According to Yahoo Finance, the sport analytics market is projected to be valued at US$ 31.4 billion by 2034[1]. This is due to the incredible prospect of improving athletic performance, team strategies, and business aspects within the sports industry through sports analytics.

The machine learning model proposed includes:

1. Feature selection for calculating swing probability
2. Swing probability prediction
3. Hit probability preditiction
4. Analysis of pitcher and batter performance

First, the model will identify the most important features through various feature selection techniques from both game situations (ie. score, inning, runners, count, outs, pitch hand, bat side, etc.) and pitch features (pitch location, release point, and Statcast metrics, etc.) to calculate the swing probability. This model will use Recursive Feature Elimination (RFE) using a logistic regression estimator to identify the most relevant features by recursively removing less important features. The swing probability calculated will then be used as an input along with pitch features to calculate a hit probability. Finally, we will use logistic regression to find the swing and hit probability, accuracy, and F1 score.

Logistic regression applies a sigmoid function to represent the probability:

$$\text{Sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

Where z is the output of the linear equation, w are the weights, x are the feature data, b is the bias:

$$z = \mathbf{w}^T \mathbf{x} + b$$

Using this information, both pitcher and batter performance can be analyzed in comparison to these probabilities. This means we could predict a positive pitch or hit from various features and potentially know the play before it happens, thus improving athletic performance and game strategy.

[1] Yahoo Finance, https://finance.yahoo.com/news/sport-analytics-market-expected-reach-180000614.html