NBA PPG Leader Prediction

Team Members: Frankie Chukwudolue(Student ID: fchukwud/000983511), Ryan Shah (Student ID: rshah20/000940805)
IT4773
12/6/23

Introduction

Overview

The NBA Player Stats dataset is a tabular dataset that contains detailed information about the performance of NBA players in the 2023-24 season. It includes data on various player statistics such as points per game, rebounds per game, assists per game, field goal percentage, three-point percentage, and various other player performance metrics.

The overall objective of this project was to develop a regression model that predicts which star NBA player (which by the NBA's definition is a player that has made an all-NBA or all-star selection in the last 3 years.) will lead the league in points per game based on players' performance statistics. This model will be used to identify the top-scoring star players and make a prediction as to which will finish the season with the highest points per game average.

Components explored in constructed pipeline: Preprocessing and Analysis

Initial hypothesis

We hypothesized based on research of stats from previous seasons that Joel Embiid will lead the NBA in points per a game this season given his upward scoring trajectory and increased field goal attempts due to roster changes.

Preliminary Analysis

Dataset Description

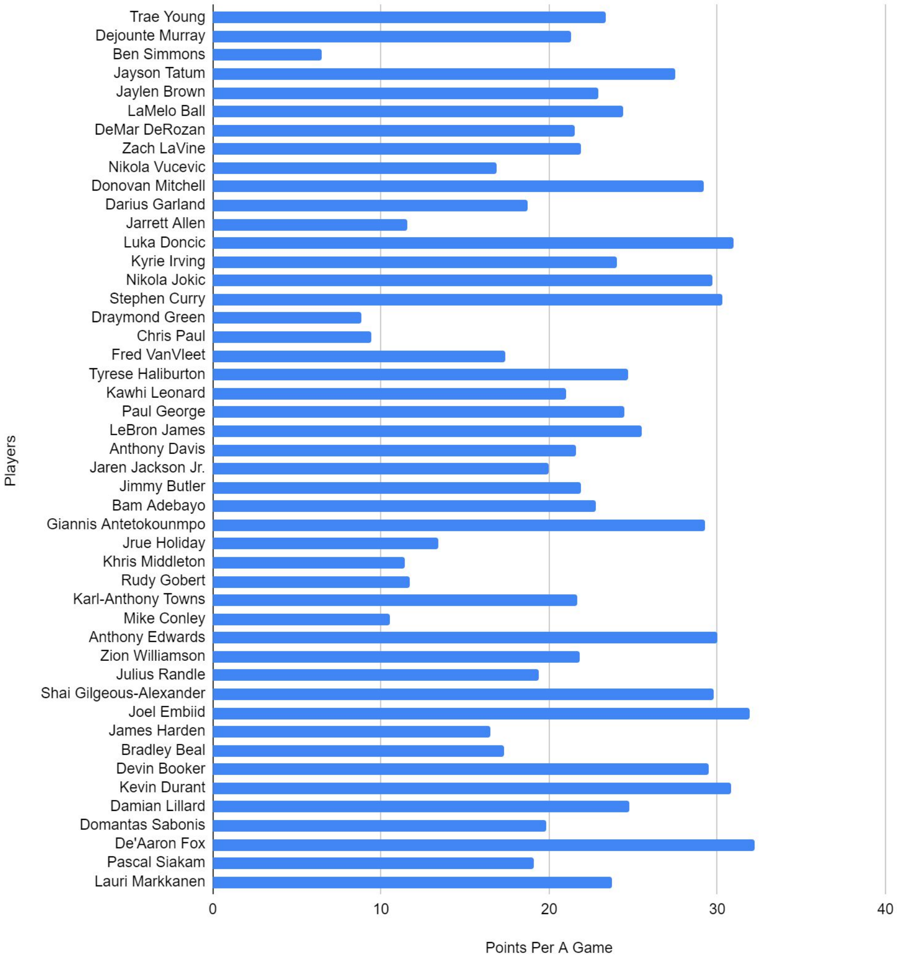
The dataset contains 2023-24 regular season NBA player stats per game. (https://www.kaggle.com/datasets/vivovinco/2023-2024-nba-player-stats/)

We decided to make a subset with only players defined as a “star” due to the dataset including hundreds of players that are in the NBA which would be too hard to handle so we narrowed it down to fewer players because the NBA has given a clear definition of how they define a star this season and the stars are the big names of the league.

Subset and explanation

List of players included in the subset: Trae Young, Dejounte Murray, Ben Simmons, Jayson Tatum, Jaylen Brown, LaMelo Ball, DeMar DeRozan, Zach LaVine, Nikola Vucevic, Donovan Mitchell, Jarrett Allen, Darius Garland, Luka Doncic, Kyrie Irving, Nikola Jokic, Stephen Curry, Draymond Green, Andrew Wiggins, Chris Paul, Fred VanVleet, Tyrese Haliburton, Kawhi Leonard, Paul George, LeBron James, Anthony Davis, Jaren Jackson Jr., Jimmy Butler, Bam Adebayo, Giannis Antetokounmpo, Jrue Holiday, Khris Middleton, Rudy Gobert, Karl-Anthony Towns, Mike Conley, Anthony Edwards, Zion Williamson, Julius Randle, Shai Gilgeous-Alexander, Joel Embiid, James Harden, Bradley Beal, Devin Booker, Kevin Durant, Damian Lillard, Domantas Sabonis, De'Aaron Fox, Pascal Siakam and Lauri Markkanen.

The visualization of the subset showcases the star players listed above and taking there points per a game data from the dataset to show the top scoring star players so far this season



Data Preprocessing

Step 1:

Create a subset for the data for ‘star’ players

Step 2:

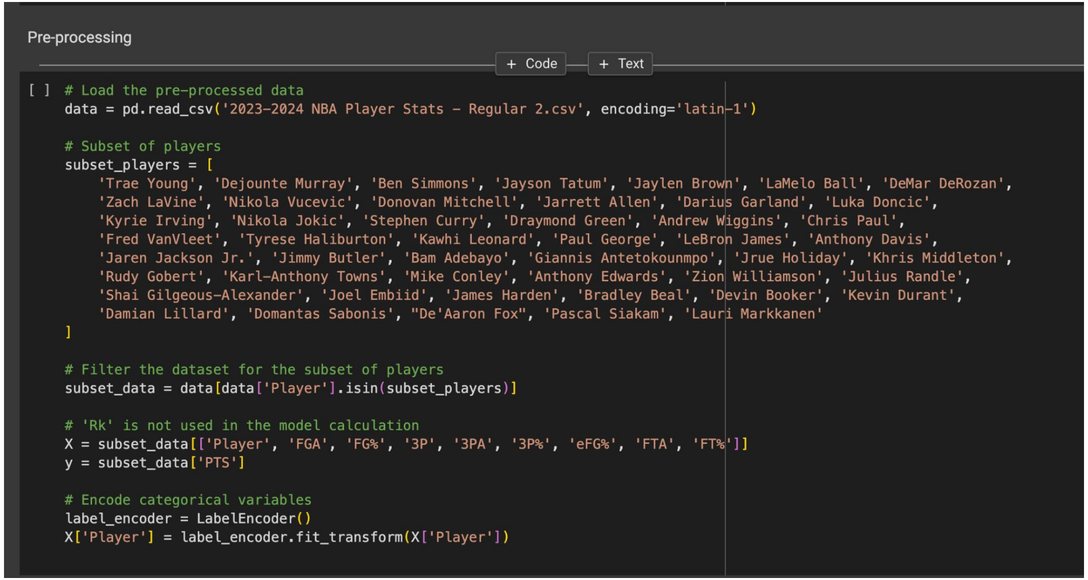
Filter the dataset for the data of the ‘star’ players listed in the subset

Step 3:

Input variables that impact players points per game to be used in the model calculation

Step 4:

Encode categorical variables



```
# Load the pre-processed data
data = pd.read_csv('2023-2024 NBA Player Stats - Regular 2.csv', encoding='latin-1')

# Subset of players
subset_players = [
    'Trae Young', 'Dejounte Murray', 'Ben Simmons', 'Jayson Tatum', 'Jaylen Brown', 'LaMelo Ball', 'DeMar DeRozan',
    'Zach LaVine', 'Nikola Vucevic', 'Donovan Mitchell', 'Jarrett Allen', 'Darius Garland', 'Luka Doncic',
    'Kyrie Irving', 'Nikola Jokic', 'Stephen Curry', 'Draymond Green', 'Andrew Wiggins', 'Chris Paul',
    'Fred VanVleet', 'Tyrese Haliburton', 'Kawhi Leonard', 'Paul George', 'LeBron James', 'Anthony Davis',
    'Jaren Jackson Jr.', 'Jimmy Butler', 'Bam Adebayo', 'Giannis Antetokounmpo', 'Jrue Holiday', 'Khris Middleton',
    'Rudy Gobert', 'Karl-Anthony Towns', 'Mike Conley', 'Anthony Edwards', 'Zion Williamson', 'Julius Randle',
    'Shai Gilgeous-Alexander', 'Joel Embiid', 'James Harden', 'Bradley Beal', 'Devin Booker', 'Kevin Durant',
    'Damian Lillard', 'Domantas Sabonis', "De'Aaron Fox", 'Pascal Siakam', 'Lauri Markkanen'
]

# Filter the dataset for the subset of players
subset_data = data[data['Player'].isin(subset_players)]

# 'Rk' is not used in the model calculation
X = subset_data[['Player', 'FGA', 'FG%', '3P', '3PA', '3P%', 'eFG%', 'FTA', 'FT%']]
y = subset_data['PTS']

# Encode categorical variables
label_encoder = LabelEncoder()
X['Player'] = label_encoder.fit_transform(X['Player'])
```

This code loads the data, creates the subset of ‘star’ players and the dataset is filtered. We then take that most impact scoring which we felt were the shooting and free throw metrics found in the dataset.

The output training and test datasets are now preprocessed and ready for model development.

Model Selection, Training and Fine Tuning

3 selected Training Models:

Linear Regression, Random Forest, and Gradient boosting

Justification

Linear Regression: We can use linear regression to predict the points per game based on various features such as the player's age, height, weight, position, and previous season's performance. Linear regression assumes a linear relationship between the features and the target variable, which can be a reasonable assumption for predicting points per game.

Random Forest: We can use a random forest regression model to predict the points per game. Random Forest can handle a large number of features and capture complex relationships between them, which can be beneficial when predicting NBA player performance.

Gradient Boosting: We can use a gradient-boosting regression model to predict the points per game. Gradient Boosting can effectively capture non-linear relationships between features and the target variable, which can be useful when predicting NBA player performance.

Pipeline of Model Training and Fine tuning

Method for Splitting Training/Validation/Test sets

The Holdout method- Involves randomly dividing the dataset into training and test sets with the test set being used to evaluate the training set.

List of Hyperparameters :

Linear regression model

- Fit intercept (True/False)
- Normalize input variables (True/False)

Random forest model

- Number of trees (n_estimators)
- Maximum depth of each tree (max_depth)
- Minimum number of samples required to split an internal node (min_samples_split)
- Minimum number of samples required to be at a leaf node (min_samples_leaf)

Gradient boosting model using the training data and tune the following hyperparameters:

- Number of boosting stages (n_estimators)
- Learning rate (learning_rate)
- Maximum depth of each tree (max_depth)
- Minimum number of samples required to split an internal node (min_samples_split)
- Minimum number of samples required to be at a leaf node (min_samples_leaf)

Evaluation, Results, and Analysis

Results of testing

Random Forest Evaluation:

Mean Squared Error: 7.917010302904162

Mean Absolute Error: 2.4229513989097327

R^2 Score: 0.7620772915515646

Linear Regression Evaluation:

Mean Squared Error: 0.9447601194384059

Mean Absolute Error: 0.8364855019782476

R^2 Score: 0.9716079835884011

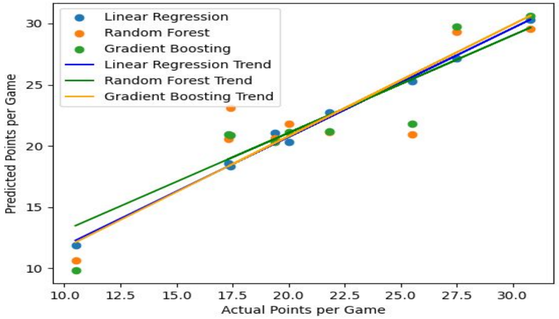
Gradient Boosting Evaluation:

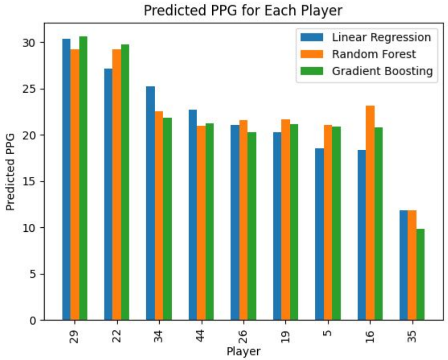
Mean Squared Error: 5.180805928172628

Mean Absolute Error: 1.8367889390834111

R^2 Score: 0.8443059524724401

Visualization





Model predictions

The Linear Regression and Gradient Boosting Models both predicted that Kevin Durant (represented by the ‘Rk’ value 110) would be the PPG leader. The Random Forest model predicted that Jayson Tatum(‘Rk’ 402) would lead the league in scoring.

Analysis

-From evaluation metrics given we can see that out of the three selected algorithms ,the Linear Regression model had the best metrics for this given task. Linear Regression Model had the lowest Mean Squared Error and Mean Absolute Error out of the three which suggests its a better overall fit for the data. It also had the Highest R^2 score indicating once again that the model fits the data better than the other two.

Recommended model:

Linear Regression model

Conclusion

Summary

In conclusion, based off of evidence gathered during this investigation we determined that the linear regression model is best suited to solve the posed question of who will lead the league in points per game this season. The model predicted that player ‘Rk’ 110(which is the ‘Rk’ of Kevin Durant) will lead the league in points per game for the 2023-24 NBA season thus disproving our initial hypothesis.

Challenges faced

-Using a dataset with many variables that are not tangible

-Choosing which variables are more significant than others and do they impact the accuracy of the calculation

-Time Management while working on this project

Thank You for Your Time.