## Problem

Can we predict the player salaries in the Indian Premier League (IPL), given aggregate data (such as the strike rate, economy rate, and average) of the player by season?

## Context

The IPL has become the premier destination for global viewership of league cricket. With editions of the IPL having been played since 2008, and player salaries having exploded since then, there is a large demand for understanding the most important aspects that go into various IPL franchises' auction strategy.

The objective for this project is to develop a model that can accurately predict the salary of players based on their performance as measured by aggregate statistics. IPL franchises can use this information to make data-driven decisions in the auction. Furthermore, this can help franchises identify undervalued and overvalued players.

## Criteria for Success

The primary focus of my capstone will be on finding the correlation of various statistical aggregate measures to the yearly salary of the player. The project will be considered successful if there is found a statistically significant correlation between the statistical aggregate data of the player and the salary of the player.

## Solution Methodology

1. *Data Collection and Cleaning*: I will collect data (IPL 2008 to 2022 All Match Dataset | Kaggle, IPL player salary | Cricmetric) and clean it to remove any missing or irrelevant values.
2. *Exploratory Data Analysis and Feature Engineering:* I will perform Exploratory Data Analysis to find patterns and new features in the data. I will engineer these features using existing metrics to better capture the performance of the player. The following features will be considered:
    - Player nationality/international status
    - Player base price
    - Player role
    - Player performance
    - Player aggregates based on game situation
3. *Model Selection and Training:* To find the most accurate prediction, I will use cross-validation techniques to select the best performing Machine Learning algorithm. I will use Mean Absolute Error and Mean Squared Error as metrics to evaluate the performance of the algorithms. Some algorithms that will be considered are:
    - Linear Regression

    - Random Forest
    - Gradient Boosting

4. *Model Interpretation:* After determining the best-performing model, I will determine the most important features in predicting player salary.
5. *Deliverables:* I will deploy the best performing model through a web application, built through Flask or Django, where a franchise can input a player's statistics to get an expected salary. The final deliverables will be a web application, and a GitHub repository containing the notebooks for each step of the process, a slide deck, and a project report.

## Potential Constraints

- The salary data is only available from 2008 to 2021, so I will focus on those seasons.
- Player performance is measured by aggregate statistics, which may not tell the full story of a player's impact on the outcome of a game. Additional features based on player role and game situation may be added through my analysis which will mitigate this.
- The salary of players may be inflated or deflated depending on reputation, which will not be reflected in the statistical data of the player. Additional features based on player international status may be added through my analysis to mitigate this.

## Conclusion

This project will develop a Machine Learning model to predict the player salary based on statistical aggregate features. In doing so, I will develop a framework for the IPL franchises to make more informed decisions during player auctions on the expected salary of the players. This can help them identify overvalued and undervalued players, allowing them to build more balanced teams.