Final Report of Predicting FIFA 2019 Players' Wages

Group 2A: Yihuan Huang, Taoyu Jiang, Huimin Zhang, Mengyu Zhang

March 2020

1 Abstract

With more than two thousand years of history, soccer is considered almost the most influential sports all over the world. And FIFA, the Federation Internationale de Football Association, was founded to oversee international competition among nations. This research targets on one of the most famous sports game FIFA, and we are going to predict soccer players' wages based on several attributes, including nationalities, clubs, positions and so on. Methods we are going to use include literature review, exploratory analysis on training and testing dataset to perform variable selection, validation of models through plotting as well as the discussion about limitations and future trend.

Here are the results of our final model:

Team Name	: Kaggle Lec 2 A
R-Square for training data	: 0.9985
R-Square for testing data	: 0.95156
Final Rank	: 30
Number of predictors	: 12 (7 categorical and 5 numerical predictors)
Number of βs	: 17
BIC of the final MLR model	: 19

2 Introduction

Our goal was to use the best valid model we created based on training data to predict FIFA soccer players' wages from testing data. The data size for us is 12745*80 provided by FIFA Website. 80 predictors include players' overall, special, and potential scores, reflecting player's strength and ability. Demographic information about players are also included in the dataset, such as players' nationality, position and club, as well as several categorical variables indicating players' soccer habits, such as real.face, right or left legs. The rest of the variables are scores about players' attributes and their scores when they are on each position. These are all variables related to player's strength. On the first sight, we noticed that variables may be correlated because several variables convey same information of the player. For example, we assume that the combination of players' attribute scores can also indicate players' strength and ability, serving the same function as players' overall or special scores. This leads us to spend the majority of our time on selecting variables and creating new variables to make the best use of these variables.

Before performing variable selection, our first step is data exploration and NA cleaning. When examining the response variable *WageNew*, we found that it ranges from 6 to 650305 with mean value 11433. This large range implies that there are some outliers that we need to examine more closely. The players with extreme low wages are those who have NA in the variable *Club*. This finding suggests that NAs in the variable *Club* are useful and *Club* is a good candidate to predict the response variable *WageNew* because those players who do not have club have mean wage significantly lower than others who are in clubs. Therefore, instead of deleting NAs in *Club*, we created a new category "None' and replace the NA values by "None". In other words, we are treating those people with no clubs as a new category "None". To make the model more valid and not influenced by extreme values, we removed the players with wages less than 200. Such a low annual wage does not make sense and is not reasonable. Therefore the final data size for us to generate the model was 80*12707.

This report will be divided into 4 parts as follows: first we are going to talk about the methodology we used to build models, including recoding variables, transformation, and adding weights. After that, we will talk about the validation of our model, including calculating VIFs, and plotting mmps graphs. Followed by is the section presenting our results, involving the summary and ANOVA table, and the discussion section where we summarize the methods we have tried. Last but not least, we will talk about limitations of our model and arrive at conclusions.

3 Methodology

In this section, we will discuss our motivation of creating new predictors after exploratory analysis and how we built our model.

3.1 Exploratory Analysis

Before actually creating new predictors, we did some exploratory analysis in order to select predictors that are highly correlated with the response variable *WageNew*.

First of all, we used correlation matrix to examine the correlation between numerical predictors and the response variable. Results are shown in Figure 1. From the figure, we can see that variables *Overall, Potential, Special* are highly correlated with the response variable, indicating that they can be good variable candidates for our model. However, we should also noticed that the correlation between variables *Potential, Special*, and *Overall* are also high. This serves as a warning that we may need to create new variables in order to make better use of these three variables, otherwise multicollinearity issues may occur. This also corresponds to our assumption at the beginning of our exploration introduced in the Introduction section.

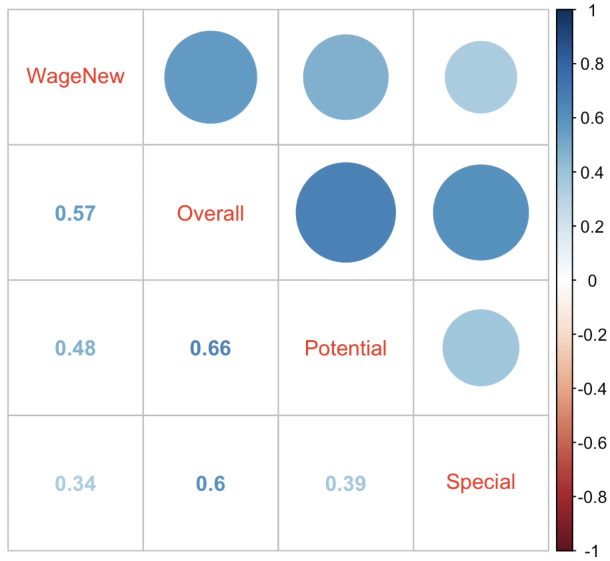


Figure 1: Correlation plot between Numerical Variables

3.1.1 Age and Potential

After numerous trials, we figured out that the variable *Overall* can mostly reflect a player's strength and has the strongest predicting ability among these three candidates (relationship can be seen in the Figure 6). Therefore, we would like to include the variable *Overall* in the model.

For the other two variables (*Potential,* and *Special*), we did further analysis into the dataset (mainly the rest of the variables) and literature review in order to extract more information from these two variables. We found that the *Potential score* is especially informative for younger players. This also makes sense because younger players can serve for a team for longer time than older players and their future developments are more important. Therefore, we created a new categorical variable *young* ("yes" when the player is younger or equal to 20 and "no" if older) and an associated interaction term *young:Potential*. In our final model, we only included the interaction term.

3.1.2 Special and Position

Besides the variable *Potential*, we also investigated further about the variable *Special* through literature review. *Special* reflects each player's strength and can be treated as an overall combination of all the other attributes (such as, GKDiving, Reaction) in the remaining dataset. However, we do not have any information with regards to the calculation of the variable *Special*. This is what makes *Special* variable hard to analyze and one of the reasons why we do not want to directly include the variable *Special* into the model. Moreover, it is worth to notice that players of different positions have different skill sets. For example, from our literature review, we found that only *GKDiving, GKHandling, GKPositioning, GKReflexes*, and *GKKicking* are relevant for Goal keepers(GK), while the rest of the attributes in the dataset are related to Attacking (LAM, CAM, RAM). An overall *Special* score may be too vague to be put in the model. Therefore, we would like to create our own special score by adding relevant attributes according to the requirements of position.

With regards to position, we specifically look at Goal Keepers and Attacking (LAM, CAM, RAM). From Figures 2 and 3, we can see that people on Attacking *Position* have higher *log(WageNew)* than non-Attacking, and Goalkeepers have lower *log(WageNew)*. Moreover, these two type of positions require two different types of skill sets as mentioned before. These reasons justify our choices of creating two new categorical variables *Goalkeeper* and *attacking*, and two new numerical variables

GKSkill and *AttackSkill* to reflect their corresponding position strength. After several model fittings and the analytical results, we decided to include *Goalkeeper* and *AttackSkill*:*attacking* in our model.

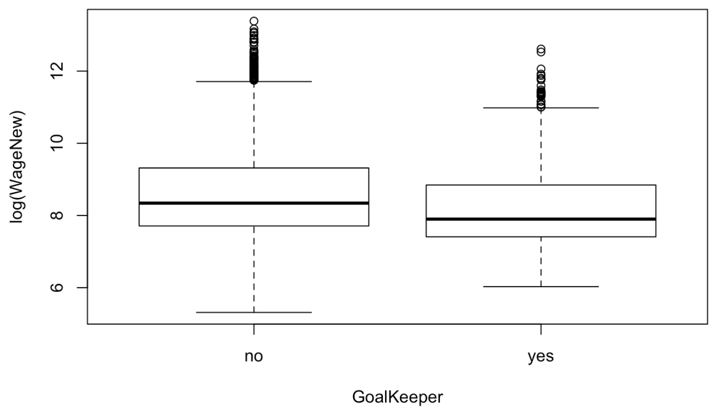


Figure 2: Boxplot between WageNew and Position, specifically about Goal Keepers and None Goal Keepers.

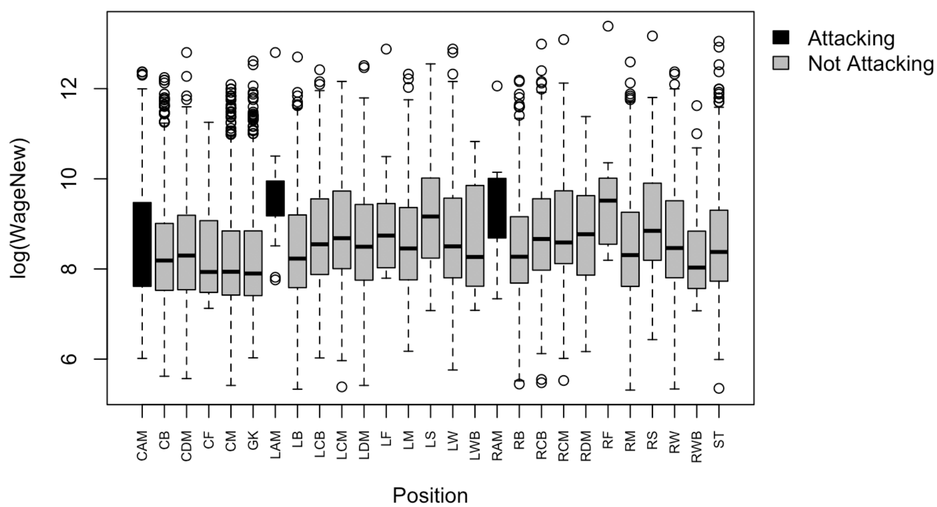


Figure 3: Boxplot between WageNew and Position. Black represents Attacking and Grey represents non Attacking.

3.1.3 Club and Nationality

After that, we focused on categorical variables and created box-plots to examine their predicting ability. There are two categorical variables that have lots of categories, *Nationality*, and *Club*. Through analysis, we figured out that people in different clubs and with different nationalities have different mean *WageNew*. However, directly adding these variables into models significantly increases the complexity of the model and may overfit the data. Therefore, we planned to recode categories of *Nationality* and *Club* by grouping categories and decreasing dimensions. For *Nationality*, we decided to divide nationality into English or not because after several model fits, we figured out that we always underestimate the wage for English people.

Club is a really important variable, to which we put our most effort. First of all, we recoded the *Club* variable by grouping clubs that have similar wage together, based on Figure 4. Colorful lines create 8 intervals of WageNew, according to which we created 8 categories of club and created a new variable *ClubLevels* with 8 levels according to the plot. We added lines in a way that clubs in one interval have similar *WageNew*. Besides that, we also added a new numerical variable *ClubRank* into

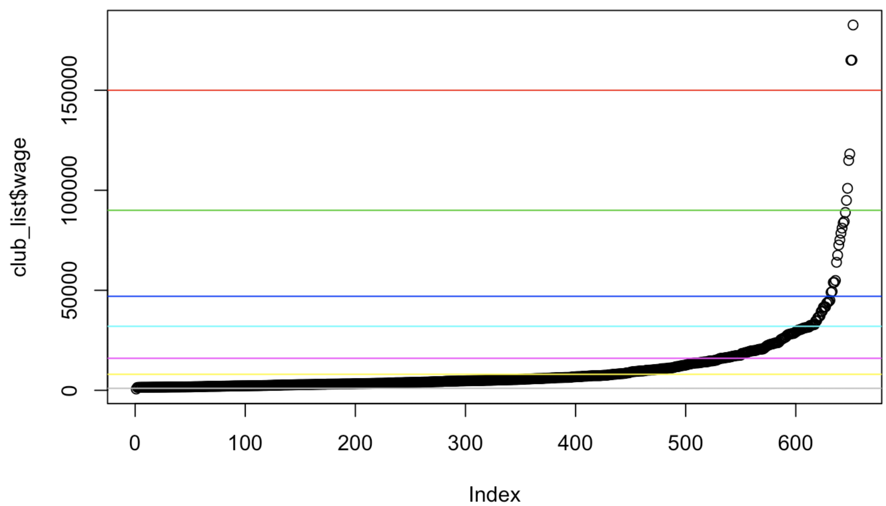


Figure 4: Relationship between Club and WageNew. Club are ordered according to their mean Club wage.

the model. *ClubRank* represents each club's rank from 1 to 652 (from lowest mean wage to highest). This variable provides additional and more accurate information about Club Rank. Moreover, this variable is numericaland we only need one β for this variable, which ensures the simplicity of the model. After we ran models with the above two variables *ClubLevels* and *ClubRank*, we found that our prediction always underestimate those clubs that have above average wage. Therefore, we include two more variables *Over90* and *Between70and90* to adjust for this underestimation. *Over90* represents clubs that have mean wage in the top 90 percentile, and *Between70and90* represents clubs that have mean wage in the top 70 to 90 percentile. To sum up, we created the variable *nationalityEng* to extract information from the variable *Nationality*, and *ClubLevels*, *ClubRank*, *Over90*, and *Between70and90* to recode the variable *Club*.

3.2 Real.Face

Through our exploratory analysis, we found that the categorical variable *Real.Face* helps explains much variance of the response variable *WageNew*. This can be verified in Figure 5, from which we can see that there is a huge gap between people falling in Real.Face category and those who don't. This is also verified in the ANOVA table in Figure 13 because *Real.Face* as a two-level categorical variable explained a decent amount of variance.

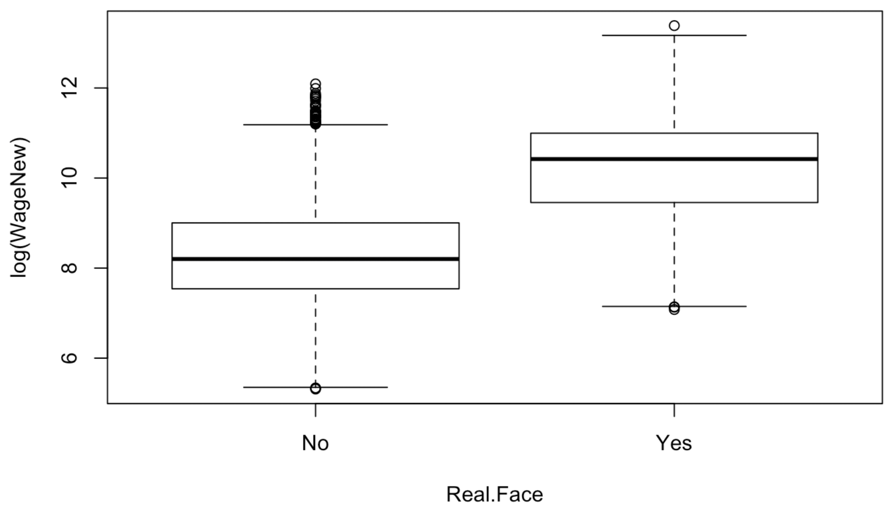


Figure 5: Relationship between log(WageNew) and Real.Face

3.3 Summary of Variables

In summary, we used 8 variables and 2 interaction terms. We used 17 betas (including the intercept term). The matrix plot is provided in Figure 6. We can conclude that all of our variables are good candidates for predicting the response variable *WageNew*.

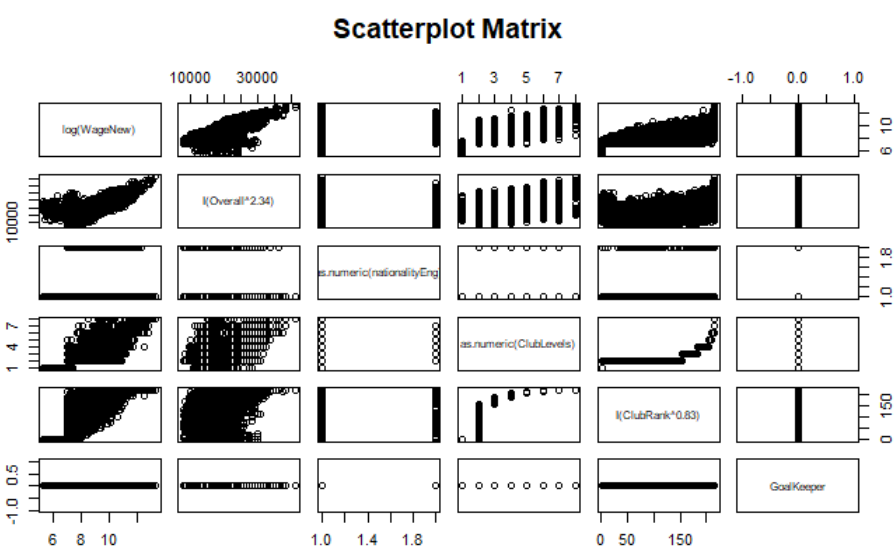


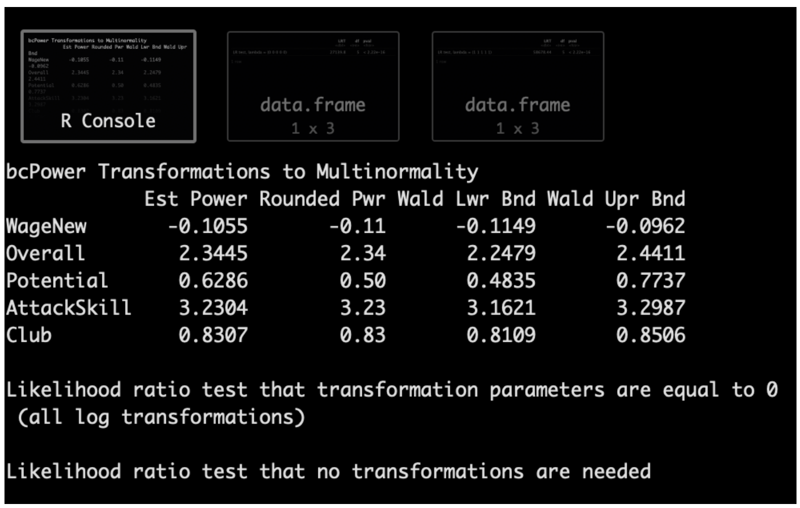
Figure 6: Scatterplot Matrix of all Predictors

3.4 Transformation

After selecting variables, we performed power transformation, and the suggested λs are presented in Figure 7. We used the λs in Table 1 according to our judgmental call. We do not use the inverseResponsePlot. Though the inverse transformation will decrease SSE of the model, it decreases the interpretability of the model. After discussion, we balanced the trade-off by not using the inverse transformation.

Variables	Lambda
1. WageNew	0
2. Overall	2.5
3. Potential	0.5
4. AttackSkill	3.25
5. Club	1

Table 1: Power Transformation




```
bcPower Transformations to Multinormality
            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
WageNew       -0.1055       -0.11      -0.1149      -0.0962
Overall        2.3445        2.34       2.2479       2.4411
Potential      0.6286        0.50       0.4835       0.7737
AttackSkill    3.2304        3.23       3.1621       3.2987
Club           0.8307        0.83       0.8109       0.8506

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

Figure 7: Results of Power Transformation

3.5 Weighted Least Squares

After variable selection and transformation of the model, we plotted the diagnostics plots in order to examine the validity of the model. Among the diagnostic plots, we were aware of the violation of non-constant variance (in Figure 8). The non-constant variance issue is especially severe when the fitted values is small. One of the methods taught in class to deal with this issue is to use Weighted Least Squares. After numerous trials, we decided to use weights equal to $Overall^{10}$. Adding weights allows the errors covariance matrix to be different from an identity matrix. In other words, WLS can also solve the non-constant variance issue by adding different weights on observations and treating data points differently. WLS is generally more flexible and can fit the model better. This helps fix the non-constant variance violation, which can be verified in the third plot of Figure 11. However, we do not have a complete explanation why 10 works better because we have not dived deep into the weighted least squares in the multiple linear regression case in class.

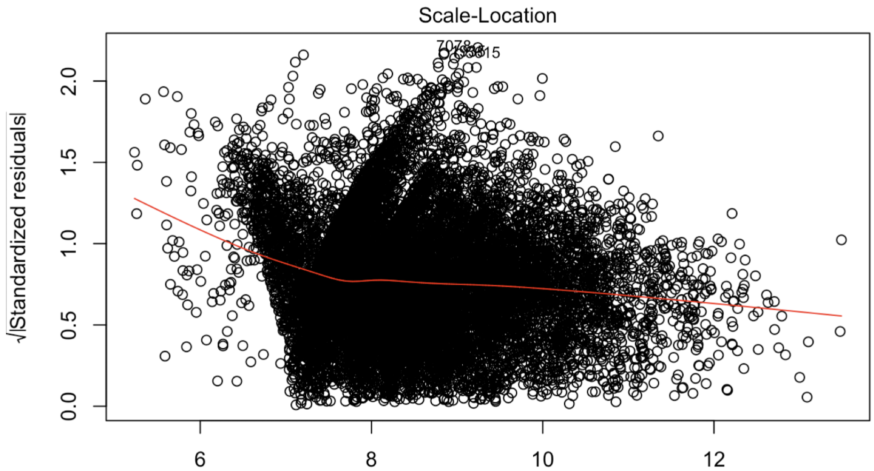


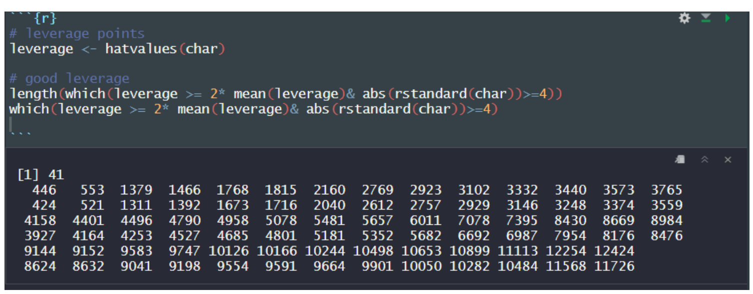
Figure 8: Diagnostic Plots of the Unweighted Model

4 Validity of the model

In this section, we will provide evidence to validate our model, including listing leverage points, presenting diagnostic plots and marginal model plots, as well as showing the variance inflation factor (VIF).

4.1 Leverage Points

After the full model was built, we searched for all leverage points which made large influence to the regression model. The good leverage points are shown in Figure 10, and bad leverage points are shown in Figure 9. We deleted all bad leverage points from our model and fitted it again as our final model. There are many good leverage points, indicating that our model is good because it is heavily influenced by good leverage points.

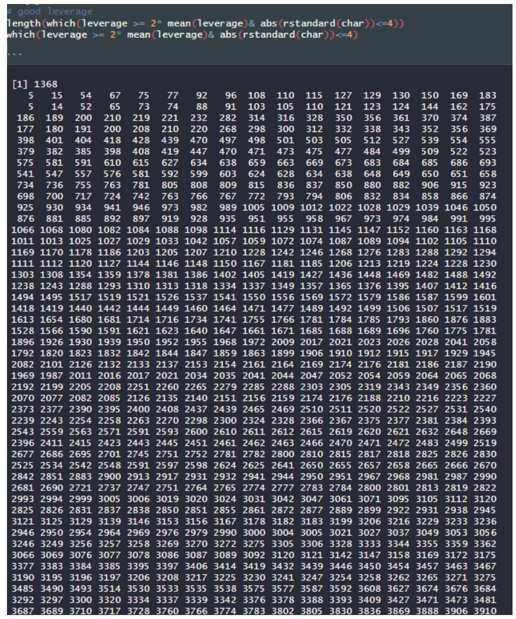


```
```{r}
# leverage points
leverage <- hatvalues(char)

# good leverage
length(which(leverage >= 2* mean(leverage)& abs(rstandard(char))>=4))
which(leverage >= 2* mean(leverage)& abs(rstandard(char))>=4)
```
```

```
[1] 41
  446   553  1379  1466  1768  1815  2160  2769  2923  3102  3332  3440  3573  3765
  424   521  1311  1392  1673  1716  2040  2612  2757  2929  3146  3248  3374  3559
 4158  4401  4496  4790  4958  5078  5481  5657  6011  7078  7395  8430  8669  8984
 3927  4164  4253  4527  4685  4801  5181  5352  5682  6692  6987  7954  8176  8476
 9144  9152  9583  9747 10126 10166 10244 10498 10653 10899 11113 12254 12424
 8624  8632  9041  9198  9554  9591  9664  9901 10050 10282 10484 11568 11726
```

Figure 9: All bad leverage points



```
# good leverage
length(which(leverage >= 2* mean(leverage)& abs(rstandard(char))<=4))
which(leverage >= 2* mean(leverage)& abs(rstandard(char))<=4)
...

[1] 1368
    5   15   54   67   75   77   92   96  108  110  115  127  129  130  150  169  183
    5   14   52   65   73   74   88   91  103  105  110  121  123  124  144  162  175
  186  189  200  210  219  221  232  282  314  316  328  350  356  361  370  374  387
  177  180  191  200  208  210  220  268  298  300  312  332  338  343  352  356  369
  398  401  404  418  428  439  470  497  498  501  503  505  512  527  539  554  555
  379  382  385  398  408  419  447  470  471  473  475  477  484  499  509  522  523
  575  581  591  610  615  627  634  638  659  663  669  673  683  684  685  686  693
  541  547  557  576  581  592  599  603  624  628  634  638  648  649  650  651  658
  734  736  755  763  781  805  808  809  815  836  837  850  880  882  906  915  923
  698  700  717  724  742  763  766  767  772  793  794  806  832  834  858  866  874
  925  930  934  941  946  973  982  989 1005 1009 1012 1022 1028 1029 1039 1046 1050
  876  881  885  892  897  919  928  935  951  955  958  967  973  974  984  991  995
 1066 1068 1080 1082 1084 1088 1098 1114 1116 1129 1131 1145 1147 1152 1160 1163 1168
 1011 1013 1025 1027 1029 1033 1042 1057 1059 1072 1074 1087 1089 1094 1102 1105 1110
 1169 1170 1178 1186 1203 1205 1207 1210 1228 1242 1246 1268 1276 1283 1288 1292 1294
 1111 1112 1120 1127 1144 1146 1148 1150 1167 1181 1185 1206 1213 1219 1224 1228 1230
 1303 1308 1354 1359 1378 1381 1386 1402 1405 1419 1427 1436 1448 1469 1482 1488 1492
 1238 1243 1288 1293 1310 1313 1318 1334 1337 1349 1357 1365 1376 1395 1407 1412 1416
 1494 1495 1517 1519 1521 1526 1537 1541 1550 1556 1569 1572 1579 1586 1587 1599 1601
 1418 1419 1440 1442 1444 1449 1460 1464 1471 1477 1489 1492 1499 1506 1507 1517 1519
 1613 1654 1680 1681 1714 1716 1734 1741 1755 1766 1781 1784 1785 1793 1860 1876 1883
 1528 1566 1590 1591 1621 1623 1640 1647 1661 1671 1685 1688 1689 1696 1760 1775 1781
 1896 1926 1930 1939 1950 1952 1955 1968 1972 2009 2017 2021 2023 2026 2028 2041 2058
 1792 1820 1823 1832 1842 1844 1847 1859 1863 1899 1906 1910 1912 1915 1917 1929 1945
 2082 2101 2126 2132 2133 2137 2153 2154 2161 2164 2169 2174 2176 2181 2186 2187 2190
 1969 1987 2011 2016 2017 2021 2034 2035 2041 2044 2047 2052 2054 2059 2064 2065 2068
 2192 2199 2205 2208 2251 2260 2265 2279 2285 2288 2303 2305 2319 2343 2349 2356 2360
 2070 2077 2082 2085 2126 2135 2140 2151 2156 2159 2174 2176 2188 2210 2216 2223 2227
 2373 2377 2390 2395 2400 2408 2437 2439 2465 2469 2510 2511 2520 2522 2527 2531 2540
 2239 2243 2254 2258 2263 2270 2298 2300 2324 2328 2366 2367 2375 2377 2381 2384 2393
 2543 2559 2563 2571 2591 2593 2600 2610 2611 2612 2615 2619 2620 2621 2632 2648 2669
 2396 2411 2415 2423 2443 2445 2451 2461 2462 2463 2466 2470 2471 2472 2483 2499 2519
 2677 2686 2695 2701 2745 2751 2752 2781 2782 2800 2810 2815 2817 2818 2825 2826 2830
 2525 2534 2542 2548 2591 2597 2598 2624 2625 2641 2650 2655 2657 2658 2665 2666 2670
 2842 2851 2883 2900 2913 2917 2931 2932 2941 2944 2950 2951 2967 2968 2981 2987 2990
 2681 2690 2721 2737 2747 2751 2764 2765 2774 2777 2783 2784 2800 2801 2813 2819 2822
 2993 2994 2999 3005 3006 3019 3020 3024 3031 3042 3047 3061 3071 3095 3105 3112 3120
 2825 2826 2831 2837 2838 2850 2851 2855 2861 2872 2877 2889 2899 2922 2931 2938 2945
 3121 3125 3129 3139 3146 3153 3156 3167 3178 3182 3183 3199 3206 3216 3229 3233 3236
 2946 2950 2954 2964 2969 2976 2979 2990 3000 3004 3005 3021 3027 3037 3049 3053 3056
 3246 3249 3256 3257 3258 3269 3270 3272 3275 3305 3306 3328 3333 3344 3355 3359 3362
 3066 3069 3076 3077 3078 3086 3087 3089 3092 3120 3121 3142 3147 3158 3169 3172 3175
 3377 3383 3384 3385 3395 3397 3406 3414 3419 3432 3439 3446 3450 3454 3457 3463 3467
 3190 3195 3196 3197 3206 3208 3217 3225 3230 3241 3247 3254 3258 3262 3265 3271 3275
 3485 3490 3493 3514 3530 3533 3535 3538 3575 3577 3587 3592 3608 3627 3674 3676 3684
 3292 3297 3300 3320 3334 3337 3339 3342 3376 3378 3388 3393 3409 3427 3471 3473 3481
 3687 3689 3710 3717 3728 3760 3766 3774 3783 3802 3805 3830 3836 3869 3888 3906 3910
```

Figure 10: Some good leverage points, 1368 good leverage points as total

4.2 Diagnostic Plots

Six diagnostics plots are show in Figure 11. Upon careful examination, we can see from the third plot that after standardization, the error is constant and randomly scattered around 0. Most of the residuals follow the normal distribution and not violating our normality assumption. There are no bad leverage points according to the calculation of cook's distance.

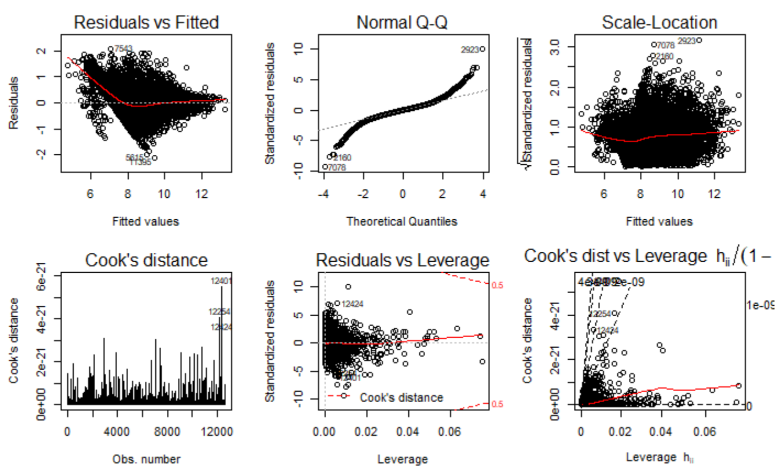


Figure 11: Diagnostic Plots of the Final Model

4.3 Variance inflation factor

Table 2 shows the VIF scores for this model. We can see that all variables have VIF scores smaller than 5. This indicates that our model does not have multicollinearity issues.

Variables	Lambda
1. I($Overall^{2.5}$)	3.874460
2. I($Potential^{0.5}$)	3.499512
3. I($AttackSkill^{3.25}$)	1.388974
4. I(ClubRank)	1.600336

Table 2: VIF of our model

4.4 Marginal Model Plotting

Figure 12 shows the MMPS plots for our model. We can see that the blue line and the red line are almost the same, except some minor deviation when the fitted values are small. This implies that our model fits the data well.

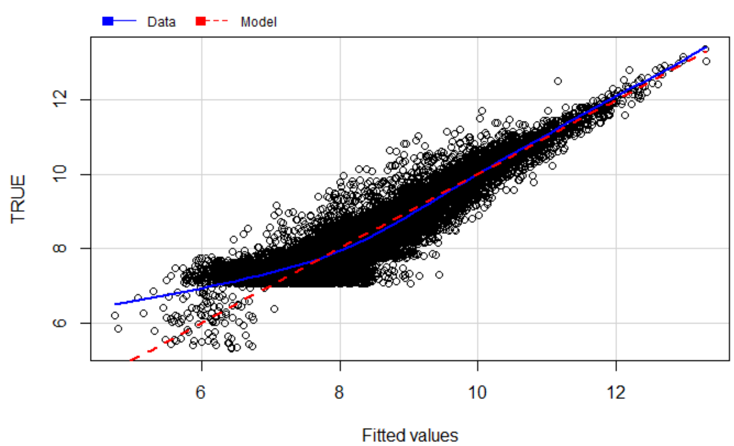


Figure 12: mmp Plots of the Final Model

5 Results

Our final model using the Box-Cox method transforms both the response and the predictor variables.

$$\begin{aligned} log(WageNew) = 0 + Overall^{2.5} + ClubLevels + Real.Face \\ + young : Potential^{0.5} + nationalityEng + attack : AttackSkill^{3.25} \\ + ClubRank + Goalkeeper + Over90 + Between70and90, \\ weights = Overall^{10} \end{aligned}$$

There are 7 categorical predictors in our model, which are *attack, young, Real.Face, nationalityEng, Goalkeeper, Over90*, and *Between70and90*. There are 5 numeric predictors in our model, which are *Overall, Potential, AttackSkill, ClubRank*, and *ClubLevels*. In total, we used 17 βs (including the intercept).

We used nature log transformation on the response variable *WageNew* and we transformed *Overall* to the power of 2.5. We also add 2 interaction terms age:$Potential^{0.5}$ and attack:$AttackSkill^{3.25}$.

The R-squared of our model in R is 0.9985, Residual standard error is 6038000000 on 12642 degrees of freedom (after adding weights). The summary results and ANOVA table are shown in Figures 14 and 13. All variables are significant and explain a decent amount of variance in the response variable *WageNew*. Moreover, upon careful examination, we can see that none of the variables have weird coefficients and all make sense.

```
Analysis of Variance Table

Response: log(WageNew)
                            Df     Sum Sq    Mean Sq    F value    Pr(>F)
I(Overall^2.34)              1 2.9567e+24 2.9567e+24 8.1111e+06 < 2.2e-16 ***
ClubLevels                   8 3.9279e+22 4.9099e+21 1.3469e+04 < 2.2e-16 ***
Real.Face                    1 1.1600e+20 1.1600e+20 3.1821e+02 < 2.2e-16 ***
nationalityEng               1 8.1034e+19 8.1034e+19 2.2230e+02 < 2.2e-16 ***
I(ClubRank^0.83)             1 3.2353e+21 3.2353e+21 8.8753e+03 < 2.2e-16 ***
GoalKeeper                   1 1.9200e+20 1.9200e+20 5.2671e+02 < 2.2e-16 ***
Over90                       1 5.1062e+19 5.1062e+19 1.4008e+02 < 2.2e-16 ***
Between70and90               1 8.9623e+18 8.9623e+18 2.4586e+01 7.198e-07 ***
young:I(Potential^0.5)       2 3.8755e+20 1.9377e+20 5.3158e+02 < 2.2e-16 ***
attack:I(AttackSkill^3.23)   2 4.2206e+19 2.1103e+19 5.7892e+01 < 2.2e-16 ***
Residuals                12642 4.6083e+21 3.6452e+17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 13: ANOVA table

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
I(Overall^2.34)                      1.162e-04  1.966e-06  59.115  < 2e-16 ***
ClubLevels1                          6.208e+00  1.660e-01  37.406  < 2e-16 ***
ClubLevels2                          7.360e+00  1.615e-01  45.578  < 2e-16 ***
ClubLevels3                          7.268e+00  1.637e-01  44.400  < 2e-16 ***
ClubLevels4                          7.371e+00  1.653e-01  44.595  < 2e-16 ***
ClubLevels5                          7.482e+00  1.667e-01  44.893  < 2e-16 ***
ClubLevels6                          7.676e+00  1.673e-01  45.876  < 2e-16 ***
ClubLevels7                          7.973e+00  1.687e-01  47.258  < 2e-16 ***
ClubLevels8                          8.142e+00  1.682e-01  48.423  < 2e-16 ***
Real.FaceYes                         9.026e-02  1.014e-02   8.902  < 2e-16 ***
nationalityEngYes                    1.633e-01  1.427e-02  11.447  < 2e-16 ***
I(ClubRank^0.83)                     1.129e-02  1.210e-04  93.371  < 2e-16 ***
GoalKeeperyes                       -1.409e-01  1.699e-02  -8.290  < 2e-16 ***
Over90yes                            2.940e-01  2.554e-02  11.510  < 2e-16 ***
Between70and90yes                    1.056e-01  2.026e-02   5.213 1.88e-07 ***
youngold:I(Potential^0.5)           -2.878e-01  2.253e-02 -12.772  < 2e-16 ***
youngyoung:I(Potential^0.5)         -3.103e-01  2.144e-02 -14.469  < 2e-16 ***
attackattack:I(AttackSkill^3.23)     7.841e-12  7.965e-13   9.844  < 2e-16 ***
attacknot_attack:I(AttackSkill^3.23) 6.807e-12  6.713e-13  10.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 603800000 on 12642 degrees of freedom
  (46 observations deleted due to missingness)
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9985
F-statistic: 4.332e+05 on 19 and 12642 DF,  p-value: < 2.2e-16
```

Figure 14: Summary Report

6 Discussion

Here, we will talk about the methods we have investigated before our final method and give some explanations why these methods do not work.

With regards to the *Special* score, we were thinking of recreating a new variable to more accurately reflect each player's strength in their own position. We noticed that each player has their score in each position. (In the data sets, there are 25 positions, such as "LS", the Left Striker; "RWB", the Right Wing Back, and so on. Each position has a numeric value representing the score of the particular player.) For each player, we tried to find the maximum score of all these position scores and defined the max score as our new *Special* variable because we assume that each player is placed in the position in which he has the highest score. This approach is reasonable and saves our time in coding. However, the new *Special* variable is not significant when fitting the model. We inferred the reason why this variable does not work is that our newly-created variable *Special* is highly correlated with the variable *Overall*, leading to multicollinearity issues. This also motivates us to calculate *AttackSkill* mentioned above.

Additionally, we tried to check if the variable *Joined* (Time when the player joined the club) has a linear relationship with Wage and can be a good candidate. Time may matter because the longer the player stayed in a club, the more experienced he may be. Therefore, we made a hypothesis that the response variable *WageNew* is positively correlated with the variable *Joined*. Cleaning the variable *Joined* takes a long time because the joined date have different input styles and we have to transform the data to be predictable and reasonable. We divided the joined date to before 2016 and after 2016. We did not need to know the specific date, we only need the year. However, we found out that the joined year is a poor predictor because the variance explained by the joined year may have explained by Age. Therefore, the need to add the *Joined* variable is limited.

7 Limitations and Conclusions

We are using *RStudio* to help analyze all the data we have. The analyzing tool we used can tell us when what should consider when taking action, but it cannot tell what action to take. As mentioned before, although we have already choosing the predictors cautiously after rigorous discussion, it is still to make the promise that all the rest variables that we did not use in the model are statistically meaningless to appear in the prediction model. Due to a large number of the origin variables as well as a lacking of the real life knowledge in *FIFA* Rule and football playing skills, the predictors we took are acceptable but may not be the best combination over the total number of 79 variables. This might be a weakness of our final model. More reference materials about football are required to make a more accurate and precise selection. Moreover, though we have made the best use of our current statistical knowledge, the knowledge is still not enough and limited to make the fittest model based on the dataset.

Nevertheless, our model makes a valid model with good approaching of the overall prediction. The model satisfied all the requirements which are necessary to a mature linear regression. By applying the the data, we successfully predict most of the player's wage with a acceptable tolerance. Combining the all the predictors together with the several specific interaction terms to get the best fitted model with highest efficiency.

References

[1] Blake, A. (2019) FIFA 19 Career Mode guide: Negotiate contracts, successfully scout, and choose the best team. *MSN Sport*. http://www.msn.com/en-gb/sport/football/fifa-19-career-mode-guide—negotiate-contracts-successfully-scout-and-buy-the-best-players/ar-BBNHtNa.

[2] FIFA. *The FIFA/Coca-Cola World Ranking.* http://www.fifa.com/fifa-world-ranking/.

[3] Licata, A. (2019) The 12 Highest-Paid Soccer Players in the World. *Business Insider*. http://www.businessinsider.com/highest-paid-soccer-players-footballers-world-2019-6.