NBA PLAYER STATISTICAL ANALYSIS AND PREDICTION PROJECT

COLLABORATORS:

JACOB EVANS
KARAN ANAND
MERT OZTOP
PRATIK PUROHIT
PRIYA MARINGANTI



OVERVIEW



This project aims to develop a basketball player statistics analysis and prediction system using machine learning techniques with NBA player datasets. The system utilizes historical player data to provide insights into player performance and predict player stats for the upcoming season, by leveraging the power of machine learning algorithms and the comprehensive NBA player statistics dataset.

Objectives

- Predict basketball player stats for the upcoming season based on historical data
- Leverage player stats from the previous year to forecast player performance
- Identify suitable machine learning algorithms for scalability, accuracy, and interpretability in predicting player performance
- Evaluate the performance of the prediction system and ensure its effectiveness in real-world scenarios, such as team selection and player scouting

METHODOLOGY



Methodology



- 1. Data Collection: Gather a large dataset of NBA player statistics from the Kaggle dataset
- 2. Data Preprocessing: Clean, normalize, encode, and engineer features from the NBA dataset
- 3. Exploratory Data Analysis: Gain insights, patterns, and analyze feature distributions
- 4. Model Training and Evaluation: Experiment with various machine learning algorithms, fine-tune models for accurate predictions
- 5. User Interface Development: Create an intuitive interface for users to input player data and view predictions
- 6. Testing and Validation: Ensure the accuracy, robustness, and scalability of the prediction system

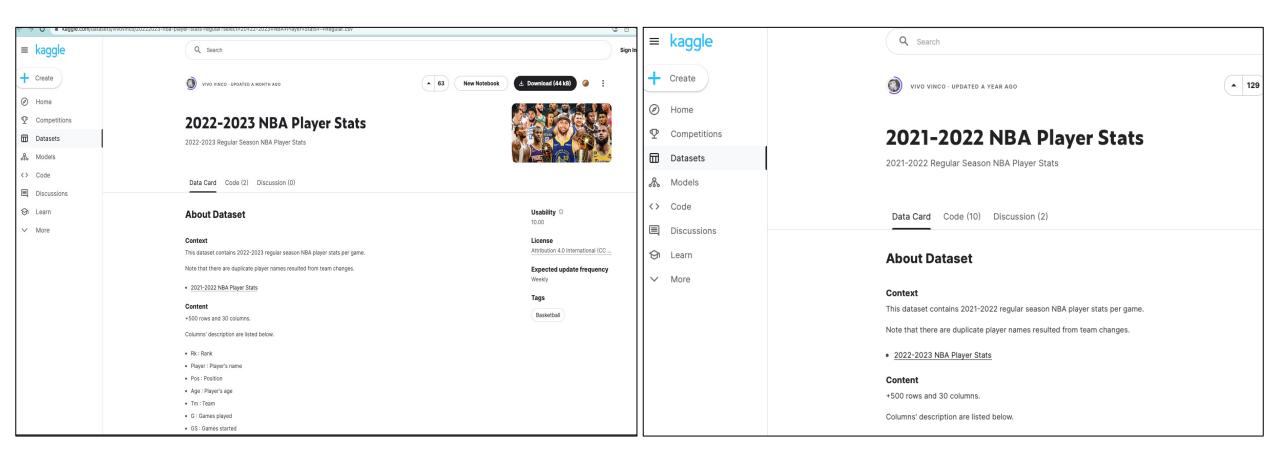
SOURCE OF DATA

- 1. <u>Kaggle (2022-23)</u>
- 2. Kaggle (2021-22)
- Loodibee (Team Logos)
- 4. NBA (Head Shots)



DATA SOURCE - KAGGLE

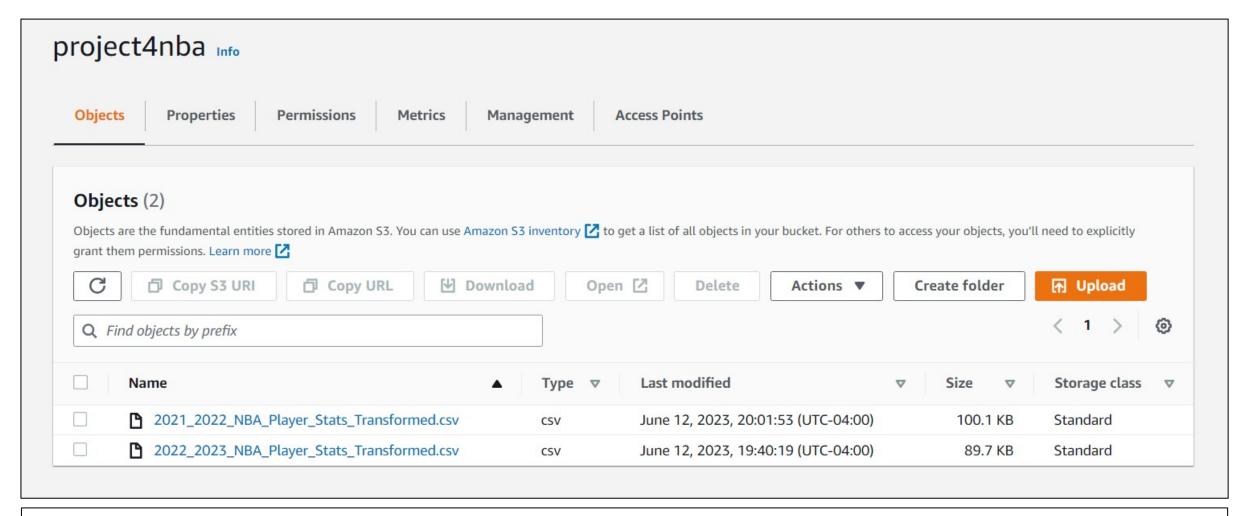




Retrieving last two seasons' data from Kaggle

DATA ON AWS





Uploading the data sets on AWS for the main code

SETTING UP ENVIRONMENT



```
import os
 # Find the latest version of spark 3.x from http://www.apache.org/dist/spark/ and enter as the spark version
 # For example:
 # spark version = 'spark-3.3.1'
 spark version = 'spark-3.3.2'
 os.environ['SPARK VERSION']=spark version
 # Install Spark and Java
 !apt-get update
 !apt-get install openjdk-11-jdk-headless -qg > /dev/null
 !wget -q http://www.apache.org/dist/spark/$SPARK VERSION/$SPARK VERSION-bin-hadoop3.tgz
 !tar xf $SPARK VERSION-bin-hadoop3.tgz
 !pip install -q findspark
 # Set Environment Variables
 os.environ["JAVA HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
 os.environ["SPARK HOME"] = f"/content/{spark version}-bin-hadoop3"
 # Start a SparkSession
 import findspark
 findspark.init()
 Hit:1 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ InRelease
 Hit:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2004/x86 64 InRelease
 Get:3 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
 Hit:4 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu focal InRelease
 Hit:5 http://archive.ubuntu.com/ubuntu focal InRelease
 Hit:6 http://archive.ubuntu.com/ubuntu focal-updates InRelease
 Hit: 7 http://ppa.launchpad.net/cran/libgit2/ubuntu focal InRelease
 Get:8 http://archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
 Hit:9 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu focal InRelease
 Hit:10 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu focal InRelease
 Hit:11 http://ppa.launchpad.net/ubuntugis/ppa/ubuntu focal InRelease
 Fetched 222 kB in 2s (130 kB/s)
 Reading package lists... Done
```

Installed the necessary tools and environment of Spark and Java

CREATING SPARK SESSION



```
# Import packages
    from pyspark.sql import SparkSession
    from pyspark.sql import Row
    from pyspark.sql.types import StructType, StructField, StringType, DateType, IntegerType
    from pyspark import SparkFiles
    import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LinearRegression
    from sklearn.ensemble import RandomForestRegressor
    from pyspark.sql.functions import col
    import numpy as np
    import matplotlib.pyplot as plt
    # Create a SparkSession
    spark = SparkSession.builder.appName("NBA_Prediction").getOrCreate()
[ ] url2022 = 'https://project4nba.s3.amazonaws.com/2022 2023 NBA Player Stats Transformed.csv'
    url2021 = 'https://project4nba.s3.amazonaws.com/2021 2022 NBA Player Stats Transformed.csv'
[ ] # SparkFiles will allow you to resolves paths to files added through `SparkContext.addFile`
    spark.sparkContext.addFile(url2022)
    spark.sparkContext.addFile(url2021)
    data2022 = spark.read.option('header', 'true').option("encoding", "utf-8").csv(SparkFiles.get("2022 2023 NBA Player Stats Transformed.csv"), inferSchema=True, sep=',')
    data2021 = spark.read.option('header', 'true').option("encoding", "utf-8").csv(SparkFiles.get("2021_2022_NBA_Player_Stats_Transformed.csv"), inferSchema=True, sep=';')
[ ] # Show DataFrame
```

Created a Spark Session

data2022.show()

SPARK DF TO PANDA DF



```
[ ] #Converting Spark DF to Pandas DF
    nba 2022 2023 df = data2022.toPandas()
[ ] #Converting Spark DF to Pandas DF
    nba 2021 2022 df = data2021.toPandas()
[ ] nba_2022_2023_df

    Precious Achiuwa

                                    TOR 55 12 20.7 3.6 7.3 ... 0.702 1.8 4.1 6.0 0.9
               Steven Adams
                                29 MEM 42 42 27.0 3.7 6.3 ... 0.364 5.1 6.5 11.5 2.3 0.9 1.1 1.9 2.3 8.6
                                   MIA 75 75 34.6 8.0 14.9 ... 0.806 2.5 6.7 9.2 3.2 1.2 0.8 2.5 2.8 20.4
                Bam Adebayo
                Ochai Agbaji
                                22 UTA 59 22 20.5 2.8 6.5 ... 0.812 0.7 1.3 2.1 1.1 0.3 0.3 0.7 1.7 7.9
                Santi Aldama
                                22 MEM 77 20 21.8 3.2 6.8
                                                              ... 0.750 1.1 3.7 4.8 1.3 0.6 0.6 0.8 1.9
    674 535 Thaddeus Young
                               34 TOR 54 9 14.7 2.0 3.7 ... 0.692 1.3 1.8 3.1 1.4 1.0 0.1 0.8 1.6 4.4
    675 536
                                    ATL 73 73 34.8 8.2 19.0
                                                             ... 0.886 0.8 2.2 3.0 10.2 1.1 0.1 4.1 1.4 26.2
                 Trae Young
              Omer Yurtseven
                                    MIA 9 0 9.2 1.8 3.0 ... 0.833 0.9 1.7 2.6 0.2 0.2 0.2 0.4 1.8 4.4
    677 538
                 Cody Zeller
                                     MIA 15 2 14.5 2.5 3.9
                                                              ... 0.686
                                                                       1.7 2.6 4.3 0.7 0.2 0.3 0.9 2.2 6.5
    678 539
                 Ivica Zubac
                                25 LAC 76 76 28.6 4.3 6.8 ... 0.697 3.1 6.8 9.9 1.0 0.4 1.3 1.5 2.9 10.8
    679 rows x 30 columns
[ ] nba_2021_2022_df
```

Converted Spark DF to Panda DF

DATA CLEANING



```
#Merging the two datasets together
joined_data = nba_2022_2023_df.merge(nba_2021_2022_df, on='Player', how='outer')
```

```
#Dropping Nulls
joined_data.dropna(inplace = True)
```

```
joined_data = joined_data.set_index('Player')
```

Cleaned the data

DATA MODELLING



```
# Select the features (X) and target variables (y) for points (PTS)

features = joined_data.drop(['PTS_x', 'Pos_x', 'Pos_y', 'Tm_x', 'Tm_y'], axis = 1)

target_pts = joined_data[['PTS_x']]

# Split the data into training and test sets
X_train, X_test, y_train_pts, y_test_pts = train_test_split(features, target_pts, test_size=0.25, random_state=4:

# Create separate models for points
model_pts = LinearRegression()

# Train the models
model pts.fit(X_train, y_train_pts)
```

Modeled the data

DATA ACCURACY



```
# Calculate R-squared for the model
r_squared = model_pts.score(X_train, y_train_pts)
r_squared
```

0.9998864936377487

Calculated the R-squared for the model

MAKING PREDICTIONS



```
# Make predictions for points
  pts_predictions = model_pts.predict(full_set)
  player_names = full_set.index.unique()
  player names
 Index(['Tyus Jones', 'Goran Dragi?', 'Larry Nance Jr.', 'Evan Fournier',
         'Davon Reed', 'James Harden', 'Devin Vassell', 'Luka Don?i?',
        'Eugene Omoruyi', 'Ish Smith',
        'Thanasis Antetokounmpo', 'Dejounte Murray', 'Monte Morris',
        'Bobby Portis', 'Khem Birch', 'Klay Thompson', 'Cory Joseph',
        'Darius Garland', 'Garrett Temple', 'Isaiah Jackson'],
       dtype='object', name='Player', length=428)
  # Import Random
  import random
  # Make prediction for randomly selected player
  player names = joined data.index.unique()
  select_player = random.choice(player_names)
  player row = joined data[joined data.index == select player]
  player_features = player_row.drop(['PTS_x', 'Pos_x', 'Pos_y', 'Tm_x', 'Tm_y'], axis = 1)
  # Get the actual points for the player
  pts_actual = player_row[['PTS_x']].values[0]
  pts_actual_2021 = player_row[['PTS_y']].values[0]
  # Predict points and assists for the player
  pts predicted = model pts.predict(player features)[0]
  print("Player:", select_player)
  print("2021-2022 Points:", pts actual 2021)
  print("2022-2023 Points:", pts_actual)
  print("2023-2024 Predicted Points:", pts_predicted)
Player: Luke Kornet
2021-2022 Points: [2.]
2022-2023 Points: [3.8]
2023-2024 Predicted Points: [3.71710539]
```

Made predictions for the points

CREATING A FILE FOR TABLEAU



```
# Create a copy of pandas df
pandas df with predictions = joined data.copy()
# Add the predictions as new columns
pandas df with predictions['2023-2024 Predicted Points'] = pts predictions
# Set negative predictions to 0
pandas_df_with_predictions.loc[pandas_df_with_predictions['2023-2024 Predicted Points'] < 0, '2023-2024 Predicted
#Reset index and add as column
pandas df with predictions.reset index(drop = False)
# Specify the path where you want to save the file
from google.colab import drive
drive.mount('/content/drive')
file path = '/content/drive/MyDrive/NBA Predicted Data.csv'
# Save the DataFrame as a CSV file
pandas df with predictions.to csv(file path, encoding='utf-8', index=False)
```

Downloaded the final data set for Tableau

TABLEAU DASHBOARD - PLAYER STATS AND PREDICTION



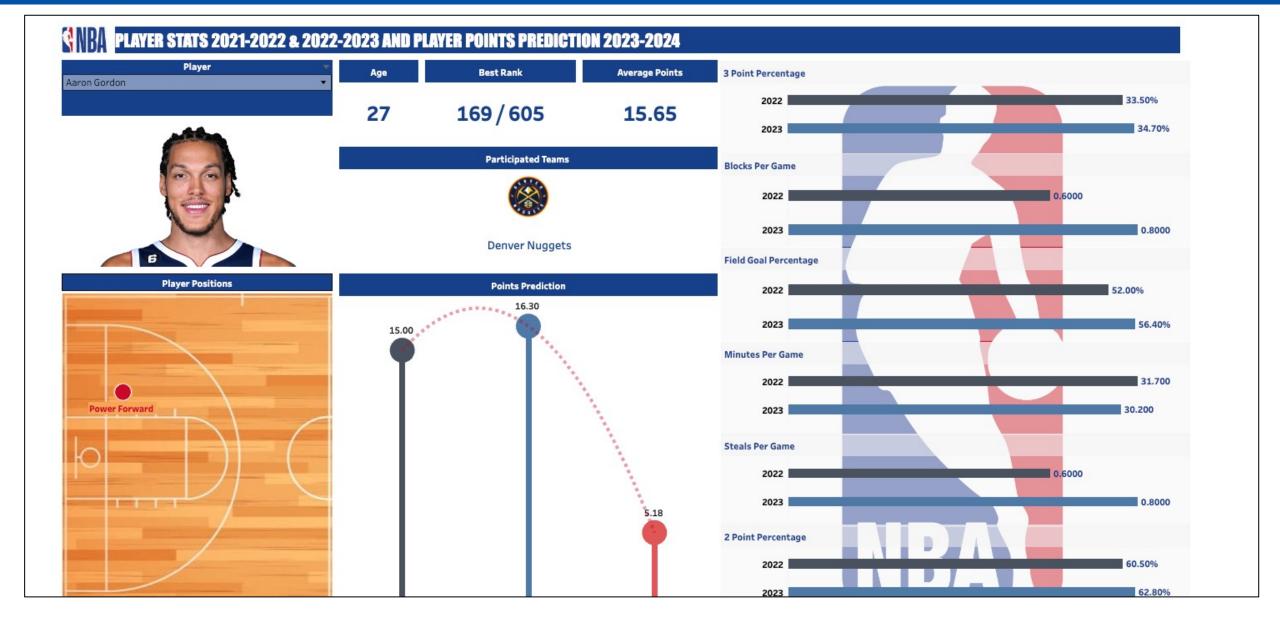


TABLEAU DASHBOARD – LEADERBOARD







Average Points (2021-2022 & 2022-2023)



Joel Embiid 31.850



Giannis Antetokounmpo 30.500



Luka Don?i? 30.400



Shai Gilgeous-Alexander 27.950



Stephen Curry 27.450



Devin Booker 27.300



Zach LaVine 24.600



De'Aaron Fox 24.100



Brandon Ingram 23.700



Karl-Anthony Towns 22.700

Average Assists (2021-2022 & 2022-2023)



James Harden 10,425



Trae Young 9.950



Chris Paul 9.850



Tyrese Haliburton 8,900



Nikola Joki? 8.850



Luka Don?i? 8.350



Darius Garland 8,200



LaMelo Ball 8,000



Dejounte Murray 7.650



Russell Westbrook 7,425

Average Three Point Percentage (2021-2022 & 2022-2023)



Jay Scrubb



Jordan Schakel 58%



Richaun Holmes 51%



50%



Luke Kennard 48%



Vlatko ?an?ar 48%



Patrick Williams



Gary Payton II



Terry Taylor 45%



Joe Harris 45%

Average Two Point Percentage (2021-2022 & 2022-2023)





















CHALLENGES



- Bad Encoding: The original data had encoding issues that we struggled to handle. We had to apply encoding techniques to ensure proper handling and interpretation of the data.
- External Factors: While player statistics provide valuable insights, it's important to note that other factors can influence a player's performance on the court. Factors such as injuries, team dynamics, coaching strategies, and external circumstances were not included in our analysis. Considering these external factors could further enhance the accuracy and predictive power of the model.
- •Outliers: To handle these outliers, we implemented a post-processing step where we replaced any negative predicted values with zeros. This approach allowed us to address the outliers and ensure that the predicted statistics remain within a valid range. By zeroing out the negative values, we mitigated the impact of outliers on the model's performance and ensured that the predicted player statistics align with the expectations of NBA player performance.

RESULT



The prediction system achieved **R-squared value of 0.9998**, indicating a high level of accuracy in predicting player statistics based on the historical data.

The NBA Player Statistics Analysis and Prediction System leverages historical player data, applies machine learning techniques, and provides valuable insights and predictions on player performance. The system can assist with team selection, player scouting, and forecasting player statistics for the upcoming season.

THANK YOU!

QUESTIONS ARE WELCOME!

COLLABORATORS:

JACOB EVANS
KARAN ANAND
MERT OZTOP
PRATIK PUROHIT
PRIYA MARINGANTI

