# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *Football players market value prediction* |
| **Requestor** | *Marcos Edu Munoz Quintasi* |
| **Date of Request** | *2023-07-17* |
| **Target Quarter for Delivery** | *August 2023* |
| **Epic Link(s)** | [*..\datasets\DATA\_DICTIONARY.xlsx*](../datasets/DATA_DICTIONARY.xlsx)  [*Presentation.pptx*](Presentation.pptx) |
| **Business Impact** | *High* |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

The football player market value prediction project presents a significant business opportunity for clubs, agents, and stakeholders in the football industry. The project seeks to find the most significant variables and introduce a systematic way of predicting the market value of football.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

* Rising importance of data-driven decision making: In recent years, the use of data analytics and predictive modeling in football has gained significant traction.
* Football clubs spent a record $1.57 billion (£1.27bn) on transfers in the January transfer window.
* A total of 4,387 international transfers were completed in men's football in January
* The number of transfers in men's football increased by 14.4% compared to the same period in 2022, with the total value of transfer fees rising by 49.4% compared to the amount spent in January 2022.

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

By accurately estimating a player's market value, informed decisions can be made regarding transfers, contract negotiations, and investments. This predictive analysis can help:

* Clubs identify undervalued players, negotiate favorable deals, and optimize their player recruitment strategies.
* Agents can leverage market value predictions to negotiate better contracts for their clients.
* Stakeholders can utilize this information to assess investment opportunities in players with high growth potential.

Overall, the ability to predict football player market values provides a competitive edge, enhances decision-making, and maximizes returns in the dynamic and lucrative football market.

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

* What are the most influential variables for predicting the market value of soccer players?
* How well can we predict the market value of football players using available data and predictive modeling techniques?
* How do the market value predictions for football players align with actual transfer fees in the football market?

Hypothesis:

* Players with better performance indicators have higher market value valuation
* Age plays a key role in predicting players' market value

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

The player’s club information (CLUB\_NAME, LEAGUE\_COUNTRY columns) will be the last team they joined/played in the 2022-2023 season (This only applies for the players who have played for more than one team in the season. The player’s statistics will be aggregated so the information about his performance will not be lost.)

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |

* Mean Squared Error (MSE): MSE measures the average squared difference between the predicted and actual values. Lower MSE indicates better model performance.
* R-squared (R²): R² represents the proportion of the variance in the dependent variable (market value) that is predictable from the independent variables (player attributes). Higher R² indicates better explanatory power of the model.
* Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and provides a measure of the average prediction error. Lower RMSE suggests better model accuracy.
* Mean Absolute Error (MAE): MAE calculates the average absolute difference between the predicted and actual values. Smaller MAE signifies better model performance.
* Adjusted R-squared: Adjusted R² accounts for the number of predictor variables in the model, providing a more robust measure of model fit. It penalizes the inclusion of unnecessary variables.
* Model complexity and interpretability: Consideration of the complexity and interpretability of the models is important. Simpler models that can still achieve comparable performance may be preferred for practical implementation and ease of understanding.

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** Linear regression, Decision trees models and Gradient Boosting.

The initial approach will be to use a linear regression model to determine which player’s variables (age, height, nationality..) are most significant related to predict his market value. We will also use other techniques to verify our findings.

**Methodology:** We will start by collecting historical data on football players, including their attributes, performance metrics, and market values. The dataset will cover a specific period, 2022-2023 season.

Next, we will preprocess the data by cleaning it, handling missing values, and normalizing numerical features. We will also conduct exploratory data analysis to gain insights into the distribution and relationships between variables.

To build the prediction model, we will employ a regression-based approach. We will use multiple linear regression as the initial model, with the player's attributes as predictor variables and the market value as the response variable. After building the initial model, we will assess the significance and importance of each predictor variable by examining the coefficients and p-values. This will help us identify the variables that have the most impact on a player's market value.

Additionally, we will perform feature engineering by creating new variables or transforming existing ones to enhance the model's predictive power. This could involve combining attributes, creating interaction terms. To further refine the model, we will explore other regression techniques such as decision trees, random forest, and gradient boosting. These methods can capture non-linear relationships, interactions, and complex patterns that may be missed by linear regression.

To evaluate the performance of the models, we will use appropriate metrics such as mean squared error (MSE) or R-squared. We will also conduct cross-validation to assess the models' robustness and generalizability.

**Output:** The output of the football player market value prediction project is the estimated market value for individual players. This could be in the form of a numerical value, indicating the expected transfer fee or market valuation. The predictions may also include confidence intervals or ranges to account for uncertainty.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection:** Bundesliga (Germany), La Liga (Spain), Premier League (England), Seria A (Italy) and Ligue 1 (France) male players.

**Observation window:** 2022-2023 season

**Inclusions:** -

**Exclusions:** Players who played in the “Goalkeeper” position

**Data Sources:** Transfermarkt (<https://www.transfermarkt.com/>) and FBref (https://fbref.com/en/)

**Audience Level:** Football fans, Football recruiters, Recruitment agents, Club’s Transfer Departments

**Variable Selection:** Please refer to the DATA\_DICTIONARY.xlsx file

**Derived Variables:** -

**Assumptions and data limitations:**

* *The player’s club information (CLUB\_NAME, LEAGUE\_COUNTRY columns) will be the last team they joined/played in the 2022-2023 season (This only applies for the players who have played for more than one team in the season. The player’s statistics will be aggregated so the information about his performance will not be lost.)*

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| Risk | Likelihood (based on historical data) | Delay (based on historical data) | Impact |
| *Domain expertise and interpretation* | *Low* | *-* | *Incorporating domain expertise into the analysis and interpretation of the results can provide valuable insights and enhance the accuracy and relevance of the predictions.* |
| *Feature selection and engineering* | *Medium* |  | *Selecting the right set of player attributes, considering interaction terms are important for capturing the factors that drive player market values.* |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item | Major Events / Milestones | Description | Deliverables | Days | Date |
| 1. | Project definition | *Clearly defining the objectives, scope, and deliverables of the project, outlining the problem statement, and setting the direction for subsequent steps.* |  | *7* | *2 Jul – 8 Jul* |
| 2. | Data gathering | *Relevant data is collected from various sources, ensuring its quality and reliability, and preparing it for analysis.* |  | *7* | *9 Jul – 15 Jul* |
| 3. | Data Cleaning  Exploratory Data Analysis | *The collected data is cleaned, transformed, and preprocessed to remove inconsistencies and errors. Exploratory data analysis techniques are applied to gain insights and understand the underlying patterns in the data.* | *17 Jul:*  *Analysis Plan / Data Finalization* | *7* | *16 Jul – 22 Jul* |
| 4. | Feature Engineering  Modeling  Prediction | *Features are selected or created from the data to build predictive models. Regression algorithms are applied to train the models and make predictions or classifications based on the available data.* | *24 Jul:*  *Data Exploration* | *14* | *23 Jul – 5 Aug* |
| 5. | Data Governance  Conclusions | *This phase involves ensuring data quality, integrity, and security. It also includes validating the models, interpreting the results, and drawing meaningful conclusions from the analysis.* | *7 Aug:*  *Modeling*  *12 Aug:*  *Governance* | *7* | *6 Aug – 12 Aug* |
| 6. | Presentation | *The findings, insights, and conclusions obtained from the project are communicated to stakeholders through effective visualizations, reports, and presentations, highlighting the key takeaways and recommendations for action.* | *13 Aug:*  *Peer feedback*  *14 – 16 Aug:*  *Documentation and*  *Presentation*  *18 Aug:*  *Portfolio* | *7* | *13 Aug – 19 Aug* |