Project 2: IPL (Indian Premier League) Match Predictive Analysis

Author: Anjani Bonda

Date: 7/22/2023

Table of Contents

Business Problem

The Indian Premier League 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).

Background/History

Since the launch of the IPL in 2008, it has attracted viewers all around the globe. Great level of uncertainty and crunch matches has urged fans to watch and increase view count. Within a short span, IPL has become the highest revenue generating league of cricket. Per the stats, the IPL is currently #2 league in the world next to NFL in USA. In a cricket match, we might have seen the small prediction ticker on screen showing the probability of the team winning based on the current match situation. This is where the data analytics and data science come into play, and we plan to build a prediction model for the same.

Data Explanation

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.

Analysis

Since this is a classification problem, multiple algorithms can train the classifier according to the data being fed and using the pattern, we would predict the outcome. We will be trying Decision Tree Classifier, Random Forest Classifier, Logistic Regression and possibly SVM or K-fold and finally choose the algorithm best suited for this use case data.

This should include the modeling analysis with accuracy score calculated for all the models.

Accuracy**:** Accuracy represents the number of correctly classified data instance over the total number of data instances.

This should also include feature analysis using various methods to find the best features from the dataset. Best feature in the dataset which shows higher impact to the target variable “Match Winner” compared to others present in the dataset.

Conclusion

The models being employed for this prediction model are Logistic Regression, SVM, Decision Tree and Random Forest and calculated the accuracy score for each of these models. Out of these 4 models, Logistic Regression and SVM had higher accuracy score of 68.75% than other algorithms for this data distribution. Even though the accuracy is not high enough to be useful, it gives an idea of the strategies and methodologies employed in designing a solution to the machine learning problem.

Assumptions

The datasets being considered may not have all the required features to support the model. I have taken the best possible dataset from the available sources. Also, some other features which may not be relevant will also be excluded.

No supplemental datasets are being used with the assumption of the summary dataset is good enough to get started and build a model.

Limitations

The dataset considered for this prediction model is to be considered a fictional dataset as it may not represent real-world or factual data. The same goes for the supplemental datasets as well. Also, the number of features included in the current dataset may not be enough for an accurate prediction model although good enough to build one.

Challenges

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 60-70% can be reasonably considered to be good.

More features might be required to enhance this model.

Future Uses/Additional Applications

While this may not exactly represent the real-world data, this model is still similar and can be run against real-world datasets to all other similar cricket leagues around the world to gain useful insights.

Recommendations

This model predicts the match/league winner and relevant useful features that impact the match prediction with better accuracy with a caveat that the model should be regressed when more or better real-world data is available.

Implementation Plan

As stated in the recommendations, this model can be implemented to predict the winner of a cricket match/tournament along with evaluation of other useful features that may impact the same outcome.

Data exploration is initially performed to clean up the dataset and remove any unwanted features.

Data visualizations are done on top of the key features and factors that contribute to the target variable.

Feature engineering is done on top of this to consider only the relevant features and establish relationships and correlation.

Finally, the models are employed and calculated the accuracy scores for the prediction.

Ethical Assessment

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.

References

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