

# **A REPORT**

**ON**

**REVOLUTIONIZING FOOTBALL METRICS: A COMPREHENSIVE STUDY IN  
UNVEILING PLAYER PERFORMANCE THROUGH EXPLAINABLE AI**

**BY**

**SRIVATHSAN V  
2021SC04865**



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE  
Pilani (Rajasthan) INDIA**

March, 2024

# **Revolutionizing Football Metrics: A Comprehensive Study in Unveiling Player Performance through Explainable AI**

DISSERTATION

Submitted in partial fulfillment of the requirements of the  
MTech Data Science and Engineering Degree programme

by

**Srivathsan V**  
**2021SC04865**

under the supervision of

**Balaji Thirumalai Vinjamur**  
**Delivery Manager**

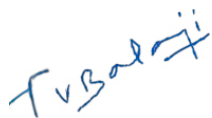
BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE  
Pilani (Rajasthan) INDIA

March, 2024

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**CERTIFICATE**

This is to certify that the Dissertation entitled “Revolutionizing Football Metrics: A Comprehensive Study in Unveiling Player Performance through Explainable AI” submitted by Mr./~~Ms.~~ Srivathsan.V IDNo. 2021SC04865 in partial fulfillment of the requirements of DSECLZG628T Dissertation, embodies the work done by him/her under my supervision.



Signature of the Supervisor

Place: Chennai

Date: 08-03-2024

Name: Balaji Thirumalai Vinjamur

Designation: Delivery Manager

# ACKNOWLEDGEMENT

Above all, I would like to thank my M.Tech. supervisor, Mr. Balaji Thirumalai Vinjamur, for his guidance and unflinching support during my studies. In addition, I would like to express my gratitude to Mr. Sankar Rajamani for serving as the course mentor and Mr. Karthik Rajkumar Kannan for his astute assistance with technical concerns; without them, finishing this dissertation would have required an enormous effort. I am grateful to all of the members of my Accenture Solutions Private Limited (India) team for their unwavering encouragement and assistance which was essential to my academic success. Finally, I want to express my gratitude to my parents for their unwavering emotional support throughout my academic career.

**SRIVATHSAN V**  
2021SC04865

# TABLE OF CONTENTS

<b>Acknowledgement</b>	<b>iii</b>
<b>Table of Contents</b>	<b>iv</b>
<b>Abstract</b>	<b>1</b>
<b>List of Tables</b>	<b>2</b>
<b>List of Figures</b>	<b>3</b>
<b>1 Introduction</b>	<b>4</b>
1.1 Dissertation Structure .....	5
<b>2 Literature Survey</b>	<b>7</b>
2.1 Estimating transfer fees of professional footballers using advanced performance metrics and machine learning.....	7
2.2 Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques .....	7
<b>3 Methodology</b>	<b>8</b>
3.1 Step 1: Data Collection.....	8
3.2 Step 2: Data Pre-processing .....	9
3.3 Step 3: Data Transformation .....	10
3.4 Step 4: Feature Selection.....	10
3.5 Step 5: Data Split .....	11
3.6 Step 6: Modelling.....	11
3.7 Step 7: Model Evaluation .....	11
3.8 Step 8: Model Anatomization .....	12
<b>4 Result</b>	<b>14</b>
4.1 Model Evaluation .....	14
4.1.1 Transfermarkt + SoFIFA.....	14
4.1.2 Transfermarkt + SoFIFA + FBRef + FPL.....	15
4.2 XAI Interpretation of the ML Model .....	15
4.2.1 Transfermarkt + SoFIFA.....	16
4.2.2 Transfermarkt + SoFIFA + FBRef + FPL.....	17
<b>5 Conclusion</b>	<b>20</b>
<b>6 Future Improvisations</b>	<b>21</b>
<b>7 Appendix A – Feature List</b>	<b>22</b>
7.1 Transfermarkt + SoFIFA .....	22
7.2 Transfermarkt + SoFIFA + FBRef + FPL .....	22

<b>8 Appendix B – Hyperparameter Values</b>	<b>23</b>
8.1 Transfermarkt + SoFIFA .....	23
8.2 Transfermarkt + SoFIFA + FBRef + FPL .....	24
<b>9 Appendix C – Abbreviations</b>	<b>25</b>
<b>10 References</b>	<b>26</b>
<b>11 Checklist</b>	<b>27</b>

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**  
**FIRST SEMESTER 2023-24**

**DSECLZG628T DISSERTATION**

**Dissertation Title** : Revolutionizing Football Metrics: A Comprehensive Study in Unveiling Player Performance through Explainable AI  
**Name of Supervisor** : Balaji Thirumalai Vinjamur  
**Name of Student** : Srivathsan V  
**ID No. of Student** : 2021SC04865

**Abstract**

The study establishes a well-designed Machine Learning models, in combination with Ensemble models and Explainable AI, using SHAP, for global and local interpretability to predict the market valuation of Football Players. Data of 15000+ players from Transfermarkt, SoFIFA, FBRef and Fantasy Premier League is being utilized. Post feature selection and model execution, it was understood that LightGBM excelled in prediction accuracy with RMSE 772,311.039 and  $R^2$  0.995038. Apart from the rating, attributes related to skill, mentality, movement added to the influence of the market valuation. The insights derived using the Explainable AI can then be used by the clubs, agents, and audience/supporters to make informed decisions about the player valuation. Nevertheless, the limitations of the study are not limited to dataset availability for fantasy games for other leagues, popularity report of the player, but also includes the financial statement of the buying and selling club, underestimation of highly ranked players, etc. The future work relies on the improvisation of data collection for additional players from different leagues and exploring the valuation prediction and the model's applicability.

**Keywords:** Sports Analytics, Explainable Artificial Intelligence (XAI), Player Performance Evaluation, Data-driven, Decision Making, Machine Learning Algorithms

# LIST OF TABLES

4.1.1	Different model executions with respective R2 and RMSE values for Transfermarkt + SoFIFA	14
4.1.2	Different model executions with respective R2 and RMSE values for Transfermarkt + SoFIFA + FBRef + FPL	15
5	Final list of Features post Feature Selection	20
7.1	Different features used during the model prediction – Transfermarkt + SoFIFA	22
7.2	Different features used during the model prediction – Transfermarkt + SoFIFA + FBRef + FPL	22
8.1	Hyperparameter values – Transfermarkt + SoFIFA	23
8.2	Hyperparameter values – Transfermarkt + SoFIFA + FBRef + FPL	24



# LIST OF FIGURES

3.8	Flowchart of the implementation of the overall design, feature reduction, model training, evaluation, XAI interpretation	13
4.2.1	Bee swarm plot of SHAP Values. Higher the SHAP value, higher the market value – Transfermarkt & SoFIFA	16
4.2.1	Market value prediction of ‘Falaye Sacko’ depicted by SHAP force plot	17
4.2.1	Market value prediction of ‘Raul Garcia’ depicted by SHAP force plot	17
4.2.2	Market value prediction of ‘Philippe Coutinho’ depicted by SHAP force plot	17
4.2.2	Market value prediction of ‘Caio Henrique Oliveira Silva’ depicted by SHAP force plot	18
4.2.2	Market value prediction of ‘Adrien Silva’ depicted by SHAP force plot	18
4.2.2	Bee swarm plot of SHAP Values. Higher the SHAP value, higher the market value along with FBRef & FPL	19

# CHAPTER 1

## INTRODUCTION

Football (aka Soccer), referred to as ‘the beautiful game’, is the most popular game in the world that encompasses around 3.5Bn fans watching the game from extremely diverse cultural, topographical, and financial backgrounds starting from the East to the West. Any sport, given the popularity, have variety of revenue generating streams like merchandise sales (majorly shirt sales), stadium tickets, broadcasting rights, investors/sponsors, the transfer market, etc.

Of all the revenue generating streams, the one under heavy perusal is the transfer market which puts a lot of restriction on the expenditure by a team. This is done in order to add a constraint to the teams to avoid them from spending more than they earn. FFP – Financial Fair Play regulations – was brought into place for the same reason and to put a cap on the expenditure so that all the teams get an equal treatment.

The transfer market goes both ways, wherein the team(s) can both sell and buy player(s). Here the selling of a player adds to the earnings and purchase amounts to the expenditure of the club. It is for this reason that estimating a player’s financial worth/valuation is of utmost importance to help in facilitating a smooth flow of financial operations within the clubs and the football industry, both at local and global scale. Additionally, the market value of a player is also directly proportional towards the wage of the player, meaning highly-rated and high-profile players demand more wage compared to the others. The same section of finance also influences the overall team valuation which will be helpful in financial negotiations or ownership transitions and helps extensively in the financial planning of the clubs. Notably, the player’s valuation is also influenced by the number of clubs interested in the player and the financial status of the interested clubs, which in turn increases the valuation of the player. With the ever-growing factors influencing a player’s market valuation and the increase in number of talented players, and the competition from other clubs, there is a need for technology to be put into play to make well-informed decisions at a faster pace.

In the era dominated by big data and analytics, accurate analytical methodologies are becoming an inevitable part of a club, acting like an additional club staff who can help analyzing multitudes of data in a jiffy. As the world is moving towards the digital landscape, so is the sports industry and its affiliates. There are multiple platforms that are sophisticated enough to provide the details about a player, the whole team and club, manager, every game, etc. Few such platforms that provide a plethora of panoramic player data are Transfermarkt, SoFIFA, FotMob, StatsBomb, WhoScored, etc. The mentioned data sources have seen multiple machine learning models being worked on them to improve the accuracy of the player market valuation. The notable one being McHale and Holmes[1], who worked on combining the data from Transfermarkt, instats, sofifa, etc. to come up with models like Plus-Minus rating and other regression models which showed that XGBoost was better at making predictions.

There have been multiple other researches focused on improving the machine learning model accuracy and valuating the players as better as possible, for example a paper by Al-Asadi & Sakir [2] worked on the video game-based data to improve the accuracy of the football player's value. But the lack of interpretability of the data and prediction has posed a need to develop a transparent and an interpretable model to identify and provide valuable insights of to understand the factors helping in valuing a player.

This study incorporates ensemble ML models, like XGBoost, CatBoost, LGBM, etc., and using the SHAP (SHapley Additive exPlanations) method to interpret the predictions of the best performing model from both the global and individual (local) perspective.

## **1.1. Dissertation Structure**

This section describes the upcoming sections and the information provided in each section.

### **Chapter 2: Methodology**

Describes all the actions that were taken, step-by-step.

### **Chapter 3: Result**

Provides the results from the model training and prediction with the  $R^2$  and RMSE scores.

#### Chapter 4: Conclusion

Concludes with the findings and pointing towards the direction of future scope.

#### Chapter 5: Future Scope

Shares information about the additional linear and non-linear features that can be considered in making better predictions.

#### Chapter 6: Appendix A – Feature list

List of all the features that were included in the final dataset.

#### Chapter 7: Appendix B – Hyperparameter values

List of hyperparameters used along with the respective values.

#### Chapter 8: Appendix C – Abbreviations

List of abbreviations used in this study.

#### Chapter 9: References

List of research papers referenced during literature survey

# CHAPTER 2

## LITERATURE SURVEY

### **2.1. Estimating transfer fees of professional footballers using advanced performance metrics and machine learning**

The paper by Ian G. McHale and Benjamin Holmes worked towards estimating the transfer fees of football players and presented a model for the same. Their work involved two different steps. Firstly, they captured the players' ability to rate them and, secondly, they worked towards improving the accuracy of the predictions by implementing machine learning models. Advanced performance metrics along with linear regression improved the predictions. Data from multiple sources like Transfermarkt for transfer details and the transfer history, Sofifa for overall and potential ratings, match event data for 36 leagues from Instat were combined to use in the models and to eventually provide the ultimate result of improved predictions.

### **2.2. Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques**

The paper serves as a baseline model to help during transfer negotiations by simplifying the process of transfer value estimation. The authors completely relied on the data available from the Sofifa, which holds the data of the football players from the game's perspective. Models like linear regression, decision trees, random forests and multiple linear regression were implemented wherein random forests showed greater potential. Finally, an RMSE value of 1,649,921 and R<sup>2</sup> value of 0.95 was reported proving to be better than the other existing models, at that time.

# CHAPTER 3

## METHODOLOGY

### 3.1. Step 1: Data Collection

The study employs the usage of datasets from Transfermarkt, SoFIFA, FBRef, Fantasy Premier League (FPL) websites by combining the datasets based on the last name and date of birth of the player between 2015 and 2023 (based on the data availability for each player).

1. **Transfermarkt** offers information about each player over the years and the games they have appeared with the number of goals, assists, yellow and red cards, etc. along with the club related stats.
2. **SoFIFA** provides detailed information about the meticulous yet critical traits of a player by rating them between 0 and 100 for each attribute. It also provides the current overall rating of the player and the potential of the player, along with other attributes like shot power, long and short passing, dribbling, pace, etc. Apart from the player stats, individual team stats are also made available, but our focus would be on the player level analysis.
3. **FBRef** shares with us very intricate and statistically calculated information like xG (expected goals), xA (expected assists), npxG (non-penalty goals), etc. The data can be segregated into multiple facets like ‘possession’, ‘shooting’, ‘passing’, ‘goal and shot creation’, ‘standard’, ‘pass types’, ‘defensive actions’, which serve to be very important information. This information is provided to FBRef by Opta Sports, popularly known as Opta (a sports analytics platform).
4. **Fantasy Premier League**, most popularly called as **FPL**, is a platform that allows its users to act as managers and select their own teams by buying players and each user is given \$100M, which is a fake money (like monopoly). Points are assigned to the players based on multiple characteristics like clean sheet maintenance, number of minutes player, assists, goals additionally, bonuses are also provided to top 3 players for that week. This is NOT a betting

platform, but a ‘fantasy’ world. This platform provides details like ‘influence’, ‘threat’, ‘selected by’ (a metric that shows how many managers have chosen that particular player in their team), ICT index, etc. But FPL data is available only for English Premier League.

### **3.2. Step 2: Data Pre-processing**

The data extracted do not always have the complete detail for all the records and they either need to be dropped or imputed. This step aims at finding such records and based on the importance of the player, by analyzing the various attributes, the records are either dropped or imputed.

Transfermarkt provides game wise, player wise, club wise, etc. details. Game wise analysis is made for each player and the club the player has played in to collate the details of the goals scored, assists provided, yellow and red cards received, minutes played. This data is then aggregated year wise to get the stats of the player for the whole year. The initial data is made available for 43 competitions.

Redundant columns like name, age, player id, club id, etc. are removed and the numerical columns are converted to either integer or float. The missing values are either removed or values are calculated based on other attributes. Footedness is not taken into consideration, unlike the way Bryson, Frick, and Simmons [3] considered, and is part of the future scope of this study.

Similarly, SoFIFA too contains details about the player’s attributes like ability of short passing, long passing, shot power, dribbling, mentality, interception, contract expiration date, release clause, etc.

With FBRef, the details of the Big 5 European Leagues (English Premier League, LaLiga, Bundesliga, Ligue 1, Serie A) alone are extracted and analyzed. The number of competitions is reduced to 5, due to lack of Opta data for other leagues/competitions. A few players do not have the data for certain important fields and they are considered as 0, to aid the analysis.

Fantasy Premier League focusses only on English Premier League, the league that is said to be the most famous in the world with 20 teams competing against each other to clench the title. The missing data is imputed or deleted, based on the importance of the player and the type of data.

### **3.3. Step 3: Data Transformation**

The data in the transfermarkt dataset was used in transforming the data by deriving calculated columns like goals conceded by the team, goals conceded by the player, goals scored by the team, goals scored by the player, number of matches played by the team, number of matches played by the player, goal contribution percentage, minutes played percentage, overall game minutes(restricted to 90 and not considering extra time), match result for both home and away teams, etc. Categorical columns like player position, match result, etc. were split into different columns. After combining SoFIFA, FBRef, and FPL data, the final dataset, before feature selection, consisted of 700+ number of players with 250+ number of features.

The Data transformation underwent a multiple-iteration setup in order to arrive at the best combination of the datasets. The initial setup began with just using the Transfermarkt data and then SoFIFA was combined with the Transfermarkt which had around 150+ features. The dataset combination showed great improvement in the RMSE along with the R2 values.

To further the reduction of the RMSE, FBRef was brought into the field and the dataset contributed additional 120+ features.

Though FBRef dataset proved the point that additional factors about a player could be of great help, it can also be seen that there was a necessity to reduce the RMSE value further and hence the Fantasy Premier League data was introduced. The Fantasy Premier League data provides the popularity of the player among the fantasy managers and deducing the importance of the player by calculating the number of fantasy managers who have selected the given player. The FPL data provided 20+ features.

### **3.4. Step 4: Feature Selection**

There was a total of 250+ features after merging the SoFIFA, Transfermarkt, FBRef, FPL datasets based on the last name and dob of the player. More than 200+ features were manually removed to avoid multi-collinearity, as they were used in deriving a calculated column to find the overall impact of the player. Also, taking into consideration the computational complexities, the features were meticulously valued and rejected.

The final dataset, post feature selection, consisted of 55 columns where 54 columns were features



and the 1 column was the label – valuation of the player.

Transfermarkt and SoFIFA contributed to 22 features and 1 label, FBRef contributed to 25 features and FPL data contributed to 7 features of the overall list.

### **3.5. Step 5: Data Split**

The number of players from Transfermarkt and SoFIFA started off with 15000+ players and gradually reduced to 1000+ players, due to the narrow downing of the number of competitions from 40 to 1 considering the quality of the dataset.

The clean dataset was then randomly split into 80% for training and 20% for testing using the ‘train\_test\_split’ function from the ‘sklearn.model\_selection’ library.

### **3.6. Step 6: Modelling**

The modelling part involved multiple regression models coupled with ensemble learning algorithms and these models were chosen based on the potential efficacy as analyzed from the previous studies of McHale and Holmes [1]. The algorithms used were AdaBoost, RandomForests, LightGBM, GBDT, CatBoost, XGBoost.

The split dataset was used to train the models and test the same using the test dataset. The hyperparameters were tuned manually to find the best combination of the parameters that can provide the best possible results.

### **3.7. Step 7: Model Evaluation**

The trained models were related to regression and hence  $R^2$  and RMSE metrics were calculated for the test dataset, to understand the model’s prediction accuracy.

### 3.8. Step 8: Model Anatomization

Machine learning models have a special name, black box, due to their nature of lack of interpreting the underlying contraptions behind the model's prediction accuracy. Lundberg and Lee came up with the SHapley Additive exPlanations (SHAP) [<https://arxiv.org/abs/1705.07874>] approach to overcome the black box nature of the models. SHAP provides with the ability to interpret the predictions at both global and local levels that helps in analyzing the individual feature's contribution towards the prediction.

The local interpretability involved the implementation of the force plot to explicate the predicted market values of the player(s). Due to high computational time requirement, only 0.1% of the test dataset is used for the analysis for the first scenario, whereas the second scenario analyzes the data on the whole dataset. The graphical representation has a horizontal scale with the base value and how the prediction varies. The scale represented the features that influenced the value both positively and negatively, where the positive influence is represented in red and the negative influence is represented in blue. The granular level analysis of the features affecting a player's valuation is the most transparent and interpretable display.

The global interpretability involved the implementation of the bee swarm plots which showed the detail of the most effective metrics/features in the whole dataset. The x-axis represents the SHAP values while the y-axis represents the different features with each feature being represented by colored dots where the red color represents the high values and the blue color represents the low values thus providing the importance of the feature on the output of the model.

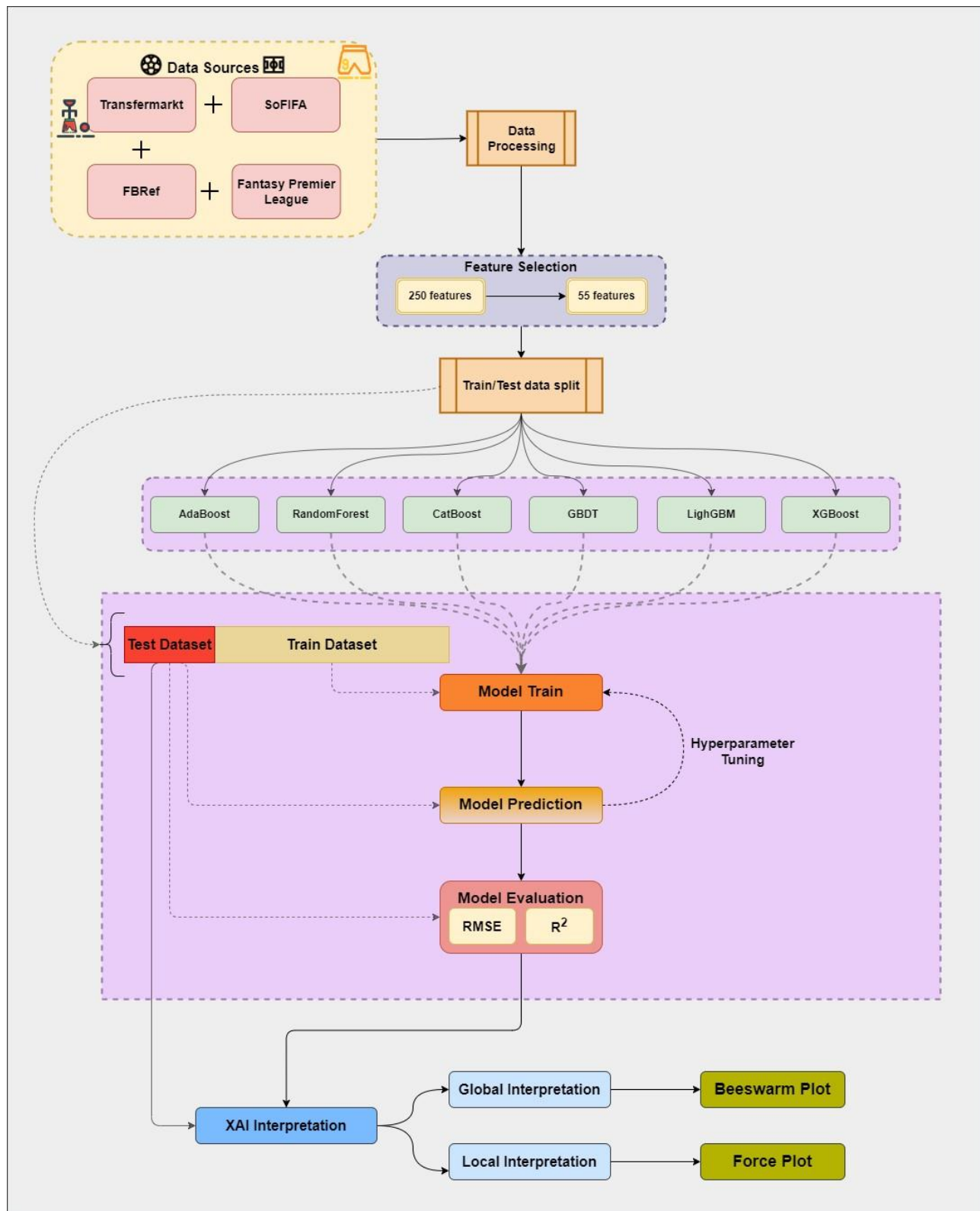


Figure 1: Flowchart of the implementation of the overall design, feature reduction, model training, evaluation, XAI interpretation

# CHAPTER 4

## RESULT

### 4.1. Model Evaluation

The models were evaluated under 2 different scenarios: one with Transfermarkt and SoFIFA data and the other with the addition of FBRef and FPL data.

#### 4.1.1. Transfermarkt + SoFIFA

The details about the  $R^2$  and RMSE values of different models are shared in the below table:

ModelName	R2(accuracy)	R2 Model Rankning	RMSE	RMSE Model Rankning	MSE
AdaBoostRegressor	0.886142	6	2695017.783	6	7263120848920.680
RandomForestRegressor	0.972193	5	1331850.442	5	1773825599389.000
CatBoostRegressor	0.981542	1	1085092.907	1	1177426616555.627
GradientBoostingRegressor	0.977999	4	1184669.337	4	1403441437269.794
LightGBM Regressor	0.978165	3	1180203.541	3	1392880399311.619
XGBoost Regressor	0.978189	2	1179555.466	2	1391351097569.409

Table 1: Different model executions with respective  $R^2$  and RMSE values for Transfermarkt + SoFIFA

As it can be seen, amongst the trained ensemble models the CatBoost model show exemplary results with an  $R^2$  value of 0.9815 and RMSE value of 1085092.907 compared to the other models, and the AdaBoost regressor model portrayed a bad R2 value of 0.886 and RMSE value of 2693776, making it the least preferred model. The GBDT model stood second with an  $R^2$  value of 0.9786 followed by XGBoost with an  $R^2$  value of 0.97818. LightGBM stood behind XGBoost with  $R^2$  value of 0.97816 by a very small margin of 0.00002 and finally RandomForest stands as last but one with an  $R^2$  value of 0.9722.

The RMSE values are also arranged in the same order of  $R^2$  values where GBDT posed an RMSE value of 1167908.034 followed by XGBoost with a value of 1179555.466, LightGBM with

1180203.5414, and RandomForest with 1331024.3441.

#### 4.1.2. Transfermarkt + SoFIFA + FBRef + FPL

The introduction of FBRef and Fantasy Premier League datasets proved to have improved the prediction accuracy and it can be seen from the below table:

	ModelName	R2(accuracy)	R2 Model Ranking	RMSE	RMSE Model Ranking	MSE
0	AdaBoostRegressor	0.925913	6	2984182.959	6	8905347934384.926
1	RandomForestRegressor	0.990172	5	1086892.265	5	1181334796056.695
2	CatBoostRegressor	0.992036	4	978435.710	4	957336438432.698
3	GradientBoostingRegressor	0.994810	3	789876.391	3	623904712511.274
4	LightGBM Regressor	0.995038	1	772311.039	1	596464340952.536
5	XGBoost Regressor	0.995017	2	773949.890	2	598998432721.127

Table 2: Different model executions with respective  $R^2$  and RMSE values for Transfermarkt + SoFIFA + FBRef + FPL

As it can be seen, amongst the trained ensemble models the LightGBM Regressor model showed exemplary results with an  $R^2$  value of 0.9950 and RMSE value of 772311.039 compared to the other models, and the AdaBoost regressor model portrayed a bad  $R^2$  value of 0.9259 and RMSE value of 2984182.96, making it the least preferred model. The XGBoost model stood second with an  $R^2$  value of 0.995017 followed by GBDT with an  $R^2$  value of 0.994810. CatBoost stood behind GBDT with  $R^2$  value of 0.992036 and finally RandomForest stands as last but one with an  $R^2$  value of 0.990172.

The RMSE values are also arranged in the same order of  $R^2$  values where XGBoost posed an RMSE value of 773949.890 followed by GBDT with 789876.391 as the RMSE which is further succeeded by CatBoost and RandomForest with RMSE values of 978435.710 and 1086892.265 respectively.

On the whole, the addition of FBRef and FPL dataset have proven to improve the predicting accuracy of the ensemble models.

## 4.2. XAI Interpretation of the ML model

Similar to the model evaluation, the XAI model implementations was also implemented for the 2 different scenarios.

### 4.2.1. Transfermarkt + SoFIFA

The SHAP bee swarm plot denotes the importance of the features of the best performing model, the CatBoostRegressor, as illustrated in the following figure:

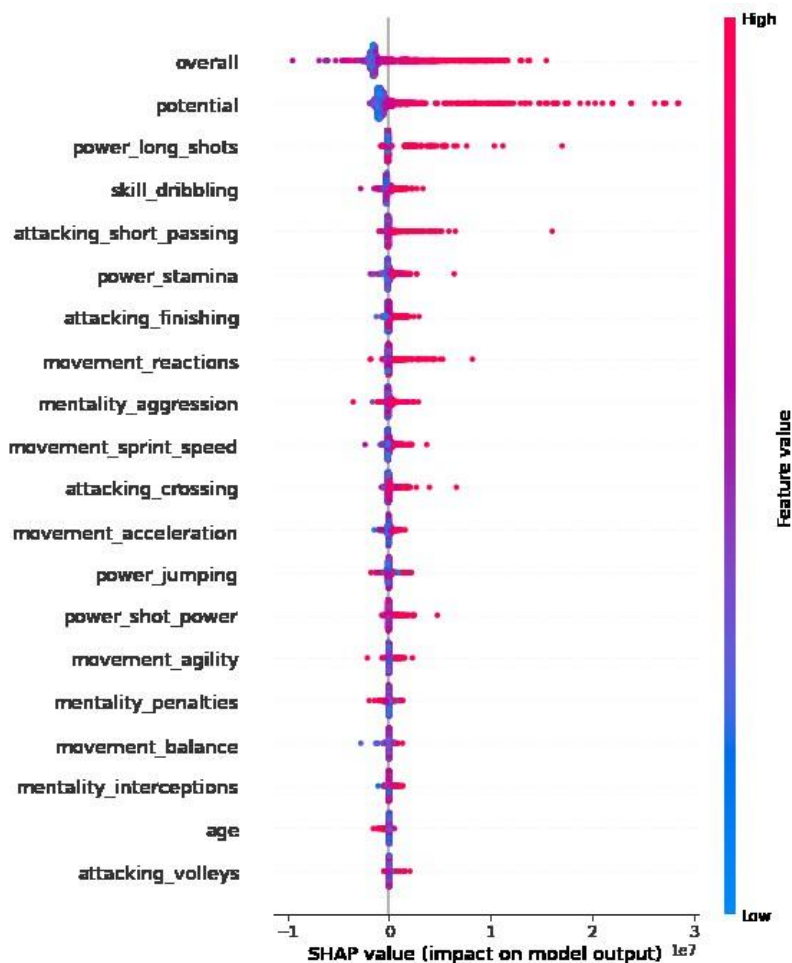


Figure 2: Beeswarm plot of SHAP Values. Higher the SHAP value, higher the market value – Transfermarkt & SoFIFA  
Note: The most valued features are from the fraction of data and not the whole data.

The Force Plot enforces that the predictions seem to be closer to the actual valuations indicating that the model has performed better in the provided circumstances:

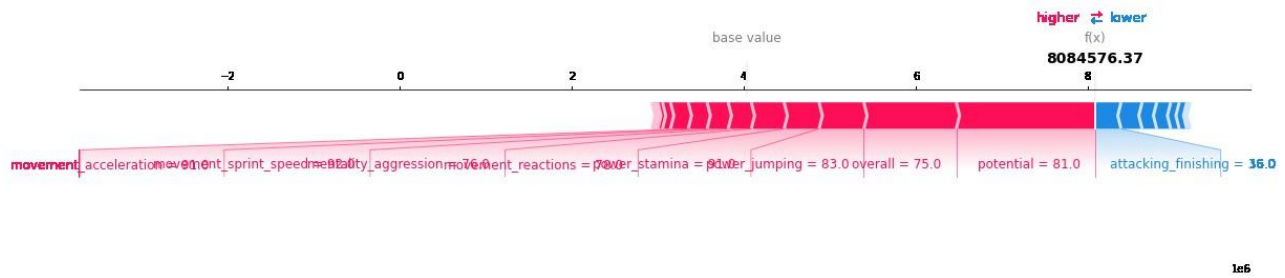


Figure 3: Market value prediction of 'Falaye Sacko' depicted by SHAP force plot.



Figure 4: Market value prediction of 'Raul Garcia' depicted by SHAP force plot.

As it can be seen from the figure 3, the actual valuation of 'Falaye Sacko' was 8M Euros. The predictions by CatBoost model show the market value as 8.08M Euros which is close enough to the actual valuation. Figure 4 shows that 'Raul Garcia' was estimated to be 17M Euros worth and the prediction shows the market value as 16.7M Euros which is close enough to original valuation. The red values show the features and the percentage of contributions towards pushing the market value higher while the counterpart blue values show the features that reduce the market valuation of the player.

#### 4.2.2. Transfermarkt + SoFIFA + FBRef + FPL

The Force Plot enforces that the predictions seem to be closer to the actual valuations indicating that the model has performed better in the provided circumstances:

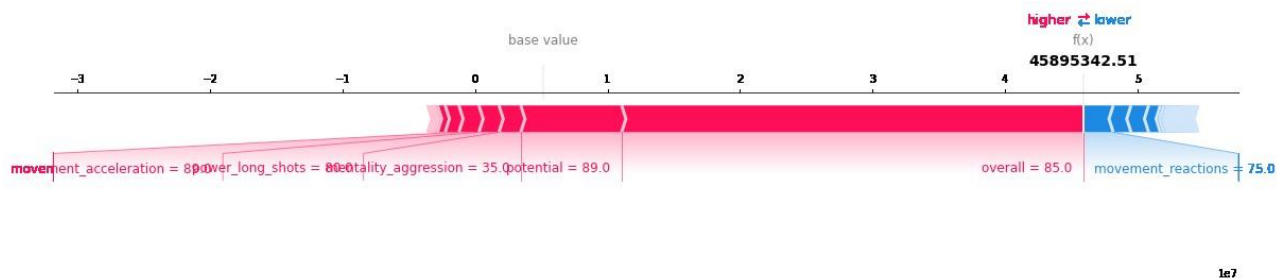


Figure 5: Market value prediction of ‘Philippe Coutinho’ depicted by SHAP force plot.



Figure 6: Market value prediction of ‘Caio Henrique Oliveira Silva’ depicted by SHAP force plot.

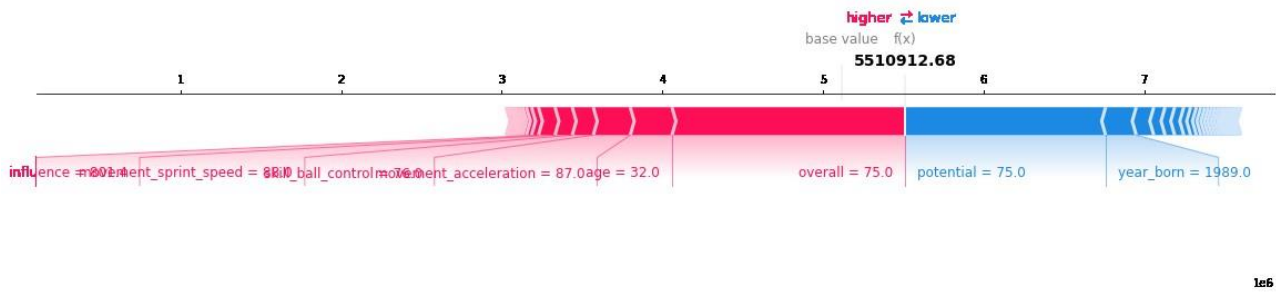


Figure 7: Market value prediction of ‘Adrien Silva’ depicted by SHAP force plot.

The force plots conclude that the Overall and Potential ratings still seem to have the upper hand in deciding the market valuation of the player at any give time. Also, Figure 5 shows the details of the player Philippe Coutinho who has a valuation of 44M Euros and the predication came close with 45.9M Euros. Figure 6 and 7 depict the predicated valuations of Oliveira Silva and Adrien Silva as 18.4M Euros and 5.51M Euros, while their actual valuations were 18.5M Euros and 5.5M Euros respectively.

The SHAP bee swarm plot including the FBRef and FPL denotes the importance of the features of the best performing model, the LightGBM Regressor, as illustrated in the following figure:



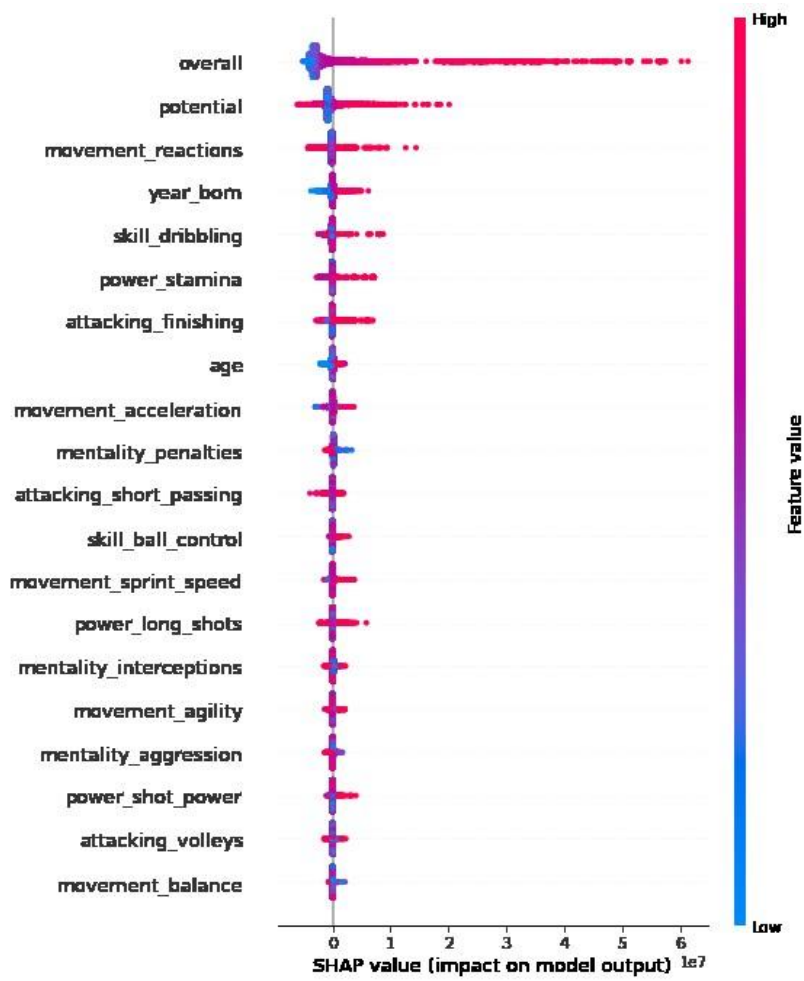


Figure 8: Bee swarm plot of SHAP Values. Higher the SHAP value, higher the market value along with FBRef & FPL

# CHAPTER 5

## CONCLUSION

As part of the conclusion, the study aimed at integrating well advanced machine learning techniques and analysis of important features to provide better, transparent, and interpretable understanding of the player's market valuation in the football empire. The study has fulfilled what it aimed and it is seen from the results of the different models and their accuracies where CatBoost outperformed the other models to provide the details about the highly influential features that helped in the accurate prediction in combination with the SHAP model.

The key features that influenced the market valuation of the player(s) are as distributed over various aspects and they are grouped as follows:

Rating	Attacking	Skill	Movement	Power	Calculated
overall	crossing	dribbling	acceleration	shot_power	SoT%
potential	finishing	ball_control	sprint_speed	jumping	Take-Ons Succ%
	short_passing		agility	stamina	Take-Ons Tkl%
	volleys		reactions	long_shots	Challenges Tkl%
			balance		Total Passes Cmp%
					PK%
<b>Mentality</b>	<b>Performance</b>	<b>Expected</b>	<b>Others</b>	<b>Fantasy</b>	xG%
aggression	SCA	npG+xAG	age	total_points	G/Sh%
interceptions	GCA	npG/Sh	overall_impact	creativity	G/SoT%
penalties	Touches	G-xG		influence	Pass %
	Tkl	npG-xG		bonus	Tackles%
	ImpactPos	A-xAG		bps	
	Foul Involvement			ict_index	
				selected by percent	

Table 3: Final list of Features post Feature Selection

With details about the player, the model could work in making better predictions with much lesser RMSE and higher  $R^2$  values. More details about the future scope are discussed in Chapter 5.

# CHAPTER 6

## FUTURE IMPROVISATIONS

As seen from the above predictions, it is ultimately evident that there is good amount of scope to add more features. Linearly dependent factors can take into consideration the player rating in each match, number of awards won, actual transfer fee, release clause value, years remaining in contract, etc., while non-linear factors may include the selling club's worth, interested club's worth and financial stability, Football Fantasy Premier League stats for other leagues, Legal Betting, player popularity, shirt sales, footedness etc. can be considered.

Considering these features during training may help in providing the expected results by pushing to contribute to the ever-growing advancements of ML applications in the sport sector.

# CHAPTER 7

## APPENDIX A – Feature List

The below are the list of final features post feature engineering that were used in the model prediction during both scenarios:

### 7.1. Transfermarkt + SoFIFA

Rating	Attacking	Skill	Movement	Power	Mentality
overall	crossing	dribbling	acceleration	shot_power	aggression
potential	finishing	ball_control	sprint_speed	jumping	interceptions
	short_passing		agility	stamina	penalties
	volleys		reactions	long_shots	
			balance		

Table 4: Different features used during the model prediction – Transfermarkt + SoFIFA

### 7.2. Transfermarkt + SoFIFA + FBRef + FPL

Rating	Attacking	Skill	Movement	Power	Calculated
overall	crossing	dribbling	acceleration	shot_power	SoT%
potential	finishing	ball_control	sprint_speed	jumping	Take-Ons Succ%
	short_passing		agility	stamina	Take-Ons Tkld%
	volleys		reactions	long_shots	Challenges Tkld%
			balance		Total Passes Cmp%
					PK%
Mentality	Performance	Expected	Others	Fantasy	xG%
aggression	SCA	npG+xAG	age	total_points	G/Sh%
interceptions	GCA	npG/Sh	overall_impact	creativity	G/SoT%
penalties	Touche	G-xG		influence	Pass %
	Tkl	npG-xG		bonus	Tackles%
	ImpactPos	A-xAG		bps	
	Foul Involvement			ict_index	
				selected by percent	

Table 5: Different features used during the model prediction – Transfermarkt + SoFIFA + FBRef + FPL

# CHAPTER 8

## APPENDIX B – Hyperparameter Values

### 8.1. Transfermarkt + SoFIFA

Model Name	Hyperparameter	Value
AdaBoost	n_estimators	50
	learning_rate	0.1
RandomForest	n_estimators	1200
	max_depth	25
	min_samples_leaf	5
	min_samples_split	5
CatBoost	iterations	300
	learning_rate	0.3
	depth	16
	subsample	0.8
GBDT	n_estimators	900
	learning_rate	0.2
	max_depth	5
LightGBM	n_estimators	1500
	learning_rate	0.3
	max_depth	6
	num_leaves	64
XGBoost	n_estimators	1500
	learning_rate	0.2
	max_depth	5

Table 6: Hyperparameter values – Transfermarkt + SoFIFA

## 8.2. Transfermarkt + SoFIFA + FBRef + FPL

Model Name	Hyperparameter	Value
AdaBoost	n_estimators	50
	learning_rate	0.1
RandomForest	n_estimators	700
	max_depth	25
	min_samples_leaf	5
	min_samples_split	5
CatBoost	iterations	500
	learning_rate	0.2
	depth	16
	subsample	0.8
GBDT	n_estimators	900
	learning_rate	0.2
	max_depth	5
LightGBM	n_estimators	1500
	learning_rate	0.1
	max_depth	6
	num_leaves	64
XGBoost	n_estimators	1500
	learning_rate	0.09
	max_depth	5

Table 7: Hyperparameter values – Transfermarkt + SoFIFA + FBRef + FPL

# CHAPTER 9

## APPENDIX C - Abbreviations

**R<sup>2</sup>** – R-Squared

**RMSE** – Root Mean Square Error

**ML** – Machine Learning

**XAI** – Explainable AI

**AI** – Artificial Intelligence

**LGBM** – Light Gradient Boosting Machine

**XGBoost** – Extreme Gradient Boosting

**CatBoost** – Category (Categorical) Boosting

**GBDT** – Gradient Boosting Decision Trees

**AdaBoost** – Adaptive Boosting

**FPL** – Fantasy Premier League

**M** – Million

**SHAP** – Shapley Additive Explanations

# CHAPTER 10

## REFERENCES

**[1] A Journal Paper:**

Ian G. McHale and Benjamin Holmes. Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. *European Journal of Operational Research*, 306(1):389–399, April 2023. ISSN 03772217. doi:10.1016/j.ejor.2022.06.033. URL <https://www.sciencedirect.com/science/article/pii/S0377221722005082>.

**[2] A Journal Paper:**

Mustafa A. AL-ASADI and Sakir Tasdemir. Predict the Value of Football Players Using FIFA video game data and Machine Learning Techniques. *IEEE Access*, pages 1–1, January 2022. doi:10.1109/access.2022.3154767. MAG ID: 4214589397 S2ID: b85f0efabc3fdce4bd997449f22eefe56e50b319

**[3] A Journal Paper:**

Bryson, A., Frick, B., & Simmons, R. (2013). The Returns to Scarce Talent: Footedness and Player Remuneration in European Soccer. *Journal of Sports Economics*, 14(6), 606-628. <https://doi.org/10.1177/1527002511435118>



# CHECKLIST

- |   |   |
|---|---|
| a) Is the Cover page in proper format?  | Y |
| b) Is the Title page in proper format?  | Y |
| c) Is the Certificate from the Supervisor in proper format? Has it been signed? | Y |
| d) Is Abstract included in the Report? Is it properly written?                  | Y |
| e) Does the Table of Contents page include chapter page numbers?                | Y |
| f) Does the Report contain a summary of the literature survey?                  | Y |
| i. Are the Pages numbered properly?   | Y |
| ii. Are the Figures numbered properly?  | Y |
| iii. Are the Tables numbered properly?  | Y |
| iv. Are the Captions for the Figures and Tables proper?                         | Y |
| v. Are the Appendices numbered?   | Y |
| g) Does the Report have Conclusion / Recommendations of the work?               | Y |
| h) Are References/Bibliography given in the Report?                             | Y |
| i) Have the References been cited in the Report?                                | Y |
| j) Is the citation of References / Bibliography in proper format?               | Y |



(Signature of Student)

Date:08/03/2024



(Signature of Supervisor)

Date: 08/03/2024

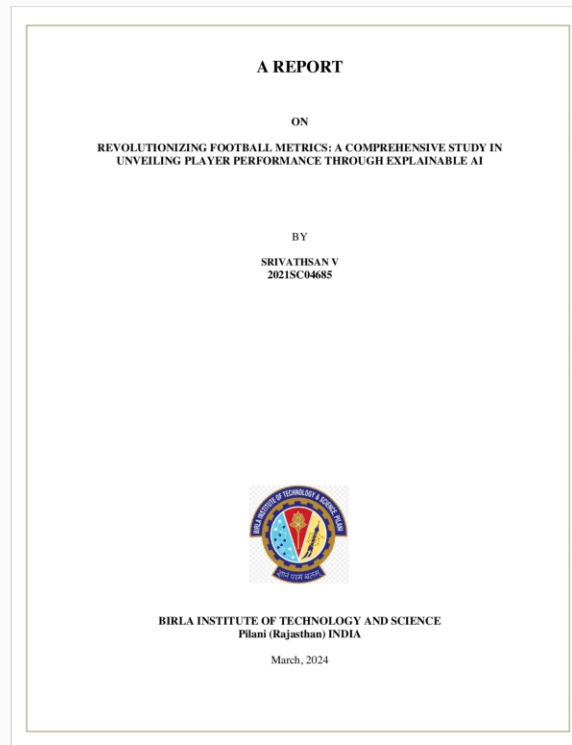


## Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author: 2021SC04865-SRIVATHSAN V .  
Assignment title: Dissertation(S1-23\_DSECLZG628T)-Final - 3  
Submission title: 2021SC04865.pdf  
File name: 2021SC04865.pdf  
File size: 1.01M  
Page count: 33  
Word count: 4,900  
Character count: 25,866  
Submission date: 13-Mar-2024 10:30AM (UTC+0530)  
Submission ID: 2316683710



## ORIGINALITY REPORT

8%

SIMILARITY INDEX

5%

INTERNET SOURCES

2%

PUBLICATIONS

4%

STUDENT PAPERS

## PRIMARY SOURCES

1	Submitted to Birla Institute of Technology and Science Pilani Student Paper	3%
2	Hai Tao, Omer A. Alawi, Raad Z. Homod, Mustafa KA. Mohammed et al. "Data driven insights for parabolic trough solar collectors: Artificial intelligence-based energy and exergy performance analysis", Journal of Cleaner Production, 2024 Publication	<1%
3	www.mdpi.com Internet Source	<1%
4	github.com Internet Source	<1%
5	dspace.cuni.cz Internet Source	<1%
6	www.coursehero.com Internet Source	<1%
7	ecommons.usask.ca Internet Source	<1%