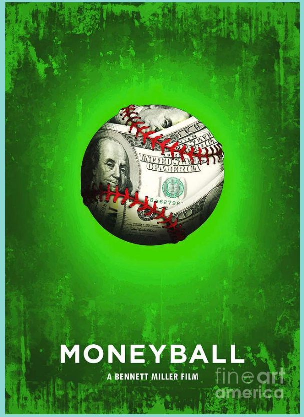
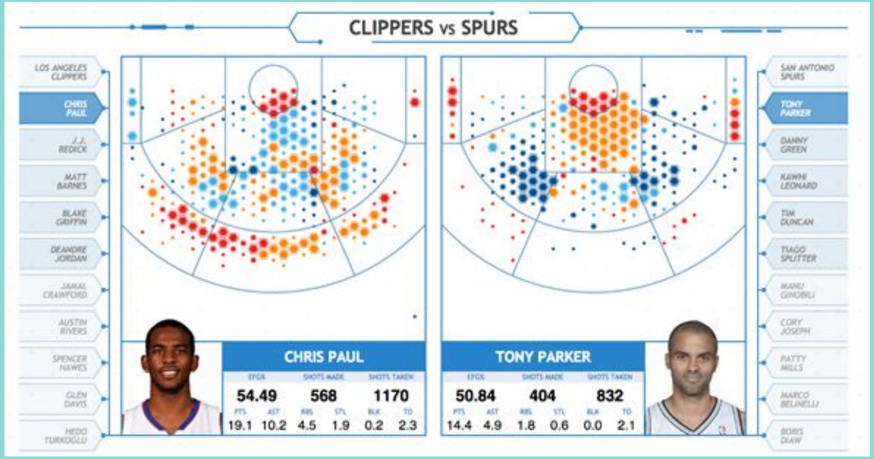
Data Mining for NBA Shot Predictions

Emma Pullen, Matthew Byam, & Matthew Spirio

Introduction- Sports & Analytics

- Analytics have played a massive role in professional sports in the last two decades
    - Moneyball (2002 Oakland Athletics)
- Allows teams to put their players in situations where statistically they are more likely to succeed
- Raises the level of competition; sports are being played at a higher level, attracting more fans to support and attend games
- Rise of Legal Sports betting→ $





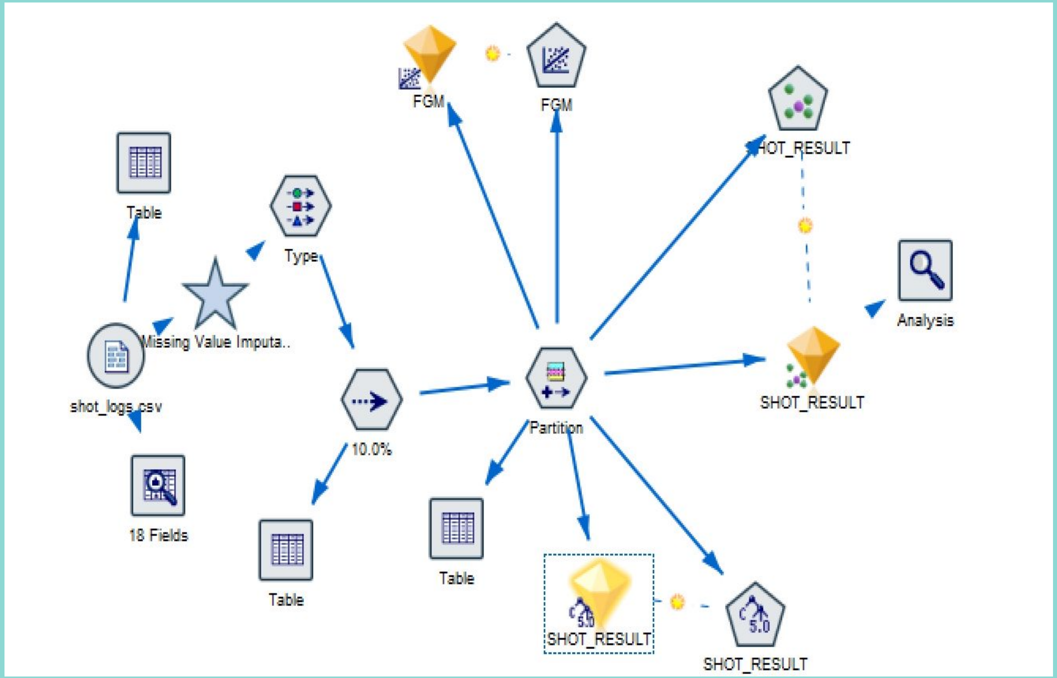
Dataset

- 128,070 records
- 2016-2017 Shot Logs
- Important Fields:
  - Shot Number
  - Shot Clock
  - Dribbles
  - Touch Time
  - Shot Distance
  - Points Type
  - Shot Result
  - Closest Defender Distance
  - FGM

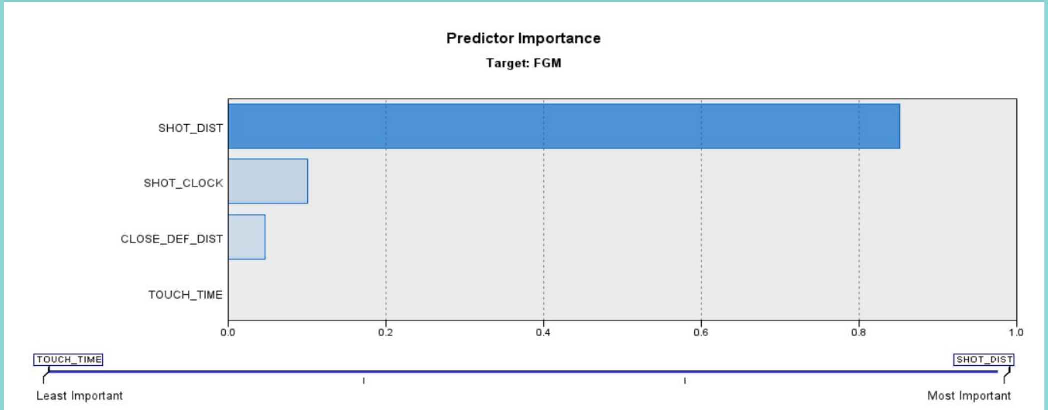
Field	Measurement	Values	Missing	Check	Role
GAME_ID	Continuous	[21400001,21400908]		None	Input
MATCHUP	Typeless			None	None
LOCATION	Flag	H/A		None	Input
W	Flag	W/L		None	Input
FINAL_MARGIN	Continuous	[-53,53]		None	Input
SHOT_NUMBER	Continuous	[1,38]		None	Input
PERIOD	Continuous	[1,7]		None	Input
GAME_CLOCK	Continuous	[00:00:00,12:00:00]		None	Input
SHOT_CLOCK	Continuous	[0.0,24.0]		None	Input
DRIBBLES	Continuous	[0,32]		None	Input
TOUCH_TIME	Continuous	<Current>		None	Input
SHOT_DIST	Continuous	[0.0,47.2]		None	Input
PTS_TYPE	Continuous	[2,3]		None	Input
SHOT_RESULT	Flag	missed/made		None	Input
CLOSEST_DEFENDER	Typeless			None	None
CLOSEST_DEFENDER_PLAYER...	Continuous	[708,530027]		None	Input
CLOSE_DEF_DIST	Continuous	[0.0,53.2]		None	Input
FGM	Continuous	[0,1]		None	Input
PTS	Continuous	[0,3]		None	Input
player_name	Typeless			None	None
player_id	Continuous	[708,204060]		None	Input

Process

- Explore Data→ Missing Values
- 10% Sample
- Partition into 70% Training & 30% Testing Data
- FGM Stepwise Linear Regression
- Shot Result KNN
- Shot Result Decision Tree



Linear Regression



- Shot Distance is leading Predictor
- Then, Shot Clock, Closest Defender Distance, and Touch Time

KNN

- 57.3% Model Accuracy
  - Seems low?
    - “Good” shooting percentage is around 50%
    - Average of top 50 shooters this season was 0.509
    - 450 players in NBA
- Performance Metrics:
  - Recall: **31.9%**
  - **Precision: 56.42%**
  - **FP Rate: 78.96%**
  - **Specificity: 21.04%**

Results for output field SHOT_RESULT

Comparing $KNN-SHOT_RESULT with SHOT_RESULT

'Partition'	1_Training		2_Testing	
Correct	5,336	59.64%	2,246	57.3%
Wrong	3,611	40.36%	1,674	42.7%
Total	8,947		3,920	

Coincidence Matrix for $KNN-SHOT_RESULT (rows show actuals)

'Partition' = 1_Training	made	missed
made	1,378	2,663
missed	948	3,958

'Partition' = 2_Testing	made	missed
made	576	1,229
missed	445	1,670

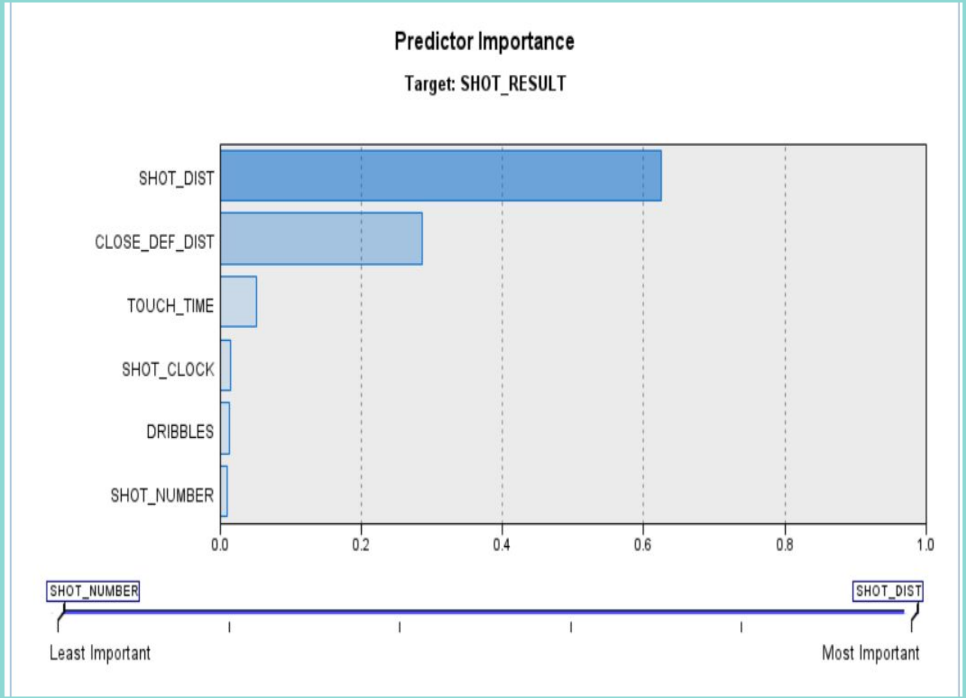
Performance Evaluation

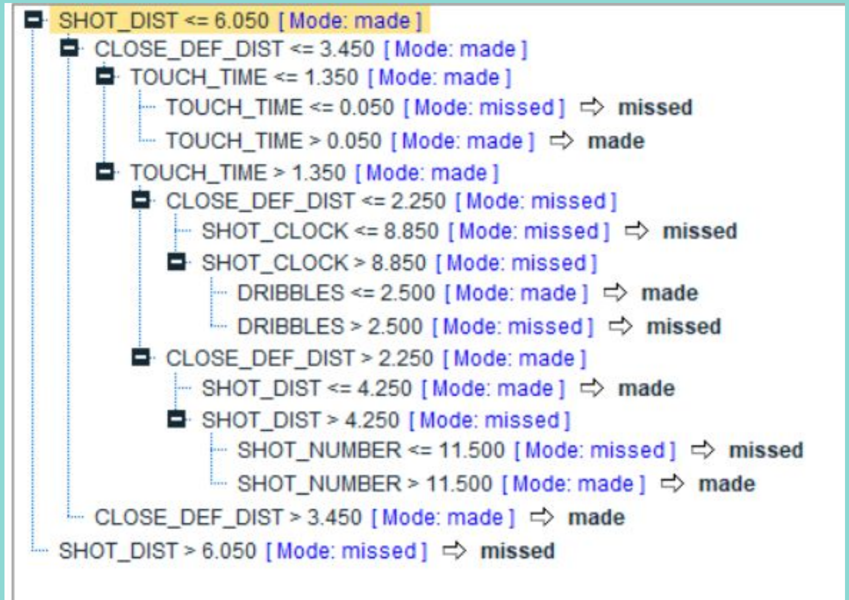
'Partition' = 1_Training	
made	0.271
missed	0.086

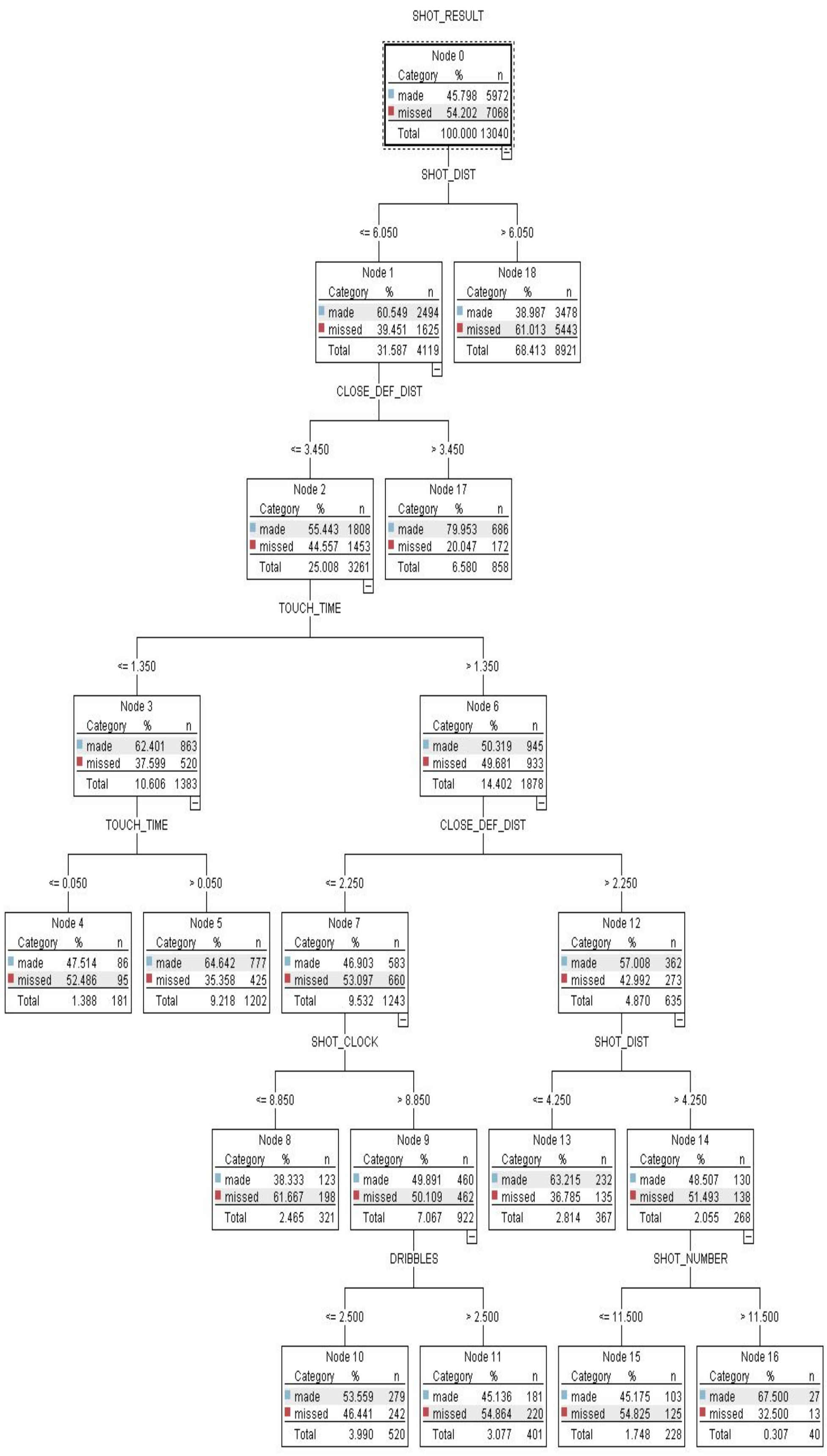
'Partition' = 2_Testing	
made	0.203
missed	0.065

Decision Tree Predictors

- Largest impact predictors
  - SHOT_DIST
  - CLOSE_DEF_DIST
- Decision Tree is pruned up to SHOT_NUMBER and Dribbles due to the rest of the predictors being insignificant






SHOT_RESULT
Node 0
Category % n
made 45.798 5972
missed 54.202 7068
Total 100.000 13040
SHOT_DIST
<= 6.050
> 6.050
Node 1
Category % n
made 60.549 2494
missed 39.451 1625
Total 31.587 4119
Node 18
Category % n
made 38.987 3478
missed 61.013 5443
Total 68.413 8921
CLOSE_DEF_DIST
<= 3.450
> 3.450
Node 2
Category % n
made 55.443 1808
missed 44.557 1453
Total 25.008 3261
Node 17
Category % n
made 79.953 686
missed 20.047 172
Total 6.580 858
TOUCH_TIME
<= 1.350
> 1.350
Node 3
Category % n
made 62.401 863
missed 37.599 520
Total 10.606 1383
Node 6
Category % n
made 50.319 945
missed 49.681 933
Total 14.402 1878
TOUCH_TIME
<= 0.050
> 0.050
CLOSE_DEF_DIST
<= 2.250
> 2.250
Node 4
Category % n
made 47.514 86
missed 52.486 95
Total 1.388 181
Node 5
Category % n
made 64.642 777
missed 35.358 425
Total 9.218 1202
Node 7
Category % n
made 46.903 583
missed 53.097 660
Total 9.532 1243
Node 12
Category % n
made 57.008 362
missed 42.992 273
Total 4.870 635
SHOT_CLOCK
<= 8.850
> 8.850
SHOT_DIST
<= 4.250
> 4.250
Node 8
Category % n
made 38.333 123
missed 61.667 198
Total 2.465 321
Node 9
Category % n
made 49.891 460
missed 50.109 462
Total 7.067 922
Node 13
Category % n
made 63.215 232
missed 36.785 135
Total 2.814 367
Node 14
Category % n
made 48.507 130
missed 51.493 138
Total 2.055 268
DRIBBLES
<= 2.500
> 2.500
SHOT_NUMBER
<= 11.500
> 11.500
Node 10
Category % n
made 53.559 279
missed 46.441 242
Total 3.990 520
Node 11
Category % n
made 45.136 181
missed 54.864 220
Total 3.077 401
Node 15
Category % n
made 45.175 103
missed 54.825 125
Total 1.748 228
Node 16
Category % n
made 67.500 27
missed 32.500 13
Total 0.307 40

Decision Tree Examples

- If shot distance ≤ 6.050 feet, touch time > 1.350 seconds, closest defender ≤ 2.250 feet, shot clock < 8.850 seconds → shot: MISSED
    - Support: 198/13040; Confidence: 198/321
- If shot distance ≤ 4.250 feet, touch time > 1.350 seconds, 2.25 < closest defender≤ 3.45 →shot: MADE
    - Support: 232/13040; Confidence: 232/367
- If shot distance ≤ 6.050 feet, closest defender > 3.450 feet away→ shot : MADE
    - Support: 686/13040; Confidence: 686/858

Decision Tree Performance Metrics

Model Accuracy: 61.05%

Recall: 36.3%

Precision: 61.3%

FP Rate: 81.2%

Specificity: 18.8%

Results for output field SHOT_RESULT

Comparing $C-SHOT_RESULT with SHOT_RESULT

'Partition'	1_Training		2_Testing	
Correct	5,551	61.67%	2,408	61.05%
Wrong	3,450	38.33%	1,536	38.95%
Total	9,001		3,944	

Coincidence Matrix for $C-SHOT_RESULT (rows show actuals)

'Partition' = 1_Training	made	missed
made	1,518	2,563
missed	887	4,033

'Partition' = 2_Testing	made	missed
made	645	1,130
missed	406	1,763

Performance Evaluation

'Partition' = 1_Training	
made	0.331
missed	0.112

'Partition' = 2_Testing	
made	0.31
missed	0.103

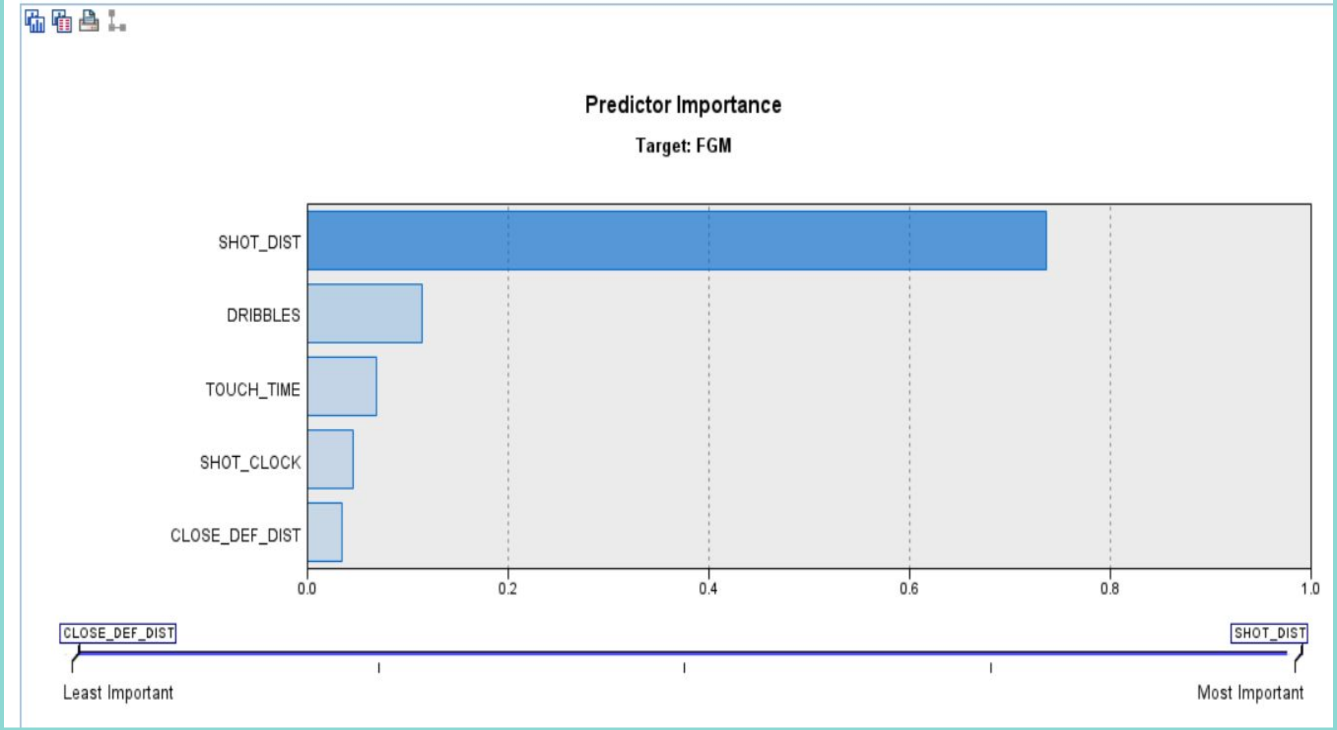
Player Comparison

- Applying analytic results to real world situations
- Comparing two successful teammates with similar skill sets
- Goal is to figure out the best way to use these players based on their shot making ability

Manu Ginobili

Tony Parker

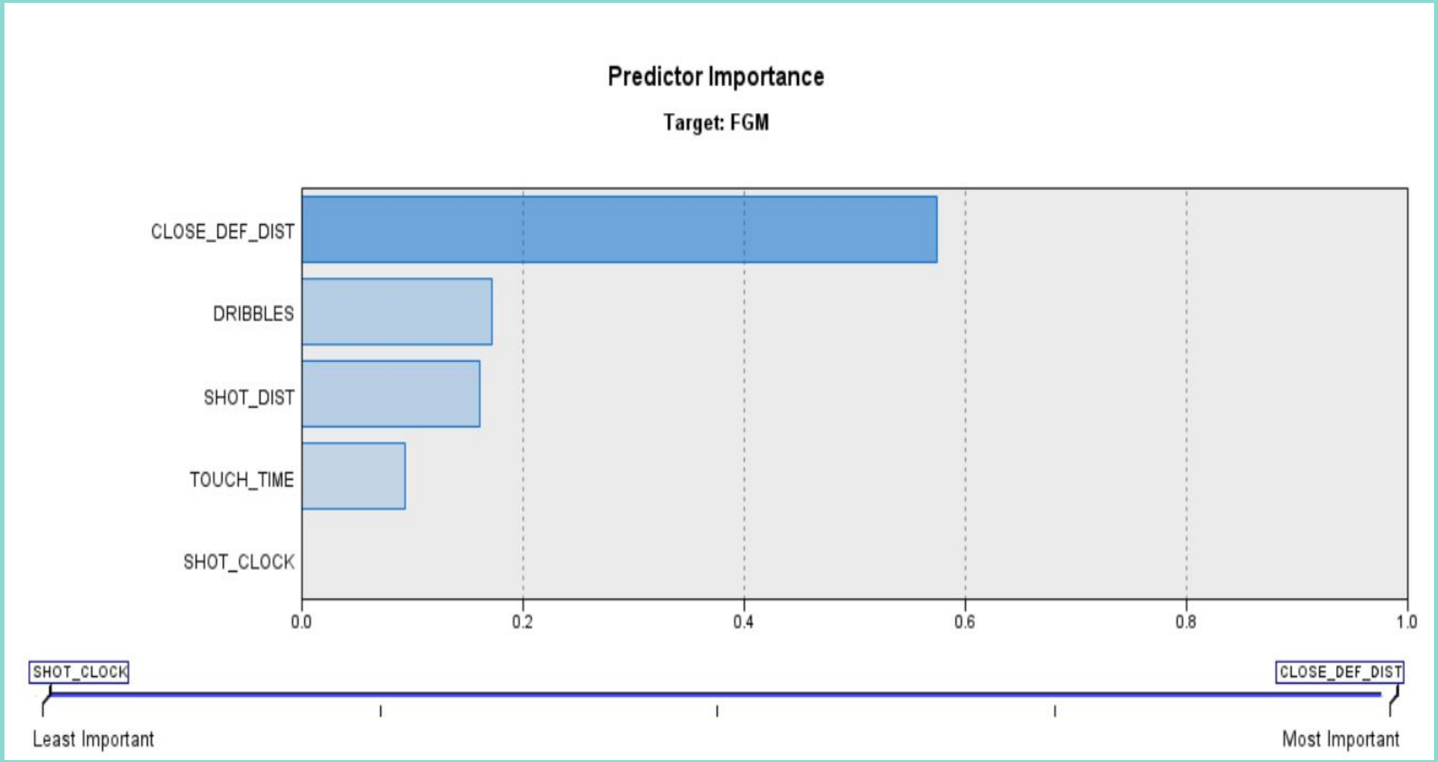
Tony Parker Predictor Importance



- Relies heavily on shot distance as an indicator
    - Closer he is, the more likely he is to make shots

- Wouldn't this be true of all players?

Manu Ginobili Predictor Importance



- Biggest indicator is closest defender distance
  - Thrives on open shots away from the defense
- Confirms that not all players are created equal

Tony Parker Linear Regression

- Analysis
  - SHOT_CLOCK * -0.00453 +
  - DRIBBLES * 0.008565 +
  - TOUCH_TIME * -0.01497 +
  - SHOT_DIST * -0.0141 +
  - CLOSE_DEF_DIST * 0.026 +
  - 0.6242
- Fields
- Build Settings
- Training Summary

Manu Ginobili Linear Regression

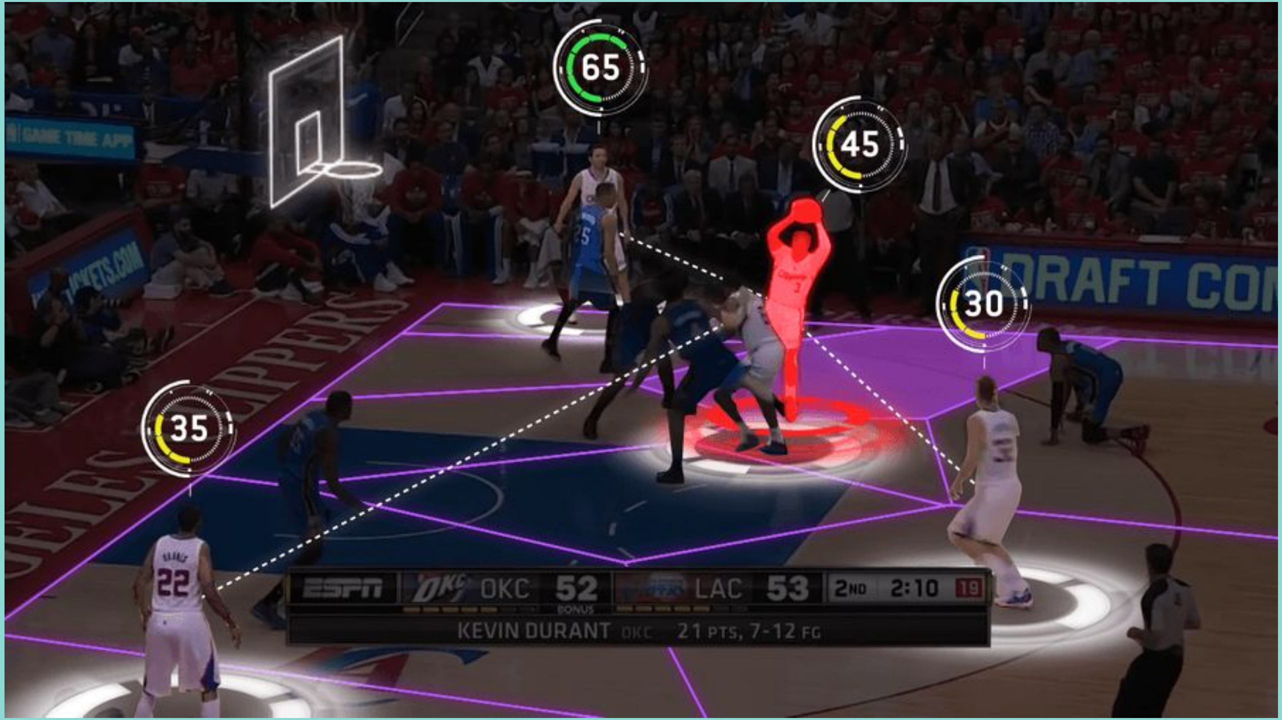
- Analysis
  - SHOT_CLOCK * -0.0001422 +
  - DRIBBLES * -0.007893 +
  - TOUCH_TIME * -0.00444 +
  - SHOT_DIST * -0.01381 +
  - CLOSE_DEF_DIST * 0.01573 +
  - 0.5931
- Fields
- Build Settings
- Training Summary

- Biggest Takeaway- Parker actually makes more shots when he takes more dribbles, the one true difference in their linear regression equations
- So what would be the best way to use these players?

Conclusion

- Shot distance, distance to closest defender, and Shot Clock are biggest influencers
- Models Created using Data Mining have a higher accuracy than most players
- The rise of Data Analytics in sports makes way for the production of curated plays
  - Both offensively and defensively → raise level game play
- Comparing individual Players performance stats tells coaches & fans their strengths and weaknesses

Thank you!!



65
45
30
35
DRAFT COM
GAME TIME APP
ESPN
OKC
52
BONUS
LAC
53
2ND
2:10
19
KEVIN DURANT OKC 21 PTS, 7-12 FG
22