Miles A. Peña
DSC 550 – Data Mining
Brett Werner
Final Project
June 1, 2024

Introduction:

In Major League Soccer (MLS), player salaries are intricately tied to their individual abilities and contributions on the field. As in many professional sports leagues, player salaries in MLS are often reflective of a player's skill level, experience, and performance. Talented players who consistently demonstrate exceptional abilities, such as goal-scoring prowess, creative playmaking, or defensive solidity, often command higher salaries compared to their counterparts. This study seeks to evaluate the compensation levels of players in MLS relative to their performance during the 2021 season and their guaranteed salaries for 2022. Focusing solely on player performance in 2021, the study aims to identify instances of players potentially being either overpaid or underpaid. It will investigate whether enhanced performance correlates with projected salary increases for the subsequent season. Additionally, the study intends to analyze seasonal trends for teams to inform roster decisions for the upcoming season.

The dataset that was used for this study comprises publicly accessible information on soccer players competing in Major League Soccer during the 2021 season. It includes performance metrics such as total wins, losses, draws, goals scored, assists, accurate passes, successful tackles, and saves, among others. Additionally, the dataset provides essential player details such as name, club affiliation, compensation, and position.

Analyzing MLS players' salaries is essential for understanding the financial landscape of the league and its clubs. By examining salary distribution, along with market trends, and player valuation, clubs can make informed decisions regarding recruitment, retention, and budgeting. Understanding salary data provides clubs with a competitive edge in attracting and retaining talent while ensuring compliance with league regulations. Moreover, transparency in salary information fosters fan engagement and facilitates league governance, enabling stakeholders to participate in discussions about the financial aspects of the sport. Overall, analyzing MLS players' salaries serves as a vital tool for clubs, fans, and league administrators in managing and navigating the intricacies of professional soccer in the United States.

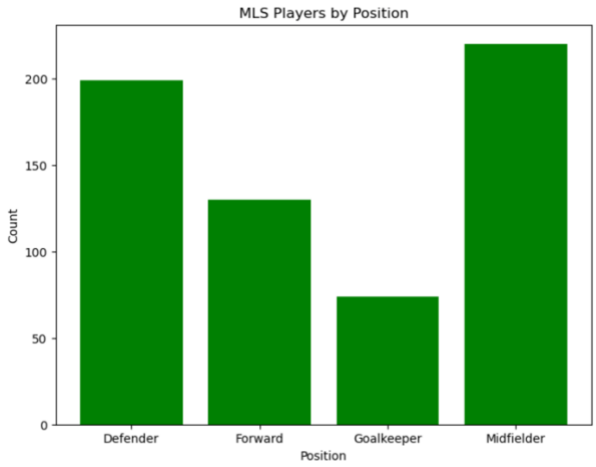
ANALYSIS OF MLS SALARIES

Milestone 1: Data Selection and EDA

To secure stakeholder buy-in for analyzing MLS players' salaries, we propose an initial milestone focused on creating graphical analyses of the dataset. This milestone serves a dual purpose. First, it initiates the exploration of the dataset, enabling us to uncover insights and trends relevant to the questions posed in the introduction. Secondly, by employing visualizations, we aim to communicate complex data in an accessible and engaging manner, facilitating a deeper understanding of the underlying trends and patterns.

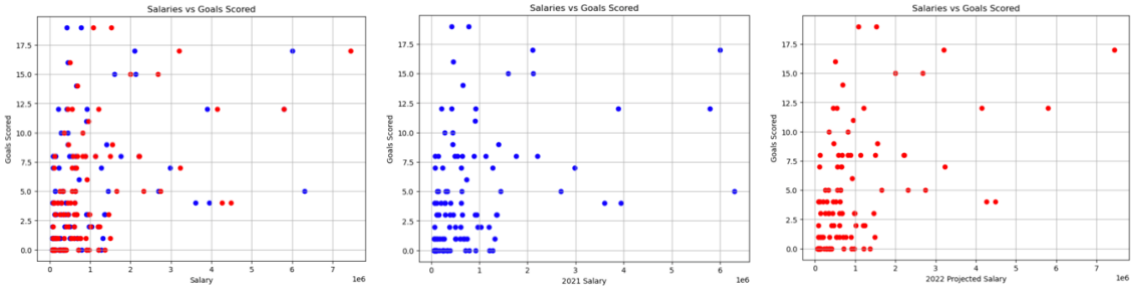
Graph Analyses:

Graph 1: I began with a straightforward visualization, aiming to illustrate the distribution of player positions in Major League Soccer (MLS). The graph reveals a notable disparity in representation among positions, with goalkeepers being the least represented. This observation aligns with the inherent nature of soccer, where each team typically fields only one goalkeeper at a time.

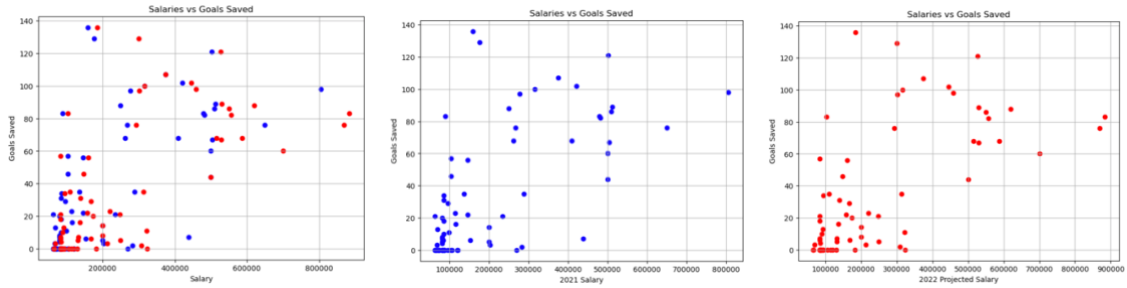


Graphs 2-4: The analysis commenced by narrowing down the dataset to include only players occupying the forward position. This subset focused on a key performance metric associated with forwards: goals scored. Graph 3 illustrates both the 2021 salaries and the projected 2022 salaries on a single plot, aiming to identify any noteworthy shifts. It is worth noting that there were a few exceptional cases that stood out from the rest, mostly involving players who had either international experience or recognition. Such players often command higher salaries due to their marketability and ability to attract fans and sponsors which further explain the large salary gap. Upon closer inspection, it became evident that the top 10 forwards with the highest goal tallies experienced salary increases from 2021 to their projected salaries for 2022. Graph 3 delves into the same data but

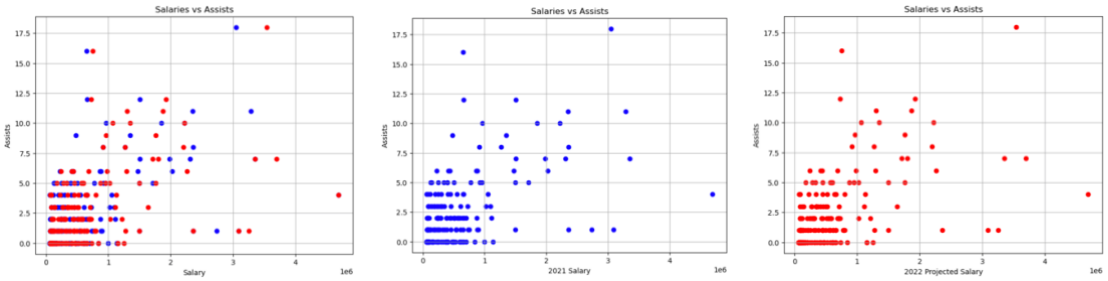
exclusively for the 2021 season, while Graph 4 provides insights into projected salaries for the 2022 season.



Graphs 5-7: The analysis of goalkeeper salaries requires further refinement, given the multitude of metrics that can be employed to gauge performance beyond saves made. However, for the present investigation, we're focusing on salaries in relation to the number of saves made. Graph 5 depicts both the 2021 salaries and the projected 2022 salaries on a unified plot. While the graph indicates some salaries projected to rise in 2022, a detailed examination of the top 10 goalkeepers ranked by saves made reveals no definitive trend correlating these metrics. Surprisingly, there are instances where salaries are anticipated to remain constant, and in the case of Brad Guzan, one goalkeeper's projected salary for 2022 is lower. Similar to the previous analysis, Graph 6 isolates data exclusively for the 2021 season, while Graph 7 provides insights into projected salaries for the 2022 season.



Graphs 8-10: Similarly, the analysis for midfielders focuses solely on assists made for this study, though it's important to acknowledge the availability of other metrics that can be used to measure performance for this position. Graph 8 presents both the 2021 salaries and the projected 2022 salaries on a consolidated plot. While the graph highlights projected salary increases, the data also reveals three midfielders (Maxi Moralez, Albert Rusnak, and Mauricio Pereyra) anticipating a decrease in pay. Graph 9 isolates data exclusively for the 2021 season, while Graph 10 offers insights into projected salaries for the 2022 season.



Graphs 11-13: The chosen performance metric for assessing defenders in the MLS was goals conceded in the box. Initially, this metric may not appear to be the most definitive measure for making determinations since the scatterplot fails to provide a clear depiction of whether salary increases for defenders correlate with the number of goals they concede in the box. However, upon scrutinizing the data sorted by the least number of goals conceded in the box, a trend emerges. Among the top 10 defenders with the lowest number of goals conceded in the box, most are expected to receive a pay increase in 2022, except for Ben Sweat of Austin FC who is expecting a decrease in pay. As with previous analyses, Graph 9 focuses exclusively on data from the 2021 season, while Graph 10 sheds light on projected salaries for the 2022 season.

Based on the insights gleaned from the graphical analysis of the selected performance metrics, a direct correlation between performance and salary increases in Major League Soccer (MLS) does not seem evident. It appears that individual players' salary adjustments are influenced by factors beyond performance alone, potentially including factors such as popularity or marketability. While there are instances where performance does appear to impact salary adjustments, it does not appear to be the sole determining factor, as anticipated. Further in-depth analysis is necessary to explore whether other performance metrics may offer better correlation, if any, and to investigate the allocation of resources by a particular team concerning player salaries and performance.

Milestone 2: Data Preparation

The data preparation milestone involved a comprehensive six-step process aimed at enhancing the dataset's cleanliness and readability. Initially, we began by standardizing the column headers throughout the dataset, rectifying numerous capitalization inconsistencies for uniformity. Subsequently, we replaced whitespace in column names with underscores to adhere to Python conventions, thereby enhancing readability and compatibility. Notably, three features were deemed non-essential and were consequently removed from the dataset. "club_abb" was eliminated due to redundancy, as club affiliation was adequately represented elsewhere. Similarly, "birth_city" was discarded as it lacked substantive relevance for our modeling objectives, while "foot" was omitted since foot dominance exhibited negligible influence on salary variations, our primary focus. Following this, feature selection was conducted on the MLS dataset, identifying the top 10 features exhibiting the most substantial absolute correlation with the '2022_Guaranteed_Compensation' column. These selected features were stored for further analysis. To augment our dataset, we engineered a new feature - Body Mass Index (BMI), as a general indicator of players' physical fitness and body composition, crucial for on-field performance. Although BMI alone does not dictate soccer prowess, it contributes to factors influencing player evaluation and performance optimization. Additionally, in addressing missing data, we opted to denote such instances as "Unknown," ensuring clarity and dataset continuity without necessitating the removal of entire columns or rows. Lastly, dummy variables were created to deal with categorical columns. Overall, while the dataset primarily exhibited cleanliness, these refinements and adjustments were essential to optimizing its utility for subsequent analyses.

Milestone 3: Model Building and Evaluation

The third and final milestone comprised three distinct steps, each crucial for model development. Initially, the data was partitioned into training and test sets. To achieve this, we utilized a 80-20 split, allocating 80% of the data for training and reserving the remaining 20% for testing. This division ensures a robust evaluation of the model's performance on unseen data. Subsequently, we examined the shapes of the resultant datasets to ascertain their dimensions:

```
X_train shape: (498, 826)
X_test shape: (125, 826)
y_train shape: (498,)
y_test shape: (125,)
```

In the subsequent step, we generated a correlation matrix to discern the relationships between features within the training dataset. This involved computing the absolute correlations among all pairs of features. Subsequently, we introduced a new column named "pairs" to the dataset, comprising tuples of the variable names, facilitating a clear identification of correlated pairs. The data frame index was then set to the "pairs" column, and the correlation column was renamed as "correlation" for clarity. To eliminate redundant correlations, pairs where both variables were identical (having a correlation of 1 which indicates self-correlation) were excluded. Finally, we isolated pairs exhibiting correlations above 75% and below 95%, yielding the following results:

pairs	level_0	level_1	correlation
(Ground_Duel, Duel_Lost)	Ground_Duel	Duel_Lost	0.949012
(Duel_Lost, Ground_Duel)	Duel_Lost	Ground_Duel	0.949012
(Saves, Good_Claim)	Saves	Good_Claim	0.947999
(Good_Claim, Saves)	Good_Claim	Saves	0.947999
(Accurate_Goal_Kicks, Good_Claim)	Accurate_Goal_Kicks	Good_Claim	0.946048
...	...	...	...
(Interception, Outfielder_Block)	Interception	Outfielder_Block	0.750469
(Times_Tackled, Possession_Lost)	Times_Tackled	Possession_Lost	0.750250
(Possession_Lost, Times_Tackled)	Possession_Lost	Times_Tackled	0.750250
(Total_Shots_at_Goal, Ground_Duel_Lost)	Total_Shots_at_Goal	Ground_Duel_Lost	0.750240
(Ground_Duel_Lost, Total_Shots_at_Goal)	Ground_Duel_Lost	Total_Shots_at_Goal	0.750240

898 rows × 3 columns

Ultimately, while we were able to find many pairs that were correlated with one another, it does not appear that any of them have a direct correlation with the target variable ("2022_Guaranteed_Compensation"). Based on this dataset, it does not seem that the statistics for each player have much to do with their next year compensation. We would need more information to make any determination.

At the culmination of this milestone, we delved into model construction, aiming to pinpoint the most suitable fit among a variety of models. Below, we present a thorough overview of the results obtained from this endeavor:

Linear Regression Evaluation:

Training Set Results:

- Mean Squared Error: `1.6429678971300458e-10` which is extremely close to zero. This indicates that the model fits the training data almost perfectly.
- R^2 Score: `1.0` which is the maximum possible value, indicating a perfect fit.

Testing Set Results:

- Mean Squared Error: `113140582109.82799` which is significantly large. This indicates that the model's predictions on the testing set have substantial errors.
- R^2 Score: `0.8858145875357414` which is quite high, suggesting that the model explains a good portion of the variance in the testing data, but not as perfectly as it does in the training data.

Conclusion for Linear Regression:

The perfect scores on the training set (MSE close to zero and R2 = 1.0) suggest that the model fits the training data extremely well. However, this might indicate overfitting, where the model has learned the training data too well, including its noise and outliers. The discrepancy between the training set and testing set results (particularly the high MSE on the testing set) further supports the possibility of overfitting. Although the R2 score on the testing set is still high, the large MSE indicates that the model's predictions are not as accurate for unseen data.

Decision Tree Regressor Evaluation:

Mean Squared Error:

- Value: `333710816178.544`
- This is a large value, indicating that the model's predictions have substantial errors when compared to the actual values in the testing set. A high MSE suggests that the model's performance is not very accurate.

R^2 Score:

- Value: `0.6632074320411205`
- This indicates that approximately 66.32% of the variance in the dependent variable (2022_Guaranteed_Compensation) is explained by the model. While this is a reasonable proportion, it also indicates that around 33.68% of the variance is not explained by the model.

Conclusion for Decision Tree Regressor:

The decision tree model shows moderate performance, explaining about 66% of the variance in the target variable but with a relatively high mean squared error. The results suggest that the model is somewhat effective but may benefit from improvements.

Random Forest Regressor Evaluation:

Mean Squared Error (MSE):

- Value: `228301915168.71487`
- This is a large value, indicating that the model's predictions have significant errors when compared to the actual values in the testing set. A high MSE suggests that the model's performance has room for improvement.

R^2 Score:

- Value: `0.7695897628967969`
- This indicates that approximately 76.96% of the variance in the dependent variable (2022_Guaranteed_Compensation) is explained by the model. This is a good proportion, suggesting that the model captures the underlying patterns in the data fairly well.

Conclusion for Random Forest Regressor:

The random forest model shows reasonable performance, explaining about 77% of the variance in the target variable, but with a relatively high mean squared error. The results suggest that the model is effective but may still benefit from improvements.

Lasso Regression Evaluation:

Best Alpha:

The cross-validation process determined the best alpha value to be `13303053405.370127`. This value is relatively high, indicating a strong regularization effect.

Training Set Evaluation:

- Mean Squared Error: `64845217965.63188`
- R^2 Score: `0.8311771696199165` The high R^2 score indicates that the model explains 83.12% of the variance in the training data, which is quite good. However, the MSE indicates that there is still some error in the predictions.

Testing Set Evaluation:

- Mean Squared Error: `73199550324.00851`
- R^2 Score: `0.9261244666583776` The R^2 score of 92.61% for the testing set suggests that the model generalizes well to unseen data, explaining a high proportion of the variance. The MSE for the testing set is slightly higher than for the training set, but the difference is not very large, indicating a good fit.

Conclusion for Lasso Regression:

Model Performance: The Lasso regression model performs well, as indicated by the high R^2 scores on both the training and testing sets. The model is able to explain a significant proportion of the variance in the target variable.

Regularization: The high alpha value indicates that a strong regularization was needed to avoid overfitting, which can be common with high-dimensional data.

Error Metrics: While the MSE values are relatively high, the R^2 scores suggest that the model predictions are still closely aligned with the actual values.

This evaluation shows that the Lasso regression model is effective in handling the given dataset, balancing between bias and variance through the selected regularization strength.

Lasso Regression w/Hyperparameter Tuning Results and Evaluation:

```
Best alpha: 1
Lasso Regression - Training set - Mean Squared Error: 209183.01955822468
Lasso Regression - Training set - R^2 Score: 0.9999994553974751
Lasso Regression - Testing set - Mean Squared Error: 46730280590.2103
Lasso Regression - Testing set - R^2 Score: 0.9528381747357105
```

Based on the results obtained from the Lasso regression model with hyperparameter tuning, with the best alpha value of 1, the evaluation of the model's performance on both the training and testing datasets reveals the following:

- The mean squared error (MSE) for the training set is remarkably low, indicating a close fit between the actual and predicted values. The MSE value of approximately 209,183 suggests that, on average, the squared difference between the actual and predicted target variables is quite small.
- The coefficient of determination (R^2) score for the training set is extremely high, nearly equal to 1. This indicates that the model explains almost all the variability in the target variable, demonstrating an excellent fit to the training data.

However, when evaluating the model on the testing dataset:

- The MSE for the testing set is substantially higher compared to the training set, indicating that the model's performance degrades when applied to unseen data. The MSE value of approximately 46,730,280,590 suggests that the model's predictions deviate significantly from the actual values in the testing dataset.
- Despite the high R^2 score for the training set, the R^2 score for the testing set is considerably lower. While the R^2 score of approximately 0.953 indicates that the model explains a significant portion of the variability in the testing data, it is notably lower than the score achieved on the training data, suggesting that the model may be overfitting to the training data and unable to generalize well to unseen data.

Conclusion for Lasso Regression w/Hyperparameter Tuning:

In conclusion, while the Lasso regression model with hyperparameter tuning achieves exceptional performance on the training dataset, demonstrating a close fit and high explanatory power, its performance on the testing dataset is suboptimal. The substantial disparity between the MSE and R^2 scores of the training and testing datasets suggests overfitting, highlighting the need for further model refinement and possibly the exploration of alternative modeling approaches to improve generalization performance.

Project Conclusion:

The analysis provides a comprehensive evaluation of several machine learning models, focusing primarily on a Lasso regression model with hyperparameter tuning. Below is a comprehensive breakdown of the insights derived from the analysis:

- **Overfitting Concerns:** The evaluation highlights a common issue in machine learning, overfitting. The models, particularly the Lasso regression model, perform exceptionally well on the training dataset, with near-perfect scores. However, this performance doesn't translate well to unseen data, as indicated by significantly higher MSE and lower R^2 scores on the testing dataset. This discrepancy suggests that the model has learned the training data too well, capturing its noise and outliers, but fails to generalize to new data.

- **Model Performance:** Despite the overfitting concerns, the Lasso regression model demonstrates strong performance on the training dataset, with low MSE and high R^2 scores. This indicates that the model can effectively explain the variability in the target variable within the training data.
- **Generalization Challenges:** The performance degradation on the testing dataset suggests that the model struggles to generalize its learned patterns to unseen data. This could be due to the model being too complex or capturing noise from the training data, leading to poor performance on new observations.

The Lasso regression model displays potential in identifying underlying data patterns, yet its susceptibility to overfitting poses a significant challenge in predicting unseen data accurately. Addressing this issue is crucial for deploying a reliable and robust predictive model; until this is resolved, the model will not be ready to deploy. The analysis suggests the necessity for refining the model further and investigating alternative modeling strategies to mitigate overfitting and enhance generalization performance. This might involve techniques such as feature selection, feature engineering, or trying different algorithms that may be more resilient to overfitting.