How Much is Your Favorite Soccer Player Worth?

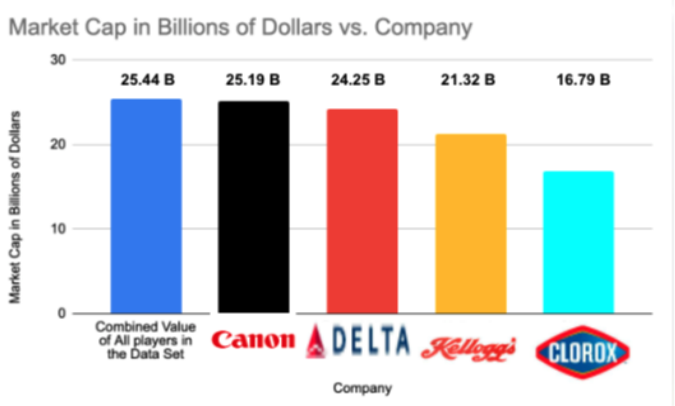
Brain Station Capstone Project: Final Report

Ben Katz
benkatz21@gmail.com

4/4/2022

My capstone project asks the following: **using data analytics and machine learning, how might I predict the transfer market values of European soccer players and determine the primary statistical predictors of those market values?**

In club soccer, a *transfer* is when one player leaves one club to join another. Usually, this involves a *transfer fee*, a sum of money the acquiring club pays to the selling club. A player's *transfer market valuation* is an estimate of how much a team should pay for a player's rights. My dataset has 1,494 observations, each representing one soccer player. The combined value of these athletes is higher than the market caps four globally recognized companies (**Figure 1)**.



Soccer clubs spent $4.86 billion dollars in transfer fees last year. My goal is to create models that will help clubs more accurately value their players and potential transfer targets, leading to more on-field success and increased revenues.

My data set had 67 columns before engineering any features, each one representing a specific statistic, demographic or positional information. I obtained the data from two sources: transfermarkt.us and fbref.com. Using *Beautiful Soup*, I scraped the following information from transfermarkt.us: player name, position, age, nationality, club, and market value. I only scraped information from the Big 5 European Soccer leagues: the *English Premier League,* the *German Bundesliga,* the *Italian Serie A,* the *Spanish La Liga,* and the *French Ligue 1*. I scraped the information of **1,600** soccer players. There were **51** players for whom I manually entered information into the dataset. I removed goalkeepers from the dataset because of vastly different roles and statistical data.

I downloaded all statistics from the 2020-2021 and 2021-2022 seasons (up to 2/27/2022) from fbref.com. I then joined these statistics by player names and summed them together to create a full picture of all raw statistics from the previous two seasons. Then I joined the scraped data and the statistical data. I created per 90-minute statistics to represent production on a per-game basis. I also created success rate statistics to measure what percentage of action was completed successfully. For example, if a player attempted 100 passes and completed 80 of them, his passing success rate would be 80%. I used per 90-minute statistics and success rates to compare all players on the same scale. In total, there are 60 columns representing statistical information, 7 representing demographic and positional information, and 1 representing the market value of each player (**Figure 2**).

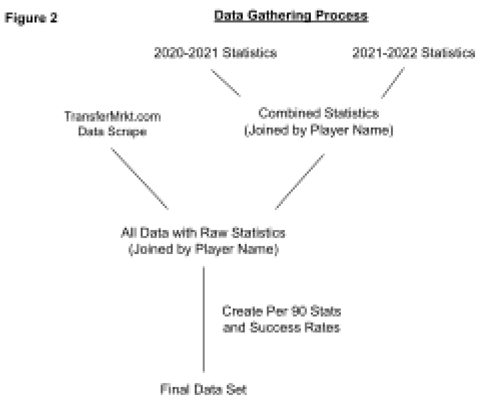


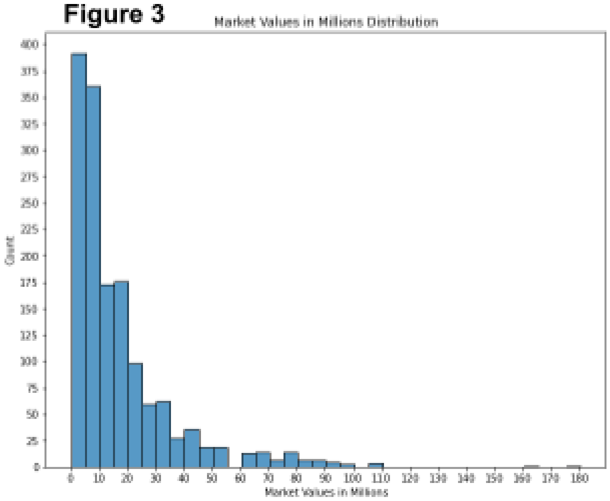
Figure 2
Data Gathering Process
2020-2021 Statistics
2021-2022 Statistics
TransferMrkt.com Data Scrape
Combined Statistics (Joined by Player Name)
All Data with Raw Statistics (Joined by Player Name)
Create Per 90 Stats and Success Rates
Final Data Set

There are two primary concerns with this dataset. First, the dataset is small. There are only so many players in the Big 5 European soccer leagues. This dataset nonetheless is informative. Theoretically, a team could run a player's statistics through the model and determine what their value would be if they played for a Big 5 team. Second, there are some outliers. There are young players who made a Big 5 roster, but have barely played over the previous two seasons. Therefore, their values are strictly based on potential. This caused some performance issues, but the models adapted well to these players.

I engineered four features: *position group, position subgroup, European, and club tier.* I used *position group* to split the dataset into three groups: defenders, midfielders, and attackers. These groups play distinct roles, so I conducted my analyses on each group separately to gain insights on what statistics are important for each specific *position group*. *Position subgroup* (types of players within *position groups*) and *European (*a binary feature) did not end up being significant predictors in any of the models. In the end, *club tier* was the only impactful engineered feature. *Club tier* was created by comparing the sum of market values of all players from each team; the more value a team has, the higher its tier. I used domain knowledge as well. As a soccer fan, I have an understanding of which clubs should fall into which tiers.

In my second notebook, I focused on EDA and hypothesis testing. I wanted to get more insights into how market values were distributed across different positions and leagues. I noticed there was a large skew in the dataset. There were many players with relatively low values, and few elite players on the high end of the market value spectrum (**Figure 3)**. This skew was consistent throughout all *position groups*, and present in some features also.

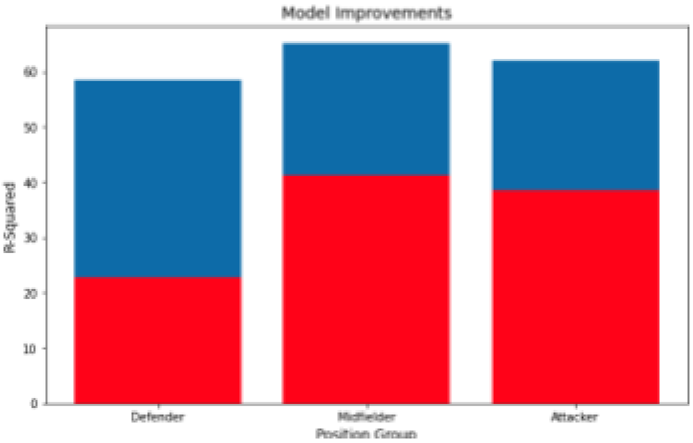
Figure 3



I ran hypothesis tests on market values across positions and leagues.The *Premier League* had a significantly higher average value than every other league. Center backs had a significantly higher average market value than wingbacks.

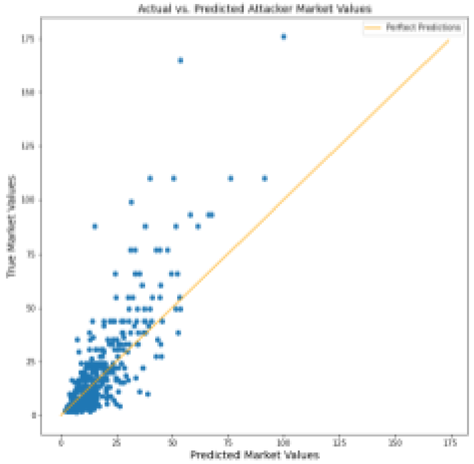
Interpretability is as important as results. To gain insights into which statistics offered the most value, I followed the same procedure for each of the position groups. I created a baseline model using just statistical data, and then introduced my engineered features. I used linear regressions for my baseline model, and lasso regressions for my optimized models. I used r-squared as my accuracy metric. I converted my target variable, Market Values in Millions, to log space. This made the distribution less skewed, and also prevented the model from predicting negative market values, which did not make sense. I chose lasso regressions to help remove variables through L1 regularization, and to account for collinearity between features. If the lasso model produced coefficients that were contradicting, for example, if Pass Cmp/90 had a positive coefficient and Short Pass Cmp/90 had a negative coefficient, I examined the contradicting features and removed those with weaker correlations. My engineered features strengthened the model as the r-squared value for my optimized models were all higher than my baseline model (**Figure 4)**.

Figure 4



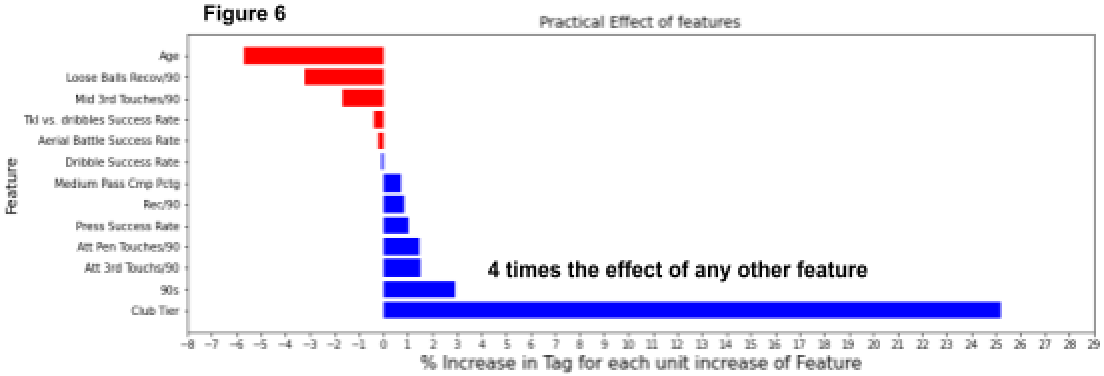
There were two major problems with the models. Each model undervalued players, especially elite players (**Figure 5)**. The The majority of points are above the orange line, showing their actual values are much higher than their predicted values. Elite soccer players are scarce, and clubs have to pay a premium to obtain them. I call this the “Elite Player Tax”. My models did not adequately account for this tax.

Figure 5



The second issue is that *club tier* dominated all the models (**Figure 6)**. I engineered a strong feature that predicts a large portion of the variance in the dataset. However, it prevented me me from seeing which statistics are actually the most important. European soccer is an oligopoly; rich clubs buy all the elite players and win all the trophies. The best teams possess the ball more often and so their players have more opportunities to generate statistics. The *club tier* feature accounted for all of this, and in turn swallowed up the effects of other statistics.

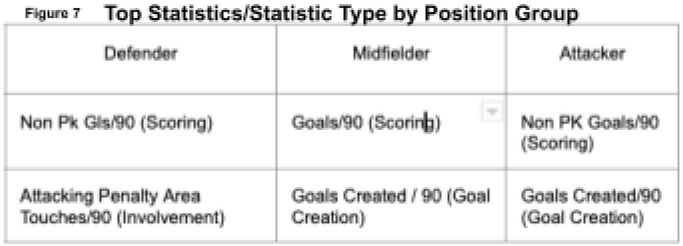
Figure 6



*Age, 90s (*games played) and *club tier* (discussed above) were the only features present in all three *position group* models. The best players play, so it makes sense that *90s* would be a consistent predictor *Age* was the model solving the outliers issue. These players' values derive strictly from their potential. The negative *Age* coefficients are representative on the model picking up this trend.

To glean which statistics were important, I removed *club tier* from the dataset and ran statistical linear regressions on each of the *position groups*. The most important statistical category for each *position group* was scoring. For attackers and midfielders, the second most important statistical category was goal creation. For defenders, it was touches in the attacking penalty area, which I will file under offensive involvement (**Figure 7)**. Across all position groups, scoring goals was the most important statistic. Goals are hard to come by, and if a player can score his value will increase. For attackers and midfielders, the second most important statistic was whether or not the player created goals for their teammates. Interestingly, for defenders, there were no defensive statistics present at the end of the model. This is not because these statistics are not important, but rather the best defenders do not have to defend as much because they play for the best teams.

Figure 7 **Top Statistics/Statistic Type by Position Group**



Defender	Midfielder	Attacker
Non Pk Gls/90 (Scoring)	Goals/90 (Scoring)	Non PK Goals/90 (Scoring)
Attacking Penalty Area Touches/90 (Involvement)	Goals Created / 90 (Goal Creation)	Goals Created/90 (Goal Creation)

I ran preliminary grid searches on K-Nearest Neighbors, decision tree, and support vector regressor models to see if any were worth exploring in the future. The only one that had any predictive power was the support vector regressor. This will be the first thing I will look into post boot camp.

These models are not ready for deployment, but I do have a few ideas on how to improve performance. I will look into more powerful models, such as XGBoost and neural networks with the hope that the added power will be able to model elite players. I want to create composite scores for each statistical category. I got caught up in seeing which individual statistics mattered the most. From focussing on individual statistics, I realized many of these statistics tell similar stories. If I create composite scores for each statistical category, I could maintain interpretability and improve performance.