Shot Results Prediction

National Basketball Association

By Amy Mykityshyn, Krystin Sinclair, Sam Kupfer, Vishnupriya Venkateswaran, Ye-in Jeon

April 25, 2018

Overview

- Background of Dataset
- Data Exploration
- Data Visualization
- Methodology
- Model
- Results
- Future Research
- Tools and References

Background

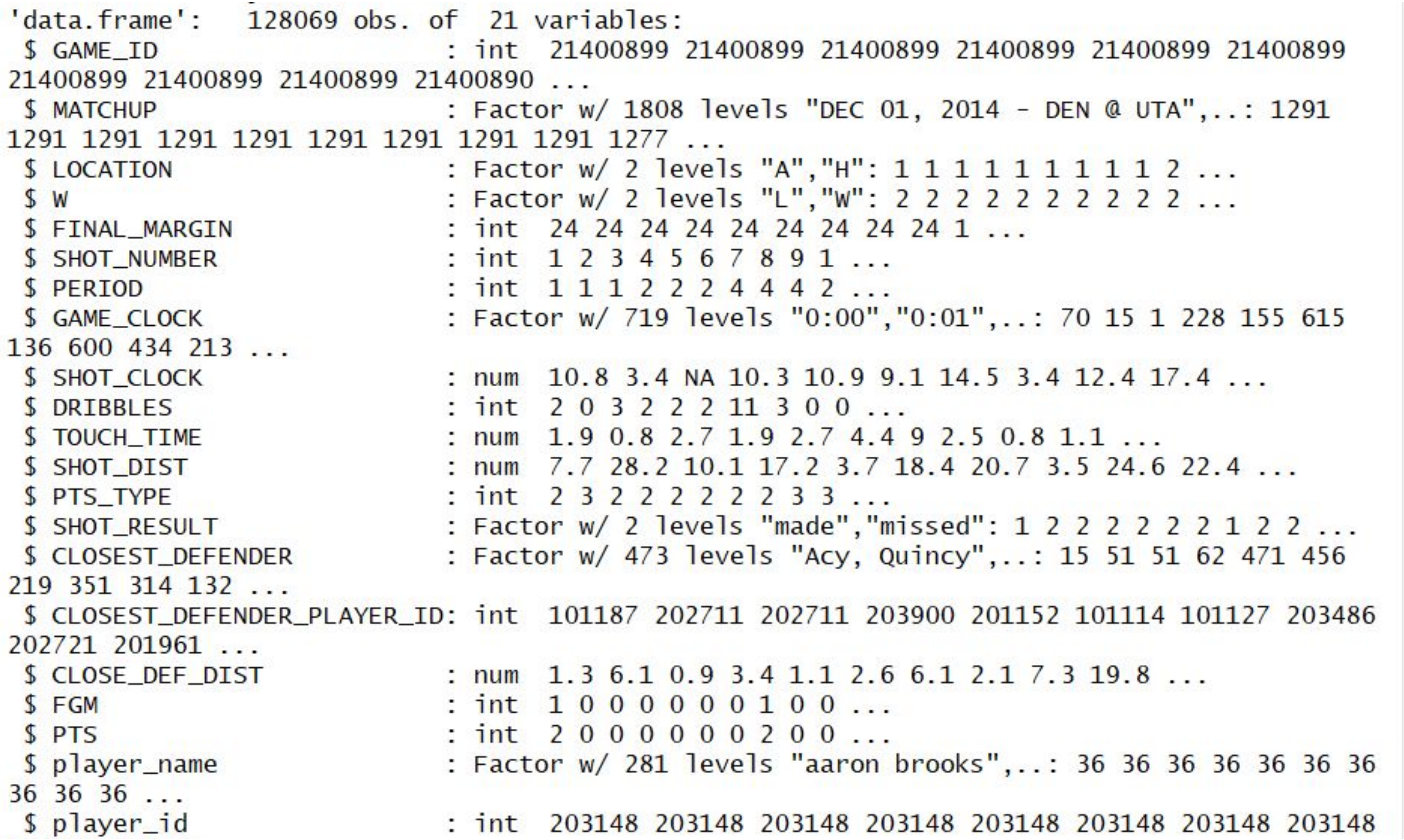
- Use of data science has become increasingly widespread in sports, including basketball, in recent years
    - Enabled by more data being recorded with new technology
- Results of this data science used by
    - Teams, for strategy and player evaluation
    - Media, for producing insights that can be presented by analysts and writers
    - Gamblers, for gaining an advantage in predicting outcomes
- NBA shot logs with every shot in the 2014-2015 season (128,069 shots) from NBA's website (via kaggle)

Question Development and Relevant Research

- What are the chances of a given shot going in?

- Impact of Research
  - Warriors and others proved that taking a higher proportion of 3 point shots can be beneficial despite previous strategies
- Data Preparation
  - Recoded game clock to be time in seconds
  - Handled null shot clock values
- Data Limitation
  - No free throws
  - Limited number of variables

Data Structure



```
'data.frame':   128069 obs. of  21 variables:
 $ GAME_ID                    : int  21400899 21400899 21400899 21400899 21400899 21400899
21400899 21400899 21400899 21400890 ...
 $ MATCHUP                    : Factor w/ 1808 levels "DEC 01, 2014 - DEN @ UTA",..: 1291
1291 1291 1291 1291 1291 1291 1291 1291 1277 ...
 $ LOCATION                   : Factor w/ 2 levels "A","H": 1 1 1 1 1 1 1 1 1 2 ...
 $ W                          : Factor w/ 2 levels "L","W": 2 2 2 2 2 2 2 2 2 2 ...
 $ FINAL_MARGIN               : int  24 24 24 24 24 24 24 24 24 1 ...
 $ SHOT_NUMBER                : int  1 2 3 4 5 6 7 8 9 1 ...
 $ PERIOD                     : int  1 1 1 2 2 2 4 4 4 2 ...
 $ GAME_CLOCK                 : Factor w/ 719 levels "0:00","0:01",..: 70 15 1 228 155 615
136 600 434 213 ...
 $ SHOT_CLOCK                 : num  10.8 3.4 NA 10.3 10.9 9.1 14.5 3.4 12.4 17.4 ...
 $ DRIBBLES                   : int  2 0 3 2 2 2 11 3 0 0 ...
 $ TOUCH_TIME                 : num  1.9 0.8 2.7 1.9 2.7 4.4 9 2.5 0.8 1.1 ...
 $ SHOT_DIST                  : num  7.7 28.2 10.1 17.2 3.7 18.4 20.7 3.5 24.6 22.4 ...
 $ PTS_TYPE                   : int  2 3 2 2 2 2 2 2 3 3 ...
 $ SHOT_RESULT                : Factor w/ 2 levels "made","missed": 1 2 2 2 2 2 2 1 2 2 ...
 $ CLOSEST_DEFENDER           : Factor w/ 473 levels "Acy, Quincy",..: 15 51 51 62 471 456
219 351 314 132 ...
 $ CLOSEST_DEFENDER_PLAYER_ID: int  101187 202711 202711 203900 201152 101114 101127 203486
202721 201961 ...
 $ CLOSE_DEF_DIST             : num  1.3 6.1 0.9 3.4 1.1 2.6 6.1 2.1 7.3 19.8 ...
 $ FGM                        : int  1 0 0 0 0 0 0 1 0 0 ...
 $ PTS                        : int  2 0 0 0 0 0 0 2 0 0 ...
 $ player_name                : Factor w/ 281 levels "aaron brooks",..: 36 36 36 36 36 36 36
36 36 36 ...
 $ player_id                  : int  203148 203148 203148 203148 203148 203148 203148 203148
```

Summary Statistics (Continuous Variables)

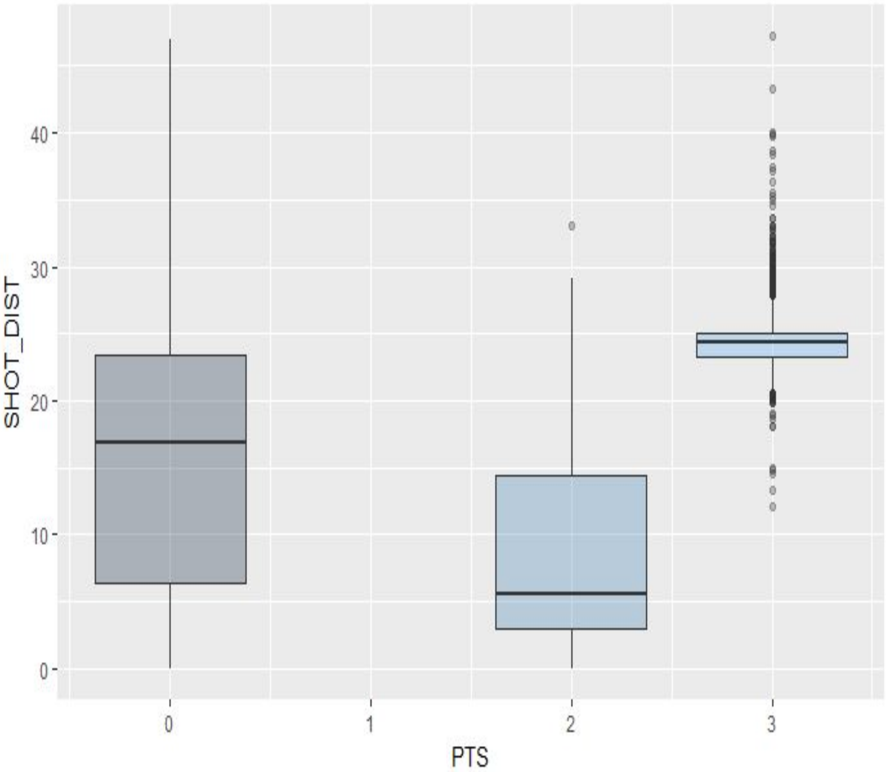
Variable	Count	Mean	Std. Dev	Min	25%	Median	75%	Max
Final Margin	128069	0.2087	13.233	-53.000	-8.000	1.000	9.000	53.000
Shot clock	122502	11.910	----	0.000	7.500	12.000	16.400	24.000
Dribbles	128069	2.023	3.477	0.000	0.000	1.000	2.000	32.000
Touch Time	128069	2.766	3.044	0.000	0.900	1.600	3.700	24.900
Shot Distance	128069	13.570	8.889	0.000	4.700	13.700	22.500	47.200
Shot number	128069	6.507	4.713	1.000	3.000	5.000	9.000	38.000
Closest defender distance	128069	4.123	2.756	0.000	2.300	3.700	5.300	53.200
Period	128069	2.469	1.139	1.000	1.000	2.000	3.000	7.000
Game Clock	128069	351	207.731	0.000	172	352	531	720

Summary Statistics (Categorical Variables)

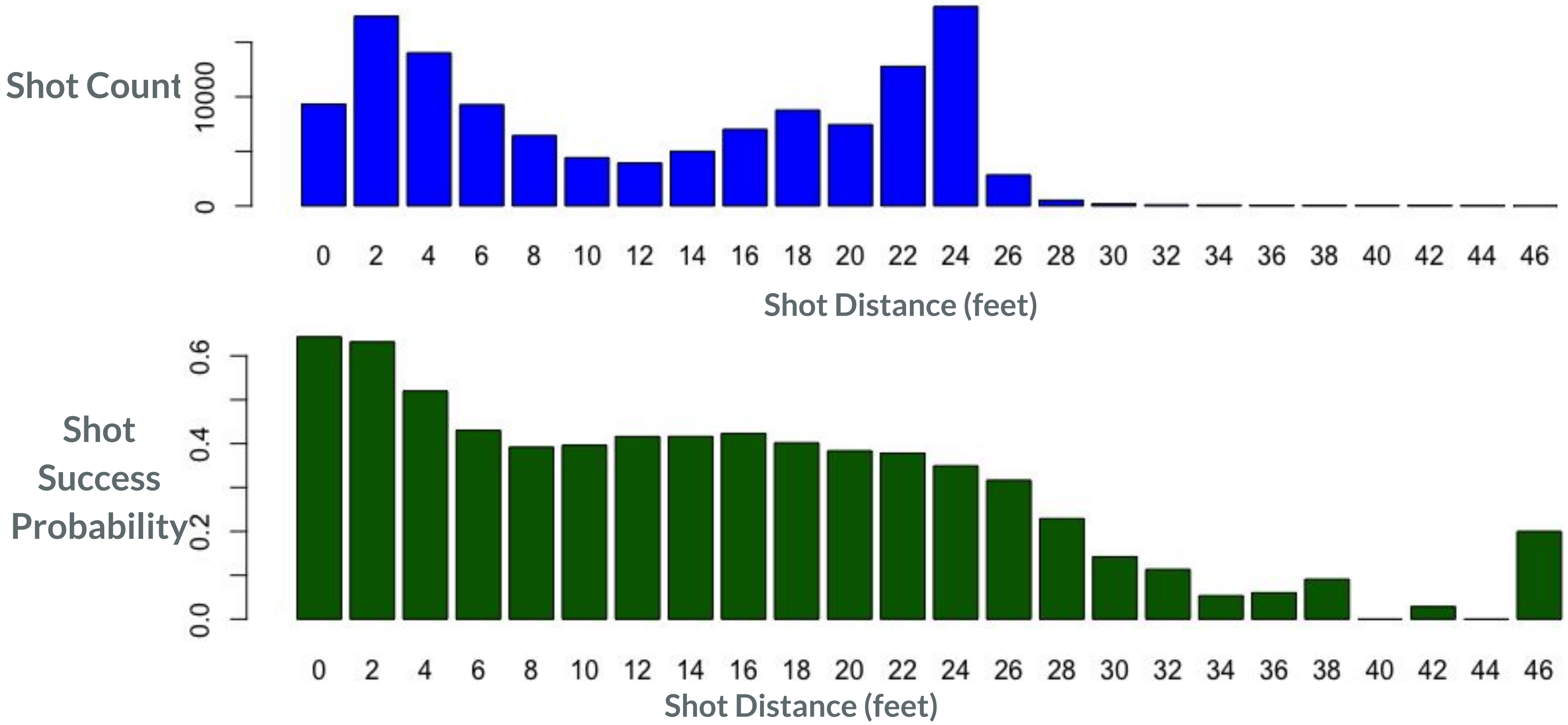
Variable	# of categories	Category Name(%)	Category Name(%)	Category Name(%)
Location	2	Home(50%)	Away(50%)	
WIN	2	Win(50%)	Loss(50%)	
Points	3	Zero(55%)	Two(36%)	Three(9%)
Point Type	2	Two(74%)	Three(26%)	
Field Goal Made	**2**	**Yes(45%)**	**No(55%)**	

Exploratory Data Analysis

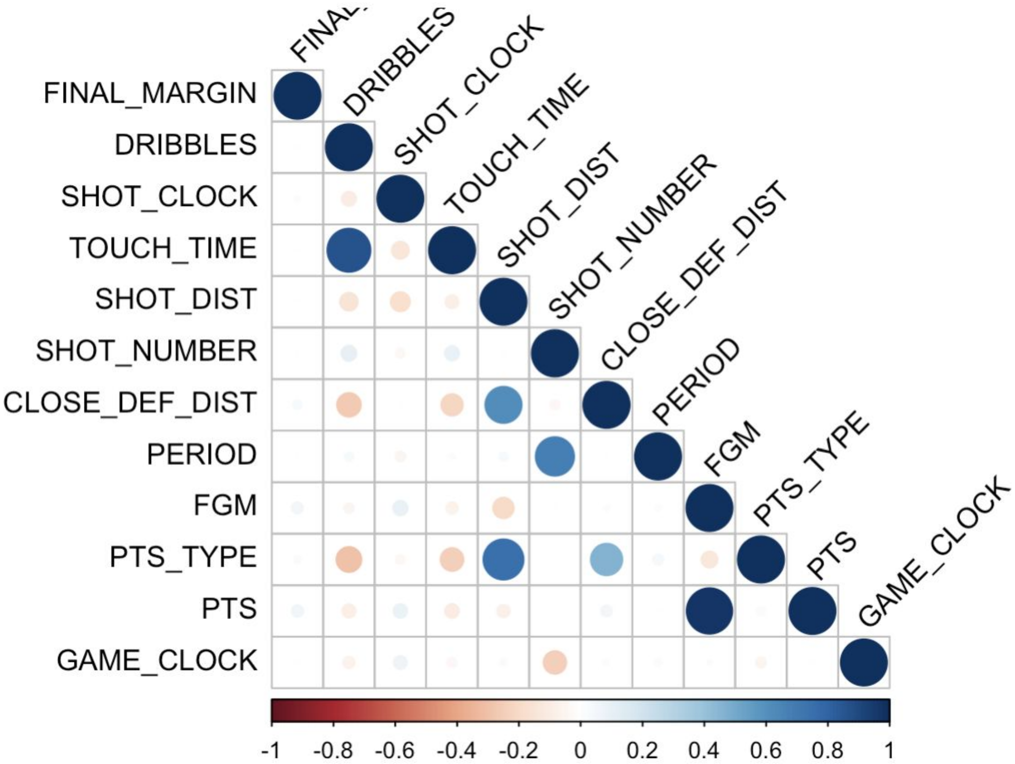
Field Goals	Made %	Missed %	Shot Type %
2-pointers	49%	51%	74%
3-pointers	35%	65%	26%
All Field Goals	45%	55%	128,069



Shot Distance



Correlation



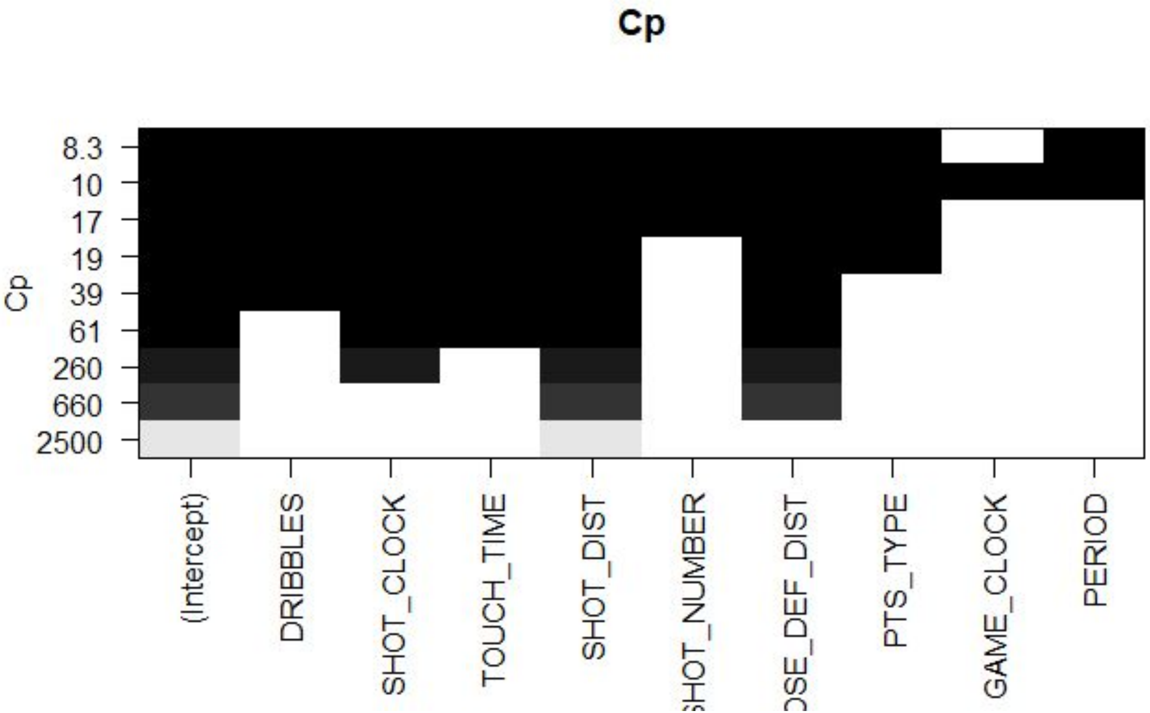
Feature Selection and Methodology

- Focused on variables pertaining to specific shots
    - Exclude outcome variables
    - Exclude descriptive variables
    - Utilized Cp to select a simple model of shot variables

- Methodology
    - Logistic regression
        - Dependent Variable = Shot Result
            - Binary
        - Built a predictive model of likelihood of shot success

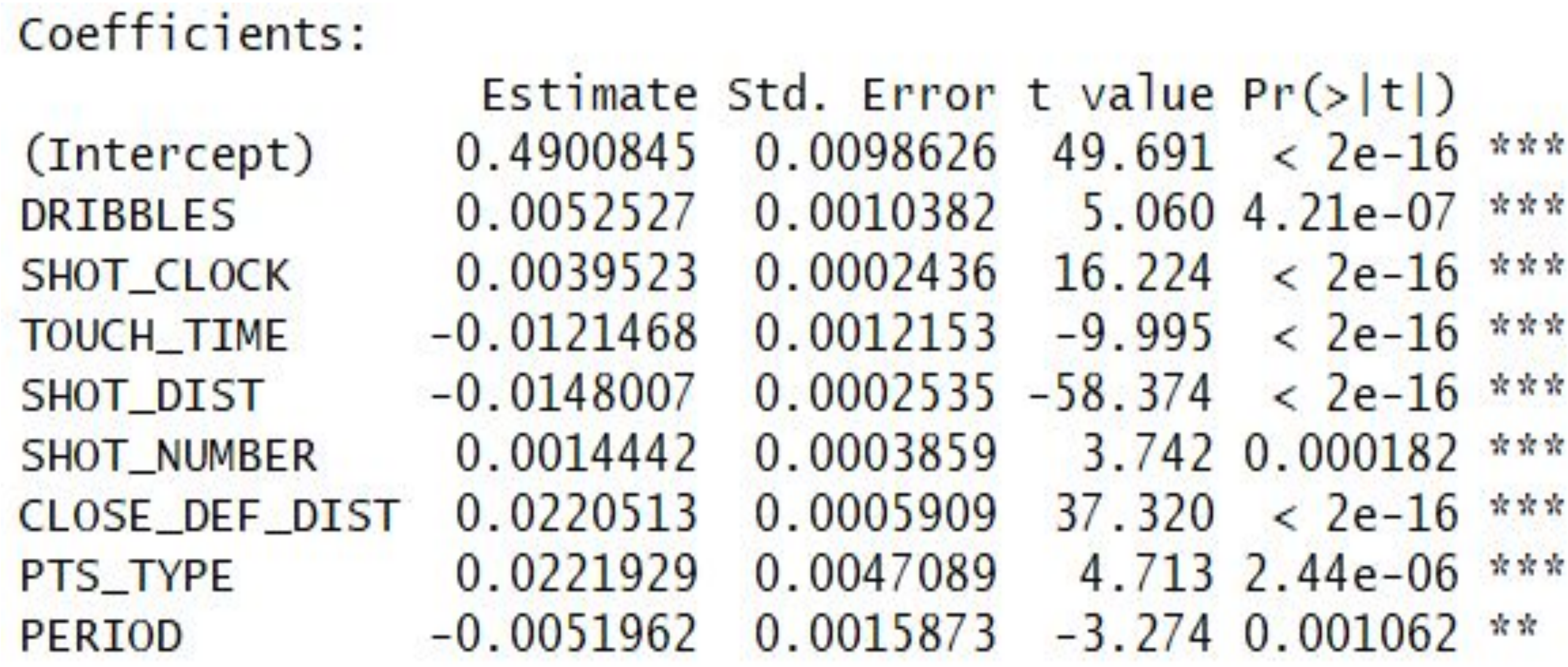
Choosing the Best Model

- Best Simple Model
  - Lowest Cp
    - Dropped Game Clock



Logistic Regression Model

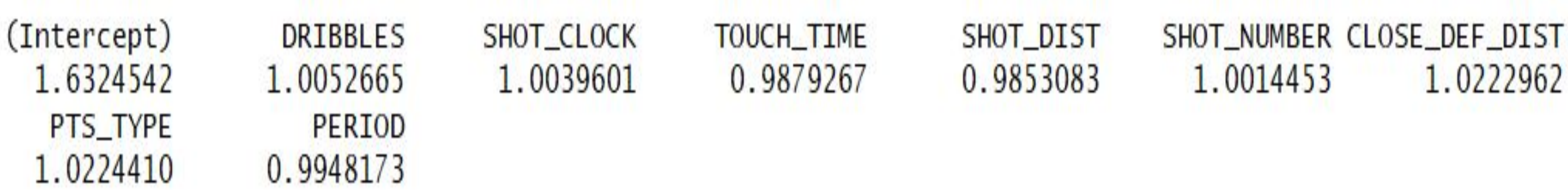
Each variable's coefficient is statistically significant



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.4900845  0.0098626  49.691  < 2e-16 ***
DRIBBLES         0.0052527  0.0010382   5.060 4.21e-07 ***
SHOT_CLOCK       0.0039523  0.0002436  16.224  < 2e-16 ***
TOUCH_TIME      -0.0121468  0.0012153  -9.995  < 2e-16 ***
SHOT_DIST       -0.0148007  0.0002535 -58.374  < 2e-16 ***
SHOT_NUMBER      0.0014442  0.0003859   3.742 0.000182 ***
CLOSE_DEF_DIST   0.0220513  0.0005909  37.320  < 2e-16 ***
PTS_TYPE         0.0221929  0.0047089   4.713 2.44e-06 ***
PERIOD          -0.0051962  0.0015873  -3.274 0.001062 ** 
```

Prediction

How much does one foot decrease in shot distance improve likelihood of shot success?



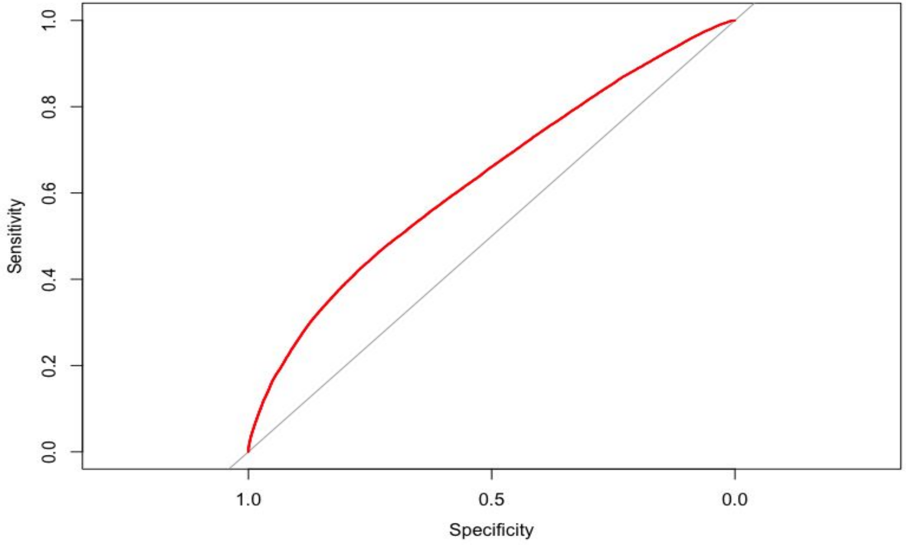
```
(Intercept)     DRIBBLES   SHOT_CLOCK   TOUCH_TIME    SHOT_DIST  SHOT_NUMBER CLOSE_DEF_DIST
  1.6324542    1.0052665    1.0039601    0.9879267    0.9853083    1.0014453      1.0222962
   PTS_TYPE       PERIOD
  1.0224410    0.9948173
```

Shot Distance odds-ratio is .985

A player is 1.5% more likely to make a shot for every foot closer to the basket they are, when they take the shot.

Results

- Are the results reliable?
    - Hosmer and Lemeshow Goodness of fit test
        - P-value is very small
            - Significant difference between model and observed data
    - ROC and AUC
        - Area under the curve is .6313



Future Prospects

- Model Improvements
  - Join additional data from other datasets
    - Player information: Height, age, years playing professionally, …
    - Both X and Y coordinates of shot location, rather than just shot distance
  - Models based on individual players

- Other Questions??
  - Model based on game win
    - Need Free Throw information

Tools

- **R**
    - **Packages**
        - corrplot
    - **Charts**
        - Histogram
        - Correlation Matrix
        - Boxplot
    - Models
        - Logistic Regression

References

National Basketball Association. NBA Advanced Stats. April 4, 2018 from https://stats.nba.com/player/201956/shots-dash/

Cohen, B. (2016). The Golden State Warriors Have Revolutionized Basketball. April 6, 2016. https://www.wsj.com/articles/the-golden-state-warriors-have-revolutionized-basketball-1459956975

Questions?