NBA Career Duration

An exploration into NBA Rookie’s 1st year performance impact on the longevity of their career

614 Term Paper

Authors: Ryan Hubley, Matthew Kilgariff, Matthew Sullivan

Abstract

This paper will aim to uncover statistical relationships between an NBA rookie’s first year stats and the length of their career. The dataset used in this analysis was found on Kaggle and was originally derived from stats found on nba.com. The paper will begin with a careful data exploration to uncover any correlations between our features and any other inherent patterns. It will then move to analyzing a linear regression fit, along with binary classification with logistic regression. The goal of the fit produced by linear regression is to predict career length, given rookie year data. We found this fit to be not very meaningful, which led us to use the logistic regression models that predict career length over/under 5 years, as a binary label. Through omission of certain data points and feature engineering, we were able to produce an accurate logistic regression model that was able to predict whether or not a rookie’s career would last 5 years or more.

Introduction

To play professional basketball in the NBA is no easy feat: only about 4% of draft-eligible college basketball athletes in the NCAA are selected to go pro[1]. Beyond that, even fewer professional basketball players maintain a career longer than 5 years. As an incoming rookie to the NBA, it might be prudent to understand, based on basketball players' historical performance within their first year on the court, what stats will be most important in optimizing their career longevity.

Our main goal will be to develop a model using rookie player statistics that can be used to predict whether or not a rookie player's career will extend beyond 5 years or not. We will be looking to understand what statistics are most important to a player in achieving this goal, as well as how those statistics interact with each other.

The dataset we explore consists of rookie basketball player historical performance between the years of 1980 and 2016. It contains 22 variables, including games and minutes played per game, points per game, shot data (attempts, successes, success rate for field goals, 3-pt shots, and free throws), game statistics (assists, steals, blocks, etc.), and a binary variable indicating if the player's career extended beyond 5 years or not. Note that these are all per-game averages across the first year of performance.

Some work has been done to shed light on this subject.

- Coates and Oguntimein (2008) look at NBA career length as a function of draft placement, and draft placement itself as a function of college player productivity. They find that college productivity explains NBA career productivity well, even after controlling for draft position.
- Martin, Arundale, Kluzek (2021) examine the effect that injury and illness in the rookie year has on career longevity in the NBA. They find that injury and illness alone do not account for career longevity; rather, that career longevity must be multifactorial.

This work suggests that a relationship may exist between NBA rookie year performance and career longevity. Our research will seek to uncover that in the given data.

Data Exploration

The dataset contains 22 features, including one index, one feature for the player's name, 18 performance statistics, and our success metric for a career 5 years or longer in the NBA.

Taking a look at the overall health of the dataset, we see that it is relatively clean: there are no cells containing no data. We remove duplicate lines for accidental repeated players, we ensure all numeric features are of the same data type, and check the success rate for our target metric to ensure it is not too small. We see that the probability for a rookie to maintain a career in the NBA for at least 5 years is 62%, which is a comfortable success rate for analysis later on.

[1] According to www.ncaa.org, of 4,181 draft-eligible Division I players in the NCAA, only 52 were chosen in the 2019 NBA draft.

Now, we examine our data for outliers. Knowing that in any sport, the NBA including, there may be a few stellar players whose statistics may be far beyond the average player's. Should we encounter this in our analysis, we will inspect the influence of these players on our statistics, but likely will not remove them from our analysis. After all, these stellar performers likely have careers in the NBA beyond 5 years, so this data will be important to our overall analysis.

By plotting each performance statistic by the index of the player, we look to see if there are any metrics in which we should examine further for potential outliers. See the plots in Table 1 of Appendix A.

For the most part, there is no cause for investigation into any outliers across the features. The only possible outliers that crop up are in the percentage metrics for field goals, 3-pointers, and free throws, where a few players achieved 100% and 0% shots made of shots attempted. These events make up a negligible portion of our data, so we assume they have negligible impact on our analysis and leave them be.

We also look at the correlation between each variable. Table 2 in Appendix A displays the correlation between all player statistics. Minutes played has a strong positive relationship with many variables, of which field goals and free throws stand out. Games played shares these relationships, although at a muted level. Turnovers also correlates strongly with activity during the game (minues played, shots attempted and made, etc.). We will examine efficiency statistics that penalize players for turnovers during gameplay.

In Table 3 of Appendix A, there are numerous matrices comparing features to our target metric. These features are continuous variables, so we group them up relative to their minimum and maximum values and compare between players who did and did not have a career lasting 5 years or longer.

Let's look at the relationship with our features to our target metric. Starting with games played and minutes played per game, we expect that those players who end up having longer careers would be playing more in their first year than those who do not. While not monotonic, the densities of games played and minutes per game do increase in their upper tails for players with longer careers, with higher mean values than those with shorter careers. This leads us to conclude that larger values of games played and minutes played increase probability of a longer career.

Points will likely be an important factor then; with more time in the game, we might expect players who end up having longer careers to be scoring more points in their first year. And we see exactly that: players with longer careers are likely to be scoring more points on average per game. Looking at games played and points per game together, we see clearly that players who score more average points per game are likely to play more games in their first season.

From this, we might suspect that shots made are important features to look into next. To look at the shot data (field goals, 3-points) together, keeping free throws separate, we combine them into a single shot metric. This will weight 3-points made as 1.5x the importance of field goals made, as it is in gameplay. From their respective matrices, that much is clear: players with longer careers have higher mean shots made and shots attempted. This could be due to greater playing time, as shown before, but there might be unseen skill variables.

To investigate that, let's look at the percentage of shots made. Interestingly, the densities between the two groups are actually similar, with the shot percentage marginally better for players with longer careers. However, this makes sense in practice: players with longer careers and marginally better shot percentages play more games and longer in games on average, leading to more shots attempted and made, finally to more points per game. We might expect shot percentage to not be a standout predictor for career length, though.

Shots aren't the only metric that defines a quality player, though. Looking at rebounds, for example, we see again that players with longer careers perform better. While this is an important statistic in leading to more points scored, the difference between the two groups is marginal, so we might not expect it to be a good predictor.

Now onto assists, blocks, steals, and turnovers. From the matrices, we see that the densities are not too different between groups, however the same trend is still present with the greatest difference in assists. One thing to note here is that an increase in turnover rate is not a positive; an increase in turnovers means the opposing team has a greater chance to score. This is likely a function of playing more games and having more playing time, but should be noted as we move to model fitting.

After further investigation, we were able to add the actual number of years the rookie's career lasted to our dataset (not just the binary indicator for over/under 5 years). Briefly, we wanted to discuss some of our findings with the addition of this variable.

The first thing we looked at with this variable was the distribution. Table 4 displays a right-skewed histogram, highlighting the asymmetric nature of the number of years played.

We also uncovered a glaring issue with our dataset with the addition of years played; rookies who were drafted in 2016, for example, will have a max of 1 year for years played (data was collected in 2016). Similarly, rookies from 2015 will have a max of 2 years, rookies from 2014 will have a max of 3 years, and so on. Table 5 shows this problem, graphing the number of years played versus year drafted. There is a clear boundary on the right side of the graph, which is this constraint. To deal with this problem, we will likely omit the players from 2013-2016, as they have not been in the NBA long enough for us to have an accurate number for years played. The tradeoff is that we are decreasing our sample size, but we feel that the accuracy of our fits will be compromised with the presence of these samples.

The addition of year drafted also allows us to examine how the statistics of rookies evolved. For most statistics, there is no discernable trend over the years. This is especially true in defensive stats like blocks, steals, and defensive rebounds. Table 6 in Appendix A shows the only statistics with a noticeable trend: three-pointers attempted and three-pointers made. Surprisingly, three-point percentage does not seem to be increasing, suggesting that the increase in three-pointers made is just a result of increased volume. With the success of the "Splash Brothers" Steph Curry and Klay Thompson in the mid-2010s, the rest of the NBA was forced to adapt to the high-scoring three-point offense. In our data, there were zero players drafted in 2016 who did not attempt a single three-pointer. Compare this to 1986, where there were 16 different rookies who never once shot behind the three-point line.

The modern day NBA looks a lot different than it did a few decades ago, and thus, there may be inaccuracies in our predictions for recent rookies. Even for "big men", there is an expectation that young seven-footers have some range. Mid-range shooters and true centers are endangered, and the inclusion of hall-of-famers like Michael Jordan, Kobe Bryant, Shaquille O'Neal, and Patrick Ewing in our analysis may misrepresent predictions for modern day rookies.

Model Selection and Fitting

There are a few options for the response variable (or target variable), which will determine which type of model we want to fit. Given the number of years a rookie's career lasted (integer) we could choose to fit a linear regression model predicting the number of years, or we could choose to fit a logistic regression model by deriving categories from the years played. For example, the over/under 5-year indicator (mentioned above) can easily be derived from the number of years played. In this phase of our project, we will aim to explore each of these different types of fits.

First, we will look at a linear regression fit to our data, aiming to find some sort of relationship between the rookie's first season stats and the number of years they will play in the NBA. This fit will be done with data that omits rookies that were drafted from 2000-2016. This is a large sample to omit, but Table 5 shows us that at year 2020, the "cap" on years played disappears. This model uses all of the available features.

```
Residual standard error: 4.269 on 571 degrees of freedom
Multiple R-squared:  0.2919,	Adjusted R-squared:  0.2671
F-statistic: 11.77 on 20 and 571 DF,  p-value: < 2.2e-16
```

The results of the linear regression model are shown above (full details in Appendix B Table 5). They indicate that the overall fit is not so great (Adjusted $R^2 = 0.27$). This makes sense because the overall range of years that we are trying to predict is quite large (0-20 years), and there are a wide variety of external factors that are not taken into consideration, like injuries. It is also worth noting that the coefficients with significant p values (< 0.05) were games played, 3-point attempts, and 3-point percentage. These results will motivate us to shift our focus to predicting categories in order to make the predicting a little easier.

Next, we will look at a binomial logistic regression fit to our data. We define our "success" class as playing for at least 5 years and our "failure" class as playing for less. The original data set has 62% of players lasting for at least 5 years, so we will use this percentage as a baseline accuracy for our models. We will use Akaike Information Criteria (AIC), ROC curves and associated AUC to compare our models, as well as residual plots and Wald statistics for feature selection and transformations. Again, we remove rookies from 2013 to 2016, since the data on years played may be inaccurate.

To begin, we fit a naive binomial logistic regression on all 22 available variables, and create a model that is 73.25% accurate with an AIC of 1354.3, a null deviance of 1565.8, and a residual deviance of 1314.3. The model has a sensitivity of 46.83%, and a specificity of 86.56% (Appendix B Table 1).

Without any data manipulation or feature selection, we construct a model that beats our benchmark and has a high specificity. There are few false positives (111), so if the naive model predicts a player will have a long career, there is a good chance he did. However, it has a low sensitivity, and there are nearly double the amount of false negatives (218) than false positives. Some variables, however, are highly correlated with each other, similarly to the results in our data exploration.

We then try to reduce the number of parameters by selecting ones with a low p-value associated with the Wald statistic and ones that are somewhat uncorrelated with each other. We regress games played, minutes, field goal percentage, three-point percentage, assists, blocks, and turnovers. This new model performs similarly to the naive, with a slightly lower accuracy of 71.46% and a similar AIC, deviance, sensitivity and specificity. This was achieved with very low correlations between the seven variables.
The seven variables selected have a basis in reality. Games played and minutes per game both show how NBA-ready a rookie is, field goal and three-point percentage show how a rookie's shooting ability transition from college to the pressure and defense in the league, assists are a proxy for playmaking and selfishness, blocks fill in the gaps for defense, and turnovers quantify how many mistakes a rookie makes.

To assess if the relationship between each variable and prediction is linear, we will look at correlations and residual plots, and try to correct any nonlinearities. Table 3 in Appendix B suggests that there may be nonlinearity in blocks, three-point and field goal percentage, and transformations may be appropriate. We also add back in rebounds and free-throw percentage, as these are important aspects for shooters and non-shooters alike.

Our final model with the original data regresses games played, minutes, assists, rebounds, free throw percentage, field goal percentage squared, and logarithm-transformed three point percentage and blocks[2]. This results in an accuracy of 73.50%, a reduced AIC of 1345.8, a sensitivity of 46.34% and an increased specificity of 87.07%. The ROC curve has moved slightly, and the AUC increased to .761 (Appendix B Table 4).

One interesting column in our data is three-point percentage. Especially in early years like 1986, there were multiple players that either made zero or all of their attempted three's. In some years, over a dozen rookies did not attempt a single three. Removing these data points increased the proportion of 5+ year careers to two-thirds and decreased the accuracy.

Our dataset reports each point as

---

[2] $\beta_0 + \beta_1 GP + \beta_2 MIN + \beta_3 AST + \beta_4 REB + \beta_5 FT\% + \beta_6 FG\%_2 + \beta_7 \log(3PT\%) + \beta_8 \log(BLK)$

per-game, as in, the rebounds stats are rebounds per game, not total over the season. Because of this, players with the same production will have different stats if their minutes-per-game are different. To correct, we divide each column (except for games played, minutes, percentage-stats, and our target variable) by minutes. Now, each stat is listed on a per-minute basis and show how efficient a rookie is regardless of how much play time he got. This helps correct for whether a rookie is a starter or a bench player, among other factors.

The naive-per-minute model came with an accuracy equivalent to the “best” per-game model, with a higher AUC, lower AIC, lower deviance, and similar sensitivity and specificity (Appendix B Table 5).

With the assistance of the Wald test applied to each parameter, along with the tools used on the per-game data, we picked our best overall model to be regressed on games played, minutes, assists, rebounds, free throw percentage squared, field goal percentage squared, and log transformations of three point percentage and blocks[3]. These variables all had Wald statistics with p-values less than .1. This results in our best model with a 73.66% accuracy, an AUC of 0.767, an AIC of 1336.1, and a residual deviance of 1318.1.

The features of this best per-minute model (and best overall) are nearly identical to the “best” per-game model, except with a squared free throw percentage term. This suggests that the variables we did select are the most important indicators for whether or not a rookie has a long career.

[3] $\beta_0 + \beta_1 GP + \beta_2 MIN + \beta_3 AST + \beta_4 REB + \beta_5 FT\%_2 + \beta_6 FG\%_2 + \beta_7 \log(3PT\%) + \beta_8 \log(BLK)$

Conclusion

We were able to create a binomial logistic regression model, when using per-minute transformations of the data, that predicts whether or not a player lasted five years in the NBA with an accuracy of 73.65%.

All of the models we created had a low sensitivity, meaning they produced a lot of false negatives. As logistic regression produces a “probability” of being in one class over another, we used 0.5 as the cutoff for below and above five years. Our dataset had around 62% above five, when we should expect a roughly 50-50 split[4]. A different cut off may have produced models with higher sensitivity. Another possible explanation could be there was not enough negative data. We also believe the exclusion of recent rookies from 2013 to 2016 may mean predictions we make for new rookies are inaccurate, as the skill sets and statistics of a modern day NBA player have evolved.

In the future, we hope to use ordinal regression to predict how long rookies last, with potential levels at 1, 2, 3, 4, and 5+ seasons. We also hope to apply machine learning techniques like binomial and multi-class support vector machines, random forests, and k-nearest neighbors, and compare the results of machine learning with the results of the statistical models we created in this paper.

[4] The average NBA career lasts 4.5 seasons.

References

Coates, Dennis and Babatunde Oguntimein. “The Length and Success of NBA Careers: Does College Production Predict Professional Outcomes?” International Journal of Sport Finance 5 (2008): 4-26.

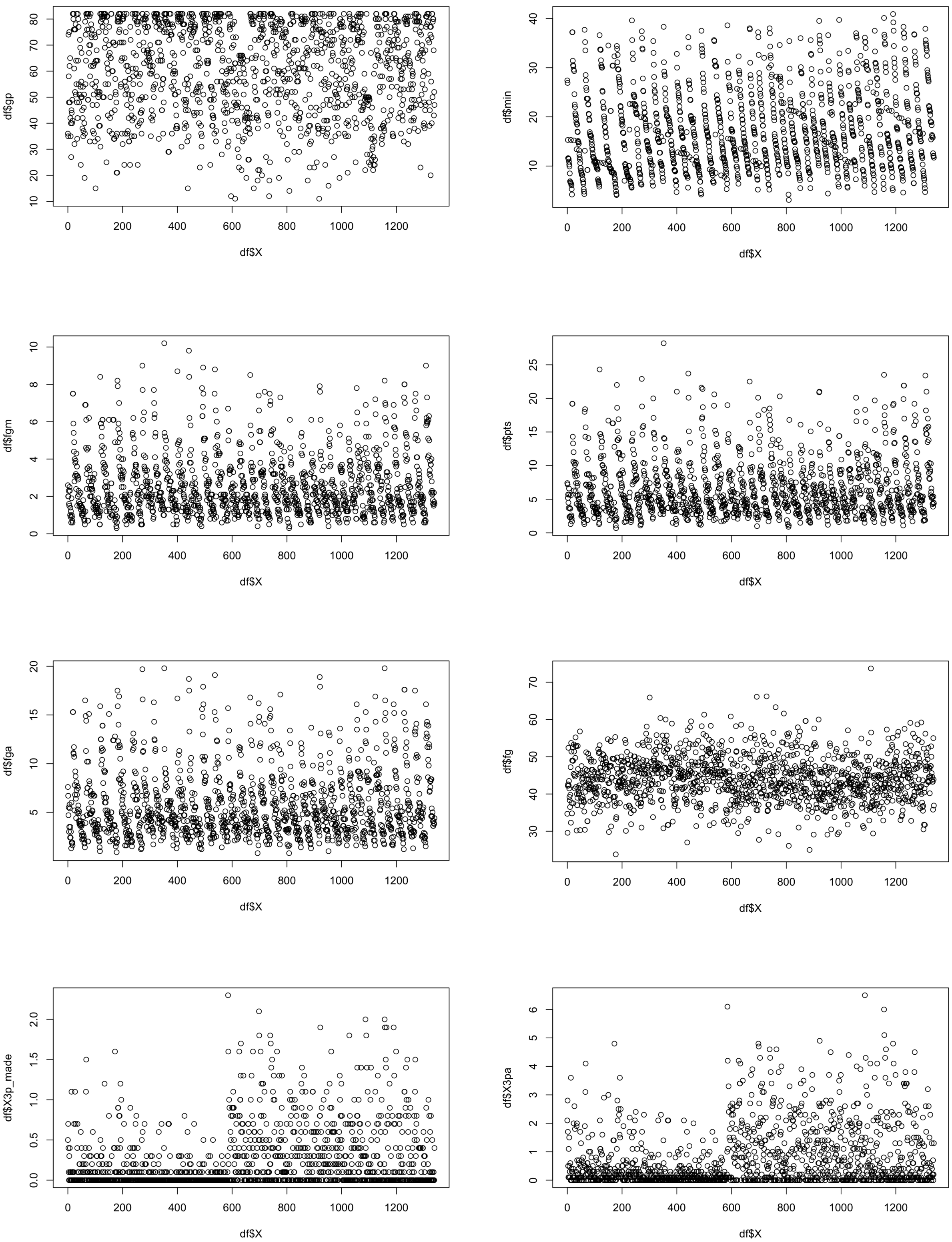
Eschlbeck, S. (2021, October). NBA Players Career Duration. Kaggle.com. Retrieved April 30, 2022, from https://www.kaggle.com/datasets/sveneschlbeck/nba-players-career-duration

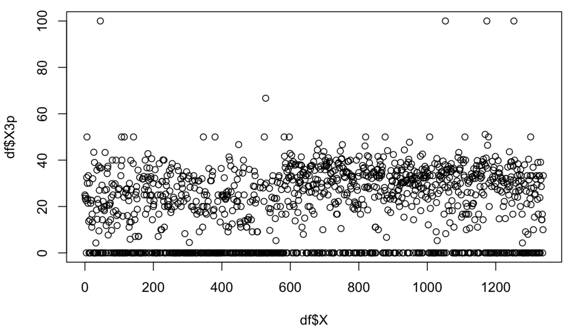
Lopez, A. J. (2010, June 10). Life after NBA comes sooner than many players think. NBA.com. Retrieved April 30, 2022, from https://www.nba.com/nuggets/features/junior_bridgeman_20100610.html

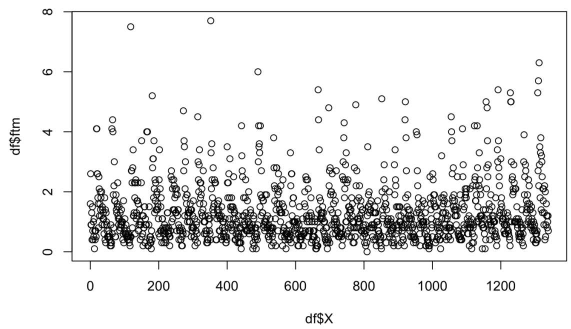
Martin CL, Arundale AJH, Kluzek S, Ferguson T, Collins GS, Bullock GS. Characterization of Rookie Season Injury and Illness and Career Longevity Among National Basketball Association Players. JAMA Network Open. 2021;4(10):e2128199. doi:10.1001/jamanetworkopen.2021.28199

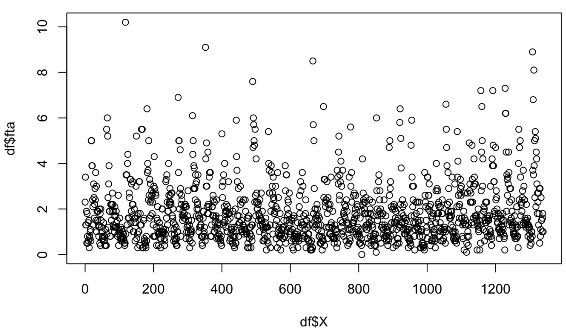
Appendix A

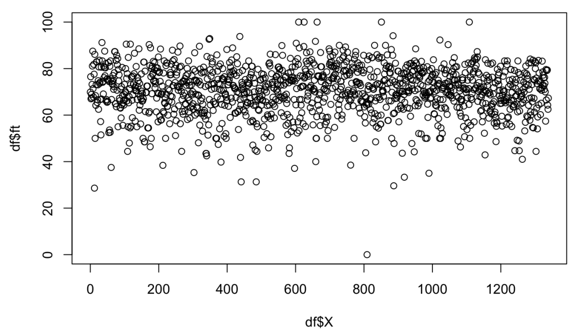
Table 1 (Performance Statistics by Index plots)

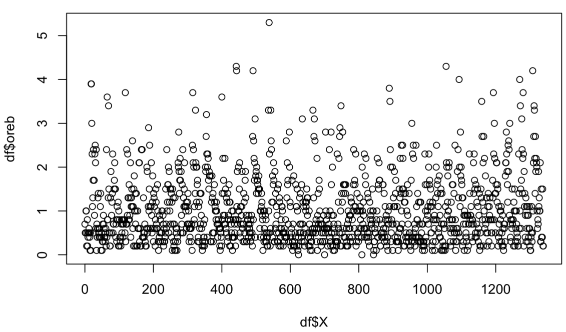


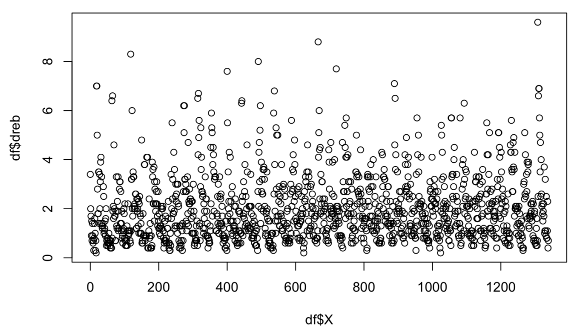

df$X3p
df$X

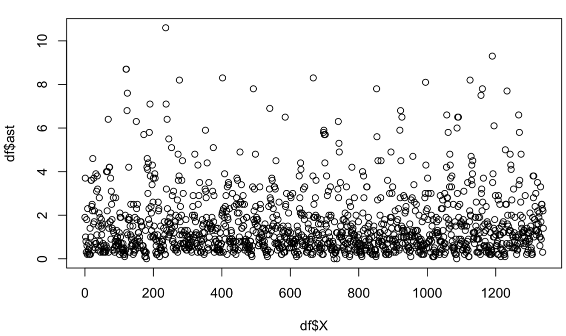

df$ftm
df$X

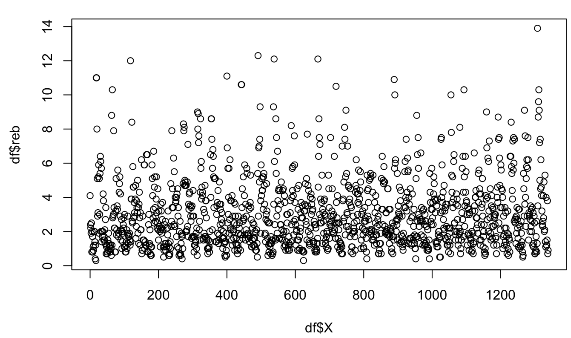

df$fta
df$X

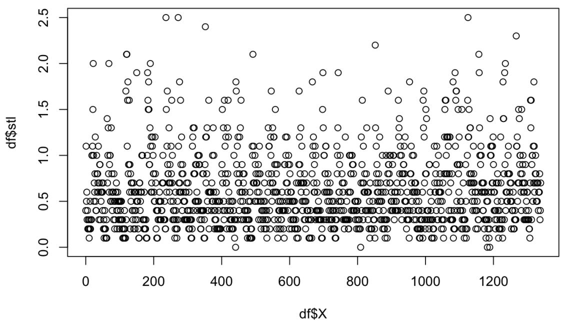

df$ft
df$X

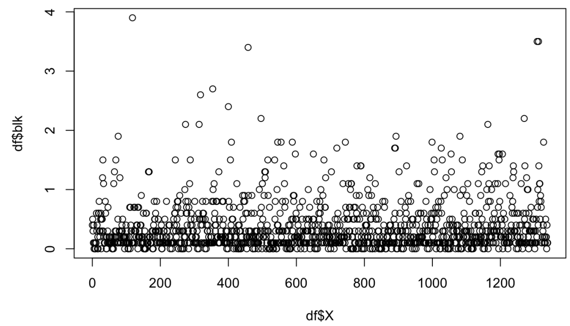

df$oreb
df$X


df$dreb
df$X


df$ast
df$X


df$reb
df$X





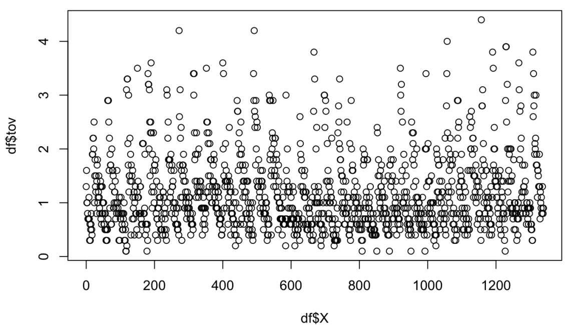


Table 2 (Correlation Plot)

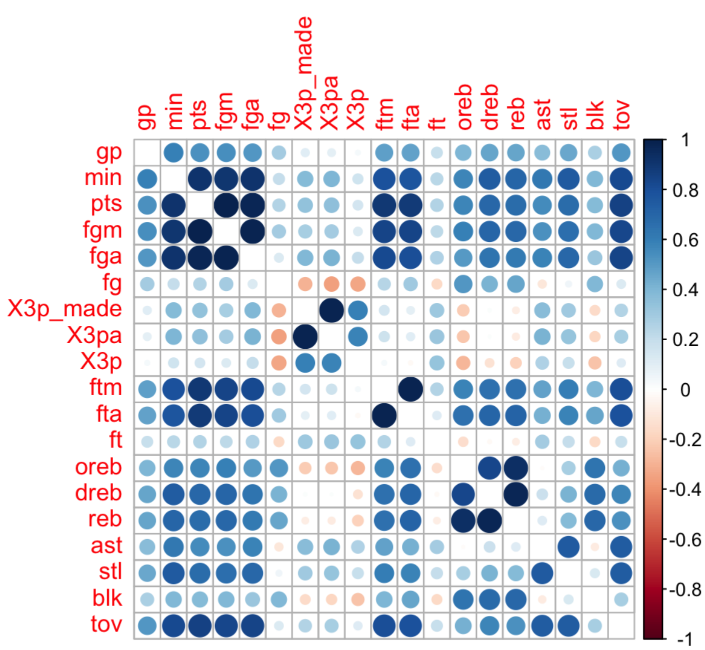


Table 3 (comparison matrices, with mean of group 0 and mean of group 1)

```
grouped_gp   0   1
  (10,15]    8   0
  (15,20]    7   3
  (20,25]   22   5
  (25,30]   13   8
  (30,35]   39  16
  (35,40]   55  36
  (40,45]   55  42
  (45,50]   49  54
  (50,60]  100 111
  (60,70]   72 137
  (70,80]   73 262
  (80,100]  16 157
[1] 51.5776
[1] 65.82671
```

```
grouped_min   0   1
    (0,5]    10   5
    (5,10]  140 115
    (10,15] 165 168
    (15,20] 100 174
    (20,25]  56 143
    (25,30]  23 104
    (30,35]  12  74
    (35,40]   3  46
    (40,45]   0   2
[1] 14.25226
[1] 19.69025
```

```
           grouped_pts
grouped_gp (0,5] (5,10] (10,15] (15,20] (20,25] (25,30]
  (10,15]      7      1       0       0       0       0
  (15,20]      9      1       0       0       0       0
  (20,25]     21      5       1       0       0       0
  (25,30]     16      5       0       0       0       0
  (30,35]     40     13       2       0       0       0
  (35,40]     74     17       0       0       0       0
  (40,45]     80     15       2       0       0       0
  (45,50]     61     29       9       4       0       0
  (50,60]    144     54       9       4       0       0
  (60,70]     82     92      25       9       1       0
  (70,80]     51    182      69      25       8       0
  (80,100]     6     86      48      21      11       1
```

```
grouped_pts   0   1
    (0,5]   319 272
    (5,10]  153 347
    (10,15]  28 137
    (15,20]   8  55
    (20,25]   1  19
    (25,30]   0   1
[1] 5.042829
[1] 7.8787
```

```
grouped_shotsm   0   1
        (0,2]  284 251
        (2,4]  174 320
        (4,6]   30 156
        (6,8]   19  72
        (8,10]   2  29
        (10,12]  0   3
[1] 2.290668
[1] 3.435319
```

```
grouped_shotsa   0   1
        (0,5]  271 293
        (5,10] 187 305
        (10,15] 34 147
        (15,20] 16  66
        (20,25]  1  19
        (25,30]  0   1
[1] 5.673969
[1] 7.899398
```

```
grouped_fta   0   1
    (0,2]   428 490
    (2,4]    72 262
    (4,6]     6  62
    (6,8]     2  12
    (8,10]    0   4
    (10,12]   0   1
[1] 1.321611
[1] 2.1284
```

```
grouped_ftm   0   1
     (0,2]  476 637
     (2,4]   29 163
     (4,6]    3  28
     (6,8]    0   3
     (8,10]   0   0
[1] 0.9233792
[1] 1.526955
```

```
grouped_shots   0   1
    (0,0.2]    27  13
    (0.2,0.4] 444 699
    (0.4,0.6]  36 116
    (0.6,0.8]   2   3
    (0.8,1]     0   0
[1] 0.3076925
[1] 0.3220481
```

```
grouped_ft     0   1
  (0,0.2]      0   0
  (0.2,0.4]    5   8
  (0.4,0.6]   95  97
  (0.6,0.8]  338 576
  (0.8,1]     70 150
[1] 0.6885855
[1] 0.7118339
```

```
grouped_rebs   0   1
     (0,2]   288 251
     (2,4]   166 314
     (4,6]    43 160
     (6,8]     8  68
     (8,10]    2  25
     (10,12]   2   9
     (12,14]   0   4
[1] 2.247544
[1] 3.516486
```

```
grouped_orebs   0   1
        (0,2]  487 700
        (2,4]   18 125
        (4,6]    0   6
        (6,8]    0   0
        (8,10]   0   0
        (10,12]  0   0
[1] 0.7182711
[1] 1.187726
```

```
grouped_drebs   0   1
       (0,2]  389 424
       (2,4]  106 302
       (4,6]   11  84
       (6,8]    3  18
       (8,10]   0   3
       (10,12]  0   0
[1] 1.531238
[1] 2.32864
```

```
grouped_ast   0   1
    (0,2]   421 600
    (2,4]    72 157
    (4,6]     9  46
    (6,8]     3  19
    (8,10]    1   7
    (10,12]   0   1
[1] 1.221022
[1] 1.752347
```

```
grouped_stl   0   1
      (0,1] 474 681
      (1,2]  32 139
      (2,3]   1   9
      (3,4]   0   0
      (4,5]   0   0
[1] 0.4982318
[1] 0.6921781
```

```
grouped_blk   0   1
      (0,1] 417 686
      (1,2]  15  71
      (2,3]   1   7
      (3,4]   0   4
      (4,5]   0   0
[1] 0.2534381
[1] 0.4391095
```

```
grouped_tov   0   1
      (0,1] 348 360
      (1,2] 138 340
      (2,3]  18  98
      (3,4]   5  30
      (4,5]   0   3
[1] 0.9422397
[1] 1.347533
```

Table 4 (distribution of Years variable)

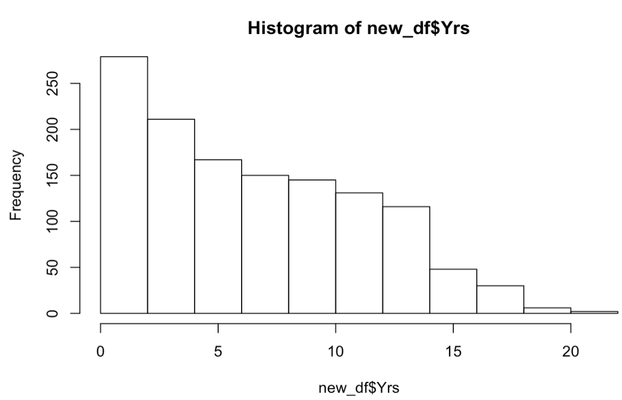


Table 5 (Years Played vs. Year Drafted)

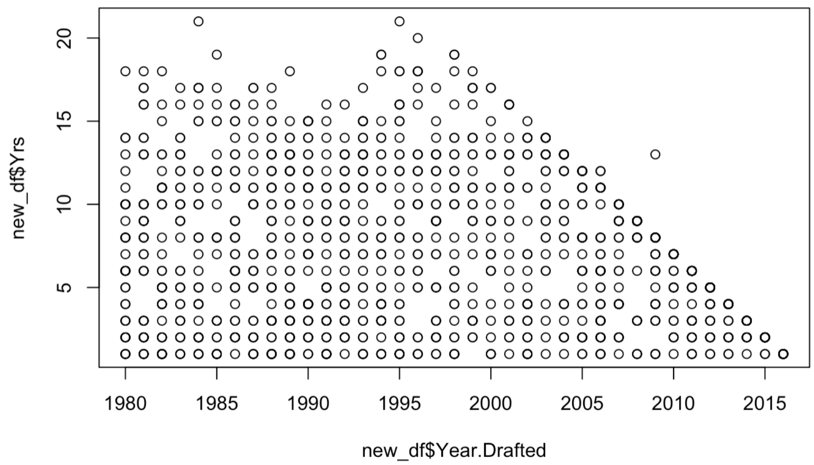
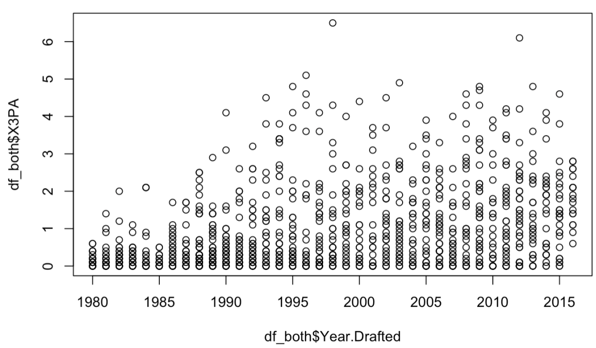
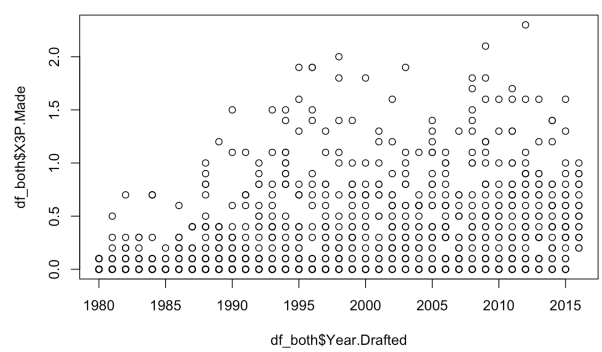
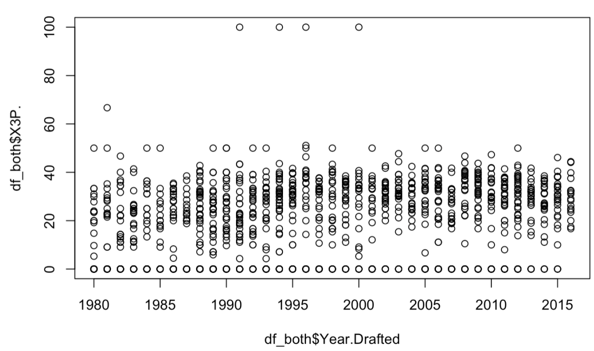


Table 6 (Three-point Statistics vs. Year Drafted)







Appendix B

Table 1 - Logistic Regression on all 22 variables

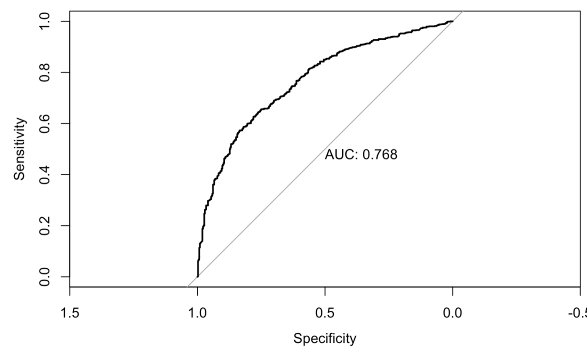
Regression Summary

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.970584   1.243542  -3.193  0.00141 **
GP           0.027480   0.005209   5.275 1.32e-07 ***
MIN         -0.074812   0.037409  -2.000  0.04552 *
PTS          0.013861   0.937465   0.015  0.98820
FGM         -0.144901   1.855451  -0.078  0.93775
FGA          0.140668   0.255643   0.550  0.58215
FG.          0.032805   0.022638   1.449  0.14730
X3P.Made     0.995309   1.460344   0.682  0.49552
X3PA        -0.136249   0.474233  -0.287  0.77388
X3P.         0.007063   0.005518   1.280  0.20057
FTM          0.591057   1.101320   0.537  0.59149
FTA         -0.427395   0.504563  -0.847  0.39696
FT.          0.009722   0.009977   0.974  0.32981
OREB        -0.171379   1.362268  -0.126  0.89989
DREB        -0.315298   1.360384  -0.232  0.81672
REB          0.681182   1.351121   0.504  0.61415
AST          0.364384   0.128804   2.829  0.00467 **
STL          0.189648   0.354822   0.534  0.59300
BLK          0.740163   0.315310   2.347  0.01890 *
TOV         -0.501880   0.300928  -1.668  0.09536 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1565.8  on 1229  degrees of freedom
Residual deviance: 1314.3  on 1210  degrees of freedom
AIC: 1354.3
```

Confusion Matrix

	Reference	
Prediction	0	1
0	192	111
1	218	709

ROC Curve



Correlations between features

	(Intercept)	GP	MIN	PTS	FGM	FGA	FG.	X3P.Made	X3PA	X3P.	FTM	FTA	FT.	OREB	DREB	REB	AST	STL	BLK	TOV
(Intercept)	1.00	-0.04	0.05	0.20	0.00	-0.65	-0.82	-0.09	-0.01	-0.13	0.10	-0.42	-0.58	-0.03	-0.02	0.02	-0.02	0.02	-0.03	-0.06
GP	-0.04	1.00	-0.18	-0.02	0.02	0.03	-0.05	0.00	0.05	-0.02	0.06	-0.07	-0.13	0.01	0.00	0.00	0.01	-0.02	-0.01	-0.06
MIN	0.05	-0.18	1.00	0.06	-0.04	-0.21	-0.08	-0.16	0.06	-0.02	-0.07	0.02	-0.02	-0.03	-0.07	0.01	-0.34	-0.32	-0.06	0.06
PTS	0.20	-0.02	0.06	1.00	-0.96	-0.19	-0.17	-0.53	-0.12	-0.05	-0.79	-0.10	-0.09	0.03	0.04	-0.04	0.01	0.01	-0.03	-0.01
FGM	0.00	0.02	-0.04	-0.96	1.00	-0.09	-0.07	0.46	0.19	0.03	0.76	0.07	0.07	-0.04	-0.05	0.05	-0.02	0.00	0.03	0.02
FGA	-0.65	0.03	-0.21	-0.19	-0.09	1.00	0.76	0.31	-0.29	0.06	0.10	0.12	0.07	0.05	0.05	-0.04	0.11	-0.01	0.02	-0.12
FG.	-0.82	-0.05	-0.08	-0.17	-0.07	0.76	1.00	0.09	-0.02	0.15	0.12	0.06	0.06	0.04	0.04	-0.04	0.05	-0.03	-0.02	0.02
X3P.Made	-0.09	0.00	-0.16	-0.53	0.46	0.31	0.09	1.00	-0.75	-0.11	0.40	0.08	0.07	-0.01	-0.01	0.02	0.07	0.10	0.01	-0.03
X3PA	-0.01	0.05	0.06	-0.12	0.19	-0.29	-0.02	-0.75	1.00	0.06	0.11	-0.01	-0.01	0.00	-0.02	0.01	-0.06	-0.10	0.02	0.06
X3P.	-0.13	-0.02	-0.02	-0.05	0.03	0.06	0.15	-0.11	0.06	1.00	0.06	-0.02	-0.09	-0.04	-0.04	0.05	0.01	-0.06	0.05	0.03
FTM	0.10	0.06	-0.07	-0.79	0.76	0.10	0.12	0.40	0.11	0.06	1.00	-0.51	-0.36	-0.04	-0.05	0.05	-0.03	0.00	0.03	0.03
FTA	-0.42	-0.07	0.02	-0.10	0.07	0.12	0.06	0.08	-0.01	-0.02	-0.51	1.00	0.68	0.03	0.05	-0.05	0.09	-0.03	-0.02	-0.14
FT.	-0.58	-0.13	-0.02	-0.09	0.07	0.07	0.06	0.07	-0.01	-0.09	-0.36	0.68	1.00	0.00	-0.01	0.01	0.00	0.03	0.07	0.03
OREB	-0.03	0.01	-0.03	0.03	-0.04	0.05	0.04	-0.01	0.00	-0.04	-0.04	0.03	0.00	1.00	0.96	-0.98	0.07	0.00	0.00	-0.06
DREB	-0.02	0.00	-0.07	0.04	-0.05	0.05	0.04	-0.01	-0.02	-0.04	-0.05	0.05	-0.01	0.96	1.00	-0.99	0.03	0.05	-0.04	-0.09
REB	0.02	0.00	0.01	-0.04	0.05	-0.04	-0.04	0.02	0.01	0.05	0.05	-0.05	0.01	-0.98	-0.99	1.00	-0.02	-0.02	0.00	0.07
AST	-0.02	0.01	-0.34	0.01	-0.02	0.11	0.05	0.07	-0.06	0.01	-0.03	0.09	0.00	0.07	0.03	-0.02	1.00	-0.30	0.17	-0.55
STL	0.02	-0.02	-0.32	0.01	0.00	-0.01	-0.03	0.10	-0.10	-0.06	0.00	-0.03	0.03	0.00	0.05	-0.02	-0.30	1.00	0.00	-0.03
BLK	-0.03	-0.01	-0.06	-0.03	0.03	0.02	-0.02	0.01	0.02	0.05	0.03	-0.02	0.07	0.00	-0.04	0.00	0.17	0.00	1.00	-0.06
TOV	-0.06	-0.06	0.06	-0.01	0.02	-0.12	0.02	-0.03	0.06	0.03	0.03	-0.14	0.03	-0.06	-0.09	0.07	-0.55	-0.03	-0.06	1.00

Table 2 - Logistic Regression on 7 variables

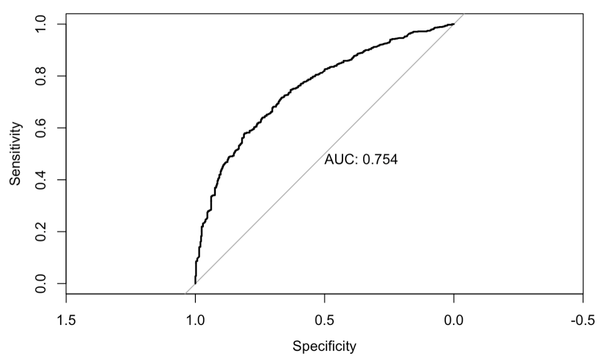
Regression Summary

```
Call:
glm(formula = target_5yrs ~ GP + MIN + FG. + X3P. + AST + BLK +
    TOV, family = "binomial", data = df_combined)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4530  -1.0184   0.5323   0.8353   1.8225

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.566381   0.571262  -6.243 4.29e-10 ***
GP           0.027139   0.005013   5.414 6.17e-08 ***
MIN          0.051895   0.018535   2.800 0.005112 **
FG.          0.033814   0.012812   2.639 0.008308 **
X3P.         0.009428   0.004699   2.007 0.044802 *
AST          0.137439   0.095443   1.440 0.149867
BLK          0.948079   0.287717   3.295 0.000984 ***
TOV         -0.297971   0.233267  -1.277 0.201467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1565.8  on 1229  degrees of freedom
Residual deviance: 1338.6  on 1222  degrees of freedom
AIC: 1354.6
```

Confusion Matrix

	Reference	
Prediction	0	1
0	175	116
1	235	704

ROC Curve



Correlations between features

	(Intercept)	GP	MIN	FG.	X3P.	AST	BLK	TOV
(Intercept)	1.00	-0.18	0.12	-0.89	-0.38	-0.09	0.09	-0.03
GP	-0.18	1.00	-0.24	-0.17	-0.01	-0.05	-0.01	-0.06
MIN	0.12	-0.24	1.00	-0.14	-0.22	-0.17	-0.35	-0.54
FG.	-0.89	-0.17	-0.14	1.00	0.29	0.16	-0.13	0.02
X3P.	-0.38	-0.01	-0.22	0.29	1.00	-0.07	0.20	0.11
AST	-0.09	-0.05	-0.17	0.16	-0.07	1.00	0.39	-0.53
BLK	0.09	-0.01	-0.35	-0.13	0.20	0.39	1.00	-0.10
TOV	-0.03	-0.06	-0.54	0.02	0.11	-0.53	-0.10	1.00

Table 3 - Residual Plots of Seven Variables

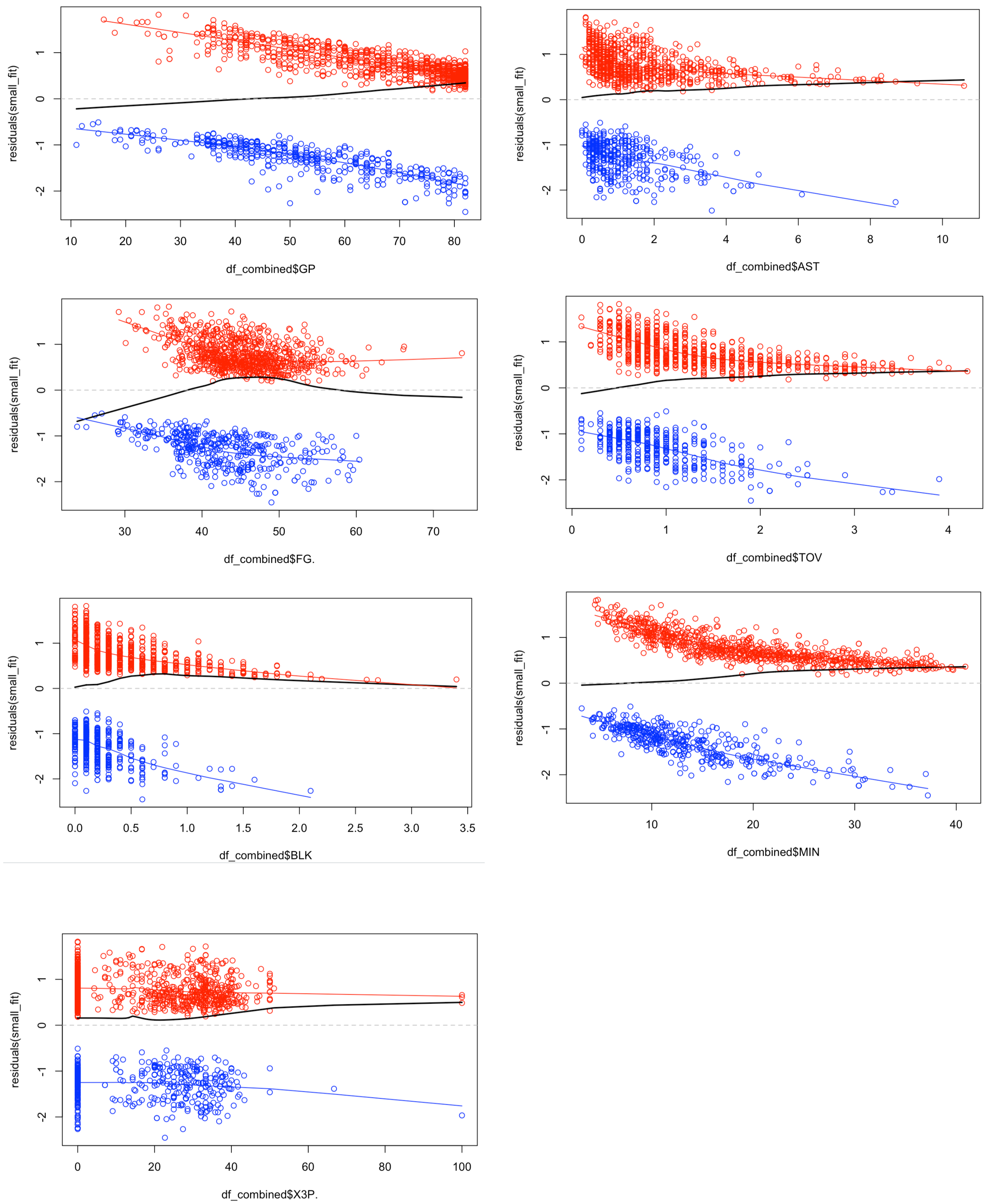


Table 4 - “Best” Regression Model Diagnostics

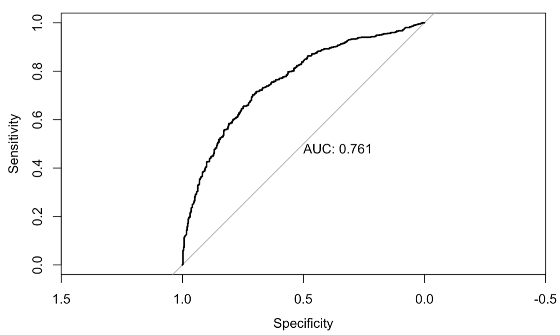
Regression Summary

```
Call:
glm(formula = target_5yrs ~ GP + MIN + I(FG. * FG.) + log(X3P. +
    0.001) + AST + log(1 + BLK) + REB + FT., family = "binomial",
    data = df_combined)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9842  -0.9998   0.5259   0.8393   1.8163

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.5679086  0.5537543  -6.443 1.17e-10 ***
GP                 0.0250978  0.0050772   4.943 7.68e-07 ***
MIN               -0.0166726  0.0227478  -0.733   0.4636
I(FG. * FG.)       0.0002792  0.0001483   1.883   0.0597 .
log(X3P. + 0.001)  0.0359691  0.0174588   2.060   0.0394 *
AST                0.1887459  0.0921666   2.048   0.0406 *
log(1 + BLK)       1.0904710  0.4816299   2.264   0.0236 *
REB                0.2863702  0.0935179   3.062   0.0022 **
FT.                0.0174476  0.0069861   2.497   0.0125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1565.8  on 1229  degrees of freedom
Residual deviance: 1327.8  on 1221  degrees of freedom
AIC: 1345.8
```

Confusion Matrix

	Reference	
Prediction	0	1
0	190	106
1	220	714

ROC Curve



Correlations

	(Intercept)	GP	MIN	I(FG. * FG.)	log(X3P. + 0.001)	AST	log(1 + BLK)	REB	FT.
(Intercept)	1.00	-0.18	0.09	-0.41	0.03	-0.02	-0.12	-0.01	-0.80
GP	-0.18	1.00	-0.20	-0.14	0.00	-0.09	-0.04	-0.04	-0.12
MIN	0.09	-0.20	1.00	0.01	-0.29	-0.69	-0.07	-0.71	-0.19
I(FG. * FG.)	-0.41	-0.14	0.01	1.00	0.30	0.07	-0.04	-0.16	0.01
log(X3P. + 0.001)	0.03	0.00	-0.29	0.30	1.00	0.05	0.09	0.26	-0.14
AST	-0.02	-0.09	-0.69	0.07	0.05	1.00	0.19	0.41	0.04
log(1 + BLK)	-0.12	-0.04	-0.07	-0.04	0.09	0.19	1.00	-0.36	0.13
REB	-0.01	-0.04	-0.71	-0.16	0.26	0.41	-0.36	1.00	0.14
FT.	-0.80	-0.12	-0.19	0.01	-0.14	0.04	0.13	0.14	1.00

Table 5 - Naive Per Minute Regression Model Diagnostics

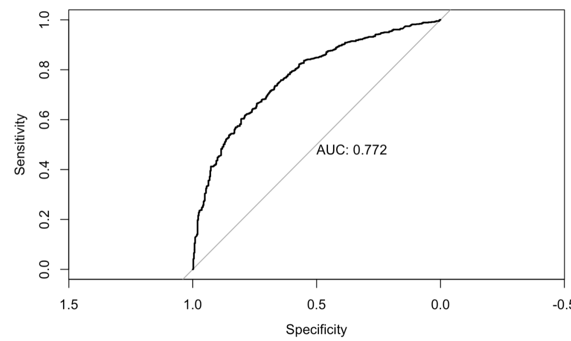
Regression Summary

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.610988   2.140260  -1.687   0.0916 .
GP            0.025830   0.005244   4.925 8.42e-07 ***
MIN           0.051963   0.012316   4.219 2.45e-05 ***
PTS           9.561895  11.601529   0.824   0.4098
FGM          -3.538696  21.748090  -0.163   0.8707
FGA          -6.316685   5.954235  -1.061   0.2887
FG.          -0.016708   0.040777  -0.410   0.6820
X3P.Made     -8.774942  18.667506  -0.470   0.6383
X3PA          3.483083   5.964048   0.584   0.5592
X3P.          0.010879   0.005999   1.813   0.0698 .
FTM           9.331809  14.688605   0.635   0.5252
FTA         -11.303722   8.009840  -1.411   0.1582
FT.           0.003868   0.012343   0.313   0.7540
OREB         -2.806312  15.065526  -0.186   0.8522
DREB         -3.500385  14.901199  -0.235   0.8143
REB          10.756606  14.780931   0.728   0.4668
AST           5.056879   2.030780   2.490   0.0128 *
STL           2.522936   5.436997   0.464   0.6426
BLK          10.826538   4.547507   2.381   0.0173 *
TOV          -3.996861   3.972368  -1.006   0.3143
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1565.8  on 1229  degrees of freedom
Residual deviance: 1308.3  on 1210  degrees of freedom
AIC: 1348.3
```

Confusion Matrix

	Reference	
Prediction	0	1
0	189	106
1	221	714

ROC Curve



Correlations

	(Intercept)	GP	MIN	PTS	FGM	FGA	FG.	X3P.Made	X3PA	X3P.	FTM	FTA	FT.	OREB	DREB	REB	AST	STL	BLK	TOV
(Intercept)	1.00	-0.08	0.01	0.42	0.08	-0.85	-0.88	-0.18	-0.07	-0.07	0.00	-0.41	-0.49	-0.06	-0.06	0.05	-0.11	-0.06	-0.01	-0.08
GP	-0.08	1.00	-0.48	-0.05	0.01	0.07	0.02	0.00	0.08	-0.05	0.06	-0.03	-0.09	0.04	0.01	-0.02	-0.05	-0.03	-0.02	-0.01
MIN	0.01	-0.48	1.00	-0.05	0.03	0.01	0.00	-0.07	0.09	-0.08	-0.01	0.05	0.02	-0.03	-0.07	0.04	-0.12	-0.03	0.01	0.19
PTS	0.42	-0.05	-0.05	1.00	-0.82	-0.40	-0.39	-0.50	-0.12	-0.04	-0.63	-0.18	-0.17	0.03	0.04	-0.03	0.03	0.01	0.00	-0.08
FGM	0.08	0.01	0.03	-0.82	1.00	-0.18	-0.17	0.36	0.18	0.04	0.56	0.10	0.10	-0.09	-0.09	0.09	-0.06	0.00	0.05	0.08
FGA	-0.85	0.07	0.01	-0.40	-0.18	1.00	0.92	0.31	-0.14	-0.01	0.19	0.16	0.13	0.09	0.10	-0.09	0.08	-0.01	-0.06	-0.05
FG.	-0.88	0.02	0.00	-0.39	-0.17	0.92	1.00	0.21	0.01	0.04	0.20	0.14	0.12	0.09	0.09	-0.10	0.07	-0.03	-0.08	0.02
X3P.Made	-0.18	0.00	-0.07	-0.50	0.36	0.31	0.21	1.00	-0.75	-0.20	0.37	0.03	-0.01	-0.02	-0.04	0.03	0.07	0.07	0.00	-0.03
X3PA	-0.07	0.08	0.09	-0.12	0.18	-0.14	0.01	-0.75	1.00	0.09	0.03	0.09	0.11	0.02	0.01	-0.01	-0.07	-0.08	0.02	0.10
X3P.	-0.07	-0.05	-0.08	-0.04	0.04	-0.01	0.04	-0.20	0.09	1.00	0.03	0.00	-0.01	-0.06	-0.06	0.08	0.00	-0.04	0.06	0.05
FTM	0.00	0.06	-0.01	-0.63	0.56	0.19	0.20	0.37	0.03	0.03	1.00	-0.62	-0.50	0.00	-0.02	0.02	0.00	0.00	-0.03	0.03
FTA	-0.41	-0.03	0.05	-0.18	0.10	0.16	0.14	0.03	0.09	0.00	-0.62	1.00	0.77	-0.01	0.01	-0.02	0.01	-0.05	0.04	-0.05
FT.	-0.49	-0.09	0.02	-0.17	0.10	0.13	0.12	-0.01	0.11	-0.01	-0.50	0.77	1.00	-0.03	-0.04	0.04	-0.05	0.03	0.12	0.11
OREB	-0.06	0.04	-0.03	0.03	-0.09	0.09	0.09	-0.02	0.02	-0.06	0.00	-0.01	-0.03	1.00	0.93	-0.97	0.12	0.00	-0.03	-0.15
DREB	-0.06	0.01	-0.07	0.04	-0.09	0.10	0.09	-0.04	0.01	-0.06	-0.02	0.01	-0.04	0.93	1.00	-0.98	0.06	0.06	-0.07	-0.16
REB	0.05	-0.02	0.04	-0.03	0.09	-0.09	-0.10	0.03	-0.01	0.08	0.02	-0.02	0.04	-0.97	-0.98	1.00	-0.04	-0.02	0.03	0.15
AST	-0.11	-0.05	-0.12	0.03	-0.06	0.08	0.07	0.07	-0.07	0.00	0.00	0.01	-0.05	0.12	0.06	-0.04	1.00	-0.32	0.15	-0.50
STL	-0.06	-0.03	-0.03	0.01	0.00	-0.01	-0.03	0.07	-0.08	-0.04	0.00	-0.05	0.03	0.00	0.06	-0.02	-0.32	1.00	0.00	0.01
BLK	-0.01	-0.02	0.01	0.00	0.05	-0.06	-0.08	0.00	0.02	0.06	-0.03	0.04	0.12	-0.03	-0.07	0.03	0.15	0.00	1.00	-0.02
TOV	-0.08	-0.01	0.19	-0.08	0.08	-0.05	0.02	-0.03	0.10	0.05	0.03	-0.05	0.11	-0.15	-0.16	0.15	-0.50	0.01	-0.02	1.00

Table 6 - Best Overall - Per Minute Regression Model Diagnostics

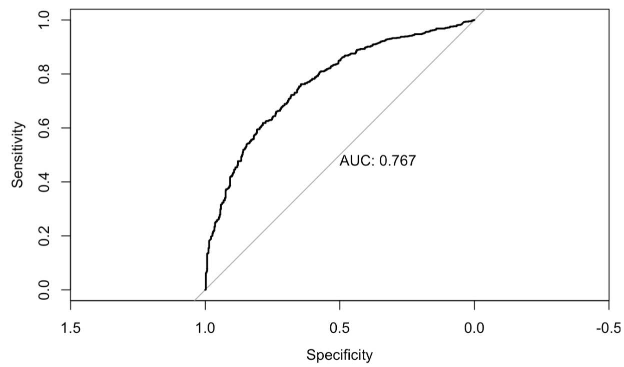
Regression Summary

```
Call:
glm(formula = target_5yrs ~ GP + MIN + I(FG. * FG.) + log(X3P. +
    0.001) + AST + log(1 + BLK) + REB + I(FT. * FT.), family = "binomial",
    data = df_min)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8179  -0.9865   0.5249   0.8226   2.0307

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.649e+00  5.216e-01  -8.912  < 2e-16 ***
GP                2.515e-02  5.097e-03   4.935 8.01e-07 ***
MIN               5.799e-02  1.171e-02   4.950 7.42e-07 ***
I(FG. * FG.)      2.389e-04  1.504e-04   1.589  0.11209
log(X3P. + 0.001)  4.782e-02  1.812e-02   2.639  0.00832 **
AST               3.552e+00  1.556e+00   2.283  0.02241 *
log(1 + BLK)      1.073e+01  4.660e+00   2.302  0.02135 *
REB               6.719e+00  1.579e+00   4.255 2.09e-05 ***
I(FT. * FT.)      1.682e-04  5.313e-05   3.165  0.00155 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1565.8  on 1229  degrees of freedom
Residual deviance: 1318.1  on 1221  degrees of freedom
AIC: 1336.1
```

Confusion Matrix

	Reference	
Prediction	0	1
0	194	108
1	216	712

ROC Curve



Correlations

	(Intercept)	GP	MIN	I(FG. * FG.)	log(X3P. + 0.001)	AST	log(1 + BLK)	REB	I(FT. * FT.)
(Intercept)	1.00	-0.21	0.09	-0.35	-0.22	-0.51	-0.12	-0.53	-0.57
GP	-0.21	1.00	-0.51	-0.15	0.01	-0.12	-0.02	-0.03	-0.09
MIN	0.09	-0.51	1.00	-0.13	-0.17	-0.02	0.00	-0.04	-0.09
I(FG. * FG.)	-0.35	-0.15	-0.13	1.00	0.26	0.10	-0.02	-0.18	-0.01
log(X3P. + 0.001)	-0.22	0.01	-0.17	0.26	1.00	0.05	0.09	0.34	-0.10
AST	-0.51	-0.12	-0.02	0.10	0.05	1.00	0.16	0.44	0.06
log(1 + BLK)	-0.12	-0.02	0.00	-0.02	0.09	0.16	1.00	-0.27	0.14
REB	-0.53	-0.03	-0.04	-0.18	0.34	0.44	-0.27	1.00	0.22
I(FT. * FT.)	-0.57	-0.09	-0.09	-0.01	-0.10	0.06	0.14	0.22	1.00

Table 7 (linear regression results)

```
Call:
lm(formula = Yrs ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.9789  -3.2503  -0.4489   3.2100  15.5158 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.86050    3.26268  -1.796   0.0730 .  
GP           0.07375    0.01424   5.179  3.1e-07 ***
MIN         -0.18440    0.08572  -2.151   0.0319 *  
PTS         -0.32422    3.63341  -0.089   0.9289    
FGM          1.58571    4.88260   0.325   0.7455    
FGA         -0.26426    2.08852  -0.127   0.8994    
FG.          0.03760    0.05811   0.647   0.5179    
X3P.Made    -6.84130    4.11628  -1.662   0.0971 .  
X3PA         3.74156    1.35925   2.753   0.0061 ** 
X3P.         0.02623    0.01302   2.015   0.0443 *  
FTM         -0.55913    2.96788  -0.188   0.8506    
FTA          1.04829    2.17761   0.481   0.6304    
FT.          0.07224    0.02930   2.465   0.0140 *  
OREB         4.43881    3.61308   1.229   0.2198    
DREB         3.70767    3.64539   1.017   0.3095    
REB         -2.95057    3.77371  -0.782   0.4346    
AST          0.84210    2.02306   0.416   0.6774    
STL          1.86515    2.19530   0.850   0.3959    
BLK          2.45326    2.09989   1.168   0.2432    
TOV         -0.81223    2.16030  -0.376   0.7071    
EFF         -0.38141    2.01876  -0.189   0.8502    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.269 on 571 degrees of freedom
Multiple R-squared:  0.2919,    Adjusted R-squared:  0.2671 
F-statistic: 11.77 on 20 and 571 DF,  p-value: < 2.2e-16
```