Final Report of Fundamental Data Science

Data Analysis and Comparison of Traditional Against Modern Goalkeeping Styles in Football

Made By

Abdul Moiz

Nicholaus Santo

Group Members and Roles

- **Abdul Moiz**
  - Documentation
  - Dateline Management
  - Presentation Assistant
- **Nicholaus Santo Agnus Dei - 2602174415**
  - Lead Coder
  - Data Gathering and Processing
  - Report Building and Finalization
  - Presentation Leader

1. Problem Analysis

a. The Problem That We Want to Solve

In this day and age, football is a very big sport with a very big community. On a daily basis, there can be a multitude of football games that are played every day. Each and every football game that is played has its own unique results and outcomes with each of its own unique data and statistics.

As there are more games and more technology that are related to the said games, the culture of football is starting to change as well. It can be said the game of football is being more advanced and a lot of changes are happening because of it. One of those changes can be clearly seen very obviously in goalkeepers. Throughout the years, the role of being a goalkeeper is no longer the same as more tactics and strategies are being developed to be more advanced and modern. Some of the obvious changes that can be seen in goalkeepers would include the likes of passing, ball playing, distributions, and more overall control of the ball with their feet and not just with their hands. These changes and demands have forced many traditional goalkeepers to be more modernized as well as to keep their relevance in the game of football.

Each type of goalkeeper has their own statistics and data regarding their performances and the amount of achievements, wins, and trophies that they have accumulated throughout their football careers. Those data can be used as a telltale of what can happen and the chances of stuff happening in a game based on previous games that have been played or other games that are similar, hence the importance of those statistics for the people such as the fans, coaches, trainers, and even the player themselves. However, on paper, those statistics are just numbers and letters. To discern those numbers and letters, it would take a long time and it would be very hard to visualize them based on just numbers and letters. Therefore, we are going to make an analysis and visualization of those said types of data and we will use those said data to predict if a player would be able to win or earn a trophy in the following season alongside a comparison of the types of goalkeepers that have emerged in the following days of football.

2. Related Works

a. **Football Dribbling Skills with Elo System**

https://towardsdatascience.com/evaluating-football-dribbling-skill-by-utilizing-theelo-algorithm-9c6aa384b991

b. **Current Best Striker in Football**

https://towardsdatascience.com/i-need-a-striker-for-my-team-who-is-he-going-to-beexamining-the-dataset-in-tableau-187e4c3f9692

c. **Game Predictions**

https://towardsdatascience.com/epl-analysis-and-gameweek-22-prediction91982b809802

d. **Comparison of Football Team Performance**

https://www.footballytics.ch/post/analytics-practice-compare-team-performancefairly

e. **Comparison of Football Players Performance**

https://www.footballytics.ch/post/data-analytics-practice-comparing-players-fairly

3. Dataset and Preprocessing

a. **The Data**

The data that we used can be found in this website:

https://fbref.com/en/

Using the said website, we have found and exported statistics regarding performances of various goalkeepers based on the conditions that we wanted to visualize and compare. We have mainly gathered their performances throughout their careers that would also include statistics, such as passing, shots faced, matches played, trophies earned and won, clean sheets, saves, goals conceded, and many alike.

b. Preprocessing

The data that we have gathered consisted of around 10 goalkeepers that are deemed to possess or would fit the modern style of goalkeeping and another 10 goalkeepers that are deemed as traditional goalkeepers. Examples of one of the dataset for a goalkeeper would be as follow:

Season	Age	Squad	Country	Comp	LgRank	MP	Starts	Min	90s	GA	GA90	SoTA	Saves	Save%	W	D	L	CS	CS%	PKatt	PKA	PKsv
2014-2015	18	Southport	eng ENG	5. Conf Pre	19th	16	16	1,440	16	31	1.94	107	76	71	6	2	8	5	31.3			
2014-2015	18	Blackburn	eng ENG	2. Champi	9th	2	2	180	2	2	1	4	2	50	2	0	0	1	50			
2015-2016	19	Blackburn	eng ENG	2. Champi	15th	5	5	450	5	5	1	12	7	58.3	0	3	2	1	20			
2016-2017	20	Blackburn	eng ENG	2. Champi	22nd	5	5	450	5	2	0.4	17	15	88.2	3	2	0	3	60	0	0	0
2017-2018	21	Blackburn	eng ENG	3. League (	2nd	45	45	4,050	45	39	0.87	175	136	78.9	28	12	5	0	0	2	2	0
2018-2019	22	Blackburn	eng ENG	2. Champi	15th	41	41	3,690	41	64	1.56	180	118	66.7	13	11	17	10	24.4	7	4	1
2019-2020	23	Brentford	eng ENG	2. Champi	3rd	46	46	4,140	46	38	0.83	136	99	74.3	24	9	13	16	34.8	3	3	0
2020-2021	24	Brentford	eng ENG	2. Champi	3rd	42	42	3,780	42	36	0.86	126	91	74.6	23	14	5	16	38.1	4	4	0
2021-2022	25	Brentford	eng ENG	1. Premier	13th	24	24	2,160	24	27	1.13	103	76	77.7	10	5	9	8	33.3	4	4	0
2022-2023	26	Brentford	eng ENG	1. Premier	9th	38	38	3,420	38	46	1.21	197	154	77.7	15	14	9	12	31.6	2	2	0
2023-2024	27	Arsenal	eng ENG	1. Premier	4th	15	15	1,350	15	16	1.07	39	22	61.5	8	3	4	5	33.3	2	1	1

The data image above is a collection of statistics from a modern goalkeeper named David Raya, and there are more of these types of statistics and files that need to be merged together into one dataset that can be used to work with properly without any ambiguity alongside cleaning the dataset that are filled with NaN values and ambiguous repetition.

Table 1: Dataset description of the passing statistics.

Season	The season of when the players played the games
Age	The players age correlating to the season that they play
Squad	The club that the players are apart of during the season they play
Country	The country that the club recedes during the season that they play
Comp	The type of competition or tournament that the club and the players partake in the season that they play
LgRank	The final finishing spot of the players' club in the season that they play
90s	The amount of time the players have played a full 90 minute game of football respective to the season that they play

Cmp	The number of completed passes that the players have made throughout the whole season that they played
Att	The number of attempted passes that the players have made throughout the whole season that they played
Cmp%	The percentile of the completed passes compared to the attempts respective to the season that they play
TotDist	The total distance that are covered by the passes made by the players respective to the season that they play in measurements of yards
PrgDist	The progressive distance that are covered by the passes made by the players respective to the season that they play in measurements of yards
Awards	The indication of whether the players have achieved an award at the end of the season that they play.

Table 2: Dataset description of the shot stopping statistic.

MP	The amount of matches that are played by the player respective to the season that they play
Starts	The amount of matches that the player started in the season that they play
Min	The cumulative amount of minutes that the players have played in the season that they played
GA	The amount of goals that are conceded by the player in the season that they play
GA90	The ratio for the goals that the player conceded for every 90 minute that they play respective to the season that they play
SoTA	The amount of shots that are faced by the player in the season that they play
Saves	The amount of saves that the player has made in the season that they play
Save%	The percentile of the saves that they made respective to the season that they played

W	The amount of wins that the player has achieved in the season that they played
D	The amount of draws that the player has achieved in the season that they played
L	The amount of losses that the player has achieved in the season that they played
CS	The amount of clean sheets that the player has achieved in the season that they play
CS%	The percentage of clean sheets that the player has against the matches that they play in the season that they play
PKatt	The amount of penalty kicks that the player has faced in the season that they play
PKA	The amount of penalty kicks that the player has conceded in the season that they play
PKsv	The amount of penalty kicks that the player has saved in the season that they play
PKm	The amount of penalty kicks where the kicker missed the penalty against the player in the season that they play
Top	The indication of whether the player is seen or awarded as the best performer in the season that they play

4. Model and Techniques

a. Models / Modules:

i. Python

The python language will be the main language that can be used to create and code the visualization as the language is easy to work with and provides many resources regarding data analysis.

ii. **Matplotlib**

Matplotlib is an open library that can be utilized with python to create static, animated, and interactive visualizations of a dataset.

iii. **Pandas**

Pandas is also a library similar to matplotlib with the difference being that pandas are made and used for manipulating data and analyzing them. Not for visualization.

iv. **Sklearn (Classifiers and Regressors)**

The Sklearn module provides classification and regression techniques that allows us to train the machine learning model to be able to predict the trophy winning performance that are set by the goalkeepers. By having two models, we can compare whether it is better to use classifiers or regressors to train the machine.

v. **Shap**

The shap python module allows us to see which features are effective for training the machine and visualizes the effectiveness of those features.

b. **Techniques:**

i. **Bar Charts**

The bar charts are used to visualize the amount of the number of passes, saves, clean sheets, and other contributing attributes from the goalkeepers of each playstyle.

ii. **Line Charts**

The line charts are used to visualize the difference, especially in a one to one direct comparison between the modern goalkeepers against the traditional ones. They are also used to help the bar charts in visualizing the values in an easier way.

iii. **Pie Charts**

The pie charts were most-often used as a way to show the preference or the playstyle of the goalkeeper.

5. Evaluation Methods

a. **Precision**

The precision of the predictive model or the trained machine will be evaluated to be able to find the true positives out of the true and false positives. The closer number is to 1, the better the machine performs

b. **Accuracy**

The accuracy is the total number of times that the machine predicted correctly regarding the data and the more the data is predicted correctly, the better the machine performs. The accuracy rate is also measured by the number 1 and the closer it is to the better the accuracy of the machine or the model is.

c. **Recall Score**

The recall score is a metric that is used to find the correct answers out of all the correct answers and the false incorrect answers. The model is also expected to have a score that is closer to 1.

d. **F1 Score**

The F1 score is a metric that is used to find the balance of the recall and the precision of the model or the average performing rate of the model based on the score of the recall and the precision. A number that is closer to 1 is also expected and considered to be better.

6. Results

a. Processed Dataset

i. Modern Goalkeepers

(1 - 30, the full dataset can be found in the repository)

	Season	Age	Squad	Country	Comp	LgRank	MP	Starts	Min	90s	GA	GA90	SoTA	Saves	Save%	W	D	L	CS	CS%	PKatt	PKA	P
0	2014	21	Internacio	brÂ BRA	1.Â SÃ©rie	3rd	11	11	990	11	0	0	0	0	0	7	1	3	2	18.2	1	1
1	2015	22	Internacio	brÂ BRA	1.Â SÃ©rie	5th	26	26	2,266	25.2	0	0	0	0	0	12	7	7	15	57.7	4	3
2	2016	23	Internacio	brÂ BRA	1.Â SÃ©rie	17th	1	1	90	1	0	0	1	1	100	0	1	0	1	100	0	0
3	2017-2018	24	Roma	itÂ ITA	1.Â Serie A	3rd	37	37	3,330	37	28	0.76	135	105	81.5	22	8	7	17	45.9	5	3
4	2018-2019	25	Liverpool	engÂ ENG	1.Â Premie	2nd	38	38	3,420	38	22	0.58	96	74	77.1	30	7	1	21	55.3	1	0
5	2019-2020	26	Liverpool	engÂ ENG	1.Â Premie	1st	29	29	2,543	28.3	23	0.81	80	58	72.5	23	3	3	13	44.8	1	1
6	2020-2021	27	Liverpool	engÂ ENG	1.Â Premie	3rd	33	33	2,970	33	32	0.97	115	82	75.7	18	8	7	10	30.3	8	4
7	2021-2022	28	Liverpool	engÂ ENG	1.Â Premie	2nd	36	36	3,240	36	24	0.67	99	76	75.8	27	7	2	20	55.6	0	0
8	2022-2023	29	Liverpool	engÂ ENG	1.Â Premie	5th	37	37	3,330	37	43	1.16	147	105	72.1	19	9	9	14	37.8	4	2
9	2023-2024	30	Liverpool	engÂ ENG	1.Â Premie	1st	18	18	1,620	18	15	0.83	62	48	77.4	11	6	1	6	33.3	1	1
10	2006-2007	23	Real Socie	esÂ ESP	1.Â La Liga	19th	29	29	2,610	29	29	1	151	122	80.8	8	7	14	8	27.6	0	0
11	2008-2009	25	Real Socie	esÂ ESP	2.Â Segunc	6th	32	32	2,880	32	28	0.87	104	76	73.1	14	11	7	13	40.6	0	0
12	2009-2010	26	Real Socie	esÂ ESP	2.Â Segunc	1st	25	25	2,156	24	22	0.92	110	88	80	13	7	4	8	32	0	0
13	2010-2011	27	Real Socie	esÂ ESP	1.Â La Liga	15th	38	38	3,420	38	66	1.74	205	139	67.8	14	3	21	9	23.7	0	0
14	2011-2012	28	Real Socie	esÂ ESP	1.Â La Liga	12th	37	37	3,330	37	51	1.38	195	144	73.8	12	10	15	12	32.4	0	0
15	2012-2013	29	Real Socie	esÂ ESP	1.Â La Liga	4th	31	31	2,790	31	40	1.29	142	102	71.8	16	10	5	9	29	0	0
16	2013-2014	30	Real Socie	esÂ ESP	1.Â La Liga	7th	37	37	3,330	37	54	1.46	178	124	69.7	16	11	10	12	32.4	0	0
17	2014-2015	31	Barcelona	esÂ ESP	1.Â La Liga	1st	37	37	3,330	37	19	0.51	89	70	78.7	30	3	4	23	62.2	0	0
18	2015-2016	32	Barcelona	esÂ ESP	1.Â La Liga	1st	32	32	2,878	32	22	0.69	107	86	80.4	24	4	4	16	50	1	1
19	2016-2017	33	Barcelona	esÂ ESP	1.Â La Liga	2nd	1	1	90	1	2	2	2	0	0	1	0	0	0	0	0	0
20	2016-2017	33	Mancheste	engÂ ENG	1.Â Premie	3rd	22	22	1,968	21.9	26	1.19	60	34	56.7	12	5	5	5	22.7	1	0
21	2017-2018	34	Mancheste	engÂ ENG	1.Â Premie	1st	3	2	226	2.5	1	0.4	5	4	80	2	0	0	2	100	0	0
22	2019-2020	36	Mancheste	engÂ ENG	1.Â Premie	2nd	4	3	347	3.9	7	1.82	15	8	53.3	2	0	1	1	33.3	0	0
23	2020-2021	37	Betis	esÂ ESP	1.Â La Liga	6th	20	20	1,800	20	25	1.25	75	50	72	8	8	4	7	35	5	4
24	2021-2022	38	Betis	esÂ ESP	1.Â La Liga	5th	17	17	1,456	16.2	19	1.17	56	39	67.9	8	4	5	5	29.4	1	1
25	2022-2023	39	Betis	esÂ ESP	1.Â La Liga	6th	12	12	1,080	12	9	0.75	42	33	81	5	5	2	4	33.3	1	1
26	2023-2024	40	Betis	esÂ ESP	1.Â La Liga	7th	7	7	630	7	5	0.71	20	16	75	3	4	0	2	28.6	0	0
27	2015-2016	16	Milan	itÂ ITA	1.Â Serie A	7th	30	30	2,628	29.2	29	0.99	107	78	72.9	12	9	7	10	33.3	0	0
28	2016-2017	17	Milan	itÂ ITA	1.Â Serie A	6th	38	38	3,420	38	45	1.18	189	144	78.8	18	9	11	12	31.6	10	5
29	2017-2018	18	Milan	itÂ ITA	1.Â Serie A	6th	38	38	3,420	38	42	1.11	132	91	69.7	18	10	10	12	31.6	3	2

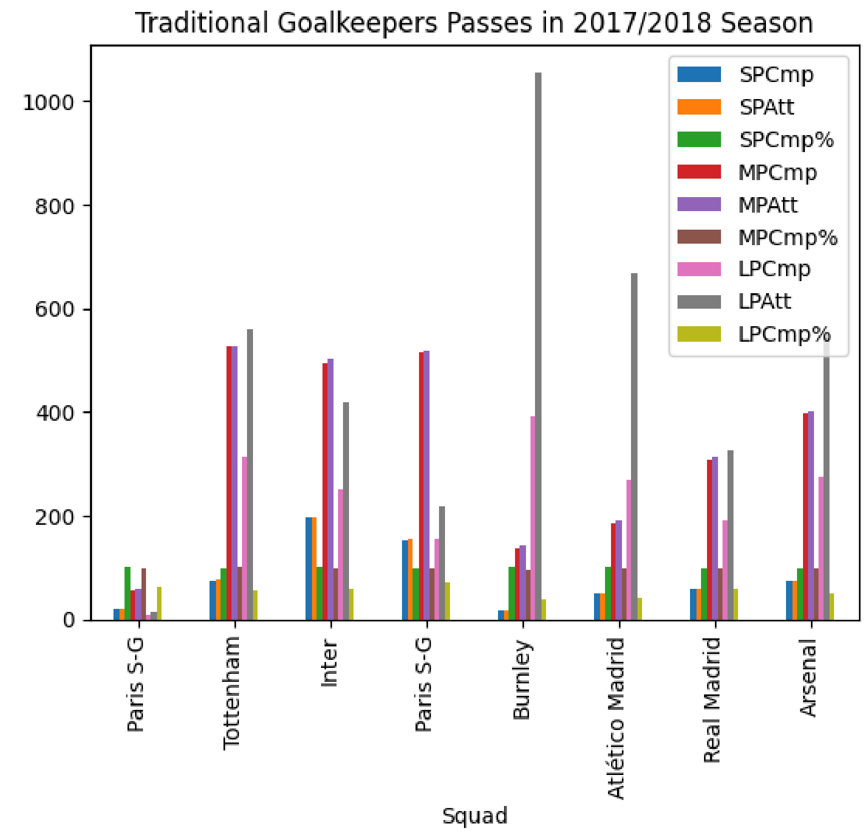
ii. Traditional Goalkeepers

(1 - 30, the full dataset can be found in the repository)

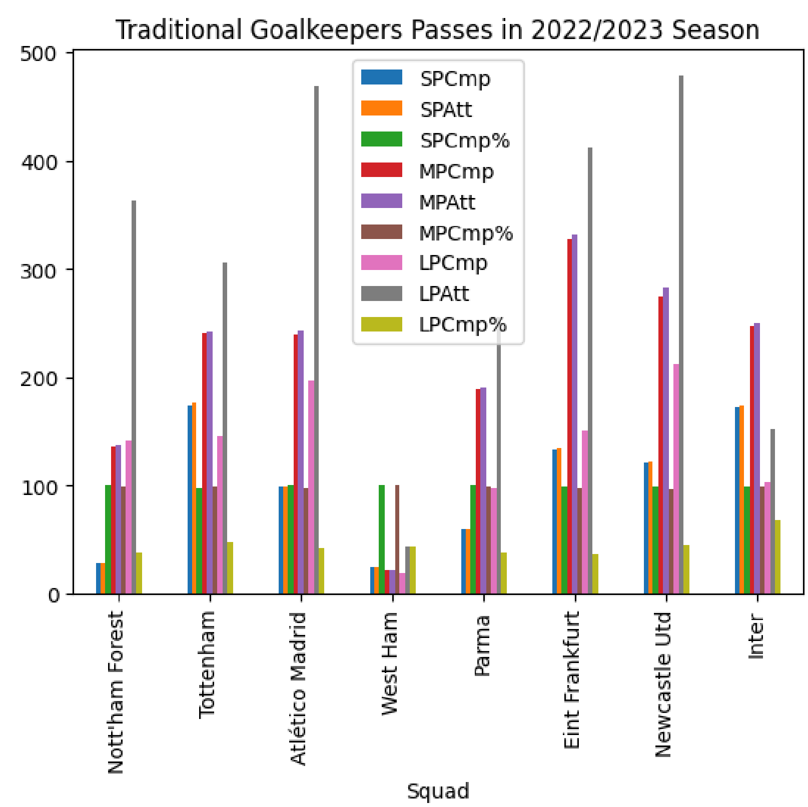
	Season	Age	Squad	Country	Comp	LgRank	MP	Starts	Min	90s	GA	GA90	SoTA	Saves	Save%	W	D	L	CS	CS%	PKatt	PKA	Pl
0	2012-2013	19	Paris S-G	frÂ FRA	1.Â Ligue 1	1st	2	1	103	1.1	1	0.87	4	3	75	0	2	0	1	100	0	0
1	2013-2014	20	Lens	frÂ FRA	2.Â Ligue 2	2nd	34	34	3,030	33.7	33	0.98	119	86	72.3	16	12	5	14	41.2	0	0
2	2014-2015	21	Bastia	frÂ FRA	1.Â Ligue 1	12th	35	35	3,150	35	42	1.2	140	98	70	12	9	14	11	31.4	0	0
3	2015-2016	22	Villarreal	esÂ ESP	1.Â La Liga	4th	32	32	2,880	32	26	0.81	104	80	76	17	8	7	15	46.9	2	1
4	2016-2017	23	Paris S-G	frÂ FRA	1.Â Ligue 1	2nd	15	14	1,297	14.4	14	0.97	35	21	62.9	9	2	3	6	42.9	1	1
5	2017-2018	24	Paris S-G	frÂ FRA	1.Â Ligue 1	1st	34	34	3,060	34	25	0.74	102	77	76.5	26	6	2	17	50	3	1
6	2018-2019	25	Paris S-G	frÂ FRA	1.Â Ligue 1	1st	21	21	1,890	21	17	0.81	66	50	80.3	16	3	2	11	52.4	4	4
7	2019-2020	26	Paris S-G	frÂ FRA	1.Â Ligue 1	1st	3	3	270	3	2	0.67	3	1	33.3	2	0	1	2	66.7	0	0
8	2019-2020	26	Real Madr	esÂ ESP	1.Â La Liga	1st	4	4	360	4	5	1.25	14	9	78.6	3	1	0	1	25	2	2
9	2020-2021	27	Fulham	engÂ ENG	1.Â Premie	18th	36	36	3,240	36	48	1.33	161	114	73.9	5	13	18	9	25	6	6
10	2021-2022	28	West Ham	engÂ ENG	1.Â Premie	7th	1	1	90	1	1	1	3	2	66.7	0	0	1	0	0	0	0
11	2022-2023	29	West Ham	engÂ ENG	1.Â Premie	14th	5	2	309	3.4	7	2.04	16	9	68.8	0	1	1	1	50	2	2
12	2023-2024	30	West Ham	engÂ ENG	1.Â Premie	6th	17	17	1,530	17	24	1.41	91	67	76.9	8	4	5	4	23.5	4	3
13	1998-1999	20	Parma	itÂ ITA	1.Â Serie A	4th	34	34	3,060	34	36	1.06	186	150	80.6	15	10	9	11	32.4	0	0
14	1999-2000	21	Parma	itÂ ITA	1.Â Serie A	5th	32	32	2,880	32	37	1.16	146	109	74.7	14	10	8	12	37.5	0	0
15	2000-2001	22	Parma	itÂ ITA	1.Â Serie A	4th	34	34	3,060	34	31	0.91	146	115	78.8	16	8	10	16	47.1	0	0
16	2001-2002	23	Juventus	itÂ ITA	1.Â Serie A	1st	34	34	3,060	34	23	0.68	134	111	82.8	20	11	3	18	52.9	0	0
17	2002-2003	24	Juventus	itÂ ITA	1.Â Serie A	1st	32	32	2,827	31.4	23	0.73	108	85	78.7	19	9	4	14	43.8	0	0
18	2003-2004	25	Juventus	itÂ ITA	1.Â Serie A	3rd	32	32	2,880	32	41	1.28	132	91	68.9	19	6	7	11	34.4	0	0
19	2004-2005	26	Juventus	itÂ ITA	1.Â Serie A	1st	37	37	3,285	36.5	25	0.68	109	84	77.1	25	8	3	19	51.4	0	0
20	2005-2006	27	Juventus	itÂ ITA	1.Â Serie A	20th	18	18	1,619	18	12	0.67	46	34	73.9	11	7	0	6	33.3	0	0
21	2006-2007	28	Juventus	itÂ ITA	2.Â Serie B	1st	37	37	3,253	36.1	22	0.61	131	109	83.2	25	9	2	20	54.1	0	0
22	2007-2008	29	Juventus	itÂ ITA	1.Â Serie A	3rd	34	34	3,050	33.9	30	0.89	133	103	77.4	19	10	5	16	47.1	0	0
23	2008-2009	30	Juventus	itÂ ITA	1.Â Serie A	2nd	23	23	2,025	22.5	26	1.16	92	66	71.7	11	8	4	8	34.8	0	0
24	2009-2010	31	Juventus	itÂ ITA	1.Â Serie A	7th	27	27	2,378	26.4	34	1.29	102	68	66.7	13	5	8	7	25.9	0	0
25	2010-2011	32	Juventus	itÂ ITA	1.Â Serie A	7th	16	16	1,362	15.1	17	1.12	57	40	70.2	5	6	4	4	25	0	0
26	2011-2012	33	Juventus	itÂ ITA	1.Â Serie A	1st	35	35	3,150	35	16	0.46	97	81	83.5	21	14	0	21	60	0	0
27	2012-2013	34	Juventus	itÂ ITA	1.Â Serie A	1st	32	32	2,880	32	19	0.59	94	75	79.8	23	5	4	16	50	0	0
28	2013-2014	35	Juventus	itÂ ITA	1.Â Serie A	1st	33	33	2,866	31.8	20	0.63	107	87	81.3	28	3	2	20	60.6	0	0
29	2014-2015	36	Juventus	itÂ ITA	1.Â Serie A	1st	33	33	2,970	33	20	0.61	83	63	75.9	23	8	2	18	54.5	0	0

b. Data Models

i. Traditional Goalkeepers

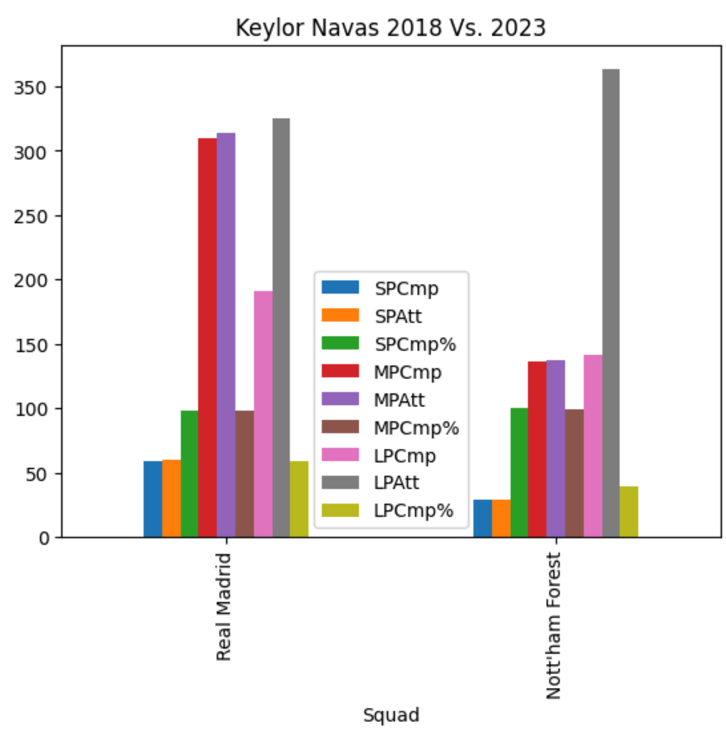

Traditional Goalkeepers Passes in 2017/2018 Season
1000
800
600
400
200
0
SPCmp
SPAtt
SPCmp%
MPCmp
MPAtt
MPCmp%
LPCmp
LPAtt
LPCmp%
Paris S-G
Tottenham
Inter
Paris S-G
Burnley
Atlético Madrid
Real Madrid
Arsenal
Squad

1.


Traditional Goalkeepers Passes in 2022/2023 Season
500
400
300
200
100
0
SPCmp
SPAtt
SPCmp%
MPCmp
MPAtt
MPCmp%
LPCmp
LPAtt
LPCmp%
Nott'ham Forest
Tottenham
Atlético Madrid
West Ham
Parma
Eint Frankfurt
Newcastle Utd
Inter
Squad

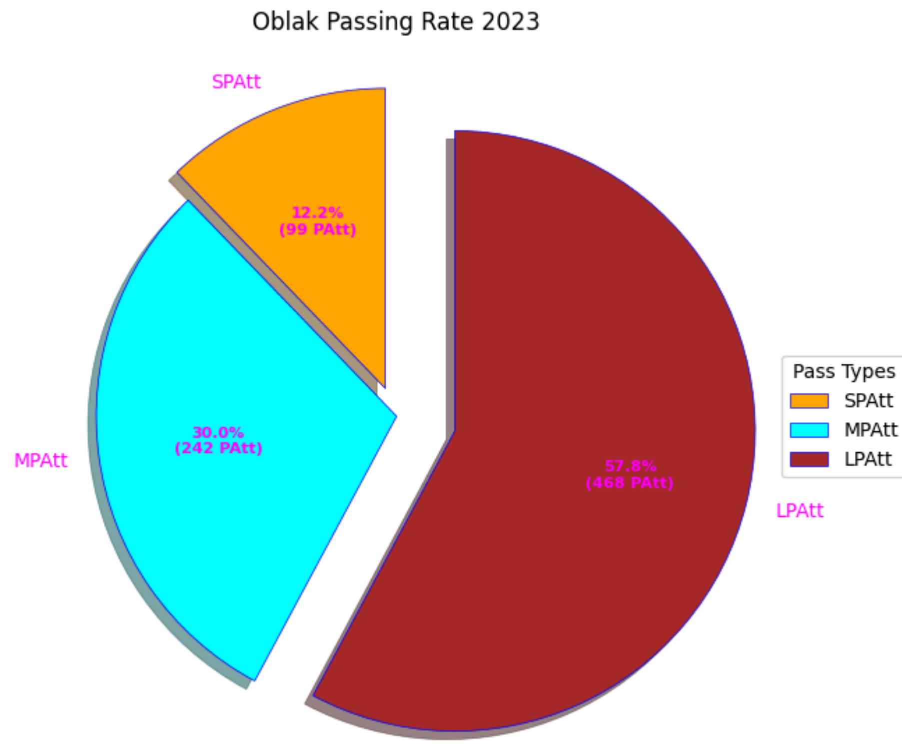
2.

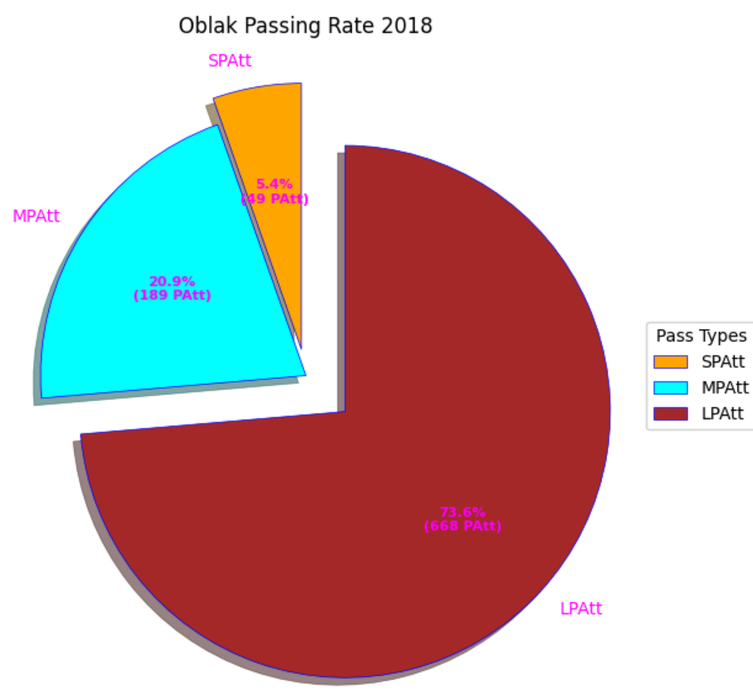
3. **Passes Attempted by Traditional Goalkeepers Over The Last 5 Years**



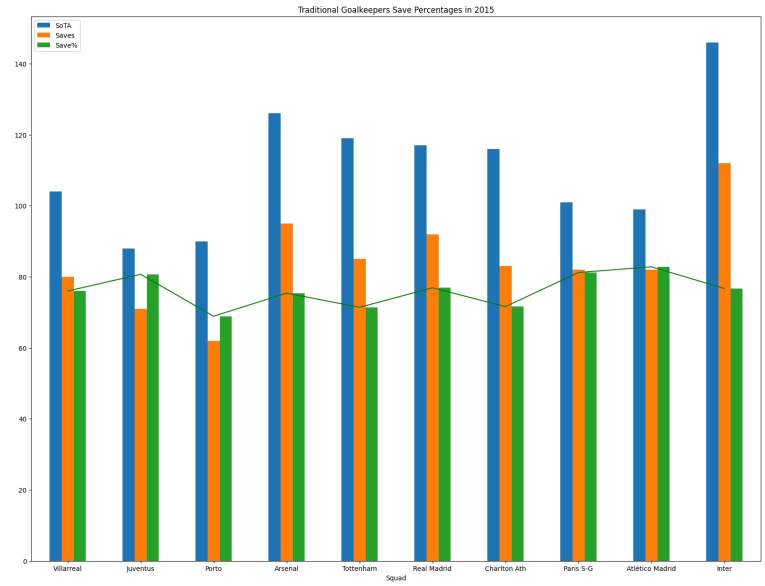
4.

5. Comparison of Player Play Style 2018 vs. 2023

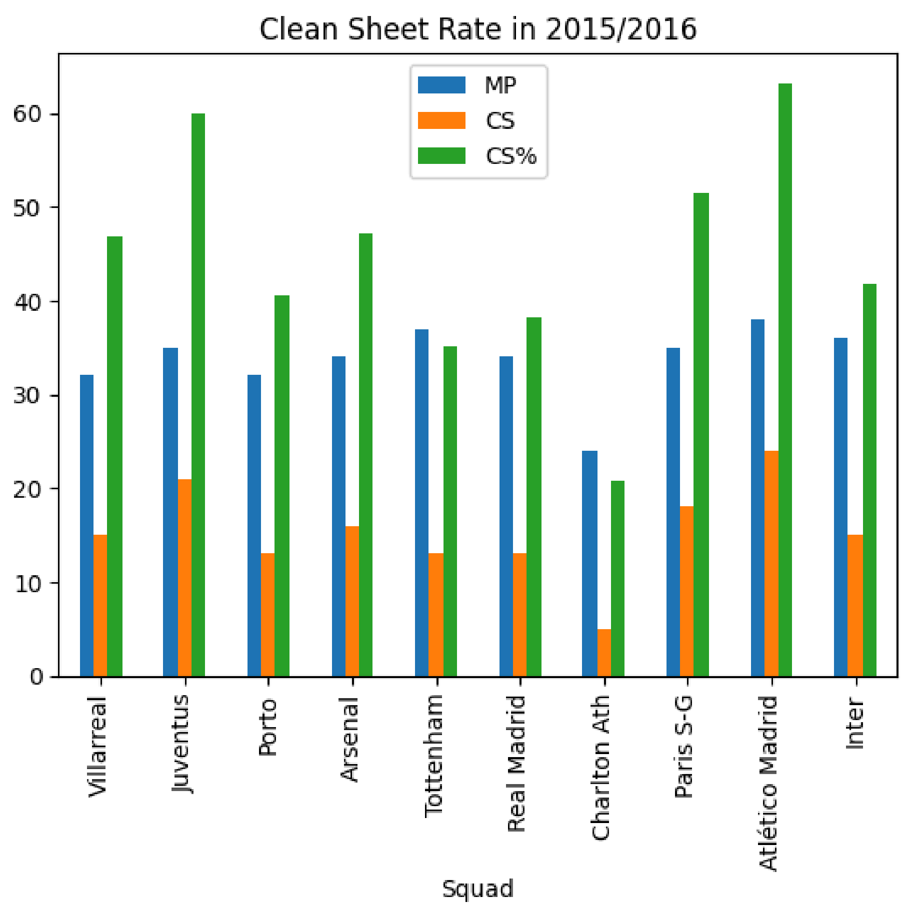




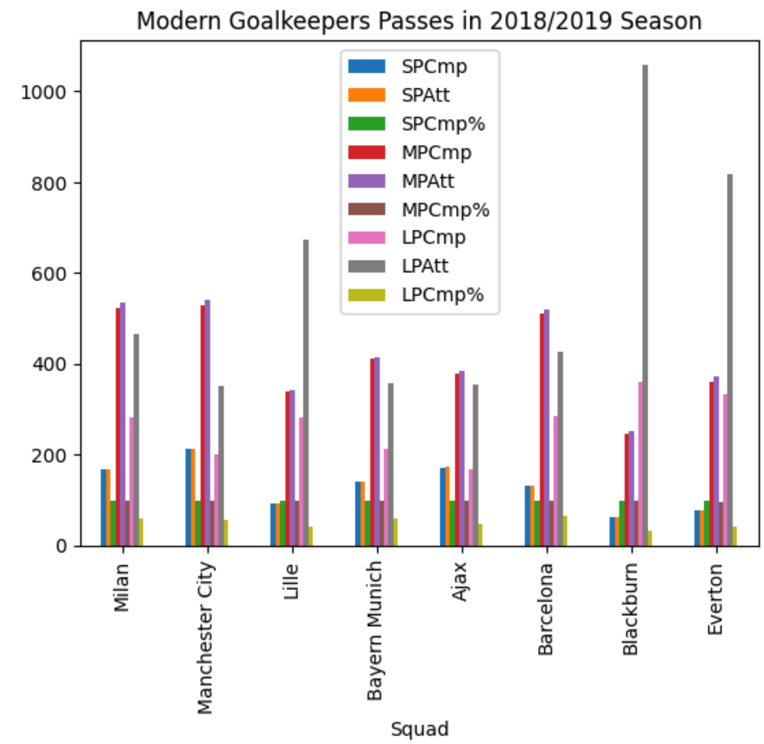
6. Saves Percentage of Traditional Goalkeepers



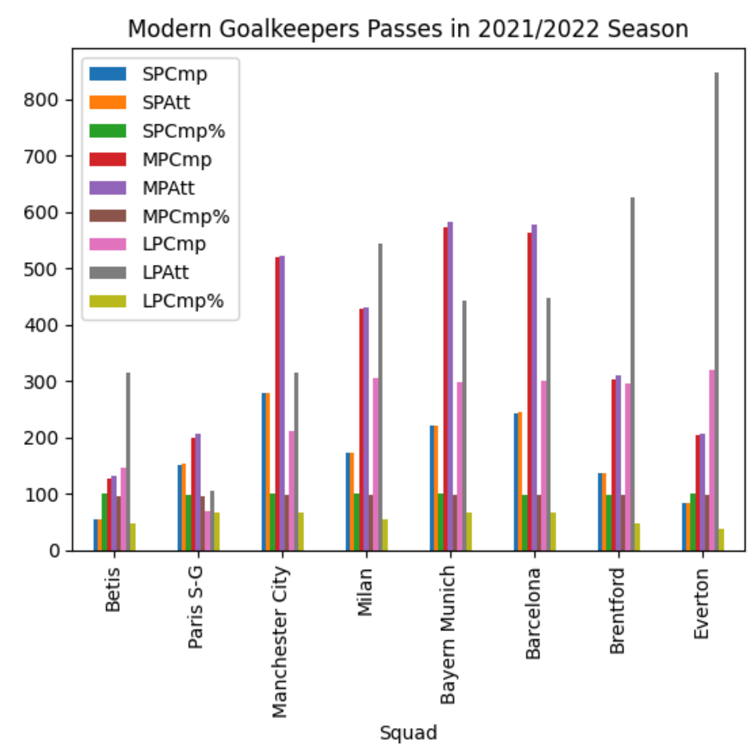
7. Clean Sheet Rate Based of Saves



ii. Modern Goalkeepers

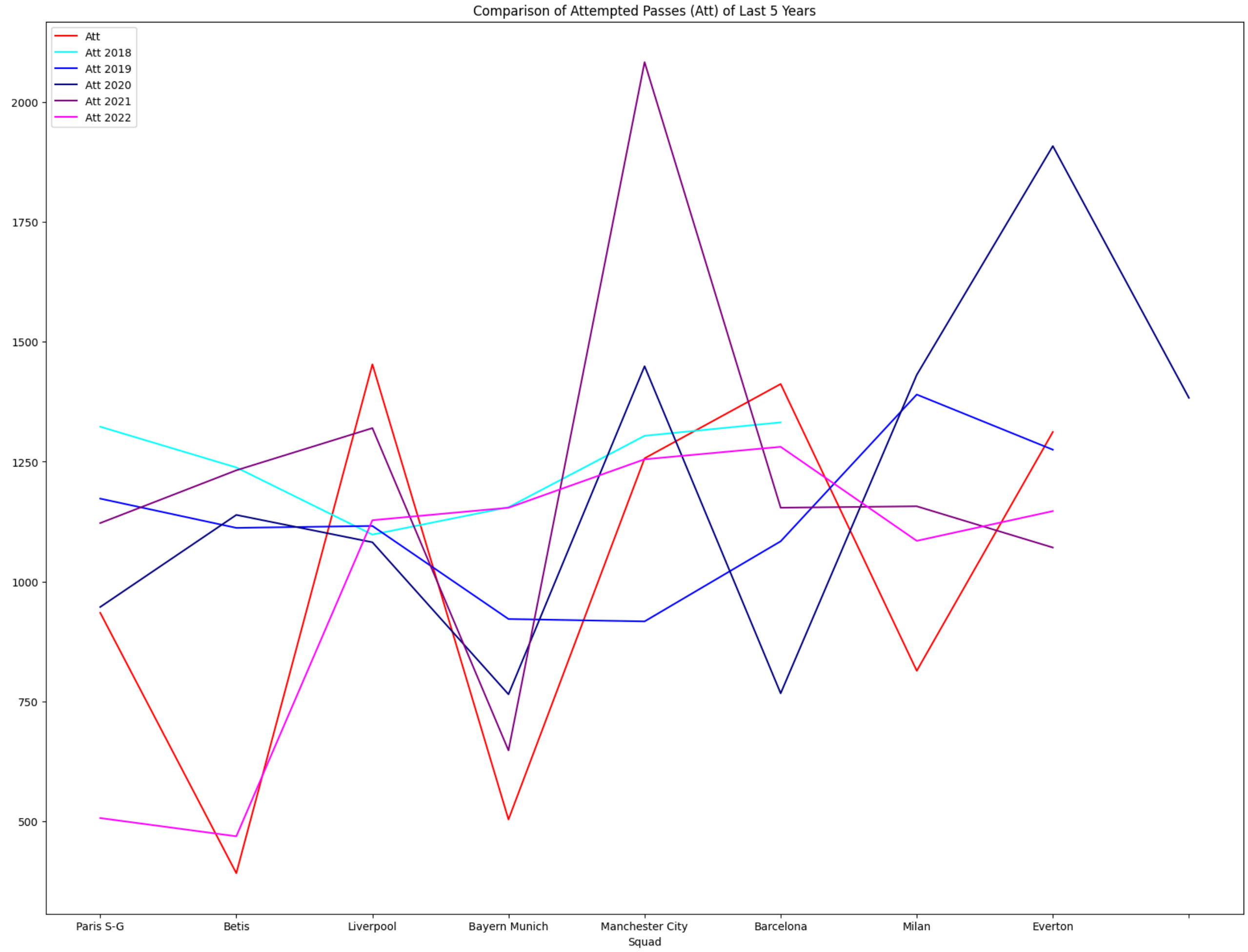


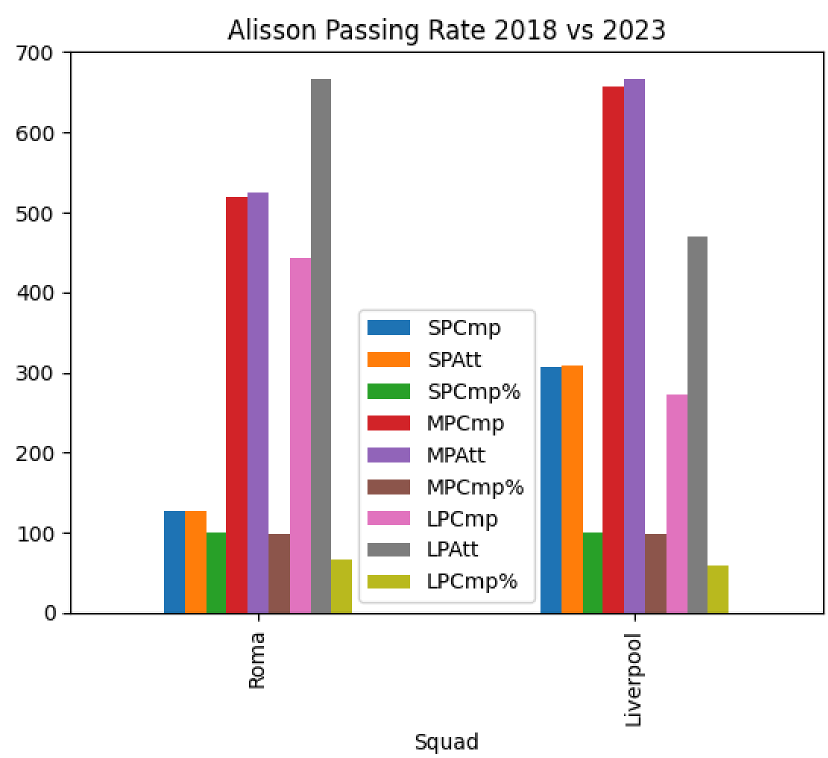
1.



2.

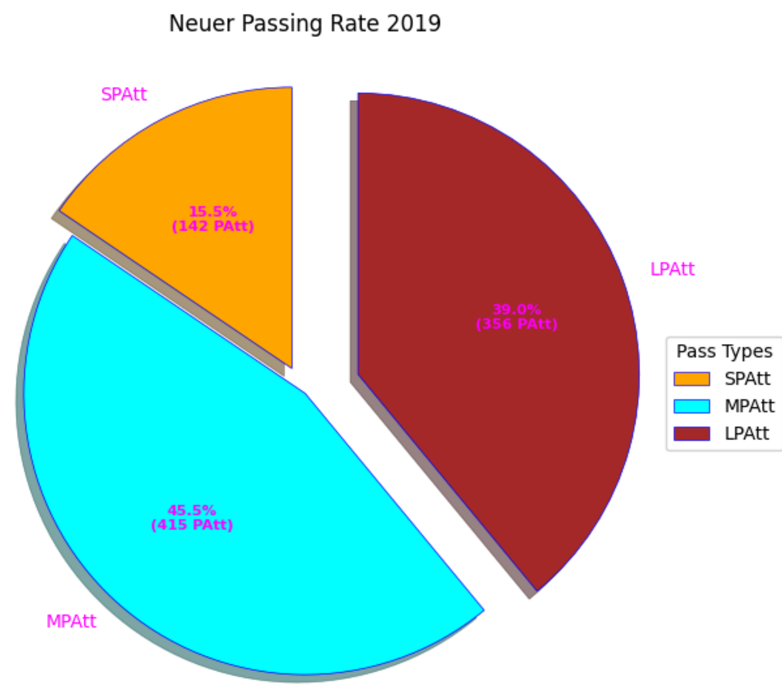
3. **Comparison of Attempted Passes of Modern Goalkeepers**

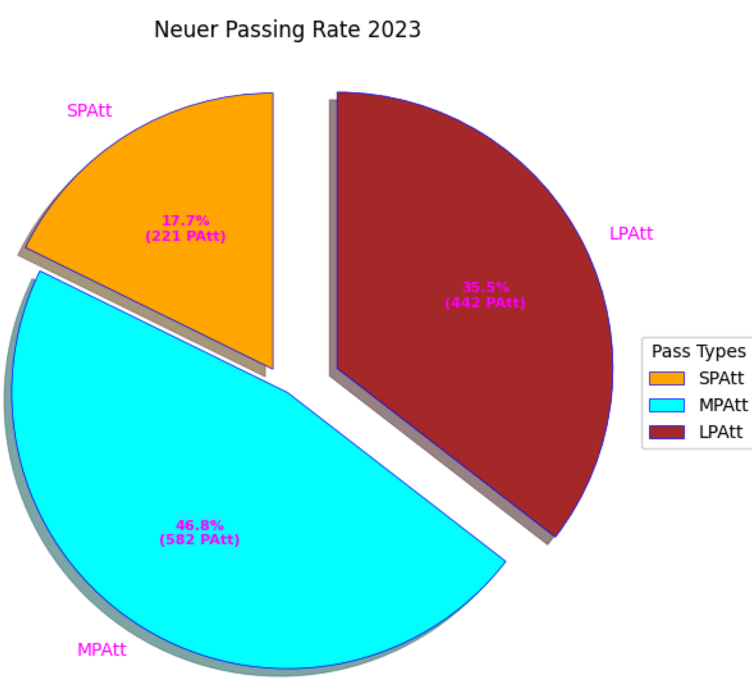

Comparison of Attempted Passes (Att) of Last 5 Years
Att
Att 2018
Att 2019
Att 2020
Att 2021
Att 2022
2000
1750
1500
1250
1000
750
500
Paris S-G
Betis
Liverpool
Bayern Munich
Manchester City
Barcelona
Milan
Everton
Squad


Alisson Passing Rate 2018 vs 2023
700
600
500
400
300
200
100
0
SPCmp
SPAtt
SPCmp%
MPCmp
MPAtt
MPCmp%
LPCmp
LPAtt
LPCmp%
Roma
Liverpool
Squad

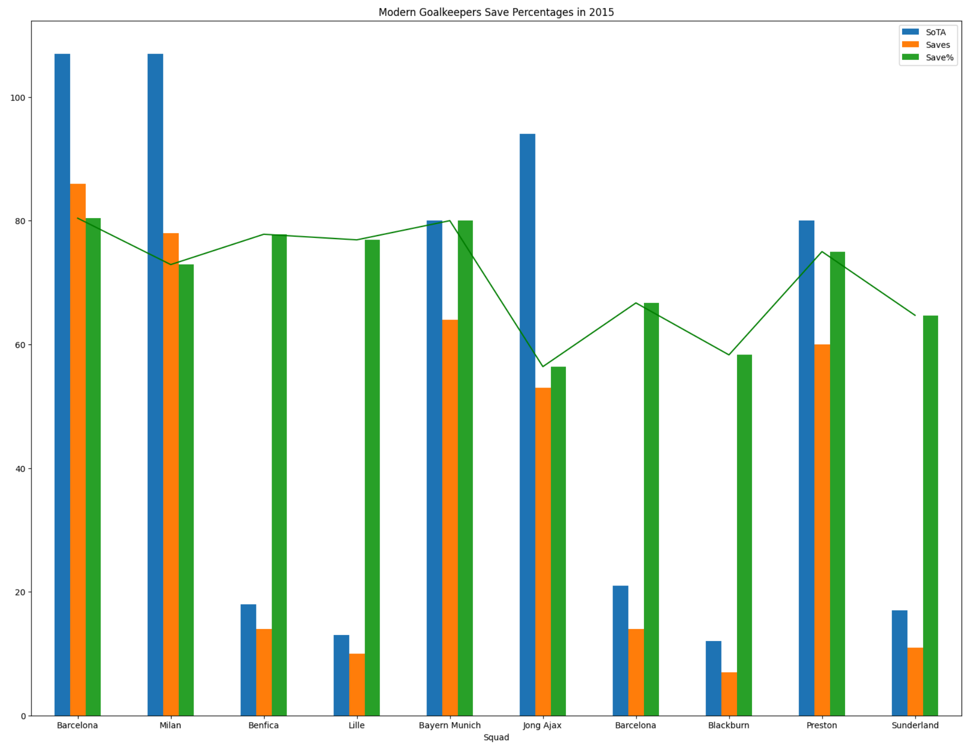
4.

5. Comparison of Player Play Style

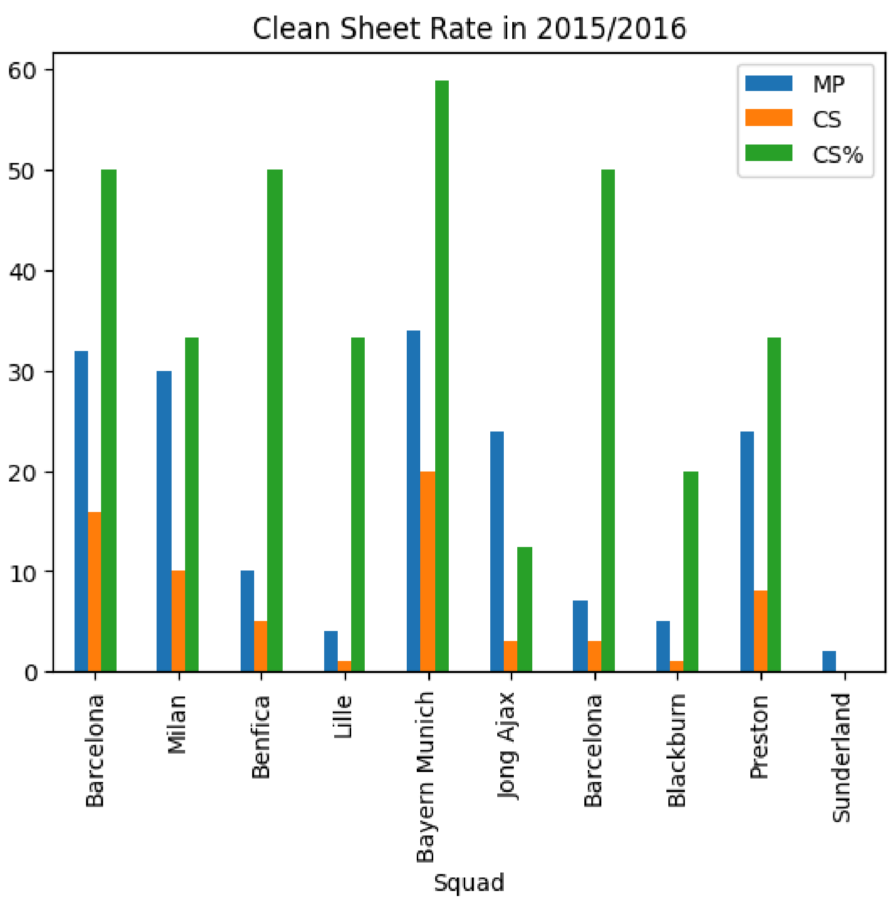




6. Saves Percentage of Modern Goalkeepers

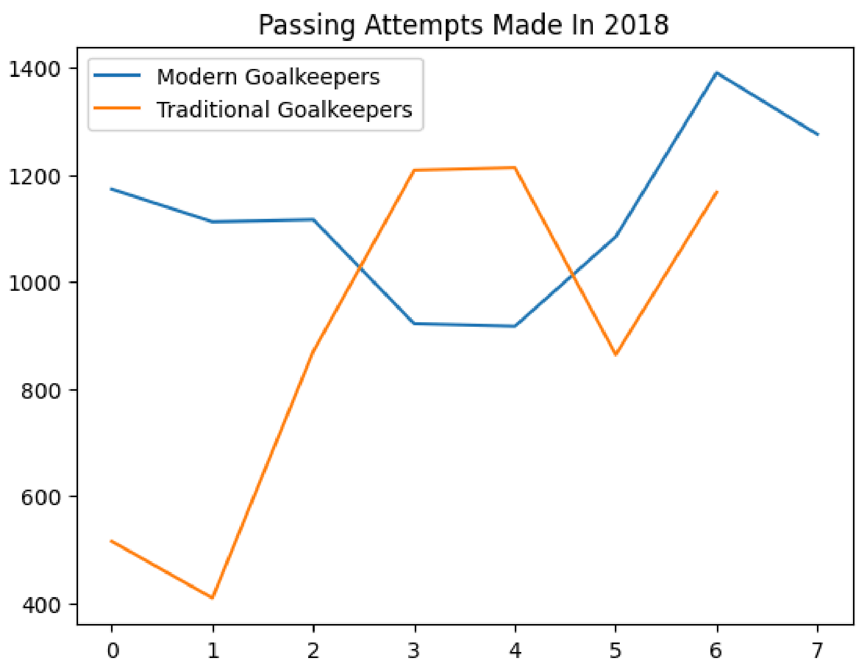


7. Clean Sheets Rate Based of Saves

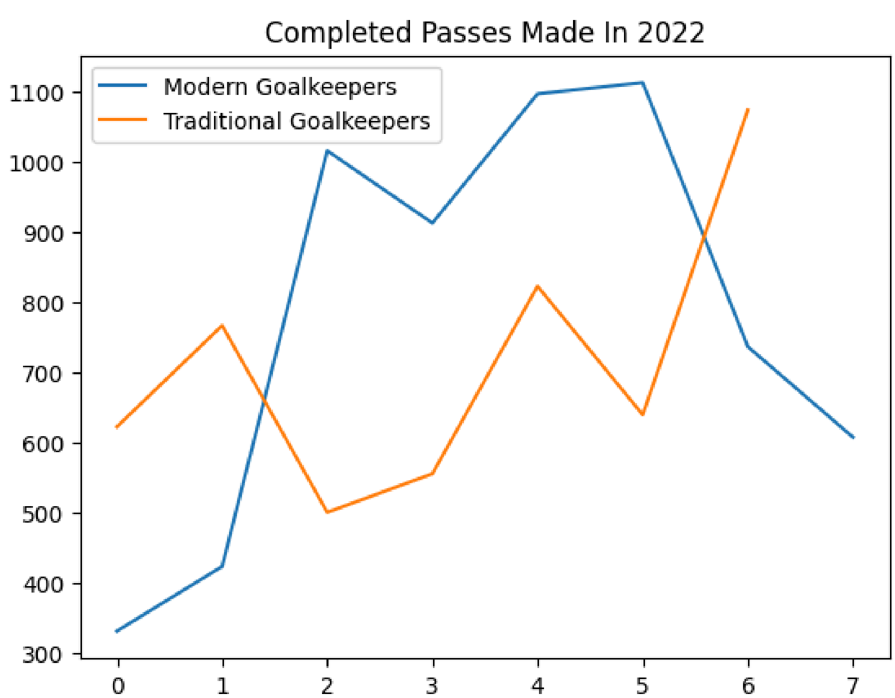


iii. Traditional Vs Modern

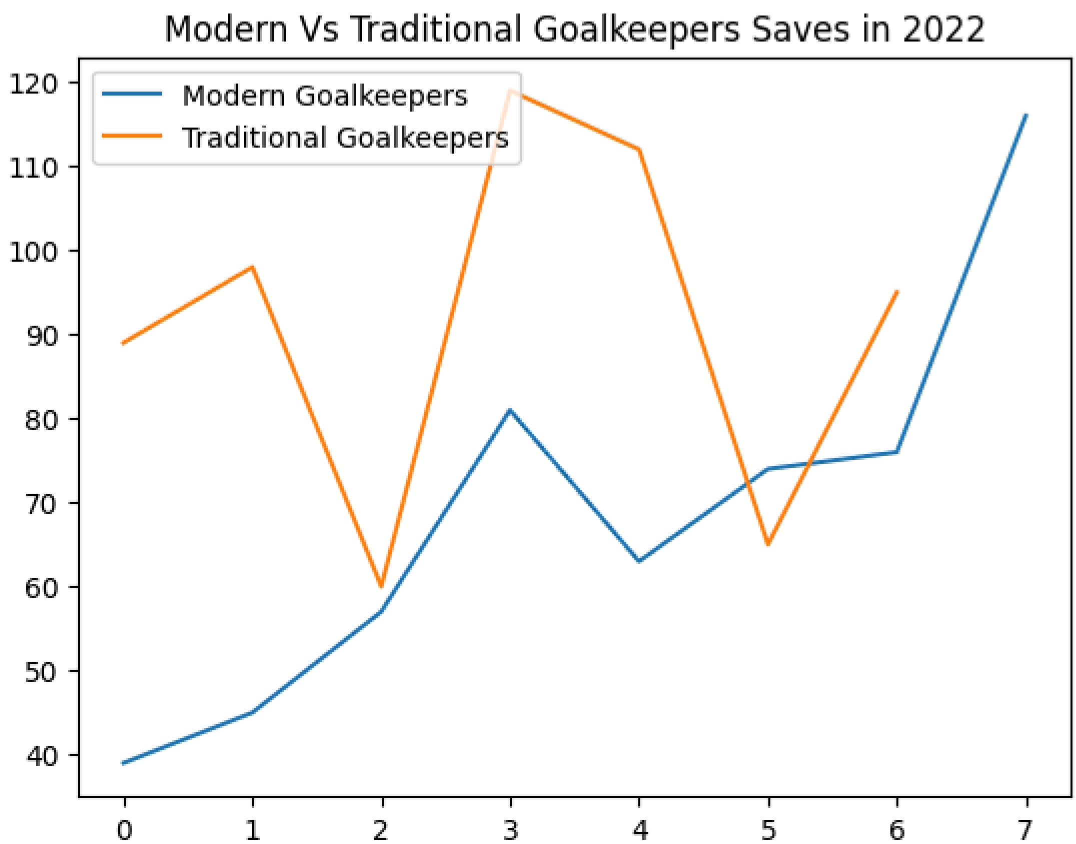
1. Attempted Passes in 2018



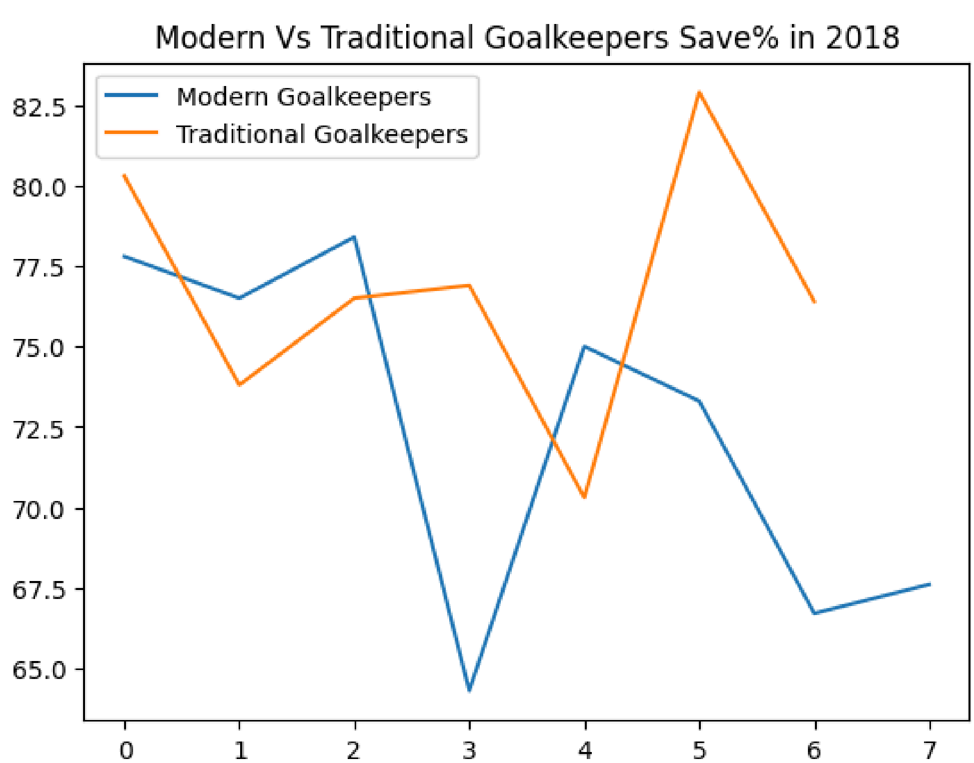
2. Completed Passes in 2022



3. Saves Made in 2022



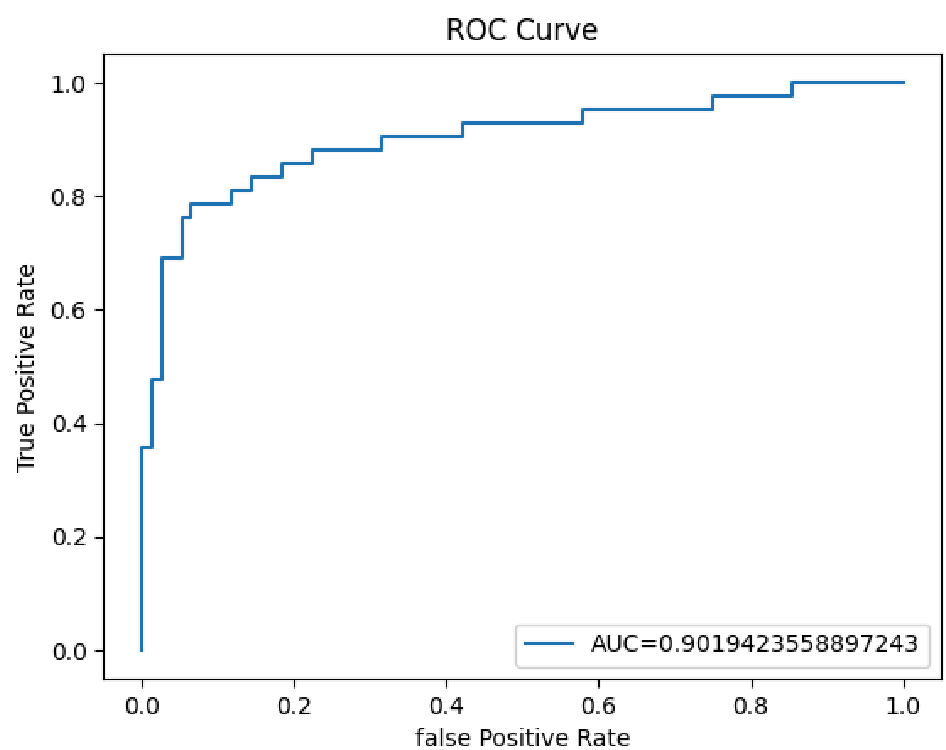
4. Save Percentage in 2018



c. Prediction Model

i. Regressors

The regressor model that we have chosen for our prediction model is the Logistic Regression as it is one of the most common models that is usually used to make a prediction model in python. It is also more fitting for us to use the Logistic Regression model, because our main goal is to make a prediction model that can predict whether they would win or not, which is also similar to black and white, not how many trophies that can be won. The result of the prediction model is as follow:

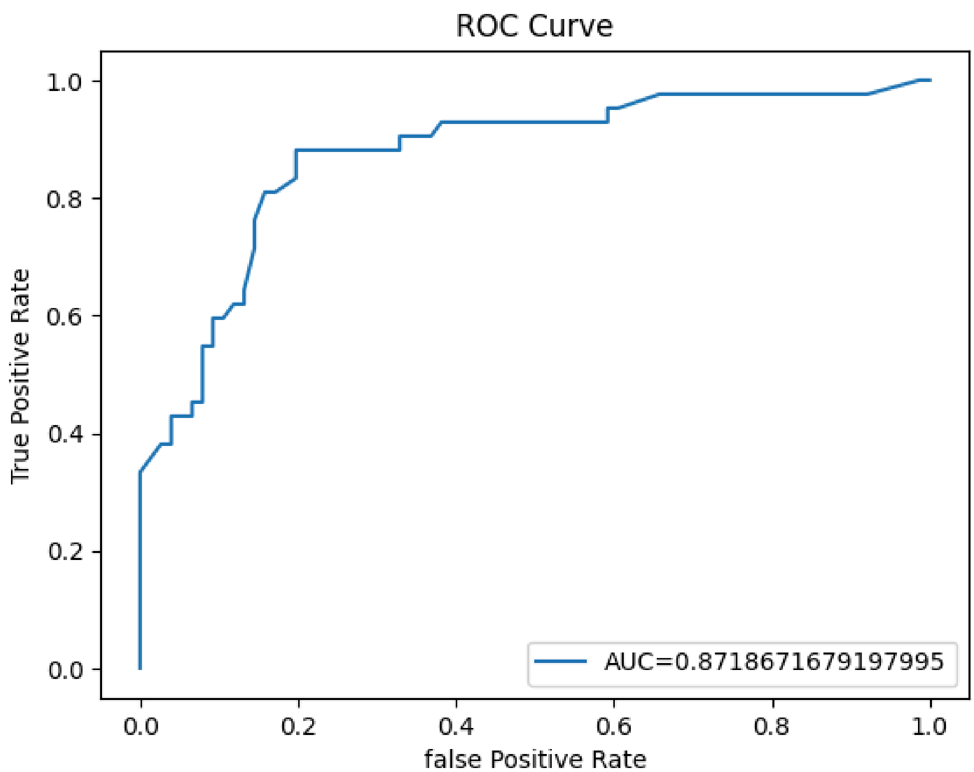


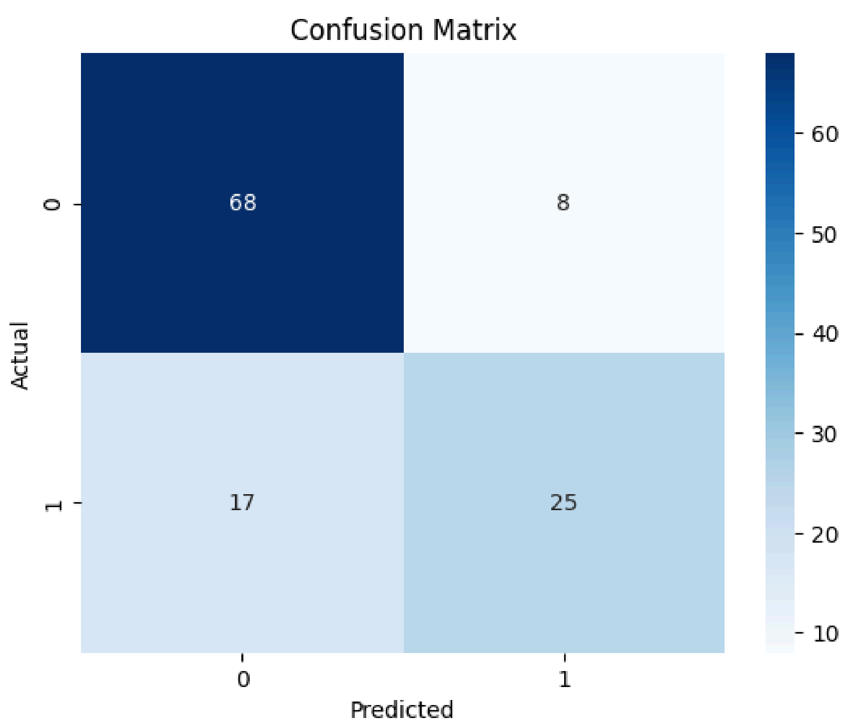
```
Accuracy:  0.8728813559322034
Recall:  0.6904761904761905
Precision: 0.9354838709677419
CL Report:               precision    recall  f1-score   support

           0       0.85      0.97      0.91        76
           1       0.94      0.69      0.79        42

    accuracy                           0.87       118
   macro avg       0.89      0.83      0.85       118
weighted avg       0.88      0.87      0.87       118
```

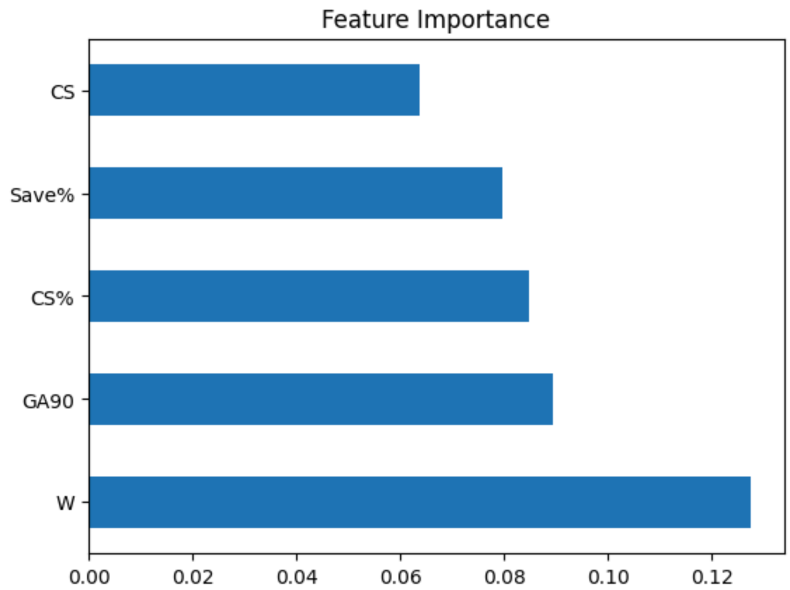
ii. Classifiers and Feature Importances

Another model we used to train the machine is the Random Forest Classifier. The reason why we used the Random Forest Classifier is because of how the Random Forest is good with high accuracy, robustness, feature importance, versatility, and scalability. The Random Forest model combines a collection of decision trees, and each tree in the collection resembles data samples that are drawn from the training set. The result of the classifier alongside the feature importances are as follows:

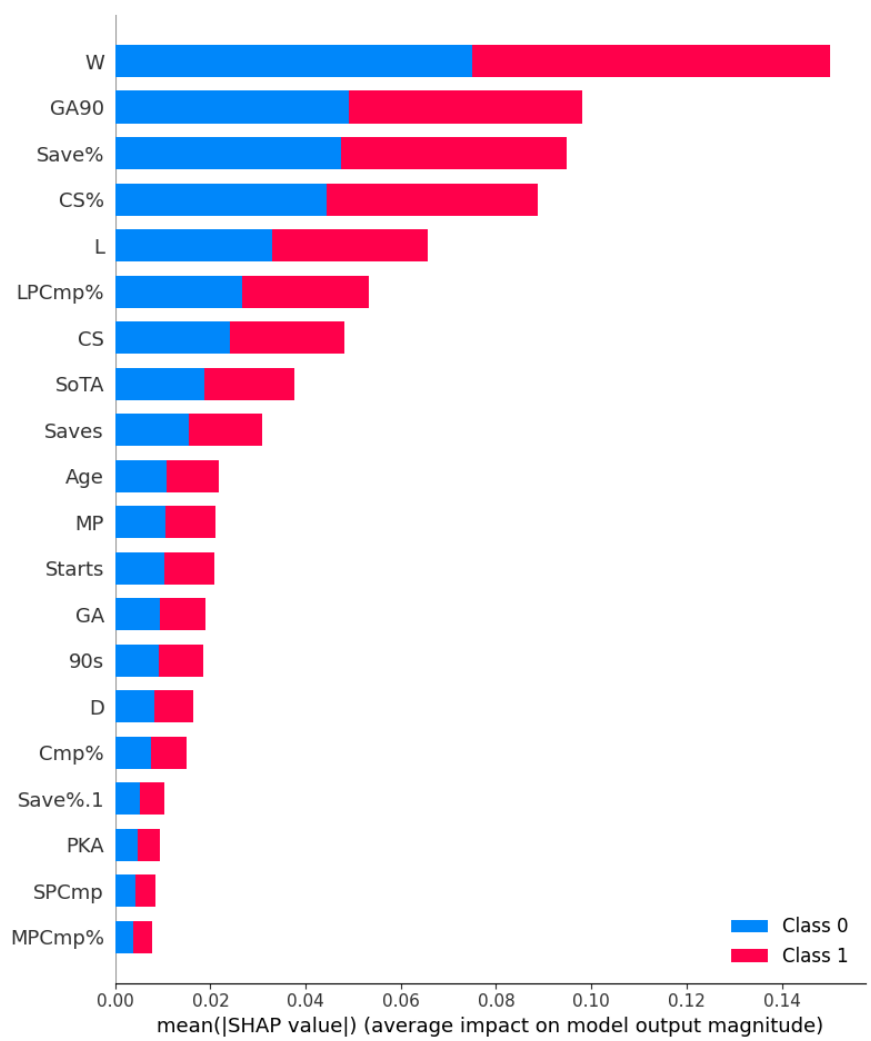
```
Model Accuracy        : 0.788135593220339
Model Precision       : 0.7575757575757576
Model Recall Score    : 0.5952380952380952
Model F1 Score        : 0.6666666666666667
```


ROC Curve
True Positive Rate
false Positive Rate
AUC=0.8718671679197995
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix
Actual
Predicted
68
8
17
25
0
1
10
20
30
40
50
60



Feature Importance
CS
Save%
CS%
GA90
W
0.00
0.02
0.04
0.06
0.08
0.10
0.12



W
GA90
Save%
CS%
L
LPCmp%
CS
SoTA
Saves
Age
MP
Starts
GA
90s
D
Cmp%
Save%.1
PKA
SPCmp
MPCmp%
Class 0
Class 1
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
mean(|SHAP value|) (average impact on model output magnitude)

7. Discussion

a. Traditional Goalkeepers

As seen from the data models, the traditional goalkeepers are very inconsistent when compared to their own data with only 5 years of difference and this proves that their playstyles are being forced to the more modern game where goalkeepers are demanded to be able to play with the ball at their feet way more than just being able to stop shots and save goals from happening.

There are also more proofs of change in regards to the playstyle of the traditional goalkeepers as there more attempted passes that has been made as the years progressed from 2018 towards 2023, signifying the fact that these types of goalkeepers are rushed by time and the era that they are playing in to be more adaptive towards them and not as they were meant to be. These changes are very visible especially in the data models of two selected goalkeepers that best represents the traditional style of goalkeeping, Keylor Navas and Jan Oblak. In the pie chart and the bar chart player comparison, it is visible that both players attempted more total passes going forward from the year of 2018.

Regardless of being put through time and are more demanded to be able to play the ball more with their feet, these types of goalkeepers are still very reliable when it comes to their actual main job, which is stop goals and saving the net from the shots that they are facing, as the data model has shown that they have a high number of saves that are made alongside the amount of shots that they have faced, meaning that they have a very high save rate.

b. Modern Goalkeepers

Unlike the data model of the traditional goalkeepers, the modern goalkeepers have a fairly consistent progress throughout the years. When compared to their own data from previous years, it can be visibly denoted that they are fairly similar to one another despite having quite a range of time in between each data.

The fact that the data models of the modern goalkeepers show little to no major changes is proof that the game of football is indeed revolving around having goalkeepers that are so called ‘modern’ or goalkeepers that can play with the ball as well as their outfield players, hence why there are no major changes in the data models unlike the traditional goalkeepers that are forced to adapt or evolve their play style into a more modern version of their original play style. This proof would then be further validated by a selected goalkeeper of the modern play style to represent the similarity in statistics over the following years, the selected goalkeeper is named Manuel Neuer. In the pie chart of the selected goalkeeper, it is very visible that there is almost no change in the percentage or rate of the passes that are made by the goalkeeper despite each data taken from two different timelines.

However, like the traditional goalkeepers, the modern goalkeepers have also shown great capabilities when it comes to stopping shots and keeping the goal safe. The data models for the saves and the save percentage are also consistent and proven to be absolute, meaning that the modern goalkeepers are why they are currently the standards of the game for they have the capabilities to not only keep the goal safe, but also help the outfield players by giving out key passes with their ball playing abilities.

c. Modern Vs Traditional

In the first two data models, we can see that the modern goalkeepers are better at passing and having the ball at their feet as the data models have shown that the modern goalkeepers have made more passes in the year of 2022 and they have also attempted more passes in the year of 2018, concluding the fact that the modern goalkeepers are indeed better with the ball at their feet.

However, the last two data models have also proven the fact that even though the traditional goalkeepers are worse with the ball on their feet, they are better when it comes to stopping shots and making saves. The last two data models have shown that the traditional goalkeepers have a higher percentage of saving shots in the year of 2018 compared to the modern goalkeepers and that the traditional goalkeepers have overall made more saves in the year or 2022.

Therefore, from the data models, we can say that the traditional goalkeepers are better at stopping shots and making saves, while the modern type of goalkeepers are better at passing and ball playing.

d. Prediction Model

The prediction model that was made using the Logistic Regression is a very successful model considering the fact that it has an AUC score of 0.9 with the worst score for the recall being 0.69 which can be rounded up to 0.7 which is still considered to be a good score nonetheless. All of this means that we can use the prediction model to predict whether a goalkeeper can win a trophy in the following seasons or not by the goalkeepers' performances.

The prediction model that is built with Random Forest Classifier can also be considered as a success as it has an AUC score of 0.87 with an accuracy score of 0.78 and the worst recall score being 0.59 which can be rounded up to 0.6. With the classifier, we have also found out about the feature importance and the data model has shown that some of the most important features that helped the machine in building the prediction model are the wins, the save percentages, the amount of goals that are allowed every 90 minutes, the clean sheet percentages, and the number of clean sheets themselves. However, the most impactful feature to the prediction model is the amount of wins in the season.

The confusion matrix that was built using the classifier has shown that the prediction model is better at predicting the seasons on which they did not win a trophy or an award. This means that the machine targets the seasons in which the goalkeepers did not win a single trophy and eliminates the trophyless seasons to predict the winning seasons.

8. Conclusion and Recommendation

a. Conclusion

In conclusion, traditional goalkeepers are better at saving and stopping shots, while modern goalkeepers are better at passing and playing with the ball at their feet. However, due to the effect of modern goalkeeping, traditional goalkeepers have been made to adapt and change their play style becoming similar to the modern goalkeepers in order to keep them relevant to the game. Regardless of the style of play that are used by the goalkeepers, both types have shown the capabilities to be able to win awards and trophies with their performances. Based on their performances, data have been gathered and collected to be then processed and turned into interfaces for comparison and prediction. The predictions are made using a trained machine by creating prediction models using regression and classifiers. The prediction model has been evaluated and it has shown that it can predict the outcome of winning trophies or not in a really good manner with a high and effective enough accuracy. The prediction model works by predicting the likelihood of not winning a trophy and then eliminating those likelihoods for an opposite outcome. The prediction model was able to find out the likelihood of not winning a trophy by calculating using the amount of wins, save ratios, clean sheets, goals conceded ratios, clean sheet ratios that have been accumulated by the goalkeepers respective to the season in which they play. Therefore, we now know the comparison between traditional and modern goalkeepers alongside having a predictive model that can help predict whether the goalkeepers can win a trophy or not based on their performances.

b. Recommendation

Recommendation for future works would include suggestions, such as making the prediction model to be able to predict the amount of trophies that can be won in a season and not just whether they can win a trophy at the end of the season or not. Another suggestion would be to use more types of regressors and classifiers that can be used as a base for the prediction model and do a comparison of each and find out which model is the best one out of all.

9. Link

a. **GitHub Repository (Containing Codes, CSVs, Goalkeeper Lists):**

   https://github.com/SAD-Nich/FundamentalDataScience/tree/0a2c90ad54e38b7ea3dd7a05db79d8c0aa72467d/Final%20Project

b. **Main Data Source (FBREF):**

   https://fbref.com/en/

Bibliography

[1] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, doi: https://doi.org/10.1109/mcse.2007.55.

[2] Pandas, “Python Data Analysis Library — pandas: Python Data Analysis Library,” *Pydata.org*, 2018. https://pandas.pydata.org/

[3] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[4] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning Word Vectors for Sentiment Analysis,” *ACLWeb*, Jun. 01, 2011. https://aclanthology.org/P11-1015/

[5] J. I. E. Hoffman, “Logistic Regression,” *Biostatistics for Medical and Biomedical Practitioners*, pp. 601–611, 2015, doi: https://doi.org/10.1016/b978-0-12-802387-7.00033-0.

[6] S. E. R, “Understand Random Forest Algorithms With Examples (Updated 2024),” *Analytics Vidhya*, Jun. 17, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=A.

Questions and Answers

1. Why aren't older goalkeepers who are actually 'traditional' used in the dataset?

   The data is not available for older goalkeepers that played in the 90s until the 2010s.

2. Why are there multiple goalkeepers used and not just a comparison between a representative of each style?

   The data will be too small if only one goalkeepers from each style are chosen.