Term Project

Yue Wang May.12

1. Data Description

EPL.csv:
200 data from Season 2006/2007 to 2015/2016 were collected. Squad size, average age, number of foreign players, total market value of the team, goals for, goals against, total points, final rank, whether it is championship were all included. For some data, the team's last year's rank as well as its most important addition/depart were also listed.

La Liga.csv:
To make comparison between these two leagues, 100 data from Season 2008/2009 to 2012/2013 were collected. Similarly, squad size, average age, number of foreign players, total market value of the team, goals for, goals against, total points, final rank, whether it is championship were all included.

2. Results and Discussion

2.1 Linear Regression for Pts and Rank

In this part, we simply set final points and team rank as responses to investigate the effect of other variables. Here's the results:

```
Call:
lm(formula = Pts ~ . - Rank - Championship, data = EPL)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4440 -3.0124 -0.2673  2.6822 13.2331

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        59.750294   7.609358   7.852 2.76e-13 ***
Squad               0.031982   0.065893   0.485    0.628
Age                -0.351568   0.254397  -1.382    0.169
Foreign.Players    -0.020166   0.020890  -0.965    0.336
Total.Market.Value  0.006225   0.004044   1.539    0.125
Goals.For           0.595867   0.032562  18.300  < 2e-16 ***
Goals.Against      -0.609798   0.032997 -18.481  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.363 on 193 degrees of freedom
Multiple R-squared:  0.9348,    Adjusted R-squared:  0.9328
F-statistic: 461.3 on 6 and 193 DF,  p-value: < 2.2e-16
```

From this, we can see that final points are mainly determined by goals for and goals against of the team. More goals for and less goals against may lead to final higher points, which makes sense.

```
Call:
lm(formula = Rank ~ . - Championship, data = EPL)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8146 -1.2223 -0.0759  1.2234  3.3337 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        19.195616   3.363395   5.707  4.3e-08 ***
Squad               0.003619   0.025371   0.143  0.88673    
Age                 0.124887   0.098374   1.270  0.20579    
Foreign.Players     0.002675   0.008058   0.332  0.74024    
Total.Market.Value  0.001331   0.001566   0.850  0.39635    
Goals.For           0.006220   0.020722   0.300  0.76438    
Goals.Against       0.059296   0.021130   2.806  0.00553 ** 
Pts                -0.298366   0.027698 -10.772  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.679 on 192 degrees of freedom
Multiple R-squared:  0.9186,	Adjusted R-squared:  0.9156 
F-statistic: 309.6 on 7 and 192 DF,  p-value: < 2.2e-16
```

Team rank is dependent on both team points and goals against. It's obvious that higher team points can result in a better rank. And the result shows goals for is insignificant but goals against is significant. This fact proves that defense, rather than offense, wins championships.

```
Call:
lm(formula = Rank ~ . - Championship - Pts, data = EPL)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2660 -1.4110  0.0952  1.5053  5.0781 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.3681724  3.6991412   0.370   0.7119    
Squad              -0.0059236  0.0320324  -0.185   0.8535    
Age                 0.2297834  0.1236699   1.858   0.0647 .  
Foreign.Players     0.0086922  0.0101555   0.856   0.3931    
Total.Market.Value -0.0005263  0.0019659  -0.268   0.7892    
Goals.For          -0.1715666  0.0158293 -10.839   <2e-16 ***
Goals.Against       0.2412391  0.0160407  15.039   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.121 on 193 degrees of freedom
Multiple R-squared:  0.8694,	Adjusted R-squared:  0.8654 
F-statistic: 214.2 on 6 and 193 DF,  p-value: < 2.2e-16
```

However, we know team rank is correlated with final points. To be careful, we excluded points and found both goals for and goals against are significant. The coefficient of goals against and goals for are 0.24 and -0.17. As the absolute value of 0.24 is greater than that of -0.17, we can still conclude that defense is more important.

2.2 Logistic Regression for Championship

```
                     Estimate   Std. Error    z value   Pr(>|z|)
(Intercept)        -1.142170e+02 54.951417553 -2.0785087 0.03766253
Squad               1.972804e-01  0.174279957  1.1319740 0.25764538
Age                 2.128459e+00  1.203224872  1.7689616 0.07690028
Foreign.Players     3.674851e-02  0.157842970  0.2328169 0.81590362
Total.Market.Value -5.684102e-03  0.007884575 -0.7209142 0.47096232
Goals.For          -2.448538e-02  0.077842605 -0.3145498 0.75310348
Goals.Against       9.773772e-02  0.126747588  0.7711210 0.44063523
Pts                 6.503647e-01  0.306890863  2.1192052 0.03407313

champion.pred   0   1
            0 187   2
            1   3   8
```

We can see that championship is actually determined by points, which fits the reality. And this model is capable for accurate prediction, the accuracy of which is 195/200=97.5%

```
                     Estimate   Std. Error     z value    Pr(>|z|)
(Intercept)        -2.719904e+01 13.458524110 -2.02095265 0.043284667
Squad               1.677019e-01  0.126755968  1.32302979 0.185825471
Age                 7.790674e-01  0.449465795  1.73331847 0.083039079
Foreign.Players    -1.843816e-01  0.163885613 -1.12506303 0.260562325
Total.Market.Value  2.449296e-04  0.005567566  0.04399223 0.964910599
Goals.For           1.408175e-01  0.046242573  3.04519147 0.002325322
Goals.Against      -1.608944e-01  0.068442841 -2.35078464 0.018733873

champion.pred2   0   1
             0 188   5
             1   2   5
```

Similarly, we excluded points and modeled it again. Results show that both goals for and goals against have significant effect on championship. But goals for seems to be more significant, which opposite to our previous conclusion. This model has a lower accuracy of 193/200=96.5% than above one.

Therefore, if we knew the points of the team (usually we do know), we could apply the first model to predict whether it would win a champion. If not, the second one is our only choice. From my point of view, the first model is preferable due to its high accuracy.

2.3 Variable Subsetting

```
Subset selection object
Call: regsubsets.formula(Pts ~ . - Rank - Championship, data = EPL,
    nvmax = 6)
6 Variables  (and intercept)
                   Forced in Forced out
Squad                  FALSE      FALSE
Age                    FALSE      FALSE
Foreign.Players        FALSE      FALSE
Total.Market.Value     FALSE      FALSE
Goals.For              FALSE      FALSE
Goals.Against          FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         Squad Age Foreign.Players Total.Market.Value Goals.For Goals.Against
1  ( 1 ) " "   " " " "             " "                "*"       " "
2  ( 1 ) " "   " " " "             " "                "*"       "*"
3  ( 1 ) " "   " " " "             "*"                "*"       "*"
4  ( 1 ) " "   "*" " "             "*"                "*"       "*"
5  ( 1 ) " "   "*" "*"             "*"                "*"       "*"
6  ( 1 ) "*"   "*" "*"             "*"                "*"       "*"
```

All 6 variables appear to be irreplaceable. And we can find that goals for is most important, goals against is second important. Team value and average age also matter a lot but the number of foreign player and squad size seem to be unimportant. Besides, forward and backward search gave us the same result.

```
Step:  AIC=593.3
Pts ~ Goals.For + Goals.Against + Total.Market.Value + Age

                   Df Sum of Sq    RSS    AIC
<none>                          3695.3 593.30
+ Foreign.Players  1   16.4595 3678.8 594.41
+ Squad            1    3.2034 3692.1 595.13
```

```
Step:  AIC=593.3
Pts ~ Age + Total.Market.Value + Goals.For + Goals.Against

                     Df Sum of Sq     RSS    AIC
<none>                             3695.3 593.30
- Age                 1      46.1  3741.4 593.78
- Total.Market.Value  1      53.1  3748.4 594.15
- Goals.For           1    6439.8 10135.1 793.09
- Goals.Against       1    7260.0 10955.3 808.65
```

```
Step:  AIC=593.3
Pts ~ Goals.For + Goals.Against + Total.Market.Value + Age

                     Df Sum of Sq     RSS    AIC
<none>                             3695.3 593.30
- Age                 1      46.1  3741.4 593.78
- Total.Market.Value  1      53.1  3748.4 594.15
+ Foreign.Players     1      16.5  3678.8 594.41
+ Squad               1       3.2  3692.1 595.13
- Goals.For           1    6439.8 10135.1 793.09
- Goals.Against       1    7260.0 10955.3 808.65
```

Step AIC was also applied. Forward-, backward-, and two-way methods all proved that number of foreign player and squad size have little contribution to the team points. Therefore, we can simplify our model by omitting these two variables.

2.4 Clustering

```
> km.out$cluster
  [1] 1 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 3 3
 [46] 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 3 3 3 3 2 2
 [91] 2 2 2 2 2 2 2 2 2 2 1 1 1 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 3 3 3 2 2 2 3 2 2 2 2 2
[136] 2 2 2 2 2 1 1 1 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 1 1 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2
[181] 1 1 1 1 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2
```

```
> cutree(hc.complete, 3)
  [1] 1 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 2 2 2
 [46] 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 2 2 2 2 3 3
 [91] 3 3 3 3 3 3 3 3 3 3 1 1 1 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 2 2 2 3 3 3 2 3 3 3 3 3
[136] 3 3 3 3 3 1 1 1 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 1 1 2 1 1 2 2 3 3 3 3 3 3 3 3 3 3 3 2 3
[181] 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
```

```
> cutree(hc.average, 3)
  [1] 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2
 [46] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2 2
 [91] 2 2 2 2 2 2 2 2 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
[136] 2 2 2 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[181] 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

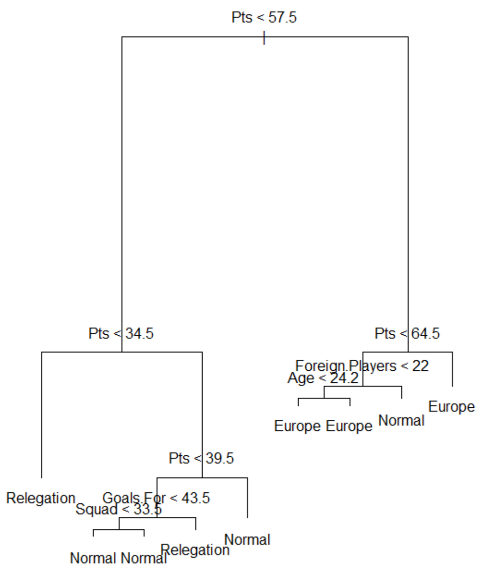
Our data are supposed to be classified into three groups: teams that are qualified to play European Champion League and Europa League, teams that are relegated, and other teams.

K-means method gave us poor classification result. This model seems to classify all data into three groups that are championship, relegated teams, and others as most teams were classified as the third kind while only few of them belong to the first kind.

Complete hierarchical clustering lead to a different classification. Championship, other teams that are qualified to play European Champion League and Europa League, and remaining teams are three groups. This is because that complete linkage does classification by evaluating maximal intercluster dissimilarity. Championships are very unique as they have the Championship variable value equal to 1 while others are all 0. Other teams that are qualified to play European Champion League and Europa League usually have similar behavior as championships except their Championship variable values are 0, which makes them differ from remaining teams.

Moreover, average hierarchical clustering can only distinguish championships from data due to its classification property to classify data based on their mean intercluster dissimilarity.

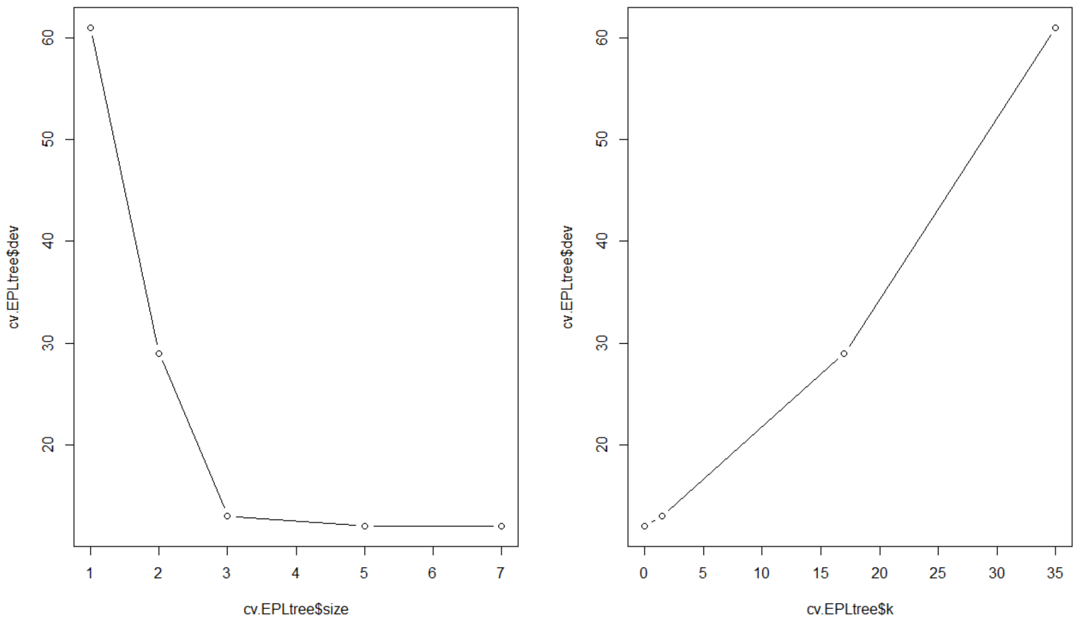
2.5 Decision Tree

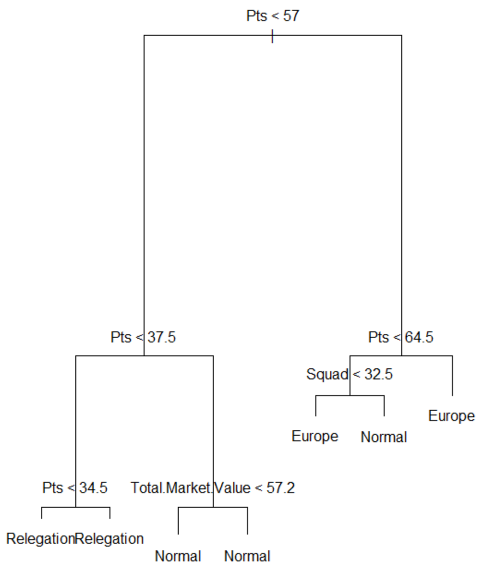


tree.pred	Normal	Europe	Relegation
Normal	42	4	1
Europe	2	16	0
Relegation	7	0	8

Decision tree without cross-validation gained (42+16+8)/80=82.5% accuracy. Teams have points higher than 64.5 can have a guaranteed qualification for European Champion League and Europa League and teams have less than 34.5 points are very likely to be relegated. For teams have points between 34.5 and 64.5, introducing less foreign players may have a positive effect on their performance. That is, the quality, instead of quantity, of foreign players matters.

```
$size
[1] 7 5 3 2 1

$dev
[1] 12 12 13 29 61

$k
[1] -Inf  0.0  1.5 17.0 35.0

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```





```
tree.pred2   Normal Europe Relegation
  Normal         42      4          1
  Europe          2     16          0
  Relegation      7      0          8
```

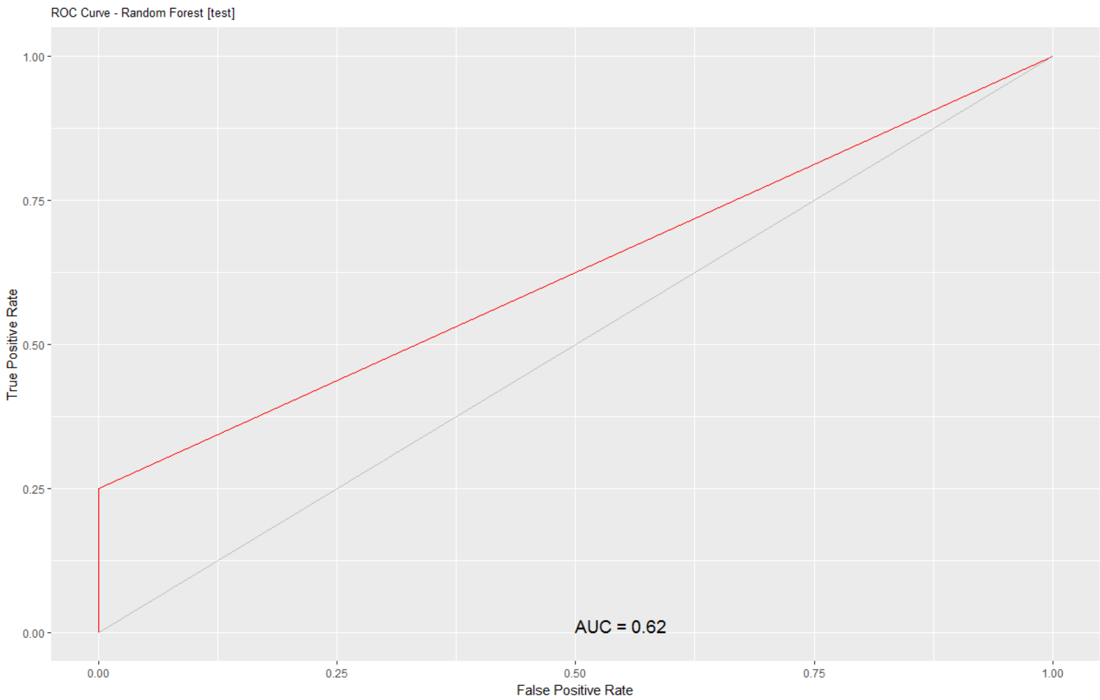
Trees with 5,6, and 7 terminal nodes resulted in the lowest cross-validation error rate, with 12 cross-validation errors. Here we set terminal nodes as 7. The case for teams have points higher than 64.5 or less than 34.5 is the same as before. However, this model indicates that foreign players contribute little to team performance. Instead, squad size matters a lot now. A smaller size could lead to a better performance, which again proves that quality is more important. Interestingly, the accuracy of this model is also 82.5%.

2.6 Random Forest

```
Random Forest

200 samples
  9 predictor
  3 classes: 'Normal', 'Europe', 'Relegation'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 180, 180, 180, 180, 180, 180, ...
Resampling results across tuning parameters:

  mtry  Accuracy  Kappa
  2     0.895     0.8170359
  4     0.900     0.8264906
  7     0.900     0.8282353

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 4.
    obs
pre  1  2  3
  1 42  1  1
  2  3 19  0
  3  6  0  8
```

First, we assumed there are three groups named “normal”, “Europe”, “Relegation”, which was in fact the same as we did in previous parts. Cross-validation told us that mtry shoule be 4. The accuracy of this model is (42+19+8)/80=86.25%, which is slightly higher than our decision tree model.

```
Random Forest

200 samples
  8 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 180, 180, 180, 180, 180, 180, ...
Resampling results across tuning parameters:

  mtry  Accuracy  Kappa
  2     0.975     0.5642857
  4     0.975     0.6590226
  7     0.980     0.7285714

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 7.

           observation2
prediction2  1  2
          1 76  3
          2  0  1
```



Then we tried to divide all data into "Championship" and "Not Championship". Now the best mtry was 7 and the accuracy was greatly enhanced. A (76+1)/80=96.25% accuracy value and 0.62 AUC suggest that this new model is especially good at predicting championship.

2.7 EPL vs. La Liga

The reason we choose La Liga to do comparison is that it also has 20 teams. Moreover, 6 of them can play European Champion League and Europa League and 3 of them will face relegation every season, which is the same as EPL does.

2.7.1 linear regression

```
Call:
lm(formula = Pts ~ . - Rank - Championship, data = LaLiga)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.1036  -4.6166  -0.0835   3.4813  15.5838 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        24.126302  22.199766   1.087   0.2799    
Squad              -0.006209   0.193791  -0.032   0.9745    
Age                -0.341747   0.748312  -0.457   0.6490    
Foreign.Players     0.028289   0.148229   0.191   0.8491    
Total.Market.Value  0.013664   0.012204   1.120   0.2658    
Goals.For           0.680898   0.081675   8.337 6.68e-13 ***
Goals.Against      -0.017006   0.010101  -1.684   0.0956 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.064 on 93 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8669 
F-statistic: 108.4 on 6 and 93 DF,  p-value: < 2.2e-16
```

It can be seen that goals for is most important in La Liga, and goals against can only slightly influence the points. This is nearly the opposite to EPL, in which defense matters a lot and offense is much less important.

```
Call:
lm(formula = Rank ~ . - Championship, data = LaLiga)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1852 -1.4578  0.1134  1.7980  6.1310 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        23.229300   8.642192   2.688  0.00853 ** 
Squad              -0.069885   0.074967  -0.932  0.35367    
Age                 0.446349   0.289804   1.540  0.12695    
Foreign.Players    -0.073165   0.057353  -1.276  0.20527    
Total.Market.Value  0.022688   0.004753   4.773  6.8e-06 ***
Goals.For           0.015493   0.041765   0.371  0.71152    
Goals.Against       0.004311   0.003966   1.087  0.27990    
Pts                -0.476049   0.040114 -11.867  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.346 on 92 degrees of freedom
Multiple R-squared:  0.8489,	Adjusted R-squared:  0.8374 
F-statistic: 73.86 on 7 and 92 DF,  p-value: < 2.2e-16

Call:
lm(formula = Rank ~ . - Championship - Pts, data = LaLiga)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0255 -2.7208 -0.0373  2.7769  6.7997 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        11.743988  13.588392   0.864   0.3897    
Squad              -0.066929   0.118619  -0.564   0.5740    
Age                 0.609038   0.458039   1.330   0.1869    
Foreign.Players    -0.086632   0.090731  -0.955   0.3421    
Total.Market.Value  0.016183   0.007470   2.166   0.0328 *  
Goals.For          -0.308648   0.049993  -6.174 1.73e-08 ***
Goals.Against       0.012407   0.006183   2.007   0.0477 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.712 on 93 degrees of freedom
Multiple R-squared:  0.6177,	Adjusted R-squared:  0.593 
F-statistic: 25.04 on 6 and 93 DF,  p-value: < 2.2e-16
```

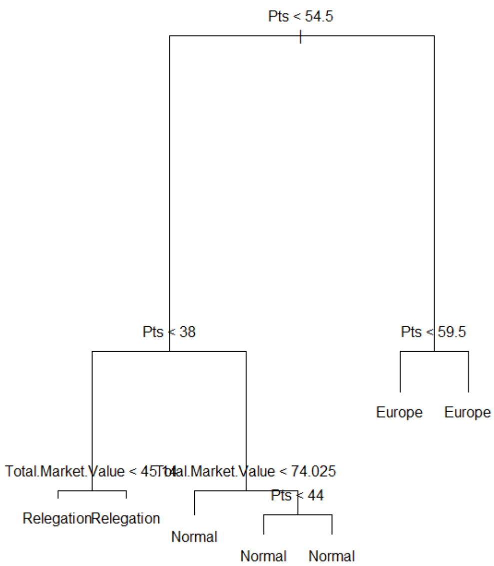
As for team rank, we modeled the data with points included first. Results show that total market value and points are most important but it is weird that both team value and goals for have negative effects on team rank. Therefore, this model appears to be inaccurate.

After points variable was excluded, goals for, and goals against became significant. Though team value still has negative effect on rank, goals for has positive impact now. This is consistent with previous “points” model, suggesting this “rank” model had been improved by excluding points as preditor.

2.7.2 Logistic regression and Decision Tree

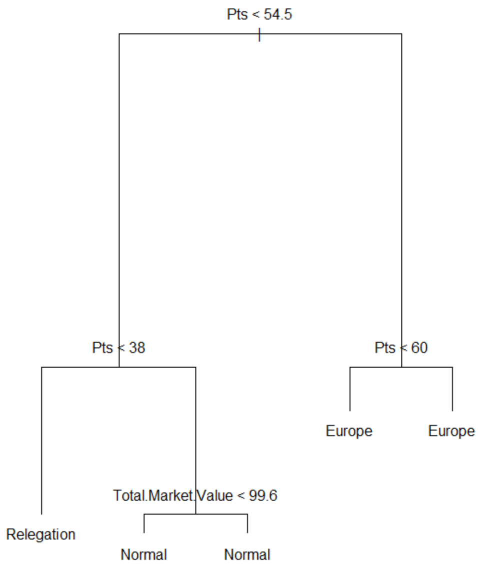
```
champion2.pred  0  1
             0 95  0
             1  0  5
```

Logistic regression model results in a 100% prediction accuracy, which means predicting La Liga championship is easier. In fact, most La Liga championships were shared by FC Barcelona and Real Madrid during these years. For EPL championship, Manutd, Mancity, Chelsea, Liverpool, Arsenal and Tottenham Hotspur are all possible competitors, which makes it hard to do right prediction.



All Europe-qualified teams are those teams who have points higher than 54.5 and all relegated teams are those teams who have points less than 38. Besides, other predictors seem to have little impact on final performance.

```
$size
[1] 5 3 2 1

$dev
[1]  8  8 15 37

$k
[1] -Inf    0    7   21

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```




```
tree.pred2   Normal Europe Relegation
  Normal         23      0          2
  Europe          2      6          0
  Relegation      1      0          6
```

Then we performed cross-validation. Trees with 3, 4, and 5 terminal nodes resulted in the lowest cross-validation error rate, with 8 cross-validation errors. However, this new tree had the same classification behavior as previous one. The accuracy of this model is (23+6+6)/40=87.5%, indicating La Liga is more predictable than EPL.

2.8 Sir Alex Ferguson vs. Wenger

Sir Alex Ferguson had managed Manchester United for 26 years and Wenger has been manager for Arsenal for more than 20 years. We can say that they are presentative figures of EPL and thus we chose them for comparison.

```
Call:
lm(formula = Pts ~ . - Rank, data = SAF)

Residuals:
      2      22      42      62      82     102     123     142     162     184 
 2.4715 -3.8201  4.3263 -2.2277  0.8597 -1.5868  2.0345 -3.2914  3.6345 -2.4007 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)  
(Intercept)        421.18701  206.78906   2.037   0.1345  
Squad               -2.17107    1.41427  -1.535   0.2223  
Age                -14.61044    8.43527  -1.732   0.1817  
Foreign.Players      1.93357    1.84365   1.049   0.3713  
Total.Market.Value   0.02107    0.03788   0.556   0.6169  
Goals.For            0.80054    0.20752   3.858   0.0308 *
Goals.Against       -0.57442    0.41217  -1.394   0.2577  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.215 on 3 degrees of freedom
Multiple R-squared:  0.9151,	Adjusted R-squared:  0.7453 
F-statistic: 5.389 on 6 and 3 DF,  p-value: 0.09754
```

```
Call:
lm(formula = Rank ~ . - Pts, data = SAF)

Residuals:
      2      22      42      62      82     102     123     142     162     184
-0.5720  0.2144 -0.7544  0.7390 -0.3929  1.2684 -1.1069  1.2376 -1.3194  0.6864

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -64.059048  65.640812  -0.976    0.401
Squad               0.412840   0.448931   0.920    0.426
Age                 2.966500   2.677597   1.108    0.349
Foreign.Players    -0.642120   0.585227  -1.097    0.353
Total.Market.Value  0.001415   0.012026   0.118    0.914
Goals.For          -0.149865   0.065874  -2.275    0.107
Goals.Against       0.144123   0.130835   1.102    0.351

Residual standard error: 1.655 on 3 degrees of freedom
Multiple R-squared:  0.797,     Adjusted R-squared:  0.3911
F-statistic: 1.963 on 6 and 3 DF,  p-value: 0.3097
```

Though Sir Alex Ferguson retired from management at the end of the 2012–13 season, we still included other three seasons' data. From these results, we can conclude that offense means a lot in his team and the team position is relatively stable during these years as no predictors can influence the team rank.

```
Call:
lm(formula = Rank ~ . - Pts, data = Wenger)

Residuals:
       4       24       45       65       85      105      125      144      165      183
 0.17623 -0.70618  0.22681  0.16113 -0.06282 -0.48767  0.64785  0.34890  0.39574 -0.69998

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        10.34863   13.57832   0.762    0.501
Squad              -0.12602    0.24467  -0.515    0.642
Age                -0.20131    0.62266  -0.323    0.768
Foreign.Players     0.04028    0.41574   0.097    0.929
Total.Market.Value -0.00121    0.01116  -0.108    0.920
Goals.For          -0.01929    0.07874  -0.245    0.822
Goals.Against       0.07859    0.07774   1.011    0.386

Residual standard error: 0.824 on 3 degrees of freedom
Multiple R-squared:  0.5371,    Adjusted R-squared:  -0.3888
F-statistic:  0.58 on 6 and 3 DF,  p-value: 0.7391
```


```
Call:
lm(formula = Rank ~ . - Pts, data = Wenger)

Residuals:
       4       24       45       65       85      105      125      144      165      183
 0.17623 -0.70618  0.22681  0.16113 -0.06282 -0.48767  0.64785  0.34890  0.39574 -0.69998

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        10.34863   13.57832   0.762    0.501
Squad              -0.12602    0.24467  -0.515    0.642
Age                -0.20131    0.62266  -0.323    0.768
Foreign.Players     0.04028    0.41574   0.097    0.929
Total.Market.Value -0.00121    0.01116  -0.108    0.920
Goals.For          -0.01929    0.07874  -0.245    0.822
Goals.Against       0.07859    0.07774   1.011    0.386

Residual standard error: 0.824 on 3 degrees of freedom
Multiple R-squared:  0.5371,    Adjusted R-squared:  -0.3888
F-statistic:  0.58 on 6 and 3 DF,  p-value: 0.7391
```


As for Wenger, both team points and rank are nearly unchangeable. Actually, Arsenal mostly ranked 3 or 4 during these ten seasons, which supports our conclusion.

2.9 Effect of Important Addition

```
Call:
lm(formula = Rankchange ~ Last.year.s.rank + Most.Important.Addition.Depart,
    data = EPL_addtion)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6753 -1.8542  0.7555  2.1263  8.3078

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -1.880095   0.706178  -2.662  0.00917 **
Last.year.s.rank               0.205600   0.080650   2.549  0.01247 *
Most.Important.Addition.Depart 0.005799   0.008754   0.662  0.50933
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.405 on 91 degrees of freedom
Multiple R-squared:  0.06809,   Adjusted R-squared:  0.04761
F-statistic: 3.324 on 2 and 91 DF,  p-value: 0.04041
```

High-priced stars are not capable to change the team rank. The changes in rank are mainly determined by last year’s rank.

```
Call:
lm(formula = Rankchange ~ Last.year.s.rank + Valuechange, data = EPL_addtion)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8264 -1.9179  0.5774  2.1339  8.2053

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.58922    0.64735  -2.455   0.0160 *
Last.year.s.rank  0.20730    0.08289   2.501   0.0142 *
Valuechange      -0.63016    1.72359  -0.366   0.7155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.411 on 91 degrees of freedom
Multiple R-squared:  0.06497,   Adjusted R-squared:  0.04442
F-statistic: 3.161 on 2 and 91 DF,  p-value: 0.04705
```

To be more accurate, we normalized the additional value by dividing the value incensement by teams’ total market value. Similarly, the results show that most important additions have litter contribution to improve team rank but last year’s rank could heavily influence the change in rank.