Validation and Governance

1. Variable Level Monitoring
   1. Model Build Variable Level Statistics (e.g., mean, median, std or distribution of categories)

Model build variable level statistics involve the comprehensive analysis and characterization of individual input variables utilized in constructing the football player market value prediction model. This process entails calculating essential statistical measures such as mean, median, and standard deviation.

These statistics serve as a foundational reference, enabling the model's builders to gain insights into the central tendencies and variability within each input feature. By establishing a clear understanding of the variables' inherent characteristics, potential outliers, and data distributions during the model's construction, practitioners can make informed decisions about feature engineering, preprocessing techniques, and the model's overall sensitivity to changes in input data.

The ongoing monitoring of these statistics throughout the model's lifecycle ensures that any deviations from the established norms can be promptly identified and addressed, contributing to the model's reliability and performance stability over time.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| AGE | 2384 | 26.2 | 4.4 | 17 | 23 | 26 | 29 | 41 |
| MIN\_PLAYING | 2384 | 1328.4 | 948.4 | 1 | 475.25 | 1286 | 2108 | 3420 |
| DIST\_STANDARD | 2384 | 17.5 | 4.9 | 2.8 | 14.2 | 17.2 | 20.5 | 37.5 |
| DEF 3RD\_TOUCHES | 2384 | 233.5 | 273.0 | 0 | 35 | 127 | 341 | 1665 |
| ATT 3RD\_TOUCHES | 2384 | 206.7 | 204.5 | 0 | 43 | 147 | 314 | 1270 |
| ATT PEN\_TOUCHES | 2384 | 31.2 | 37.3 | 0 | 6 | 19 | 41 | 302 |
| ATT\_TAKE | 2384 | 27.0 | 31.4 | 0 | 5 | 16 | 38 | 306 |
| TOTDIST\_CARRIES | 2384 | 2648.4 | 2350.8 | 0 | 703.75 | 2149 | 3969.75 | 14430 |
| CPA\_CARRIES | 2384 | 6.3 | 10.9 | 0 | 0 | 2 | 8 | 140 |
| MIS\_CARRIES | 2384 | 21.7 | 21.3 | 0 | 5 | 16 | 31 | 139 |
| SUBS\_SUBS | 2384 | 6.4 | 5.4 | 0 | 2 | 5 | 9 | 29 |
| UNSUB\_SUBS | 2384 | 5.8 | 6.5 | 0 | 1 | 4 | 9 | 35 |
| PLUS\_PER\_\_MINUS\_\_TEAM.SUCCESS | 2384 | 0.0 | 12.7 | -46 | -6 | -1 | 5 | 58 |
| TOTDIST\_TOTAL | 2384 | 9168.7 | 9035.4 | 0 | 1951.8 | 6344 | 13902.8 | 54839 |
| CRSPA | 2384 | 3.0 | 4.8 | 0 | 0 | 1 | 4 | 37 |
| FLS | 2384 | 17.8 | 14.4 | 0 | 6 | 16 | 27 | 81 |
| FLD | 2384 | 16.8 | 16.2 | 0 | 4 | 13 | 24 | 122 |
| OFF | 2384 | 2.7 | 4.6 | 0 | 0 | 1 | 3 | 34 |
| CRS | 2384 | 25.6 | 39.9 | 0 | 2 | 9 | 33 | 393 |
| WON\_AERIAL | 2384 | 20.5 | 23.9 | 0 | 4 | 12 | 29 | 214 |
| LOST\_AERIAL | 2384 | 20.7 | 20.4 | 0 | 6 | 16 | 29 | 191 |
| ATT 3RD\_TACKLES | 2384 | 3.2 | 3.5 | 0 | 0 | 2 | 5 | 26 |
| LOST\_CHALLENGES | 2384 | 12.4 | 11.7 | 0 | 3 | 9 | 18 | 75 |
| PASS\_BLOCKS | 2384 | 11.9 | 10.2 | 0 | 3 | 10 | 18 | 52 |
| TKL+INT | 2384 | 37.9 | 33.8 | 0 | 9.75 | 29 | 59 | 195 |

* 1. Acceptable Ranges

The concept of establishing acceptable ranges for input variables is a pivotal component in ensuring the accuracy and robustness of the football player market value prediction model. To determine these ranges, the descriptive statistics derived from the initial model build, such as the minimum and maximum values extracted from the describe table, are leveraged as benchmarks.

By encapsulating the natural variability and boundaries present within the training data, these minimum and maximum values provide a pragmatic reference point for identifying outliers or unexpected fluctuations during ongoing monitoring. Adhering to these predefined ranges empowers the model to operate within the context of realistic and historically grounded input conditions, mitigating the potential impact of extreme values that might skew predictions.

As part of a vigilant validation and governance strategy, consistently upholding these acceptable ranges not only bolsters the model's predictive accuracy but also enhances its resilience to dynamic shifts in the input data landscape.

* 1. Missing Values

To address the occurrence of missing values within the input variables, a systematic approach of imputation will be implemented, specifically involving the utilization of the median value from the training dataset. Missing data can disrupt the model's predictive capabilities, and by imputing with the median, we ensure a robust and statistically sound replacement strategy.

The median, as a measure of central tendency, is less sensitive to outliers compared to the mean, making it a reliable choice for imputation. This approach maintains the integrity of the data distribution and aligns with the existing variability present in the training set. By consistently imputing missing values with the median, we ensure the model's ability to make accurate predictions while minimizing potential distortions caused by incomplete input information.

Regular monitoring will track the proportion of missing values and validate the effectiveness of the imputation strategy, further enhancing the model's overall reliability.

* 1. Variable Drift Monitoring Tolerance

Variable drift monitoring is a critical aspect of maintaining the stability and validity of the football player market value prediction model. It involves vigilant observation and assessment of changes in the distribution of input variables over time. As the football player market landscape evolves, the characteristics of player attributes might naturally shift due to factors such as changing player demographics or evolving market dynamics.

By establishing a predefined tolerance level for variable drift, the model can detect and signal when input variables exhibit significant deviations from their historical patterns. This proactive monitoring not only safeguards against potentially inaccurate predictions resulting from shifts in data distribution but also serves as an early warning mechanism for changes that might require model retraining or adjustment.

Continuous variable drift monitoring ensures that the model remains attuned to the dynamic nature of the market, enabling timely interventions to uphold its accuracy and effectiveness in reflecting real-world conditions.

1. Model Monitoring, Health & Stability
   1. Initial Model Fit Statistics

The assessment of initial model fit statistics, particularly Root Mean Squared Error (RMSE) and R-squared (R2), plays a pivotal role in evaluating the performance and predictive power of the football player market value prediction model.

RMSE quantifies the average magnitude of prediction errors, providing a comprehensive measure of how well the model's predictions align with actual market values. A lower RMSE indicates better predictive accuracy. On the other hand, R2 offers insights into the proportion of variance in the market values that the model can explain.

A higher R2 value signifies a stronger ability to capture underlying trends and patterns. Together, these statistics provide a comprehensive view of the model's precision and explanatory capability, guiding the initial assessment of its effectiveness.

Regular monitoring and comparison of these metrics against established benchmarks will ensure that the model's performance remains consistent and reliable as it adapts to changing player market dynamics.

The selected model in this project has an R2 of 0.64.

1. Risk Tiering (e.g., no action, report, refit, rebuild)

Risk tiering is a strategic framework that guides decision-making in response to the varying degrees of risk identified through continuous monitoring and evaluation of the football player market value prediction model. This approach involves categorizing potential risks into distinct tiers, each with predefined actions to address them effectively.

|  |  |  |
| --- | --- | --- |
|  | Actual Value | Minimum acceptable |
| RMSE | 0.83 | 0.87 |
| R2 | 0.64 | 0.6 |

• In the low-risk tier, no intervention will be required if the drift remains below 3%.

• Situated within the moderate-risk tier, drift falling between 3% and 7% can prompt the application of optimization techniques such as hyperparameter tuning, ensemble learning methodologies, and principal component analysis to fine-tune the model.

• Earning placement in the high-risk tier, drift ranging from 7% to 12% necessitates the refitting of the model with the updated dataset, given its potential indication of evolving data patterns over time.

• Crossing into the unacceptable risk tier, drift surpassing 12% will result in the model being reconstructed, as it signals a fundamental shift in the data's structure that compromises the model's predictive capabilities.