THE WORLD IS NOT LINEAR?

Benchmarking OLS with non-linear models:
A "supervised" investigation on players' compensation exploiting basketball data

20630 - Introduction to Sport Analytics
Team 2, Final Presentation

OUR TEAM

Federico Leonardi

Alberto Allegri

HC:

Beatrice Guidotti

Leonardo Yang

Jakob Schlierf

Tiziano Paci

TABLE OF CONTENTS

01 Recap

OUR GOAL

$$y = f(x) + \varepsilon$$

Salary

Model

Performance

Market dynamics, NBA rules etc.

Let's try to reduce the noise as much as possible.

ADDRESSING THE SALARY CAP

PROBLEMS	SOLUTIONS
Salary Cap Growth	Take salaries expressed as percentage of salary cap each year as our target variable
Soft Cap/Bird Rights	Add a dummy variable for contracts signed using bird rights
Rookie/Maximum Contracts	Add a dummy variable for rookie/maximum contracts

MAX / SUPERMAX CONTRACTS

Look-Ahead-Bias: Using Information that would not be available at the time of analysis
⇒ Could undermine strength of our model (model leakage), but we should also consider:

Proxy for Team Impact

Accounting for:

- Star Players have large impact on team
- Veteran / locker room presence

Eligibility Requirements

Accounting for:

- Seniority
- Individual accomplishments (e.g. MVP, All-NBA)
- Past performance

Proxy for Monetary Impact

Accounting for:

- Personal brand of player
- Marketing impact on team
- Monopsony Rent

Despite the risk, we decided to keep the Dummy Variables for Max & Supermax Contracts.

DATA COLLECTION

Source

BASKETBALL REFERENCE

spotrac

Statistics

Per-Game &
Advanced statistics

Target

Every season statistics and **Contract Type** for **every player** that started playing in Season 1999-00 (up to season 2020-21)

Some Numbers

61 Variables
6870 Rows
1202 Players

OUR VARIABLES (1/3)

PER-GAME

1. **Season** – Year the season ended
2. **Age** -- Player's age on February 1st.
3. **Tm** -- Team
4. **Lg** -- League
5. **Pos** -- Position
6. **G** -- Games
7. **GS** -- Games Started
8. **MP** -- Minutes Played Per Game
9. **FG** -- Field Goals Per Game
10. **FGA** -- Field Goal Attempts Per Game
11. **FG%** -- Field Goal Percentage
12. **3P** -- 3-Point Field Goals Per Game
13. **3PA** -- 3-Point Field Goal Attempts Per Game
14. **3P%** -- 3-Point Field Goal Percentage
15. **2P** -- 2-Point Field Goals Per Game
16. **2PA** -- 2-Point Field Goal Attempts Per Game
17. **2P%** -- 2-Point Field Goal Percentage
18. **eFG%** -- Effective Field Goal Percentage
19. **FT** -- Free Throws Per Game
20. **FTA** -- Free Throw Attempts Per Game
21. **FT%** -- Free Throw Percentage
22. **ORB** -- Offensive Rebounds Per Game
23. **DRB** -- Defensive Rebounds Per Game
24. **TRB** -- Total Rebounds Per Game
25. **AST** -- Assists Per Game
26. **STL** -- Steals Per Game
27. **BLK** -- Blocks Per Game
28. **TOV** -- Turnovers Per Game
29. **PF** -- Personal Fouls Per Game
30. **PTS** -- Points Per Game

OUR VARIABLES (2/3)

ADVANCED

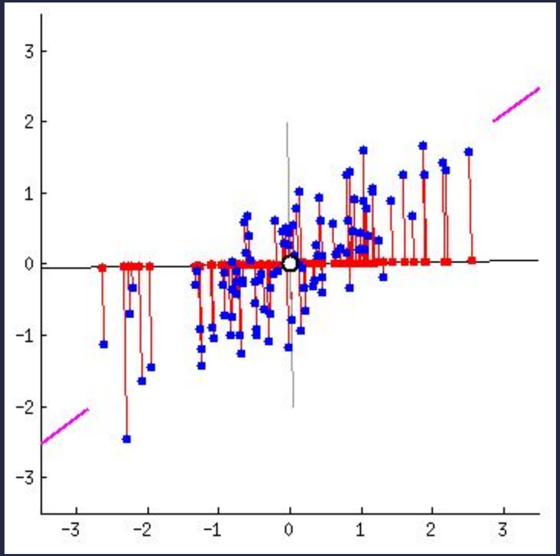
1. PER -- Player Efficiency Rating
2. TS% -- True Shooting Percentage
3. 3PAr -- 3-Point Attempt Rate
4. FTr -- Free Throw Attempt Rate
5. ORB% -- Offensive Rebound Percentage
6. DRB% -- Defensive Rebound Percentage
7. TRB% -- Total Rebound Percentage
8. AST% -- Assist Percentage
9. STL% -- Steal Percentage
10. BLK% -- Block Percentage
11. TOV% -- Turnover Percentage
12. USG% -- Usage Percentage
13. OWS -- Offensive Win Shares
14. DWS -- Defensive Win Shares
15. WS -- Win Shares
16. WS/48 -- Win Shares Per 48 Minutes
17. OBPM -- Offensive Box Plus/Minus
18. DBPM -- Defensive Box Plus/Minus
19. BPM -- Box Plus/Minus
20. VORP -- Value Over Replacement Player

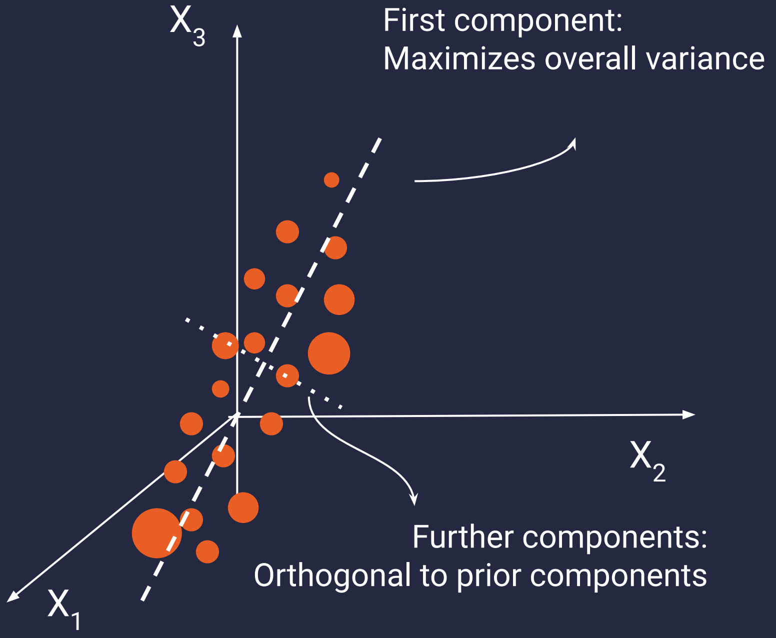
OUR VARIABLES (3/3)

CONTRACTS & OTHERS

1. PLAYER -- Player's Name
2. IMAGE_LINK -- Link of BR image
3. US_PLAYER -- Dummy for US Players
4. SALARY -- Player's Salary
5. SALARY_CAP -- Cap of that season
6. SALARY_CAP_% -- Player's Cap %
7. CONTRACT_TYPE -- Player's CT
8. ROOKIE_CONTRACT -- Dummy for Rookies
9. BIRD_RIGHTS -- Dummy for Bird Contracts
10. MAXIMUM_CONTRACT -- Dummy for Max Contracts
11. SUPER_MAX_CONTRACT -- Dummy for Super-Max Contracts

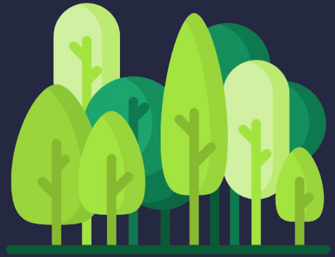
DIMENSIONALITY REDUCTION: PCA





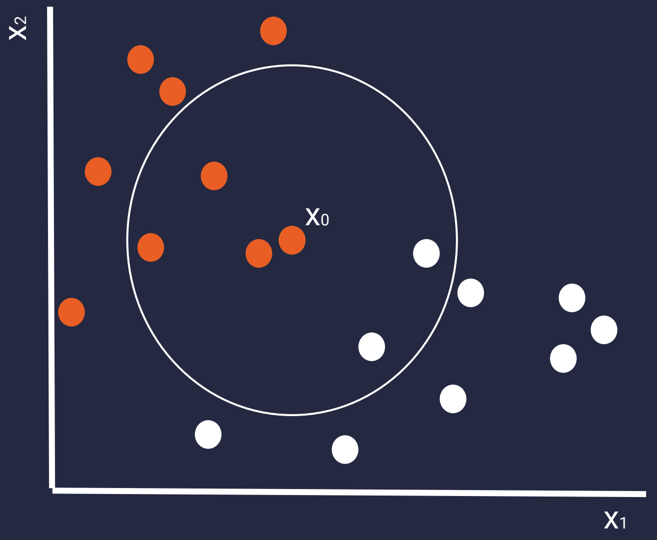
OUR MODELS

K-Nearest Neighbors



Random Forest

K-NEAREST NEIGHBORS



K = 5

- 2 white neighbors, i.e. $P(white|x_0) = 0.4$
- 3 orange neighbors, i.e. $P(orange|x_0) = 0.6$

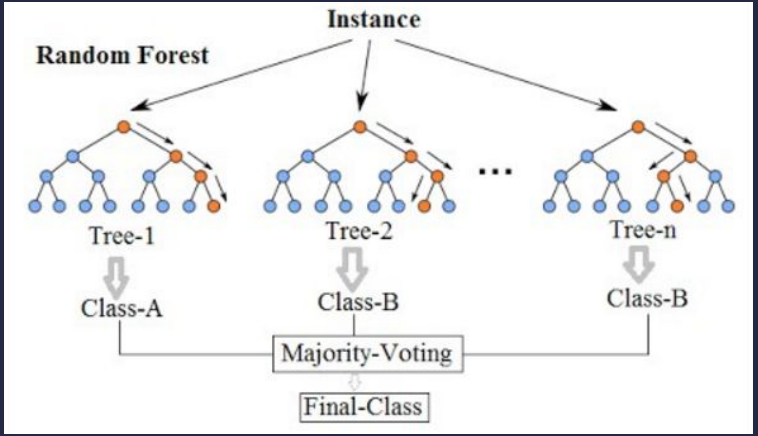
MAJORITY VOTE

↓

ORANGE CLASS

The value of a data point is determined by the data points around it.

RANDOM FOREST



1. **Bootstrap:** Pick **at random and with replacement** n data points from the original dataset
2. **Training:** Build the decision tree associates with the newly constructed dataset
3. **Build a Forest:** Repeat steps 1 and 2 B times, with B being the number of trees in the forest
4. **Ensemble:** Predict the target of a new data point by combining the different predictions coming from the B trees

GRADIENT BOOSTING

- Random Forest lowers variance by randomly changing inputs to trees
- What if instead of randomly, we focus on areas where we underperform?
- This is the idea of **Boosting**

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$
2. For $b = 1, 2 \ldots, B$ repeat:
   a. Fit tree to residual-weighted training data
   b. Update $\hat{f}$ with new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \hat{\lambda}^{b}(x)$$
   c. Update residuals
3. Output boosted model:
$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(x)$$

<u>**Three tuning parameters:**</u>

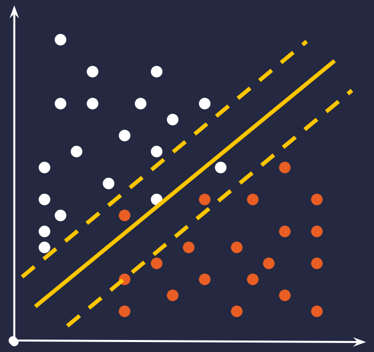
- B - Number of trees (overfitting possible)
- λ - shrinkage parameter (controls learning rate)
- d - Number of splits in each tree

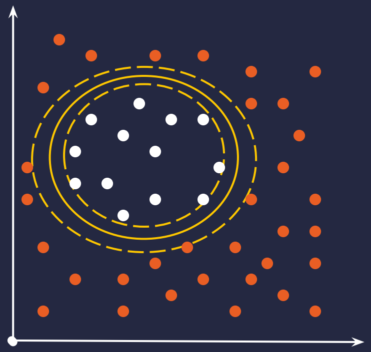
Main Idea

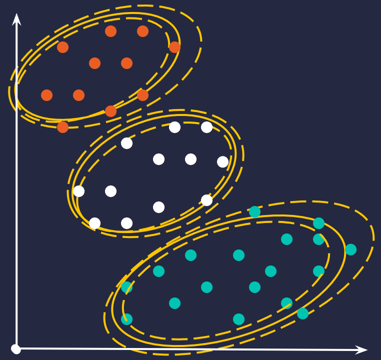
Details

Implementation

SUPPORT VECTOR MACHINES







Non-Perfect Separation

- Allow for violations of border through parameter C

Non-Linear Clusters

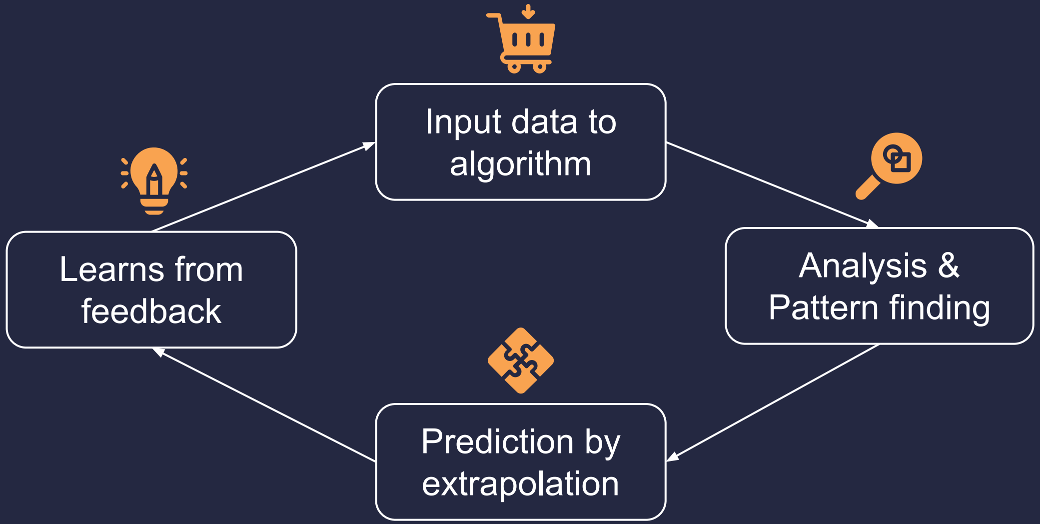
- Using Kernels, we can achieve non-linearity

Multiple Clusters

- We can use OVA & OVO to separate multiple clusters
  - OVA: One vs. All
  - OVO: One vs. One

TRIAL & ERROR: WHY?

STEPS INVOLVED IN MACHINE LEARNING



Failure is essential.

The rationale behind ML is exactly based on allowing the model to **understand patterns** and then trying to adapt for the subsequent step in order to **minimize a certain error**.

SEPARATION OF WORK

Sportylytics

Dataset

OLS

KNN

Random Forest

Gradient Boosting

SVM

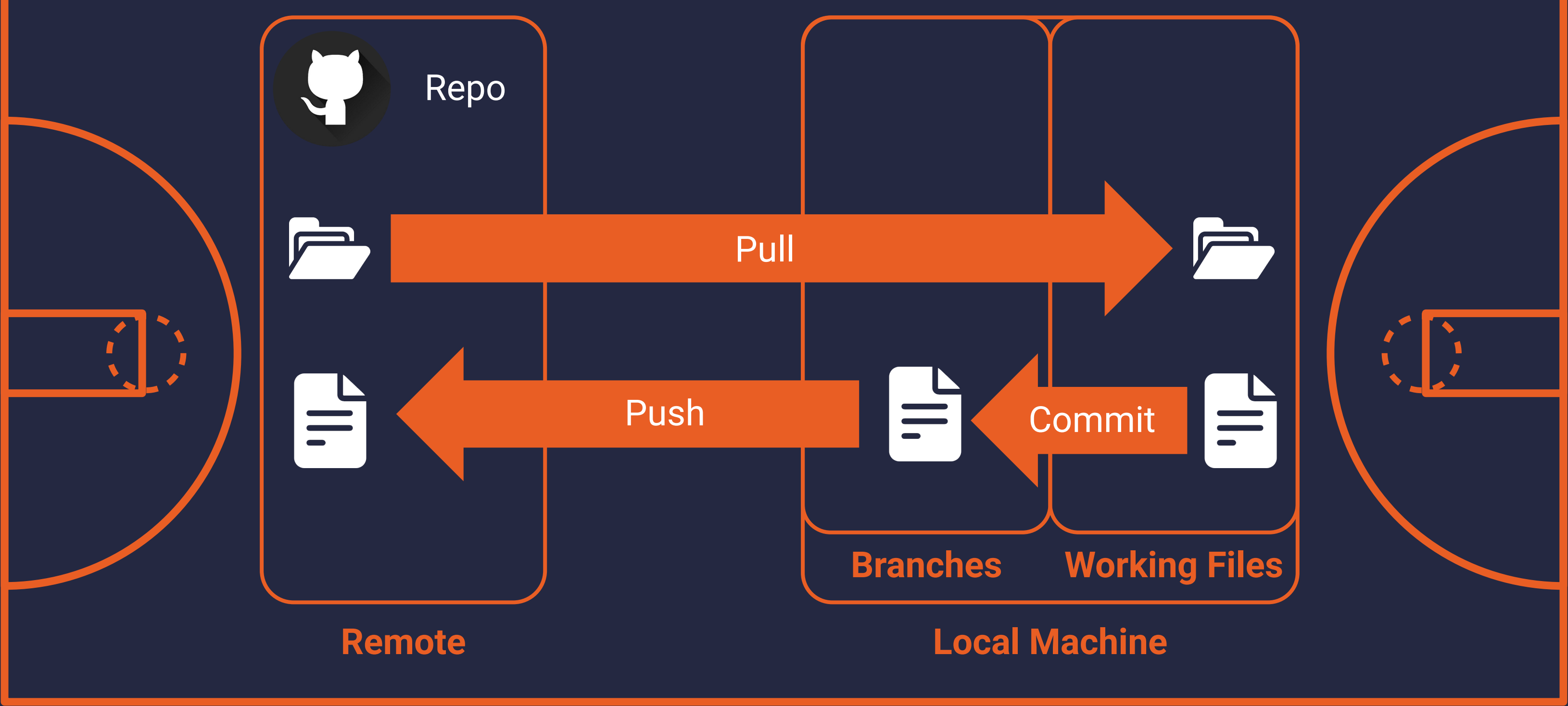
Numerous workstreams, split among team members

BUT

All workstreams need to rely on the correct, current dataset

⇒ Syncing data across team is vital for comparability of model performance metrics

INTRODUCTION TO GITHUB



02

DATA SCRAPING

WEB SCRAPING 1/4

BBR PACKAGE

To scrape each player **Basketball Reference** “slug”

(source: https://github.com/mbjoseph/bbr)

player	slug
Álex Abrines	abrinal01
Precious Achiuwa	achiupr01
Alex Acker	ackeral01
Quincy Acy	acyqu01
Hassan Adams	adamsha01
Jaylen Adams	adamsja01
Jordan Adams	adamsjo01

R CODES

First letter of the Slug

"https://.../players/", **initial**, "/",**slug**,”.html”

Slug scraped with BBR function “get_player”

WEB SCRAPING 2/4

Scraped Elements: BASKETBALL REFERENCE

- **Player Name**
- **Image Link**
- **Tables (Per-Game st., Advanced st., Cap History)**

Per Game

Bold indicates league leader Share & Export ▾ Glossary

Regular Season | Playoffs

Season
2018-19
2019-20 ★
2020-21 ★
2021-22 ★
Career

Advanced

Bold indicates league leader Share & Export ▾ Glossary

Regular Season | Playoffs

Season	Age	Tm	Lg	Pos	G	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	US							BPM	VORP
2018-19	19	DAL	NBA	SG	72	2318	19.6	.545	.433	.409	4.0	21.9	13.0	31.6	1.6	0.9	15.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3.9	3.4
2019-20 ★	20	DAL	NBA	PG	61	2047	27.6	.585	.431	.448	4.1	25.0	14.7	45.7	1.5	0.6	14.8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	8.4	5.4
2020-21 ★	21	DAL	NBA	PG	66	2262	25.3	.587	.406	.349	2.7	22.9	12.8	44.1	1.4	1.5	15.3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6.8	5.1
2021-22 ★	22	DAL	NBA	PG	65	2301	25.1	.571	.406	.349	2.7	26.0	14.3	46.0	1.6	1.4	15.3	**37.4**	3.8	3.8	7.6	.159	6.4	1.8	8.2	5.9
Career			**NBA**		**264**	**8928**	**24.3**	**.573**	**.418**	**.386**	**3.3**	**23.9**	**13.7**	**41.7**	**1.5**	**1.1**	**15.1**	**35.1**	**17.5**	**11.5**	**29.0**	**.156**	**5.8**	**0.9**	**6.8**	**19.8**

Salary Cap History

Sha

Year	Salary Cap	2021 Dollars
1984-85	$3,600,000	$9,083,539
1985-86	$4,233,000	$10,479,694
1986-87	$4,945,000	$11,812,089
1987-88	$6,164,000	$14,143,899
1988-89	$7,232,000	$15,832,740
1989-90	$9,802,000	$20,360,496
1990-91	$11,871,000	$23,652,090
1991-92	$12,500,000	$24,172,990
1992-93	$14,000,000	$26,300,391
1993-94	$15,175,000	$27,784,491
1994-95	$15,964,000	$28,431,190
1995-96	$23,000,000	$39,797,368
1996-97	$24,363,000	$41,185,792
1997-98	$26,900,000	$44,787,566
1998-99	$30,000,000	$48,871,354

WEB SCRAPING 3/4

spotrac

Current & Previous Contract Types

CONTRACT: 5 yr(s) / $201,158,790	**AVG. SALARY:** $40,231,758	**GTD AT SIGN:** $201,158,790	**SIGNED USING:** Designated Player Veteran Extension/Bird	**FREE AGENT:** 2022 / UFA

2013-2016

CONTRACT: 4 yr(s) / $44,000,000	**AVG. SALARY:** $11,000,000	**SIGNED USING:** Rookie Extension/Bird	**FREE AGENT:** 2017 / UFA

2009-2012 ENTRY LEVEL

CONTRACT: 4 yr(s) / $12,700,262	**AVG. SALARY:** $3,175,066	**SIGNED USING:** Entry Level/Rookie

WEB SCRAPING 4/4

Columns **Season** and **Player** used to merge the two scraped datasets, obtaining a final dataset with **61 variables** and **6870 rows**

season	Player	Image_Link	US…ayer	age	tm	lg	pos	g	Salary_Ca	Salary_Ca	Sala
1999-00	Jeff Foster	https://www.b	1	23	IND	NBA	C	19	0.0252282	34000000	8
1999-00	Steve Francis	https://www.b	1	22	HOU	NBA	PG	77	0.0888388	34000000	30
1999-00	Kenny Thomas	ht									
1999-00	Richard Hamilton	ht									
1999-00	Baron Davis	ht									
1999-00	Elton Brand	ht									
1999-00	Jason Terry	ht									
1999-00	Corey Maggette	ht									
1999-00	James Posey	ht									
1999-00	Shawn Marion	ht									
1999-00	Chucky Atkins	ht									

season	player_name	signed_using	contract_type	rookie_co	bird_right	maxi
1999-00	Kevin Garnett	Rookie Extension	Standard	0	0	
1999-00	Tim Duncan	Entry Level	Rookie Contract	1	0	
1999-00	Shaquille O'Neal	Standard	Standard	0	0	
1999-00	Rashard Lewis	Entry Level	Rookie Contract	1	0	
1999-00	Chris Webber	Standard	Standard	0	0	
1999-00	Juwan Howard	Standard	Standard	0	0	
1999-00	Alonzo Mourning	Standard	Standard	0	0	
1999-00	Michael Finley	Rookie Extension	Standard	0	0	

03

OUR RESULTS & CODE DEMO

Model Evaluation

RMSE

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}}$$

Standard way to measure the error of a model in predicting quantitative data

R^2

$$R^2 = 1 - \frac{RSS}{TSS}$$

Statistical measure that represents the proportion of the variance of a dependent variable that's explained by the model

Model Summary: OLS

VERSIONS:

- Pure OLS on the entire dataset
- Variable exclusion for correlation reduction
- PCA: 18 out of 56 components explain most of the variability
- Forward, Backward and Bi-directional variable selection: **Forward selection based on AIC works the best**

The importance of CONTRACT VARIABLES is confirmed in every alternative

Best RMSE:

0.040

Best R^2:

0.715

Computational Costs

Train Score

Model Summary: KNN

VERSIONS:

- Simple KNN (K = 5)
- Resampled (25 bootstrap rep.), cross validated K (K= 9)
- Resampled, centered, scaled, cv'd (K = 9)
- Resampled, centered, scaled, cv'd (K = 11)
- **Last 7 seasons, centered, scaled, cv'd (K = 9)**

When only looking at the last 7 season, the chosen K drops back to 9

Best RMSE:

0.033

Best R^2:

0.808

Computational Costs

Model Summary: RF

VERSIONS:

- Random Forest using full dataset
- Random Forest keeping only players who played more than 20 games in a season
- **Random Forest keeping only last 7 seasons (from 2014/2015 to 2020/2021)**

Focusing only on the last 7 seasons, BIRD RIGHTS became the most important variable

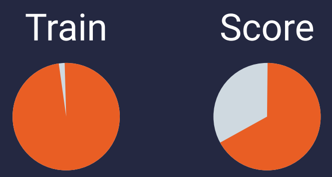
Best RMSE:

0.035

Best R^2:

0.765

Computational Costs



Model Summary: GB

VERSIONS:

- Gradient Boosting, full dataset
- Gradient Boosting keeping only players who played more than 20 games in a season
- **Gradient Boosting keeping only last 7 seasons (from 2014/2015 to 2020/2021)**

Most increased variable importance: THREE POINT PERCENTAGE (from 0.056 to 0.648, 11x increase)

Best RMSE:

0.036

Best R^2:

0.777

Computational Costs

Train Score

Model Summary: SVM

VERSIONS:

- Linear, **Radial** and Polynomial Kernels
- SVM keeping only players who played more than 20 games in a season
- **SVM keeping only last 7 seasons (from 2014/2015 to 2020/2021)**

<u>Best Setting:</u> Radial Kernel, keeping only last 7 seasons (from 2014/15 to 2020/21)
<u>Tuning:</u> C = 1.5, Sigma = 0.005

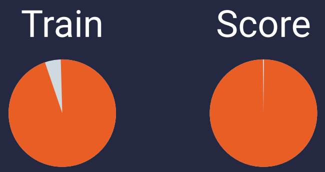
Best RMSE:

0.032

Best R^2:

0.821

Computational Costs



Models Summary

Model	RMSE	R^2
OLS	0.040	0.715
KNN	0.033	0.808
Random Forest	0.035	0.765
Gradient Boosting	0.036	0.777
SVM	0.032	0.821

Model Summary II

OLS

Need to eliminate Multicollinearity prior to model being usable

KNN

Good results, but not quite as interpretable

Random Forest

Decent results, ability to interpret results and what drives performance

Gradient Boosting

Similar to Random Forest, but improved performance

SVM

Best results achieved, interpretable results

Results Interpretation

- If we fit well-performing model to only the last 7 seasons (in line with the experts opinion on when the increased focus for 3pt shots happened), we see an increase in model performance
- Most important variables proved to be minutes played, points scored, age as well as our contract variables
- Advanced Metrics do not perform as well as expected, instead, we see a market reliance on standard metrics
- If we compare variable importance for the whole dataset and only the last 7 seasons, we see that the biggest increase in importance is 3pt% (11x), Box Plus-Minus (8x), and Defensive Win Shares (7.9x)

“Agent” Effect: Chandler Parsons

Season	Predicted Salary	True Salary
2014/2015	10.113% ($6,377,544)	23.309% ($14,700,000)
2015/2016	11.030% ($7,720,861)	21.945% ($15,361,500)

“I end up hiring Dan Fagen, the only reason because he said 'I can get you out of that fourth year.' And no one else could. How he did it is he basically used leverage. He went to the GM, he went to the owners and said 'I'll get you Dwight Howard, but you're not picking up Parsons' contract. So instead of getting paid $920k I got bumped to a max and we got Dwight Howard. Agents get you paid but Dan Fagan got you overpaid.”

— Chandler Parsons, answering on how he got his contract

Behavioral Problems: DeMarcus Cousins

Season	Predicted Salary	True Salary
2018/2019	17.283% ($17,605,589)	5.239% ($5,337,000)

DeMarcus Cousins Audio Allegedly Threatening to Shoot Baby Mama Before Wedding

DEMARCUS COUSINS
ALLEGEDLY THREATENED TO KILL BABY MAMA
'Bullet In Your F'ing Head'

ESPN Stats & Info
@ESPNStatsInfo

DeMarcus Cousins
Since 2010-11, leads NBA in

Technical fouls 105
Times fouling out 46
Ejections 12

3:00 PM · Feb 20, 2017 · TweetDeck

Feeling the Pressure: Ryan Anderson

Season	Predicted Salary	True Salary
2017/2018	7.828% ($7,756,916)	19.758% ($19,578,455)
2018/2019	4.577% ($4,662,805)	20.047% ($20,421,546)

"It was a new thing for me, because I had sort of always been the underdog, overachieving and now I was sort of the overpaid guy who was underachieving from what they wanted. It was hard for me to be the guy that was like, 'You need to do more and we're paying you a lot for this,' rather than before it was like, 'Wow, we got a steal for this guy.' It really affected me at home. I felt like every time I was in Houston, I was letting down the fans, or something like that. Houston's one of those sports cities where just the pressure is always on you, and that's all people want to talk about with you."

— Ryan Anderson, on performing after signing his new contract

Most Underpaid Players by Model

Gradient Boosting	Random Forest	KNN
Andre Drummond	Andre Drummond	Blake Griffin
Carl Landry	Damion Lee	Bobby Portis
David West	**DeMarcus Cousins**	Chris Boucher
DeMarcus Cousins	Jusuf Nurkic	JaKarr Middleton
Kemba Walker	Kemba Walker	**Michael Carter-Williams**
Kenny Thomas	Khem Birch	Montrezi Harrell
Marc Jackson	Marco Belinelli	Reggie Bullock
Michael Carter WIlliams	Michael Carter -Williams	Spencer Dinwiddie

- Quite some difference between Models
- Gradient Boosting and Random Forest being comparable seems intuitive, since both are tree-based models

04

GM's Little Helper

R-SHINY APP

https://sportylytics-predictions.herokuapp.com/

HEROKU

THANK YOU!