Analysis of IPL Auction Dataset Using Explainable Machine Learning with Lime and H2O AutoML

Aradya Garg

Amity Institute of Information Technology

Amity University Noida

Uttar Pradesh, India

gargaradya@gmail.com

Alka Chaudhary

Amity Institute of Information Technology

Amity University Noida

Uttar Pradesh, India

achaudhary4@amity.edu

Abstract— The global sports market, one of the biggest markets in the world, grew from \$354.96 billion in 2021 to \$496.52 billion in 2022, according to research from a business research organization. Sports teams are becoming increasingly devoted to investing in sports data analytics to gain a competitive edge as spending on the global sports market rises; as a result, it is predicted that the sports analytics industry will exceed \$4.5 billion by 2025. It is the study of athletic performance and business health to maximize a sports organization's processes and results. Explainable machine learning (XML) is a key part of machine learning and AI because it explains how machine learning models create predictions. To establish trust and ensure accountability in AI systems, it is crucial to be able to comprehend and evaluate a model's predictions.

Keywords— Explainable Machine Learning, H2O, LIME, R, Sports Analytics, Cricket, IPL, Stacked Ensemble.

I. INTRODUCTION

XML allows you to understand how a machine learning model produces predictions by explaining its judgements in a human-readable and interpretable format. The data is first prepared and divided into training and testing sets. Then, we use the lime package to describe the predictions made by a machine learning model that was trained using the h2o package. Lastly, to better understand how the model generated its predictions, we show the explanations [1]. The methods described in this project offers a useful manual for carrying out XML in R using LIME and H2O [2]. Data scientists and machine learning experts who wish to create more transparent and reliable AI systems may find this to be helpful.

A. Sports Analytics

Sports analytics is a burgeoning scientific field with potential applications in a variety of industries. Some of them include predicting an athlete's or a team's performance, determining an athlete's skill level and market value, and predicting an injury that might occur. To improve their plans, teams and coaches are increasingly willing to use such tools into their training. Technology innovation has made it simple and easy to acquire detailed data, which has sparked advancements in machine learning and data analytics. This supports marketing initiatives made by sporting organizations

to increase merchandise sales and build their fan base. Also, it allows them to run simulation games for matches they have yet to play and collect sponsorships. Sports organizations deploy data analysis to evaluate the performance of its athletes and determine the extent of personnel necessary to boost team performance. Additionally, it determines the weak and strong parts of the opposition, enabling coaches to select the best course of action. They may increase revenue, reduce costs, and guarantee superior investment returns by simple using data.

B. Explainable Machine Learning

The ability to illustrate the events occurring in the machine learning model is called explainability. It overcomes the black box dilemma and makes models transparent. The other term for this method is explainable AI (XAI), and it is used for all artificial intelligence. XAI refers to techniques that aid human specialists in comprehending AI- developed solutions. It is popular to use the terms "explainability" and "interpretability" interchangeably. Despite the fact that they both seek to "understand the model. Christoph Molnar defines interpretability in his book "Interpretable Machine Learning" as the extent to which a person can reliably forecast the outcomes of an ML model or comprehend the reasons behind a decision [6]. The following three criteria for model explainability are crucial: Transparency, Questioning skills and Ease of comprehension.

C. H2O AutoML

Automatic machine learning, also known as AutoML, is the technique of automating data pre-processing, hyperparameter tuning, model selection, and evaluation in the machine learning pipeline. H2O AutoML includes the cutting-edge and distributed implementation of numerous machine learning techniques. These algorithms are accessible in the programming languages Java, Python, Spark, Scala, and R. Moreover, H2O offers a web GUI that implements these techniques using JSON. The models developed using H2O AutoML can be quickly deployed on the Spark server, AWS, etc. The key benefit of H2O AutoML is that it automates procedures such as basic data processing, model training and tuning, Ensemble and stacking of several models to deliver the

optimum performance for the models, allowing developers to focus on other tasks such as data collecting, feature engineering, and model deployment.

D. LIME

The term is an acronym for Local Interpretable Model Agnostic Explanation. The use of the word "local" describes how it is used to explain certain predictions generated by a machine learning model. LIME (Local Interpretable Modelagnostic Explanations), an explainable AI technique, aids in illustrating a machine learning model and making each prediction's specific implications evident. The method is suitable for local explanations because it describes the classifier specifically for a single case. LIME modifies the initial information to create a sequence of artificial data that only partially preserves the original attributes. As a result, numerous duplicates of the original text are formed, for example, in the case of text data, in which a specific number of various, randomly selected words are deleted. Afterwards, other classifications are given to these new synthetic data (classified). As a result, we may determine the impact of particular keywords on the classification of the chosen text by observing their presence or absence.

II. LITERARY REVIEW

Vangelis Sarlis and Christos Tjortjis addressed the idea of explainable machine learning in their work Sports analyticsassessment of team performance and basketball players and distinguished between transparency, interpretability, and explainability. Also, they talked on the potential for influencing model design decisions as well as the process of using domain expertise and a posteriori consistency checks to interpret algorithmic outputs. By using a few recent exemplary studies in the field of machine learning in the natural sciences, they provided a more detailed characterisation of the various stages of explainability [3,7]. Vaishak Belle and Ioannis Papantonis carried out research for their work, Principles and Practice of Explainable Machine Learning, in order to better clarify the topic of explainable machine learning and apply the appropriate tools for industry practitioners (but also data scientists in general). The last sections create a story around a hypothetical data scientist and explore how they may approach describing her models by posing the appropriate queries. After motivating the subject broadly from an organisational perspective, they highlight the important breakthroughs, such as the notions that allow us to distinguish between transparent and opaque models and model-specific or model-neutral post-hoc explainability approaches. Before concluding with a discussion of prospective future study areas, they also briefly examine deep learning models [4,6]. Konstantinos Apostolou and Christos Tjortjis conducted trials utilizing needed algorithms, primarily on dataset related to football, to forecast a player's position on the field in their article Sports Analytics algorithms for performance prediction. For the purpose of trying to predict a player's goal scoring performance in the future season, statistics from prior seasons were gathered as well as information on the

number of shots a player takes during a game, which is believed to be related to goal scoring likelihood. Findings demonstrate good accuracy and are quite encouraging, especially given that the anticipated and actual goal totals were fairly close [1,2,5].

III. PROPOSED METHOD

In this paper we are going to apply explainable machine learning using LIME and H2O AutoML. We have extracted IPL Auction data from Kaggle, which has 20 columns describing various features of a player like base price and previous IPL teams. After importing the various libraries and dataset, we proceeded with EDA. Then we initiated the H20 and trained the models with dataset and choose an algorithm to explore the explanation. In the final steps we decided the metric by which explanation will be performed and provided the lime for explainer.

Table 1: Dataset Atrriubutes Table

Pinyan	1000 Set	Country	State Association	Age	Specialism	Butting Style	Bowling Style	Test cape	OOX caps	Titl cape	-	Previous IP(Teambé	9001 Years	2001 Ft.	GNIA	Reserve Price Rs Laith	Price Pald	Sean.	Ber
T. Authorite	w	Tefa	790A	35	ALI POUNDER	five .	PROFE WAY OF Spin.	84.0	112.0	91.6	167.0	CSK, RSR, KKIP, DC	00	13.0	Capped	200	5000008.0	PR.	See
Treet Boult	w	New Zeeland	*	30	BOHLER	no.	LEFT ARM Feet Wedlum	74.0	90.0	64.0	62.0	SPH, KHIR, DO,ME	M	14.0	Capped	200	1,000000	nn.	Det
Pat Cummins	w	Australia		10	ALL FOUNDER	rva .	RIGHT WAN Fast	36.0	60.0	17.6	814	00.WL00R	100	7.0	Capped	200	79100008.6	100	See
Quinton the Knoth	w	South Africa		29	WONEDWERF	LHB		94.0	126.0	61.0	710	SPH, DO, PCB, MI	M	11.0	Capped	200	6700008.0	1.00	See
Shidhar Ohawan	w	Pefin	00CA	×	BATSMAN	LHB		34.0	146.0	44.0	190.0	00H, MI, 68H, 00	00	16.0	Capped	200	80100000.0	PK	See
Full Du Pleases	w	South Africa	*	327	BATSMAN	Pris .	RIGHT WWW Lag Spin.	66.0	143.0	98.0	100.0	APS, CISK	CSK	16.0	Capped	200	7000008.0	PCB	Date
Shrayen Yer	w	rule	WGA	22	BATSMAN	rva .	RiGHT MW Lag Spin.	14	20.0	38.6	87.0	00	00		Capped	200	122500064	100	See
Kagiso Rabada	w	South Africa		24	BOMLER	LHB .	RIGHT WAN Fast	80.0	60.0	46.0	86.0	00	00	15.0	Capped	200	90'000008.6	PK	bee
Mohammad Shami	w	India	CAB	102	BOWLER	RHS.	RIGHT ARM Fast	67.0	79.0	17.6	me	DOIL HHR. DD. PSHS	гека	14.0	Capped	200	62500008.6	67	546
Devid Warner	w	Autralia	*	35	BATSMAN	LHB		91.0	126.0	88.0	190.0	00,5RH	SAH	8.0	Capped	200	42100000.0	00	See
Shinnes Helimper	641	West Index		29	BATSMAN	LHB .	NGHT MW Lag Spin.	16.0	67.0	42.0	31.0	ACE DC	00	140	Capped	198	850000 E	PR.	bee
Dentel Miller	641	South Africa		32	BATSMAN	LHB		64	136.0	95.0	86.0	KOSE MR	nn.	9.0	Capped	100	3000008.6	61	bee
Devotati Padhilud	847	inde	MSGA.	21	BATSMAN	LHB	RIGHT WAN OR Spin.	6.0	0.0	2.0	25.0	ACB	PCB	14.0	Capped	299	7790000.0	PR	546
Manish Panday	847	India	KSCA	32	BATSMAN	PHG .	RIGHT WAY OR Spin.	9.0	29.2	38.0	154.0	MLRCE, PML KKR, SPH	SAH	8.0	Capped	100	4600000.0	156	See
Suresh Keine	647	India	UPCA	39	BATSMAN	UB.	NGHT WW OR Spin.	16.0	3363	79.0	205.0	GSK, GL	CSK	12.0	Capped	260	9.6	DOM	e Dide
James Roy	647	England		an	BATSMAN	Red.		54	96.0	59.0	184	GL DC 59H	SRH	54	Capped	200	2000008.0	61	Dog
Stave Smith	847	Autoria		32	BATSMAN	RHS		60.0	126.0	9.0	100.0	PWLRPS, RR, DC	00	80	Capped	200	9.6	Uda	e Ude
Robin Uthappe	8A7	Pda	RCA	36	BATSMIN	PHS .		0.0	462	13.0	190.0	MI, PW, ROE, KIR, RR, CSK.	CSK	40	Capped	200	2000008.0	CSK	See
Shakib N Hasan	4,1	Berglebsh		35	AL FOLKER	LHB	LEFT ARM Slow Orthodox	56.0	215.0	94.0	71.0	584,400	1001	80	Capped	200	9.6	Uda	e Urbe
Designe Brass	4,1	West Index		38	AL-ROUNDER	Red.	NGrT WW Feet Medium	400	164.0	91.0	151.0	W. G., CSK	CSK	11.0	Capped	200	440000.0	CSK	See
Warindu Hasarange	46,5	SriLania		24	ALI-ROUNDER	RHB	RiGHT ARM Lag Spin	40	29.0	33.0	2.0	ACB	PCB .	2.0	Capped	100	15750008.0	PCB.	500
Janua Holder	463	West Index		30	ALI-ROUNDER	PHB .	PIGHT HIM FWE	50.0	124.0	30.0	26.0	CSK, KRPI, SPH	SRH	80	Capped	198	670000.0	1.50	See
Mitchell Marsh	4,1	Australia		30	ALCOHOLD	Red.	RGrT WW Feet Medium	30.0	60.0	36.0	210	DOIL PMI, RPS, SPH	Do No Play	0.0	Capped	200	4500000.0	00	See
Mohammad Natio	4.1	Alphanotes		32	ALCOHOLD	PriS.	RIGHT WWW ON Sele-	3.0	127.0	86.0	17.0	SNH	SPH	3.0	Capped	100	1000008.0	1001	See
Knowl Pandys	46.5	India	96A	91	ALL-ROUNDER	LHB	LEFT ARM Slow Orthodox	0.0	5.0	19.0	840	w	M	130	Capped	200	1000000	1.99	See
Harshall Patrol	46.1	refe	HCA	31	ALL FOUNDER	nea .	PICHT RPM Feet Medium	0.0	0.0	2.0	65.0	DC.908	PICB .	15.0	Capped	200	157900005.0	PCB	See
Nition Force	44.1	India	DOCA	28	ALL ROUNDER	LHB	TROUT ARM OF Sex.	0.0	1.0	2.0	TTO	W. 607	10th	17.0	Capped	100	1000000	1001	See
Washington Sundar	44.1	India	TNOA	22	ALL ROUNDER	LHB	TROUT ARM OF Sex.	4.0	1.0	21.0	40.0	nce	nce	60	Capped	190	6700000.0	55H	See
Joney Bainstow	880	England		100	WONEDWERFER	rva .		800	80.0	10.0	20.0	Silve	68st	7.0	Capped	100	470mm r	PK	See

This flow chart defines when to perform each step and the prerequisites to ensure before moving onto the next step. The flow chart of my implantation of the methods is mentioned below:

Table 2: Steps for Creating Model

Step Number	Step Name
1	Importing the Libraries
2	Importing Dataset
3	Data Pre-Processing
4	Initialize H2o And Split The Data
5	Training Models
6	Leaderboard & Model Selection
7	Metrics
8	Explainer
9	Explanation

IV. IMPLEMENTATION

The implementation is performed in the following steps:

E. Importing the Libraries

The libraries used for this project are:

- tidyverse- it is used for data science
- h2o- it is used for running h2o and communicate with it
- lime- it is used for creating model explanations
- recipes- it is used for data pre-processing.

F. Importing Dataset

In this step we imported the Auction dataset from Kaggle and loaded into the R studio.

G. Data Preprocessing

In this step we used functions like:

- recipe_data- this method is used to build the recipe for EDA.
- step_rm- this method is used to remove specific columns.
- step_scale- this method is used to normalize data so that it has standard deviation of 1.
- step_center- this method is used to normalize data so that it has mean of 0.

H. Initialize H2O and Split the data

In this step we used functions like:

- h2o.init- it is used to start the h2o cluster.
- splits- it is used to divide the dataset into multiple parts.

I. Training Models

In this step we used h20.automl to train the machine learning models.

J. Leaderboard Selection

In this step we select the model by analysing the leaderboard results and select an algorithm of our choice.

K. Metrics Formulation

In this step we select the metric by which explanations will be performed. The functions used are:

- perf
- metrics

L. Explainer

This is the final step in which we define the explainer and explanation. These two functions are defined to give a particular formatting to the results.

V. RESULT ANALYSIS

As in this paper we implement explainable ML. These two figures are the final explanations obtained. By observing figure 2 we can see each individual case and identify which metric is more important for a player and in case 1 we can see, the metric which has a positive relation for the player being picked is its team and reserve price. Whereas age has a negative relation with it. In figure 3 we observe the combine explanation of metrics given importance to and the player cases.

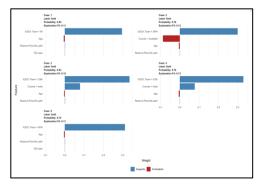


Fig. 2: Explainer's Explanation

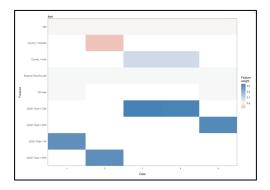


Fig. 3: Explanation's Explanation

VI. CONCLUSION AND FUTURE SCOPE

In recent years, the use of explainable machine learning has been growing significantly as sophisticated algorithms are used to make judgements that have an influence on people's lives in the real world. Using LIME and H2O to implement explainable machine learning in R can give stakeholders a framework for comprehending the predictions of complex models and can boost stakeholder confidence in these technologies. As we can see through the findings presented explainable ML can provide deep and useful insights for a subject. Sports and data have been inextricably connected. Like newspapers report box scores and baseball cards display a player's career stats, radio broadcasters have historically used facts to provide context to their commentary, such as the average amount of yards a running back has gained in each game they have played. Players are assessed by general managers and coaches using a combination of stats, such as batting average, points or yards thrown and a subjective gut feeling, such as "This player is out for a hit." Beyond those superficial statistics, sports analytics with the use of explainable ML can provide concrete evidence on the selection or rejection of a player. In conclusion, Explainable Machine Learning with LIME and H2O in R offers data scientists and practitioners a useful tool to create and comprehend sophisticated machine learning models.

REFERENCES

[1] Konstantinos A, and Christos T, "Sports analytics algorithms for performance prediction," 2019 10th International Conference on

- Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900754.
- [2] Starting H2O H2O 3.40.0.2 documentation.(n.d.-b). https://docs.h2o.ai/h2o/latest-stable/h2o-docs/starting-h2o.html.
- [3] Vangelis S, & Christos T., "Sports analytics Evaluation of basketball players and team performance. Information Systems", 93, 101562. https://doi.org/10.1016/j.is.2020.101562, 2020.
- [4] Belle, Vaishak, and Ioannis Papantonis, "Principles and Practice of Explainable Machine Learning", Frontiers in Big Data, vol. 4, 2021. Frontiers,https://www.frontiersin.org/articles/10.3389/fdata.2021.68896 9
- [5] Onose, E. (2023, April 19). Explainability and auditability in ML: Definitions, techniques, and Tools. neptune.ai. https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools
- [6] Manocha, T., & Sharma, V. (2020). Study on the readiness among Youth towards Industry 4.0. Scopus Indexed Journal 'International Journal of Advanced Science and Technology', 29(3).
- [7] Sharma, V., & Manocha, T. (2023). Comparative Analysis of Online Fashion Retailers Using Customer Sentiment Analysis on Twitter. Available at SSRN 4361107