

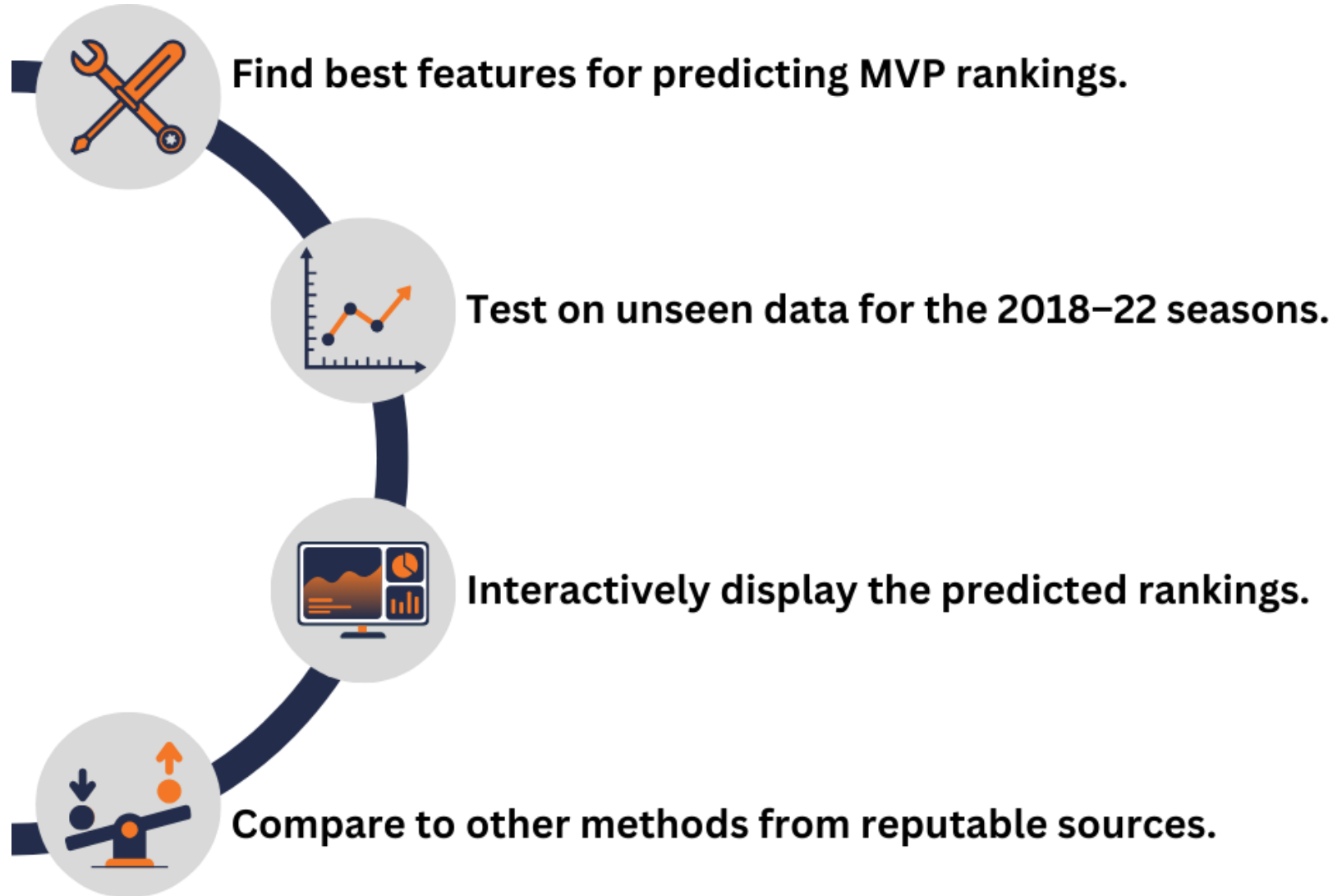
MVP Feature Importance

Predicting the NBA's Most Valuable Player

Group 7: Sarah Elmasry, Ran Gao, Wyatt Scott, and Rishi Sharma



Goals



Dataset

- The dataset from [JK-Future](#) includes statistics relevant to determining the MVP.
- After cleaning and preprocessing, the dataset contains **11,282** entries with **56** columns spanning the 1980–2022 seasons, with variables covering:
 - **Player Info**: name, height, weight, season year, age, position.
 - **Game Stats**: games played, minutes, field goals, 3-pointers, etc.
 - **Advanced Stats**: efficiency rating, win shares, plus/minus metrics, VORP.
 - **Other**: MVP ranking, team wins, conference, etc.

```
df.columns
```

```
Index(['name', 'height', 'weight', 'Season', 'Age', 'Pos', 'G', 'GS',  
'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%', 'eFG%',  
'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV',  
'PF', 'PTS', 'PER', 'TS%', '3PAr', 'FTr', 'ORB%', 'DRB%', 'TRB%', 'AST%',  
'STL%', 'BLK%', 'TOV%', 'USG%', 'OWS', 'DWS', 'WS', 'WS/48', 'OBPM',  
'DBPM', 'BPM', 'VORP', 'Rank', 'mvp_share', 'Trp Db1', 'conference', 'W',  
'Rk_Year', 'Overall'], dtype='object')
```

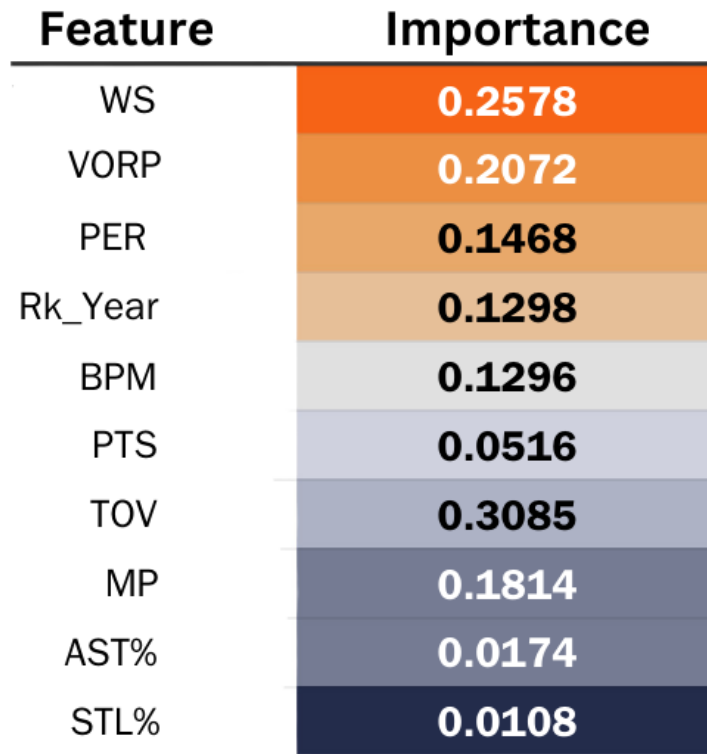


Data Preprocessing

Data Cleaning	Handle missing values, encode categorical features, scale numeric features, combine features where possible
Train/Test/Validate	Split data into training, testing, and validation sets
Feature Selection	Reduce model and index complexity via feature selection and variable importance analysis
Model Validation	Leverage five-fold cross validation for model stability and reliability



Feature Importance, Model Comparison



Scale (Jenks natural breaks, 10 steps)



Model Performance Comparison

Others Best Model

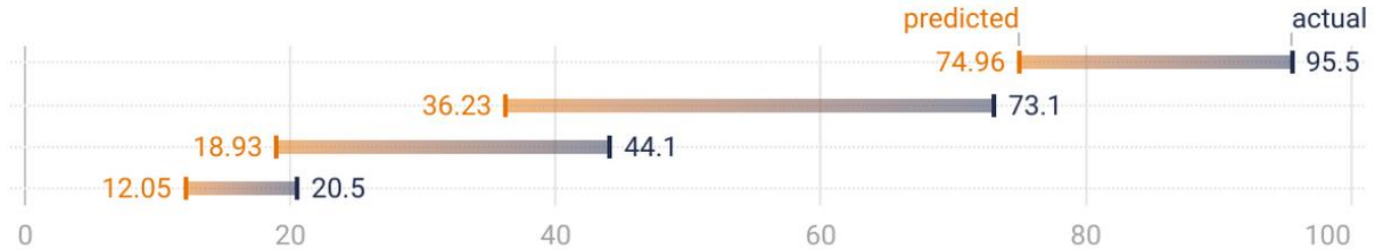




Predicted vs. Actual (mvp_share)

2018

James Harden
LeBron James
Anthony Davis
Damian Lillard



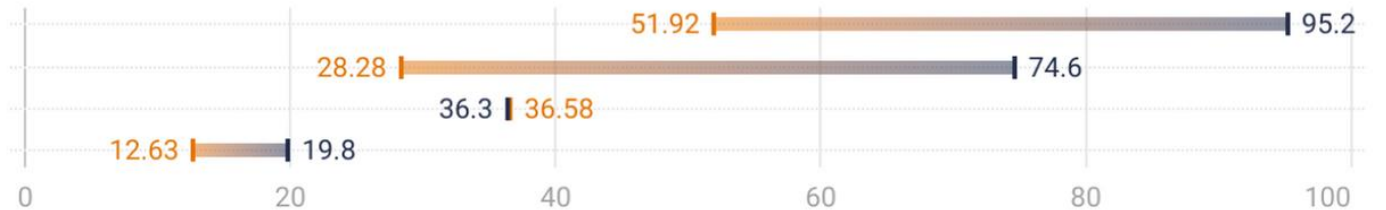
2019

G. Antetokou...
James Harden
Paul George
Nikola Jokic



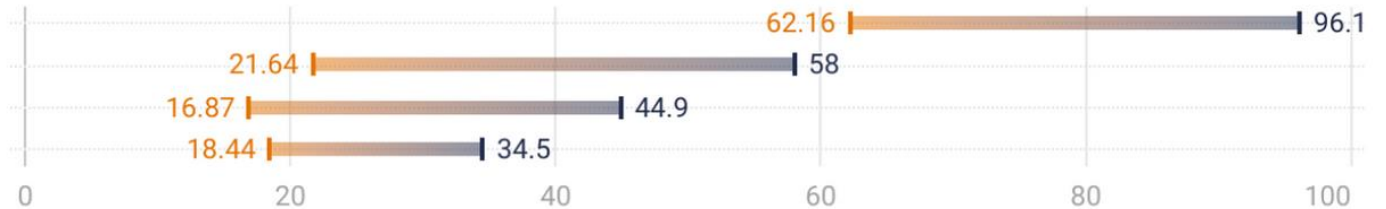
2020

G. Antetokou...
LeBron James
James Harden
Luka Doncic



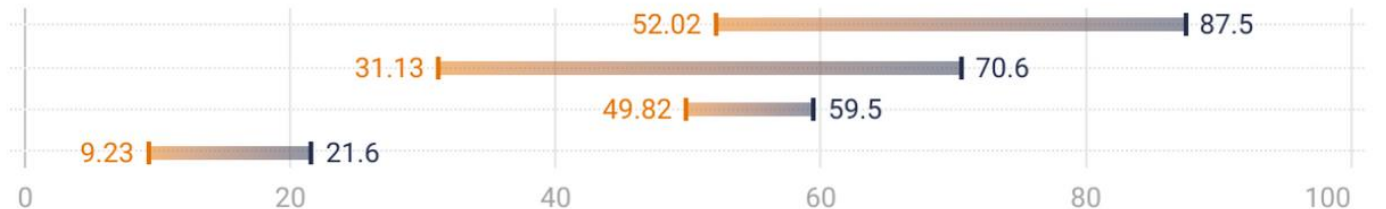
2021

Nikola Jokic
Joel Embiid
Steph Curry
G. Antetokou...

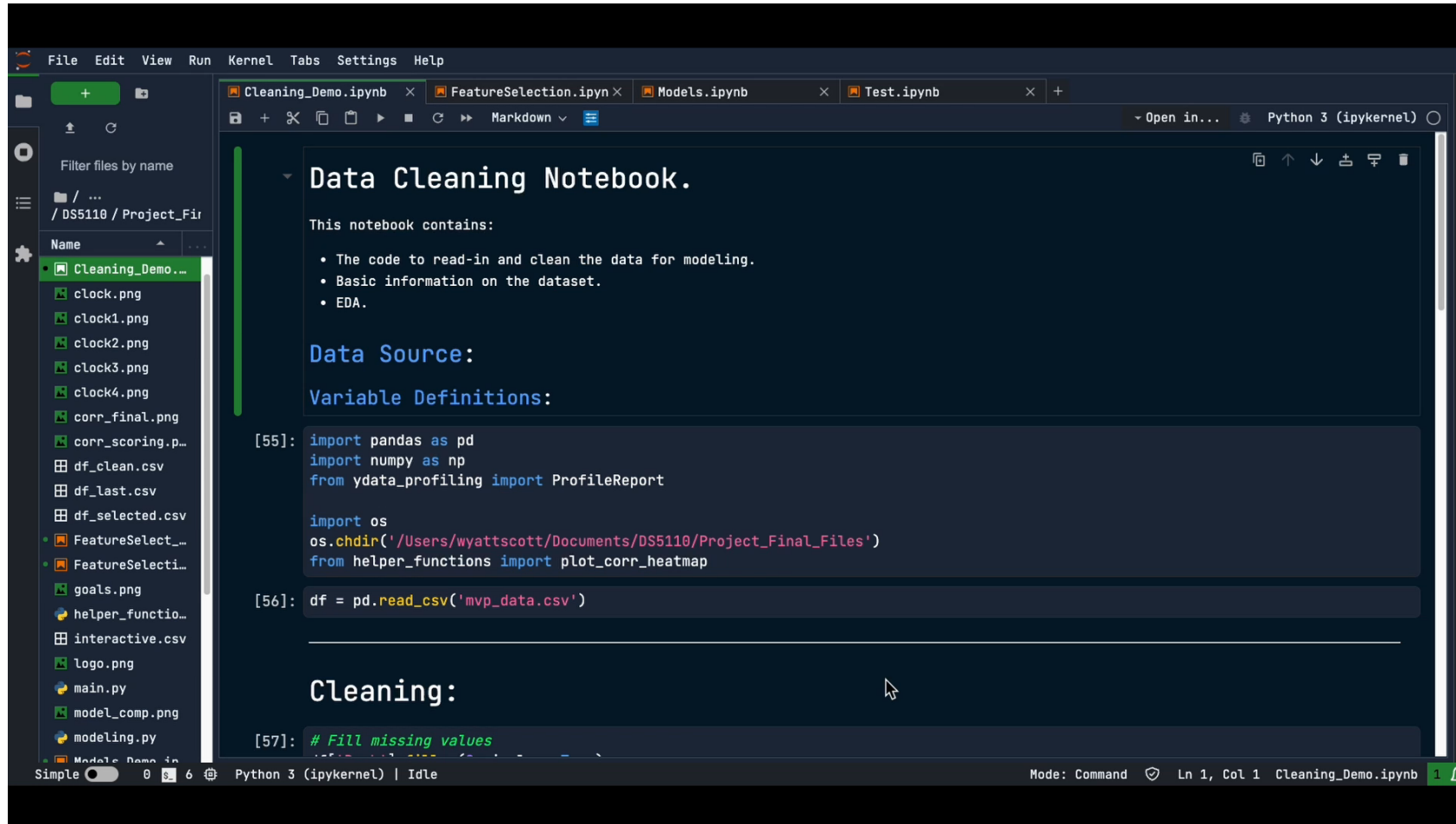


2022

Nikola Jokic
Joel Embiid
G. Antetokou...
Devin Booker



Code and Visualizations Demo



The screenshot shows a Jupyter Notebook interface with a dark theme. The left sidebar displays a file explorer for the directory `/DS5110 / Project_Final`. The main area shows the 'Cleaning_Demo.ipynb' notebook. The notebook content includes a title 'Data Cleaning Notebook.', a list of contents, a 'Data Source:' section, 'Variable Definitions:', and a 'Cleaning:' section with code cells.

Data Cleaning Notebook.

This notebook contains:

- The code to read-in and clean the data for modeling.
- Basic information on the dataset.
- EDA.

Data Source:

Variable Definitions:

```
[55]: import pandas as pd
import numpy as np
from ydata_profiling import ProfileReport

import os
os.chdir('/Users/wyattscott/Documents/DS5110/Project_Final_Files')
from helper_functions import plot_corr_heatmap
```

Cleaning:

```
[57]: # Fill missing values
```

The status bar at the bottom indicates 'Python 3 (ipykernel) | Idle' and 'Mode: Command'.



Conclusions

Ways to Improve:

Predict a complete voting distribution, compare to other historical methods

Ways to Expand:

Assign points for media member vote ranks, predict top 10 players in consideration for the award

Adding functionality:

Handle the weights of media members' top 5 votes

Used for others:

Another model to predict MVP



Thank You!