Analysis to the data of NFL

Jiazhang Cai

4/26/2020

Abstract

Football is one of the most popular sport in USA. This report present some analysis to the data of the attendence of each game, the evaluation of each team and the situation of each game. The content is consisted of three main parts, relating to three different datasets. First is about the study of the attendence. We aim to find the variable that effect the attendence most. Second is about the evaluation of every team. Here uses whether the team play the playoffs as the response to partially standing for the strength of a team. Finally is about the estimation to the game's result. We aim to find if there are any variable that may effect the result except the strength of the teams. These three parts are independent with each other although they used each datasets. For every topic, we assume that the data is the origin data and we won't use the model or variable created in other parts.

Basic information about the dataset

The main data is about the NFL stadium attendence from "https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-02-04". There are three tables in the dataset.

The first one is the overview of the attendence and the dictionary of this dataset is:

variable	class	description
team	character	team city
team_name	character	team name
year	integer	season year
total	double	total attendence across 17 weeks (1 week = no game)
home	double	total home attendence
away	double	total away attendence
week	character	week number (1-17)
weekly_attendence	double	weekly attendence

The second one is the information about each team and the dictionary of this dataset is:

variable	class	description
team	character	team city
team_name	character	team name
year	integer	season year
wins	double	wins (0-16)
loss	double	losses (0-16)
points_for	double	points for offensive performance
points_against	double	points for defensive performance
points_differential	double	points_for-points_against
margin_of schedule	double	(points scored-points allowed)/game played
strength_of_schedule	double	average quality of opponent as measured as measured by SRS

variable	class	description
simple rating	double	team quality relative to average as measured by SRS
offensive_ranking	double	team offense quality relative to average as measured by SRS
defensive_ranking	double	team defense quality relative to average as measured by SRS
playoffs	character	made playoffs or not
sb_winner	character	won superbowl or not

The last one is the information of every games and the dictionary of this dataset is:

variable	class	description
year	integer	season year
week	character	week number (1-17 and playoffs)
home_team	character	home team
away_team	character	away team
winner	character	winning team
tie	character	same for both team
day	character	day of week
date	character	date without year
time	character	time of game start
pts_win	double	points by winning team
pts_loss	double	points by lossing team
yds_win	double	yards by winning team
turnovers_win	double	turnovers by winning team
yds_loss	double	yards by losing team
turnovers_loss	double	turnovers by losing team
home_team_name	character	home team name
home_team_city	character	home team city
away_team_name	character	away team name
away_team_name	character	away team city

The additional data is from "https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/national/totals/", the *United States Census* website. The data is about the population and population change in every state of the United States.

Data analysis

1.Attendance

1.1 Basic idea

To study the problem thoroughly, we can start from different aspects. The datasets provide sufficient information for each game, which allows us to find as much as possible variables that may effect the attendance.

We mainly devide these potential effects into three aspects, the location, the time and the players. The location is about the information of where the games played. It's obvious that the more people in one state, the more attendance tends to be because more people means more potential audiences, more support to the

team, even more money for the improvrment to the team or the bigger stadium. According to our dataset, the population of each state is the best variable to quantify the location.

The time is about the information of when the games played. There are many dimensions to study the time effect. Different year, different week, different day, different time of a day may all influence the attendance. For example, there might be more audience in the weekend than the weekday, or there might be more audience at the end of the season than the start of the season. Each year's situation may also be different.

The team is the information of who played the games. Many audiences have their supporting teams, they would go to support their favorite team in person no matter the team played as the home team or away team. So the games played by the popular teams may tend to have a higher attendance.

1.2 Data preprocess

Based on the basic idea we have, we then clear up the data and combine different parts together as the working dataset. To build the model, what we need to do first is to quantify each variable. Our response in this part is the weekly attedance, or in other word, the attendance for each game. Here we just study the attendance for the home team, because the attendance is the information of the stadium not the team.

As mentioned before the locate can be expressed by the population of the state. We also create the variable called *attendance_rate*, which is the rate of the yearly attendance of the home team and the attendance of the away team. The difference between the attendance of home team and the away team is also a character of the stadium, however, we cannot use the absolute value because it is obviously correlated with the weekly attendance. This variable can represent the popularity of one team in its home city to some extent.

For the time information, we have *year*, *week*, *day* and *time*. They are all potential significant effects for the attendance. We can still separate these effects into two parts. One is the attention to the games will change from time to time, like in different stages of the season or different years, people may hold different interest to the games, especially during the playoffs or the Super Bowl. The other is the personal condition of every audience. For the common people, they may can't reach the stadium in person if the game is played during the work time. Therefore the attendeance of the games played during the weekend or evening may higher than the work time. Because the *day* and *time* are nominal variable, to simplify the problem, here transform them as binary variable. Set *day* equals to 1 if the game is played on weekend and equals to 0 if not. Set *time* equals to 1 if the games is played after 18:00 and equals to 0 if not.

For the team information, here mainly uses the *simple_rating*, the team quality relative to average as measured by SRS (Simple Rating System), to indicate the strength of the home team and the away team. As we know, the popular team may lead to more attendance and the strength of one team would influence its popularity a lot. Hence we also put the strength of the team into our consideration and let it be the quantified variable for the team information.

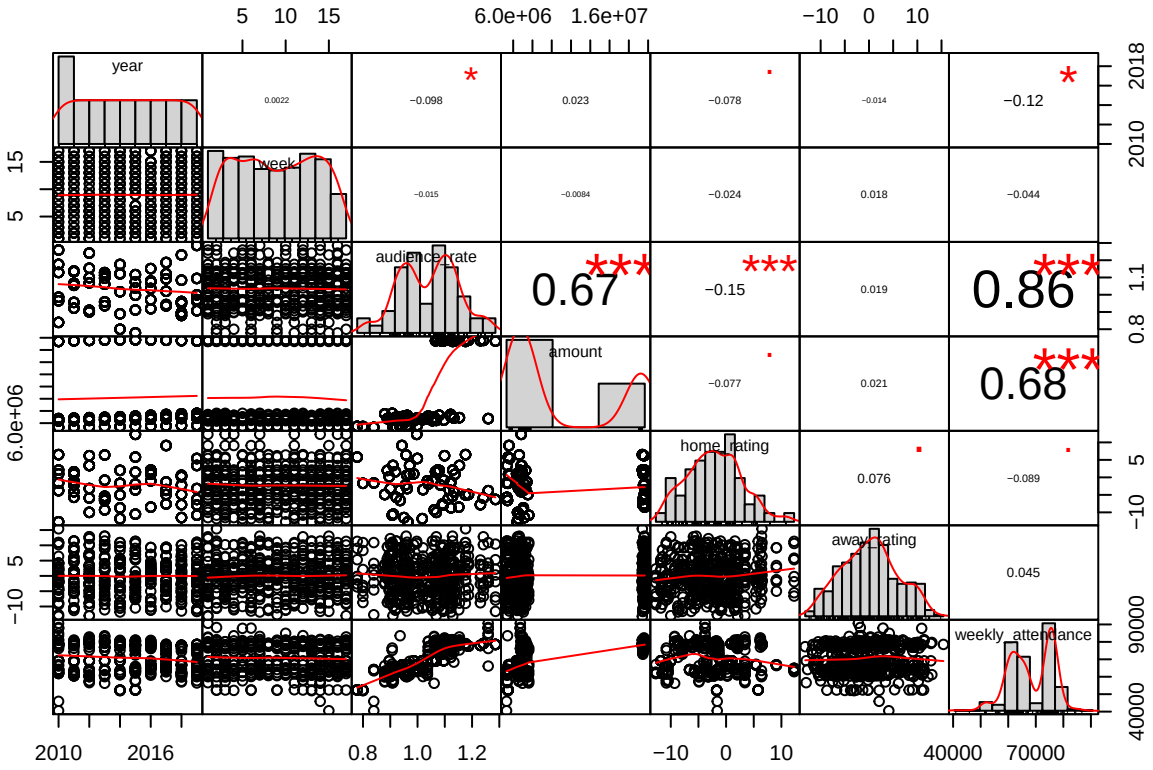
1.3 Model construct

After preprocessing the origin datasets, we get the working dataset with 8 effects and 1 response:

```
##   weekly_attendance year week day time audience_rate  amount home_rating
## 1             62439 2010    3   1    0     0.9647746 6407172       -12.7
## 2             62621 2010    5   1    0     0.9647746 6407172       -12.7
## 3             61857 2010    8   1    0     0.9647746 6407172       -12.7
## 4             61904 2010   10   1    0     0.9647746 6407172       -12.7
## 5             62308 2010   12   0    1     0.9647746 6407172       -12.7
## 6             61874 2010   13   1    0     0.9647746 6407172       -12.7
##   away_rating
## 1         0.2
## 2         2.3
```

```
## 3        -0.6
## 4        -9.4
## 5        -5.8
## 6        -6.7
```

Only *time* and *day* are nominal variables, the rest are all continuous variables. Next we will check the correlation between each effect and response, and find if we can remove any irrelevant variables to simplify the model:



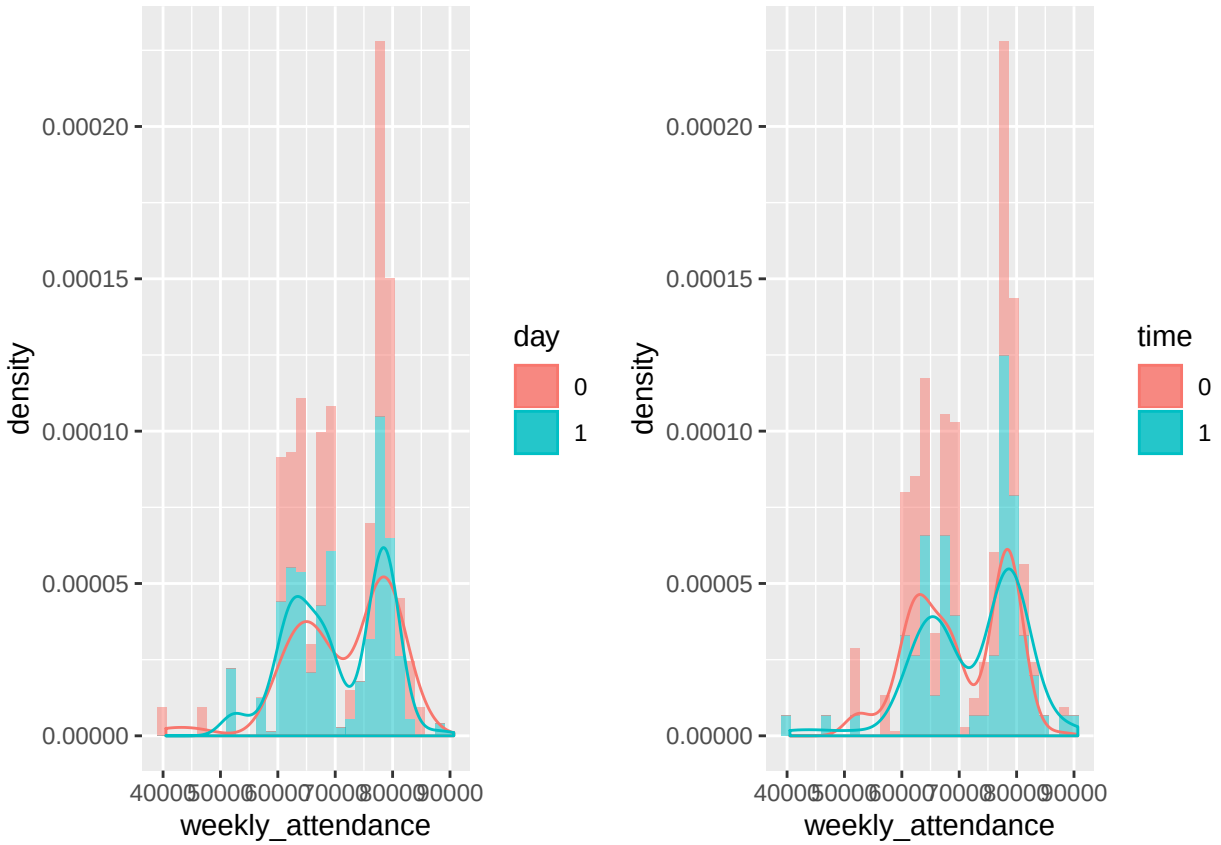
From the plot above, the most left column indicates the correlation between every numeric effect and the response and the row at the bottom indicate the relationship between each effect and the response. We can learn that the most significant effects to the response are *audience_rate*, *amount* and *year*.

It's a little surprised that the strength of the teams doesn't influence the attendance a lot. But on the other hand we can find that the *home_rating* and *audience_rating* are highly correlated, which means that the strength of the home team can be partially explained by the audience rate. And the audience rate is one of the most significant effect to the response, which means that the strength of one team is not completely uncorrelated with the attendance.

What's more, notice that *amount* and *audience_rating* are highly coreelated, which indicates the difference between big state and small state. The bigger the state, or we say the bigger population of a state, the more support there would be to the home team.

Then for the binary variable *day* and *time*, we aim to find the variable that has significantly different performance in two different category. This is the distribution of attendance in two different categoty:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It's hard to tell if the distributions of two categories are different. To quantify it, here uses the Kolmogorov-Smirnov Tests. KS test can quantify the difference between two distributions. If we set the confidence degree equals to 0.1, then we can choose the relative variable based on the degree we set:

```
##   variable    p-value if choose
## 1      day 0.57938415        No
## 2     time 0.07109758       Yes
```

So finally the features we select are *year*, *audience_rate*, *amount* and *time*. Then we constuct the linear regression model as follow:

```
##
## Call:
## lm(formula = weekly_attendance ~ year + audience_rate + amount +
##     time, data = attendance_working)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19469.0  -2870.9     50.9   2466.5  19634.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.052e+05  1.330e+05   2.294   0.0222 *
## year          -1.451e+02  6.588e+01  -2.202   0.0281 *
## audience_rate  5.261e+04  2.243e+03  23.453  < 2e-16 ***
## amount         2.786e-04  4.098e-05   6.797 3.21e-11 ***
```

```
## time           7.631e+02  4.846e+02   1.575   0.1160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4095 on 475 degrees of freedom
## Multiple R-squared:  0.7633, Adjusted R-squared:  0.7613
## F-statistic: 382.9 on 4 and 475 DF,  p-value: < 2.2e-16
```

From the result of the regression, the R-squared is 0.7633, which indicates that the performance of the regerssion is not bad. The most significant effects are *amount* and *audience_rate*. It's not surprised that the population of a state will influence the attendance, because more people means more potential audiences. However, the audience rate is also significant. This indicates that the rate of the yearly attendance of home team and away team will also influence the attendance. That means, in the state will larger population, people will support their team more than the state with small population. The reason might be that people in the big city pay more attention to the football game as a kind of interest or entertainment.

2.Team strength

2.1 Basic idea

In this part, we aim to find something about the estimation to one team. The dataset has provided some relative estimation from SRS (Simple Rating System). There is a formula on how to calculate this score for every team and it is based on some basic information like wins, losses, points for each game. What we want to do is to find which of these effects influence the estimation most.

To study whether the estimation from SRS is correct, we need to find a response and set the estimation from SRS as the comparison group. Here uses whether the team made the playoffs as the response. We can also use whether the team won the superbowl as the response, but as we know, football games depend on fortune sometimes and also depend on the state of the players and it is hard to judge a team's strength by one game. Whether made the playoffs is a good choice. It not only can indicate the average strength of a team but also can simplify the problem. Besides, the data of it is more sufficient, which will lead to a more reasonable result.

2.2 Data preprocess

As mentioned, we aim to find how the response, whether the team made the playoffs, is affected. Therefore, we plan to use two models, one uses the estimation of SRS, the other uses the origin information. To compare these two models, we use cross validation to test their relative performance.

For the origin data, we first set the response *playoffs* equals to 1 if the team made the playoffs and equals to 0 if not. In this part, we remove the variable *year* because we assume that the estimation only depends on the performance of the team and would not be influenced by other effects like location and time. Or in other word, the performance of team has included some information of the other effects. The performance of the team is the direct effect to the final estimation to the team.

To implement the cross validation, we devide the dataset into two equal parts randomly as the training set and the test set.

2.3 Model constuct

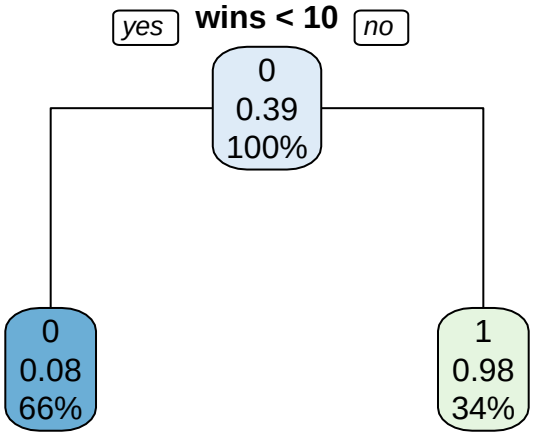
Here we use two kinds of models to study this problem. For the estimation from the SRS, there are totally 5 variables: *simple_rating*, *margin_of_victory*, *strength_of_schedule*, *offensive_ranking* and *defensive_ranking*, where *simple_rating* = *margin_of_victory* + *strength_of_schedule* = *offensive_ranking* +

defensive_ranking. To consist as many effects as possible, here uses all the variables expected *simple_rating* because it is the linear combination of others.

Logistic regression is suitable for the binary response. Here construct the logistic regression model with the training data:

```
##
## Call:
## glm(formula = playoffs ~ ., family = "binomial", data = standings_working_logistic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0288  -0.4807  -0.1018   0.3629   2.8055
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.1414     0.2986  -3.823 0.000132 ***
## margin_of_victory     -8.6322     4.2018  -2.054 0.039937 *
## strength_of_schedule  -8.8855     4.2122  -2.109 0.034905 *
## offensive_ranking      9.1366     4.2222   2.164 0.030469 *
## defensive_ranking      9.0925     4.2278   2.151 0.031505 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 213.64  on 159  degrees of freedom
## Residual deviance: 101.17  on 155  degrees of freedom
## AIC: 111.17
##
## Number of Fisher Scoring iterations: 6
```

For the model with origin information, here uses the decision tree because the decision tree can show us clearly which variables is the most effective variable to the response. This is the decision tree after pruning:



It's surprised that if we set whether the team made the playoffs as a standard of a team's strength, the response only depends on the wins.

Then we calculate the accuracy, sensitivity, specifity, precision and F1-score for both model with the test data to compare their performance on the test data. This is the final result:

```
##           model accuracy sensitivity specifity precision  F1-score
## 1      logistic  0.87500   0.7931034 0.9215686 0.8518519 0.8214286
## 2 desicion tree  0.91875   0.8620690 0.9509804 0.9090909 0.8849558
```

We can find the decision tree even perform better than the logistic model. As a result, we can say that the main effect to whether a team would made the playoffs. However, it's obvious that the wins would influence whether it would make the playoffs. This response may can't indicate everything of a team's strength because wins depends on too many other factors and it's also hard to judge whether the estimation to a team is correct or not.

3. Games

3.1 Basic idea

Although the three parts analysis are independent with each other, we can still take advantages of others idea to do the analysis. Similar to the study of attendance, we also separate all the origin information into three parts, location, time and the team. However, this time, the response is whether the home team won the game and the attendance becomes one of the effects.

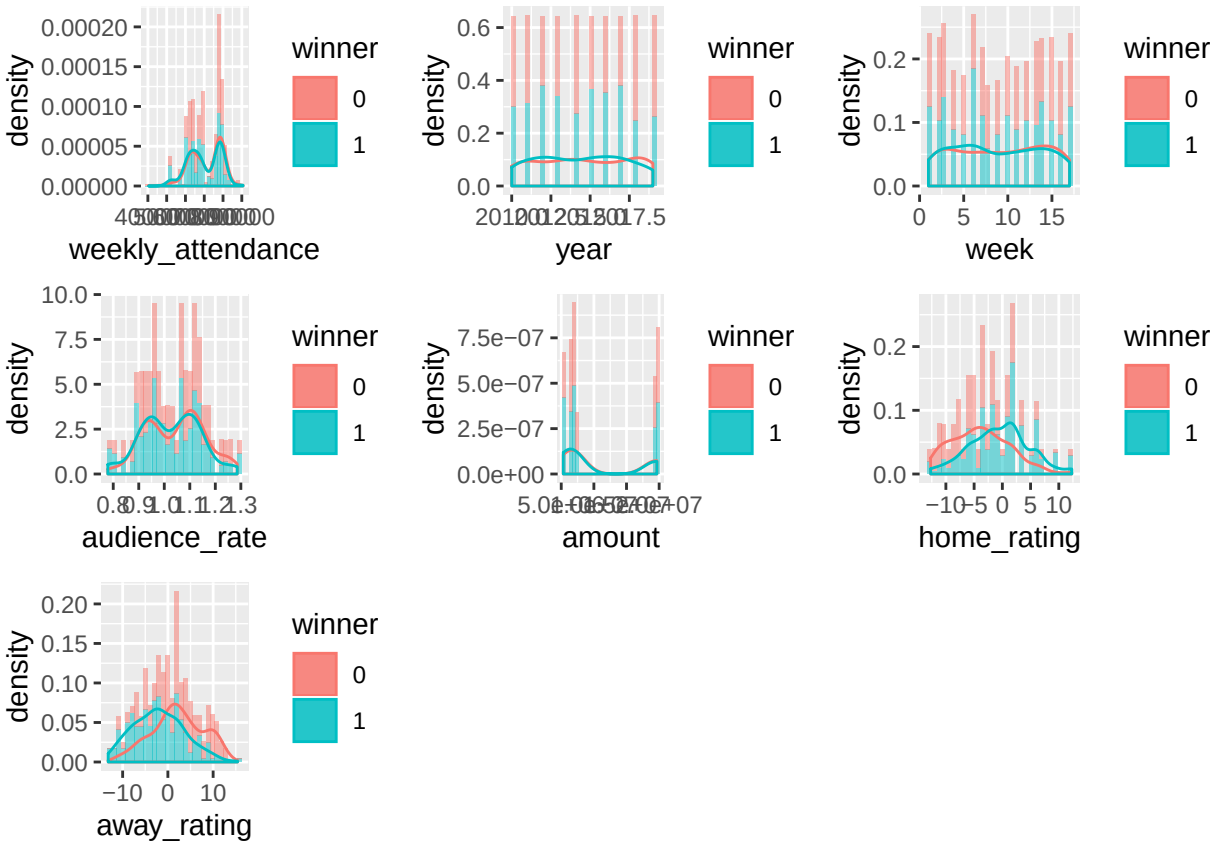
3.2 Data preprocess

Similar to attendance, transform the variables into relative form.

3.3 Model construct

Similar to attendance, we still want to do the features selection for the model. Unlike attendance, the response here is a binary response. For the continuous variables, we want to see if the variable has the same distributions for each outcome category:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Instead of checking the correlation between each effect and response, we want to test if the distributions of each variable in two outcome category are same. Here uses KS test for continuous variables and P test for binary response to check each variable's distributions in two different group. If we set the confidence degree equals to 0.1, then we will get:

```
##            variable      p-value if choose
## 1 weekly_attendance 3.982317e-01        No
## 2              year 3.508425e-01        No
## 3              week 9.347015e-01        No
## 4     audience_rate 1.538806e-01        No
## 5            amount 1.739122e-01        No
## 6       home_rating 2.592123e-08       Yes
## 7       away_rating 2.917666e-12       Yes
## 8               day 2.433308e-02       Yes
## 9              time 9.525411e-02       Yes
```

From the result we get, we would use *home_rating*, *away_rating*, *day* and *time* as the main effects to the result of one game. In fact, the strength of the team is the main factor that may affect the result of the game. Other factor like location still is not significant enough to influence the result of the game. The interesting thing is that the time of the game can influence the result of the game more or less. It may because that players have different state at different time.

Then we build the logistic model using the features we get:

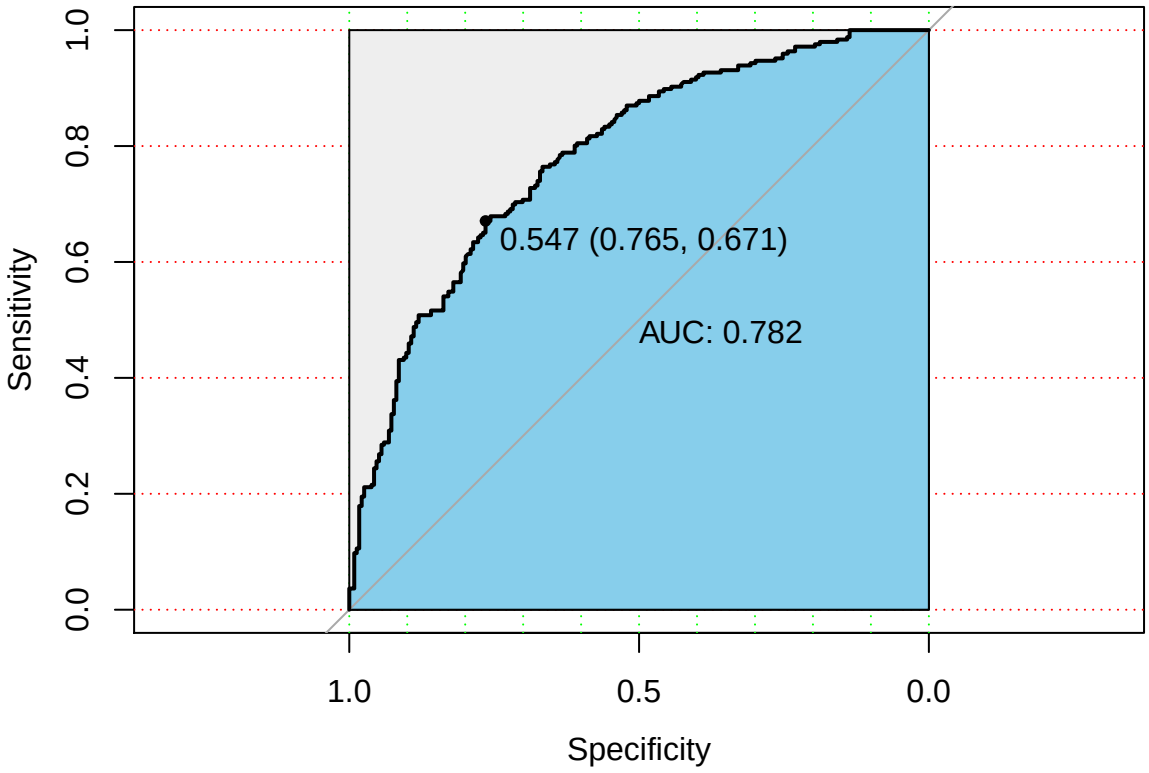
```
##
## Call:
## glm(formula = winner ~ home_rating + away_rating + day + time,
```

```
##     family = "binomial", data = games_working)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2601  -0.9362   0.3617   0.9137   2.0325
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.38443    0.52650  -0.730   0.4653
## home_rating  0.15800    0.02165   7.298 2.92e-13 ***
## away_rating -0.15642    0.02013  -7.771 7.77e-15 ***
## day          0.86845    0.51046   1.701   0.0889 .
## time         0.16608    0.42919   0.387   0.6988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.12  on 479  degrees of freedom
## Residual deviance: 537.52  on 475  degrees of freedom
## AIC: 547.52
##
## Number of Fisher Scoring iterations: 4
```

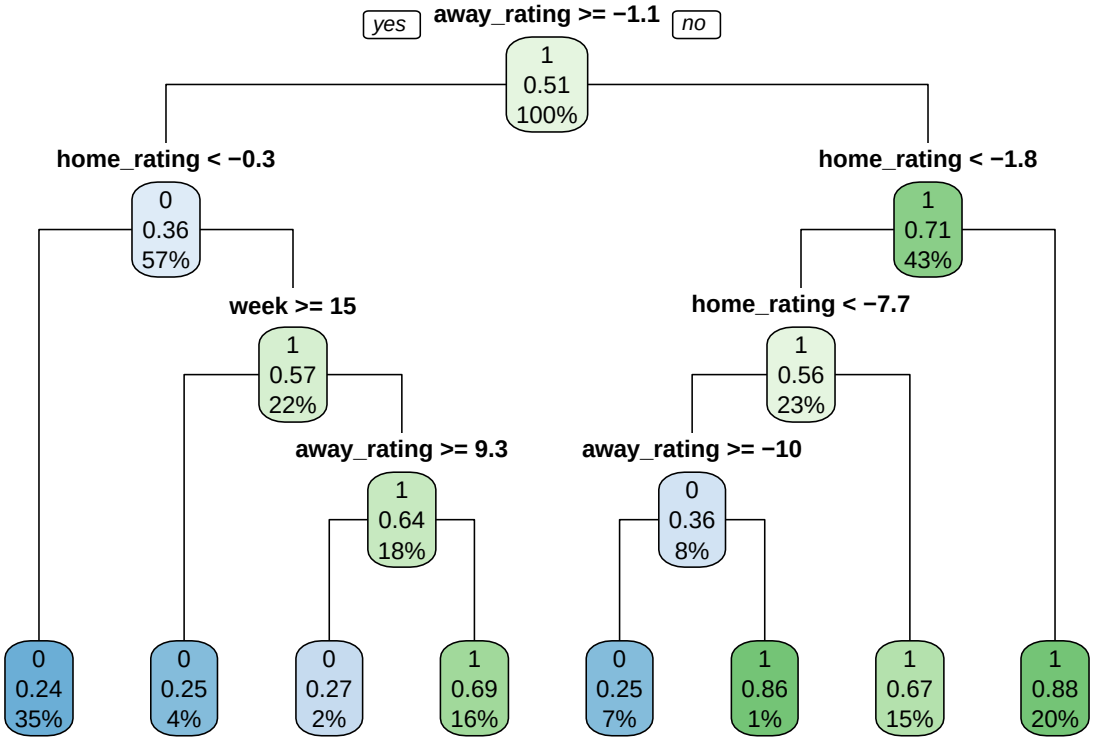
This is the ROC curve and relative AUC. AUC here equals to 0.782 which is an ordinary result.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Next we use a decision tree with all the original information to check how the effects influence the final result. This is the tree after pruning:



Still, the strength of the teams is the mainly factor that affect the result of a game. However, the interesting thing is that *week* becomes one of the main factor, instead of the *day* or *time*.