**English Premier League Prediction Engine**

**Under the guidance of Sean Mondesire**

**Problem:**

Develop a model that generates accurate predictions of 10 matches from the English Premier League.

**Background/Literature Review:**

The English Premier League has the most lucrative broadcast television deal among European soccer leagues. According to a report from Deloitte’s Sports Business Group, the league reported a profit of $2.42 billion in the 2017/2018 season, and in the previous season, reported a record revenue of nearly $6.4 billion (CNN articles cite). Due to the league’s popularity, a lot of league’s inflow of cash comes from gambling, and predicting the outcome of matches for this league in particular has become more common worldwide.

**Rationale for the choice of DM and ML techniques:**

* **Data Mining:** Linear regression analysis.
* **Machine Learning:** eXtreme Gradient Boosting (XGBoost)

**Description of Design:**

For predicting the result, a linear regression analysis will be implemented along with an eXtreme Gradient Boosting technique. The data from the CSV. file will be transferred into the model and subsequently it will be able to generate results for further analysis.

**Explanation of data (file format, independent/dependent variables, sources):**

The data is organized in a CSV. Excel file with the historical and current season’s data. The metrics and variables used are geared more towards goals, since they directly impact a match’s result, which is the dependent variable. A CSV. File facilitates the transfer of data into the model.

**Identification of Training and Testing data**

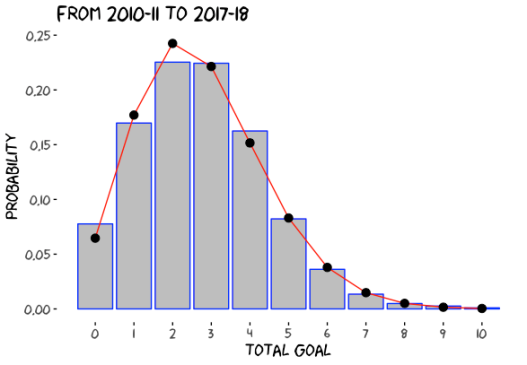
* **Training data:** from the entire sample of data, 80% is being allocated towards the training phase (i.e. 1,000 rows of data)
* **Testing data:** the remainder 20% of the data from the entire sample is going to be tested and later use the output and compare it with previous matches’ results to determine degree of approximation.

**Testing and Validation Procedure:**After training the data allocated for the training set, the sample testing set will be used to test the model, and the model should produce an output with relative certainty to the real-life result

**Results and Incites:**

The model produces predictions that closely approximate to previous scores and predicts outcomes with a certain level of accuracy.

**Graphs:**

****

**Citations (as a footnote):**

* Garcia, A. (2018). Premier League revenues hit record high $6.4 billion. *CNN Business*. Retrieved April 19, 2018, from https://money.cnn.com/2018/04/19/news/companies/premier-league-record-revenue/index.html.
* Barnard, M., Dwyer, M., Wilsom, J., & Winn, C. (2018). *Annual Review of Football Finance 2018* (pp. 1-36, Rep.) (D. Jones, T. Bridge, & M. Green, Eds.). Manchester, UK: Deloitte.