FIFA ‘19 FINAL PROJECT WRITE UP

A Term Report by

Cady Stringer, James Kistner, Justo Garcia, and Miranda Heyer

Chapman University

Orange, CA

Argyros School of Business and Economics

Submitted in partial fulfillment of the requirements for the course MGSC 310.

Statistical Models in Business Analytics

Course Instructor:

Dr. Jonathan Hersh

December 2019

1. Problem

We built regression models to predict a FIFA soccer player’s wage based on player attributes. The business application of this would be that team management could accurately predict what a player is worth before negotiating a contract and properly allocate their market cap on total player salary expenditure to different players depending on the value the player can bring to the team. We used the FIFA 2019 video game dataset. These values are based on real-life assessments of player skill and demographic information, but because the data has been scaled for the video game it has lost some of its integrity. This analysis is still valuable because this dataset is the most comprehensive FIFA data available on all players and teams around the world.

1. Summary Statistics

Europe and Africa were the most popular continents, with 7707 and 3108 clubs respectively. The mean wage is 10,000 Euros per week and the average age was 25 years old. The average weight was 166.2 pounds and the average height was 71.31 inches. Right footed players were more common by a margin of 7304 players. A medium/medium work rate was most common by far. See appendix for the output of all summary statistics.

1. Data Cleaning

We began with data cleaning and first addressed that we had over 200 factor levels for the countries column. We downloaded a list of countries and their corresponding continents and did a right join between our dataset to create a continent column to use instead. Next we removed variables that were not compatible with regression models or were only applicable to the FIFA video game. We removed 42 attributes, including: X, ID, Name, Nationality, Value, Release Clause, Photo, Jersey Number, Joined, Flag, Club Logo, Special, Body Type, Real Face, Loaned From, Contract Valid Until, and all 25 of the position skill levels besides the players preferred positions. Then, we stripped additional characters like the Euro symbol, lbs, and K from the numeric variables wage, value, and weight. We created a height to inches function and changed the height values written as 5’6 to 66 inches. We factor lumped our clubs into the top 10, which became the top 28 because there were ties.

1. Data Exploration

Our cleaning left us with 13,910 rows. We selected a 70/30 testing training split and began data exploration. After reviewing a series of plots investigating relationships between wage and other variables, the most interesting were wage and overall skill rating, wage and age, and wage density by club. Figure 1 in the appendix displays player age plotted against wage and illustrates that there is a “prime age” for players to maximize wage, between 25 and 32. The two pink outliers at the top of the plot are the wages of FIFA superstars Lionel Messi and Cristiano Ronaldo. Figure 2 shows wage against overall player skill rating and shows little to no increase in pay as skill rating increases until a skill rating of about 80, when the graph becomes exponential and wage grows quickly for even a small increase in skill. It is important to note that it is very difficult to improve your skills at this level, and players are compensated accordingly. Figure 3 reveals that the majority of clubs are paying a lot of players very little, and it is much rarer to have high wages.

1. Modeling

We built a linear regression model, lasso, post lasso estimator, elastic net, random forest, and a neural network. The linear regression predicted wage used all other variables. The lasso model also predicted wage in terms of everything else and used 10 folds. The lambda that minimized error chose 96 variables, and lambda at 1 standard error picked 8. Initially we ran the random forest with 1000 trees then cross validated and found that error does not decrease beyond 100 trees, so we changed Ntree to be 100. We used a modified dataset that did not include the Club variable for our random forest model because it had more than 53 categories and wasn’t compatible. The post lasso estimator was a linear model using the 4 lasso-selected variables international reputation, overall, potential, and club. For our elastic net, we created an alpha grid with values between 0 and 1 in 0.25 increments, cross-validated to select our optimal alpha of 0.75 and used lambda.1se. For the neural network, we read in the same training and testing data from the R file into a python document (we realize that we could’ve just made it in R). We then normalized the data so it was on a common scale before it was inputted into the network. We made a basic neural network with some basic dense layers and a dropout layer. We used MSE as a performance metric, but there are better metrics we should’ve used. We trained the network for 75 epochs with a batch size of 107 because 107 times 75 is the same size as our data.

1. Model Performance

Running diagnostics for our models revealed interesting patterns and performances. Figure 4 displays a bar chart of the R2 values by model, Figure 5 shows RMSE by model, and Figure 6 shows MAE. Lasso was our worst-performing model with an R2 of 0.633, RMSE of 14.442, and MAE of 6.317. The post-lasso estimator came next with an R2 of 0.649, RMSE of 14.036, and MAE of 6.269. The elastic net performed only slightly better than the lasso variations, with an R2 of 0.658, RMSE of 13.859, and MAE of 6.333. Interestingly, a linear regression out-performed lasso, post lasso, and an elastic net with an R2 of 0.659, RMSE of 13.839, and MAE of 6.395. It was the third best model, just after our neural network and random forest. It should be noted that the error of the linear model was heteroskedastic. Both the neural network and random forest greatly improved predictive power and raised R2 by about 0.1. As we expected, random forest was our best model with an R2 of 0.788, RMSE of 11.365, and MAE of 4.558. Neural network came in a close second with an R2 of 0.746, RMSE of 11.875, and MAE of 5.106.

To test the business application of our random forest model, we identified the top 10 most overvalued and 10 most undervalued players by calculating our residuals and finding the top and bottom 10, then matched the name of the player with their corresponding residual. The top 10 overvalued players are R. Marchizza, F. Garcia, J. Foyth, L. Comas, O. Al Soma, D. Lemos, M. van der Werff, Bebe, Pedraza, and J. Stryger Larsen. These players are being paid too much, and it may be beneficial for management to pay these players less and reevaluate how they’re spending their market cap, or find cheaper players to replace them, since they aren’t bringing as much value to the team as they’re being paid for. The top 10 undervalued players are M. Wostry, A. Limbombe, V. Tsygankov, F. Aguilar, Kevin Rodrigues, Luan, G. Burdisso, M. Herrera, M. Gulde, and Paulo Vitor. These players are a good deal, and it would be beneficial for management to hold on to these players and try to keep their salaries low because they add a lot of value to the team.

1. Results and Analysis

We learned that a few variables matter a lot, and besides those variables everything else doesn’t really have an impact. The most important variables across all models included overall skill rating, potential rating, club, international reputation, reactions, and position. The post-lasso estimator revealed that a linear model using only the 4 lasso-selected variables, international reputation, overall, potential, and club performed almost as well as a linear model using all the variables. This establishes that those 4 variables are able to explain almost all the variation in the data. These variables were repeatedly chosen by other models as some of the most important, which further confirms their predictive power.

Our analysis identified a few ways that management can maximize their use of the market cap. The first method is by hiring someone highly skilled in a low-earning position, because certain positions are disproportionately paid much more than others. Management can underpay the skilled player in a low-paying position and add a lot of value to the team. As we saw in Figure 1, wage increases exponentially after the skill rating of 80, so teams could save on salaries by hiring a player with a skill level of 75-80 rather than one with 80-85, because after ratings of 80 the cost of the player increases exponentially and the return on investment is much worse. For example, an increase in skill rating from 60 to 80 adds a lot of skill to the team, with little to no increase in pay. Management could create a powerful team of players with skill ratings around 80 and save a lot on player salary. In addition, teams with lower international reputations can hire highly skilled players just under a rating of 80, because these teams generally pay much smaller salaries than those with better reputations of 4 or 5.

1. Conclusion

In sum, we found that a random forest model using 100 trees was able to most accurately predict FIFA player wage, and that the most important player attributes were overall skill rating, potential skill rating, international reputation, and club. We found through our data exploration that players are paid most between 25 and 32, and that pay has little to no increase between an overall skill rating of 0 and 80 and increases exponentially after that. Management can hire undervalued players, players with a skill level of 80, and highly skilled players in low-paying positions to maximize team skill using their market cap. Teams can identify their overvalued players using the residuals from a random forest predictive model and either pay them less or hire different players, and find their undervalued players and hold on to them.

Appendix

**SUMMARY STATISTICS**

Continent Age Overall Potential Club

Africa :1090 Min. :16.00 Min. :47.00 Min. :48.00 Other :13270

Asia :1189 1st Qu.:21.00 1st Qu.:63.00 1st Qu.:68.00 Atlético Madrid: 33

Europe :7707 Median :25.00 Median :67.00 Median :72.00 CD Leganés : 33

North America: 547 Mean :25.18 Mean :66.98 Mean :71.93 FC Barcelona : 33

Oceania : 269 3rd Qu.:28.00 3rd Qu.:71.00 3rd Qu.:76.00 Frosinone : 33

South America:3108 Max. :45.00 Max. :94.00 Max. :95.00 RC Celta : 33

(Other) : 475

Wage Preferred.Foot International.Reputation Weak.Foot Skill.Moves

Min. : 1.00 : 0 Min. :1.000 Min. :1.00 Min. :1.000

1st Qu.: 1.00 Left : 3303 1st Qu.:1.000 1st Qu.:3.00 1st Qu.:2.000

Median : 3.00 Right:10607 Median :1.000 Median :3.00 Median :2.000

Mean : 10.56 Mean :1.129 Mean :2.96 Mean :2.398

3rd Qu.: 10.00 3rd Qu.:1.000 3rd Qu.:3.00 3rd Qu.:3.000

Max. :565.00 Max. :5.000 Max. :5.00 Max. :5.000

Work.Rate Position Height Weight Crossing

Medium/ Medium:7645 ST :1628 Min. :61.00 Min. :110.0 Min. : 5.00

High/ Medium :2383 GK :1528 1st Qu.:69.00 1st Qu.:154.0 1st Qu.:39.00

Medium/ High :1235 CB :1358 Median :71.00 Median :165.0 Median :55.00

Medium/ Low : 696 LB :1019 Mean :71.31 Mean :166.2 Mean :50.46

High/ High : 690 CM : 971 3rd Qu.:73.00 3rd Qu.:176.0 3rd Qu.:64.00

High/ Low : 578 RB : 963 Max. :81.00 Max. :236.0 Max. :93.00

(Other) : 683 (Other):6443

Finishing

Min. : 2.00

1st Qu.:30.00

Median :49.00

Mean :46.02

3rd Qu.:62.00

Max. :95.00

HeadingAccuracy ShortPassing Volleys Dribbling Curve FKAccuracy

Min. : 4.0 Min. :11.00 Min. : 4.00 Min. : 4.00 Min. : 6.00 Min. : 3.00

1st Qu.:45.0 1st Qu.:55.00 1st Qu.:31.00 1st Qu.:50.00 1st Qu.:35.00 1st Qu.:31.00

Median :56.0 Median :63.00 Median :45.00 Median :62.00 Median :49.00 Median :42.00

Mean :52.8 Mean :59.54 Mean :43.59 Mean :56.14 Mean :47.94 Mean :43.47

3rd Qu.:65.0 3rd Qu.:69.00 3rd Qu.:58.00 3rd Qu.:69.00 3rd Qu.:63.00 3rd Qu.:57.00

Max. :94.0 Max. :93.00 Max. :90.00 Max. :97.00 Max. :94.00 Max. :94.00

LongPassing BallControl Acceleration SprintSpeed Agility Reactions

Min. : 9.0 Min. : 5.00 Min. :12.00 Min. :12.00 Min. :15.00 Min. :30.0

1st Qu.:44.0 1st Qu.:55.00 1st Qu.:57.00 1st Qu.:57.00 1st Qu.:56.00 1st Qu.:57.0

Median :57.0 Median :64.00 Median :67.00 Median :68.00 Median :66.00 Median :63.0

Mean :53.4 Mean :59.25 Mean :64.77 Mean :64.86 Mean :63.67 Mean :62.6

3rd Qu.:65.0 3rd Qu.:70.00 3rd Qu.:75.00 3rd Qu.:75.00 3rd Qu.:74.00 3rd Qu.:69.0

Max. :93.0 Max. :96.00 Max. :97.00 Max. :96.00 Max. :96.00 Max. :96.0

Balance ShotPower Jumping Stamina Strength LongShots

Min. :16.00 Min. : 3.00 Min. :15.00 Min. :12.00 Min. :17.00 Min. : 3.00

1st Qu.:56.00 1st Qu.:47.00 1st Qu.:58.00 1st Qu.:56.00 1st Qu.:58.00 1st Qu.:34.00

Median :66.00 Median :60.00 Median :66.00 Median :67.00 Median :67.00 Median :52.00

Mean :63.96 Mean :56.37 Mean :65.21 Mean :63.36 Mean :65.56 Mean :47.88

3rd Qu.:74.00 3rd Qu.:69.00 3rd Qu.:73.00 3rd Qu.:74.00 3rd Qu.:74.00 3rd Qu.:63.00

Max. :96.00 Max. :95.00 Max. :95.00 Max. :96.00 Max. :95.00 Max. :94.00

Aggression Interceptions Positioning Vision Penalties Composure

Min. :11.0 Min. : 3.00 Min. : 2.00 Min. :10.00 Min. : 5.00 Min. : 3.00

1st Qu.:44.0 1st Qu.:26.00 1st Qu.:39.00 1st Qu.:44.00 1st Qu.:39.00 1st Qu.:52.00

Median :59.0 Median :53.00 Median :56.00 Median :55.00 Median :50.00 Median :60.00

Mean :56.1 Mean :47.12 Mean :50.48 Mean :53.88 Mean :49.04 Mean :59.34

3rd Qu.:70.0 3rd Qu.:65.00 3rd Qu.:65.00 3rd Qu.:65.00 3rd Qu.:61.00 3rd Qu.:67.00

Max. :94.0 Max. :92.00 Max. :95.00 Max. :94.00 Max. :92.00 Max. :96.00

Marking StandingTackle SlidingTackle GKDiving GKHandling GKKicking

Min. : 3.00 Min. : 6.00 Min. : 6.00 Min. : 1.00 Min. : 1.00 Min. : 1.00

1st Qu.:30.00 1st Qu.:27.00 1st Qu.:24.00 1st Qu.: 8.00 1st Qu.: 8.00 1st Qu.: 8.00

Median :53.00 Median :55.00 Median :53.00 Median :11.00 Median :11.00 Median :11.00

Mean :47.68 Mean :48.05 Mean :45.99 Mean :16.57 Mean :16.34 Mean :16.15

3rd Qu.:64.00 3rd Qu.:67.00 3rd Qu.:64.00 3rd Qu.:14.00 3rd Qu.:14.00 3rd Qu.:14.00

Max. :94.00 Max. :93.00 Max. :91.00 Max. :90.00 Max. :92.00 Max. :91.00

GKPositioning GKReflexes

Min. : 1.00 Min. : 1.00

1st Qu.: 8.00 1st Qu.: 8.00

Median :11.00 Median :11.00

Mean :16.34 Mean :16.66

3rd Qu.:14.00 3rd Qu.:14.00

Max. :90.00 Max. :94.00

**FIGURE 1**

A screenshot of a cell phone

Description automatically generated

**FIGURE 2**

A close up of a map

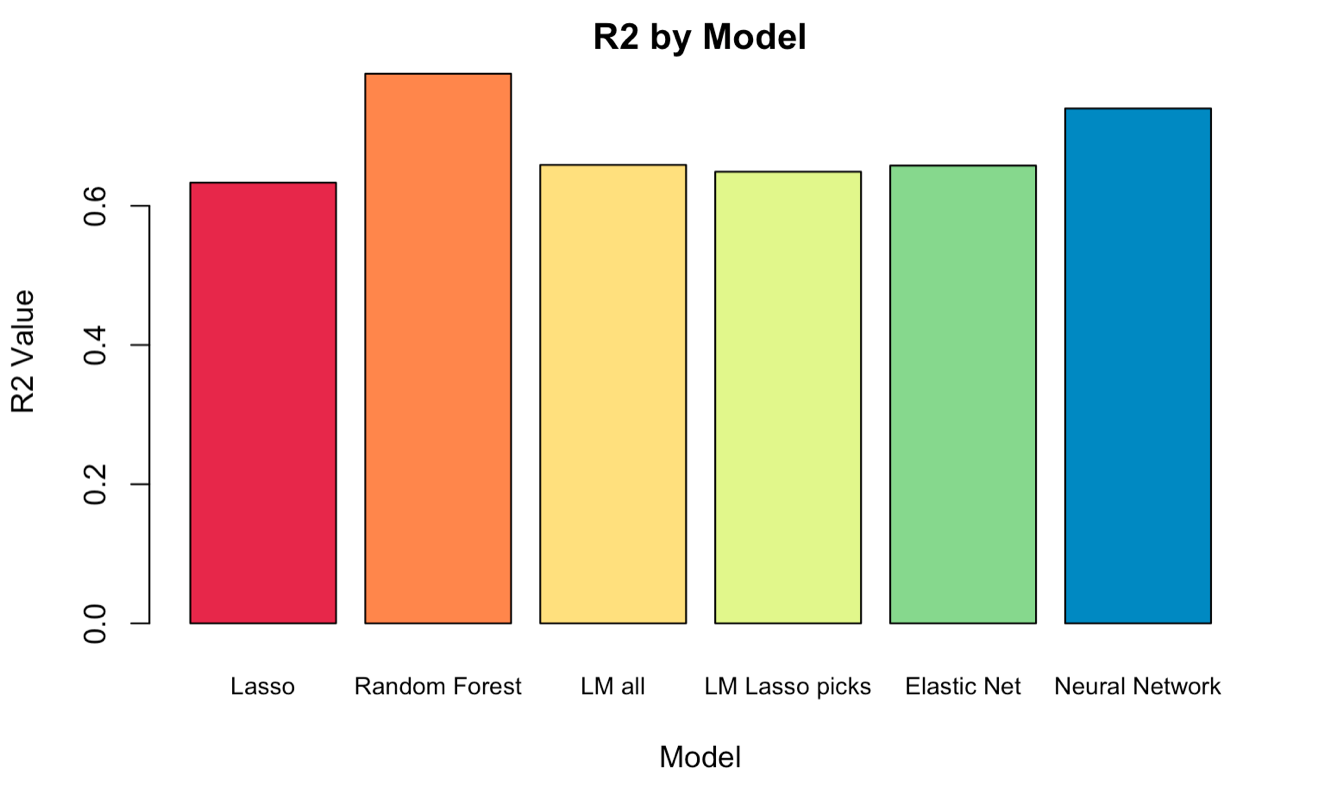
Description automatically generated

**FIGURE 3**

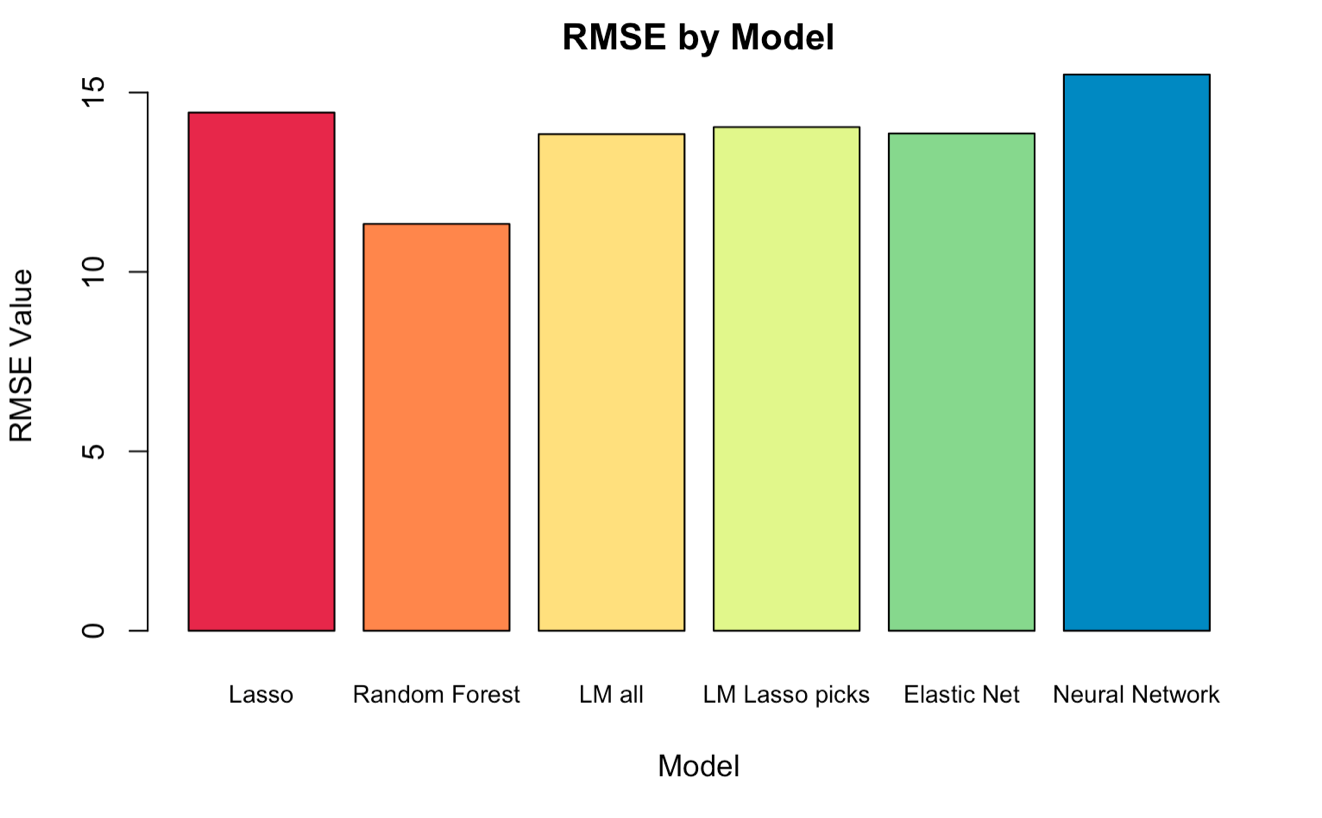
A screenshot of a cell phone

Description automatically generated

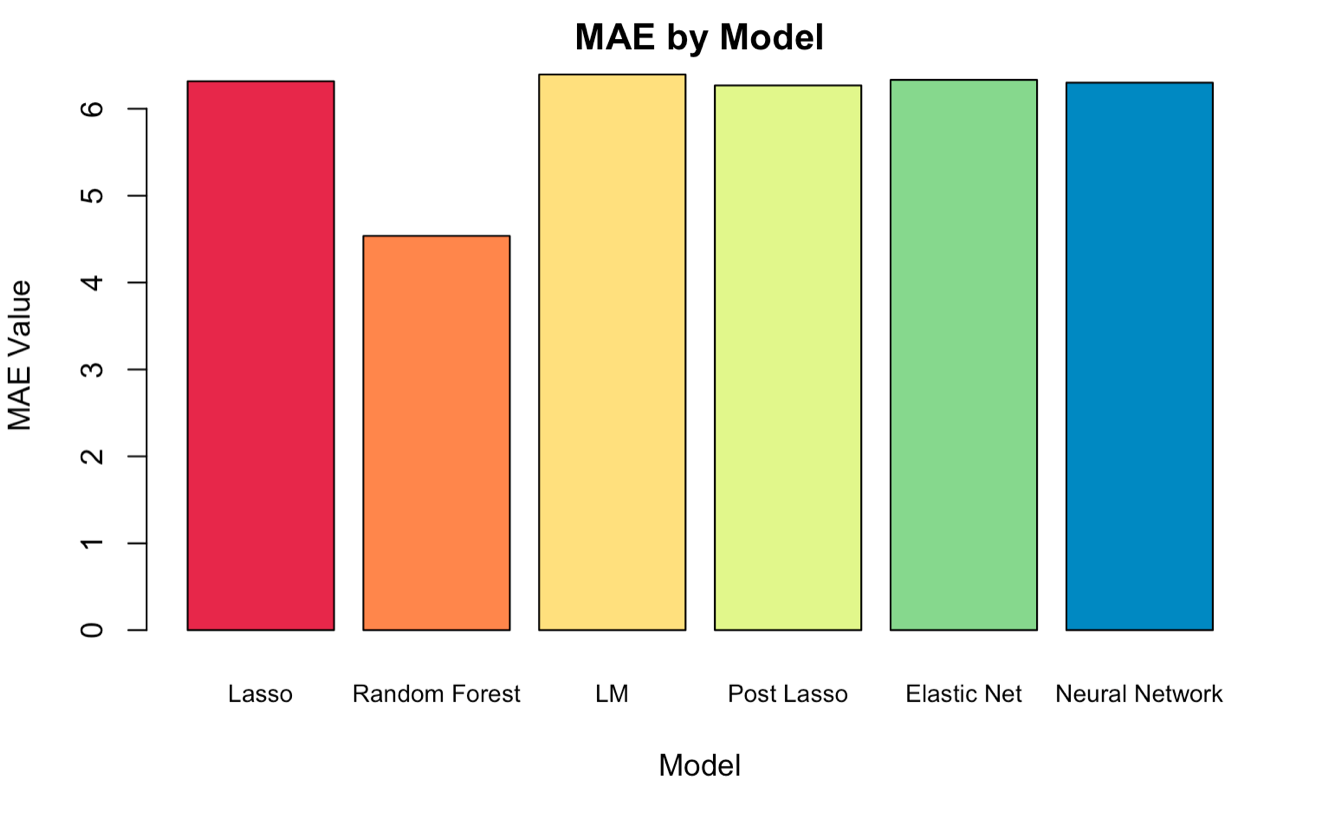
**FIGURE 4**



**FIGURE 5**

****

**FIGURE 6**

****