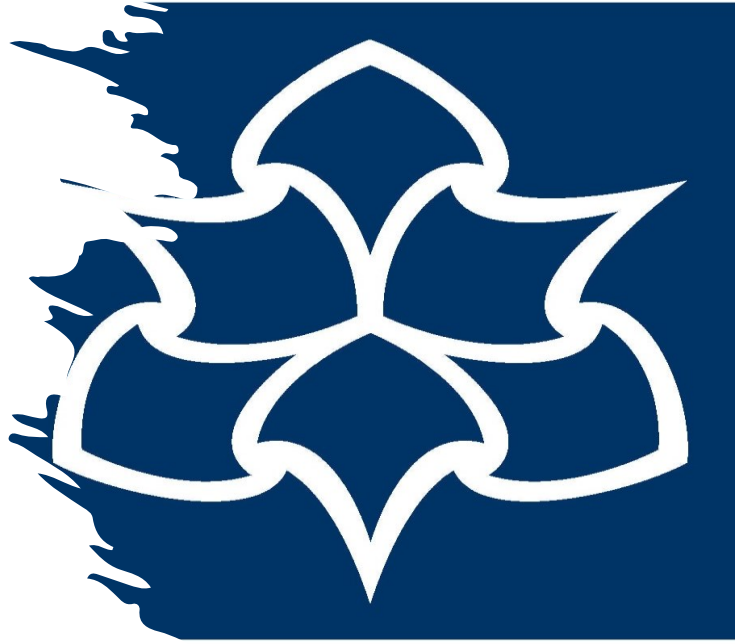
IDENTIFYING UNDERVALUED FOOTBALL PLAYERS IN TRANSFER MARKETS THROUGH NLP-BASED NEWS ANALYSIS



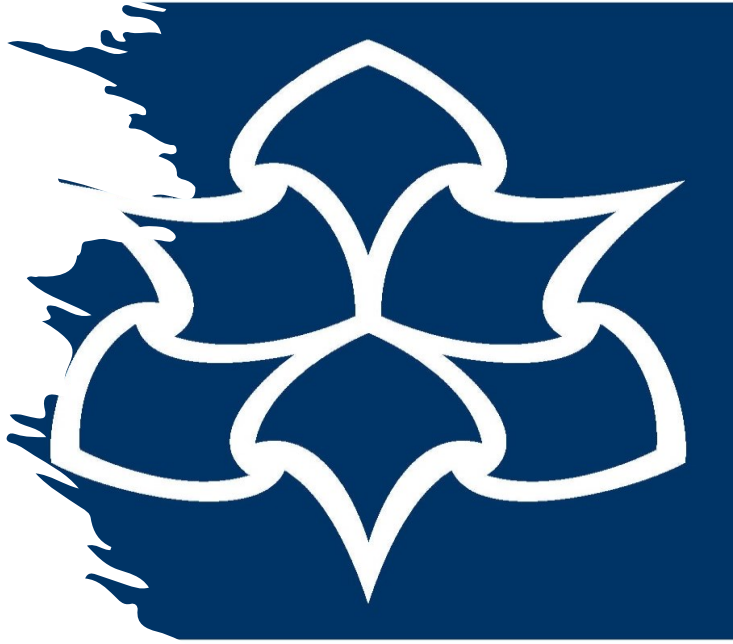
by

CHIDIOGO MAUREEN MADUKA

Introduction

Natural Language Processing and Machine Learning approaches have become powerful tools leveraged by organizations, institutions, government or policy makers in making decisions.

The positives likewise the negatives are yet to be fully unravelled.



Benefits of NLP and ML application in sport

- Injury detection and prevention
- Performance analysis
- Talent scouting
- Effective Financial decision
- Marketing and Ads
- Fans engagement
- Many more…

Dataset

Provided by Typewind LTD

- Collection of extracted English football news
- Player details
- Player transfer history
- Player market value

The common feature of merge was the date column.

Players = “Saka Bukayo", "Rashford Marcus", "Erling Haaland", "Henry Kane", "Kevin De Bruyne", "Foden Phil", "Victor Osimhen", "Antony Santos", "kylian Mbappé", "Gabriel Jesus", "Vinicius Junior", "Jordan Pickford"

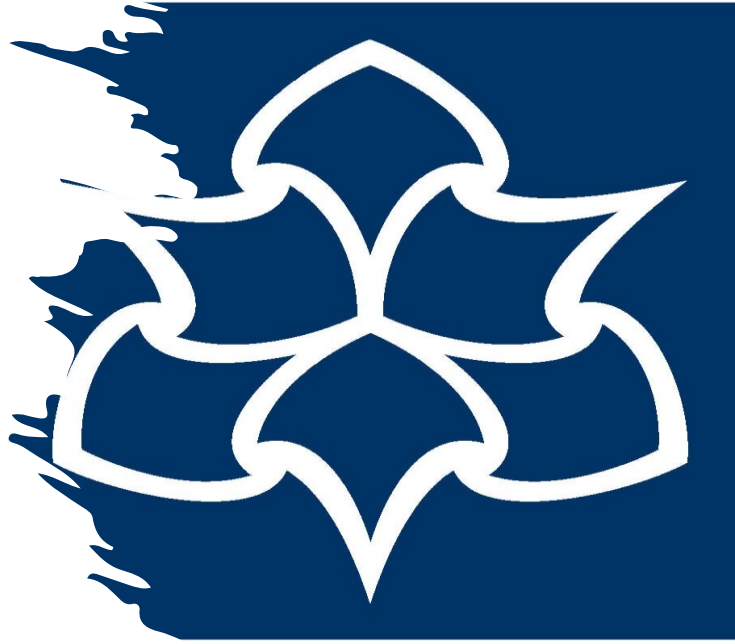
The names above make approximately 10% subset of the entire news data that was used for the NLP and ML analysis

NLP techniques implemented

- Coreference resolution
- NER
- Sentiment analysis
- TD-IDF

ML algorithms

- Random Forest Regressor
- Gradient Boosting regressor
- k Nearest Neighbour regressor
- Adaptive Boosting Regressor
- Support Vector Regressor



Feature Engineering

- PCA
- Polynomial Feature
- Encoding

Evaluation Metrics

- Mean Absolute Error
- Mean Square Error
- Root Mean Square Error
- R_squared

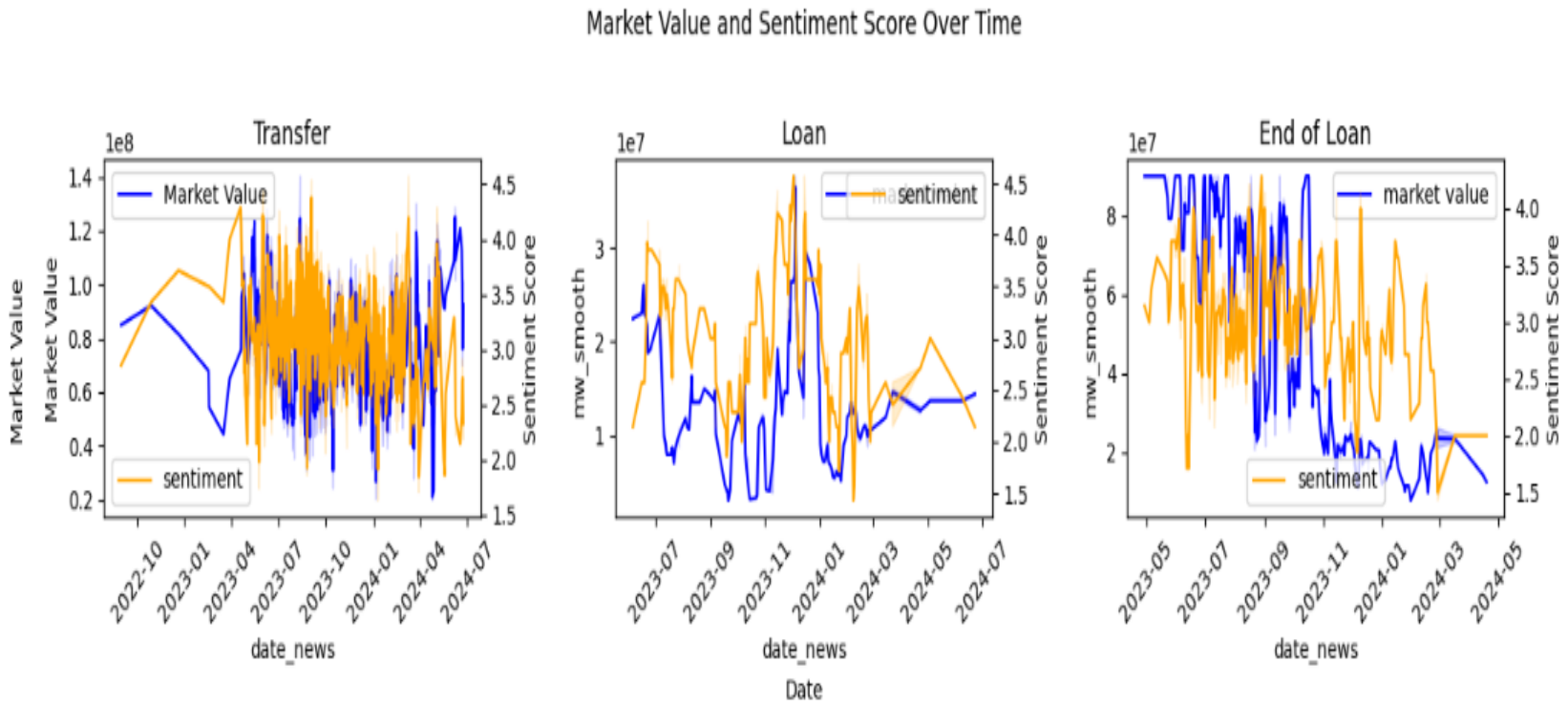
Tools for Implementation

- Spacy pretained model
- Hugging face pretrained transformer model

Player specific sentiment polarity assignment

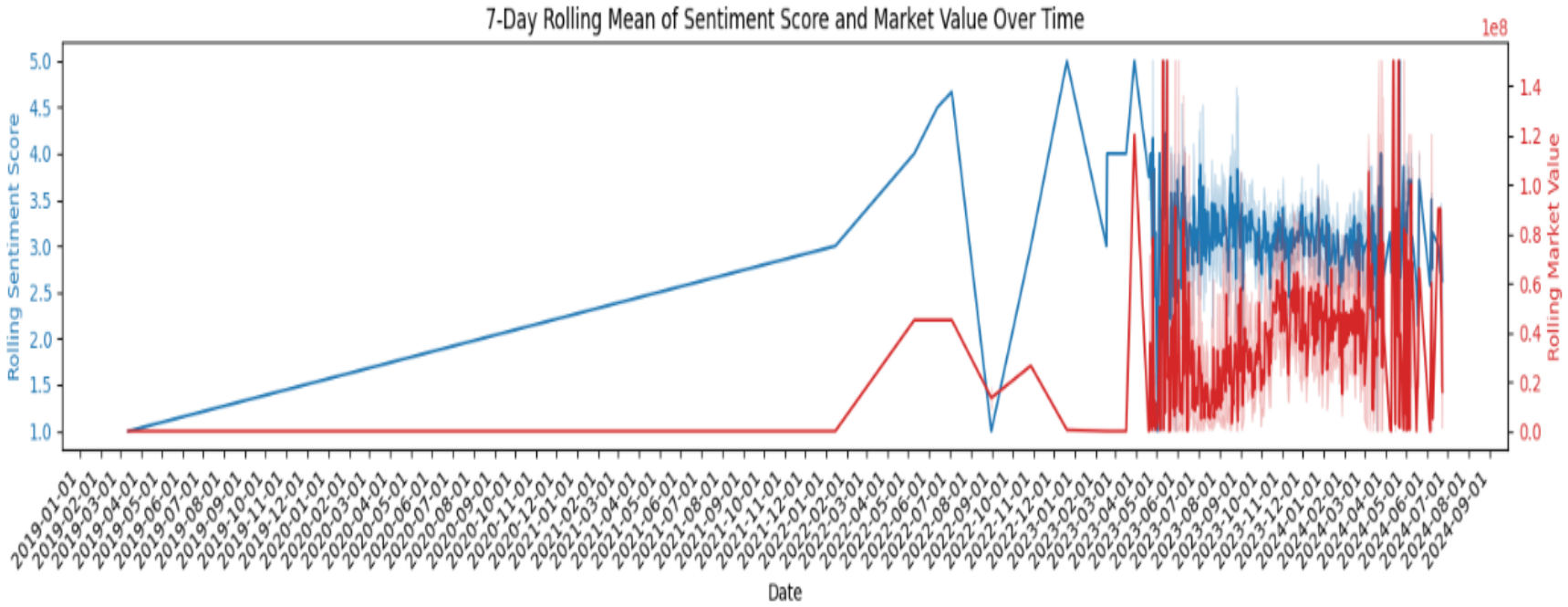
```
                                {'Rashford': 2}
{'Kalvin Phillips': 4, 'Jack Grealish': 4, 'Na...
{'Billy Gilmour': 3, 'Dominic Solanke': 2, 'Mi...
{'Pep Guardiol': 4, 'Eddie Howe': 1, 'Tino Liv...
 {'Simon Jordan': 3, 'Unai Emery': 3, 'Simon': 3}
                      ...
{'Julie Bradbury': 4, 'Iga Swiatek': 4, 'Erlin...
{'Declan Rice': 2, 'Mikel Arteta': 2, 'Gabriel...
{'Mason Greenwood': 4, 'Jadon Sancho': 4, 'Eri...
{'Benjamin Sesko': 3, 'Erling Haalan': 2, 'Se'...
{'Gary Neville': 3, 'Harry Lange': 2, 'Declan ...
```

Market value and sentiment trend



Trend

A patterns appears around the opening of transfer window and after its close



Evaluation and analysis

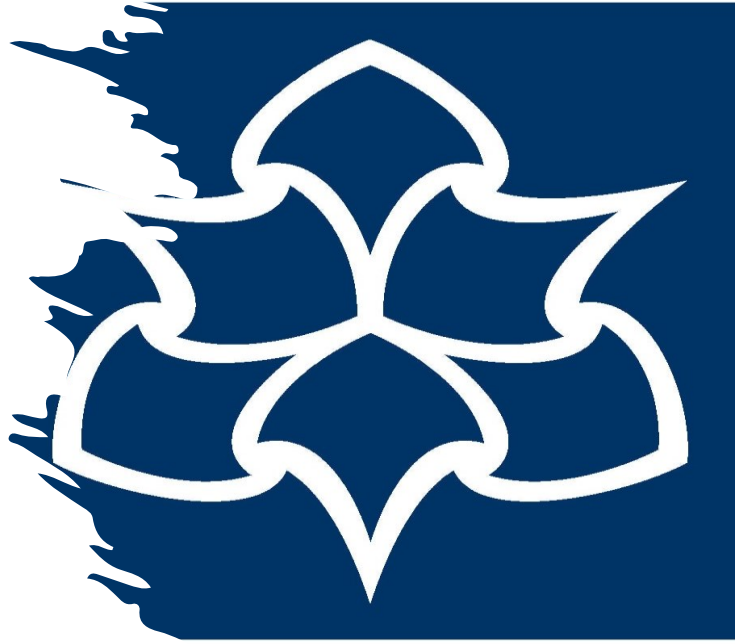
Without TD-IDF with polynomial features and PCA					
Default	**RF**	**GB**	**KNN**	**AdaBoost**	**SVR**
R^2 Score	0.995	0.980	0.933	0.807	-0.096
MAE	0.024	0.128	0.103	0.508	1.282
MSE	0.097	0.035	0.116	0.577	1.890
RMSE	0.009	0.188	0.341	0.333	1.375
Hyperparameter Tunning					
R^2 Score	0.905	0.838	0.926	0.688	0.893
MAE	0.219	0.427	0.132	0.564	0.261
MSE	0.165	0.279	0.128	0.538	0.185
RMSE	0.406	0.528	0.358	0.733	0.430
TD-IDF, with polynomial features and PCA					
Default					
R^2 Score	0.992	0.980	0.577	0.752	-0.538
MAE	0.020	0.113	0.425	0.543	1.427
MSE	0.012	0.031	0.641	0.377	2.334
RMSE	0.109	0.176	0.801	0.614	1.528
Hyperparameter Tunning					
R^2 Score	0.269	-	0.372	-	-
MAE	0.742	-	0.600	-	-
MSE	1.109	-	0.953	-	-
RMSE	1.053	-	0.976	-	-

Table 5.1: Compilation of results

Predictions

From the table below, most of the players are undervalued by the model which would mean the reverse from a market perspective that they are overvalued except for Kalvin Philips with a 7.5% undervaluation in the transfer market.

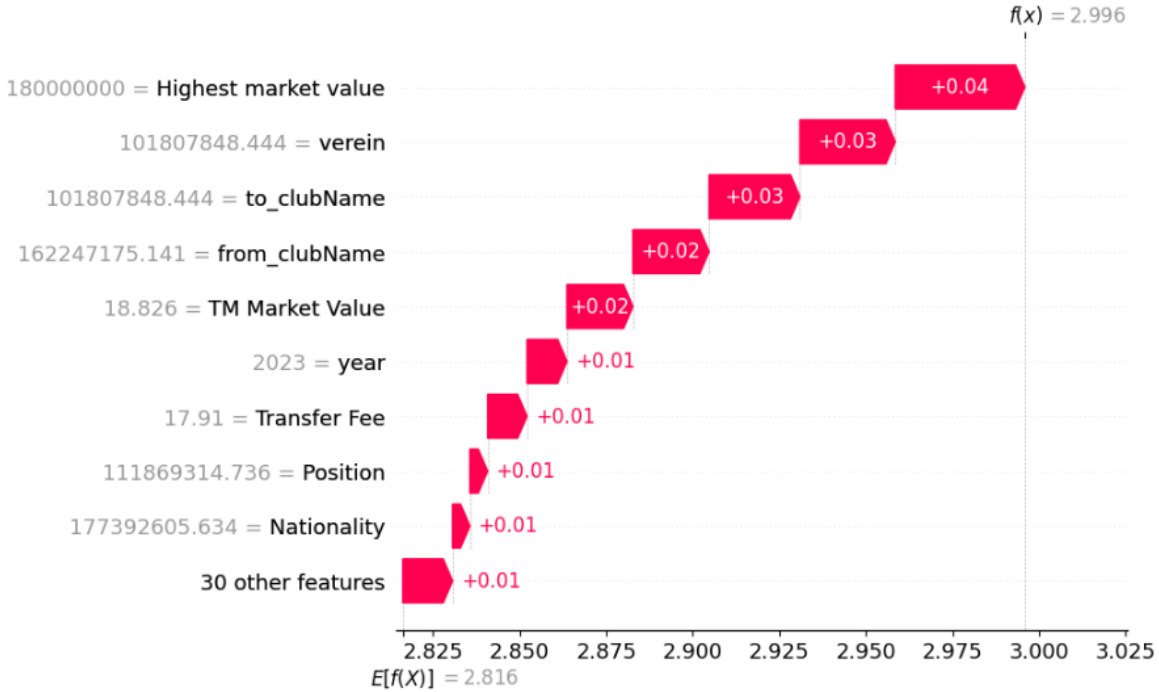
	player_name	predicted_market_value	actual_market_value	undervaluation
2869	Erling Haaland	18.984266	19.008467	-0.024202
16786	Marcus Rashford	18.126130	18.132999	-0.006869
28514	Kalvin Phillips	17.223005	17.147715	0.075289
9631	Rodri	18.315287	18.420681	-0.105394
1469	Bruno Fernandes	17.731449	18.132999	-0.401549



Limitations

- Insufficient timeline aligned dataset
- Possible bias towards more mentioned players in the media
- Trained and tested on only a particular demography. E.g. Male English league football players
- Computational cost led to implementing NLP techniques on Title rather than content

SHAP Additive Explanation



Conclusion

In summary, this research focused on the regression task of predicting market values of players and in comparing the training of 5 ML models, consequently identified undervalued football players.

Random Forest topped the chart with the highest accuracy rate of above 90% with the least error values in all scenario.