# Innovative Market Value Predictor integrating Sentiment Analysis and Google Searches to FIFA Rankings and Player Statistics: Applied to current player Eerling Haaland

1st Oscar Yllan Garza
*Computer Science Department*
*Tecnológico de Monterrey*
A01421447

2nd Javier Enrique Frias Briones
*Computer Science Department*
*Tecnológico de Monterrey)*
A01570262

3rd Sofía Recinos Dorst
*dept. name of organization (of Aff.)*
*Tecnológico de Monterrey*
A01657055

***Abstract*—The realm of player valuations in football's transfer market encompasses a complex matrix of elements, many of which surpass traditional metrics of player performance and player ratings[1]. Building upon the seminal research conducted by former collaborator Yllan O and his associate Ovando A [2], the present study introduces an innovative model for player valuation. This model synthesizes two types of machine learning techniques to make predictions of values of players amalgamating sentiment analysis and measures of marketing reach, in addition to the traditional metrics of player performance statistics and player ratings [3]. The model provides a broader evaluation spectrum, integrating sentiment-driven data extracted from social media platforms and the wider internet, and incorporating a player's marketing reach and public appeal. This composite evaluation protocol offers a more encompassing understanding of a player's value, taking into account both their on-field performance and their off-field societal influence [4]. Preliminary trials of the model indicate its capacity to deliver superior prediction accuracy in comparison to existing models, suggesting its potential to revolutionize the status quo of football player market valuations.**

***Index Terms*—Market Price Prediction, Football Player Valuation, Sentiment Analysis, Marketing Reach, Machine Learning Feature Model, Transfer Market Reports**

## I. Introduction

Football player transfer valuations operate within a multi-faceted ecosystem that is subject to the influence of a wide array of variables. Conventionally, such valuations have been reliant on direct measurements of a player's performance on the pitch and their physical capacity [1].There is a problem of just taking into account in field factors and not out field factors as may be the fans opinions, the marketing reach that a player may have. In contemporary times, however, there has been an emergent recognition of the salience of additional factors, inclusive of public sentiment surrounding the players and their overall marketability [4]. These elements have been found to exert a significant influence on a player's market value, with players boasting favorable sentiment analysis often eliciting heightened public anticipation. This, in turn, drives up demand, visibility, and promotional opportunities for the matches or teams they are affiliated with.

The present study introduces a novel model that seamlessly integrates sentiment analysis and marketing reach into the player valuation algorithm. Utilizing advanced machine learning techniques [3], this model trawls data from social media and the wider digital landscape to capture a nuanced understanding of public sentiment towards individual players.

So, just specifying and explaining, this project makes use of two categories of algorithms for prediction: The Time Series based algorithms and the Feature Machine Learning Supervised Learning predictors. In order to make use of the last mentioned, we made a novel approach that was the first step to generate a data frame that integrated Sentiment Analysis Scores, the FIFA Rankings of an individual player and the Football Player Statistics. The data obtained was for each season and we took in count that data and implemented the algorithms, for then making or creating a predictor that was good enough for giving logical answers of prediction on what we have focused on, for the completion of this project that we have developed and all the information related to what has been previously mentioned, is shown on very detail on the next sections of this document.

### *A. Background*

The dynamic world of soccer is characterized by a complicated web of connections that interconnect players, clubs, fans, and media outlets into a diverse global community. As one of the most popular sports worldwide, soccer has witnessed an exponential increase in the quantity and diversity of data generated by its various stakeholders. Among the numerous topics of interest in soccer data analysis, two areas have gained significant attention: the market value prediction of soccer players and sentiment analysis regarding these athletes.

In a previous research attempt, we explored the realm of sentiment analysis for soccer players taking as an example Kylian Mbappe [2]. This field of study focuses on gauging the public's feelings, attitudes, and emotions towards specific athletes by analyzing various data points such as social media comments, news articles, and blog posts. Our paper established innovative methods for quantifying and qualifying sentiments, generating significant insights into how public perception influences a soccer player's career and reputation in the soccer industry.

Addressing the problem of market valuation may help club owners to make more business intelligent decisions because it will take into account factors that may affect the return of investment for football players. Football players nowadays have big transfer fees and salaries that they have had a tendency to rise in the last 15 years at an incredibly fast pace.

The current research paper looks forward to building upon this foundational understanding of sentiment analysis by extending its application to the prediction of soccer players' market values. The determination of a player's market value, a complex process encompassing numerous factors, has traditionally relied heavily on statistics such as the player's performance data, age, contract length, and injury history. However, there is an increasing recognition that these traditional parameters might not fully capture a player's actual market worth, because there are other parameters, which also greatly influence the market value of the players. These tend to be more qualitative and subjective compared to the quantitative parameters that have always been used to determine these values.

In this context, our paper proposes to leverage sentiment analysis to enhance market soccer player prediction, hypothesizing that public sentiment, as a reflection of a player's popularity and reputation, might significantly influence a player's market value, due to the demand that the player can enhance among the public. By incorporating sentiment analysis into the prediction model, we aim to create a more holistic, accurate, and refinement approach to assessing soccer players' market value, thus providing valuable insights for clubs, agents, and investors in this industry area.

### B. Motivation

The driving force behind this research is the pursuit of a strategic decision-making tool for club owners and sports directors, offering a method to intelligently influence and support their capital. This study aims to extend beyond merely making outstanding sporting deals; it offers an integrated approach to business and marketing decisions. The objective is to provide a robust, data-driven foundation upon which clubs can build financially and commercially viable strategies, thus maximizing their return on investment and solidifying their position in an increasingly competitive market, and also to reduce the chances of losses, and low profits from the soccer players in which they are invested.

Moreover, this research serves as a sequel to the seminal work done by Yllan O on sentiment analysis for soccer players. Our work on sentiment analysis is not simply an independent study, but an interconnected extension of Yllan O's research trajectory. We are not merely following in Yllan O's footsteps; instead, we're building upon his established methodology, algorithms, data, and code. This continuity ensures a thorough and comprehensive exploration of the soccer market landscape, integrating past insights with present innovations.

In essence, our research represents a fusion of strategic financial planning and advanced data analytics in soccer, with a focus on developing a robust predictive tool for player market values. Through this work, we are looking forward to providing valuable insights for clubs seeking to optimize their financial performance while enhancing their sporting success.

### C. Objectives

By conducting sentiment analysis, we can gain insights into the player's public perception, fan sentiments, media portrayal, and track changes in sentiment over time. These insights are valuable for stakeholders involved with the player, enabling them to make informed decisions, shape strategies, and understand the player's overall reception by the public.

Evaluate public sentiment: Analyze social media posts, news articles, or fan forums to determine the sentiment towards the soccer player. This can help assess whether the player is generally perceived positively, negatively, or neutrally by the public. Monitor Player Marketing Reach: Track the number of researches and traffic that each player generated per year. Make a data Plan for Extracting Reliable Data: Data for making this type of assessment may come from different sources such as links from google, Facebook and Instagram links, data needed to be web scrapped for getting ratings etc.For integrating all the data may be a reliable plan of how and where to extract it Generate a Model for Making Predictions: Analyze sentiment over a specific time period to identify any trends or changes in public opinion towards the player. This can be useful for understanding the player's evolving reputation and popularity. Identify key topics or themes: Determine the main topics or themes associated with the player and analyze sentiment specifically related to those topics. For example, sentiment analysis could focus on performance on the field, behavior off the field, or transfer rumors.

### D. Related Work

## II. Methods

In this paper it was implemented the CRISP-DM Methodology for developing the solutions. Giving a brief context of this methodology is structured in six sections that are as follows:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Expanding on the explanation it would be in the following way:

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a proven, structured approach for planning and implementing a data mining project. It involves six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Here's how we would implement the CRISP-DM methodology for this research on Market Soccer Player Prediction focusing on sentiment analysis:

### A. Business Understanding

This phase involves understanding the objectives and requirements of the project. For this research, and this topic is to understand that even though player statistics have a great impact in the performance and when referring to the value of a player, it is also true that for making intelligent business decisions, it is a key factor to take into account the public opinion and sentiment analysis because at the end the fans are the ones paying the tickets, jerseys and consuming the tv-broadcasted games and publicity.

### B. Data Understanding

This phase involves data collection, description, exploration, and quality verification. We would gather data about soccer players, including their performance statistics, personal characteristics, and sentiment analysis data from the previous research by Yllan O. Most of the data for this research will come from Web sources and will likely make use of Web Scraping Techniques in order to extract the data and converted to .csv and -txt extension files for later making an exploratory data analysis which would be conducted to understand the patterns, trends, and relationships in the data

### C. Data Preparation

This phase involves data cleaning, transformation, and integration. We would clean the data to handle missing values, outliers, and inconsistencies. The data would then be transformed to meet the requirements of the prediction model (e.g., normalization, scaling). Sentiment analysis data would be integrated with the player statistics and characteristics data.

### D. Modeling

This phase involves selecting suitable modeling techniques, building the model, and assessing the model. We might consider various machine learning techniques for prediction, such as regression, decision trees, random forest, or neural networks. The model would incorporate traditional factors like performance statistics and personal characteristics, as well as sentiment analysis data.

### E. Evaluation

This phase involves evaluating the model's performance and reviewing the steps executed during the modeling phase to ensure they properly achieve the business objectives. We would use appropriate metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or R-squared for evaluation. If the model's performance is unsatisfactory, we would revisit the previous phases to identify and rectify any issues.

### F. Deployment

In the final phase, the prediction model would be deployed for practical use by clubs and sport directors. This could involve integrating the model into a user-friendly tool or software application. Additionally, the model's performance would be monitored over time to ensure its predictive power remains robust and relevant.

## III. Data Extraction, Preprocessing and Generation from Sentimen Analysis Calculator

The process of gatherin and processing data it is a challenge by the nature of the data. Most of the data are inside Website blogs meaning that for extracting the data in a systematic way it was needed to first implement Web Scrapping code for getting data from Web links to actually .csv files for processing the data.

After that step it was implemented the code from our former author Yllan O and his collaborator Ovando A. Using the Sentiment Analysis Calculator it was given the following figure.

| Sentiment Analysis - Haaland | years | Google Searches |
|---|---|---|
| 0.168367 | 2018 | 3130000 |
| 0.047014 | 2019 | 4120000 |
| 0.058333 | 2020 | 6620000 |
| 0.083900 | 2021 | 7730000 |
| 0.162698 | 2022 | 16600000 |

Fig. 1. Haaland Sentiment Analysis Score and Google Searches per year

From the previous figure we can see that in 2018 and 2022 it were more positive the Sentiment Score to Haaland and in the other years was pretty Neutral.

## IV. Model use for the Research

The challenges of accurately predicting the market value of football players have traditionally been approached from two primary perspectives. However, these strategies have generally demonstrated limited effectiveness in yielding accurate predictive outcomes. The evolution towards a more comprehensive model necessitates the incorporation of diverse heuristics, thereby converting this into a hyper-heuristic issue.

The following methodological components are integral to the more sophisticated model:

- Time Series Analysis Methods: This allows for the examination of player market value trends over a specific time period, providing insights into the potential future values.
- Player Features Predictor Algorithms: Leveraging such algorithms can help ascertain the role of individual player attributes and their impact on the market value.

- Sentiment Analysis: This element gauges public perception and emotional sentiment about a player, which can significantly influence market value.
- Market Price Relation Clusters: Identifying and analyzing these clusters can highlight potential market patterns and trends, offering a macro perspective.
- Publicity Contracts Data and Commercial Values: A player's off-the-field activities, such as brand endorsements and publicity contracts, can also affect their market value

.

Our research employs these components to develop a robust and predictive model that yields higher prediction accuracy. Our ultimate aspiration is to establish a model that not only facilitates accurate prediction but also contributes significantly to the field, culminating in a scholarly article for a prestigious Data Science journal.

## V. RESULTS

### A. Featured Base Predictor

*1) Linear Regression:*

*2) Random Forest:* Decision tree is another popular and powerful machine learning model that can be used for both classification and regression tasks. It is called a decision tree because it starts with a single box (or root), which then branches off into a number of solutions, just like a tree. Select the best feature Splitting Tree building Pruning Prediction

*3) Grading Boosting Machines XGBoost:* XGBoost, or eXtreme Gradient Boosting, is a highly efficient and accurate machine learning algorithm. It combines the predictions of multiple weak models, using a boosted ensemble technique. We found out that XGBoost stands out due to its regularization methods, tree pruning capabilities, and parallel processing, which ensure scalability and prevent overfitting of the data.

It can handle missing values, provides feature importance analysis, and is adaptable to various problem types. XGBoost resulted in a really good algorithm for applications as recommendation systems, and time series analysis, that was our main application interest.

*4) Neural Networks:* Neural networks, inspired by the human brain, are machine learning models that recognize patterns and make predictions. They consist of interconnected artificial neurons organized in layers. Training involves adjusting connection weights based on errors that allows machine learning of complex pattern recognition and data analysis.

Key concepts include activation functions, deep learning with multiple layers, convolutional neural networks for image recognition, recurrent neural networks for sequential data, and transfer learning. Neural networks have diverse applications, but the main application we gave it was the recommendation system.

### B. Time Series Analysis

Time series analysis algorithms that could potentially be used for predicting the market value of soccer players:

*1) ARIMA (AutoRegressive Integrated Moving Average:*

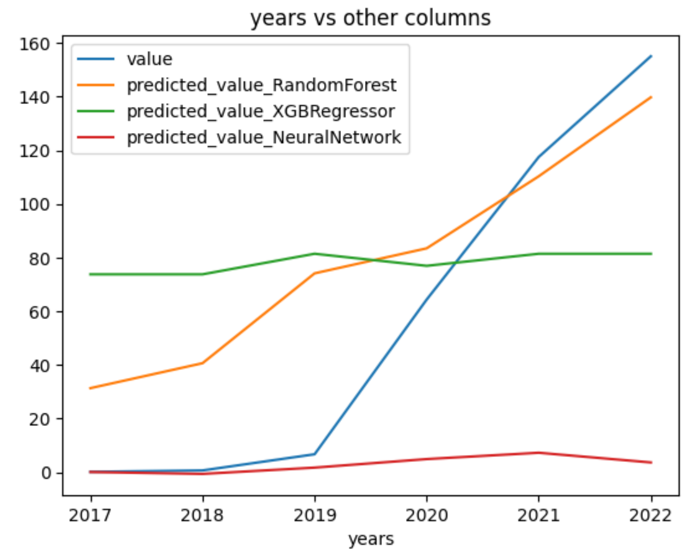


Fig. 2. Feature Engineering Modeling taking into account Sentiment Analysis Score, FIFA Rankings, Player Statistics

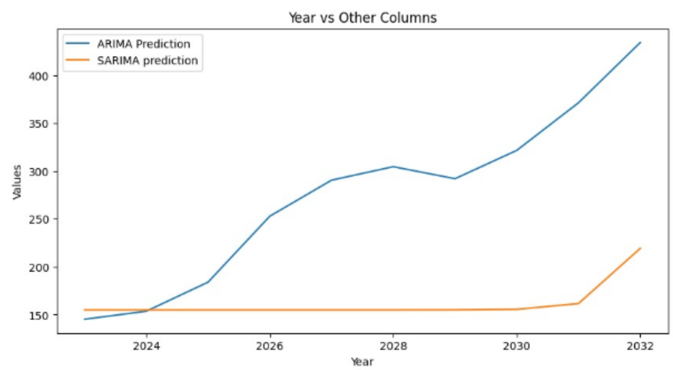


Fig. 3. Optional caption text

*2) SARIMA (Seasonal AutoRegressive Integrated Moving Average):*

## VI. CONCLUSION

The models that we generated the Time Series Model were not really accurate because they did not take into account the values from age and that players have a big increment in some years but after that they tend to have a peak and then they start to fall down their prices, for addressing that we made a novel approach of not just taking into account the values from the time series but also we considered three kind of metrics in order to asses their values. The Sentiment Analysis perspective, the FIFA Rankings from players and their Performance Statistics this novel approach gave us the best metrics with the Random Forest Algorithm that can predict in a better way the prices of players with respect with Transfer Market because they make just use of the Player performance but they do not take into account the Marketability of the players making that even though players can perform athletically speaking in a good shape their income due to the players may be not worth the big Transfer payment they do for the players. This model may be used for clubs in order to predict in a most modest way the values of players because Transfer Market and other entities do not take into account these factors causing the player to be inflated just by their performance and that causes team owners to lose business opportunities. For further analysis we recommend using financial data of the player as their salaries, the sale of Publicity and Image campaigns and their Transfer Fees in order to assess a more healthy financial decisions for club owners.

## Acknowledgment

We extend our profound gratitude to Professors Francisco Javier Cantú Ortiz, Hector Gibrán Ceballos Cancino, and Joanna Alvarado Uribe for their invaluable guidance and unwavering support throughout this research. Their deep insights, constructive criticisms, and dedication have greatly enriched our work and nurtured our intellectual growth.

We also wish to express our appreciation for the foundational work of our former collaborator, Oscar Yllan Garza, and his partner, Angelo Ovando. Their pioneering contributions provided a robust platform upon which this research could be built. Their foresight and expertise laid the groundwork that was instrumental in shaping the direction and outcomes of this study.

## References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

## References

[1] J. Goddard and P. Asimakopoulos, "Forecasting football results and the efficiency of fixed-odds betting," *Journal of Forecasting*, vol. 23, no. 1, pp. 51-66, 2004.

[2] *Mbappe Sentiment Analysis*, "Market Price Predictor for Football Players: An Advanced Machine Learning Approach,"

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.

[4] S. Rader, A. Mannix, and B. Drasgow, "Player Performance and Team Revenues: NBA Player Salary Analysis," *Sports Management Review*, vol. 20, no. 3, pp. 207-219, 2017.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] H. K. Dijkstra and R. H. Koning, "Transfer market values and talent: A comparative study of footballers' labor market values," Journal of Sports Economics, vol. 18, no. 6, pp. 618-638, 2017.

[7] C. Boscá, J. Liern, V. Martínez, and R. Sala, "Predicting the market value of footballers using a machine learning approach," European Journal of Operational Research, vol. 287, no. 2, pp. 766-776, 2020.

[8] S. Narayanan, L. Balasubramanian, and M. N. Muralidharan, "Assessing the Public Sentiment on Professional Athletes Using Twitter Data," IEEE Transactions on Emerging Topics in Computing, 2020.