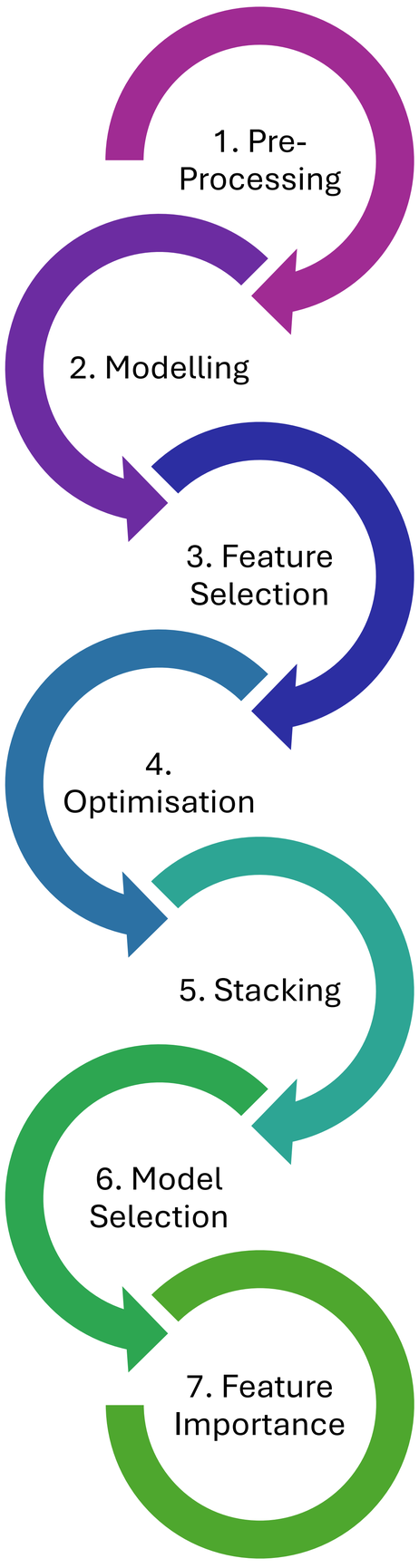
What gives a country better international football tournament-winning odds?

Sam Ryder, 18317496

ACM40960 Projects in Maths Modelling, MSc Data and Computational Science

Background: Winning international football tournaments is crucial for nations as it increases national pride, global prestige and national economics. Understanding the key factors that contribute to success can help nations boost their odds of future tournament success.



1. Pre-Processing and Data Exploration: Data was sourced as multiple datasets. Desired data from each dataset were isolated and grouped. Additional features were engineered. This data was explored using correlation and clustering. The target Variable was most highly correlated with: **FIFA Ranking, Market Value, Win Percentage, Goals Scored.** Clustering methods identified: **FIFA Ranking, Titles, Market Value, Win Percentage** and **Tournament Odds** to be the most discriminatory.

2. Modelling: The results were stored in a data frame. Models varied in complexity, structure, and the assumptions they made about the data. In order to find the best performing model a variety of models were run using train and test data. The following regression models were run: Linear, Ridge, Lasso, Trees, Random Forests, Support Vector Machines, Neural Networks, Generalised Linear Models, ZIP, Generalised Additive Models.

3. Feature Selection: Forward and Backward feature selection was performed to find the best subset of predictive features. Features were removed or added based on the largest reduction of AIC (Alkaline Information Criterion). The disparity here is large. Models were run on these and the results were stored.

Forward Selection	Backward Selection
FIFA Rank	FIFA Rank
Caps	Manager Age
Titles (in European Champs)	Titles (in European Champs)
Hierarchical cluster	Months as manager
% of clean sheets qualifying	Player average age
Expected goals (Q_xGF)	Player average height

Important features by selection method

4. Model optimisation: The same models previously run, were re-run with the Grid Search Cross Validation optimisation of some select hyperparameters. This allows for a restrained parameter space and efficient optimisation. The optimum hyperparameters were identified using the training set

5. Model Stacking: The optimised models were stacked, to combine their predictive power and generate a new model. XGBoost was chosen as the meta-regressor because it is generally considered to be high performing, robust to overfitting and able to handle complex relationships. This model was then tested and run on the testing data.

6. Model Selection: The best performing model from steps 2-5 (left) was identified as the base ZIP model, with its base hyperparameters.
The results, shown in the table to the right, show a **strong performance.**

Mean Squared Error	0.003642
R-Squared	0.923917

Zip Model Results

7. Feature Importance: Feature importance was performed using SHAP on the best model (ZIP model). SHAP (SHapley Additive exPlanations) provides a ranking of feature importance in model outcomes and highlights the traits that make these features important.

Main finding

Figure 1 shows the features (and traits) identified as important.

A team wanting to have good odds of winning should try to obtain:

- Market Value (High)
- Expected Goals in Qualifying (Q_xGF) (High)
- FIFA ranking (Low/Top Ranking)
- Qualifying Clean Sheets % (Low)
- Caps (Low)

Conclusion

In conclusion, machine learning has enabled a ranking of feature importance in model outcomes and highlights the traits that make these features important.

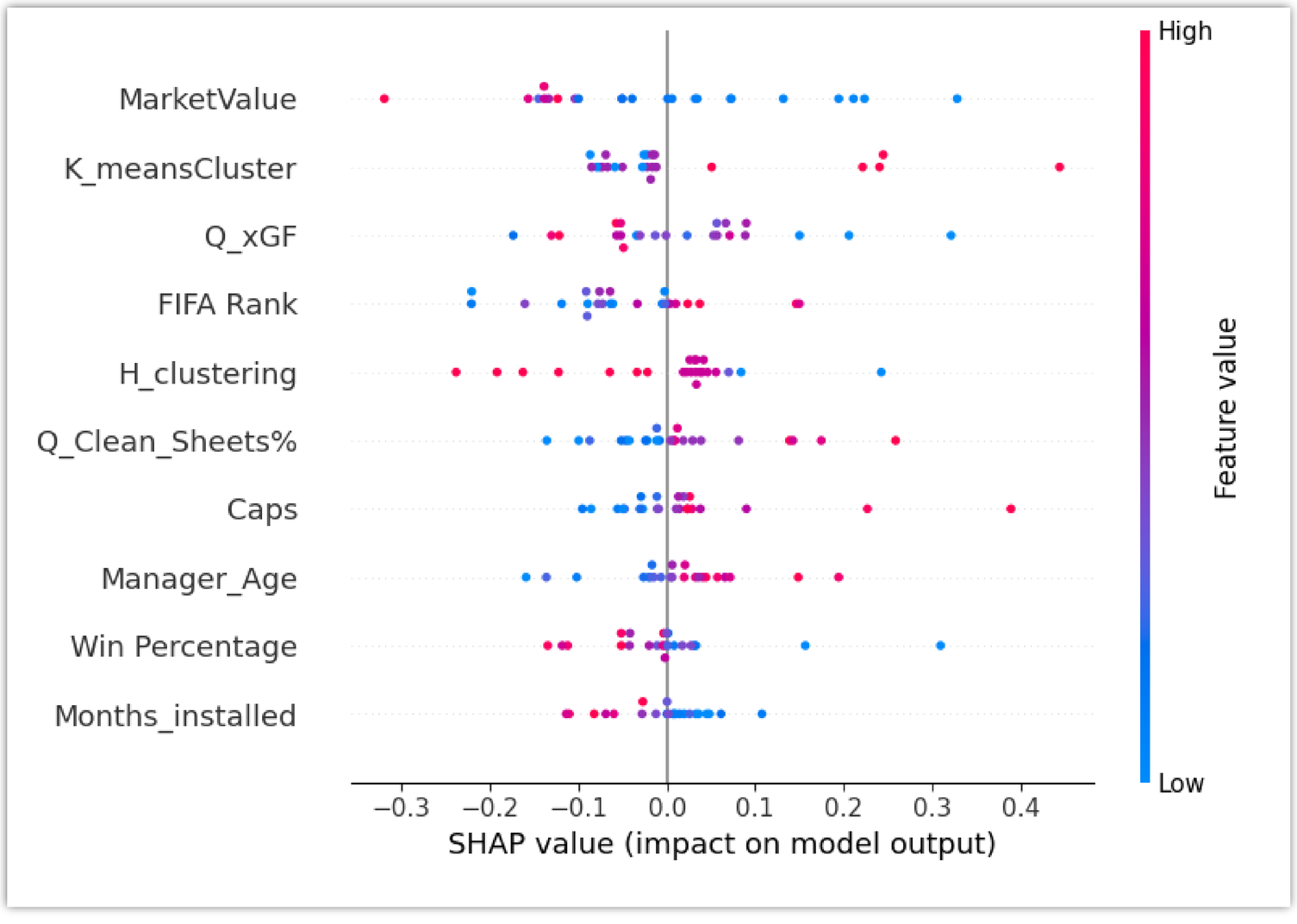


Figure 1: Visualisation of Feature Importance and Value