How Well Does FIFA Ratings Contribute Towards Predicting Player's Actual Market Value

Bjorn Teo, Ignacio Arevalo, Prat Kothiyal

Introduction

Soccer is the most popular sport in the world with an estimated 3.5 billion fans and an annual global revenue of about 28 billion USD. Soccer clubs around the world play in domestic and continental competitions where revenue is generated from match-day ticket sales, merchandise sales and broadcasting contracts. Naturally, better performance on-field would attract a larger fanbase as well as increased revenue with a hierarchical monetary reward in the competitions. Soccer players play a huge part in the fortunes of their club, not only contributing with their on-field performance but also revenue from merchandise sales and advertising contracts. All these factors contribute to the extremely high transfer fees for players where prospective clubs have to fork out these sums to their parent clubs.

The FIFA video game franchise is one of the best-selling and is well known for having extremely well researched and documented in depth player ratings for players' physical attributes and technical soccer skills, along with demographic statistics like club, age etc. For our project, we studied how the amalgamation of these features aided in predicting the player's actual market value. We used past editions of the game data from 2015 to 2021 to build our model. We hope that through our project, soccer clubs will be able to make informed decisions on future transfers and have a sense of how their investment in new players might pan out financially, particularly if there is predicted to be significant profit or loss to be made from the transfer.

Pre-processing

We chose to filter out rows of players to keep only those who have a non-zero market value recorded. Those players with a market value of zero recorded in the data are assumed to have unknown market value, and thus we omit them from our analysis. We found that 31% of our dataset has players with unknown market value, but altogether our non-omitted records are plentiful, being between 10,000 to 15,000 per year over five years of data. For all other null values we found, we replace these with zero. The reason we handle our nulls in this manner is that null values are found entirely in one of two cases - first, players who are outfielders have entirely different skills than goalkeepers, and so their goalkeeper skills are null values, and vice versa. As they are being evaluated separately, we are able to separate these two classes of players entirely.

We also choose to transform the market value field to its natural-log, as the net worth varies by several orders of magnitude, and using the natural-log would make a linear trend more apparent. Finally, we decided to add two features that were not player skill related: year and league rank. To account for how player market values differ across seasons, we incorporated year as a feature in our model. This ordinal feature classifies the data according to the year of the game edition the data was from, with 2015 - 0, 2016 - 1 and 2021-6. Generally speaking, as the year progresses, the market value will always increase for a player. We also recognize differences in the resources that different leagues have, which would influence the market values of players in those leagues and so we created a feature called League Ranking. We grouped leagues that exist in the current dataset into 5 categories, with tier 1 containing the most powerful leagues and tier 5 containing players that are free agents without a club. This ranking system was based on the study by Habib (2021).

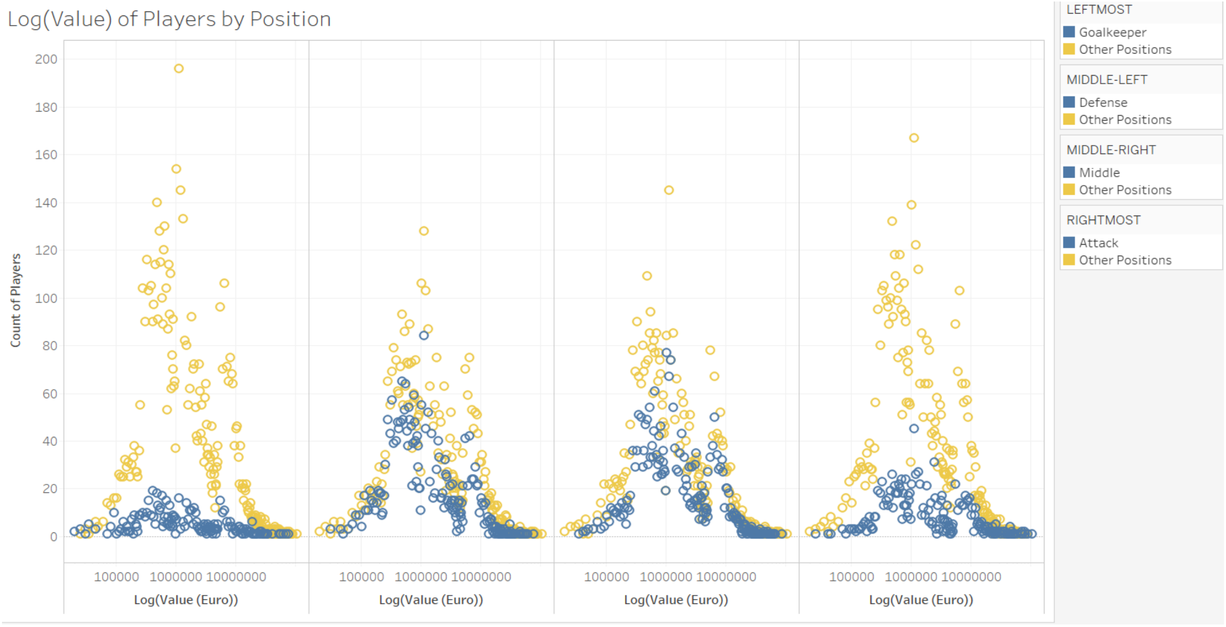
Previously, we had also transformed a column called player_traits into a many-hot vector, and planned to include this in our analysis. The player_traits column contains a set of named traits associated with each player character in the video game (ex, Diver, Long Passer, etc.). Players can have from anywhere between none and many player traits, which is why we decided to create a many-hot vector to represent this. However, ultimately we decided that rather than use player traits because the dimensionality of the columns would be too expansive, and compared to league ranking, we believed it would be a poorer predictor of player value.

Below is a summary of all features we incorporated into our model. We note that the data also has several sub-skills, which are all components of each of the six skills, and whose weighted average equals the above skills. For example, movement_acceleration and movement_sprint_speed are the two sub-skills that make up pace; if we include them along with pace in our input features, it would be redundant as together they represent a linear combination of the pace column. Finally, we included two other ranking features in addition to the six core skills: Overall score and Potential score. Overall represents the average score of the six core skills per player. Although they are related, we decided to keep the feature in order to investigate whether it would be possible to predict a player's value based on an overall ranking. Potential ranking represents what sports analysts believe is possible for a player to achieve in future overall ranking. We can interpret this as potential value that a player could provide if he performed well in the future, which would affect what his market value would be.

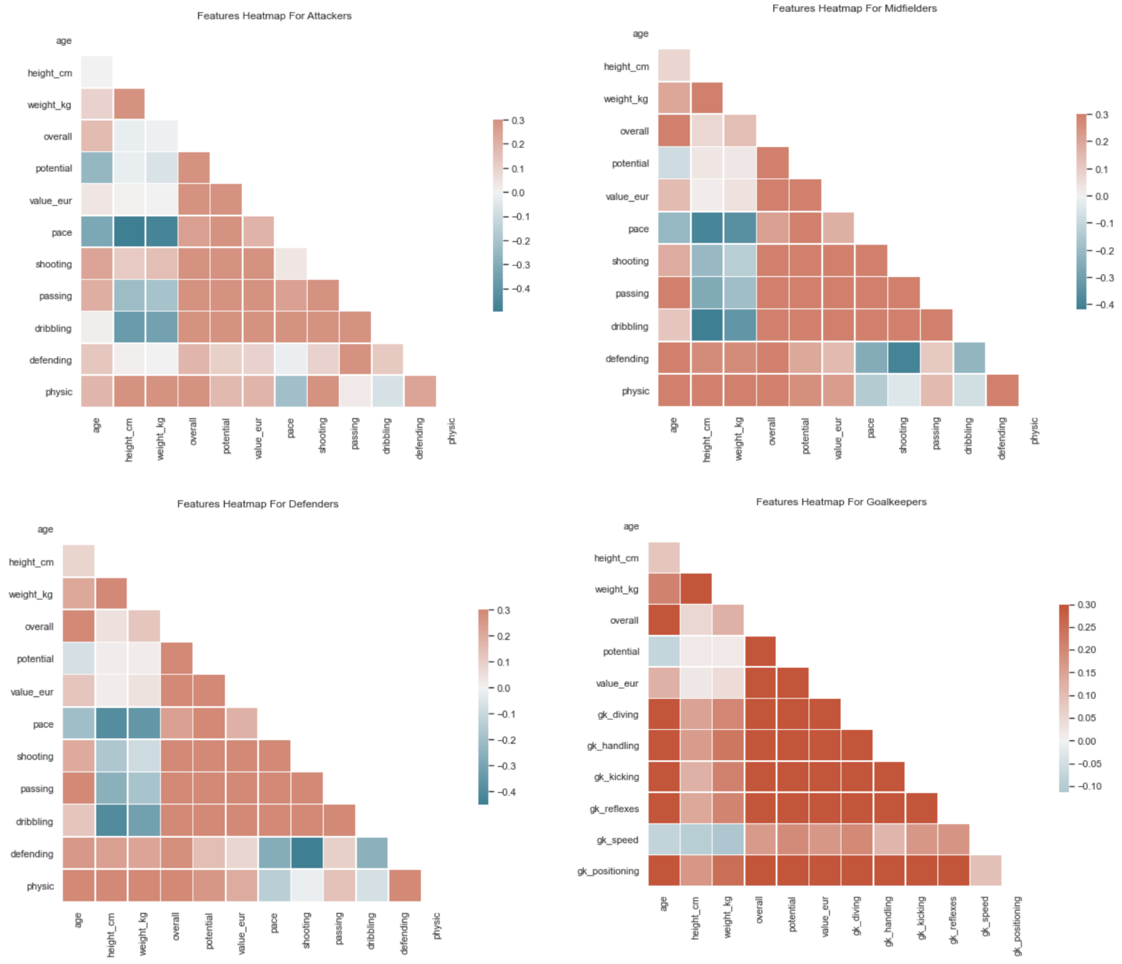
Feature	**Description**	**Example**
Ranking*	Rating on a scale of 1-100 for a player's skill	96
Year	The FIFA year for the player record, in units of years since the first year in the dataset, which is 2015	0
Age	Player's age for a version of FIFA	24
Weight	Player's weight in kilograms for a given record	75
Height	Player's height in centimeters for a given record	180
League Ranking	Rank of the soccer league that a player plays for in a given record	1

**We included the six basic rankings for each player. These rankings are different for outfielders and goalkeepers. These are: [pace, shooting, passing, dribbling, defending, physical] for outfielders, and [diving, kicking, handling, reflexes, speed, positioning] for goalkeepers.*

Exploratory Data Analysis



Above, we used Tableau to visualize the log of market value across all players in our dataset who belong to one of four general positions (in order with the charts below): goalkeepers, defenders, midfielders, and attackers. In general, player value appears normally distributed across all positions. Since there is only one goalkeeper per team, we see a smaller concentration of goalkeeper counts around the mean, but there still is a normal distribution for the goalkeeper market value scatter plot.



Correlation Heatmaps - Attackers, Midfielders, Defenders, Goalkeepers

The above correlation heatmaps show the relationship between the features we have chosen for the subgroups of players - Attackers, Midfielders, Defenders and Goalkeepers. Many of the features share a positive correlation, which we believe is to be expected as most of these athletic traits are shared with more skilled players. Height and weight are negatively correlated with many offensive stats, which we also believe is to be expected in soccer players. Due to the nature of many positive correlations, we may see that our linear model can be built using a smaller number of our initial proposed features.

Based on our findings from the correlation heatmaps, we decided to split our data into Outfielders and Goalkeepers to build two separate models. Outfielders comprise of the attackers, midfielders and defenders and we can see that the heatmaps for these 3 categories of players are extremely similar, while goalkeeper’s heatmap is vastly different with vastly different statistics.

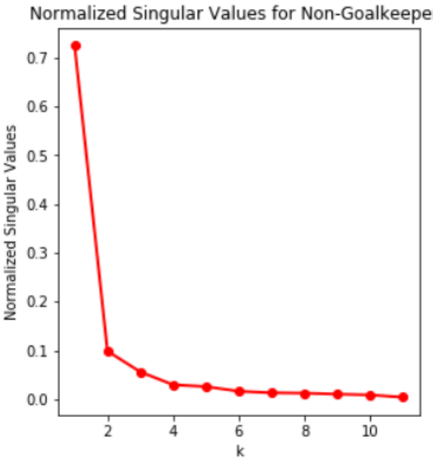
Fairness of Feature Selection

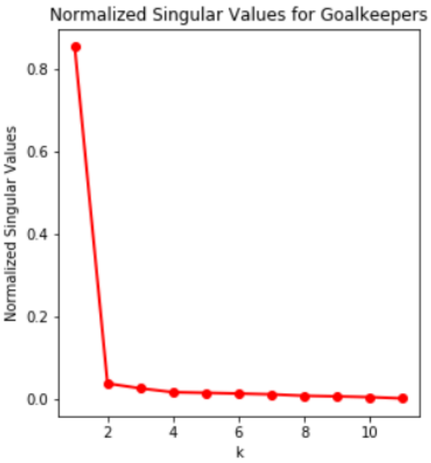
Ideally, the value of the players would be dependent solely on their physical attributes and their skill level. However, this is not true in the real world as there are various other features that affect player values. One example would be the amount of public exposure or the popularity of a player, which would entice more fans to make shirt purchases or increase the brand image of the club that would correlate to a higher valuation. To account for the disparity, we introduced an additional feature 'league ranking' which separated players from different tiers of leagues.

Preliminary Model and Results

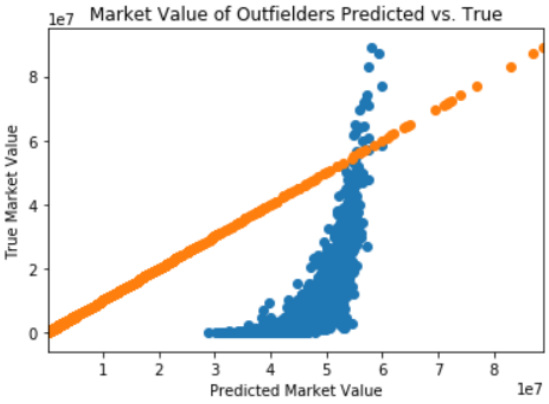
To do our PCA we decided to merge our data sets from FIFA 2015 - 2020 to try and predict market value using one year's features to try and predict the respective player's market value next year. To do so we separated the whole data set into random subsets of 65% training, 15% validation, and 20% testing. We did this twice: once for the dataset of all players not including goalkeepers, and then again for the dataset but only keeping goalkeepers since they have six different core skill areas.

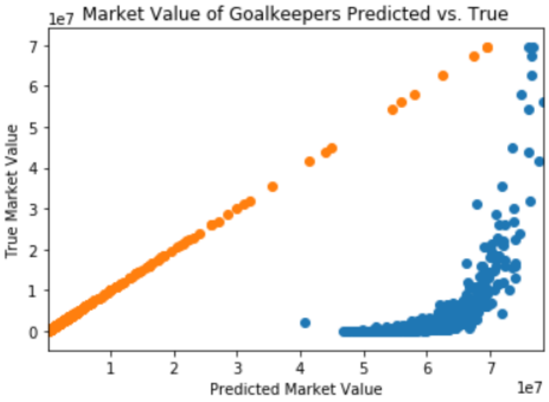
After separating the datasets then we found the SVD of the training set and did some analysis on the singular values including finding the respective eigenvalues. Here we see the normalized singular values of which we can see that for outfielders (non-goalkeepers) we should consider keeping 5-7 features as that is when the singular starts flattening out, while for goalkeepers we consider keeping 3-6 features for the reasoning. Choosing from this subset allows us to avoid over and underfitting while choosing the best possible features from our selection to predict future market value.

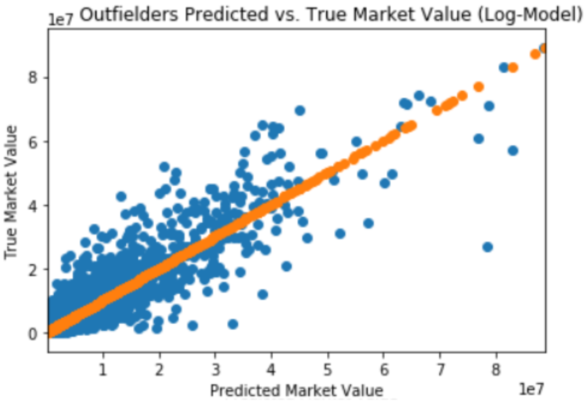

Normalized Singular Values for Non-Goalkeepers

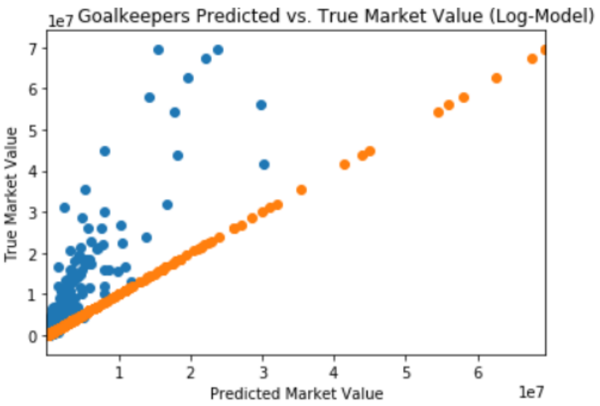

Normalized Singular Values for Goalkeepers

We then used our validation set to see how well our training set fits a linear regression model using 5-7 features of the non-goalkeepers, and 3-6 features of goalkeepers. Then based on the validation set output of mean squared errors we kept the number of features that managed to get the smallest MSE. For the case of non-goalkeepers the validation set chose 7 features to keep, while for goalkeepers the validation set chose to keep 6 features. We ran this not only on the market value but also the natural-log of it as described above and found that the validation set still chose the same number of features for non-goalkeepers and goalkeepers, respectively. We found that the linear regression fit against the natural-log of market value performed the best for non-goalkeepers and goalkeepers.


Market Value of Outfielders Predicted vs. True


Market Value of Goalkeepers Predicted vs. True



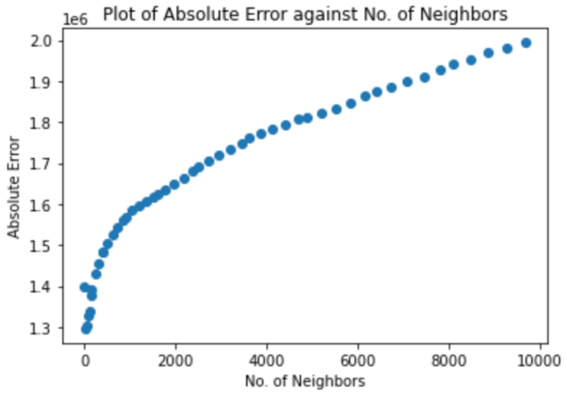


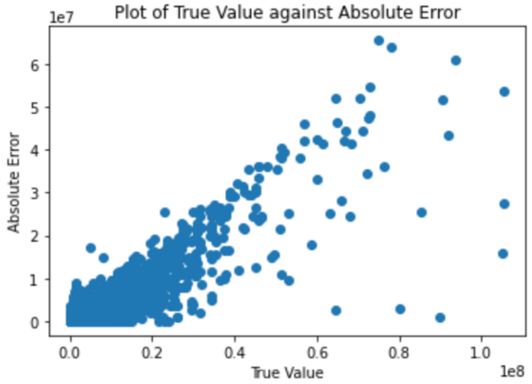
Here we see the plots for predicted market value (x-axis) vs true market value (y-axis) where on the left we see outfielders, and on the right we have goalkeepers. And on the top row we see the model outputs from performing SVD on the market value, while on the bottom row we see the model output from performing SVD on the log(market value). And along the middle we have plotted the true y values that way we can clearly see how much deviation there is. The bottom row can be seen to look better than the top row (taking the log of market value performs better) as it follows the middle points more closely of the true value.

Even though the log(market value) models fit considerably better than not taking the natural-log, after taking the exponential we see that our mean absolute error is still along the magnitude of $1.1 million, for outfielder (non-goalkeepers), and $2.3 million for goalkeepers (compared to $41 million, and $58.8 million, respectively). The reason for us choosing to look at mean absolute error rather than MSE is the amount of outliers present in our data. This output makes us think that performing a linear regression is not the best way to go about it, and hence we will now try some classifiers.

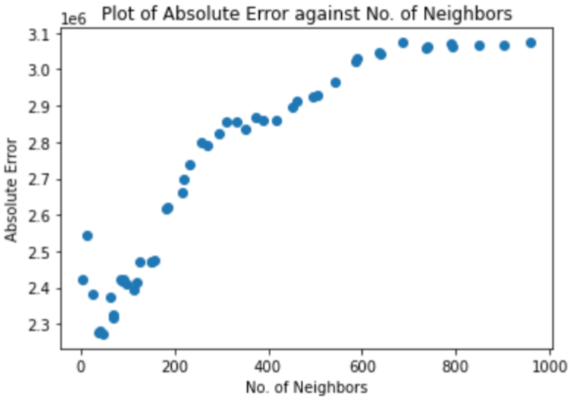
Classifier #1: k-NN Model

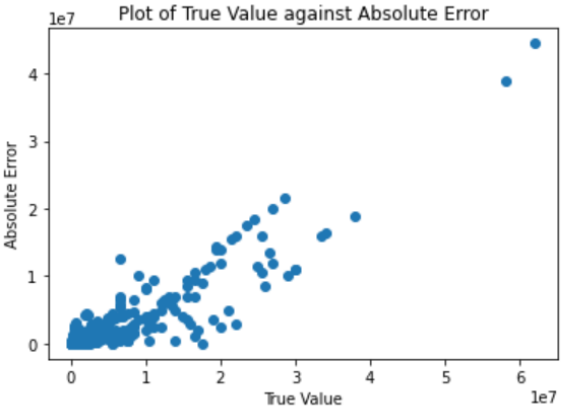
After trying out linear regression, we decided to try a supervised learning method - k Nearest Neighbors. To achieve this, we first split the data into 2 different sets, outfielders and goalkeepers, due to the difference in the player statistics between these 2 groups and selected 13 features (including age, skill ratings etc.) We then split the data into 65% training, 15% validation and 20% test sets. We then made a for loop to run sklearn k-NN for a range of neighbors with increasing step sizes.





For outfielders, we obtained the above plots, which show that as the number of neighbors increases, so does the absolute error, with a decrease in gradient after about 2000 neighbors. We found that the ideal number of neighbors with the least absolute error was 30, with a mean absolute error of $1.3 million. The scatter plot on the right shows that for players with higher true values, generally, the model performs poorer, with a higher absolute error (difference between validation and predicted data), intuitively this makes sense as with larger player values a smaller percentage deviation would result in a larger absolute error.





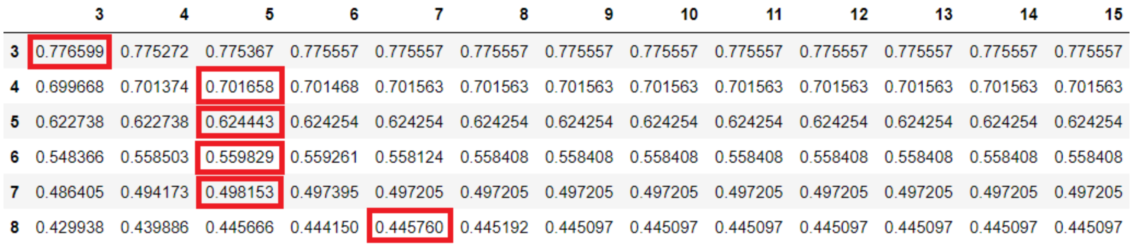
For goalkeepers, we obtained the above results which do not have as nice of a plot, due to a much smaller set of data. We observed some outliers with a higher absolute error for a low number of neighbors and as expected, the absolute error increases with the number of neighbors. We found that the ideal number was 46 neighbors with a mean absolute error of $1.7 million. The scatter plot on the right confirms the findings from outfielders that players with higher true value tend to have a higher absolute error.

The findings from the k-NN were not ideal and we decided to try an unsupervised model of Bagging of Decision Trees.

Classifier #2: Bagging (Decision Tree) Model

For our last model we decided to try out a Bagging of Decision Trees classifier. To do so we created bins based on market value. So, for the case of 3 bins we chose the bins: < $600,000; $600,000 - $2,500,000; $2,500,00 <. We created bins like this for the cases of 3 - 8. We chose these numbers of bins as we wanted to group as many players that are like each other while also not having so many groups that the model has problems identifying factors that separate these groups.

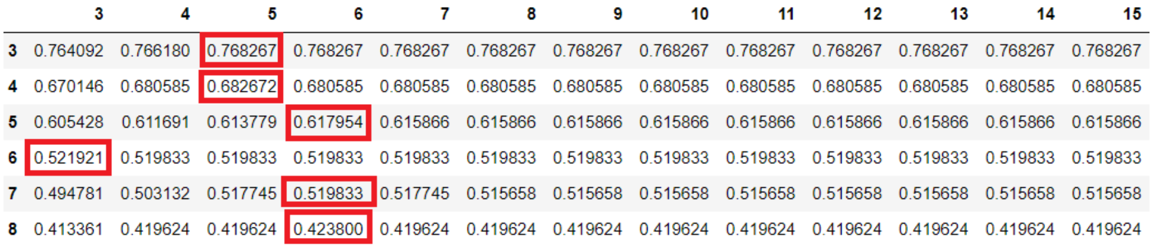
We used the same dataset of features that we used to fit on the kNN and PCA methods to fit this model, creating 2 outputs one for outfielders and another one for goalkeepers. We again split our model into 65% training, 15% validation and 20% test set. The purpose of our validation set was to verify how well the model built from the training set predicted the true values. We then output these values in a table to see which depth performed the best per number of bins and then based on the output we would choose the model that would give us as many bins as possible while taking the correctly predicted percentage into account. For the case of outfielders our table looked as follows:



	3	4	5	6	7	8	9	10	11	12	13	14	15
3	0.776599	0.775272	0.775367	0.775557	0.775557	0.775557	0.775557	0.775557	0.775557	0.775557	0.775557	0.775557	0.775557
4	0.699668	0.701374	0.701658	0.701468	0.701563	0.701563	0.701563	0.701563	0.701563	0.701563	0.701563	0.701563	0.701563
5	0.622738	0.622738	0.624443	0.624254	0.624254	0.624254	0.624254	0.624254	0.624254	0.624254	0.624254	0.624254	0.624254
6	0.548366	0.558503	0.559829	0.559261	0.558124	0.558408	0.558408	0.558408	0.558408	0.558408	0.558408	0.558408	0.558408
7	0.486405	0.494173	0.498153	0.497395	0.497205	0.497205	0.497205	0.497205	0.497205	0.497205	0.497205	0.497205	0.497205
8	0.429938	0.439886	0.445666	0.444150	0.445760	0.445192	0.445097	0.445097	0.445097	0.445097	0.445097	0.445097	0.445097

This table has the number of bins as the rows (3 - 8 tested bins), and the max tree depth per Decision Tree as the columns. We have also placed a red rectangle around the best performing max tree depth per number of bins. Based on this output we can see that from 3 to 4 bins we drop 7.5% in accuracy, from 4 to 5 bins we drop 7.7%, from 5 to 6 bins we drop 6.5%, from 6 to 7 we drop 6.2%, and from 7 to 8 bins we drop 5.2% in accuracy. Based on these drops as well as the numbers themselves we chose to fit on 4 bins with a max tree depth of 5 as we thought that we would be willing to take a 7.5% drop but not an additional 7.7% drop to get even more bins. We found that our test set attained a 70.2% correct prediction rate which is quite close to the 70.17% correctly predicted rate from our validation set.

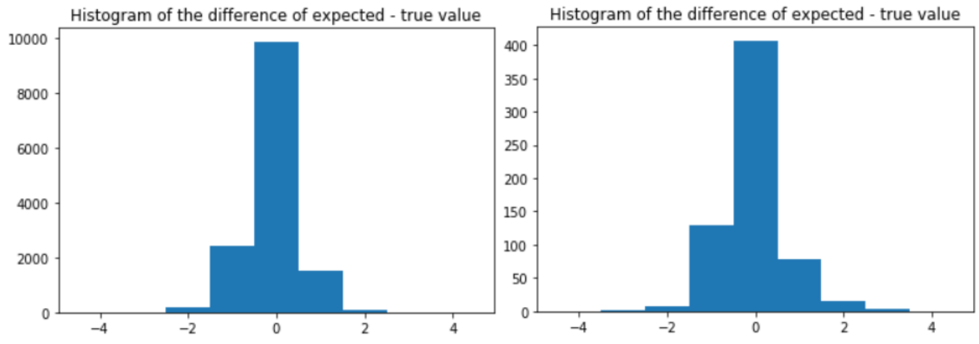
We repeated this process for the goalkeeper dataset with the same bin count and bin sizes, and found that the results were decently similar as follows:



	3	4	5	6	7	8	9	10	11	12	13	14	15
3	0.764092	0.766180	0.768267	0.768267	0.768267	0.768267	0.768267	0.768267	0.768267	0.768267	0.768267	0.768267	0.768267
4	0.670146	0.680585	0.682672	0.680585	0.680585	0.680585	0.680585	0.680585	0.680585	0.680585	0.680585	0.680585	0.680585
5	0.605428	0.611691	0.613779	0.617954	0.615866	0.615866	0.615866	0.615866	0.615866	0.615866	0.615866	0.615866	0.615866
6	0.521921	0.519833	0.519833	0.519833	0.519833	0.519833	0.519833	0.519833	0.519833	0.519833	0.519833	0.519833	0.519833
7	0.494781	0.503132	0.517745	0.519833	0.517745	0.515658	0.515658	0.515658	0.515658	0.515658	0.515658	0.515658	0.515658
8	0.413361	0.419624	0.419624	0.423800	0.419624	0.419624	0.419624	0.419624	0.419624	0.419624	0.419624	0.419624	0.419624

The output overall looks fairly similar but for the goalkeepers data we found that the results were more sporadic regarding the correct prediction rate. We believe this comes from the fact that the goalkeeper dataset only has approximately 2000 data points in our training set compared to the 50,000+ in our outfielders dataset. Anyways, regarding this data we chose to fit on 5 bins as the correct prediction rate drops 9.6% accuracy from 5 to 6 bins, but from 4 to 5 it only dropped 6.5%. Our test set then attained a 63.5% accuracy using these 5 bins at a max tree depth of 6.

We then decided to take a look at a histogram of difference of expected - true market value.



Outfielders *Goalkeepers*

We found that for both models our bagging classifiers tend to slightly under predict than over predict (if true market value is in the 2nd bin, if it predicts incorrectly it tends to predict 1st over 3rd bin). And for the cases of a difference of 2 or greater we see that this is rare as can be seen in the histograms below.

Weapon of Math Destruction

Our prediction outcomes are easily measurable as we can compare them to the player value approximations available on transfermarkt.com or reference them to actual player sales. Arguably, our predictions should not have significant negative influence on players as their values are not set in stone but can vary with each season as an improvement in performance could increase their predicted values, this variability also prevents the creation of self-fulfilling feedback loops. Thus our models are definitely not weapons of math destruction.

AutoML, Conclusion and Future Work

After evaluating all models, we found that the best performing model was our Bagged Decision Tree Classifier for both Outfielders and Goalkeepers, with the model using 4 bins at max depth 5 attaining a 70% prediction rate on our test set for Outfielders and our model using 5 bins at max depth 6 attaining a prediction rate of 63.5% for Goalkeepers. We believe that this model performed much better than the regression via SVD and k-NN classifier due to our definition of the label class being defined as market value bins rather than predicting the actual market value itself. For the other models, even with log-scaling, there was large variance in the market value data, which would affect the performance of the models. In future work, one may consider building and training a neural network as they are better at classification tasks through use of its hidden layers.

After building our models, we decided to use TPOT, an automated machine learning tool, to observe whether there is potential for greater accuracy using regression and classification algorithms on our current feature and label sets. When predicting actual market value, the test set errors all were close or just below $1 million. Though all of these errors were for the most part better than those we had obtained through building our kNN and regression models, which ranged from $1.1 to $2.3 million, their magnitudes were still not ideal, especially in a dataset where many players' values are in the tens or hundreds of thousands. In contrast, predicting market value tiers using the best TPOT classifier yields similar results to our own Bagged classifier, outperforming over all bins but still having comparable scores (TPOT's 4-bin Outfielder and 5-bin Goalkeeper scores outperformed our equivalent models by 5%). Observing all these results help support our hypothesis that the current features we used to build our models limit the accuracy of our predictions, regardless of model type.

TPOT Analysis - Predicting Actual Market Value		
	Outfielders	Goalkeepers
Classifier	ExtraTreesClassifier Test Set Score: $1.1 million	RandomForestClassifier Test Set Error: $1.0 million
Regressor	ExtraTreesRegressor Test Set Error: $1.0 million	RandomForestRegressor Test Set Error: $1.0 million

TPOT Analysis - Predicting Market Value Tier		
Bins	Outfielders	Goalkeepers
3	81%	78%
4	75%	71%
5	68%	65%
6	57%	59%
7	56%	58%
8	52%	51%

Given these results, we don't believe that our model is mature enough to predict player value using these features alone. Although there is a correlation between market value and player skills, the factors by which a player's market value changes can occur over the course of a season, depending on other factors not captured in our current model, such as their accident record per season, and their own team's win / loss record. This information is not available in our current FIFA dataset, but we believe that finding and including these features into our analysis would help build a more robust predictor. The challenge would be finding information from lesser known leagues and players - while prominent European leagues would have many sports analysts tracking their statistics, data from less popular leagues may come from varied data sources, and may be limited or outdated depending on the popularity and resources those leagues have, which is why we were not able to include them in the scope of our current project.

References

Habib, D. (2021, January). *Top 50 Best Football Leagues in the World Ranking 2021*. The Football Lovers. https://thefootballlovers.com/best-football-leagues-in-the-world-ranking/

Leone, S.(2020), *FIFA 21 Complete Player Dataset* https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset