Kobe Bryant Shot Selection

K.I. Chung, K.J. Yang
Supervised by: Prof. Guo, Mei-Hui

January 14, 2017

About the data

- From the Kaggle
- Containing all the Kobe's shots during his career
- 30697 shots and 25 variables
- 25697 training data and 5000 test data

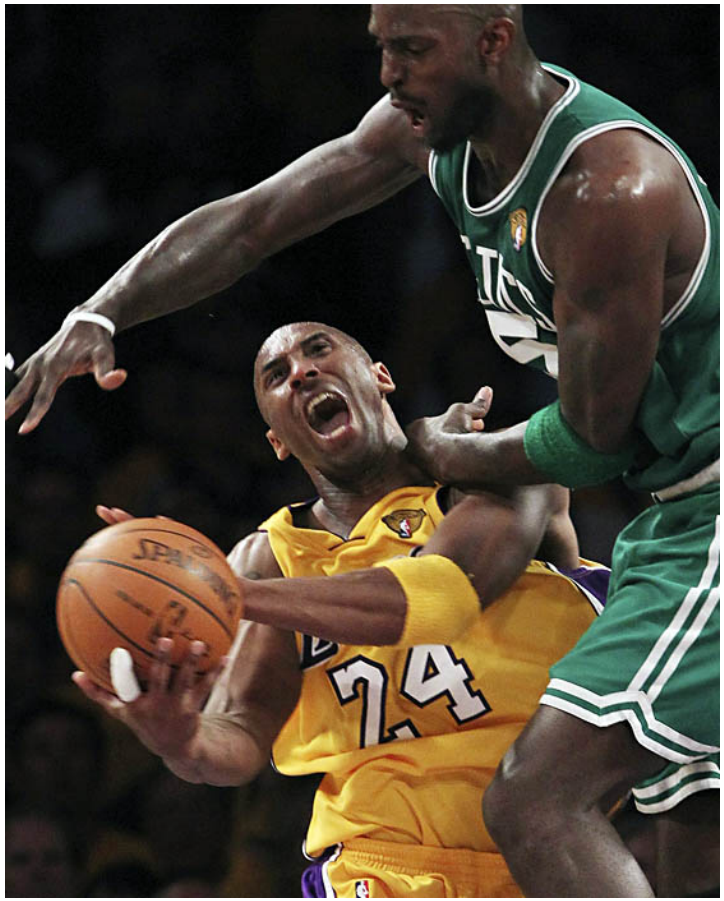


Figure: Kobe Bryant was blocked

Response

- Whether a shot is made or not
- 1 and 0 are denoted for a shot is made or not
- NA is denoted for the test data

Figure: Kobe Bryant was in a rage

Predictors

- 24 variables
- Factors: shot types, shot zone, match up, playoffs?
- Continuous: location(x, y), time remaining?

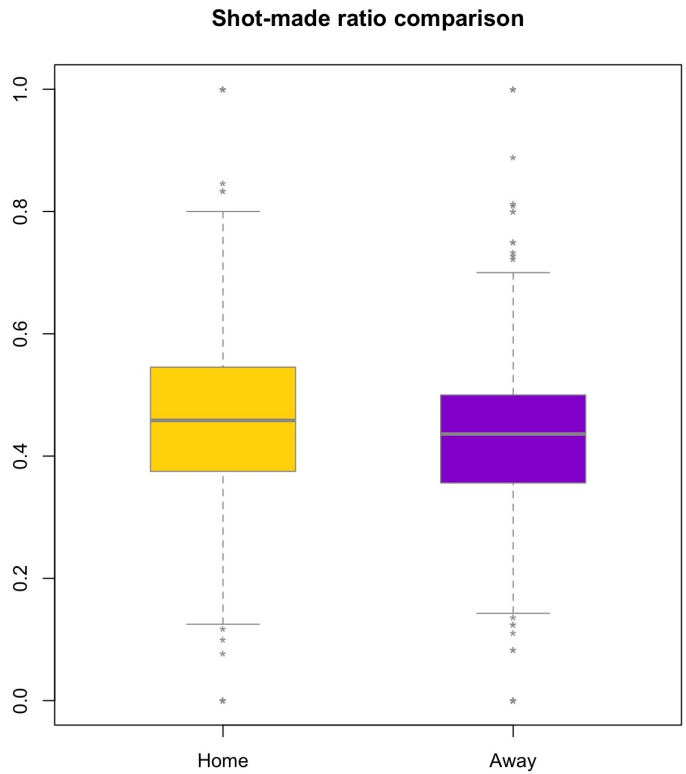
Figure: Kobe Bryant did not care

```
> summary(df)
     season           game_date         game_id                     action_type
 2005   : 1924   2016-04-13:   43   21501228:   43   Jump Shot           :15836
 2002   : 1852   2007-03-30:   41   20601081:   41   Layup Shot          : 2154
 2008   : 1851   2002-11-07:   39   20200069:   39   Driving Layup Shot  : 1628
 2007   : 1819   2006-01-22:   39   20500591:   39   Turnaround Jump Shot:  891
 2009   : 1772   2008-01-14:   37   20700553:   37   Fadeaway Jump Shot  :  872
 2001   : 1708   2010-01-08:   36   20900527:   36   Running Jump Shot   :  779
 (Other):14771   (Other)   :25462   (Other) :25462   (Other)            : 3537
 combined_shot_type     loc_x             loc_y        shot_distance   three_pt
 Bank Shot:  120    Min.   :-250.000   Min.   :-44.00   Min.   : 0.00   0:20284
 Dunk     : 1056    1st Qu.: -67.000   1st Qu.:  4.00   1st Qu.: 5.00   1: 5413
 Hook Shot:  127    Median :   0.000   Median : 74.00   Median :15.00
 Jump Shot:19710    Mean   :   7.148   Mean   : 91.26   Mean   :13.46
 Layup    : 4532    3rd Qu.:  94.000   3rd Qu.:160.00   3rd Qu.:21.00
 Tip Shot :  152    Max.   : 248.000   Max.   :791.00   Max.   :79.00

 time_remaining ot        home      playoffs     opponent     shot_made_flag
 Min.   :   0   0:25380   0:13212   0:21939   SAS    : 1638   0:14232
 1st Qu.: 720   1:  317   1:12485   1: 3758   PHX    : 1535   1:11465
 Median :1378                                 HOU    : 1399
 Mean   :1397                                 SAC    : 1397
 3rd Qu.:2171                                 DEN    : 1352
 Max.   :2874                                 POR    : 1292
                                              (Other):17084
```

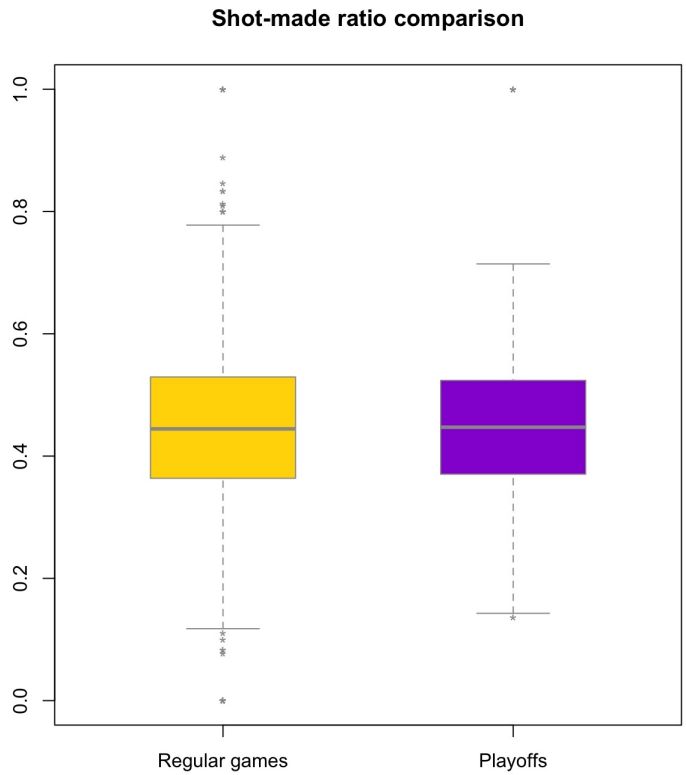
Home v.s. Away

- paired: FALSE
- p-value of variance test: 0.7318
- var.equal: TRUE
- p-value of two sample t-test:0.0007103
- result: different means



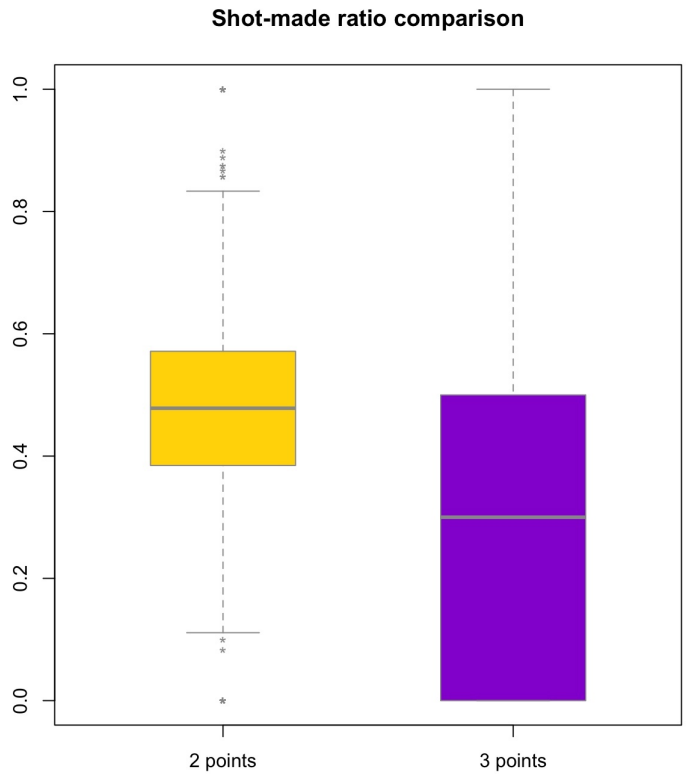
Regular games v.s. Playoffs

- paired: FALSE
- p-value of variance test: 0.2206
- var.equal: TRUE
- p-value of two sample t-test:0.5658
- result: same means

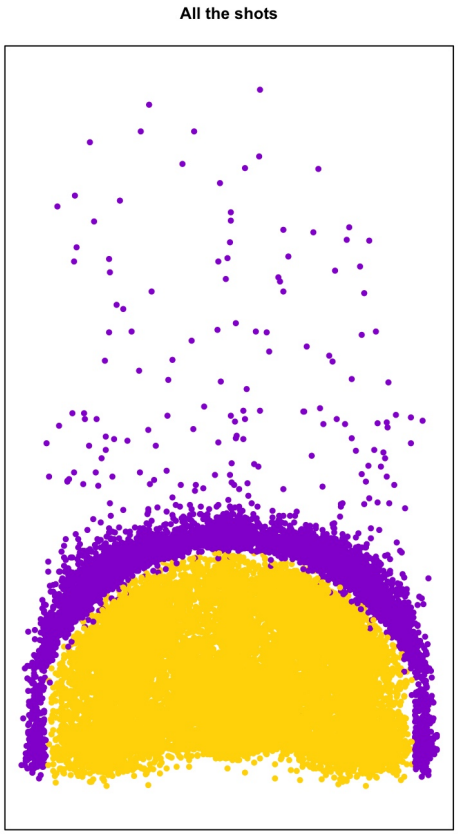


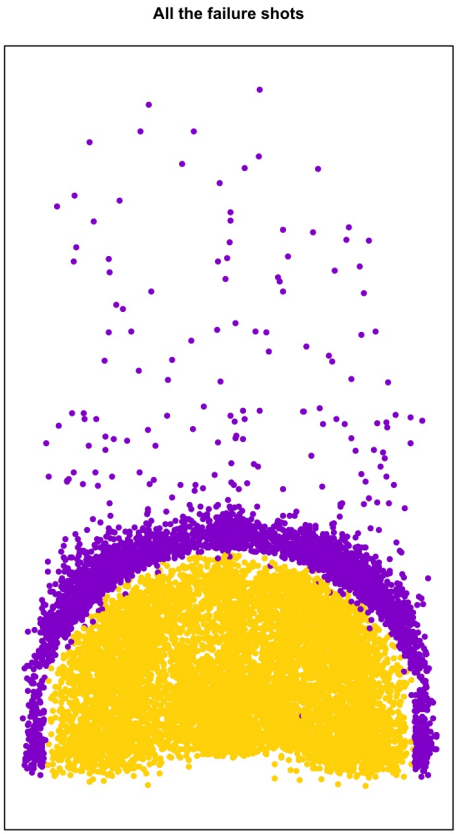
2-pt. v.s. 3-pt.

- paired: TRUE
- p-value of variance test: 0
- var.equal: FALSE
- p-value of two sample t-test: 0
- result: different means

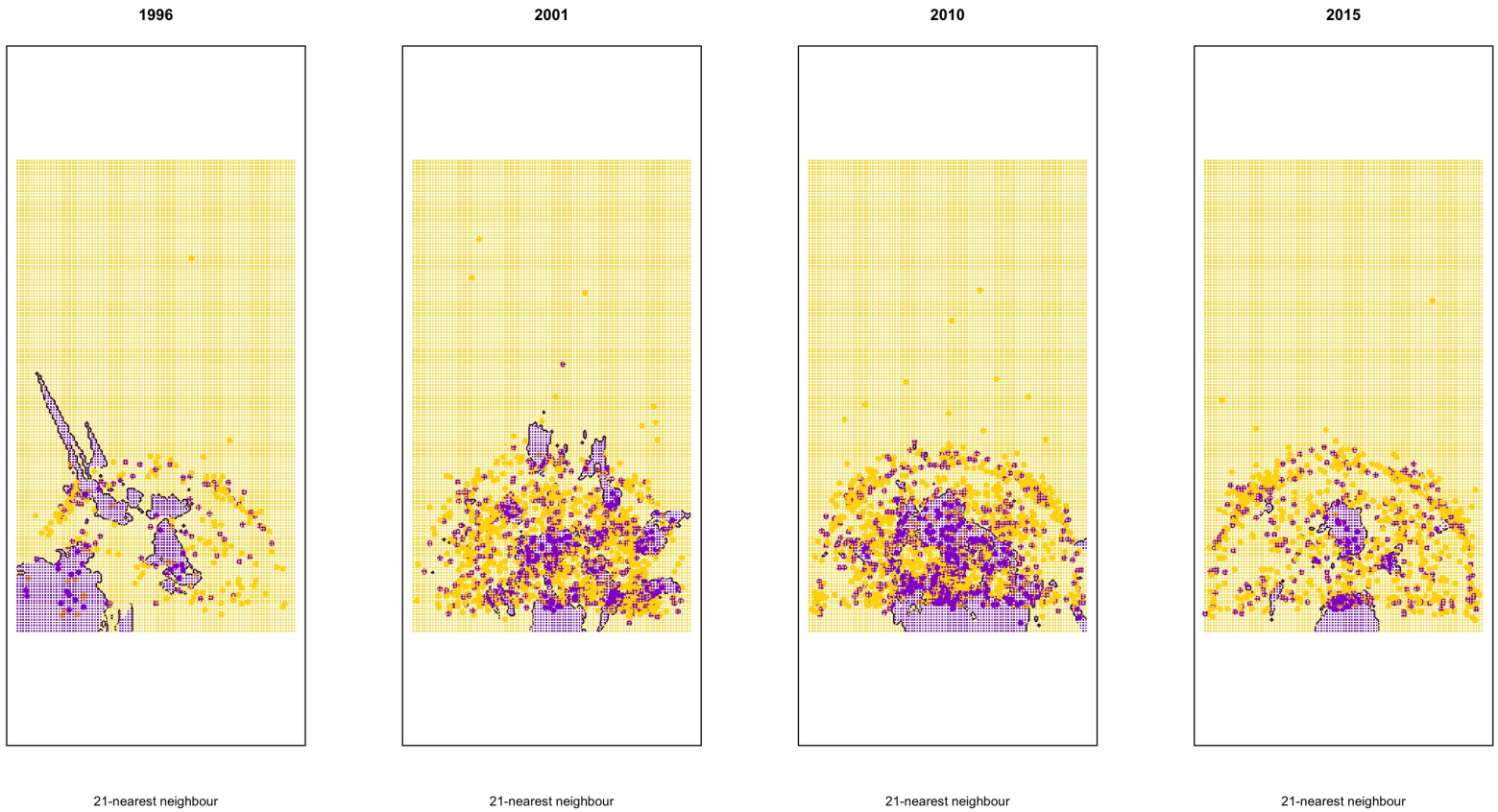


Made Shot Location Comparison



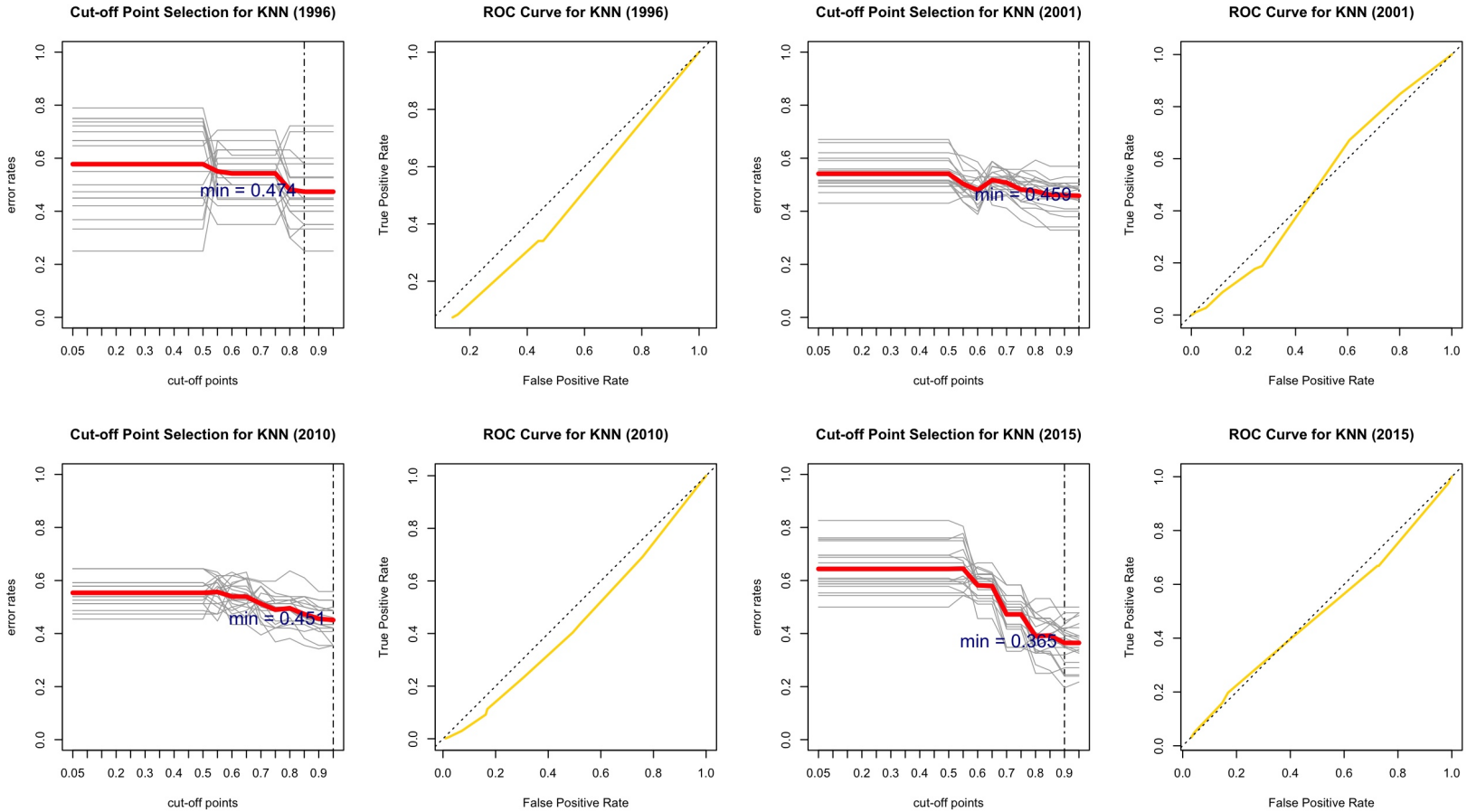


Yearly Trend of the KNN Classifier



KNN Results with Different k

20-folds Cross Validation Result in Different Season



With the mixed selection method , the following predictors were chosen.

- season
- combined_shot_ type
- loc_y
- shot_distance
- three_pt
- time_remaining
- home

Now we check the collinearity, the result is shown below. We can noticed that the predictors location y, shot distance and three point have relatively high collinearity.

$$GVIF' = GVIF^{\frac{1}{2df}}$$

predictor	GVIF	df	GVIF'
season	1.0772	19	1.0019
shot type	2.9098	5	1.1127
location y	3.0372	1	1.7428
shot distance	6.9311	1	2.6327
three point	2.2918	1	1.5139
time remaining	1.0176	1	1.0088
home	1.0090	1	1.0045

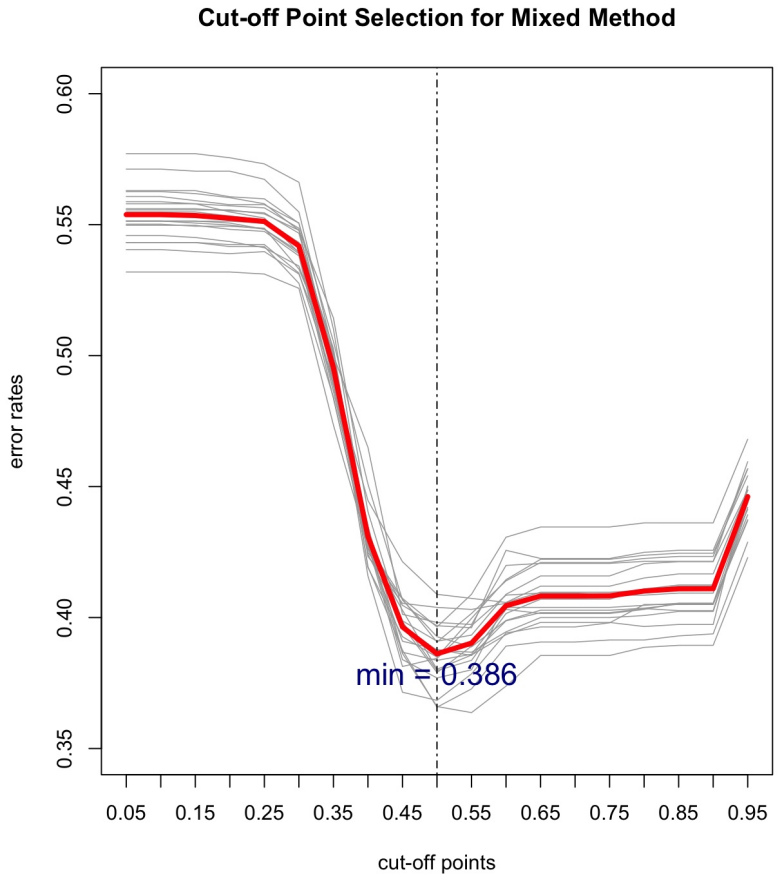
To eliminate the the effect of the collinearity, we will perform the "leave-one-in" model selection which provide us the least AIC model with one of three collinear predictors.

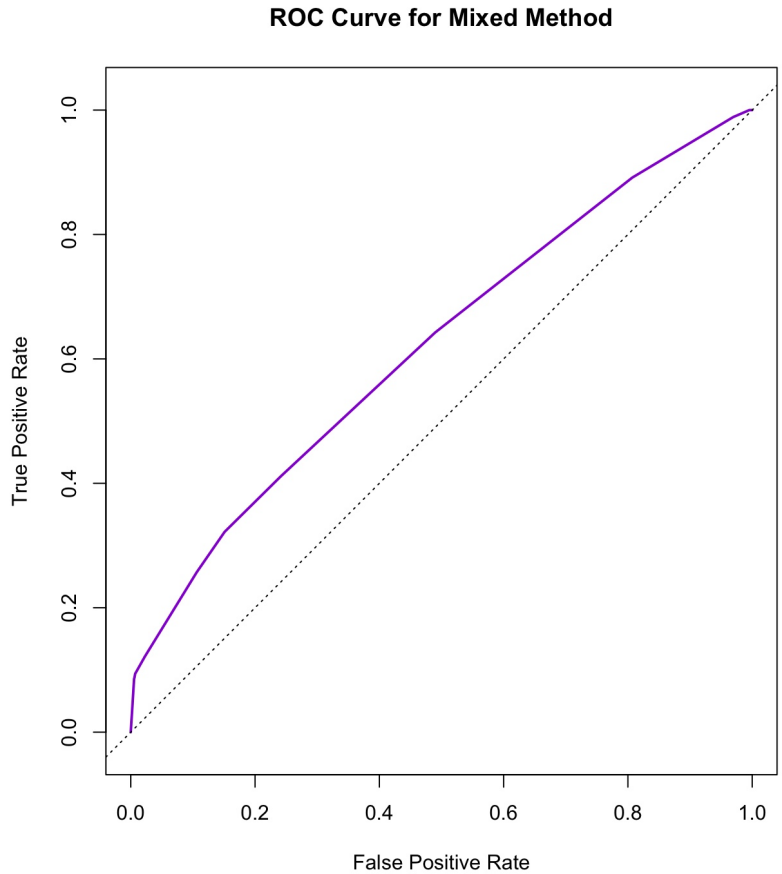
Table: default

predictor	location y	shot distance	three point
AIC	33540.53	33466.69	33493.91

Table: Coefficients of the Selected Predictors without Collinearity

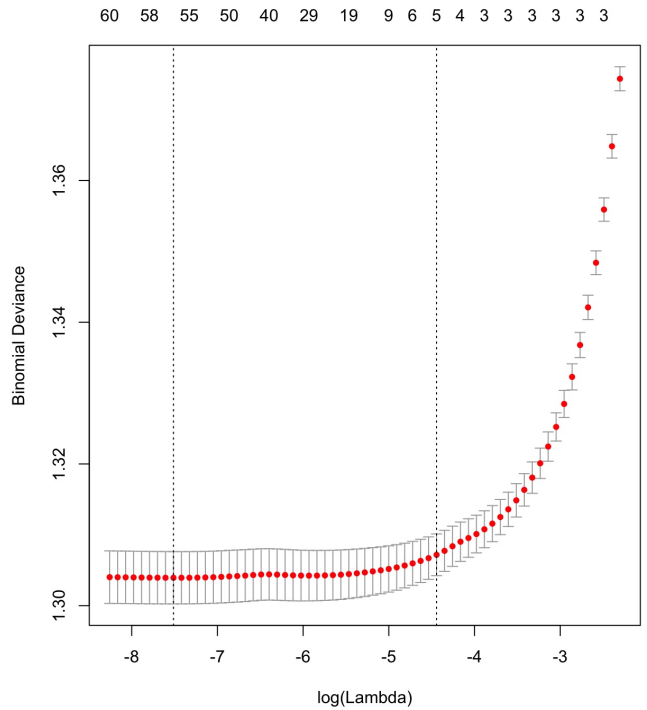
Intercept	1997	1998	1999	2000
1.2131	-0.0367	0.1537	0.1611	0.1874
2001	2002	2003	2004	2005
0.1397	0.1017	0.0515	0.1003	0.2651
2006	2007	2008	2009	2010
0.2487	0.2421	0.2673	0.2188	0.2018
2011	2012	2013	2014	2015
0.1482	0.244	0.0589	-0.0236	-0.042
Dunk	Shot	Jump Shot	Layup	Tip Shot
1.0643	-1.2381	-1.5344	-1.2354	-2.1032
Distance	Time	Home		
-0.0241	0.0004	0.0417		

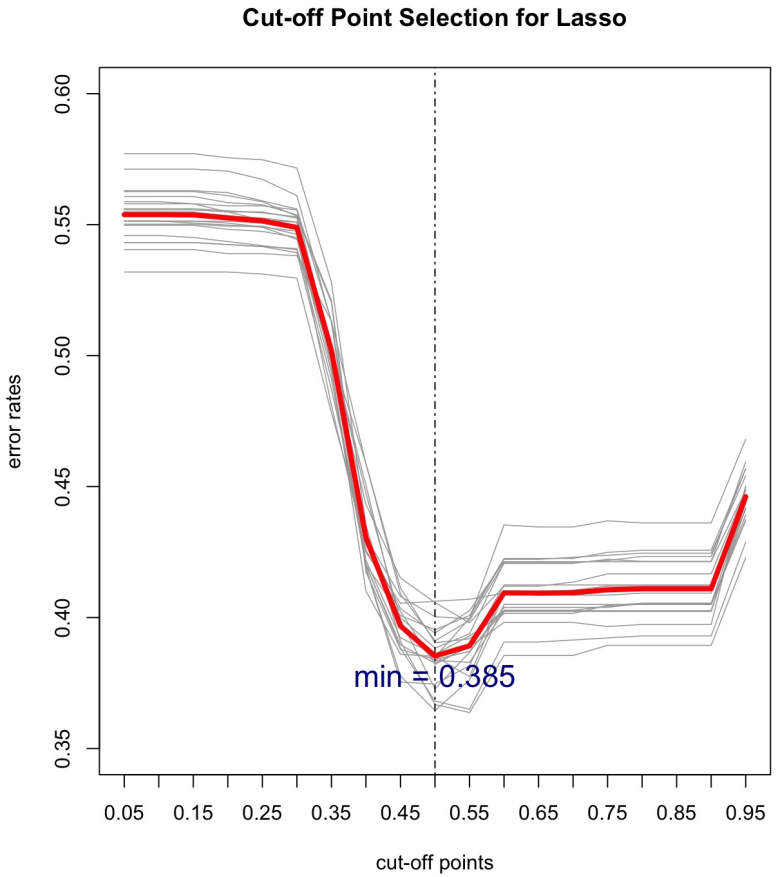

Cut-off Point Selection for Mixed Method
error rates
0.35
0.40
0.45
0.50
0.55
0.60
min = 0.386
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
cut-off points

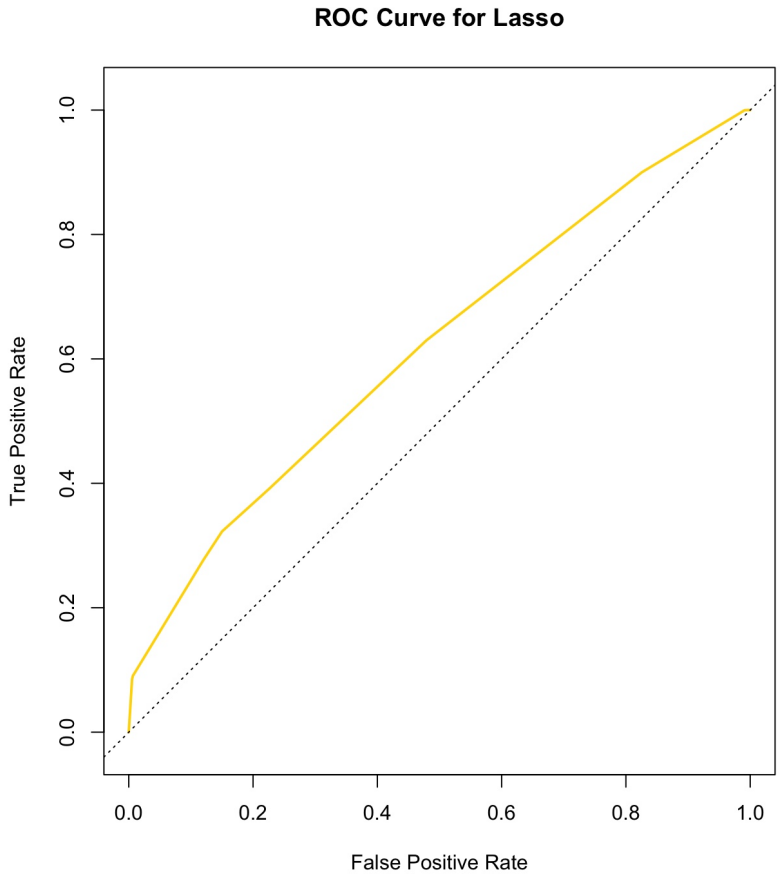

ROC Curve for Mixed Method
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

Now we introduce another model selection method, LASSO (least absolute shrinkage and selection operator) regression. After performing cross validation, we have the best $\lambda = 0.000546$ which has the model provide the least test error rate.

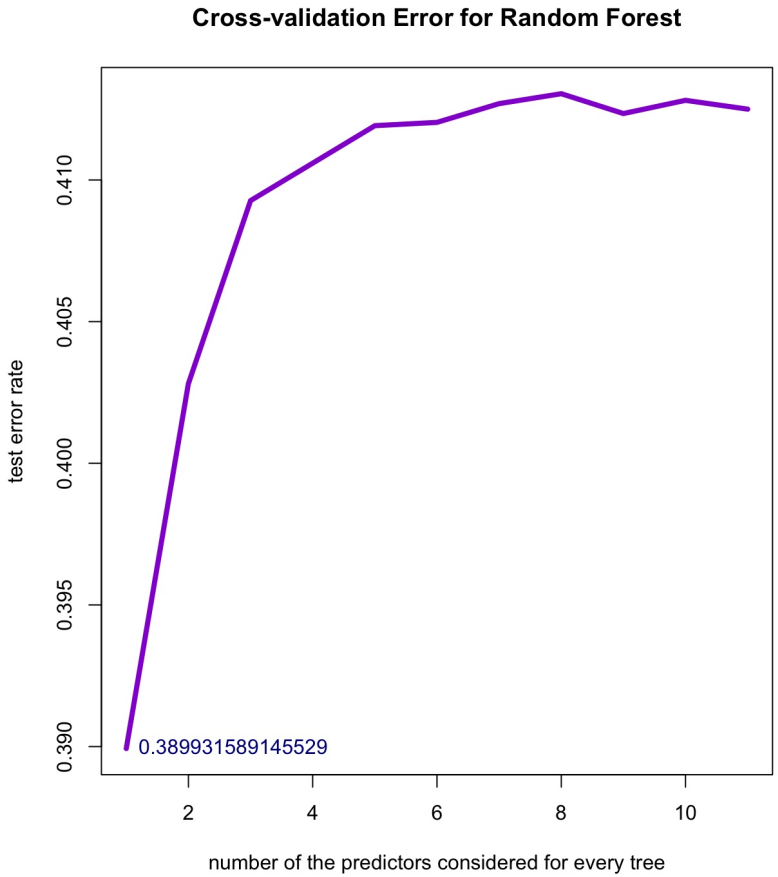
predictor	coefficient
Intercept	0.19113
Dunk	-1.69536
Jump Shot	-0.28377
Tip Shot	-0.20894
Shot Distance	-0.02038
Time Remaining	0.00002

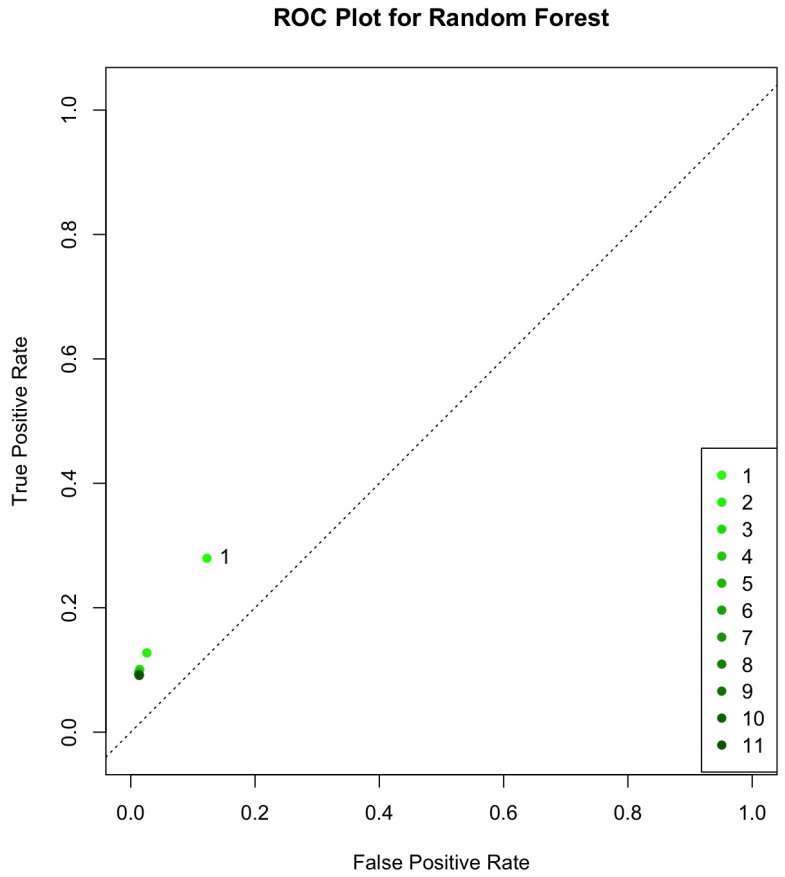



Cut-off Point Selection for Lasso
error rates
0.35
0.40
0.45
0.50
0.55
0.60
min = 0.385
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
cut-off points


ROC Curve for Lasso
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

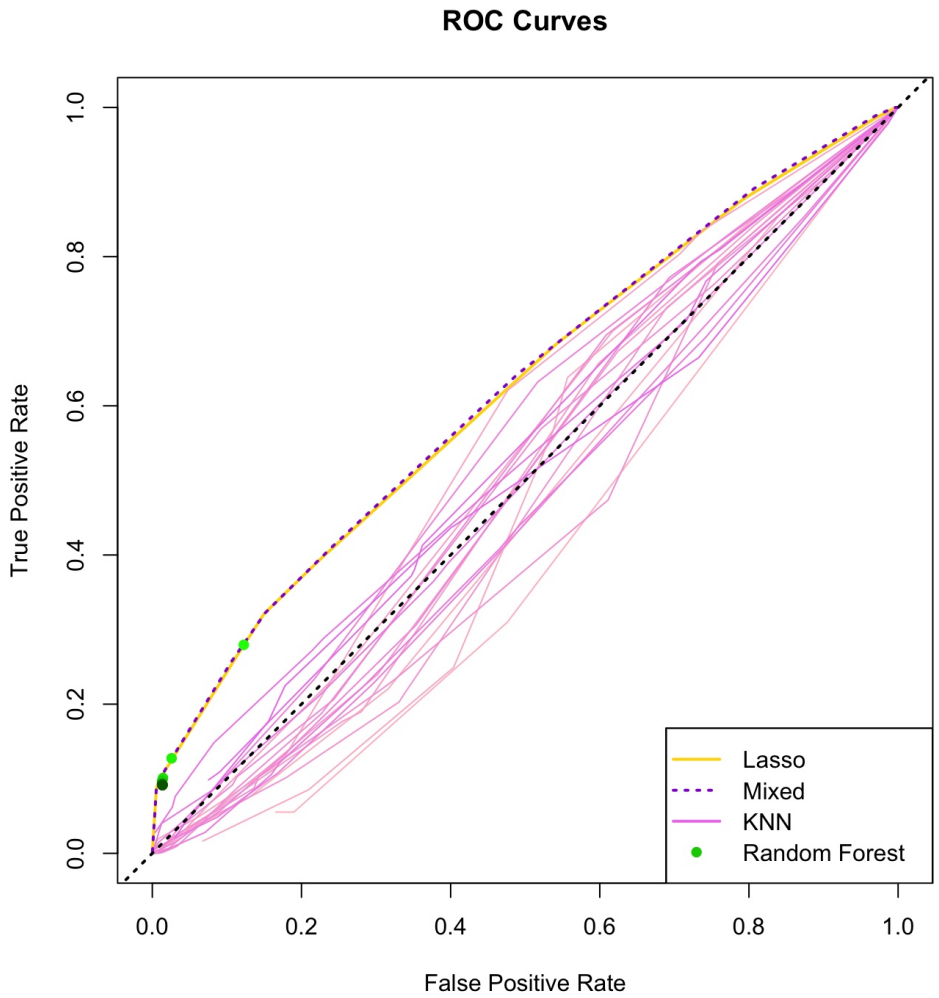
We applied random forest with 500 trees and derived 20-folds cross validation test error rate. The figure in the next slide shows that test error rates under the different m, the numbers of variables randomly sampled as candidates at each split. Notice that the Bagging(bootstrap and aggregating) method is a special case of the random forest with $m = 11$, the number of the predictors.


Cross-validation Error for Random Forest
test error rate
0.390
0.395
0.400
0.405
0.410
0.389931589145529
2
4
6
8
10
number of the predictors considered for every tree


ROC Plot for Random Forest
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
1
1
2
3
4
5
6
7
8
9
10
11
False Positive Rate

The random forest with $m = 1$ performed the least test error rate among the 11 forests. Yet, the result is slightly worse than the one of the logistic regression.

Comparing the ROC curves, we noticed that curves of all the methods excluding the KNN's coincide. Also the cross validation test error rates are about 0.385. Thus, we speculate that the models have reach their irreducible error.



- Consider the support vector machine
- Collect data with more variable
- Give up

1 James, G., Witten, D., Hastie, T., Tibshirani, R., *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013

2 Kaggle, *Kobe Bryant Shot Selection*, https://www.kaggle.com/c/kobe-bryant-shot-selection