Predicting Offensive Goals Added Per 90 Minutes for MLS Players

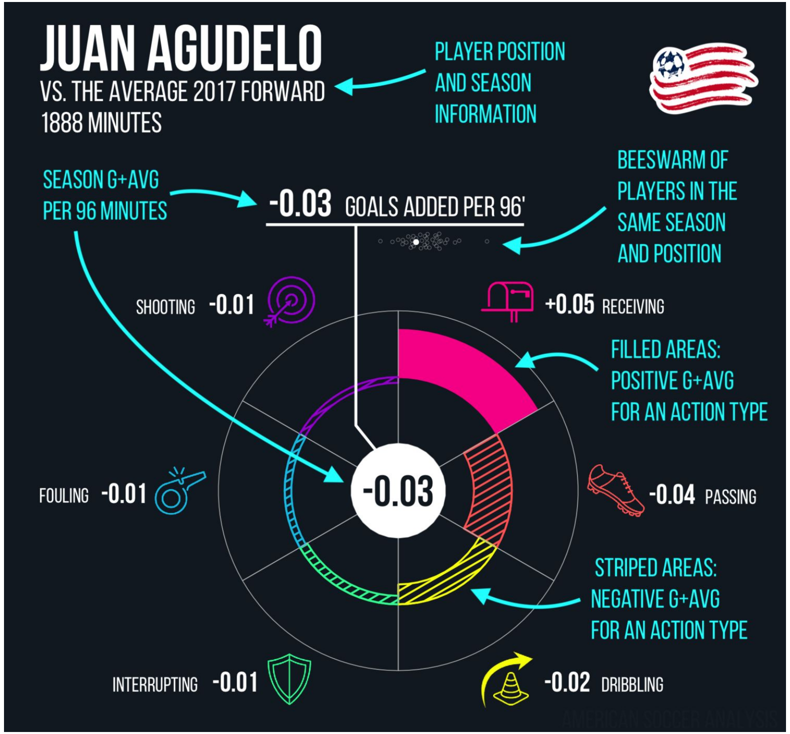
Alexander Booth

5/17/2024

Introduction

The objective of this project is to leverage machine learning to predict the offensive goals added per 90 minutes for MLS players in the 2023 season, based on their performance in prior seasons.

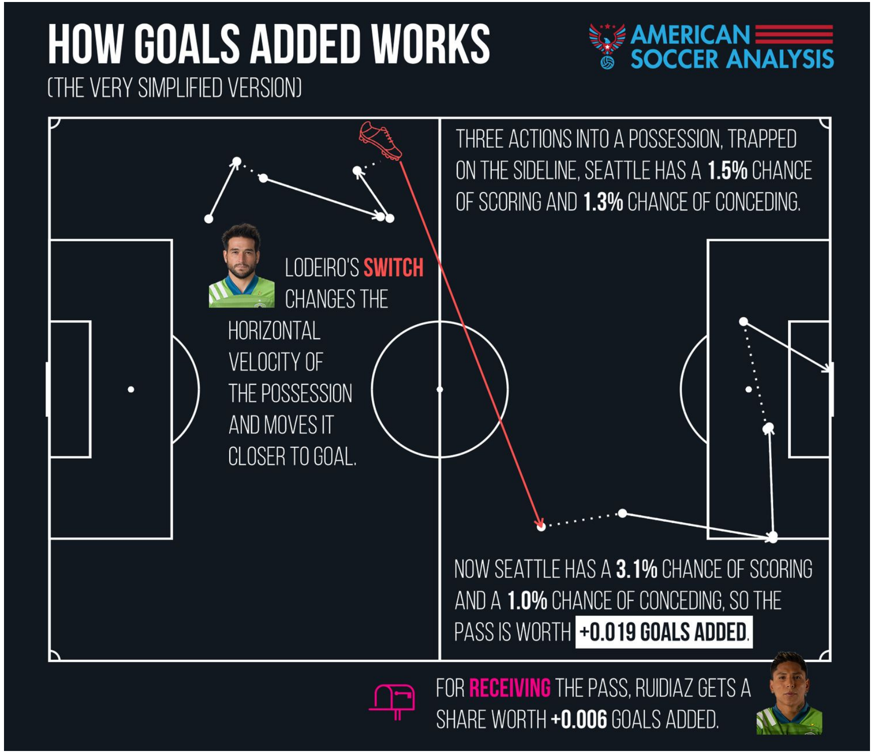
Accurate predictions can help teams optimize player performance, make informed decisions on player acquisitions, and strategize game plans.



Introduction

Goals added (g+) measures a player's total on-ball contribution in attack and defense. It does this by calculating how much each touch changes their team's chances of scoring and conceding across two possessions.

Source: https://www.americansocceranalysis.com/what-are-goals-added



HOW GOALS ADDED WORKS

(THE VERY SIMPLIFIED VERSION)

THREE ACTIONS INTO A POSSESSION, TRAPPED ON THE SIDELINE, SEATTLE HAS A **1.5%** CHANCE OF SCORING AND **1.3%** CHANCE OF CONCEDING.

LODEIRO'S **SWITCH** CHANGES THE HORIZONTAL VELOCITY OF THE POSSESSION AND MOVES IT CLOSER TO GOAL.

NOW SEATTLE HAS A **3.1%** CHANCE OF SCORING AND A **1.0%** CHANCE OF CONCEDING, SO THE PASS IS WORTH **+0.019 GOALS ADDED**.

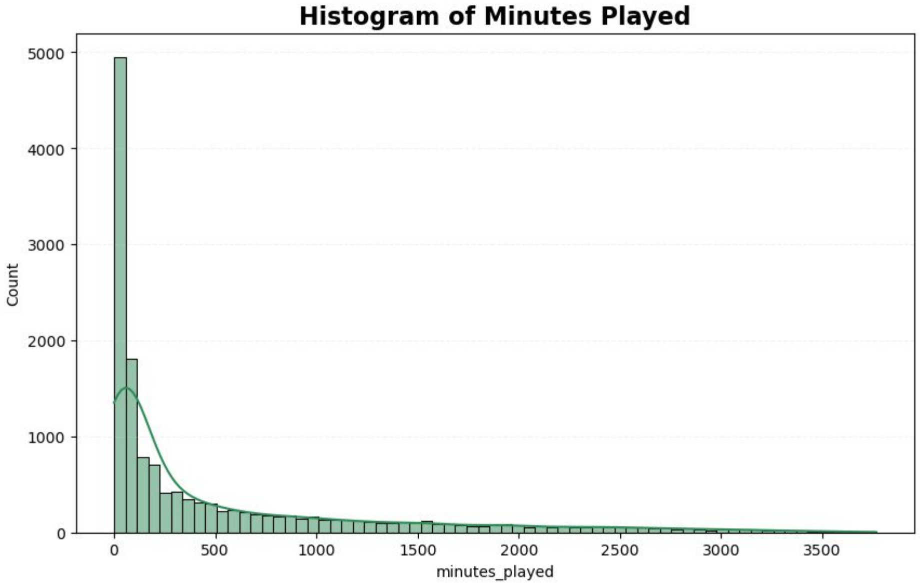
FOR **RECEIVING** THE PASS, RUIDIAZ GETS A SHARE WORTH **+0.006** GOALS ADDED.

Data Overview

Data sources include official MLS statistics and historical performance data from https://app.americansocceranalysis.com.

Key variables include:

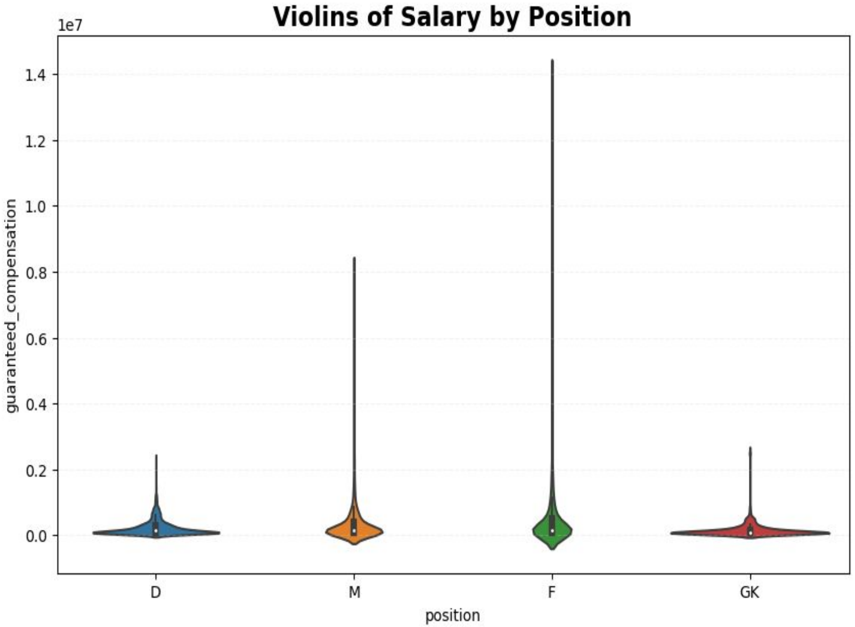
- Minutes Played
- Primary Position
- Passes Attempted, Expected, and Completed per Game
- Salary
- Age



Data Preprocessing

Data cleaning was performed to prepare the data for modelling.

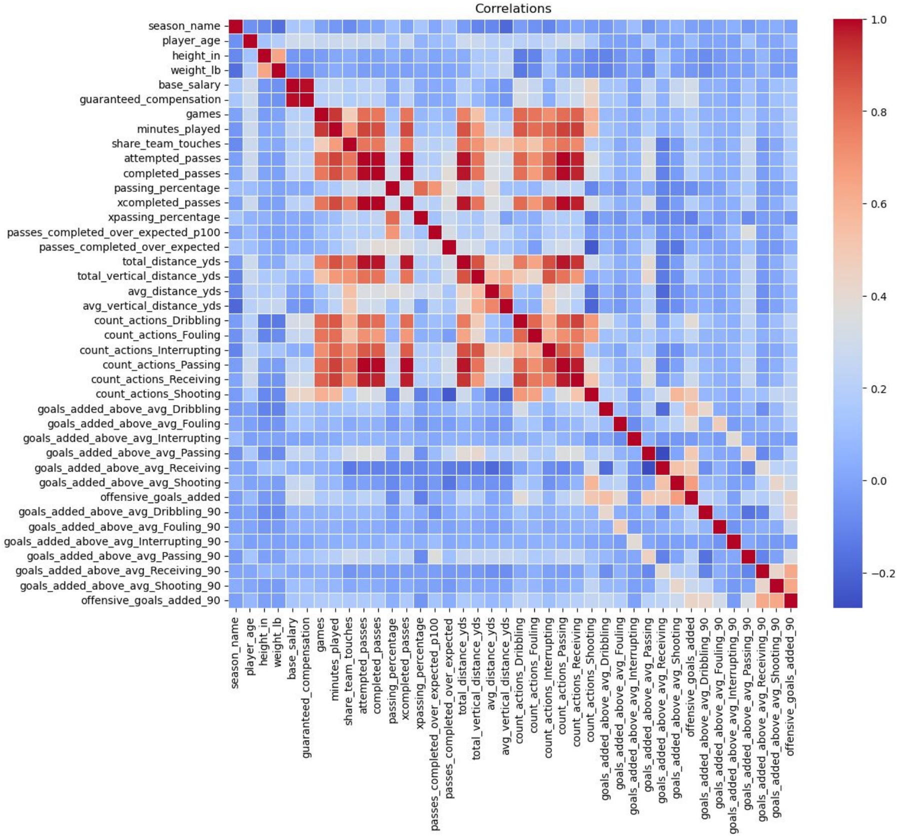
- Aggregated G+ counting stats by player/season and found G+ per 90
- Calculated Age and Height for each player
- Calculated season passing statistics per player
- Handled missing values by imputation using median values for salary, average values for age, height and weight, and the most common (USA) category for nationality
- Merged all disparate datasets together



Data Preprocessing

Feature engineering was performed to calculate historical statistics. Based on the correlation analysis, certain features were selected for modeling.

For a given projection year, e.g. 2022, all feature data for that player from 3, 2, & 1 years prior, e.g. 2019-2021, was pulled. A player making their debut in the target projection year was not considered or included in this training set.

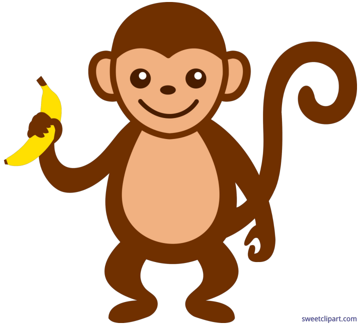


Marcel the Monkey

Marcel Projections are a basic forecasting system for counting stats. It typically uses 3 years of season data, with the most recent data weighted heavier. It regresses towards the mean. And it has an age factor. As a baseline, they are typically pretty good and are fairly reliable compared to more complicated systems.

A Marcel projection provides a good comparison for any machine learning systems, and are typically more explainable.

More information: https://tangotiger.net/archives/stud0346.shtml



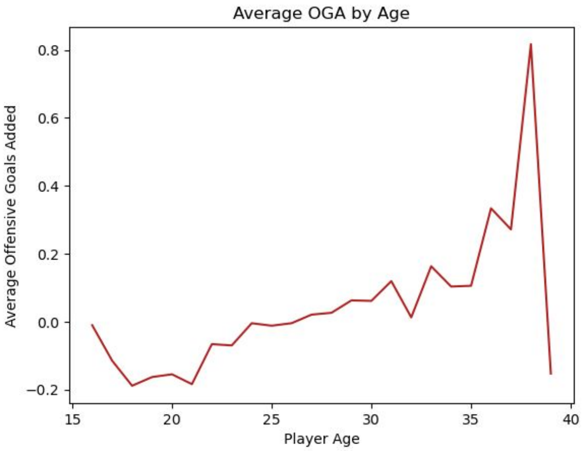
sweetclipart.com

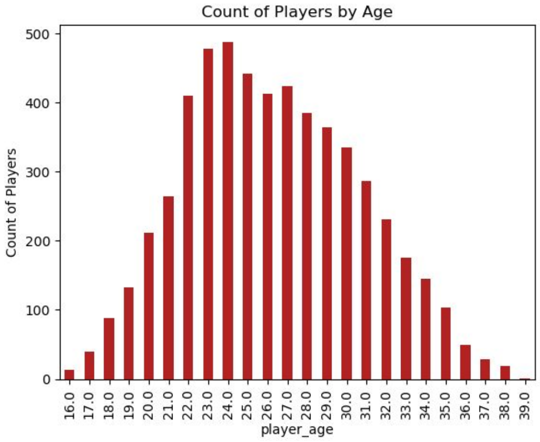
Aging Curves

Offensive Goals Added increases YoY, however, once players reach their 30s, survival bias kicks in. Only the best players continue to perform.

It is interesting that there is no obvious macro-decrease in performance as players age, it is a continual increase. For future work, it would be worth breaking down by position.

This information was used to create a custom age adjustment for this dataset.





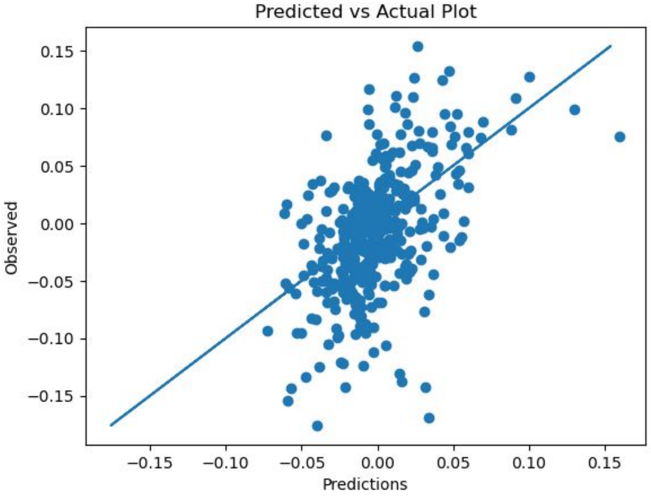
Marcel Results

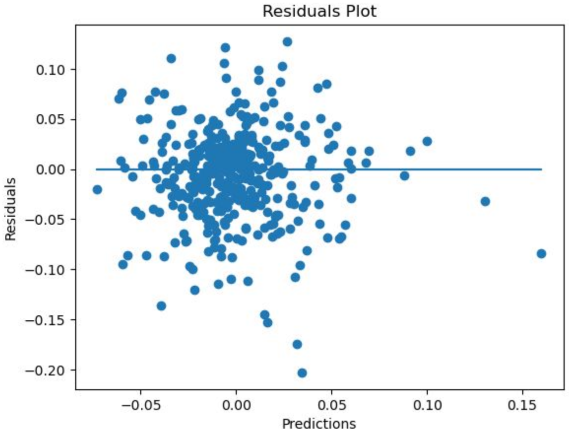
As a validation set, all players with over 500 minutes in 2022 (395) were examined. OGA_90 was calculated by dividing the projected OGA by the projected Minutes Played and multiplying by 90.

R2: 0.241

MAE: 0.0340

RMSE: 0.0458

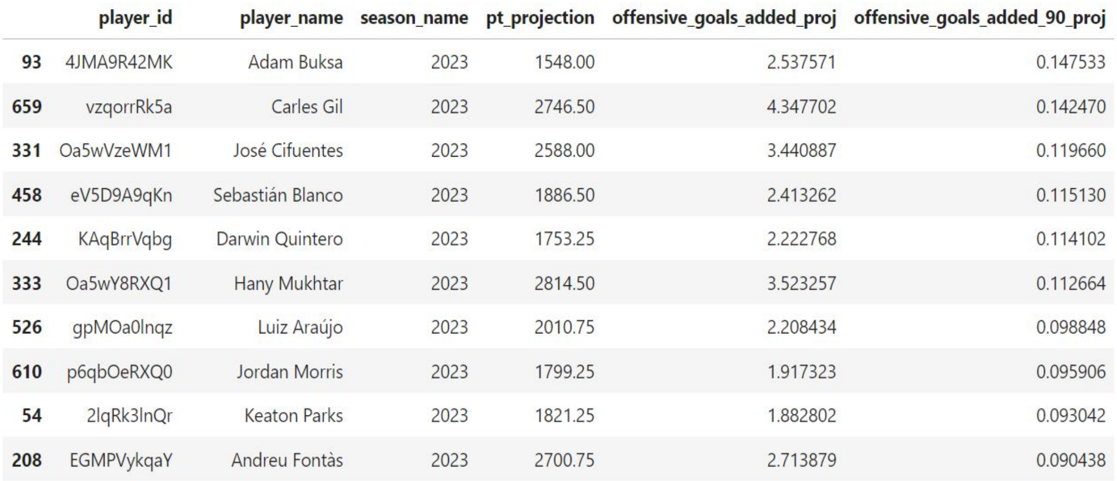




Marcel Results - 2023

Limitations: Heavy reliance on past performance. Tough to accurately project playing time. Weights are somewhat arbitrary.

- Adam Buksa was transferred to Ligue 1 for $$$
- Carles Gil was an All-Star and scored 11 goals in 2023
- Hany Mukhtar was 2022 MVP and scored 15 goals in 2023
- José Cifuentes was transferred to Scotland Premier League



	player_id	player_name	season_name	pt_projection	offensive_goals_added_proj	offensive_goals_added_90_proj
93	4JMA9R42MK	Adam Buksa	2023	1548.00	2.537571	0.147533
659	vzqorrRk5a	Carles Gil	2023	2746.50	4.347702	0.142470
331	Oa5wVzeWM1	José Cifuentes	2023	2588.00	3.440887	0.119660
458	eV5D9A9qKn	Sebastián Blanco	2023	1886.50	2.413262	0.115130
244	KAqBrrVqbg	Darwin Quintero	2023	1753.25	2.222768	0.114102
333	Oa5wY8RXQ1	Hany Mukhtar	2023	2814.50	3.523257	0.112664
526	gpMOa0lnqz	Luiz Araújo	2023	2010.75	2.208434	0.098848
610	p6qbOeRXQ0	Jordan Morris	2023	1799.25	1.917323	0.095906
54	2lqRk3lnQr	Keaton Parks	2023	1821.25	1.882802	0.093042
208	EGMPVykqaY	Andreu Fontàs	2023	2700.75	2.713879	0.090438

Model Training

Two additional approaches were taken to predict OGA_90 with Machine Learning.

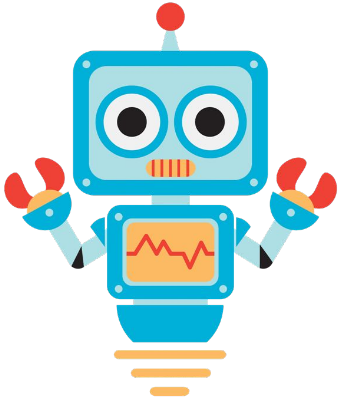
First, separate models were used to predict Minutes Played and OGA, respectively. These predictions were then combined into OGA_90.

Second, a model was trained exclusively on OGA_90.

Initially, 2022 was withheld to act as a validation set.

Cross-validation was used to ensure robust performance estimates. Key metrics included Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Considered several models: linear regression, random forests, gradient boosting. A StandardScaler was used for the linear models.

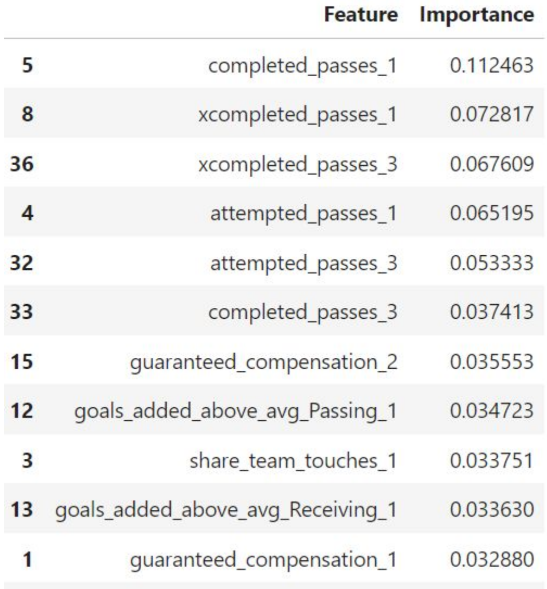


Model Selection

For all three models, a Ridge regression was selected due to its superior cross-validated MAE and RMSE on all the target variables. While the boosted tree models had expected feature importances, they suffered from overfitting.

Interestingly, completed, expected, and attempted passes in the previous years were the top features for OGA and OGA_90, in addition to salary and share of team touches.

For minutes played, past minutes played as well as compensation, were top features. Teams want to play their popular and well-paid players!



	Feature	Importance
5	completed_passes_1	0.112463
8	xcompleted_passes_1	0.072817
36	xcompleted_passes_3	0.067609
4	attempted_passes_1	0.065195
32	attempted_passes_3	0.053333
33	completed_passes_3	0.037413
15	guaranteed_compensation_2	0.035553
12	goals_added_above_avg_Passing_1	0.034723
3	share_team_touches_1	0.033751
13	goals_added_above_avg_Receiving_1	0.033630
1	guaranteed_compensation_1	0.032880

Model Performance

As a validation set, all players with over 500 minutes in 2022 (395) were examined. OGA_90 was calculated by dividing the projected OGA by the projected Minutes Played and multiplying by 90 for the first model. The second model predicted OGA_90 directly.

The second model outperformed the Marcel projections in all metrics.

Combined Minutes/OGA Model

R2: 0.111

MAE: 0.0352

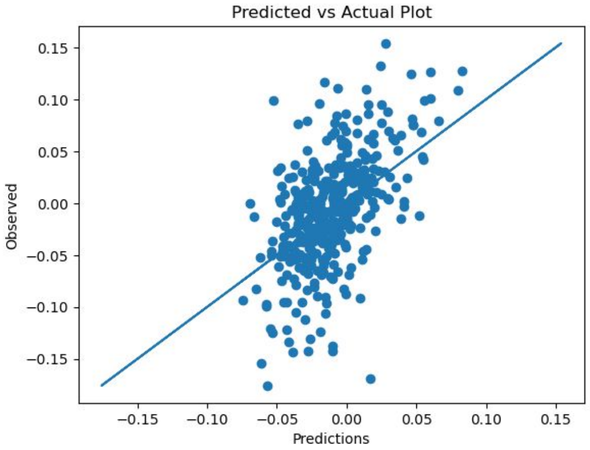
RMSE: 0.0495

OGA_90 Model

R2: 0.278

MAE: 0.0336

RMSE: 0.0446

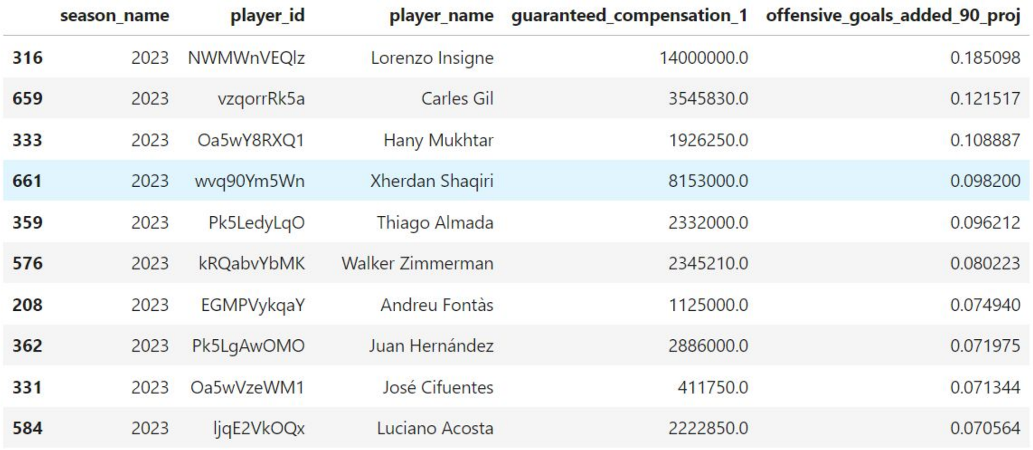


Predictions for 2023

The OGA_90 model was retrained on all data, including 2022. Then, 2023 projections were generated for all players who had at least 1 minute played in 2022, as seen below.

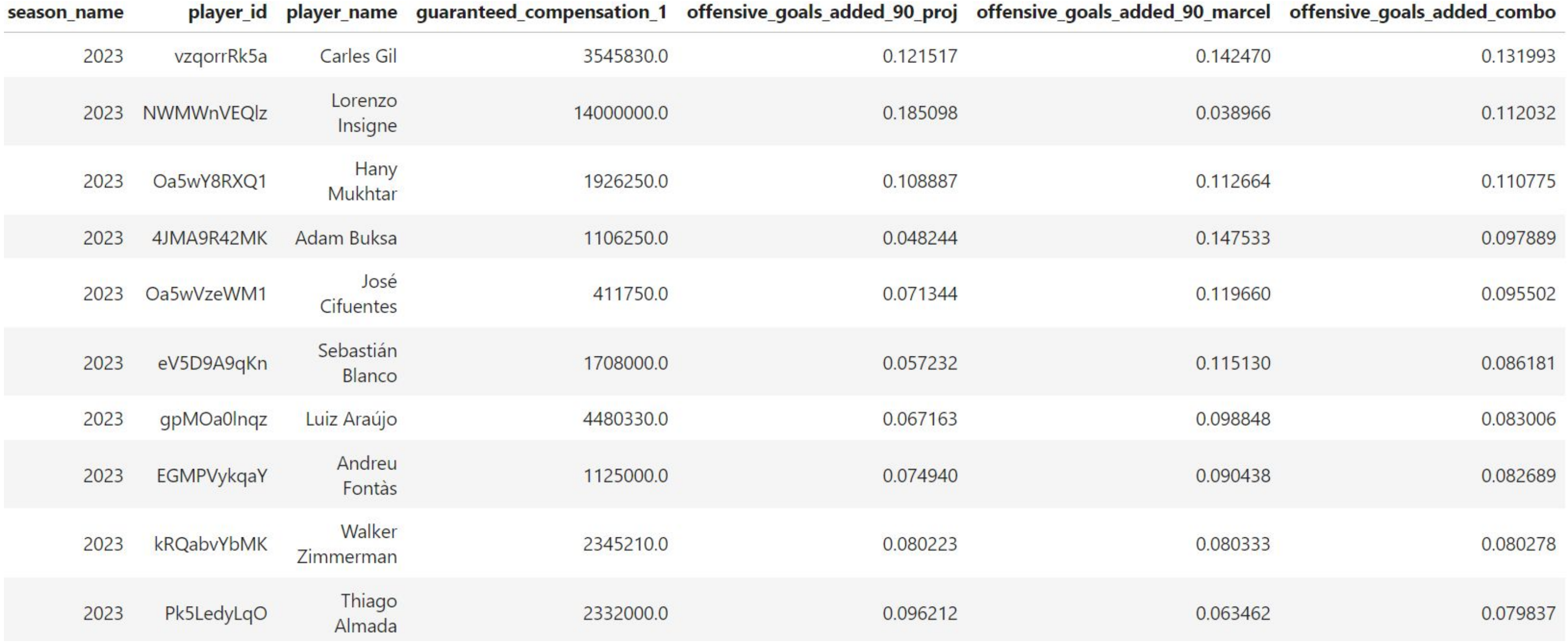
However, it was noted that these model predictions seemed to rely heavily on salary as well as skewed slightly high.

Since all projection systems have limitations, a combination of multiple typically has a stronger prediction power than one alone. Final projections averaged both the Model + Marcel outputs.



	season_name	player_id	player_name	guaranteed_compensation_1	offensive_goals_added_90_proj
316	2023	NWMWnVEQlz	Lorenzo Insigne	14000000.0	0.185098
659	2023	vzqorrRk5a	Carles Gil	3545830.0	0.121517
333	2023	Oa5wY8RXQ1	Hany Mukhtar	1926250.0	0.108887
661	2023	wvq90Ym5Wn	Xherdan Shaqiri	8153000.0	0.098200
359	2023	Pk5LedyLqO	Thiago Almada	2332000.0	0.096212
576	2023	kRQabvYbMK	Walker Zimmerman	2345210.0	0.080223
208	2023	EGMPVykqaY	Andreu Fontàs	1125000.0	0.074940
362	2023	Pk5LgAwOMO	Juan Hernández	2886000.0	0.071975
331	2023	Oa5wVzeWM1	José Cifuentes	411750.0	0.071344
584	2023	ljqE2VkOQx	Luciano Acosta	2222850.0	0.070564

Predictions for 2023



season_name	player_id	player_name	guaranteed_compensation_1	offensive_goals_added_90_proj	offensive_goals_added_90_marcel	offensive_goals_added_combo
2023	vzqorrRk5a	Carles Gil	3545830.0	0.121517	0.142470	0.131993
2023	NWMWnVEQlz	Lorenzo Insigne	14000000.0	0.185098	0.038966	0.112032
2023	Oa5wY8RXQ1	Hany Mukhtar	1926250.0	0.108887	0.112664	0.110775
2023	4JMA9R42MK	Adam Buksa	1106250.0	0.048244	0.147533	0.097889
2023	Oa5wVzeWM1	José Cifuentes	411750.0	0.071344	0.119660	0.095502
2023	eV5D9A9qKn	Sebastián Blanco	1708000.0	0.057232	0.115130	0.086181
2023	gpMOa0lnqz	Luiz Araújo	4480330.0	0.067163	0.098848	0.083006
2023	EGMPVykqaY	Andreu Fontàs	1125000.0	0.074940	0.090438	0.082689
2023	kRQabvYbMK	Walker Zimmerman	2345210.0	0.080223	0.080333	0.080278
2023	Pk5LedyLqO	Thiago Almada	2332000.0	0.096212	0.063462	0.079837

Predictions for 2023

Top predicted performers include the following:

José Cifuentes looks like a steal for that salary - perhaps why he went to Scotland Premier League in 2023

Carles Gil was an All-Star and scored 11 goals in 2023

Lorenzo Insigne highest paid MLS player - the "Italian Messi"

Hany Mukhtar was 2022 MVP and scored 15 goals in 2023

Adam Buksa was transferred to Ligue 1 for $$$

Luiz Araújo transferred to Serie A in Brazil for 9 million euros

Walker Zimmerman - 2023 All-Star, USA National team cap

Lorenzo Insigne in 2014

Value Add

Team Strategy and Game Planning:

Coaches can use these predictions to tailor game strategies, focusing on players expected to contribute significantly to offensive play.

Player Development:

Training staff can identify players with potential for growth and tailor their development programs to enhance their performance.

Data-Driven Decision Making:

Moves team decisions from intuition-based to data-driven, improving overall accuracy and outcomes.

Player Acquisition and Trading:

General managers can make informed decisions about which players to acquire or trade based on their predicted future performance and current salary.

Injury Management:

Medical teams can manage player workloads to minimize injury risk by understanding player value and contribution in various leverage scenarios.

Limitations and Future Work

Model limitations include an over-reliance on salary, past performance and potential overfitting. Further, the observed decrease in minutes played in 2020 due to the Covid-19 pandemic were not taken into account in any of the models.

The outputs of the training notebooks indicated a flaw in the ElasticNet and Stochastic Gradient Descent (SGD) Regressor. Since the linear models outperformed the trees algorithms, it is expected that either of these should also outperform the Ridge model, specifically the SGDRegressor.

Future work could include the use of external tools to compare those additional linear models in addition to performing hyper-parameter tuning.

Additional future work could involve incorporating more advanced features and external data sources like player health or weather conditions. Finally, the player's team was considered as a feature but not implemented.

Conclusion

The goal of this project was to leverage machine learning to predict offensive goals added per 90 minutes for MLS players in the 2023 season based on their performance in prior seasons.

Using a robust data preprocessing pipeline, I considered several machine learning models, and ultimately selected a Ridge Regression model with Marcel due to its strong performance and interpretability.

The model demonstrated fairly reliable accuracy in predicting player performance, identifying key predictors such as passing performance and salary.

Jose Cifuentes in 2023

Conclusion

Teams can use this model to make more informed decisions on player acquisitions, game strategy, and player development. Identifying undervalued players and optimizing player performance leads to cost-efficient team management.

Next Steps:

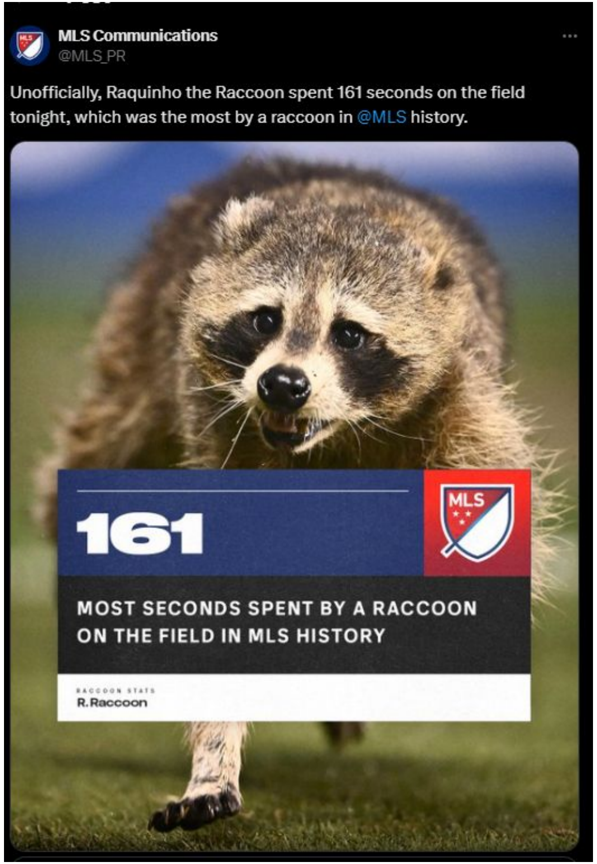
Integrate the model into team decision-making processes, ensuring that insights are actionable and accessible to all stakeholders. I would also explore the potential for extending the model to predict other performance metrics in addition to OGA_90.

Carles Gil playing in 2015

Conclusion

Perhaps Raquinho can be featured in the model next year after almost 3 minutes of playing time!

(Kidding)



MLS Communications
@MLS_PR

Unofficially, Raquinho the Raccoon spent 161 seconds on the field tonight, which was the most by a raccoon in @MLS history.

Questions?

Thank you for your attention! For further inquiries, you can reach me at adbooth01@gmail.com