Modeling the NBA Leap

Ryan Lewis

Table of Contents

What is “The Leap”?

“A player’s identity typically begins to crystallize in his third or fourth NBA season. Young players have learned the ropes, and veterans have departed or aged, vacating heavy-duty roles that need filling. Everyone involved — players, agents, executives — looks to see what emerges as a player nears the expiration of his rookie contract.”
- Zach Lowe, NBA Analyst

Business Question

Based off an NBA players first three seasons, can you predict if they will make '**The Leap**' to become an All-NBA player in seasons four through six?

Business Question Importance

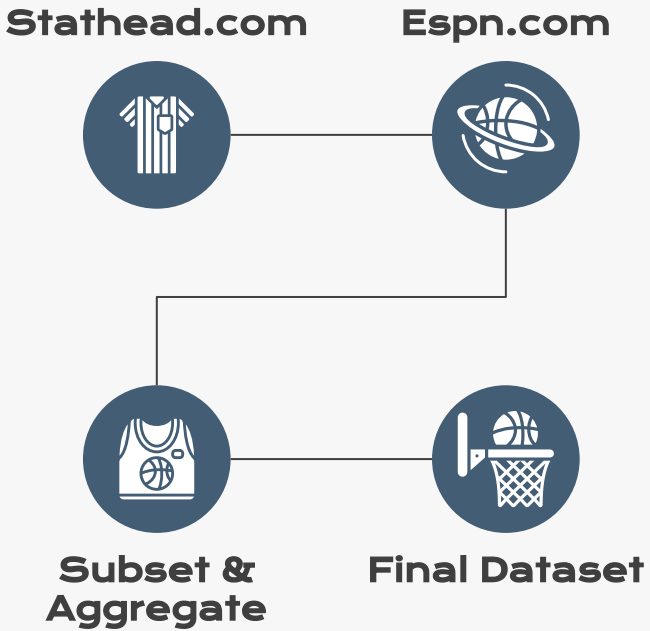
Importance

Front office financial planning: NBA players drafted in the first round can command contract extensions of up to 25-30% of a team's salary cap in 2020-21 season

Focus

In order to **minimize false positives**, our models will focus on the 'precision' metric -- this puts a higher emphasis on front offices correctly identifying actual true positives

Dataset



Stathead.com

Using Selenium, web scraped all seasonal and advanced player statistics dating back to 1947

ESPN.com

Merged in all player & team awards pulled from ESPN.com

Subset & Aggregate

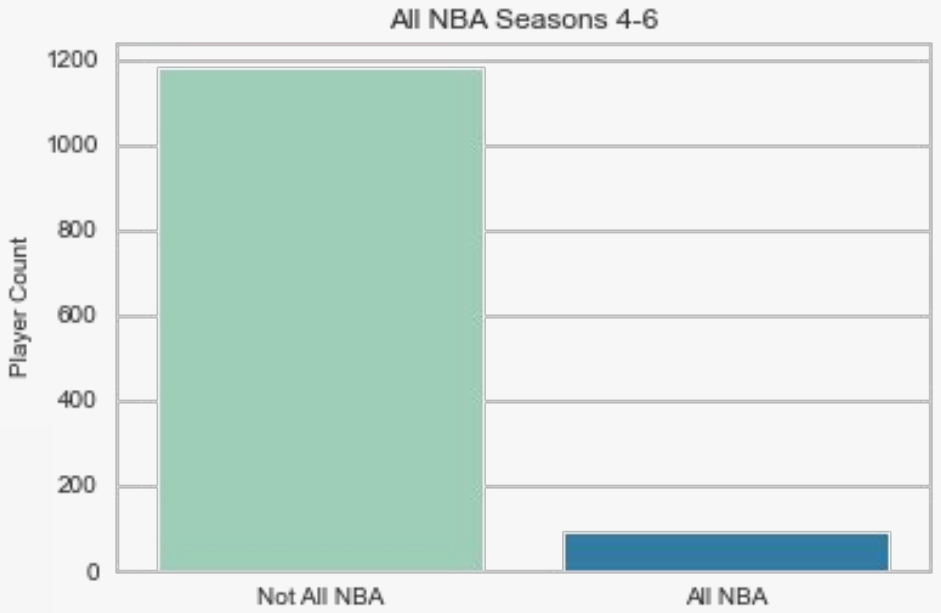
- Include only seasons after 1977
- Players who have played at least 6 seasons
- Aggregated so each row represented one player

Final Dataset

- 1273 qualified players
- 221 features; total & seasonal statistics

Exploratory Data Analysis

7%
Reach All-NBA in seasons 4-6



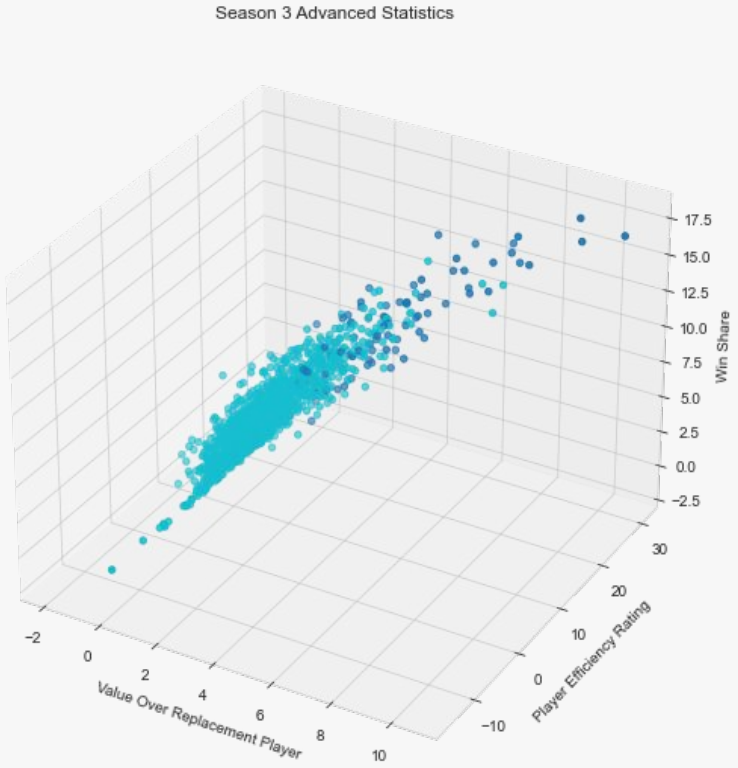
EDA - Basic Totals Pairplot

Statistic total from players first three seasons in the league

All-NBA players separate themselves from their peers statistically

EDA - Season 3 Advanced Statistics

- Season three is when players start to differentiate themselves
- Advanced statistics such as **'Win-Share'**, '**Player Efficiency Rating**' and '**Value Over Replacement Player**' are key identifiers of future stars



Modeling Techniques - RFE

After Feature Engineering

- 200+ columns of player data

Feature Selection

- Correlation Analysis
- Random Forest Feature Importance
- **Recursive Feature Elimination**

After RFE

- **VORP_3**
- **PER_3**
- **PTS**
- VORP_2
- PER_2
- WS_2
- Rookie_Year

Modeling

Interpretability

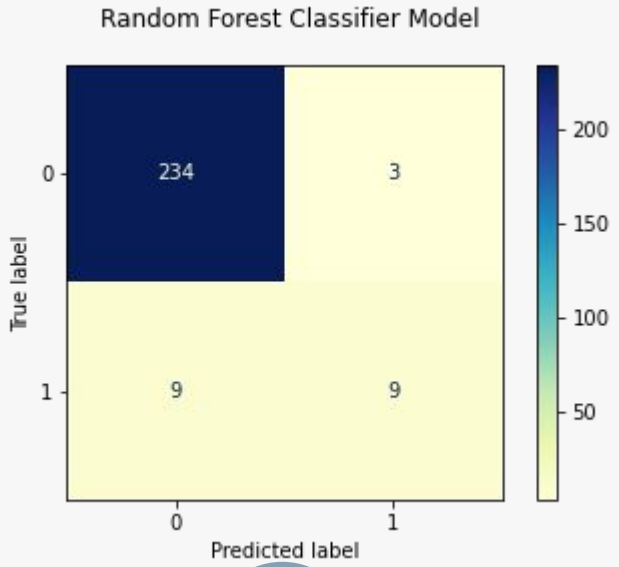
Only focused on models that are easily interpretable

Precision

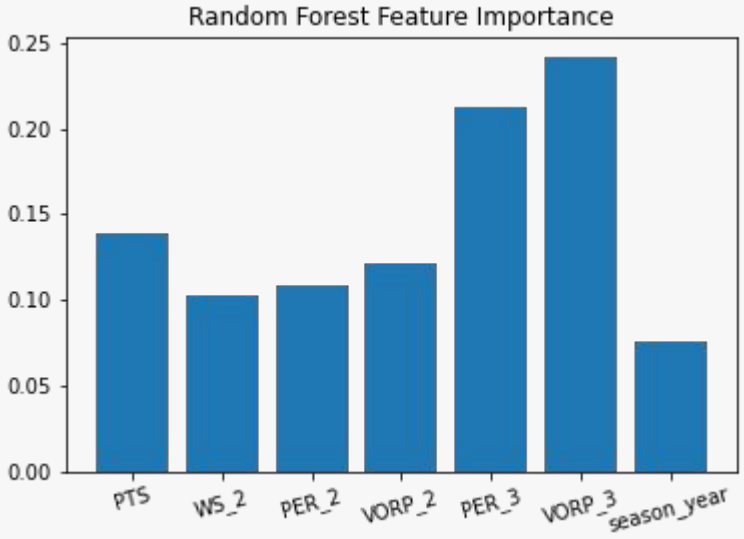
Most important metric was precision for our business question

Model/Metric	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.945098	0.555556	0.625000	0.588235
LR w/ resampling	0.894118	0.777778	0.378378	0.509091
Random Forest	0.952941	0.500000	0.750000	0.600000
RF w/ resampling	0.921569	0.722222	0.464286	0.565217
Decision Tree	0.898039	0.388889	0.318182	0.350000
DT w/ resampling	0.913725	0.611111	0.423077	0.500000

Modeling Results



Confusion Matrix



Feature Importance

Next Steps

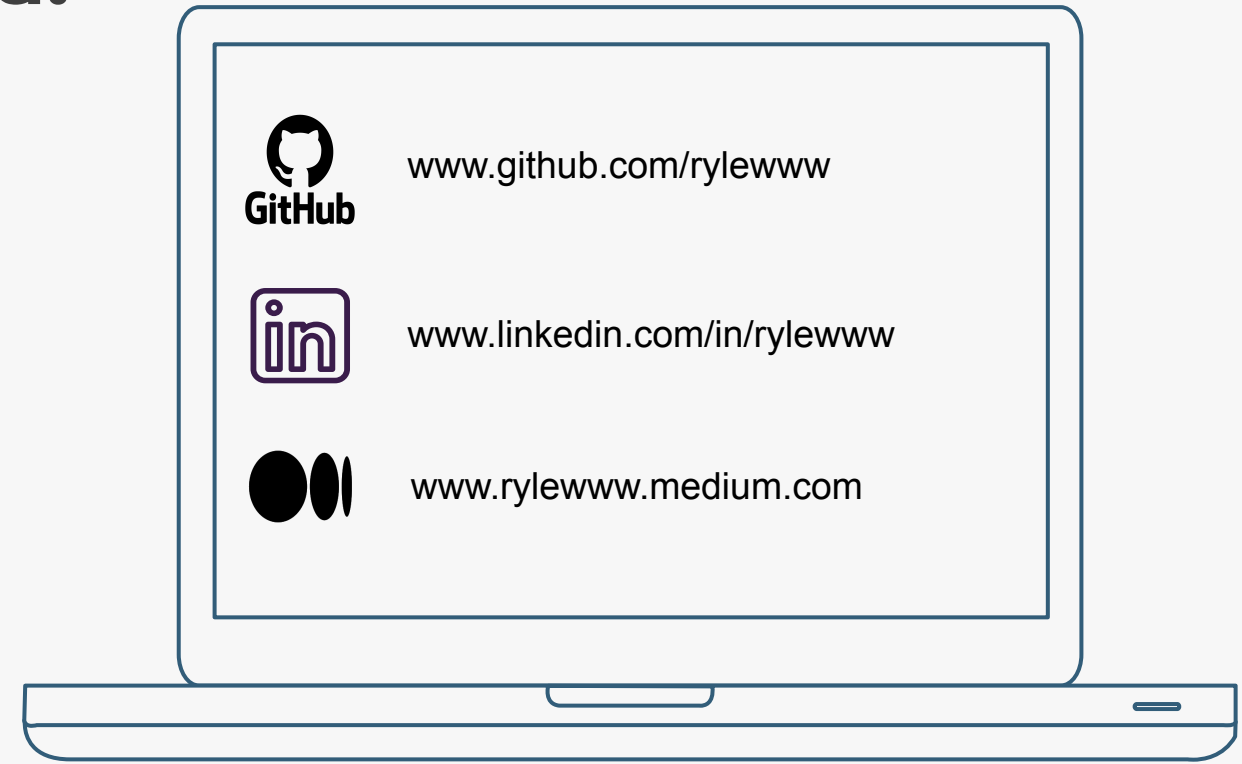
Implement other resampling techniques to deal with class imbalances

Pull in additional categorical data such as draft pick position or team success

Build out the interpretability of these models

Look into generational trends, did All NBA players look statistically different in the 80's vs the 90's

Thank you!


GitHub
www.github.com/rylewww
www.linkedin.com/in/rylewww
www.rylewww.medium.com