IS THE WORLD TRULY LINEAR?

Benchmarking OLS with non-linear models:
A “supervised” investigation on players’
compensation exploiting basketball data

20630 - Introduction to Sport Analytics
Team 2, Mid-Term Presentation

OUR TEAM

Federico Leonardi

Alberto Allegri

Jakob Schlierf

Tiziano Paci

HC:

Beatrice Guidotti

Leonardo Yang

TABLE OF CONTENTS

01
BACKGROUND
CBA, salary cap & more: How to deal with them

02
OUR DATA
Literature review, data collection & final dataset

03
OUR APPROACH
ML models & regularization techniques: A theoretical introduction

04
NEXT STEPS
ML as a trial-and-error process: Dependent variable and predictors alternatives

01 Background

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.

RULES OF THE GAME: THE CBA

Collective Bargaining Agreement (CBA):

- Signed by the NBPA (Players' Union) and the NBA
- Sets out the terms and conditions of employment, as well as the respective rights and obligations, of the NBA Clubs, the NBA, and the NBPA
- Dictates the rules of player contracts, trades, revenue distribution, the NBA Draft, and the **salary cap**, among other things

NBPA meeting before signing the last CBA, 2017

CBA: THE SALARY CAP

Salary Cap:

- **Maximum** total amount of money that NBA teams are allowed to pay their players
- Calculated yearly by multiplying projected Basketball-Related income (TV deals, ticket purchases and merchandise sales) by 44.74% and then dividing by the number of teams
- Teams are **required to spend at least 90%** of the salary cap during a season

Adam Silver, NBA commissioner

SALARY CAP ROADMAP

How CBA and Salary Cap affect our project:

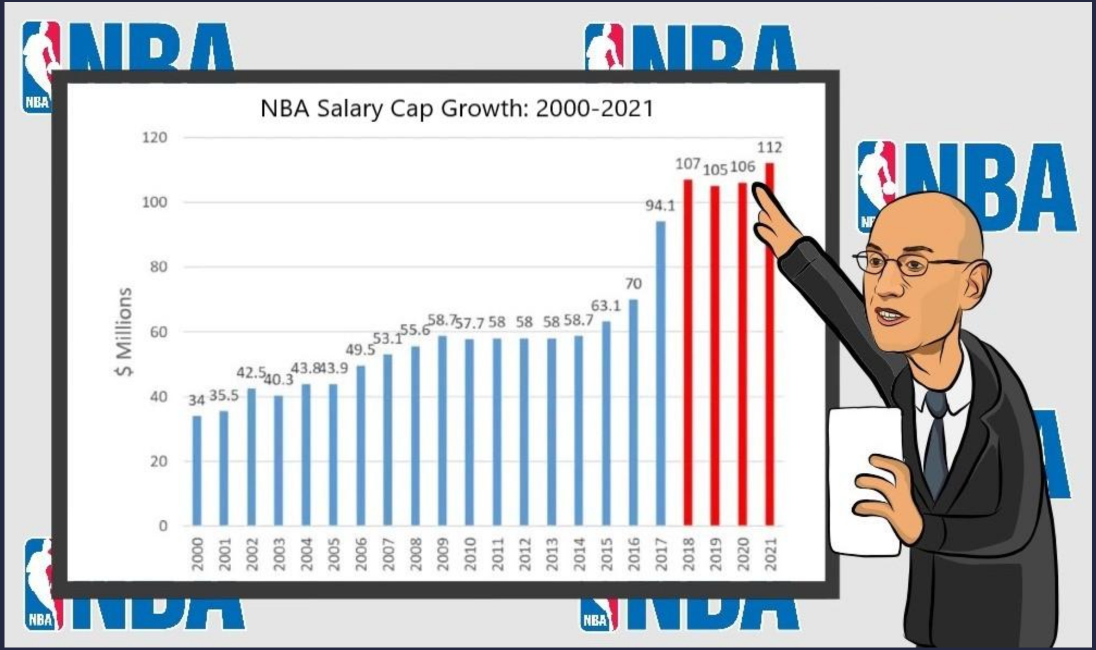
- **NBA "Inflation"**: Cap growth over the years
- **Soft Cap**: Ways to get over the cap
- **Special Contracts**: Rookie and Maximum

SALARY CAP GROWTH

From $**34**M. in 2000 to $**112**M. in 2021

New TV deal with ESPN in 2016

COVID-19 effect in 2020



SALARY CAP GROWTH ($ Million)

Season	Average	Avg. Infl. Adjusted	Median	Maximum
2000/2001	2.92	2.92	1.93	19.6 (K. Garnett)
2010/2011	4.64	3.65	2.82	24.8 (K. Bryant)
2020/2021	8.25	5.02	4.02	43.0 (S. Curry)

TAKEAWAY

We should not use nominal or real-world inflation adjusted salaries as target variable.

As we have seen, those numbers are biased towards recent years due to NBA “Inflation”.

OUR SOLUTION

Take salaries expressed as salary cap percentage each year as our target variable

Most contracts have already “priced-in” expectations of future salary cap growth. For this reason, taking the percentage year-by-year does not actually underestimate contracts signed in the past.

SALARY CAP %

Season	Salary Cap	N. Jokic	N. Jokic (Cap %)
2018/2019	101,869,000	25,467,250	25.00%
2019/2020	109,140,000	27,504,630	25.20%
2020/2021	109,140,000	29,542,010	27.07%
2021/2022	112,414,000	31,579,390	28.09%

SALARY CAP ROADMAP

How CBA and Salary Cap affect our project:

- **NBA "Inflation"**: Cap growth over the years
- **Soft Cap**: Ways to get over the cap
- **Special Contracts**: Rookie and Maximum

SOFT & HARD CAP

- A **hard** salary cap does not allow the total payroll for the team to be exceeded for any reason (NFL, NHL, MLS).

- The NBA, compared to other professional american leagues, has what is known as a **“soft”** salary cap.

- The NBA's cap contains so many **exceptions** that very few teams are ever under the cap for a season.

LUXURY TAX

- Teams that go over the league's cap will pay a "**luxury" tax** depending on different tiers.

- The tiers are set according to how much money that team has gone over the cap and for how many years.

- Tax money are equally distributed among non-tax paying teams.

- Owners willing to pay luxury taxes may get a competitive advantage.

SALARY CAP BY TEAM (TOP 5)

Team	Total Cap	Cap Maximum	Lux. Tax Thr.	Cap Space
	184,024,769	112,414,000	136,606,000	-71,610,769
	172,815,092	112,414,000	136,606,000	-60,401,092
	166,008,910	112,414,000	136,606,000	-53,594,910
	165,361,473	112,414,000	136,606,000	-52,947,473
	161,851,801	112,414,000	136,606,000	-49,437,801

SALARY CAP BY TEAM (BOTTOM 5)

Team	Total Cap	Cap Maximum	Lux. Tax Thr.	Cap Space
SAN ANTONIO SPURS	126,419,926	112,414,000	136,606,000	-14,005,926
CHARLOTTE HORNETS	120,798,884	112,414,000	136,606,000	-8,384,884
NEW YORK KNICKS	120,376,240	112,414,000	136,606,000	-7,962,240
	115,994,102	112,414,000	136,606,000	-3,580,102
OKC	90,342,857	112,414,000	136,606,000	+22,071,143

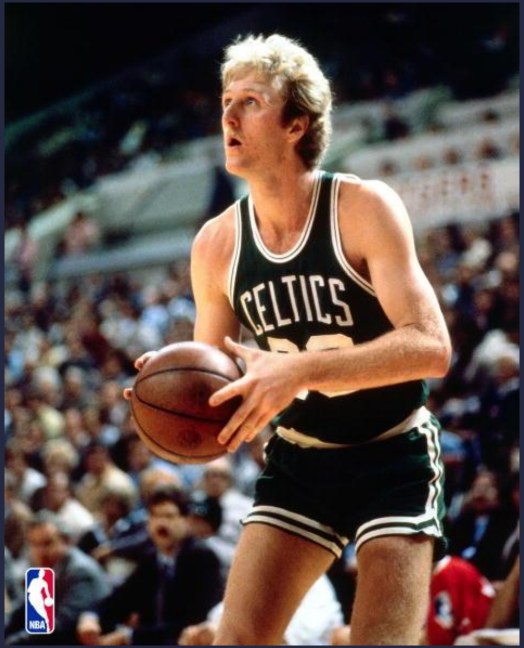
WAYS TO GET OVER THE CAP

Teams can use the following exceptions to exceed the salary cap:

- **Veteran Free Agent Exceptions (“Bird” or “Early Bird” rights)**
- Traded Player Exception
- Mid-Level Salary Exceptions
- Bi-Annual Exception
- Rookie Exception
- Minimum Player Salary Exception
- Disabled Player Exception

BIRD RIGHTS

- Bird rights are by far the **most commonly used exception** to exceed the salary cap
- Introduced in the 1983 CBA and used for the first time by the Boston Celtics to keep their franchise player Larry Bird
- **Reward loyal players**, allowing teams to sign bigger contracts without worrying about the salary cap
- To do so, a team has to “earn” Bird rights on a player



BIRD RIGHTS

Type	Years under contract necessary	Salary over the cap allowed
Non-Bird Rights	1	120% previous contract
Early Bird Rights	2	175% previous contract
Full Bird Rights	3+	Maximum

TAKEAWAY

Teams may be willing to overpay players if they own their Bird rights.

The salary cap is a budget constraint, but if you can sign certain players ignoring their cap hit, it becomes way more difficult to predict how performances are affecting your decision.

OUR SOLUTION

Add a dummy variable for contracts signed using bird rights.

Using Bird rights as an additional regressor we potentially add another source of information.

SALARY CAP ROADMAP

How CBA and Salary Cap affect our project

- **NBA “Inflation”**: Cap growth over the years
- **Soft Cap**: Ways to get over the cap
- **Special Contracts**: Rookie and Maximum

SPECIAL CONTRACTS

- **Rookie** Contracts
- **Maximum and Supermax** Contracts

Nonlinearity Concerns

ROOKIE CONTRACTS

- A rookie contract is given to any player that has never before played in the league regardless of age.
- Players must be selected (“drafted”) by a team during the NBA Draft.
- The value and length are tied to when a player is drafted (“draft position”).
- **Given the draft position, rookie contracts are fixed and do not need to be negotiated.**

2021-22 DRAFT CLASS SALARY

Pick	Year 1	Year 2	Year 3 (Option)	Year 4 (Option)
1	8,375,100	8,794,000	9,212,700	11,617,215
2	7,493,500	7,868,100	8,243,000	10,402,666
3	6,729,300	7,065,600	7,402,300	9,356,507
4	6,067,100	6,370,600	6,673,800	8,442,357
5	5,494,200	5,768,700	6,043,500	7,657,115
...	...	...	...	...

TAKEAWAY

Rookie contracts, by definition, do not depend on performances.

Rookies, especially top picks, are often severely underpaid compared to players with similar performances.
Fixed contracts with a relatively low salary have a huge outlier potential.

ROOKIE CONTRACTS

 Luka Dončić		 Royce O'Neale
8,049,360 $	Salary 20/21	8,800,000 $
27.7 / 8.0 / 8.6	Pts./Reb./Ast.	7.0 / 6.8 / 2.5
Rookie	Contract	Veteran Extension

OUR SOLUTION

Add a dummy for players on rookie contracts.

Adding a control for players on rookie contracts, we can eventually decide to filter them out.

MAXIMUM CONTRACTS

- The CBA regulates the maximum amount of money a team can pay an individual player, his "**maximum contract**".
- Depends on seniority and accomplishments.
- The last CBA introduced the "**supermax**" contract, to reward loyalty.
- Right now, only **6** players have signed a supermax contract, while around **50** players have standard max contracts.

MAXIMUM CONTRACTS

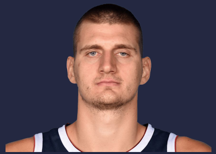
Years of Service	Maximum Salary (Salary Cap %)
6 Years or Less	25%
7-9 Years	30%
10+ Years	35%

TAKEAWAY

Maximum contracts put the top ~50 players on the same salary level.

Maximum contracts artificially create a ceiling over how much an individual player can get paid, regardless of how he performs.
In an "open" market we would expect a significant difference between how much the first and the fiftieth best players earn.

MAX CONTRACTS

 Nikola Jokić		 K. Porziņģis
29,542,010 $	Salary 20/21	29,467,800 $
26.4 / 10.8 / 8.3	Pts./Reb./Ast.	20.1 / 8.9 / 1.6
Maximum	Contract	Maximum

OUR SOLUTION

Add a dummy for players on maximum contracts.

Again, we would like to investigate how our model performs controlling for maximum contracts.
Given the nature of special contracts, we believe non-linear models might be more suited to our purposes.

RECAPPING

PROBLEMS	SOLUTIONS
Salary Cap Growth	Take salaries expressed as salary cap percentage 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

02

OUR DATA

“In God we trust;
all others have to bring data.”

—W. Edwards Deming

American engineer, statistician & professor (1900 - 1993)

DATA COLLECTION

Source

BASKETBALL REFERENCE

Statistics

Per-Game &
Advanced statistics

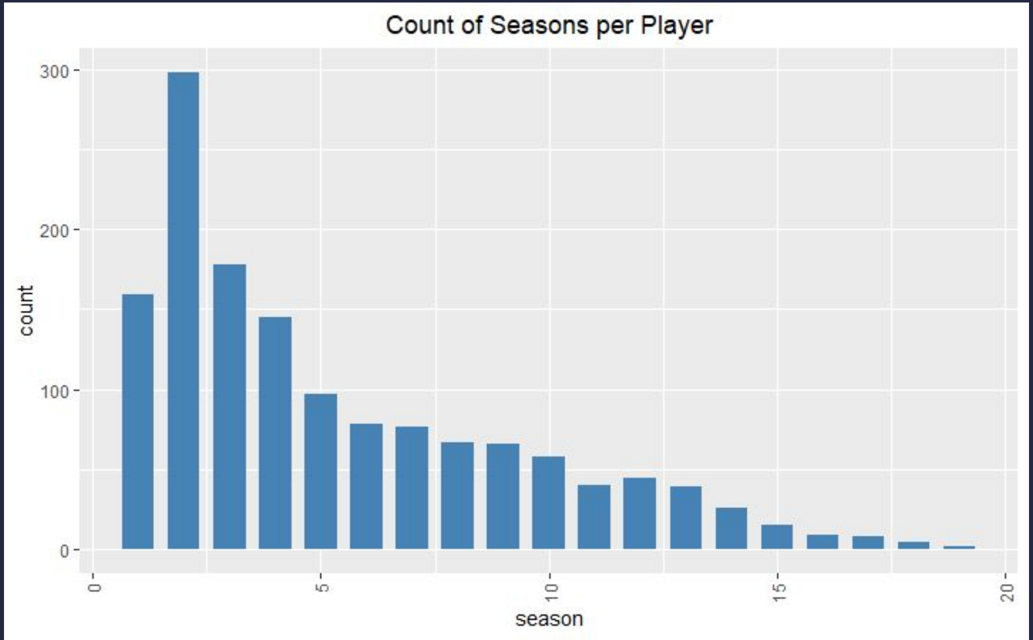
Target

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

Some Numbers

52 Variables
7555 Rows
1412 Players

COUNT OF SEASONS



5.35

Average across players

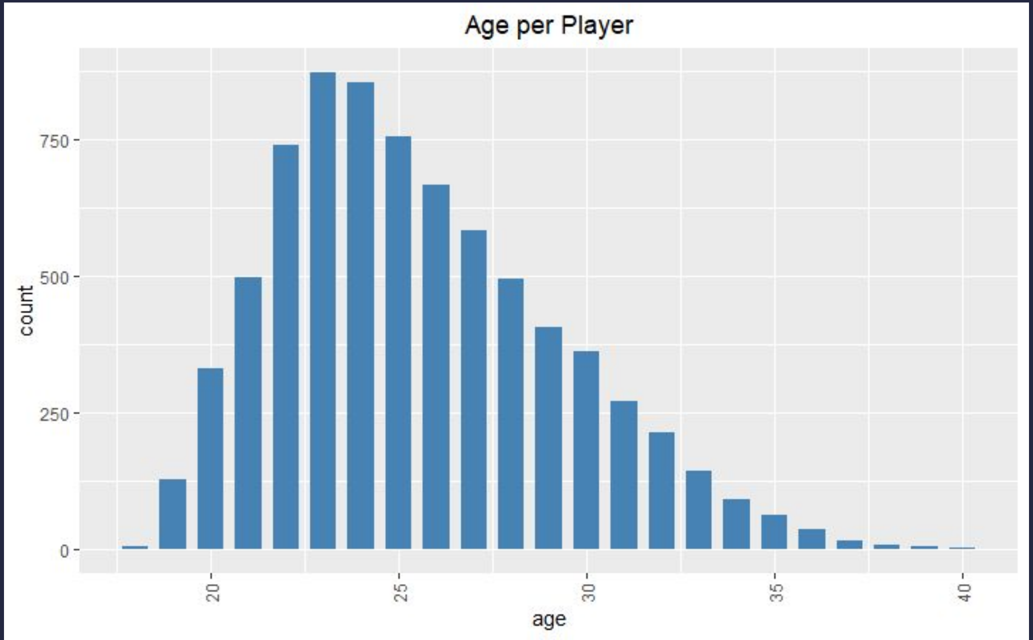
Tyson Chandler

Jason Terry

19

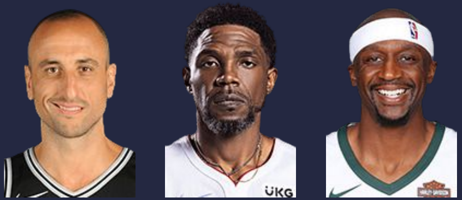
Max number of seasons for a single player

AGE DISTRIBUTION



25.5

Average across players



Manu Ginóbili

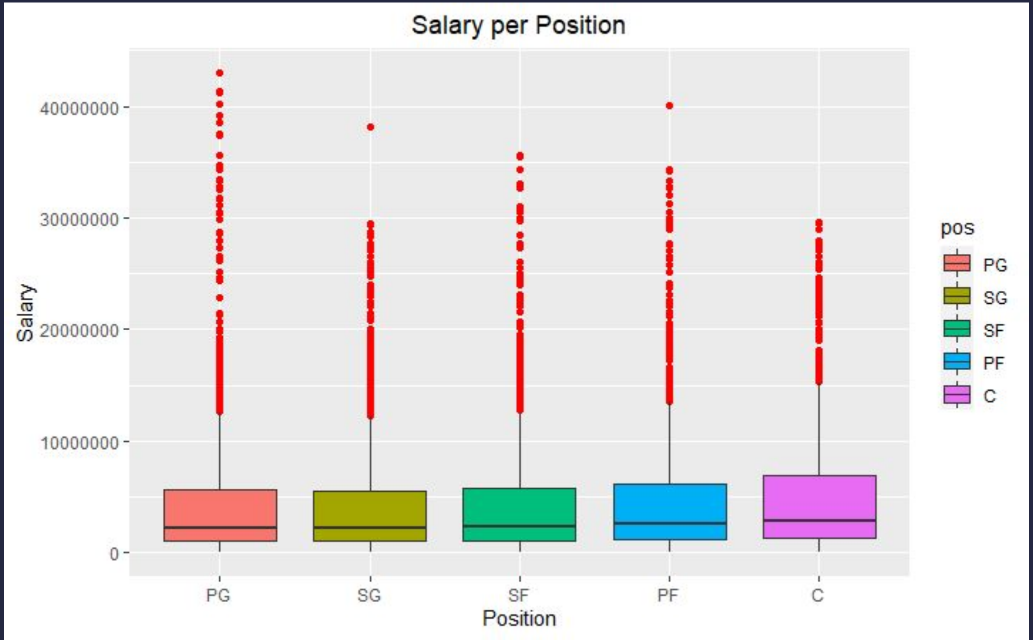
Udonis Haslem

Jason Terry

40

Max age in a season for a single player

DATA FOR COMPENSATION



$4,703,898

Average across players

Stephen Curry

$43,006,362

Max salary in a season for a single player

QUICK LOOK TO COMPENSATION PER POINTS SCORED (ACROSS POSITION)

X-Axis

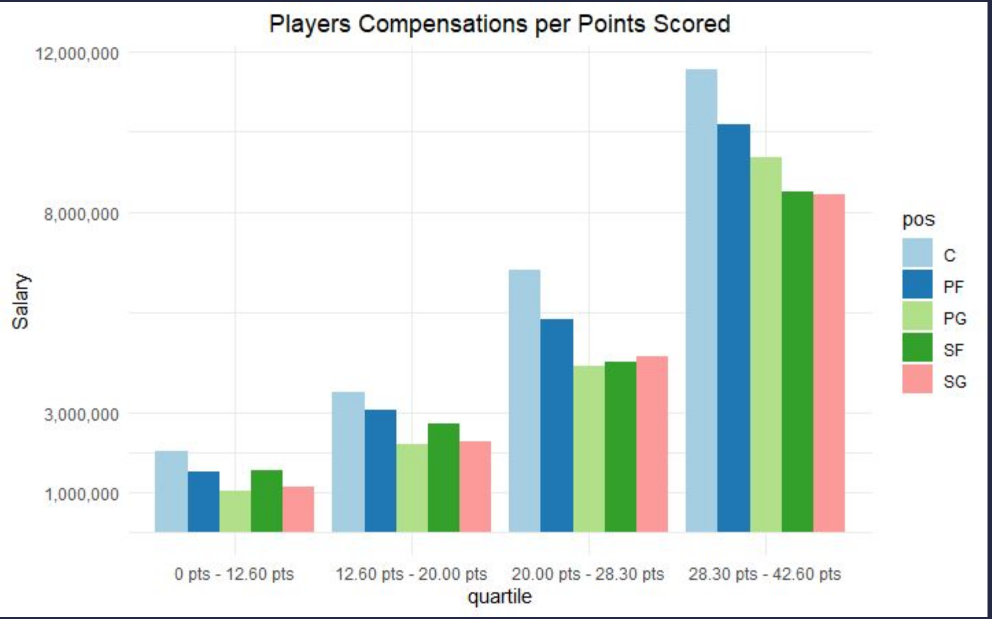
Quartiles of Points/Game in a season

Y-Axis

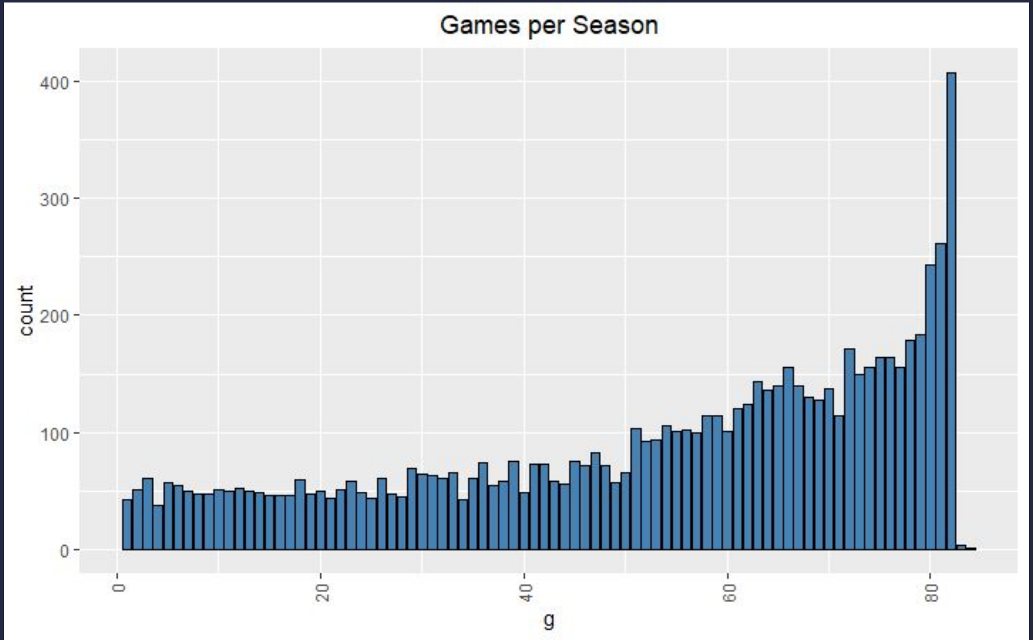
Players Average Compensation

Fill

Player Position on the field



DATA FOR GAMES x SEASON



53.63

Average across players

Joe Ingles

77.86 in 7y

Max average of Games x Season

OUR VARIABLES (1/2)

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 (2/2)

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

LITERATURE REVIEW

When Statistics, Basketball and ML tools are combined, interesting elements may be derived.

- *Rosson J.*: **Normalization** of salary as a % of salary cap & distinction among **basic**, regular and **advanced stats.**
- *Fleenor A. T.*: Pay attention to the **outliers**!
- *Papadaki I. and Tsagris M.*: Avoiding overfitting; **Non-Linearity**; Variable Selection with **LASSO** & ML (**Random forests**); in-game stats are the best performer.
- *Wu W. et al.*: Specific players' examples & prediction in terms of **salary ranges.**

Sources: 1) NBA Salary Predictions using Data Science and Linear Regression, 2) Predicting National Basketball Association (NBA) Player Salaries, 3) Are NBA Players' Salaries in Accordance with Their Performance on Court?, 4) Classification of NBA Salaries through Player Statistics.

QUINTESSENTIAL ELEMENTS

NORMALIZATION

Statistics may lead to biased results in many dimensions when not normalized.

STATS: BASIC vs ADVANCED

It's not gold all that glitters.

FLAWS & OUTLIERS

Something's missing and something's "strange".

ML MODELS

Introducing non-linearity: Is OLS the best solution?

REGULARIZATION TECHNIQUES

(Sometimes) Less is More.

TRIAL & ERROR

Don't expect the first model to be the best performer.

03 OUR METHOD

SUPERVISED LEARNING

Learning from labeled data

Classification
Predict a label/class

Regression
Predict a continuous quantity

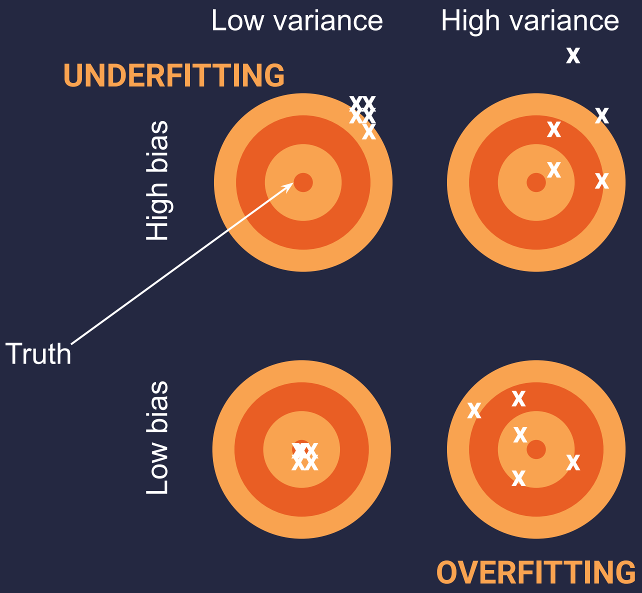
ML CHEAT SHEET

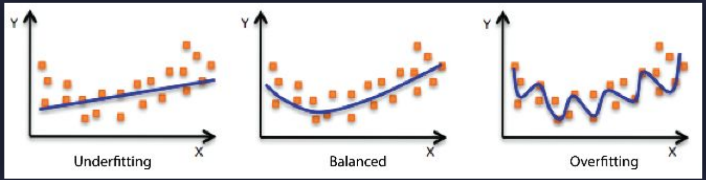
BIAS

Difference between the **average prediction** of our model and the **correct value** which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model.

VARIANCE

Variability of model prediction for a given data point. Model with high variance pays a lot of attention to training data and does not generalize well.





IN THIS CHAPTER

Dimensionality Reduction

Reduce the number of variables without losing predictive power

Models

Explain & predict player's salaries using Machine Learning models

DIMENSIONALITY REDUCTION

Allows to reduce the number of attributes in a dataset while **keeping as much of the variation** in the original dataset as possible.

We still lose some percentage of the variability of the original data, but there are **many advantages**:

- Less training time and computational resources
- Increases in performance
- Reduces the risk of overfitting
- Takes care of multicollinearity
- Makes multi-dimensional data plottable
- Removes noise

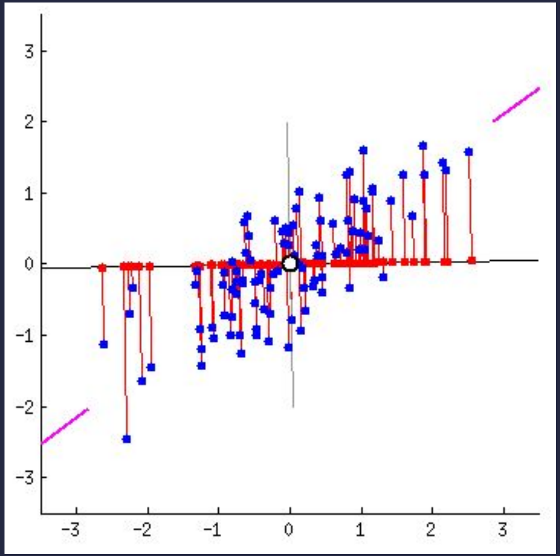
PRINCIPAL COMPONENT ANALYSIS

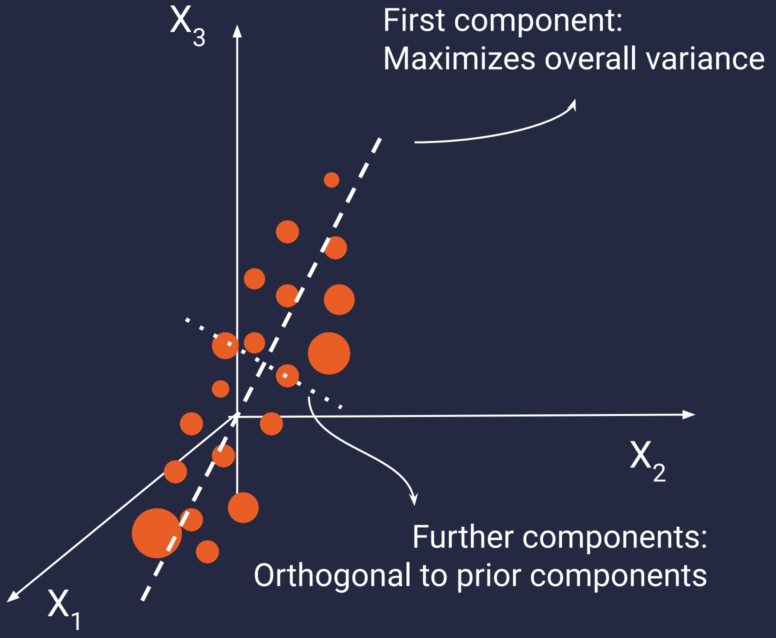
Imagine there are three features in our dataset:

$$PC1 = a_1x_1 + a_2x_2 + a_3x_3$$
$$PC2 = b_1x_1 + b_2x_2 + b_3x_3$$
$$PC3 = c_1x_1 + c_2x_2 + c_3x_3$$

Principal Component Analysis allows for the summary of variables, capturing most information by **linearly combining** multiple variables into a single one.

PCA


3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


X_3
First component:
Maximizes overall variance
X_2
Further components:
Orthogonal to prior components
X_1

PCA

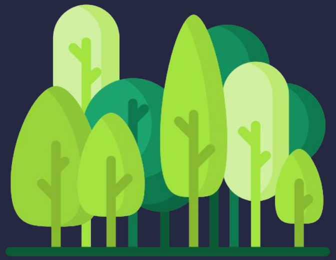
PROS

- Removes correlated features
- Reduces overfitting

CONS

- Interpretability concerns
- Information loss

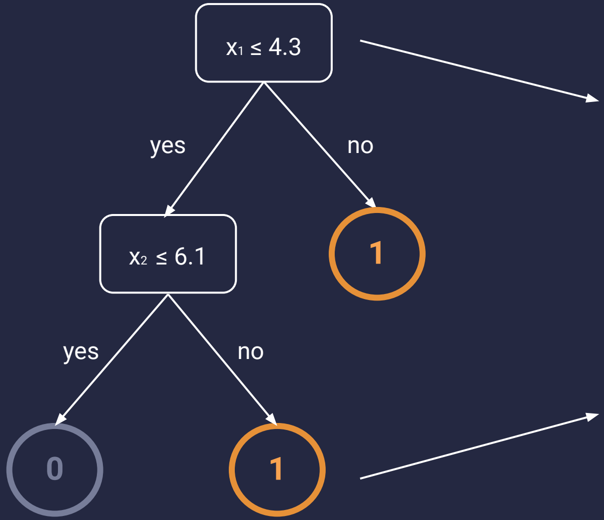
OUR MODELS



Random Forest

K-Nearest Neighbors

DECISION TREES

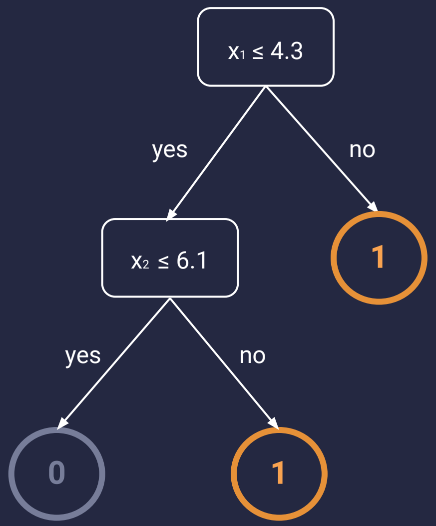


Non-terminal nodes are used to make local decisions based on the local information they possess.

Terminal nodes make the final decision.

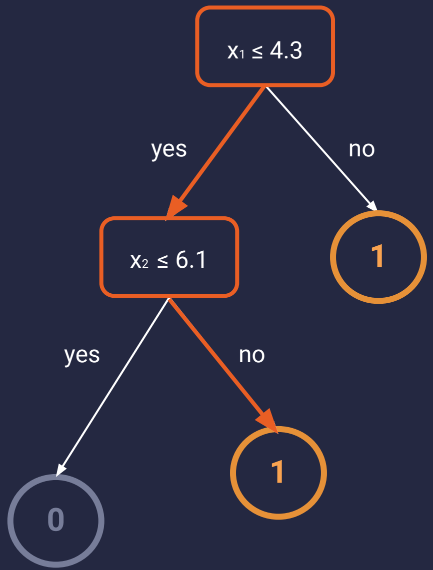
DECISION TREES

ID	y	x_1	x_2
1	0	4.3	4.9
2	0	3.9	6.1
3	1	6.6	4.4
4	0	2.7	4.8
5	1	6.5	2.9
6	1	2.7	6.7

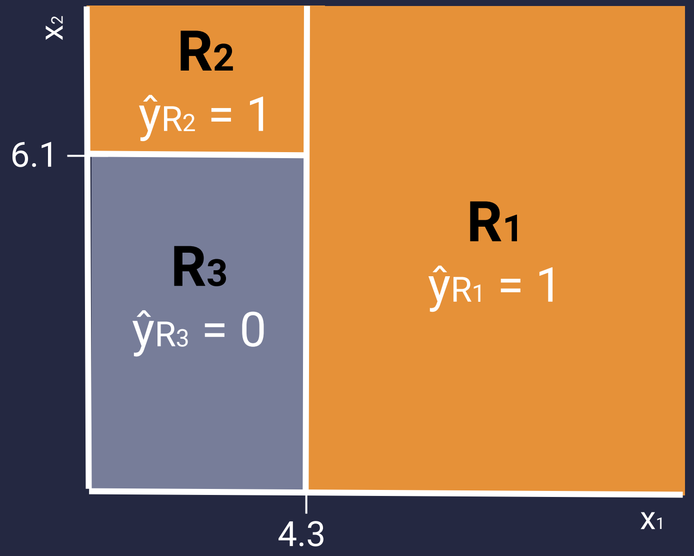


Sample of observations

ID	y	x1	x2
1	0	4.3	4.9
2	0	3.9	6.1
3	1	6.6	4.4
4	0	2.7	4.8
5	1	6.5	2.9
6	1	2.7	6.7



Feature space



y	x_1	x_2
1	2.3	8.1

DECISION TREES

PROS

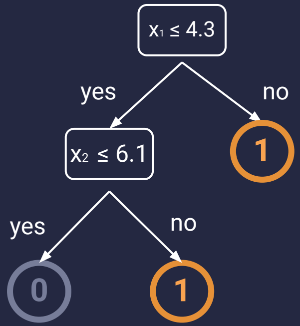
- Very easy to explain, interpret
- Mirror humans decision making
- Handle qualitative data

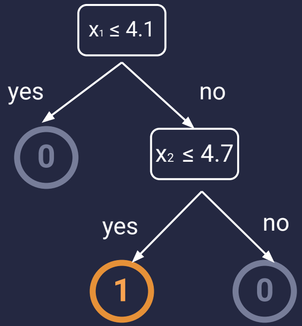
CONS

- Low predictive accuracy
- Non-robust

ID	y	x_1	x_2
1	0	4.3	4.9
2	0	3.9	6.1
3	1	6.6	4.4
4	0	2.7	4.8
5	1	6.5	2.9
6	1	2.7	6.7

ID	y	x_1	x_2
1	0	4.3	4.9
2	0	6.5	4.1
3	1	6.6	4.4
4	0	2.7	4.8
5	1	6.5	2.9
6	1	2.7	6.7

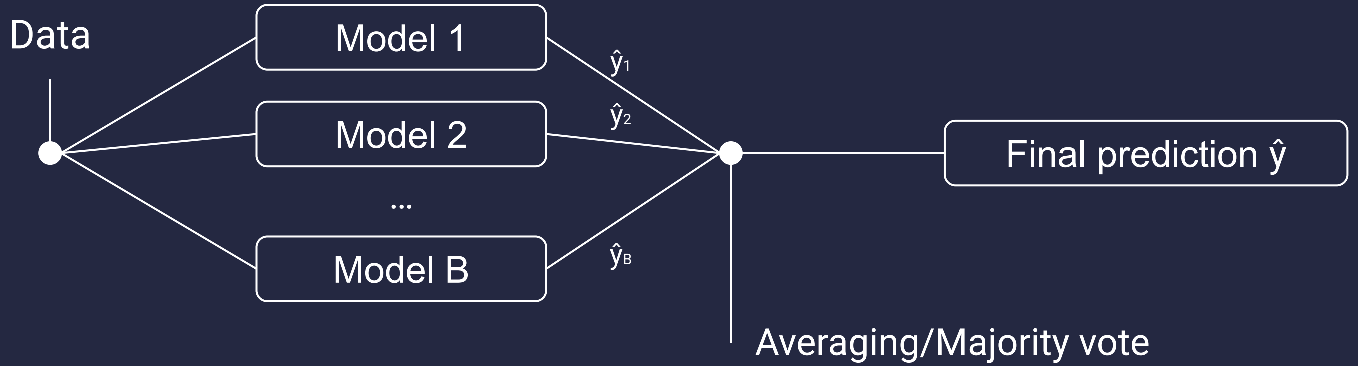




A small change in the data can cause a **large change** in the final estimated tree.

ENSEMBLE LEARNING

Combat overfitting by combining the predictions of many models.



ENSEMBLE LEARNING

To understand the motivation for averaging, consider a set of **uncorrelated** random variables $Y_1,\dots,Y_n$ with common **mean** **$E[Y_i] = \mu$** and variance **$Var(Y_i) = \sigma^2$**.

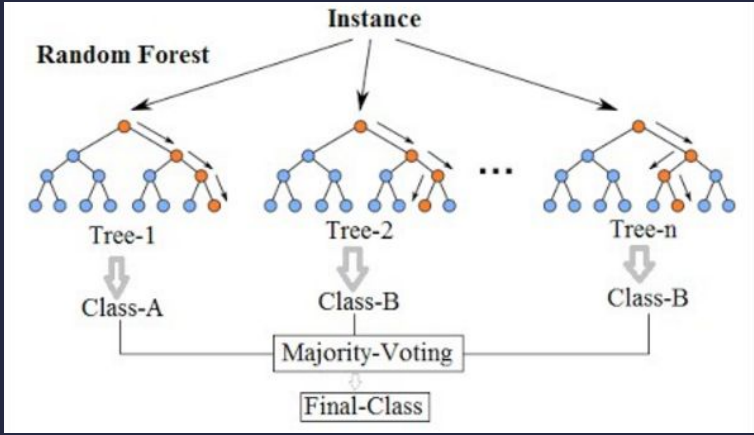
The average of these random variables is the sample mean $\bar{Y}$, whose expected value is the same of the individual Y_i's, but whose variance is **lower**:

$$\mathbb{E}(\bar{Y}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \mu \qquad Var[\bar{Y}] = Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{\sigma^2}{n} < \sigma^2$$

However, real-world predictions will not be completely uncorrelated: given pairwise correlation ρ it can be proven that:

$$Var[\bar{Y}] = \frac{\sigma^2(1-\rho)}{n} + \rho\sigma^2$$

RANDOM FOREST



Two sources of randomization to reduce correlation among the trees:

- **In the dataset** used to train the single tree: We cannot sample from the population, but we can generate B bootstrap samples from the original data.
- **Per-split feature** randomization: For each tree in the forest, randomly select m (rule of thumb: $m \approx \sqrt{p}$) inputs to be considered at each split of that tree.

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

RANDOM FOREST

ID	y	x_1	x_2	x_3	x_4
1	0	4.3	4.9	4.4	4.7
2	0	3.9	6.1	5.9	5.5
3	1	6.6	4.4	4.5	3.9
4	0	2.7	4.8	4.1	5.0
5	1	6.5	2.9	4.7	4.6
6	1	2.7	6.7	4.2	5.3

Sample size: n = 6
Number of features: p = 4
Binary target variable

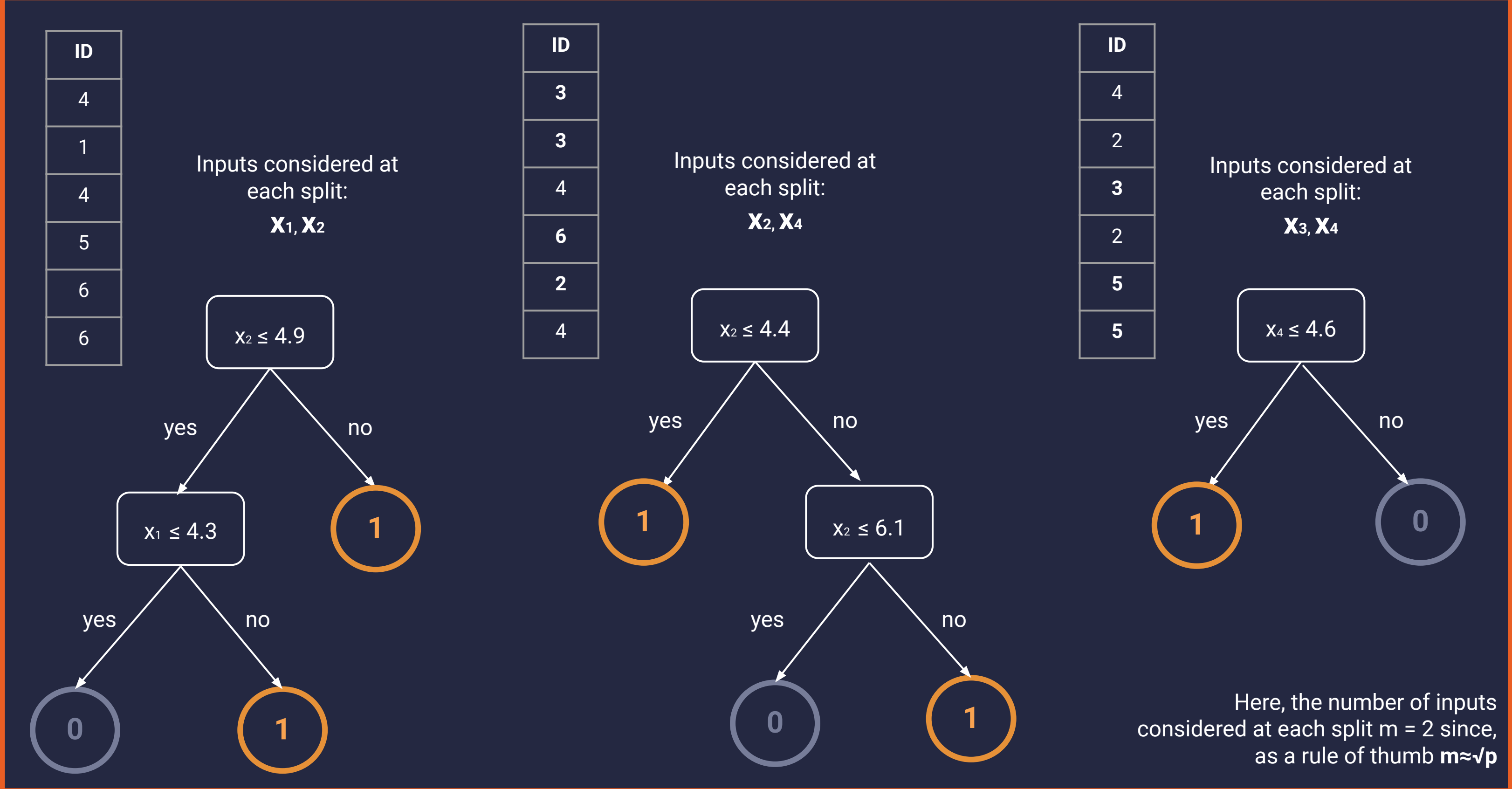
ID
4
1
4
5
6
6

ID
3
3
4
6
2
4

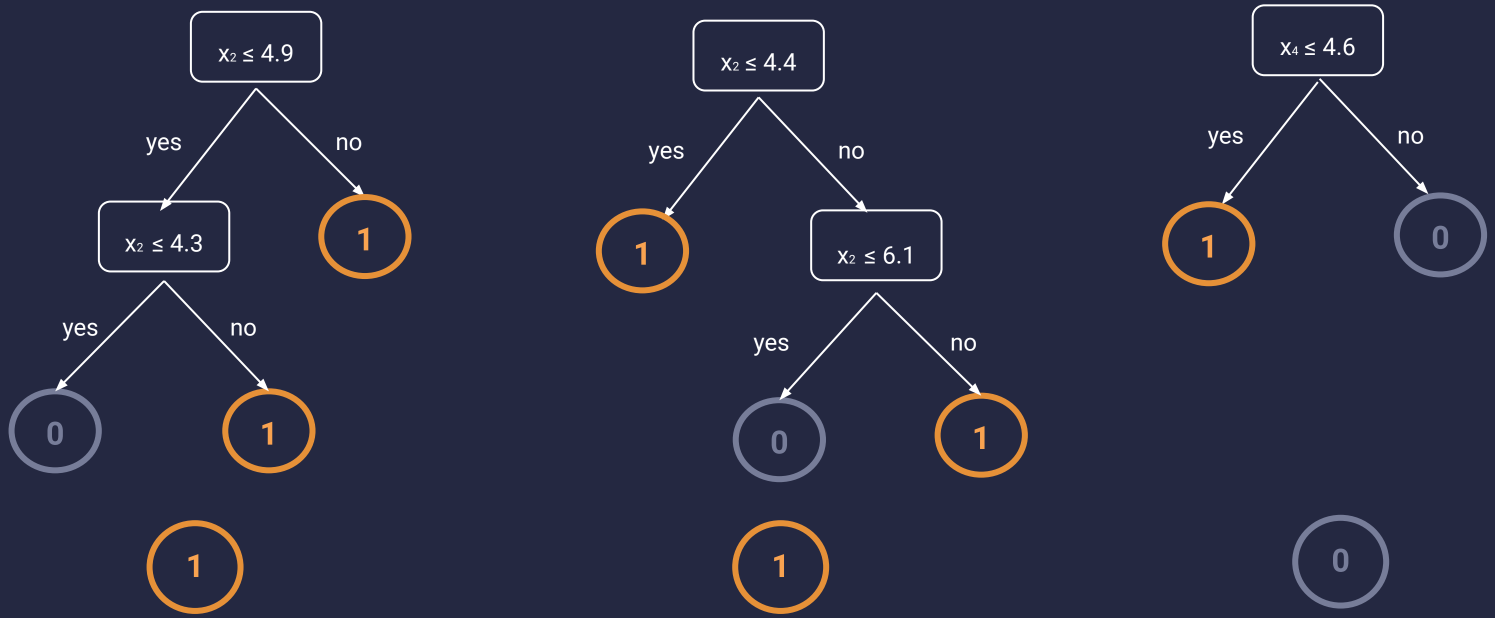
ID
4
2
3
2
5
5

Randomization 1:
BOOTSTRAPPING

Randomization 2: **FEATURE RANDOMIZATION**



Here, the number of inputs considered at each split m = 2 since, as a rule of thumb **m≈√p**



y	x_1	x_2	x_3	x_4
1	2.3	8.1	5.3	4.5

RANDOM FOREST

PROS

- It reduces overfitting
- Classification and regression
- Categorical and continuous variables
- Normalization of data is not required

CONS

- Computational expensive
- Due to the ensemble of decision trees, it also suffers interpretability

K-NEAREST NEIGHBORS

Similar things exist in close proximity

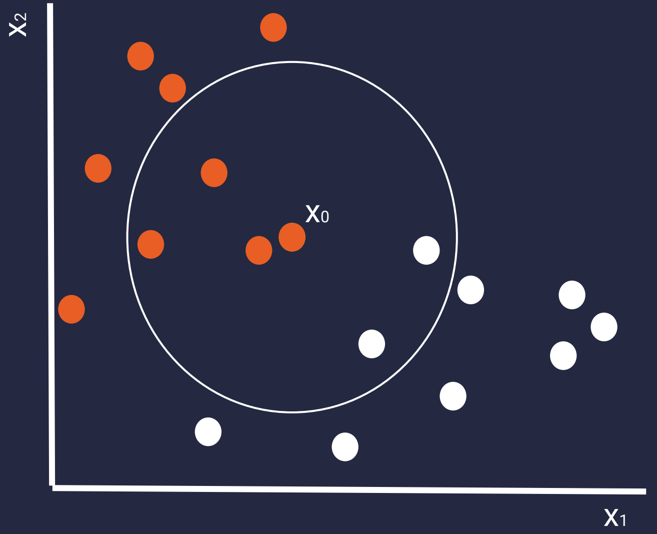
Given an integer k and a test observation x_0:

1. Identify the **k training points** that are closest to x_0, represented by N_0
2. **Estimate** the **conditional probability** of x_0 to be assigned to class j as the fraction of points in the neighborhood N_0 whose target value is equal to j:

$$P(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

3. **Classify** the test observation to the class with the largest probability.

K-NEAREST NEIGHBORS



K = 5

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

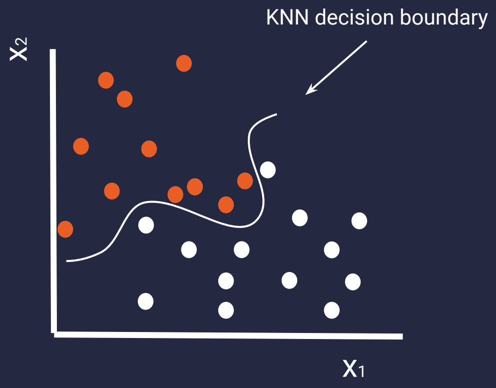
MAJORITY VOTE

↓

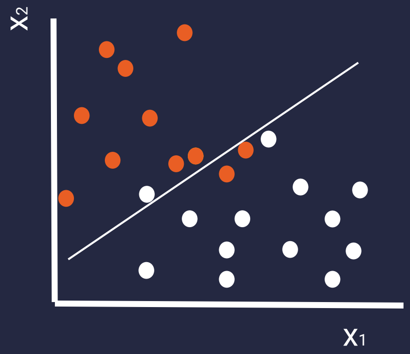
ORANGE CLASS

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

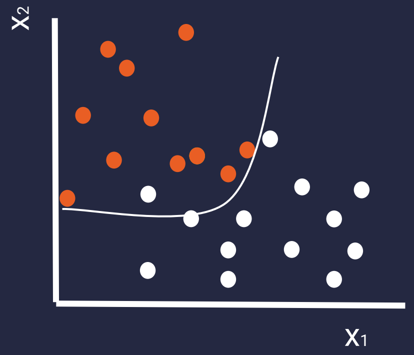
How to choose k?



Small k: low bias & high variance
OVERFITTING



Large k: high bias & low variance
UNDERFITTING



Best k: Controls the balance between overfitting and underfitting

KNN

PROS

- Learning and implementation is extremely simple and intuitive
- Flexible decision boundaries
- No prior knowledge about data distribution is required

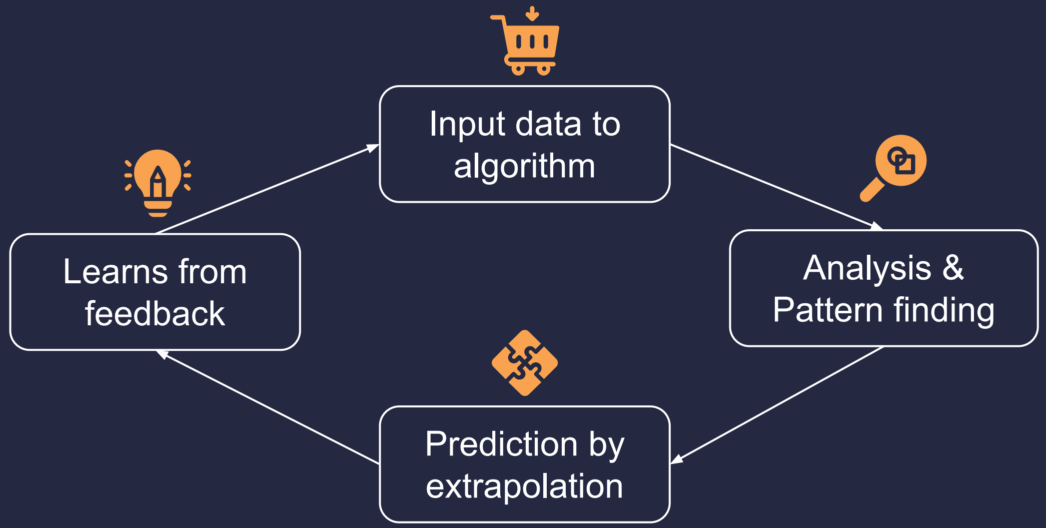
CONS

- Irrelevant or correlated features have high impact and must be eliminated
- Typically difficult to handle high dimensionality
- Computational costs

04 NEXT STEPS

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**.

TRIAL AND ERROR: WHAT?

SALARY

% vs 10

Salary as a percentage of the CAP vs Salary as nominal value

ML MODEL

vs

KNN vs Random Forest

STATISTICS

vs

Basic vs Advanced Stats

THANK YOU!

…And we’re looking forward to show you our results on the final presentation.