NBA MATCH PREDICTION

**COMP-7901-011 INDEPENDENT STUDIES**

**NAME:** LAKSHMI PRATHYUSHA CHELEMELLA

**UID:** U00846770

**MAJOR:** DATA SCIENCE

ABSTRACT

The goal of this project is to improve the accuracy of machine learning models for predicting NBA games by building upon existing techniques. To achieve this, we used Python to implement various machine learning models and worked with a dataset, containing team statistics for both home and visitor teams in each matchup. Additionally, we engineered two supporting features to aid the models’ performance. We evaluated six models on the training set, Logistic Regression, Random Forest Classifier, Support Vector Classifier and XGBoost Classifier. Random Forest Classifier and XGBoost Classifier outperformed the other models, and we refined them further to build a GUI to predict the outcome.

Dataset is taken from Kaggle, https://www.kaggle.com/code/stevejeongs/nba-my-data-analysis/input?select=games\_details.csv

INTRODUCTION

The National Basketball Association(NBA) is the world’s largest basketball association that generates an estimated 10 billion dollars annually. With such a massive amount of revenue, comes an abundance amount of data that is valuable for both teams and players. Analytics and modelling have become essential tools for sports teams and organizations to gain a competitive advantage. Decisions made during games are increasingly based on statistical analysis, and the outcome can often help to know the insights from the derived data. Predictive models have become more relevant in the industry, and the analysts now provide pre-game predictions for the winning team. To ensure that the betting process is fair, an accurate model is necessary to predict games and quantify all metrics in a basketball game.

LITERATURE REVIEW

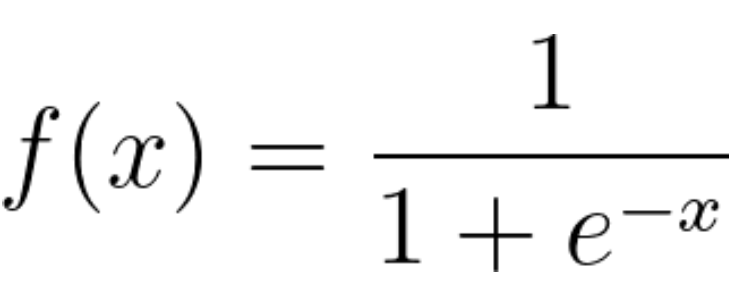
One of the fundamental machine learning techniques necessary for creating a successful model is cross-validation. It is important because it produces an output that is less biased than other methods and ensures that the entire dataset is used for testing. Additionally, the user can choose to stratify the data to ensure that each fold has an equal proportion of observations within a given value.

MACHINE LEARNING MODELS

The problem of determining which team will win in a basketball game can be classified as a binary classification problem, as it has two possible outcomes- either a win or a loss. As we are using four models to classify the outcome.

Logistic Regression :

Logistic regression is a statistical method that uses the logistic function to map y as a sigmoid function of x. It is basically employed to forecast a dependent variable that is categorical. The predicted outputs in a binary logistic regression model are limited to two possibilities, which are represented as 1 and 0. This model can also determine the ratio of success to failure or log odds.



Linear Regression:

Linear regression is a statistical method used to define a relationship between a response variable and a predictor variable. Generally, this method is used to to check the collinearity between two or more explanatory variables.

Random Forest Classifier :

Random Forest Classifier is a popular supervised learning algorithm used for both classification and regression problems. It tends to combine hundreds of decision trees and then trains each decision tree on a different sample of observations. It helps to find important features in the dataset. In our project, the classifier algorithm will create a decision tree for each sample selected, and voting is performed for every predicted result by using the method ‘mode’. The final prediction is developed from the most voted prediction.

Chart, radar chart

Description automatically generated

Support Vector Classifier :

SVM is a supervised machine learning technique that divides the datasets into classes to find the maximum marginal hyperplane. The support vectors are the data points that are closest to the hyperplane and the hyperplane is the plane that which is divided to make a decision for the model. SVM is a capable model, and it can handle multiple continuous and categorical variables.

XGBoost Classifier

XGBoost stands for eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. It is a classification algorithm that uses each desisior tree to verify the previous tree. The dataset is split into train-validation-test sets and XGBoost model is trained to evaluate probabilities on the validation of the dataset. The propability is then calibrated using Isotonic Regression.

DATA PREPROCESSING

String values in the dataset have been converted to numerical values by time mapping feature. Feature selection of relevant features(unique values) from the dataset is done by importing the training data to obtain accurate predictions. SelectKbest function is used to obtain the top 10 features form the dataset. Features such as Team\_name, L(loss), GP(games participation) & W(win) had the highest score.

DATA VISUALISATION

By using seaborn library to develop correlation matrix with heatmap, features which are most related to the target variable are identified. The correlation has coefficient has values between -1to 1. Feature value closer to 0 implies low correlation, a value close to 1 implies higher positive correlation and a value closer to -1 implies higher negative correlation. This method is used to check for any multi collinearity in the data.

Missing values and quality of the data: the Dataset has been cleaned before processing and has been checked for any null values. As there are no missing values, the quality of the dataset is good.

Data imbalance: There is no data imbalance in this dataset. If there is any imbalance in the observations, by eliminating the outliers, we can balance the data.

FEATURE ENGINEERING

Feature engineering is a technique that leverages data to create new variables that are not in the training set. It helps in the creation of new features for supervised and unsupervised learning, that speed up the data transformation and simplification of the data, resulting in better model accuracy. There are few features engineering techniques that helps to prepare the data for performing machine learning without any missing values that contribute to data flow interruptions and impact the performance of the machine learning models. Numerical imputation techniques is used to fill any gaps of information in the data X =X.fillna(0) is used to fill any missing values with 0

Outlier handling: this is used to remove any outlier from the dataset to produce more accurate data. This procedure is meant to be completed. Prior to model. There is no abnormal or inconsistent data observations in the dataset, as I have replaced the missing values with o.

One hot encoding is used to encode an element of a finite dataset, represented by the index 1 and all other elements are within the range (0,n-1) ensures each element has possible a unique value. To determine the relationship between features and the target, I have used matplotlib.pyplot to view the noise in the correlation of the data. There is a pattern between the total points scored and the average of the teams last 50 games played.

PRE-EXISTING MODEL RESEARCH AND COMPARSION

Point-based models aim at estimating the probability of winning single points within a match and then derive expressions for the prediction of the overall match. For example, historical match data to predict single points and calculate the probability of the outcome of the entire match based on a Markov chain. To analyse a Markov model that yields a betting return og about 4%, point-based models are used by a Bayesian hierarchical approach for match prediction.

RESULTS

The use of machine learning techniques is more of a novel area in sports prediction. Forecasting the result of a game is very complex and involves dealing with uncertain factors such as injuries, clutch performance etc,. Hence, a model that can predict the result of the game with an accuracy of over 50% is commendable. Despite a wide spectrum of approaches, data, calibrations, and evaluation metrics, a prediction accuracy of around 58% and 99% was observed for Logistic regression and Random Forest. Most studies agree that models are generally not able to beat the predictions, as data not more than one year are used for analysis.

To predict the outcome of NBA games, training and testing classifier models help to give the best prediction. Out of the five models used, Random forest, XGBoost and SVM give the best prediction, with the outcome 1, predicting the wins of the home team. As the location of the game matters, we have calculated the win for the home team.

MODEL DEPLOYMENT

All models are trained first and then the data is fed into the trained model, which improves the prediction accuracy, and the biases are minimised. After validation of various models, XGBoost is known to be the best operational model. We need to integrate the model into an environment, such as API/UI. This is achieved by incorporating the end user’s system via API/UI. Then the data is accessed, and the output is implemented. On the main page, we need to select the two teams that we need to predict, and then select the season type below and then opt for enter. We are then taken to a new page, where the winning team is declared, after calculating the previous years data.

FUTURE IMPROVEMENTS

A potential future implementation would be to fully deploy the model to a live statistics, to predict the current day’s NBA games. This would require the data to be updated concurrently, but it would be a realistic use case for the model.

|  |
| --- |
|  |
|  |  |

VIDEO LINK:

<https://youtu.be/2DhRSE0h7vs>

GITHUB LINK:

REFERENCES

1. Papadopoulos, Aris. “Machine Learning with the NBA's ‘Shooting Machine.’” Medium, Towards Data Science, 5 Sept. 2018, towardsdatascience.com/nba-data-science93e0314bb45e.
2. “Machine Learning Systems in Sports.” 12th Anniversary, NBAstuffer, 2020, [www.nbastuffer.com/analytics101/machine-learning-systems-in-sports/](http://www.nbastuffer.com/analytics101/machine-learning-systems-in-sports/).
3. Wang, Kuan-Chieh, and Richard Zemel. "Classifying NBA offensive plays using neural networks." Proceedings of MIT Sloan Sports Analytics Conference. Vol. 4. 2016.
4. Wheeler, Kevin. "Predicting NBA Player Performance." (2012).
5. Schumaker, Robert P., Osama K. Solieman, and Hsinchun Chen. "Predictive modeling for sports and gaming." Sports Data Mining. Springer, Boston, MA, 2010. 55-63.
6. Bunker, Rory P., and Fadi Thabtah. "A machine learning framework for sport result prediction." Applied computing and informatics 15.1 (2019): 27-33.
7. Caliwag, Jasmin A., et al. "Predicting Basketball Results Using Cascading Algorithm." Proceedings of the 2018 International Conference on Information Science and System. 2018.
8. Franks, Alexander, et al. "Counterpoints: Advanced defensive metrics for nba basketball." 9th Annual MIT Sloan Sports Analytics Conference, Boston, MA. 2015.