Data Analysis of the Indian Premier League

Utkarsh Agiwal
Dept. of Mechanical Engineering
Indian Institute of Technology, Bombay
Mumbai, India
200020154@iitb.ac.in

***Abstract*—In recent times, before COVID-19 was at its peak, cricket has been a very popular sport in the South-east Asia and other parts of the world. One format that unites players of all countries and has turned out to be one of the highest revenue generating is the IPL or the T20 format. Therefore it is imperative to induce predictions in this field to test out the algorithms. In the following paper, I have applied various Machine Learning and Deep Learning Models to predict the final score and winner of a match depending upon certain parameters to a certain accuracy.**

I. Introduction

Sports analytics are a collection of relevant, historical, statistics that can provide a competitive advantage to a team or individual. Through the collection and analyzation of these data, sports analytics inform players, coaches and other staff in order to facilitate decision making both during and prior to sporting events. It is gaining popularity since it gives advantage to both players as well as the various stakeholders of the game. It would also help in creating strategies which provide better chances to win. Recently, various sources of game prediction have been brought up to facilitate audience in predicting various parameters of the game. The spectators have shown a keen interest too and this field is exceeeding its limitations day by day.

A. About Cricket

Cricket is a bat-and-ball game played between two teams of eleven players each on a field. The game proceeds when a player on the fielding team, called the bowler, "bowls" the ball from one end of the pitch towards the wicket at the other end, with an "over" being completed once they have legally done so six times. Forms of cricket range from Twenty20, with each team batting for a single innings of 20 overs and the game generally lasting three hours, to Test matches played over five days. A batsmans' role is to score maximum runs while saving their wicket whereas bowlers duty is to knock out as many wickets as possible. Analytics can be used to influence the toss decisions, win rate, playing eleven, predicted score in accordance with the against team, stadium and weather conditions. With correct formula, it can bridge gap between a non-funded and a funded strategic team.

B. About Indian Premier League

The Indian Premier League (IPL) is a professional men's Twenty20 cricket league and is the most-attended cricket league in the world. In 2014, it was ranked sixth by average attendance among all sports leagues. There have been fourteen seasons of the IPL tournament. Each team plays each other twice in a home-and-away round-robin format in the league phase. At the conclusion of the league stage, the top four teams will qualify for the playoffs. The top two teams from the league phase will play against each other in the first Qualifying match, with the winner going straight to the IPL final and the loser getting another chance to qualify for the IPL final by playing the second Qualifying match.

II. Datasets & Feature Engineering

The datasets used for analysis and prediction were collected from kaggle. Overall, three datasets were used. Two of them had data for IPL matches right from the first season i.e. 2008 to 2020 and the third one (testing dataset) contained the matches happened in the latest season. The first two dataset were merged giving 193,468 instances each described by 34 features.

The dataset consisted of a large number of unwanted features. A thorough exploratory and descriptive data analysis supported by statistical hypothesis testing reduced the number of features to 16 of which 12 features were used for ML modelling. Several categorical features like names of the teams were categorically encoded in increasing magnitude based on their performance concluded from EDA. Since the data consisted of the games right from the first season, teams like Delhi Daredevils - Delhi Capitals, Kings XI Punjab - Punjab Kings, Rising Pune Supergiants - Pune Warriors were considered same during encoding. One hot encoding was used for features like eliminator, toss decision and match result if it was normal or D/L.

Similar pre-processing and feature engineering was carried out on the 2021 dataset (primary test set). It consisted the data of 21 matches happened in first half of the year. The data of the other half of the season which happened in september was not available at the time of making of the report.

III. Analysis Pipeline

As in analytics, some analysis is carried out earlier to find out various conclusions regarding players, teams and their luck. This analysis is helpful for teams to make long-term predictions. Such type of analysis can be effective only before the start of the match and can't be changed once the match started. So, we start out with Exploratory Data Analysis to judge various parameters necessary.

A. *Exploratory Data Analysis*

Meaningful Expolaratory Data Analysis is what helps us in preventing problems like overfitting and underfitting.

1) Total Matches played in a season : As seen in Fig. 1, total matches increased in the years 2011, 2012 and 2013. The main reason behind this was addition of teams in IPL. Hence these three seasons shouldn't be used as a testing dataset. Thereon, the total games are almost constant with only slight differences mainly due to calling off the match due to weather conditions.

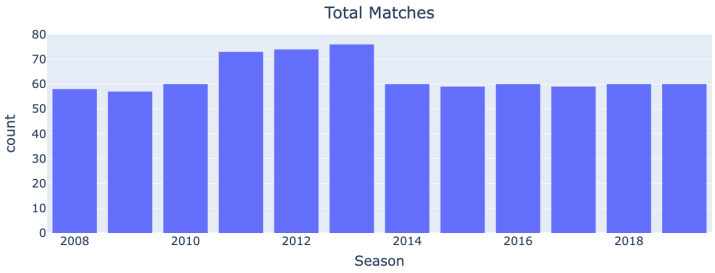


Fig. 1. Total Matches played in each season

2) Most Valuable Players : This gives an insight about the top players in the Indian Premier League.

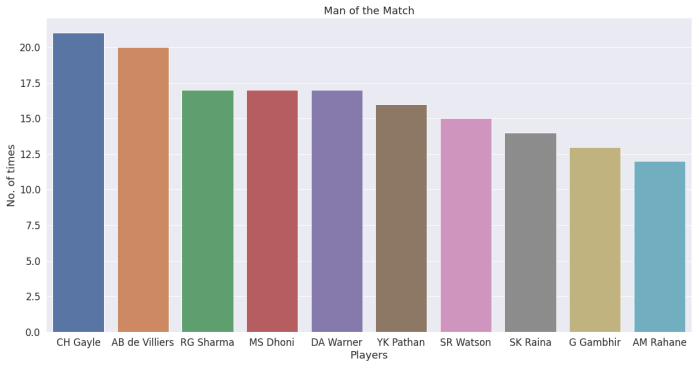


Fig. 2. Most Valuable Players

3) Max Wins : Current and Past strongest Teams.

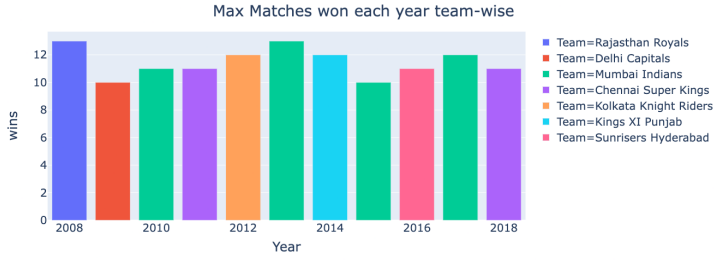


Fig. 3. Winners of different seasons

4) Top Cities : Revenue from such tournaments isn't only generated by sponsorship and marketing. Tourism revenue comes in handy with such tournaments. As seen in Fig. 4. Mumbai is the most favourable hosting city.

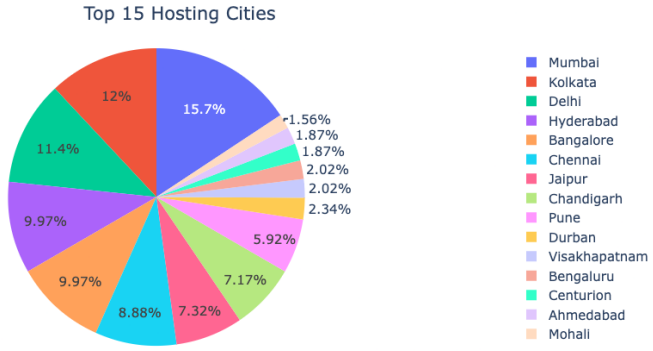


Fig. 4. Top Cities

5) Toss Analysis : In fig. 5, as we can see, the trend of toss decision has reversed over the years. Upto 2013, batting used to be considerate option but in the recent years fielding has completely dominated its counterpart.

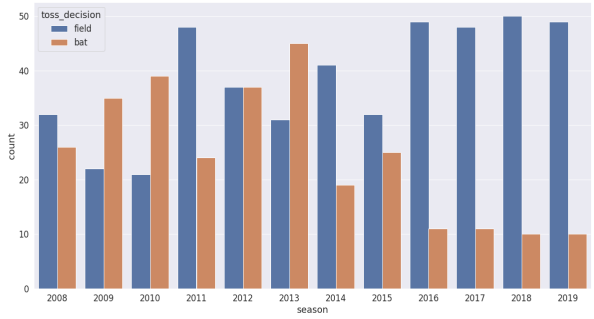


Fig. 5. Toss Trends

6) Toss Influence : As it can be observed in fig. 6, winning the toss doesn't help much as teams were only 52.3% times they were right in their decision.

7) Toss and win comparison : Here we checked team-wise relationship between their match winning and toss winning percentages. Observations include it really comes down to players strength. First half of the teams in the graph have winning% greater than toss% while the other half just has reverse of it.

8) Team rivalry : Year wise statistics of teams rivalry are also plotted. It can help a team identify against whom are they strongest and weakest in the recent years providing them a win/loss ratio. Fig. 8. includes CSK vs MI analysis over different seasons.

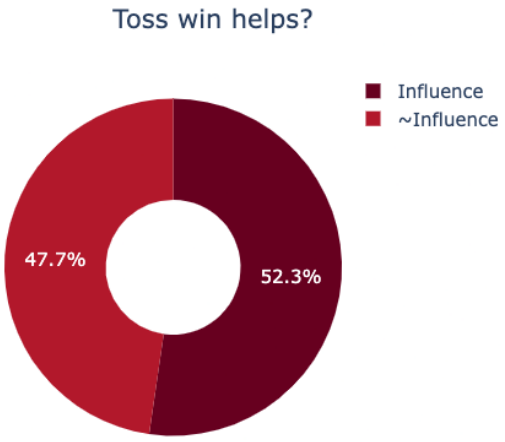


Fig. 6. Does it matter?

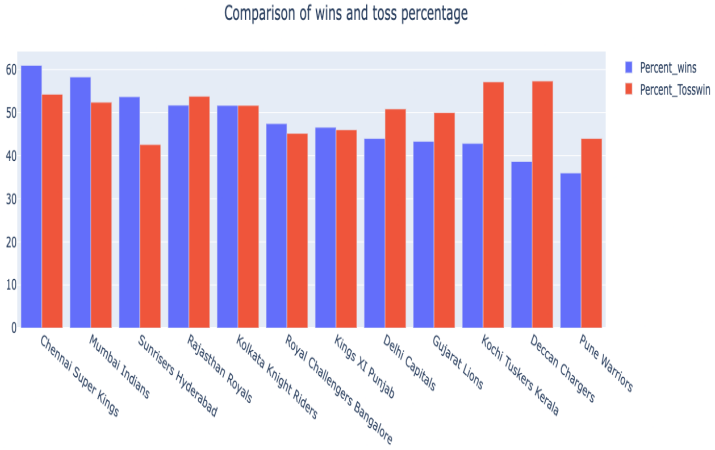


Fig. 7. Team toss-wise analysis

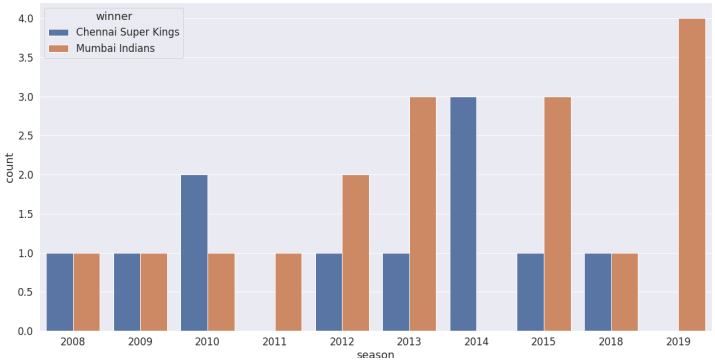


Fig. 8. CSK vs MI

IV. Evaluation

The EDA helped us to reduce the number of features from 34 to 16 of which 12 will be used for modelling.

A. Metrics Used

1) Machine Learning: For prediction of total runs, R^2 score and MSE (Mean Square Error) were used. A new accuracy metric was defined which gave accuracy if the score is within a limit of +-12. This accuracy function was used throughout the total runs prediction model.
For predicting the winner, classifiers were trained and the default accuracy model of scikit learn was used. Teams were used as classes which were label encoded on the basis of their strength concluded from EDA. Therefore, regressors were trained and floor of the output was taken to be the class it belonged to.
2) Deep Learning: For predicting total runs, the accuracy function was used same as in the Machine Learning. In addition, validation and training loss were calculated at each epoch.
For predicting the winner, NN classification was used and its default accuracy model. Similarly, validation and training loss was calculated at each epoch.

B. Models Used - Score Prediction

1) Support Vector Regression (SVR): Since number of features were small and training examples were intermediate/large compared to features, it makes sense to use support vector machine.
Hyperparamter Tuning: The paramter C and degree are the most crucial parameters for a support vector machine. Hyperparameter tuning was performed on the dataset with 5-fold cross validation strategy on C, degree and epsilon. Best values of were found to be 2, 1 and 0.1 respectively.
Results: Choosing the best parameters obtained from GridSearch CV, we get accuracy as 60.27% and R^2 score as 0.752 on 2019 IPL(Fig. 10.) data whereas 57.14% and R^2 score as 0.769 on 2021 data(Fig. 9.).

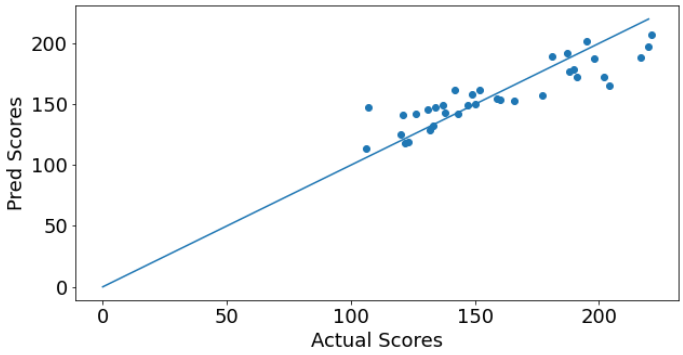


Fig. 9. IPL 2021

2) Random Forest Regressor: The tree growing in Random Forests happen in parallel which saves a lot of time

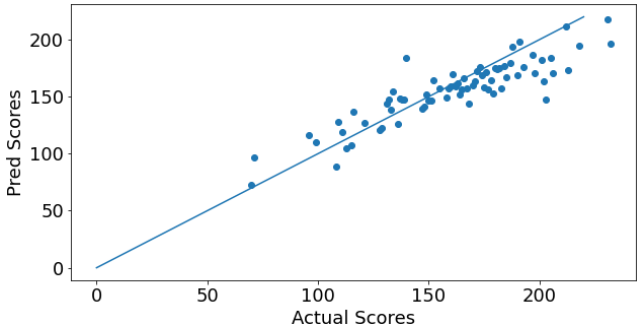


Fig. 10. IPL 2019

if the dataset size is large. But with the data chosen, the same for SVR, it was highly overfitting as depicted in Fig. 11. even when the n_estimators ranged from 1-5000.

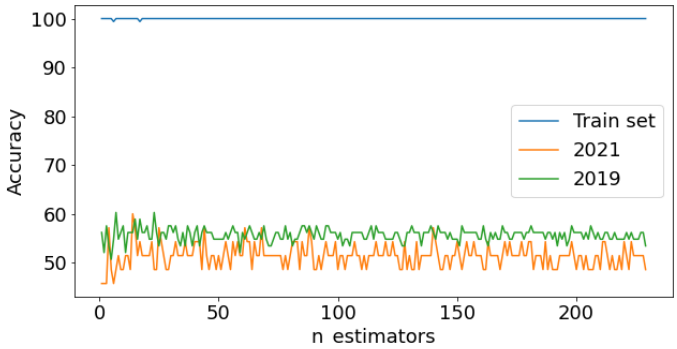
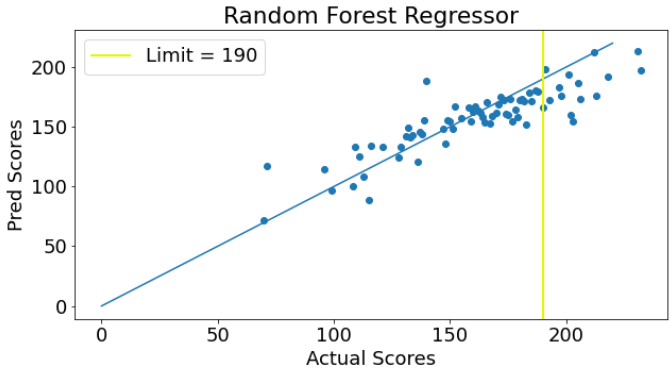


Fig. 11. Random Forest

Hyperparameter Tuning: As even with the tuning, the model was highly overfitting. So the dataset size was increased i.e. instead of the data from overs 11-16, 7-16 was chosen to be trained upon. However, the result turned out to be the same. Hyperparameter tuning gave best parameter as n_estimators = 70.



Results: Selecting the parameters obtained from GridSearchCV, we get accuracy as 56.16% and R^2 score as 0.74 on 2019 IPL data (figure below) whereas 54.285% and R^2 score of 0.785 on 2021 set. It was noted that regressor was going off-limit for actual scores over 190.

3) Linear Regression : Linear regression is the simplest of model. This model was chosen to avoid any chances of overfitting.
Results: As expected, results weren't much fascinating with 50% accuracy on the 2019 as well as the 2021 dataset and R^2 score 0.70 and 0.77 respectively.

4) Neural Networks(NNs): Neural Networks are the modern prediction models. Analogies similar to human nervous systems are drawn. Each unit is considered as a neuron which fires an output through an activation function upon receiving an input. Several layers of units are stacked together to predict the results.
Input layer has number of units equivalent to the number of features which are passed to the second layer through activation functions. Functions like sigmoid, RelU and Leaky RelU are used. Adam Optimizer and SGD (Stochastic Gradient Descent) are used for back-propagation. SGD is usually not chosen since it suffers vanishing gradients problem. Dropout regularization is used to prevent overfitting.

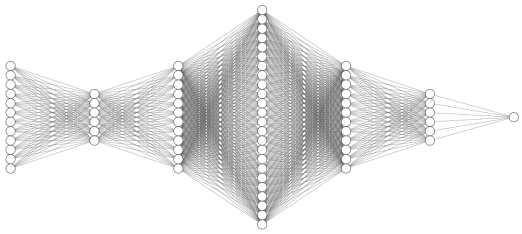


Fig. 12. Neural Network

7 layers were build with 5 of them being hidden layers. A total of 85k parameters were trained. Training the model on the same input as used in previous models didn't come out to be of much use as an accuracy of around only 50% came. Therefore, the neural network

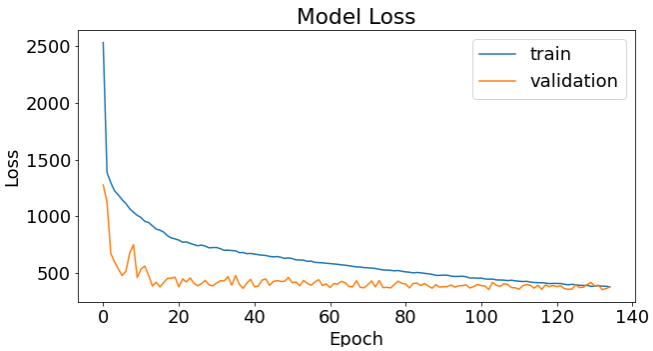


Fig. 13. Loss

model was decided to be trained upon individual balls to

see how it performs over that. Accuracy score of about 53% was acheived which was only slightly better and it wouldn't be even of much practical use. Limited amount of data is the primary reason why neural network isn't performing well over the dataset.

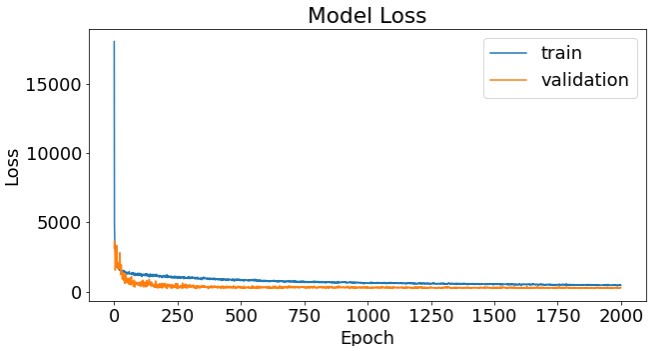


Fig. 14. Loss

C. Models Used - Winner Prediction

1) Support Vector Regression (SVR): Strategy described earlier in metrics used were used. Default parameters turned out to be the most useful. The winner prediction was done over 2021 IPL matches data.
Hyperparameter Tuning: C is one of the most important parameter and gridsearchCV gave optimized parameter C = 2.
Results: Accuracy score : 91.52%
2) Random Forest Regression: Metrics and input dataset are all same.
Results: Accuracy score : 91.52%
3) Support Vector Classifier (SVC): Since it is a classification problem, classifiers must be checked and not just trying out regression methods with custom defined accuracy functions.
Results: Accuracy score : 81.355%
4) Random Forest Classifier: All parameters are same.
Results: Accuracy score : 32.9%
5) Neural Networks: While neural networks weren't fruitful in predicting final score, quite the opposite was true in classification domain. A 6 layer model was trained with 4 hidden units with the final activation function being softmax. Stochastic gradient descent was used as the optimizer. It was trained over for only 30 epochs since after that overfitting was happening. (Fig. 15.)
Results: Accuracy score : 90.47%

V. Conclusion and Future Scope

This paper provides useful insights from IPL dataset about various teams & players along with the toss analysis and its importance in winning games for the strongest and weakest teams. Sponsors can do marketing in only certain cities citing highest spectators involved. The prediction of final score is done on the basis of teams involved in the match, the current

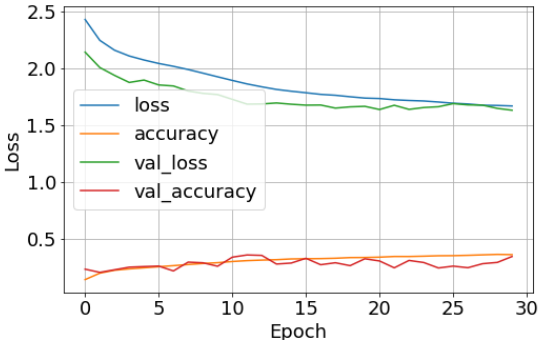


Fig. 15. Neural Network - winner

overl/ball, current score as well as wickets fallen, toss winner and its decision. The conclusion is as follows:

Model	Accuracy
SVR	60.50%
Random Forest	56%
Linear Regression	50%
Neural Networks	50%

Fig. 16. Conclusion - Score Prediction

SVR turns out to be the best algorithm for predicting the total runs. The prediction of final winner was also done on the same input parameters. The winner prediction like score prediction works from overs 11-16. The conclusion is as follows:

Model	Accuracy
Random Forest Regression	91.52%
Support Vector Regressoin	91.52%
Neural Networks	90.48%
Support Vector Classifier	81.35%

Fig. 17. Conclusion - Winner Prediction

Future scope including the batting averages and bowler economies of all players team-wise, taking account other T20 formats to predict the outcomes thereby accumulating more data for Neural Networks.

Acknowledgment

I would like to thank Analytics Club, IIT Bombay for hosting WIDS(Winter in Data Science) Bootcamp and providing a bunch of students with a mentor and a project along with great appreciation to my mentor, Prayas Jain to provide with the best resources enriching this experience.

References

[1] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition
[2] IPL Cricket Match prediction, Praveen Sridhar
[3] IPL Score prediction using Deep Learning, url: https://www.geeksforgeeks.org/ipl-score-prediction-using-deep-learning/