**Project Plan: Predicting the Ideal Striker for Arsenal**

**1. Introduction**

This project aims to identify the striker who best fits Arsenal’s style, philosophy, and tactical structure using a data-driven approach. The goal is to combine tactical, performance, and market data to recommend a striker who can deliver consistent goal contributions (20–25 goals per season), assists, and effective link-up play while aligning with Arsenal's identified gaps and transfer strategy for the summer 2025 window.

**2. Objectives & Key Performance Indicators (KPIs)**

**Objectives**

* **Identify a striker** that best fits Arsenal’s tactical requirements.
* **Ensure consistent performance** in terms of goals, assists, and overall contribution.
* **Incorporate tactical analysis** such as movement, pressing, and link-up play.
* **Factor in market feasibility** and cost-effectiveness for a realistic transfer target.

**KPIs**

* **Performance Metrics:** Goals, assists, conversion rates, expected goals (xG), and expected assists (xA).
* **Tactical Fit:** Composite "fit score" based on quantified data (movement, pressing, link-up play) sourced from Instat/WyScout.
* **Availability:** Percentage of minutes played, injury history, and reliability indicators.
* **Cost Metrics:** Market value, transfer fee, wage demands, and cost-efficiency ratios.
* **Benchmarking:** Comparison against top strikers in the top 5 leagues (Premier League, La Liga, Serie A, Bundesliga, Ligue 1) to identify Arsenal’s current gaps.

**3. Current Analysis of Arsenal’s Standards & Gaps**

* **Internal Analysis:** Review historical striker performance, tactical reports, and expert analyses.
* **Benchmarking:** Compare key performance metrics against strikers from top teams.
* **Gap Identification:**
  + Determine areas such as finishing, physicality, and link-up play where Arsenal underperforms.
  + Use these insights to refine the tactical "fit score" for candidate evaluation.

**4. Data Sources & Collection Strategy**

**4.1 Tactical Data Integration**

* **Data Sources:**
  + **Instat or WyScout:** Extract data on player movement, pressing, and link-up play.
* **Purpose:**
  + Quantify tactical nuances to create a composite "fit score" for each striker.
  + Compare each striker’s style against the tactical gaps identified in Arsenal’s current performance.

**4.2 Performance Data**

* **Data Sources:**
  + **Sportradar, Soccerway, Transfermarkt:** Scrape detailed performance metrics, market values, and injury histories.
* **Purpose:**
  + Evaluate historical performance and reliability.

**4.3 Social Media Sentiment Analysis**

* **Data Sources:**
  + **Twitter, Reddit, and relevant forums:** Scrape public sentiment, trends, and buzz.
* **Purpose:**
  + Capture qualitative aspects of player performance and market opinion.
  + Tie sentiment data into the overall analysis to compare popular candidates versus under-the-radar "hidden gems".

**4.4 Clustering Techniques for Outlier Detection**

* **Methodology:**
  + Apply clustering algorithms (e.g., k-means) on performance metrics to identify outliers.
* **Purpose:**
  + Detect players who consistently deliver strong performances but may be undervalued by mainstream media.

**5. Data Warehousing & Integration**

**5.1 Data Storage**

* **Primary Storage:**
  + **GitHub Repository:**
    - Store cleaned and processed datasets as CSV or JSON files.
    - Use GitHub’s raw file URLs to integrate with PowerBI.
* **Note:**
  + This is a cost-effective, portfolio-friendly solution. Real-time reporting is not required for this project.

**5.2 ETL Pipeline**

* **Steps:**
  1. **Data Acquisition:**
     + Extract data via APIs and web scraping from selected sources.
  2. **Data Cleaning & Normalization:**
     + Remove duplicates, handle missing values, and standardize metrics (e.g., per 90 minutes).
  3. **Data Integration:**
     + Merge tactical, performance, and sentiment data using unique player identifiers.
* **Modularity:**
  1. Ensure pipelines are modular to allow future expansion to other positions or teams.

**6. Exploratory Data Analysis (EDA) & Feature Engineering**

**6.1 Descriptive Analysis**

* **Techniques:**
  + Use histograms, box plots, scatter plots, and correlation matrices to explore data distributions.
* **Tools:**
  + Python (Pandas, Matplotlib, Seaborn) for initial EDA.

**6.2 Feature Engineering**

* **Derived Metrics:**
  + **Tactical Fit Score:**
    - Combine Instat/WyScout metrics (movement, pressing, link-up play) into a composite index.
  + **Availability Percentage:**
    - Calculate as the ratio of minutes played to total possible minutes.
  + **Performance Consistency Score:**
    - Compute the standard deviation of key performance metrics over the past 5 seasons.
  + **Cost-Efficiency Index:**
    - Ratio of market value/transfer fee to performance contributions.
* **Clustering:**
  + Use clustering algorithms to identify outlier performers (hidden gems) versus popular high-performers.

**7. Predictive & Comparative Analysis**

**7.1 Predictive Modeling**

* **Methods:**
  + **Regression Models:**
    - Forecast future goal contributions (linear or Poisson regression).
  + **Classification Models:**
    - Evaluate player fit with Arsenal’s tactical requirements (Random Forest, SVM).
  + **Ensemble Techniques:**
    - Combine multiple models for robust predictions.

**7.2 Comparative Benchmarking**

* **Approach:**
  + Compare candidate metrics against top strikers from the best teams in the top 5 leagues.
  + Analyze tactical fit scores, cost-effectiveness, and performance consistency.
* **Cost-Benefit Analysis:**
  + Develop decision matrices to evaluate transfer scenarios and ROI.

**7.3 Validation**

* **Backtesting:**
  + Validate models using historical data to assess if they would have predicted successful past signings.
* **Expert Input:**
  + Although detailed scouting reports are omitted, incorporate any available domain insights for validation.

**8. Insights Extraction & Reporting**

**8.1 Key Insights**

* **Performance Trends:**
  + Identify candidates with consistent high performance and tactical compatibility.
* **Tactical Alignment:**
  + Highlight players whose movement, pressing, and link-up play address Arsenal’s tactical gaps.
* **Cost vs. Benefit:**
  + Balance market value with performance and reliability indicators.
* **Hidden Gems:**
  + Use clustering results to uncover undervalued players who may have been overlooked by mainstream sentiment.

**8.2 Reporting**

* **Deliverables:**
  + A comprehensive report detailing methodology, data sources, analysis steps, and recommendations.
  + Clear documentation of data cleaning, integration, and feature engineering processes.
  + An executive summary aligning insights with real-world transfer strategy principles.

**9. Data Visualization & Dashboard Development**

**9.1 PowerBI Dashboard Components**

* **Overview Dashboard:**
  + Summary metrics for each candidate including composite fit scores.
* **Performance Trends:**
  + Time series charts showing goals, assists, and tactical fit metrics over 5 seasons.
* **Cost vs. Performance:**
  + Scatter plots or bubble charts comparing market value with performance indices.
* **Comparative Benchmarking:**
  + Radar or spider charts comparing Arsenal’s tactical gaps versus top league benchmarks.
* **Outlier Identification:**
  + Visualizations highlighting hidden gems identified through clustering.

**9.2 Integration with Data Warehousing**

* **Data Source:**
  + Use raw file URLs from GitHub for data integration in PowerBI.
* **Interactivity:**
  + Design the dashboard to be self-explanatory and interactive, making it ideal for a portfolio project.

**10. Additional Considerations**

**Scalability**

* **Modular Design:**
  + Ensure that data pipelines are modular to easily extend the project to other positions or teams.

**Data Privacy & Ethics**

* **Compliance:**
  + Ensure all collected data is anonymized where necessary.
  + Adhere to terms of use for any scraped or licensed data.

**Integration with Transfer Strategy**

* **Alignment:**
  + Clearly integrate findings with real-world transfer strategy principles.
  + Use insights to demonstrate how data can drive practical transfer decisions.

**11. Project Wrap-Up**

**Final Deliverables**

* **Documentation:**
  + Complete project documentation hosted in a GitHub repository including all code, datasets, and detailed methodology.
* **Portfolio Presentation:**
  + A PowerBI dashboard and written report summarizing the project’s approach, insights, and recommendations.

**Future Expansion**

* **Scalability:**
  + Maintain modular pipelines and clear documentation to enable future expansion to additional positions or teams.
* **Continuous Improvement:**
  + Plan periodic reviews to update models and data pipelines as new data becomes available.