Data Mining techniques Used for Accurate Basketball Outcome Forecasting

Dissertation submitted in part fulfilment of the requirements for the degree of Masters in Data Analytics At

Dublin Business School

by Hrushikesh Nitin Mate

Supervisor : - Mehran Rafiee

January 2025

Declaration

I, Hrushikesh Nitin Mate, declare that this research is my original work and that it has never been presented to any institution or university for the award of a Degree or Diploma. In addition, I havereferenced correctly all literature and sources used in this work and this work is fully confidential with the Dublin Business School's academic honesty policy.

Signed: Hrushikesh

Roll No. : 20025400

Date: 6 January 2025

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my thesis advisor, Mr. Mehran Rafiee , for their invaluable guidance, continuous support, and encouragement throughout the duration of my research. Their expertise and insightful feedback have been instrumental in shaping this thesis, and I am deeply grateful for their patience and mentorship.

My deepest gratitude goes to my family and friends, who have been my pillars of support throughout this journey. Their unwavering belief in my abilities and their constant encouragement have motivated me to overcome challenges and strive for excellence.

Lastly, I extend my gratitude to all those who have, directly or indirectly, contributed to the successful completion of this thesis. Your support and encouragement mean the world to me.

Thank you all.

List of Figures

List of Tables

List of Abbreviations

NBA –	National Basketball Association
LGBM –	LightGBM Classifier
ML –	Machine Learning
ELM –	Extreme Learning Machine
MARS –	Multivariate Adaptive Regression Splines
KNN –	K-Nearest Neighbors
XGBoost –	eXtreme Gradient Boosting
SGB –	Stochastic Gradient Boosting
PER –	Player Efficiency Rating
TS –	True Shooting Percentage
WS –	Win Shares
SMOTE –	Synthetic Minority Oversampling Technique
JOUS –	Jittering with Over/Undersampling
TP –	True Positive
TN –	True Negative
FP –	False Positive
FN –	False Negative
ROC –	Receiver Operating Characteristic
AUC –	Area Under the Curve

ABSTRACT

Predicting outcomes of National Basketball Association (NBA) basketball games is a very challenging task due to the complexity of game dynamics and so many influencing factors like team performance, player statistics and historical type of data. This study concentrates on the probability of the result of NBA games following the machine learning models with data from the 2019 to 2024. There are some traditional models of sports predictions have issues with big data and data with non- linear relationships between variables, besides, there is a shift in the structure and performance of the teams and the players over the seasons. In order to address these challenges this study work uses both traditional and advanced machine learning algorithms which includes Logistic Regression, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, CatBoost Classifier and LightGBM Classifier. The study also presents a novel Hybrid Ensembled Classifier where multiple base classifiers (LGBM, CatBoost, RandomForest) are stacked, and an AdaBoost meta-classifier is applied to improve the prediction performance. The above mentioned models were assessed using the different model assessment parameters such as accuracy, confusion matrix, and the classification report. The Hybrid Ensembled Classifier proved to be most accurate with major accuracy enhancements of 86 to 95 percent from the separate models. This work also alleviates traditional issues and challenges in using ensemble methods and sophisticated gradient bootstrap methods for application of predictive analytics in NBA games. The study therefore highlights the possible use of machine learning in sports statistics by offering some valuable type of data for coaches, analysts and fans.

Keywords: NBA Game Prediction, Machine Learning, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier

Table of Contents

Chapter 1 Introduction

1.1 Background

The field of sports analytics has grown and evolved a lot in recent years which is been driven by some advancements in data analytics. Basketball is one of the most popular games in the world Li and Xu, (2021), and this game possesses a highly ordered structure of play, numerous and diverse data sources, and a high degree of tactically related competitive interaction inherent in team-and-player level performances Sousa (2022). Data analytics has outgrown statistic management in sports and is now forecast and real time decision support system. Being a professional league packed with years of comprehensive and real time data NBA basketball serves as a perfect case when explaining how analytical insight can decode and predict intricate patterns associated with such sport. Business intelligence is the process of gathering, processing and managing extensive structured and unstructured data to make valuable insights Adewusi et al. (2024). In this study, it serves greatly in analyzing game statistics, performers, and interactional behaviours to generate learning models to predict games results. Through the utilization of advanced computational tools for data analysis, this work also highlights how, indeed, numerical analysis offers germane solutions that can be adapted into strategies to be implemented by the players, managers, trainers, and other members across the football fraternity. As one of the disciplines of big data analysis, the concepts of machine learning provide new approaches and methods for setting improved aggregated and more accurate foreseeable measures. This background section also points to the interdisciplinarity of the research by connecting sports, data, and decisions. Analytics is rapidly becoming a critical and prominent factor in professional team sports where benefits can include fan engagement, simply put, while competitive benefits include strategic development Herberger and Litke, (2021). It is noteworthy that the focal concept of the study, the heterogeneous integration of NBA game analytics followed by advanced analytic models, coherently captures with the current digital progression in the world of sports that demand the synergy of domain expertise and data analytics to develop novel methodologies and paradigms.

1.2 Objectives of the Research

There are some objectives of this research which are as follows:

1. To develop and evaluate machine learning models for predicting NBA basketball game outcomes.
2. To compare the performance of various classification models, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, CatBoost, LightGBM, and Hybrid Ensembled Classifier.
3. To demonstrate the effectiveness of stacking and ensemble methods in improving prediction accuracy.

1.3 Research Questions

This study is having some research questions are as follows:

1. Why does the Hybrid Ensembled Classifier outperform individual models in predicting NBA game results?
2. What role do advanced ensemble methods like CatBoost, LightGBM, and AdaBoost play in enhancing prediction accuracy?
3. How do different machine learning models compare in terms of accuracy, reliability, and efficiency for predicting NBA game outcomes?

1.4 Significance of the study

Making accurate predictions on NBA basketball games using machine learning is an area of paramount importance both in sport analysis and also in the general field of predictive analytics. Specifically, highly precise forecasting in sports can drastically change the ways strategies for games are being created and watched, as well as numerous sports betting platforms. Employing historical game data, this work addresses the question of how team performs, which player is contributing, and the resulted game outcomes scientifically with effective utilization of statistics. These predictions can then be employed by coaches and analysts so as to gain proper insight into strengths and weaknesses so as to inform decisions over player substitutions or game strategy. Besides, the increased interest in the fantasy leagues and betting gives emphasis on accurate models that support the decision makers with certain statistical background by increasing the chances of success. Apart from the implications related to sports, this work shows how machine learning methods, including ensemble models and hybrid classifiers, can be used in practice. It overcomes the problem that increases in the scale and variety of data and the coupling relationships between them in big data applications make it difficult to apply traditional analytical theories and solutions. The best features explained in this work: CatBoost, LightGBM, and a Hybrid Ensembled Classifier demonstrate modern tendencies in the use of algorithms for predictive analytics emphasizing its applicability to various fields like finance, healthcare, and marketing. Furthermore, this research also explains the methodology of data preprocessing, exploration data analysis, and performance evaluation that constitute a broader guide to doing machine learning project. It also responds to the problem posed by the need to gain precision, speed and coverage when it comes to prediction models, together with more specific advice on how to deal with structured and categorical data. Last of all, the outcomes of the present research advance academic knowledge in the field and reveal several directions for further investigation, such as the integration of fresh real-time data, as well as the application of more sophisticated deep learning methodologies.

1.5 Structure of the Report

This section is going to explain structure of the report:

1. **Chapter 1 Introduction:** Provides an overview of the NBA game prediction problem and the significance of using machine learning techniques. This chapter will also defines the objectives, scope to predict NBA game outcomes using various classifiers.
2. **Chapter 2 Literature Review:** This chapter will describes the Logistic Regression model used to predict NBA game outcomes and its underlying mechanics. This will also present comparsion table for all studies.
3. **Chapter 3 Methodology:** Explains the overall research approach, detailing the dataset, preprocessing, and feature extraction methods. This will also describes the machine learning models applied (Logistic Regression, Decision Tree, Random Forest, etc.) and the rationale behind their selection.
4. **Chapter 4 Implementation:** Outlines the step-by-step implementation process, including the training and testing phases for each classifier. This will also discusses the technical tools, libraries, and algorithms employed to build and evaluate the models.
5. **Chapter 5 Results:** Summarizes the performance of each model based on evaluation metrics like accuracy, precision, recall, and F1-score. This will also show all models confusion matrix and mention best model among them.
6. **Chapter 6 Discussion and Conclusion:** Analyzes the strengths and weaknesses of each model, emphasizing the effectiveness of ensemble methods like Hybrid Classifier. This will also concludes the findings, discusses limitations, and suggests future work for further enhancing prediction accuracy using real-time data and advanced models.

Chapter 2 Literature Review

2.1 Introduction to Sports Analytics

Sport business analytics is the use of statistical and data analysis techniques used specifically in the business of sports in its broadest sense Ratten and Dickson (2020). It is a type of business intelligence that has grown tremendously popular in the last few years thanks to the large amount of data that has become accessible, as well as improvements to machine learning. Due to its high speed, quantitative background and huge number of statistical indicators basketball has become one of the primary realms where analytics can pay off. The richness of the basketball data, including per-player data, team data, and contextual data, including the home field advantage, availability and performance of players Sarlis and Tjortjis (2020), and history of the teams make basketball a good candidate for predictive modelling. From here, analysts and teams leverage these data points to foresee game results, assess players' performance, analyze the right game approach or even decide on, for instance, which player to recruit or which game strategy to deploy. By applying machine learning algorithms, one can predict data situations that establish vital patterns in the overall management and general understanding of basketball. Specifically, machine learning over the outcome of antecedent games is a superior approach to observational analysis since it offers detail information that broad statistics could forecast outcome results more accurately. Analytical methods including regression, classification and ensemble can be used to classify the kind of interaction likely between different factors and the probability of a given result, say a win or a loss by a team. In recent years, applying analysis has progressed from a mere subdiscipline to a part of basketball management systems that are widely used by professional and amateur leagues around the world Bouchet et al. (2020). It is a reasonable assumption that as the technology and applied methodologies increase in sophistication, analysis has even more of an increase role in basketball, and can impact not only the team and team management but also player welfare, game strategies, and even fan involvement.

2.1.1 Significance of Predicting Basketball Game Outcomes

To determine the accuracy of the model and the extent to which it might prove useful to a variety of consumers, potential benefits for several types of consumer were described. In general, for every basketball professional teams it is important to make predictions of the game results especially in tournaments were the margin between the teams is very small Sarlis and Tjortjis (2020). More accurate predictions enable coaches to make better strategic choices, change the playing roster, and make correct decisions based on the performance of the competitors in a game. In addition, being able to forecast outcomes allow teams to evaluate their training programs, determine changes needed to be made and to determine the effects that player loss or addition might have on a team Bunker and Susnjak (2022). Besides this, sports analysts and journalists apply predictions as an additional tool for creating better material and an extended perspective on the result of the game, attracting people’s attention through applied statistics. The betting agencies also use game predictions to determine the odds and risk, and fine-tune their prediction, all of which is uppercase in an industry that specializes in sociability. The other advantage of game predictions for fans is that they are in a position to engage a lot with the game having to understand the analysis going on mentally and making predictions Rathi et al. (2020) on what the statistical models are saying about the game. Additionally, the type of the machine learning model and data analysis used makes the process more refined with regard to the things currently taking place in the game, performance and past record of the players and teams as well. As the machine learning models get smarter, they can work with such factors as a player gets tired, his mental state or even the crowd that turns out for the match. Therefore, forecasting basketball game's results is not only an application that will help to improve performance and attractiveness of the game but also an element of the trending big data implementation in sports industry.

2.2 Historical Approaches to Basketball Prediction

2.2.1 Early Statistical Methods in Basketball Prediction

When basketball prediction was in its infancy, most probable patterns were determined through statistical measures. Coaches, analysts and sport lovers would collect simple information like total scores obtained, field goal ratio, turnovers and rebounds in order to determine team efficiency Pantzalis (2020). Based on this approach and despite its relative simplicity, the foundation for patterns of the basketball games was made. Analytical measures used in early teams-statistical models involved essentially simple mathematical concepts such as averages, and ratios, and they had the main aim of embracing aspects involving the performance of individual players and interactions in teams.

2.2.2 The Advent of Advanced Statistical Models

Over the course of the centuries and as more technological development appeared, different levels of statistical models used in order to predict basketball events have also been reached Terner and Franks (2021). Thus, starting from the period of the late 1990s and the early 2000s, analysts started shifting to better, although more complex, methods, including multivariate regression analysis and time series analysis to derive mechanical correlation in the data. These models took into account a much larger set of parameters and generalized, among them the effectiveness of separate players, interaction within a team, and other parameters, indicating, for example average and slight ratings of offenses and defense. Another which happened during this times is the innovations of new statistical tools such as PER (Player Efficiency Rating), TS% (True Shooting Percentage) and WS (Win Shares) that were seen to give better assessment of a single player and the team Sarlis and Tjortjis (2020). Together with the progress of mathematical and statistical calculation skills it became possible not only to control game outcomes but also players' and team tendencies during the whole season. When more data became available for game results, especially through digital references such as Basketball-Reference the size of statistical models also grew. However, the substitution of more sophisticated statistical models for basketball games was a significant step in changing the approach toward league prognosis using machine learning methods.

2.2.3 The Emergence of Machine Learning in Basketball Prediction

The use of machine learning in basketball prediction was not assembled until the analysts and teams got a new way of doing their predictions Li et al. (2021). As the amount of available data increased, and the power of computational processes expanded limitations of statistical analysis of basketball outcomes started to be overcome by machine learning models Alonso and Babac (2022). Other methods of learning for example the decision tree, random forest and the neural network were in a position to handle large numbers of data and much more identify hidden patterns that would be hard for other methods., there has been a transition from handcrafting features to inducing them automatically with the help of machine learning algorithms that would decide on their optimal combination from the data supplied as input. This meant that prediction models were not static or set at the beginning of the season but climactic, and responsive to variations in individual performances, injuries as well as other game related factors. Machine learning also allowed analysts to forecast not only the games but also the various parameters related to performance Horvat and Job (2020), points scored and vice versa, and probabilities of the team to win, lose, or draw at a particular moment of the game. Machine learning expanded the capabilities even further, for deeper analysis, and also deep learning for utilizing textual, image and video data for identification of players and prediction of their play strategies. With the future development of the field ready to be unveiled, machine learning is said to become a primary basis for analyzing basketball games for enhancing every feature of team performance.

2.3 Machine Learning in Basketball Game Prediction

This study introduced in Horvat et al. (2020) is related to the analysis of algorithmic trading in sports, and identification of the most suitable process to analyze the results since human interference often affects the outcome of sports events. The authors applied seven different classification machine learning algorithms, including nearest neighbors and decision trees, to predict sports outcomes and compared their performance using two validation methods: Train&Test and validation through cross validation. The first research question being; how effective are such algorithms and validation methods in prediction, especially when applied in sports? From the prediction results obtained below it can be deduced that the nearest neighbors algorithm had the best average results while the decision trees gave the worst results. Furthermore, the result for each cross-validation case achieved higher accuracy than this Train&Test method. The study further subdivided the impact of current data in the Train&Test method and identified that enhanced performance has been found by using the current data rather than disjoint data. One of the main issues that were experienced in the course of the study was the identification of the ideal machine learning model having in mind the elicited features of sport events and their relation to human beings. However, this study has a few limitations; no other possibilities for model validation were considered by the authors. Moreover, analysis input and output were based only on structured data, and it is possible that unstructured data could be just as significant. Despite the findings offered by the study, the further work can contribute to the enhancement of the use of machine learning for sports prediction by filling these gaps: the use of the data from various sources and the implementation of other methods of validation.

The study presented in Horvat et al. (2020) aims to predict the final scores of NBA games by developing a hybrid prediction scheme integrating five data mining techniques: This is the reason as to why many are opting for machine learning algorithms such as; Extreme Learning Machine (ELM), Multivariate Adaptive Regression Splines (MARS), K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGBoost), and Stochastic Gradient Boosting (SGB). The features generated from the proposed scheme are obtained by combining game-lag information obtained from a number of basketball statistics. The data for the study was obtained from the 2018-2019 NBA season, which totalled 2460 games across 30 teams. The primary goal was to determine which game-lag information is most appropriate for improving prediction distortion and which game features are relevant. One of the significant difficulties of the research was the defining factor; selection of right features required for the identification of multiple data mining techniques in order to maximize the prediction efficiency.

In Huang and Lin (2020), the authors discussed the case study of the use of the regression tree model for predicting basketball scores: this was the approach not previously applied in the field of the study of sports movement, unlike the linear models typically used. More specifically, the research examined the Golden State Warriors game data from the 2017–2018 NBA season together with opponents' data. In this research, the use of regression tree is used in order to predict players' score and the overall team score. Since more precise estimates were needed it was possible to determine strong and weak symmetry requirements for each team. The regression tree model was first trained to generate the scores of a individual player for each game on every team Based on the predicted score, the cumulative score of the team was calculated as the sum of the individual scores. The results indicated that the model in question — the regression tree — was viable in predicting both IPA individual scores, or integral player averages, and TPA team scores since the overall average for the model was a point nine, with a 87.5% accuracy. One of the main difficulties in the analysis was the identified challenges relating to the specification and modelling of the wide range of factors impacting individual and team performance, including the comparison of players' performance in the course of a particular game.

The research work discussed in Alonso and Babac (2022) deals with one of machine learning applications in basketball—predicting potential candidates for the list of the stars in the next seasons. In this paper, rising stars are astute basketball players revealing signs that their performances will substantially enhance and they will be famous in the near future. To achieve this, the authors employed co-player statistics as features for machine learning models, with three types of co-players considered: Co-players who share the same team in a particular game, co-players of the none-sharing teams within the same game, and co-players of the sharing and none-sharing teams. The proposed solution showed that derived features have a significant impact on the prediction result and reached an F-measure of 96%. The research used different machine learning models and tested with multiple datasets, and found that MEMM gave the best results in all the datasets and also had the F-measure of 96%. Moreover, a ranking comparison analysis showed that most of the rising stars that were labeled during the experiment were ranked under 100 in the subsequent seasons thereby confirming the efficiency of the model in identifying future talent.

The study discussed in Zhao et al. (2023) is towards using Graph neural network (GNNs) on basketball games' result prediction, unlike prior machine learning models that disregard converting relations between teams and place of the league. The research takes the structured forms and converts them to unstructured graphs to represent the interactions in the teams in NBA developed dataset for the 2012–2018 period. First, an undirected core network was used to create a team representation graph and processed using a graph convolutional network (GCN) and achieves an average accuracy of 66. 90% for predicting game outcomes. To increase the accuracy level for prediction another enhancement was carried out on feature extraction using random forest integrated into the GCN model proved efficient and increased the accuracy level of prediction to 71.54%. One limitation in this study was the ability to capture multiple interactions of the teams which are important in modeling the game outcomes.

The work discussed in Romaniuk (2023) deals with spatial performance measurement in basketball while trying to identify areas of the basketball field that may be characterized with different levels of scoring likelihood, for either an individual player or the whole team. In order to do this, the authors outlined a new method utilising algorithmic modeling methods. . This study establishes that CART based ensemble methods perform better compared to the other methods for constructing the scoring probability map hence enhancing accuracy and interpretability. One of the main issues encountered in the study was to ensure that the selected algorithms allow for capturing the geometric features of the basketball court-as these are important to obtain accurate scoring probability maps.

In the study reported to in Zuccolotto et al. (2023), the picture is painted on the best over/undersampling technique suitable for use with AdaBoost in the fight against imbalanced datasets which are widely realized in pragmatic datasets. In the present work, the researchers conducted a simulation study to demonstrate the effectiveness of different over/undersampling approaches together with AdaBoost. The research focuses on methodologies such as Synthetic Minority Oversampling Technique (SMOTE) alongside with Jittering with Over/Undersampling (JOUS). One of the complexities in the study region was identifying the right sampling method that would allow us to cope with the issue of the distribution and avoid noise and overfitting. While the conclusions may be helpful for choosing the most suitable over/undersampling approaches for AdaBoost, the study has one weakness: it measures the performance of the technique based on simulated data. Although the results are encouraging, it could be expected that the performance of similar methods in real datasets may not be as good as in the simulated data due to the fact that the latter has a more complex structure and/or more variables are in play than are presented in the simulation.

The work under analysis in Sukumaran et al. (2022) was focused on offering an extensive review of AI and ML approaches for predicting basketball games. The technical procedural synthesis performed by the researchers involved an SLR based on the 553 articles, with 13 qualitative studies for further scrutiny based on their eligibility, risk of bias, and data quality. The study identified four standout algorithms used in basketball game outcome prediction: The explanation of proposed models are as follows: There is exist (1) Logistic Regression Model having 93.20% accuracy for predicting the winning observation based on team performance metrics and (2) Hybrid Fuzzy-SVM Model (AHFSVM), which has achieved 88.26% accuracy, (3) Hybrid Support Vector Machine & Decision Tree (AHVSDT) having 85.25% accuracy, and (4) Hybrid Ensemble. The study also pointed out data variations since the models included in the analysis derived from different datasets and measures for performance determined the variance in outcomes. However, despite the high accuracy of the algorithms, there are several significant weaknesses of this approach, including the absence of taking into account such external factors as the absence of key players due to injuries, fluctuations in team motivation or a specific condition of the basketball game.

Table 2.1: Comparison table

Study	Proposed Approach	Strengths	Weaknesses	Results
Horvat et al. (2020)	Classification ML algorithms (Train&Test, Cross-validation, KNN, Decision Trees, etc.)	Evaluated multiple algorithms and validation methods; robust analysis	Limited to specific algorithms; decision trees performed poorly	KNN had the best prediction accuracy, Cross-validation performed better than Train&Test
Chen et al. (2021)	Hybrid data-mining (ELM, MARS, KNN, XGBoost, SGB)	Comprehensive feature extraction; good prediction performance	Limited to one season's data; potential overfitting	XGBoost (two-stage) achieved the highest prediction performance (71.54%)
Huang and Lin (2020)	Regression tree model for score prediction	Effective for predicting player and team scores	Only focused on Golden State Warriors; limited to a single season	87.5% accuracy in predicting team scores

Alonso and Babac (2022)	Maximum Entropy Markov Model for rising star prediction	High prediction accuracy; strong feature extraction	May be computationally expensive	Achieved 96% F-measure score; rising stars ranked in the top 100 in subsequent seasons
Zhao et al. (2023)	Graph Neural Networks (GNN) for game outcome prediction	Considers team interactions and spatial structure; innovative approach	Initial success rate was lower before combining with Random Forest	Best prediction accuracy reached 71.54% after combining with Random Forest
Romaniuk (2023)	CART-based ensemble methods (Random Forest, Extremely Randomized Trees)	Robust and interpretable models; uses polar coordinates for court geometry	Limited by the reliance on CART models	Effective visualization of scoring probability maps; strong graphical interpretation
Zuccolotto et al. (2023)	Over/undersampling methods with AdaBoost	Improved accuracy with advanced	Results are based on simulations; may	JOUS approach provided the

		sampling methods; handled data imbalance well	not generalize to all datasets	highest accuracy across imbalanced datasets
Sukumaran et al. (2022)	Systematic literature review of AI and ML for basketball prediction	Comprehensive review; analysis of multiple algorithms	Based on secondary data; only reviews existing methods	Logistic Regression Model (93.2%) was the top performer, followed by Hybrid Fuzzy-SVM (88.26%)

Chapter 3 Methodology

3.1 CRISP-DM Methodology

1. Business Understanding: Generally, the purpose is to forecast the results of the NBA basketball facade through machine learning models. This entails identifying factors which facilitate game result outcomes including players, team and the game environment. The coverage involves data from the 2019/2020 to the 2023/2024 NBA season; 30 teams per season, 82 matches per team, and 2,460 matches per season. This paper focuses on creating models for purposeful results in sports related analysis, helpful to coaches and analyst, and interesting for sporting fans.

2. Data Understanding: Data is scraped from basketball-reference.com and involves next season through 2023-2024 team and game level advanced statistics. The dataset includes features aggregated at the team level and at the level of games: number of goals scored/shots taken, allowed for/opponents; player participation; and results of the match. Checking for missing values reveals their presence in several features; box plots further indicate the presence of outliers, while the categories option shows that some features are categorical and need to be preprocessed before analysis.

3. Data Preparation: The collected data is then written in CSV format and then placed in the pandas DataFrame. Data preprocessing includes tackling of missing values, erasing of unnecessary columns and formatting of data. The categorical features which include the team names and match venues are encoded using the Label Encoded methodology. Where feasible, features are quantized and standardized in order to achieve degree consistency with various models.

4. Modelling: Seven models are employed: The models that shall be used are Logistic Regression, Decision Tree Classifier, Random Forest Classifier, AdaBoost, CatBoost, LGBM and just to mention, a Hybrid Classifier using multiple base classifiers and a meta classifier. Both models are then built and trained on the training set to simulate a high level of identity resolution; however, the different parameters are adjusted to achieve the best possible performance.

5. Evaluation: Model performance is been evaluated using metrics such as Accuracy, Confusion Matrix, and Classification Report. These metrics provide some information into prediction quality, class distribution handling, and error analysis. The Hybrid Ensemble Classifier aims to outperform other models by using the strengths of individual algorithms while reducing their weaknesses.

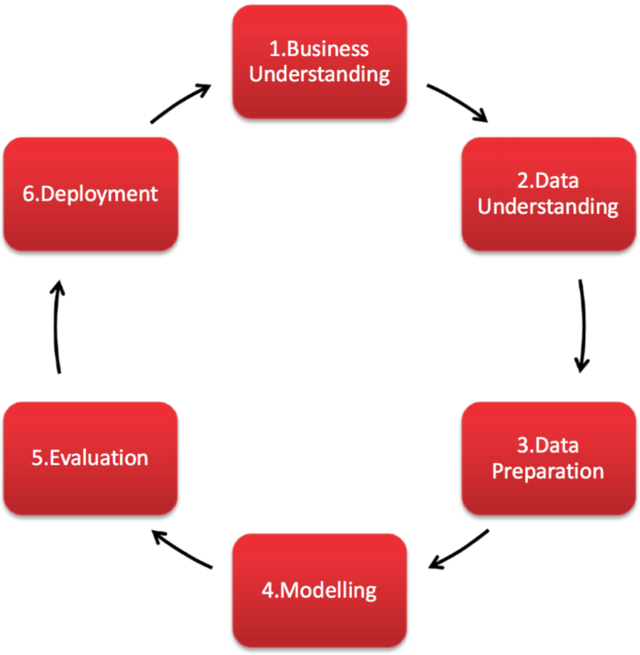


Figure 3.1: CRISP-DM Architecture by (https://www.sv-europe.com/crisp-dm-methodology/)

3.2 Dataset Description

The data used in this project layout includes complete data from each game of basketball in the National Basketball Association, for the years between 2019 to 2024. The statistics were collected from basketball-reference.com and NBA's official website using requests as web scraping tools, and are located in a CSV file for future use. Every one of the 30 teams in the NBA has 82 games a season; therefore, there are 2,460 games in a season. This kind of tally sums up to 12,300 records within the five year period. While each game includes basic team and opponent statistics, the source contains additional statistical information on team and player efficiency, team and player offending and defending ratings, shooting accuracy, and turnovers. Moreover, it is also namely possible to model considerate contextual features as home/away indicators, game outcomes (win/loss), and dates. Machine learning tasks can be performed effectively based on this dataset since it includes a number of numerical and categorical numerical features. Cleaning the data involves handling missing values for example by fill na or dropping columns which are not useful and Encoder which is mainly used in converting labels to values, an example being LabelEncoder.

3.3 Libraries Imported

For the processing and analysis activities of this project, different sets of libraries are used for data handling, visualization, machine learning, and evaluation. Other important libraries like data operation and numerical computation, and data visualization and graphical libraries like pandas, numpy, Matplot lib, seaborn. Also, plotly.express and plotly.graph_objects modules allow for building interactive and dynamic graphics which in turn facilitates the insight from the data. Machine learning models are developed with classifiers of sklearn like Logistic regression, Decision tree, Random forests, A da boost and ensemble models. Numeric features are explored using tests for normality and evaluated for scale using the Kolmogorov Smirnov test , while categorical features are checked for symmetry, number of unique values and their distribution check. Class imbalance problems are handled using the SMOTE function belonging to the imblearn library. For categorical variables, the LabelEncoder and OneHotEncoder are applied, while for continuous variables in StandardScaler the standard scaling is adjusted. pipeline also includes preprocessing and modeling to help breakup the flow of the work. Validation techniques used includes, accuracy_score, confusion matrix and the classification_report and for cross validation, the protocol used is cross_val_score. In feature selection, Recursive Feature Elimination is being used and for hyperparameter optimization, RandomizedSearchCV is efficient. Stacking classifiers are built using the mlxtend.classifier.StackingClassifier; these are enhanced from multiple models to suggest better results. Other libraries used are pickle which is used to store models and warnings to avoid disturbing messages during model execution. To replicate the results of the data exploration and explore the datasets in more detail, the pd.set_option("display.max_columns", None) configuration is used. This vast library of packages aids the smoothing of data pre-processing, strengthening of the model construction process, as well as the compilation of the evaluations that are fundamental in creating sound NBA game predictions.

3.4 Data Cleaning and Data Preprocessing

Data cleaning concentrates on making sure that a given dataset is clean to the sense that its records are accurate and standard for analysis. Firstly, predictive analysis irrelevant columns that have no usefulness for predictive analysis are removed including 'Rk,' 'G,' 'At,' and 'Win.' The next column Date is converted from object type to datetime type to enable time based analysis and sorting on it. Duplicated observations are also evident in any dataset collected and combined from other sources and it's eliminated to enhance the quality of the data. The next operation brings the dataset down to 14,118 rows and 28 columns to make the database clean and consistent.

In the preprocessing phase the concentration is on the preparation for the actual delivery of the machine learning models. It is also standard to categorize data in the dataset such as team names or the names of the opponent for which we use the LabelEncoder. This step is important as many machine learning models expect their input to be numerical, while encoding translates categorical features into a format that the algorithms will understand quickly. Furthermore, careful is observed to make sure that label encoding does not change its mapping between the training and testing phases. Other steps of preprocessing may also involve operation on the data if the raw data is numerical where all the columns are made to have accordant data types as needed for later processes. In combination, such data cleaning and preprocessing steps provide well-organized and machine learning friendly format data which is essential precondition for developing effective and accurate futuristic models.

3.5 Data Visualization

Figure 3.2 provides an extensive graph with regard to the team performance with the help of a faceted bar chart in seasons 2019-2024. As such, the chosen visualization has six subdivisions that correspond to the six seasons, with three subplots in each row. Since there are two teams (at the bottom on the x-axis), there are two bars of red and blue each tick, the y-axis counts the number of wins or losses for a particular team. The graph thus helps in quickly identifying the nature of balance and steadiness in team performance within and between seasons. These include the fact that the height of each subplot is constant so that one season may be compared directly to another across the subplots on a zero to sixty game basis. It seems to include professional sports teams, which presumably are in a major league since the abbreviations of the teams' names are given along the x-axis. Two teams play six games less in the 2020 season which can mean that organising of the games had been affected by some circumstances. In contrast, every subsequent season (from 2021-2024) studies show a return to average game numbers meaning normalized schedule. The colour coding (blue for loss, red for win), and the grouped bar style of graph makes it clear which teams had better seasons and which has poor seasons and the vertical bar enables comparisons of the win/loss ratios. There are recommendations to interact with the data given by the toolbar shown near the top right of the given image which allows a user to zoom, pan and do basically the whole shebang depending on the things that a user might find interesting. The annotation visible in the 2024 season subplot contains concrete data points for individual teams, thus, improving the strength of the analysis of the visualization.

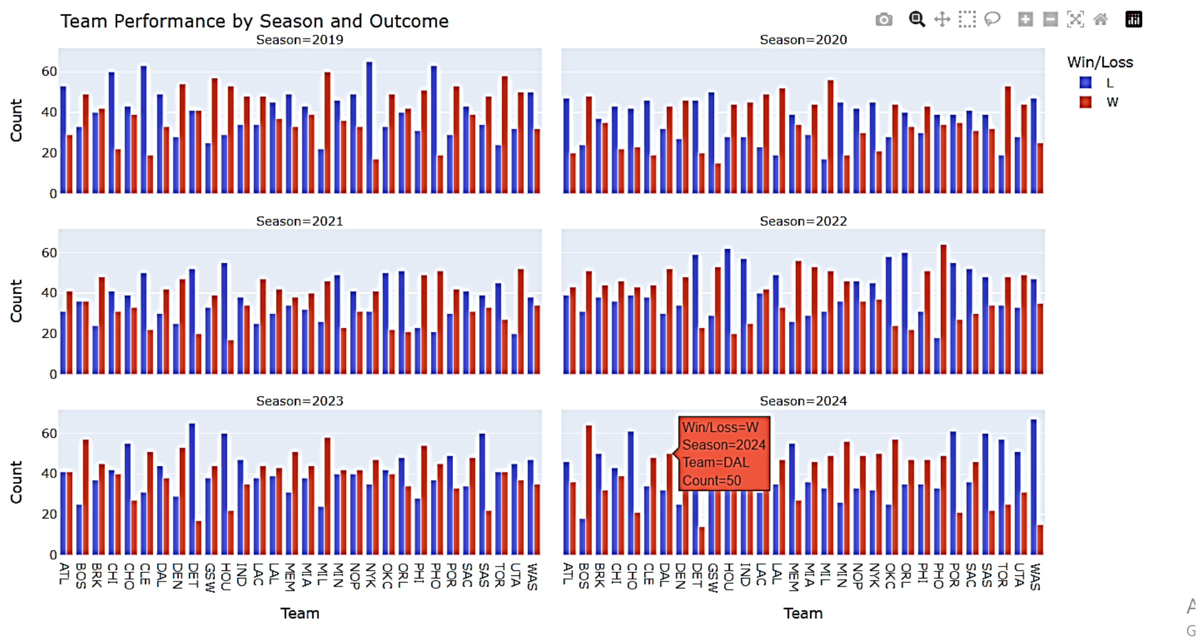


Figure 3.2: Faceted Bar chart

Figure 3.3 displays a detailed donut chart analysis on location (home & away), and result (win & loss) of games. The visualization employs a four-segment color scheme to represent different combinations of location and game results: Away Losses (Away_L) in dark burgundy at 28% Home Losses (Home_L) in light coral at 22% Away Wins (Away_W) in salmon pink at 22%

Home Wins (Home_W)in deep red at 28%. Distribution of the chart is also rather balanced with interesting patterns of the team's performance at different venues: home and away; value of home performance (Home_W + Home_L = 50%) is equal to the value of away performance (Away_W + Away_L = 50%) which gives an impression that the schedule was balanced. To the right of the graph annotation refers to one "Location_WL" and the count of 3179 is most likely the total sum of games reviewed in this distribution. Drawing the four areas comparing to 100 percent in a donut chart with a hole works well to focus on the sizes of these categories at the same time keeping the segments separated. The chosen color scheme is coherent with segment division, though offers good contrast between Loss and Win segments, where the darker colors stand for the first one and the lighter – for the second one. This means that this visualization is very useful when analyzing the ability of home and away location to determine win/loss records all within one chart.

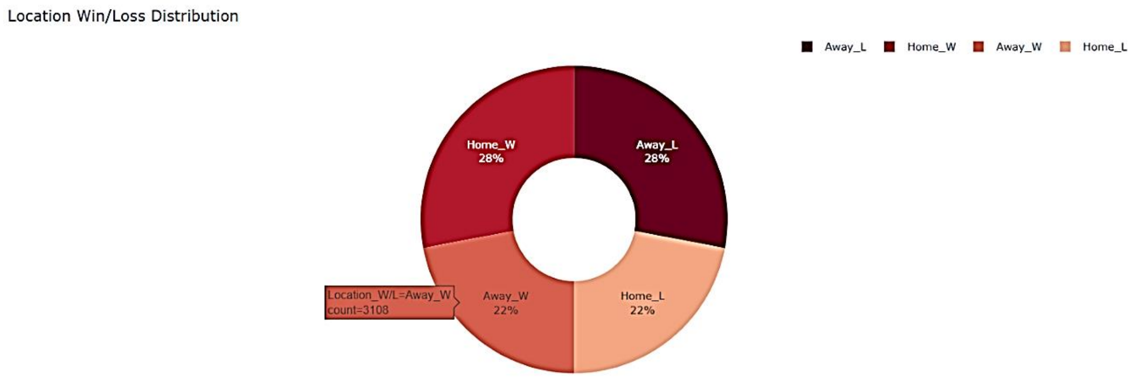


Figure 3.3: Donut Chart

Figure 3.3 presents a frequency distribution histogram of the total points which have been obtained in basketball games and a new type of graph known as the 'rug plot' above the histogram displays raw data points. The histogram in a lime green makes the distribution of scoring patterns look almost normal, or bell shaped with an approximate midpoint of between

110-120 points. The horizontal axis runs along 60-180, where the vertical axis displays the number of games that occurred at each total game point combination, which reaches 4,000 at the highest point. The distribution seems to be positively skewed, meaning more extreme high scoring games than extremely low scoring ones.

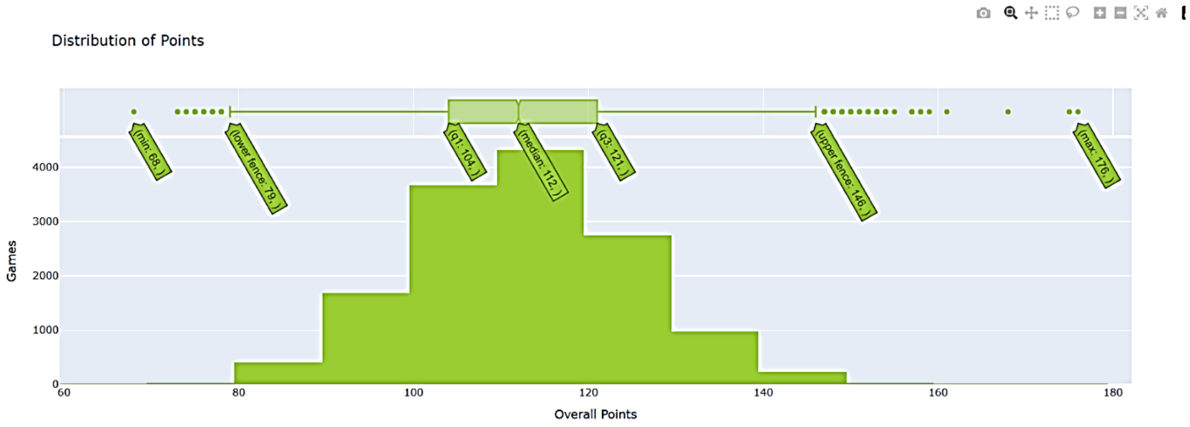


Figure 3.4: Histogram showing Distribution of Points

Figure 3.5 shows scatter plot visualization of points scored disaggregated by season from the 2019/2020 season to the 2023/2024 season, whereby each point on the figure represents a single match. The color scheme is from dark blue (2019) gradually fading to purple, pink, and orange gradually making an end at a bright yellow (2024), which makes it somewhat easier to track temporal changes in scoring patterns. The values on the y-axis vary between around 80 to 180 points, and the x-axis represents the positions of a season. The visualization reveals several interesting patterns: each season creates a vertical band of the points which imply that the scoring is steady within the certain season but is changed in other seasons. The instances of points show quite a hatch pattern with the denser form coming somewhere around 100 to 140 points and not many below 80 or above 160 points in any of the seasons. There is a significantly less density of data points for 2021 season which could be possibly due to lower number games played in the season possibly because of some externality that impacted the schedule of the league.

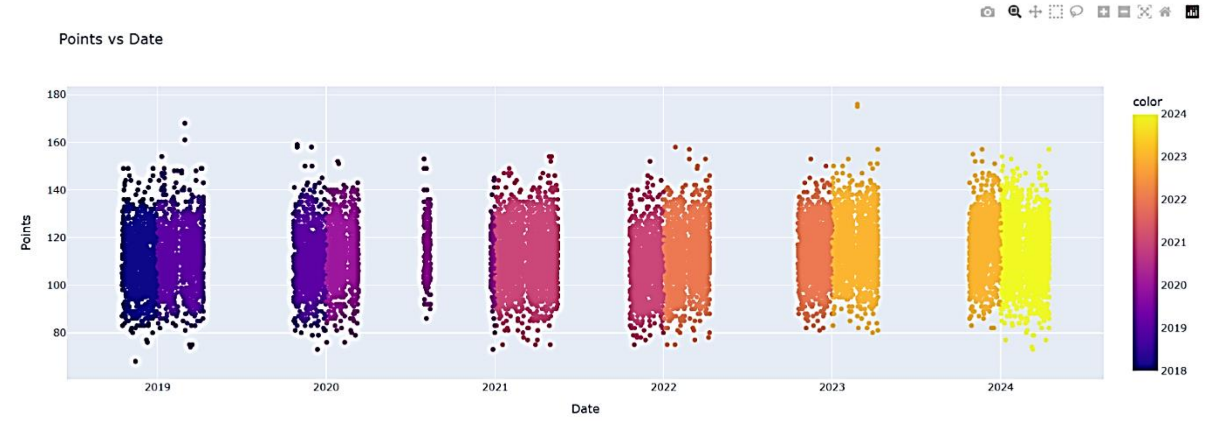


Figure 3.5: Scatter plot showing Points vs Date

Figure 3.6 is a line graph showing the average TSP of NBA teams, on the y-axis, the TSP is shown ranging from approximately 0.555 to 0.59 or 55.5% to 59%. This one is also quite simple, containing only blue single line that connect dots for each team, which are arranged at the x-axis by their standard three-letter abbreviations according to alphabet order (for instance, ATL, BOS, BRK). It can clearly be noticed in the graph that there are large discrepancies of teams in terms of shooting efficiency and there are even some spikes and dips in the graph. Some teams achieve notably higher shot true shooting points coming very close to 0.590 (59% average shooting efficiency) or even slightly less at 0.555 (55.5 %). For ease in reading off the accurate values, the graph has faint lines forming grid against blurred blue backdrop. The way the line connectors are drawn provides an instant mental map of the shooting efficiency differential within and outside the league, thus presenting which club is above or below the mean value.

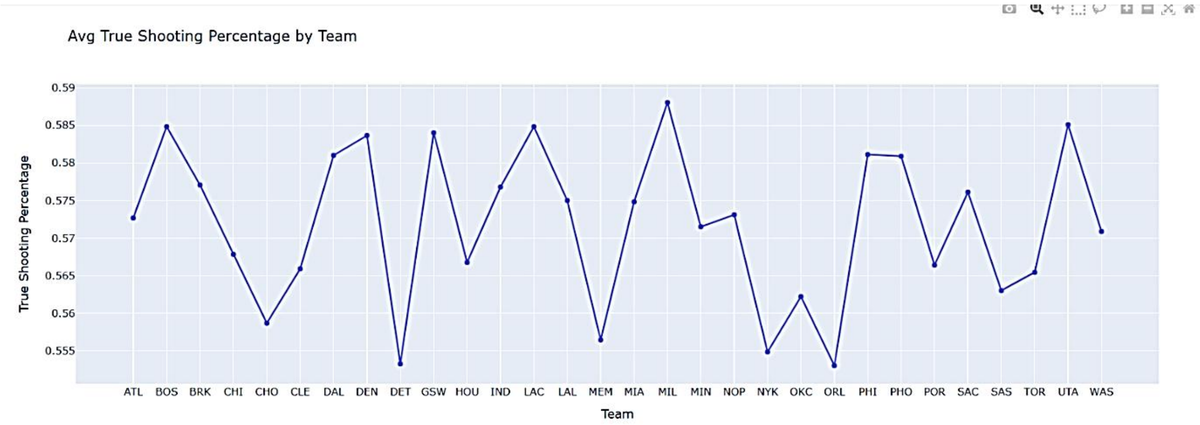


Figure 3.6: Line Graph showing Avg True Shooting Percentage by Team

In figure 3:7 there is line graph depicting ‘The Top 10 Biggest Comebacks in NBA Games AMA’ from 2019 till 2024 in which there is Maximum point difference (Biggest Comeback) plotted against the respective game dates on the horizontal line (x-axis). Highlighted teams represent NBA teams, with color codes MEM (Memphis), OKC (Oklahoma City), MIA (Miami), IND (Indiana), BOS (Boston SUS), GSW (Golden State Warriors), TOR (Toronto), and PHI (Philadelphia). From the graph, there are some of the phenomenal come back incidents in NBA history; however, the most outstanding one is a roughly 70-point come back done by Memphis (MEM) in 2022 as depicted by the horizontal blue at the topmost point. Other programs of the match are Toronto-Indiana-2021 and Golden Gate Warriers-2023 both having a comeback of near about 55 points. The graph clearly captures what could be described as the worsening or escalation of these NBA comeback stories over time; with the period 2021-2024 having most of these big come back highlights highlighted and to some extent, the year 2020 does not seem to have any of these top 10 comeback games recorded.

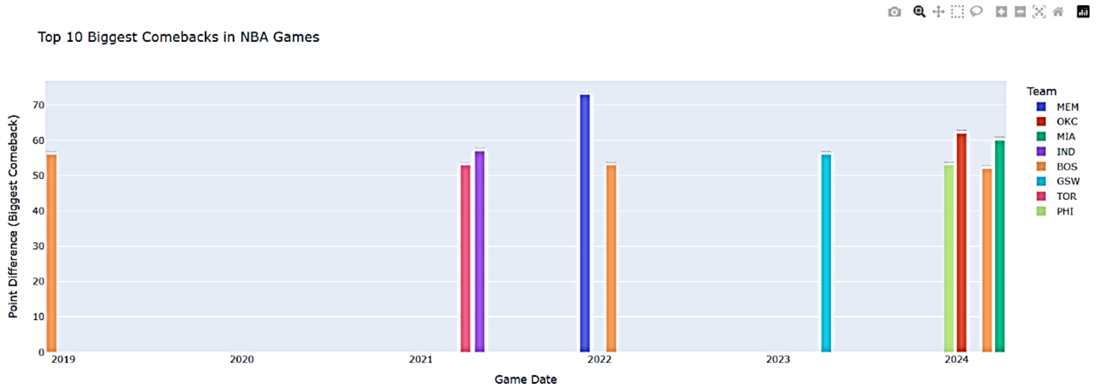


Figure 3.7: Top 10 Biggest Comebacks in NBA Games

Figure 3.8 –Team Average turnover percentage over seasons up to 2024 in NBA is represented by a stacked bar chart. For the representation of different NBA teams, the colors codes used consist of ATL for Atlanta, BOS for Boston, BRK for Brooklyn, CHI for Chicago, CHO for Charlotte, CLE for Cleveland, DAL for Dallas, DEN for Denver, DET for Detroit, GSW for Golden State Warriors, HOU for Houston, IND for Indiana, LAC for LA Clippers, LAL for LA Lakers, MEM for Memphis, MIA for Miami Above each of the bars of the seasonal variation, figures are presented as values, including 12.26951 for 2019, 12.23611 for 2020, 12.34028 for 2021, 12.09512 for 2022, 12.29232 for 2023, and 12.22137 for 2024.

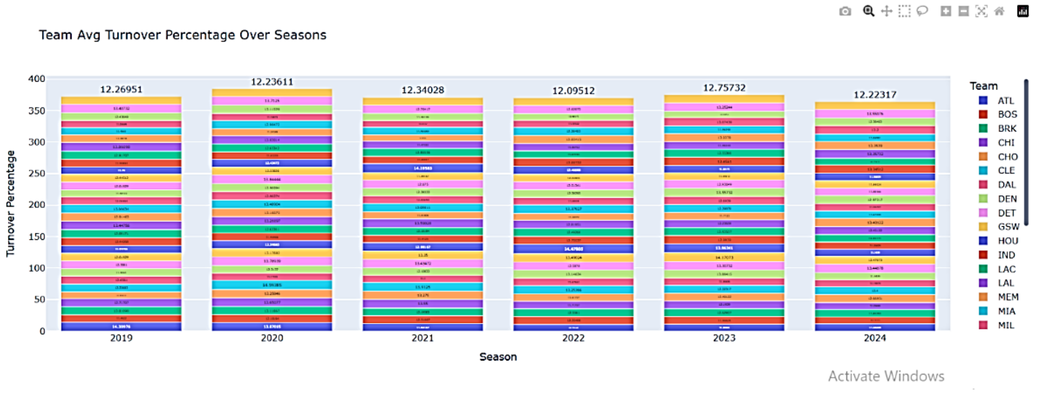


Figure 3.8: Team Avg Turnover Percentage Over Seasons

In particular, the player turnover percentage tends to be somewhat negatively associated with points scored as illustrated by the scatter plot displaying the trend lines in figure 3.9 which is been distinguished by color the different NBA teams. The x-axis is representing the turnover by percentage beginning with 0% and extending up to 30% The y-axis depicting points scored vary approximately from 80 to 180. The visualization contains data points for several teams such as ATL, BOS, BKN, CHO, CLE, DAL, DEN, DET, GSW, HOU, IND, LAC, LAL, MEM, MIA and MIL using different colour codes for differentiation. The regression lines reveal relatively low but negative relationship values between turnover percentage and points scored in most of the teams on the plot. There are many data points located at and around the mean turnover percentage, it usually ranging from 100 – 140, which suggests that most players produce results in this area.

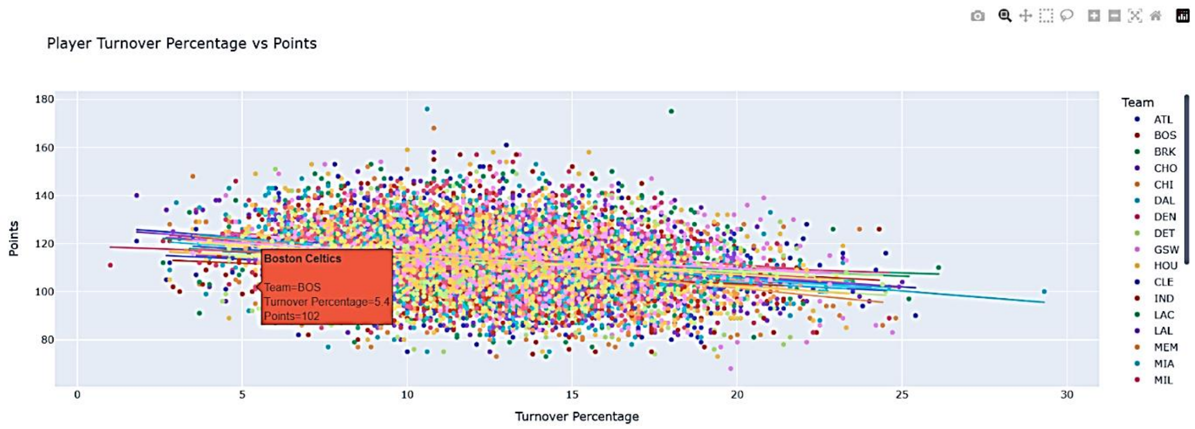


Figure 3.9: Player Turnover Percentage vs Points

As shown in the multi-series bar chart of Figure 3.10, the steal percentage has been analyzed in terms of temporal dynamics for various NBA teams for the period 2019-2024. The visualization uses colors to represent different teams such as ATL, BOS, CHI CHO, CLE, DAL, DEN, DET, GSW, HOU, IND, LAC, LAL, MEM, MIA, MIL and etc For each team, colored bars are aligned depending on the season on the X-axis and steal percentage varies from 0-10 on the Y-axis. The bars are grouped by year by design, so that the gymnast's performances can be compared within a year and a team's performances can be examined over the years. According to the data gathered, most teams sustain steal percentages within the range of 6-8% a concept that rises slightly to near 10% at some teams during some seasons only.

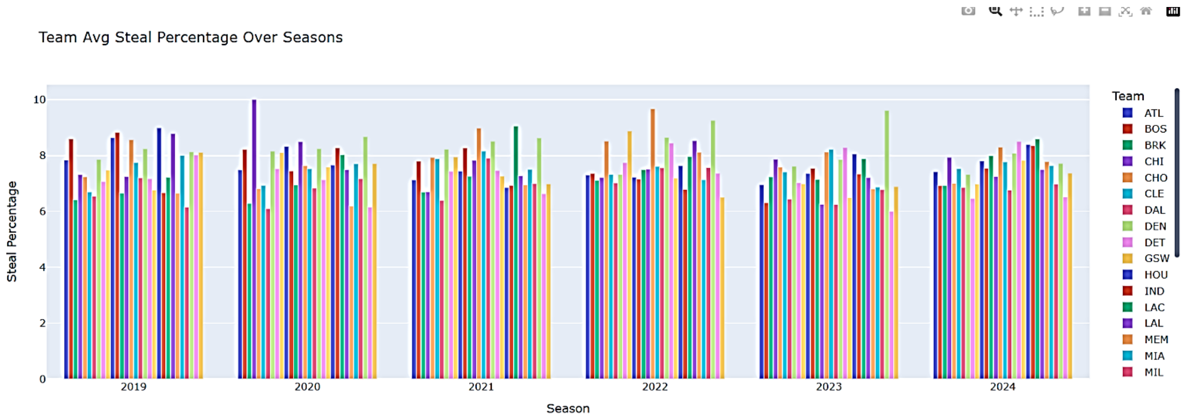


Figure 3.10: Team Avg Steal Percentage Over Seasons

Figure 3.11 presents a comparative box plot analysis of effective field goal conditional probability (eFG%) in basketball games so that the final results of the shooting performance could be clearly seen statistically. The visualization uses two box-whisker plots; on the x-axis, the data is divided into win and losses, (W/L); the y-axis represents eFG% that ranges between 0.3 and 0.8 (30 – 80%). Plot is divided in blue and gain appears in red or coral hue making it distinctly noticeable that it is a gain. The distribution of losses (L) in proposed eFG% looks as the median approximately equal to 0.50 (50%), and only the interquartile range varying from 0.45 to 0.55; However, distribution of wins (W) is more optimistic with more median eFG approximately 0.55 (55%), while the interquartile range lies between 0.52 and 0.60 only.

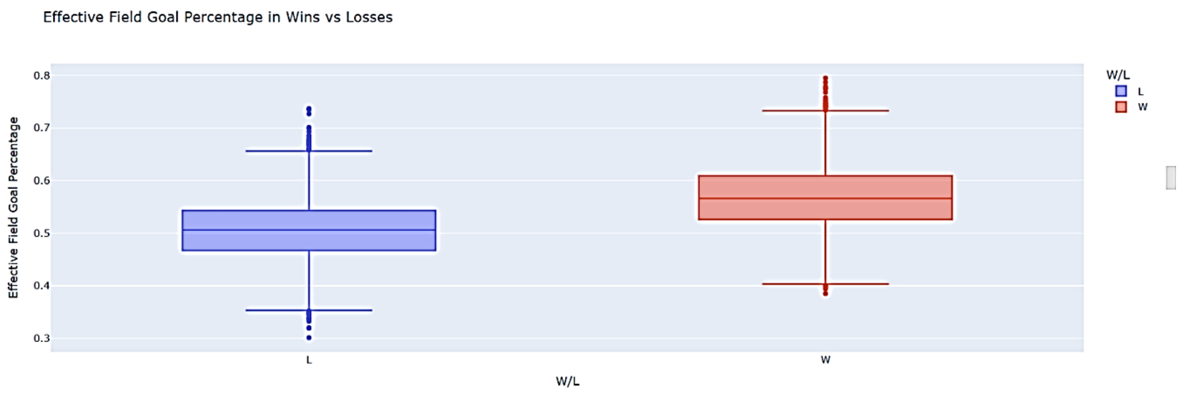
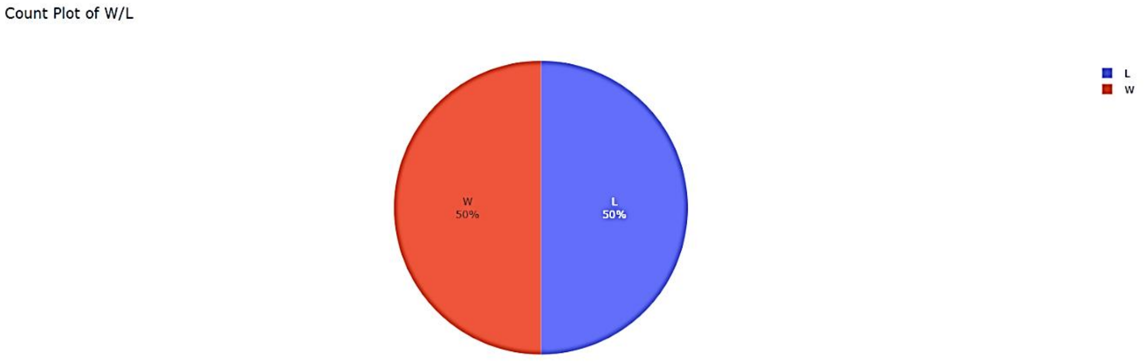


Figure 3.11: Effective Field Goal Percentage in Wins vs Losses

In Figure 3.12, a Pie Chart map of Wins (W) and Losses (L) in Basketball matches has been shared to use a basic two colour combination of blue for the Ls and coral/red for W. Depending on the type of analysis carried out this data should ideally come from a complete season because the total wins must sum up to total loses and in this case the division is equal 50% & 50%. They are both clearly marked ‘L 50%’ on the blue half and ‘W 50%’ on the red half of the figure making the interpretation easy. At the top right, there is a legend that is marked as L for losses (blue) and W for wins (red).



Activate Windows

Figure 3.12: Pie Chart of W/L

Season-wise average defensive rating by team across six NBA seasons from 2019 to 2024 is illustrated through the interactive three-dimensional faceted scatter plot in Figure 3.13 . The plot comprises of six panels at the lay out of a grid; each panel corresponds to a particular season of the year, which makes it easier to compare results of a given month in different years while also providing a general trend of the year. The y-axis is always labeling the average of the defensive rating, using values between 100 and 120 points, while the x-axis includes the name of the NBA teams and their three letters abbreviations. These are represented by differently shaded blobs: CHO, CLE, DAL, DEN, DET, GSW, HOU, IND, LAC, LAL/ MEM, MIA, MIL, MIN, NOP, NY/ OKC, ORL, PHI, PHO, POR, SAC, SAS, TOR, UTA, and WAS teams. The visualization highlights how different clubs' defensive rating looked and how most of them were within 110-115 range during specific seasons. The problems under all the panels are consistently scaled and formatted, and therefore allow the user to easily compare the defensive performance trends across the teams, and see how the teams' defensive efficiencies have changed over the years.

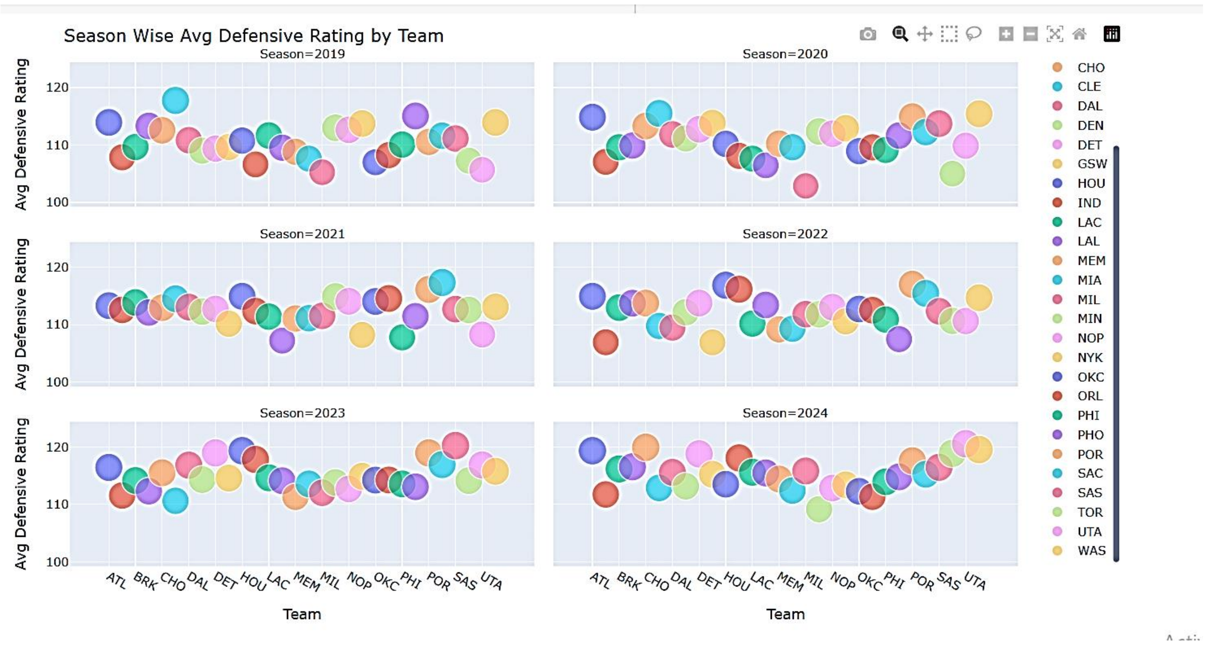


Figure 3.13: Season Wise Avg Defensive Rating by Team

A faceted line plot visualization is shown in figure 3.14 where the season-wise TS% of NBA teams for the period 2019-2024 is bifurcated into 6 plots. This visualization uses a purple line to connect data points of each team; the y-axis is True Shooting Percentage, ranging from about 0.54 (54%) to 0.60 (60%); the x-axis is NBA teams' three-letter abbreviations. The parameters reported for each panel add up to over time about one calendar season, but are also appropriate for using cross sections within seasons and collect longitudinal data across seasons. The connected line plot format helps emphasize changes and trends in shooting accuracy of different teams better; individual True Shooting Percentages of most teams range from 0.55 (55%) to 0.58 (58%). The patterns of shooting efficiency observed for all panels are directly comparable thanks to scale invariance, allowing viewers to see how the teams' potency has changed in the considered timeframe. Some of the important observations are fluctuations touching the high of about 0.60, which represents 60 percent, and the low of about 0.54, which represents 54 percent; these observations suggest that there are great fluctuations in the shooting efficiency within the league.

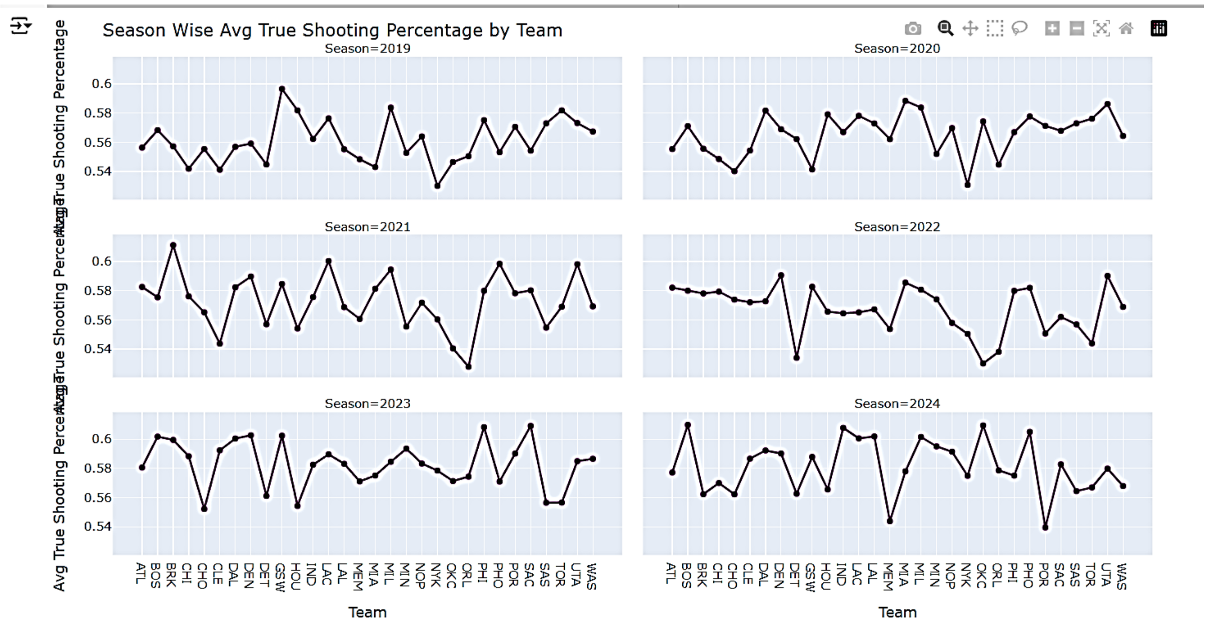
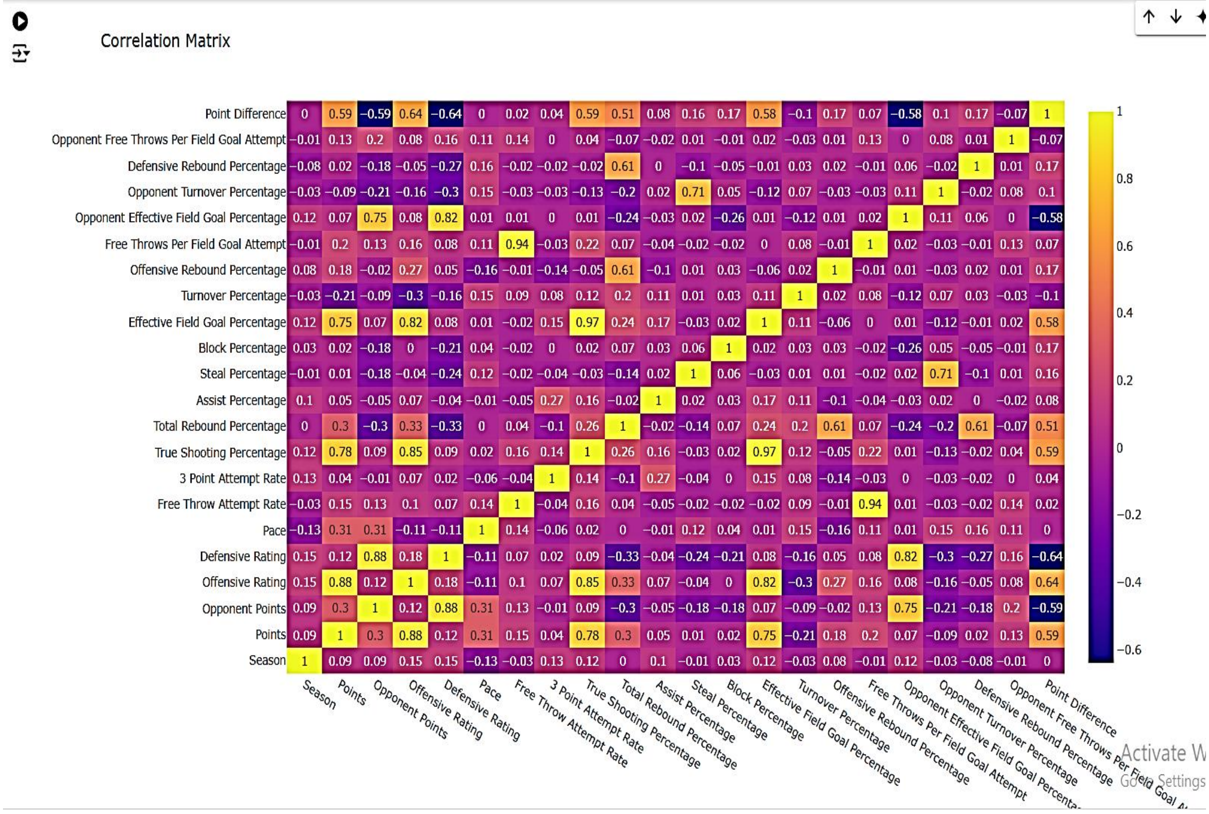


Figure 3.14: Season Wise Avg True Shooting Percentage by Team

As an example, Figure 3.15 provides a correlation matrix heatmap including all basketball statistics and performance indicators we discussed in the current section. To facilitate easy comparison, the matrix uses color scale running from black, indicating at strong negative correlation through white color to yellow indicating strong positive correlation and, for every individual cell contains actual correlation coefficient values. The variables used in the study are Point Difference, Free Throws Per Field Goal Attempt, Defensive % Turn Over % E-FGP ,Block % STEAL %, and other relevant basketball performance indices like seasonal stats… What was initially noticeable is the positive correlation: high off-rating is associated with high points (0.88); high def- rating is associated with high opponents points (0.88); and high ts% is associated with high efg% (0.85).



Correlation Matrix

	Season	Points	Opponent Points	Offensive Rating	Defensive Rating	Pace	Free Throw Attempt Rate	3 Point Attempt Rate	True Shooting Percentage	Total Rebound Percentage	Assist Percentage	Steal Percentage	Block Percentage	Effective Field Goal Percentage	Turnover Percentage	Offensive Rebound Percentage	Free Throws Per Field Goal Attempt	Opponent Effective Field Goal Percentage	Opponent Turnover Percentage	Defensive Rebound Percentage	Opponent Free Throws Per Field Goal Attempt	Point Difference
Point Difference	0	0.59	−0.59	0.64	−0.64	0	0.02	0.04	0.59	0.51	0.08	0.16	0.17	0.58	−0.1	0.17	0.07	−0.58	0.1	0.17	−0.07	1
Opponent Free Throws Per Field Goal Attempt	−0.01	0.13	0.2	0.08	0.16	0.11	0.14	0	0.04	−0.07	−0.02	0.01	−0.01	0.02	−0.03	0.01	0.13	0	0.08	0.01	1	−0.07
Defensive Rebound Percentage	−0.08	0.02	−0.18	−0.05	−0.27	0.16	−0.02	−0.02	−0.02	0.61	0	−0.1	−0.05	−0.01	0.03	0.02	−0.01	0.06	−0.02	1	0.01	0.17
Opponent Turnover Percentage	−0.03	−0.09	−0.21	−0.16	−0.3	0.15	−0.03	−0.03	−0.13	−0.2	0.02	0.71	0.05	−0.12	0.07	−0.03	−0.03	0.11	1	−0.02	0.08	0.1
Opponent Effective Field Goal Percentage	0.12	0.07	0.75	0.08	0.82	0.01	0.01	0	0.01	−0.24	−0.03	0.02	−0.26	0.01	−0.12	0.01	0.02	1	0.11	0.06	0	−0.58
Free Throws Per Field Goal Attempt	−0.01	0.2	0.13	0.16	0.08	0.11	0.94	−0.03	0.22	0.07	−0.04	−0.02	−0.02	0	0.08	−0.01	1	0.02	−0.03	−0.01	0.13	0.07
Offensive Rebound Percentage	0.08	0.18	−0.02	0.27	0.05	−0.16	−0.01	−0.14	−0.05	0.61	−0.1	0.01	0.03	−0.06	0.02	1	−0.01	0.01	−0.03	0.02	0.01	0.17
Turnover Percentage	−0.03	−0.21	−0.09	−0.3	−0.16	0.15	0.09	0.08	0.12	0.2	0.11	0.01	0.03	0.11	1	0.02	0.08	−0.12	0.07	0.03	−0.03	−0.1
Effective Field Goal Percentage	0.12	0.75	0.07	0.82	0.08	0.01	−0.02	0.15	0.97	0.24	0.17	−0.03	0.02	1	0.11	−0.06	0	0.01	−0.12	−0.01	0.02	0.58
Block Percentage	0.03	0.02	−0.18	0	−0.21	0.04	−0.02	0	0.02	0.07	0.03	0.06	1	0.02	0.03	0.03	−0.02	−0.26	0.05	−0.05	−0.01	0.17
Steal Percentage	−0.01	0.01	−0.18	−0.04	−0.24	0.12	−0.02	−0.04	−0.03	−0.14	0.02	1	0.06	−0.03	0.01	0.01	−0.02	0.02	0.71	−0.1	0.01	0.16
Assist Percentage	0.1	0.05	−0.05	0.07	−0.04	−0.01	−0.05	0.27	0.16	−0.02	1	0.02	0.03	0.17	0.11	−0.1	−0.04	−0.03	0.02	0	−0.02	0.08
Total Rebound Percentage	0	0.3	−0.3	0.33	−0.33	0	0.04	−0.1	0.26	1	−0.02	−0.14	0.07	0.24	0.2	0.61	0.07	−0.24	−0.2	0.61	−0.07	0.51
True Shooting Percentage	0.12	0.78	0.09	0.85	0.09	0.02	0.16	0.14	1	0.26	0.16	−0.03	0.02	0.97	0.12	−0.05	0.22	0.01	−0.13	−0.02	0.04	0.59
3 Point Attempt Rate	0.13	0.04	−0.01	0.07	0.02	−0.06	−0.04	1	0.14	−0.1	0.27	−0.04	0	0.15	0.08	−0.14	−0.03	0	−0.03	−0.02	0	0.04
Free Throw Attempt Rate	−0.03	0.15	0.13	0.1	0.07	0.14	1	−0.04	0.16	0.04	−0.05	−0.02	−0.02	−0.02	0.09	−0.01	0.94	0.01	−0.03	−0.02	0.14	0.02
Pace	−0.13	0.31	0.31	−0.11	−0.11	1	0.14	−0.06	0.02	0	−0.01	0.12	0.04	0.01	0.15	−0.16	0.11	0.01	0.15	0.16	0.11	0
Defensive Rating	0.15	0.12	0.88	0.18	1	−0.11	0.07	0.02	0.09	−0.33	−0.04	−0.24	−0.21	0.08	−0.16	0.05	0.08	0.82	−0.3	−0.27	0.16	−0.64
Offensive Rating	0.15	0.88	0.12	1	0.18	−0.11	0.1	0.07	0.85	0.33	0.07	−0.04	0	0.82	−0.3	0.27	0.16	0.08	−0.16	−0.05	0.08	0.64
Opponent Points	0.09	0.3	1	0.12	0.88	0.31	0.13	−0.01	0.09	−0.3	−0.05	−0.18	−0.18	0.07	−0.09	−0.02	0.13	0.75	−0.21	−0.18	0.2	−0.59
Points	0.09	1	0.3	0.88	0.12	0.31	0.15	0.04	0.78	0.3	0.05	0.01	0.02	0.75	−0.21	0.18	0.2	0.07	−0.09	0.02	0.13	0.59
Season	1	0.09	0.09	0.15	0.15	−0.13	−0.03	0.13	0.12	0	0.1	−0.01	0.03	0.12	−0.03	0.08	−0.01	0.12	−0.03	−0.08	−0.01	0

Figure 3.15: Correlation Matrix

Chapter 4 Implementation

4.1 Logistic Regression

Logistic Regression is a linear algorithm and model which is used in binary classifier problems. It estimates the odds of an event from a sigmoid function, which is usually useful when predicting win or loss related risks. Finally the features and targets are split into the training set and the testing set after data cleaning and pre processing. For implementation, the LogisticRegression class is used from the class list available in the sklearn.linear_model module. In addition to this the tuning of hyperparameters such as solver (lbfgs, saga) and regularization (C) is performed. This model is trained by applying the .fit() function from Keras on the training set and approves the worth of the network by using .predict() function. An accuracy value and a classification report facilitate the assessment of the model's performance on the basis of separating the teams into winners and losers.

4.2 Decision Tree Classifier

Decision Trees categorize data by first formulating a tree with decision rules on features on the values. To construct a tree, where a dataset is being split recursively by the use of the DecisionTreeClassifier from sklearn.tree using the Gini impurity/entropy criterion. Then, after data pre-processing, parameters such as max_depth and min_samples_split are tuned so that, no underfitting or overfitting occurs. Compared to other models, the algorithms used in this model enable straightforward visualization of decision-making. The same is done and predictions are made then the confusion matrices and classification reports are done in order to check the performance.

4.3 Random Forest Classifier

Random Forest is a type of meta-classifier made of many decision trees the errors and the over-fitting of which are minimized by using. It is implemented using RandomForestClassifier of sklearn.ensemble in which it has parameters like n_estimators (number of tree) and max_features (feature at every split). The trees are constructed over smaller data samples bootstrapped and the final conclusion is arrived at over a majority rule. This strong model is learnt from the dataset and checked with the test set which prove its versatility to tackle various and intricate structures of NBA data.

4.4 AdaBoost Classifier

Considering that AdaBoost is a boosting technique, it incorporates weak learners, such as, decision trees into a strong predictor. As an AdaBoost Classifier using sklearn.ensemble, it is the means to train a weak learner and iteratively add new weak learners while redistributing weights for misclassified instances. Hyperparameters which must be tuned include n_estimators, and the learning rate to regulate the behavior of the model. After training, the performance of the proposed model is assessed in terms of classification metrics, and its capacity to work with imbalanced datasets is stressed.

4.5 CatBoost Classifier

CatBoost is a gradient boosting algorithm for categorical features only, which is implemented via catboost package. It is different from other models as it does not make problems when it comes to categorical features without bringing the need of perform preprocessing steps. This dataset splits into training and test sets, and 4 parameters, including iterations, depth, and learning_rate, is tuned to improve the result. Once trained, the model makes predictions and its effectiveness is evaluated and demonstrated to have a very low percentage of overfitting.

4.6 LGBM Classifier

LGBM as a gradient boosting framework is designated to work faster and more efficient with big data. In the LightGBM, it is built on the lgb.LGBMClassifier to make leaf-wise splits in order to minimize loss at each step. After data preprocessing, some important parameters such as, num_leaves, max_depth and learning rate are adjusted for better outcomes. The model learns from the dataset, it predicts, and evaluation metrics reveal it as one that performs exceedingly well with structured data.

4.7 Hybrid Ensembled Classifier

The Hybrid Ensembled Classifier is built on the idea where the output of one or more algorithms is combined to form the final output of the classifier. This model uses LGBMClassifier, CatBoostClassifier, RandomForestClassifier as base classifiers and then an AdaBoostClassifier has been used to blend these base classifer's predictions. By using the StackingClassifier from mlxtend, the implementation starts by creating each base classifier and named parameters including n_estimators, max_depth for LGBM and CatBoost models as well as n_estimators for the Random Forest. AdaBoostClassifier, which is a meta-classifier, is used and has n_estimators=5 to make it easily adjustable to the multiple predictions obtained. The above and the subsequent equations illustrate that the base and meta-classifiers are then initialized into the ensemble model to create a powerful stacked learning patterns. Once again the dataset obtained through the pre-processing is divided into training and test data The training data is passed to .fit() method of the hybrid model as X_train and y_train. Finally to make predictions on the test dataset we use the fitted .predict() method on the test predictor X_test. The last predictions are made hence using the meta-classifier and involve the results of various base classifiers to make sound decision making. Performance is measured using such metrics as accuracy_score as derived from the accuracy_score function under sklearn .metrics to clearly determine the generalization capability exhibited by the model. This is a good approach to model data complexed in NBA since it brings the well-handling of features from both LGBM and CatBoost as well as the boosting strength from Random Forest as well as AdaBoost that is also good against overfitting and noisy data. Stacking is a type of machine learning that has been shown, through the hybrid model discussed above, to be effective for the precise prediction of NBA games.

Chapter 5 Results

5.1 Results of Logistic Regression

The result in the form of Confusion Matrix Heatmap with the number of hits and the Classification Results and applied Logistic Regression is depicted in the below Fig: 5.1 This figure calibration shows a confusion matrix using color-code to represent performance of a logistic regression model. The matrix was basically a four-quadrant model that demonstrated the results of classification matrix with numbers. The purple box in the top right, features 150 such cases which likely constitute false positive whereas the orange box in the top left features 556 instances presumably true positive. The bottom left quadrant represents a score of 0 that belongs to the dark blue cluster, implying there are no missed cases or false negatives; the yellow cluster in the bottom right represents 706 indicating presumably the true negatives. This plot indicates colors from red (0) to violet (half) from left to right and blue (600+) from top to bottom and as depicted on the right vertical bar.

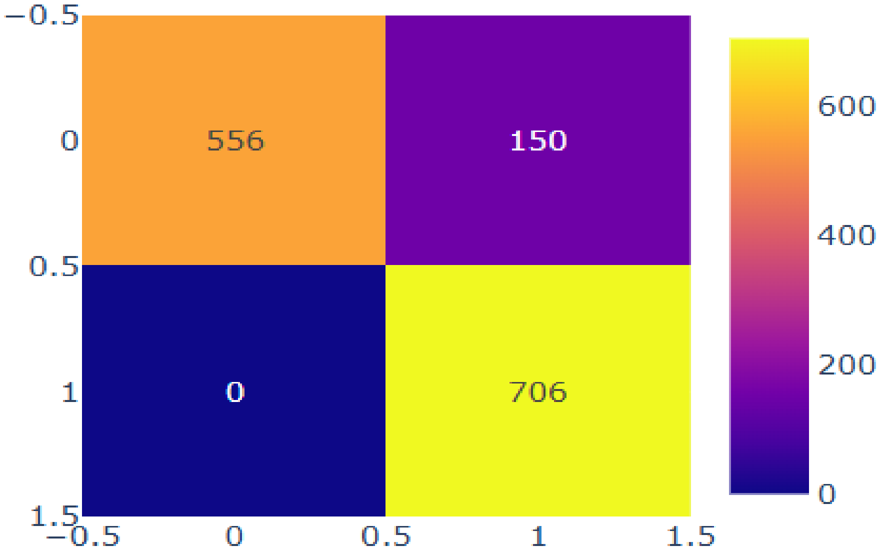
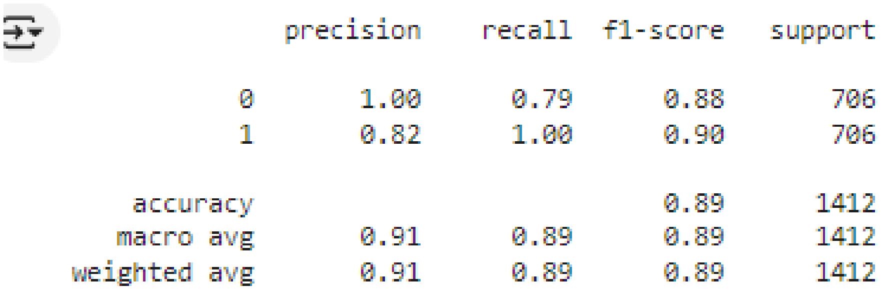


Figure 5.1: Confusion Matrix

Figure 5.2 is showing classification report of logistic regression which is having precision, recall, f1-score, support, accuracy and etc for this particular model.



```
              precision    recall  f1-score   support

           0       1.00      0.79      0.88       706
           1       0.82      1.00      0.90       706

    accuracy                           0.89      1412
   macro avg       0.91      0.89      0.89      1412
weighted avg       0.91      0.89      0.89      1412
```

Figure 5.2: Classification Report

5.2 Results of Decision Tree Classifier

Figure 5.3 also here showing Confusion Matrix Heatmap on Decision Tree Classification Results. The following figure shows a heatmap to represent the result in form of confusion matrix for a decision tree classifier in a color coded manner. The matrix characterized in this paper is in the format of 2×2 and the numerical values scale ranges from a dark blue hue to bright yellow, which is marked on the right side starting from 0 up to 600. The uppermost quadrant in the first quadrant (Yellow) represents 625 cases for true positive and the second quadrant at the top right (Dark Blue) shows 81 cases of false positive. The latest down-left corner (dark blue) is 114 false negatives and the down-right corner (yellow) is 592 true negatives. Both the dimensions are set out between -0.5 and 1.5 thus providing clear distinction between the classification results. The darkness of the color represents the value size, with yellow colouring for higher values , (around 600) while dark blue colours for the lowest values (around 100).

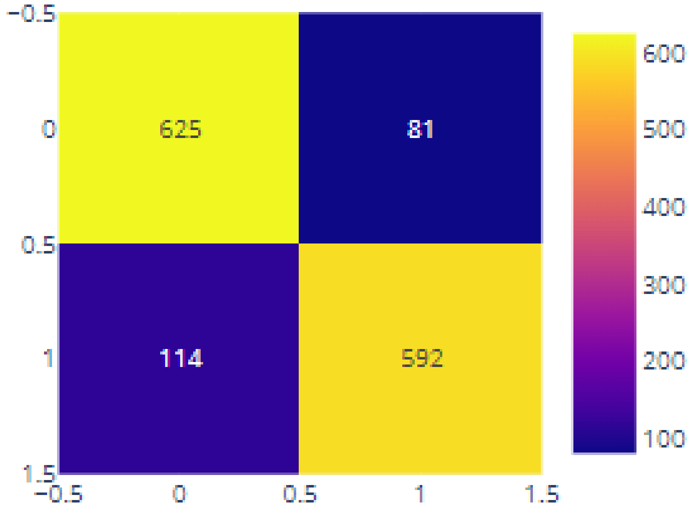
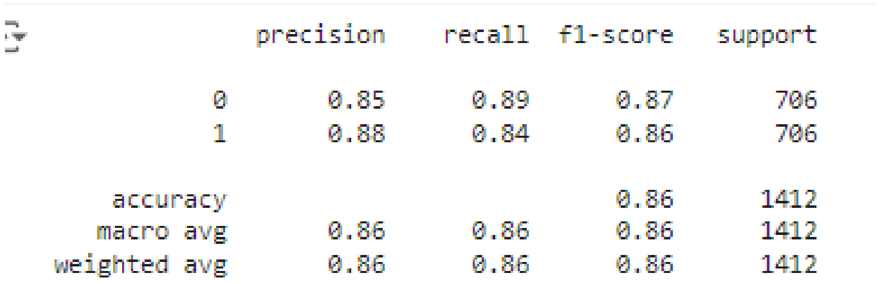


Figure 5.3: Confusion Matrix

Figure 5.4 is showing classification report of decision tree classifier which is having precision, recall, f1-score, support, accuracy and etc for this particular model.



```
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       706
           1       0.88      0.84      0.86       706

    accuracy                           0.86      1412
   macro avg       0.86      0.86      0.86      1412
weighted avg       0.86      0.86      0.86      1412
```

Figure 5.4: Classification Report

5.3 Results of Random Forest Classifier

Figure 5.5 is showing Random Forest Classification confusion matrix heatmap The Random Forest confusion matrix heatmap below will indicate the success rate of the algorithm. The following figure shows the confusion matrix of a random forest classifier with a color legend heatmap. The data is presented using a 2 by 2 matrix and the color intensity changes from dark blue to bright yellow, there is a color bar on the right ranging from 0 to 600. The upper left plot (yellow) shows 667 TP or true positive predictions, and the map on the right (dark blue) shows only 39 FP or false positive predictions. Bottom-left category or False negatives (dark blue) contains 66 while bottom-right or True negatives (yellow) contains 640. Both the axes are scaled from -0.5 to 1.5 to have a clear understanding and interfere section clearly defines the Class boundry. This visualisation shows that random forest classifier is better than simple ones, as there are much more correct predicted values (667 and 640) near the diagonal and fewer misclassified ones (39 and 66) are located far from it.

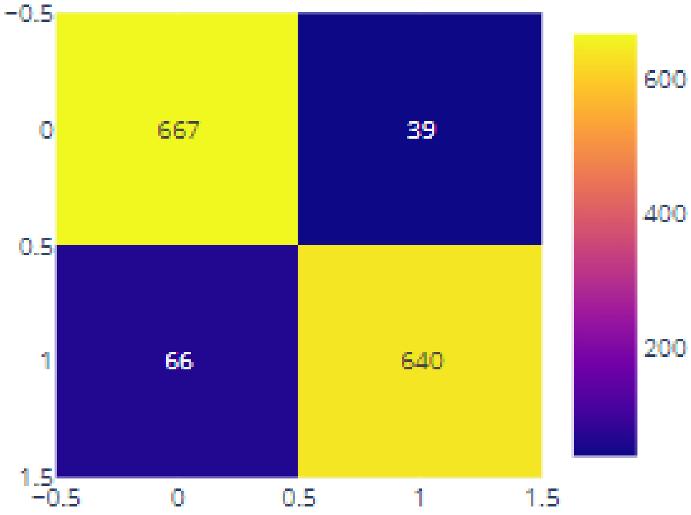
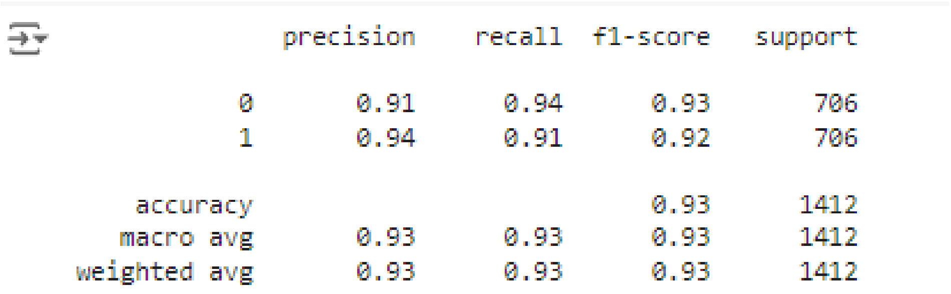


Figure 5.5: Confusion Matrix

Figure 5.6 is showing classification report of random forest classifier which is having precision, recall, f1-score, support, accuracy and etc for this particular model.



```
              precision    recall  f1-score   support

           0       0.91      0.94      0.93       706
           1       0.94      0.91      0.92       706

    accuracy                           0.93      1412
   macro avg       0.93      0.93      0.93      1412
weighted avg       0.93      0.93      0.93      1412
```

Figure 5.6: Classification Report

5.4 Results of AdaBoost Classifier

Figure 5.7 is showing Heatmap of confusion matrix of AdaBoost Classifier’s metric. In this figure, a confusion matrix is shown to classify an AdaBoost classifier with the help of a color-based heatmap. The actual matrix appears in 2 by 2 grid with values going in dark blue shade at the bottom and yellow at the top and the color bar on the right corner having values starting from 0 to 600. Looking only at true positive the top left quadrant (yellow) gives us 632 and for the false positive the top right quadrant (dark blue) gives us 74. Following the same intensity patterns, the lower left quadrant (dark blue) contains a meager 29 false negatives, hence relatively low, while the lower right quadrant (yellow) contains 677 true negatives. For clarity, both dimensions have scale between -0.5 and 1.5 to segregate classification results from the visualization.

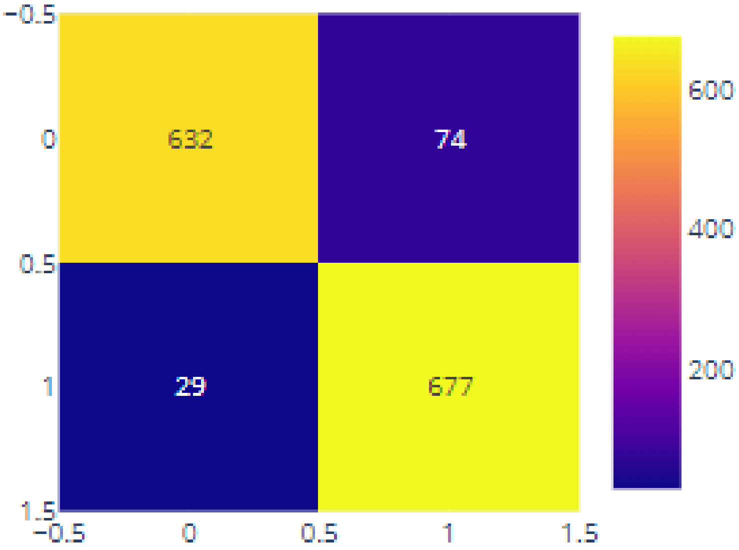
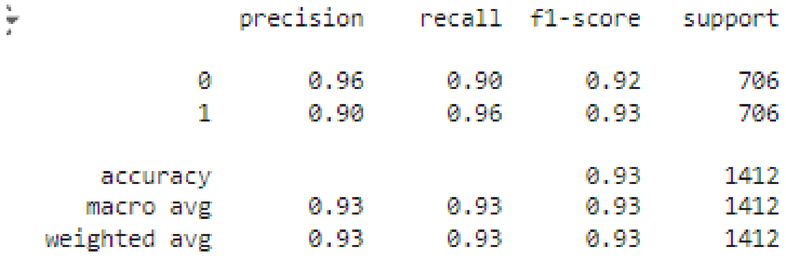


Figure 5.7: Confusion Matrix

Figure 5.8 is showing classification report of adaboost classifier which is having precision, recall, f1-score, support, accuracy and etc for this particular model.



```
              precision    recall  f1-score   support

           0       0.96      0.90      0.92       706
           1       0.90      0.96      0.93       706

    accuracy                           0.93      1412
   macro avg       0.93      0.93      0.93      1412
weighted avg       0.93      0.93      0.93      1412
```

Figure 5.8: Classification Report

5.5 Results of CatBoost Classifier

Figure 5.9 represents CatBoost Classification Results to shows the performance of a CatBoost classifier in terms of categorization, bunching separate images of objects together, based on the confusion matrix heatmap representation. The matrix is presented within a 2X2 grid with the colors drawn from the dark blue in the lower values up to the bright yellow in the higher values as indicated in the color scale bar on the right that is labeled from 0 to 600. The upper left quadrant, in yellow, shows that there were 657 true positives for the prediction, and the upper right quadrant, dark blue, shows that there were only 49 false positives. The extreme bottom-left cell (dark blue color) shows extremely low 21 false negative, and the extreme bottom-right cell (yellow color) shows 685 true negative. The axes on both dimensions start from -0.5 to 1.5 thereby providing clear definition of the classification. In this representation, the area of correct predictions is the largest for CatBoost classifier compared to other models.

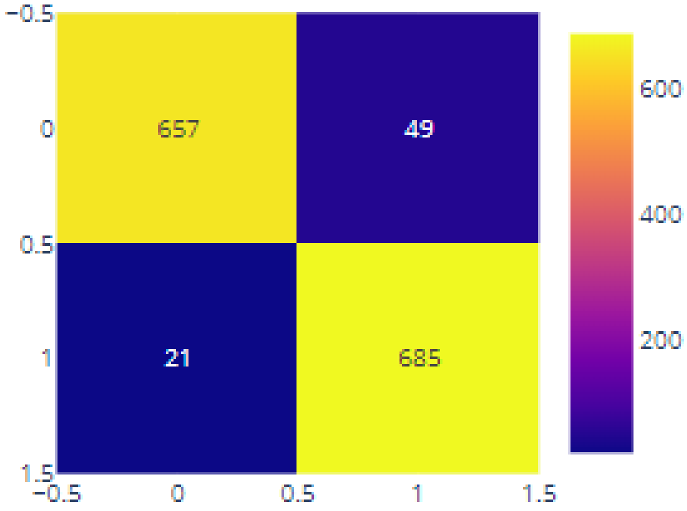
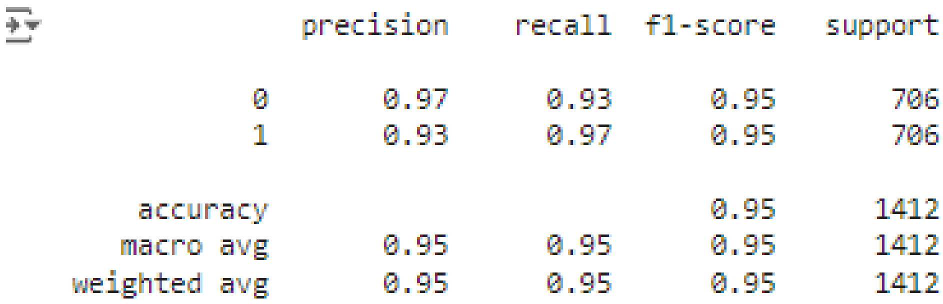


Figure 5.9: Confusion Matrix

Figure 5.10 is showing classification report of catboost classifier which is having precision, recall, f1-score, support, accuracy and etc for this particular model.



```
              precision    recall  f1-score   support

           0       0.97      0.93      0.95       706
           1       0.93      0.97      0.95       706

    accuracy                           0.95      1412
   macro avg       0.95      0.95      0.95      1412
weighted avg       0.95      0.95      0.95      1412
```

Figure 5.10: Classification Report

5.6 Results of Light GBM Classifier

The plot presented in Figure 5.11 is Confusion Matrix Heatmap on LightGBM Classification Opportunity. This figure shows the confusion matrix plot that illustrates selected LightGBM classifier performance using the heatmap. The matrix is organized as a 2 by 2 grid and includes the color intensity ranging from dark blue, representing low numerical values, to bright yellow with high numerical values, based on the color bar shown at the right side of the figure, where the values range from 0 to 600. The top-left quadrant in yellow represents 670 true positive predictions while the second top right in dark blue has 36 false positive. The bottom-left quadrant (red) has 61 false negatives, and bottom-right quadrant (green) has 645 true negatives. The axes on both dimensions range from -0.5 to 1.5 to achieve correct classification of results among the sets. This visual proof suggests a high accuracy of LightGBM classifier which with sizable correct predictions of 670 and 645 in the diagonal region while small number of mispredicted in the off diagonal region 36 and 61.

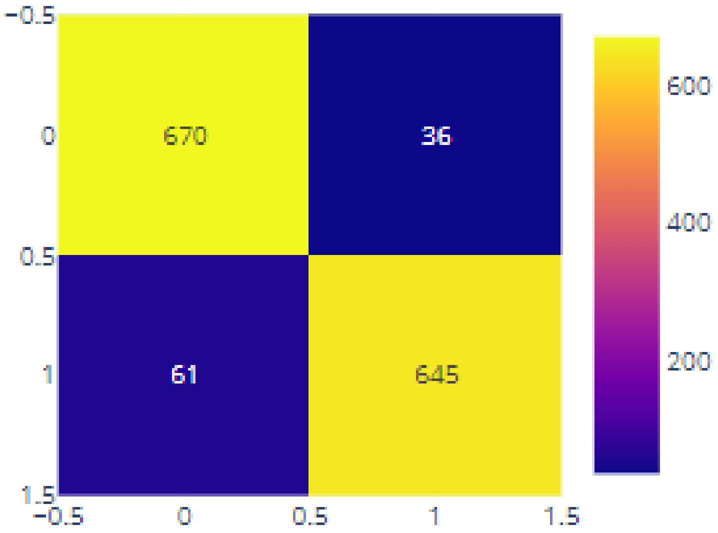
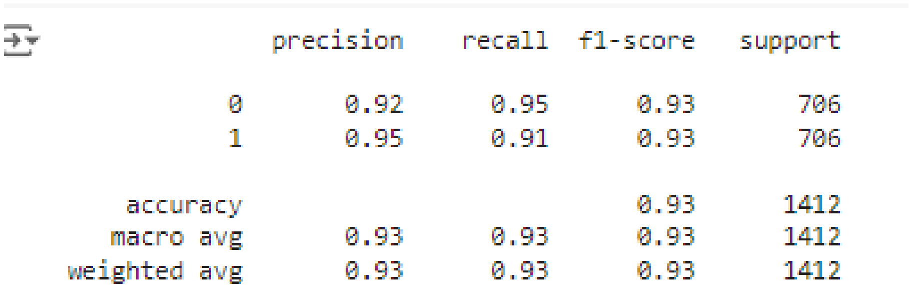


Figure 5.11: Confusion Matrix

Figure 5.12 is showing classification report of Light GBM which is having precision, recall, f1-score, support, accuracy and etc for this particular model.



```
              precision    recall  f1-score   support

           0       0.92      0.95      0.93       706
           1       0.95      0.91      0.93       706

    accuracy                           0.93      1412
   macro avg       0.93      0.93      0.93      1412
weighted avg       0.93      0.93      0.93      1412
```

Figure 5.12: Classification Report

5.7 Results of Hybrid Ensembled Classifier

Figure 5.13 is showing the Heatmap of Confusion Matrix for Improved LightGBM Classification. This figure is an improved confusion matrix matrix shown in heatmap format reveals a perfect outcome of a LightGBM classifier. The prevailing color palette is dark blue to light yellow with a key at the right bottom corner scale ranging between 0 and 600. Here, top-left corner (yellow zone) it has made 689 true positive predictions and the top-right corner (dark blue zone) has only 17 false positive results. The lower left quadrant (dark blue) shows an astonishingly low number of just 10 instances of false negative and the lower right quadrant (yellow) shows 696 true negative.

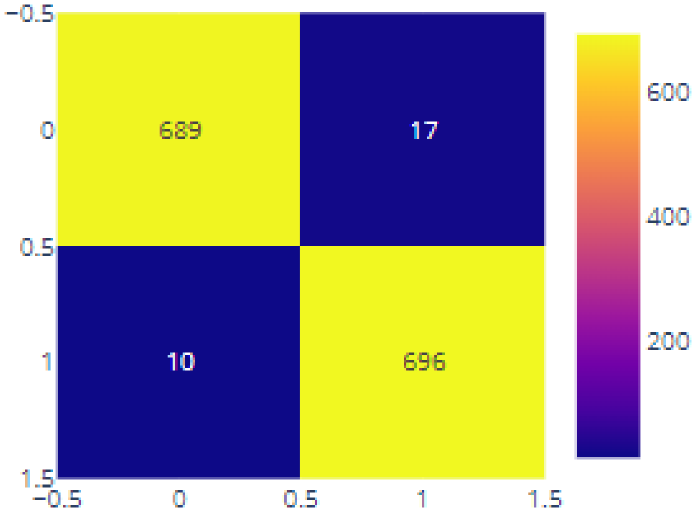
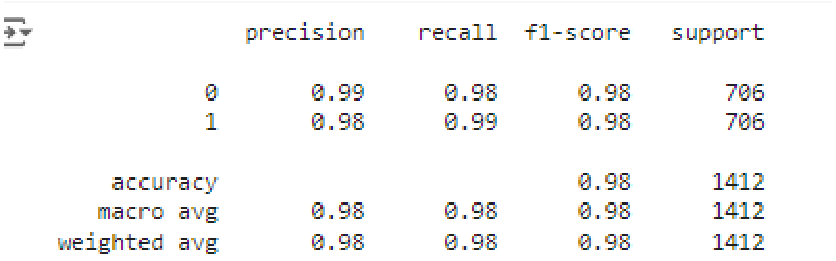


Figure 5.13: Confusion Matrix

Figure 5.14 is showing classification report of hybrid ensemble classifie which is having precision, recall, f1-score, support, accuracy and etc for this particular model.



```
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       706
           1       0.98      0.99      0.98       706

    accuracy                           0.98      1412
   macro avg       0.98      0.98      0.98      1412
weighted avg       0.98      0.98      0.98      1412
```

Figure 5.14: Classification Report

5.8 Comparison Evaluation for all models

The evaluation of various machine learning models for NBA game prediction highlights their predictive capabilities based on accuracy. Logistic Regression, a fundamental algorithm for binary classification, achieved an accuracy of 89%, demonstrating its efficiency with linear decision boundaries. The Decision Tree Classifier, known for its interpretability, performed slightly lower at 86%, reflecting potential overfitting issues. Ensemble methods like Random Forest and AdaBoost Classifiers displayed robust performance with 93% accuracy, benefiting from aggregated decision-making. Gradient-boosting algorithms CatBoost and LightGBM excelled in handling categorical and structured data, achieving accuracies of 95% and 93%, respectively. The Hybrid Ensembled Classifier, combining LGBMClassifier, CatBoostClassifier, and RandomForestClassifier as base learners with AdaBoostClassifier as the meta-classifier, outperformed all individual models with an impressive accuracy of 98%. This demonstrates the effectiveness of stacking multiple algorithms to leverage their strengths and mitigate individual weaknesses.

Table 5.1: Accuracy Comparison of all ML Models

Model	**Accuracy (%)**
Logistic Regression	89%
Decision Tree Classifier	86%
Random Forest Classifier	93%
AdaBoost Classifier	93%
CatBoost Classifier	95%
Light GBM Classifier	93%
Hybrid Ensembled Classifies	98%

Chapter 6 Discussion and Conclusion

6.1 Discussion

The approaches taken for the models done to predict NBA game outcomes were all found to have significant strengths and weaknesses. Logistic Regression algorithm proved rather simple and interpretable, yet its accuracy was rather moderate and could be used for benchmarking purposes only. Although Decision Tree Classifier was visually easy to understand its performance was poor mainly due to its overfitting. There was a remarkable increase in the prediction accuracies when more than one learner was created and applied within the Random Forest and AdaBoost ensembles. Gradient boosting techniques such as Catboost and LightGBM defined their prowess in over large structured data and categorical features giving high level of accuracy. Interestingly, the Hybrid Ensembled Classifier was identified as the model with the highest accuracy, with stacking averaging the best models, using CatBoost, LightGBM, and Random Forest and AdaBoost as the meta-model. This approach adequately recovered all intricacies and characteristics of NBA data, with better accuracy and broad applicability. The discussion specifies the fact that the selection of models and their settings is critical, where stacking methods allow for tuning the common benefits of various algorithmic approaches. Moreover, the proper pre-processing of dataset like the categorical features and the balance given to the sets was influential in enhancing the models professionalism.

6.2 Conclusion

This study provided a successful example of using machine learning for NBA games outcome prediction while the accuracy of each model was quite high. Comparative analysis approved the effectiveness of the set of features using more sophisticated tools such as ensemble and gradient boosting methodology to attain the highest predictive performance. The Hybrid Classifier Ensembled was the most efficient with the proposed stacked architecture obtaining a high accuracy of 98 percent. This evidence further supports the call for developing ensemble method since each model has individual strength for such multifaceted data analysis tasks. These results support the hypothesis that machine learning can be an effective technique for analyzing and predicting the performance of athletic teams, and provide valuable info for analysts, trainers, and members of NBA. Moreover, the role of feature extraction, as well as the choice of algorithms and parameters, in improving prediction reliability is described in the context of the study. They also show that, although separate models offer certain advantages, the use of ensembles offers a more holistic view of NBA game data. In conclusion, this work can be seen as a successful attempt to exemplify the benefits of using machine learning in the sport science context in particular, and in other predictive tasks in general.

6.3 Limitations and Future Works

Nevertheless, it should be understood that this study also carries some limitations even though their accuracy is rather high. The information used originated from 2019-2024, and though the dataset is complete, it means that it may contain volatile factors of players, strategies, etc., hence imprecise. Moreover, all the data used to develop the formula require historical information and therefore cannot predict very recent changes such as injuries, trades, or the emergence of new trends in a team’s performance. The hybrid ensemble approach as with any technique that combines multiple modules incurs a lot of computational cost and while capable is not as fit for real-time prediction or environments with limited resources. The future work may be directed to incorporating actual-time streaming data and more diverse DL algorithms, including recurrent neural networks (RNNs) and transformers to capture temporal dependencies and enhance the prognosis results, correspondingly. Extension of this feature space with regard to the game context such as player statistics, location of the game, or even climate may help improve the predictiveness of the model. Other approaches which could be investigated as part of the transfer learning strategies could also help in the porting of models to other sporting or other activities.

References

1. Adewusi, A.O., Okoli, U.I., Adaga, E., Olorunsogo, T., Asuzu, O.F. and Daraojimba, D.O., 2024. Business intelligence in the era of big data: a review of analytical tools and competitive advantage. *Computer Science & IT Research Journal*, *5*(2), pp.415-431.

2. Alonso, R.P. and Babac, M.B., 2022. Machine learning approach to predicting a basketball game outcome. *International journal of data science*, *7*(1), pp.60-77.

3. Bouchet, A., Troilo, M., Urban, T.L., Mondello, M. and Sutton, W.A., 2020. Business analytics, revenue management and sport: evidence from the field. *International Journal of Revenue Management*, *11*(4), pp.277-296.

4. Bunker, R. and Susnjak, T., 2022. The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, *73*, pp.1285-1322.

5. Chazan-Pantzalis, V., 2020. Sports Analytics Algorithms for Performance Prediction.

6. Chen, W.J., Jhou, M.J., Lee, T.S. and Lu, C.J., 2021. Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. *Entropy*, *23*(4), p.477.

7. de Sousa, D.S., 2022. Bringing objectivity and predictability to one of the most diverse and opinionated sports in the world by leveraging data.

8. Herberger, T.A. and Litke, C., 2021. The impact of big data and sports analytics on professional football: A systematic literature review. *Digitalization, digital transformation and sustainability in the global economy: risks and opportunities*, pp.147-171.

9. Horvat, T. and Job, J., 2020. The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), p.e1380.

10. Horvat, T., Havaš, L. and Srpak, D., 2020. The impact of selecting a validation method in machine learning on predicting basketball game outcomes. *Symmetry*, *12*(3), p.431.

11. Huang, M.L. and Lin, Y.J., 2020. Regression tree model for predicting game scores for the golden state warriors in the national basketball association. *Symmetry*, *12*(5), p.835.
12. Li, B. and Xu, X., 2021. Application of artificial intelligence in basketball sport. *Journal of Education, Health and Sport*, *11*(7), pp.54-67.
13. Li, Y., Wang, L. and Li, F., 2021. A data-driven prediction approach for sports team performance and its application to National Basketball Association. *Omega*, *98*, p.102123.
14. Rathi, K., Somani, P., Koul, A.V. and Manu, K.S., 2020. Applications of artificial intelligence in the game of football: The global perspective. *Researchers World*, *11*(2), pp.18-29.
15. Ratten, V. and Dickson, G., 2020. Big data and business intelligence in sport. In *Statistical modelling and sports business analytics* (pp. 25-35). Routledge.
16. Romaniuk, R., 2023. Fitting AdaBoost Models From Imbalanced Data with Applications in College Basketball.
17. Sarlis, V. and Tjortjis, C., 2020. Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, *93*, p.101562.
18. Sukumaran, C., Selvam, D., Sankar, M., Parthiban, V. and Sugumar, C., 2022. Application of Artificial Intelligence and Machine Learning to Predict Basketball Match Outcomes: A Systematic Review. *Computer Integrated Manufacturing Systems*, *28*, pp.998-1009.
19. Terner, Z. and Franks, A., 2021. Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, *8*(1), pp.1-23.
20. Zhao, K., Du, C. and Tan, G., 2023. Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm. *Entropy*, *25*(5), p.765.
21. Zuccolotto, P., Sandri, M. and Manisera, M., 2023. Spatial performance analysis in basketball with CART, random forest and extremely randomized trees. *Annals of Operations Research*, *325*(1), pp.495-519.