# Decoding NBA Greatness

A Data-Driven Journey from Metrics to Mastery

Presented by Wenli

# Introduction

### Overview

This project analyzes NBA playoff player statistics from 2005 to 2024 using ML techniques. Through PCA for dimensionality reduction, K-Means Clustering for player grouping, and Linear Regression, Random Forest & NN for performance prediction, we aim to uncover patterns and provide actionable insights.

### Objective

- Identify distinct playing styles.
- Enhance predictions of Player Impact Estimate.
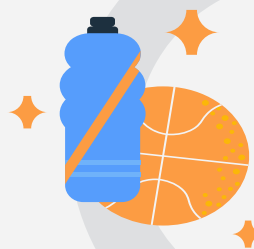- Support player development, scouting, and strategy decisions.

# 01

# Data Exploration & Diagnostics

# Dataset Summary

- **Years**
  - 2005–2024 (~210 data points/year)
- **Player Metrics**
  - Minuted Played, Offensive Rating, Defensive Rating, Effective Field Goal Percentage, Shooting Percentage, Usage, Assistant Percentage, Offensive Rebounding, Rebounding, Pace, Age.
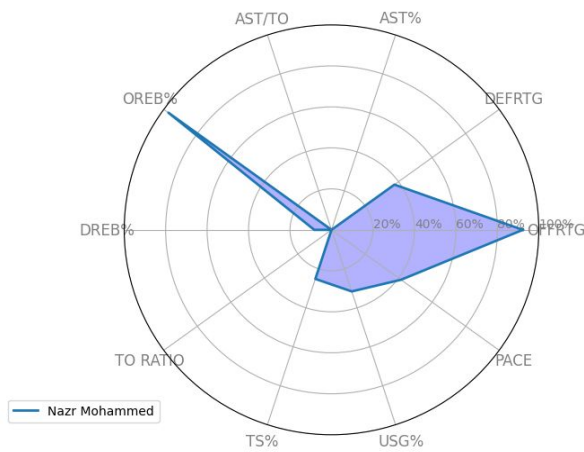- **Target**
  - Player Impact Estimate



Top 10 Players by Player Impact Estimate (PIE)
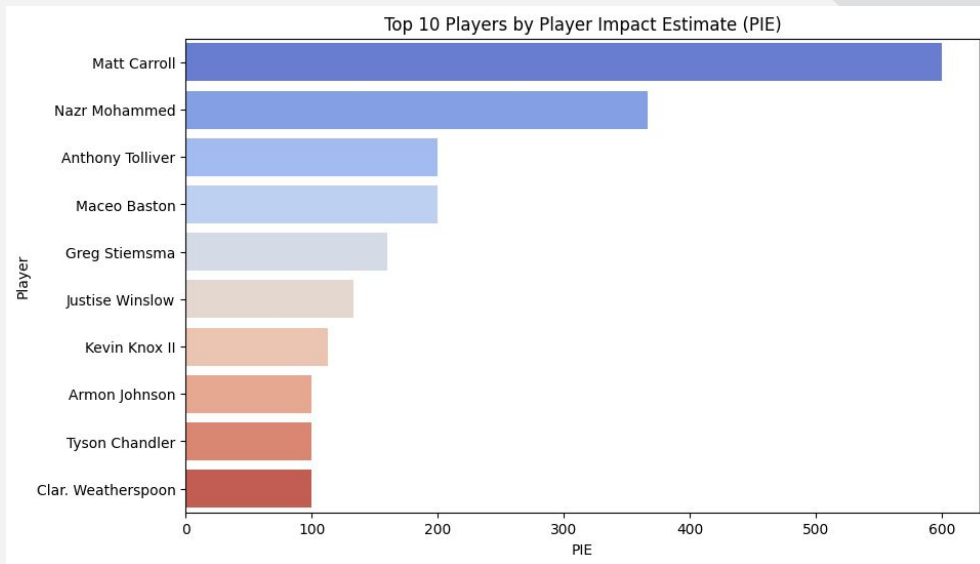
# Selected Top Player Profile
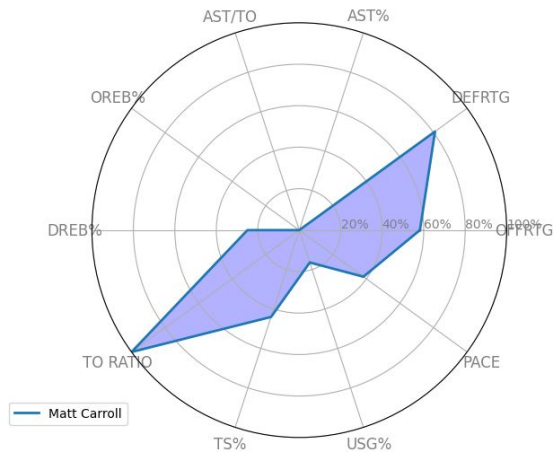


*Strong performance in Offensive Rating and Usage.*

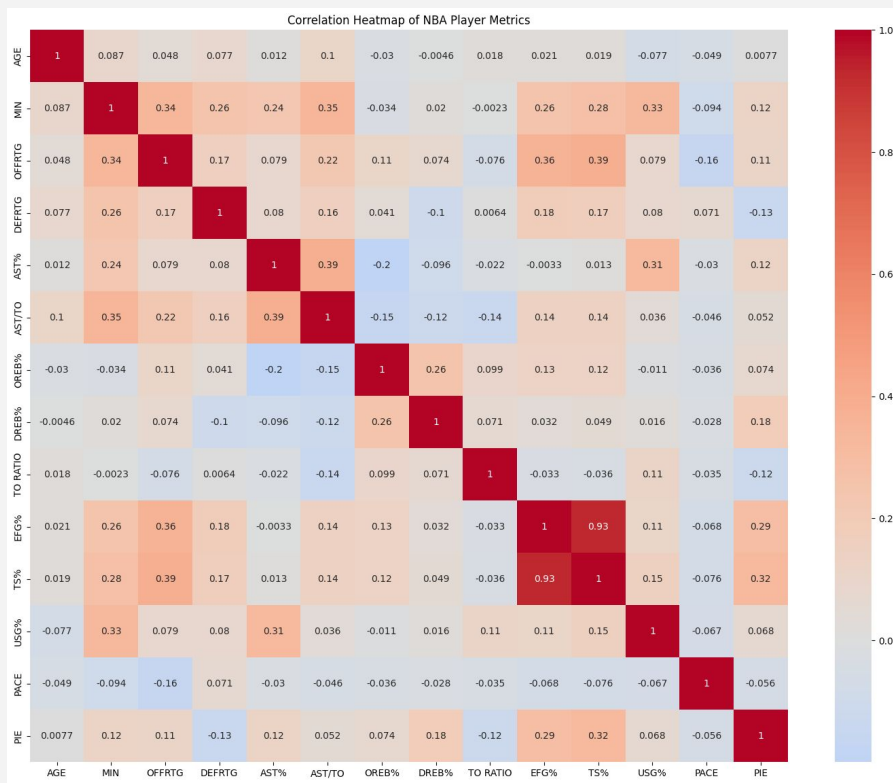*Strong performance in Offensive Rebounding.*
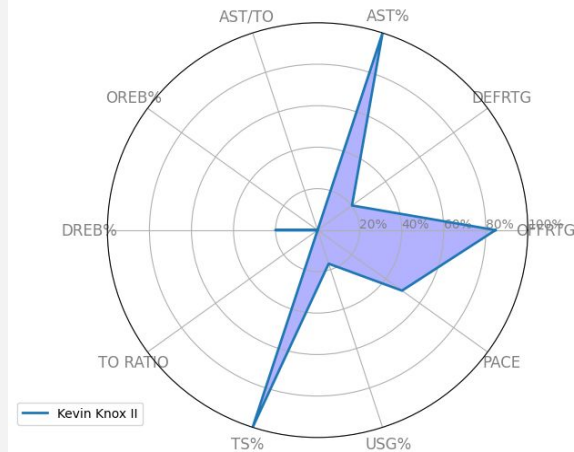
*Strong performance Defensive Rating and Usage.*

# Correlation Heatmap of Player Metrics
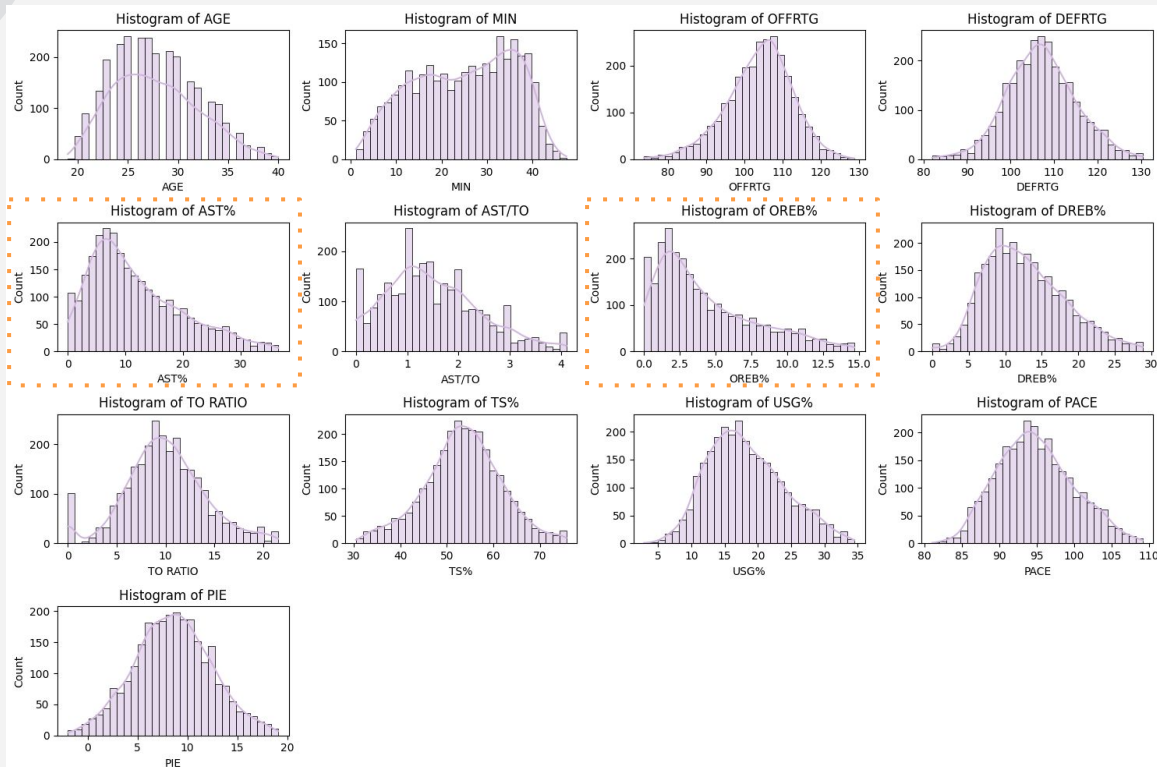


Correlation Heatmap of NBA Player Metrics

*High correlation between EFG% and TS%.*

# Outlier Analysis
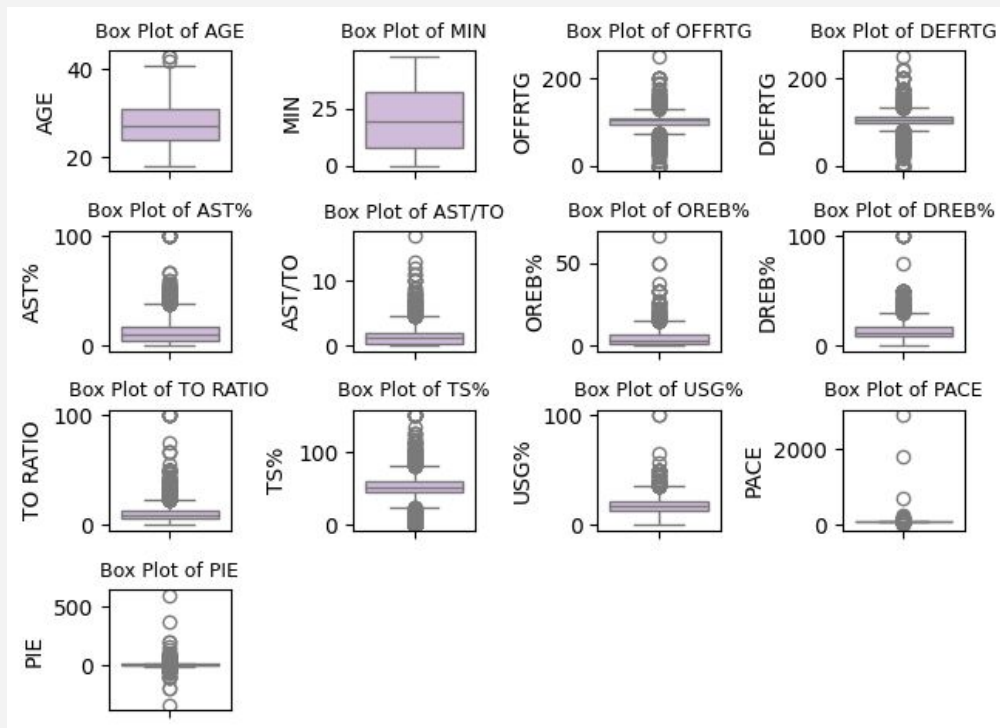


*Significant variability in most Player Metrics, especially for offensive/defensive ratings, assist percentages, and rebounding metrics.*
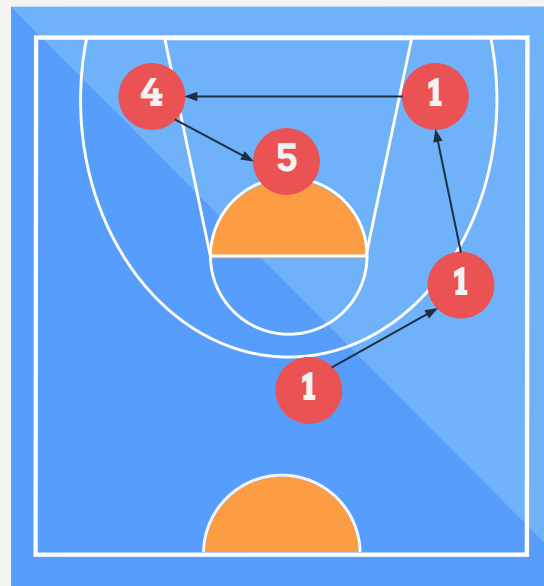
# Skewness Analysis



| Variables | Skewness Value |
|-----------|----------------|
| OREB% | 0.95 |
| AST% | 0.84 |
| AST/TO | 0.52 |
| DREB% | 0.52 |
| USG% | 0.41 |
| AGE | 0.38 |
| PACE | 0.28 |
| DEFRTG | 0.08 |
| TO RATIO | 0.07 |
| PIE | 0.06 |
| TS% | -0.08 |
| MIN | -0.2 |
| OFFRTG | -0.37 |

*OREB% and AST% show moderate skewness.*

# 02

# Base Model

# Linear Regression Summary

| Variable | Coefficient | Std Error | t-Statistic | P-Value |
|---|---|---|---|---|
| const | 5.27 | 0.77 | 6.80 | 0.00 |
| AGE | -0.02 | 0.01 | -3.27 | 0.00 |
| MIN | 0.06 | 0.00 | 16.69 | 0.00 |
| OFFRTG | -0.04 | 0.00 | -10.01 | 0.00 |
| DEFRTG | -0.09 | 0.00 | -23.86 | 0.00 |
| AST% | 1.28 | 0.04 | 34.94 | 0.00 |
| OREB% | 0.48 | 0.04 | 13.28 | 0.00 |
| DREB% | 0.26 | 0.01 | 41.38 | 0.00 |
| TO RATIO | -0.22 | 0.01 | -29.59 | 0.00 |
| TS% | 0.26 | 0.00 | 64.73 | 0.00 |
| USG% | 0.22 | 0.01 | 34.81 | 0.00 |
| PACE | -0.02 | 0.01 | -3.91 | 0.00 |
| R-squared | | | | **84%** |

Note: AST/TO is removed from the regression model with a p-value of 0.18.
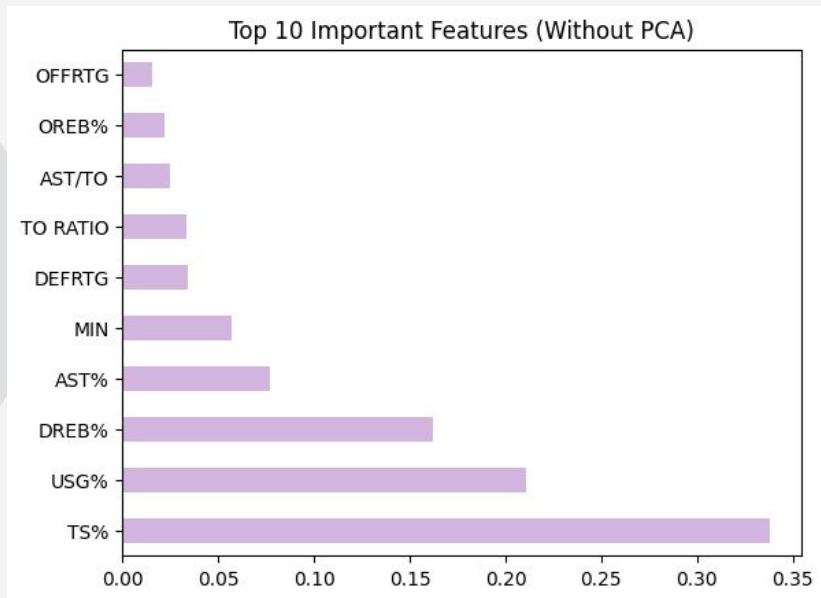
# Model Diagnostics



Residuals vs Fitted Values
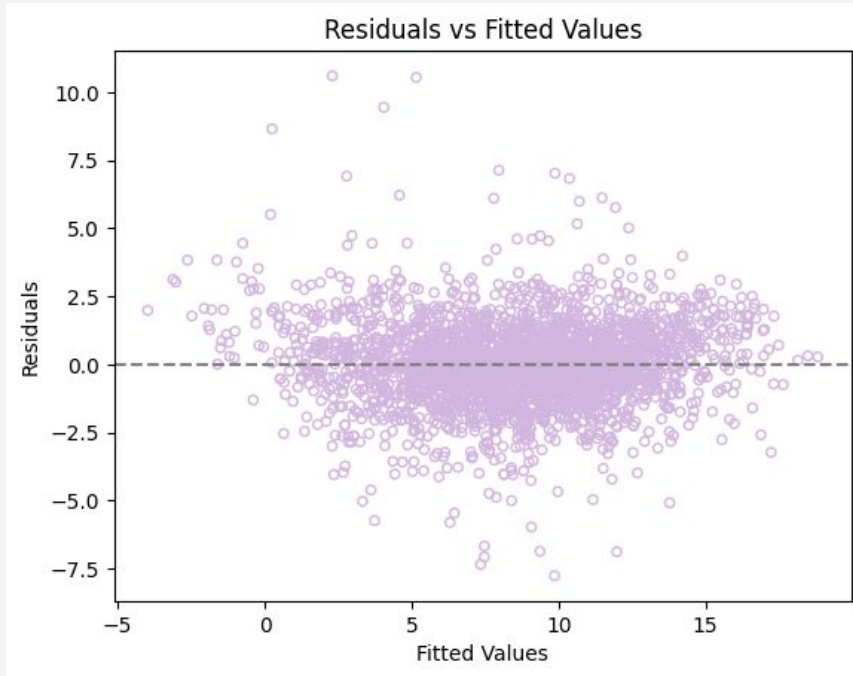
*Residual has constant variance.*

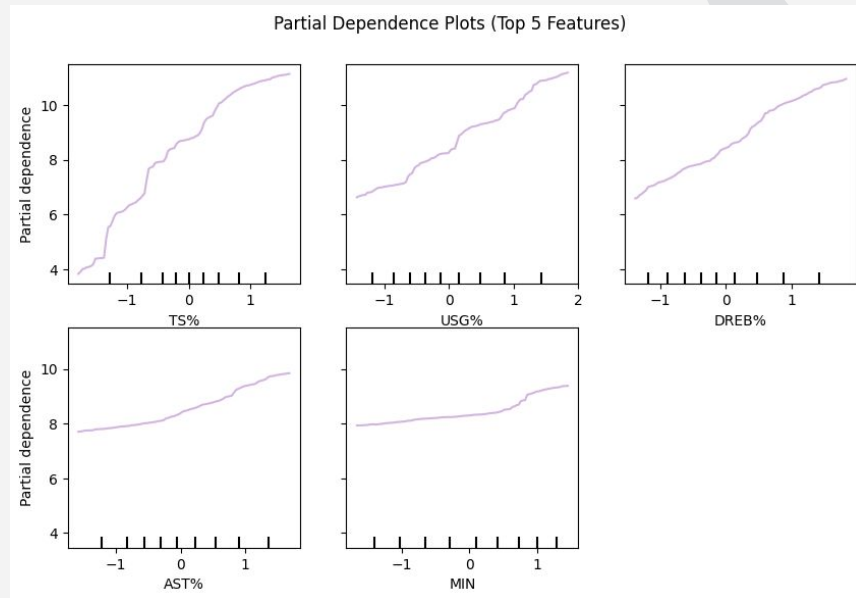| Feature | VIF |
|---------|--------|
| const | 680.59 |
| AGE | 1.04 |
| MIN | 1.56 |
| AST% | 1.53 |
| DEFRTG | 1.09 |
| OFFRTG | 1.37 |
| DREB% | 1.36 |
| OREB% | 1.46 |
| TS% | 1.24 |
| TO RATIO | 1.09 |
| USG% | 1.49 |
| PACE | 1.16 |

*No alarming multicollinearity violation.*

# 03

# Alternative Model 1

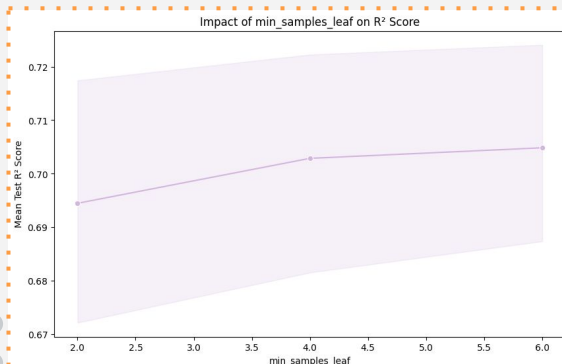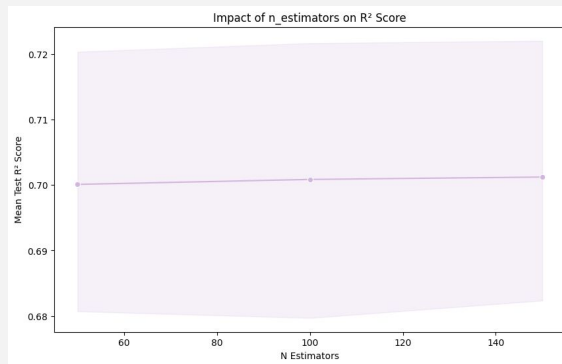# Random Forest Model Summary



Top 10 Important Features (Without PCA)

*True shooting contributes the most to the model's predictive power.*



Partial Dependence Plots (Top 5 Features)

*Key player metrics significantly contribute to higher player performance.*
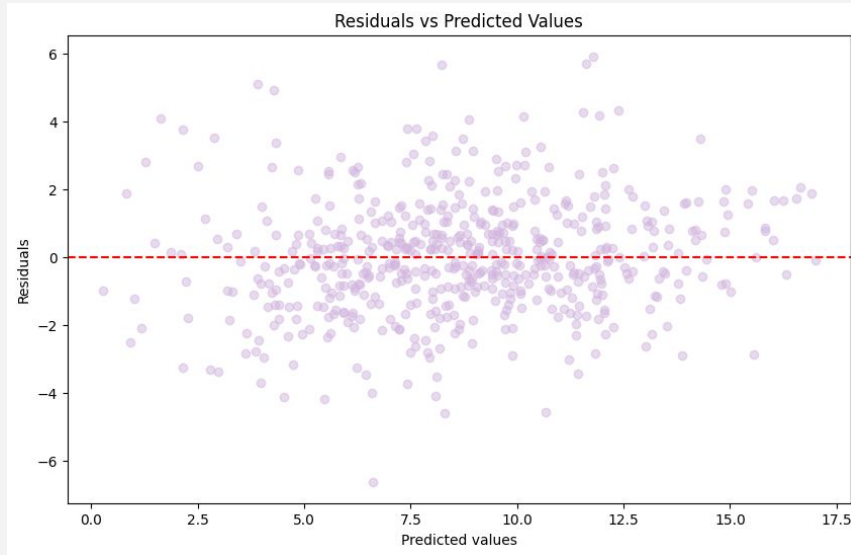
# Hyperparameter Tuning
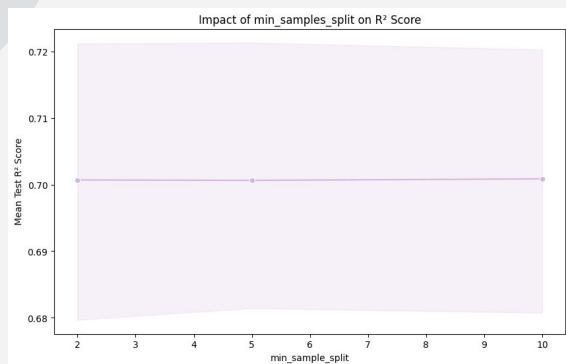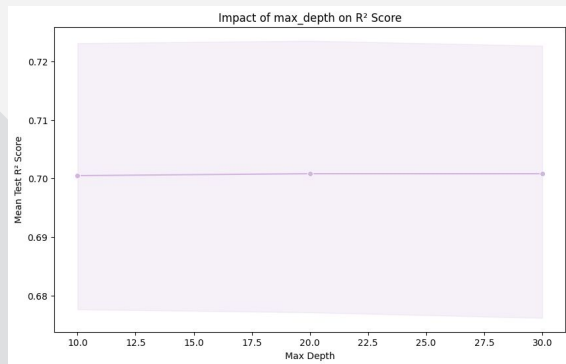


**Best Model**

- Bootstrap with 0 max depth, 2 leaf node, 2 minimal sample split and 200 number of trees in forest
- R² Scores: 80%
- Mean Squared Error: 2.8

# Model Diagnostics



Residuals vs Predicted Values
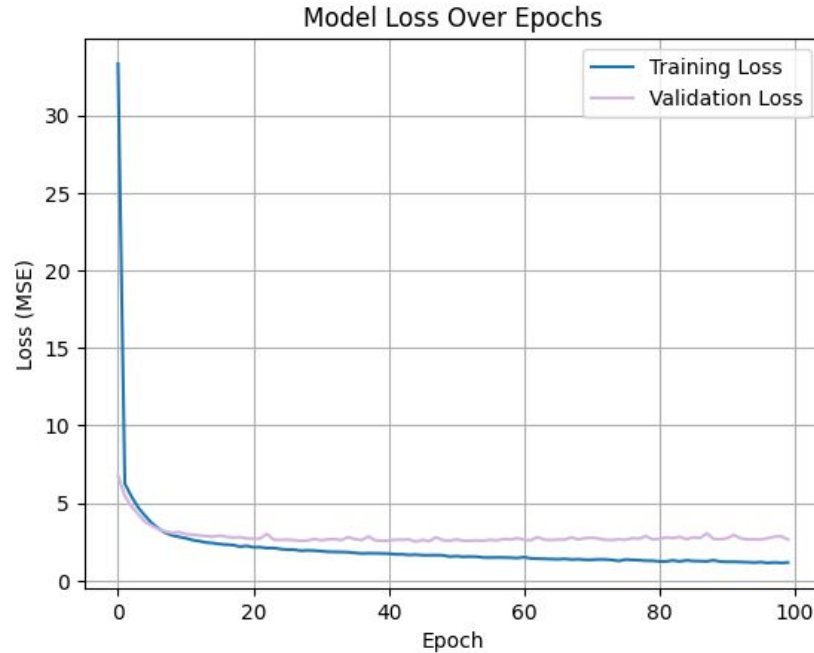
*Residual has constant variance.*

**5 Fold Cross Validation**

- R² Scores range from 77% to 84%
- Mean R²: 80%
- Standard deviation of R²: 0.028

# 03

# Alternative
# Model 2

# Neural Network Model



Model Loss Over Epochs

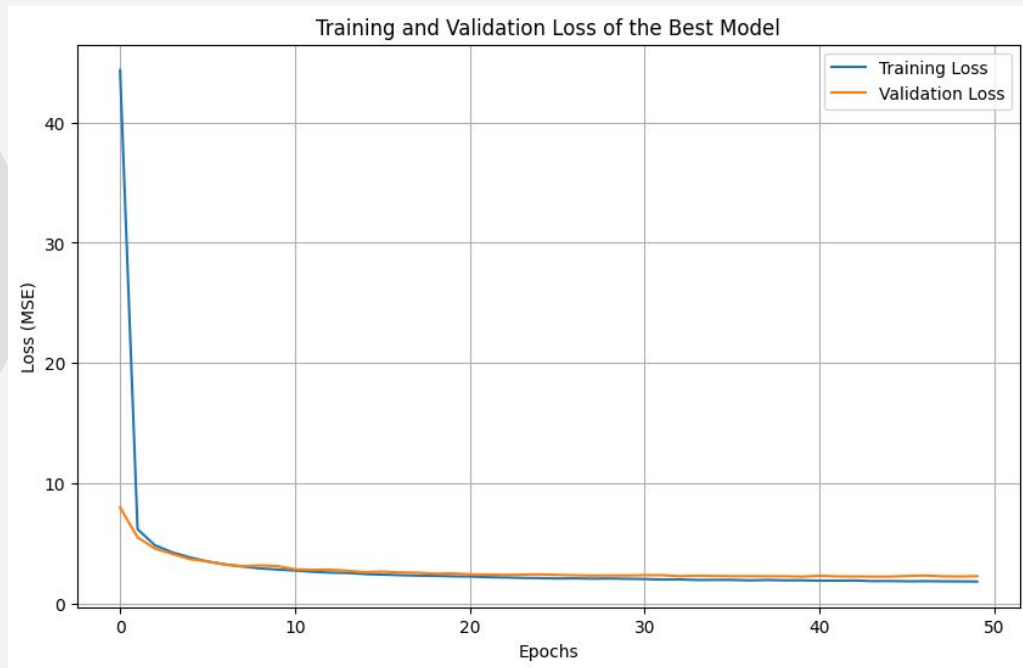## Architecture

- 1 input layer: 128 neuros, L2 reg and relu activation
- 2 hidden layers: 64 & 32 neurons with L2 reg
- 1 output layer: 1 neuron and linear activation

## Training and Evaluation

- 100 epochs & 32 batch size
- Mean Squared Error: 2.2
- R² Score: .84.7%

# Hyperparameter Tuning



Training and Validation Loss of the Best Model

**Best Model**

- R² Scores: 86.4%
- MSE: 1.96

# Conclusions

| Models | Pro | Con |
|---|---|---|
| Linear Regression | - High $R^2$<br>- Interpretable | - May not capture non-linear relationships as effectively<br>- Sensitive to multicollinearity |
| Random Forest | - Interpretable<br>- Robust to multicollinearity<br>- Capture complex, non-linear relationships | - Lowest $R^2$ |
| Neural Network | - Highest $R^2$<br>- Capture complex, non-linear relationships<br>- Robust to multicollinearity | - Challenge to interpret |

# Thank you!

# Thanks!

Do you have any questions?

youremail@freepik.com

+91 620 421 838

yourcompany.com

# Desktop app

You can replace the image on the screen with your own work. Just right-click on it and select "Replace image"

# Icon pack

# Alternative resources

Here's an assortment of alternative resources whose style fits the one of this template

## Vectors

- Flat design basketball twitch banner with hoop
- Flat-hand drawn curly hair types collection
- Flat design basketball twitch background with balls