CSE 6242 Final Report

Generating Betting Insight on NBA Player Performance using Machine Learning

Group 42: Sahil Bishnoi, Josh Garretson, Avery Girsky, Oliver Hewett, Hardik Patel, Atticus Rex

Introduction

The group is interested in identifying profitable bets given historical professional basketball data scraped from the NBA API, at the player and team levels. Once this data is gathered, the team is interested in training a model to output a probability distribution of the five main player stats, which are points (PTS), assists (AST), rebounds (REB), blocks (BLK) and steals (STL), given the model outputs a predicted value for each of these measures. This will help us compute a probability that a player will score more or less than a particular value. With these attributes computed in the backend, the team seeks to display these insights in a digestible, interactive format. These visualizations include the results of the model's training, probability distributions for current predictions, and sliders and inputs that help the player identify profitable bets.

Problem Definition

The goal of this project is to develop a predictive model that can forecast a player's future performance to inform statistically optimal basketball betting strategy and provide a visually appealing interactive dashboard of trends in player-level data. We hope that our final product will be a necessary companion for any NBA fan interested in watching basketball and sports betting. The use of interactive visualization around machine learning prediction should help to distill advanced data science concepts for everyone to receive data-driven advice to assist their intuition.

Literature Survey

(Ref. 1) Professors Yu and Stasko design and evaluate two interactive NBA statistic visualization systems, primarily to aid journalists. Many interesting visualization styles are presented, which will be inspirational for our representation. Our project addresses the visualization of predicted statistics, while this paper only addresses previous game data. **(Ref. 2)** The authors analyze various daily fantasy basketball predictions. We seek to perform similar predictive capabilities to the models referenced in the paper. Hence the regression based testing of forecasts presented in the paper would be a good method to validate our own predictions. **(Ref. 3)** The authors of this paper provide many interesting NBA visualizations, primarily at the game level. Taking inspiration, we can display player statistics as a similar time series to team win/loss. In terms of improvement, we will focus on player-level data and also seek to provide simpler visuals, even if at the expense of information quantity. **(Ref. 4)** This paper is a broad review of advances in predicting outcomes of sporting events. It includes data for a broad range of sports and shows that American Football and Basketball had the highest accuracies for outcome prediction. This paper provides excellent insight into what has worked in the past and what might be useful information to include in examining *specific* players' performances. **(Ref. 5)** This paper uses Support Vector Machines (SVMs) to classify the outcome of basketball games. This paper gives detailed methodology on data preprocessing as well as a top-down mathematical explanation on why Hybrid-Fuzzy SVMs are particularly well-equipped to handle problems like this. This is a very narrow paper in terms of the models examined, though. **(Ref. 6)** This paper examines the use of CART and Random Forest techniques to predict basketball outcomes. This shows how to use the Box Score statistics of a team to accurately make informed predictions about the outcome of the game, similar to what the team plans on doing. **(Ref. 7)** This paper talks about different methods for using machine learning to make predictions from the large wealth of data that is available from the NBA. It focuses on predicting a player's future performance, which is very similar to what our group wants to achieve with this project. This paper lacks on the visualization front, which we aim to make a focus of our project. **(Ref. 8)** This paper focuses on methods for the visualization aspect of modeling NBA team and player data. This paper will be very useful for our project, as there are few papers that focus on the visual representation of NBA data. We aim to improve on this by future outcomes of NBA player statistics. **(Ref. 9)** This paper focuses on applying machine learning to three areas with professional basketball: All-Star Prediction, Playoff Prediction and Hot Streak Fallacy. This article gives us helpful information on how to handle NBA data for modeling. We aim to go one step further than this paper and predict player success on a game-by-game level. **(Ref. 10)** This paper provides an interesting machine learning perspective in fusing convolutional networks and random forests in game prediction. We can seek to use a similar ML structure on player prediction. **(Ref. 11)** The author presents a variety of models to use in game prediction. Notably, the use of player statistics as variables is important towards our goal. The paper shortcomings revolve around a lack of data-specific visuals. **(Ref. 12)** The authors implement a strong ML pipeline for game level prediction that could influence our modeling stage, but fall short in describing the techniques used. **(Ref. 13)** This paper showcases the enhancement of predictive models for soccer match outcomes through domain-specific knowledge, focusing on recency feature extraction and rating feature learning. These methodologies emphasize the importance of considering sport-specific dynamics, like team form and opposition strength, to improve model accuracy. Adapting these approaches to basketball could significantly benefit NBA player performance predictions. **(Ref. 14)** This reference provides an example of domain-specific knowledge we could incorporate into our model. Using match data, the authors built a weighted directed graph of players to analyze the offensive behaviors of basketball teams, a method we could utilize to engineer a feature that provides information on the quality of player teams. **(Ref 15.)** This study identified several features that have the greatest predictive power of player performance. The authors found that minutes played, usage percentage, and difference in team quality were the main factors of variance in individual point scores. **(Ref 16.)** This paper compares present data of football teams and their matchups to similar matchups in the past to predict game outcomes. This is useful as in our project we can try comparing past players and current players with similar styles of play to make specific predictions regarding the player. One of the drawbacks of the paper is that it makes a broad prediction of who will win the game compared to our in depth predictions of the specific statistics of each player. **(Ref 17.)** The authors predict soccer match outcomes by assigning ratings based on algorithms for each team. The GAP method that assigns a rating for offense, defense, and home or away games to predict success in a matchup can be a useful way to organize our statistics for each player as it has seen success. The drawback of this paper is that it focuses on overall team statistics and does not take into account specific players that could make a difference. Our data will be player-specific as well. (**Ref 18.)** This paper predicts the rating of a basketball player based on how their team performs when they are and are not on the floor. This idea will be useful to our project as we will need to take into account which players are on the floor as that will affect a player's numbers based on who is injured or not. However, this approach has only been proven to be accurate for offensive statistics. We will need to take into account the opposing team's offensive statistics to predict defensive numbers for players.

Proposed Method

1. Intuition - This project seeks to combine team and player-level statistics as well as state-of-the-art calibration methods to create a more holistic, probability-based model that predicts individual statistics. The team seeks to implement an algorithm to identify profitable wagers based on expected value maximization, and to utilize effective visualization methods to display i) historical data ii) a machine learning model's predictions for player performance, and iii) expected returns of real-time betting lines scraped from the internet.

2. Description - We describe our data collection and processing, model building, training, and selection, and visualization approaches.

A. Data Collection - The data used for this project comes from the NBA statistical database and is accessed via the **`nba_api`** python module. This module came with its share of challenges as the server often became overloaded and the connection timed out making data collection a time-consuming and unreliable process. The api also was not able to gather large amounts of player-data at one-time, meaning the team had to query the database individually for each active player. Server-side improvements on this would greatly increase the speed and reliability of the querying process.

B. Data Processing - Once the player and team-level data was gathered from the API, the following manipulations were performed:
   a. The player data was grouped by player and season, and a running average of quantitative statistics was calculated for each game a player participated in. This helped give a "snapshot" of a player's performance at a particular point in time. The first game of each season was disregarded.
   b. The team-level data was grouped by team and season, and a similar running average of quantitative team box score statistics was calculated for each game the team participated in to help give a snapshot of the team performance at a particular point in time.
   c. Once these running average calculations were made, the player and team data were merged into one dataframe where the player's running average statistics, the player's team running average statistics, and the opposing team's running average statistics formed the features of the model.

C. Machine Learning - In predicting player statistics based on the given features, the following models were trained:
   a. Ridge and Lasso regression were trained with regularization constant 9e-3. The efficiency and explainability of these models were important considerations for model selection.
   b. A K-Nearest Neighbors (KNN) regression model was implemented by selecting the average result among 25 neighbors.
   c. Tree-based methods including AdaBoost and Random Forests were used with 250 trees and exponential loss and 25 trees with a maximum depth of 3 and maximum features of 10, respectively. These models were examined for their flexibility and insights into feature importances.
   d. The team trained a simple regularized Multilayer Perceptron (MLP) neural network model with 2 layers of 50 hidden neurons and Rectified Linear Unit (ReLU) activation, because of its computational efficiency and universal function approximation of nonlinear relationships.

D. Model Selection - Cross-Validation (CV) was used in the model selection process and performance was evaluated with Mean Absolute Error (MAE). Lasso regression was chosen due to its accuracy, interpretability, and efficiency both in training in evaluation. While ensemble and gradient-based models like AdaBoost, Random Forest and MLP took minutes to train, Ridge and Lasso regression needed less than a second. Performance was comparable across all models, though the MLP did generalize the best by a negligible margin after enough training. From all the data, the team synthesized a conditional probability distribution based on true variability of an individual statistic given a prediction from the model. These probabilities were used to determine the *probability* of a player scoring more or less than a certain amount.

Test MAE By Model, Statistic	PTS	REB	AST	STL	BLK
Ridge Regression	4.866	2.031	1.402	0.570	0.744
Lasso Regression	4.866	2.001	1.401	0.572	0.745
Neural Network	**4.589**	**1.998**	**1.399**	**0.561**	**0.731**
KNN Regression	4.943	2.101	1.538	0.610	0.754
AdaBoost	4.712	2.062	1.477	0.590	0.746
Random Forest	4.640	2.048	1.492	0.577	0.737

E. Data Visualization/User Interface - We are building an interactive app that displays a select player's historical statistics and the ML model's predictions for the upcoming game. This allows the user to not only judge player performance based on past events, but also examine the statistical properties gleaned from model-based analysis. The user also has an opportunity to input a player's over/under for a certain statistic, the odds for the bet, and the amount wagered. The application then computes the expected return for a bet based on the probability of success.. We leverage user input of a player and statistic to provide two primary visuals:
   a. A time series plot of the target players' previous ten performances for the target statistic combined with the projection for the upcoming game.
   b. A distribution of all prior performances for a player that was projected the same amount of the target statistic by the model. The model projection and current line from DraftKings Sportsbook will be displayed as well. The percentage of performances based on model projection above/below the given line will function as a confidence level in an over/under recommendation by our model.

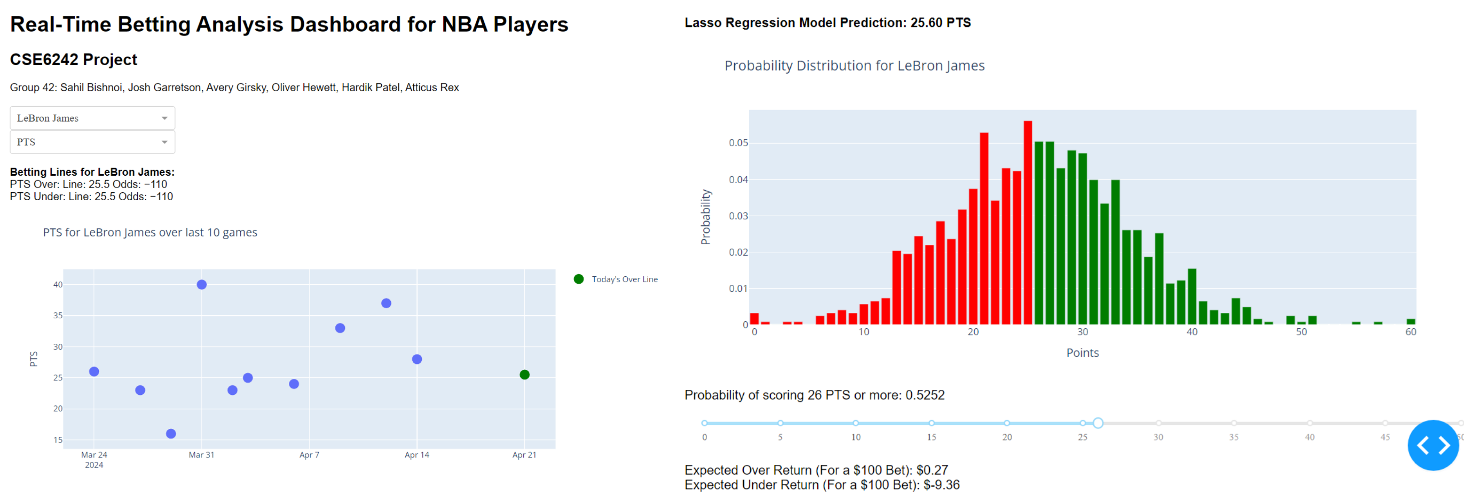

Figure 1: Visualization showcasing Lebron James past games and future prediction for steals

Figure 1 shows selected player **LeBron James** and selected stat **STL**. We observe the DraftKings Sportsbook line, model prediction, previous performances, over/under probabilities, and expected returns.

Experiments and Evaluation

The team sought to answer two questions: how can we best collect and model player and team level data to predict player statistics. We expect that a running average for a player, their team, and their opponent can provide the best results. With these results, the next question becomes whether our model can predict with a high enough accuracy to be confident in odds against a sportsbook in the long run. While this question will be harder to answer within the timeframe of the project, we hope our advanced visualizations provide enough innovation to make our product useful, even if the model does not measure up to those deployed by sportsbooks.

The model was evaluated using CV and holding out 20% of total data to assess how well the model generalized to new data. The model's accuracy in predicting quantitative player statistics was assessed using Mean Absolute Error (MAE) for its better interpretability in the context of the statistic e.g. on average, the model deviates by 4.59 points from the true amount a player actually scores.

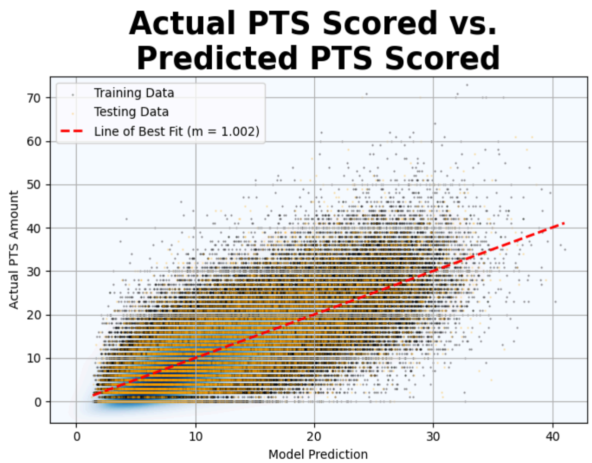

Figure 2: Model points prediction vs. points scored

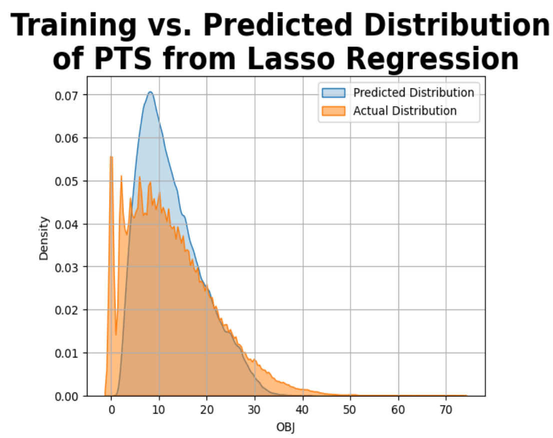

Figure 3: Actual vs. Simulated Points Distribution

Figure 2 Depicts the results of a regression of the previously described features and the number of points a player will score in a given game. The black dots represent the training data and the yellow dots represent the validation/holdout/testing data. Clearly, there is a positive, linear relationship between what the model predicts and what the player ends up scoring. **Figure 3** Demonstrates the differences in distribution between actual PTS outcomes and model predicted outcomes. We see the model acts conservatively, often failing in its prediction of low point outcomes and high point outcomes.

Further, the team was able to approximate probability distributions of the actual counts of a particular statistic, say points, for example, given the model outputted a certain value. For example, the distribution of points a player is likely to score when the model predicts a player will score 5 points is very different from when the model outputs 20 points. Generally speaking, a large number of players score zero points and hence there is a spike at zero for the lower predictions. This will allow the team to effectively produce a CDF against which to test betting lines using the following formula:

$$\mathbb{E}[\text{Bet}] = P_w \cdot W + (1 - P_w) \cdot L$$

where W is the proportion of a bet if the outcome is a win, L is the loss incurred if a bet is lost, and P_w is the probability of the bet winning, which will be calculated from the aforementioned CDFs. If the expected value of a bet is positive, this indicates that, on average, a particular line will be profitable. In our initial tests, simpler machine learning paradigms performed similarly to more sophisticated ones—Lasso Regression worked just about as well as a many-layered regularized neural network. This is desirable for the low-number and interpretability of its parameters.

Through self-assessment of our visualization product, we believe that our project provides significant innovation to the casual sports betting space yet still has room for improvement. The product allows for many user input opportunities, allowing a user to observe information for any current player-statistic combination and over/under probabilities for any statistical output. For a fan that wants to bet on a player-statistic line, they will be able to verify their intuition with our model and better understand through the visual component.

However, the room for improvement comes in the form of how a user can find their desired information. We believe that the product could be improved through a dashboard that specifically highlights bets that the model feels very confident about, or that have high expected outcomes rather than make the user search player by player for these cases. Additionally, we began to include a live component by displaying the stat for a player in their current game (if existing), but this could be expanded by adjusting the model to make live predictions for live in-game betting opportunities.

Conclusions and Discussion

Our project aimed to develop a predictive model for forecasting NBA player statistics with a focus on optimizing betting strategies. A variety of machine learning models were tested, arriving on a LASSO Regression model, used to predict five key player statistics: points, assists, rebounds, blocks, and steals. The decision to select the LASSO model was driven by its balance of accuracy, interpretability, and computational efficiency.

The model outputs were used to generate conditional probability distributions for player performance across each stat, which were then applied to create a decision framework for betting. The probability distribution allowed us to calculate expected return for a given betting line, hence identifying profitable bets. Our interactive dashboard presents these insights in a digestible manner, showing past player performances, model predictions, and the expected returns of given betting lines.This project contributes to the field of sports analytics by integrating game data, player data, probability measures for quantitative statistics, and novel visualization techniques.

One of the primary limitations of our application is that it does not consider the variability introduced by player availability in NBA games. Due to the length of the season, it is common for players to rest or become injured. The absence of a teammate or opponent can significantly affect a player's performance, a feature unseen using our methodology. Further, the model is only trained on regular season statistics, potentially limiting its ability to accurately predict performance during different times of the year e.g. preseason, summer league, play-in, playoffs and championship games.

Looking ahead, several enhancements could be made to improve both the model's accuracy and user experience. Integrating additional data sources such as player injuries and real-time game conditions could provide more complete features and potentially improve prediction accuracy. A leaderboard of bets could show the most promising betting opportunities identified by our model so users can easily identify the best bets of the day. In short, our project presents a step forward in the application of machine learning for sports betting. Sports analytics is a growing field, and there is a lot of potential for impact. We hope that our work will inspire further research and innovation, paving the way for more sophisticated analytical tools in the sports industry.

All team members have contributed a similar amount of effort.

References

1. Fu, Y., & Stasko, J. (2022). Supporting Data-Driven Basketball Journalism through Interactive Visualization. CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3502078
2. Döpke, J., Köhler, T. & Tegtmeier, L. (2023). Are they worth it? – An evaluation of predictions for NBA 'Fantasy Sports'. J Econ Finan. https://doi.org/10.1007/s12197-023-09646-7
3. Chen, W., Lao, T., Xia, J., Huang, X., Zhu, B., Hu, W., & Guan, H. (2016). GameFlow: Narrative Visualization of NBA Basketball Games. IEEE Transactions on Multimedia, 18(11), 2247–2256. https://doi.org/10.1109/tmm.2016.2614221
4. Horvat, T. (2022). The use of machine learning in sport outcome prediction: A Review. WIREs Data Mining and Knowledge Discovery, 12(2). https://doi.org/10.1002/widm.1445
5. Jain, S., & Kaur, H. (2017). Machine learning approaches to predict basketball game outcome. *2017 3rd International Conference on Advances in Computing,Communication & Automation (ICACCA) (Fall)*. https://doi.org/10.1109/icaccaf.2017.8344688
6. Migliorati, M. (2020). Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms. *Electronic Journal of Applied Statistical Analysis*. https://doi.org/doi.org/10.1285/i20705948v13n2p454
7. Nguyen, N. H., Nguyen, D. T., Ma, B., & Hu, J. (2021). The application of machine learning and Deep Learning in sport: Predicting NBA players' performance and popularity. Journal of Information and Telecommunication, 6(2), 217–235. https://doi.org/10.1080/24751839.2021.1977066
8. Shi, Z., Li, M., Wang, M., Shen, J., Chen, W., & Luo, X. (2022). NPIPVIS: A visualization system involving NBA visual analysis and Integrated Learning Model Prediction. Virtual Reality & Intelligent Hardware, 4(5), 444–458. https://doi.org/10.1016/j.vrih.2022.08.008
9. Wang, J., & Fan, Q. (2021). Application of machine learning on NBA data sets. Journal of Physics: Conference Series, 1802(3), 032036. https://doi.org/10.1088/1742-6596/1802/3/032036
10. Kai Zhao, Chunjie Du, & Guangxin Tan, (2023). Enhancing Basketball Game Outcome Prediction through Fused Graph Convolutional Networks and Random Forest Algorithm https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10217531/
11. Junwen Wang. 2023. Predictive Analysis of NBA Game Outcomes through Machine Learning. In The 6th International Conference on Machine Learning and Machine Intelligence (MLMI 2023), October 27-29, 2023, Chongqing, China. ACM, New York, NY, USA, 14 Pages. https://doi.org/10.1145/3635638.3635646
12. Tomislav Horvat, Robert Logozar, Caslav Livada (2023), Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games https://www.mdpi.com/2073-8994/15/4/798
13. Berrar, D., Lopes, P. & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. Mach Learn 108, 97–126. https://doi.org/10.1007/s10994-018-5747-8
14. Tao, Z., Hu, G., & Liao, Q. (2013). Analysis of offense tactics of basketball games using link prediction. 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS), 207-212. https://ieeexplore.ieee.org/document/6607842
15. Casals, Martí & Martinez, Jose. (2013). Modeling player performance in basketball through mixed models. International Journal of Performance Analysis in Sports. 13. 64-82. https://www.tandfonline.com/doi/abs/10.1080/24748668.2013.11868632
16. Leung, C. K., & Joseph, K. W. (2014). Sports Data Mining: Predicting Results for the College Football Games. Procedia Computer Science, 35, 710–719. https://doi.org/10.1016/j.procs.2014.08.153
17. Wheatcroft, E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics*, *7*(2), 77–97. https://doi.org/10.3233/JSA-200462
18. Fearnhead, P., & Taylor, B. M. (2011). On Estimating the Ability of NBA Players. Journal of Quantitative Analysis in Sports, 7(3), 1298–1298. https://doi.org/10.2202/1559-0410.1298