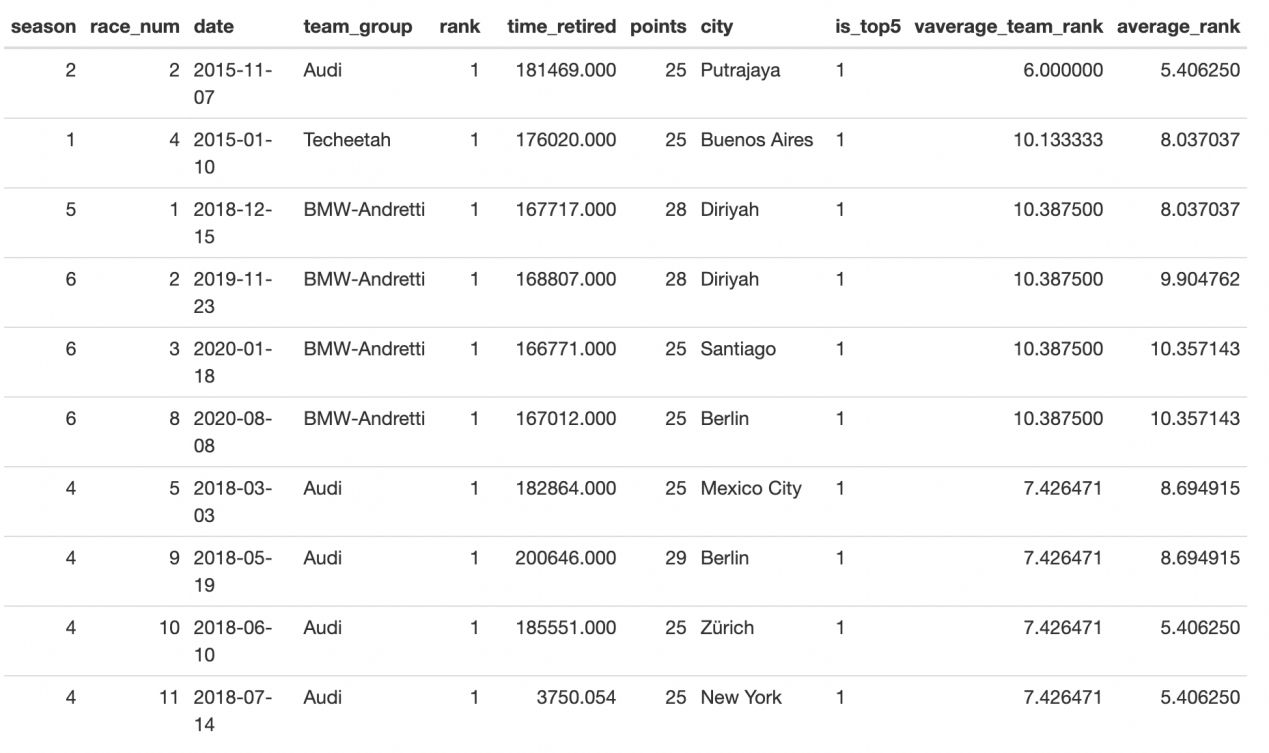
Firstly, based on the previous data processing, the data format was processed into a more reasonable model. The tasks include turning the time\_retired into minutes, creating the split-box operation between different cities, and setting the definition of a new variable is\_top5 which means whether the rank ranking is in the top five. The corresponding data were converted to the appropriate format that NA values were removed and the appropriate variables were filtered out to construct the new data set "data".

The code is :

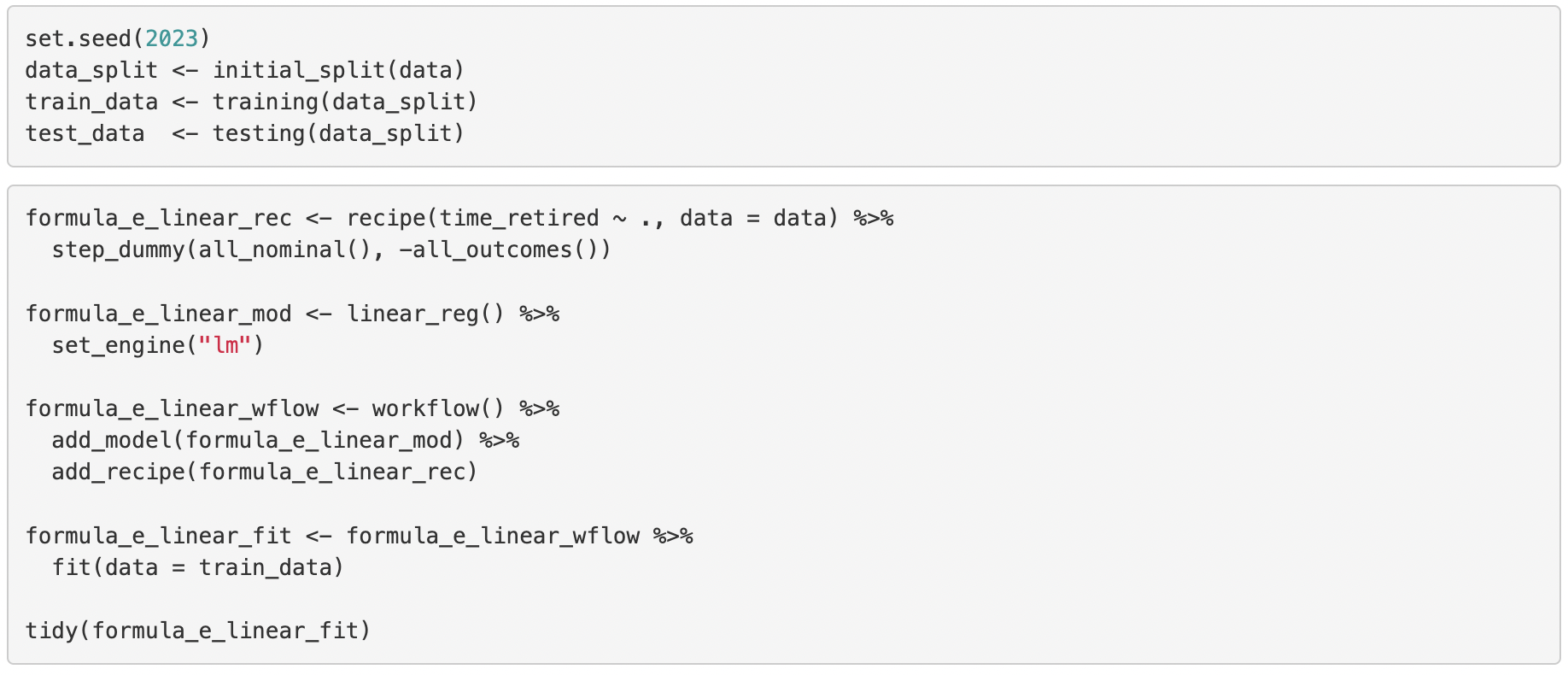


The output is ：

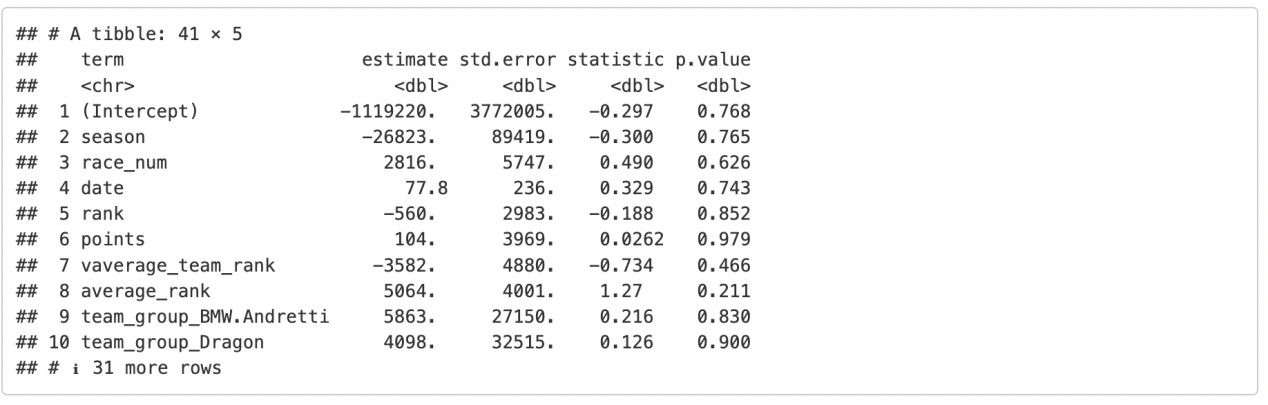


Secondly, the data is cut into training and test sets in the ratio of 3:1. Meanwhile, time\_retired is used as the dependent variable, and the rest of the variables in data are used as the independent variables.

The code is :

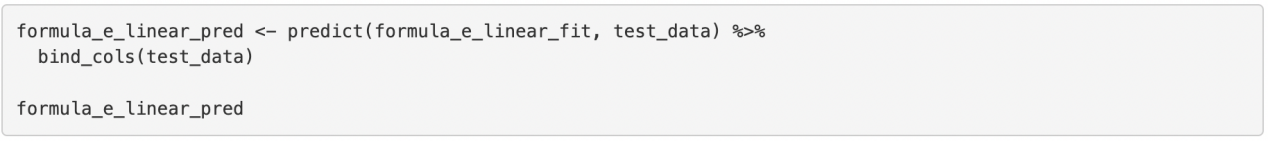


The output is ：

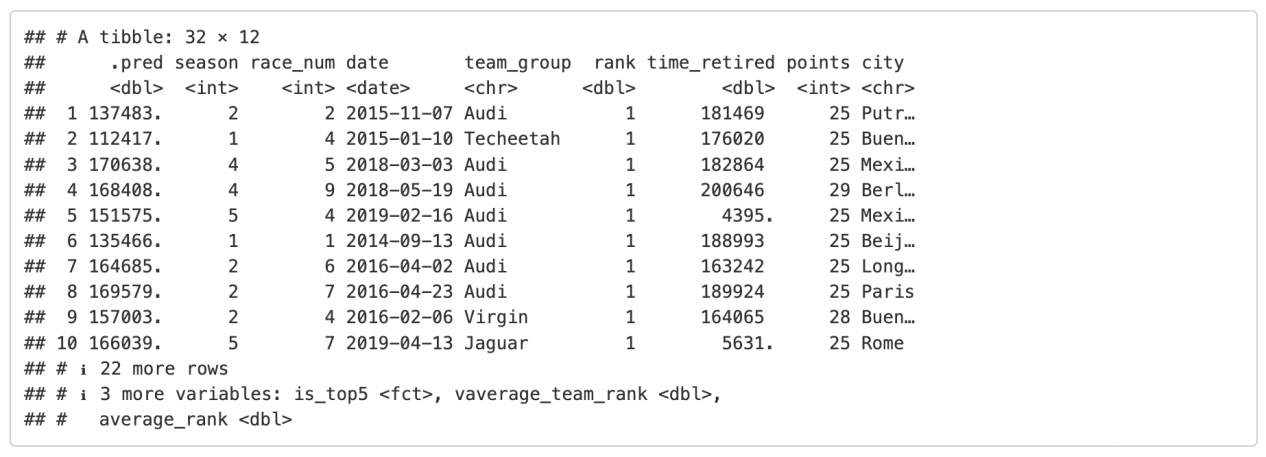


It is obvious that Time\_retired is a continuous numerical variable, so we use the linear model to do the initial modelling.

The code is :



The output is :

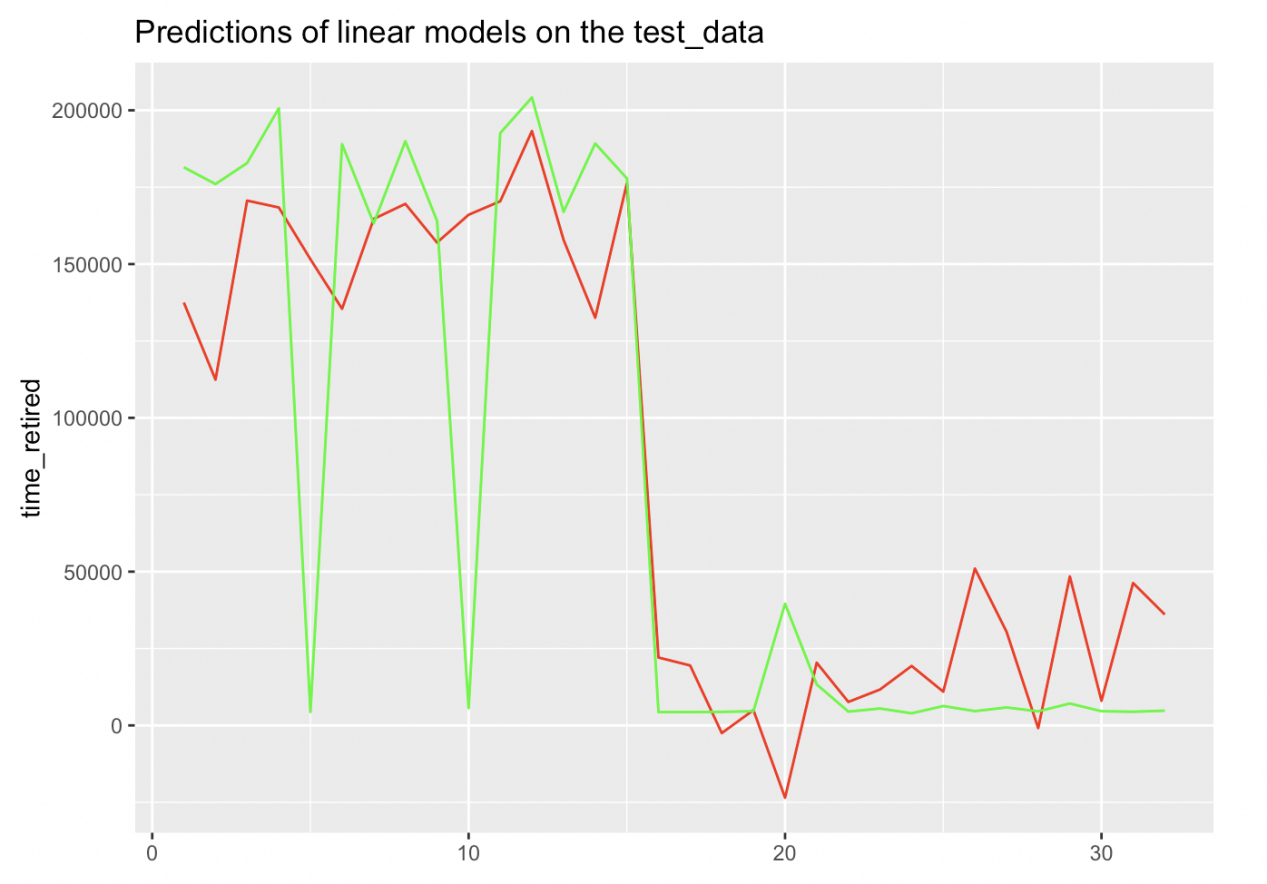


After constructing the recipe and further adopting the model for fit, The curves fitted by the model on the test set are shown in the pictures:

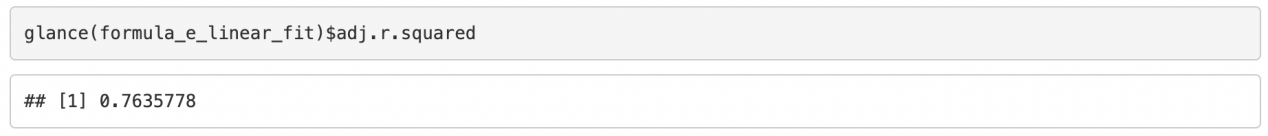
The code is :



The output is :



The code and output of the adjusted R^2:



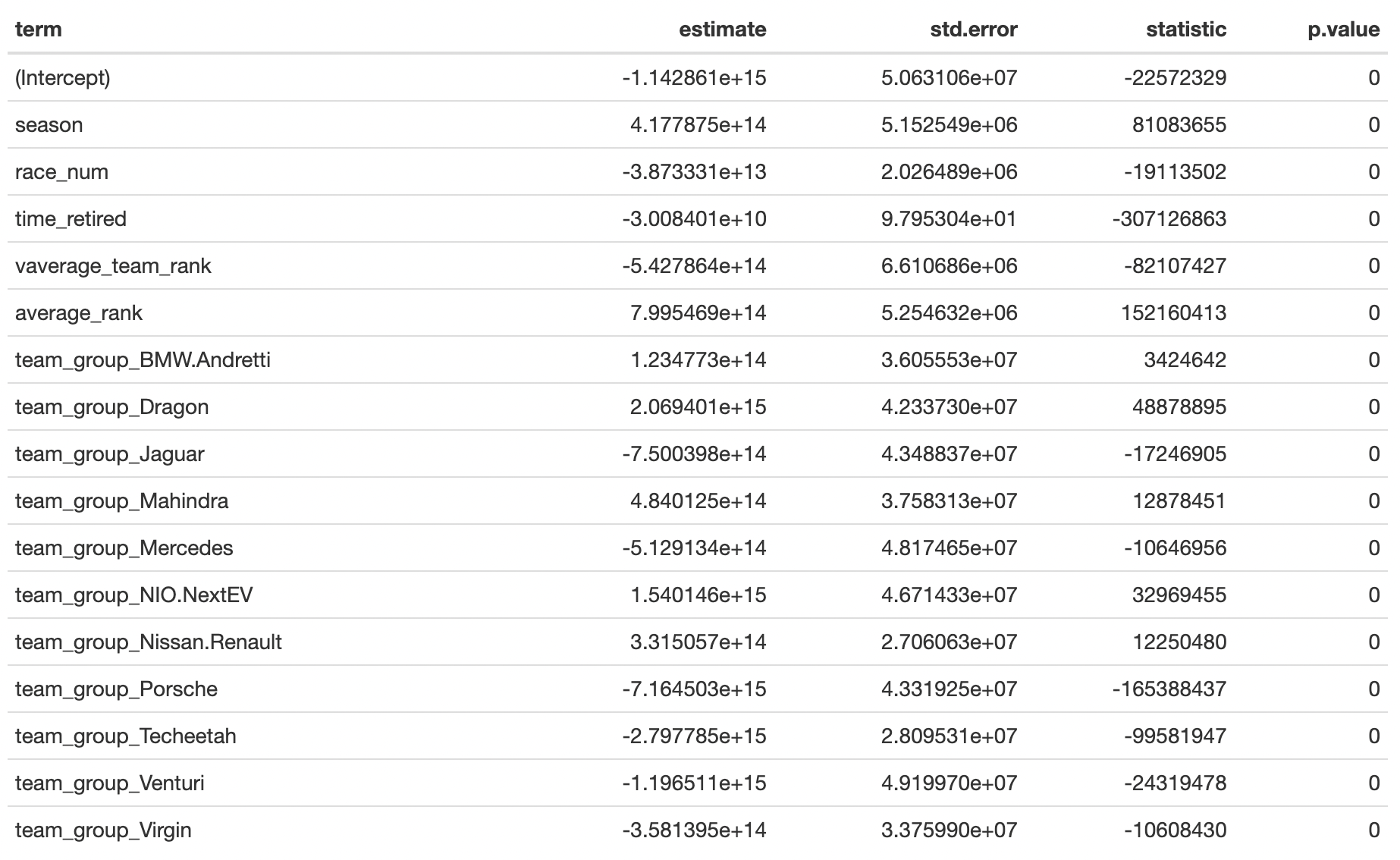
It can be seen that the linear model is able to broadly identify trends in the data as well as specific interpretable mathematical formulas, and the adjusted R-squared of the model fit is 0.764. But it must be improved if there are the more precise dimensions.

Finally, the is\_top5 variable created earlier is used as the dependent variable, and other variables in the dataset "data" such as season, race\_num, etc., except for rank, date, points, and city, are used as independent variables for the next step. Considering that is\_top5 is a category data with two values, which is a typical binary classification problem, we adopt logistic regression as the model.

The code is :

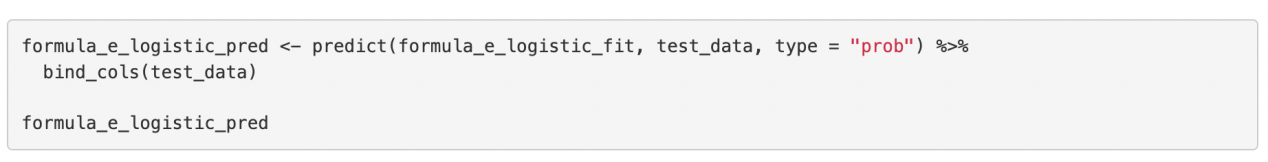


The output is :

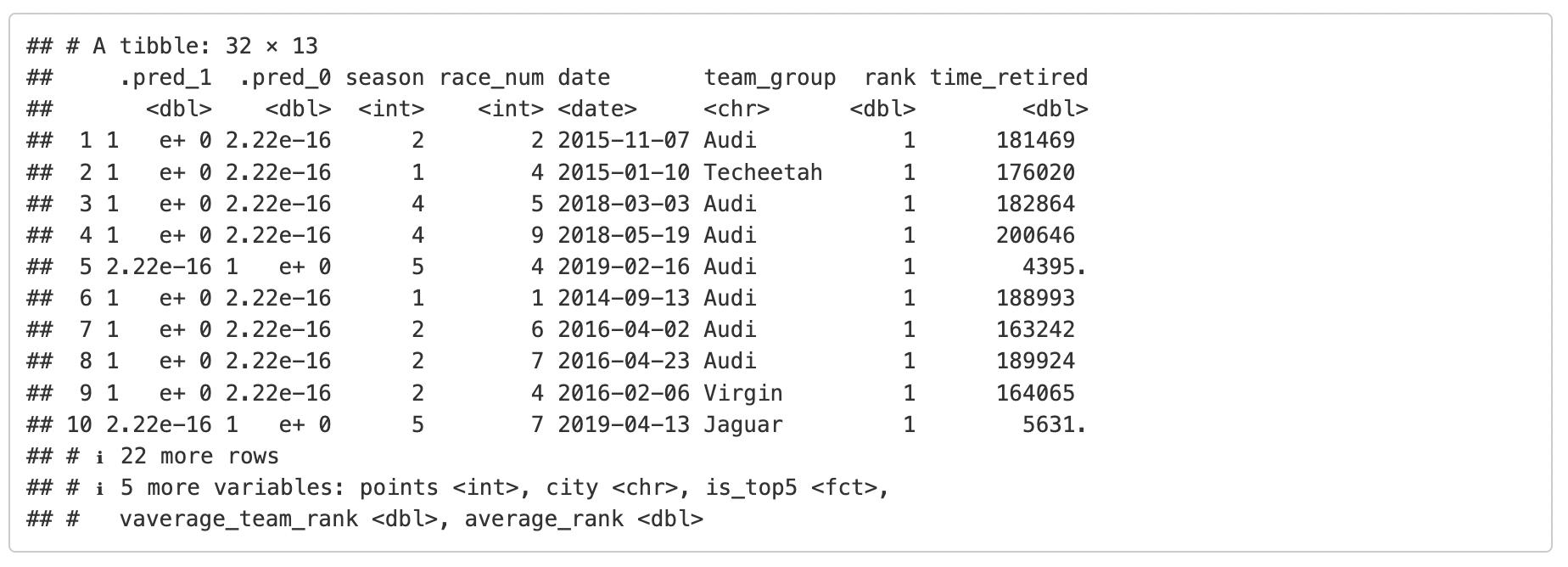


The similar way that build recipe and then use the model to fit.

The code is :



The output is :

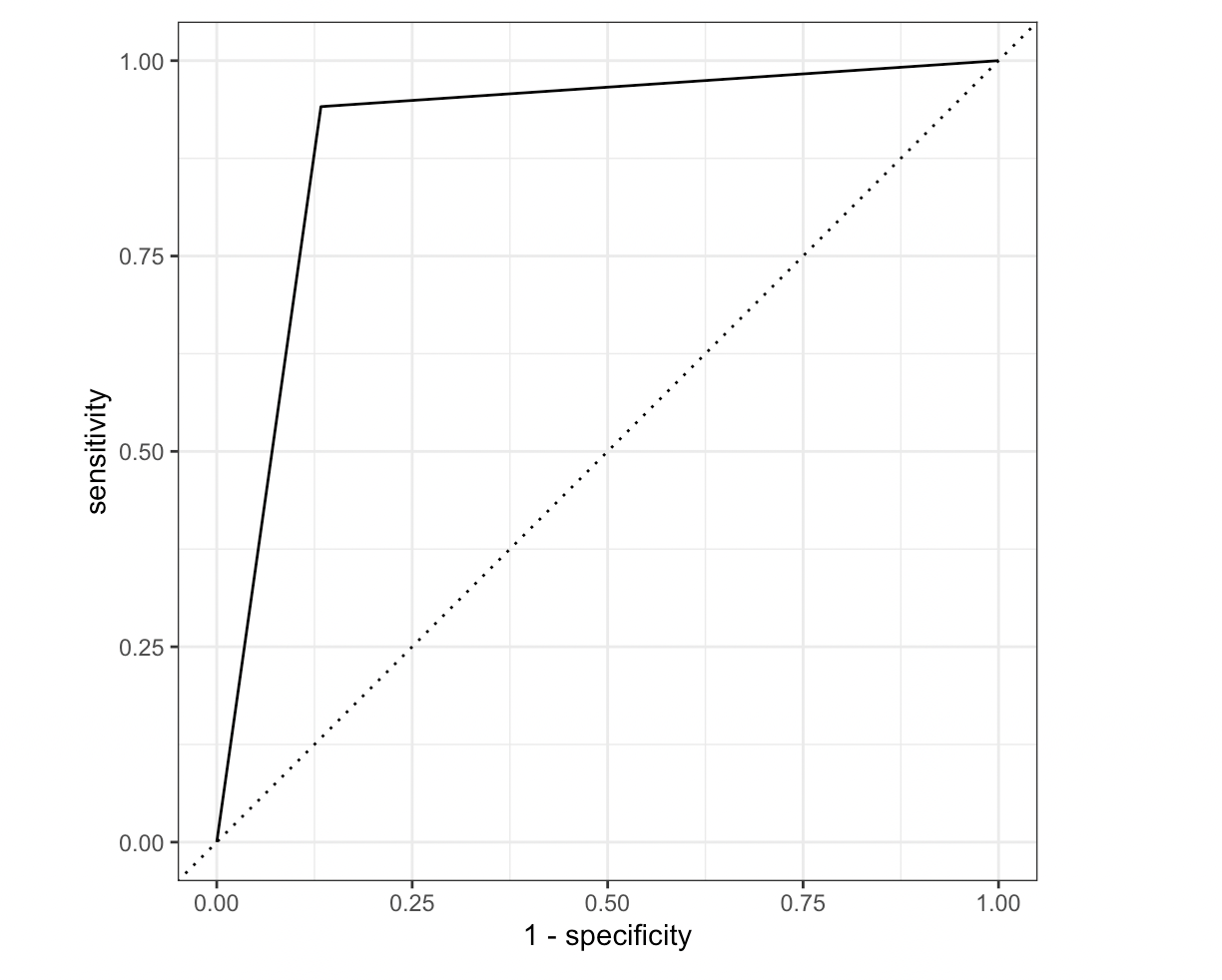


As can be seen from the roc curve, the model fit is relatively good.

The code is :



The output is :



and it corresponds to an AUC value of 0.904.

The code is :



The output is :

