Predictive Analysis of NBA Championship Using Machine Learning Algorithms

EM – 623 Data Science and Knowledge Discovery

Prof. Feng Liu

Ishan Bhatt

Priya Rao

Table of Contents

Introduction	
Uses of Predictive Analysis of NBA Championship	
Problems faced in NBA Championship:	
Solutions by using Predictive Analysis of NBA championship:	
Dataset Description	
Data Visualization	
Bar Plot	
Scattered Plot	
Line Plot	
Heat Map Plot	
Model Selection	
Model Evaluation	
Summary	
Refrence	

Introduction

The NBA has been one of the most popular sports leagues in the world for decades. Winning a championship is the ultimate goal for every NBA team, and the pursuit of this goal has led to some of the most intense and exciting moments in sports history. However, predicting which team will win the championship in any given season can be a difficult task. That's where machine learning comes in.

In this project, we will explore the use of machine learning algorithms to predict the NBA championship winner for a given season. We will start by gathering data on various factors that may contribute to a team's success, such as regular season record, player statistics, and team performance in previous playoff rounds. We will then preprocess and clean the data, and use feature engineering techniques to create meaningful features for the machine learning models.

Uses of Predictive Analysis of NBA Championship

Picking the Best Team: The NBA championship, the top honor in the world of professional basketball, is given to the squad that wins the NBA Finals. It means that the winning team has established itself as the league's finest for that specific season.

Reputation and status: The NBA champion team, its players, coaches, and entire organization gain reputation and status. Their standing in the sports world frequently rises as a result, solidifying their position in the annals of basketball.

Player Legacies: The NBA title has a big impact on how players are remembered. A player's reputation can be improved, they can be considered one of the all-time greats, and it can increase their chances of getting inducted into the Basketball Hall of Fame if they win numerous championships.

Financial Effects: The NBA club and its players may benefit financially from winning the title. Victories in championships frequently result in an increase in the organization's revenue generating overall as well as in merchandise sales, ticket demand, sponsorship opportunities, and so on. Following a championship victory, athletes may also bargain for better compensation or secure sponsorship deals.

Fan Engagement and City Pride: NBA victories strengthen bonds between neighborhoods and instill a spirit of enthusiasm and pride among supporters. The city or region that the winning team represents may see increased fan participation, memorable experiences, and long-lasting effects as a result of the team's title.

Legacy of the Franchise: An NBA title can influence a franchise's legacy. It can contribute to the team's entire history and reputation, aid in the development of a winning culture, and draw in top talent in the future.

Inspiration and Motivation: The NBA championship provides players, teams, and supporters with inspiration and motivation. It spurs athletes to raise their game, drives teams to strive for excellence, and inspires fans to fervently support their favourite players and teams.

Problems faced in NBA Championship:

Like any professional sports league, the NBA has its share of difficulties. Here are some of the frequent problems the NBA has:

Competitive Balance: The NBA constantly faces the difficulty of keeping teams competitive with one another. Talent, market, and financial disparities can lead to imbalances that result in dominant teams and uncompetitive contests.

Referee choices and Officiating: Players, coaches, and fans may get dissatisfied with referee choices due to their subjectivity and occasionally arising controversy over officiating. The league continues to struggle with ensuring impartiality and accuracy in the officials' calls.

Globalization and International Expansion: The NBA continues to draw international players and broaden its worldwide presence. However, when advertising the league globally, difficulties can occur due to cultural differences, logistical problems, and accommodating varied fan bases.

Salary Cap and Financial Management: To ensure competitive balance and financial viability, the NBA operates under a salary cap system. Complex jobs can include managing player pay, boosting revenue sources, and making sure all teams are financially viable.

Fan Engagement and Changing Viewing Patterns: The NBA, like many sports leagues, must find ways to engage fans and modify their viewing patterns. The league needs to be creative to attract and hold fan attention in the age of streaming services and digital media.

Player mobility and Team Dynamics: Frequent player mobility through trades, free agency, and player empowerment can interfere with team dynamics and fan loyalty. In an era of player movement, it can be difficult to maintain stability and preserve the identity of teams.

Solutions by using Predictive Analysis of NBA championship:

These are a some of the obstacles and issues the NBA has to deal with using Predictive Analysis. In order to provide players, clubs, and fans with the greatest experience possible, the league continuously resolves these problems. In order to make wise judgments and raise the likelihood of winning the championship, predictive analysis might be a useful tool in the NBA. Here are a few ways that predictive analysis can assist in addressing issues and locating solutions:

Performance Analysis: Predictive algorithms can assess player and team performance information to pinpoint strengths, weaknesses, and trends. Coaches can use this information to design game plans that are specifically suited to taking advantage of opponents' weaknesses, as well as to optimize lineups and make tactical modifications.

Predicting Team Success: By examining historical data and different performance criteria, predictive models can calculate a team's likelihood of success in terms of regular-season performance or playoff outcomes. This knowledge can help teams set reasonable objectives, modify their rosters, and allocate their resources wisely.

Fan interaction and Marketing: Teams can use predictive data to pinpoint fan preferences, interests, and interaction trends. To increase engagement and fan satisfaction, teams can customize marketing campaigns, promotional events, and fan experiences by evaluating data on fan behaviour, social media interactions, and ticket sales.

Revenue Optimization: Predictive models can help organizations estimate revenue sources like ticket sales, item sales, and sponsorships. Teams can optimize pricing strategies, create focused marketing activities, and maximize income generation prospects by evaluating market trends, fan behaviour, and other factors.

Game Simulation and Strategy: Predictive analysis can model game situations and assist teams in assessing various strategies, such as line-up combinations, offensive plays, or defensive methods. Teams can find the best strategy and sharpen their decision-making during games by simulating different scenarios and outcomes.

In many facets of the NBA, predictive analysis can help teams overcome obstacles and improve their chances of winning championships by offering insightful information and guiding decision-making. Predictive analysis should, however, be utilized in conjunction with professional knowledge and judgment from coaches, players, and management because it is not 100% accurate.

Dataset Description

We used two datatset for our project one is historical data of NBA statistics from 1997 to 2022 and the second dataset is the current season's dataset. The below figure shows how our dataset looks like:

	No	Team	G	GW	GL	WP	Min	Pts	Reb	Ast	...	3PA	Pct.1	FTM	FTA	Pct.2	Eff	Deff	YS	YE	Championship
0	1	Chicago	103	62	20	0.756	48.4	96.0	44.1	23.1	...	12.0	0.323	18.7	25.2	0.741	111.6	17.5	1997	1998	Won
1	2	Utah	102	62	20	0.756	48.3	98.6	40.8	24.7	...	8.4	0.368	23.8	30.9	0.768	116.3	17.5	1997	1998	Lost
2	3	Indiana	98	58	24	0.707	48.4	95.3	38.6	22.4	...	12.9	0.387	20.3	26.6	0.763	109.0	12.2	1997	1998	Lost
3	4	L.A.Lakers	95	61	21	0.744	48.3	104.8	42.9	24.3	...	17.3	0.350	22.8	33.7	0.675	120.8	13.2	1997	1998	Lost
4	5	Seattle	92	61	21	0.744	48.2	100.0	38.5	24.0	...	19.1	0.390	18.4	25.4	0.724	114.5	9.4	1997	1998	Lost

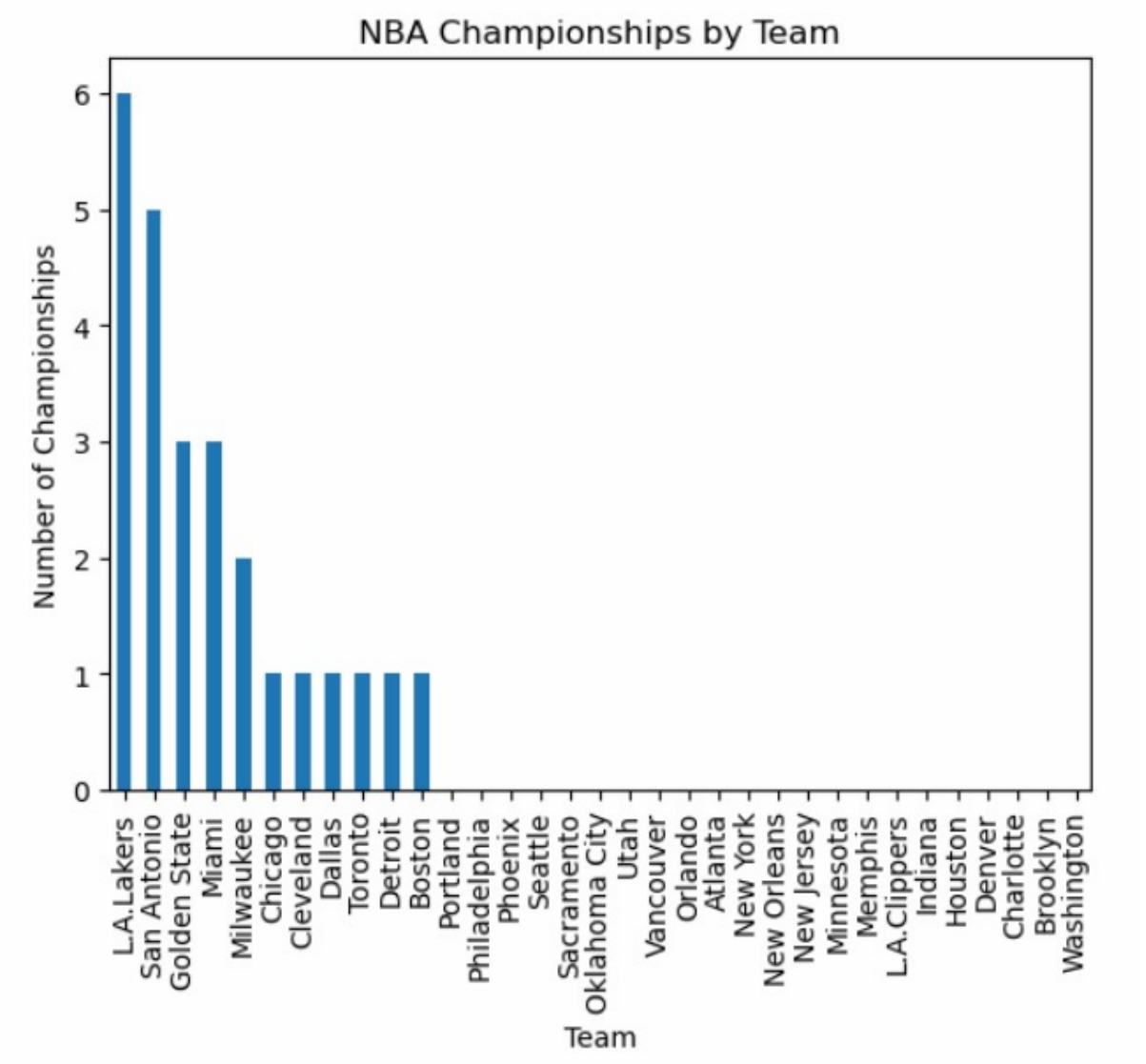
The dataset contains 30 columns in total, ranging from name of teams that played from 1997 to 2022 in NBA. Number of games played, games won, games lost and win percentage each year by each team. There are average data of minutes played, points scored that season and assists scored. Then on there are offensive statistics of team of that year as well as defensive statistics such as Rebounds, 3 points attempted, 3 points scored, etc.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 725 entries, 0 to 724
Data columns (total 30 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   No            725 non-null    int64
 1   Team          725 non-null    object
 2   G             725 non-null    int64
 3   GW            725 non-null    int64
 4   GL            725 non-null    int64
 5   WP            725 non-null    float64
 6   Min           725 non-null    float64
 7   Pts           725 non-null    float64
 8   Reb           725 non-null    float64
 9   Ast           725 non-null    float64
 10  Stl           725 non-null    float64
 11  Blk           725 non-null    float64
 12  To            725 non-null    float64
 13  Pf            725 non-null    float64
 14  Dreb          725 non-null    float64
 15  Oreb          725 non-null    float64
 16  FGM           725 non-null    float64
 17  FGA           725 non-null    float64
 18  Pct           725 non-null    float64
 19  3PM           725 non-null    float64
 20  3PA           725 non-null    float64
 21  Pct.1         725 non-null    float64
 22  FTM           725 non-null    float64
 23  FTA           725 non-null    float64
 24  Pct.2         725 non-null    float64
 25  Eff           725 non-null    float64
 26  Deff          725 non-null    float64
 27  YS            725 non-null    int64
 28  YE            725 non-null    int64
 29  Championship  725 non-null    object
dtypes: float64(22), int64(6), object(2)
memory usage: 170.0+ KB
```

The below figure shows the number of entries, columns, the datatypes used and the memory usage. In the dataset there are 725 entries ranging from 0 to 724 and a total of 30 columns. There are 2 types of variables used the int and the float type of variable. The total memory usage is 170.0+ Kb. There are 0 null values in the entire dataset.

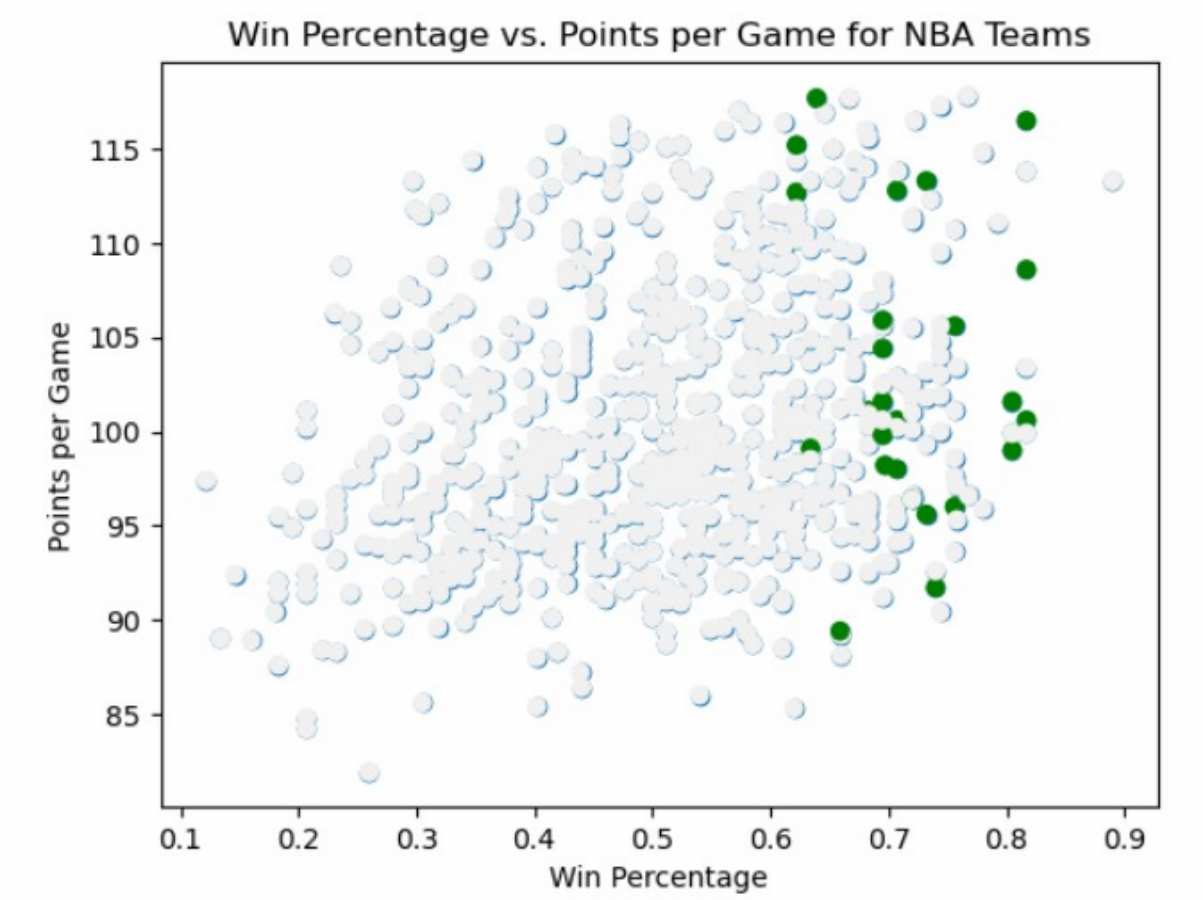
Data Visualization

Bar Plot



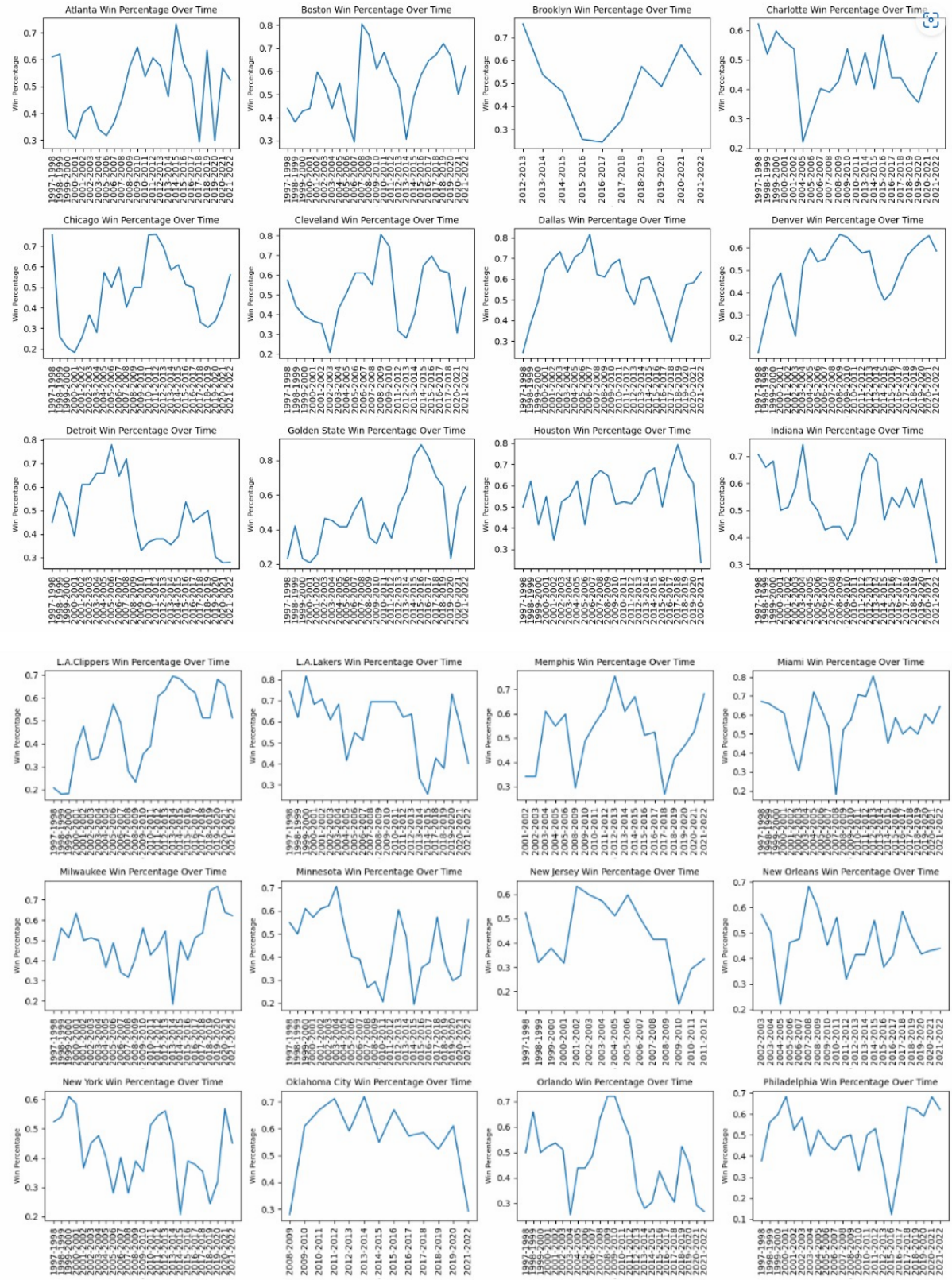
The above figure which is a histogram shows a graph of team vs the number of championships won by the team from 1997-2022.The teams Boston, Detroit, Toronto, Dallas, Cleveland and Chicago won the championship once. The team Milwaukee won the championship twice, Miami and Golden State won the championship thrice, San Antonio five times and LA Lakers six times.

Scattered Plot



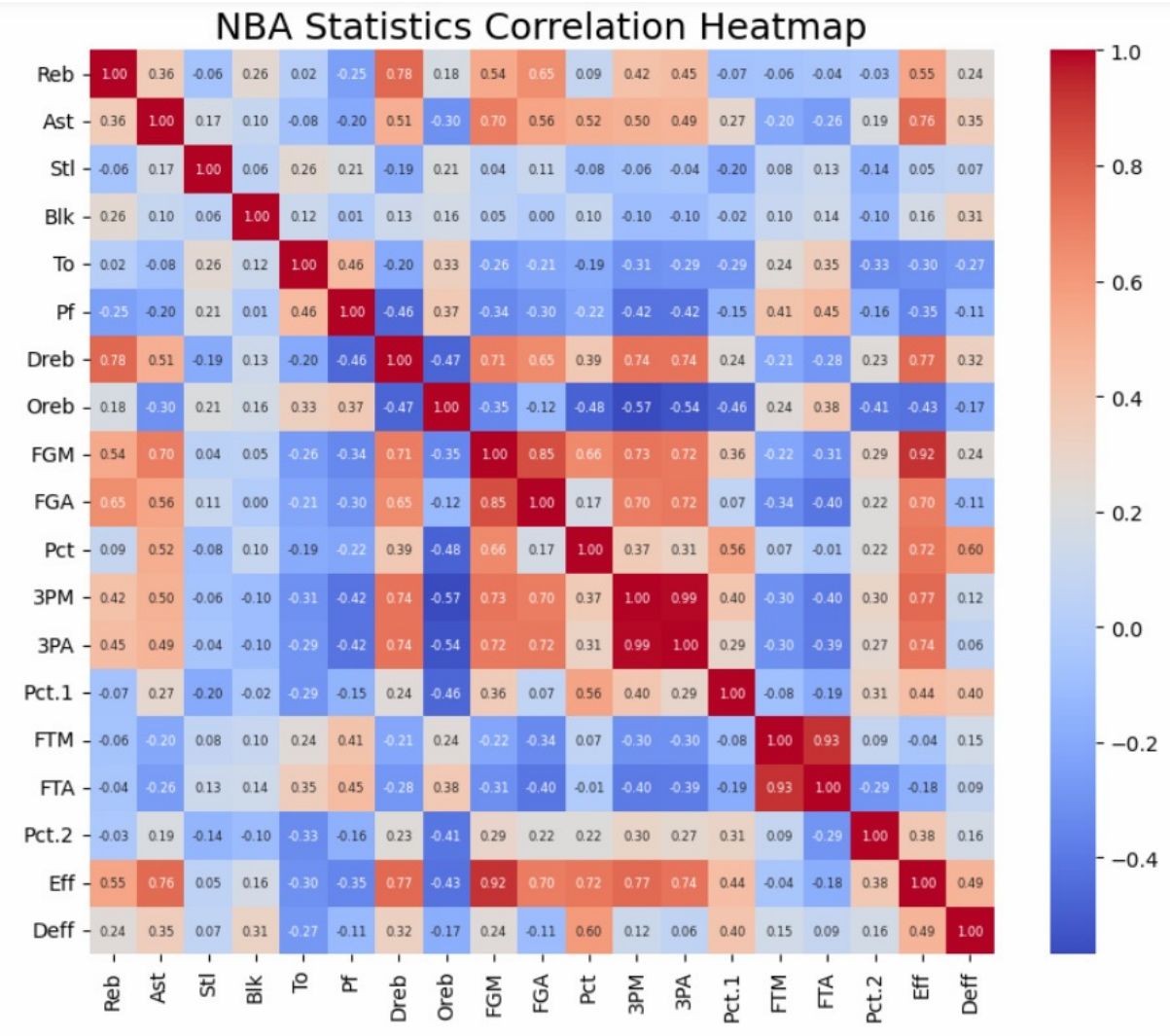
The above scattered plot shows the relation between win percentage of teams over the years and average points per game. The green plots displays the team that won championship. It can be easily identified from the plot that higher the number of points per game, higher winning percentage. Also as seen majorly the teams that won championships has higher win percentage and have better points per game throughout the season.

Line Plot


Atlanta Win Percentage Over Time
Boston Win Percentage Over Time
Brooklyn Win Percentage Over Time
Charlotte Win Percentage Over Time
Chicago Win Percentage Over Time
Cleveland Win Percentage Over Time
Dallas Win Percentage Over Time
Denver Win Percentage Over Time
Detroit Win Percentage Over Time
Golden State Win Percentage Over Time
Houston Win Percentage Over Time
Indiana Win Percentage Over Time
L.A.Clippers Win Percentage Over Time
L.A.Lakers Win Percentage Over Time
Memphis Win Percentage Over Time
Miami Win Percentage Over Time
Milwaukee Win Percentage Over Time
Minnesota Win Percentage Over Time
New Jersey Win Percentage Over Time
New Orleans Win Percentage Over Time
New York Win Percentage Over Time
Oklahoma City Win Percentage Over Time
Orlando Win Percentage Over Time
Philadelphia Win Percentage Over Time
Win Percentage

The above graph shows the performance of each team over the period of years from 1997-2022. In the first graph team atlanta shows a good performance in the years 2014-2015 and 2018-2019. Team Boston shows good performance in the year 2008-2009 and 2018-2019. Team Brooklyn shows a rise in performance in the years 2012-2013 and 2020-2021. The team of Chicago has won in the years 1997-1998 and 2011-2012. Team Cleveland shows a rise in performance in the years 2008-2009 and 2015-2016. Dallas has performed well in the year 2005-2006 and there was a dip in performance in the year 1997-1998 and 2017-2018.The performance of the team Houston is varying with respect to years. Team Houston has won the game in the years 2017-2018. There are such analysis for the entire teams for better visualisation of performance of each team.

Heat Map Plot



Heatmaps can be useful for finding correlations between NBA statistics, including rebounds (Reb), assists (Ast), steals (Stl), blocks (Blk), turnovers (To), personal fouls (Pf), defensive rebounds (Dreb), offensive rebounds (Oreb), etc. Here the red blocks dignifies the dominating statistics with compared to other where as the blue block dignifies the non-dominating statistics.

Model selection

Machine learning predictive models were utilized to predict NBA championships by analysing large amounts of historical data and identifying patterns and relationships between different variables. By processing vast amounts of data quickly and accurately, machine learning models produced more reliable predictions compared to human analysis. These models can also save time and resources by automating the analysis of large datasets.

Moreover, machine learning models can adapt to new data and adjust their predictions accordingly, which can improve their accuracy over time. This makes them particularly useful for predicting outcomes in dynamic environments such as the NBA, where team performance can change rapidly due to player injuries, trades, or changes in strategy.

We used five models for the predictive analysis which were Decision Tree Classifier, Random Forest Classifier, K-Neighbours classifier, Gaussian Classifier and SV Classifier. And they produced the following results:

```
Model: DecisionTreeClassifier
Accuracy: 0.9379
Confusion Matrix:
[[134   7]
 [  2   2]]

Model: RandomForestClassifier
Accuracy: 0.9517
Confusion Matrix:
[[138   3]
 [  4   0]]

Model: KNeighborsClassifier
Accuracy: 0.9724
Confusion Matrix:
[[141   0]
 [  4   0]]

Model: GaussianNB
Accuracy: 0.8276
Confusion Matrix:
[[118  23]
 [  2   2]]

Model: SVC
Accuracy: 0.9724
Confusion Matrix:
[[141   0]
 [  4   0]]
```

For each model, the code reports the accuracy of the model on the data, which is the proportion of correctly classified samples out of the total number of samples. The code also reports the confusion matrix for each model, which is a table that shows the number of true positive, true negative, false positive, and false negative predictions made by the model.

We can see that the K-Neighbours Classifier and SVC models have the highest accuracy, with both achieving an accuracy of 0.9724. The Decision Tree Classifier and Random Forest Classifier models also have high accuracy, with 0.9379 and 0.9517 respectively. However, the Gaussian NB model has a lower accuracy of 0.8276, indicating that it may not be the best choice for this dataset. Additionally, we can see from the confusion matrices that the K-Neighbours Classifier and SVC models have no false positives, while the other models do have false positives.

The Random Forest Classifier had a lower accuracy than the K-Nearest Neighbours Classifier and the Support Vector Machine (SVM), but it was still chosen over The Decision Tree Classifier and the Gaussian Naive Bayes (GNB) Classifier was because accuracy is an important metric, but it is not the only factor that determines the choice of a model.

The SV Classifier and KN Classifier had a high accuracy on the training data, but it performed poorly on the test data, indicating that it might have overfit the training data. In contrast, the Random Forest Classifier uses multiple decision trees and ensemble methods to reduce overfitting and improve generalization performance.

Model Evaluation

As we used Random Forest Classifier, we used the historical data from 1997 to 2022 of NBA Team's performance and used it on current season's statistics. With Random Forest Classifier it gave us the following results:

	Team	Predicted Champion
0	Boston	Won
1	Golden State	Lost
2	Houston	Lost
3	Utah	Lost
4	Philadelphia	Lost
5	Toronto	Lost
6	New Orleans	Lost
7	Indiana	Lost
8	Milwaukee	Lost
9	Washington	Lost
10	Oklahoma City	Lost
11	San Antonio	Lost
12	Minnesota	Lost
13	Miami	Lost
14	Portland	Lost
15	Denver	Lost
16	L.A.Lakers	Lost
17	L.A.Clippers	Lost
18	Detroit	Lost
19	Charlotte	Lost
20	New York	Lost
21	Brooklyn	Lost
22	Dallas	Lost
23	Orlando	Lost
24	Memphis	Lost
25	Atlanta	Lost
26	Sacramento	Lost
27	Chicago	Lost
28	Utah Jazz	Lost
29	Washington Wizards	Lost

The Boston Celtics have a better probability of winning the NBA championship this season than the other teams competing in the current season, according to Random Forest Classifier's Predictive Analysis on NBA Championship utilizing Machine Learning Models with historical data utilized on the current season.

Summary

As a students in a Data Science and Knowledge Discovery class, We had the opportunity to learn a variety of techniques related to data analysis, statistical modelling, and machine learning. These techniques can be applied in many different industries and domains, making this class a valuable foundation for my future career.

Some of the topics covered in this class include data pre-processing and cleaning, exploratory data analysis, statistical modelling, machine learning, data visualization, and data ethics and privacy. Data pre-processing and cleaning involves techniques for handling missing data, dealing with outliers, and transforming data into a format suitable for analysis. Exploratory data analysis involves techniques for visualizing and summarizing data to gain insights into patterns, trends, and relationships. Statistical modelling involves techniques for building models to capture relationships between variables and making predictions or inferences from the data. Machine learning involves techniques for building models that can learn from data and make predictions or decisions based on new data. Data visualization involves techniques for creating effective visualizations to communicate insights and findings to others. Data ethics and privacy involves considerations for ethical and responsible use of data, including issues related to privacy, bias, and fairness.

To keep up with emerging technologies and techniques, we may also need to learn about deep learning, natural language processing, big data technologies, explainable AI, and responsible AI. Deep learning involves techniques for building neural networks with multiple layers that can learn more complex relationships between variables. Natural language processing involves techniques for processing and analysing text data, such as sentiment analysis, topic modelling, and language translation. Big data technologies involve techniques for handling and analysing large-scale datasets, such as distributed computing, cloud computing, and NoSQL databases.

In conclusion, a Data Science and Knowledge Discovery class provided a strong foundation for building skills in data analysis and modelling that can be applied in many different industries and domains. By learning these techniques and keeping up with emerging technologies, we can stay relevant and valuable in the ever-evolving field of data science.

Reference

NBA Team Stats - dataset by etocco | data.world

2022-23 NBA Power Index | ESPN

2022-23 NBA Player Stats: Totals | Basketball-Reference.com

National Basketball Association - Wikipedia

How to Build a Predictive Model for NBA Games – Predictive Hacks

2021-22 NBA Predictions | FiveThirtyEight

How data analytics is revolutionizing the NBA - Digital Innovation and Transformation (harvard.edu)

Predicting NBA Win Percentage. How I built a multiple regression model… | by Robert Alterman | Towards Data Science

What is the best advanced statistic for basketball? NBA executives weigh in | HoopsHype