NBA Play-By-Play
Final presentation

Jonas Lindner, Ege Arikan, Žan Stanonik, Gregor Caf

Motivation

- Gain deeper understanding of NBA
- How did the game developed over the years?
- Distribution of player performance
- Why/How are some teams/players better than others?
- Is it possible to predict the outcome a game?

Outline

- Data format and loading pipeline
- Exploration of data on different levels:
  - Player
  - Seasons
  - Game
- Prediction tasks:
  - Game result
  - All-NBA

The raw data

Raw Play-By-Play base data set

- 19 seasons (2000-01 – 2018-19), data split by season
- 10 389 755 plays/events (shot, foul, turnover, …)
- 35 columns
- nominal, continuous, time series

	A
1	,EVENTMSGACTIONTYPE,EVENTMSGTYPE,EVENTNUM,GAME_ID,HOMEDESCRIPTION,NEUTRALDESCRIPTION,PCTIMESTRING,PERIOD,PERSON1TYPE,PERSON2TYPE,PERSON3TYPE,PLAYER1_ID,PLAYER1_NAME,PLAYER1_TEAM_AE
2	0,0,12,0,0021600229,,,12:00,1,0.0,0,0,0,,,,,,0,,,,,,0,,,,,,,,,7:41 PM
3	1,0,10,1,0021600229,Jump Ball Okafor vs. Lopez: Tip to Gibson,,12:00,1,4.0,5,5,1626143,Jahlil Okafor,PHI,Philadelphia,1610612755.0,76ers,201577,Robin Lopez,CHI,Chicago,1610612741.0,Bulls,201959,Taj Gibson,CHI,Chicago,161(
4	2,57,1,2,0021600229,,,11:39,1,5.0,0,0,201577,Robin Lopez,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,2 - 0,-2,Lopez 12' Driving Hook Shot (2 PTS),7:41 PM
5	3,1,5,3,0021600229,Bayless Bad Pass Turnover (P1.T1),,11:28,1,4.0,5,0,201573,Jerryd Bayless,PHI,Philadelphia,1610612755.0,76ers,2548,Dwyane Wade,CHI,Chicago,1610612741.0,Bulls,0,,,,,,,,Wade STEAL (1 STL),7:42 PM
6	4,41,1,4,0021600229,,,11:25,1,5.0,5,0,2548,Dwyane Wade,CHI,Chicago,1610612741.0,Bulls,200765,Rajon Rondo,CHI,Chicago,1610612741.0,Bulls,0,,,,,,4 - 0,-4,Wade 1' Running Layup (2 PTS) (Rondo 1 AST),7:42 PM
7	5,1,2,5,0021600229,MISS Covington 26' 3PT Jump Shot,,11:12,1,4.0,0,0,203496,Robert Covington,PHI,Philadelphia,1610612755.0,76ers,0,,,,,,0,,,,,,,,,7:42 PM
8	6,0,4,6,0021600229,,,11:10,1,5.0,0,0,201959,Taj Gibson,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,,,Gibson REBOUND (Off:0 Def:1),7:42 PM
9	7,4,6,8,0021600229,,,11:02,1,5.0,4,1,201577,Robin Lopez,CHI,Chicago,1610612741.0,Bulls,1626143,Jahlil Okafor,PHI,Philadelphia,1610612755.0,76ers,0,,,,,,,,Lopez OFF.Foul (P1) (T.Brothers),7:42 PM
10	8,5,5,9,0021600229,,,11:02,1,5.0,0,1,201577,Robin Lopez,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,,,Lopez Foul Turnover (P1.T1),7:42 PM
11	9,1,2,10,0021600229,MISS Covington 25' 3PT Jump Shot,,10:42,1,4.0,0,0,203496,Robert Covington,PHI,Philadelphia,1610612755.0,76ers,0,,,,,,0,,,,,,,,,7:43 PM
12	10,0,4,11,0021600229,,,10:40,1,5.0,0,0,200765,Rajon Rondo,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,,,Rondo REBOUND (Off:0 Def:1),7:43 PM
13	11,46,1,12,0021600229,,,10:35,1,5.0,5,0,202710,Jimmy Butler,CHI,Chicago,1610612741.0,Bulls,200765,Rajon Rondo,CHI,Chicago,1610612741.0,Bulls,0,,,,,,6 - 0,-6,Butler 18' Running Jump Shot (2 PTS) (Rondo 2 AST),7:43 PM
14	12,80,2,13,0021600229,MISS Ilyasova 17' Step Back Jump Shot,,10:12,1,4.0,0,0,101141,Ersan Ilyasova,PHI,Philadelphia,1610612755.0,76ers,0,,,,,,0,,,,,,,,,7:43 PM
15	13,0,4,14,0021600229,,,10:09,1,5.0,0,0,200765,Rajon Rondo,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,,,Rondo REBOUND (Off:0 Def:2),7:43 PM
16	14,39,5,15,0021600229,,,10:01,1,5.0,0,0,201959,Taj Gibson,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,,,Gibson Step Out of Bounds Turnover (P1.T2),7:43 PM
17	15,1,2,17,0021600229,MISS Bayless 14' Jump Shot,,9:48,1,4.0,0,0,201573,Jerryd Bayless,PHI,Philadelphia,1610612755.0,76ers,0,,,,,,0,,,,,,,,,7:44 PM
18	16,0,4,18,0021600229,,,9:46,1,5.0,0,0,2548,Dwyane Wade,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,,,Wade REBOUND (Off:0 Def:1),7:44 PM
19	17,80,1,19,0021600229,,,9:38,1,5.0,0,0,2548,Dwyane Wade,CHI,Chicago,1610612741.0,Bulls,0,,,,,,0,,,,,,8 - 0,-8,Wade 16' Step Back Jump Shot (4 PTS),7:44 PM
20	18,1,9,20,0021600229,76ERS Timeout: Regular (Full 1 Short 0),,9:37,1,2.0,0,0,1610612755,,,,,,0,,,,,,0,,,,,,,,,7:44 PM

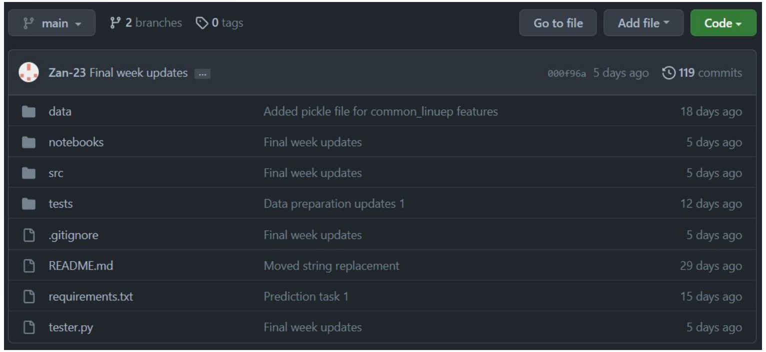
Tools

- Main workhorse
  - Python 3 + jupyter notebooks

- Data acquisition
  - beautiful soup, selenium, tdqm, webdriver-manager

- Data analysis and prediction
  - numpy, pandas, plotly, sklearn, pickle

- Version control
  - GitHub



Data loader pipeline

Base data set of 1.6 GB

- long time to load
- especially with extra processing

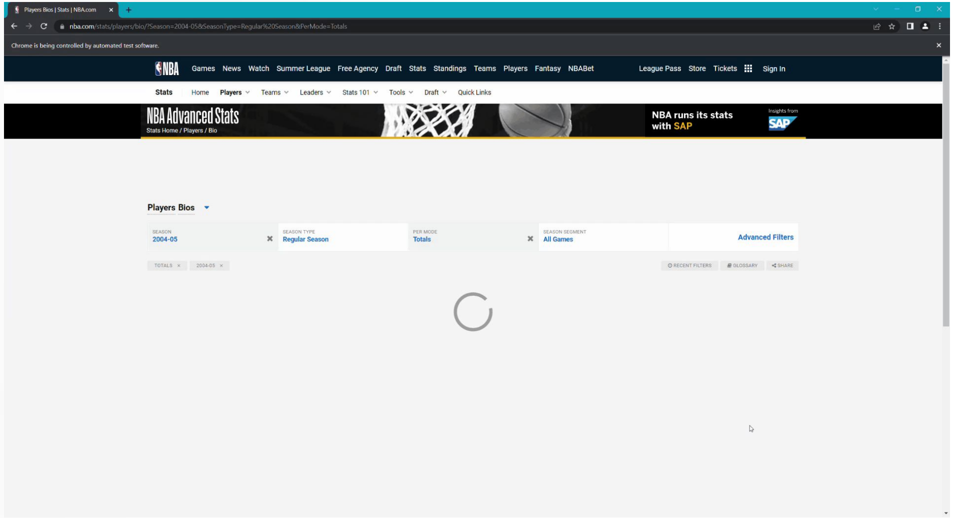
Solution:

- multiple data loaders created
- special parameter for loading
- intense use of pandas masks
- storing computed results in intermediate pickle files
- ~1650 lines of code
- data loading optimized from minutes to seconds or milliseconds

Player data

When searching for player data no complete dataset were found

1. We created our own data scraper
2. Used selenium and chrome driver to simulate browser usage since static loading failed
3. Now it can retrieve all player data available on the NBA website

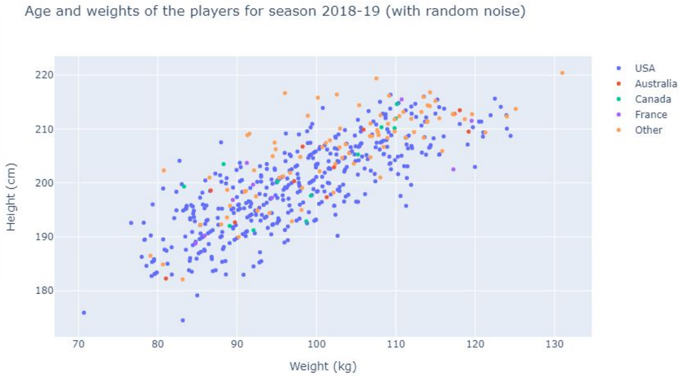


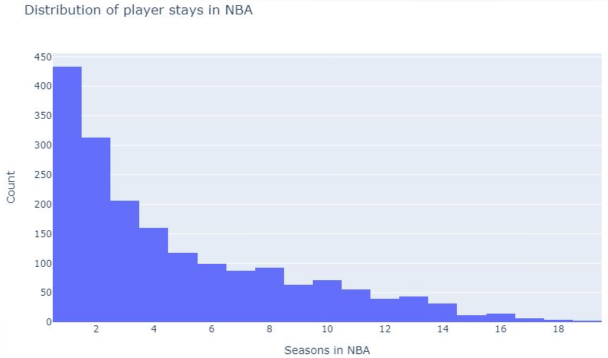
Exploring player characteristics

We managed to learn some common characteristics of NBA players

Average statistics about players through all seasons

	Age (years)	Height (cm)	Weight (kg)
Mean	27	200	100
Standard deviation	0.35	0.26	0.94

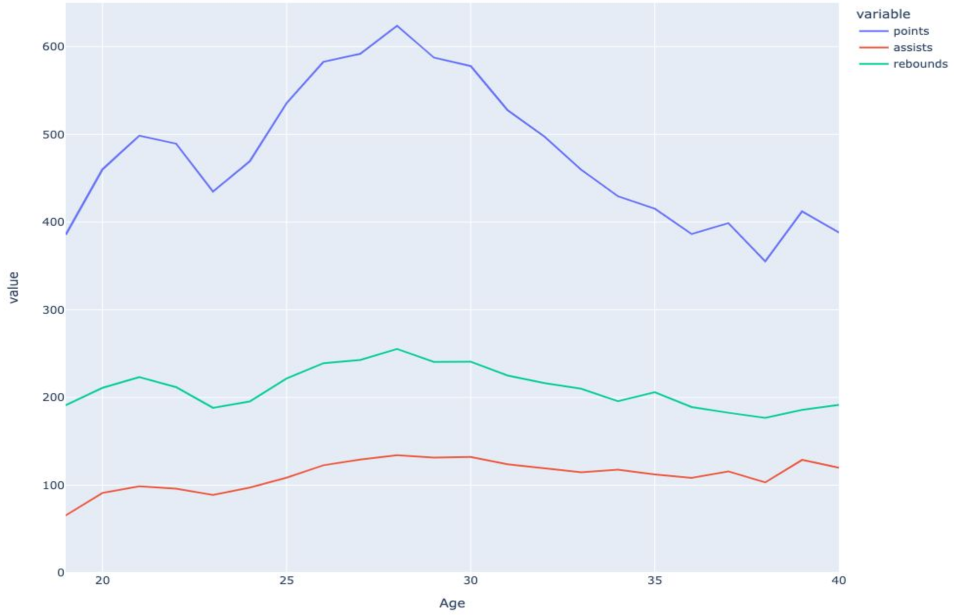




Combining age with extracted performance

Answering questions that our raw data cannot directly answer:

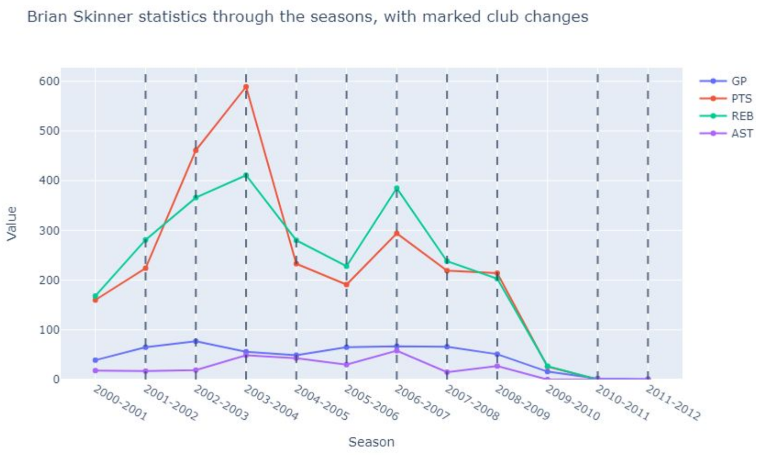
Where is the peak point of players performance during their careers?

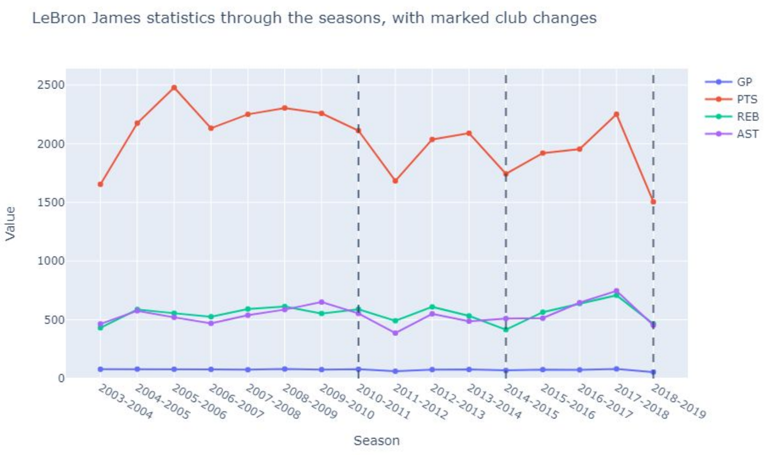


Diving into individual players

Impact of variables on player performance

- club changed negatively impacted player performance (points)
- problems because of incomplete data

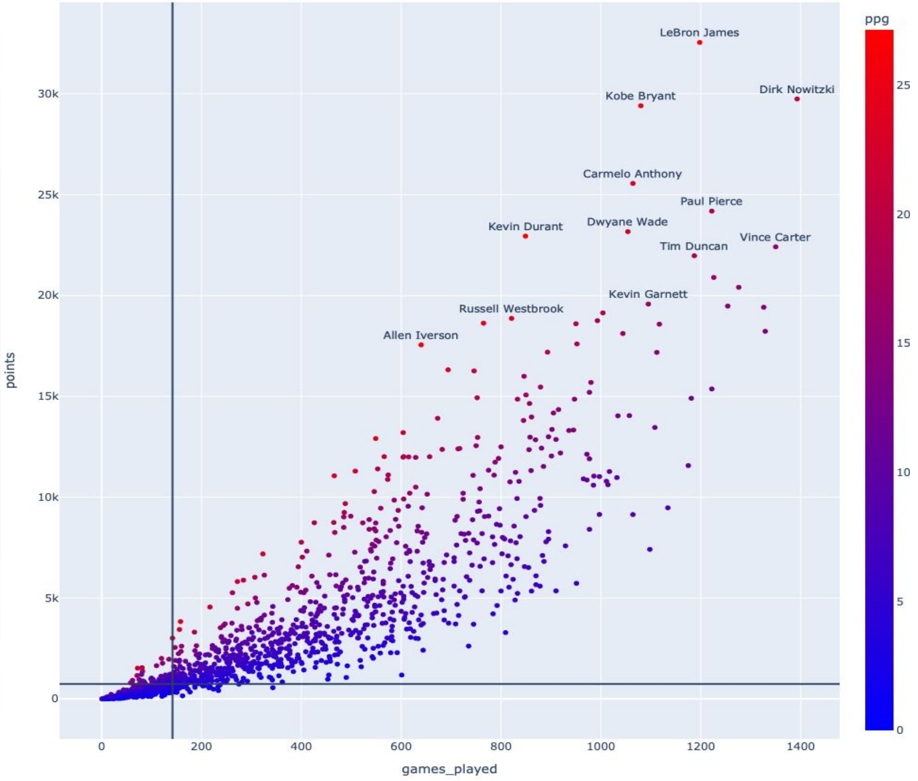




Star players dominate the statistics

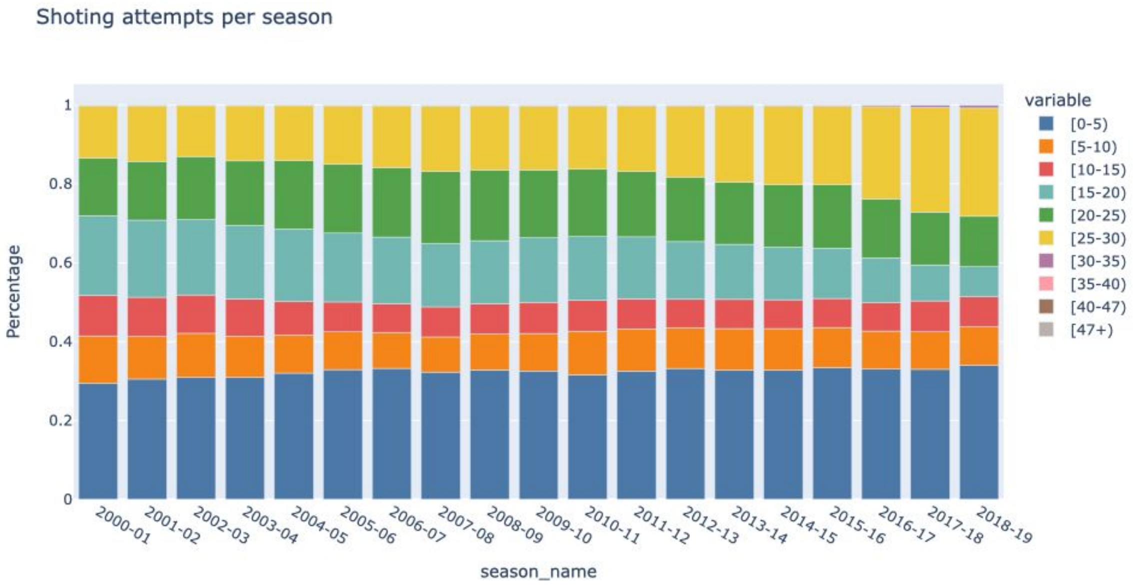
Few players played many games or scored a lot of points

- The median number of points scored is 742
- The median number of games played is 142



Distribution of shot distances per season

How game changed over years or what has changed in the game?

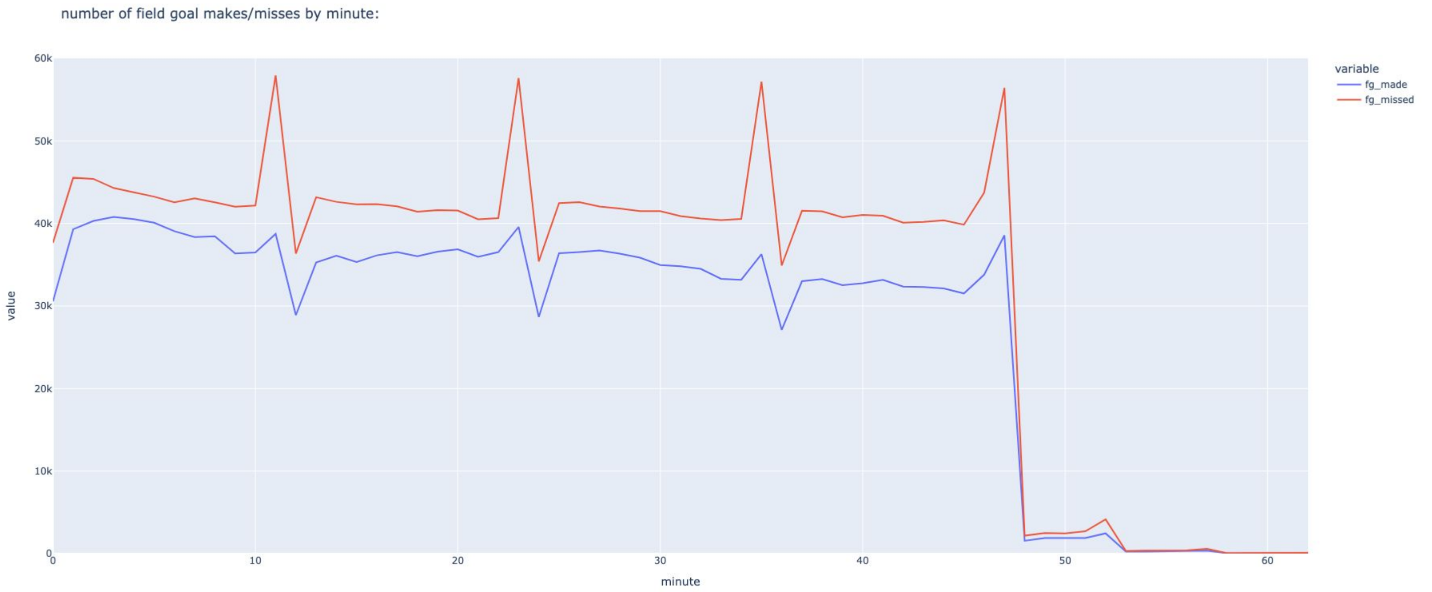


Shift in play-style confirmed by data

Goodbye to mid-range shots!!

Analysis of games

- Large corpus of 22,965 games (every game in 19 consecutive regular seasons)
- Extract the essence of an NBA game -> Contrast with domain knowledge
- Heartbeat of a game:



Using data to answer questions

- Why do players shoot worse at the end of close games?
  - Defense vs. Pressure/Fatigue
- Use shot types as indicator:

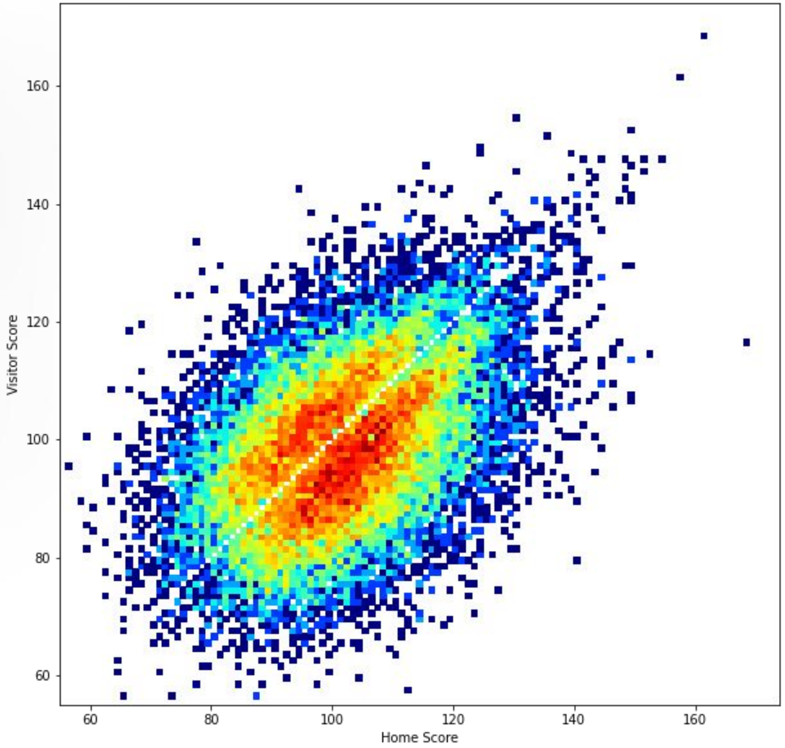
Field goal accuracy

Free throw accuracy

Shows power of combining data analysis with domain knowledge

Final game score/result

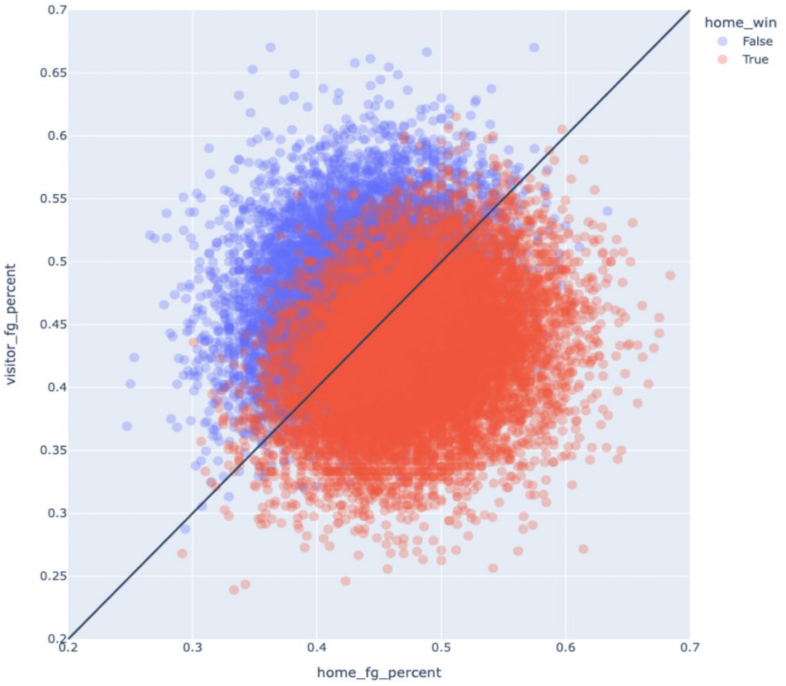
- Most important/interesting feature of a game
- Product of an artificial competition
- Unique distribution:



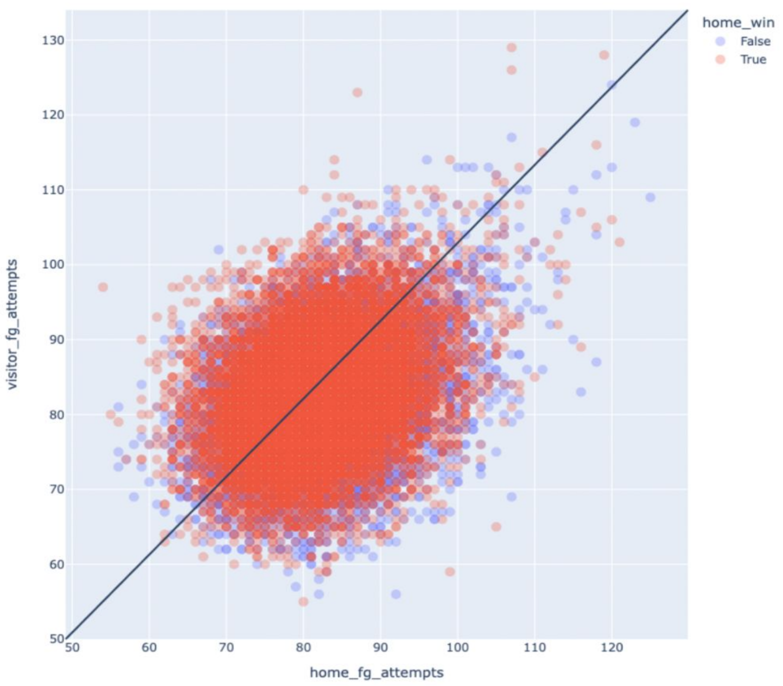
- What impacts winning?
- Can we predict a game's result?

Finding features correlated with wins

Some do (Field-goal accuracy)

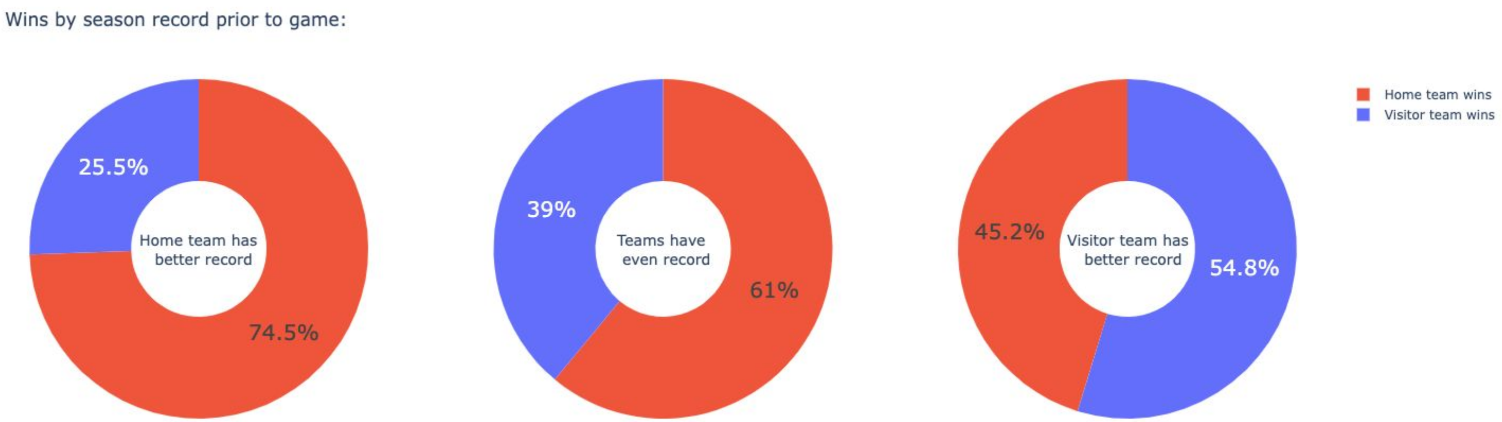


Some don't (Field goal volume)



Putting games into temporal context

- Order games by date in time -> Series of games
- Past performance is indicative of future success:



➔ Central for game result prediction

Predictive mining

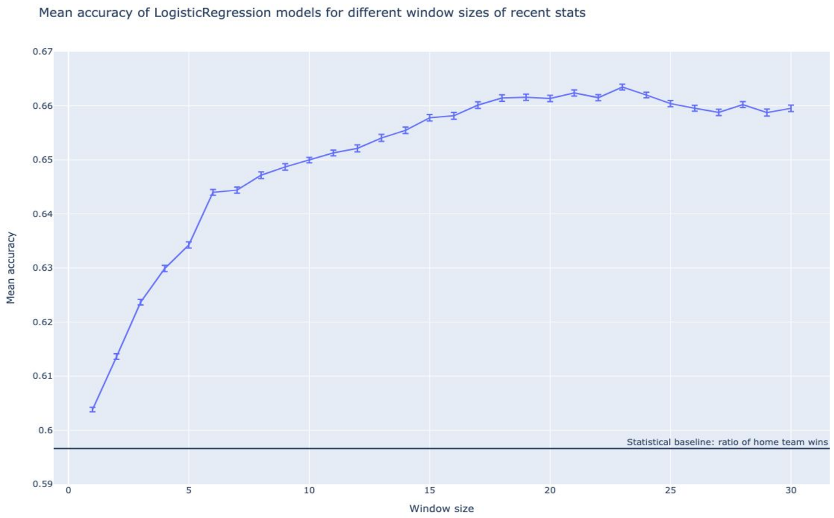
1. **Winner of the game**
   - game level
   - sports betting?

2. **All-NBA team**
   - season level
   - justifying the journalists choices / who deserved the reward?
   - find minimal production stats for achieving All-NBA

- Importance of data loader pipeline
  - easy to add new features

Winner of the game - overview

- Statistical baseline
  - 60 % games won by home team
- Window size (rolling averages)
- Unbalanced data
- Scaled features
- Labeled team IDs
- Hyperparameter search
  - regularization, kernel, max depth
- Feature selection
  - manual selection, RFE, select k best, fastener



Winner of the game - comparison of models

Random forest

Accuracy	Precision	Recall	f1-score
0.65	0.64	0.62	0.61

SVM

Accuracy	Precision	Recall	f1-score
0.66	0.64	0.64	0.64

SVM hyperparameters

Parameter	Value
C	24.8
kernel	"linear"
decision_function_shape	"ovo"
random_state	0

Gradient boosting classifier

Accuracy	Precision	Recall	f1-score
0.65	0.64	0.63	0.64

Logistic regression

Accuracy	Precision	Recall	f1-score
0.65	0.65	0.65	0.64

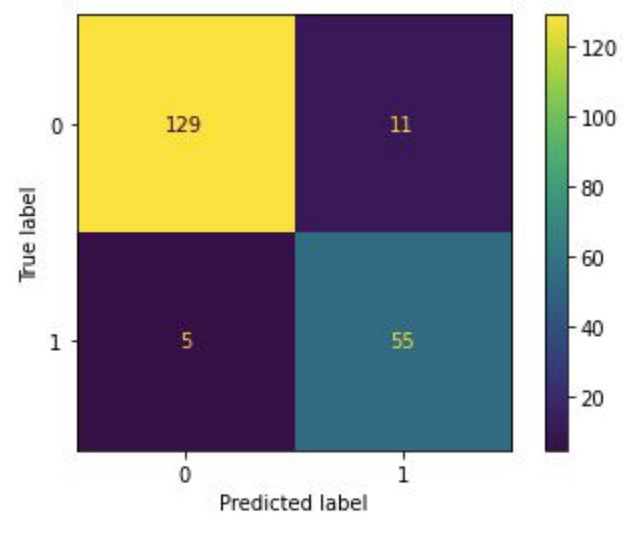
All-NBA team - overview

- 15 players selected by the journalists
- Closely related to previously extracted features/analysis
- Scraped additional data
- Filtering by number of games played (145 players per season)
- Very unbalanced –> needed balancing
- Logistic regression

1. First 15 seasons as training set, predict last 4 seasons
2. Dataset is shuffled, seasons are not relevant

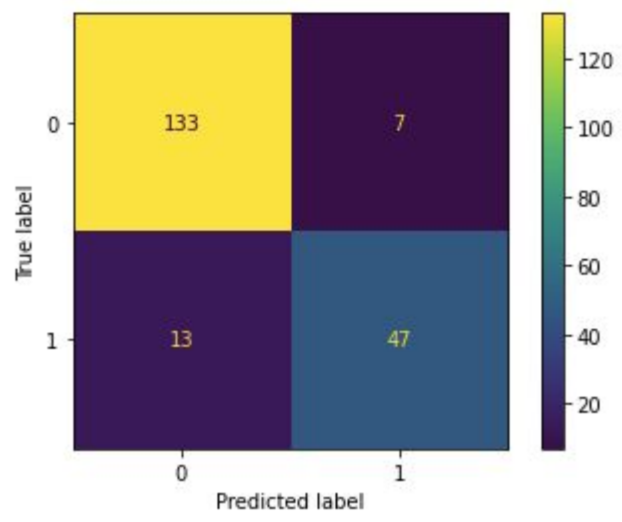
All-NBA team - evaluation of model

Chronological data



Accuracy	Precision	Recall	f1-score
0.92	0.83	0.92	0.87

Shuffled data



Accuracy	Precision	Recall	f1-score
0.90	0.87	0.78	0.82

Conclusion

- Play-By-Play data allows for multiple views on data
  - Extracted many features on different levels
  - Per game, player, season …
- Strong data-loading pipeline crucial for efficient working in a team
- Combining previous domain knowledge with data analysis allowed us to answer interesting questions in a data-driven manner
- Putting games into order reveals temporal dependencies
- Prediction of game result / All-NBA is possible

DANKESCHÖN, HVALA, TEŞEKKÜRLER