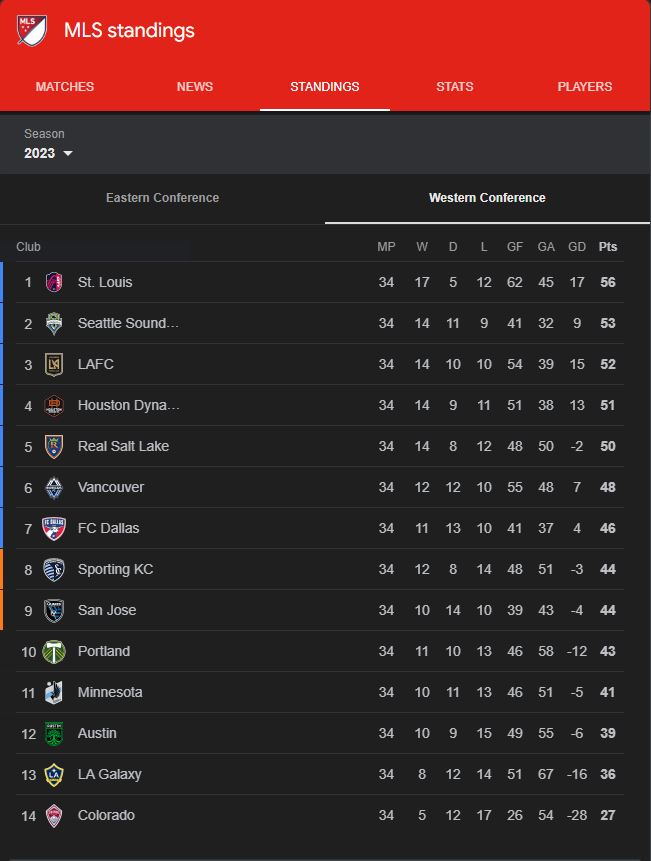
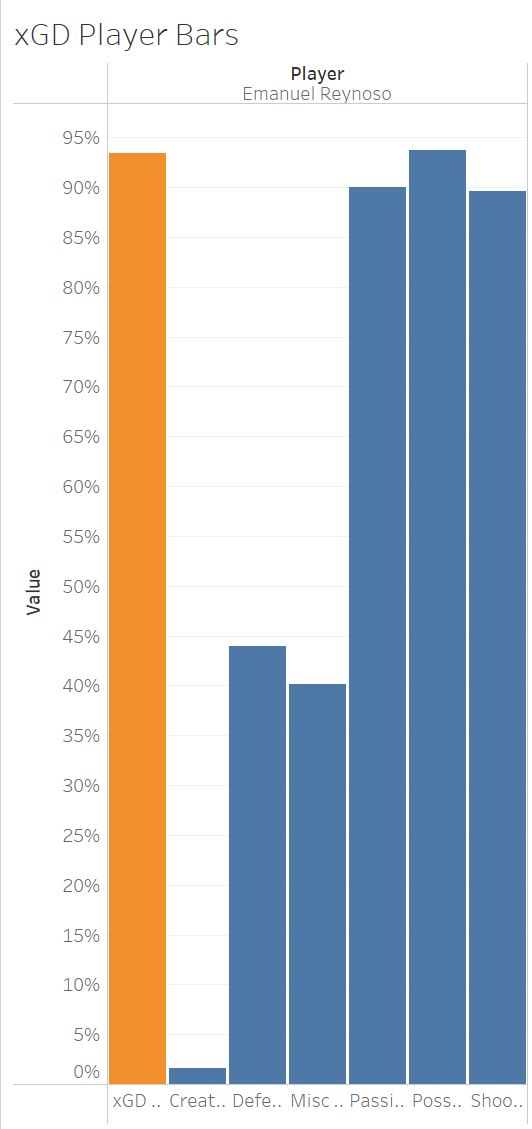
**A player’s expected goal differential (xGD) attempts to predict whether a player will have a positive or negative goal impact for his team across a season. Essentially, any number above zero means the player is expected to have a positive contribution to his team while any number below zero means the player is having a negative contribution to his team.**

****

**Take a look at the standing table above. Let’s use Minnesota United as an example. Minnesota had a goal differential of -5. If we take Minnesota’s goal differential and divide it by 34 - the number of games played in the season, we would get the number we expect them to lose by on average for each individual game. So on a per game basis, Minnesota was expected to lose by -0.147 goals.**

**What xGD is attempting to do is to bring this goal differential number down to the player level. The model attempts to assign a player an average goal differential per game, based on the stats that he has accumulated across the season on a “per 90” basis. The stats span the categories of shooting, passing, possession, defense, goal and shot creation, and miscellaneous stats from FBref.com. In soccer, stats are often transformed into a per 90 minute basis to help normalize results between starters and subs. If a sub comes on for only 4 minutes he still gets marked down for a match played. If we took that sub’s average stats per game, his contribution could be understated since he only played 4 minutes. If we normalize a player’s stats to be represented on his “per 90” minute contributions we don’t run into a player being misrepresented on a per game basis.**

****

**Let's go back to the Minnesota United Example. Above I have a screenshot of a “player card” that I have developed for Emmanuel Reynoso based on the results of the xGD model for the 2023 season. In order to make player evaluation more intuitive for someone viewing the data, I have transformed this xGD for each player into a ranked percentile where the 100th percentile is the highest performing player and the 0th percentile is the worst performing player. That would mean anyone at or around the 50th percentile is completely average. Reynoso’s xGD for the 2023 season was 0.12 which puts him in the 93rd percentile (focus on the orange bar). So even though Minnesota as a team was expected to lose by -0.147 goals per game, based on the stats Reynoso produced across the 2023 season he was expected to help his team win more often than not. So the higher the xGD a player has, the more the model thinks that he produces positive results for his team.**

**For those who are interested, I will go a little more into detail about exactly how a player’s xGD is calculated. I have player data ranging from 2018 (this is when advanced stats like xG started to be tracked for MLS players) to 2023. Each player represents a row. Within a player row of data there are stats that are normalized to a per 90 basis for that specific player. Each row identifies the season the player played in and the team that the player played for. Another stat within this player row is the average goals scored by his TEAM in that season. So in the Minnesota United example, Emmanual Reynoso’s row for the 2023 season would have a value of 1.36 goals for (46 divided by 34). This process is carried out for every player and season from 2018 to 2023. The model takes in the player’s stats produced on a per 90 basis for that season and then attempts to predict how many goals his team scored based on each individual player’s production. The model also tells us what individual player stats are expected to produce goals for a team. By interpreting the model coefficients, we can quantify the impact/importance of each stat on the expected average team goals per game.**

**A similar method is carried out for goals against. With a different set of stats (more on how this is decided below) the model attempts to predict how many goals against a player’s team is expected to concede based on his individual stat production per 90 in that season.**

**So essentially the xGD model is the result of two models doing work. There’s a goals for model that predicts the average number of goals a player’s team is expected to produce per game based on his individual per 90 stats and a goals against model that predicts the average number of goals that his team will concede per game based on his individual per 90 stats. The final result of the xGD model is simply taking a player’s predicted goals for result and then subtracting the player’s predicted Goals Against result. Positive numbers would indicate that a player is expected to produce positive outcomes for his team where negative numbers would indicate a player is expected to produce negative outcomes for his team. Or more simply, a player with positive numbers is more likely to help his team win while a player with negative numbers is more likely to “help” his team lose.**

**To give an example of the xGD process, for the 2023 season, Minnesota United was expected to produce 1.58 goals for and concede 1.46 goals against based on Reynoso’s individual per 90 stats and the results of the goals for and goals against models. Taking the expected goals for result and subtracting it by the expected goals against result, we arrive at Reynoso’s xGD value of 0.12.**

**This may seem confusing since we're taking an individual player’s stats and trying to “predict” his team level stats. What the model is really trying to do is identify players and the stats players produce that are likely to result in positive (or negative) outcomes for his team.**

**The results of this model that were tested and trained on the 2018 to 2023 MLS seasons are then applied to the 2024 season so we can understand who is having the best and worst impacts for his team (so far).**

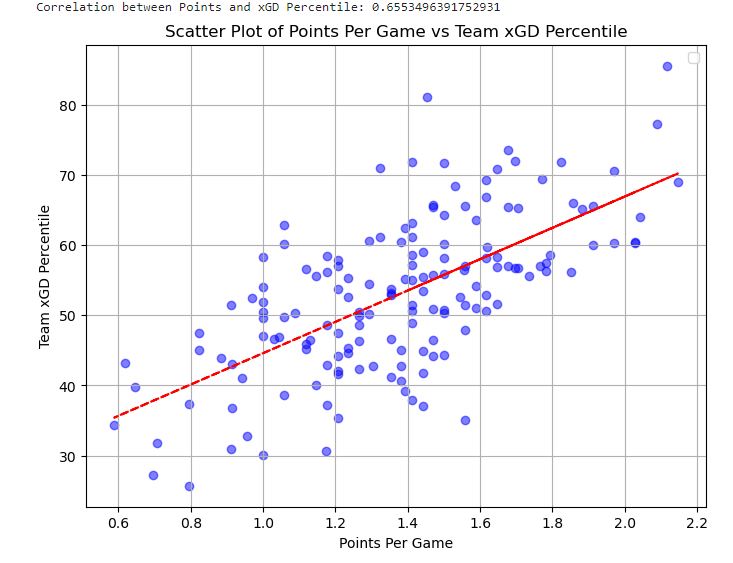
**xGD Model Validation**

**Okay cool, that all makes sense but does this model actually work? Does it do a good job of rating players? I attempted to answer this both subjectively and objectively.**

**Subjectively, I can look at the top players in the model since 2018 and see if that lines up with gut feel. Some of the top rated players include the likes of Carlos Vela in 2019 (his amazing MVP season), Zlatan Ibrahimovic in both 2018 and 2019, Alberth Elis in 2019, Darwin Quintero in 2018, and Cucho in 2023. Funny enough, Ilsinho’s 2019 season comes out as the top season ever in the model (limited to player’s playing more than ten 90s) but this is likely due to him averaging impressive marks across 12.2 90s (essentially 12.2 full games worth of playing time) across a full 34 game season. He predominantly played in a sub role and produced impressive numbers while doing so. I’m not familiar with Ilsinho’s play but if any Philadelphia fans can attest to his performance during the 2019 season I would love to hear.**

**Based on who the model thinks are the best players at the top, I’m feeling pretty good about its ability to identify good players. But feel free to take a look at how the model rates any player you can think of and see if it lines up with your sentiment and gut feel towards that player. I have attached a link at the bottom of this document that contains various visualizations and interactive tools that I have produced for my xGD model. The model is certainly far from being perfect and will likely over rate or under rate certain players. The model can only rate on the stats it has access to. We are certainly far away from being able to quantify everything a good (or bad) player does on a spreadsheet, so there are bound to be hits and misses in the model based on this inability to measure everything a player does on the field in a game.**

**As for objectivity, at first I was stumped on how to measure the model’s performance. Eventually, I settled on comparing the weighted average of a team’s xGD percentiles to the average number of points they got in a season. (The player percentiles were weighted by the number of 90s each player played, so a player’s xGD with a higher number of 90s played were given more weight. I.e. A player with 0.6 90s and a 96th percentile xGD was given less weight into his team’s average xGD than a player with 26.2 90s and a 96th percentile.) Good teams consist of good player’s right? If that’s the case, then teams with a higher weighted average of xGD players should get more points in a season than team’s with lower weighted average of xGD players; *if* my model is any good at rating player performance.**

****

**And that’s exactly the exercise I carried out in the above chart. To reiterate, the y-axis measures the weighted average xGD of all players on a team for a single season. The x- axis represents how many points per game a team earned in a single season. So each point on the chart shows a team’s weighted xGD average of all players on that team and the number of points per game they earned in that season. Drawing a line of best fit, we can see that there is a moderate to strong positive correlation between the two stats. So this does in fact reinforce two things:**

1. **Good teams are comprised of good players**
2. **With this positive correlation between my models xGD output and team performance in a season, my model does a pretty solid job of rating individual players**

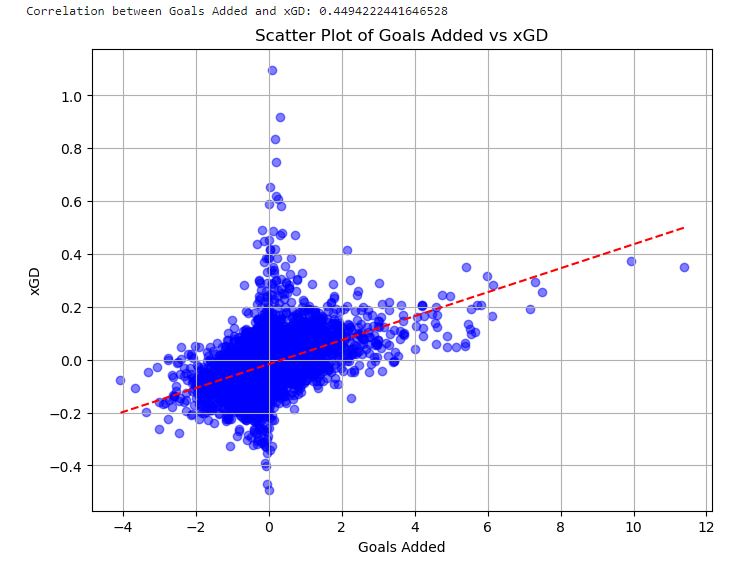
**I believe this subjective and objective approach reinforces that there is some value-add from my model. However, if anyone has any critique of my validation methodologies or any additional ways in which I can validate it, I’m all ears. In fact, that does remind me of one additional approach I used in which to validate my model.**

**American Soccer Analysis is a website/blog that posts research articles and stats related to any form of, well, American soccer. A handful of their brilliant contributors came up with a metric that is attempting to do pretty much the same thing that I am attempting to do: isolate a player’s actions and determine how much he helps or hurts his team. They came up with the catchy name “Goals Added” (g+).**

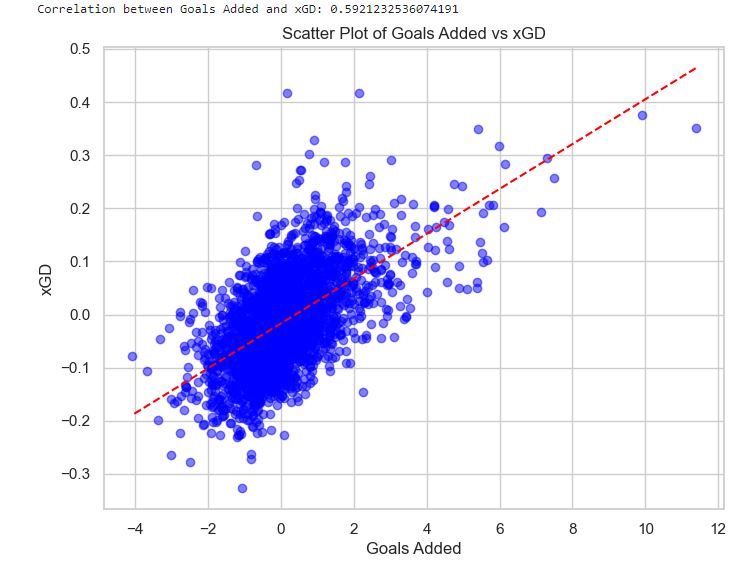
**To be frank, it’s a lot more sophisticated than the approach I use. They utilize something similar to a Markov chain to evaluate *every single*** **action that a player on the ball makes within every game. Each action is given a certain numerical value that is either positive or negative based on the probability change from one event to the next. So for example, Player A passes the ball into the box to teammate B who is now in a more threatening position to score. Player A is credited with a goals added value for that specific event by taking the probability that his team scores after his pass is received by Player B and then subtracting the probability of his team scoring before he gave the pass to Player B. So lets say that Player A is on the edge of the box before he passes to Player B and his team has a probability of 0.12 to score from where Player A has the ball. Now, after his pass to Player B, his team has a probability of 0.62 to score. Player A would be credited with 0.5 goals added for his pass to Player B. He improved his team's ability to score by 0.5.**

**Every action is scored in this way on the field. These actions are then summed across minutes in a game, across games, and eventually across whole seasons. So to put it simply - it's a really advanced approach that does a really good job of evaluating a player and the actions he takes on the field. For a more in depth read on how their goals added model works, you can click** [**here**](https://www.americansocceranalysis.com/what-are-goals-added)**. If you aren’t familiar with American Soccer Analysis, please do yourself a favor and check out the great work they’re doing over on their website** [**here**](https://www.americansocceranalysis.com/)**. I’m extremely envious of the in-game event data that they are able to work with and hope to be able to work with similar data in the future. If only I could afford it.**

**Right so where were we? Ah yes additional validation. So to get another pulse on my model’s performance, I decided to compare my model’s rating of a player within each season to American Soccer Analysis’ goals added metric.**

****

**From the above chart we can see a pretty solid positive relationship between my xGD metric (y-axis) and ASA’s Goals Added metric (x-axis). If you look at the relationship between ASA’s Goals Added value of around 0, there’s a pretty big straight line going up and down for a variety of values of xGD. My intuition tells me that ASA’s goals added model has a lot of players with a small amount of minutes played and therefore little opportunity to produce actions that add to their goals added. My intuition also tells me that since my model is based on a player’s statistics per 90 minutes that there is going to be a wild amount of variation in the model’s performance for players with not much playing time. So to get a better sense of my validating my model with ASA’s goals added metric, I will reduce the sample size to an arbitrary number of at least 600 minutes played.**

****

**Turns out my intuition is correct. We no longer see a big straight line at 0 goals added and we now have an even more positive relationship between xGD and goals added. So in general, when my model likes a player ASA likes a player. When my model dislikes a player, ASA also tends to dislike that same player. Obviously there are exceptions to this and I think that’s where a lot of the fun in this analysis comes.**

**For example, Yaw Yeboah is someone that my model is pretty high on. His XGD metric comes in at 0.112 as of matchday 9, which puts him around a percentile of 82%. My model thinks that Yaw Yeboah has a positive contribution to his team. ASA on the other hand has Yaw Yeboah with a goals added value of -0.35. Their model thinks that Yaw Yeboah has had a negative impact on his team.**

**I can point directly to my model and tell you exactly why it thinks that Yaw is having a positive impact on his team. The model sees him as a player who provides a lot of value in terms of possession and creation for his team. He specifically has impressive marks in touches inside the penalty area, touches inside the opponents final 3rd, progressive carries, progressive distance from carries, shot creating actions, and goal creating actions per 90. We expect players who produce high numbers per 90 in those categories to produce more goals for his team than player’s who don't produce high numbers.**

**ASA agrees with my model's interpretation of Yaw’s ability to create off dribbling and possession but views him as a player who hurts his team off his passing decisions. My model doesn’t think Yaw is a glorious passer of the ball; it views him as slightly above average. ASA’s model also views Yaw as someone who hurts his team defensively, something my model completely aligns with (percentile of 21.5% defensively).**

**Essentially, it seems as though my model thinks that Yaw more than makes up for his poorer defensive play with his ability to generate valuable opportunities with his dribbling/possession and his ability to create valuable shots for himself or his teammates. Meanwhile ASA’s goals added model thinks that his dribbling value add does NOT make up for his poor defensive play. There’s additional components to both the xGD model and goals added model such as shooting, fouling, and receiving but Yaw seems to be about average in those categories and I figured a discussion around just a couple metrics would simplify the comparison between the two models.**

**So you might be thinking: what does this all mean? Whose model is better? Is Yaw a good player or not? And my answer is probably a little cliche and will likely induce some eye rolls but I think the answer to these questions isn’t really an answer at all. I think if you wanted to get as much value-add out of these tools/metrics as possible, you’d simply take these preconceived notions that both models have about Yaw, store them in your head, and watch a few games where Yaw is playing. The answer is that Yaw is likely valuable to his team in ways that both models CAN capture and probably valuable in ways that they also CAN’T capture. All in all, I think models, data, and analytical tools are all helpful items to use to evaluate a player but I don’t believe you should ever solely rely on them. Analytics are just yet another useful piece of the puzzle in trying to figure out what players are good and what exactly helps teams win soccer games. Also, ASA doesn’t necessarily think Yaw is a bad player, his goals added value isn’t that far below zero. I simply used him as an example as he’s one of the players that both models don’t necessarily agree on.**

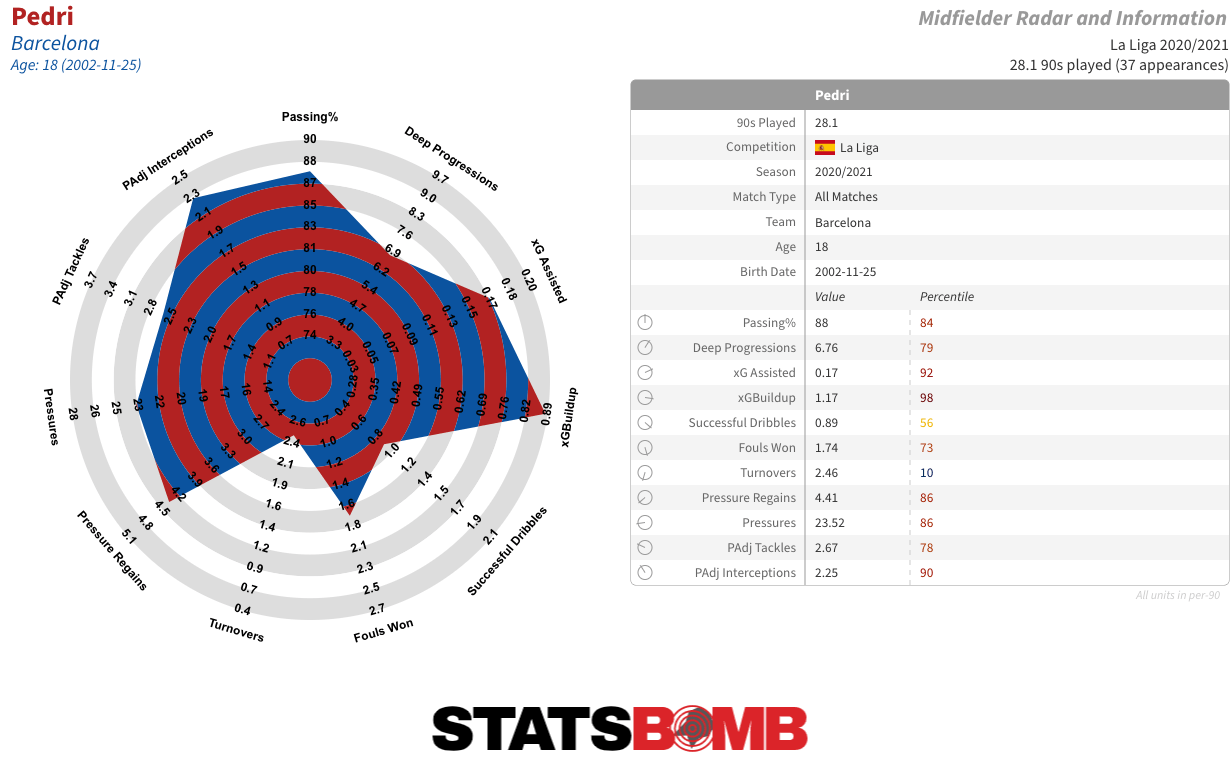
**Now there are absolutely shortcomings in my model. I think the biggest caveat that needs to be mentioned is the model’s ability to properly value defensive play. To be fair to the model, this is a widespread problem in soccer analytics in general though. A lot of the value in defensive play comes from proper positioning and being in the right place at the right time which is most often happening OFF THE BALL. Remember, the model is only valuing ON THE BALL actions. A really good defensive player is going to put himself in positions where the ball doesn’t come his way because he is positioned properly. Without player tracking data capturing a player’s location at all times of a match, a model like mine is not going to properly value defensive players and contributions.**

**That being said, I am very confident in the goals for model performance and think the model does a great job of properly valuing offensive contributions players make. The goals against model isn’t horrible - it returned a decent model performance - but it isn’t quite at the same level as the goals for model. When looking at the final xGD model results, just keep this in mind.**

**So to sum it all up: I think this player xGD metric I have created does a pretty solid job of telling you what players help their teams win. As a fan of MLS, I love diving into my model’s results before I turn a game on on Saturday night and seeing which players I need to keep an eye on (in a good way OR a bad way). I hope that some other people get some enjoyment out of the data I’ve compiled and find the results useful in their fandom of the beautiful game. If anyone has any critique, comments, or questions about my model I am more than happy to respond. Or if you have any requests or ideas on things you’d like to see data wise related to the model I’d love to hear them! If you’ve made it this far, thanks for your attention. I hope you got some form of enjoyment out of my rambling.**

**I have lots of ideas for further enhancements to my xGD tools and future analyses related to MLS, so if you think any of this stuff is cool I have some resources below where you can connect with me.**

**Partial Motivation for the Creation of the xGD Model**

****

**Part of my motivation for creating this model was my frustration of coming across charts that displayed player stats, like the StatsBomb one above - a common sentiment shared by author and soccer writer Ryan O'Hanlon. On page 171 of his book “Net Gains,” he states the following about these charts: “You’ll see a slick black background with various fire-colored vectors shooting off in all different directions, with slight variations to each line - a squiggle in the middle, a different shape on the end - to signify the different kind of action the line is representing. It’s like if Jackson Pollock was a graphic designer who cared only about soccer. I’ll find myself marveling at the detail and the design before butting up against an intellectual wall: What does this actually tell me? How am I learning anything here beyond ‘This player or this team sure did a lot of stuff’? And then I realize I’m not; it’s a display of the company’s computing power, a call to convince more clients to sign up.”**

**A paragraph in a book has never resonated with me so deeply. I mean look at the chart of Pedri above. What *does* it tell you? You can come across similar charts on all sorts of soccer stat websites. These charts do a great job of telling you a player does a lot of things on the field - but it really doesn’t do much outside of that. With my model and project, I want users to understand that when you’re looking at visualizations and players, that a higher percentile means that a player contributes more actions that help his team WIN. So when you look for example at Messi’s chart, just remember that his high percentile value for shooting means that he executes a high number of shooting actions that are expected to help his team win. These charts are directly showing you an estimated amount of how much a player helps his team win. They aren’t just colors occupying space.**

**Appendix**

**Alright I’ll make this section short but I am sure there’s a handful of people who are interested in the actual methods used for modeling. Looking at fbref.com, if you go to the MLS league page and hover over “Squad & Player Stats,” there’s a handful of stat categories. For my model, using my intuition, I figured the stats stored in shooting, passing, goal and shot creation, defensive actions, possession, and miscellaneous stats were the ones that were most impactful to understand what stats players contributed that helped their teams win.**

**Where I couldn’t use simple intuition was choosing the plethora of stats that exist within each of these categories. Just within shooting stats there are 17 different measures. This can be overwhelming to cherry pick stats that seem like they would be statistically significant in measuring expected goals for or expected goals against. Mashing together 100 variables into a model would not work well for many reasons, a main concern being overfitting.**

**So I employed the help of a machine learning algorithm/method known as a GLM LASSO model. Similar to my description about creating the goals for and goals against models, I took the per 90 stats in each stat category on FBRef for each player and set it against that player’s team’s average goals for that season.**

**A GLM LASSO model is a powerful algorithm that can be used for variable selection. Without going too far “under the hood,” according to IBM, LASSO is a “regularization technique that applies a penalty to prevent overfitting and enhance accuracy of statistical models.” This penalty that the model applies shrinks certain predictor variable’s coefficients (variables trying to predict a team’s goals for in our case) towards zero or even sets the coefficient to zero as it is training. That’s what the S’s stand for in LASSO: shrinkage and selection.**

**In an overly simplified description, what we essentially have is the LASSO model deciding what variables have some signal in explaining the variation of our target variable (average goals for or average goals against) for each stat category on FBRef (like shooting). After the GLM LASSO model was done training, I got a set of variables that either had coefficients above 0, below 0, or set at 0. I selected the variables that had coefficient values above 0 or below 0. This process was then repeated for every single stat category on FBRef that I talked about above for both the goals for model and goals against model. So the variable selection process was carried out by 12 separate GLM LASSO models. I then combined all the variables the LASSO models determined as having a positive or negative coefficient into a simple OLS regression for both GF and GA.**

**So to make it more simple: one GLM LASSO model for each stat category selects what stats within that category are important for both average goals for and average goals against data. All those “important” stats are combined into separate OLS regression models for both average goals for and average goals against. Then for each player, we subtract the predicted average goals for value by the predicted average goals against value. That gives us a final xGD number for each specific player. That final xGD number is converted to a ranked percentile to make it more intuitive for people viewing the data. Oulia, there you have it. A more detailed look into the modeling portion. That’s xGD.**

**In an effort to visualize and hopefully help you understand the modeling process I have a diagram attached** [**here**](https://drive.google.com/file/d/1lWrxRMg32ezMVVHAQNqz1TpwjUDLRD1g/view?usp=sharing)**.**

**For more details on LASSO regression please refer** [**here**](https://www.ibm.com/topics/lasso-regression#:~:text=Lasso%20regression%20is%20a%20regularization,the%20accuracy%20of%20statistical%20models.)**.**

**Link to xGD Dashboards:**

**Email:** [**zrodeheffer@gmail.com**](mailto:zrodeheffer@gmail.com)

**My substack:**

**Works Cited**

**“What Is Lasso Regression?” *IBM*, IBM, 27 Mar. 2024, www.ibm.com/topics/lasso-regression#:~:text=Lasso%20regression%20is%20a%20regularization,the%20accuracy%20of%20statistical%20models.**

**O’Hanlon, Ryan. *Net Gains: Inside the Beautiful Game’s Analytics Revolution*. Abrams Press, 2022.**

**Muller, John. “What Are Goals Added (G+)?” *American Soccer Analysis*, www.americansocceranalysis.com/what-are-goals-added. Accessed 7 May 2024.**

**“Football Statistics and History.” *FBref.Com*, Sports Reference, fbref.com/en/. Accessed 7 May 2024.**