Project2/Milestone 1: Proposal & Data Selection

Author: Anjani Bonda

Date: 7/08/2023

## **Topic**

IPL (Indian Premier League) Match Predictive Analysis

## **Business Problem**

The [Indian Premier League](https://www.iplt20.com/) is a professional Twenty20 cricket league in India. There are 10 franchises in the league representing different cities/state in India. IPL features many international and domestic players and usually happens between march and may every year. IPL (Indian Premier League) is the 2nd biggest sports league in the world next to NFL in USA. This league is based on the sport – Cricket. For a better picture, here is the list of biggest sports leagues in the world.

|  |  |  |  |
| --- | --- | --- | --- |
| **League Rank** | **League Name** | **Country** | **Value** |
| 1. | National Football League | America | $16 billion |
| 2. | Indian Premier League | India | $10 billion+ |
| 3. | Major League Baseball | America | $10 billion |
| 4. | National Basketball Association | North America | $8 billion |
| 5. | English Premier League | England | $5.3 billion |
| 6. | National Hockey League | North America | $4.8 billion |
| 7. | La Liga Santander | Spain | $4.5 billion |
| 8. | Bundesliga | Germany | $4.3 billion |
| 9. | Serie A | Italy | $2.3 billion |
| 10. | UEFA Champions League | Switzerland | $2 billion |

IPL has only started in 2008 and it is already one of the fastest growing sports leagues in the world taking 2nd spot recently. The viewership and revenue have been in an uptrend too. Data Analytics has been a part of sports entertainment for a long time. As a sports enthusiast, I am curious to build a predictive analysis model to predict the winning team in an IPL match/tournament using various available stats available for all seasons (2008-2022).

## **Datasets**

The datasets are extracted from Kaggle website.

<https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?resource=download&select=all_season_summary.csv>

Source: <https://www.espncricinfo.com/>

Data has been scraped and transformed into following files. The data provided in match level summary as well as ball-by-ball details format for all matches from 2008 till 2022 season.

* all\_season\_summary.csv - Summary of all matches across all seasons
* all\_season\_details.csv - Ball-by-ball details of all matches across all seasons
* all\_season\_batting\_card.csv - Batting performance of players, all matches across all seasons
* all\_season\_bowling\_card.csv - Bowling performance of players, all matches across all seasons
* points\_table.csv - Overall points table of teams across seasons

The main dataset to be used in this model is “all\_season\_summary.csv” which has about 45 columns and 958 records summarizing every single match from the beginning of season (2008 – 2022).

## **Methods**

Below algorithms or model techniques will be utilized on the dataset to determine which features are related to our target variable “winner”. Since the output of winner prediction is a categorical value, the problem which we are trying to solve is a Classification problem.

1. Logistic Regression
2. Random Forest
3. Decision Tree
4. Support Vector Machine (SVM) Classifier
5. K-fold (if required)

Logistic regression is a statistical analysis method used to predict a binary outcome such as yes or no based on prior observation of the data set. Here, “Purchase” feature present in the dataset has only binary values and will be used as target for the model. This model falls under supervised learning as the data is well labelled and has a target variable, a column in the data representing values to predict from other columns in the data. Under supervised learning, this dataset falls under classification model as it reads the input and generates an output that classifies the input into two categories: one having purchase as “Yes” and “No”. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

Support vector machines so called as SVM is a *supervised learning algorithm* which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is used for smaller dataset as it takes too long to process. In this set, we will be focusing on SVC.

## **Ethical Considerations**

There are no possible ethical aspects to this model as the data is public info and doesn’t really include any consumer or personal related information.

## **Challenges/Issues**

Although the datasets taken from Kaggle have great deal of information, we can only assume that this is not an accurate dataset and not being fact checked.

The models employed may not be highly accurate but given the data, anything more than 50-60% can be reasonably considered to be good.

More features might be required to enhance this model.

## **Reference**

<https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons/versions/27?resource=download>

<https://www.kreedon.com/top-10-biggest-sports-leagues-in-the-world/>

<https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589>