Final Project – Statistics and Machine Learning ACA Chris Gochis

The world of professional sports has always provided us a lot of data. One of the interesting things about how sports have evolved is not on the field, but off. Data is becoming more and more important in the way that sports clubs make decisions. These decisions, good or bad, could cost you a game, a player, or even an entire season. Knowing the data available, and using it to help make these important decisions will continue to grow in importance.

Professional soccer is the largest sport in the world. With its massive amount of fans comes ravenous followings both for their city and for their country as a whole, depending on what tournament may be at play. When dealing with a sport of this size, money becomes increasingly important. How well your team performs usually translates into larger amount of money incoming for your organization. With that pressure, making the right calls for your team is crucial. This is where using our data and some statistical models should be able to shine some light into areas that may be worth exploiting, hopefully to score some goals for the team.

The data that I have found is actually based on the FIFA 2019 video game. It was collected from Kaggle, and the creator of the dataset cites the information coming from a scraping of the website, sofifa.com. This website compiled the game-used statistics into a large, user-friendly interface, so fans could research their favorite players, or build their own favorite fantasy teams. Now, video game statistics may not seem so important to the real world, but FIFA is not your average video game. It is built using real world player statistics. This provides the developers the groundwork for the player models, things like height, weight, speed, accuracy, etc. The FIFA game also has a team manager simulation, which allows players to play with real life contract data, such as player value and salaries. One could get lost for hours, or months, trying to formulate the best attack plan for your made up squad.

What I found rather interesting about the idea that this data is that it is based on real life people. Could you use this video game data, to help your real life soccer club make player decisions? Could this data be used to make some scouting calls as to which players are worth signing, will be the useful, or even predict the overall quality of players that may be new to the professional playing field. It may even be useful to see how certain traits impact overall player value and performance. Can this information help team managers better predict financial needs? I feel that this analysis would be useful for the analytics teams employed by the clubs. It may help them answer questions about player scouting or even gain insight on opponents.

This data set provides us with a lot of useful information. There are just over 18,000 players included with great detail provided for each. Basic player information includes data such as Name, Age, Nationality, Club, Preferred Foot, Height, and Weight. Each player is also assigned a large collection of rankings. These rankings usually range from 0, the worst, to 100, the best. The rankings cover the player’s abilities. We have information such as acceleration, accuracy, stamina, sprint speed, shot power, composure, ball control, penalty kicks. We even have information for goalies, such as handling, kicking, positioning and reflexes. Each player also receives rankings for their abilities in the on-field positions. Positions such as striker (ST), right wing (RW), center back (CB), left middle (LM). Along with position and skill ranking, the data also provides us with an overall player rating, as well as an international reputation, weak foot, skill moves, and work rate ranking. Each player is also provided with some financial information. These attributes include overall player value, yearly wage, and even a contract release clause amount.

I had to do some cleaning of the data before I was able to import the set into R for analysis. The changes made were in an effort to simplify the use of these attributes and reformat them into usable pieces of data. Originally the weight class contained an “lbs” after each integer, that was removed. I converted the height from feet and inches (F’In”) into just total inches. The body type category was divided into 3 factors, lean, normal, stocky, with the exception of 6 players, which were given their own names as a body type. I assume this was done because these players are considered to be the 6 best in the world, so the game developers made specific models for each. I assigned them an appropriate body type based on the 3 factors. There were a few attributes for image links, those were removed. The currency values were originally listed in the format like “$100.5M”. I converted these to the format 100500000, for example, for easier analysis. The position rankings were also provided with a “progress over season” addition. For example a player in the CM position would have a ranking of 85+2. This would mean that over the 2019 season they could have increased their ranking by 2. I viewed this as excess information used in the video game and felt it wouldn’t pertain well to the analysis so I removed the “+X” from each, and left each player with their season start ranking. The great thing about this dataset is that is has a lot of player stats for all different kinds of metrics. The one thing I wish it had was a few more categorical variables, ones with a smaller number of categories, for some more comparative analysis.

A quick overview of our player set shows us some descriptive information. Ages vary from 16-45, with an average of about 25 years old. The average weight is 166lbs and the top 6 nationalities are England, Germany, Spain, Argentina, France, and Brazil. Figure 1 shows the majority of players prefer their right foot to their left 77% to 23%. Figure 2 shows us that a majority, 59% of players fall into the normal body type category. 35% are lean, and 6% are stocky. This would make sense, soccer is an endurance game, stocky builds may not fair so well running such long distances (speaking from experience!). Analyzing the position break down in Figure 3 we see that most players are strikers (ST). I was surprised to see that followed by goal keepers (GK) and center backs (CB). The next statistic is interesting. Figure 4 shows the work rate of each player, showing how much effort they put into the attacking / defending portions of the game or field. I was surprised to see so many well rounded players. I would have expected to see more High/Low or Low/High, based on the fact that the most common positions are Striker and Goalkeeper.

**MODELS**

Before beginning our statistical analysis and running our models, I first need to take a sample of the dataset. Using roughly 18,000 players would be too many for our needs. I will use a sample of 1,000 players, selected randomly.

For our first set of models, I will build a series of T-Tests. The goal of these tests is to see if there are some statistical differences between the player’s preferred foot, this is the perfect binary variable to run these test. I would like to see if there is any differences made to the players Overall, International, and Penalty kicking ratings, based on what foot they prefer. The goal is to try and help scouts focus on players with a preferred foot, depending on the overall recruiting goal. Results are discussed below.

Next, ANOVA models will be helpful in running comparisons between attributes with multiple categories, such as Work Rate and Body Type. The goal I’m trying to achieve is to give the recruiters a good baseline to go on for some basic characteristics for the players. The Work Rate and Body Type fields provide us with some good categorical values that we can place against some larger rankings as well as physical attributes. I will be testing Overall Rank and International Reputation compared to Work Rate as well as Sprint speed, Agility, and Stamina compared to Body Types. I hope that this will give our scouts some good player profiles to focus on. Results are discussed below.

Next I would like to build a regression analysis. The goal of this analysis is to provide some of the team managers insights into drivers of overall player value. What makes a player worth more? This is important because if they develop a player well from a young age, they could use this as financial leveraging to gain advantages in further seasons with trades or sales etc. Player Value will be our dependent variable. An since my table has over 80 attributes per player, I have isolated a handful that I feel have the potential to be important metrics based on a well-rounded player skillset. I have chosen Age, Field Kick Accuracy, Heading Accuracy, Ball Control, Aggression, Vision, Composure, and Sliding Tackle. These attributes will serve as my independent variables. I have chosen these variables because they provide good numeric metrics and they should provide us with some clear relationships.

The regression analysis is going to need analyzing to maximize its effectiveness. To help with this process, I want to first try and eliminate any sub-optimal attributes that I may have chosen. I will complete this using the stepwise regression tool in R. I feel that this is a more accurate way of analyzing variables instead of looking to see which were significant or not, and removing them by hand. Next I will run a multicollinearity test, this will show me if any of the attributes I have chosen are too closely related and therefore become redundant. I will then create a diagnostic report and asses my residuals. This should provide me with the further steps needed to create a powerful and useful regression analysis. The results of this process are detailed below.

Team management may want to know if there is a statistical relationship between both body type, and work rate. This could be important knowing what type of player, by build, would be apt to working on different sides of the field. Since both of these values are categorical, I believe that using a Chi-Squared analysis would be best. If we can find that the null hypothesis is true, that these player attributes are independent, then it is one less thing for our scouts to look for. To perform this analysis these items need to be set in a table. Luckily with the power of R we can build this inside the function. I did run into an issue when trying to run this too, one of my Work Rate values had zero instances matched across all Body Types. I had to remove this column from my table to run this test.

Following the Chi-Squared, I would like to run some tests on the Skill Moves rating. However, I would like to cut it into two categories, high and low. Currently the metric is on a scale of 1-5 but I would like to convert that. I would like to create a new attribute containing 1-2 as a LOW rating and 3-5 as a HIGH rating. The reason why I am cutting after 2 is because of how the data is spread out. If I made 1-3 the low category, the sample selection would be heavily weighted in that direction. Originally I wanted to run this test on the International Reputation, but I found that since there are so many young players in this dataset, there was a very large section of players still in the level 1 reputation. I also felt that this was not a good indicator because so much of reputation happens over the career length, not just based on the skill statistics. Using the Skill Moves rating, I believe it will be easier to predict based on the skills metrics in our dataset. Using the rating 2 as a dividing point, I get two balanced groups to test.

Since we are now working with machine learning methods, I need to be able to train the models and analyze the accuracy. I will take our sample set, and create both a train and test subset based on the 75/25% set division. With our train and test set, we can build our logistic model. I am going to use the individual field stills variables section of our dataset to start with. The section that includes Crossing, Accuracy, Slide Tackles, etc, will give us a good starting point and we can cut down that attributes from there based on how significant they are. The goal of this overall test is to provide the management team a way to possibly predict what skills will potentially push a player into the HIGH Skill Move territory. This is important for teams to know, because players with higher skill moves have the ability to create more chances for their teams to score, possibly leading to more wins. The results are below.

Following the Logistic Regression, I would like to see if the previous insights could be improved upon or expanded using a decision tree or a random forest. Since my dataset does not contain much useful binary categories for prediction, I want to see if we can compare these model types to see what may be more useful for our scouting team. Using the logistic regression, when cleaning our model, it eliminated a lot of variables that seemed to be nonsignificant for how the model works. I would like to see if a decision tree or a random forest might take into account a few more field skill rates to possibly enhance our level of Skill Move LOW/HIGH prediction.

I would like to construct the decision tree and random forest using the original set of all the field skill attributes, the 34 variables we started with. I will analyze our models using a new train and test sample set from our starting sample set as before, and I will also use the same 75/25% train/test creation format as before. This way we can see which variables that each method finds to be useful. The overall goal is to try and create the broadest player profile for our scouts to use. I felt like making judgement calls for recruitment based only on a few attributes did not give that team the analytical power it needed to best predict the type of player they may be signing.

**RESULTS**

Here we will begin to discuss the results our models generated and how we reached the endpoint. I will walk through the steps I took to run the models, analyze the models for best use, discuss some limitations as well as interpreting results and making suggestions to our management teams.

I would like to run a few T-Tests to get started. The first set of test I would like to see how the dominant foot affects both overall player rating, and national reputation. Are lefties more often to be higher or lower? The T-Test produced a p-value of 0.3862, therefor there is no significant difference between preferred foot when compared to Overall player ranking. So what about International Ranking? The T-Test Provides us with a p-value of 0.288, so there is no significant difference between feet for International Ranking either. Finally I would like to run a T-Test to see if a preferred foot make a difference in the players penalty kicking score. If a team is looking to find a player with a high penalty kick rating, maybe they should look at preferred foot. Running the T-Test we see that there is a significant different between the feet, the p-value is 0.048. One of the limitations of using T-Tests is that you don’t get to see directly which foot does better. However, we can see that the Left foot group mean is 50.20642 and the Right is 47.80563. This could offer us a clue that Left foot groups may actually be the stronger group for penalty kicks. This may be something recruiters should keep in mind when scouting.

Running the ANOVA models we see some interesting results. Figure 5 and 6 show when comparing the Overall Rating and International Reputation to Work Rate, we see there is a significance in both situations. It appears that the Work Rate does make a difference. The TukeyHSD multiple means comparison for these show that there is only a significant difference when comparing Medium/Medium and High/High for the international rating, however there were 9 pairings that showed a significant difference when looking at Overall ratings, see Figure 10. These results are showing that the High/High work rates for players are making more of a positive impact on their overall ratings. This was to be expected, but it’s good to know that finding players with an equally high work rate for both attacking and defending appear to score higher on both overall rating and international reputation.

Using ANOVAs to do some physical skills comparisons, I have found that body type is significant when Sprint Speed and Agility are concerned, but not significant in regards to stamina. Figures 7, 8, and 9 show this. We can see from the results of our TukeyHSD tests that having a Lean body type provides an advantage for both sprint speed and agility. Figure 11 shows this. This is helpful information for the scouts. If they are looking for players with increased levels of speed and agility, they can focus on players who fit the lean description. If they are looking for good stamina, body type is less significant.

Next, we build our regression model. As described above I would like to use Value as our dependent variable and Age, Field Kick Accuracy, Heading Accuracy, Ball Control, Aggression, Vision, Composure, and Sliding Tackle as our independent variables. I will build this model and run it through our stepwise regression tool. Upon first look at the summary table, Figure 12, we can see that there are some less than significant attributes, so let’s run the stepwise regression to sort them out.

The stepwise regression tool suggests a model containing the attributes, which we can see are essentially the same significant attributes as I saw in the original run, Composure, Age, Vision, and Ball Control. I will alter the original regression model and run it again. I was surprised to see that roughly half of these attributes were not significant factors into a players overall value.

This new model provides us with some more focused results. The vif() results are also very good. All values are way below 10. Our regression diagnostic is shown in Figure 13 and our summary table in Figure 14. Our R-squared value comes in around 0.228, so not the best, but not terrible. We can still draw some inferences from the coefficients. It actually looks like we are getting some good results from this regression model. Our Residuals line is fairly straight. It may have a very slight U but not enough to cause me much concern. I did attempt to add the squared values of our negative coefficients (Age and BallControl) in an attempt to address a possible issue with quadratic term. This did not change the shape of the residuals, only shifting the slight U across the X axis. This may just have to do with some coefficients being negative and others being positive. We can now begin to review some of our results. My expected results were to see them all have a positive relationship with Value, with the exception of age, at least for a while. As age increases it should increase with value, because skill maturity would increase, but overall an older the player gets, it will begin to bring their value down. We see this with a decrease of 155,000 per year. Seems harsh but, it’s business. What I was surprised to see was the negative coefficient attached to BallControl. This seems like such a crucial skill but according to this data, it causes a decrease in player Value, interesting. Overall these results followed my expectations and I believe that knowing some of these crucial field skills will help a players overall value, a scout or manager can make sure to look for skills like Vision and Composure in their players. It’s also possible that these results show us that off field factors play into value as well.

I think a crucial limitation of running a players value through a regression on this dataset is that it is not a complete picture. I believe there is a lot more that goes into a players overall value, even including metrics that are collected off field. Leadership skills and community engagement are some examples. Also this data does not include statistics such as goals scored over a season, or saves made, balls stolen, plays broken up. Those are also important metrics. However for the dataset that we do have, I feel like we’ve seen what we overall expected for the players value regression analysis. The better a player performs overall, the more they are worth.

Next we can run the Chi-Squared test. As mentioned above, I made a table using body type and work rate to see if there was a significant relationship. I had to remove the Low/Low work rate since it contained no instances of body type to compare with. What we found is listed in Figure 15. We can see that our p-value is 0.0201. Since this is below our 0.05 significance level we can reject the null hypothesis that says that body type and work rate are independent. We can see that there is evidence of a relationship between these two variables. Figure 16 shows our Pearson Residuals showing us where these values are attracted to each other. As we can see, players with LOW and MEDIUM attack ratings tend to be stocky, and lean players are most likely to have a work rate of HIGH/LOW. This is what I was expecting to see. I predicted that various body types would have an impact on a players performance on different ends of the field. Stocky players are better suited for defensive positions and lean players are more in line with the offensive side of the field.

I think the issue we run into with a test like this is it really only shows us if they are independent. However, you can see that in the table and graph that there is a decent amount of spread between work rates and body types. Seeing as a majority of players fall in to the normal body type and the medium/medium work rates, they somewhat dominate the relationship, but other than that you can see there is some spread. In all other work rate categories, the normal body type has more. We can expect then that player with normal body types are actually the best choice overall, they don’t suffer some of the limitations that the other body types have. They tend to match better with higher ratings on both sides of the field simultaneously. I would use this information to tell our scouts to focus on those normal body types, and to look outside there when a specific need is to be met or a special case presents itself.

As stated above, we are now moving into some predictive modeling, using machine learning. First I want to build and test a Logistic Regression model. We will be testing the individual skills attributes against our Skill Moves metric, grouped into HIGH and LOW categories. When I first ran the model, we can see that there are a number of these skills that do not have a significant relationship with our dependent variable. The initial results are show in Figure 17. As we can see, for this model, the significant variables are Dribbling, FKAccuracy, BallControl, and GKPositioning. We will rebuild the model with these attributes and run it again. The new results are shown in Figure 18. As we can see by our odds ratios, we can start to make some inferences. It looks like all 4 individual skills increase the odds for outcome increase in Skill Moves ranking. Higher field skills correlate well with the higher Skill Moves rating. This is expected and what we like to see! Let’s do some tests to make sure this result is clean. First I want to run an multicollinearity test. Since all levels are below five, I am not concerned about this issue having a negative impact. Now it is time to test our model accuracy before interpreting results. Figure 19 shows us our confusion matrix, and our accuracy percentage of 91%. We have a pretty accurate model! Let’s run this again on our test set. Figure 20 shows us that we have a solid model. The test set resulted with 90.8% accuracy! This is now a usable model for our scouts who want to know how a players Skill Move ranking will behave based on the 4 significant field skills we outlined above.

All of this shows that there is a good correlation between a players skill move ranking and how they perform on certain specific field tasks. I believe these takeaways can help our scouts identify specific on field skills if they are looking for players with an increased level of Skill Move rankings. I think one of the limitations that we see from this data set, is that for this model, 4 of the 34 field skills were a significant contributor to the overall Skill Move ranking. This may have to do with how the data is being collected, or it may have to do with some of the other implications we’ve mentioned before. There is more that goes into those scores than we can see in our database. However, it is interesting to see that we can make some predictions on a handful of field skills.

As stated in our model formulation, we are building a decision tree based on comparing the Skill Moves rating to the field skills variables. Creating this tree based on the standard complexity parameter, we see in Figure 21 that we only ended up using 2 variables, and since this is less than the logistic regression, I would like to adjust this CP to allow for more variables. Adjusting the CP to 0.0001 we see in Figure 22, that the model chose 5 variables. Which is one more than our logistic regression had. Figure 23 shows the complexity parameter graph, you can see that after the second variable, it does increase a bit showing where the tree has probably reached its peak effectiveness. Figure 24 shows a plot of the tree we created. We can now test the accuracy of our tree, similar to how we tested our logistical model. Figure 25 shows the confusion table as well as our accuracy percentage. 93.2%! We have actually improved slighty over our logistic regression. It appears that having more field skills to predict with helps us a bit.

To go a step further, I would also like to attempt to make a random forest with the original comparison set. I feel like this may even further help us make predictions, because of the sheer number of attributes we are using to predict. My prediction was correct, using random forest we have also slightly increased our prediction accuracy. Using the train/test analysis method from the last two models, Figure 26 shows that we have increased our accuracy to 94.8%. Figure 27 shows how our model has improved by the number of trees added. It looks like around 100 trees is where the improvement starts to level out, and model stops improving. As we can see on Figure 28, the two most important variables are Dribbling and BallControl, followed by Positioning and Agility and so on. What I find interesting about this model is how it compares to our logistic regression. In that first test, we saw that GKPositioning was one of the most significant variables, but in this forest, we see it’s the second to least important. It is interesting to see how there are differences between the three models we have tested for this prediction set.

After reviewing all of the results for these three models, I would recommend using the Random Forest since it gave us the most robust model. It had the highest level of accuracy and allowed for the most variables to contribute to its overall prediction. This will be helpful to the scouting teams, because they can use more data to make better predictions about the skill moves potential of the player they hope to sign.

Overall this dataset has provided us with a lot of interesting ways to explore important statistical information. I believe that these models have provided a lot of useful information and solid recommendations for our team management to make informed decisions. The regression analysis helped us make some financial predictions, and our other tests helped us lay the groundwork for proper player profiling. This should be helpful for our managers to make proper signing decisions based on the information that the scouts bring in from their observations on the field. One final recommendation I would make to this team is to take this dataset and expand on it. As I had mentioned a few times, even though we had some strong models, I don’t feel like they really encapsulated the entire picture for some of the overall ratings. Since so much of a players reputation and worth happen in terms of team relationship and goal productivity, it was hard to really provide a full picture. However, using the data that we have, I believe we have given them a lot to consider, and hopefully data will continue to level up the game of soccer.