Baseball Awards and Salaries

Tom Cooper, Lindsey Dodd Gooch, Josh Harwood,
Grace Wang, Danielle Williams

Overview: Lahman Baseball Data

Salaries	• Player salary Data
Master	• Player's demographics (name, DOB, etc)
Batting	• Batting statistics
Fielding	• Fielding data such as position, assists, double plays, etc.
AwardsPlayers	• Player award information
Pitchers	• Pitchers pitching statistics

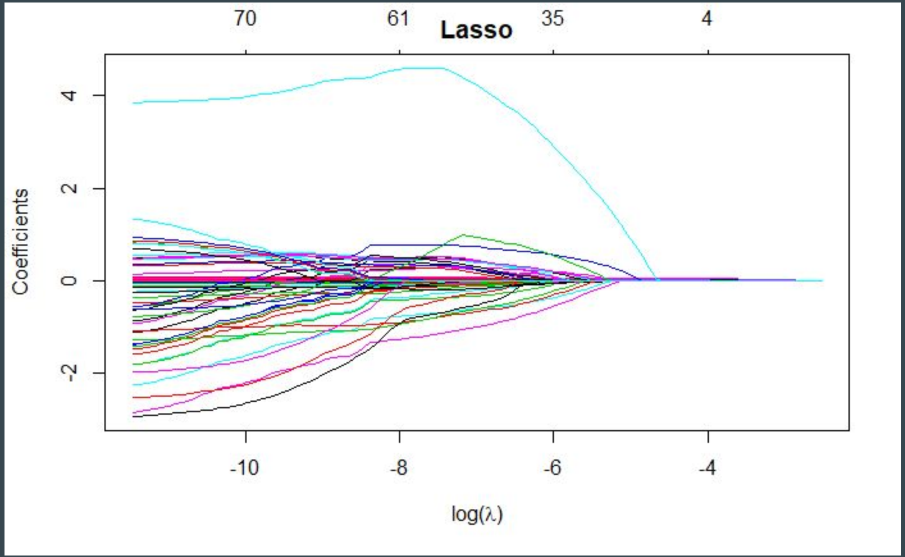
Baseball Awards

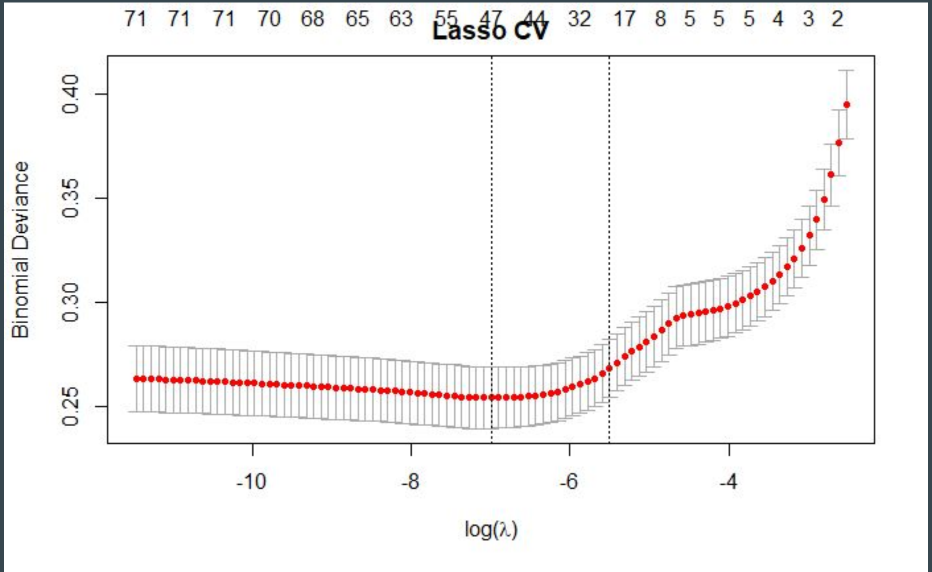
Preparing the data

```
#subset player awards and define "number of awards" variable
Winners<- AwardsPlayers %>%
            select(playerID, awardID, yearID) %>%
            filter(between(yearID, 2010, 2016)) %>%
            group_by(playerID, yearID) %>% mutate(num_awards = n())

Winners$playerID= as.factor(Winners$playerID)
Winners$yearID = as.factor(Winners$yearID)

#further subset to drop duplicates (players winning more than one award
winners<- Winners %>%
          distinct(playerID, yearID, num_awards)


#join to salary data set established for regression problem
class<- salData %>%
             left_join(winners,
                       by =c("playerID", "yearID"))
class[is.na(class)]<-0

class %>%
  group_by(num_awards) %>%
  summarise(no_rows = length(num_awards))

#create binary variable 0=no award, 1= at least one award
class<- class %>%
    mutate(award=(if_else(num_awards>0, 1, 0)))
```

	GVIF	Df	GVIF^(1/(2*Df))
yearID	1.249109	5	1.022492
POS	85477.228744	6	2.576247
G.x	41.403393	1	6.434547
GS	391.028681	1	19.774445
InnOuts	677.625691	1	26.031245
PO	34.830712	1	5.901755
A	58.986169	1	7.680245
E	4.287932	1	2.070732
DP	32.075485	1	5.663522
salary	1.987468	1	1.409776
height	1.653824	1	1.286011
weight	1.751745	1	1.323535
age	1.714796	1	1.309502
lgID.y	1.147982	1	1.071439
AB	118.850897	1	10.901876
R	40.534707	1	6.366687
H	104.627197	1	10.228744
X2B	11.845038	1	3.441662
X3B	2.644842	1	1.626297
HR	14.295384	1	3.780924
RBI	34.094181	1	5.839022
SB	4.219480	1	2.054137
CS	3.669873	1	1.915691
BB	7.438540	1	2.727369
SO	8.709011	1	2.951103
IBB	2.793288	1	1.671313
HBP	1.833786	1	1.354174
SH	1.698568	1	1.303291
SF	3.100570	1	1.760844
GIDP	4.480024	1	2.116607

Logistic Regression: Lasso CV





Logistic Regression: Lasso CV

Best Lambda

```
      lasso.best
 [1,] "yearID"
 [2,] "GS"
 [3,] "A"
 [4,] "E"
 [5,] "DP"
 [6,] "salary"
 [7,] "age"
 [8,] "H"
 [9,] "X3B"
[10,] "HR"
[11,] "RBI"
[12,] "SB"
[13,] "CS"
[14,] "BB"
[15,] "SO"
[16,] "IBB"
[17,] "HBP"
[18,] "SH"
[19,] "SF"
[20,] "GIDP"
[21,] "teamIDATL"
[22,] "teamIDBAL"
[23,] "teamIDBOS"
```

```
[24,] "teamIDCHN"
[25,] "teamIDCLE"
[26,] "teamIDCOL"
[27,] "teamIDDET"
[28,] "teamIDKCA"
[29,] "teamIDLAA"
[30,] "teamIDLAN"
[31,] "teamIDMIA"
[32,] "teamIDMIL"
[33,] "teamIDMIN"
[34,] "teamIDOAK"
[35,] "teamIDPIT"
[36,] "teamIDSDN"
[37,] "teamIDSFN"
[38,] "teamIDSLN"
[39,] "teamIDTBA"
[40,] "teamIDTEX"
[41,] "teamIDTOR"
[42,] "teamIDWAS"
[43,] "POS2B"
[44,] "POS3B"
[45,] "POSC"
[46,] "POSP"
[47,] "handComboBR"
```

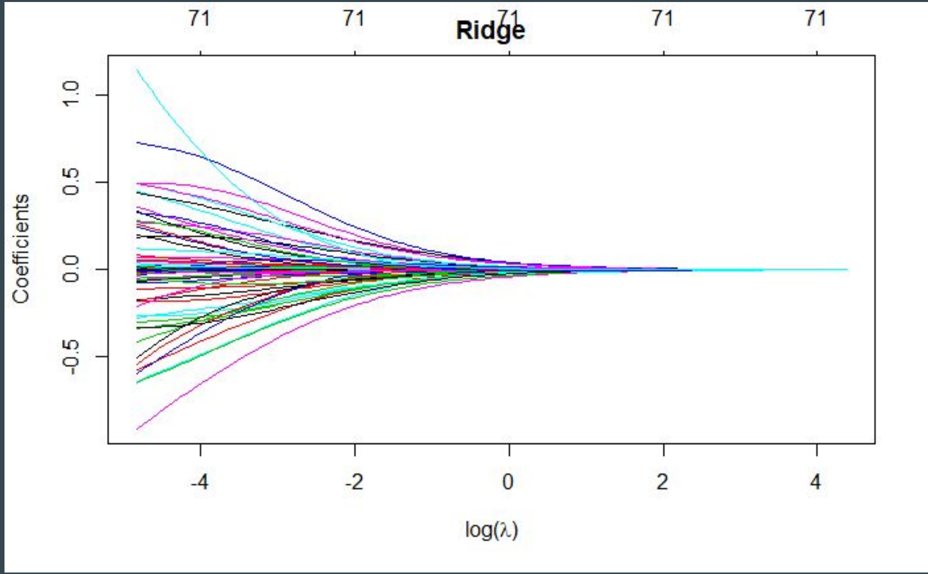
Best Lambda +1se

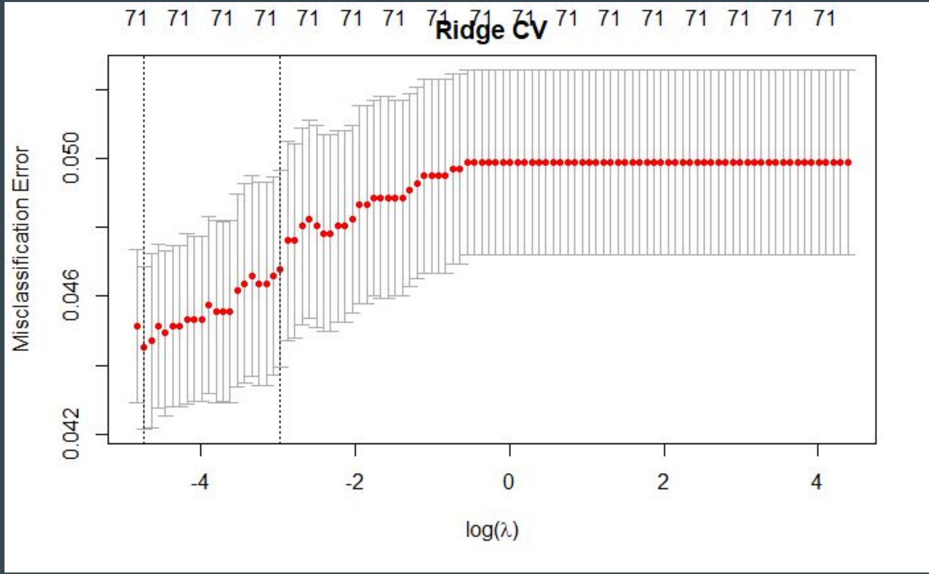
```
      lasso.1se
 [1,] "GS"
 [2,] "A"
 [3,] "E"
 [4,] "DP"
 [5,] "salary"
 [6,] "age"
 [7,] "H"
 [8,] "HR"
 [9,] "RBI"
[10,] "SO"
```

```
[11,] "IBB"
[12,] "SH"
[13,] "teamIDBAL"
[14,] "teamIDKCA"
[15,] "teamIDLAA"
[16,] "teamIDMIL"
[17,] "teamIDOAK"
[18,] "teamIDSFN"
[19,] "teamIDSLN"
[20,] "POSC"
[21,] "POSP"
```

	Misclassification error <dbl>
Lasso Best	0.03754941
Lasso 1se	0.04347826

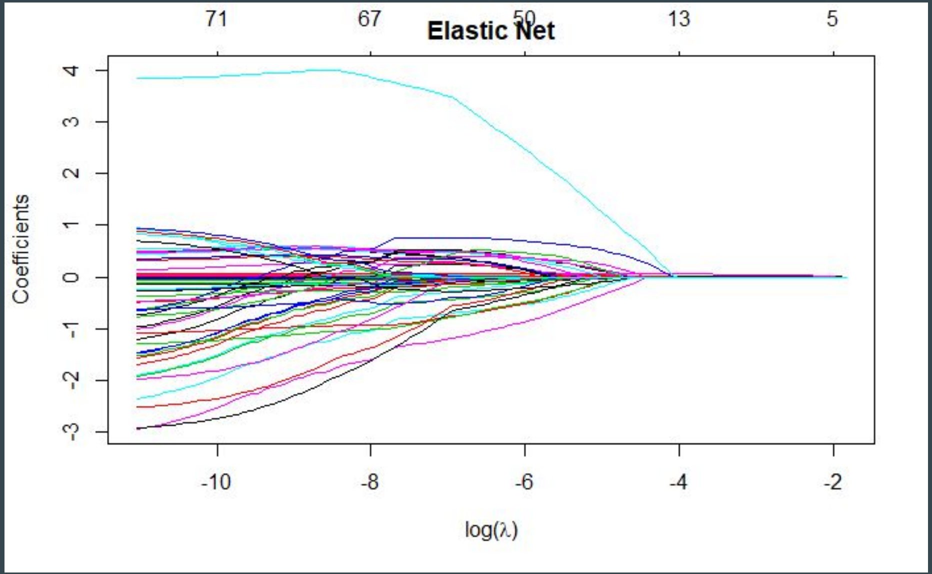
Logistic Regression: Ridge CV

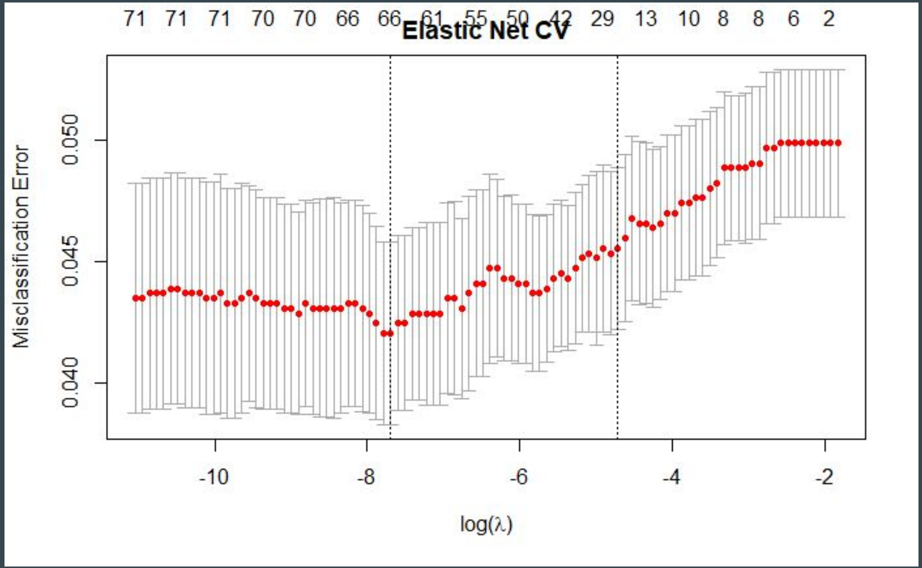




	Misclassification error <dbl>
Ridge Best	0.03952569
Ridge 1se	0.04150198

Logistic Regression: Elastic Net CV





Logistic Regression: Elastic Net CV

Best Lambda

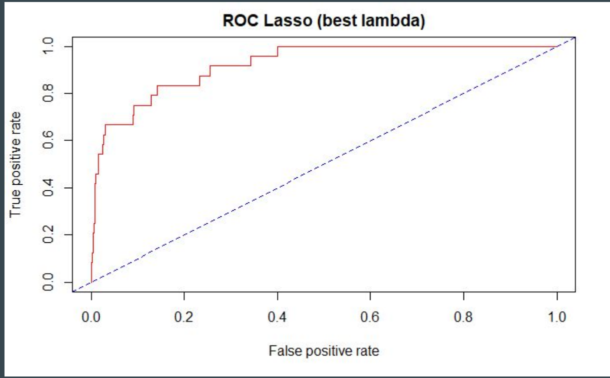
```
      enet.best
 [1,] "yearID"
 [2,] "G.x"
 [3,] "GS"
 [4,] "PO"
 [5,] "A"
 [6,] "E"
 [7,] "DP"
 [8,] "salary"
 [9,] "weight"
[10,] "age"
[11,] "AB"
[12,] "R"
[13,] "H"
[14,] "X3B"
[15,] "HR"
[16,] "RBI"
[17,] "SB"
[18,] "CS"
[19,] "BB"
[20,] "SO"
[21,] "IBB"
[22,] "HBP"
[23,] "SH"
[24,] "SF"
[25,] "GIDP"
[26,] "teamIDATL"
[27,] "teamIDBAL"
[28,] "teamIDBOS"
[29,] "teamIDCHA"
[30,] "teamIDCHN"
[31,] "teamIDCIN"
[32,] "teamIDCLE"
[33,] "teamIDCOL"
[34,] "teamIDDET"
[35,] "teamIDFLO"
[36,] "teamIDHOU"
[37,] "teamIDKCA"
[38,] "teamIDLAA"
[39,] "teamIDLAN"
[40,] "teamIDMIA"
[41,] "teamIDMIL"
[42,] "teamIDMIN"
[43,] "teamIDNYA"
[44,] "teamIDNYN"
[45,] "teamIDOAK"
[46,] "teamIDPHI"
[47,] "teamIDPIT"
[48,] "teamIDSDN"
[49,] "teamIDSFN"
[50,] "teamIDSLN"
[51,] "teamIDTBA"
[52,] "teamIDTEX"
[53,] "teamIDTOR"
[54,] "teamIDWAS"
[55,] "lgID.xAL"
[56,] "lgID.xNL"
[57,] "POS2B"
[58,] "POS3B"
[59,] "POSC"
[60,] "POSOF"
[61,] "POSP"
[62,] "POSSS"
[63,] "handComboBR"
[64,] "handComboLL"
[65,] "handComboRL"
[66,] "handComboRR"
```

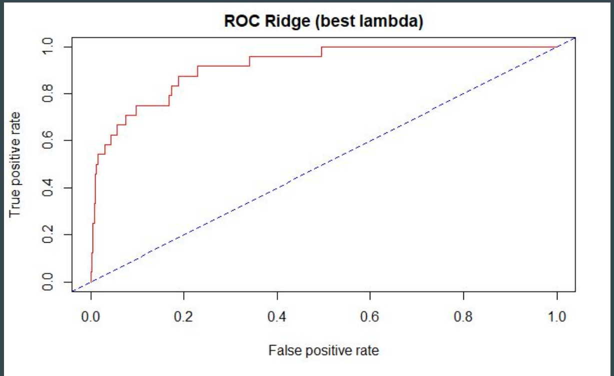
Best Lambda +1se

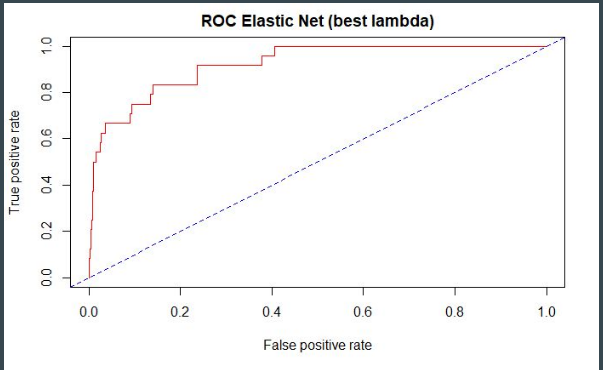
```
      enet.1se
 [1,] "GS"
 [2,] "InnOuts"
 [3,] "A"
 [4,] "E"
 [5,] "DP"
 [6,] "salary"
 [7,] "age"
 [8,] "R"
 [9,] "H"
[10,] "X2B"
[11,] "HR"
[12,] "RBI"
[13,] "SB"
[14,] "SO"
[15,] "IBB"
[16,] "SH"
[17,] "teamIDBAL"
[18,] "teamIDDET"
[19,] "teamIDKCA"
[20,] "teamIDLAA"
[21,] "teamIDOAK"
[22,] "teamIDSLN"
[23,] "POSP"
[24,] "handComboBR"
```

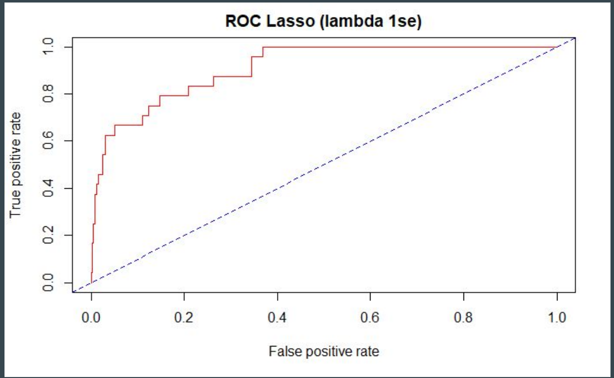
	Misclassification error <dbl>
ENet Best	0.03557312
ENet 1se	0.04150198

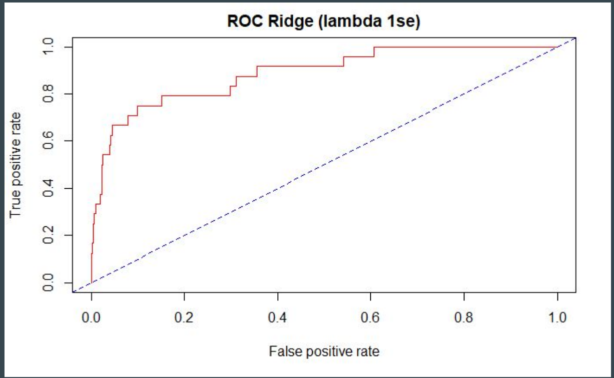
Comparison of ROC curves (they all look basically the same)

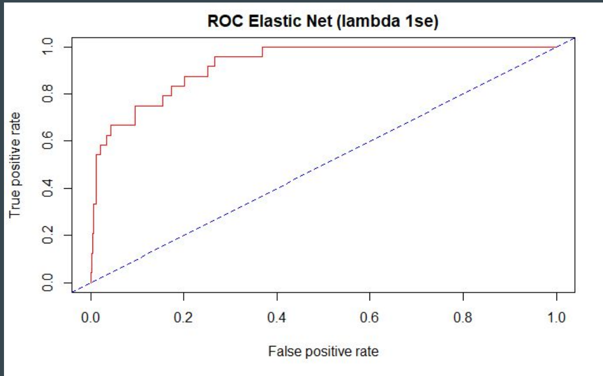












Conclusions: Logistic Regression for Awards

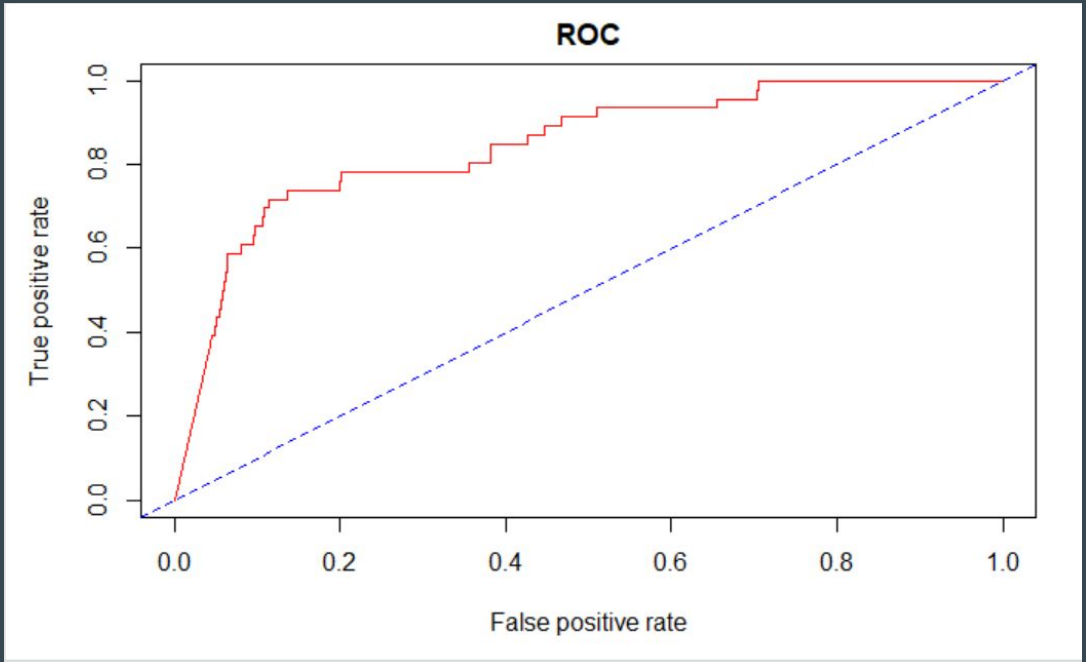
	Regressors <dbl>	AUC <dbl>	Misclass. Error <dbl>
Lasso Best	47	0.9228043	0.03754941
Lasso 1se	21	0.9108748	0.04347826
Ridge Best	71	0.9170989	0.03952569
Ridge 1se	71	0.8871888	0.04150198
Elastic Net Best	66	0.9213347	0.03557312
Elastic Net 1 se	24	0.9250519	0.04150198

Naive Bayes Full Model

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 610  12
         1 151  34

               Accuracy : 0.798
                 95% CI : (0.7686, 0.8252)
    No Information Rate : 0.943
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2235

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8016
            Specificity : 0.7391
         Pos Pred Value : 0.9807
         Neg Pred Value : 0.1838
             Prevalence : 0.9430
         Detection Rate : 0.7559
   Detection Prevalence : 0.7708
      Balanced Accuracy : 0.7704

       'Positive' Class : 0
```

AUC

```
[1] 0.8457264
```

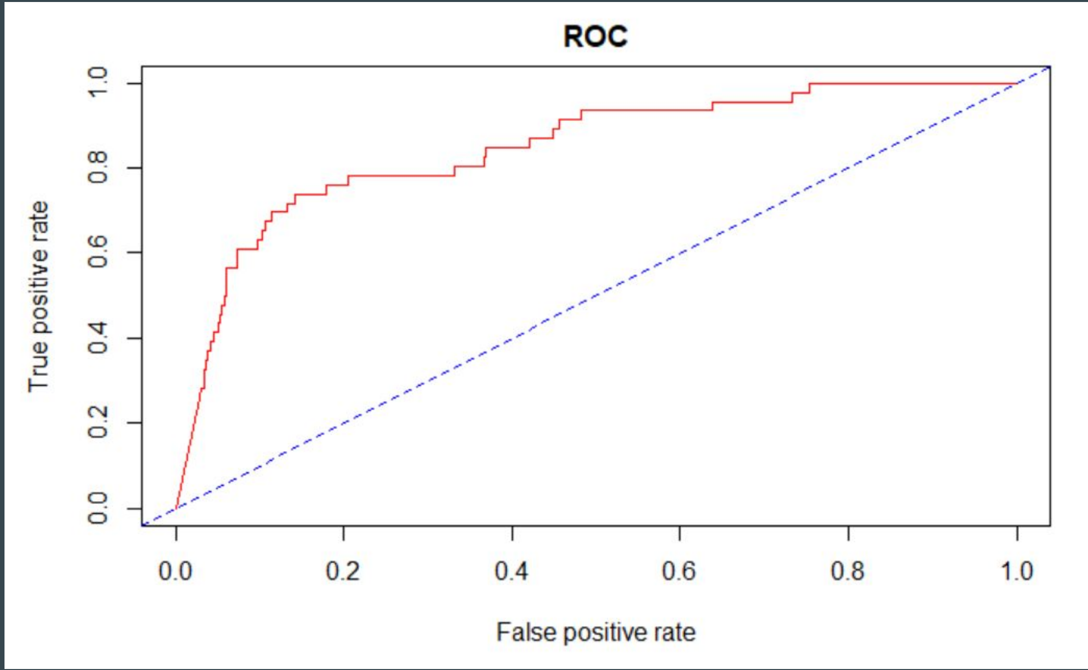


Naive Bayes Reduced Model

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 612  11
         1 149  35

               Accuracy : 0.8017
                 95% CI : (0.7725, 0.8287)
    No Information Rate : 0.943
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2345

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8042
            Specificity : 0.7609
         Pos Pred Value : 0.9823
         Neg Pred Value : 0.1902
             Prevalence : 0.9430
         Detection Rate : 0.7584
   Detection Prevalence : 0.7720
      Balanced Accuracy : 0.7825

       'Positive' Class : 0
```

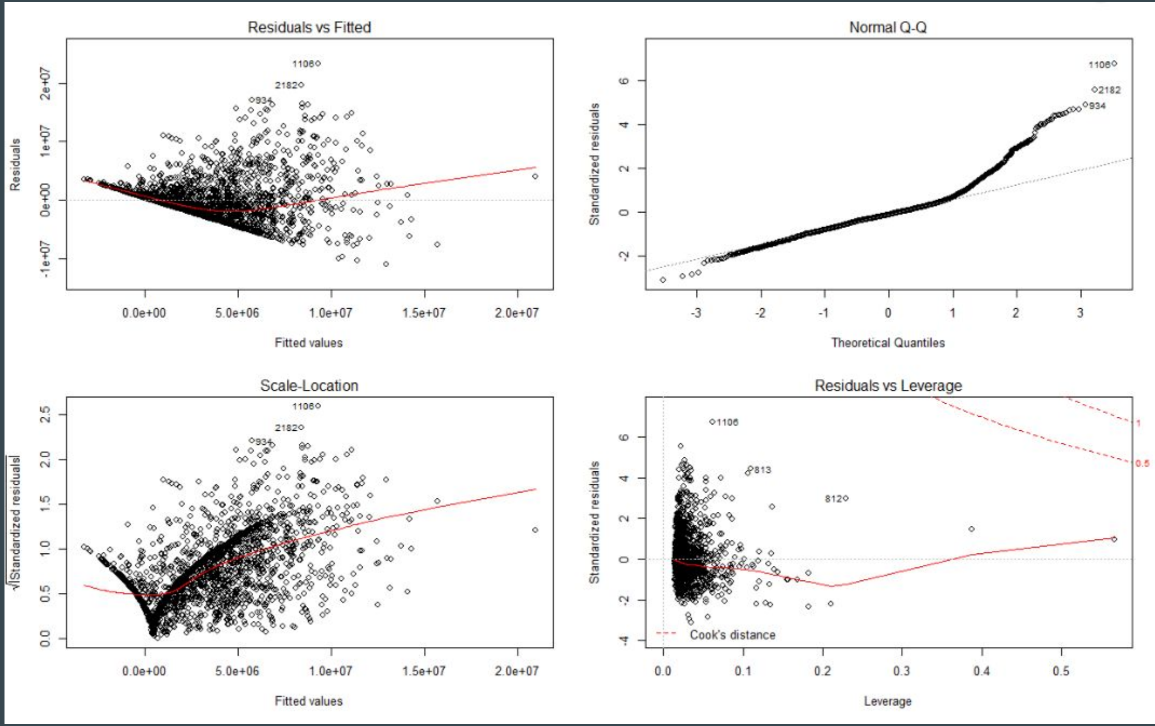
AUC

```
[1] 0.8469548
```



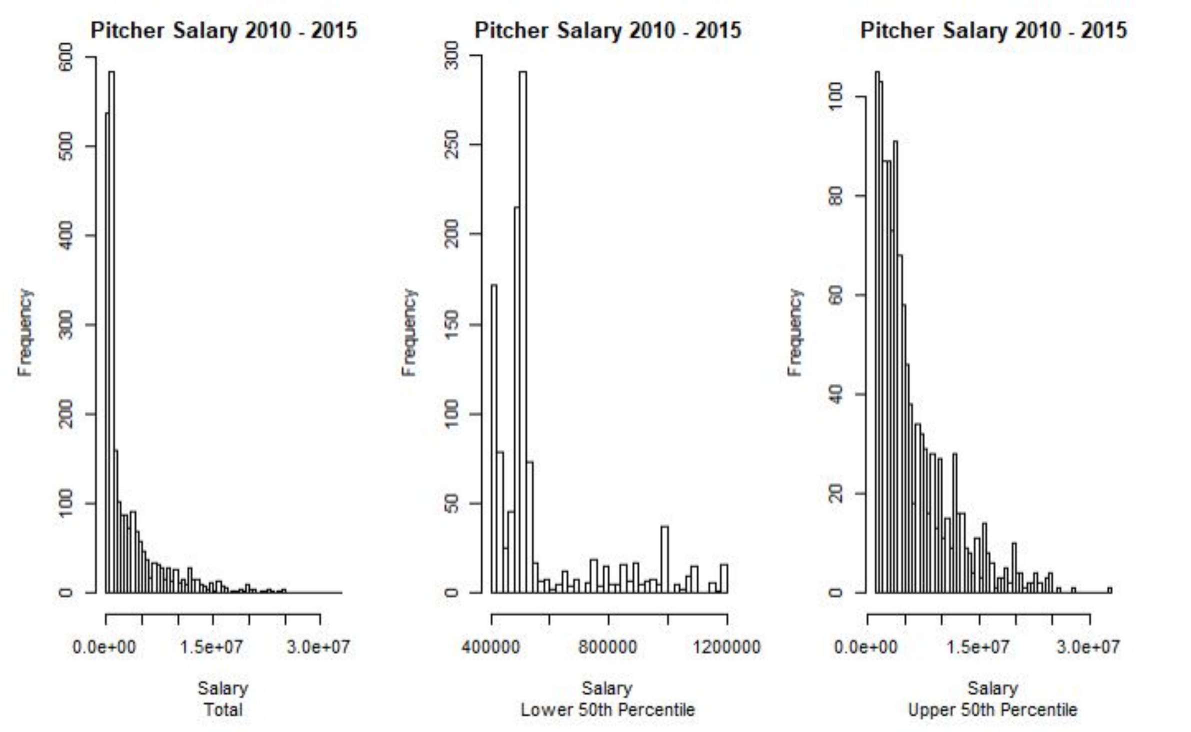
Pitcher Salaries

Multiple Linear Regression Model?...



...I don't think so

Pitcher Salary Distribution:



- Bimodal below Median
- Possibly Gamma above Median
- Violates Normality Assumption
- Perhaps GLM with Gamma Error, Poisson Regression?
- Decided on Logistic Regression: Above or Below Median

Multicollinearity:

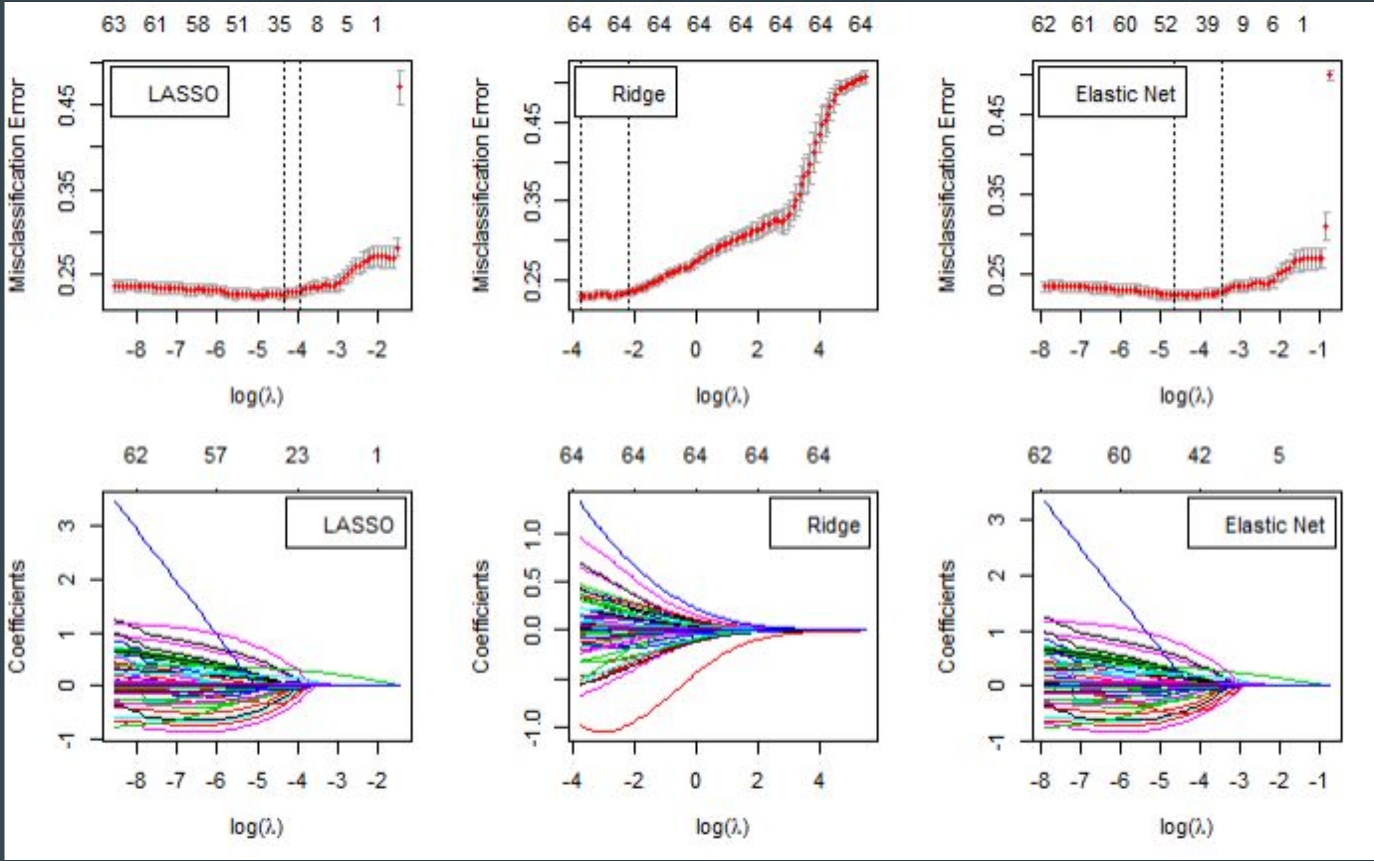
	GVIF	Df
yearID	1.171545	1
teamID	107.980900	30
lgID	36.762282	1
PO	3.276636	1
A	6.245499	1
E	1.422173	1
DP	1.775827	1
throws	1.366465	1
height	1.469680	1
weight	1.488256	1
age	1.153116	1
AB	37.079359	1
R.x	4.678012	1
H.x	12.176161	1
X2B	2.371188	1
X3B	1.189561	1
HR.x	1.995133	1
RBI	4.044633	1
SB	1.151083	1
CS	1.156434	1
BB.x	2.203889	1
SO.x	14.779505	1
HBP.x	1.191194	1
SH.x	4.243424	1
SF.x	1.245339	1
GIDP.x	1.520876	1
G.Bat	1229.677843	1
W	7.892225	1
L	6.348057	1
G	1243.986895	1
GS	91.186419	1
CG	3.681091	1
SHO	2.778764	1
SV	10.743192	1
IPouts	5387.060977	1
H.y	618.794523	1
ER	193.307494	1
HR.y	8.051504	1
BB.y	57.398088	1
SO.y	18.479856	1
BAOpp	3.173196	1
ERA	2.145984	1
IBB	1.620769	1
WP	1.960207	1
HBP.y	3.061349	1
BK	1.170837	1
BFP	10076.980668	1
GF	17.728092	1
R.y	219.879597	1
SH.y	3.373972	1

```
"GS"    "BFP"    "IPouts" "H.y"    "R.y"    "ER"     "SO.y"   "A"      "W"      "BB.y"   "AB"     "L"      "SO.x"   "H.x"    "CG"     "GF"     "G.Bat"
```

	GVIF	Df
yearID	1.092462	1
teamID	73.041535	30
lgID	36.333594	1
PO	2.858961	1
E	1.360894	1
DP	1.649132	1
throws	1.199673	1
height	1.445757	1
weight	1.460545	1
age	1.100903	1
R.x	3.363498	1
X2B	2.136739	1
X3B	1.151605	1
HR.x	1.938051	1
RBI	3.555429	1
SB	1.118960	1
CS	1.139796	1
BB.x	2.094206	1
HBP.x	1.178022	1
SH.x	3.028290	1
SF.x	1.224835	1
GIDP.x	1.387674	1
G	1.921870	1
SHO	1.307737	1
SV	1.332481	1
HR.y	2.970490	1
BAOpp	2.130492	1
ERA	1.910086	1
IBB	1.503353	1
WP	1.576547	1
HBP.y	1.985657	1
BK	1.151106	1
SH.y	2.847400	1
SF.y	1.964894	1
GIDP.y	3.554631	1

- Multicollinearity Issues Present as evidenced by VIF
- Removed Variables with correlation >= 0.75 with other Predictors
- League ID and Team ID are retained

CV for Lambda Selection



LASSO VS. Elastic Net

```
teamIDBAL teamIDCHA teamIDCIN teamIDCOL teamIDDET teamIDFLO teamIDLAA teamIDMIA teamIDMIL teamIDMIN teamIDNYA teamIDPHI teamIDSDN
        0         0         0         0         0         0         0         0         0         0         0         0         0
teamIDSFN teamIDSLN teamIDTBA teamIDTEX teamIDTOR     lgIDNL        PO         E        DP    weight       X2B       RBI        SB
        0         0         0         0         0         0         0         0         0         0         0         0         0
       CS      BB.x     HBP.x      SH.x      SF.x    GIDP.x        BK      SF.y
        0         0         0         0         0         0         0         0
```

```
teamIDFLO teamIDLAA teamIDMIA teamIDMIN teamIDPHI teamIDTEX teamIDTOR    lgIDNL        DP       RBI      SH.x      SF.x      SF.y
        0         0         0         0         0         0         0         0         0         0         0         0         0
```

LASSO

```
          Reference
Prediction   0   1
         0 150  48
         1  43 159

               Accuracy : 0.7725
                 95% CI : (0.7282, 0.8127)
    No Information Rate : 0.5175
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5448
```

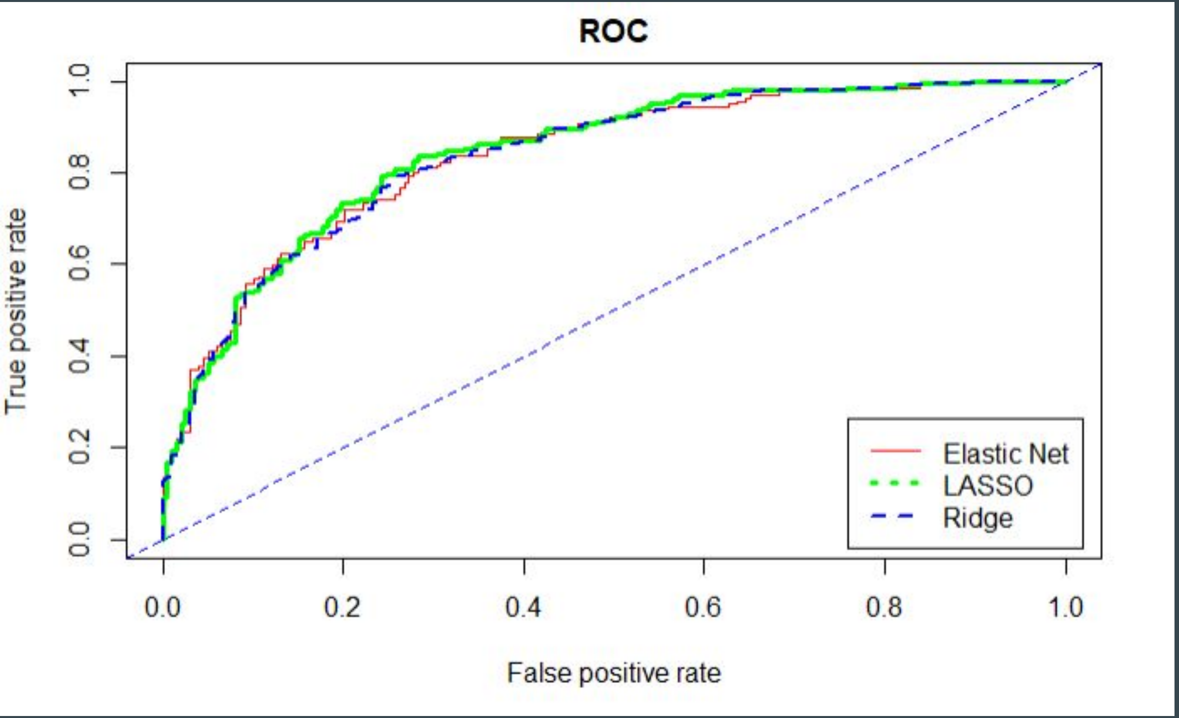
Ridge

```
          Reference
Prediction   0   1
         0 143  55
         1  42 160

               Accuracy : 0.7575
                 95% CI : (0.7124, 0.7987)
    No Information Rate : 0.5375
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5146
```

Elastic Net

```
          Reference
Prediction   0   1
         0 144  54
         1  42 160

               Accuracy : 0.76
                 95% CI : (0.7151, 0.801)
    No Information Rate : 0.535
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5197
```

ROC Curve



- Each Model Performs Reasonably Well on the Test Data
- Each Model is comparable in terms of AUC:
    - LASSO AUC: 84.3%
    - Ridge AUC: 82.6%
    - Elastic Net AUC: 83.6%
- ROC points to LASSO

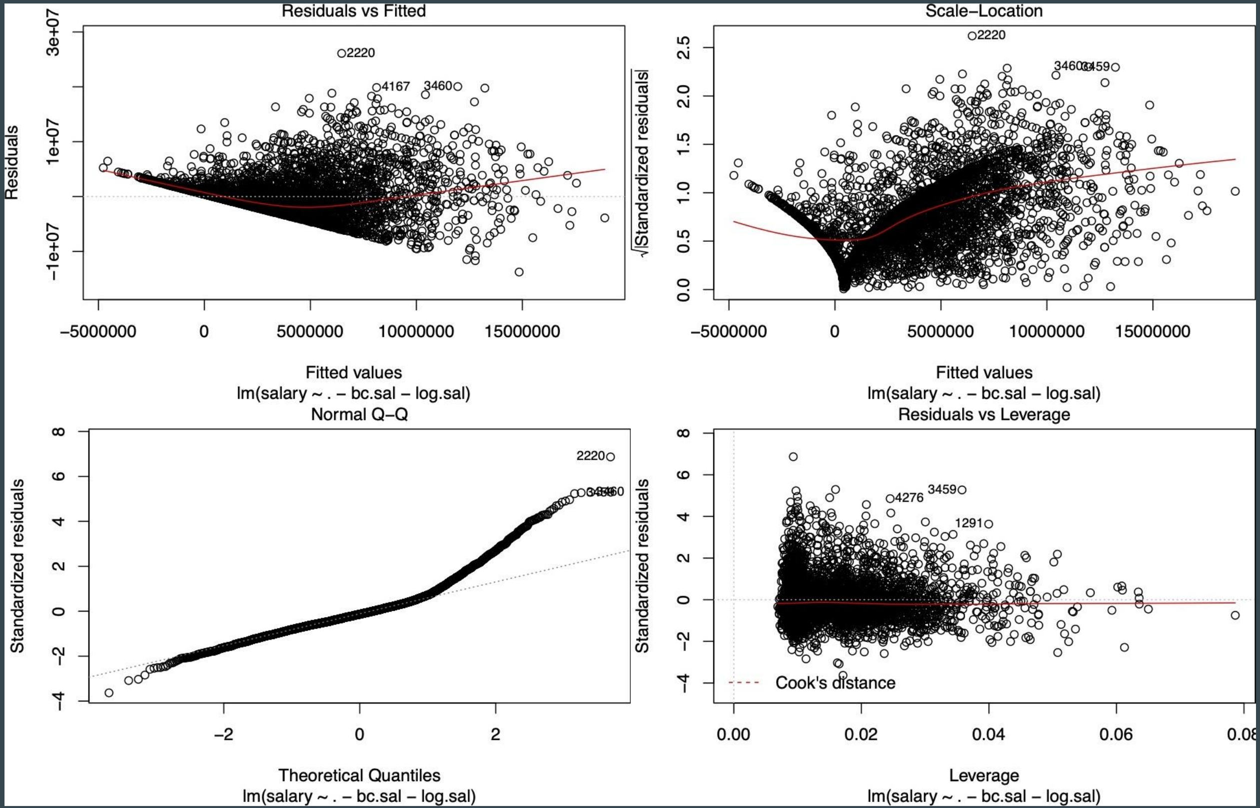
Conclusion

Feature	Coefficient
(Intercept)	-65.72272422
yearID	0.027094268
teamIDATL	-0.127754906
teamIDBOS	0.032070611
teamIDCHN	0.248088682
teamIDCLE	-0.084995161
teamIDHOU	-0.460343013
teamIDKCA	0.278368697
teamIDLAN	0.070507483
teamIDNYN	-0.34221448
teamIDOAK	-0.06248534
teamIDPIT	-0.250948293
teamIDSEA	-0.206128471
teamIDWAS	0.071299069
throwsR	-0.054049985
height	0.017374822
age	0.324490967
R.x	0.072615269
X3B	0.483900795
HR.x	0.052897799
G	-0.003003752
SHO	0.179814186
SV	0.044065605
HR.y	0.034776602
BAOpp	-0.293282746
ERA	-0.020337088
IBB	-0.057195333
WP	0.01498598
HBP.y	0.064372503
SH.y	0.030055974
GIDP.y	0.020651474

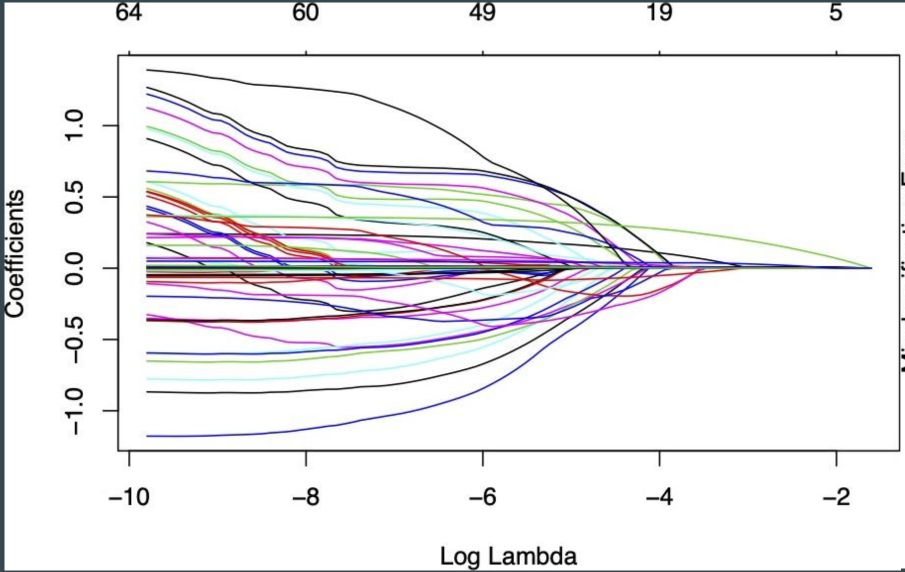
- Being Right-Handed slightly decreases your log odds of being above the median
- Log odds increase with age
- Pitching for the KC Royals increase your log odds the most while Houston decreases them the most
  - True in 2016 but no longer True! Better to pitch for Houston in 2019
  - Probably best not to include team in the future
- Shutouts increase log odds significantly (surprise, surprise...)
- Being able to hit triples greatly increases your log odds (x3b)
- Opponents Batting Average increasing greatly decreases log odds (BAOpp, surprise, surprise...)

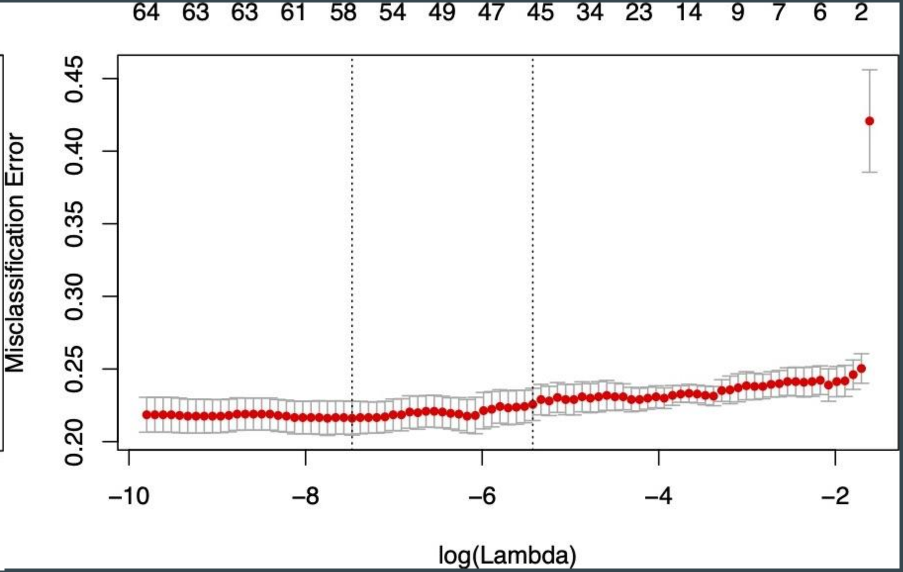
Fielder Salaries

Violation of linear regression assumptions



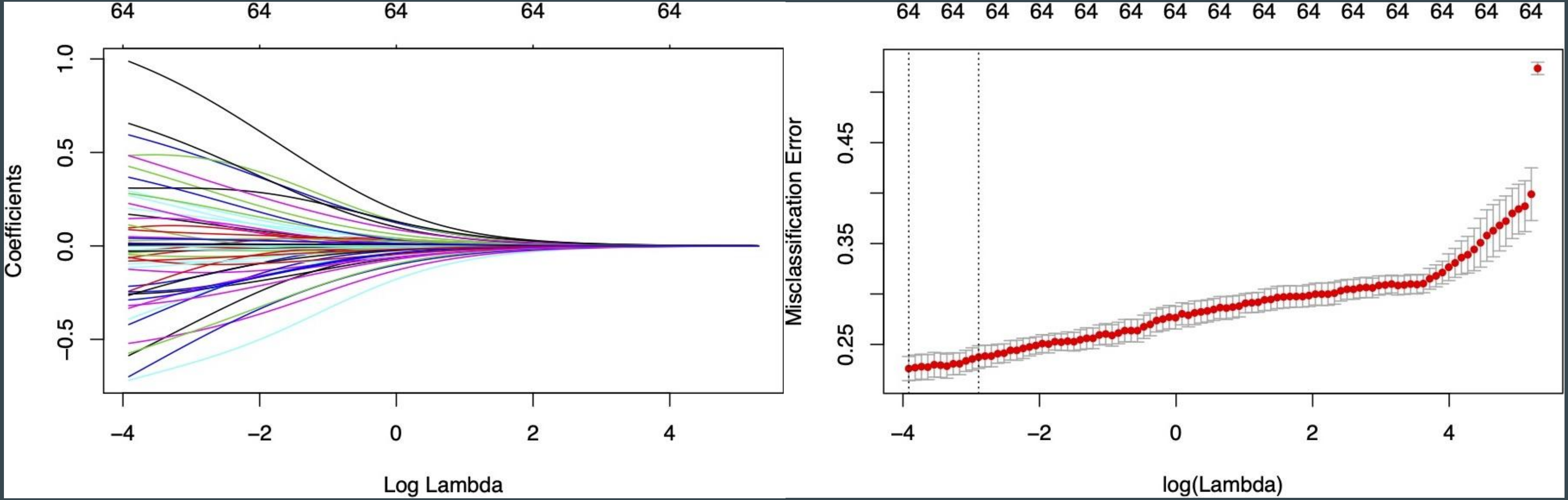
Logistic regression: Lasso



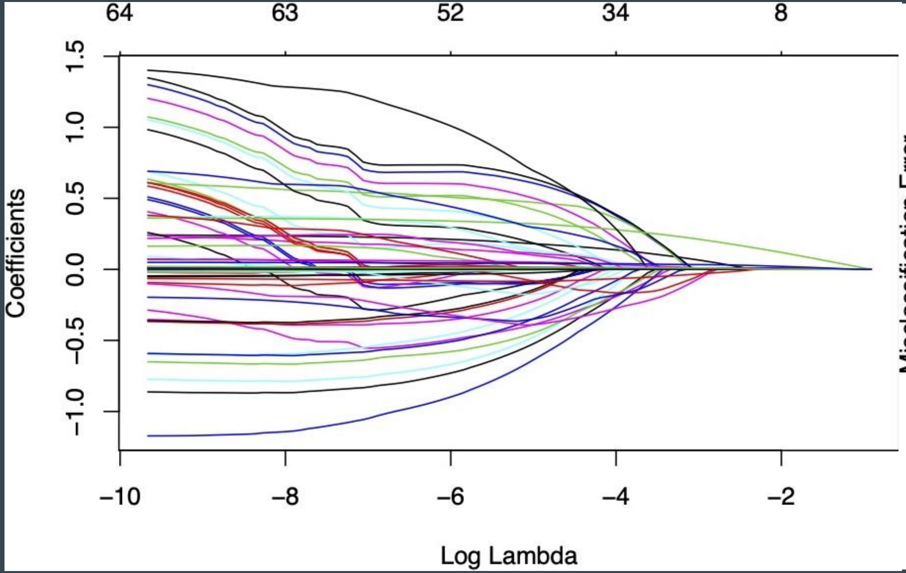


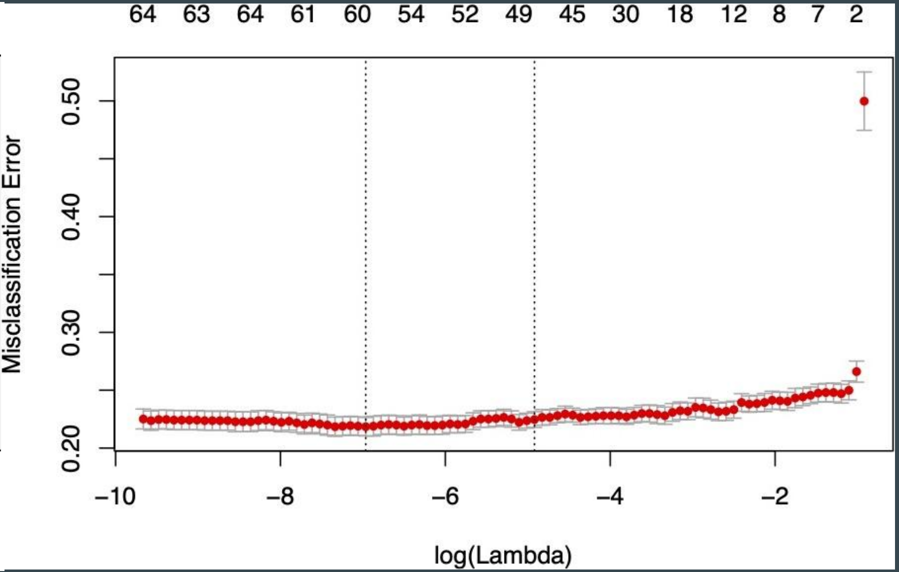

```
##    teamIDATL   teamIDCLE   teamIDLAA           E handComboLL          HR
##            0           0           0           0           0           0
##          HBP
##            0
```

Logistic regression: Ridge



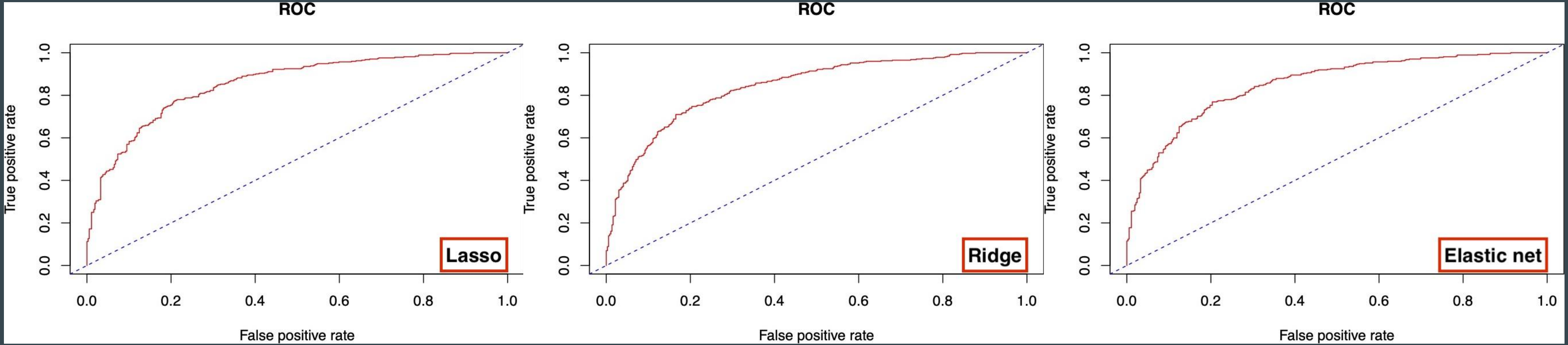
Logistic regression: Elastic net






```
## teamIDCOL teamIDLAA teamIDPIT        HR
##         0         0         0         0
```

Comparing the approaches



##	Regressors	Misclassification Rate	AUC
## Lasso	58	0.2346041	0.8700633
## Elastic Net	61	0.2375367	0.8706825
## Ridge	65	0.2404692	0.8643870

Conclusions

The bottom five: Oakland A's ($88M), Baltimore Orioles ($78M), Pittsburgh Pirates ($77M), Miami Marlins ($71M million) and Tampa Bay Rays ($61M).

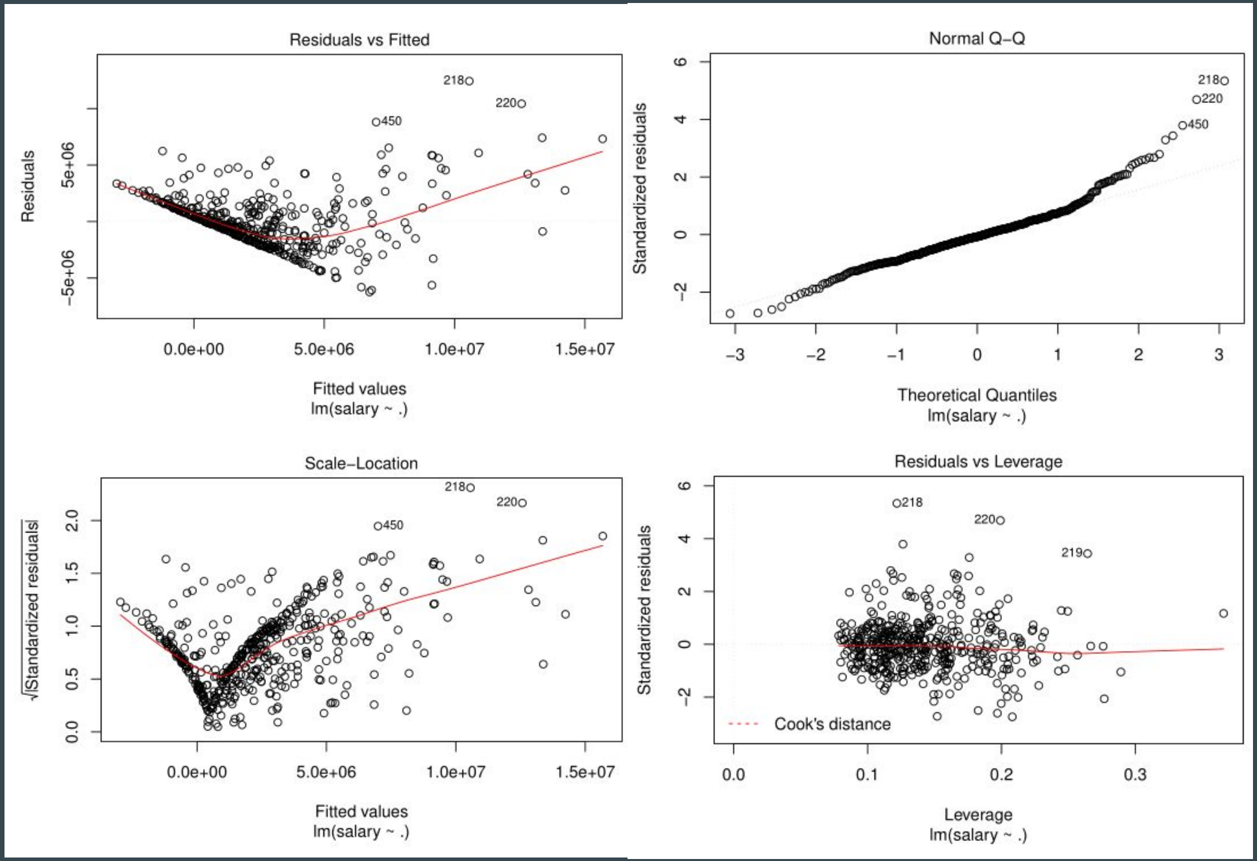
```
coef = lasso.field.coef[-1]

coef[which.min(coef)]

## teamIDMIA
## -1.080435

coef[which.max(coef)]

## handComboRL
##      1.2233
```

Negative impact:
Playing for Miami Marlins
Odds of making above the median salary decreases by over 60%.

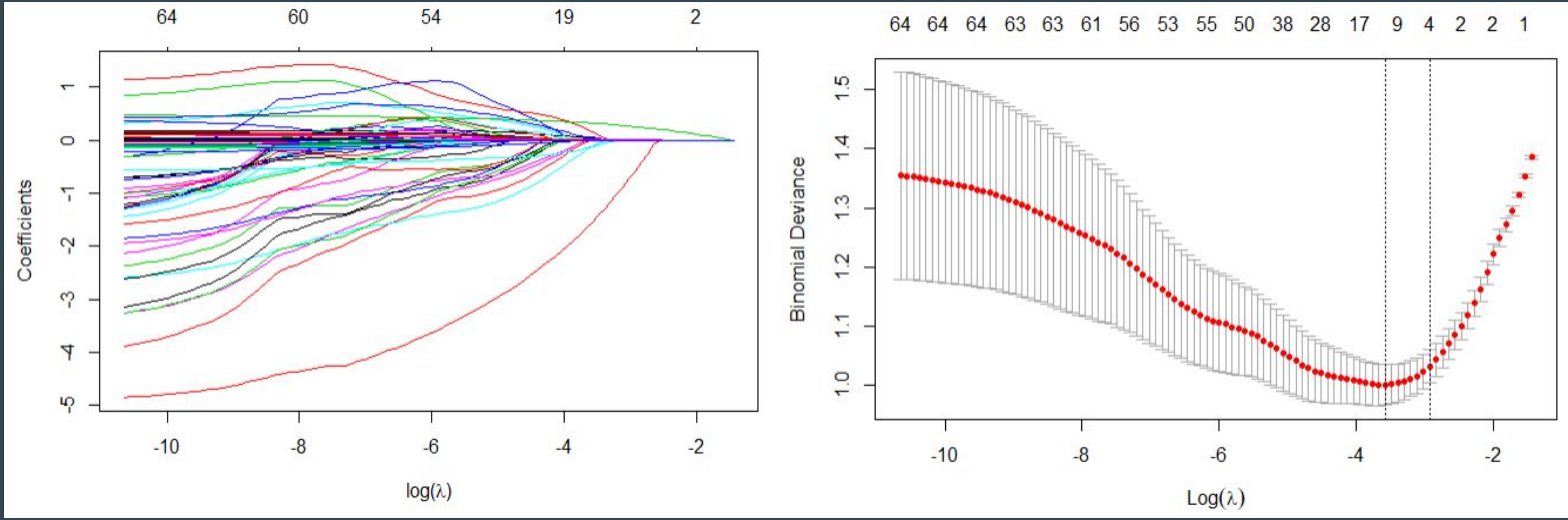
Positive impact:
Batting with right hand, throwing with left
Odds of making above the median salary increases by over 200%!

Catcher Salaries

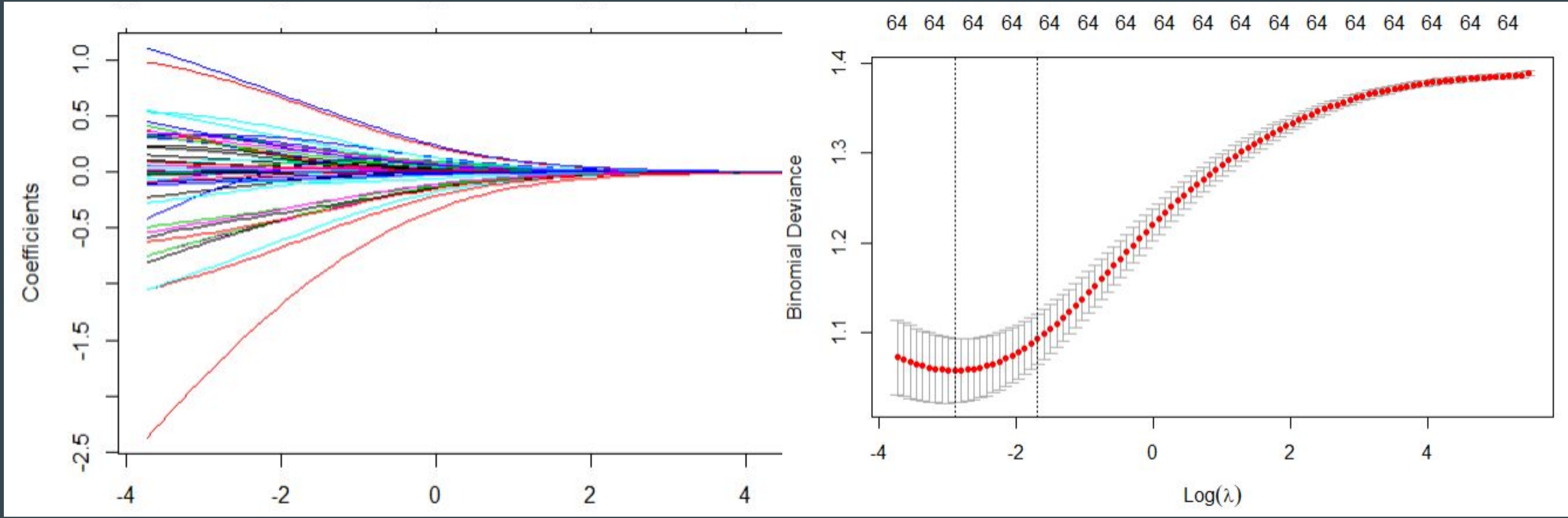
Linearity Assumptions



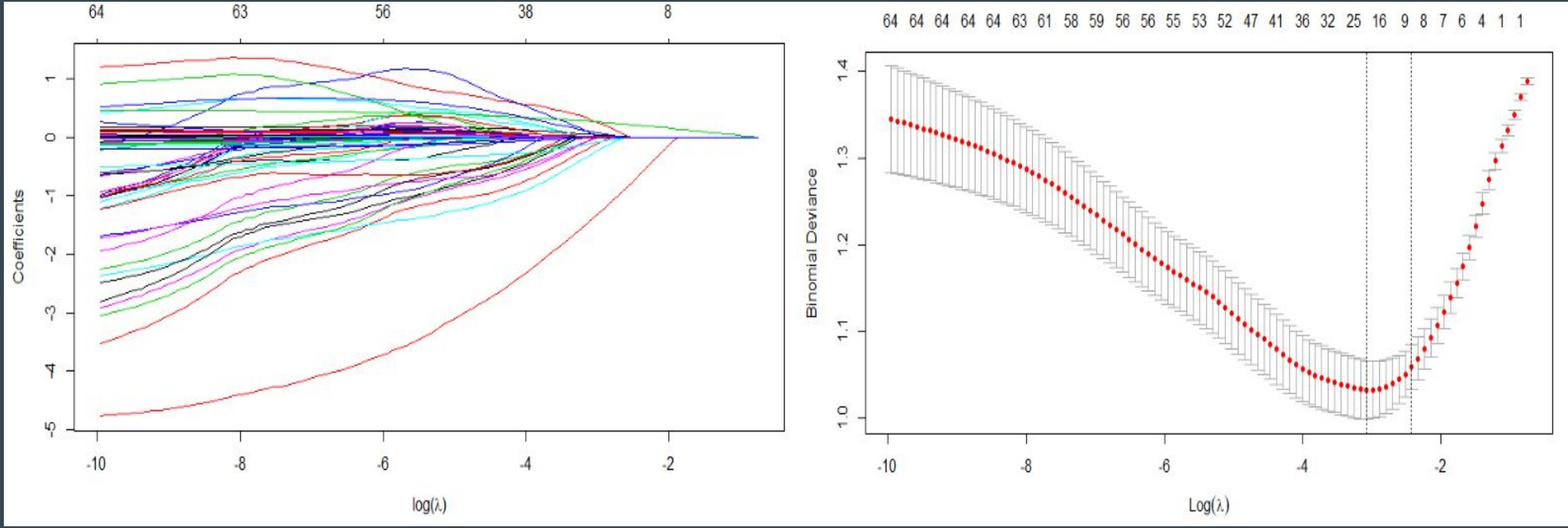
Lasso Regression



Ridge Regression

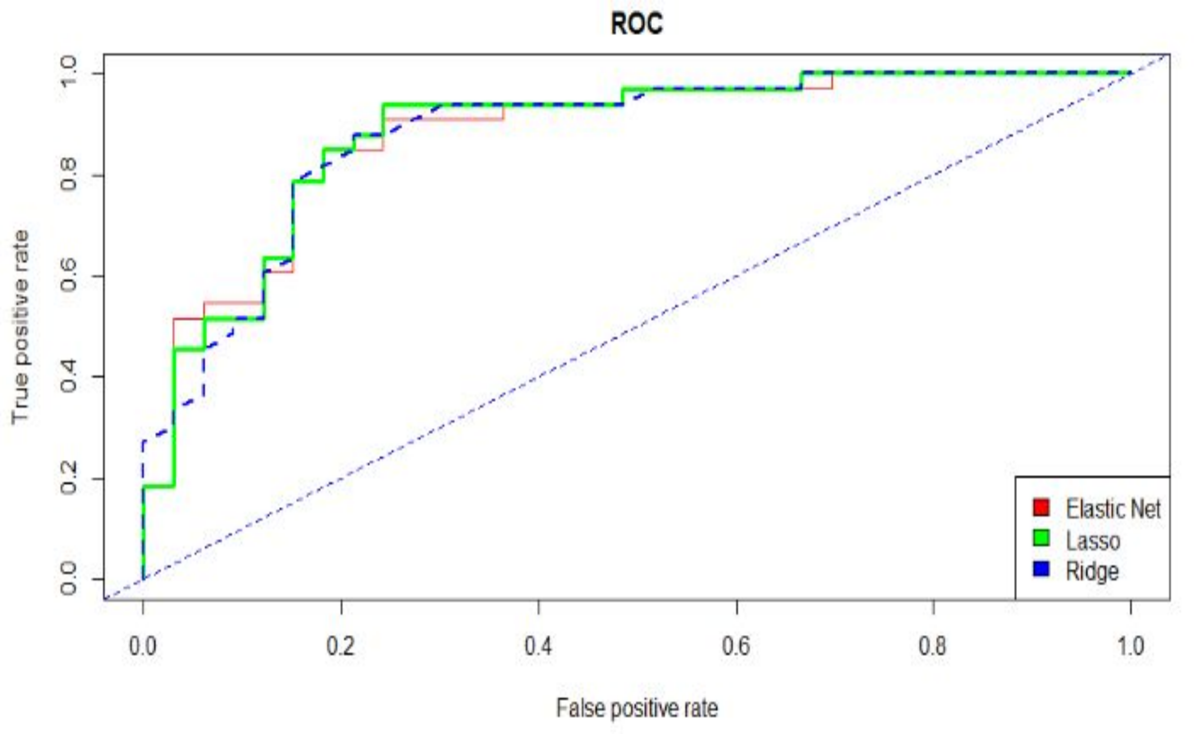


Elastic Net



64 63 56 38 8
Coefficients
1 0 -1 -2 -3 -4 -5
-10 -8 -6 -4 -2
log(λ)
64 64 64 64 64 63 61 58 59 56 56 55 53 52 47 41 36 32 25 16 9 8 7 6 4 1 1
Binomial Deviance
1.4 1.3 1.2 1.1 1.0
-10 -8 -6 -4 -2
Log(λ)

Comparisons



Regression	AUC
Lasso	88.34%
Ridge	87.14%
Elastic	87.79%

Lasso has the highest AUC at 88.34% indicating that the ROC points to Lasso Regression as the best shrinkage method

Conclusions

Catcher is one of the worst paid positions in Baseball

Ways to increase salary:

- Play for the Atlanta Braves
- Be good at batting

Things to Avoid:

- The New York Mets
- Being right-handed for batting and throwing

Variable	Coefficient
intercept	-4.61E+01
yearID	1.80E-02
teamIDATL	3.57E-02
teamIDNYN	1.30E+00
GS	3.03E-03
age	3.03E-01
handComboLR	1.69E-02
handComboRR	6.68E-02
AB	3.51E-03
H	5.58E-04