A Practical Introduction to SHAP for Making Sense of Machine Learning Models

Insights Family

Luciano Gerber

Senior Lecturer
Department of Computing and Mathematics
Faculty of Science and Engineering
Manchester Metropolitan University

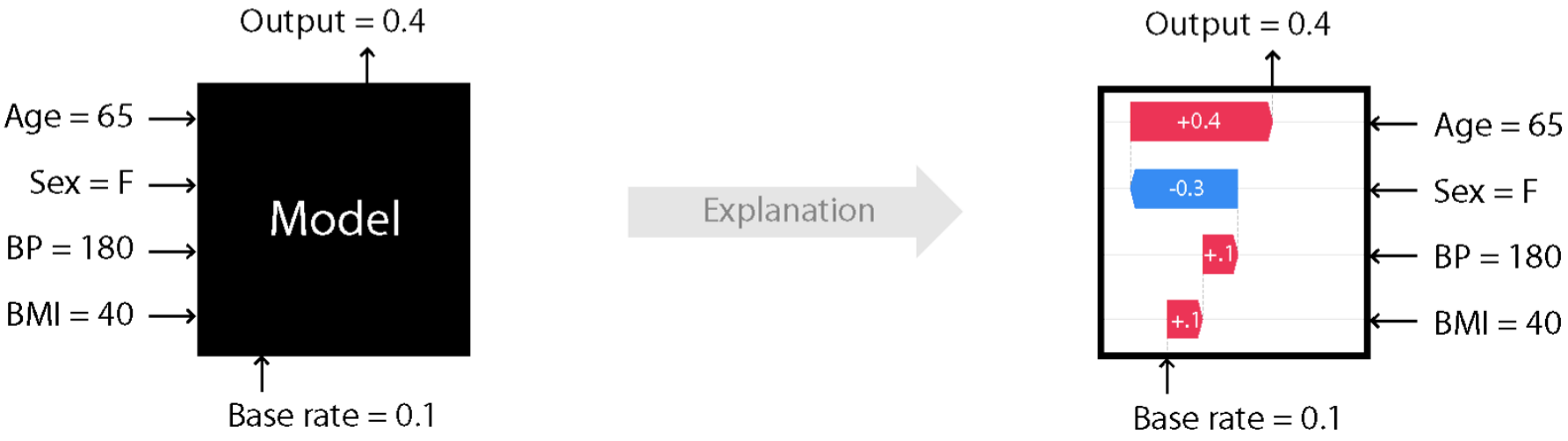
21-Jul-22

Intro

Motivation

- Increasingly, models with better performance but higher complexity; opaque; too many moving parts; high cognitive load.
- Tree Ensembles, Neural Networks, ...
- Making sense of models becomes a task of probing, of using proxy structures for observing a model's predictive patterns, of measuring the effect of the inputs (i.e., features and values) in the outcomes.
- SHAP is one of the most popular, recent techniques for doing so.

The Gist of SHAP



[1]

[1] https://github.com/slundberg/shap

Making Sense of Models via Local and Global Explanations

- explain model predictions to stakeholders (e.g., domain experts, data scientists, end user (?)).
- **local**: for a prediction on a single instance, which (quantified) effect does each feature has on the outcome?
- **global**: what are the most significant features in terms of the prevalence and magnitude of the effect on the outcomes?

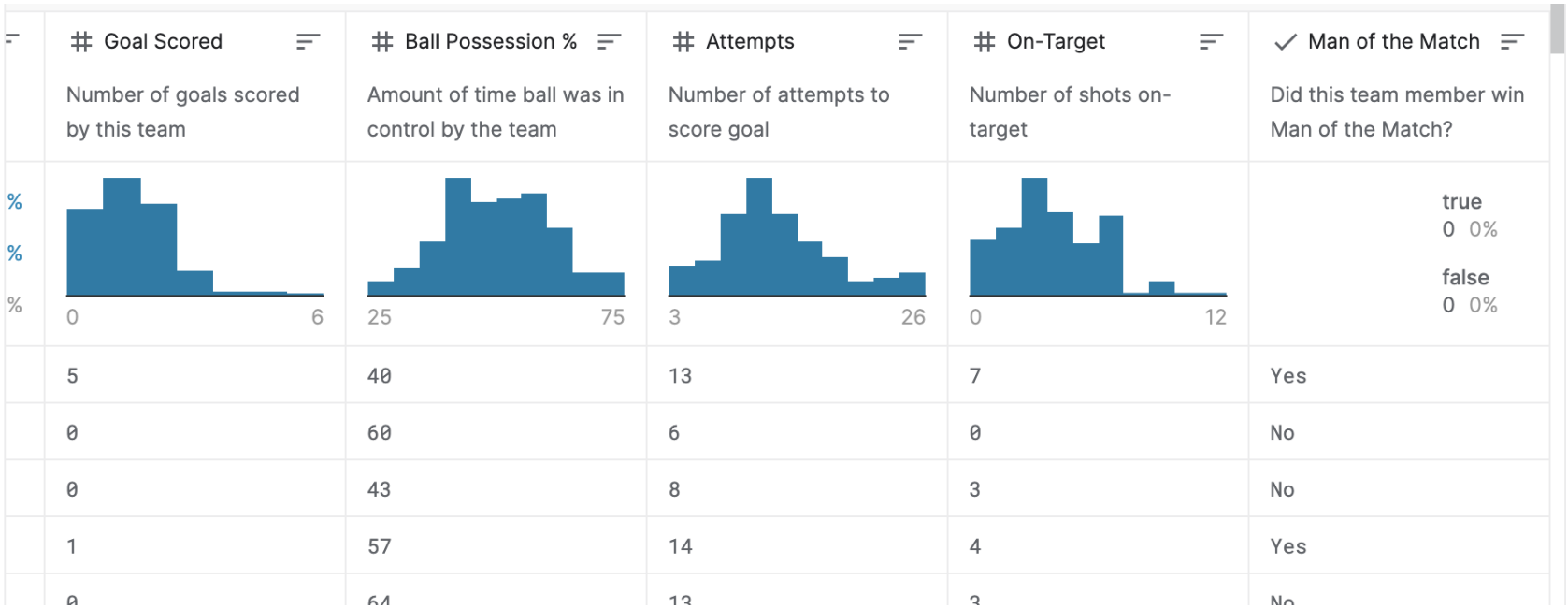
Useful Properties

- **Model-agnostic** (however, model-specific versions are offered for better quality/performance)
- The quantification of the effects is on the **same unit** as the predictions (e.g., probabilities, risk, house prices, salaries).
- **Additive**: for a single prediction, outcome = base value + sum of each feature's contribution.
- Python package with some quite useful pre-defined plots.

Example of Use and Application

Example: Dataset[2]

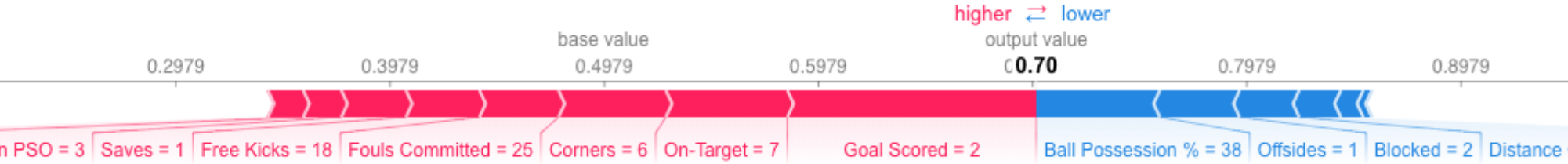
- my apologies for the men's football example, but hopefully it is easily understood.
- probability of `man-of-the-match` based on team/match info and statistical data.



Goal Scored	Ball Possession %	Attempts	On-Target	Man of the Match
Number of goals scored by this team	Amount of time ball was in control by the team	Number of attempts to score goal	Number of shots on-target	Did this team member win Man of the Match?
0 – 6	25 – 75	3 – 26	0 – 12	true 0 0%; false 0 0%
5	40	13	7	Yes
0	60	6	0	No
0	43	8	3	No
1	57	14	4	Yes
0	64	13	3	No

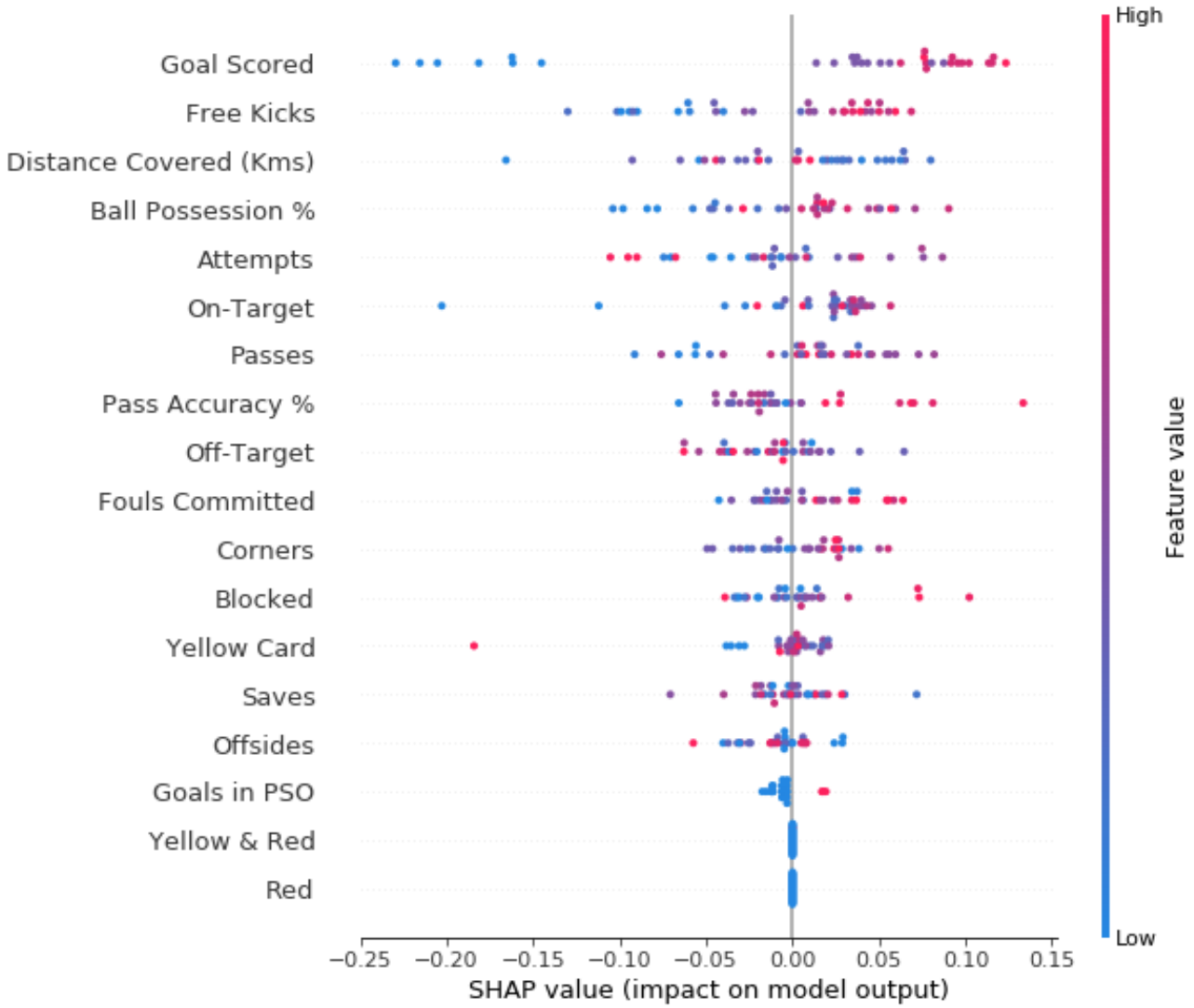
[2] Kaggle's FIFA 2018 Man of the Match at https://www.kaggle.com/mathan/fifa-2018-match-statistics

Example: Local (One Instance) Explanation[3]



[3] https://www.kaggle.com/dansbecker/advanced-uses-of-shap-values

Example: Global Explanations (Many Features, Many Instances)[4]



[4] https://www.kaggle.com/dansbecker/advanced-uses-of-shap-values

How It Works

Shapley Values (from Game Theory)[5]

Shapley Values Defined in Game Theory

How to Divide Money Between Game Players in a Fair Way?

Fairness properties:

- Additivity (Amounts sum up to the final game result)
- Consistency (More effect – not less money)

Theorem: The only fair way to distribute the money is:

Shapley Value for player i in game f = Average over all players' subsets $S \subseteq M/\{i\}$ [Marginal contribution $f(S \cup \{i\}) - f(S)$]

f – game
M – all players
S – subset of players
i – specific player

Money difference with and without this player

PayPal

Lloyd S Shapley. "A value for n-person games" (1953)

5

[5] SHAP Values for ML Explainability, Adi Watzman, https://www.youtube.com/watch?v=0yXtdkIL3Xk&t.

SHAP: Adapting Shapley Values to ML[6]

SHAP Values for Local Feature Attributions

How to Assign Contribution Between Features in a Human-Intuitive Way?

SHAP Value for feature i, contribution of feature i to model prediction $f(x)$ = Average over all features' subsets $S \subseteq M/\{i\}$ $\left(f(S \cup \{i\}) - f(S) \right)$

f – ~~game~~ model
M – all ~~players~~ features
S – subset of ~~players~~ features
i – specific ~~player~~ feature
x – instance being explained

Exponential time

How to run a model with missing features?

By following Shapley definitions, SHAP is proved to be the unique solution that satisfies additivity and consistency → More contribution of i → greater or equal SHAP value for i

Naïve solution: Train a model for every subset of features

PayPal

Lundberg et al: NeurIPS (2017) & arXiv preprint (2019) 6

[6] SHAP Values for ML Explainability, Adi Watzman, https://www.youtube.com/watch?v=0yXtdkIL3Xk&t.

Relationship with Permutation Importance[7]

```
import eli5
from eli5.sklearn import PermutationImportance
```

Weight	Feature
0.1750 ± 0.0848	Goal Scored
0.0500 ± 0.0637	Distance Covered (Kms)
0.0437 ± 0.0637	Yellow Card
0.0187 ± 0.0500	Off-Target
0.0187 ± 0.0637	Free Kicks
0.0187 ± 0.0637	Fouls Committed
0.0125 ± 0.0637	Pass Accuracy %
0.0125 ± 0.0306	Blocked
0.0063 ± 0.0612	Saves
0.0063 ± 0.0250	Ball Possession %
0 ± 0.0000	Red
0 ± 0.0000	Yellow & Red
0.0000 ± 0.0559	On-Target
-0.0063 ± 0.0729	Offsides
-0.0063 ± 0.0919	Corners
-0.0063 ± 0.0250	Goals in PSO
-0.0187 ± 0.0306	Attempts
-0.0500 ± 0.0637	Passes

[7] https://www.kaggle.com/dansbecker/permutation-importance.

Conclusion

Further Considerations

- Can these tick "right to explanation" to different stakeholders?
- **Not causation**.
- Counterfactuals?
- Specialised explainers for trees, DNNs.
- NLP, images.

Resources

- the original paper on SHAP at NeurIPS. More accessible resources are:
  - This section of the Interpretable Machine Learning book.
  - The examples and explanations on The GitHub SHAP repo.
  - This brief, hands-on Kaggle tutorial.