Basketball Analysis

Samantha Waters

FINAL PROJECT – CPSC 392

Overview

Dataset: 2020 – 2021 NBA Player Stats (Totals)

Predictive Models:

- Logistic Regression – predict whether a player plays in Center position
- Linear Regression – predict total number of points scored by a player

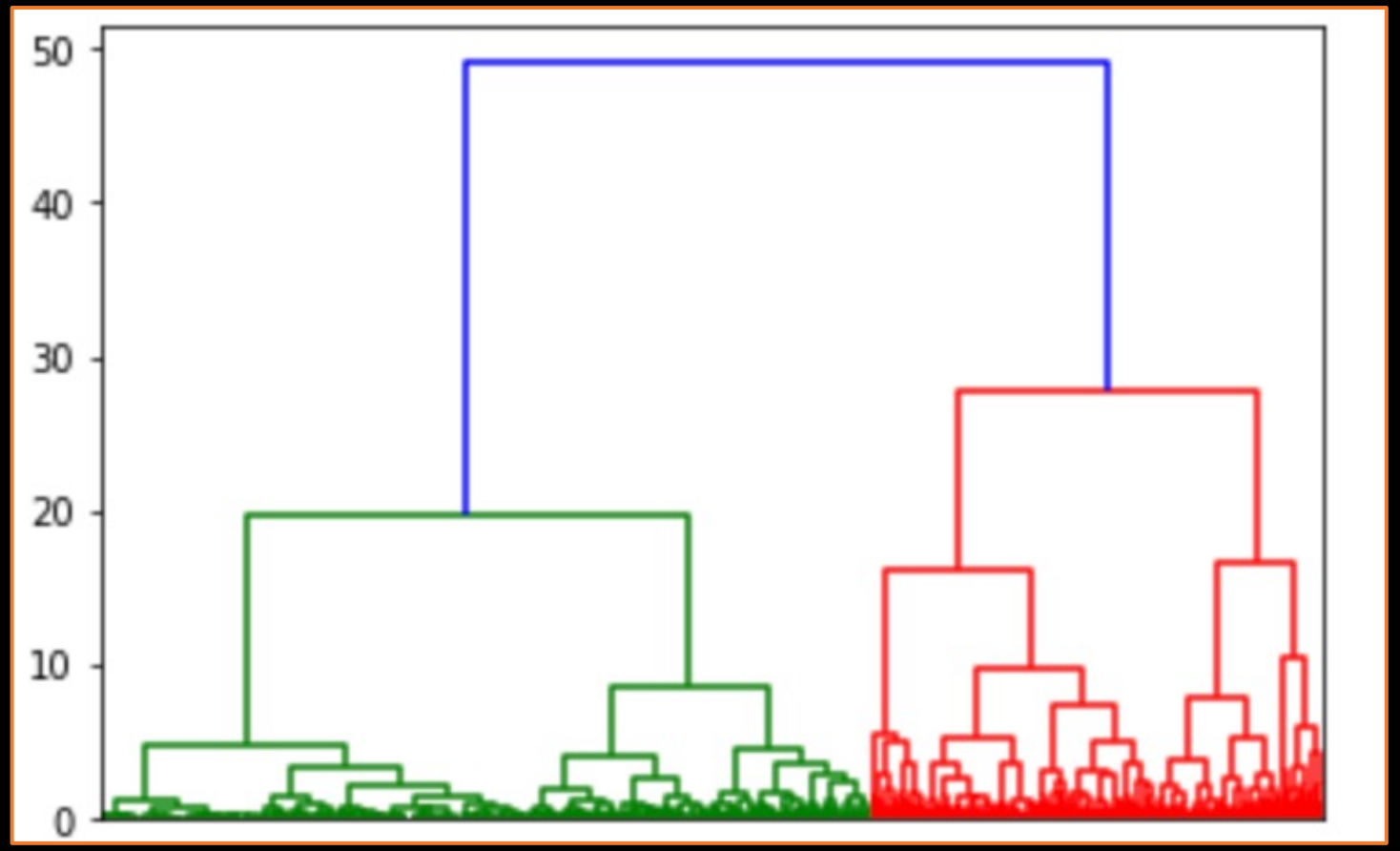
Clustering Method:

- Hierarchical Agglomerative Clustering – discovering distinct groups (clusters) in data

Variables

- **Player** - player name
- **Pos** – position
- **Age** (in years) – player's age on February 1 of the season
- **Tm** - team
- **G** - games
- **GS** – games started
- **MP** (in minutes) – minutes played per game
- **FG** – field goals
- **FGA** – field goal attempts
- **FG%** - field goal percentage
- **3P** – 3-point field goals
- **3PA** – 3-point field goal attempts
- **3P%** - 3-point field goal percentage
- **2P** – 2-point field goals
- **2PA** – 2-point field goal attempts
- **2P%** - 2-point field goal percentage
- **eFG%** - Effective Field Goal Percentage
- **FT** – free throws
- **FTA** – free throw attempts
- **FT%** - free throw percentage
- **ORB** – offensive rebounds
- **DRB** – defensive rebounds
- **TRB** – total rebounds
- **AST** – assists
- **STL** – steals
- **BLK** – blocks
- **TOV** – turnovers
- **PF** – personal fouls
- **PTS** – points

When considering points, blocks, minutes played, and 3-points scored, what types of clusters emerge and what characterizes these clusters?

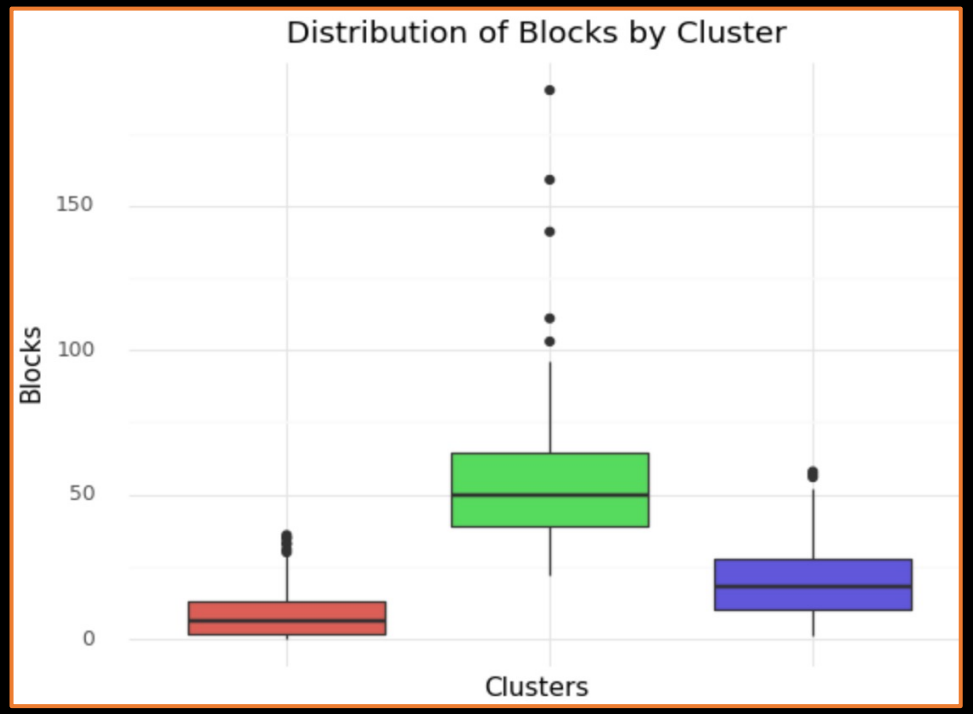


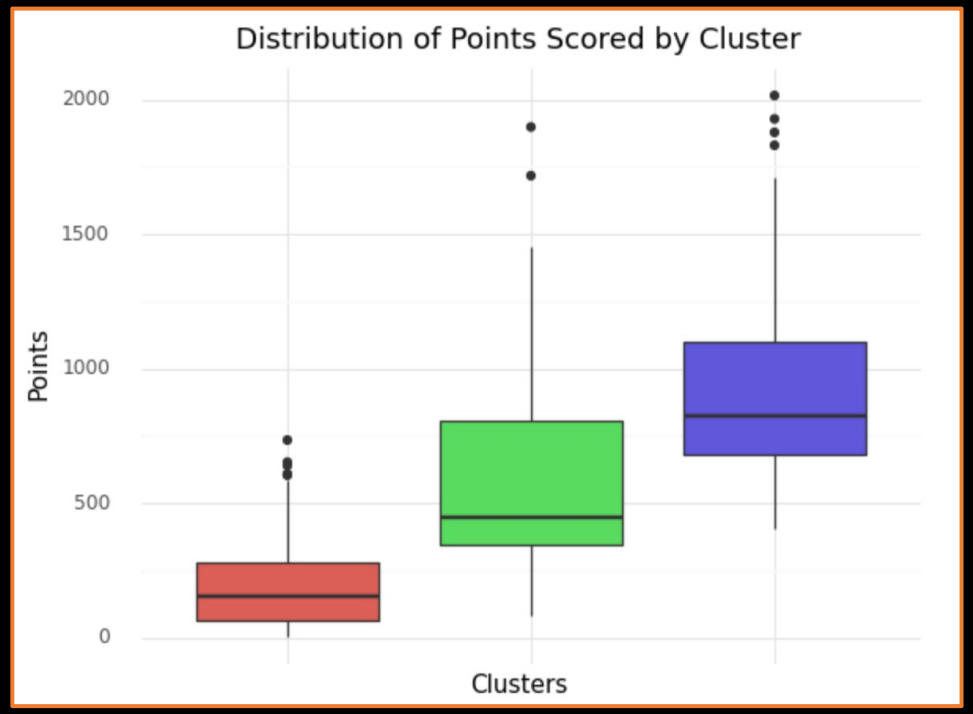
Silhouette Score:

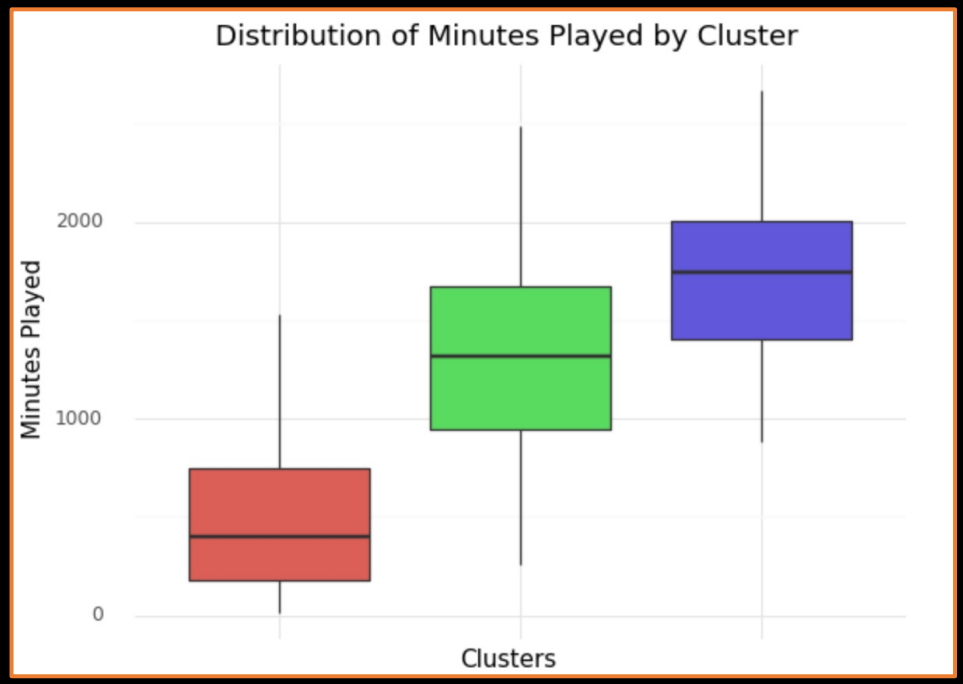
~ 0.495

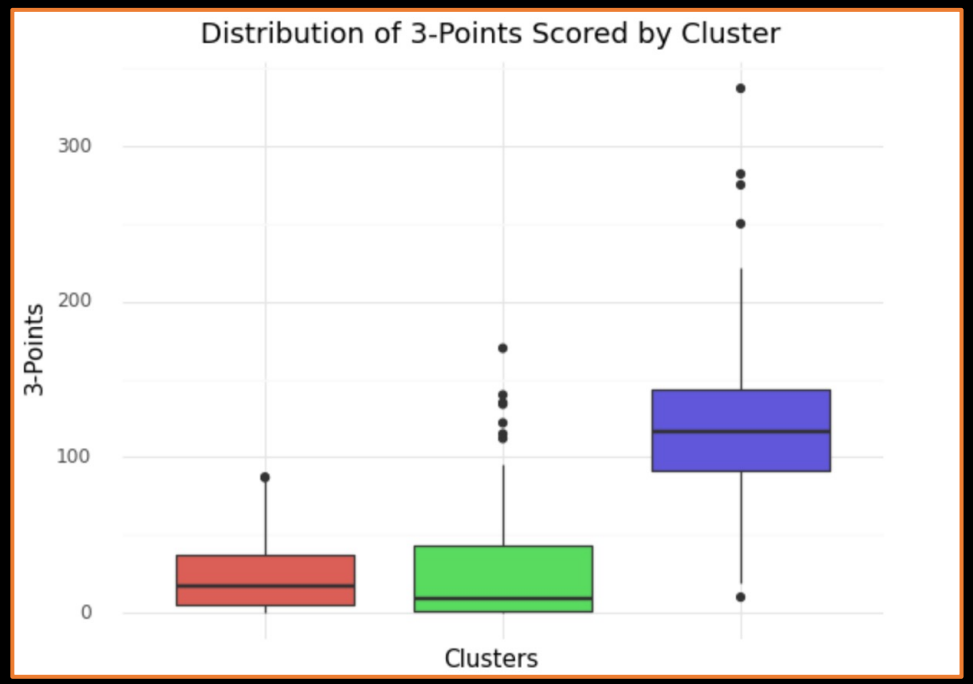
<---------- Dendrogram

Clustering Method Results (HAC)







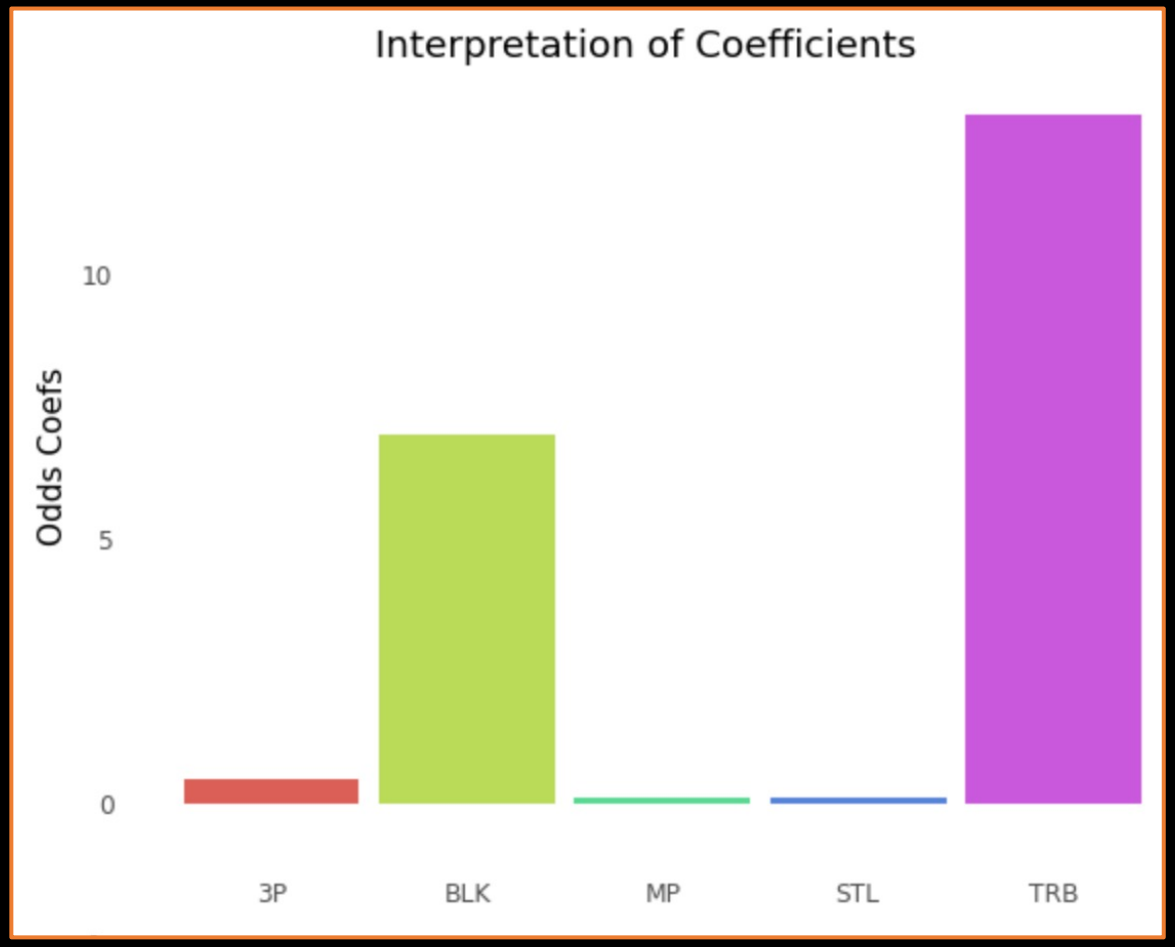


How well does a model perform when classifying whether someone is a Center based on blocks, rebounds, steals, 3-point shots, and minutes played?

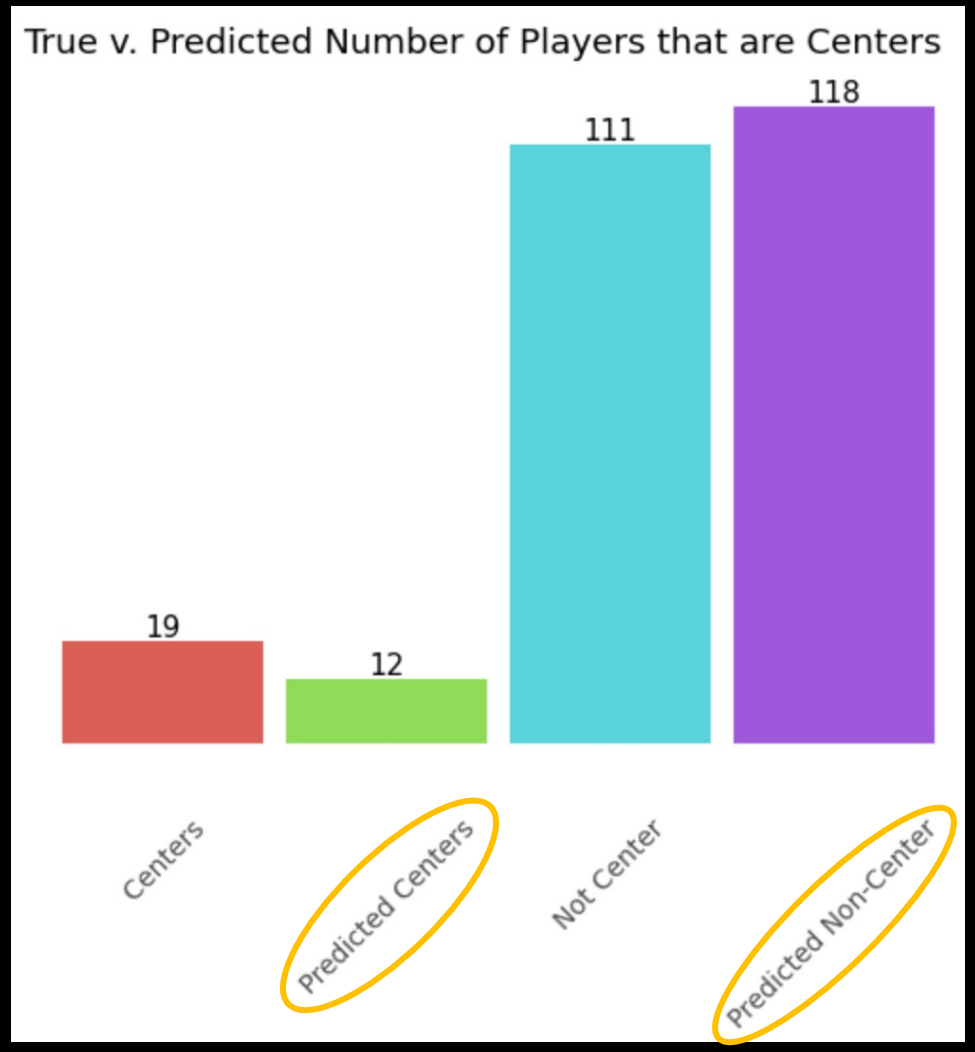
- Logistic Regression Model
- Top Coefficients:

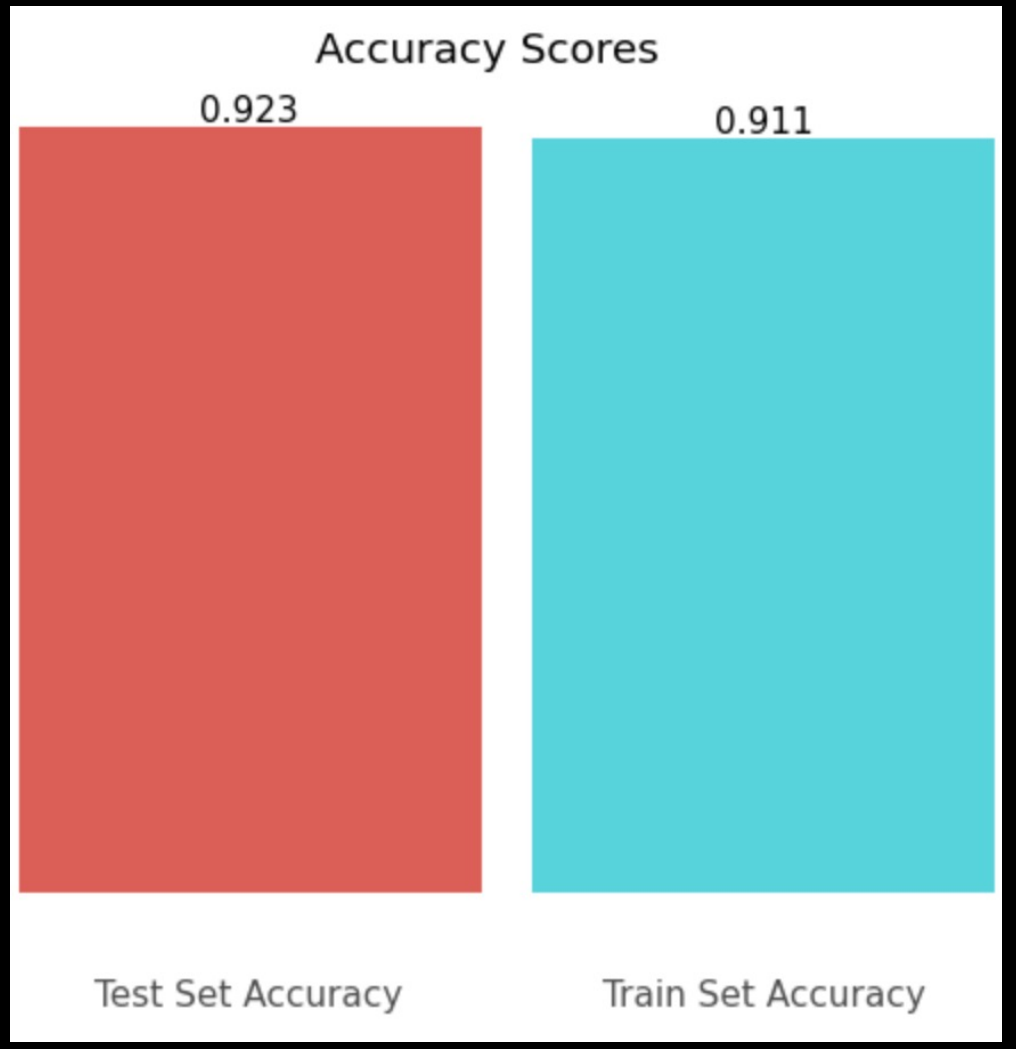
  (in terms of odds)

  1. Total Rebounds – 13.05x
  2. Blocks – 7.02x
  3. 3-points – 0.477x



Logistic Regression Accuracy Results





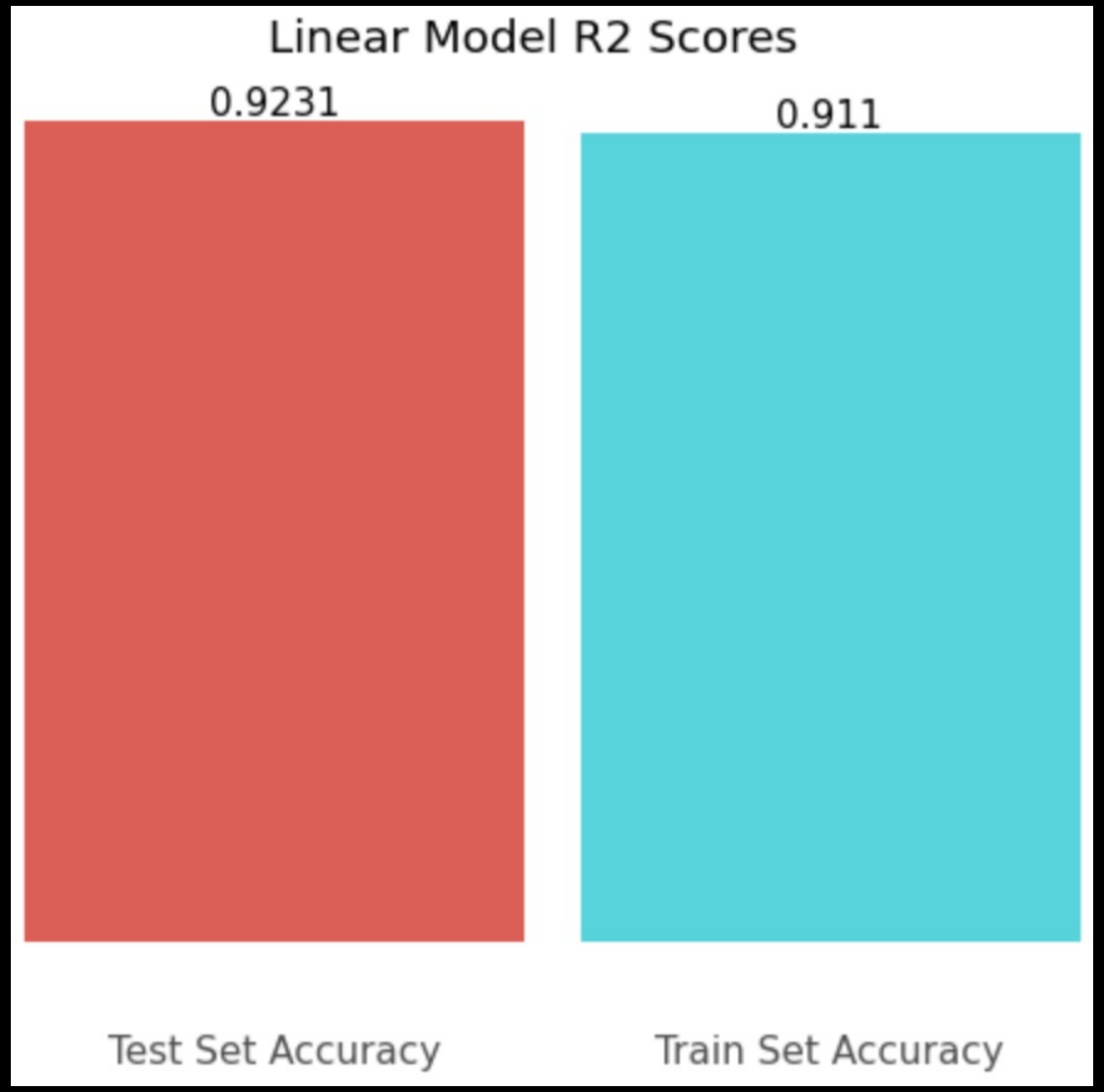
When comparing a model using LASSO to a model not using LASSO to predict a player's total number of points scored, how does each model perform, and which model would you choose?

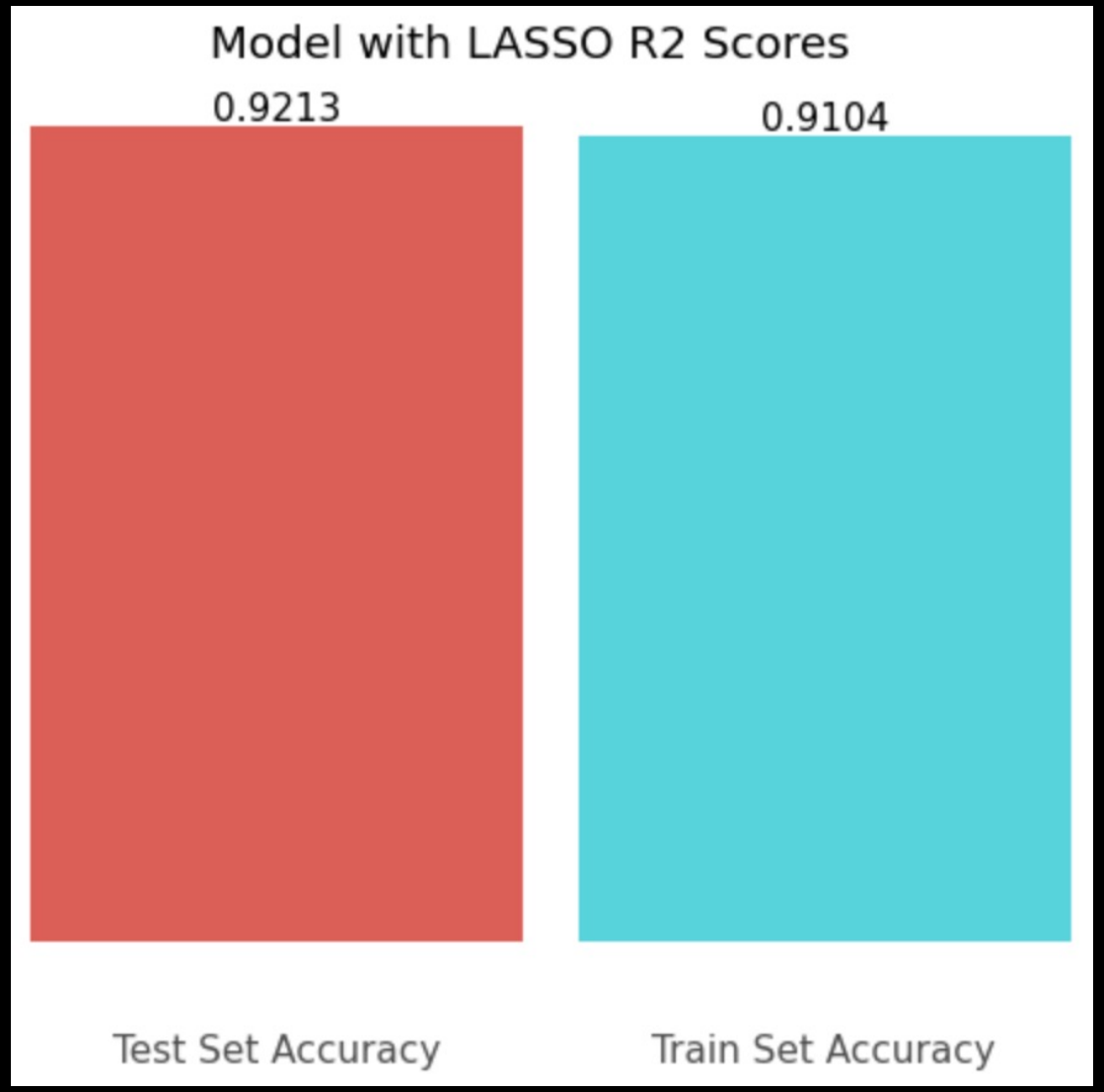
- Linear Regression Model
- LASSO – Regularization
- R2 - % of the variation explained by our model

Variables Involved:

- *Age* – Years of Age
- *G* - games played
- *GS* – games started
- *MP* – minutes played
- FG% - field goal percentage
- eFG% - effective field goal percentage
- *ORB* – offensive rebound
- *DRB* – defensive rebounds
- *TRB* – total rebounds
- *AST* - assists
- *STL* - steals
- *BLK* - blocks
- *TOV* - turnovers
- *PF* – personal fouls

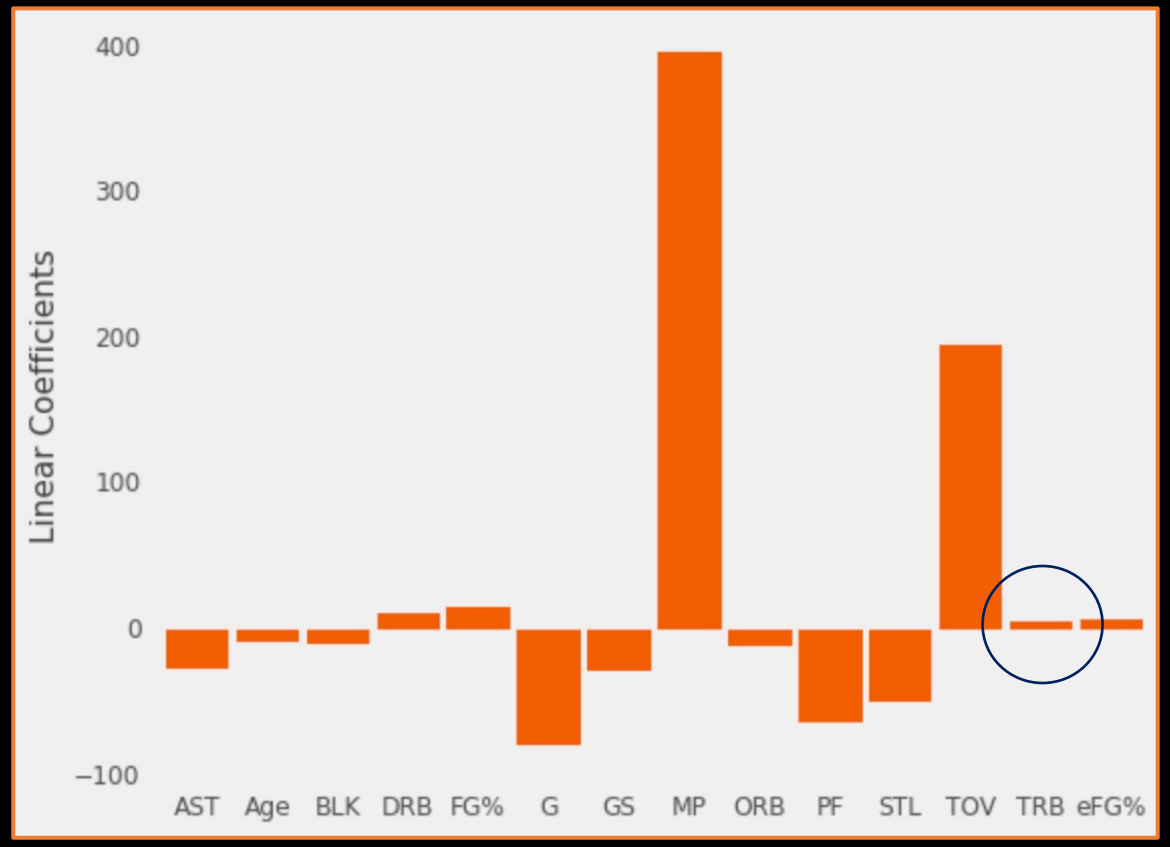
Performance of Both Models



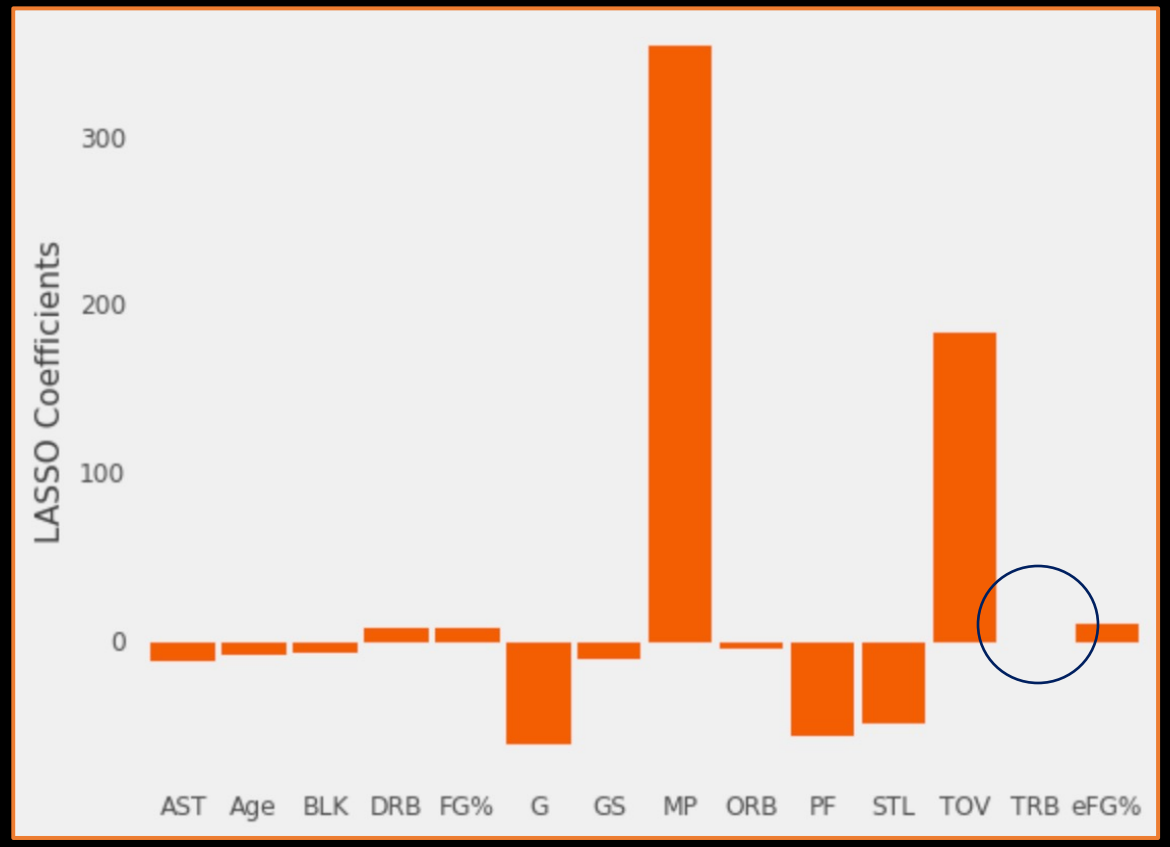


Interpretation of Coefficients

Linear Model



Linear Model using LASSO



Thank You!