Predicting Game Winners

Javier Herbas, Aug 2020

Introduction & Objectives

- Predict NBA game winners based on Home Court Advantage (Home Team)

How?

- Evaluating different ML Classification Algorithms and shortlisting a set of stats that could translate into wins

Questions raised…

- ~70 years of NBA data available
- Over 100 stats per team are available
- The game has evolve through the years

- How far behind should the models go?
- Which combinations could translate into wins?
- Can we get any insight from this evolution to help us predict the winners?

Data Gathering and Conditioning

Data sources:

- https://stats.nba.com/teams/
- https://www.espn.co.uk/nba/stats/_/view/team
- https://www.basketball-reference.com
- https://watchstadium.com
- https://www.si.com
- https://www.basketball-reference.com/

- 1 year data (2018 – 2019)
- 10 year data (2010 – 2020)
- 20 year data (2000 – 2020)
- 30 year data (1990 - 2020

Target:

	Team1	Team1Score	Team2	Team2Score	year	month	Game_Result
0	Charlotte Hornets	106	Atlanta Hawks	82	2000	10	0
1	Cleveland Cavaliers	86	New Jersey Nets	82	2000	10	0

Team1 = Visiting
Team2 = Home
Game Result = 0 >>> Game Lost by Home Team
Game Result = 1 >>> Game Won by Home Team

	TEAM	year	month	PTS	FGM	FGA	FG_P	3PM	3PA	3P_P
0	Atlanta Hawks	2010	10	107.3	37.3	77.0	48.5	6.3	17.3	36.5
1	Los Angeles Lakers	2010	10	111.0	41.7	93.0	44.8	9.0	22.3	40.3

	TEAM	year	month	GP	W	L	MIN	OFFRTG	DEFRTG	NETRTG	AST_P	AST/TO	ASTRATIO	OREB_P	DREB_P	REB_P	TOV_P	EFG_P	TS_P	PACE
0	Miami Heat	2010	10	4	3	1	192	103.6	90.0	13.6	58.9	1.43	15.9	26.0	69.0	49.1	14.7	49.5	54.6	90.00
1	Dallas Mavericks	2010	10	3	2	1	144	102.5	91.2	11.2	68.5	1.49	19.7	22.8	69.4	48.9	18.0	53.3	56.8	94.67

	Game_Result	Team1_OFFRTG	Team1_DEFRTG	Team1_AST/TO	Team1_REB_P	Team1_FG_P	Team1_FGA	Team1_PACE	Team1_3PA	Team1_3P_P
0	1	106.3	107.5	1.83	51.1	43.5	93.1	104.39	35.9	33.8
1	1	102.4	105.0	1.30	50.8	42.8	93.8	106.08	29.7	27.5
2	0	113.2	98.2	1.57	53.4	48.0	91.4	106.00	40.6	38.4
3	1	108.0	111.4	1.46	48.7	45.4	87.5	99.15	35.8	38.1
4	1	104.9	101.5	1.73	47.9	43.5	82.7	98.67	29.0	36.8

Data Gathering and Conditioning

Data sources:

- https://stats.nba.com/teams/
- https://www.espn.co.uk/nba/stats/_/view/team
- https://www.basketball-reference.com
- https://watchstadium.com
- https://www.si.com
- https://www.basketball-reference.com/

- 1 year data (2018 – 2019)
- 10 year data (2010 – 2020)
- 20 year data (2000 – 2020)
- 30 year data (1990 - 2020

Target:

	Team1	Team1Score	Team2	Team2Score	year	month	Game_Result
0	Charlotte Hornets	106	Atlanta Hawks	82	2000	10	0
1	Cleveland Cavaliers	86	New Jersey Nets	82	2000	10	0

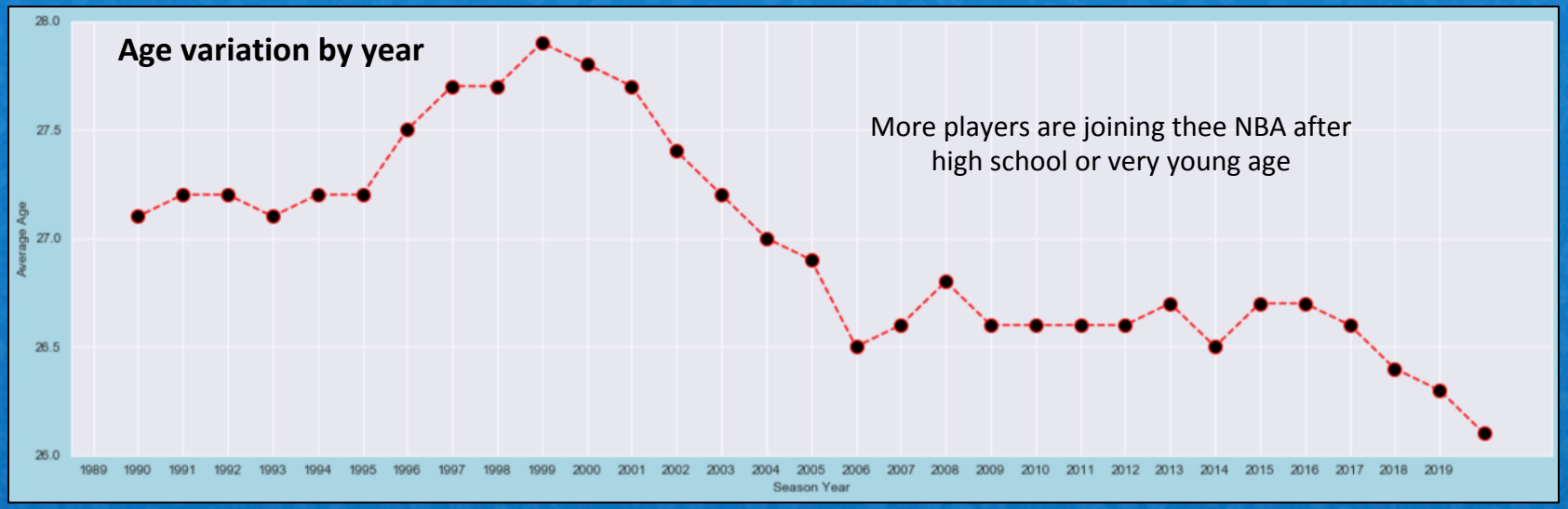
Team1 = Visiting
Team2 = Home
Game Result = 0 >>> Game Lost by Home Team
Game Result = 1 >>> Game Won by Home Team

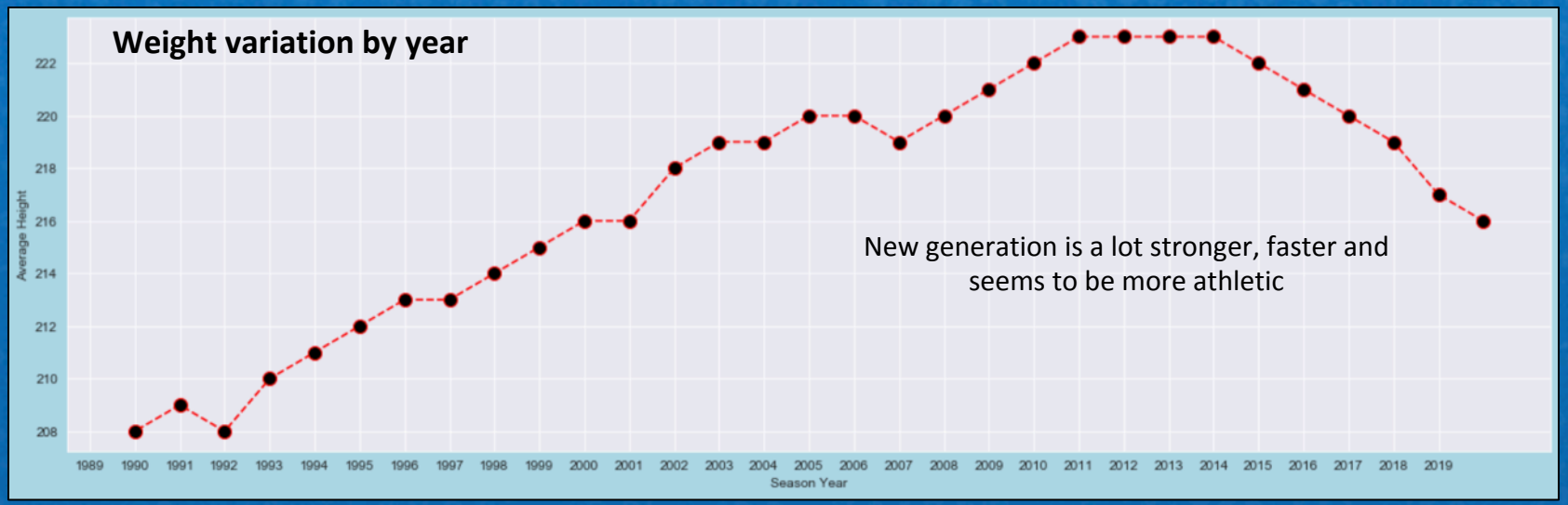
	TEAM	year	month	PTS	FGM	FGA	FG_P	3PM	3PA	3P_P
0	Atlanta Hawks	2010	10	107.3	37.3	77.0	48.5	6.3	17.3	36.5
1	Los Angeles Lakers	2010	10	111.0	41.7	93.0	44.8	9.0	22.3	40.3

	TEAM	year	month	GP	W	L	MIN	OFFRTG	DEFRTG	NETRTG	AST_P	AST/TO	ASTRATIO	OREB_P	DREB_P	REB_P	TOV_P	EFG_P	TS_P	PACE
0	Miami Heat	2010	10	4	3	1	192	103.6	90.0	13.6	58.9	1.43	15.9	26.0	69.0	49.1	14.7	49.5	54.6	90.00
1	Dallas Mavericks	2010	10	3	2	1	144	102.5	91.2	11.2	68.5	1.49	19.7	22.8	69.4	48.9	18.0	53.3	56.8	94.67

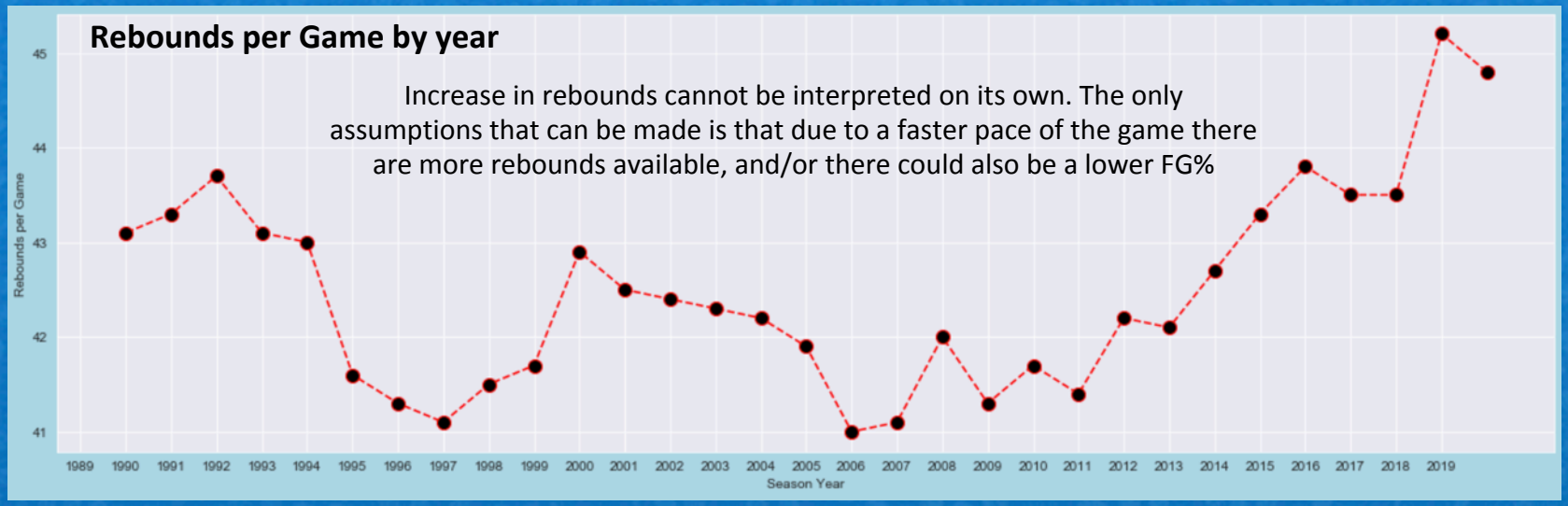
	Game_Result	Team1_OFFRTG	Team1_DEFRTG	Team1_AST/TO	Team1_REB_P	Team1_FG_P	Team1_FGA	Team1_PACE	Team1_3PA	Team1_3P_P
0	1	106.3	107.5	1.83	51.1	43.5	93.1	104.39	35.9	33.8
1	1	102.4	105.0	1.30	50.8	42.8	93.8	106.08	29.7	27.5
2	0	113.2	98.2	1.57	53.4	48.0	91.4	106.00	40.6	38.4
3	1	108.0	111.4	1.46	48.7	45.4	87.5	99.15	35.8	38.1
4	1	104.9	101.5	1.73	47.9	43.5	82.7	98.67	29.0	36.8

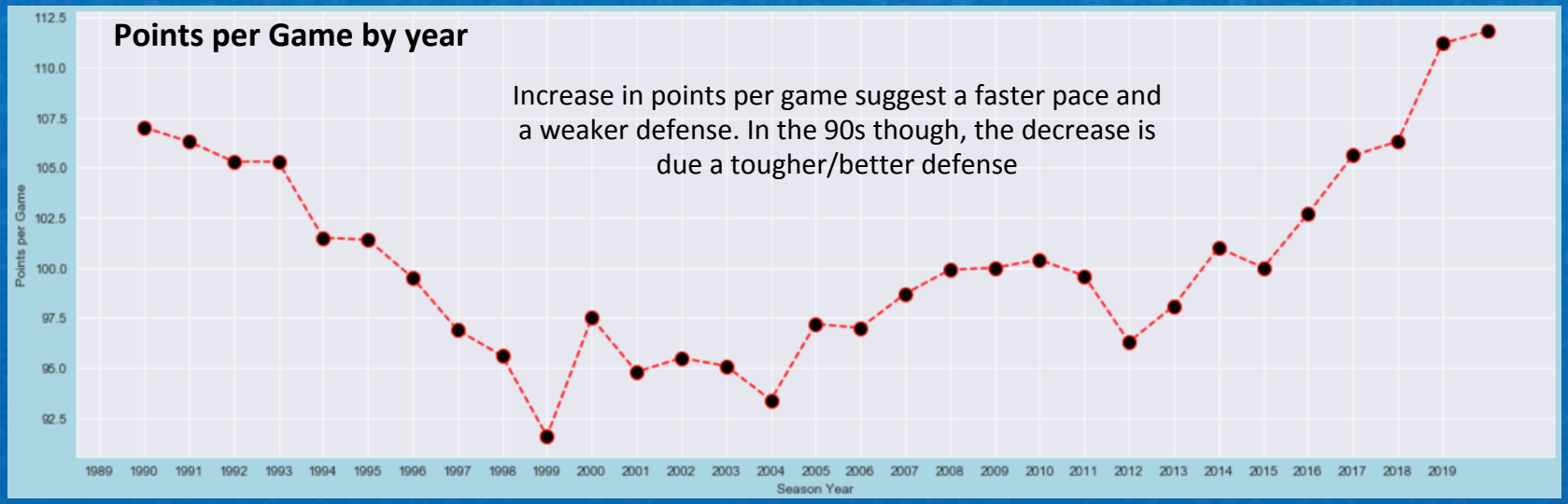
EDA – Game's evolution



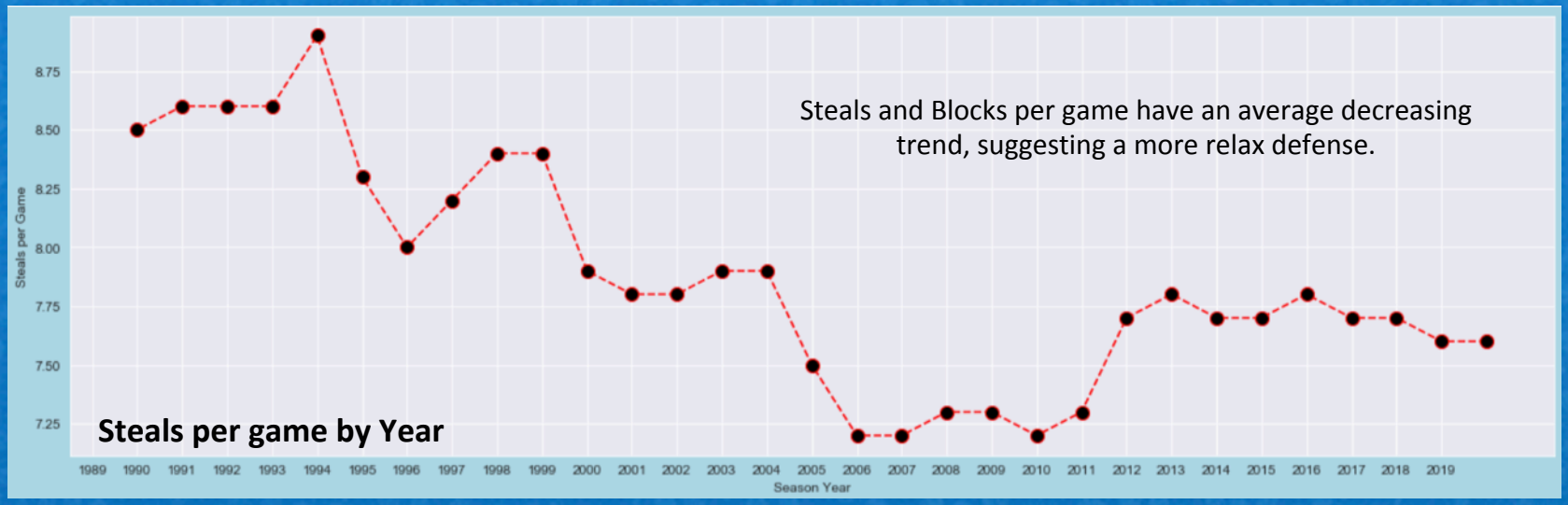


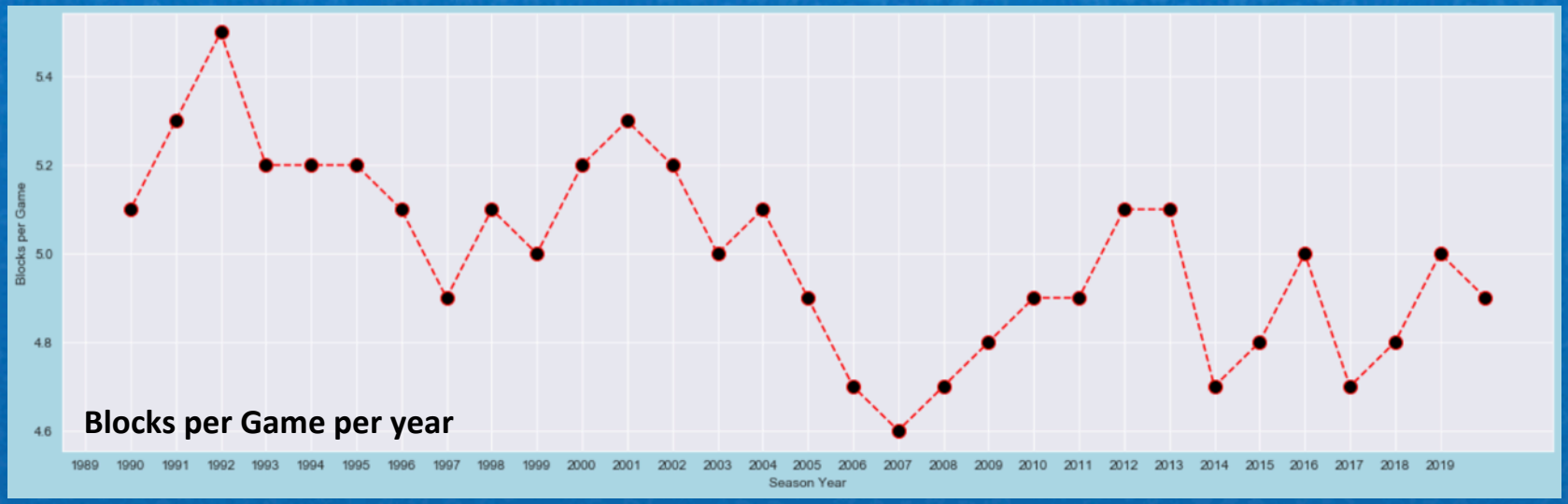
EDA – Game's Evolution





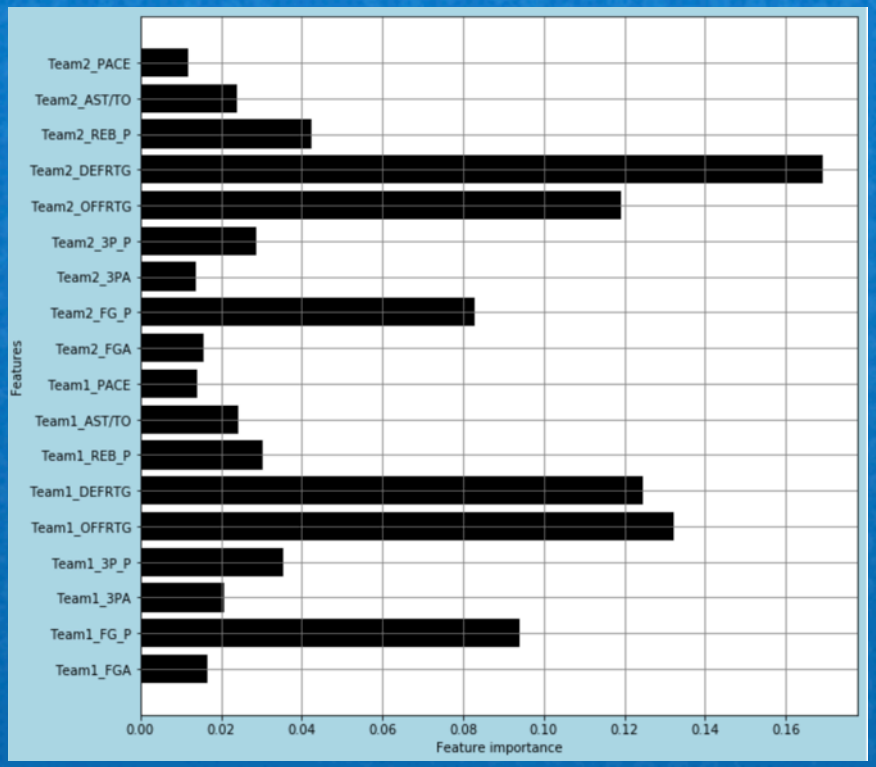
EDA – Game's Evolution



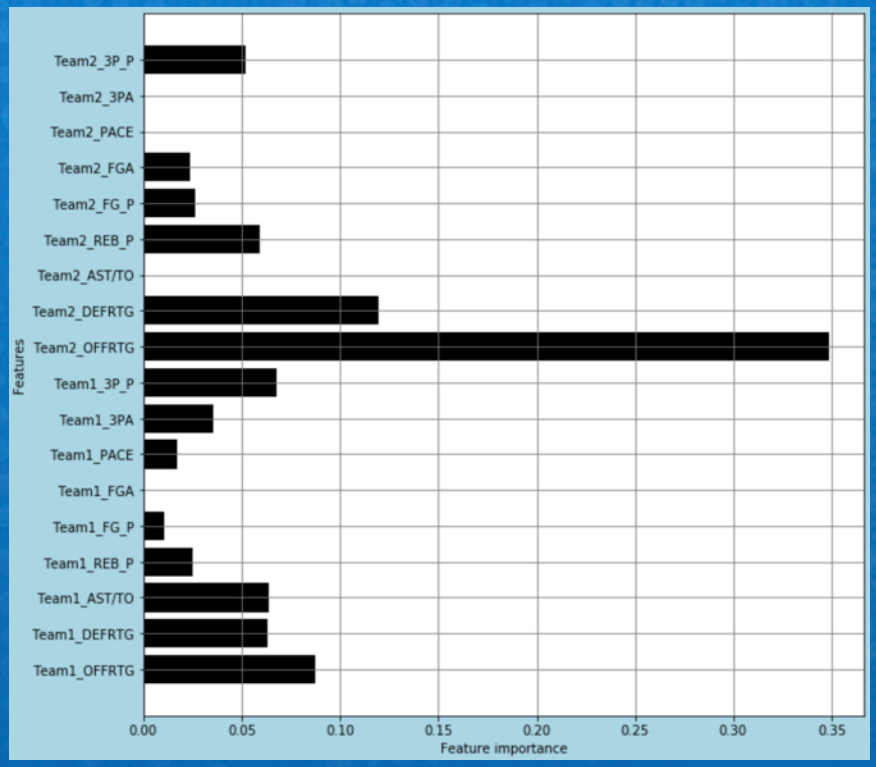


EDA – Feature Importance

2000-2010



2010-2020



Game Evolution through the past 2 decades…

Exploratory Data Analysis

Machine Learning Algorithm Tested:

- Logistic Regression Classification
- KNN Classification
- Decision Tree
- Bagged Trees
- Random Forest Classification
- SVM Linear Classification
- SVC Prediction Model
- Bagging SVC Ensemble Classification
- Adaboost Classification
- XGBoost Classification

Initial Statistical Features

Date	Game day
Team1	Visiting Team
Team1Score	Points scored by Team1
GP	Games Played
W	Wins
L	Losses
WIN_P	Win Percentage
MIN	Minutes Played
Team2	HomeTeam
Team2Score	Points scored by Team2
PTS	Points scored
FGA	Field Goals Attempted
FGM	Field Goals Made
FG_P	Field Goal Percentage
EFG_P	Effective Field Goal Percentage
3PA	3 Points Attempted
3PM	3 Points Made
3P_P	3 Point Percentage
OFFRTG	Offensive Rating
DEFRTG	Defensive Rating
NETRTG	Net Rating
OREB	Offensive Rebounds
DREB	Defensive Rebounds
REB	Rebounds
REB_P	Rebound Percentage
AST_P	Assist Percentage
AST/TO	Assit to Turnover Ratio
AST_P	Assit Percentage
Pace	Possessions per 48 min
TOV	Turnovers
FTA	Free Throws Attempted
FTM	Free Throws Made
FT_P	Free Throw Percentage
BLK	Blocks
BLKA	Blocked Field Goal Attempts
PFD	Personal Fouls Drawn

Shortlisted Features

FGA
FG_P
3PA
3P_P
OFFRTG
DEFRTG
REB_P
AST/TO
Pace

Selecting number of years for predictions

Model results – All Teams

Accuracy (%)		All Teams			
		1 year	NBA All-Teams (Average)	10 years	20 years
		2018 to 2019		2010 to 2020	2000 to 2020
		Predicted		Predicted	Predicted
Logistic Regression Class.		76.85	75%	71.57	71.75
KNN Classification		67.78		67.18	68.74
Decision Trees		61.75		61.28	62.42
Ensembles	Baseline	65.36		67.52	67.20
	Bagged Trees	66.59		70.92	69.20
	Random Forest	65.69		70.39	69.40
SVM Linear Classifier				71.88	
SVC Prediction				71.58	
Bagging SVC Ensemble Classifier				71.62	
AdaBoosst Classifier				69.35	
XGBoost Classification				70.52	

- Best predictions were using 10 years of data
- SVM Linear Classifier gave the best accuracy (71.88%)
- Bagged Trees used for real time predictions with good results (17 out of 20 matches accurately predicted)
- No model accounts for upset and matchups between players

Model Testing Results

Accuracy (%)		San Antonio Spurs					
		1 year 2018 to 2019		NBA SAS (Average)	10 years 2010 to 2020		20 years 2000 to 2020	
		True	Predicted		True	Predicted	True	Predicted
Logistic Regression Class.		78.56	72.72	81%	80.50	79.80	78.93	77.73
KNN Classification			100.00			75.96		79.83
Decision Trees			81.82			65.38		62.13
Ensembles	Baseline		67.74			72.35		72.86
	Bagged Trees		77.42			76.52		76.37
	Random Forest		74.19			78.45		78.76
Support Vector Machine								

Accuracy (%)		Miami Heat					
		1 year 2018 to 2019		NBA MIA (Average)	10 years 2010 to 2020		20 years 2000 to 2020	
		True	Predicted		True	Predicted	True	Predicted
Logistic Regression Class.		52.5	40	69%	68.7	64.07	60.91	75.00
KNN Classification			64			63.10		67.79
Decision Trees			75.00			73.53		65.25
Ensembles	Baseline		43.33			60.46		63.98
	Bagged Trees		43.33			66.34		70.62
	Random Forest		53.33			66.01		70.48
Support Vector Machine								

Real Time Results

- So far accuracy of 86.36%
- Expect the accuracy to decrease as the playoff advance
- 76ers have suffered key losses with the injury of Ben Simmons (1 of 2 top players), therefore predictions were affected

Correct Prediction
Wrong Prediction

Matchday	Matchup		Predicted	Results	
	Visiting	Home			
20/08/20	Heat	Pacers	Heat	Heat	
	Thunder	Rockets	Rockets	Rockets	
	Magic	Bucks	Bucks	Bucks	
	Portland	Lakers	Lakers	Lakers	
21/08/20	Raptors	Nets	Raptors	Raptors	
	Nuggets	Jazz	Jazz	Jazz	
	Celtics	76ers	76ers	Celtics	
	Clippers	Mavericks	Mavericks	Clippers	
22/08/20	Bucks	Magic	Bucks	Bucks	
	Heat	Pacers	Heat	Heat	
	Rockets	Thunder	Thunder	Thunder	
	Lakers	Trailblazzers	Lakers	Lakers	
23/08/20	Celtics	76res	76ers	Celtics	
	Clippers	Mavericks	Mavericks	Mavericks	
	Raptors	Nets	Raptors	Raptors	
	Nuggets	Jazz	Jazz	Jazz	
24/08/20	Bucks	Magic	Bucks	Bucks	
	Rockets	Thunder	Thunder	Thunder	
	Pacers	Heat	Heat	Heat	
	Lakers	Trailblazzers	Lakers	Lakers	
25/08/20	Jazz	Nuggets	Nuggets	Nuggets	
	Mavericks	Clippers	Clippers	Clippers	
26/08/20	Magic	Bucks	Bucks		
	Thunder	Rockets	Rockets		
	Trailblazers	Lakers	Lakers		
27/08/20	Jazz	Nuggets	Nuggets		
	Celtics	Raptors	Raptors		
	Clippers	Mavericks	Mavericks		
					86.364

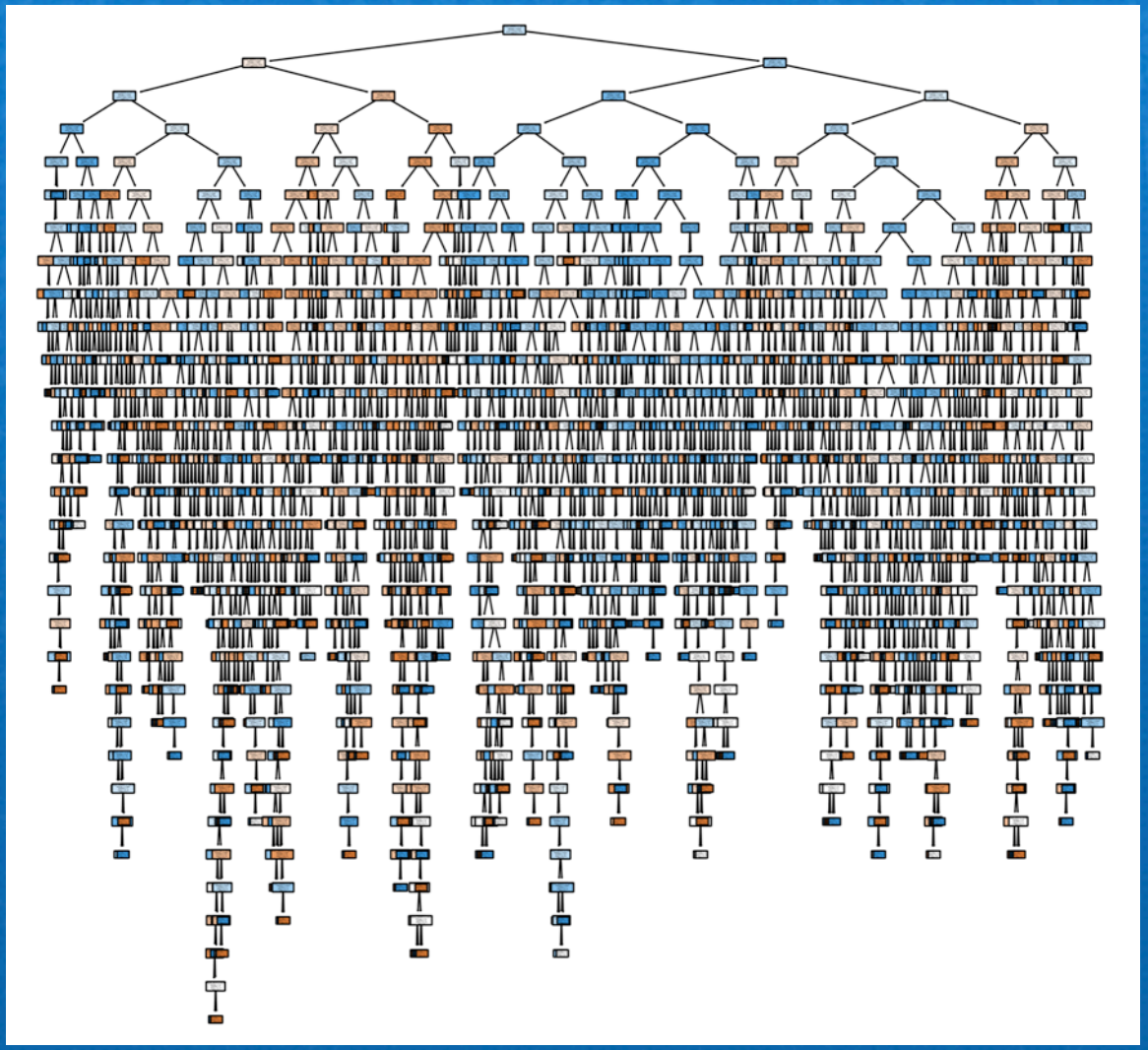
Future Work and Recommendations

- Re-run models with different year sample. Think towards 2 or 3 years
- Try to account for specific matches, meaning how a Team A plays against a particular Team B. This can be a historical stat. For example when Team A plays against Team B, they win 65% at home and 60% when visiting.
- Tiredness should be accounted for. E.g. Playing back-to-back games, or 4 tough opponents in a week.
- If any of these models are to be used during the playoffs, a factor accounting for coaching experience should be added as well as team's playoff experience
- Recently I found out that some NBA Analyst used several models for their predictions and not limit themselves to only one. Maybe that is something to think about for this project

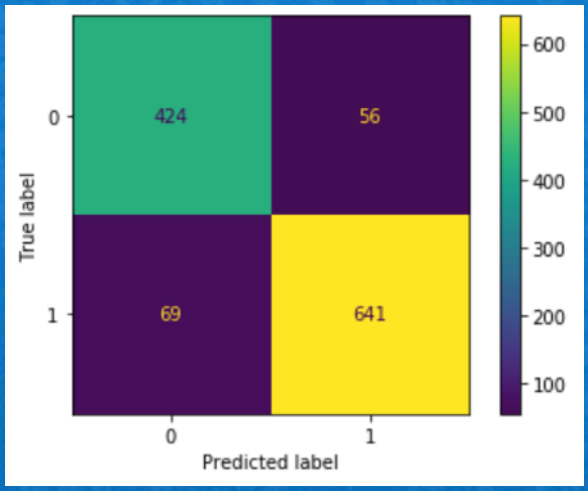

THANK YOU
Winston
Coca-Cola
Gatorade
Slamdunk Championship

Appendix

Example of one of the Decision trees:



Example of Confusion Matrix



- Confusion Matrix example comes from the 2018-2019 1 season dataset
- Decision Tree comes from the 2010 – 2020 dataset using all teams as input