Final Project

By Changhao Ying, Ruijin Jia, Jialei Chen

12/12/2019

1. Dataset Description

The theme of our group project is about football players in FIFA. The materials we use are Detailed attributes for every player registered in the latest edition of FIFA 19 database. The data are obtained from Kaggle (https://www.kaggle.com/karangadiya/fifa19) and the original ones are scraped from <https://sofifa.com/>.

Our dataset contains 18208 rows(That means 18207 football players). There are 87 variables. Including basic information like ID, name, age, photo, nationality, flag, overall(overall rating on a scale of 100), potential(potential rating on a scale of 100), club, value, wage, special, preferred foot(left/right),international reputation(on a scale of 5), weak foot(how good use weak foot, on a scale of 5), skill moves(on a scale of 5), work rate, body type, real face, jersey number, height weight, positions(26 positions including( LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB). The 26 positions have different values and have some floats(like 85+3). And there are 34 numeric variables including the Scores for Attacking (Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys), the Scores for Skill (Dribbling, Curve, FKAccuracy, LongPassing, BallControl), the Scores for Movement (Acceleration, SprintSpeed, Agility, Reactions, Balance), the Scores for Power (ShotPower, Jumping, Stamina, Strength, LongShots), the Scores for Mentality (Aggression, Interceptions, Positioning, Vision, Penalties, Composure), the Scores for defending (Marking, StandingTackle, SlidingTackle), the Scores for Goalkeeping (GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes).

Based on the information in the dataset, our group wants to explore several questions about the football players. The below table shows the problems we want to explore and the model type we used:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Question | Variables | Method |
| 1 | How is wage distributed? Can we model wages with soccer players’ features? | Full variables adding BMI and removing ID, name, flag, photo, club and joined | Linear Regression |
| 2 | Is players’ overall rating calculated from a formula of abilities rating? And what factors influence the overall rating most? | 34 variables of abilities | Linear Regression |
| 3 | Is the preferred foot of a FIFA player was affected by his abilities rating? | 34 variables of abilities | Logistic Regression |
| 4 | Whether the player’s international reputation is affected by their technical scores? | 34 variables of abilities | KNN |
| 5 | Whether the preferred foot used is affected by players’ technical scores? | 34 variables of abilities | KNN |

1. Exploratory Data Analysis

Firstly, we checked whether the index is unique and whether there is any NA values. We added a column called index whose values turn from 0 to 18206. We turned some variable names to easy type like ‘Preferred Foot’ turned out to be ‘Preferred\_Foot’. We also used lambda method to turn wage, weight and height from ‘series’ to be ‘int’ type. For example, the original value of wage is $565k, the new value of wage is 565, original height is 5'7 and the new one is 170.18. We increased one more variable BMI by calculating weight(kg)/height(m)^2. After all set, our dataset is left to be 14743 values.

And then we import necessary packages to help us do the model construction:

from sklearn.model\_selection import train\_test\_split

from sklearn import neighbors

from sklearn.metrics import mean\_squared\_error

from math import sqrt

import matplotlib.pyplot as plt

from sklearn.neighbors import KNeighborsClassifier

from sklearn import metrics

from sklearn.preprocessing import scale

from sklearn.model\_selection import cross\_val\_score

from sklearn.metrics import classification\_report

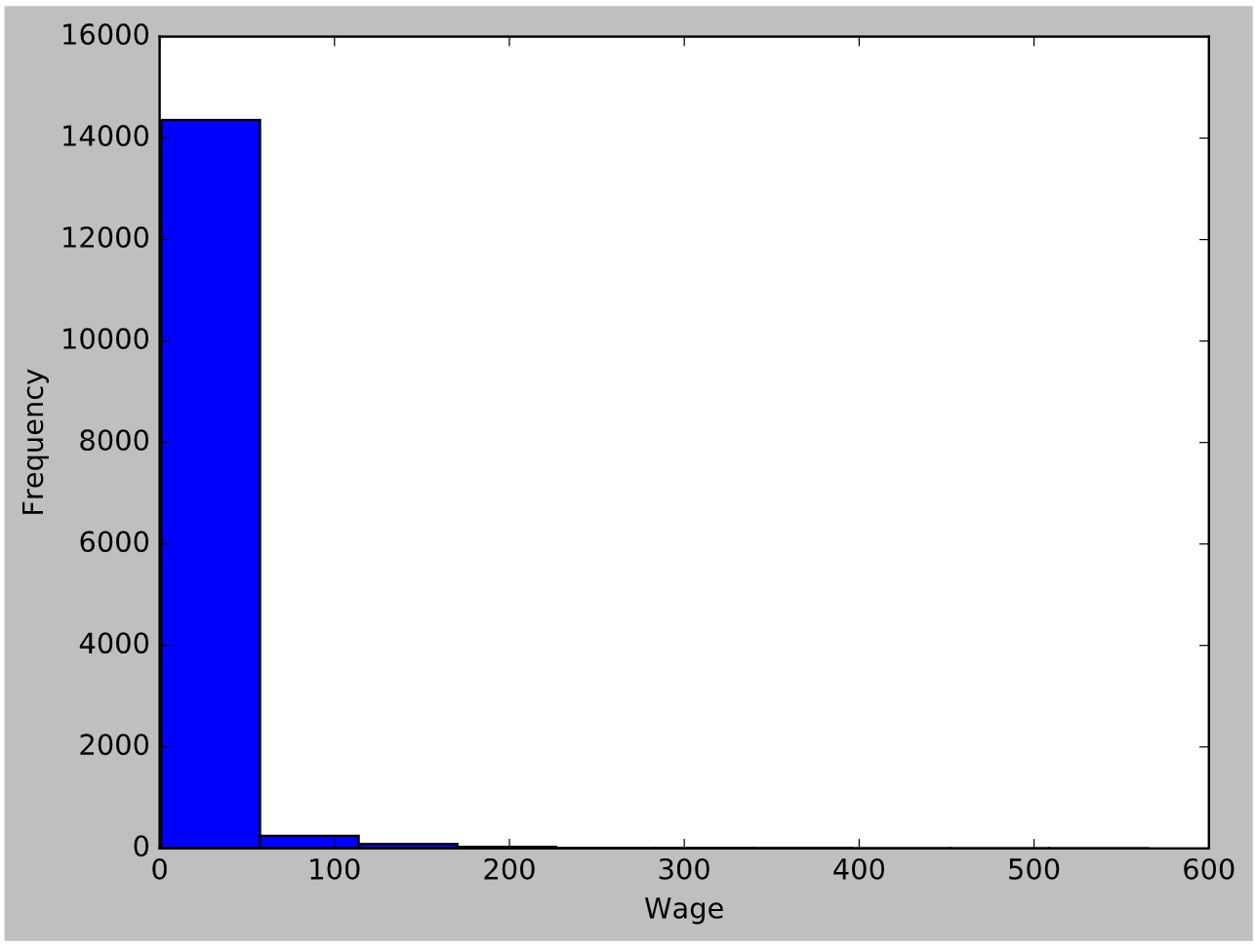
import pandas as pd

Finally, we start to do our analysis on the dataset.

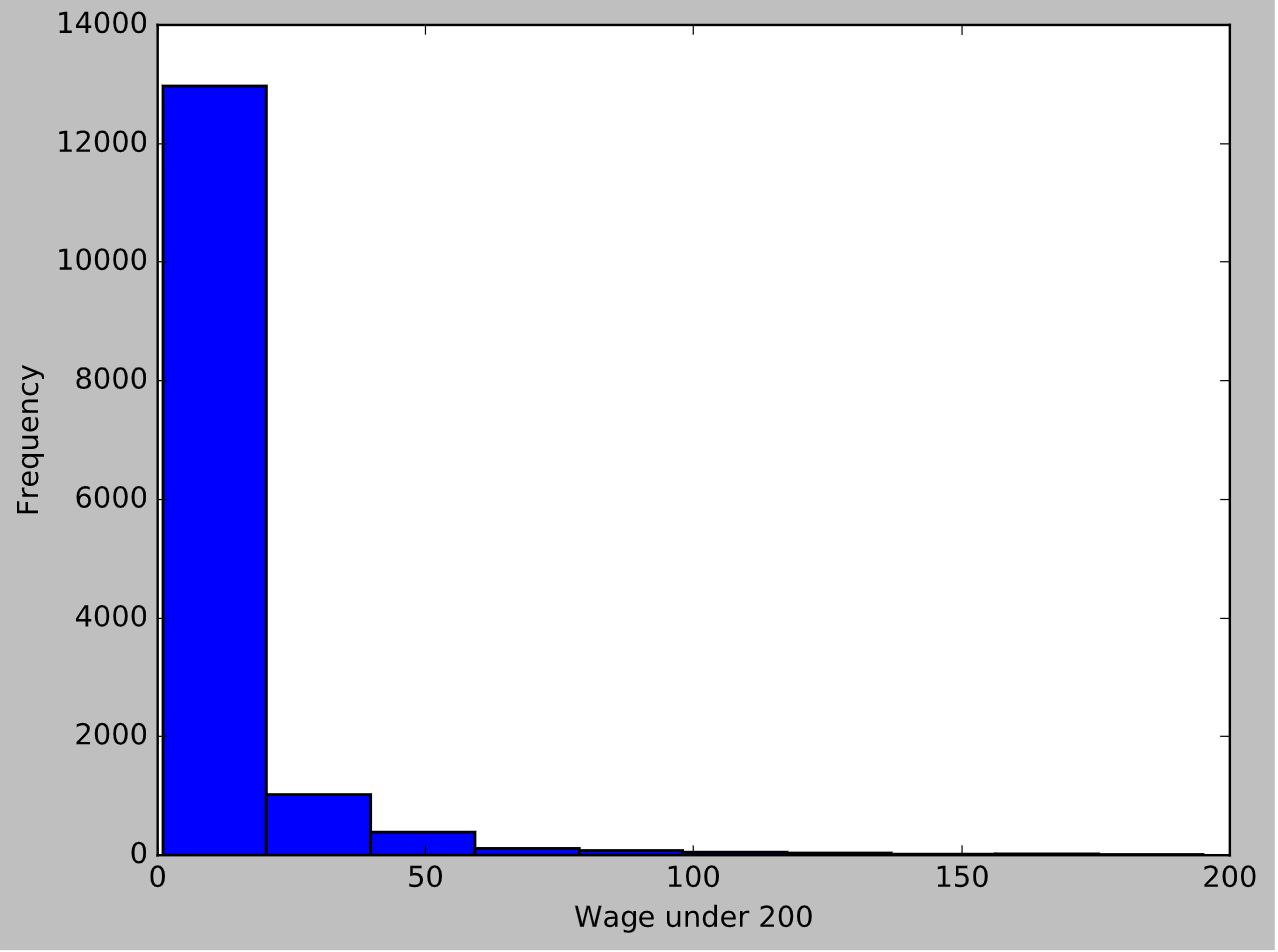
1. Data Analysis

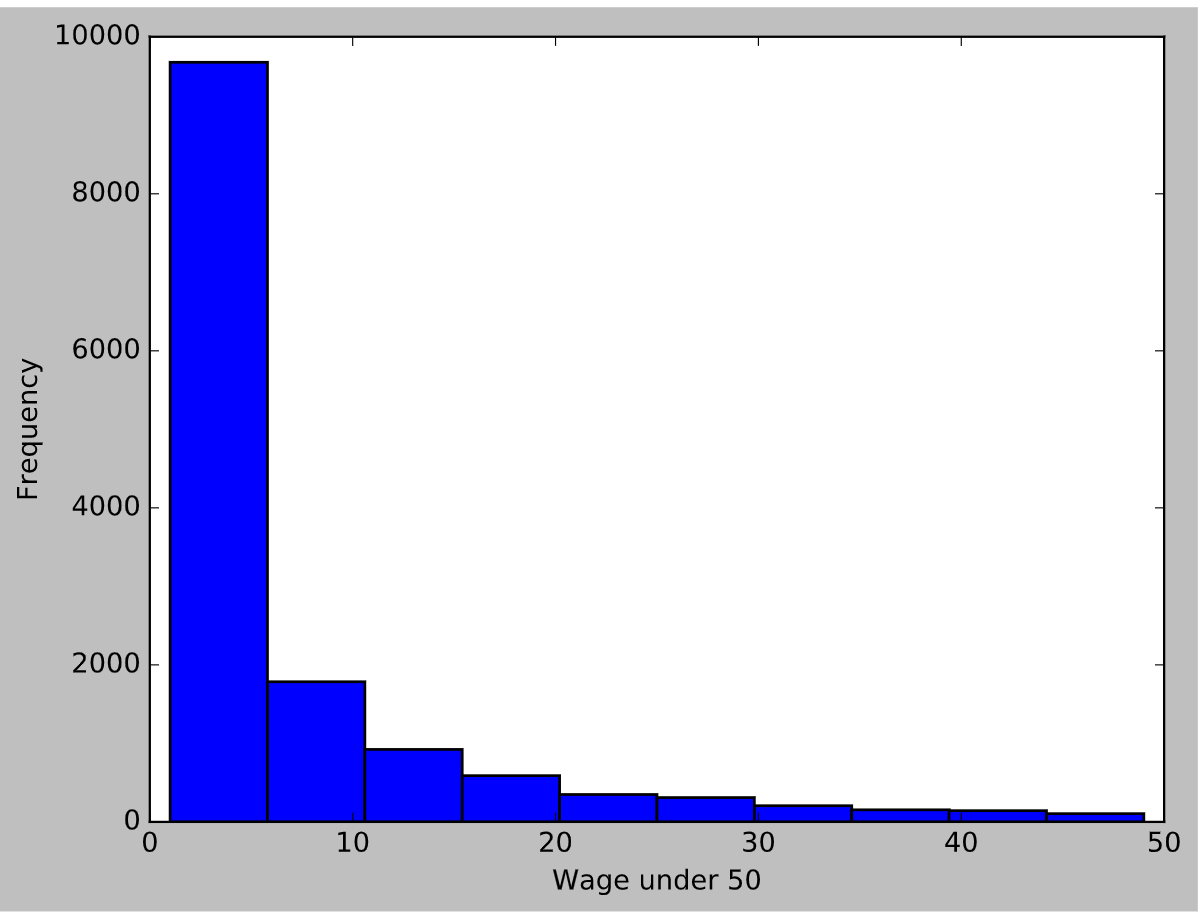
3.1 How is wage distributed? Can we model wages with soccer players’ features? (Linear Regression Model)

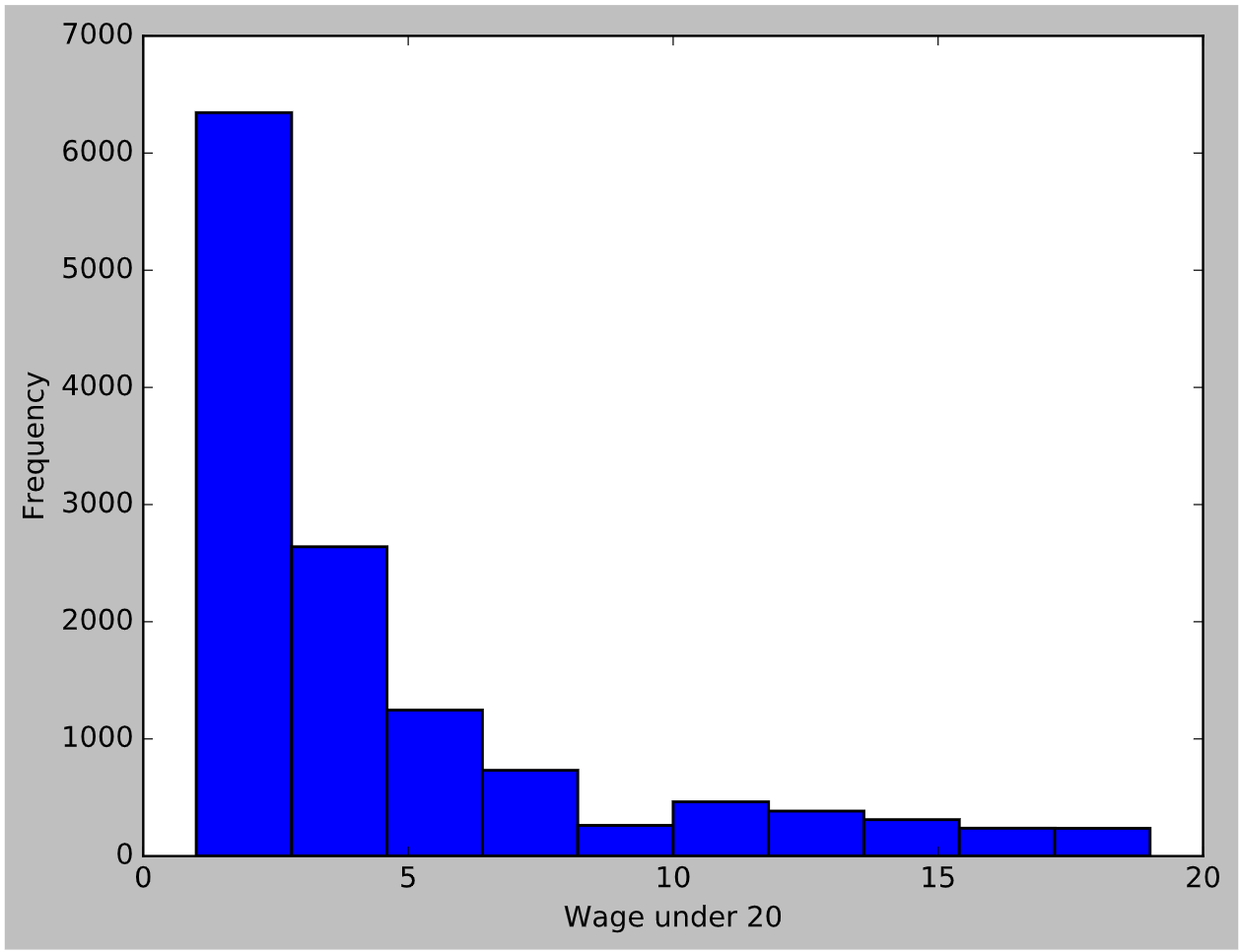
In this section, we will look into how wages are distributed for all soccer players among different group, and whether we could predict wage through linear regression model. First we would have a look at how wages are distributed for all soccer players.



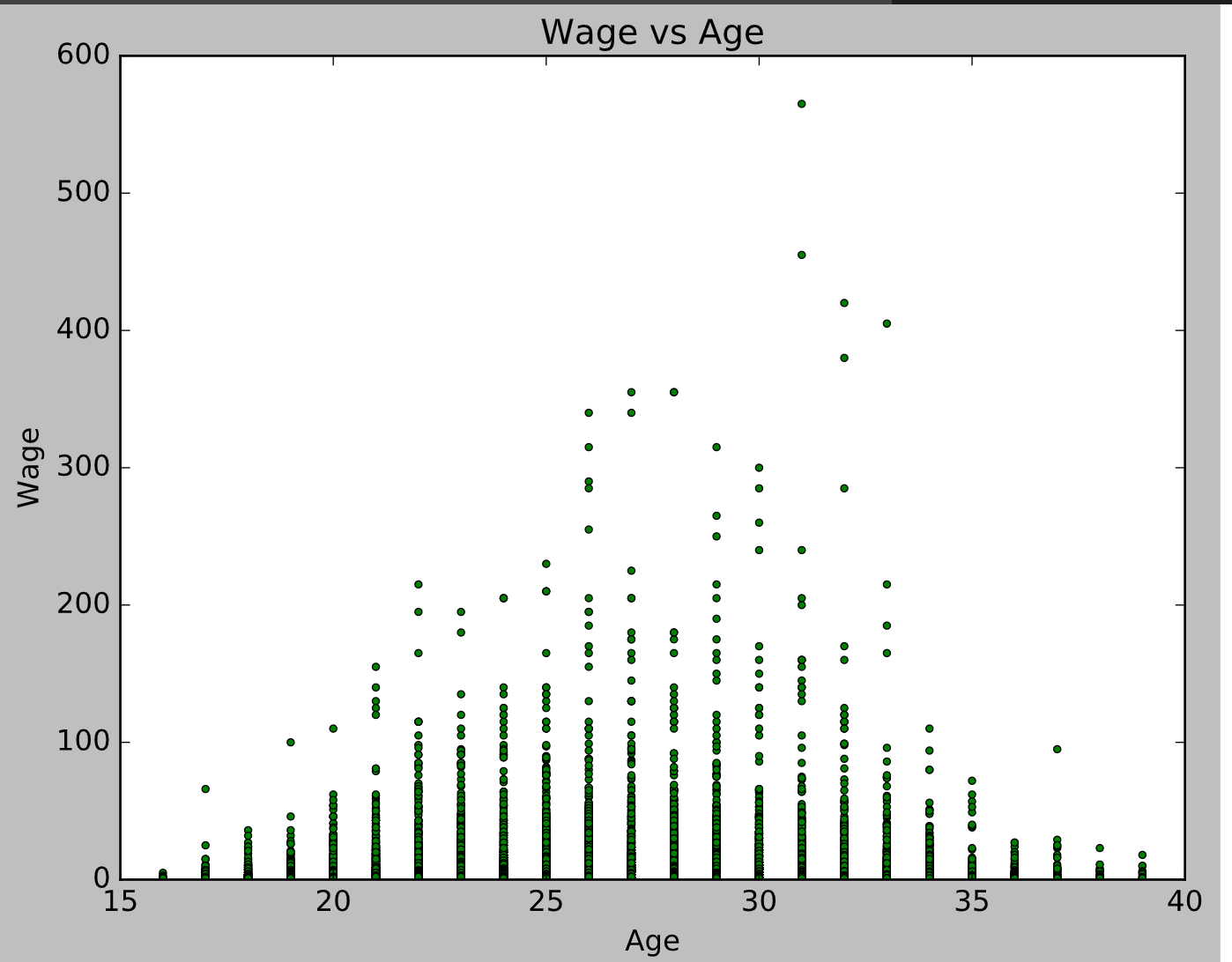
From the chart we can see most of the soccer players have wage under 50, so we cut off outliers and we can got new chart wage under 200, wage under 50 and wage under 20 below:



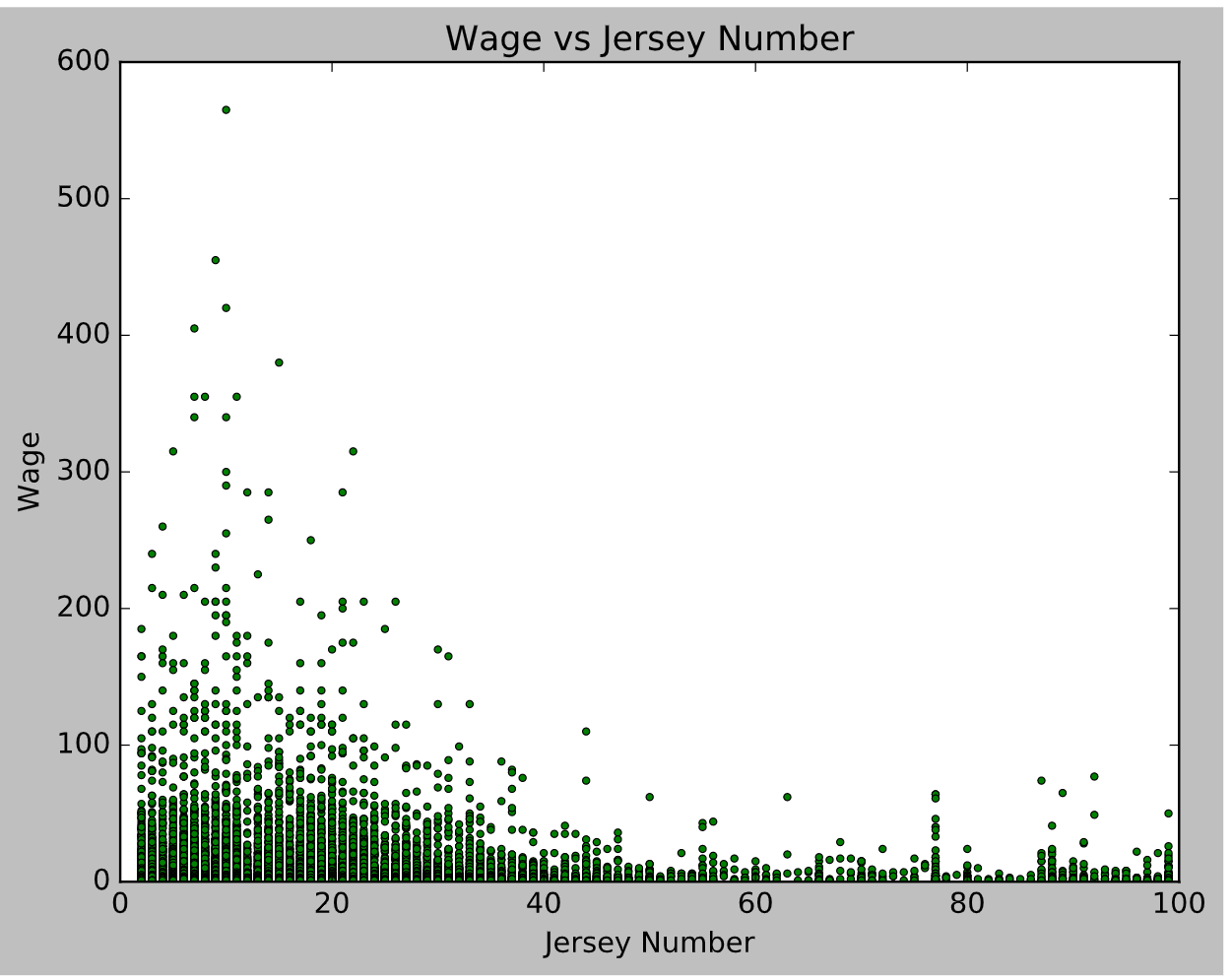




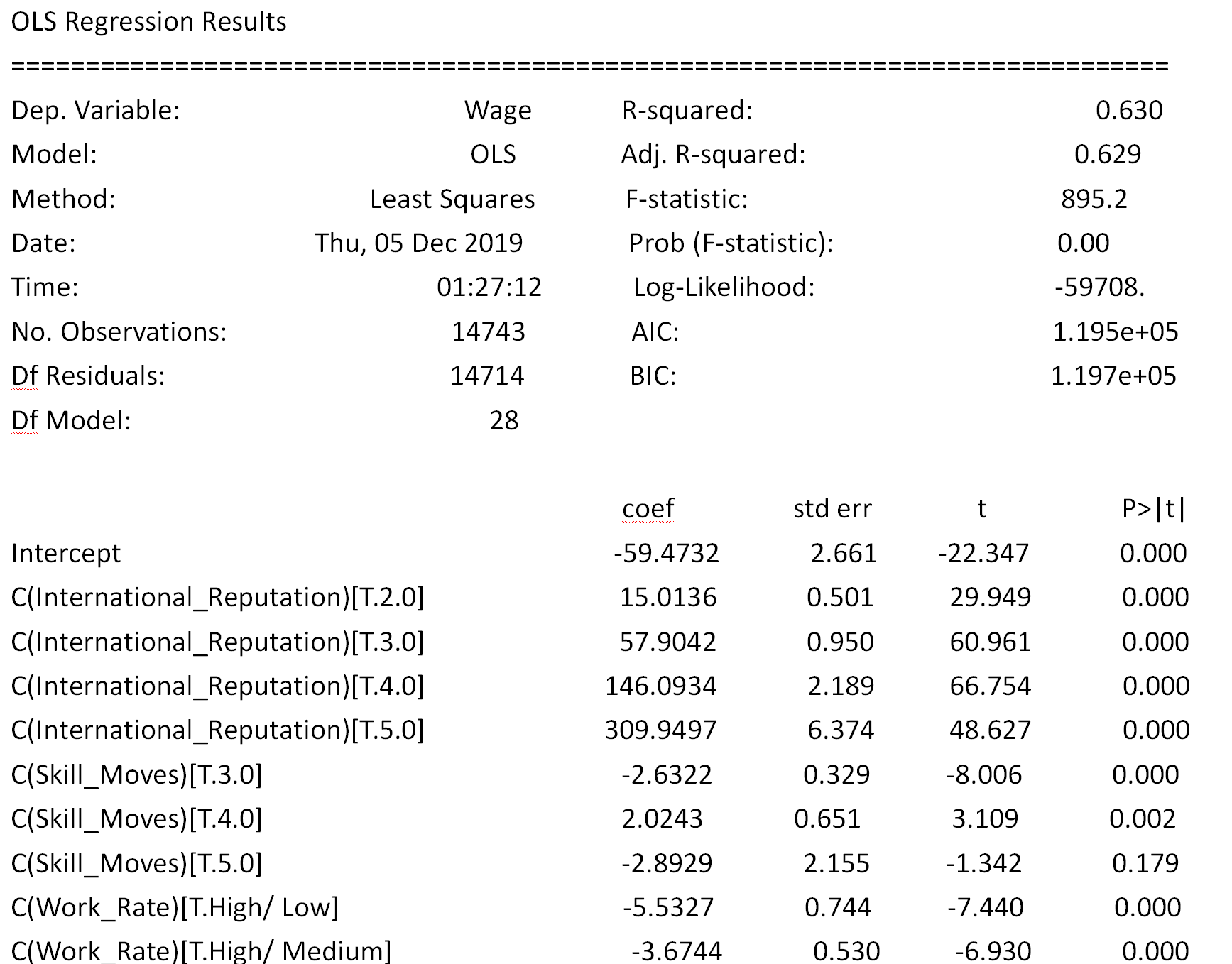
And then we have a look at how wages are distributed against age. Soccer players who are bewteen 25 to 30 have relatively higher wages in average. However, for wage outliers, they did not fall into this range. Therefore, for toppest player, age is not an important restriction factor for high wages.

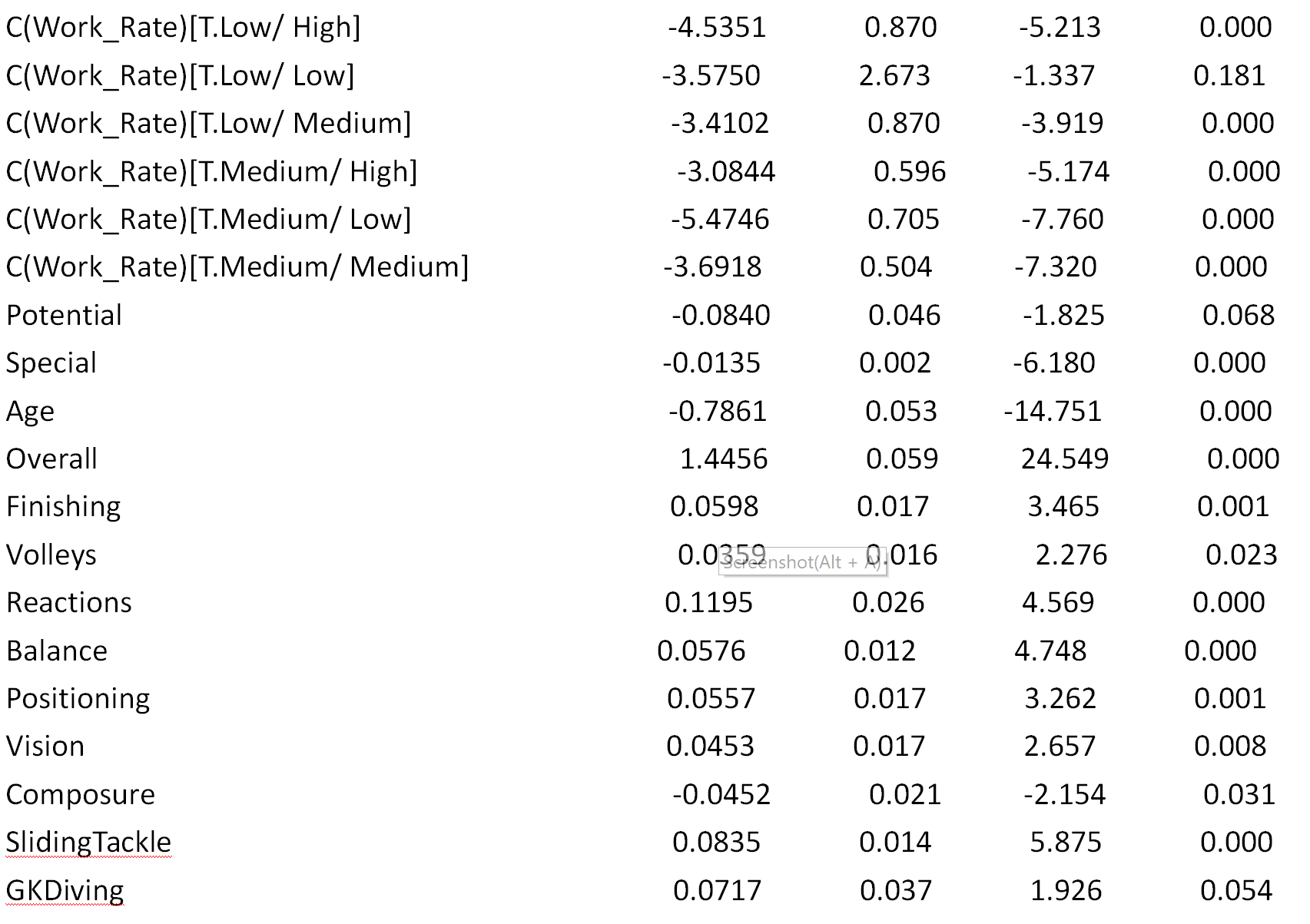


In addition, we would look into how wage is distributed against jersey numbers. Because public or some players believe the lucky number, and in the chart, wages tend to be higher for jersey number around 10, including outlier wages distributed around 10 as well. What cause this is soccers with jersey number 10 are usually forward, so they are usually the important one in the team.

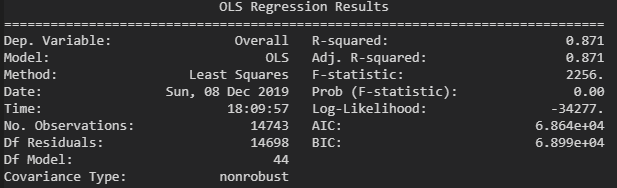


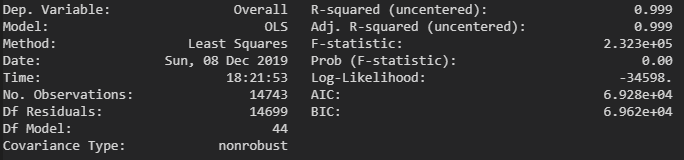
Then we build linear regression model with output wage and full variables adding BMI and removing ID, name, flag, photo, club and joined. We dropped insignificant variables which p-value is large in primary model, and now we got linear regression model below. International reputation has largest coefficient. R square is 0.629, which means 62.9% output could be explained by model.

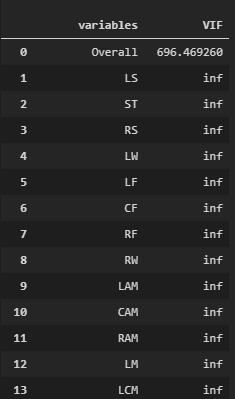




3.2: Is player’s overall rating calculated from a formula of abilities rating? Besides, what factors influences the overall rating most? (Linear model to predict the overall rating)  
For the dataset, we have 34 variables of abilities rating and an overall rating. We think there should be a relationship between them. Because FIFA is a video game, the overall rating would increase when gamer increase some of the abilities rating by training the FIFA player.

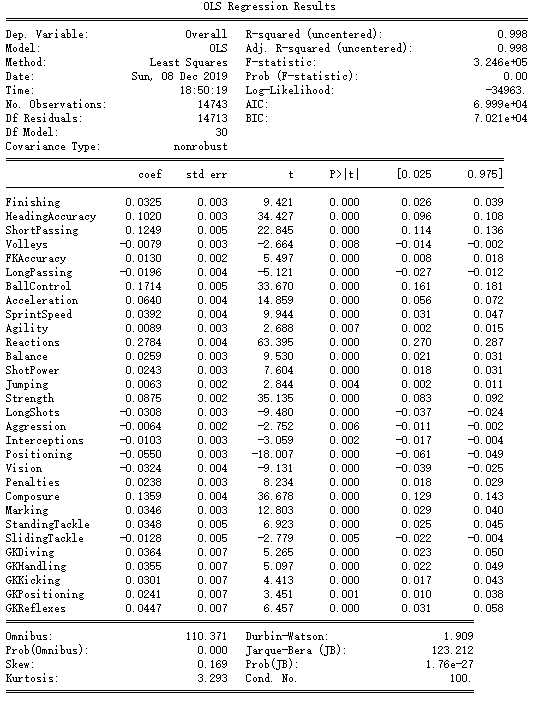
I first try the full linear model with the positions and abilities rating. The R square was 0.871. It is not like what we expect.

Then, we realized that for overall rating, the constant would not make sense because the rank of players would not be different if everyone increase 5 points on the overall rating. Next, we remove the constant from the model. The R square increase to 0.999.

Next we found that there are strong multicollinearity problems in this model. The VIF variables are very large for some of the variables. Especially for positions, the VIF are infinite for all positions. 



For this condition, we realize that positions have very strong correlations with abilities rating. In general, what abilities is good or bad will really influence the position of a player. For example, if a player is not good at running which means the speed rating is very low. Then this player would get a very low rating in positions that require running.

So we decided to remove variables of positions from the linear model.The condition number and VIF improves a lot after removing positions variables. After removing variables that p-value is lower than 5%, the final model is shown below.

All variables are significant enough and R square is 0.998 which is quite good.

From this model we can see, the BallControl and Reactions influence most to overall rating. This result implies that the soccer is all about ball control and reactions, which fits the common sense.

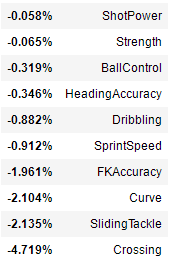
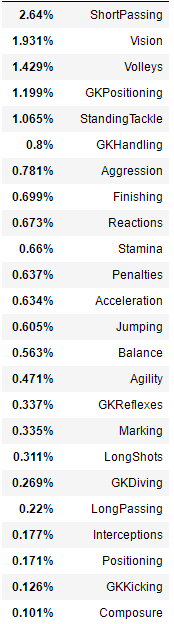
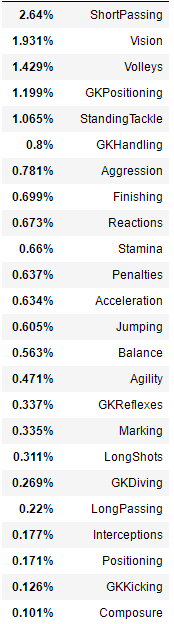
3.3 Is the preferred foot of a FIFA player was affected by his abilities rating? (Logistic regression to predict the preferred foot)

In our dataset, the preferred foot is a categorical variable. It only has two values right and left, which could be predicted by using a logistic regression.

We selected variables of 34 abilities to be a new dataframe. Then split it into train and test dataframe. The test dataset contains ¼ values of the total 14743 values, which contains 3686 values.

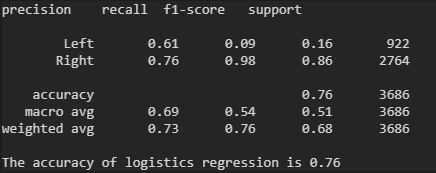
Afterwards, we build the logistic regression to predict the preferred foot. Note that 0 means the player’s preferred foot is left and 1 means that the preferred foot is right.

Here is the ranking of the variables contributes to right foot.



For example, if a player is good at ShortPassing and Vision. Then he is likely to be a right foot player. On the opposite, if a player is good at SlidingTackle and Crossing. Then he is likely to be a left foot player.

Here is the matrix. The accuracy is 76% for this logistic regression.



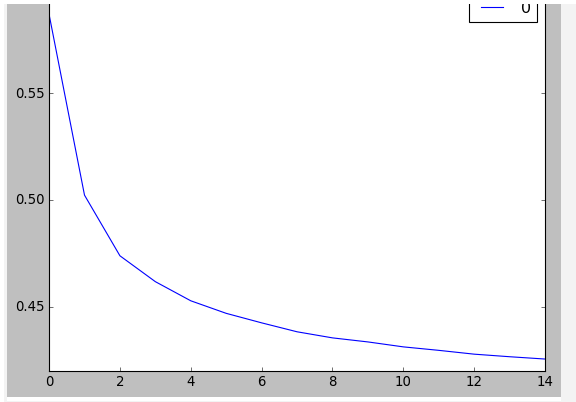
3.4 Is the preferred foot(Left/Right) used by football players affected by their scores? (KNN to predict preferred foot of a FIFA player)

Because in our dataset, the preferred is a categorical variable. The variable has two types: left and right. So I first turned the variable to be numeric. I created a new column called ‘P\_foot’. In this column, there are two types of value, 0 and 1. 0 means the player’s preferred foot is left and 1 means that the preferred foot is right. Then I set this column to be independent variable which is called yfifa1.

The 34 numeric variables including(rate from 1 to 100) crossing, finishing, Curve, Dribbling, etc were referred to be dependent variables. Though all the variables are rated from 1 to 100, I still scaled them to ensure they are normalized. Because normalizing the data is necessary in KNN method or the Euclidean distance from independent variable to the dependent variables will be similar and I cannot choose the proper K value.

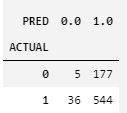
I split the test and train dataset, the test dataset contains ¼ values of the total 14743 values, which contains 3686 values.

I decided to use RMSE value as an index to decide which K value is best. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Because the expected value of the preferred foot will be a float from 0 to 1(for example, some expected values will be 0.6, 0.7). However, the actual values are 0 or 1. If I use the accuracy to decide the K value, there will be some mistakes. I run the RMSE value of K from 2 to 16, here is a visualization of my result:



We can see that the RMSE values continues to be lower but does not change a lot when the K be larger. I also took a try of K=100 as normally the best K value is the root of the observed numbers and I found out that the RMSE did not be lower too much. Finally, I decided to use K value as 15 to construct the KNN model. I used cross-validation method to find the accuracy of the model as I created 5 validation sets, the accuracy of the model is 74.7%.

I created a cross table to see the results more directly. I round the results of 0.6 to be 1 and 0.4 to be 0. I used the K=15 to build the KNN predict model from X\_test and X-train. Then I used the predict model to predict the y\_train values and compared them to y\_test. Here is the table:



Limitations:

We can see from the table that the model predict 544 values right when the preferred foot is right and with 177 values of preferred foot is right, our model predict wrongly.

With the KNN model, we cannot see how variables correlated to the independent variable like linear model that we can see estimated value of each variable. By this way, we cannot learn anything from the training dataset.

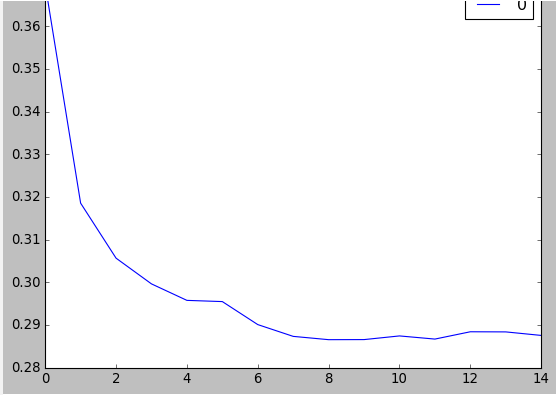
3.5 Is the international reputation of the international players can be predicted from the players’ scores? (KNN test)

The independent variable is international-reputation. This variable rates on a scale of 1-5 ( reputation 5 is the best, but there are only few players have reputation of 5). The variable type is ‘int’, so I do not need to change anything but just set it as ‘y’ variable.

The ‘x’ variables contains 34 numeric variables including crossing, finishing, accuracy, etc. These variables are all on a scale from 1-100. As I did in the last problem, I also scaled them to ensure they are normalized. The reason of it is same as above.

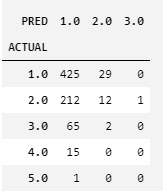
I split the dataset into test and train where the test dataset set contains ¼ of the total 14743 values.

Then I used RMSE to find the right K value. The reason for using RMSE as an index stated in the last problem. I run the K value from 2 to 16 as well, here is the visualization of the result:



We can see from the graph, the RMSE values continues to be lower with the larger of K value. However, the RMSE value does not change a lot when K- value reaches 9. With the larger of K, the RMSE is even larger. I also checked the RMSE value when K is 100 which is the square root of the observed datasets numbers and the RMSE value does not change much. So I decided to use K value =9 to construct the KNN model.

I tested the accuracy of the model when K=9, the result is 87.8%. I also created a cross table to visualize the result as I round the predicted value to be integer and compared them to the y\_test data so I can see more information. Below is the table.



Limitations:

We can see from the table that when the reputation is 1, our model predicted 425 correctly. However, our model mostly predict the reputation to be 1 when 212 values are 2, 65 values are 3, 15 values are 4 and 1 value is 5. Moreover, our model does not predict any value of 4 or 5 though there are few players’ international reputation is 4 or 5. The high accuracy of the model is mainly caused by the reason that most of the data located at the reputation of 1.

1. Conclusion

1. Wage could be predicted by the linear regression model with R square 0.630, and international reputation has the largest coefficient, so it contributes to wage most.

2. In general, overall rating of a player is most likely calculated from a formula of abilities and it is influenced most by ball control and reactions.

3. The accuracy of predicting preferred foot from logistics regression is 76% while from KNN is 74.7%. Logistic regression performs a little bit better than KNN.

4. The accuracy of predicting international reputation from knn method is 87.8% which means the reputation can be predicted by abilities with high accuracy.