CSE 343/ECE 343 : Machine Learning Project Proposal
Title : The Dream Team Creator

Anurag Yadav
anurag19150@iiitd.ac.in
Agamdeep Bains
agamdeep19009@iiitd.ac.in

Ajay Kumar
Ajay19293@iiitd.ac.in
Yash Aggarwal
yash19219@iiitd.ac.in

Abstract

Player selection is one of the most important tasks for any sport. According to a recent report, the gross revenue of fantasy sports operators stood at Rs. ~2,400 crores (US$ 340.47 million) for FY20 compared with Rs. ~920 crores (US$ 131.64 million) in FY19 ~3X YoY increase. The market has witnessed a 700% increase in the past decade in the number of fantasy sports operators and a 2,500% spike in the number of fantasy sports users. The performance of the players depends on various factors such as the opposition team, the venue, his current form, etc. The team management, the coach, and the captain select eleven players for each match from a squad of 15 to 20 players. They analyze different characteristics and the statistics of the players to select the best playing 11 for each match. The main focus of this project is to create a winning fantasy team by looking at the huge amount of data present online and predicting which team will win by analyzing player's performance history, statistics against the opponent team, home ground status, weather impact, and many other factors and attempt to predict the outcome

1. Introduction

We plan to use previous years' tournament data to train a machine learning model and then use the model in order to predict the outcomes of matches. We plan to identify several factors which could affect the results of matches and compare various Machine Learning models using our features based on metrics like Accuracy, Precision, Recall and F1 score for our problem.

2. Literature Survey

During the past, several researchers have contributed their efforts towards result prediction and the formation of an optimal team that will most likely win in a particular sport:

1) Players Performance Prediction in ODI Cricket

This paper proposes a model to predict a cricket player's upcoming match performance using machine learning algorithms like recursive feature elimination and univariate selection, linear regression, support vector machine with the linear and polynomial kernel, etc. The authors have collected data of the Bangladesh National Cricket team from trusted sports websites. Then this data is processed into numerical values to implement the algorithms mentioned above. Feature selection algorithms are applied for extracting the attributes that are more related to the output feature. This model tries to predict the runs scored by a batsman and runs conceded by a bowler in the upcoming match.

2) A Machine Learning Application for Football Player's Selection

This paper proposes a model that divides player selection criteria into four key areas: player's technique, the player's speed, the player's physical status, and the player's resistance using neural network technique to determine these significant attributes for each player. Every player will be judged based on the features mentioned above. There are some attributes that a football player may have which cannot be neglected when it comes to choosing a rightful player for a football team. This system has employed the idea of a neural network in considering this large amount of attributes needed in selecting the appropriate player for a football team.

3) Machine learning-based Selection of Optimal sports Team based on the Players Performance

This paper provides a model that can select the best playing 11 in the Indian Cricket team. There are

various factors on which a player's performance depends, like pitch type, the opposition team, the ground, etc. The model contains data from the One Day International of the past several years of team India, and this dataset is created using data from trusted sites like espn.com. The proposed model gives complete information about the batting, bowling, and fielding skills of a player. The player performance is classified into several classes, and a random forest classifier is used to predict the player's performance. The proposed work can address the issue of selecting the optimal team in cricket without any prejudice and give equal importance to all-rounders.

3. Dataset Features

We have picked two kind of data set one contains the data of all the ipl matches played in between 2008 to 2020 some of its features are:

1. id
2. city
3. date
4. player_of_match venue
5. neutral_venue
6. team1
7. team2
8. toss_winner
9. toss_decision
10. winner
11. result
12. result_margin
13. eliminator
14. method
15. umpire1
16. umpire2

With the help of this data we have calculated the toss won by teams, which venue is best for which team and other data has been collected from this data set.
This data consists of 817 rows and 17 columns

we have used another large data set that consist of 18 columns and 193469 columns
This data has details of what happened on that particular ball i.e. this is ball by ball data
features of this data are:

1. id
2. inning
3. over
4. ball
5. batsman
6. non_striker
7. bowler
8. batsman_runs
9. extra_runs
10. total_runs
11. non_boundary

12. is_wicket
13. dismissal_kind
14. player_dismissed
15. fielder   extras_type
16. batting_team
17. bowling_team

we have used this data for several purposes

1. Total_Runs
2. Total_Boundries
3. Striker_Rate
4. Average
5. Wickets
6. Economy
7. Wickets_over
8. Catches
9. Stumped
10. Balls_Faced
11. Matched_played
12. Run_Given
13. ball delivered

all this data about every single has been stored with help of this data set
This is total runs scored by some of the players:-
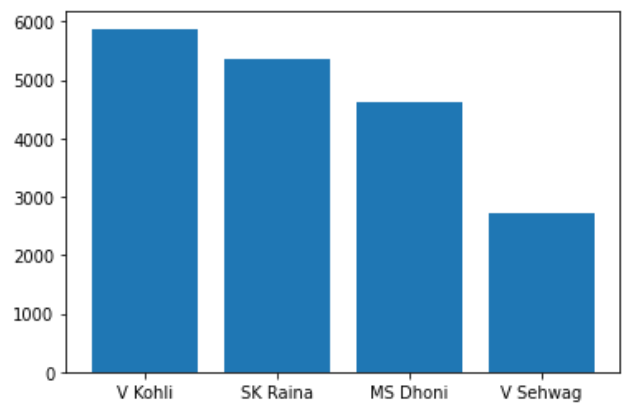
This is the total no. of matches played in the IPL every year between 2008 and 2020

This is total no. of matches played by each team from year 2008 to 2020

This is total no. of matches won by each team from year 2008 to 2020

This is the result of all IPL matches from 2008 to 2020

This is the strike rate of some of the players

PREPROCESSING AND FEATURES EXTRACTION

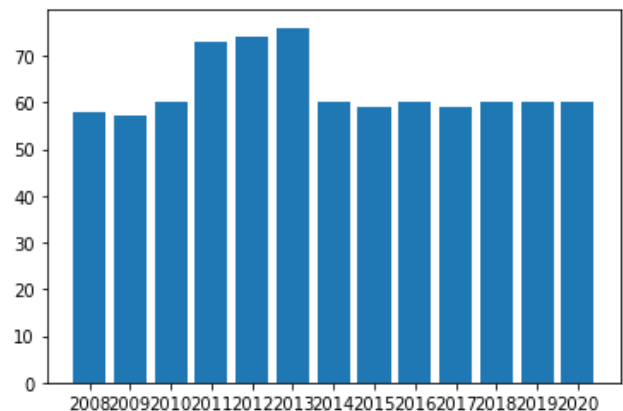### 3.1 for strike rate, economy, wickets per over

we have used the data from the ball by ball and calculated various stuff like total run scored by a batsman, total wickets were taken total ball faced and by using this we have calculated the data for
1.strike rate=total runs/total balls faced
2.economy= total run given/total ball delivered
3.wickets per over=total wickets/total overs

3.2 Reduction of dimensions of the dataset

We had some data which we didn't required so we have removed those columns as the part of data preprocessing

4. Methodology, model details

We are interested in using the Naive Bayesian model using the parameters match venue (stadium name) and what a team does in the first inning (batting/fielding). Although we do not plan to use this as a baseline model, but rather as a yardstick to measure the usefulness of our final model, we decided to use it to augment our current model in the future

We also used logistic regression (batch gradient descent) using the following parameters:
- Average (Runs per match)
- Balls Faced
- Catches
- Economy
- Runs Given
- Strike Rate
- Players Stumped
- Total Boundaries
- Total Runs
- Wickets
- Wickets Per Over

In order to create the dataset, we first find the average values of these parameters for all the players in a team. Then, we find the difference between these average values between the two teams. We then perform BGD on the resultant dataset.

5. Results and analysis

Using the Naive Bayesian model, we were able to obtain the following metrics on the entire dataset:
Accuracy: 0.6007371007371007
Precision: 0.6149425287356322
Recall: 0.7213483146067415
F1: 0.6639089968976214

When using the Logistic Regression model, the following metrics were obtained on the testing dataset with 1000 epochs, learning rate 1.0 and no regularization:
Accuracy: 0.6380368098159509
Precision: 0.6923076923076923
Recall: 0.6702127659574468
F1: 0.681081081081081

6. Work left and Learnings

Work Left
The project progress has been on schedule and a lot of work has been done

However, we wish to improve upon the base model by adjusting hyperparameter values (learning rate) and adding regularization

We also plan to use both the Naive Bayes and the Logistic Regression models in conjunction to achieve better results

We will also use the data collected and analysed using the models in order to create a "Dream Team" which would have the highest likelihood of winning against all currently existing teams

Learnings

We learnt about the nitty-gritties of logistic regression and Naive Bayes just as we did in Assignment 1. However, a new challenge we faced was the collection of data from trustworthy sources and reading material regarding our subject. We also learnt that not all the data that we wish to find is always available. Hence, we often have to compromise and make do with what we have. This is what we tried to do while preprocessing the data. We also learnt how to extract useful information from raw datasets and to process it in order to use it in the best way possible. Also, as we worked in groups, we learnt about how people work on ML projects in the industry

Work done by every member of the team
1) **Anurag Yadav**: Dataset Preprocessing- Data Extraction, Features extraction, Data collecting, Data analysis, Noise Reduction,Report writing, helped in the logistic regression model

2) **Yash Aggarwal**: Data Collection, Feature extraction, Report Writing, Data Preprocessing, helped in Naive Bayesian model and metric Collection

3) **Agamdeep Bains:** Data collection, Feature extraction, Coding Logistic Regression model and Naive Bayesian model, Metric Collection, Report Writing

4) **Ajay Kumar:** Naive Bayesian model, Report Writing, data collection

7. References

[1] Aminul Anik "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms" BRAC University, Dhaka, Bangladesh, 4th International Conference 2018 on Electrical Engineering and Information and Communication Technology

[2] Predicting Fantasy Football Performance with Machine Learning Technique by Nitin Kapania**.**

[3] Fantasy Football Trade analyzer by Jim Kim.

[4] Predicting Players' Performance in the Game of Cricket Using Machine Learning