Proposal to Build a NBA Game Predictor*

Rudra Patel[1], Amy Huang[2], Valli Paladugu[3], Sanya Shrivastava[4], and Annalise Smith[5]

***Abstract*— This project focuses on the design and implementation of a machine learning system to predict the outcome of National Basketball Association (NBA) games using historical performance data. By processing publicly available datasets, we aim to identify the features that are the most influential in determining game results. Our system will integrate with a web application for easy use. Beyond building an accurate model, this project emphasizes end-to-end development, with the intended outcome being tool that is both accurate and accessible.**

I. INTRODUCTION

A. *Objective*

The objective of this project is to design and implement a machine learning model that is capable of predicting the outcomes of NBA games by using historical game data and key performance metrics. We aim to build a robust predictive system by analyzing publicly available datasets and identifying the most influential features. This system will include a user-friendly web application, where users can select two NBA teams and receive predictions on the likely winner and loser. Ultimately, our project is geared towards creating an engaging tool for basketball enthusiasts and analysts.

B. *Motivation*

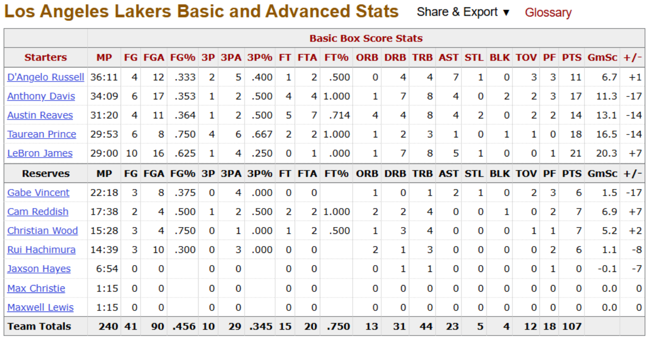
The National Basketball Associated (NBA) is the United States' premier competitive basketball league. Each year, millions of fans tune in to watch over 1,000 regular season games. With this large fan base, the NBA is one of the largest sports leagues globally, and the outcomes of its game are of large importance to many people.

Predicting the results of NBA games is difficult because it requires a combination of sports knowledge and technical skills. We chose this project because it is a way for us to apply our team's knowledge in data science, machine learning, and web development to a real-world problem. The NBA is a perfect candidate for this analysis because it has a large amount of publicly available data. The goal is not just to build a model that makes accurate predictions, but to better understand the factors that actually influence the outcomes of games. In doing so, we will have the chance to build a full-stack project from start to finish, allowing us to practice data collection, data cleaning, model training, and web application deployment.

II. DATA OVERVIEW

Because we must train a machine learning model, it is critical that we have high-quality data that we can train off of. For this project, we will obtain historical NBA game data from publicly available sources. These sources include Basketball Reference, Kaggle datasets, and the official NBA statistics database.

We will most likely opt to collect most of our data from Basketball Reference because of its easy-to-parse HTML files. For every NBA season, Basketball Reference has a HTML file for each game in the season that presents the game's data in tables. You can collect a plethora of data points, but for this project, we will collect the number of field goals, field goals attempted, field goal percentage, three pointers, three pointers attempted, three pointer accuracy, free throws, free throws attempted, free throw accuracy, offensive rebounds, defensive rebounds, assists, steals, blocks, turnovers, personal fouls, and points. This data gives us insights into the offensive and defensive performance of a team and the final outcome of the game, hopefully allowing us to determine which factors matter the most and predict the outcome of future games.



Los Angeles Lakers Basic and Advanced Stats

	Basic Box Score Stats																			
Starters	MP	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	GmSc	+/-
D'Angelo Russell	36:11	4	12	.333	2	5	.400	1	2	.500	0	4	4	7	1	0	3	3	11	6.7	+1
Anthony Davis	34:09	6	17	.353	1	2	.500	4	4	1.000	1	7	8	4	0	2	2	3	17	11.3	-17
Austin Reaves	31:20	4	11	.364	1	2	.500	5	7	.714	4	4	8	4	2	0	2	2	14	13.1	-14
Taurean Prince	29:53	6	8	.750	4	6	.667	2	2	1.000	1	2	3	1	0	1	1	0	18	16.5	-14
LeBron James	29:00	10	16	.625	1	4	.250	0	1	.000	1	7	8	5	1	0	0	1	21	20.3	+7
Reserves	**MP**	**FG**	**FGA**	**FG%**	**3P**	**3PA**	**3P%**	**FT**	**FTA**	**FT%**	**ORB**	**DRB**	**TRB**	**AST**	**STL**	**BLK**	**TOV**	**PF**	**PTS**	**GmSc**	**+/-**
Gabe Vincent	22:18	3	8	.375	0	4	.000	0	0		1	0	1	2	1	0	2	3	6	1.5	-17
Cam Reddish	17:38	2	4	.500	1	2	.500	2	2	1.000	2	2	4	0	0	1	0	2	7	6.9	+7
Christian Wood	15:28	3	4	.750	0	1	.000	1	2	.500	1	3	4	0	0	0	1	1	7	5.2	+2
Rui Hachimura	14:39	3	10	.300	0	3	.000	0	0		2	1	3	0	0	0	0	2	6	1.1	-8
Jaxson Hayes	6:54	0	0		0	0		0	0		0	1	1	0	0	0	0	1	0	-0.1	-7
Max Christie	1:15	0	0		0	0		0	0		0	0	0	0	0	0	0	0	0	0.0	0
Maxwell Lewis	1:15	0	0		0	0		0	0		0	0	0	0	0	0	0	0	0	0.0	0
Team Totals	**240**	**41**	**90**	**.456**	**10**	**29**	**.345**	**15**	**20**	**.750**	**13**	**31**	**44**	**23**	**5**	**4**	**12**	**18**	**107**		

Fig. 1. *An example data table from Basketball Reference*

III. WORK DISTRIBUTION

Because this project has two main components, we will split off into two teams based on our expertise. The first team, consisting of Amy, Rudra, and Valli, will be responsible for making the machine learning model. Rudra, with his expertise in web scraping, will collect the data from online sources. Amy and Valli, with their knowledge of machine

*This work was not supported by any organization

[1] R. Patel is with the Department of Eletrical Engineering & Computer Science, University of Tennessee rpatel12@vols.utk.edu

[2] A. Huang is with the Department of Eletrical Engineering & Computer Science, University of Tennessee ahuang16@vols.utk.edu

[3] V. Paladugu is with the Department of Eletrical Engineering & Computer Science, University of Tennessee spaladu1@vols.utk.edu

[4] S. Shrivastava is with the Department of Eletrical Engineering & Computer Science, University of Tennessee sshriva2@vols.utk.edu

[5] A. Smith is with the Department of Eletrical Engineering & Computer Science, University of Tennessee asmit494@vols.utk.edu

learning and deep learning, will take the data and train a model with it.

The second team will be responsible for the website. Analise and Sanya, with their expertise in web design and web development, will create a frontend that consumes the model and allows a player to select two NBA teams. Then, it will call upon the model to predict the outcome based on historical data.

IV. Timeline of Milestones

To make sure the project progresses throughout the semester, we plan to meet the milestones and deadlines outlined in Table I. These planned deadlines may be changed in the future based on the class deadlines for the final report and class presentation.

TABLE I
Project Milestones and Deadlines

Milestone	**Deadline**
Data collection	October 3rd
Create wireframe design of frontend	October 3rd
Research machine learning models	October 14th
Implement base machine learning model	October 24th
Implement base frontend	October 24th
Analyze results and improve model	November 7th
Connect the model and website	November 13th
Write 4–6 page final report	November 18th
Create and deliver class presentation	Late November

V. Expected Outcome and Conclusion

We expect to create a website where a user can choose any two NBA teams to generate a game matchup. This website will connect with a machine learning model in the backend to predict the winner of the matchup with over 50% accuracy. The prediction model will be a binary classification model using effective algorithms such as XGBoost or Convolutional Neural Networks (CNNs). Given the data available to us from past NBA games, the algorithms will determine which features are the most impactful and necessary to consider when predicting the outcome of a game.

In conclusion, through this project, we will combine data scraping, data science, machine learning, and web development to address the complex problem of predicting the outcomes of NBA games. By leveraging high-quality online datasets and experimenting with various machine learning models, we expect to create an accurate system that is suitable for use by fans and analysts to explore hypothetical match-ups. Upon the completion of model and web application, the team will have greatly improved their skills in all aspects of the data science life cycle, from data preparation to model deployment.