IPL Win Probability Predictor, End To End Machine Learning Problem.

Team:

Raghavendra K H (23M0013)
J Ramakrishna (23M0024)
Sujan D Silva (23M0011)

Course Instructor:

Prof. Amit Sethi

Course:

EE 769: Introduction to Machine Learning

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
Mumbai - 400076, India

ABSTRACT

Cricket is India's favorite sport, played in different ways like T20, ODI, and Test matches all over the country. The Indian Premier League (IPL) is a big cricket event where players from different parts of India and even other countries compete. The IPL is super popular because you can watch it live on TV, listen on the radio, or stream it online. People who bet money or sponsor teams are really interested in predicting who will win IPL matches. To make predictions, we look at things like which team has won the toss, what is the choice of the team after winning the toss (bat/field) and if the team is playing in the home ground or away. In a study, we have used Machine Learning Models like SVC, Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbor, Naïve Bayes and also the boosting algorithms to predict IPL match outcomes.

INTRODUCTION

Cricket is a game played outside with a bat and a ball, and it involves two teams, each with 11 players. It's all about teamwork. Cricket has three main ways it's played, and it's super popular worldwide. Just like any sport, lots of things decide who wins. Picking the team depends on how well the players perform and other stuff like where they're playing and how big the team is.

The Indian Premier League (IPL) is a big cricket event in India. Players come from different parts of India, and some even come from other countries. It's a T20 format and is owned by famous people and business folks. The IPL is run by the Board of Control for Cricket in India (BCCI). This year, there are eight teams in the IPL: Royal Challengers Bangalore (RCB), Rajasthan Royals (RR), Chennai Super Kings (CSK), Mumbai Indians (MI), Kolkata Knight Riders (KKR), Delhi Capitals (DC), Punjab Kings (PK), and Sunrises Hyderabad (SRH).

This paper is looking into some questions like: What's the chance of winning if you bat or bowl first after winning the coin toss? Does playing at home make a difference in winning?

We're using Machine Learning Algorithms like SVM, Random Forest, and Logistic Regression to figure out which team might win an IPL match based on where they're playing and what they decide to do after winning the toss.

There's a bunch of stuff that makes figuring out cricket matches tricky. Cricket comes in three types: Tests, Twenty-twenty (T20), and One Day International (ODI). It's not just a game for one country; it's played all over the world. Every ball in cricket is really important because even one ball can totally change the game.

Cricket isn't just about hitting balls and running around; there's a whole strategy behind it. Each team has to think smart about when to bat and when to bowl, and they have to adapt to the conditions of the pitch and the weather. Sometimes, a team might choose to bat first if the pitch looks good for batting, while other times they might choose to bowl first to take advantage of early moisture in the pitch. It's all about making the right decisions at the right time.

The Indian Premier League adds another layer of excitement to cricket. With players from different regions and even different countries coming together to compete, it's like a melting pot of talent and culture. Fans eagerly await each match, cheering for their favorite teams and players. The IPL

isn't just about cricket; it's also about entertainment, with flashy events, music, and celebrities adding to the spectacle. It's a chance for players to showcase their skills on a big stage and for fans to witness some electrifying cricket moments.

As technology advances, the way we analyze cricket matches evolves too. Using machine learning techniques allows us to crunch massive amounts of data to predict match outcomes more accurately. By looking at past performances, player statistics, and match conditions, these algorithms can give us insights into which team might have the upper hand in a particular match. It's a fascinating intersection of sports and technology, showing how innovation can enhance our understanding and enjoyment of the game.

METHODOLOGY

A. The dataset required to train the models was downloaded from kaggle. This datasets contained: 1. Team wise home and away dataset; 2. The Matches datasets; 3. Player's dataset; 4. Teams Datasets; 5. Deliveries dataset and 6. Most_runs_average_strikerate dataset. However, this data was modified to create the features which were required to solve the present problem.

The modified data file contained the following features:

1. Toss winner: This contained the names of the teams which won the toss in the respective matches. One hot encoding was performed on this column.

2. Toss decision: This contained the choice made by the toss winner (bat/field). Label encoding was performed on this column (bat:0, field:1)

3. Venue: This column contained the two names Home or Away depending on whether the toss winner was playing in home ground or away. Label encoding was performed on this column (Away:0, Home:1)

4. Result: This column had two names Win or Lose depending on whether the toss winner won the match or lost the match. Label encoding was performed on this column (Lose:0, Win:1). **This was the column whose results had to be predicted by the machine learning model.**

B. After the data was prepared, **x** and **y** datasets were created. **x** contained the data with the independent variables **Toss winner, Toss decision and Venue. y** contained the dependent variable **Result.**

C. Splitting the Data into test and train data: First, we divide the data into two parts: one for testing and one for training. We do this so we can check how well our model works. The training part is like a teacher helping the model learn. We use methods where the model knows the right answers to learn from them.

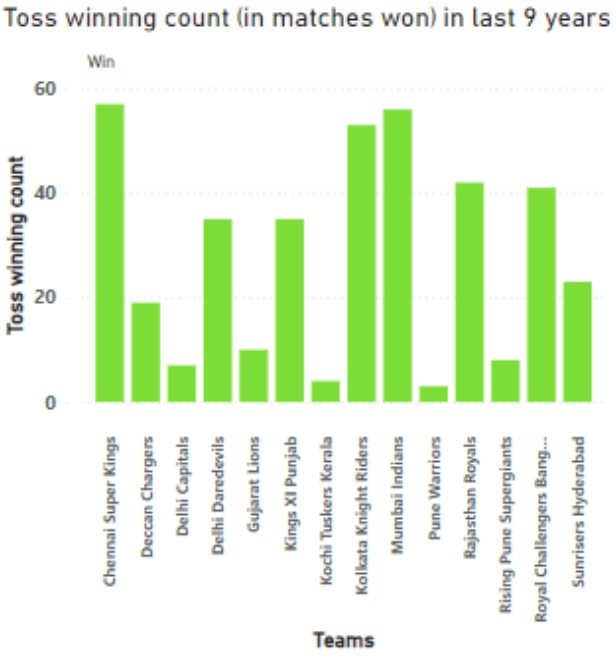
After training the model, we test it to see if it learned well. We use the testing data to see how accurate our model is in predicting results.

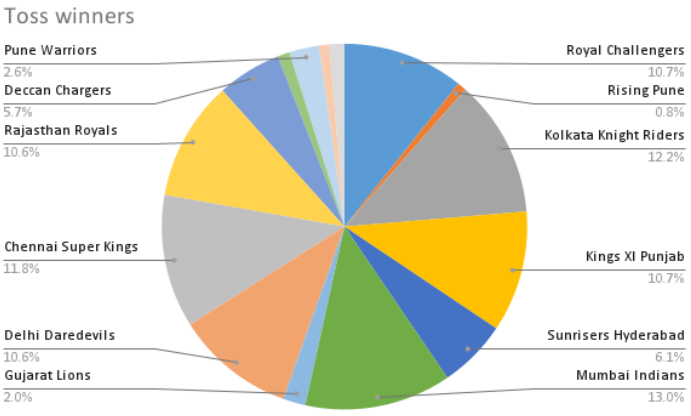
It's common to split the testing and training data with a ratio of 20:80. This split helps in ensuring that the model learns well from a variety of examples without relying too much on just one set of data.

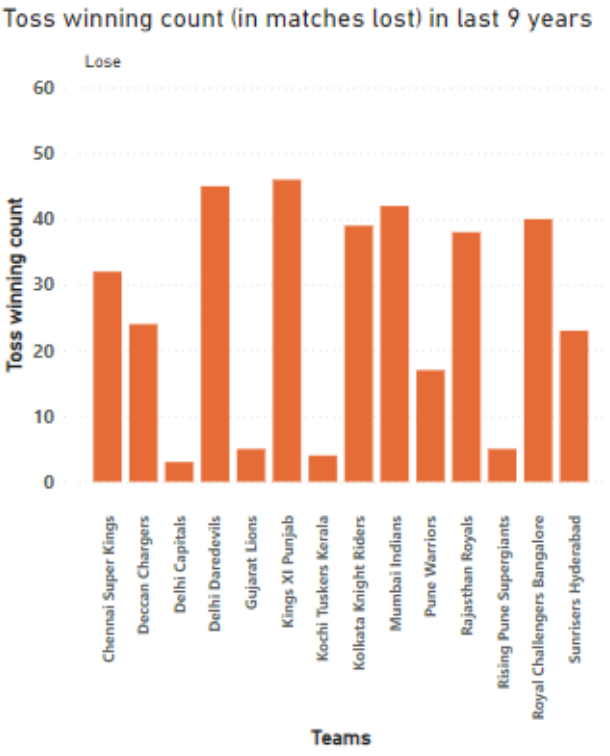
D. Training and testing the model: Training is the most important stage in

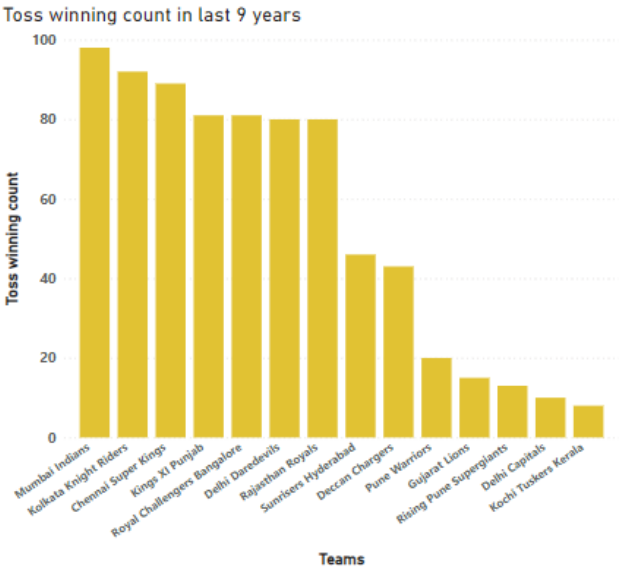
Machine Learning. In this step the model is trained using training data to find patterns and make predictions. It results in the model learning from the dataset so that it can accomplish the given task. We used SVC with RBF, Sigmoid and polynomial kernels, Naïve Bayes (Gaussian NB, Bernaulli NB and Multinomial NB), Logistic Regression, KNN Classifier, Decision Tree, Random Forest and Adaptive Boosting algorithms to train the model.

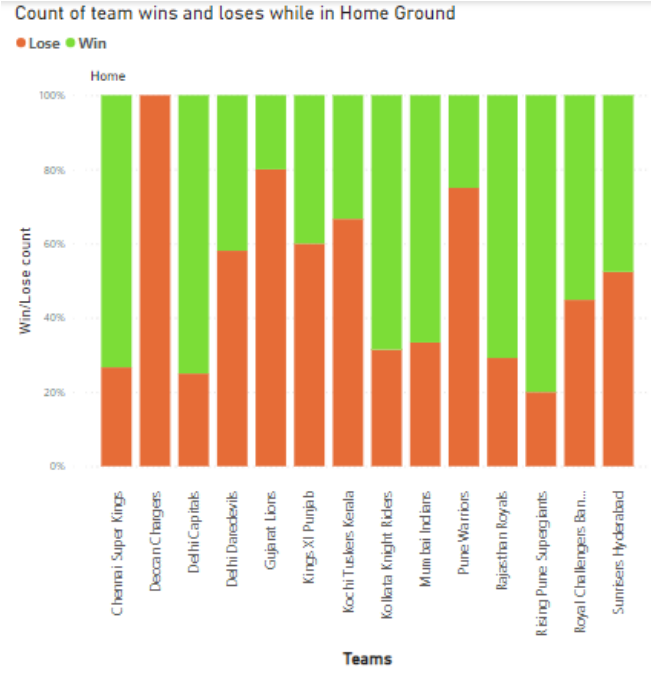
RESULTS

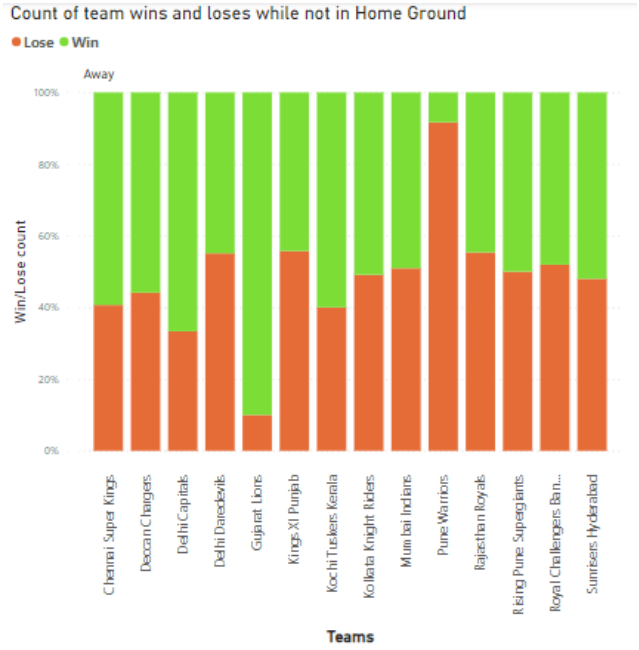












Machine Learning Models used and their Accuracy Scores:

Sl. No.	Machine Learning Model	Accuracy Score
1.	Support Vector Classifier (rbf kernel)	0.6250
2.	Support Vector Classifier (polynomial kernel)	0.6184
3.	Support Vector Classifier (sigmoid kernel)	0.6184
4.	Logistic Regression	0.6184
5.	K Neighbors Classifier	0.5526
6.	Naive Bayes (Gaussian NB)	0.5789
7.	Naive Bayes (Bernaulli NB)	0.6250
8.	Naive Bayes (Multinomial NB)	0.5723
9.	Decision Tree Classifier (criterion = 'entropy')	0.5131
10.	Decision Tree Classifier (criterion = 'gini')	0.5131
11.	Random Forest Classifier	0.5263
12.	AdaBoosting Algorithm (Decision Tree)	0.6052
13.	AdaBoosting Algorithm (Random Forest)	0.6250
14.	AdaBoosting Algorithm (Multinomial Naive Bayce)	0.5855
15.	AdaBoosting Algorithm (Bernaulli Naive Bayce)	0.4342
16.	AdaBoosting Algorithm (Gaussian Naive Bayce)	0.4342
17.	AdaBoosting Algorithm (Logistic Regression)	0.5855

CONCLUSION

Making predictions about who will win in cricket, like in the Indian Premier League (IPL), is really tough. It involves lots of complicated stuff. But now, thanks to machine learning, it can become simpler. In this paper, we found out different things that affect IPL matches. Big things like which teams are playing, where the match is happening, which city it's in, who wins the coin toss, and what they decide to do after winning the toss, all play a huge role in who wins.

We looked at lots of IPL data and used it to guess who would win based on how players perform. We used different methods like Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest Classifier, Naïve Bayes and K-nearest Neighborhood. We also used Adaptive Boosting on all of the above models to improve accuracy. Out of all these, **Support Vector Classifier (rbf kernel), AdaBoosting Algorithm (Random Forest),** and **Naive Bayes (Bernaulli NB)** did the best job at predicting the winners each with an accuracy score of **0.625**

Predicting cricket match outcomes is complicated, but machine learning helps simplify the process. By analyzing various factors like teams, venue, toss winner, and player performance, we can improve our predictions. Among the methods used, the **Naive Bayes (Bernaulli NB)** was chosen as the best model.

REFERENCES

[1] Srikantaiah KC, Khetan A, Kumar B, Tolani D, Patel H. Prediction of IPL match outcome using machine learning techniques. In 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021) 2021 Sep 13 (pp. 399-406). Atlantis Press.

[2] Kalpdrum Passi and Nirav Kumar Pandey, predicting player's performance in one day international cricket matches using machine learning, Volume 8, Computer Science & Information Technology, 2018.

[3] Haseeb Ahmad, Ali Daud, Licheng Wang, Haibo Hong, Hussain Dawood and Yixian Yang, Prediction of Rising Stars in the Game of Cricket, IEEE Access, Volume 5, PP. 4104 – 4124, 14 March 2017.

[4] Balasundaram A, Ashokkkumar S, Jayashree D, Magesh Kumar S, Data Mining based Classification of Players in Game of Cricket, proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020), IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9.