FIFA 20: A Guide to Ultimate Team

Exploring FIFA 20 to Predict Player's Potential

Table of Contents

Overview

Background

FIFA 2020 is one of the world's most popular video games. It is a football simulation video game published by Electronic Arts (EA Sports) as part of the FIFA series. It is the 27th installment in the FIFA series and was released on 27 September 2019 for Microsoft Windows, PlayStation 4, Xbox One, and Nintendo Switch.

Historically the FIFA series has been one of the main revenue generators for Electronic Arts (EA). According to sportbible, across all sports franchise titles, including FIFA and Madden, EA generated a total of $1.49 billion through the Ultimate Team platform --- which is a $120 million increase on last year's revenue total of $1.37 billion. The Ultimate team platform is a gameplay experience that allows gamers to build their team and compete against others. The goal is to create your dream squad with superstars from past and present. For fiscal 2020, EA's revenue topped $5.5 billion, with $2.7 billion of that generated from players spending money on **in-game content** or live services. Gamers looking to take advantage of in-game purchases by finding hidden gems, players with high potential, or already established superstars to build their team.

Goal

This document explores in-game players from the FIFA 20 database to understand how different characteristics, traits, demographics, and more influence players' overall rating and potential. Throughout this analysis, we will uncover insights to drive a machine learning model.

The objective of the ML model will be to predict the ***potential*** of a player in the game. Uncovering a player's potential is vital for the gameplay experience because it gives gamers insights on which players have the most upsell or player development opportunities. This strategy will give EA sports the resources to price individual players higher/lower based on their potential score, thus converting in-app purchases.

Key Data Points

Below are ***some*** of the important columns available from the dataset. These won't be the only columns used but should be sufficient to guide the initial analysis.

Column	Description
long_name	full name of player
age	player's age
dob	date of birth
nationality	country the player represents
club	club the player represents
overall	player rating (scaled from 1 - 100)
potential	player potential rating (scaled from 1 - 100)
value_eur	value in Euros
wage_eur	wage value in Euros
player_positions	positions a player can play in (can contain multiple positions)
preferred_foot	dominant foot used
skill_moves	rating from 1 to 5 for skills level
work_rate	rating from 1 to 5 for work rate level
body_type	type of physique (lean, normal, stocky)
release_clause_eur	buy out contract value in Euros
player_tags	tags a player has (can contain multiple tags)
player_traits	traits a player has (can contain multiple traits)
team_position	position a player plays for in a team
main_position	position a player prefers to play in
team_jersey_number	the number on a players club jersey
pace	player's pace score from 1 to 100
shooting	shooting score from 1 to 100
passing	passing score from 1 to 100
dribbling	dribbling score from 1 to 100
defending	defending score from 1 to 100
physic	physic score from 1 to 100

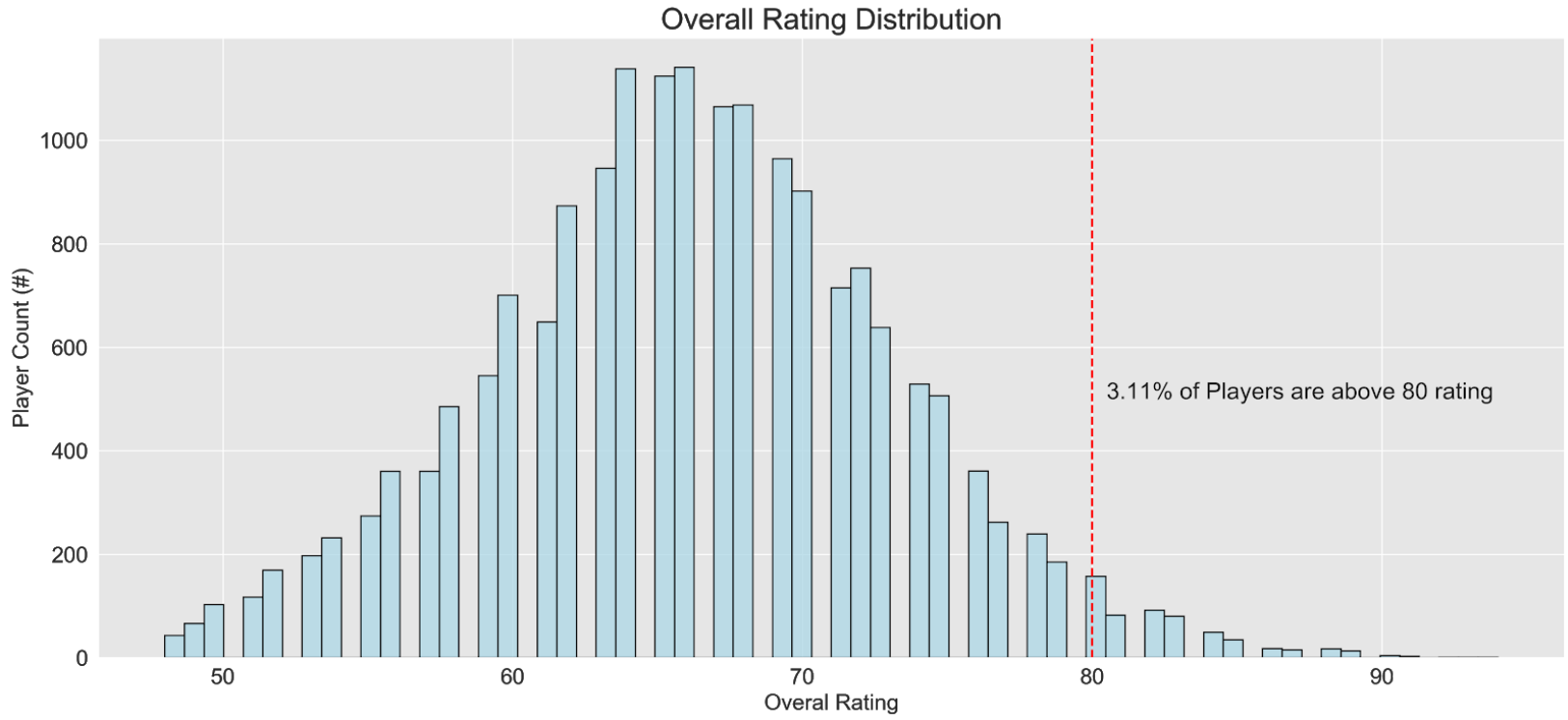
Report Summary

- 568 out of 18,278 players in the FIFA 20 database recorded ≥ 80 overall ratings. An 80 rating is generally seen as a sign that the player plays for a top team or is a better player for their respective team.
- In situations where a player could be ranked higher based on their overall rating than their overall valuation, we can find that 5,625 (31%) players have a lower valuation ranking than their overall rating.
    - For example, Lionel Messi is ranked number #1 for the overall rating, but he is ranked number #2 for valuation.
- Younger players generally have leaner body physiques compared to the older group. Specifically, players under 25 years old are the only group with more lean players than Normal or Stocky body types.
- The most common **team position** in FIFA 20 is a substitute (representing 58% of all players), which means that most players are not starters for their respective clubs.
    - However, looking at the player's preferred position, which is their **main position**, the most common is the ST (striker) position recording 2.1K players (approximately 11.5% of the player database).
- Amongst the top teams, most have a spread out overall rating between their 25th and 75th quartiles of player ratings. However, 3 top teams such as Bayern Munich, Juventus, and Real Madrid have their interquartile range of ratings much narrower, around 75 to 87 ratings.
- A player's *Potential* and *Release Clause* has a medium/high correlation coefficient at 0.60 and enables us to identify young players with the most potential for in-game play.
- There is a clear pattern where a player's shirt number and position have a strong relationship. The GK wears #1, defenders between #2 - #6, then midfielders and strikers #7 - #11.
- For ≥ 80 overall ratings, 80% of the top 10 countries represented are in Europe, except Brazil and Argentina (South America region), with Spain the most represented country.
- Building a model to predict a player's ***potential***, we can see that developing features around a player's position, traits, and characteristics proved significant in overall performance.
    - Regression Training Score: 0.8538
    - Regression Testing Score: 0.8487
- Some examples of the coefficients impacts on the model are below:
    - For every 1 increase in the trait 'tactician', expect a 1.3 **increase** in potential.
    - For every 1 increase in a players' age' expect a -2.4 **decrease** in potential.
- 50% of the model's prediction fell within -1 to 1 from the actual potential value.

Exploration

What percentage of players have a rating above 80%?

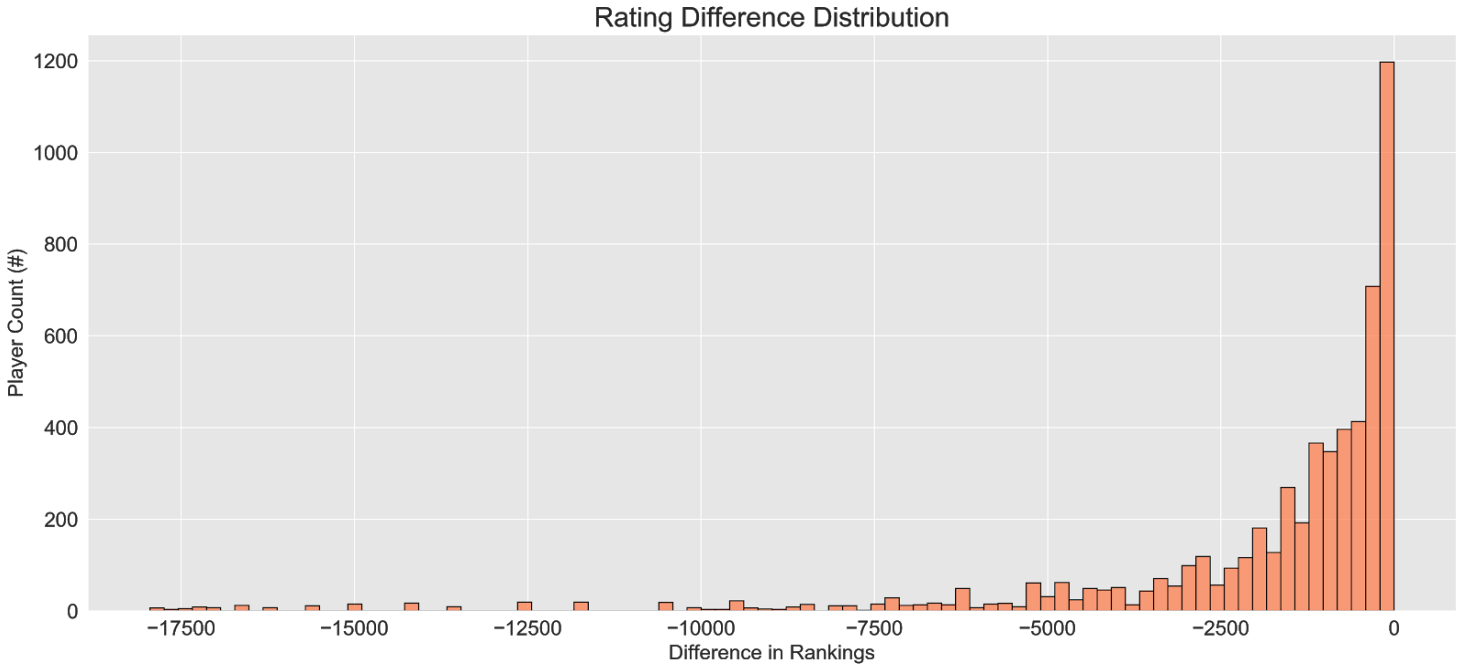
Only 3.11% (568 players) of players have an overall rating above 80. The median player rating in the database is 66. The graph below shows a normal distribution with a slightly thinner tail on the right side. It is suggesting that there are fewer top-rated players in the dataset. These ≥ 80 rated players represent 112 clubs, which is approximately 16% of all the clubs, telling us that even though the subset of customers is small at ~3.11%, the variety of clubs isn't as scarce.



Is a player's overall rating ranking similar to their valuation ranking?

5,625 (31%) players have a lower valuation ranking than their overall rating. For example, Lionel Messi is ranked number #1 for the overall rating. However, he is ranked number #2 for valuation, meaning he has a ranking difference of -1, and it is not a 1:1 match between rankings.

While 31% of players having a negative difference in ranking may seem high, many factors influence overall rating and player valuation. Due to age, older rated players won't be valued as high compared to younger players with more potential. Since we look at the ranks between individual players, the histogram below displays the 5625 (31%) players with a lower valuation ranking than their overall rating.

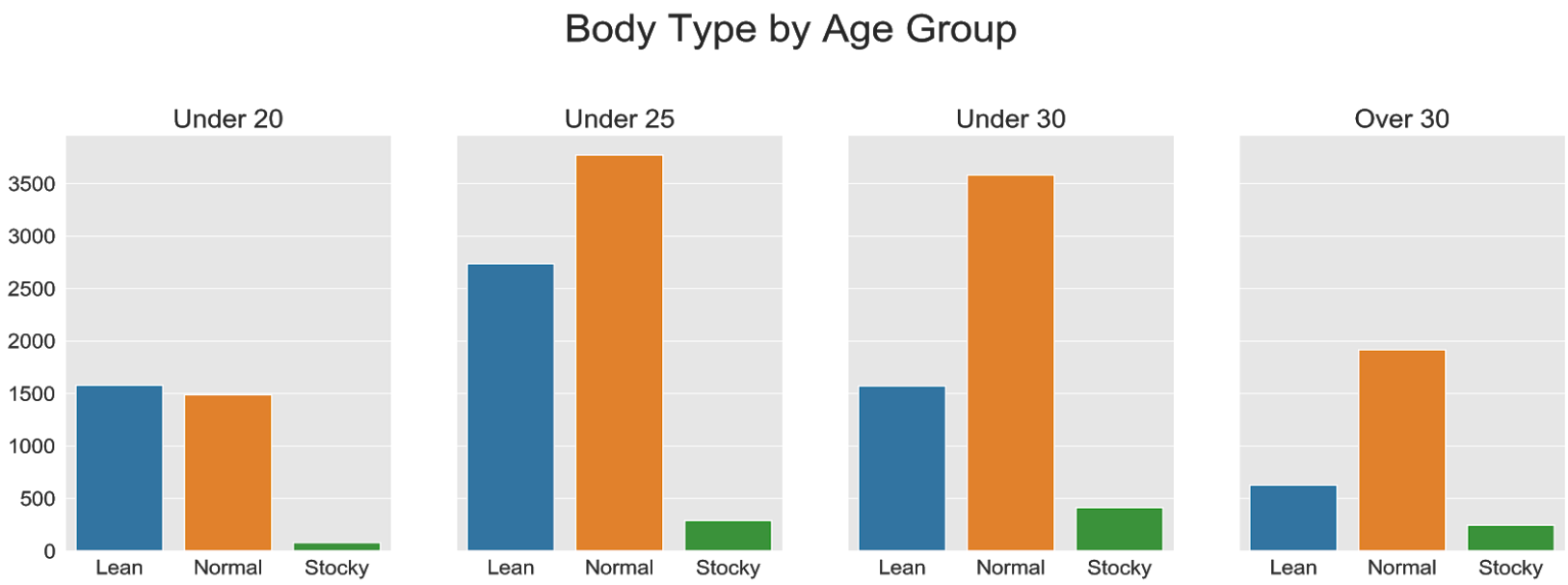


The chart is left-skewed, which is expected given we don't expect players to have astronomical differences in rankings between rating and valuation. However, interestingly there are a noticeable amount of players with a lower than -2500 difference in rankings. Suggesting these players are potentially are "hidden gems" within the game since they are highly ranked in their overall rating but lower ranked in their valuation. Essentially they could be classified as

“undervalued,” which would be an opportunity to train or upsell them in the gameplay career mode.

Are younger players typically leaner relative to their age group?

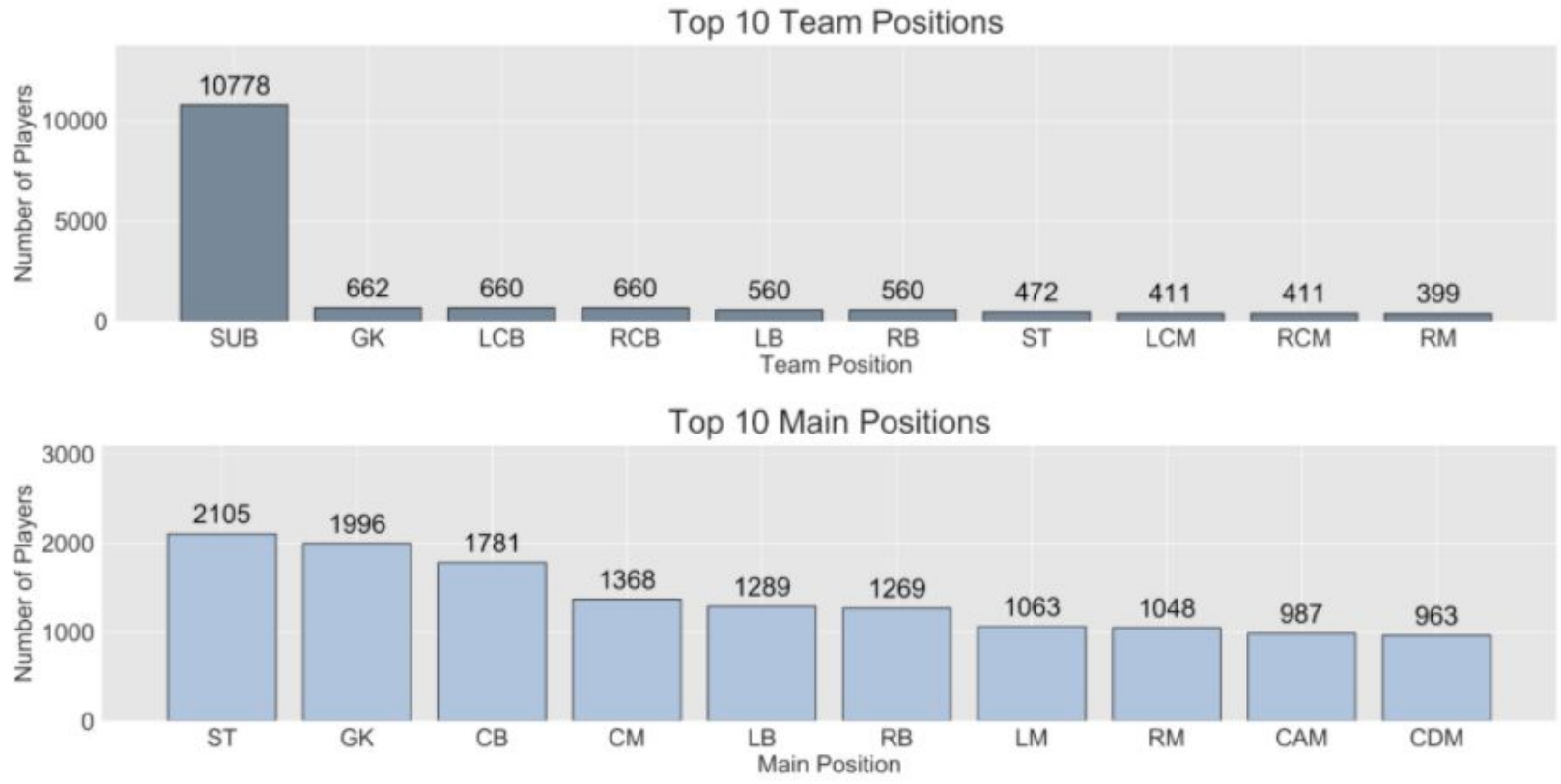
Yes, the players *Under 20* are the only age group with more Lean players relative to their group size compared to Normal or Stocky. When players are younger, they are less developed and may not be as experienced in building muscle. Unless a player under 20 is a promising talent or already a rising star, these players are likely coming up from the youth leagues or being promoted from the academy team, which means there is more room for growth and development. In contrast, amongst all the other age groups, the Normal body type is the clear leader.



Is the most common position a CB or CM?

In a football team, it is normal for a team to play with 2 CBs (center backs) and 1-2 CM (center midfielders) at once, so we would expect these positions to have a higher volume of players. However, that is not the case in the dataset available.

First, we can split positions up into two types Team Position and Main Position. Team position would be where the individual plays most of the time week in and week out. This can be represented by the top bar chart below. The most common team position is SUB (10.8K players), representing 58% of all players. The SUB stands for substitute, which means they are not starters. They are backups or reserves. Implying that most of the players available in FIFA 20 are not actively playing in the starting lineup for their teams in real life.



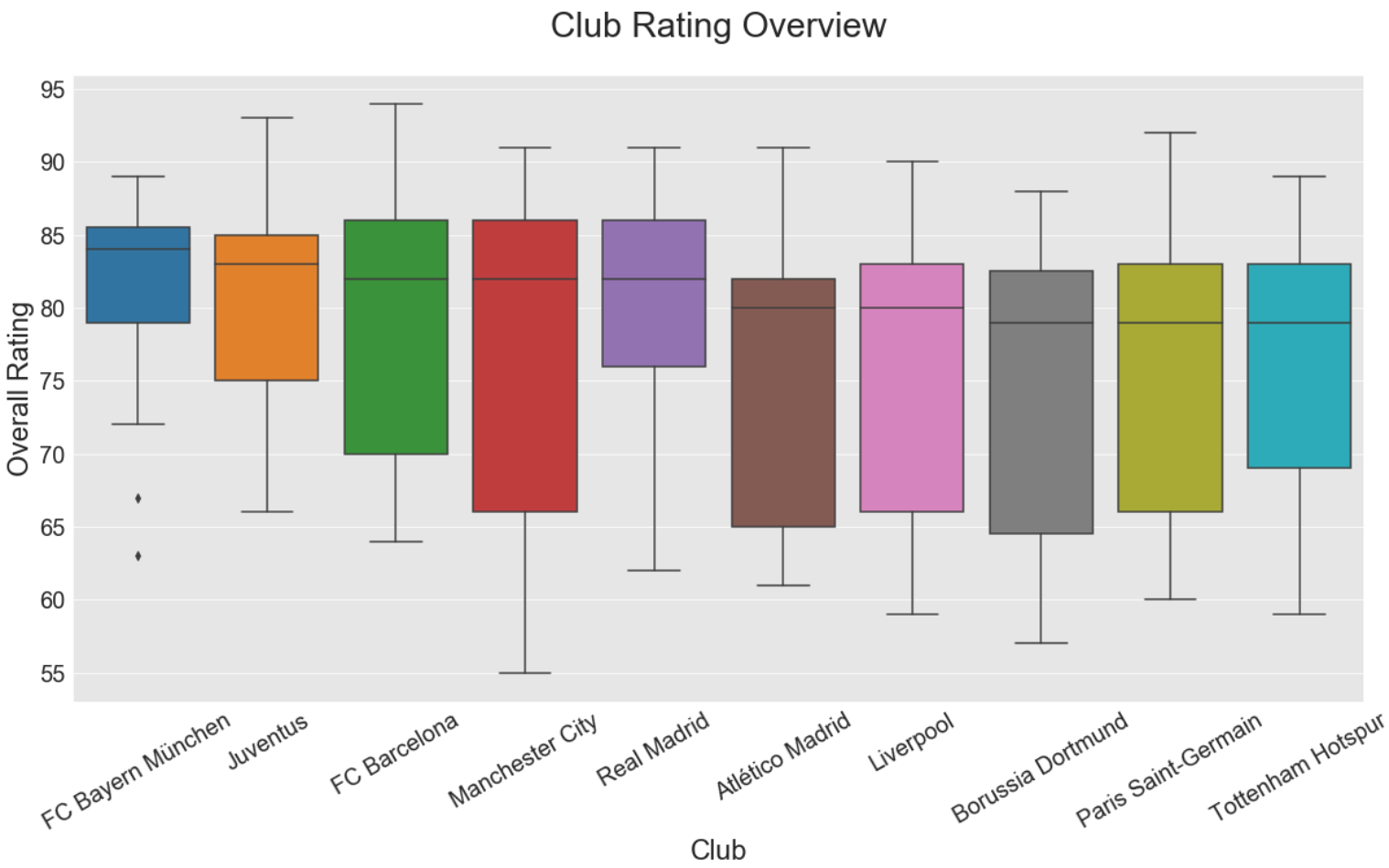
Next, we looked at the Main Position, which would be where the players would play, assuming they are not a SUB. Evidently, it is their preferred position. Now the second bar chart below, we see a shift in the positions. The ST (striker) position has 2.1K players and is the most popular position preferred. From the first graph, we know that only 473 out of 2.1K ST players are starters, with the second graph informing the remaining 1.6K (78%) ST are SUBs. We can see similar patterns with the GK position, where a considerable number of players (1.3K players or 68% of GKs) in the game are, in fact, a SUB.

On the other hand, CB and CM's Main position is relatively high with slightly more reasonable numbers. A reminder that from Team Position (LCB + RCB) = Main Position (CB) and Team Position (LCM + RCM) = Main Position (CM). The L or R represent Left or Right for their respective position, so if we calculate CB and CM's totals, we get the following.

- For Team Position: LCB (660) + RCB (660) = 1320 → compared to 1781 CB for Main position. This tells us 1320/1781 or 74% of CB are starting for their teams.
- For Team Position: LCM (411) + RCM (411) = 822 → compared to 1368 CM for Main position. This tells us 822/1368 or 60% of CM are starting for their teams.

Are there spread player ratings amongst the top teams?

Those who have played FIFA know that most top teams have players above an 80 rating across their roster. If we focused on the top teams, isolate which have more spread player ratings. Below we can see the top 10 highest rated teams, based on the numbers of players with a score above 80. The boxplot is sorted from the highest median rating (Bayern Munich - left) to the lowest median rating (Tottenham Hotspur - right). There are a few key takeaways.



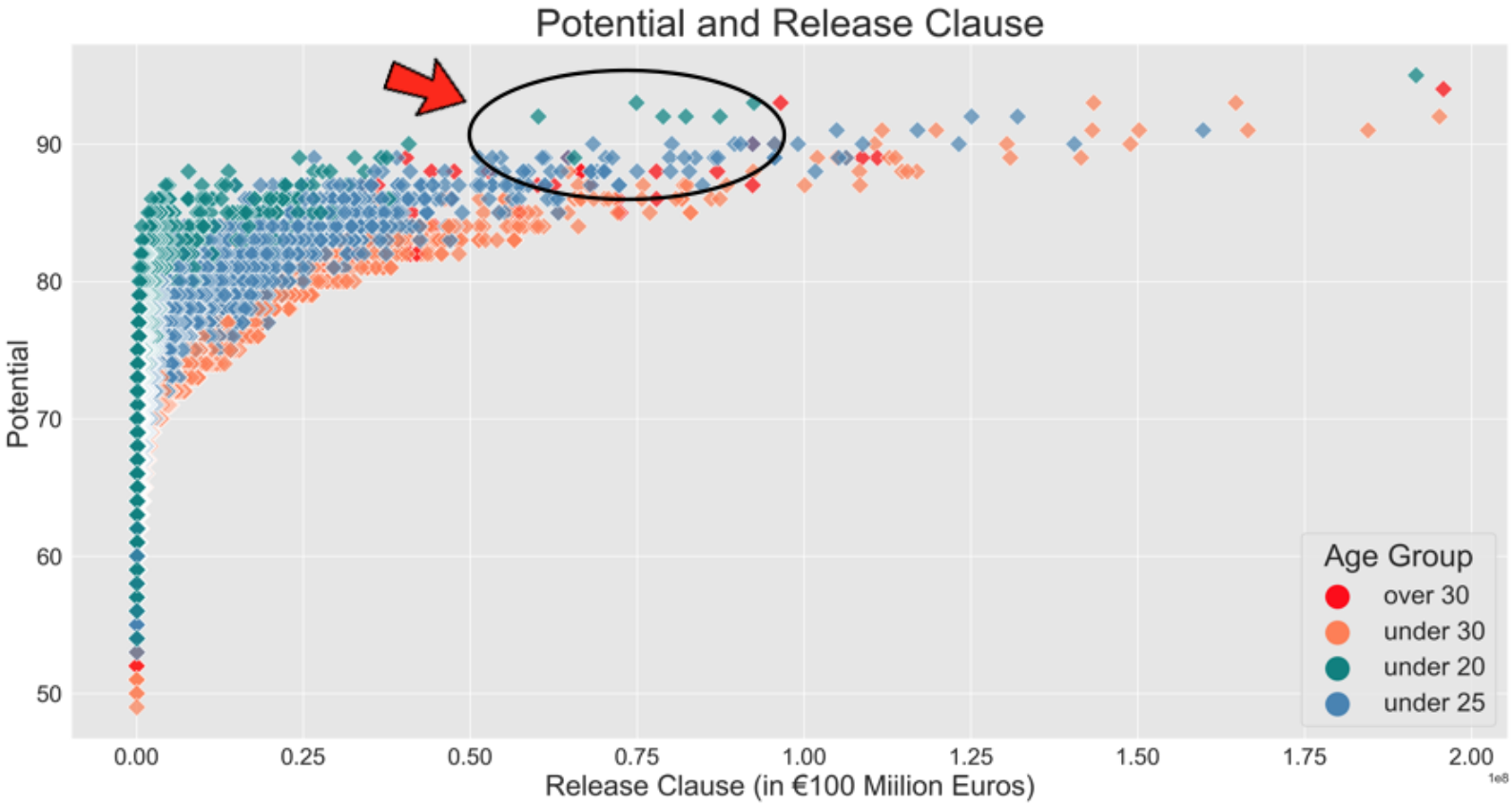
First, three teams have minimal interquartile ranges (height of the box), Bayern Munich, Juventus, and Real Madrid. This tells us that 50% of their players are very closely rated. For example, Bayern Munich has the highest median player rating (~84) while their 25th to 75th percentile is between 79 to 86 ratings, which is very high - suggesting the team is even and emphasizes the depth in their team. Coincidently, Bayern Munich won the biggest tournament

of the year, the UEFA Champions League 2020, so the data validates their results with their team player ratings. Higher rated players equal strong team performances/results.

In comparison, other teams have their player ratings in their interquartile range spread around 15 - 20 points. Focusing on Manchester City (red), they have a relatively high median player rating of ~83, while their 25th to 75th percentile ranges from 66 to 86. This tells us they have a wide variety of player ratings, which is a noticeable difference from Bayern Munich. One thing to note is Manchester City has 33 players on their roster compared to 23 for Bayern Munich. It is common for some teams to have a large squad due to academy prospects, but it still gives a strong sense that some top teams manage their players differently and have more well-rounded ratings.

Do younger players with the most potential have high release clauses?

Young players with higher potential do not have as high release clauses than older players, except for a few cases. A release clause is how much it will cost a competing club to purchase the player before their contract expires. It is expected that already established high rated players to have astronomical release clauses in their contract. For example, looking at the scatter plot below, the far-right dark red marker represents Lionel Messi is 32 years old and is the highest-rated player in the game with a release clause of 196 million euros (€).



We can see a pattern with the players where most of the higher release clause players with potential above 88 are under 25 and 30. Potential and Release Clause have a medium/high correlation coefficient at 0.60, so gamers trying to find which players would be an optimal choice to grow their team and find the best value. The circle highlights critical players in the under 20 and 25 groups, with high potential and relatively lower release clauses than some of

the older players. These players could be interpreted as “best bang for your buck,” with them ranging between 50M - 100M euro release clauses. While this is still a large release clause, purchasing them at a younger age with high potential would mean their future resale value increases. The table below represents the players from the highlighted circle area in the scatter plot.

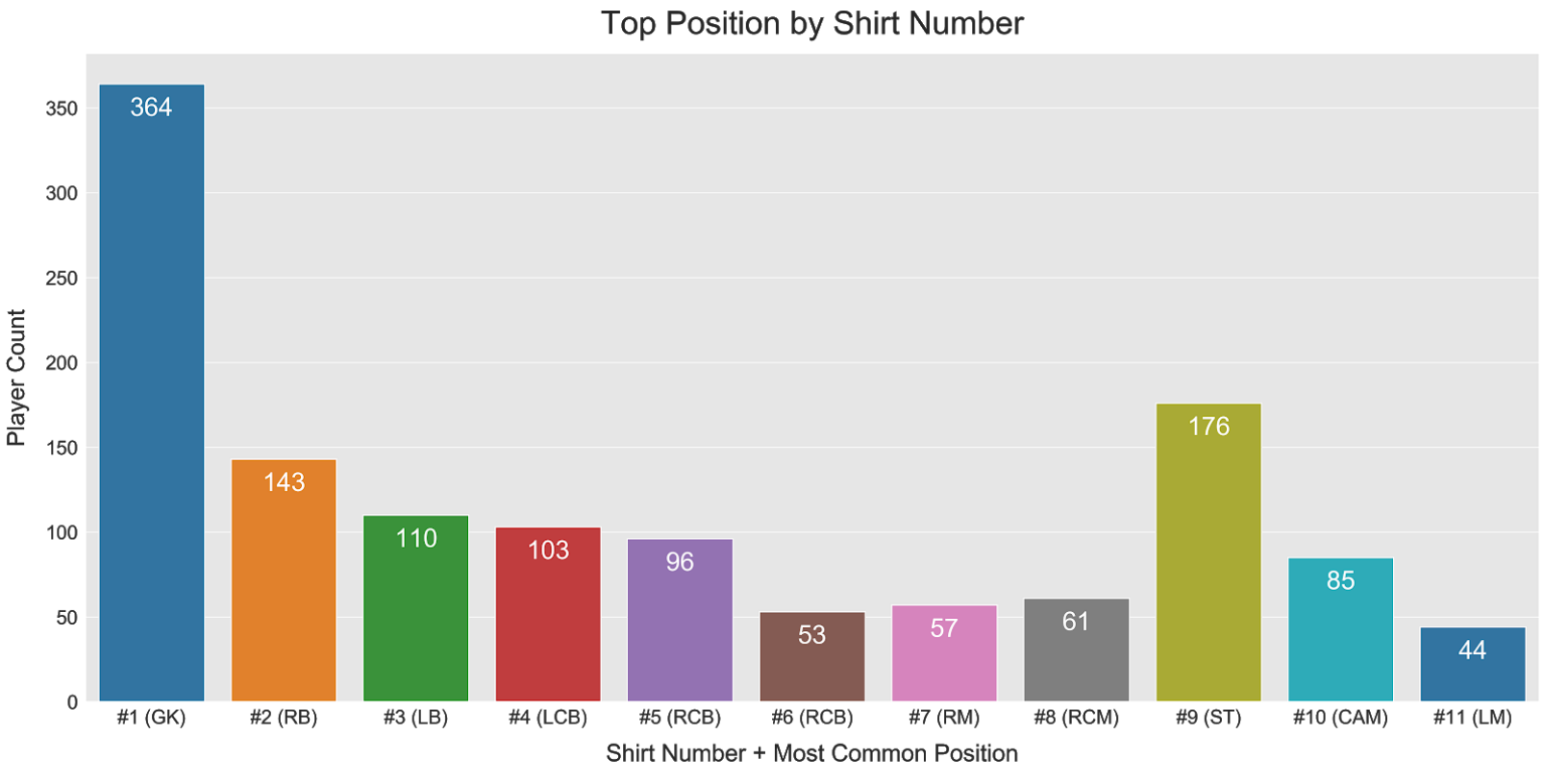
Player	Club	Potential	Release Clause (Euros)
M. de Ligt	Juventus	93	92,500,000
G. Donnarumma	AC Milan	92	78,900,000
J. Sancho	Borussia Dortmund	92	82,300,000
K. Havertz	Bayern Leverkusen/ Chelsea	92	87,400,000
João Félix	Atletico Madrid	93	74,900,000
Vinícius Jr.	Real Madrid	92	60,200,000

Can we see any patterns in the shirt numbers players wear and the position they play in?

Yes, traditional players in certain positions will wear specific numbers. This isn't a requirement, but historically players have selected specific shirt numbers based on their position. For example, famous players such as Maradonna wore the number 10 shirt and played in the CAM/ST position. David Beckham wore the number 7, playing in RM and Gianluigi Buffon wore number 1 as a GK. These are some examples of top players who made their position and shirt number in unison.

In football, the starting lineup consists of 11 players. Generally, the GK takes #1, defenders between #2 - #6, then midfielders and strikers #7 - #11. We can see this pattern across all players in the FIFA 20 database. Below is a bar chart representing the most frequent position wearing the respective number - so the leftmost bar is for shirt number #1 with the most common position GK. There is a precise sequence if we look at the different positions and as the shirt numbers increase.

- Goalkeeper: Shirt #1 with 364 positions
- Defenders: #2 - #6 held by RB, LB, LCB, RCB and RCB
- Midfielders: #7, #8, #10, #11 held by RM, RCM, CAM and LM
- Striker: Shirt #9

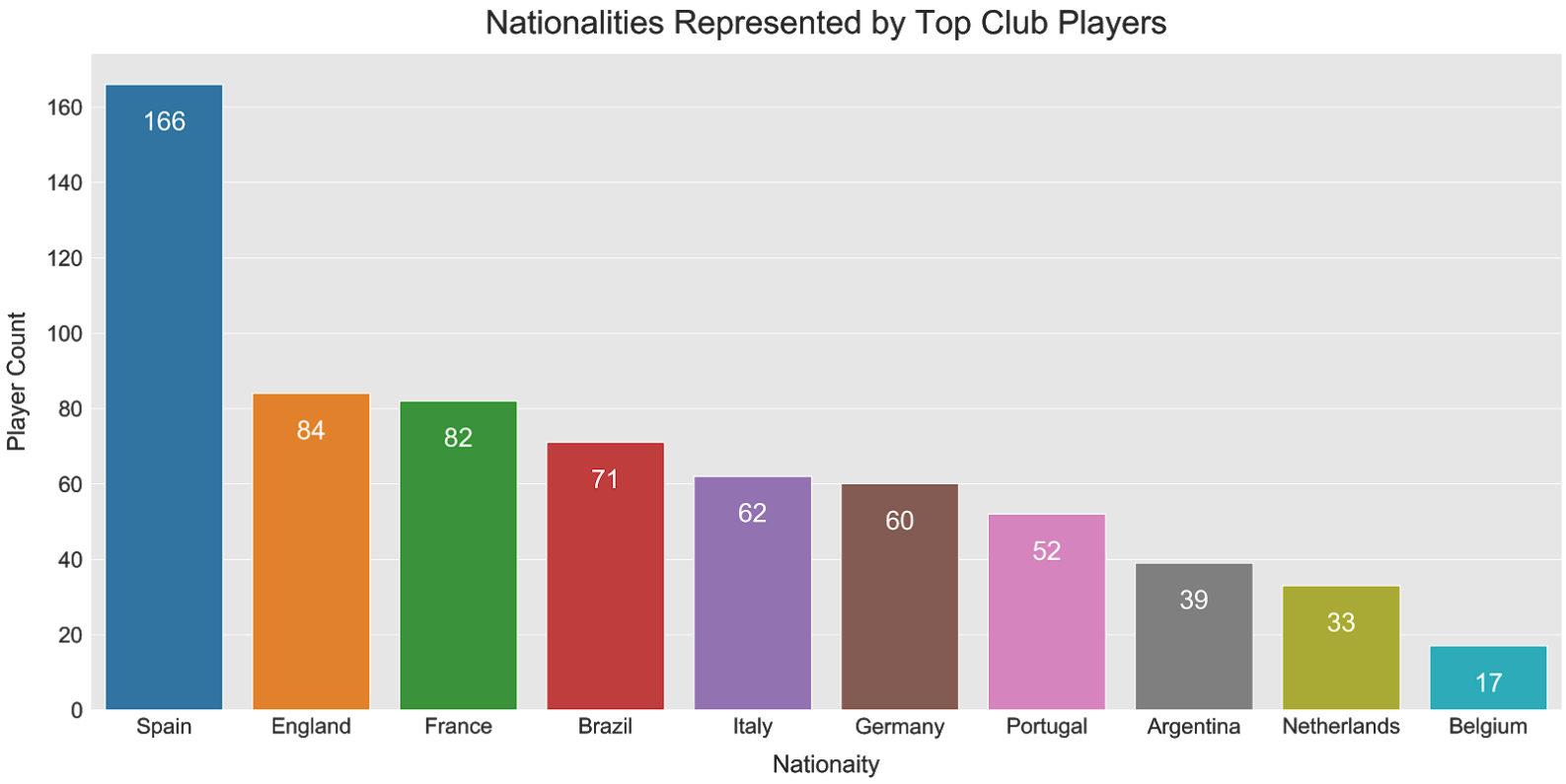


While we see a pattern, these numbers are certainly not set in stone or a requirement, but it gives a general overview that IF a player wears one of these numbers in that position, there is a higher chance they are a starter team. Below we can see some examples of the top 5 players that wear the number #9 shirts and their ratings and positions. They are all first-team starters and represent elite clubs from around the world.

Name	Club	Overall Rating	Shirt Number	Position
L. Suárez	FC Barcelona/Atlético Madrid	89	*#9*	ST
R. Lewandowski	FC Bayern München	89	*#9*	ST
E. Cavani	Paris Saint-Germain/Manchester United	88	*#9*	ST
K. Benzema	Real Madrid	87	*#9*	ST
Roberto Firmino	Liverpool	86	*#9*	ST

Which are the most represented countries amongst the top clubs?

If we look at which clubs have the most ≥ 80 overall rating players and count the number of countries they represent, most of them come from Europe. 80% of the top 10 countries represented are in Europe, except Brazil and Argentina (South America region). Amongst the top clubs, Spain is the most represented country with 166 players (+82 more players), with England next at 84 players, followed by France at 82 players.



One of the drivers for more Spanish players playing in the top teams is that Spanish clubs have a higher average of domestic talent. Amongst the top clubs, the average Spanish club has 18 Spanish players. In contrast, the average number of local, national players representing their clubs is 11 for England, 12 for French, 11 for Italy, and 11 for Germany. This informs us that more Spanish players remain local in their league, resulting in top Spain clubs better nationally represented.

Model

Overview

Throughout the analysis we have seen that players in FIFA 2020 are given statistics on their skill set, favored positions, club/nationality information, player traits, and characteristics. Finally, a player receives an overall rating that gamers use to identify the quality of the player/team they are using. In the game, certain individual players have more potential than others. It allows gamers to purchase players in the Ultimate Team gameplay. This encourages them to find players with the most potential to develop and grow them into top players and possibly sell them for the most money. Potential generally has the most significant gain for younger players in the top teams. Still, other factors come into play that influence their rating.

The model developed in this process will be used to output a final potential rating for players in the game and can be used by EA sports to price players at different tiers. The pricing tiers is how EA sports convert unpaid gamers into paid gamers. For example, if a gamer sees a FIFA 2020 player with a promising potential rating, they will be encouraged to spend real money to acquire them. The gameplay process allows in-game purchases, and having players with the most potential has the most significant upside.

Feature Engineering

Time/Dates

Introduced several new columns focused on the available date values from the given dataset - results are below.

1. Number of years between the current date and their expected contract expiration year
2. Number of days between the current date and the date they joined their club
3. Columns representing players birth day, birth month, and birth year

Positions

The original dataset provided three columns for a player's position, player_positions, team_position, and national_position. In the analysis, I discovered that most of the players had a SUB (substitute) position for their team_position column. This didn't provide enough context because being a substitute shouldn't always impact their rating. Therefore I developed a new column called main_position, which captures a player's **preferred** position if they were, in fact, a substitute. Next, I converted the new main_position and original team_position columns into dummy variables to encode a 1 or 0 depending on if a player meets the requirements.

Player Traits and Tags

There are two columns player_traits and player_tags, that store multiple text values in a single cell. This allows a player to have several traits/tags that define their characteristics. For example, a player could have several traits such as 'injury prone', 'injury free', 'tactician', 'poacher', etc. We want to ensure we capture all these traits or tags, so I created a new column to capture the **count** of traits/tags per player and convert each trait/tag into a dummy variable.

Remaining Categorical Variables

Lastly, I converted the remaining categorical columns from the dataset into dummy variables, including a players preferred_foot, work_rate, international_reputation, skill_moves, and body_types.

Stats Model

Now we have developed all the necessary features; we can proceed with building the model. The first step of the process was to use the statsmodel package to eliminate irrelevant/ insignificant features. This is to identify the variables with the highest p-values to remove them and help optimize the model manually. If we don't do this step and look to build models using scikit learn, redundant variables will be kept and slow the model down.

First, we create the train/test split data sets. We want to ensure we are training the model on a portion of our data and then saving the remaining sample for test/validation. A reminder the target (`Y`) variable we are predicting is potential.

We start with 173 columns. After running the model through statsmodel, we go through eliminating variables for variables over 0.05, which is the level of significance set. At the end of the process, we remove 64 redundant columns and keep the remaining 109 columns. The results recorded a relatively high R-square of 0.854. With the optimized model after removing the high p-values, we can feed in the same inputs into the scikit learn OLS package and obtain prediction results.

Scikit Learn - OLS

Next, we feed the significant variables from the statsmodel into the scikit learn linear regression model (OLS). Ordinary Least Squares regression is one of the more common practices for modeling. It studies the relationship from the X variables and how it affects the values y - in our case, potential. This model is a good starting point to see how it performs and see if there is anywhere that needs to be pivoted to obtain a higher-performing model based on the features. OLS is not the only optimization strategy. It is the most popular for this kind of task since the regression outputs (that are, coefficients) are unbiased estimators of the real values.

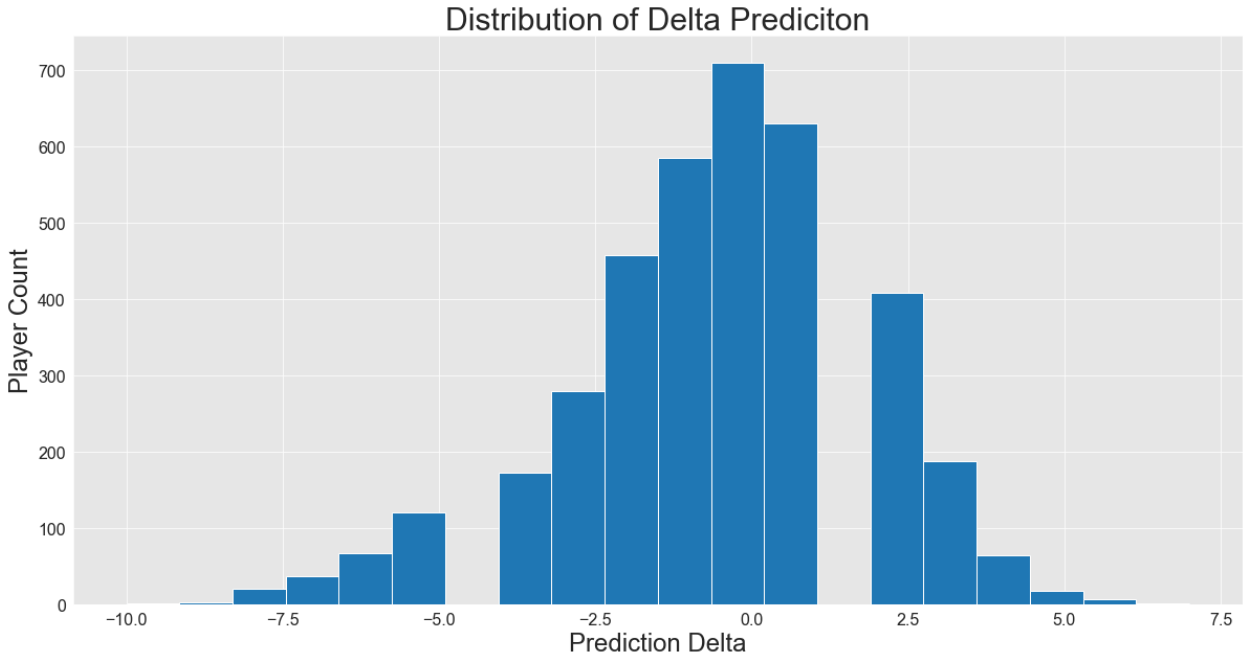
- Regression Training Score: 0.8538
- Regression Testing Score: 0.8487

Visualization Results

We obtained the y_pred variable, which is essentially the linear regression equation, combining all the coefficients we saw from the statsmodel stage into a simplified function to calculate the test model results automatically. While we can display the entire regression equation, since there are 100+ variables it would be difficult to interpret each of them. Below are some takeaways from a select few coefficient impacts on a players potential.

- For every 1 increase in the trait 'poacher' expect a 2.1 increase in potential
- For every 1 increase in the trait 'tactician' expect a 1.3 increase in potential
- For every 1 increase in the a players 'age' expect a -2.4 decrease in potential
- For every 1 increase in the a trait 'selfish' expect a -0.32 decrease in potential

Below we can visualize what the delta is between the actual prediction score and the model's prediction score.



From this histogram, we can see that 50% of the model's prediction fell within -1 to 1 from the actual potential value. While this is certainly not perfect, it gives a good sense that there are not that many players with astronomically incorrect predictions. The histogram shows that the model's worse prediction was a -9 delta from the original potential rating for only 1 player.

Appendix

- Exploration Notebook (Python)
- Model Notebook (Python)
- FIFA 20 Dataset