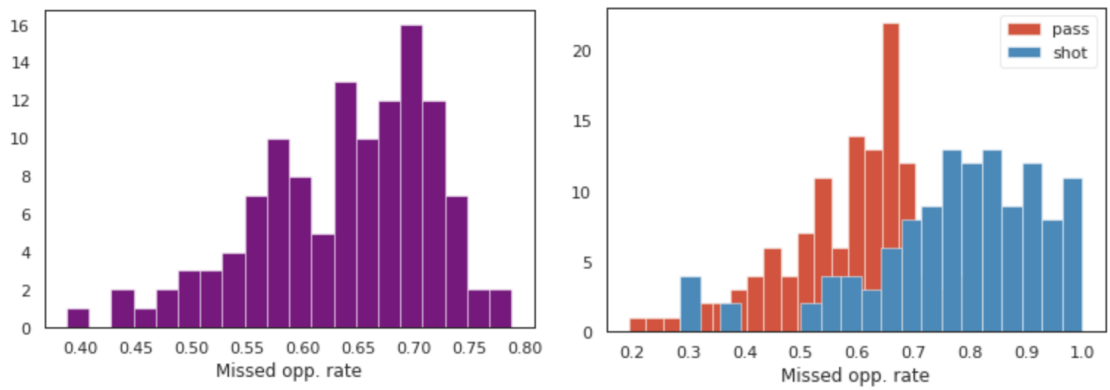

Figure 8-5: Histogram for missed opportunity rates. *Left.* All missed opportunities. *Right.* Separated into shot and pass opportunities.

make:

1. Pass the ball to the "opportunistic" high EPV player.
2. Pass the ball to other offensive player 1.
3. Pass the ball to other offensive player 2.
4. Pass the ball to other offensive player 3.
5. Keep the ball.
6. Shoot the ball.

For pass opportunities as we have defined them, actions 2-6 would be considered missed opportunities, while action 1 would be converting the opportunity. For shot opportunities as we have defined them, actions 1-5 would be considered missed opportunities, while only action 6 would be converting the opportunity. Thus, we see that at for any particular opportunity that arises, 5/6 discrete choices would result in missed opportunities. If the decisions at these junctions were made completely randomly, the missed opportunity rate would be on average exactly $5/6 \approx 0.83$. The mere fact that we observe the actual average significantly lower at 0.65 is a confirmation that indeed these decisions are not made randomly and that there is an element of decision-making ability. Of course, the goal of a good decision-making player is to

widen the gap as much as possible away from the "random decision-making" missed opportunity rate of 0.83.

Another interesting phenomenon to note is that on average the missed shot opportunity rates are higher than the missed pass opportunity rates. A potential explanation for this is that players on average are biased against shooting and would rather pass the ball to develop a possession even if they have a good shot. NBA teams often stress passing much more than shooting for an effective offense, and often prefer running set plays with predetermined passes rather than taking advantage of an unexpected open shot. Of course, this will differ from player to player, but is an interesting pattern to see nonetheless.

8.4.2 Comparisons with Missed Opportunity Rate

To actually start to construct the narrative around decision-making with players in the NBA, we looked at exactly what kinds of players had lots of opportunities and how they took advantage of those opportunities. Normalized by the amount of time spent as the ballhandler, we generated some interesting player cloud plots that allowed us to compare and contrast these metrics across players. We also separate shot opportunities and pass opportunities to further clarify the kinds of situations these players find themselves in.

As a first example, consider the player cloud depicted in Figure 8-6, plotting the number of pass opportunities a player gets per second as the ballhandler against their missed pass opportunity rate. Evidently, we see a clear positive correlation between these two variables. To put into basketball terms, the more pass opportunities a player gets per second, the more likely they are to overlook those opportunities. The plot is broken into four quadrants, indicating four separate bins of players. The players that live in the upper right are unsurprisingly very ball-dominant players, but who also happen to have high missed opportunity rates. An ideal ball-dominant player from a decision-making standpoint should be down in the lower right quadrant, though we can attribute the higher missed opportunity rate to their high ballhandler usage. On the other hand, the players in the bottom left corner all tend to be big men

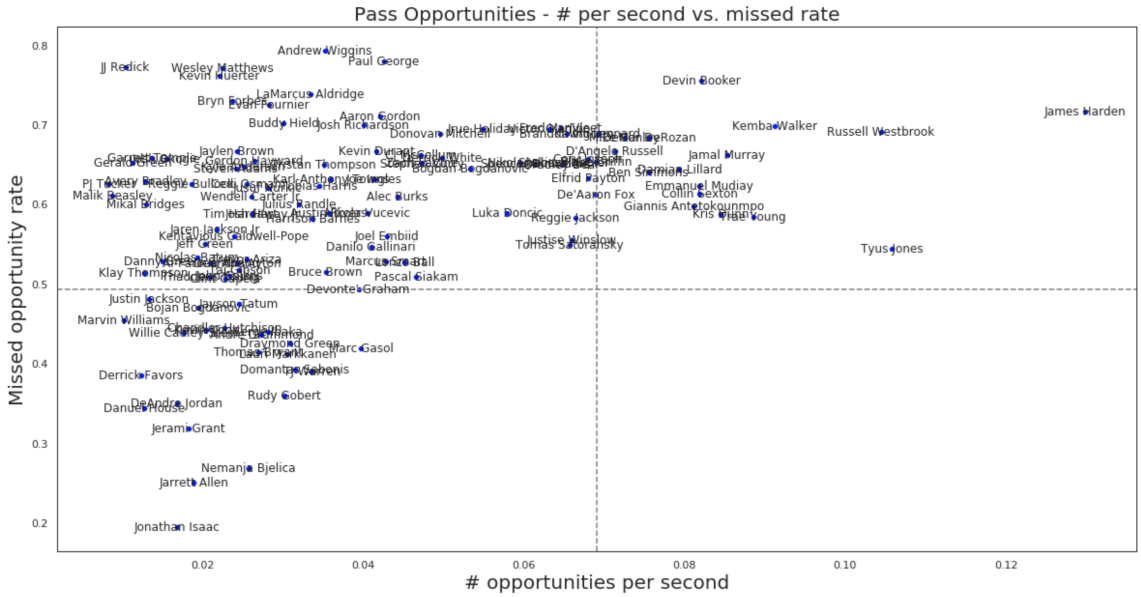

Figure 8-6: Player cloud of pass opportunities per second vs. missed pass opportunity rate.

who rarely spend time with the ball, reflected by their lack of pass opportunities per second. They also have low missed pass opportunity rates, a sign of good decision-making when they get their rare opportunities to make a good pass.

The troubling group, which also happens to be the most populous group, is the upper left quadrant. These players not only lack pass opportunities in general, but when they do get them they don't capitalize well. When we look to evaluate players' decision-making, this is the quadrant we want to focus on. A player's amount of pass opportunities is something that is out of their control, dictated by a team's strategy and the movement of their teammates. If you have lots of opportunities due to always being on the ball, it makes sense to have a higher missed opportunity rate, which we can explain away. However, if you do not have those opportunities, the only thing that prevents you from having a lower missed opportunity rate is decision-making and execution. Thus, in looking at this player cloud, we can start to make comparisons about players' decision-making along any particular vertical split of the graph. For example, we could consider all players who have 0.04 pass opportunities per second, and compare their missed opportunity rates. We would observe Andrew Wiggins and Paul George to have the highest missed opportunity rates at around

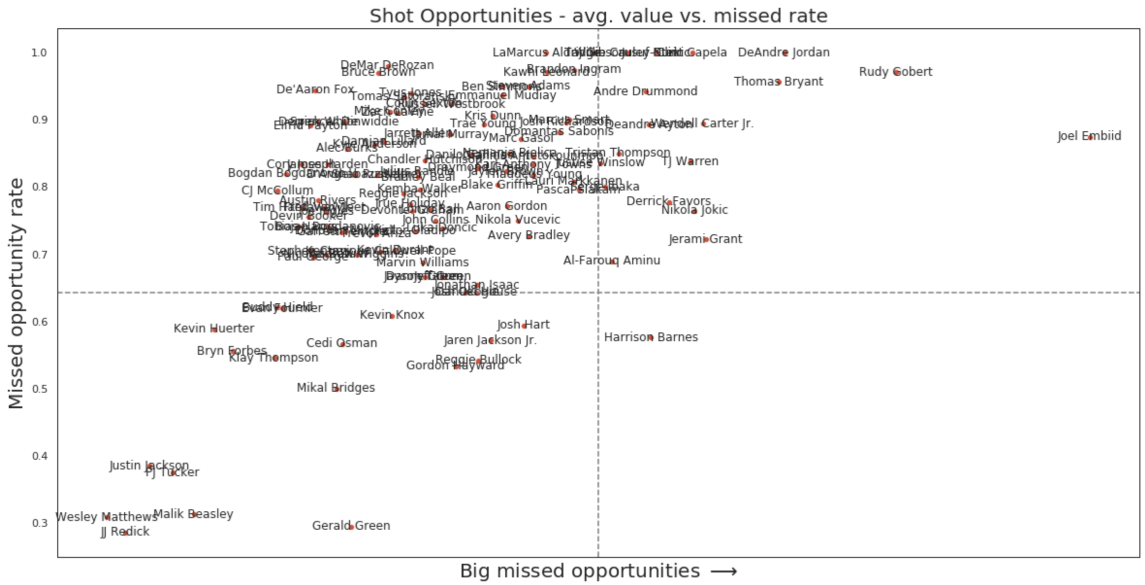

Figure 8-7: Player cloud of shot opportunity values vs. missed pass opportunity rate.

0.8, and Marc Gasol to have the lowest around 0.4. While overly simplistic to make the blanket statement "Marc Gasol is a better decision-making", this does give us exactly the framework to explore these kinds of insights. An equivalent plot for shot opportunities is located in Appendix B as Figure B-1.

In addition to looking at the pure number of opportunities per second, we can also analyze the average missed opportunity value, how bad each missed opportunity tends to be per player. Shown in Figure 8-7 is a similar player cloud plot, but this time with missed opportunity value on the x-axis. A corresponding plot for pass opportunity values can be found in Figure B-2 in Appendix B.

Interestingly, we now see somewhat of a reversal in the narrative with respect to the types of players. In particular, the big men now group together in the upper right quadrant of Figure 8-7, indicating that that they both miss a lot of shot opportunities, which also tend to be great shot opportunities in terms of expected points. Meanwhile, the players that rarely miss shooting opportunities and whose shot opportunities are usually not as good (lower left quadrant) can be generally classified as perimeter shooters who don't spend a lot of time in the paint. Again, from this plot we can construct a narrative about decision-making, specifically shot selection in the NBA. Players who classify as shooters often do not hesitate to shoot, hence the low missed

shot opportunity rate, and also will take harder shots that result in lower expected points. This is in stark contrast to big men who tend not to be as confident in their shot, and as a result squander good opportunities to get a lot of expected points from a possession. While the strategic implications of this are up to coaches to interpret and act on, being able to understand that there are points being left on the board due to shooting hesitancy from players is an integral component to evaluating how their decision-making impacts the game.

8.5 Team Opportunities

Evaluating opportunities on a per player basis can give us a lot of insight about decision-making on an individual level, which is extremely valuable for each of the teams. However, we can also use this framework to develop a more holistic view on the league in general by analyzing opportunities on a per team basis. When evaluating teams as a whole, we can think of the overall "decision-making" as the execution level of the team: how well do they capitalize on the opportunities that are generated in their offense. As such, this analysis becomes one of evaluating team strategy versus team execution. The strategy component lies in how many opportunities the team is able to set up through game-plan, while the execution component lies in how often they miss or convert these opportunities.

The easiest way to see this is through another four-quadrant plot as shown in Figure 8-8. This figure plots the number of shot opportunities against the missed shot opportunity rate, which correspond to strategy and execution, respectively. Keeping in mind that these opportunity metrics were run for the same number of possessions (400) for every team, comparing where each team lies on the plot gives us a way of comparing their levels of strategy and execution relative to other teams. Breaking the plot into four quadrants similar to what we did with the individual player plots further categorizes the teams into some combination of good/bad strategy and good/bad execution, which further emphasizes what this analysis can demonstrate.

There are some interesting takeaways from this plot that we can use in conjunction

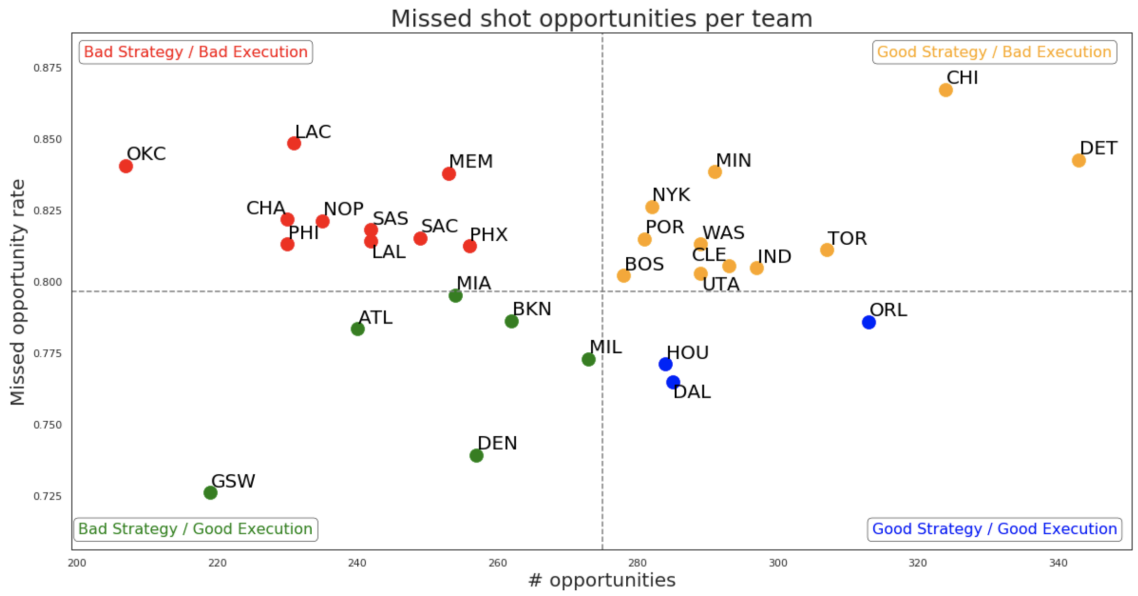

Figure 8-8: Team cloud of number of shot opportunities compared to missed shot opportunity rate. Number of opportunities increases with better strategy, missed opportunity rate increases with worse execution.

with what we know about how good these teams were in the 2018-2019 season. The Golden State Warriors (GSW) were dominant during this season, finishing with the third best record in the NBA and the highest offensive efficiency. This efficiency is indeed reflected by their vertical position in this plot, meaning that they are good at converting opportunities when presented to them. Of note though is the fact that they also are located far to the left in this plot, a sign of bad strategy. The NBA community viewed the Warriors during this time period as a flawless offensive basketball team, due to the sheer number of shots they would make and the points they could put up every game. Instead what we are seeing is that in fact their success is largely tied to the fact that they possessed three of the top shooters in NBA history in Stephen Curry, Klay Thompson, and Kevin Durant, and were inherently poised to make more shots than other teams. With regards to actually creating opportunities consistently, this plot shows us that they were lacking in that department, and were instead reliant on exceptional shooting for their success.

We can also consider a team that performed very poorly in the 2018-2019 season, the Chicago Bulls (CHI). With the fourth worst record in the NBA that season and

the second-worst offensive efficiency, it's no surprise to find the Bulls located in the upper half of this plot, as they clearly weren't executing close to a team like the Warriors. While typical stats could tell you a simple narrative like that, this plot offers some actionable insight for the coach. The Bulls were actually one of the best teams at creating consistent shot opportunities through their offensive game plan, and their lack of success can actually be almost entirely attributed to their poor execution. Identifying this exact weak point is extremely valuable to the coaches and players on the Bulls, emphasizing the need to practice shooting rather than changing up strategy.

Differentiating between game-plan success and execution success up to this point has been a very subjective task that is extremely difficult to do accurately. Teams are often curious as to whether their strategy is creating good chances to score or how the players are performing relative to their system. By setting up this framework, we have provided a method of independently evaluating the two. Generally speaking using an analysis like this, coaches can better understand if they need to modify their strategy like the Warriors, or if they need to change / train their players to perform better like the Bulls.

We can also focus more on the efficiency of execution with opportunity analysis, by looking at how teams' missed opportunity rates compare to average missed opportunity deltas. Recall that the missed opportunity delta is the difference in max EPV from the opportunity and the realized EPV during that duration. An example of this for pass opportunities is shown in Figure 8-9.

This plot tells us less about team strategy and more about how they capitalize on the opportunities presented to them on offense. Once again, teams that fall to the right of this plot tend to miss more opportunities, indicating that they are not efficient in decision-making on the offensive end of the court. Now if we consider missed opportunity delta on the vertical axis, we start to get a more complete picture on how teams actually execute. Teams in the lower part of the plot tend not to miss big opportunities. The opportunities they do miss are because the ballhandler decided to make a play that resulted in a similar EPV. However, teams with higher

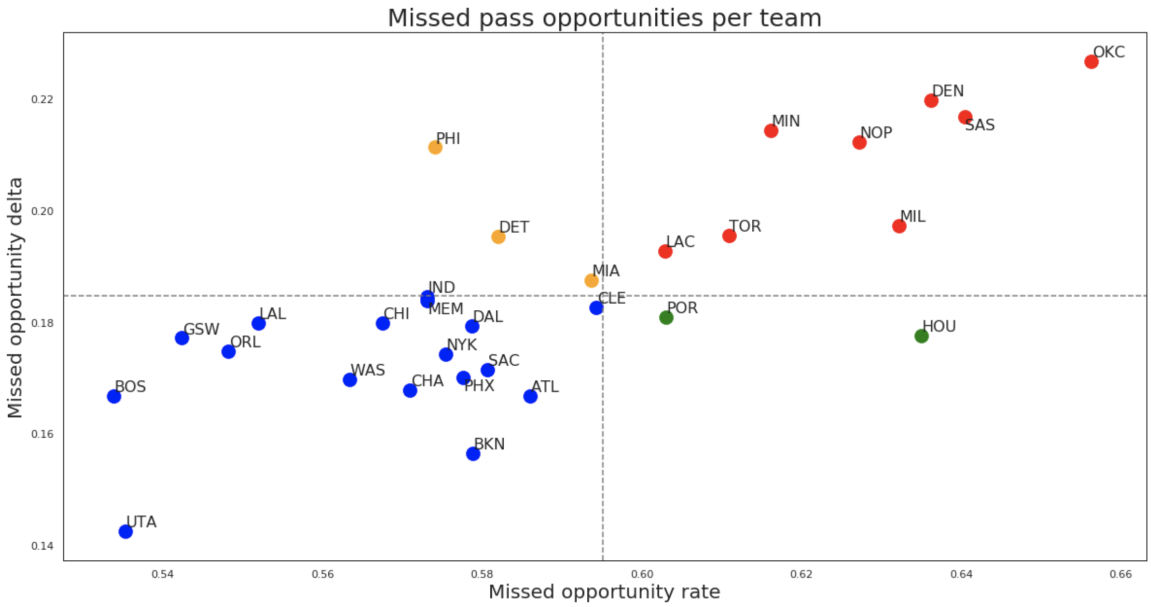

Figure 8-9: Team cloud of missed pass opportunity rate compared to average missed pass opportunity delta.

average missed opportunity delta can be thought of as being too conservative or not decisive enough to make the critical pass. While opportunities arise for these teams, what actually ends up happening is that the ballhandler will hesitate or make a much more conservative decision that results in a lower quality possession. Seeing this from the point of view of a team like the Philadelphia 76ers (PHI), for example, might serve as a tool for understanding that players need to take more risks on the offensive end and be more definitive in their passes, because the good opportunities are there. Teams on the right side of this plot can now understand that they are often missing opportunities to create good shots for their team and can quantify how much it is hurting them. This understanding translates into actionable insights for the coaches to try and push their teams down to the bottom left as much as possible.

Evidently, this kind of opportunity analysis can tell us a lot about how effective team strategy is and how well they execute. In addition, we can disentangle the two by using different opportunity metrics, thinking about to what components we can attribute a team's success to and identifying areas on offense where they can be more efficient. As such, I expect opportunity analysis to be an incredibly useful tool in NBA analytics.

Chapter 9

Discussion

Through identifying opportunities and analyzing how players and teams react to those opportunities, we have shown how using the expectation based metric EPV can help us evaluate decision-making, team strategy, and execution on the offensive end of the court with regards to passing and shooting. Because this metric relies on expectation and not on outcome, it is not straightforward to convince NBA players and coaches of its usefulness as an analytical framework. To do so, we need to translate our analysis into the language of wins and losses and points being left on the table in order for NBA personnel to fully envision the benefits. I discuss some of these quantities in this chapter, showing how missed opportunity analysis and EPV can quantify hypothetical optimal performance.

Despite its usefulness, EPV right now is still very much focused on a large but limited component of the game: passing and shooting. Not only is the analysis restricted to the offensive end of the court, but other actions such as driving and drawing fouls are not integrated into the framework I have developed, as I have discussed in Chapter 7. In this chapter I will also pose questions about how we can further advance this framework to better capture decision-making, and how we can further leverage EPV as is.

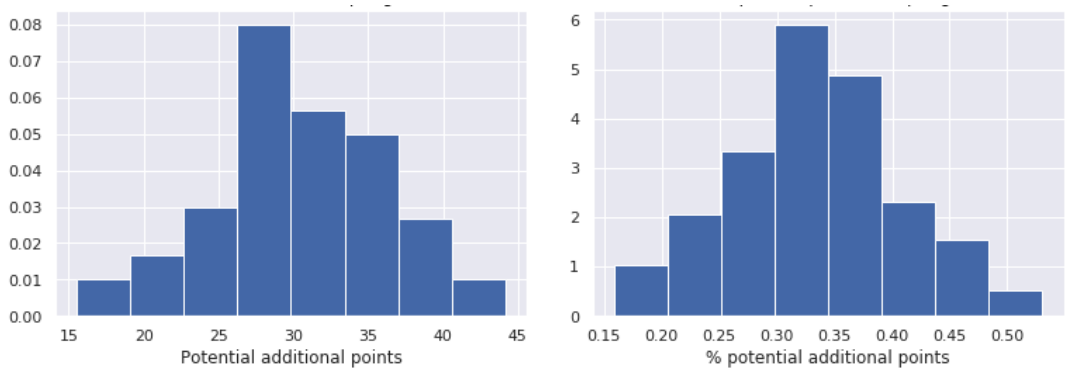

Figure 9-1: Histograms showing how many additional points could be had on expectation per game if opportunities were fully capitalized on. Left is raw additional points, right is in terms of percentage of actual points scored that game.

9.1 Point Improvement Potential

To be truly convincing to teams that EPV is useful for their analysis, we need to show how teams *could* be performing if they were fully capitalizing on their opportunities. Using our missed opportunity analysis on a game-by-game basis, we can pose the question: if a team had a 0% missed opportunity rate, how many additional points could they have scored? This is a simple question that our framework is well equipped to answer. In particular, we just sum up all the missed opportunity deltas throughout a game, capturing the differences between what teams actually did versus what they hypothetically could have achieved.

The results are quite astonishing, and are shown in Figure 9-1. These two bar charts are histograms for teams showing how many additional points can be scored if every opportunity was converted, on a per-game basis. On the left shows the raw number of additional points that are being left on the table, while the right chart puts the points in terms of percentage improvement over the actual points scored by that team in a game. The average potential point improvement for a game is 30 points, which is more than enough to flip a loss to a win. The average percentage improvement is 33%, indicating that teams are currently operating at well below optimal efficiency.

Taking these results to teams will no doubt show the efficacy of using EPV as a

means for extracting insights about inefficiencies. NBA teams revolve their decisions around winning, largely concerned with metrics and analytics that will provide them with more victories throughout a season. Not only is this analysis highlighting that teams have the ability to score many more points in a game and thus win many more games, it also provides actionable insights as to why that isn't happening currently, and what the coaching staff may need to do to address it. Clearly, through this lens, we see how massively beneficial using this framework can be for teams that are looking to improve upon their inefficiencies and consequently be more successful and competitive in the NBA.

9.2 Decision-Making Beyond EPV

While we have shown EPV to be very powerful in basketball analysis, it only scratches the surface of what can be said about decision-making of players. Limited to the simple decisions of passing or shooting, many more actions could be incorporated into the calculation of EPV to take into account other decisions that can be made on the court. On the offensive side of the ball, this could include the decision to drive to the basket, which might be a very good decision if there was a high probability of drawing a foul. We could also extend the analysis to the defensive side, looking at defenders who give up very little high EPV moments to the other team and exploring their defensive actions as well.

Something that isn't addressed in my work but is crucial to evaluating decision-making is incorporating small details about the game environment that give rise to "basketball IQ" plays. Good decision-makers in the NBA are usually considered high basketball IQ players, meaning that in the face of certain unlikely situations they tend to think creatively and decisively to make a high value play. For example, good teams will usually implement a "2-for-1" opportunity at the end of quarters. In a 2-for-1 situation, the offensive team will hold the ball until about 31 to 33 seconds left in the quarter before shooting, no matter where they are on the court. Naturally, these shots tend not to be high percentage shots as they are often forced, and EPV

would classify them as such, characterizing these shots as poor decisions. However the reason teams do this is because it forces the other team to use approximately 20-24 seconds to run their possession, before giving themselves the final opportunity to score in the quarter. Essentially the 2-for-1 tactic can be thought of as a way of guaranteeing your team the last possession, while also giving you a "free" shot with about 30 seconds left on the clock. A player's decision to take that 2-for-1 shot should be considered good for that reason. EPV isn't complex enough to handle situations like these. What's important to understand about EPV is that it is not a particularly specialized system designed to understand the game of basketball in its entirety. Rather, the development of EPV was focused on breaking down the game into the simple components of passing and shooting, quantifying the easiest probabilities we could while giving estimates for the value simple decisions can result in. This simplicity is what makes EPV so powerful, as it can be applied to all aspects of the game and works well on average. Of course there are situations in basketball such as the 2-for-1 that concoct unique conditions for what should be considered a good or bad decision, but EPV in general can give us a wonderful narrative about decision-making that we never could really characterize before.

9.3 Integration into the NBA

The last big question to ask is how exactly a system like this can be actually integrated into the NBA, and put into the hands of the teams it would help. Many teams (including the San Antonio Spurs) already have very thorough analytics platforms that they use to watch game film, store stats, and run their internal software. My goal would be to package the EPV computations into a software module that was pre-configured to identify opportunities, match them up with video, and provide our visualization from Section 6.2. After games, I imagine that the EPV module could run automatically once tracking data becomes available, and all opportunities, EPV values, and corresponding film from those instances could be readily queried in the teams' systems for analysis. This is something we have discussed with the San Antonio

Spurs and is something that they have expressed interest in using.

Chapter 10

Conclusion

Motivated by the difficult task of evaluating NBA players' decision-making abilities, my research has set a firm foundation for leveraging expectation to assess the value of hypothetical player actions. Through deep learning techniques, I was able to develop models that accurately quantify the difficulty of passes and shots, showing that these neural network approaches are comparable in performance to existing models. The novel metric Expected Possession Value was then constructed by the simple integration of these models, capturing the expected points a player will contribute to the possession if the decision is made to pass the ball to them. This highlights an important aspect of decision-making analysis that my research is one of the first to address: quantifying hypotheticals and counterfactuals. Without asking the question "what if", no decision-making evaluation can occur, as there would be no consideration of the possibilities a decision-maker thinks through. By explicitly laying out possible actions for a ballhandler in the form of passing to a particular player or shooting the ball, then quantifying each of those actions, we build an evaluation framework that replicates how a player would think.

Using this framework, we unlocked a treasure trove of tools and insights into how we can start to analyze decision-making and team efficiency. In particular, EPV allowed us to identify opportunities in possessions where on expectation, a certain action would result in a high number of points. These opportunities served as unique points in the game where we could zoom in and start to understand if players were

making good decisions or not. Knowing that the objective is to maximize expected points in each possession, I performed an in-depth analysis with opportunities, seeking to understand how often those opportunities are missed and relating that back to decision-making. What I found was that indeed there is tremendous amount of room for improvement when it comes to being more efficient with regards to these opportunities, a characteristic that we can largely attribute to decision-making. In fact, I quantified this improvement by showing that teams were leaving a large number of points on the table throughout their games, to the point where it could change the outcome. This proved the value that EPV and opportunities analysis can bring to the game of basketball, and showed its efficacy in the evaluation of player and team performance.

My research has a lot of promise moving forward as we continue to try and evaluate subjective components of athletes. A lot of natural extensions arise from what I have laid out that could further the advance the role of EPV in this kind of analysis. I discussed a few limitations of my work, including the fact that my models don't incorporate player identities, and that the action space for the ballhandler has been reduced to merely passing or shooting. These kinds of simplifications were necessary to get a working system in my time, but offer lots of untapped potential in terms of maximizing the accuracy of EPV.

As it stands, decision-making analysis is still very challenging to do objectively, and there is sparse research around it. My work on this project serves as an excellent first step towards doing just that, offering a new perspective through expectation to evaluate cognitive ability. It is my hope that this can be embraced in the NBA with their newfound adoption of analytics, and can be used as a powerful tool like never before.

Appendix A

Tables

Table A.1: Possessions table schema.

Name	Description
id	Unique ID for the possession
gameId	Game code
season	Season of the game, marked by the year of the first game of the season
period	Regulation periods 1-4, overtime periods 5 and up
startFrame	Frame ID, marking the start of the possession
endFrame	Frame ID, marking the end of the possession
startWallClock	Start time stamp provided for a frame, measured in milliseconds
endWallClock	End time stamp provided for a frame, measured in milliseconds
startGameClock	Number of seconds remaining in period at the start of the possession
endGameClock	Number of seconds remaining in period at the end of the possession
offTeamId	ID of the team on offense
defTeamId	ID of the team on defense
startType	Type of the possession based on how the possession starts
outcome	How the possession ended, selected from a specified list
offPlayerIds	List of player IDs for the offensive players on the court for the possession
defPlayerIds	List of player IDs for the defensive players on the court for the possession
homeStartScore	Amount of points the home team has at the start of the possession
awayStartScore	Amount of points the away team has at the start of the possession
ptsScored	Number of points scored in the possession
bringUpBhrId	ID of the player who brings the ball up the court for that possession
reversals	Number of ball reversals in the possession
outcomePbpId	Sequence number of the outcome event of the possession
ballInPaint	Whether the ball was in the paint at any point during the possession
endLoc	Location of last offensive ball handler in the possession

Table A.2: Passes table schema.

Name	Description
id	Unique ID for the pass
gameId	Game code
season	Season of the game, marked by the year of the first game of the season
possessionId	ID of the possession this pass occurred in
chanceId	ID of the chance this pass occurred in
period	Period this pass occurred in.
startGameClock	Time on the game clock when the pass was thrown
endGameClock	Time on the game clock when the pass was received
startFrame	Frame ID this pass was thrown at
endFrame	Frame ID this pass was received at
shotClock	Number of seconds left on shot clock
startWallClock	Time stamp provided for the start frame, measured in milliseconds
endWallClock	Time stamp provided for the end frame, measured in milliseconds
offTeamId	ID of the offensive team when the pass occurred
defTeamId	ID of the defensive team when the pass occurred
passerId	Player ID for the passer
receiverId	Player ID for the receiver. Null if the pass was never received
toReceiverId	Player ID of the player to steal the ball, if a player stole the ball
passType	Type of the pass, selected from a list of pass types
complete	True if the pass was complete to a team member, false otherwise
turnover	True if the pass was a turnover, false otherwise
ledToShot	True if the receiver shot the ball, false otherwise
assistOpp	True if the receiver shot the ball within 1 dribble and 2.5 seconds of catching it, false otherwise
secondaryAssist	True if the receiver of the pass recorded an assist without dribbling or holding the ball for more than one second
reversal	True if the pass changes the strong side of the court
inbounds	True if the pass was an inbounds pass
backcourt	True if the pass originated in the backcourt
passerLoc	Location of the passer
receiverLoc	Location of the receiver, if one exists
toReceiverLoc	Location of the player who stole the ball, if the pass was a live ball turnover
distance	Distance in feet that the pass traveled
ballStartLoc	Ball's location at the start of the pass.
ballEndLoc	Ball's location at the end of the pass.

Table A.3: Shots table schema.

Name	Description
id	Unique ID of the shot
possessionId	ID of the possession in which this event took place
chanceId	ID of the chance in which this event took place
startFrame	frame ID, denoting frame number from the start of the current period
startWallClock	Time stamp provided for a frame, measured in milliseconds
gameId	Game code
season	Season of the game, marked by the year of the first game of the season
period	Regulation periods 1-4, overtime periods 5 and up
startGameClock	Number of seconds remaining in period, at the release of the shot
location	Location of the shooter at the release
shooterId	Player ID of shooter
closestDefId	Player ID of shooter's defender
offTeamId	ID of team on offense
defTeamId	ID of team on defense
shotClock	Number of seconds left on shot clock
fouled	True if the shooter was fouled, false otherwise
outcome	True if the shot was made, false if missed
three	True for three point attempts
assisted	True if the shot was marked with an assist in the play-by-play
assistOpp	True if the shooter took <= 1 dribbles and held the ball for less than 2.5 seconds
distance	Distance from shot location to the hoop, in feet
dribblesBefore	Number of dribbles the shooter took before shooting
shotType	Shot type selected from a list
complexShotType	Shot type selected from a more complex list
catchAndShoot	Whether or not the shooter dribbled before shooting
qSQ	Expected probability the shot goes in, multiplied by 1.5 if the shot is a three pointer. Ignores identity of shooter and defender
qSP	Expected probability the shot goes in, multiplied by 1.5 if the shot is a three pointer. Considers identity of shooter
passerId	Player ID of the player that passed to the shooter. Null if no pass
passerLoc	Location of the passer
blocked	True if shot was blocked, false otherwise
blockerId	Player ID of the player who blocked the shot. Null if not blocked
createdFromPaint	Whether the ball was in the paint 2 seconds or less before shot was taken
receiverLoc	Player's location when they received the ball
putback	True if shot taken within 2 seconds of offensive rebound by same player
contesterIds	Player IDs of the contesters of the shot
contested	Whether any defenders are contesting the shot
releaseTime	Number of seconds from the start of the touch to the shooter's release
pbpId	Event number of the shot in the play-by-play feed

Table A.4: Tracking table schema.

Name	**Description**
gameId	Game code
period	Regulation periods 1-4, overtime periods 5 and up
gameClock	Number of seconds remaining in period
playerId	Player ID
homeTeamId	ID of home team
awayTeamId	ID of away team
shotClock	Number of seconds left on shot clock
timestamp	Time stamp provided for a frame, measured in milliseconds
frameIdx	Frame index of the game
ballX	X coordinate of the ball
ballY	Y coordinate of the ball
ballZ	Z coordinate of the ball
x	X coordinate of the player
y	Y coordinate of the player
z	Z coordinate of the player

Appendix B

Figures

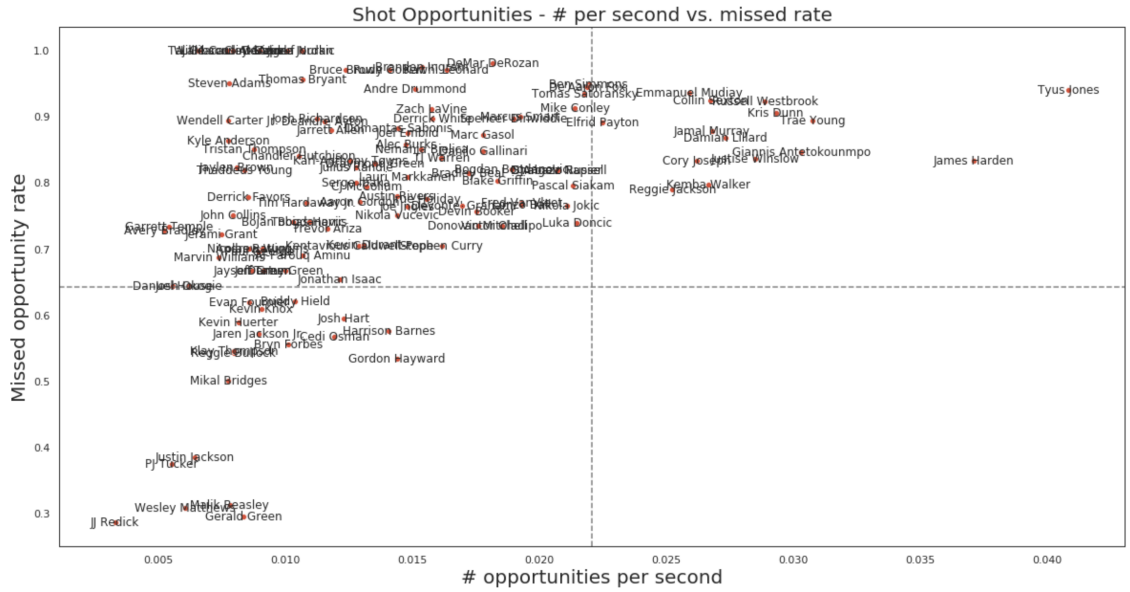

Figure B-1: Player cloud of shot opportunities per second vs. missed shot opportunity rate.

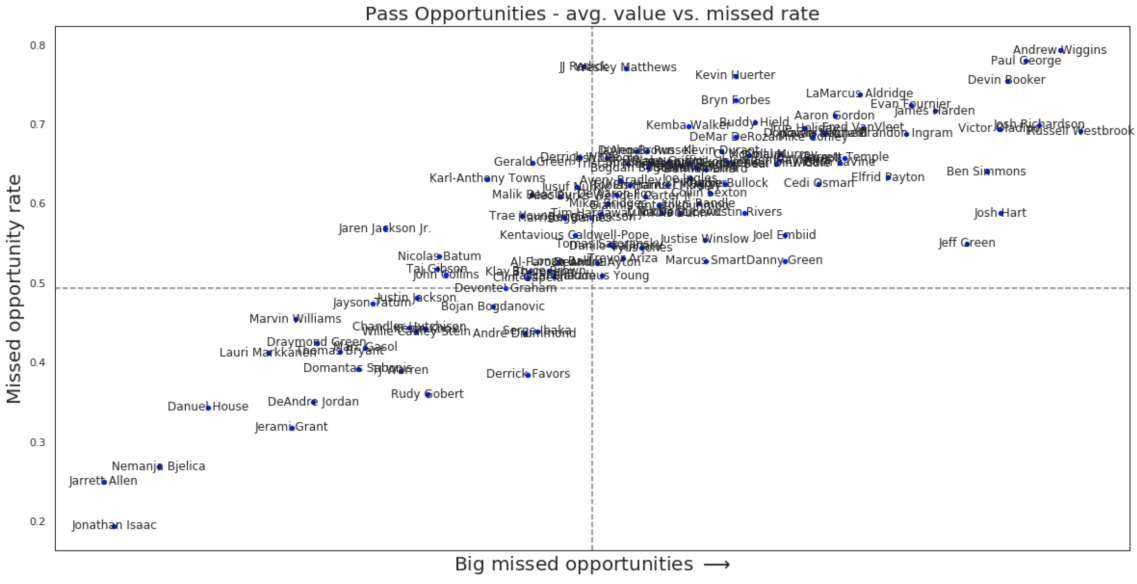

Figure B-2: Player cloud of pass opportunity values vs. missed pass opportunity rate.

Bibliography

[1] Chris Hayhurst. How NBA Data Analytics is Changing the Sport. *Dell Technologies, Transformative Leadership*, August 2020.

[2] Yu-Han Chang, Rajiv Maheswaran, Jeff Su, Sheldon Kwok, Tal Levy, Adam Wexler, and Kevin Squire. Quantifying Shot Quality in the NBA. *MIT Sloan Sports Analytics Conference*, February 2014.

[3] Kirk Goldsberry. CourtVision: New Visual and Spatial Analytics for the NBA. *MIT Sloan Sports Analytics Conference*, 2012.

[4] Rajiv Maheswaran. What advanced tracking data reveals about NBA shooters. *ESPN*, May 2012.

[5] Krishna Narsu. Introducing shot difficulty: Comparing game-winning shots in the playoffs. May 2015.

[6] Andrew Bocskocsky, John Ezekowitz, and Carolyn Stein. The Hot Hand: A new approach to an old "fallacy". *MIT Sloan Sports Analytics Conference*, February 2014.

[7] Karun Singh. Introducing Expected Threat (xT). Modelling team behaviour in possession to gain a deeper understanding of buildup play. May 2015.

[8] Aniruddha Bhandari. AUC-ROC curve in machine learning clearly explained. June 2020.

[9] Neal B. Gallagher. Savitzky-Golay smoothing and differentiation filter. *Eigenvector Research Incorporated.*

[10] Hsin-Ying Hsieh, Chieh-Yu Chen, Yu-Shuen Wang, and Jung-Hong Chuang. BasketballGAN: Generating basketball play simulation though sketching. October 2019.