Author: Alexandru Mihalache
BrainStation Data Science Capstone November 2022

Capstone Project Report
Football Transfer Performance

1. Project Problem Statement

Football transfers are big business with billions of £ being transferred each year in player moves. Football player prices have increased dramatically in the past decade, therefore the risks associated with transfers are massive. You need to get the transfer right.

However, we often see players fail at their new clubs, despite being leading players at their old club, and despite the extensive scouting and data analysis being done before the transfer.

So might there be another way to assess whether a player would continue to put out the same type of output in a new club – which is what this project aimed to answer.

2. Background on the subject matter area

Transfers usually involve video analysis, tactical profiling, personality assessments of a player, and data analysis. Despite all this effort we often see players not do well in a new club (e.g., at Tottenham Hotspur, in the men's team Tanguy Ndombele and Gio Lo Celso, are seen as failed transfers as the players could not replicate their incredible output from their previous teams, those 2 players represent a reported £90m worth of transfers).

In this project I aim at placing the player (or a representation of the player, through their data) into the new team (the models I am building) and assessing the output against the true outcomes of a transfer.

The field of football analytics and machine learning is growing at an increasing pace, as more data scientist join the area. But also, as further investment is poured in by team, agents, media and gambling organisations.

3. Details on dataset

- 3 seasons of the FA Women's Super League
- Includes every match
- Includes every action in each match
- Includes supporting data such as players, scores
- Data sourced from StatsBomb through the Soccer Action library

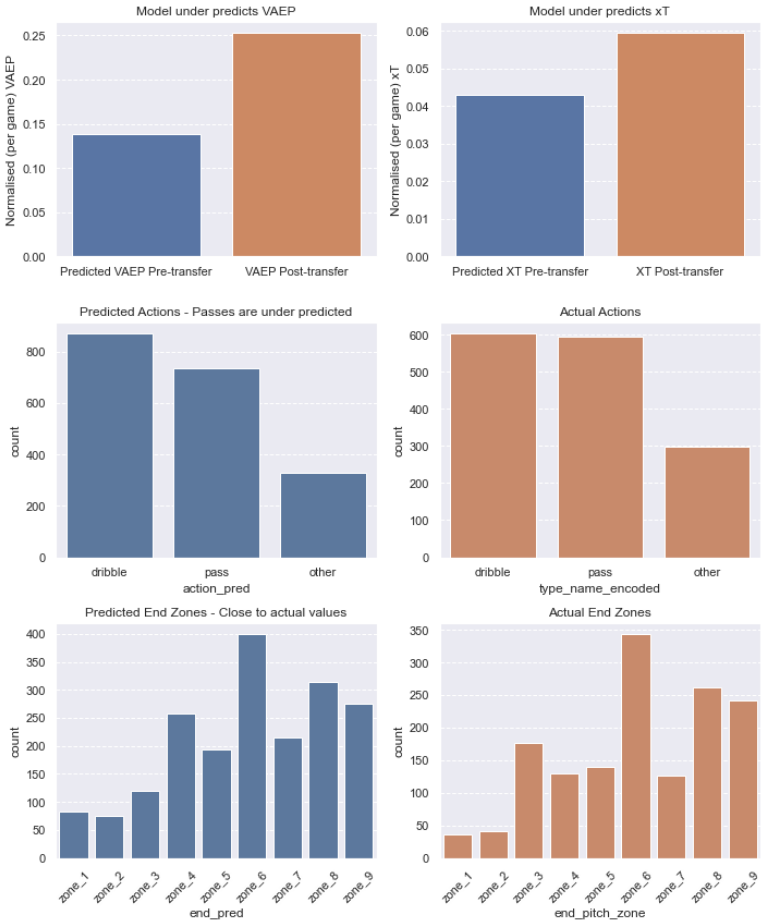
4. Summary of cleaning and pre-processing

- Using the Soccer Action library, and the SPADL (Soccer Player Action Description Language) Schema – I am getting formatted and pre-cleaned data to work with

Author: Alexandru Mihalache
BrainStation Data Science Capstone November 2022

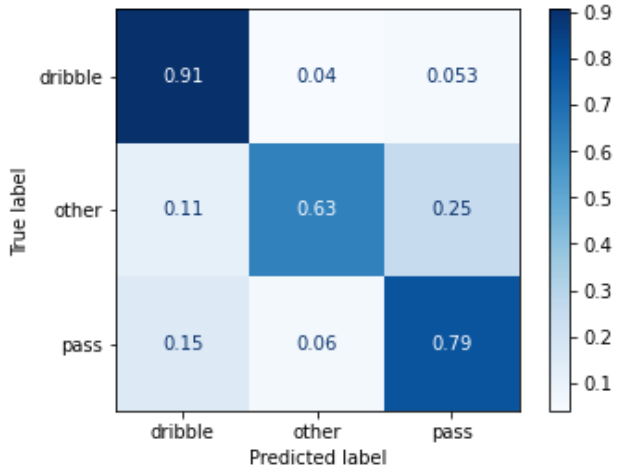
- As part of the pre-processing approach, I have created new features to describe each of the previous 5 actions
- The reason for this was to remove the sequencing of the dataset, add in the history of how each action came to be into the action itself.
- Through this I hope to capture player and team play style

5. Insights, modelling, and results



Model prediction performance – overall only the end zone model performed well, the other 3 models require further tuning and different training approaches outlined in the conclusion.

Next action proved to be a problematic model, and in numerous iterations showed a large risk of false positives in its classification.



Model Performance - Next action confusion matrix

As an example of how the models perform, we have the confusion matrix of the Next Action model – here we can see that across different classes the model performs with different levels of accuracy. This was also seen in the end zone model. This is an indication that further work needs to be done to even out the model's ability to predict without class bias.

Author: Alexandru Mihalache
BrainStation Data Science Capstone November 2022

6. Modelling approach

- Started with 2 approaches that I then tested and made a final selection from
- One approach was to fit a model to a player the other was to fit a model to a team
- I have 4 target variables, 2 classification targets and 2 regression targets
- After baselining both approaches across all 4 target variables, the team approach was more consistent so used this for my final modelling.
- The final models, were then tuned through a cross validation grid search
- All 4 final models are xGboost models, for classification and regression, depending on the target variable

7. Findings and conclusions

As a high-level conclusion, I think we have enough evidence to say that we can represent a team play style as a model, and that we can feed in a new player's data to get a view of performance.

The approach of describing each action through a history of its 5 previous moves seems to be useful in prediction, although based on model feature importance, we don't need to go more than 2 moves back.

However, we need to now narrow down into specific areas of the analysis, such as certain moves, or certain areas of the pitch, to build more target models and see if we can improve predictive performance – as this first iteration looked at a broad application of the models.

Models performed moderately well; however, they will need further tuning and iterations. Of the 4 models, end zone performed the best.

At this stage I don't believe the models are ready to be used in transfer decision making. However, they have opened new areas of problem discovery, such as looking in depth at crosses and freekicks.

Recommendation for next iteration:
- balance all classification input data
- binarize zone data, to focus on a specific zone of group of zones vs the rest
- run further hyper parameter optimisation
- optimise each model for each team, rather than a general optimisation - this is currently a big limitation of the current models, which is that they were optimised for a specific team but then they are applied to another, which may not give the best outcome