**Section 1:**

**Executive Summary**

In a world where sports dominate society both socially and economically, the growing importance and popularity of soccer in the United States is notable. As leagues such as Major League Soccer (MLS) become more of a household name in the United States, the dollar value of each team becomes an interesting value for sports fanatics and investors to investigate. This paper and the analysis conducted aims to use multiple linear regression to predict the values of each team making up the MLS. By doing so, one would be able to observe and predict the evolution of team values of the MLS in the upcoming years.

**Hypothesis**

This paper aims to investigate the research question of what effect of season level statistics such as winning percentage or shot accuracy have on Team Value. One will expect the cumulative player values and revenue to have some significance on the Team Value, but this paper aims to answer if Team Value be partially explained by how a team performs in a given season. Our team’s hypothesis is that season level statistics will have some significance in predicting Team Value. One would assume that the higher the winning percentage of a team, the more popular they are and therefore the more likely they are to sell more tickets to games, jerseys, etc.

**Data Description**

This report sources data from Transfermarkt, the MLS soccer league website, and Forbes from 2014 - 2019. Transfermarkt is a German website that publishes football information and scores including player values and average game attendance. The MLS soccer league website is the official site for the Major League Soccer professional league in the United States. The website contains information such as matches played, number of goals a season, and fouls committed a season. Lastly, every year Forbes publishes overall team value as well as revenue statistics on every MLS team. The final dataset has 103 observations and 34 variables to consider. 34 variables will not be included in the final model, but they are important to consider in building the model as there are many statistics in soccer that could prove to be significant.

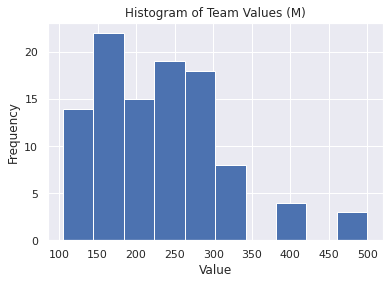
The target or Y variable used in the model is a Team Value in millions sourced from Transfermarkt. The predictor variables that will be used in regression models to be explained later on in this paper are displayed in the table on the following page.

**Table 1.** Data Set Description

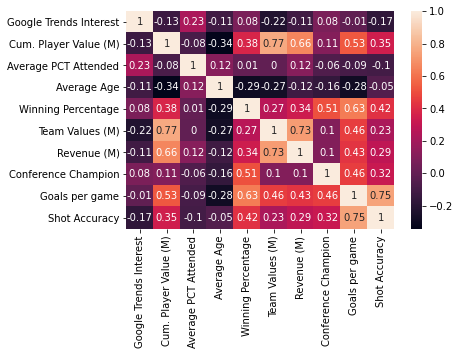
| **Variable Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Google Trends Interest | int | Determined by the relative interest or popularity of the search term, the search term being the team name, from 2004 to present. Sourced from Google Trends. |
| Cumulative Player Value | float | Sum of the individual values of each player on a team for that season. |
| Average % of Game Attendance according to Stadium Capacity | float | Average attendance divided by the stadium capacity. |
| Winning Percentage | float | Number of wins a team has for the season divided by the number of matches that season. |
| Average Age | float | Average age of players on the team. |
| Average Goals per Game | float | Total number of goals for a season divided by the number of matches. |
| Shot Accuracy | float | Number of shots for the season divided by the total number of goals a season. |
| Conference Champions | int | Dummy variable of whether a team was a Western or Eastern Conference Champion for that season. |
| Team Revenue | int | Comprises revenue streams such as ticket sales and merchandise sales for every team. |

**Exploratory Data Analysis**

With preliminary analysis of the dataset, it appears that there are some outliers in our response variable for Team Values. The mean for Team Values is $229 million. As shown in the graph to the right, there are some outliers where some team values are in the upper $300 and $400 million values. It will be important to note these outliers as we develop the model. It is possible that taking the log or square root of the variable will help create a normal distribution in the data instead of some of the skew in the histogram above.

**Figure 1:** Histogram of Team Values

In addition, after creating a heat map shown below of the variables within the dataset it can be seen that Team Value in millions is highly correlated with variables such as Revenue and Cumulative Player Variable. It would be expected to see these variables as significant within the model. Among the predictor variables within the model Shot Accuracy and Goals per Game are also very correlated. It should be avoided using both of the predictor variables within the same model to avoid issues with multicollinearity.

**Figure 2:** Variable Heat Map

**Section 2:**

**Variable Selection:**

Since the dataset contains many predictor variables to potentially include in the model, it was important to consider different methods of variable selection. More specifically, this report first used Ordinary Least Squares (OLS) Regression. OLS is a very common linear regression model, using the principle of least squares to estimate the parameters and coefficients of the variables in the model. The predictor variables that were considered significant, have a p-value of less than .05, in the OLS regression model include Google Trends Interest, Cumulative Player Value, Revenue, and Goals per Game. In addition, this report used Bayesian Information Criterion (BIC) to choose predictor variables. In linear regression models one aims to minimize BIC. In using the BIC value as a method of choosing variables, Cumulative Player Value and Revenue were the only predictor variables considered significant. Lastly, Recursive Feature Elimination (RFE) was used in selecting predictor variables. RFE considers all X variables in the data set and will recursively remove the variables with the least amount of significance. The predictor variables chosen by RFE include Cumulative Player Value, Average Age, Winning Percentage, Goals per Game, and Shot Accuracy.

Each method of variable selection chose different variables to include in the final model. Given this, our team ran multiple linear regression models including LassoCV, Ridge, and ElasticNet Regression. The reasoning behind choosing these models will be explained further in this paper. After running each model with the different predictor variables chosen by the OLS, BIC, and RFE models were analyzed using the mean squared error value. Mean squared error measures the average squared difference between the predicted values of a model and the true value. In general, there is no correct value for mean squared error but the lower the value indicates that the model’s regression line is more closely fitted to the set of points. The table to the right shows the mean squared error for each linear regression model created given the variables chosen above. It is clear that the predictor variables chosen by the OLS Regression model created linear regression models with the lowest MSE. Given this, our team further investigate the linear regression models using the predictor variables chosen by the OLS regression when considering the final model.

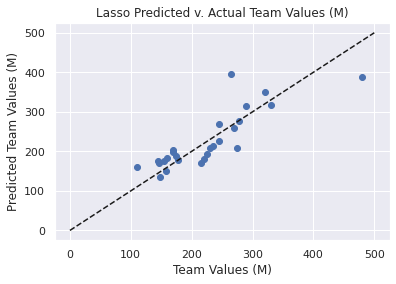
**Table 2:** Mean Squared Error of Linear Regression Models

| **MSE Values** | **OLS Regression Variables** | **BIC Variables** | **RFE Variables** |
| --- | --- | --- | --- |
| LassoCV Regression MSE | 1765.286 | 2446.303 | 2941.762 |
| Ridge Regression MSE | 1803.243 | 2413.195 | 3334.332 |
| ElasticNet Regression MSE | 1809.059 | 2865.177 | 2442.715 |

**Final Model & Assumptions:**

The final model chosen was a LassoCV model using the OLS regression variables. This was due to the fact it had the lowest mean squared error out in comparison to the other linear models including ridge regression and elastic net regression. In addition, LassoCV regression models are useful as they can aid in regularization of the model and the alpha parameter will automatically be chosen. The lasso regression models will minimize error and bias in the model as it will cause some X variable coefficients to shrink towards 0.

**Figure 3:** LassoCV Model Output



**Analysis of Fit:**

In addition to determining the mean squared error of the LassoCV model, our team also found the R2 value as well as the cross validation score of the model. It is important to consider R2 values when analyzing the fit of the model because a high R2 value indicates that the model is a good fit for the data and that the variance of the predictor variables can be explained by the model. Our team also considered the cross validation score in determining the fit of the model which in essence determines the quality or accuracy of the model.

As shown in the table below, the initial R2 and cross validation score for the model were moderate, but our team attempted to improve these scores by conducting Grid Search CV to find the optimal parameters for the LassoCV model. Grid Search CV was conducted on n\_alphas, the number of alphas on the regularization path. However, the R2 and cross validation score after a new LassoCV model was created with the optimal value for n\_alphas did not improve the model greatly as shown in the table below.

**Table 3:** Accuracy Scores of LassoCV Models

|  | **Mean Squared Error** | **R-Squared** | **Cross Validation Score** |
| --- | --- | --- | --- |
| Initial LassoCV Scores | 1765.286 | 0.705 | 0.590 |
| LassoCV Scores After Grid Search CV | 1764.690 | 0.705 | 0.590 |

Overall, the R2 value is relatively large enough to say that the model is a decent fit for the dataset and can explain a majority of the variance among the predictor variables; however, there could be other predictor variables that could be included in the model to increase this value. In addition, the cross validation score indicates that the model is fairly accurate, but certainly needs more fine tuning or perhaps our team would need to use another model to more accurately predict a team’s value.

**Coefficients & Findings:**

The coefficients for each model using the predictor variables found in the OLS regression model are shown below. While the Google Trends Interest, Cumulative Player Value, and Revenue coefficients are relatively similar, it is interesting to note the difference in the coefficient for goals per game for each model. Specifically, the LassoCV model has a zero coefficient for goals per game. This coefficient goes against our initial hypothesis that game level performance or statistics would be associated with Team Value. In addition, Cumulative Player Value is the most highly associated with Team Value with a one million increase in Cumulative Player Value being associated with a roughly four million increase in Team Value.

**Table 4:** LassoCV Model Variable Coefficients

| **Models** | **Intercept** | **Google Trends Interest** | **Cum. Player Value (M)** | **Revenue (M)** | **Goals per game** |
| --- | --- | --- | --- | --- | --- |
| LassoCV | 87.0815 | -0.075 | 4.222 | 2.388 | 0 |
| Ridge Regression | 68.095 | -0.077 | 4.331 | 2.417 | 11.013 |
| ElasticNet | 80.535 | -0.076 | 4.454 | 2.437 | 0.407 |

**Section 3:**

● Section 3: Discussion & Limitations: Idriss

As it was outlined in the last section, for our principal model and regression method, the following were decided:

- Model: Team Values (M) is identified as the Y value while the final X values are ‘Google Trends Interest’, ‘Cum. Player Value (M)’, ‘Revenue (M)’, ‘Goals per game’, ‘Shot Accuracy’. Even though ‘Shot Accuracy’ was identified as significant enough according its p-value when running the OLS regression method, the fact it has a negative correlation with the model does not make sense. Thus, for better accuracy, this variable has been removed as well.

- Regression method : the Lasso CV regression was identified as the best method so far.

The LassoCV is indeed a very good tool of prediction for the team values for the upcoming years, with an average deviation from the actual values of --.

As a reminder, the data set which was compiled for the project is based on 4 seasons from 2015 to 2019. Indeed, the 2019-2020 season was not included as a result of the current sanitary situation (covid-19) that is distorting any data.

| **Models** | **Intercept** | **Google Trends Interest** | **Cum. Player Value (M)** | **Revenue (M)** | **Goals per game** | **MSE** |
| --- | --- | --- | --- | --- | --- | --- |
| LassoCV | 81.29202638 | -0.0330604 | 3.9948736 | 2.3746304 | 0 | 1969.702 |
| Ridge Regression | 61.33782875 | -0.0365789 | 3.7702196 | 2.3623682 | 17.2988391 | 1985.643 |
| ElasticNet | 79.92092404 | -0.0334392 | 3.9639906 | 2.3812319 | 1.25522792 | 1970.933 |

The equation for our model is as follows : 81.2920263824015 (intercept) +(-0.03306037 (B1) + 3.99487361 (B2) + 2.37463042 (B3) + 0 (B4).

This equation already clashes with the initial hypothesis made that Goals Per Game were statistically significant and thus were kept. Finally, by using the best regression method, the latter shows the variable unnecessity for this study.

To assess the conformity of the regression equation with the predicted computed values (with software), let us focus on Atlanta United FC for the 2018-2019 season: the Google Trends Interest are of 292, the ‘Cum. Player Value (M)’ and ‘Revenue (M)’ are of 69.78 and 78. The predicted team value would be :

81.2920263824015 + - 0.03306037 \* 200 + 3.99487361 \* 67 + 2.37463042 \* 78 = 527.5577. Thus, the predicted value for Atlanta United is $527 million, the initial value was $500 million. The same amount was found via the software. The deviation from the original value is of 27.

To go further in our analysis and experiment the model, the New York Red Bulls value for the season 2020-2021 will be computed based on observed information from the 2018-2019 as well as current information :

- Current Google Trends interest : 200

- Cumulative players values : 27.91 (2020)

- Revenue : the same revenue as 2019 will be kept for this example

Now we integrate the values into the equation => 81.2920263824015 + -0.03306037 \* 200 + 3.99487361 \* 27.91 + 2.37463042 \* 36 = 271.6636. Hence, according to the current trends interest, cumulative plaers values and by keeping the team revenue the same as in 2019, New York Red Bulls would be valuated at $272 million. This result is below the team real valuation in 2019 (290) and its expected value (273). Indeed, even if it has a small impact (very low coefficient), Google Trends Interest have a negative impact on the final output : by increasing, it would decrease the team’s value. Another factor explaining this decrease, is the cumulative players values that is slightly below what it used to be in 2019. Nonetheless, not taking into account the error margin, the obtained results make sense and confirm the pattern of the importance of the selected variables.

It was pointed out that we decided to withdraw ‘Shot Accuracy’ from the equation as it did not make sense that the more accurate the team, the less its value would be. Afterwards, we saw that Goals Per Game also did not count. What about Google Trends Interest? The common sense would be to think that it is supposed to have a positive impact on a team’s value, not the opposite: by removing it from the equation, we see that the prediction for Atlanta United value in 2019 (500) is of 534 instead of 527 found when it was included. In the case of New York City, the expected value is of 278 instead of 273 while the real value was 290: in this case, we get better results by removing it.

Certainly, there are some hidden elements influencing this variable and how it affects the team values. An in-depth analysis of Google Trend Interests on soccer franchises should be undertaken to get a better understanding of its impact and how to interpret it.

What are the limitations of this model?

Considering the table comparing different regression methods, it appears that the ElasticNet regression method whose mean squared error of 1970 is very much close to Lasso CV’s (1969), indicates to consider goals per game for prediction. Thusly, it results difficult to gauge the relevance of this variable.

For this study, after analyzing more than 15 variables and their incidences on the Y variable, it was finally decided to only focus on 4 variables for the final model. Nevertheless, the model might have been different if additional information were available: soccer being the number one sport on the planet, a large amount of data are available on the internet. Unfortunately, since the Major League Soccer is still in ‘launch’ mode, some important information are either totally or partially missing.

As far as this analysis is concerned, the dataset could have been more complete if data on marketing, advertising, ticket sales, player trading revenues were available, for each franchise, and more importantly over the course of a significant amount of time. Furthermore, we could have incorporated the different most important costs a franchise usually has. All these variables could somehow have an incidence on the team values.

Google trends interest should also be examined thoroughly as to how illustrate it in general ways and explain its effect on teams’ values.

Another important point of thoughts for future predictions analysis would be to implement a study with variables for outstanding events, such as what is happening nowadays with the covid-19 outbreak. It is certain that such unfortunate situations can happen, and so the need to express them (quantify them) is also important in the domain of sports, not only soccer.

○ This section should include any relevant predictions and/or conclusions drawn from the model. Also, critique your own methods and provide suggestions for improving your analysis. For example, a paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project can be included.

● Section 4: Conclusion

○ In this section, you should summarize your project and highlight any final points you wish the reader to get from the project.