Optimization of Daily Fantasy Basketball Lineups via Machine Learning

James Earl

A Senior Thesis submitted in partial fulfillment
of the requirements for graduation
in the Honors Program
Liberty University
Spring 2019

Acceptance of Senior Honors Thesis

This Senior Honors Thesis is accepted in partial fulfillment of the requirements for graduation from the Honors Program of Liberty University.

_____

Dr. Mark Merry, Ph.D.
Thesis Chair

_____

Dr. Robert Tucker, Ph.D.
Committee Member

_____

Dr. Tim Van Voorhis, Ph.D.
Committee Member

_____

Dr. David Schweitzer, Ph.D.
Assistant Honors Director

_____

Date

Abstract

Machine learning is providing a way to glean never before known insights from the data that gets recorded every day. This paper examines the application of machine learning to the novel field of Daily Fantasy Basketball. The particularities of the fantasy basketball ruleset and playstyle are discussed, and then the results of a data science case study are reviewed. The data set consists of player performance statistics as well as Fantasy Points, implied team total, DvP, and player status. The end goal is to evaluate how accurately the computer can predict a player's fantasy performance based off a chosen feature set, selection algorithm, and probabilistic methods.

Optimization of Daily Fantasy Basketball Lineups via Machine Learning

Rules and Regulations

The world of sports has become an enormous section of the entertainment market; with billions of dollars being made each year, individuals and startups are constantly looking for new ways to profit in the business. One of the newest and fastest growing of these markets is fantasy sports. More than simply watching a game, playing fantasy sports allows customers to show off their skills and knowledge of the game by picking the sports players that the individual believes will perform best in games. These players score points based on specific rule sets laid out by the fantasy sports companies. At the end of the game, day, or season, depending on the league, the best lineups win a prize. Some leagues are free; the only prize the winners receive is the envy of their friends. Other leagues require the customer to buy-in and can payout quite a lot of money. Due to the inevitable randomness of player performance, issue has been raised that paid leagues are very similar to gambling, but the vast majority of states, including Virginia, have classified fantasy sports as a game of skill. In support of this claim, an elite group of professional fantasy sports players has arisen comprised of those who have devised ways to consistently profit despite the randomness of the game. The goal of this paper is to attempt to discover the nature of these strategies, however, to make any sort of sense a discussion of daily fantasy basketball must be undertaken.

Keeping Score

The game of basketball is one of the simplest contests in all of professional sports. At its core, two teams comprised of five men each attempt to score points by getting the

ball into the other team's basket while simultaneously preventing the other team from scoring on their basket. A relatively small number of statistics detailing a player's in-game performance are recorded, but they hold tremendous importance to the world of fantasy basketball. The most important is the field goal: this statistic defines how many times the player scored points either via a traditional two-point shot, a three-point shot, or free throws with each specific type recorded separately. (NBA, n.d.) Naturally though, the ball does not always fall well upon the hoop. When players miss a shot, it provides an opportunity for a rebound, the successful retrieval of the missed shot. Rebounds are broken down as either offensive, when the shooting team recovers their own miss, or defensive, when the defending team takes possession of the ball. Another common statistic is the assist, which is credited to a player who passes the ball to a teammate who scores within a small amount of time. This statistic attempts to credit teammates who facilitate scoring for another. These comprise all of the offensive statistics, however, there are also two important defensive statistics that are tracked: steals and blocks. First, a steal results when a defending player gains possession of the basketball from the opposing team in any way except from a missed shot, which would instead be a rebound. Players who record a high number of steals are known to be aggressive defenders who pressure opponents into making mistakes. Second, a block results when a defending player is able to deflect any shot attempt by the opposing team. There is one other important metric, but, unlike the rest of these stats, it is an indicator of poor performance. Turnovers result when the offensive team loses possession of the ball for any reason outside of a shot attempt. Outside of these few trackers, there are more advanced

statistics, but these play the largest role in the game of fantasy basketball because they result in fantasy points.

A Fantastic Affair

Fantasy points are a company specific statistic that daily fantasy businesses use to quantify player performance. Instead of tracking the exact number of points, rebounds, or assists players get, each of these statistics is translated into fantasy points that are credited to the player for each corresponding action. The exact conversion is slightly different from company to company. All of the play performed for this study was carried out on the daily fantasy website FanDuel so their proprietary conversion will be used. Each actual point scored from field goals is worth one fantasy point, meaning a two-pointer is two, a three-pointer is three, and free throws are one each. (FanDuel, n.d.) Rebounds are worth 1.2 fantasy points and there is no differentiation between offensive or defensive rebounds. Assists are worth 1.5 fantasy points each making them nearly as valuable as made shots. Steals and Blocks are each worth three fantasy points and are tied with three-pointers as the most valuable fantasy action. Finally, turnovers actually take fantasy points away from the player who commits them. Each turnover costs the player one fantasy point. These six statistics are the only actions that result in a change of fantasy points.

Lineup Construction

Now that the basis for scoring fantasy points is understood, the method of play can be more easily explained. Entrants on daily fantasy basketball websites compete by creating lineups of players within a given salary constraint. The player base to choose

from consists exclusively of all players scheduled to participate in a game on the current calendar day. Customers create lineups by "buying" players for their rosters. Players have a personalized cost, and the total lineup cost must not exceed an amount specified by the daily fantasy company. FanDuel sets the total lineup cost at $60000. (FanDuel, n.d.) This amount does not represent real cash but just the available pool of "money" the customer can use to create a lineup. All participants in a daily fantasy contest must ensure that their lineups are under this salary constraint.

Created lineups, as previously mentioned, must remain at or below the salary cap, and they must also be comprised of a certain type and number of players. Basketball teams consist of different positions that perform different roles, and, on daily fantasy basketball sites, players are broken into categories based on the position they typically play. There are five positions: point guard, shooting guard, small forward, power forward, and center. On FanDuel, valid lineups consist of nine players: two point guards, two shooting guards, two small forwards, two power forwards, and one center. (FanDuel, n.d.) No other combination of positions can be played, and players are only available at one position. This means that Giannis Antetokounmpo, an all-star small/power forward hybrid, is only available for play at one position (traditionally small forward) even though he actually plays multiple positions throughout a basketball game. Given these constraints, it is usually impossible to create a lineup from just the best players on each day. Star players that produce large numbers of fantasy points cost significantly more than bench players who receive far fewer minutes of playtime. In addition, the distribution of high performing players is usually not uniform across all positions. Often,

the point guard position will have numerous highly priced players, while power forward may only consist of weaker players. Thus, to create a winning lineup, the goal is not only to maximize fantasy points, but to maximize fantasy points per average cost of player. This metric will be referred to as a player's "value" from now on. FanDuel prices players based on their expected fantasy points using previous performance to create their predictions. Generally, players are priced so that their value is $200 per point. For example, a player like Kevon Looney who traditionally produces 15 to 20 fantasy points will be priced between $3,500-$4,000 while James Harden, a star shooting guard who averages nearly 60 fantasy points, would be priced close to or above $12,000. Understanding this, it makes little sense to roster a highly priced player simply because he is highly priced; the player must also "make good" on his price by producing numerous fantasy points to be a good play. Similarly, cheaper players can also be good plays even if they produce less than star players as long as their value is high. The last point to mention concerning lineup construction is that FanDuel adjusts player price after each game; this provides a semi-accurate reflection of players' current performance. (FanDuel, n.d.) If a player is priced at $4,000 and scores 30 fantasy points, he will definitely cost close to $5,000 the next time he plays. Player price tends to rise quickly and fall slowly making it important to predict high value performances and then to resist chasing their possible reoccurrence.

A Friendly Contest

With the discussion of the scoring and lineup construction systems completed, the final piece of fantasy basketball that needs explanation is the types of contests run on

FanDuel. There are many different contest structures in which customers can participate. These contests range from top heavy tournaments to more balanced game formats. The end goal of all these contests is the same: to construct a lineup that will score more fantasy points than your opponents. However, the strategy and specifics of play vary greatly between these game formats. The play carried out for this experiment was mainly performed in two types of contests: 50/50s and Guaranteed Prize Pools, henceforth shortened to GPPs or tournaments.

First, GPPs have a tournament style structure and tend to consist of many competitors, usually upwards of 10,000 unique entries. These contests payout to the top scoring 20-25% of competitors and the prizes are broken into numerous brackets. The highest scoring lineup wins a tremendous prize ranging from $25,000-$100,000. (FanDuel, n.d.) The lineups immediately below the winner receive a smaller, but still generous, payout from around $5,000-$10,000. Outside the top ten, lineups are aggregated into prize brackets with the top .5-1% scoring a certain number, the top 5-10% scoring another, and so on to the bottom of the winning percentage. The remaining players that lie below the cutline do not win anything. Due to the tremendous numbers of entries, tournaments are often extremely competitive and only a near perfect lineup can take first place. Even netting any prize money requires scoring in the top 20-25% of entries, making it rather difficult to win any money at all. However, unique to GPPs is the ability to enter multiple lineups into one contest. (FanDuel, n.d.) This requires paying an entry fee for each additional lineup, but it enables competitors to cover a wider player base and more unique combinations than a single entry allows. For serious competitors,

this is the only way to play tournaments. In fact, it is likely that the only contenders who win first place are those that play multiple entries. The maximum number of entries that can be played is proportional to the size of the contest, but for larger contests the number cannot exceed 150. Playing 150 entries can require a significant amount of money if the entry fee per lineup is high. FanDuel does run cheaper contests though with entry fees as low as 25 cents. For this experiment, the majority of tournament play was conducted in these cheaper contests, and the maximum of 150 lineups would be played each day.

The other major contest type that was researched for this study is the 50/50. (FanDuel, n.d.) This contest structure is much more straightforward and simpler than tournaments. Competitors play against each other in groups of 10-100 players. The top half of lineups win prize money and the bottom half lose everything. In these contests, the goal is to simply be better than 50% of players as there are no prize brackets. Every player, whether 1st or 50th, wins 1.8 times their entry fee. This provides a natural opportunity to talk about the difficulty of being profitable in daily fantasy basketball. The host company takes a cut of all money played on the site. On FanDuel, this "rake" is 10% for 50/50s and a similar number for tournaments. To be profitable in 50/50s, contenders must win more than 50% of the contests they participate in to overcome the rake. In fact, to merely break even, competitors must win 55.5% of the time in 50/50s. Due to this rake, 50/50s are a very competitive contest structure. Participants tend to play more safely than in tournaments as there is a high cost to losing and no benefit between placing 1st or 50th outside of personal pride.

To review, the important pieces of daily fantasy basketball are that entrants create lineups of nine players: two point guards, two shooting guards, two small forwards, two power forwards, and a center. Players score fantasy points based on in-game actions, and the highest lineups win a prize. Now, the playstyle required for each contest type will be discussed further.

Game Theory

This section details the theory behind playing smart daily fantasy basketball. It is important to have a thorough understanding of the problem one wishes to solve before entering into machine learning applications. Otherwise, the model will most likely produce errant results because the user did not provide a representative dataset. The following are theories that were developed for playing in both 50/50s and tournaments. Each will be discussed to provide a more precise understanding of the complicated challenge in playing daily fantasy sports.

GPPs

First, developing a successful strategy to play in GPPs is a challenge itself. For this experiment, the goal of playing in tournaments was simply to win first. To achieve this goal, the researcher determined strategies to cover the diverse player base, stay in tune with player injury and availability news, and create lineups in a timely manner before the day's basketball games began. The basic strategy behind all of this remains simple: pick the best players that will result in the best lineup.

Due to the large number of players and inherent randomness of the sport, it is foolish to think one can easily pick the nine best players. Yet, it is plausible to pick a

couple of the best players, about two to four. However, it is still hard to predict player performance without some intervening circumstances. In daily fantasy basketball, the easiest way to pick good players is to roster the replacement for an injured player, although this is not always possible. Often, injuries are announced before the basketball game begins but after FanDuel has opened entry into their contests. This means that FanDuel is unable to update pricing for these players which often results in a value opportunity. When a player is healthy, his backup will often not get much playtime, but when the starter is injured, the backup has an opportunity for more minutes and thereby more chances to score fantasy points. Backups receiving the starting nod are not guaranteed to be the best plays of the day, but they often are and remain some of the most predictable value plays in daily fantasy sports. Another common way to identify the best player at a position is to recognize when there is a large gap between the most expensive player at a position and the next available player. For example, there are only a few superstar power forwards in the league, and there is no one else in the league like Anthony Davis. Davis has consistently been the highest priced player at the power forward position if not the highest priced of the entire day due to tremendous performances ranging from 65-80 fantasy points. Often, this results in a large gap between his price and the next available player, usually a disparity of $4,000-$5,000. The salary gap indicates that there is expected to be a large difference between the number of fantasy points Davis will score and the number the next available player will produce. If Davis performs well, he will almost undoubtedly be a necessary piece in the best lineups due to the massive differential in points between him and any other player at

the position. Even if this expensive player does not provide great value (salary/fantasy points), the raw number of fantasy points he produces solidifies him as a viable play. Usually, value at other positions will also alleviate any value deficiencies this raw point producer may have.

Once a couple great plays have been decided upon, whether due to injury or price disparity, it is necessary to "lock" these players into the lineups, meaning that these 3-4 players are rostered in every unique entry for the tournament. The point of locking players is to reasonably cover the player base in a limited number of lineups. By locking a couple players in every lineup, an entrant effectively only needs to construct lineups from 5-6 players instead of 9 which considerably reduces the possible number of combinations. The downside of doing so is that the 3-4 locked players must perform well if an entrant is to make any money. More than that, if a tournament is going to be won, those players must be the best plays at their respective position to overcome the high level of competition in GPPs.

After selecting these players, it is still necessary to select the remaining players to finish the lineups. The process to do this up until this point has been mostly through the researcher's personal analysis. 6-8 players are selected at each position to create player pools. These players are chosen based upon who is believed to have the best opportunity to perform well. The specific criteria that leads to selection is often hard to determine and provides an opportunity for machine learning to streamline this process. Often, it is difficult to be consistent in this selection process without unequally considering different aspects of performance as more important than others. By implementing this selection

process using machine learning, the model could make balanced, consistent decisions using a selected feature set, algorithm, and learning from previous data.

One way or another, player pools are created and need to be combined with the "locked" players to finalize the lineups. Some simple code was developed to loop through each pool and create valid lineups that stay within the salary constraints of the contest. Automating the process of lineup creation in some shape or form is necessary for anyone hoping to play multiple entries in GPPs. It takes too long to create lineups by hand to be able to respond to the numerous announcements that occur throughout the day with any sort of quickness. Thankfully, FanDuel allows lineups to be edited via .csv files, a basic file type used to store data in a columnar form, which can be easily produced with a small amount of coding knowledge. In theory, if the "locked" players are correct and enough of the players from the pools perform, the result will be a winning lineup. From the researcher's past experience, it is actually possible to win tournaments. A GPP consisting of 70,000 entries was won by using this strategy of locking players and picking the remaining spots using player pools.

50/50s

Playing in 50/50s is significantly different than playing in GPPs. For one, entrants are only allowed one lineup per contest meaning there is no need for the formation of pools or a computer program to expedite the creation process. (FanDuel, n.d.) Linked to this single-entry requirement, it is much more necessary to be correct about plays. Even one poor performance by a selected player can result in a loss for the night. The general strategy is to pick the most consistent plays in addition to a couple

high value plays. The high value plays can be picked exactly like in tournaments by using injury news and price disparity to guide one's choices. However, repeatedly picking the consistent players ends up being quite the challenge. This may seem to conflict with the very definition of consistency, but it is a realistic result of having to pick 5 or 6 players each night. As mentioned before, winning in 50/50s awards the participant 1.8 times the entry fee. Thus, it is not profitable to win one day and lose the next. To drive home the level of precision required to profit in 50/50s, consider a three-day span of fantasy play. A competitor would need to win a minimum of 2 out of these 3 days to overcome the rake and make any money. Across this three-day span, the participant will have to be correct on 15/18 or a similar fraction of the players he or she selects. Thus, it is necessary to be extremely accurate in one's selection of players to profit. The selection process is similar to the formation of player pools for GPPs; it has been largely based on the researcher's personal analysis. Similarly, it suffers from human inconsistency, failure to identify favorable scenarios, and general error. The selection of these 5-6 players provides another great opportunity for a more consistent machine to make informed decisions to eliminate the human error.

A Data Driven Study

A machine learning experiment consists of four important steps. One, a dataset needs to be acquired that relates to the problem at hand. (Alpaydin, 2014) The dataset should consist of numerous predictive variables and a variable that the data scientist desires to predict. Two, relevant features must be determined from that dataset to tighten the focus of the machine learning model. Features can be considered the most predictive

variables from the dataset. Selecting features involves culling the dataset by eliminating variables that are highly correlated or have no influence on the variable to be predicted. Third, models must be selected and trained using the previously determined features. The most commonly used models perform linear regression, but other options are also popular. Finally, the results of this training must be collected and analyzed to determine the nature and accuracy of the relationship between the data and the predicted variable.

Acquisition of Data

The dataset for this experiment consisted of performance statistics from every player for the 2017-2018 NBA season. This includes all the performance-based metrics: points, rebounds, assists, steals, blocks, turnovers, and fantasy points for every game. Some other statistics were also recorded including the player's status for the day which gives an indication of whether the player was injured, a backup, or a starter. The game's implied total score and defense versus position (DvP) was also recorded. DvP is a statistic that was created for the experiment by the researcher that indicates how well the opposing team defends each position. In theory, players with a favorable DvP matchup are more likely to have a large fantasy performance than if the matchup was poor. The data set resulted in about 32,000 rows and 80 observations per player.

In its original state, the data was not usable for the machine learning project. Of course, the machine could learn from the current observations, and, indeed, it would most likely come to extremely accurate conclusions. Yet, these conclusions would yield no applications for this experiment. The main issue is that performance-based statistics cannot be known until the basketball game has been played out. This is too late to be any

help for daily fantasy contests which lock before even a single game begins. Just as an entrant into a fantasy contest is tasked with predicting player performance, the machine learning algorithm has to work with predicted data for the current day. The dataset had to be mutated to fit a prediction algorithm. This was achieved by using a rolling average to predict future performance. To calculate data for the current day, a player's past seven performances were averaged together and then recorded. This process was repeated for each game to produce the predicted dataset. Due to the notion of a rolling average, the predicted dataset reflects a player's recent performance without getting bogged down by past failures or inflated by distant successes. The first seven performances were dropped for each player as it would be impossible to compute a predicted value for these observations since seven games had not yet been played. Thankfully, the other three metrics, player status, implied total, and DvP were all known before games began and did not need to be predicted.

Feature Selection

After a usable dataset had been acquired, the most predictive variables needed to be pulled out to sharpen the focus of the model. (Alpaydin, 2014) In the case of this dataset, the features had already been selected so that the features and dataset agree, but some explanation is required for why these variables were chosen. The root of the issue is selecting variables that directly or indirectly predict the desired variable. (Shalev-Shwartz, 2014)

The directly predictive features are fairly easy to see in this experiment because they form a mathematical function. (Shalev-Shwartz, 2014) These are represented by the

performance-based statistics; a mathematic relationship exists because the desired variable, fantasy points, is calculated by multiplying scalar values to each performance statistic and then adding them together. If a player nets less field goals than average, he cannot gain the normal amount of fantasy points from this value so necessarily his fantasy points will be lower than an average performance.

The indirectly predictive features present more of a challenge. These features represent the variables that affect fantasy points by influencing the other directly predictive variables. An easy to understand example is player status. If a player is injured, then he will not play in the game so he cannot score points or record a rebound. Thus, the player will score zero fantasy points. The player's status does not contribute to the mathematical function that calculates fantasy points, but by forcing those direct variables to zero it indirectly affects fantasy points. The other indirectly predictive features include DvP and implied game total.

Training the Model

With features selected and a wealth of data to use, the researcher's next step was to train a machine learning model. Machine learning is typically performed in R or Python. For this experiment, the machine learning was performed using R, a specialized programming language that was designed to facilitate advanced statistics and machine learning. R has many tools to aid in machine learning, however, due to limits in the researcher's knowledge, one of the simplest was selected: the Caret package. Caret is a library in R that provides many different machine learning tools all unified under a simple API. To perform machine learning, the train function is used to select a model

which learns from the provided dataset. After learning from the data, the model then tests its knowledge of the relationship between the features and the predicted variable by predicting the variable using test data. Test data is simply a portion of the data set that is held back from the training session. As long as the data is unified, the training and test data should reflect the same relationship between the features and the predicted variable leading to accurate predictions. However, if there is no meaningful relationship in the data, then testing the model will yield inaccurate results.

For this experiment, the desired variable to predict was fantasy points. However, predicting fantasy points precisely would be extremely difficult even for a machine. The researcher deemed that a classification would be a better metric that the computer might be able to determine. This classification should be based on fantasy points and allow more leniency than a precise numeric datapoint. The machine was tasked with deciding if a player was going to have a high or low fantasy performance based on the predicted dataset. The definitions of high and low were determined using the concept of value. If a player's predicted fantasy points divided by his adjusted FanDuel salary produced a value of 0.0 to 0.5, then the performance is classified as low. A value above .5 is classified as high.

Using this metric for successful prediction and the mutated dataset, the only missing piece was to select a model to use. Selecting the proper model for the experiment can be difficult without first getting some initial results. To determine the best model to use, the researcher found it appropriate to test a large number of types and so use the most accurate for the larger experiment. Initial training was performed on five

different popular machine learning models: Support Vector Machines, Random Forest Classification, Linear Discriminant Analysis, K-Nearest Neighbor, and Classification and Regression Training. One of the advantages of machine learning is that data scientists are not required to have a deep understanding of the mathematics that drive these models in order to use them. Instead, by learning how to use these tools effectively, data scientists can still come to meaningful conclusions. Of course, a more complete understanding of the model can lead to more informed decisions. Some investigation by the researcher was necessary to determine the best model to use for the larger dataset. After preliminary testing it seemed to the researcher that the Support Vector Machine model would be the best fit with an accuracy of 60%. However, using accuracy alone can be misleading. Machine learning packages in R not only provide an overview of a model's accuracy, but also allow the user to view the data values that were predicted or even predict further values using new data. Upon inspection of the predicted values, it became clear that some of the models were not learning meaningful information. The Support Vector Machine and K-Nearest Neighbor models were simply predicting that every performance would be low, and, since this outcome is more likely than being high, the models were "more accurate" than the others. Note, it is more likely to be low because value is the considered metric. Players do not often return good value because FanDuel constantly adjusts a player's price whenever he has a high performance and rarely lower the price even after repeated poor outings. Thus, it is more common to return low value. While these blanket predictions were technically more accurate than the others, the real problems in using this model reveal themselves when considering the

application of this knowledge. To make money playing daily fantasy basketball, it is necessary to be correct about plays that produce good value, or, in the context of the model, above average plays. Using the blanket below average predictions, the entrant would be inclined to never pick any player because the computer is informing him that no player would ever have a good game. This prediction is certainly not always the case and would lead the competitor to abandon the sport of daily fantasy completely. Therefore, such models had to be abandoned and others inspected. After reviewing predicted data for the remaining models, it became clear that the linear discriminate analysis was providing the most useful predictions despite being less accurate than some of the others. As such, the linear discriminate analysis model was used on the complete dataset.

Results

As may have been expected, the overall accuracy of the model's predictions was not incredibly high; the average of each player's individual predictions was a little above 59%. Some players were outliers on either side. For example, players like DeWayne Dedmon refused to be predicted while Tony Snell followed the model precisely. However, the goal of this experiment was not to predict every player's fantasy performance. As mentioned above, the idea was to predict the middle tier players that are always part of a winning lineup. Therefore, it would be more meaningful to calculate how accurately the model predicted these players' performances. When these players are isolated, the model's accuracy was still 59%. This was not different than the overall prediction for all players. The results still suggested that there exists a meaningful relationship between the features and fantasy points, but it needs to be better determined.

Numbers are helpful for quantifying success, but a qualitative analysis could be more easily understood. Why was the model so inaccurate in general? An obvious explanation to at least some of the error is that the majority of the dataset had to be predicted. This was achieved via averaging a player's past 7 fantasy outings and seemed reasonable enough, but Daily Fantasy Basketball does not lend itself well to averages. It is a game of extremes with players having highlight nights and then returning to relative irrelevance for the majority of the season. By using an average to predict the data, outlier performances can pull the averages in either direction. This could lead to the model learning a relationship that does not exist. For example, a player who usually plays 5 minutes could play 20 for one game due to a mid-game substitution. The predicted data would be unaware of this implicit change in status and thus the average predicted minutes would be higher than normal for the next 7 predicted data values. Another fault of the model is that it attempts to force every player to fit the same archetype. By evaluating each player using the same features without personalization, the model treats the players as if they were more machine than man. Such a practice may be less detrimental in more correlated datasets, but basketball players are individuals. Many players are more inconsistent than others or are more affected by matchup than their counterparts. When the model does not allow for these personalized dependencies, it should be no surprise that its predictions have limited accuracy.

What improvements can be made then to enhance accuracy? One of the most important changes could be creating personalized models for each player. Due to the limited nature of this experiment, it was impossible to build custom datasets and find

unique features for each of the more than four hundred NBA players. However, this would surely lead to improved predictions. Another simpler solution could be to add more data. The dataset for this experiment consisted of a single season of NBA statistics. Therefore, no player could have more than eighty-two discrete observations as there are only eighty-two games played each season. Since this was a limited number of observations, it was difficult for the machine to detect subtle relationships that may only be reflected in singular data points. Accuracy could be improved by bringing in data from multiple NBA seasons. Finally, and maybe most obviously, the model could surely be improved by determining better features. The amount of data transformation that was performed to come up with this dataset was limited. A deeper exploration into the exact nature of the relationship between the features and the predicted variable could lead to more precise predictions overall. Beyond transforming the dataset, simply bringing in a new, possibly ignored feature might lead to better prediction. Conversely, removing one of the provided features could lead to improvements due to features being closely correlated to one another. Other features could even be non-factors that actually hinder the learning of the model. These features may not even affect the predicted variable, but their presence in the dataset misleads the model into attempting to learn from them.

Applying the Results

Given this discussion, it seems reasonable to assume that the model could be improved with some effort. Yet, it would still be practical to apply what was learned to daily fantasy play. More specifically, this section will answer the question: how could

using these results help entrants to pick the 3 to 4 desired mid-tier players, and how would doing so affect play in GPPs and 50/50s?

To win a GPP, it is necessary to select the best possible plays on a given day. Due to the limits of this success condition, winning a GPP requires highly accurate predictions. The nature of these machine learning results does not lend itself well to playing in GPPs. The prediction variable was simply a classification value determining if the player's fantasy score would be a high or low value. While it is necessary for players in winning GPP lineups to score above average fantasy points, this is not usually enough. Players need to score far beyond average to place into the upper tiers of GPPs. Thus, the model's results do not provide enough specificity to be useful for GPP play. The experiment could be re-run with a more specific success case that is only true when the player is predicted to score far above average, but this produces another difficulty. A player scoring far above average is inherently an outlier value. Since the model is trying to maximize accuracy, it tends to ignore outliers since they, by definition, do not happen often. Thus, training the models would again result in blanket low predictions. This was confirmed by performing some testing, and all models yielded such predictions. A more fine-tuned model may be able to provide more useful insight by using a different metric for success or by forcing it to minimally predict a couple successful cases.

50/50s by comparison, require a much lower measure of precision to win. Due to this imprecision, there is no issue with using the high/low prediction. Even players who perform at average are not necessarily detrimental to the success of a lineup. Only poor performers hinder 50/50 lineups. The main goal is to avoid rostering bad players more so

than ensuring the entrant selects above average performers. With this in mind, it can be seen that having an excess amount of low predictions actually provides useful insight. These predictions mean the model is being safe. However, the model also needs to be correct on its high predictions to be useful for fantasy play. In general, the model was actually slightly more accurate when only high predictions were considered with an average of 62.2% for the mid-tier players. This means that, when the model predicted a player would have an above average performance, it was correct 62.2% of the time. Applying this to actual fantasy play, this 62.2% prediction rate implies that it could be possible to be profitable using these machine learning strategies in 50/50s. The prediction rate is higher than the 55.5% that must be won to be profitable in 50/50s. While this ignores the other process of selecting the other value plays, profitability could theoretically be achieved using this model. Surely, improving the model could lead to more precise predictions.

Conclusion

Daily Fantasy Basketball is a difficult sport that requires highly accurate predictions to achieve profitability. Applying machine learning to aid in player selection demonstrated that models can semi-accurately predict fantasy performance although such models are not highly precise. As machine learning technologies continue to be developed, improvements in modeling and better feature selection could lead to more and more precise predictions for fantasy performance. These improvements in modeling could lead to more fleshed out strategies being developed to be profitable in GPPs and

50/50s. The data definitely suggests that this is possible in 50/50s, but more work will need to be done to confirm this theory.

References

Albalate, A., & Minker, W. (2011). *Semi-supervised and unsupervised machine learning: Novel strategies*. London, UK: ISTE.

Albert, J., Bennett, J., & Cochran, J. J. (2008). *Anthology of statistics in sports*. Philadelphia, PA: Society for Industrial & Applied Mathematics.

Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge, MA: MIT Press.

Alpaydin, E. (2016). *Machine learning: The new AI*. Cambridge, MA: MIT Press.

Bell, J. (2015). *Machine learning: Hands-on for developers and technical professionals*. Indianapolis, IN: John Wiley & Sons.

FanDuel (n.d.). Rules & scoring. Retrieved March 6, 2019, from https://www.fanduel.com/rules

Hughes, M., & Franks, I. M. (2015). *Essentials of performance analysis in sport*. Abingdon, UK: Routledge.

Jaynes, E. T., & Bretthorst, G. L. (2017). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Kung, S. Y. (2014). *Kernel methods and machine learning*. Cambridge, UK: Cambridge University Press.

Lampropoulos, A. S., & Tsihrintzis, G. A. (2015). *Machine learning paradigms: Applications in recommender systems*. New York, NY: Springer International Publishing.

Liu, X., Datta, A., & Lim, E. (2015). *Computational trust models and machine learning*. Boca Raton, FL: CRC Press.

Marsland, S. (2008). *Introduction to machine learning: An algorithmic perspective*. Boca Raton, FL: CRC Press.

Mohammed, M., Khan, M. B., & Bashier, E. B. (2017). *Machine learning: Algorithms and applications*. Boca Raton, FL: CRC Press.

NBA (n.d.). Stat glossary. Retrieved March 6, 2019, from https://stats.nba.com/help/glossary

Rudas, T. (2005). *Probability theory: A primer*. Thousand Oaks, CA: Sage Publications.

Rudas, T. (2008). *Handbook of probability theory and applications*. Thousand Oaks, CA: Sage Publications.

Shalev-Shwartz, S. (2014). *Understanding machine learning: From foundations to algorithms*. Cambridge, UK: Cambridge University Press.

Sra, S., Nowozin, S., & Wright, S. J. (2012). *Optimization for machine learning*. Cambridge, MA: MIT Press.

Sugiyama, M. (2015). *Statistical reinforcement learning: Modern machine learning approaches*. Boca Raton, FL: CRC Press.

Winston, W. L. (2012). *Mathletics: How gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football*. Princeton, NJ: Princeton University Press.