Major League Soccer Total Compensation Analysis

Brendan Zimmer

Major League Soccer (MLS) was founded in 1993 and at the time many of the players were poorly compensated compared to other sports in America. Today, in 2021, MLS players are now paid upwards of 7 Million dollars which is a massive increase since 1993. Along with the increase in pay, there is a better system for tracking players' stats in-game which also is a determining factor for how much they will be compensated for. Often players will not know what they may be paid because they are not concerned with their statistics, this way players are able to understand why they are paid. My reasoning behind this project is to provide a prediction on what players should be paid, determine what part of a player's stats are most important, and allow others to understand what they may be paid if they were in MLS.

I first started by gathering data about the players and their stats per season based on the year. I found this data which was the only available stats from 1997-2020 from https://www.kaggle.com/josephvm/major-league-soccer-dataset. I then collected compensation data about players from 2007-2018 which again was the only data I could find at https://data.world/dataremixed/mls-player-salaries-2010-2018. I then merged these two data frames into a larger one by player names and removed all missing data. I removed most of the columns and choose to only keep 13 columns which would be further reduced later in the project. This is the data I used to train my model for 2019 predictions.

Looking at **Figure 1**, I combined the player count and average compensation per year into a line graph. There are two y-axes each representing a line, with the right side representing the red line. Here we can observe that there is an obvious increase in players and compensation whether this is due to fans or inflation is past me. But there has been a 229.96% increase in compensation from 2007 to 2018 and the biggest increase of compensation from 2014 to 2015 of 26.8%. I am unsure of the reasoning behind the differences in increase, but I guarantee that it would be extremely interesting to see why. Compared to the compensation increase from 2007 to 2018 the player increase has been 87% with the largest being from 2010 to 2011 with a 32.7% increase. I decided on looking over the increase of players because I figured it would have the most change on compensation increase.

After learning more about the compensation I decided to do cluster analysis, because of the massive difference between each player. This is one of the shortcomings of the data is that there is too much variation in it with many different players. I reduced the columns to 6 categories that I determined were some of the most important variables on compensation GP (games played), MINS (minutes played), G (Goals), SOG (shots on goal), GWG (gaming winning goals), and FC (Fouls committed). Looking at **Figure 2** I chose 5 clusters and the inertia started at 4.05x10^9 and reduced to 1.4x10^9 at cluster 5. This shows that as the clusters increased there was more similarity between them. I did not find this very surprising as the number of positions and columns would create many variations between groupings. Though this did not help me with my predictions it led me to understand the variance. I saw it as a failure of this data and my organization of it and another time I would reduce the variance through reduction of columns and choosing similar positions of players.

Figure 3 shows the coefficient of my linear regression. I chose to do this after seeing the variance that my clustered KMeans analysis showed me. The maximum score I was able to achieve was .4203 using the columns stated on the graph. Looking at the graph games played had a coefficient of 3.3x10^6 which was the greatest positive weight while SOG was the most negative with -5.16x10^6. What surprised me was that assists had a greater weight than goals or SOG and those two were to most negative. It was unfortunate to see that these were most likely not similar to how players are compensated based on stats, as goals usually have the most positive weight.

In conclusion, my model was able to predict the salaries for 2019, unfortunately, I did not find them satisfactory and decided not to include them. I believe that my data was not organized while enough to produce my expected results but proved useful in showing me what data may be useful in the future. Overall, I would recommend anyone interested in playing in the MLS to play as many games as possible.

Figure 1: Average Compensation with Player Count

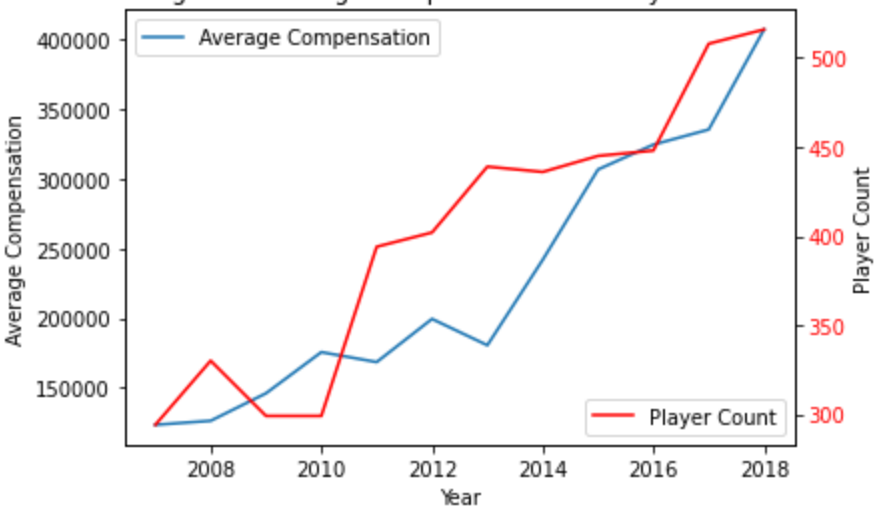

Figure 2: Cluster vs Inertia Analysis

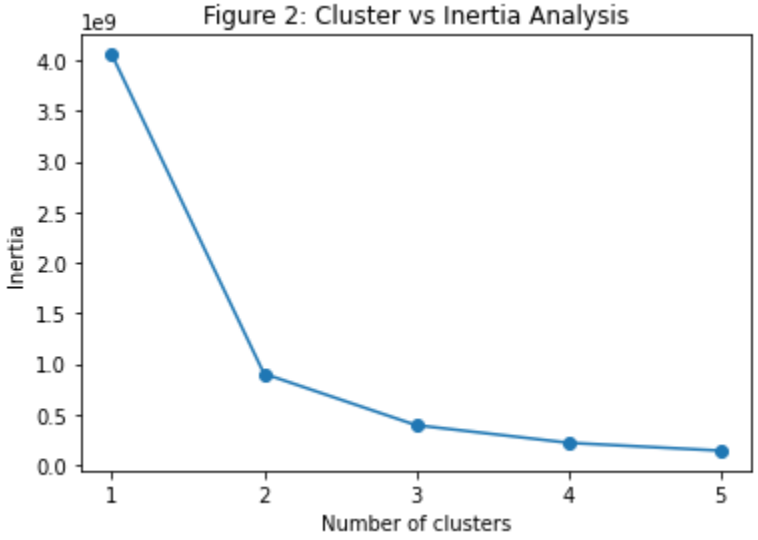

Figure 3: Logistic Regression Coefficent Weight of Features