Advancing Action-Level Soccer Analytics:

A Comparative Study of VAEP Model Enhancements Using Division 1 Women's Collegiate Soccer Event Data

---

Benjamin Thorpe

Advisor: Dr. Jerry Reiter

Acknowledgements

- My advisor: Dr. Reiter
- Leo Biral and Coach Kieran Hall
- My committee members: Dr. Tackett and Dr. Rundel
- Dr. Durso
- My parents, especially my mom

How can event-level soccer data best be utilized by a data scientist to provide actionable insights to a coaching staff?

My research looks to answer two main questions as potential solutions to the above:

1. Will a more detailed dataset improve our ability to assign quantitative values to actions?

2. Can the variables used in quantifying actions provide insights that can help soccer teams win games?

Data

Match Period	Minute	Second	Action Type	Action Location
1H	6	51	{'primary': 'duel', 'secondary': ['aerial_duel', 'recovery', 'counterpressing_recovery']}	{'x': 61, 'y': 23}
1H	6	54	{'primary': 'interception', 'secondary': ['progressive_run', 'carry']}	{'x': 77, 'y': 10}
1H	6	54	{'primary': 'duel', 'secondary': ['defensive_duel', 'ground_duel']}	{'x': 18, 'y': 89}

{'accurate': True, 'angle': 110, 'height': None, 'length': 12, 'recipient': {'id': 689757, 'name': 'J. Echegini', 'position': 'LW'}, 'endLocation': {'x': 80, 'y': 40}}

Pass

{'bodyPart': 'right_foot', 'isGoal': True, 'onTarget': True, 'goalZone': 'gt', 'xg': 0.05795, 'postShotXg': 0.07749, 'goalkeeperActionId': 1830086339, 'goalkeeper': {'id': 688133, 'name': 'H. Mackiewicz'}}

Shot

- Each observation is one action
- From all ACC women’s soccer games over the last two seasons (151 total games)
- Used first 80% of games as the training data and latter 20% for model evaluation

VAEP Model Framework

- VAEP – **Valuing Actions by Estimating Probabilities**

- Assigns a numerical value to each action based on its **impact on the likelihood of scoring or conceding** a goal

- Provides a more **objective and comprehensive measure** of a player's contribution to the team's success

Tom Decroos. 2018. Actions Speak Louder than Goals: Valuing Player Actions in Soccer.

	TIME	PLAYER	ACTION	P_{scores}	VALUE
1	92m4s	S. Busquets	pass	0.03	0.00
2	92m6s	L. Messi	pass	0.02	- 0.01
3	92m8s	S. Busquets	pass	0.03	+ 0.01
4	92m11s	L. Messi	take on	0.08	+ 0.05
5	92m12s	L. Messi	pass	0.17	+ 0.09
6	92m14s	A. Vidal	shot	1.00	+ 0.83

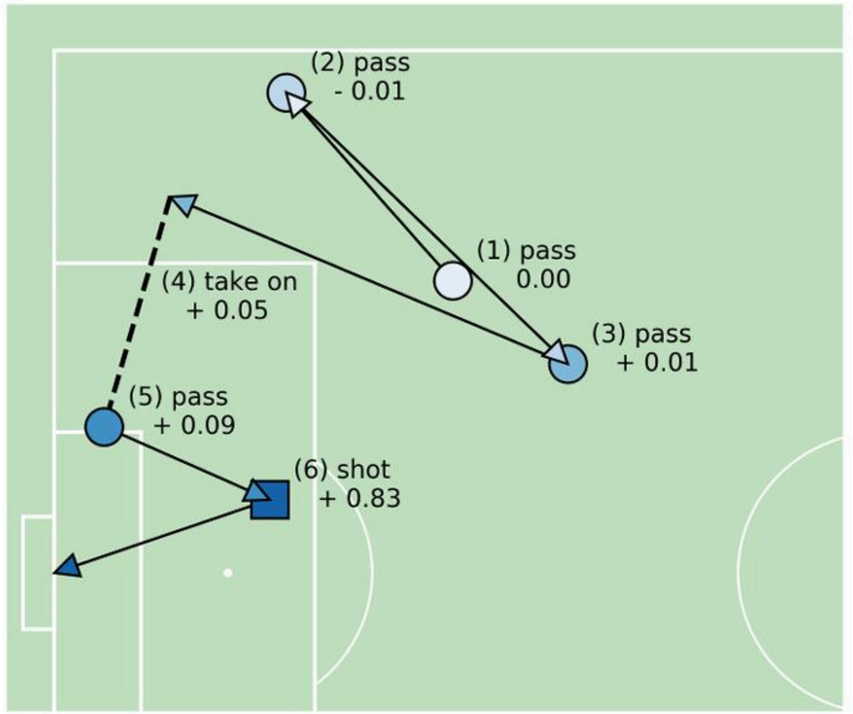


Figure 1: The attack leading up to Barcelona's final goal in their 3-0 win against Real Madrid on December 23, 2017.

VAEP Model

- Consists of two distinct models: one for **scoring** and another for **conceding**
- Considers the **type** of action, the action's **location** on the pitch, the **game context**, etc.
- **XGBoost** selected as the underlying classification model

Relevant variables:

- **S** - The game state which is a set of actions
- **t** - The team with possession during S
- **a** - An action
- **k** - the number of actions to look back to in defining the outcome variable

Value for the *i*th game state in a given soccer match

$$V(S_i) = P_{score}(S_i, t) - P_{concede}(S_i, t)$$

VAEP score for the *i*th action in a given soccer match

$$\Delta P_{\text{score}}(a_i, t) = P^k_{\text{score}}(S_i, t) - P^k_{\text{score}}(S_{i-1}, t)$$

Data Comparison

- Wyscout Version 2 vs. Version 3
- Model parameters: ***j*** and ***k***
- Decision on using AUROC or Brier Score for model evaluation

Will the more detailed model perform better?

Variable Importance Analysis

- Why it is important?
- Shapley values and “beeswarm” plots
- Three groups of modeling experiments:
  - Base Scoring and Conceding
  - Passing and Crossing
  - Random Results for Scoring and Conceding

Are there any trends in which variables are important and how they are correlated with the response variable? If so, how can a team use these to improve their overall strategy?

J values

j - a set number of actions which define a game state

j	Model Type	V2 AUROC	V3 AUROC	V3 AUROC Improvement
3	concedes	0.741794	0.793788	0.051994
3	scores	0.785398	0.777370	-0.008028
6	concedes	0.739896	0.793511	0.053615
6	scores	0.785585	0.779724	-0.005861
9	concedes	0.738152	0.793884	0.055731
9	scores	0.783248	0.776421	-0.006827

Ex: For *j*=3, game state 20 is defined as S20 = {a18, a19, a20} where a20 is the 20th action that occurs in a game

K values

k - determines the number of actions to look back to in defining the outcome variable

k	Model Type	V2 AUROC	V3 AUROC	V3 AUROC Improvement
3	concedes	0.930162	0.951997	0.021836
3	scores	0.957101	0.968783	0.011682
6	concedes	0.837349	0.886612	0.049262
6	scores	0.853636	0.856935	0.003298
10	concedes	0.741794	0.793788	0.051994
10	scores	0.785398	0.777370	-0.008028
13	concedes	0.712236	0.745270	0.033034
13	scores	0.752618	0.744161	-0.008457

Ex: For *k*=10, if a goal is scored in any action between a20 and a30 then S20 would be assigned a positive label in the scoring model

Comparing Data Versions

- **AUROC** is chosen since it measures the ability of a model to **distinguish between classes**
- Results show V3 **improves the conceding** model significantly
- Difference in **scoring** model fairly **negligible**
- **V3** model with ***j*=3 and *k*=6** chosen for use in analysis

k	Model Type	V2 AUROC	V3 AUROC	V3 AUROC Improvement
3	concedes	0.930162	0.951997	0.021836
3	scores	0.957101	0.968783	0.011682
6	concedes	0.837349	0.886612	0.049262
6	scores	0.853636	0.856935	0.003298
10	concedes	0.741794	0.793788	0.051994
10	scores	0.785398	0.777370	-0.008028
13	concedes	0.712236	0.745270	0.033034
13	scores	0.752618	0.744161	-0.008457

Variable Analysis Reasoning

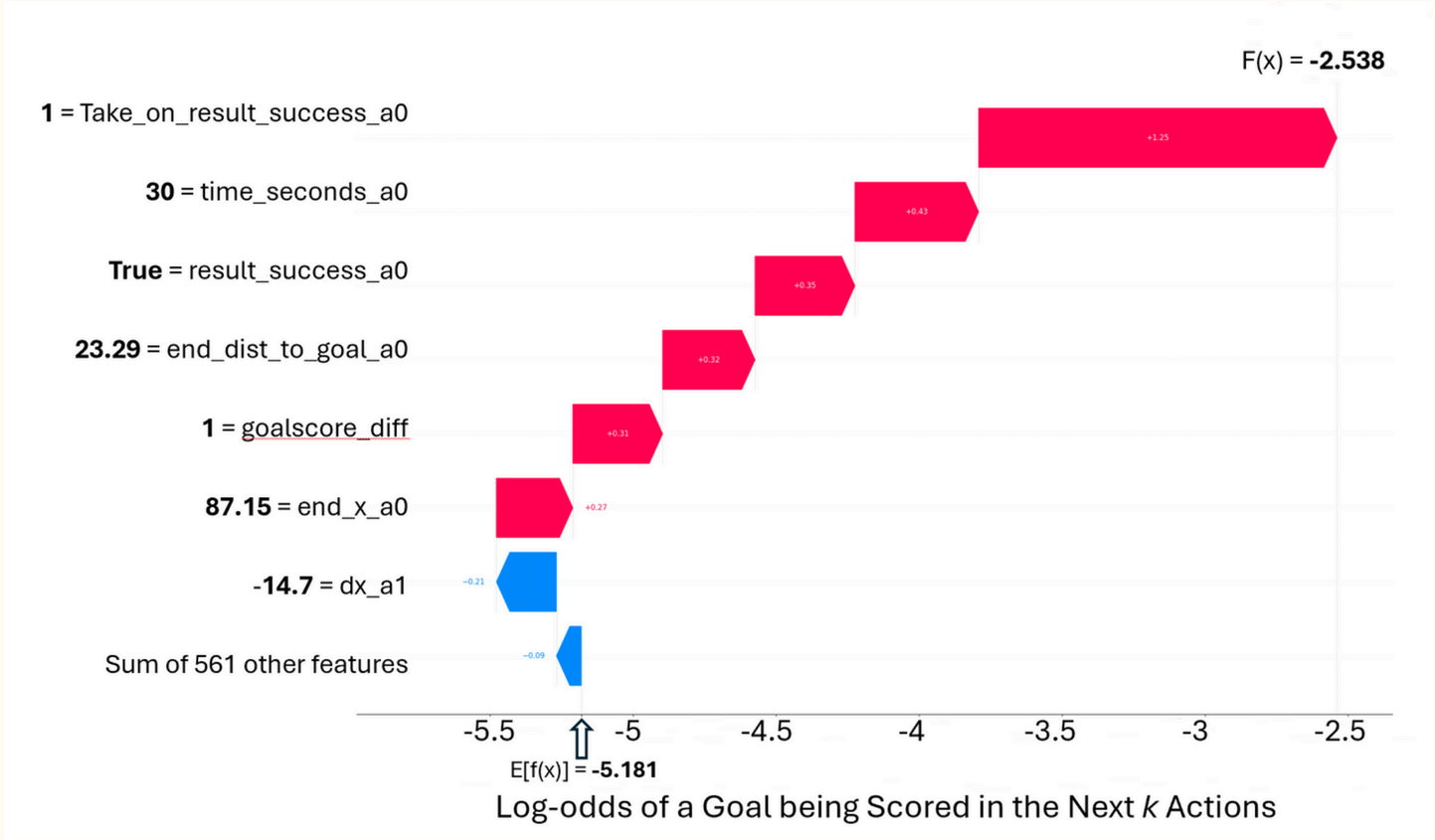
- Bridges gap between statistical models and practical insights
- Can provide further detail on player and team style and tendencies
- Contributes to interpretability in machine learning

Shapley (and SHAP) Values

- Shapley values aim to attribute the **contribution of each variable** to the prediction of a model

- **SHAP values** – Shapley values applied to a conditional expectation function of a model

- Are a good fit due to their **clarity** and **interpretability**



Beeswarm Plots

- Effectively displays the **distribution** and **impact** of Shapley values
- Each **dot** represents one **observation** for a **given feature**
- The **color** gradient helps to **correlate** the feature's **observed value** with its **effect on** the output

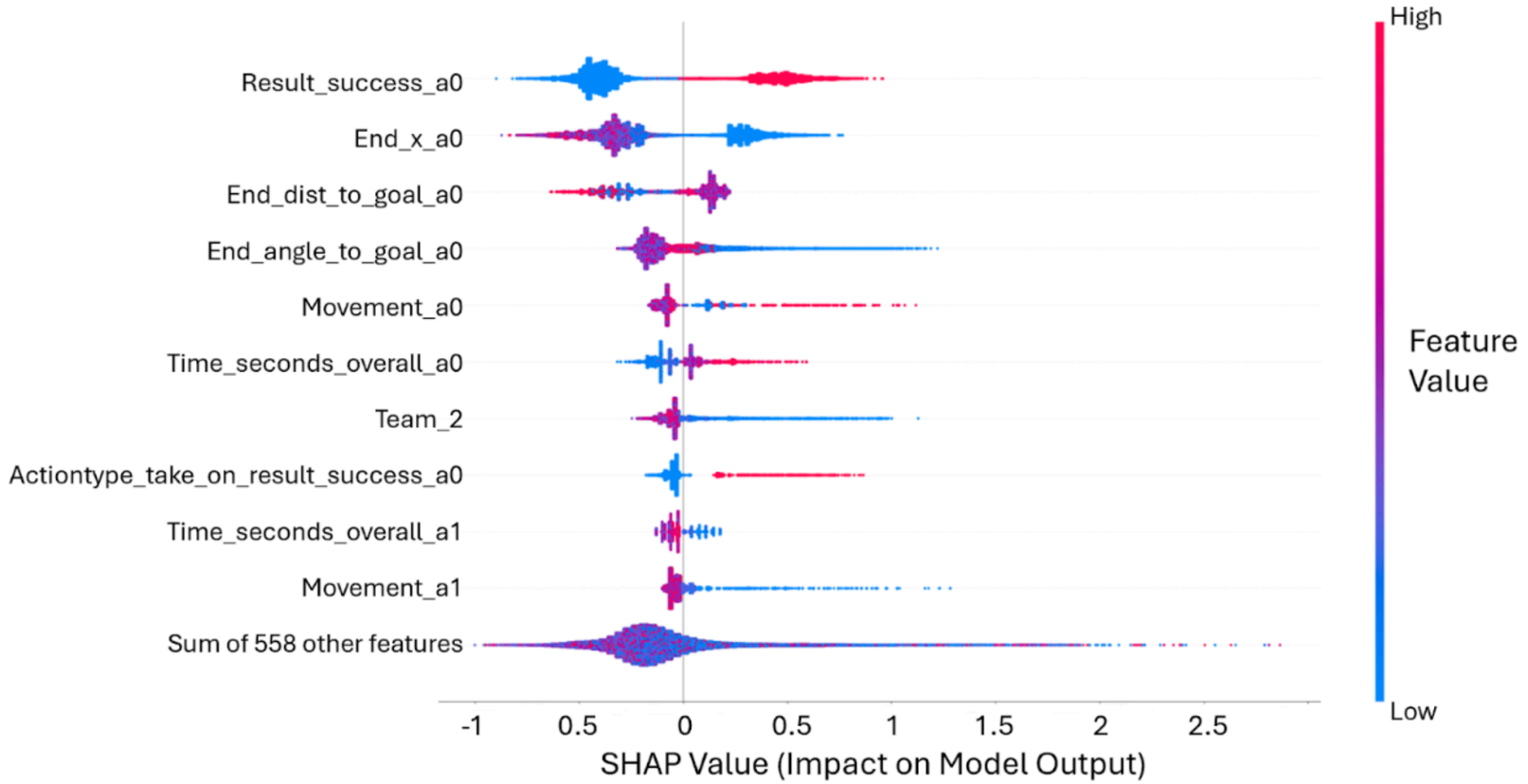


Figure 9: *j*=3 and *k*=6 Scoring Model Beeswarm Plot

Base Models

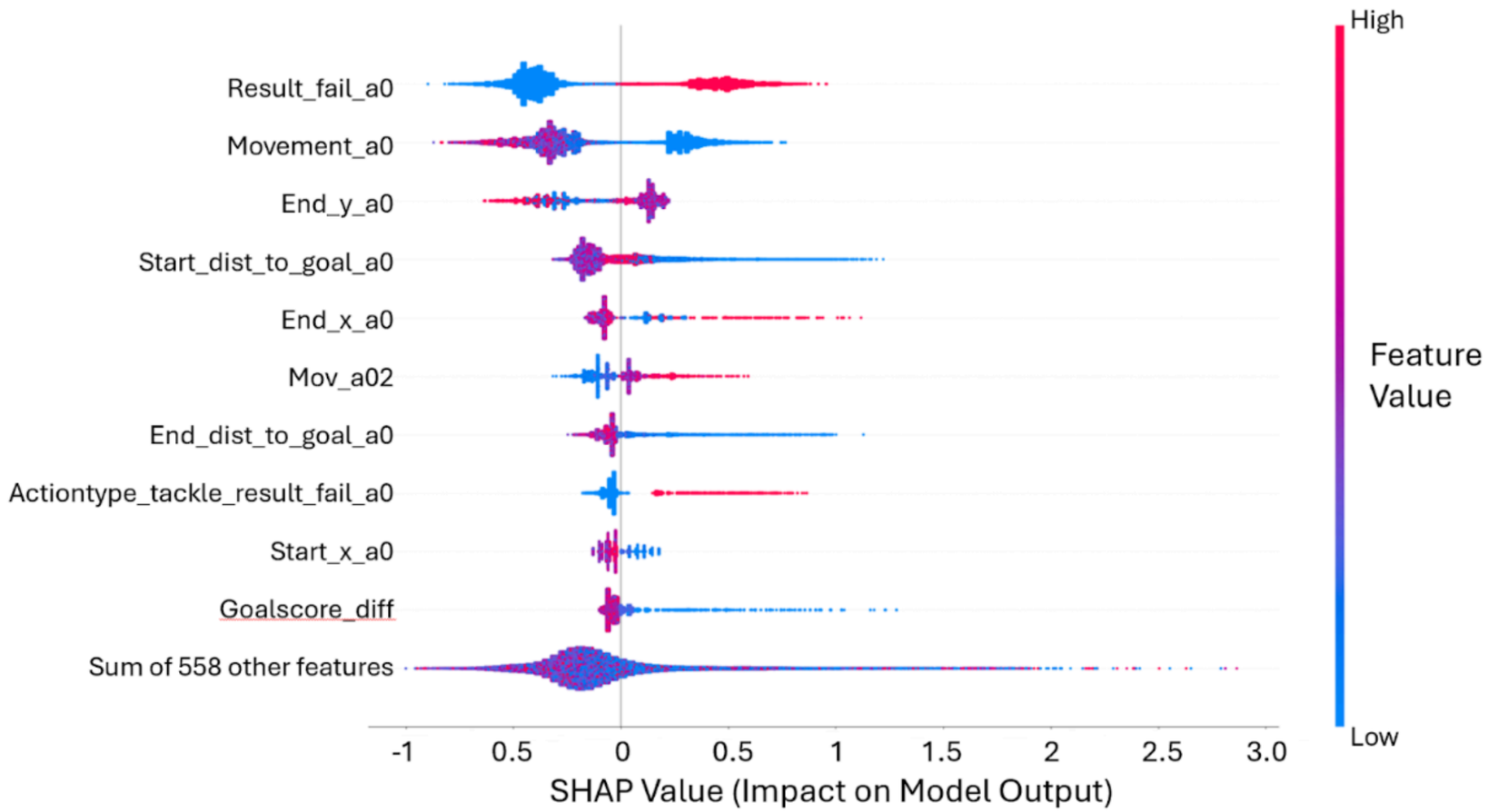


Figure 10: *j*=3 and *k*=6 Conceding Model Beeswarm Plot

Key Findings:

- Goals are more likely to come later in games
- Quick counterattacks look to be effective

Passing and Crossing Models

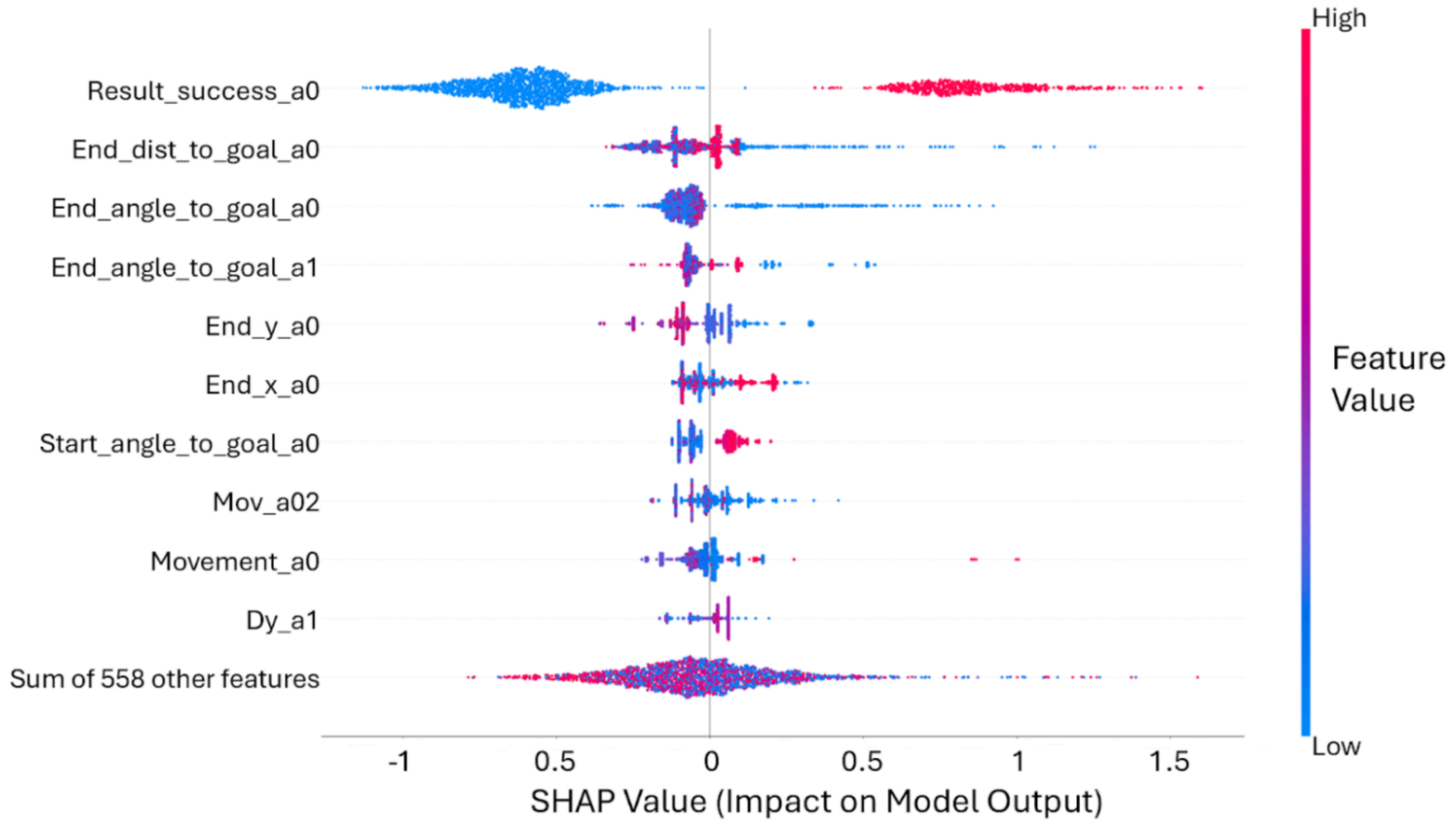


Figure 12: Crossing Scoring Model

Key Findings:

- Central play is valuable in creating opportunities
- Crosses most successful going from wide-to-center

Random Results Models

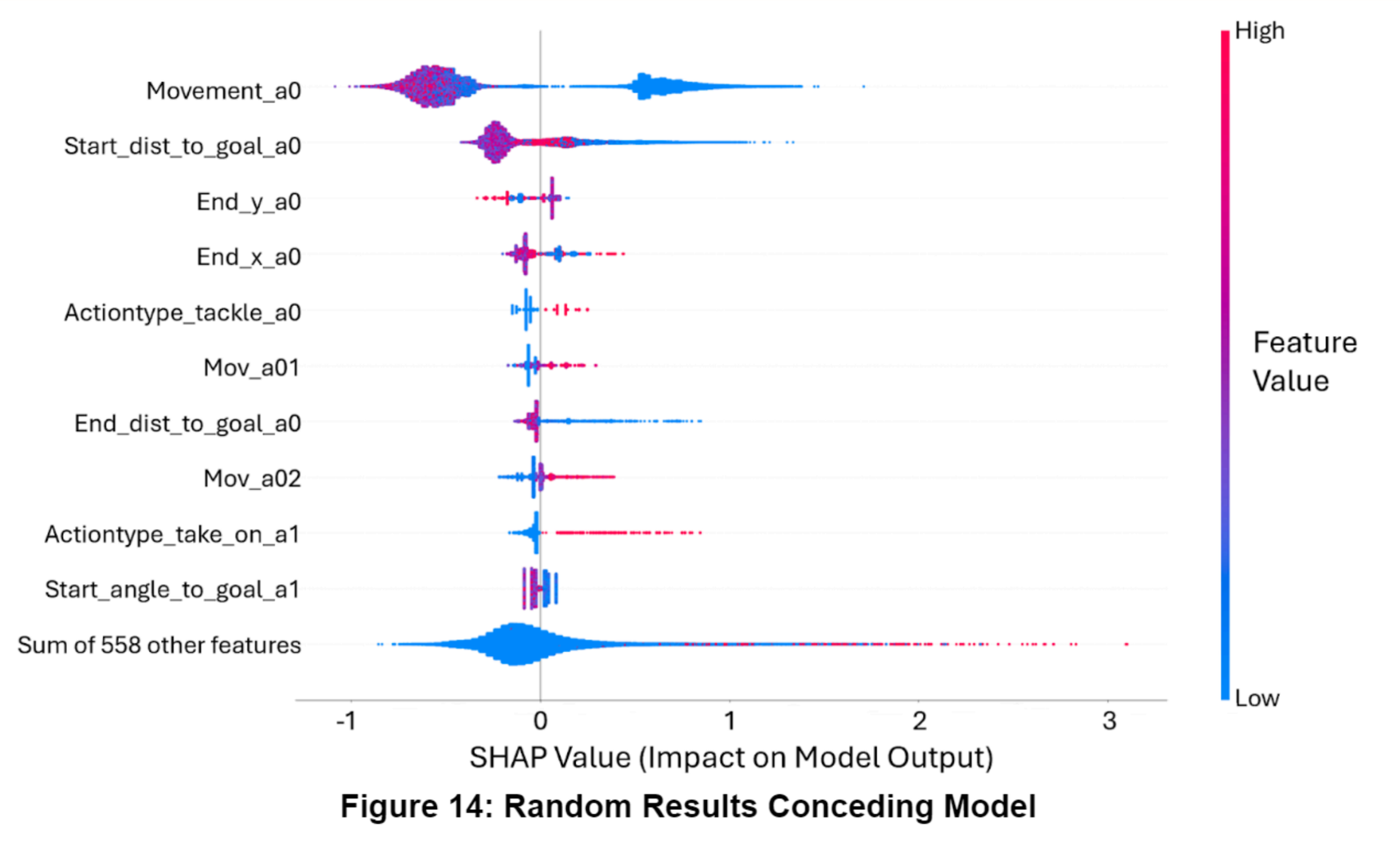


Figure 14: Random Results Conceding Model

Key Findings:

- Movement in actions becomes relevant
- Losing possession off a dribble correlates with concessions

Main Takeaways

- V3 VAEP model outperforms the V2 one underscoring the role of enhanced data quality
- SHAP values in conjunction with the VAEP model can generate actionable insights into soccer gameplay
- Game context (time and score) is an important determinant of the scoring probability of an action
- Quick counterattacks seem to be particularly effective in ACC women's soccer, with evidence appearing in each experiment

Limitations

- Dataset sample size and overall scope
- Lack of generalizability - findings only directly applicable to ACC women's soccer
- Not a causal analysis

Future Work

- Using the same approach on new datasets (reproducible code)
- Hyperparameter tuning of the VAEP models
- Using VAEP to simulate game sequences

Overall, this thesis contributes to the growing body of soccer analytics research by offering a strong mechanism for evaluating player performance and shaping game strategies through the VAEP model.

References

- Tom Decroos, Jan Van Haaren, Lotte Branson, and Jesse Davis. 2018. Actions Speak Louder than Goals: Valuing Player Actions in Soccer.

- SoccerAction. 2020.

- Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *University of Washington*

- Scott Lundberg. 2018. An introduction to explainable AI with Shapley values.

- Emily K Marsh. 2023. Calculating XGBoost Feature Importance.

Thank You!