SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

*E-ISSN: 2996-8380*
*Published by Leadership & Innovation Lab*

**FALL 2025**

# Machine Learning in Soccer: Determining the Value of MLS Players in Past Seasons Based on Statistics and How that Applies to Value as a Player in General

**By Dhruv Panchagatti**

## AUTHOR BIOGRAPHY

Dhruv Panchagatti is currently a senior at Northwood High School. He enjoys engineering and soccer with a passion for building ML projects that connect stats to the real world. He hopes to work in Formula One someday.

## ABSTRACT

This research develops a machine learning model to predict the yearly wages of Major League Soccer (MLS) players based on their performance statistics. By analyzing the "MLS season player stats and salaries" dataset that includes metrics such as goals, assists, minutes played, passing accuracy, defensive actions, and more, the study aims to identify the key performance indicators most strongly linked to salary outcomes. The model, trained on 2022 data, utilized feature selection methods to isolate the most relevant statistics, and various regression algorithms like linear and ridge were tested to determine the most accurate predictive model. Performance was evaluated using metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The findings reveal that metrics such as goals scored, goals per game, and successful dribbles have a strong correlation with player wages, while factors like defensive actions and passing accuracy are less impactful. In general, attacking outputs and positions were stronger predictors of high salary than other statistics.This research highlights the potential of machine learning in sports economics, offering teams and agents a valuable tool for making data-driven decisions in player contracts and negotiations. The study also underscores the importance of data analytics in understanding wage determinants in professional soccer. Future work could expand this approach to other leagues and incorporate additional factors like player age, experience, and marketability.

**Keywords:** *Sports Science, Machine Learning, Soccer/Football, Player, Value, Important Statistics, Goal/Assists, Positions*

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

*E-ISSN: 2996-8380*
*Published by Leadership & Innovation Lab*

**FALL 2025**

## INTRODUCTION

As technology, funding, and sports science have evolved over the decades, the use of statistics has become one of the primary ways that people determine the value of a certain athlete or sports team. Baseball was one of the premier sports to introduce a batting average stat to help teams determine how good a hitter was. Using this statistic fans and teams could now know that a 0.25 batting average (how often the batter hits the ball) is about average and closer to 0.3 are some of the best hitters in the league (Bailey, 2023).

This metric is decently reliable when it comes to assessing the skill of a hitter in baseball but what about a team sport like soccer where a majority of the team contributes to goals rather than a single hitter contributing to a home run? Some might say that goals are the best indicator of a soccer player's value. However, this raises questions about how the value of other positions come into play such as defenders and goalkeepers, and, in general, how the value of a player is measured.

First, it is important to realize that data like this is crucial, not only to the soccer clubs, but to the individual players themselves. One common example occurs during penalty kicks. The kicker generally has a substantial advantage due to the close proximity of the goal, but it is not as straightforward as that. The goalie usually has access to or has studied previous penalties of the taker (Pastuovic, 2020). How many times he went left or right, what does his run-up look like, or does he tend to look at the opposite side to the kick? Many big games, such as the 2022 FIFA World Cup, end in penalty shootouts so understanding statistics can be the difference between winning and losing.

This paper will not be focusing on penalty kicks in particular, but an array of statistics to determine what exactly makes a player valuable.

## METHODOLOGY

This study used a free dataset found online to develop a machine learning model to determine the stat that contributed the most to a player's value (Ivanovs, 2023). The dataset comes from the 2021-2022 Major League Soccer (MLS) season and it has captured almost every stat recorded for 623 players during that year. From goals to saves to touches to even positions, this dataset has over 90 different stats that capture almost all aspects of "the beautiful game".

**Table 1**

*The following is the first five rows and a few columns of the dataset (before being trimmed down)*

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

FALL 2025

| | 2022 guar. comp. | height, cm | weight, kg | games_played | game_started | mins | total_wins | goals | assist | accurate_back_zone_pass | accurate_cross | accurate_cross_nocorner |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1076000.0 | 178.0 | 70.0 | 32 | 32 | 2760 | 14 | 19 | 8 | 118 | 1 | 1 |
| 2 | 7443750.0 | 175.0 | 73.0 | 21 | 20 | 1736 | 9 | 17 | 3 | 96 | 2 | 2 |
| 3 | 3201120.0 | 169.0 | 65.0 | 26 | 24 | 2146 | 14 | 17 | 1 | 140 | 3 | 3 |
| 4 | 1762572.0 | 186.0 | 79.0 | 33 | 32 | 2875 | 14 | 16 | 9 | 432 | 1 | 1 |
| 5 | 1926250.0 | 173.0 | 72.0 | 31 | 28 | 2505 | 11 | 16 | 12 | 343 | 44 | 9 |

*Note.* The rest of the statistics are cut off. The head() function was called and shows just the top left of the entire dataset but the dataset does contain other statistics and hundreds of rows that are not shown.

To get back to the aforementioned question: How exactly do teams measure the value of a player? The dataset also included a column that showed the compensation that the players received that season. For soccer clubs, money is what keeps them afloat at the end of the day so paying a player a high wage is a good indication of the value that player brings to the team as a whole.

Before getting into the research, it would be helpful to understand some of the machine learning techniques used and how a model was able to be trained and validated using the dataset.

Machine Learning is giving a system data to create a model that is able to make predictions based on the data set. In this case, the model was given a CSV file containing a table.

Using Python, the table was modified to fit the needs of the model. First of all, the model only works with numbers so all text data such as position and preferred foot had to be accounted for. For example, the position column would have one of four entries: Forward, Midfielder, Defender, and Goalkeeper. But, using the process of one hot encoding this one column was separated into four columns with each of the positions as a column and either a 1 or a 0 in that category. As stated before, the model processes numbers, not words.

**Figure 1**

*Code for One Hot Encoding*

One Hot Encoding for position and foot columns

```
[ ]  df['Forward'] = df['position'].replace(['Forward', 'Midfielder', 'Defender', 'Goalkeeper'], [1,0,0,0])
     df['Midfielder'] = df['position'].replace(['Forward', 'Midfielder', 'Defender', 'Goalkeeper'], [0,1,0,0])
     df['Defender'] = df['position'].replace(['Forward', 'Midfielder', 'Defender', 'Goalkeeper'], [0,0,1,0])
     df['Goalkeeper'] = df['position'].replace(['Forward', 'Midfielder', 'Defender', 'Goalkeeper'], [0,0,0,1])
     df.drop(columns='position',inplace = True)
```

Using personal knowledge (a Varsity soccer player and player for ten years) as well as peer consultation (help from coaches and teammates), the 90+ stats were trimmed down to 14 considered most important. Importance was determined by omitting subcategories of existing statistics. There were several niche statistics that were generalized into one category. For example, a few statistics were right-foot goals, left-foot goals, and

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

FALL 2025

header goals, but instead of factoring all of these as different statistics, they were combined into one statistic called "goals". Through trimming methods like that, the data set dwindled down to 14 major statistics.

**Table 2**
*List of Initial Statistics (Linear Regression Model)*

| | |
|---|---|
| Total Wins | How many wins the player was a part of |
| Goals | How many goals a player scored |
| Assists | How many passes a player made that directly lead to a goal |
| Accurate Passes | How many passes made it to the destination |
| Successful Dribbles | How many dribbles a player made without losing the ball |
| Tackles Won | How many times a player gained possession of the ball |
| Touches | How many times a player touched the ball |
| Duels Won | How many times a player won the ball in a 1v1 |
| Saves | How many times a goalkeeper blocked an on-target shot |
| Forward | Player that is positioned in the upper third of the field |
| Midfielder | Player that is positioned in the middle third of the field |
| Defender | Player that is positioned in the lower third of the field |
| Goalkeeper | Player that is positioned within the 18-yard box |
| Games Played | How many games a player played |

After all of the data was converted to numbers, it was split into a training set and a validation set. The training set would be used to train the machine learning model while the validation set would be used to determine if the model was in fact accurate. Here's what the model looked like followed by the corresponding validation and training errors:

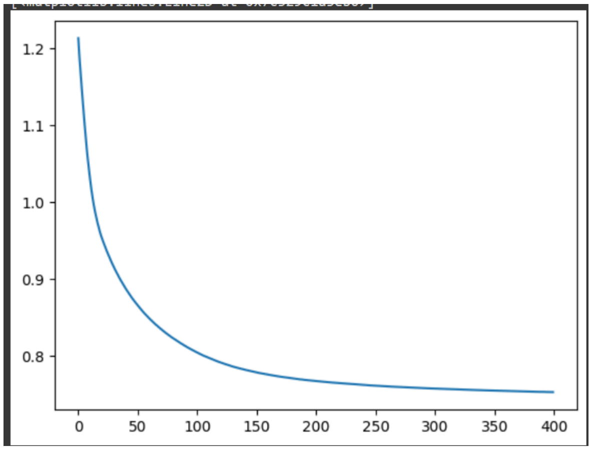
**Figure 2**
*Linear Regression Model + Errors*

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

FALL 2025

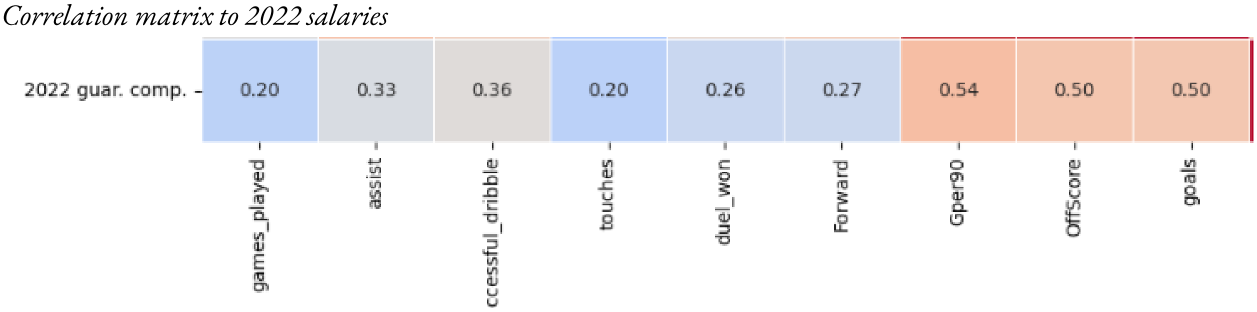



*Note.* This model produced a training accuracy of 0.75 and a validation accuracy of 0.68.

Now that the model was trained it was time to determine what remaining statistic contributed to the player's compensation the most. This was done by removing one statistic from the model at a time and monitoring the change in error that followed. A larger change in the error would mean that the model predicts better with that specific stat included. This model produced an estimated RMSE of about $741,839 which is okay given that the mean salary was around $1,165,000. However, after running the linear regression model, a ridge regression model was tested to see if this error could be improved based on feature weights and correlation. The ridge regression model improved upon the linear regression after some of the 14 statistics were omitted due to their low correlation with salary as shown through the correlation matrix. Any statistics with a correlation of less than 0.20 were omitted leaving these 9 stats (see Figure 3).

**Figure 3**

*Correlation matrix to 2022 salaries*




*Note.* Seven statistics remained out of the fourteen after the 0.20 correlation coefficient cutoff was applied (games played, assists, successful dribbles, touches, duels won, being a forward, and goals. Gper90 (number of goals per 90 minutes) and OffScore (offensive score calculated by adding goals and assists together) were introduced to bring the total to 9 different statistics.

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

FALL 2025

The Gper90 and OffScore statistics were engineered to try and find better predictors of salary. They got a correlation of 0.54 and 0.50 respectively. These features were inputted into the ridge regression model with a tuned alpha of 9 and a 5-fold cross validation. The model produced an RMSE of $647,196. This is almost 100,000 dollars better than the previous model error.

## RESULTS

**Table 3**

*Delta of Errors in Linear Regression Model*

| | games played | wins | goals | assists | accurate passes | successful dribbles | tackles won |
|---|---|---|---|---|---|---|---|
| Change in Model Error | 0.004661 | 0.006350 | 0.1091530 | -0.008330 | 0.0059911 | 0.017113 | -0.00659 |
| Importance Ranking | 6 | 3 | 1 | 14 | 5 | 2 | 13 |

| | touches | duels won | saves | forward | midfielder | defender | goalkeeper |
|---|---|---|---|---|---|---|---|
| Change in Model Error | 0.003472 | 0.003536 | 0.002172 | 0.0063234 | 0.0008420 | -0.001924 | 0.0009346 |
| Importance Ranking | 8 | 7 | 9 | 4 | 11 | 12 | 10 |

*Note.* All of the statistics were removed from the model one at a time to see the effect on the accuracy of the model. The statistics with the largest error after removal were given importance rankings close to 1 (more impactful and important for prediction).

**Table 4**

*Weights of Features in Ridge Regression Model*

| | |
|---|---|
| Games played | -20,047.90 |
| Assists | -22,646.00 |
| Successful dribbles | 11,333.49 |
| Touches | 689.59 |

| | |
|---|---|
| Duels won | -4,005.26 |
| Forward | 176,254.90 |
| Gper90 | 37,992.81 |
| OffScore | 0.00 |
| Goals | 0.00 |

Ridge regressions use feature weights to make better predictions of player salary assigning these values to each stat. Essentially, it found that for every increase in Gper90, the player salary should go up roughly $38,000. Positive weights suggest that an increase in that stat increases salary, while negative weights suggest the opposite.

According to the revised model, simply being a forward stood out as the number one contributor to a player's compensation. Next in line was Gper90. The model basically ignored OffScore and goals as they were just versions of other more important stats.

Next up are successful dribbles. Having lots of touches of the ball was almost negligible as it was very slightly positive.

Although the rest of the stats had at least 0.20 correlation coefficient with player salary in the correlation matrix, all of them were assigned negative weights in the model. This includes games played, assists, and duels won.

The ridge regression model was able to improve the initial Linear Regression Model by almost $100,000 in accuracy bringing the error in predicted salary to be $647.196.

## DISCUSSION

The two models discussed in this paper yielded slightly different results in terms of accuracy and results. The linear regression model had a $741,000 salary error margin and placed goals scored as the best predictor of salary. The ridge regression model had a salary error margin of around $650,000 and showed that being a forward was overwhelmingly the most important factor in predicting player salary. This was determined because the ridge regression model had been as forward as the highest weight magnitude with second place being just a fifth of the same weight.

The dataset used in this machine learning model was gathered during the 2022 MLS season so some of the figures and stats may not be as perfectly applicable to the present day however this model is still a decent indication of value during that season and can be used to compare to other seasons if further data is collected.

Another important note was that this season happened before the arrival of Lionel Messi, Sergio Busquets, and many other big-name players to the MLS. As big-name players come to the league, the league gains revenue incentivizing higher pay for the players.

The stats were initially analyzed using a simple linear regression model which was then improved and fine tuned as a ridge regression model to help improve the error and predictions.

Here are a few reasons or hypotheses developed as to why some statistics were deemed more or less valuable than others. Although dribbling itself does not bring team success, it was the third most valued stat. This does make sense because some of the highest-valued, and widely considered as the best players, have outstanding dribbling skills such as Lionel Messi, Neymar Jr, etc. Although Neymar was never in the MLS and Messi only joined in 2023, dribbling is generally considered highly valuable as they were the fifth and second (Knight, 2025).

As for assists producing the least weightage in both models, there are a couple of theories. A speculation here could be that the MLS clubs reward the flashy stuff like goals and dribbles/skills that help sell jerseys with that player's name on them. Overall, the level of the MLS is lower compared to the European leagues (2022 WC data showed 37 MLS players vs 133 English Premier League players) so individual flash seems to take precedence when overall team quality isn't as high (Schoinarakis, 2022). This might explain why tackles and assists, less flashy and unselfish parts of the game, are so undervalued though just as important to team success.

Future models could be trained among other leagues to see which statistics are valued and undervalued to help analyze the styles of different leagues. For example, the MLS may be all about goals but the Italian Serie A league could value possession more as they are known for team play.

Lastly, the value of the player cannot be fully captured in terms of money. There are several players who are underpaid for the impact that they make, but this was just the best way to measure value. For leagues without money, like high school soccer and club soccer, maybe value can be measured simply in the minutes or number of games played instead of that being another factor.

## REFERENCES

Bailey, M. (2023, November 7). *What Is Batting Average In Baseball: What Is A Good Number?* Baseball Bible. https://www.baseballbible.net/what-is-batting-average/

Ivanovs, V. (2023). *MLS season 2021 players stats and their salaries [Data set]*. Kaggle . https://doi.org/10.34740/KAGGLE/DSV/5078405

Pastuovic, M. (2020, December 4). *The Numbers behind Penalty Shootout in World Cup*. Northwestern Sports Analytics Group. https://sites.northwestern.edu/nusportsanalytics/2020/12/04/the-numbers-behind-penalty-shootout-in-world-cup/

Schoinarakis, A. (2022). *World Cup squad breakdown: Stats and facts about the official rosters*. Mancunian Matters. https://www.mancunianmatters.co.uk/sport/17112022-wold-cup-squad-breakdown-stats-and-facts-about-the-official-rosters/

Starling, M. (2022, October 26). *World's highest-paid football players in 2022 – in pictures*. The Week.https://theweek.com/sport/football/953916/gallery-top-ten-highest-paid-football-players