Allegheny College
Department of Computer Science

Senior Thesis

ProPredict: Unlocking Sports Insights, Creating an Interactive Dashboard with Predictive Modeling Algorithms

by

Liam Black

ALLEGHENY COLLEGE

COMPUTER SCIENCE

Project Supervisor: **Dr. Oliver Bonham-Carter**
Co-Supervisor: **Professor Douglas Luman**

7 May 2023

Abstract

This study introduces an intuitive sports analysis tool, Propredict, with advanced features that offer an interactive experience for users. ProPredict is an ideal tool for both sports teams and fans seeking to conduct their own analyses of their favorite teams and players. By utilizing machine learning algorithms, ProPredict empowers users to gain valuable insights and stay ahead of the curve, enabling them to be proactive in their approach. The tool allows users to easily import their preferred datasets and generate graphs quickly to identify patterns in the data.

This thesis highlights the effectiveness of the tool in meeting its goal, providing a flexible interface that meets users' needs, while generating accurate predictions. It also sheds light on the challenges of predicting sports data, due to the complexity and external factors of sports data . The analyses show that perfect accuracy in sports data prediction is unattainable, but continued testing and tweaking of models can lead to improving results.

The study suggests that future work and testing could extend the tool's capabilities to include various sports data, leading to the development of more effective models that enable users to identify trends and answer underlying research queries. Overall, this tool is a valuable resource for sports enthusiasts and researchers, offering a versatile platform to conduct analyses and obtain findings for the future of sports.

Table of Contents

Introduction **4**
- Motivation . . . 4
- Current State of the Art . . . 5
- Goals of the Project . . . 6
- Empowering Users with Control: The Benefits of Personalized Analysis . . . 7
- Ethical Implications . . . 10

Related work **13**
- Video Analysis Software . . . 13
- Current State of NBA Analysis . . . 15
- Popular Dashboards . . . 18
- The Moneyball Approach . . . 19

Method of approach **22**
- Data Visualization with Plotly: Converting Data into Insights . . . 23

Experiments **27**
- Experimental design . . . 27
- Linear Regression Algorithm . . . 32
- Random Forest Regression Algorithm . . . 33
- Evaluation . . . 36
- Analyzing Predictions . . . 39
- Threats to Validity . . . 39
- Summary of Results . . . 42

Future work **43**
- Future Ethical implications . . . 44

Conclusions 45

References **47**

List of Figures

1 Phoenix Suns Dataset 7
2 Selecting the dataset 8
3 Displayed Team Data 8
4 Phoenix Suns Dataset 9

5 SportsCode tool 14
6 Player Analysis tools 15
7 Three point graphic 17
8 The method behind MoneyBall 20

9 Data Points 24
10 Strong Correlation 25

11 Experiment Dataset 28
12 Kevin Durant's 3-point percentage across career 29
13 Linear Regression Model 30
14 Random Forest Model 31
15 Linear Model Flowchart 32
16 Random Forest Flowchart 34
17 Experimental Results 36
18 PTS Results 38
19 Lebron James Points Throughout His Career 38

Introduction

Motivation

Each year the field of sports brings in new athletes that are stronger, faster, and more athletic than before, continuously evolving the game alongside the players. When looking at the most winning NBA teams in the past few decades, such as the San Antonio Spurs, whom from the 2000 season to 2017 set a record of 18 consecutive 50 win seasons, questions arise of what specific strategies are teams like the Spurs using to yield high winning seasons year in and year out, and how has the evolution of strategy changed over time to either continue or break a teams legacy. The key to answering questions like these is through data analysis.

The motivation for this project stems from how unpredictable sports games can be and how many different coaching styles and new players continue to change the game. Using a platform like `streamlit` with an interactive dashboard where users can select elements of a graph based on the data set is a way not only sports fans can get a deeper understanding of the game by exploring visual trends, but also allows coaches/analysts to create in-depth analysis' backed up by measurable conclusions to see behind the scenes of what statistics stand out for players and teams, and how can players and teams change the way they strategize. Data analyst's for team's can draw conclusions and tailor similar schemes to their team's strengths to win more games based on personalized research, using data analysis tools.[14]

The goal of this study is to demonstrate how analyzing statistical data and utilizing machine learning to train and test predictions can lead to improved player and team performance in sports. Not only will my project enable users to generate queries using a dataset of their choosing, allowing for the inclusion of many competitive sports, but they will also have the ability to visually analyze the data to identify trends. To accomplish this, I've

utilized Streamlit, which is a visualization platform that allows users to take a dataset and produce a variety of graphics such as pie charts, bar graphs, and line graphs. This platform allows users to view their query results and give them the ability to personalize their input by specifying which teams and statistics they want to display. ProPredict is flexible and empowers users to refine their original queries from a broad perspective to specific statistics. Using the Streamlit interface, researchers can analyze any point of data and run through graph after graph until every query is visually answered with a plot. To take it one step further, ProPredict features two machine learning algorithms to generate trained and tested predictions that go beyond the available input data.

ProPredict also uncovers the complex ethical implications behind the use case of sports data, and how analysis of sports data can be both beneficial and detrimental to the community of sports.

By using this tool a broad spectrum of sports fans, sports analysts, and sports teams can study and examine a range of data to uncover conclusions backed up by measurable statistics, such as R-squared values to enhance the overall performance of teams or players and unveil insights that go beyond numerical data. This research paper effectively shows how ProPredict culminates the effectiveness of data analysis for improved performance in sports, while also putting data analyzers in the driver seat, letting them choose the direction of their research.

Current State of the Art

Sports data has been recorded for quite some time now, and is a product of the games being played. With constant upgrades to artificial intelligence and the addition of high speed cameras, recording sports statistics have never been easier. Sites like `nba.com` and `basketball-reference.com` have easily accessible data for users to export and turn into their own datasets. Sports data is something anyone can look at and attempt to create their own predictions, either as a fan, or even as a coach looking at their player statistics to learn how to best change their lineups. For team statistics, with the evolution of databases that can be maintained efficiently, it has become significantly easier overtime connecting research on player statistics with team statistics.

While looking at player statistics is initially important, to understand how sports like basketball has evolved overtime and investigating how different

teams have excelled at scoring on different areas of the court, it is essential to start with a broad analysis, using all 32 NBA teams offensive scoring data, before diving into specific teams.

In the 2010's using data science and machine learning really shocked the world when Germany's world cup soccer team used high speed efficient algorithms to study player movement and touch tracking data(how long did each player touch the ball per possession) to drastically improve their efficiency on offense, better than any other team, and eventually win the world cup that year.

Similar to soccer, NBA teams use a variety of techniques to analyze data for team improvement, including technical and non-technical based approaches . Many NBA teams use advanced analytics software to comb through team data, allowing them to track and analyze a wide range of statistics, such as player performance, game outcomes, and team strategy. [5]

NBA teams and players talk about improvement during post-game and post-season reviews. After each game and at the end of the season, many NBA teams conduct reviews to assess their performance and identify areas for improvement. These reviews often involve presentations of statistical data and video footage so the players can gain a better understanding of what worked well and what didn't.

However, perhaps the most important way NBA teams can improve their performance is by having analytics experts on their staff who are responsible for looking through the team's data and providing statistical based insights and recommendations to the coaching staff and front office. [8].

These analysts use data visualization tools to display charts, graphs, and tables, to help them visualize and understand complex data sets. These tools can make it easier to identify trends and patterns in the data and communicate findings to others. ProPredict uses this strategy to connect large sets of data with graphics to formulate understandable analysis that results in improving changes in team performance.

Goals of the Project

The goal of this project is to create a data analysis tool, ProPredict, that ensures personalized and versatile research studies, while achieving a low cognitive load for users who may not have a coding background.

ProPredict is personalized because it allows the user to select their own dataset they want to investigate. When users have freedom to analyze any elements of a chosen dataset against each other in graphics, such as scatter plot's, they are able to get results that are significant to themselves or their team. These goals differ from other sports analysis tools because it allows for a positive user experience for all audiences.

ProPredict is broadcasted to the full spectrum of sports fans. Fans who may not know much about a sport have freedom to generate graphs and find broad trends across all teams in a league, their hometown team, or even their favorite players by just easily selecting X and Y variables from a drop down menu. Contrastingly, users with statistical backgrounds are able to investigate specific queries and look at r-squared values from the graphics to find measurable results. These same users are also able to take their research further and use the predictive models if they so choose, and possibly create their own broad, or specific research study.

Empowering Users with Control: The Benefits of Personalized Analysis

By reading through the steps below users can see how beneficial platforms designed for user experience can be.

After obtaining the desired dataset, for example from `nba.com`, they are ready to begin their own research. [12]

BY YEAR	GP	MIN	%FGA 2PT	%FGA 3PT	%PTS 2PT	%PTS 2PT MR	%PTS 3PT	%PTS FBPS	%PTS FT	%PTS OFF TO	%PTS PITP	2FGM %AST	2FGM %UAST	3FGM %AST	3FGM %UAST	FGM %AST	FGM %UAST
2020-21	72	48.6	60.8	39.2	52.5	13.2	34.0	11.3	13.5	13.9	39.3	52.4	47.6	84.9	15.1	62.2	37.8
2022-23	23	48.2	63.3	36.7	52.8	10.9	33.0	10.0	14.2	16.8	41.9	55.3	44.7	82.4	17.6	63.2	36.8
2019-20	73	48.2	63.9	36.1	52.4	10.4	30.1	12.9	17.5	16.2	42.0	57.4	42.6	88.9	11.1	66.1	33.9
2021-22	82	48.1	64.6	35.4	55.9	12.5	30.3	10.9	13.8	15.7	43.3	55.2	44.8	83.2	16.8	62.7	37.3
2018-19	82	48.5	66.5	33.5	56.7	10.4	26.9	14.1	16.4	17.0	46.3	52.8	47.2	80.8	19.2	59.5	40.5
2017-18	82	48.1	68.0	32.0	56.1	11.8	26.9	13.8	17.0	14.4	44.3	46.7	53.3	82.3	17.7	55.3	44.7
2015-16	82	48.1	69.8	30.2	55.9	14.9	26.8	14.3	17.3	16.4	41.1	47.0	53.0	83.2	16.8	55.8	44.2
2013-14	82	48.2	70.0	30.0	55.8	15.7	26.6	17.8	17.6	16.8	40.1	38.8	61.2	82.4	17.6	49.3	50.7
2005-06	82	48.7	70.7	29.3	58.4	23.8	28.3	14.8	13.4	15.7	34.5	56.9	43.1	84.0	16.0	63.5	36.5
2014-15	82	48.5	70.9	29.1	59.1	17.3	24.9	17.8	16.0	17.3	41.8	44.3	55.7	80.2	19.8	52.2	47.8
2004-05	82	48.2	71.1	28.9	56.4	17.3	26.4	16.7	17.2	14.9	39.1	48.3	51.7	86.9	13.1	57.5	42.5
2006-07	82	48.5	71.3	28.7	57.6	17.9	26.1	15.4	16.3	15.8	39.7	55.7	44.3	85.6	14.4	62.6	37.4

Figure 1: Phoenix Suns Dataset

Figure 1 above is an example dataset that the user can download to their computer to be able to use for the streamlit application. The user should download their desired `.csv` file. Once the program is run, the streamlit web browser will be opened to a local host, where the user will be prompted to either select the dataset from my repository or their own.

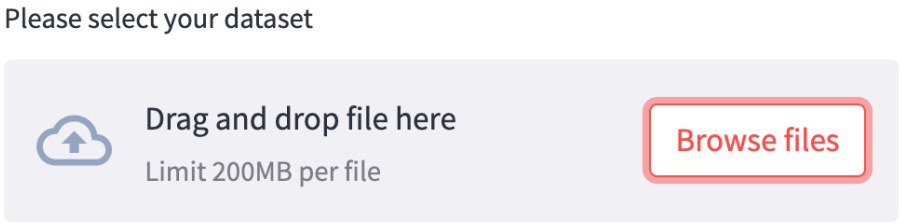
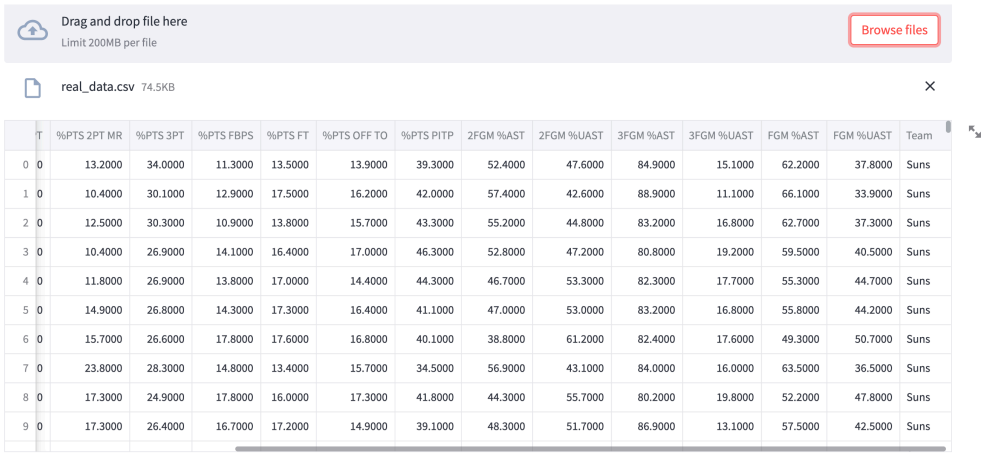


Figure 2: Selecting the dataset

Figure 2 displays what it looks like from the user end, when they select their dataset. After the datafile is selected the streamlit app will display the contents of the selected file.



	T	%PTS 2PT MR	%PTS 3PT	%PTS FBPS	%PTS FT	%PTS OFF TO	%PTS PITP	2FGM %AST	2FGM %UAST	3FGM %AST	3FGM %UAST	FGM %AST	FGM %UAST	Team
0	0	13.2000	34.0000	11.3000	13.5000	13.9000	39.3000	52.4000	47.6000	84.9000	15.1000	62.2000	37.8000	Suns
1	0	10.4000	30.1000	12.9000	17.5000	16.2000	42.0000	57.4000	42.6000	88.9000	11.1000	66.1000	33.9000	Suns
2	0	12.5000	30.3000	10.9000	13.8000	15.7000	43.3000	55.2000	44.8000	83.2000	16.8000	62.7000	37.3000	Suns
3	0	10.4000	26.9000	14.1000	16.4000	17.0000	46.3000	52.8000	47.2000	80.8000	19.2000	59.5000	40.5000	Suns
4	0	11.8000	26.9000	13.8000	17.0000	14.4000	44.3000	46.7000	53.3000	82.3000	17.7000	55.3000	44.7000	Suns
5	0	14.9000	26.8000	14.3000	17.3000	16.4000	41.1000	47.0000	53.0000	83.2000	16.8000	55.8000	44.2000	Suns
6	0	15.7000	26.6000	17.8000	17.6000	16.8000	40.1000	38.8000	61.2000	82.4000	17.6000	49.3000	50.7000	Suns
7	0	23.8000	28.3000	14.8000	13.4000	15.7000	34.5000	56.9000	43.1000	84.0000	16.0000	63.5000	36.5000	Suns
8	0	17.3000	24.9000	17.8000	16.0000	17.3000	41.8000	44.3000	55.7000	80.2000	19.8000	52.2000	47.8000	Suns
9	0	17.3000	26.4000	16.7000	17.2000	14.9000	39.1000	48.3000	51.7000	86.9000	13.1000	57.5000	42.5000	Suns

Figure 3: Displayed Team Data

Figure 3 shows how the dataset will be shown to the user, and they will be able to scroll and navigate to any part of the dataset. The next step for the user is to choose the elements they want displayed on the axis of the graph and what type of graph they would like to display their results on.

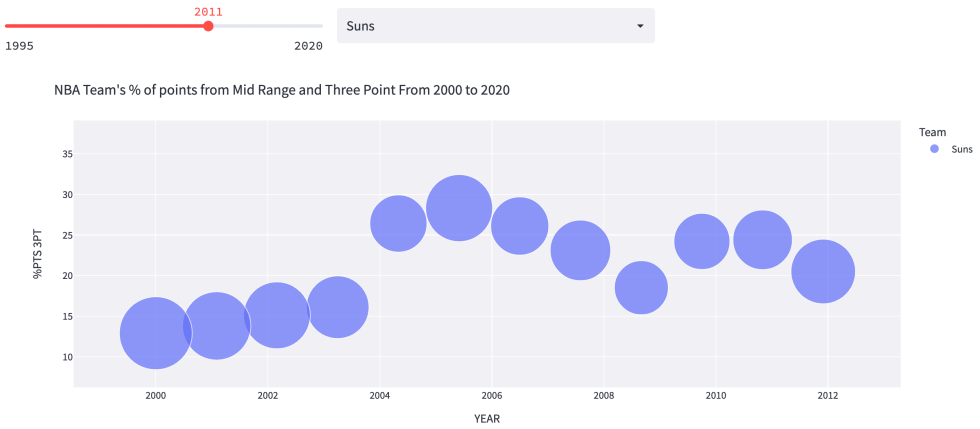


Figure 4: Phoenix Suns Dataset

Figure 4 shows the Phoenix Suns percentage of points off of three point field goals from the year 2000 to 2012. The next phase of this project is the analysis, where anyone can investigate any part of the dataset.

ProPredict's interface allows for a versatile analysis of sports team/player data. This tool allows users to create their own narrative for the study, should they choose so.

A narrative a team analyst or a coach would look at might be enhancing team performance. By analyzing data on statistics like shooting percentage, turnovers, and player efficiency, teams can identify areas where they need to improve.

A simple experiment that reflects this for basketball teams is looking at the effect of three point percentage and total wins for a season. After generating a graph comparing 3-point percentage and total wins in a season for all teams the coach notices that his team is underperforming when compared to other teams and they may want to focus on improving their shooting technique or incorporating more 3-point attempts into their practices, thus improving their shooting from behind the arc in games.

Coaches and sports analysts can discover trends in their teams to develop strategies tailored to the strengths and weaknesses of their players. Teams can then create a game plan backed up by significant data. This approach is a way for coaches to investigate weaknesses in their teams offense, defense and bring actionable change that is backed up by evidence.

To take it one step further teams can stay ahead of the curve by engaging with the predictive models. The two predictive models are trained on the data from the chosen variables graphed, and then depending on the model chosen will generate a prediction for a specific date or time. For example, a team who is just entering the offseason would follow the approach of selecting the prediction to be for the next season to get the best results.

The goal of the linear regression and random forest predictive models is to allow users to discover future insights and trends for the league, for teams, and for players. Not only does this positively affect player and team performance analysis, by seeing how likely they are to perform based on historical data, it also serves to enrich the fan experience allowing them to gain a deeper understanding of the data.

To recap, the main goals of the project are to provide a platform for coaches, fans, and experts to conduct research studies by using specific elements of data to generate graphs to analyze trends. From this point users can engage with linear and random forest regression models to stay ahead of the curve by discovering machine learning trained predictions for their data.

Overall, ProPredict uses data storage, visualizations, and machine learning predictions to meet these goals. ProPredicts dynamic interface allows users to conduct motivating research studies while keeping a low level cognitive load to broadcast the tool to all sport audiences.[10]

Ethical Implications

When conducting any research study, it is important to recognize and implement safety measures to address the ethical implications of the study. Information privacy and accuracy are two crucial aspects to consider. In this study, the user is simply selecting a CSV file to import into Streamlit to create their visualizations, so the ethical implications of this study are focused on the input data and the predictive models.

While the information selected by the user deals with sport statistics, specifically team statistics, it is important to consider the privacy of the

information. In most cases, sports statistics are not considered private information as they are nationally televised and recorded daily. However, it is still important to consider any privacy implications that may arise, such as revealing sensitive personal information or confidential team strategies.

In terms of information accuracy, the credibility and reliability of the data source should be considered. Websites such as 'nba.com' and 'basketball-reference.com' are highly reputable sources that provide up-to-date and accurate information. However, it is still important to verify the accuracy of the dataset the user inputs, as errors or inconsistencies can still occur. The data is readily available on numerous basketball sites, but not all of them have options to directly export the data into CSV or TSV formats. In these cases, manual extraction and formatting of the data or web-scraping may be required to convert the data into a readable csv file.

With the new age of sports data and data science-related studies, there has been much debate among analysts and experts about whether performance-enhancing analysis of individual players or team play is considered unfair or even cheating.

Some experts and analysts argue in favor of performance-enhancing data analysis, stating that it is part of the game and simply involves finding measurable trends in data that have the highest chance of yielding wins. They view it as using abstract interfaces to make sense of the vast amounts of data that are already available and using it to find the best strategies for coaches to use. This is where visualization tools such as Streamlit can come in handy - by making complex data more accessible to coaches, they can make better-informed decisions and improve the game.

On the other side, expert analysts such as Charles Barkley identify advanced statistical analysis as making the game boring and taking away the versatility of basketball. While both sides of the argument have valid points regarding the ethics of such studies, it is essential as a researcher to base conclusions solely on data with accurate and measurable results, such as those obtained through linear regression analysis. It is important to supply these results to individuals such as coaches with an emphasis on how these results can lead to improved team performance in a logical manner, so that they can make decisions backed up by the analysis. [6]

Additionally, it is important to note that the use of data and advanced analytics in sports is not a new phenomenon. Teams have been using data to gain an edge for decades, whether it's by scouting opponents or analyzing player performance. The use of data and analytics is simply becoming more

prevalent and sophisticated with advancements in technology. Data is a product of the game, which makes it up to individual teams and leagues to determine how they want to use the data and whether they consider it to be an unfair advantage.

Related work

Traditionally, sports performance enhancing analysis has been defined as an observational analysis task that goes from data collection to the delivery of evidence based feedback. It aims to improve sports performance by involving coaches, players, and analysts themselves. The process starts with data collection, which in the context of sports, doesn't just consist of basic statistics like points and assists, but can also include advanced statistics such as player speed values, heart rate, etc. This advanced data can provide valuable insights into a player's performance beyond what traditional statistics reveal.

Video Analysis Software

To collect and analyze this data, specialized software such as SportsCode have been developed to make it easy for coaches and analysts to develop statistical reports that are easy to dissect. According to Arastey (2019), SportsCode is a popular software in the field of sports performance analysis that allows coaches and analysts to record, replay, and analyze games. The software has a user-friendly interface that simplifies the process of data collection and analysis, and can provide real-time feedback to players and coaches.[2]



Figure 5: SportsCode tool

Sportscode(fig. 5) is an analysis tool that doesn't just conduct regression analysis on numeric values but conducts video analysis instead. Tools like Sportscode take a play-by-play analysis approach of sports footage and create a platform where analysts and coaches can visually identify what went well and what can be improved in either a training session or game by providing quick and easy ways for users to interact with the tool in their own way(Arastey, 2020).

While traditional sports statistics like points and assists have always been used to evaluate player and team performance, advanced statistics provide a more comprehensive understanding of the game.

There has been some controversy surrounding the use of advanced analytics in basketball, with some experts, like Charles Barkley, claiming that it takes away from the natural beauty of the game. Despite this, the use of advanced statistics in basketball has become increasingly popular in recent years, with many teams and analysts relying on it to gain a competitive advantage.

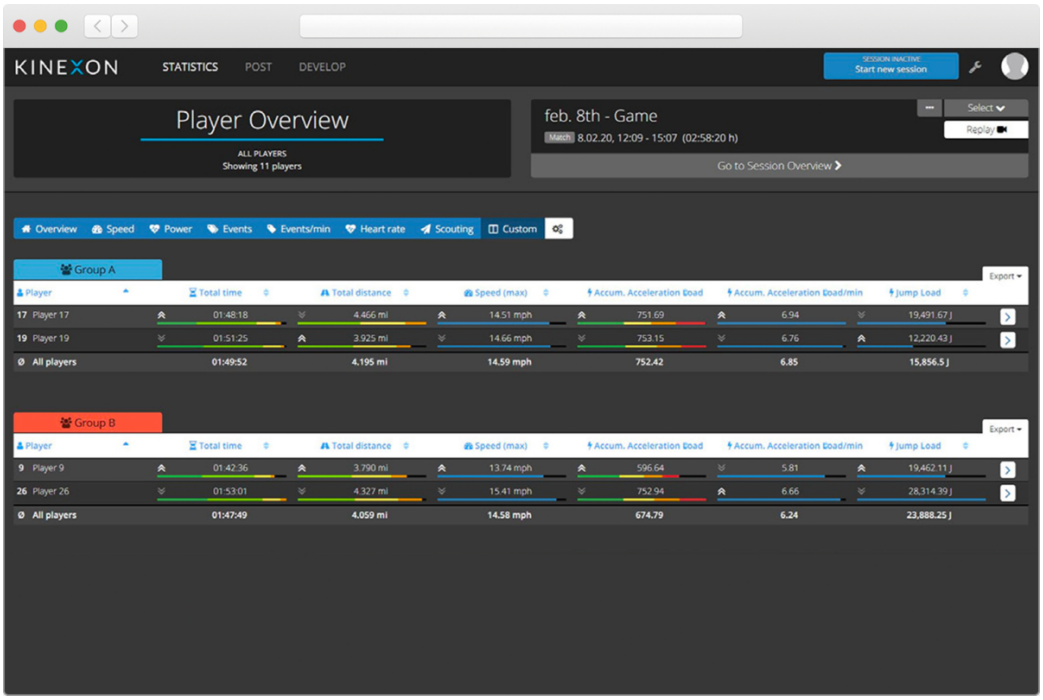


Figure 6: Player Analysis tools

Analysis software doesn't just have to contain graphics, Figure 6's graph displays a version of a soccer analysis tool without graphical representation. [3]

Current State of NBA Analysis

Sports analyst software is all about creating a platform that takes in sports data and allows the user to interact with key data points to visualize and understand significant correlations between those data points. Analysis tools turn a giant table of numbers into a story.

One big example of a popular topic to research with NBA data is analyzing how offensive strategies have changed over time.

NBA statistics and play styles have changed overtime for a multitude of reasons consisting of rule changes, the addition of more and more international players, and so much more. To give context for this research study of

analyzing basketball strategy throughout the decades we need to take note of important moments in the past 40 years of the NBA.

The biggest change the NBA has seen was the introduction of the three-point shot in 1979. The NBA(National Basketball Association) combined with the failing ABA(American Basketball Association) and with it came the three point shot. Numerous accounts of disappointment as many coaches and players at the time felt as if it was like "cheating". Famed coach and former player Gregg Popovich claims "To me, it's not basketball, but you got to use it. If you don't use it, you're in big trouble. But you sorta feel like it's cheating… I still hate it."

Even all-time great three point shooter Larry Bird commented on the manner,"Back when I played, we just didn't shoot it that much. Now, if you're not firing up thirty 3's, you're just not playing basketball." The initial addition of the three point shot may not have gone over as smoothly as the NBA would've liked, but it was a part of the game and players and coaches had to accept it and change their playstyle to fit with it. The addition of the three pointer was a drastic change that not only changed the style and even build of players in the league, but made the league enter an new era of scoring statistics.

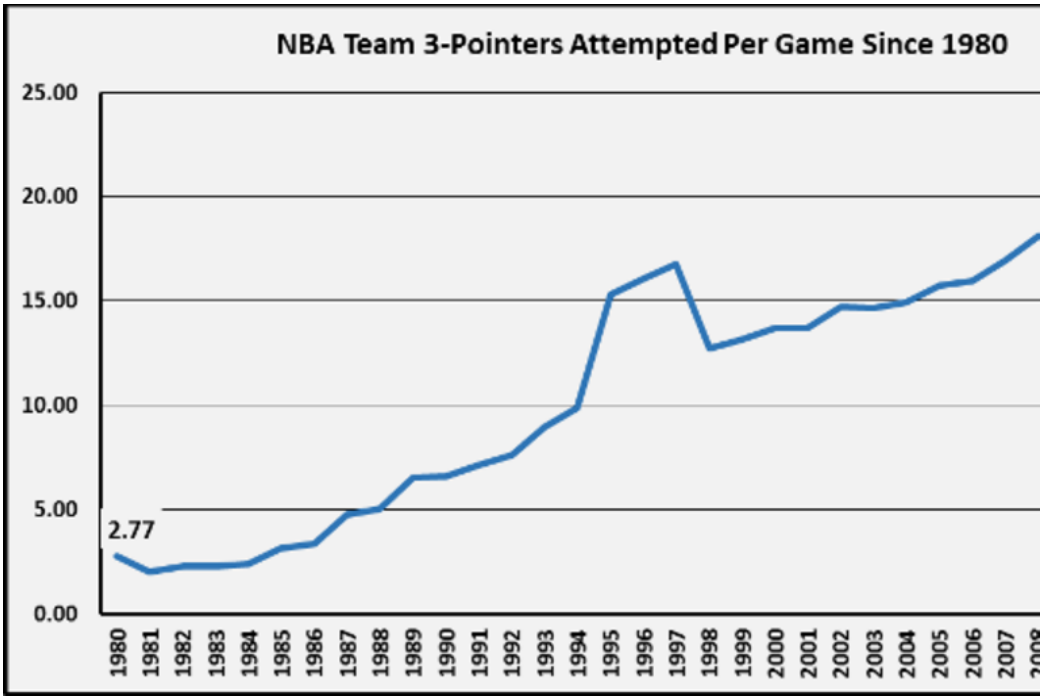


Figure 7: Three point graphic

This graph shows the drastic increase of three pointers attempted after the introduction of the three point line in the NBA. The graph highlights the initial gradual increase of three point shots attempted, and the continuous upward trend of three point shots attempted by teams around the league.

Another major rule change that affected the way defense was played and thus affected offenses was the slow abolishment of hand and forearm checking.

Aggressive defensive teams like the "Bad Boy" pistons in the 1990s heavily influenced this change as the NBA aimed to combat fighting.

Defensive contact with hands and forearms both in the back and front court was eliminated, except with offensive players who got the ball below the free throw line. Defenders also were prohibited from pushing players off the ball or stopping and interrupting offensive players setting screens. The new rules were intended to liven up the game of basketball and change it from a slow physical game to a more "flashy" game that had a rise in average points scored, field goal percentages and three-point percentages. [9]

The major rule changes led to the rise of the new 2000's strategy known

as `small ball`. Small ball deviated away from big men being constantly stationed under the post, near the basket. Small ball was all about sacrificing player size and emphasis to the low post to instead utilize smaller, agile outside scoring players. Small ball oriented their team generally around strong small forwards like `Lebron James`, `Kevin Durant`, and `Carmelo Anthony` who are surrounded by players that can switch between positions and be more versatile.

One of the great examples of a team excelling at small ball is the 2014-2015 Golden State Warriors, who for the first time in forty years won a Championship without a starting player over 6'8". Small ball had offensive tactics that allowed all five versatile players to execute quick effective passes before shooting which was essential to their success, as more athletic and agile sized players were favored.

The rise in three point shots was also a direct effect of this new style of play. With a much more spread-out court and lack of big men posting up under the basket, naturally players were encouraged to increase outside shooting leading to the rise of three-point shots and excellent shooters like League MVP `Steph Curry` that helped lead the frontline of the warriors domination in the 2010's(Grant, 2015).

One last major factor that is necessary to look at to understand how and why the game has changed over time is the increase of international players in the league. European players have had a playstyle that focuses on fundamentals and more team oriented playing versus "selfish" basketball from U.S. players. By the 2014-2015 season nearly a quarter of players represented 37 total nations of players in the NBA. San Antonio Spurs head coach Greg Popovich excelled at bringing the talent of foreign players together. The Spurs have seen major success, being the most winning NBA team since the 2000's. "With this strategy Popovich became a three-time NBA Coach of the Year and earned five Championship rings. Spurs general manager R.C. Buford claims "This is all Pop's vision. He decided there wasn't a place in the world, a street we wouldn't walk down, a rock we wouldn't turn to find a player." [9]

Popular Dashboards

There are many platforms where users like myself have used dashboards to create interactive visualizations. Some of the top platforms include `Tableau`,

`Google Analytics`, `Klipfolio`, `Power BI`, and `Datapine`.

One specific project that implemented similar approach to mine is **An Interactive Basketball Analytics Report using Tableau.** by Adhithyan R. [13]

The main objective of his project was to develop a Tableau Dashboard that can assist users in viewing the shot charts of each player during a specific season. This would be achieved by allowing users to create and execute queries and then present the results graphically. Research projects like this are beneficial for basketball fans and coaches who wish to conduct experiments and draw conclusions using user-friendly graphical interfaces. ProPredict, leverages similar ideas to projects like these, with an emphasis on the user experience, giving them a platform to study their data freely.

The Moneyball Approach

A well documented performance enhancing algorithm is The Moneyball approach, which is a data analysis tool that utilizes statistical models, such as linear regression, to predict sports performance based on player statistics. The approach was first used in Major League Baseball (MLB) in the early 2000s by the Oakland Athletics, a small-market team with limited resources.

To implement the Moneyball approach, the Athletics' front office focused on identifying undervalued players who excelled in areas that were statistically linked to team success. These areas included on-base percentage, walks, and slugging percentage.

To identify these undervalued players, the Athletics used a linear regression model to predict team success based on these statistical factors. Specifically, the team developed a model to predict a team's win-loss record based on the following equation:

```
Wins = b0 + b1 * OBP + b2 * SLG + b3 * FLD
```

where OBP represents on-base percentage, SLG represents slugging percentage, FLD represents fielding percentage, and b0, b1, b2, and b3 represent coefficients determined through the regression analysis.

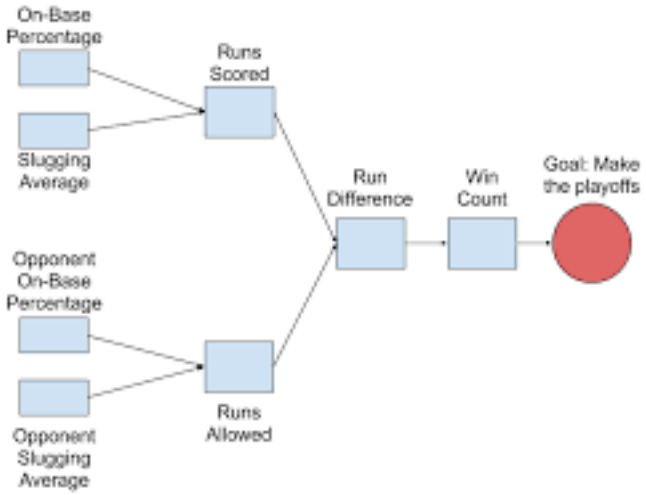


Figure 8: The method behind MoneyBall

Using this model, the Athletics were able to identify undervalued players who excelled in these areas but were not necessarily recognized by traditional scouting methods. This allowed the team to acquire these players at a lower cost than more well-known players with similar overall performance metrics.

The results of the Moneyball approach were impressive. Despite having one of the lowest payrolls in the league, the Athletics were able to compete with much wealthier teams and consistently contend for playoff spots. In the 2002 season, the team won an impressive 103 games, tying for the best record in the league.

The Moneyball approach is a prime example of how data analysis software can be used to generate winning results in sports. By leveraging statistical models such as linear regression to identify undervalued players and determine which player statistics are most correlated with team success, teams can gain a competitive edge even with limited resources.

The success of the Moneyball approach has since inspired other teams across sports to adopt similar data-driven strategies. In baseball, for example, the approach has become commonplace, with teams investing heavily in statistical analysis and hiring specialized analysts to crunch numbers and identify undervalued players. [18]

Basketball has undergone numerous changes over the years, including modifications to the rules and the way the game is played. For example, the elimination of hand-checking, the introduction of the three-point shot, and

the influx of international players have had significant impacts on the game.

Analysts who use data visualization platforms such as ProPredict can conduct various research studies like exploring the addition of the three point shot to dive deeply into team and player statistics over the years and analyze the evolution of basketball as a sport.

ProPredict replicates the consistent theme of sports analysis tools by being able to turn columns of data into evidence-based results that can contribute to the success of a sports team.

Method of approach

This project began with the task of making a simple yet versatile interactive data dashboard that experts or fans can use with minimal knowledge of code. To do this I had to map out why coaches/analysts would want to study, in this case NBA team data, and how do I give them a platform that gives statistically correct visuals and predictions.

There are several reasons why an interactive dashboard would be useful to coaches. Analyzing team data can provide insights into the complex dynamics of the sport of basketball. This can help analysts to better understand how teams win games and to identify key factors that contribute to success. They can also use team data to inform strategic and tactical decisions, such as offensive and defensive strategies and coordinate their lineup to best suit their desired strategy. This can help teams to gain a competitive advantage and to achieve better results on the court. Analysts can use team data to evaluate player and team performance, to identify strengths and weaknesses, and to make informed decisions about player development, coaching, and other key aspects of the game.

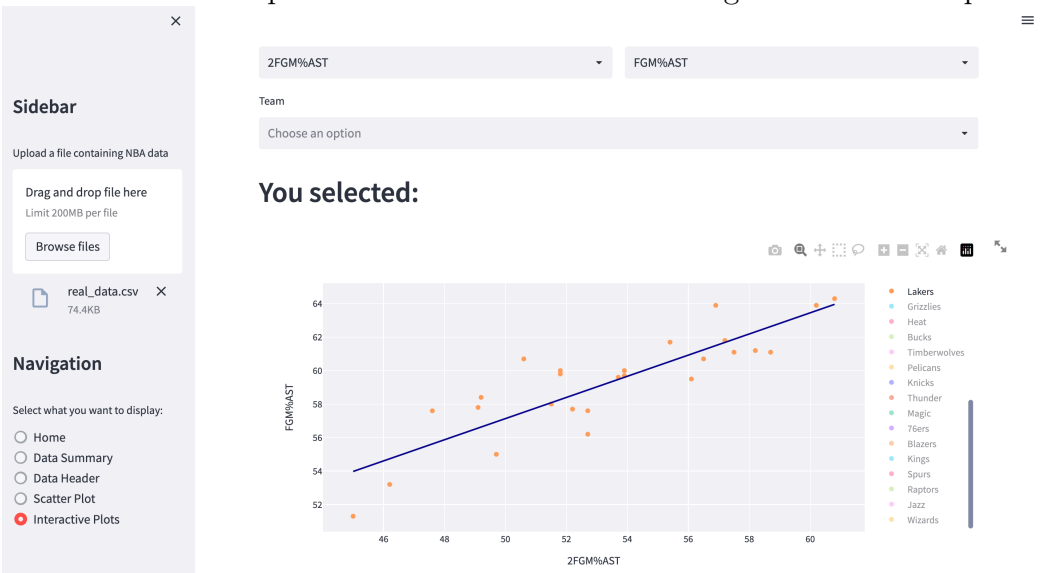
By analyzing team data over time, analysts can track trends and make predictions about future performance. This can help teams to plan for the future and to make informed decisions about the future of the sport. They can use team data to look at the impact of league rules and policies on the sport, such as the introduction of the three point line, and make recommendations for changes that would improve the game.

The design of my study analyzing NBA player and team data was done by using the python library, streamlit, and plotly, which set up a web application dashboard where users can create graphs and run experiments on any data of their choosing. Streamlit was chosen as the dashboard for my study because of its user-friendly interface, but still a strong versatility. Streamlit is an open-source Python library that allows users to create interactive,

web-based dashboards with simple Python code. Streamlit takes care of the web framework, front-end rendering, and user interaction, enabling data programmers to focus on writing the code for their data analysis. It has built-in support for popular data visualization libraries such as Plotly components that can be easily customized and extended to meet specific needs for many datasets. By using Streamlit, users can quickly prototype and deploy their data analysis dashboards, making it an accessible and powerful tool for data scientists and engineers. [16]

Data Visualization with Plotly: Converting Data into Insights

Taking a closer look into the implementation of the graphics, `plotly`, the data visualization library that enables users to analyze and present their datasets in an interactive, visually appealing way is used. Plotly includes a wide range of chart types, including line charts, scatter plots, bar charts, histograms, pie charts, and more, allowing users to choose the most appropriate type for their data. Ploty was also a good fit for this experiment because of its versatile built-in functionality for statistical analysis, displaying r-squared values and mean squared error values to measure significance of the plots.



In figure 8 you can see an example of the built in regression line in the

plotly library, where data analysts can easily see a line of best fit for their data.

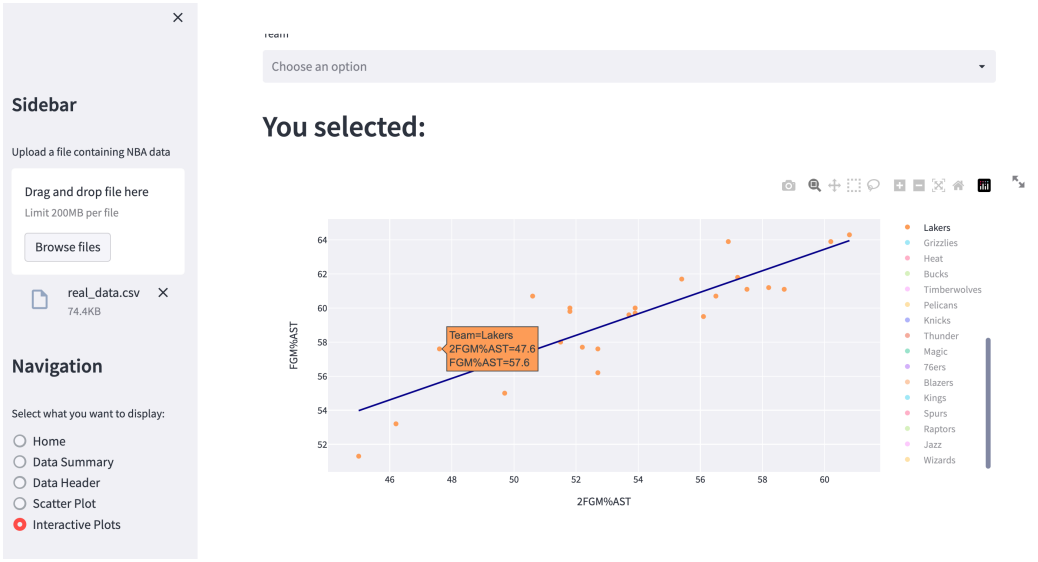


Figure 9: Data Points

To take a closer statistical look, as shown in figure 9 plotly users can hover over data points,to be able to see the exact point coordinates and be able to understand the relationship between the chosen variables.

Plotly's scatter plots are very beneficial for analyzing sports team data for many reasons. Scatter plots allow users to visualize relationships between variables, identify outliers, and track trends over time. For this study scatter plot's were effectively used in two ways, Trend analysis and outlier detection.

Scatter plots can be used to track trends in team performance over time. For example, a scatter plot can be created to show the change in a team's three point percentage over the course of a season. This can help analysts/-coaches understand if a team is improving, declining, or staying consistent over time. Comparing trends between multiple teams at once using streaml-its multi select feature with scatter plots can identify key differences between teams and develop into further research inquiries. Scatter plots allow users to easily identify outliers, or data points that are significantly different from the rest of the data. This can be useful in sports analysis to identify teams

that are performing significantly better or worse than expected, and to understand the reasons behind these anomalies.

Diving deeper into statistical analysis, I needed to include an approach that would be able to give statistical evidence that there is or isn't a correlation between X and Y values. Adding into my design was the inclusion of a line of best fit that displayed R-squared values when the user hovered over it as seen in figure 10 below.

The R-squared value (also known as the coefficient of determination) is a statistical measure that indicates the proportion of variation in a dependent variable that can be explained by the independent variable(s) in a regression model. In other words, it tells you how well the model fits the data.



Figure 10: Strong Correlation

As in figure 10, a higher R-squared value indicates a stronger relationship between the independent and dependent variables and a better fit of the model to the data. This advanced statistic is a great way to indicate a strong and significant relationship between variables, or a weak and insignificant relationship between variables, either way displaying to the user evidence behind the correlation on the scatter plot. [19]

After comparing different elements of the dataset I am able to analyze over time what statistics changed the most numerically, and back my statistical

evidence up by relevant research. For example, the 2010's brought us the NBA'S all time three point scorer, Steph Curry. Having known background knowledge of the monumental success of his team, the Golden State Warriors in the last decade, I am able to draw a graph comparing the percentage of points scored by three pointers a game against years overtime. After looking at team's with the highest correlation between these variables I am also able to compare this statistic against win count to not only display a significant statistic that has a high correlation between three pointers made, and a specific time frame, but also a statistic that correlates positively with number of wins in a season, showing an known and proveable effective strategy for a team during this time period.

To recap, Propredict's methodology creates a platform for coaches, analysts, and sports enthusiasts to be able run experiments on a dataset of their choosing without having to worry about working and maintaining code, but rather focusing their time on diving into their own analysis of sports data on a platform that allows them to input their own sports dataset of choosing. A platform where users can select variables from their dataset to get meaningful and statistically measurable results that are easy to comprehend, and in the end be able see conclusions because of the user-friendly interface of this dashboard. With these beginning methods, allowing users to research through different team and player statistics, users can get an idea of how they want to use the predictive modeling section to create and run experiments of their own with results tailored to their own research interests chosen from the dataset.

Experiments

Experimental design

The experiment for this section was chosen to reflect the versatility and user driven experience of the tool. The data for this experiment was obtained from basketballreference.com, a publicly available online database of NBA player statistics. The dataset covers the period from the start of the 2000 NBA season until the end of the 2021-2022 NBA season. The current season (2022-2023) was intentionally omitted from the dataset to serve as a benchmark for comparing predicted and actual values. To ensure a manageable sample size and representative selection of the NBA's top performers, a subset of 10 players was selected from the larger dataset. Each player had to meet the following criteria: have played in at least 55 games in the current season (2022-2023), and be ranked among the top 50 players in one of the following statistical categories: points, assists, rebounds, turnovers, and steals per game as of March 23, 2023.

	A	B	C
1	**Player**	**Statistic used**	**r-squared**
2	Joel Embid	PTS	0.74
3	Luka Doncic	PTS	0.55
4	Damian Lillard	PTS	0.56
5	Lebron James	PTS	0
6	Nikola Jokic	PTS	0.91
7	Trae Young	PTS	0.42
8	Jimmy Butler	PTS	0.56
9	Russel Westbroc	PTS	0.15
10	Julius Randle	PTS	0.79
11	Jayson Tatum	PTS	0.91
12			
13	Joel Embid	AST	0.4
14	Luka Doncic	AST	0.57
15	Damian Lillard	AST	0.58
16	Lebron James	AST	0.24
17	Nikola Jokic	AST	0.83
18	Trae Young	AST	0.81
19	Jimmy Butler	AST	0.79
20	Russel Westbroc	AST	0.28
21	Julius Randle	AST	0.76
22	Jayson Tatum	AST	0.95

Figure 11: Experiment Dataset

The entire career data for each of the 10 players was used to train and test the linear and random forest models. Their points, assists, rebounds, turnovers, and steals per game averages from the current season (2022-2023) were recorded as the actual results to be compared to the predicted values generated by the models. The number of seasons played by each player was not controlled for, as this allowed for a more comprehensive evaluation of the models' performance across different levels of experience. The primary objective of this experiment was to evaluate the accuracy of two different

models (linear and random forest) and to determine which one performs better.

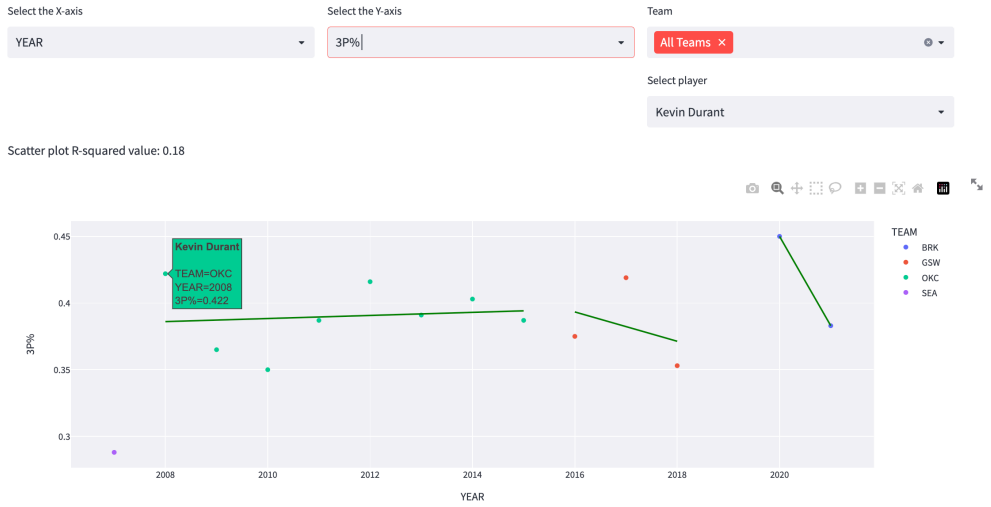


Figure 12: Kevin Durant's 3-point percentage across career

To begin the experiment, an interactive plot section within the tool was used to select a specific player and generate a graph of their statistics over time. The graph displayed the year/season on the x-axis and the selected y-variable statistic on the y-axis, providing a visual representation of any trends or patterns in the data. A line of best fit was generated using the scikit learn package, along with an r-squared value, to establish a baseline understanding of the correlation between the variables. [17]

After generating the graph and establishing the baseline correlation, the given data for each player was run through both a linear regression model and a random forest model on the predictive modeling page as shown in figure 13 below.

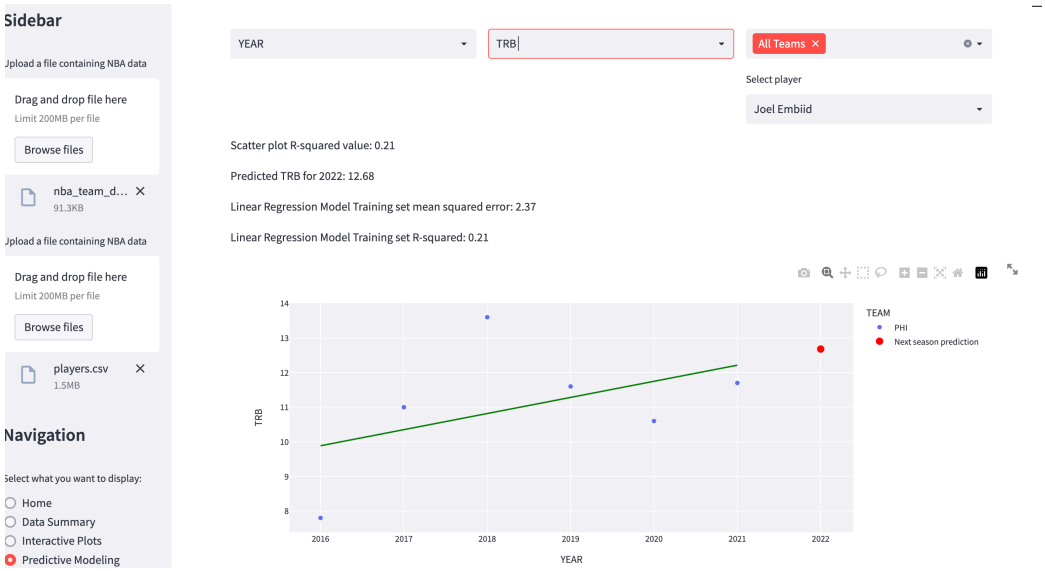


Figure 13: Linear Regression Model

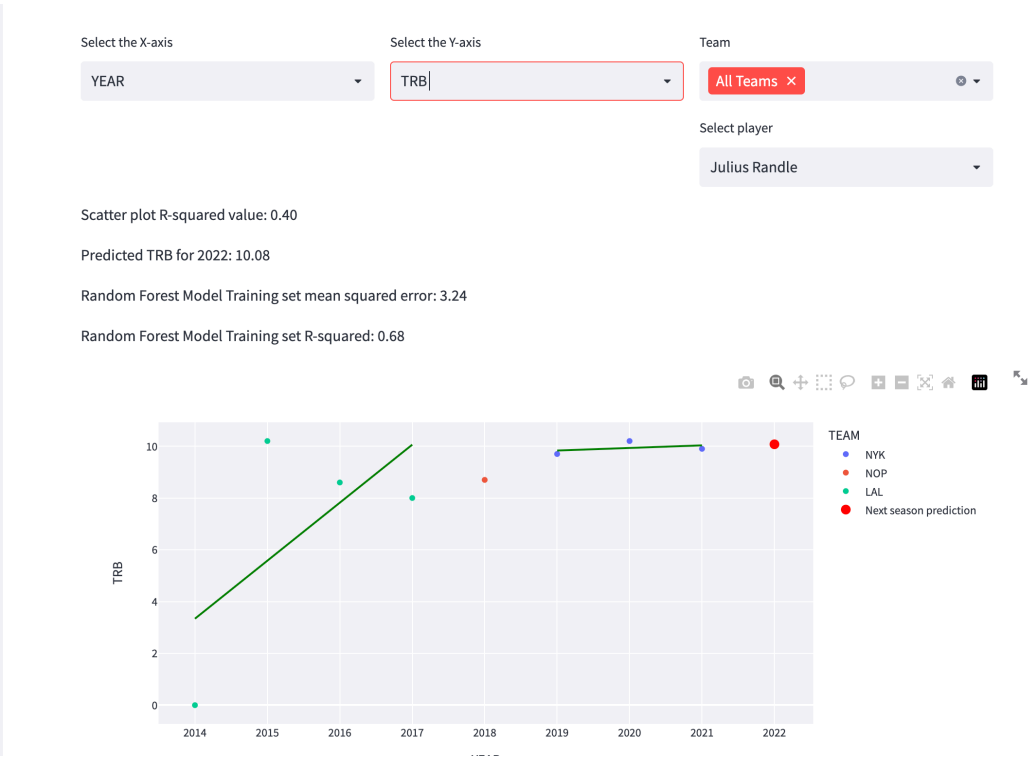


Figure 14: Random Forest Model

This was done by selecting the player name and the y-variable statistic for prediction, and the output predicted value was recorded in a table for both models to allow for easy comparison. This approach allowed for a comprehensive evaluation of the accuracy of both models, which could then be analyzed to identify which model performed better.

The experimental design allowed for a thorough evaluation of the accuracy of both linear regression and random forest models in predicting NBA player statistics. The results of the experiment can be analyzed to identify which model performed better and to identify any limitations or areas for future research.

Linear Regression Algorithm

The purpose of the linear model is to predict a player's statistics for the next season. The linear model uses the player's historical data to make this prediction.

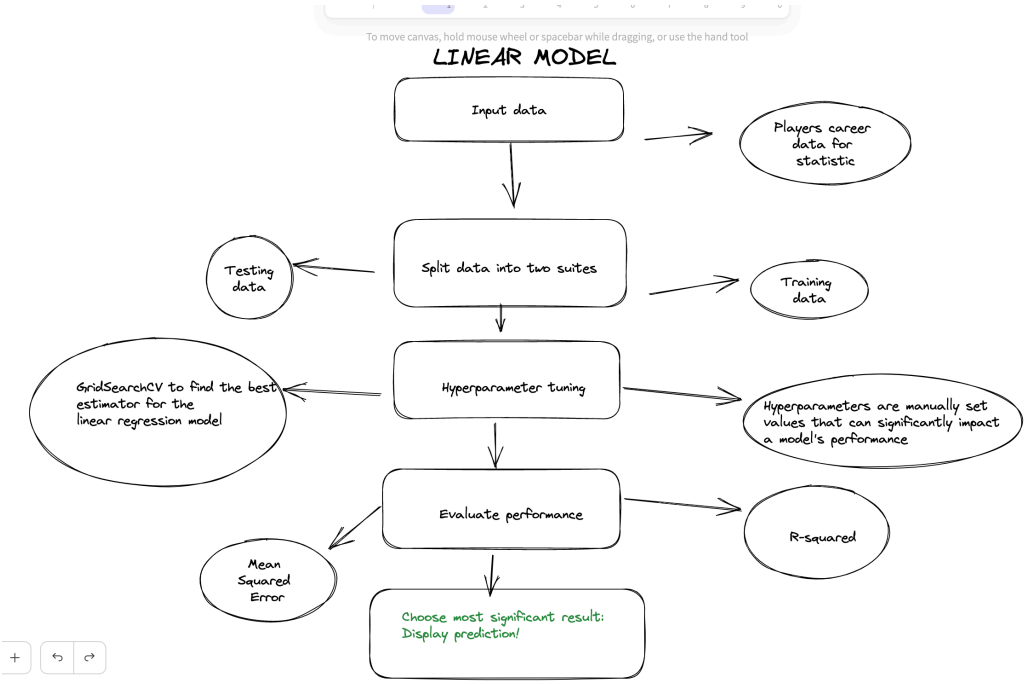


Figure 15: Linear Model Flowchart

To generate the prediction, the model first splits the data into a training set and a testing set. The training set contains all the data up until the end of the most recent season, while the testing set contains only the data for the upcoming season.

The model then performs hyperparameter tuning using GridSearchCV to find the best estimator for the linear regression model. Hyperparameters are values that cannot be learned from the data and must be set manually by the user. They can have a significant impact on a model's performance and accuracy, so it is important to choose them carefully. Hyperparameter tuning is the process of selecting the best hyperparameters for a model to

optimize its performance.

The GridSearchCV method exhaustively searches through all possible combinations of hyperparameters in the specified grid and evaluates the performance of each combination using a specified scoring metric, in this case, the R-squared value. The best combination of hyperparameters is then selected as the one that results in the highest score on the scoring metric and then used to calculate the predicted value.

After fitting the linear regression model to the training set, the model evaluates its performance using mean squared error and R-squared values. The mean squared error indicates how close the predicted values are to the actual values, while the R-squared value measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

Finally, the model generates a prediction for the player's performance in the next season using the best estimator from the linear regression model. The predicted value is then displayed on an interactive scatter plot alongside the player's historical data(shown in red), allowing for easy comparison and analysis. The model's performance can be evaluated by comparing its predicted values with the actual values in the next season.

On the positive side, linear models provide a clear and interpretable relationship between the predictor variables and the response variable. They are also simpler than their counterparts and can be easier to understand without a statistical background. Additionally, linear models can be trained more quickly, even on large datasets.

However, the linear model has its disadvantages as well. The model is based on the assumption that the relationship between the variables is linear, which means that it performs very well when the points follow a linear trend, but can be off on its prediction when they don't. This is because linear models can be sensitive to outliers, which can significantly affect the model's accuracy. Because NBA players can be subject to injuries, trades, and other factors that cause outliers in the dataset, the linear model's accuracy may be affected. [4]

Random Forest Regression Algorithm

The random forest model is a machine learning algorithm that creates multiple decision trees to predict an outcome variable (in this case, an NBA

player's statistics) based on a subset of input variables (in this case, the year). By combining the predictions from all the decision trees, the model reduces overfitting and improves accuracy.

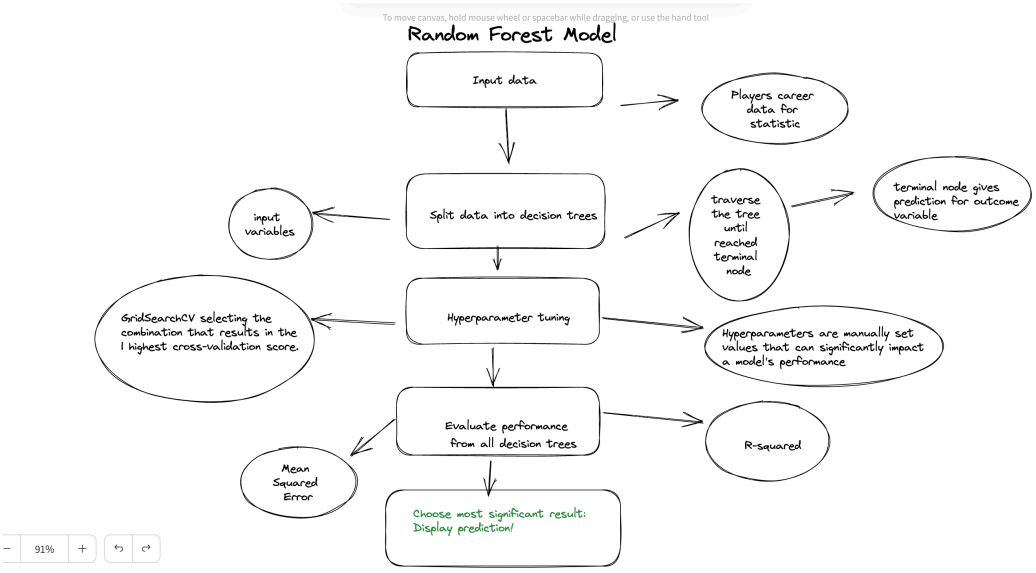


Figure 16: Random Forest Flowchart

A decision tree is a powerful and simple algorithm that recursively splits the data based on input variable values until it reaches a set of terminal nodes that make a prediction for the outcome variable.[7] [15] The tree structure makes it easy to interpret and visualize, as it shows which input variables are most important for making a prediction, and the terminal nodes provide a clear decision rule for making a prediction. Additionally, decision trees are robust to outliers and can handle both numerical and categorical input variables.

Hyperparameters of the random forest model specify how the decision trees are built. The GridSearchCV function was used to find the best hyperparameters for the random forest model by performing a grid search over a range of hyperparameters and selecting the combination that results in the highest cross-validation score. One specific hyperparameter was the number of decision trees to include in the random forest, decided between 10, 50, or 100. The exhaustive search would find which number of trees gave the

highest r-squared value and then, that subset of trees was chosen for the prediction.

The random forest model was used to predict an NBA player's statistics for the 2022 season based on their historical statistics from previous seasons. The model's performance was evaluated by calculating the mean squared error (MSE) and R-squared value on the training set and then recorded in a table for comparison.

The random forest model has its pros and cons. Unlike the linear model, the random forest model is built to be more flexible as it is non-parametric and non-linear, meaning it does not rely on assumptions about the relationships between variables. Additionally, as an ensemble of decision trees, it is less sensitive to overfitting than individual decision trees, and it is robust to outliers and missing data. However, the random forest model can be time-consuming to train when there are many input features or a large amount of data. Additionally, it can still overfit if hyperparameters are not tuned correctly, requiring careful selection and validation. Lastly, the model can be challenging to interpret, as it consists of an ensemble of decision trees, making it difficult to understand which input features are most important for predicting the output variable. [7]

Evaluation

	A	B	C	D	E	F	G	H	I	J
1	**Player**	**Statistic used**	**r-squared**	**predicted LM**	**predicted RF**	**actual**	**error LM**	**error RF**	**MAE LM**	**MAE RF**
2	Joel Embid	PTS	0.74	31.88	29.21	33.2	1.32	3.99	3.923	3.401
3	Luka Doncic	PTS	0.55	31.65	28.27	32.9	1.25	4.63		
4	Damian Lillard	PTS	0.56	24.9	26.39	32.2	7.3	5.81		
5	Lebron James	PTS	0	27.24	26.79	29.5	2.26	2.71		
6	Nikola Jokic	PTS	0.91	30.11	26.75	24.8	5.31	1.95		
7	Trae Young	PTS	0.42	31.5	27.71	26.8	4.7	0.91		
8	Jimmy Butler	PTS	0.56	17.59	21.1	22.9	5.31	1.8		
9	Russel Westbroc	PTS	0.15	23.04	20.71	15.4	7.64	5.31		
10	Julius Randle	PTS	0.79	27.47	22.3	25.5	1.97	3.2		
11	Jayson Tatum	PTS	0.91	32.27	26.4	30.1	2.17	3.7		
12										
13	Joel Embid	AST	0.4	4.03	3.66	4.2	0.17	0.54	1.223	0.773
14	Luka Doncic	AST	0.57	10	8.69	8.2	1.8	0.49		
15	Damian Lillard	AST	0.58	7.75	7.39	7.3	0.45	0.09		
16	Lebron James	AST	0.24	8.38	7.88	6.9	1.48	0.98		
17	Nikola Jokic	AST	0.83	9.73	8.1	9.9	0.17	1.8		
18	Trae Young	AST	0.81	10.35	9.53	10	0.35	0.47		
19	Jimmy Butler	AST	0.79	7.43	5.91	5.1	2.33	0.81		
20	Russel Westbroc	AST	0.28	10.31	8.23	7.5	2.81	0.73		
21	Julius Randle	AST	0.76	6.15	5.55	4.2	1.95	1.35		
22	Jayson Tatum	AST	0.95	5.42	4.23	4.7	0.72	0.47		
23										
24	Joel Embid	REB	0.21	12.68	11.3	10.2	2.48	1.1	1.387	0.693
25	Luka Doncic	REB	0.17	9.2	8.83	8.6	0.6	0.23		
26	Damian Lillard	REB	0.22	4.19	4.16	4	0.19	0.16		
27	Lebron James	REB	0.32	8.32	8.1	8.4	0.08	0.3		
28	Nikola Jokic	REB	0.67	13.43	12.25	11.9	1.53	0.35		
29	Trae Young	REB	0.03	3.8	3.88	3.1	0.7	0.78		
30	Jimmy Butler	REB	0.58	7.36	6.25	6.1	1.26	0.15		
31	Russel Westbroc	REB	0.63	11.04	8.66	6	5.04	2.66		
32	Julius Randle	REB	0.4	12.09	10.08	10.1	1.99	0.02		
33	Jayson Tatum	REB	0.97	8.9	7.72	8.9	0	1.18		
34										
35	Joel Embid	TOV	0.88	2.81	3.1	3.5	0.69	0.4	0.724	0.414
36	Luka Doncic	TOV	0.75	4.95	4.4	3.7	1.25	0.7		

Figure 17: Experimental Results

Figure 17 shows the recorded output values for both the linear and random forest models. The actual statistical values for the current season were recorded before testing, and the error difference was calculated by subtracting the absolute value of the actual value from the predicted value for each player. The mean absolute error was then calculated for each model and for each corresponding statistic. The findings from the experiment show that the random forest model outperformed the linear model in all five tests, achieving a lower mean absolute error. This indicates that the random forest model is more accurate than the linear model in predicting statistical values for this experiment.

The null hypothesis proposed in this study was that the linear model

would be more accurate in cases where the r-value for a player's statistical graph was greater than .5 and that the random forest model would be more effective for lower r-values, variables with a lower correlation and more outliers. This is based on the assumption that a significant correlation between the x and y values would result in a linear trend, which the linear model is better equipped to handle. However, the analysis of the results shows that there is no significant correlation between a high r-value and the accuracy of the linear model compared to the random forest model. Therefore, the null hypothesis can be rejected.

The random forest model proved to be more effective in this case due to its ability to capture non-linear relationships by combining decision trees with ensemble learning. By splitting the data into multiple decision trees and then aggregating the predictions from that unit of trees, the model can arrive at more accurate conclusions.

This is particularly well-suited to NBA data, which includes many outliers resulting from factors such as player injuries, trades, and changes in team performance.

Additionally, player statistics can change due to human factors, making it impossible to predict their performance perfectly every time, so a model that expects there to be changes in the trend of the data works well.

The accuracy of the model is noted by the MAE(mean absolute error column). The mean absolute error is the difference between the predicted values and the actual values. This statistic was chosen to display the accuracy of the tool so that readers can easily see the average point differential between predicted and actual values.

However, determining an accuracy percentage has its complications. To determine an accuracy percentage I would first need to set a threshold for a mean absolute error that would result in an accurate prediction. From there the total number of predictions within that threshold would be deemed accurate and the number of predictions outside the threshold would be called inaccurate, and I could get a percentage from that. Choosing the threshold is the difficult part because different predictions have different meanings.

A closer examination of the results table reveals that some statistical categories have a lower mean absolute error than others. For example, categories such as PTS per game, where the numbers typically range from the low 20s to low 30s, are more difficult to predict accurately. In contrast, statistics like STL (steals) were predicted more closely by both models, likely because no player's steals per game exceeded 2, making it easier for the model to be

accurately trained on these statistics.

Player	Statistic used	r-squared	predicted LM	predicted RF	actual	error LM	error RF	MAE LM	MAE RF
Joel Embid	PTS	0.74	31.88	29.21	33.2	1.32	3.99	3.923	3.401
Luka Doncic	PTS	0.55	31.65	28.27	32.9	1.25	4.63		
Damian Lillard	PTS	0.56	24.9	26.39	32.2	7.3	5.81		
Lebron James	PTS	0	27.24	26.79	29.5	2.26	2.71		
Nikola Jokic	PTS	0.91	30.11	26.75	24.8	5.31	1.95		
Trae Young	PTS	0.42	31.5	27.71	26.8	4.7	0.91		
Jimmy Butler	PTS	0.56	17.59	21.1	22.9	5.31	1.8		
Russel Westbroc	PTS	0.15	23.04	20.71	15.4	7.64	5.31		
Julius Randle	PTS	0.79	27.47	22.3	25.5	1.97	3.2		
Jayson Tatum	PTS	0.91	32.27	26.4	30.1	2.17	3.7		

Figure 18: PTS Results

Figure highlights an important discovery in the results which is the tendency of the model to undershoot the actual predictive value for most of the players when predicting points per game for the next season. While points per game is the statistic that usually varies the most throughout a player's career. The below scatter plot of LeBron James, a super successful NBA star, visualizes how much a top player's points per game can change across their career, making it one of the harder NBA statistics to predict accurately.

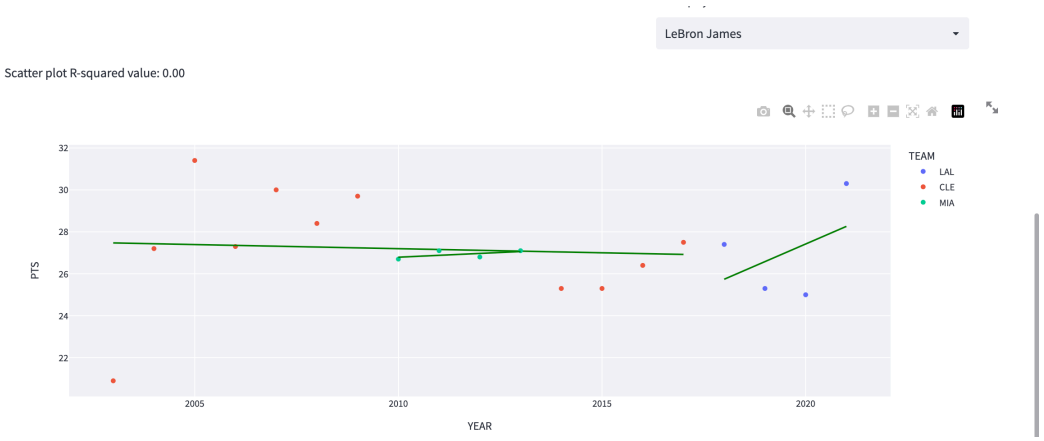


Figure 19: Lebron James Points Throughout His Career

This plot even hints that there is more to being a successful player than averaging the most points per game each year.

Analyzing Predictions

When making conclusions based on results from ProPredict it is essential to research how the prediction fits into the current date and time of the sport.

Per espn.com "The NBA experienced a notable explosion in scoring throughout the 2022-23 regular season, particularly among the league's stars. The NBA saw 15 different players combine for 25 50-point games this season, the most in a single regular season since the early 1960s." [1]

This outbreak in individual high scoring performances effectively explains why the model was undershooting the points per game prediction consistently, especially because the experiment was tailored to top players in the league.

This also further highlights the difficulty of predicting human based numerics, because these players cannot realistically put up the exact same numbers each game.

From the experiment, it is conclusive that when these models have more data to train from, and have lower numerical values to train from, they are able to make more accurate predictions. Based on these conclusions, it is impractical to choose an overall accuracy score because each statistical category's predictions hold different accuracy values. To get a valid accuracy score for each statistic, more players would need to be included in the experiment to answer the question of what is an accurate prediction. Because of this I chose to leave the accuracy metric as a subtracted difference value so that the accuracy can be examined on a statistic by statistical analysis.

Threats to Validity

The first area to look through when checking the Validity of the dashboard is the input data. The dashboard is built to handle sports data, so it is important to choose a dataset that deals with sports statistics to ensure there doesn't have to be code manipulation. The interactive plots work with current and old sports data, however the predictive model will only give reliable conclusions with current sports players. Choosing to predict points per game for next season for a player who retired ten years ago would be an example of an inaccurate conclusion. The model will still generate a prediction, because in plain terms the model sees input data and gives output data, the prediction. It is up to the researcher to analyze the type of data

the model is being trained on and to analyze what the conclusion means in the current date and time it was generated. If the conclusion is valid then it can be declared as useful, if it doesn't fit in real world time, the prediction is flawed and shouldn't be considered.

It is also crucial to analyze the output data, especially from the predictive models. When determining the accuracy of the model, analysts must take into account the type of statistic that was being predicted. A metric such as predicted steals per game for a season which most NBA players rarely exceed two , would be incorrect to declare as accurate if the predictive model is overshooting it and predicting values of 4 or 5. In contrast, a statistic such as points per game in which the highest scorers for a season are in the low 30s can be called accurate for predictions that are within 3-4 points because the points per game statistic usually varies among players throughout their careers. The best way to prove the accuracy of these models is to use the predictions, find statistics that result in more wins, change the strategy to fit the predictions and then see how effective those changes are at winning more games.

The dashboard has two main uses, the interactive plots and the predictive models. To get the most reliable results from ProPredict, and based on the way it is set up, it is most crucial to choose a dataset that deals with team and player statistics. A dataset containing years worth of team/player statistics would allow users to use the interactive plots and gain accurate insights from predictive models as intended and not need to change much, if any code. The plots are setup to exclude player names and team names, and other string variables from the dropdown menu for the plots, and instead make them into their own column for team/player selection because it would be impractical to include all player or team names on a graph, and is better to have them in a sidebar with a key identifier so users can easily see what points on the plot relate to the player or team. If no team or player statistical data can be found or used, sports fans can still get use out of the tool. The way the interactive plots are setup is that it reads in the csv file and has it saved as a dataframe. From that dataframe users can select any of their input variables for the graph as long as it is not a string value. The interactive plots page will still work as intended with datasets like ticket sales, or merchandise sales of players or teams, and you will still be able to find trends and outliers amongst your data easily.

However, the predictive models are set up to make predictions about in-game statistics for the next game, next month, or next year depending on

the input data. To get full use out of the tool users would have to make some small tweaks to the code, making sure they exclude the strings in their data set and changing the prediction value to match their dataset. For example a dataset that deals with total tickets sold for one season will work and users will be able to view the different elements of the set on a graph, but the predictive model specifically is most effective when it has more data to be trained on.

Sports data is subject to outliers due to injuries, team changes, and other external factors that can impact player performance. It is crucial to acknowledge these factors when analyzing the results of the experiment.

There is a risk of selection bias due to the small sample size of the experiment, which only involves 10 players. As a result, it may not be possible to draw conclusions about the entire population of NBA players. To mitigate this issue, it would be beneficial to include more participants in the study, which would provide a better representation of the population and a more robust conclusion regarding which model performs better for NBA data.

The use of hyperparameter tuning, which are manually set values that impact the models performance, could introduce overfitting, which means that the model may perform well on the training set but struggle when new data is added. To avoid this issue, it is essential to use a separate validation set during the tuning process to test the model's performance on data it has not yet seen.

In future experiments, incorporating more data points and testing additional models could improve the accuracy of predictions and produce more reliable results. The experiment featured a non-linear trend expecting model and a linear trend expecting model to give a baseline of which class of models are more effective with NBA data. Implementing more models will give even deeper insights to the most efficient and effective predictive algorithms for sports data. The experiment showcased 10 different NBA players who had a different amount of data points to be trained on for the predictions. The number of seasons each player has been in the NBA was purposely left uncontrolled so that the results could shed light on how predictive models work better when they have more data points to be trained on for the prediction. Players like Trae Young who has only been in the NBA for a few seasons had poor predictions when compared to players like LeBron James who has been in the league for 20 years now. While LeBron James had plenty of seasons with recorded statistics such as points per game to test on, the models can still be improved by adding more data points to be trained on. To get a more

accurate prediction in this case, a user can input datasets containing LeBron James points for each game in a season and then run the model with those sets and get a closer prediction to the actual value because of the increase in data points to be trained on.

Summary of Results

The experiment evaluated the performance of linear and random forest models in predicting statistical values for NBA players. The findings show that the random forest model outperformed the linear model in all five tests, indicating that it is more accurate in predicting statistical values for this experiment. The null hypothesis that the linear model would be more accurate for high r-values was rejected, and the random forest model was found to be more effective due to its ability to capture non-linear relationships by combining decision trees with ensemble learning. The analysis also revealed that some statistical categories are more difficult to predict accurately than others. The experiment demonstrates the benefits of using advanced machine learning models to predict player performance in the NBA.

Future work

While the tool has already demonstrated effectiveness in generating predictions based on year split statistics, there are several potential enhancements that could be made to increase its efficiency and versatility.

One area of improvement would be to explore additional regression models that could be incorporated into the tool. By testing and comparing different models, such as linear regression, logistic regression, or decision tree regression, it may be possible to identify the model that yields the most accurate predictions.

Another avenue for enhancing the tool's effectiveness would be to run experiments on sports data to determine which models work best for specific sports. By comparing model performance across different sports, it may be possible to identify which models are most effective for predicting outcomes in particular sports. This would enable the tool to be more versatile and provide more accurate predictions across a wider range of sports.

Another potential area for improvement would be to change the data set used by the tool from year by year statistics to a single season's worth of data. This would allow for more flexible and testable predictions as users could attempt to predict scores for the next game and those following it. By testing the tool's predictions against actual outcomes, it would be possible to identify areas where the tool could be improved, and further refine its predictive abilities.

To improve the predictive models, experiments set up to investigate the balance between accurate predictions can be set up. The machine learning algorithms' hyperparameters have been set up and run with both an exhaustive gridsearch and a randomized search. A grid search takes longer to run, combing through all hyperparameter combinations but has yielded more accurate results at a cost of runtime while a randomized search runs faster at a cost of predictive accuracy.

ProPredict is set up using a local host that opens up a one page streamlit browser. Users have to install the requirements file prior to running the tool. By deploying ProPredict to a website link, users would no longer have to know any code for the intended use of the tool. In addition, separating the different functions into different python files will allow the streamlit website to be separated into different pages, rather than one continuous page that takes some time to load each page.

Future Ethical implications

As with any tool that is designed to make predictions, it is important to consider the potential ethical implications of its use. In particular, if the tool were to be used to predict game outcomes and player statistics for upcoming games, it is important to recognize the potential for its use in sports betting, which can be a gray area in terms of ethical considerations. ProPredict is set up to give users free will to do their own research, and because of this it is up to the user to decide are the results meaningful, and what ways can these results improve and add to sports knowledge.

Another potential ethical concern is the potential for the tool and bigger tools using the same concept to be used to gain a competitive advantage. If a team or individual were to use the tool to make predictions about their opponents, they could potentially gain an advantage by adjusting their tactics or game plan accordingly. There is an ever growing conversation about ethical considerations of sports data, because data is a product of the people in the sport and as long as the game continues so will the data flow, and so will research and prediction based tools like ProPredict. It is important to consider a balance between ethical use of data and giving researchers the option to find and explore different stories that data may tell.

However, there are also arguments in favor of the use of prediction software. For example, proponents argue that such software can help to level the playing field between teams of different skill levels. Additionally, they argue that the use of such software can help to reduce the risk of injuries by identifying potential risk factors and allowing teams to adjust their training and playing strategies accordingly. The use of prediction software in sports is a complex issue that requires careful consideration of a number of different ethical concerns. In order to ensure that the use of such software remains ethical and transparent, it may be necessary to establish clear guidelines and

regulations governing its use. This could include requirements for teams to disclose the methods and data sources used in their predictions, as well as regular audits to ensure compliance with ethical standards. [11].

Conclusions

The primary goal of this tool is to provide users with a friendly and intuitive interface that offers advanced features. Its versatility makes it useful not only for sports teams, but also for fans who want to conduct their own experiments and learn something new about the sports they are passionate about. The tool's user-friendly interface and built-in flexibility make it effective in achieving its goals.

Users can easily import datasets of their choice by simply clicking on the downloaded file. The interactive plot page enables users to generate graphs quickly and easily to identify patterns in the data. Additionally, the predictive modeling feature allows researchers to experiment with different variables and record and compare statistical outcomes. The model testing results indicate that the random forest model was more effective at predicting accurate results when compared to the actual data.

The results of this study highlight the complexity of sports data, particularly when attempting to use models to predict player statistics. The experiments showed that there are many extraneous and unseen factors that can affect sports data, making it impossible to predict with perfect accuracy each time. Conclusions can be drawn that models, such as random forest models, which are built upon the foundation to work well with non-linear trends and outliers work more effectively with NBA data. However, the study also demonstrated that with tweaking and continuing testing of these models, improvements can be made to come closer and closer each time with more training.

In future work, it would be interesting to expand the scope of the tool to include varying sports data from physical metrics to in-game statistics. While the images and experiments for this paper included only NBA data, this tool is built for research for many sports. ProPredict is built with a goal of low cognitive load for understanding the uses of the tool, so that sports fans and analysts can focus their time on the research, on generating graphs, on experimenting with predictive algorithms. This tool is very effective with sports that have a team of players and play a game where the winner of the

game is declared by the team with the most score. Team and player datasets for sports like soccer, basketball, football, hockey, ect. work easily with the tool because it is designed for sports like this. Sports such as volleyball which play to a certain score rather than total score after time expires require extra conditions to be set up in the code so that any output data gives accurate conclusions to the rules of sports like volleyball. The inclusion of different sports datasets allows ProPredict to become more versatile and can allow for experiments across a multitude of sports. This also increases the audience of this tool and adds to the personalized features of ProPredict. The driving goal of this tool is to make it easy to use and to let people conduct self-motivating research.

ProPredict is an inclusive data analysis tool for sports enthusiasts and researchers alike, providing them with a flexible and user-friendly platform for conducting experiments and gaining proactive insights into the sports they love. ProPredict also accurately demonstrates the effectiveness of data analysis with machine learning turning statistical data into actionable results that can be used to improve sports players and teams.

References

[1] Kendra Andrews. 2023. Breaking down the NBA's historic surge of 50-point performances this season. *ESPN*.

[2] G. M. Arastey. 2019. An overview of sportscode, a key video analysis platform for performance analysts. (2019).

[3] G. M. Arastey. 2020. What is performance analysis in sport? (2020).

[4] Jason Brownlee. 2020. Linear regression for machine learning.

[5] Merrimack College (Ed.). 2021. How NBA analytics is changing basketball: Merrimack college. (2021).

[6] Tim Daniels. 2015. Charles barkley comments on role of advanced statistics, analytics in NBA. (2015).

[7] Niklas Donges. 2015. Random forest: A complete guide for machine learning. (2015).

[8] Kirk Goldsberry. 2019. *Sprawlball: A visual tour of the new era of the NBA*. Mariner Books.

[9] Horace Grant. 2022. How the game has changed. (2022).

[10] Jorge Lorenzo, Alberto Lorenzo, Daniele Conte, and Mario Giménez. 2019. Long-term analysis of elite basketball players' game-related statistics throughout their careers. *Frontiers in Psychology* 10, (2019), 421.

[11] Mike McNamee. 2015. Ethics and sports. In *Routledge handbook of the philosophy of sport*. Routledge, 131–141.

[12] NBA. 2021. Phoenix suns team info and news. (2021).

[13] Adhithyan R. 2021. An interactive basketball analytics report using tableau. (2021).

[14] Sujay Raghavendra. 2022. Visualization. In *Beginner's guide to streamlit with python: Build web-based data and machine learning applications*. Springer, 47–78.

[15] Mark R Segal. 2004. Machine learning benchmarks and random forest regression. (2004).

[16] Adrien Treuille. 2023. Main concepts. (2023).

[17] Gaël Varoquaux, Lars Buitinck, Gilles Louppe, Olivier Grisel, Fabian Pedregosa, and Andreas Mueller. 2015. Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications* 19, 1 (2015), 29–33.

[18] Ehren Wassermann, Daniel R Czech, Matthew J Wilson, A Barry Joyner, and others. 2005. An examination of the moneyball theory: A baseball statistical analysis. *The Sport Journal* 8, 1 (2005).

[19] Yuanhao Stanley Yang. 2015. Predicting regular season results of NBA teams based on regression analysis of common basketball statistics. *Under graduate Thesis, UC Berkeley* (2015).