Maximum Likelihood Estimation for Statistical modelling of winning probability in Cricket

Amrita Sundari V[1]

***Abstract*— Statistical analysis and modelling in sports are of high interest in the recent decades and predicting the winning probability of a team became the fundamental problem for such modelling. Cricket, being one of the most popular team sports in the world, it is very important to understand the factors that affect the game's outcome. Current prediction techniques use run rate as the only influential factor to predict the score in each innings of a game. In contrast to that, the outcome of a cricket match are influenced by many factors such as Toss, Innings, Run rate, Target and many dynamic factors which vary during the course of the game. Conclusively, I propose a model whose parameters are determined by the maximum likelihood estimator, and predict the outcome of the game using the model.**

Keywords: **Statistical model, Cricket, Maximum Likelihood estimator ,winning probability.**

I. INTRODUCTION

Cricket is a bat and ball team game between two teams each with eleven players. It is said to be originated from England and is the second most popular game in the world, the first being soccer. The objective of each team is to score more than the opponent team to win the game. In the first inning, one team bats with the objective of scoring as many runs as possible while the other team bowls and fields with an objective to dismiss the batsmen and thus limit the runs scored by the batting team. In the second inning, the role of the teams are interchanged. The teams play for one or two innings based on the type of match.

The team winning the coin-flipping toss decides which team bats and fields. There are various internationally accepted forms of the game: *Twenty20* in which the teams play one innings each with 20 overs, 6 deliveries in each over,*Test match* which is played for 5 days, the teams playing two innings each and *One Day International(ODI)* which is scheduled to be completed in a day or a day/night combination.

In this project, I consider the games played in one of the finest Twenty20 competition in the world of cricket, *Indian Premier League(IPL)*. The popularity of Twenty20 cricket has once more increased in 2008 by the creation of IPL, a professional league for Twenty20 cricket competition in India. The league, which is based on a round-robin and knockout format, has teams in major Indian cities. 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. Meanwhile, the third and fourth place teams from league phase play against each other in an eliminator match and the winner from that match will play the loser from the first Qualifying match. The winner of the second Qualifying match will move onto the final to play the winner of the first Qualifying match in the IPL Final match, where the winner will be crowned the Indian Premier League champions.

Over the years, researchers have successfully applied various statistical methods to cricket data for various application. An early attempt of modeling cricket batsmen data can be found in Elderton and Wood (1945)[1]. Fernando et al. (2013)[3] applied a regression-based method to address the home-field advantage in ODI games. In the recent years, Pathak and Wadhwa(2016)[4], applied modern classification algorithms like Naive Bayes, Support Vector Machines and KNNs to predict the outcome of ODI cricket. Jayalath(2018)[5] worked on quantifying different factors affecting ODI cricket match using graphical classification and regression tree(CART). For this project, the IPL data is taken from the website *www.cricsheet.org* for all the matches from 2008-2016. Using this data, I have extracted out different features and used maximum likelihood estimation to estimate the statistical model parameters that effectively fits the data.

The rest of the report is organized as follows: the next section gives the problem formulation and a brief description of the statistical model used in this project . Section III describes about the data set , section IV gives us a brief overview of the feature engineering, Section V describes the modelling and interpretation of the results and finally in Section VI the report is finished with a conclusion and future work.

II. PROBLEM FORMULATION

A. Modelling Conditional Probabilities

One of the many ways to predict the outcome of a cricket game is to model the conditional probability of winning given the influential factors of each of the game. Let us consider the target variable Y of each game to take two values 1 and 0, representing that the team under consideration won the game and did not win the game respectively. Assuming that X represents the features, our objective is to estimate the conditional probability $Pr(Y = 1/X)$. Let's assume that the above probability can be considered as a function p parameterized by θ, $P(Y = 1/X) = p(x; \theta)$ and

[1]Amrita Sundari V is with Department of Electrical and Computer Engineering, Texas AM University, College Station, TX- 77840 amrita95@tamu.edu

further assume that the observations are independent of each other.

The likelihood function is given by

$$\prod_{i=1}^{n} Pr(Y = y_i/X = x_i) = \prod_{i=1}^{n} p(x_i;\theta)^{y_i}(1 - p(x_i;\theta)^{1-y_i})$$

In this kind of model,if the probability function $p(x;\theta)$ is continuous, the value of the probability p_i that is obtained from the function will be similar to similar x_i.

B. Logistic Regression

The next step is to model the function $p(x)$. It cannot be modelled as a linear function since it is unbounded and does not consider the fact that the values of probability should be between 0 and 1. In order to overcome this shortcoming, $log(p(x)/(1 - p(x))$ is modeled as a linear function. This will ensure that the value of p(x) lies between the range. This modification is known as the *logistic* or *logit transformation*.

This alternative model is called the **Logistic Regression**. The model is given by:

$$log\frac{p(x)}{1 - p(x)} = \beta_0 + \beta.X$$

Solving this equation for $p(x)$, we get

$$p(x;b,w) = \frac{e^{\beta_0+\beta.X}}{1 + e^{\beta_0+\beta.X}}$$

This implies $p(x) > 0.5$ if $\beta_0 + \beta.X$ is positive and $p(x) < 0.5$ if $\beta_0 + \beta.X$ is negative. So Logistic regression gives a linear classifier. This value $p(x)$ can then be used to categorize whether the team under consideration will win the game or not. Therefore we choose this model to predict quantitative variables using Linear regression.

C. Maximum Likelihood Estimation

Maximum Likelihood Estimation is a method of estimating the parameters of a statistical model given observations. This method aims to maximize the likelihood function given the observation and the estimate is known as the *Maximum Likelihood Estimate(MLE)*.

For each observation, we have a vector of features, x_i, and a corresponding target variable y_i whose conditional probability we are trying to model. We can fit the model using likelihood function since logistic regression effectively estimates the probability.

The likelihood function is given by

$$L(\beta_0,\beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

The MLE can be found by maximizing the above function.The log-likelihood turns products into sums,

$$l(\beta_0,\beta) = \sum_{i=1}^{n} y_i log p(x_i) + (1 - y_i) log(1 - p(x_i))$$

$$= \sum_{i=1}^{n} log(1 - p(x_i)) + \sum_{i=1}^{n} y_i log\frac{p(x_i)}{1 - p(x_i)}$$

$$= \sum_{i=1}^{n} log(1 - p(x_i)) + \sum_{i=1}^{n} y_i(\beta_0 + \beta.X)$$

$$= \sum_{i=1}^{n} -log(1 + e^{(\beta_0+\beta.X)}) + \sum_{i=1}^{n} y_i(\beta_0 + \beta.X)$$

Therefore, in order to find the MLE we find the derivative of the above equation and set it to zero and solve for β. Let's take a derivative of the log likelihood function with respect to one of the parameters, say β_j,

$$\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{1}{1 + e^{\beta_0+\beta.X}} e^{\beta_0+\beta.X} x_{ij} + \sum_{i=1}^{n} y_i x_{ij}$$

$$= \sum_{i=1}^{n} (y_i - p(x_i;\beta_0,\beta))x_{ij}$$

Since we cannot set this equation to zero and directly solve it, we go for numerical methods like gradient descent, to solve for the parameters.

III. DATA DESCRIPTION

The data set is taken from a Kaggle Challenge which provides two csv files *deliveries.csv* and *match.csv* respectively. Originally this data set was taken from *cricsheet.org* in YAML format and have been converted into csv files. The data set provides the ball by ball data of all Indian Premier League Cricket matches between 2008 and 2016. matches.csv contains details related to the match such as location, contesting teams, umpires, results, etc.deliveries.csv is the ball-by-ball data of all the IPL matches including data of the batting team, batsman, bowler, non-striker, runs scored, etc. In total, the data set consists of 636 games that was played for 9 seasons of IPL.

For this project, I consider the matches that happened in 2010, the 3rd season of the IPL. Also, I consider only the matches that had a definite winner in the end of the game i.e the games which were not cancelled due to rain or any external factors. This consisted of around 60 matches in total. I train the model using all the matches except for the for Final match and test the model on the Final match and predict the probability of the team *Chennai Super Kings* against *Royal Challengers Bangalore*. The proposed method can be extended any type of cricket which is played for two innings.

IV. FEATURE ENGINEERING

The task of predicting the probabilities is modeled using Logistic Regression. In previous works, the run rate of the batting team and the number of players dismissed were the only feature set that have been considered[2]. In this project, I propose 10 different features that are directly influential on the outcome of any cricket game irrespective of the teams. From the data set, the features that are identified and

extracted. The features that are included in this project is given in the I .

Feature_id	Feature name
1	Cumulative Runs
2	No. of Wickets
3	Total runs of the innings
4	Total wickets
5	Run Target
6	Remaining target
7	Run Rate
8	Required Run rate
9	Run rate difference
10	Batting or Bowling

TABLE I
FEATURE SET

In order extract these features, the data from the two csv files are stored in data frame data type of Pandas package in Python. Then, these data frames are merged into one, using the common column of the two data frames namely the match_id. A description of all the feature sets are briefly discussed in the following section. Each of these feature is stored as columns of a separate data frame named *train_df*.

A. Cumulative runs

Cumulative runs indicate the sum of runs scored by each team at the end of every over when they are batting. This feature is stored in the column innings_score of the new data frame *train*

B. Number of wickets

The number of wickets indicates the number of players dismissed by the end of each over in total. This is one of the important feature since it indicates the number of resources each team still has at the end of the over.

C. Total runs of the innings

This feature indicates the total runs scored by both the teams at the end of the each innings. This will serve as a contributing factor since it evaluate the performance of the batting team in each innings.

D. Total wickets

Total wickets indicates the number of players dismissed during the end of each innings. Like mentioned before, this feature will be indicative of the performance of the bowling team in each innings.

E. Run target

The Run target is the runs that have to be scored by the team batting in the second innings in order to win the game.

F. Remaining target

Remaining target is the difference between the Run target and the runs that have been scored at the end of each over.

G. Run rate

Run rate is the ratio of the total runs scored to the total overs. This has been an important factor that has been considered in various previous researches also.

H. Required run rate

Similar to the previous feature, the required run rate is the ratio of the required runs in order to reach the target score to the remaining overs for the batting team.

I. Run rate difference

The difference between the required rate and the current rate of the batting team in the second innings in order to win the game.

J. Batting or Bowling

This is a binary feature, indicating if the team under consideration is batting or bowling.

V. Modelling and Results

Of the total 60 matches that were played in the season 3 of IPL in 2010, 59 matches are considered for training and the final match of the season is used for testing the trained model. The final match was played by Chennai Super Kings(CSK) and Royal Challengers Bangalore(RCB).

The target variable for over-by-over data points is set to be 1 for the winning team and 0 for the losing team. The probability function which is parameterized by β is assumed to be a linear model on logit transformation. The parameters of the model are estimated using Maximum Likelihood Estimation and Gradient Descent algorithm.

In order to analyze the different features and the importance of each of feature that was taken into consideration, I construct a bar plot of the absolute values of the estimated parameters with respect to the feature.

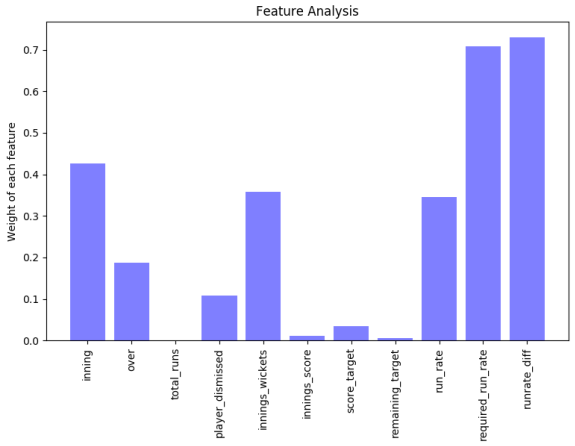

Fig. 1. Feature Importance

As we can see from Fig 1, **the required run rate** and **the run rate difference** is of prime importance among all

the other features. Also, we can clearly see that there are more number of features which are of comparable importance with the previously used like features run_rate and the wickets. The features like *total score, target_score and remaining_target* has a lower value of weight when compared to the other features. This indicates that the total score and the target score does not matter if the run rate per over is sufficiently high enough to match the required run rate.

The Table II shows the maximum likelihood estimates of the parameters and the corresponding feature.

Parameter	Feature	Estimate
β_0	Cumulative Runs	0.42
β_1	No. of Wickets	0.18
β_2	Total runs of the innings	0.005
β_3	Total wickets	0.10
β_4	Run Target	0.35
β_5	Remaining target	0.011
β_6	Run Rate	0.035
β_7	Required Run rate	0.005
β_8	Run rate difference	0.34
β_9	Batting or Bowling	0.70

TABLE II

FEATURE SET

Using the above estimates of the model parameters, the win prediction probability of CSK in the Final match of the 2010 season of IPL is computed for every over. The winning probability of any cricket game in the first innings depends on the training data, but in reality it is fully unpredictable. There is no meaning in predicting the winning probability of any team in the first innings of the game.

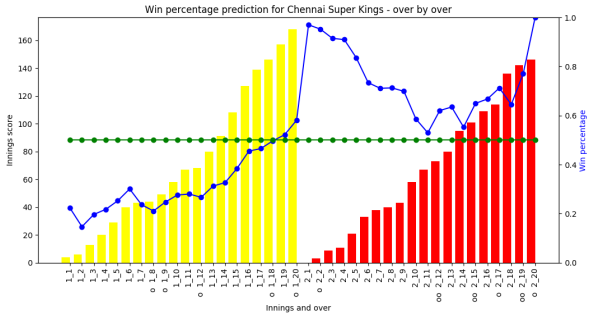

Fig. 2. Win probability of Chennai Super Kings- over by over

The Fig 2 represents the win prediction of CSK match when it played against RCB in 2010. The yellow bars represents the cumulative runs scored by CSK team in each over in first innings and similarly the red bars represents the cumulative runs scored by RCB in the second innings. The blue line connecting the dots represents the probability of CSK winning the Finals and the green line represents the 0.5 probability line. The x-axis indicates the innings and the number of the over under consideration in the format $innings\_over$. The small circles 'o' in the x-axis indicates the number 1

From the Fig 2, we can see that the probability of winning of CSK in the second innings is very high in the first few overs of the second innings. This is due to fact that the first over of the second innings, was a maiden over, which implies RCB did not score any runs in the first over. As the match progresses in the second innings, we can see that the probability of CSK winning reduces, since the run rate increased gradually from the sixth over of the second innings.

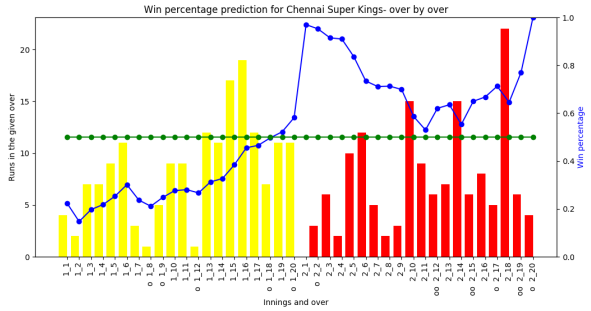

Fig. 3. Win probability of Chennai Super Kings- over by over

Fig 3 gives us a better picture of the changes in probability of winning by CSK. In this plot the yellow and red bars indicate the runs scored in each over by CSK and RCB respectively. The probability of CSK winning drops steeply whenever RCB scores a considerate amount of runs in the over. This is because, the run rate of RCB will be high in the particular over which in turn affects the winning probability of CSK. This can be seen very clearly in the 18th over of second innings where RCB scored over 20 runs in one over and the CSK winning probability reduces a bit. When the runs scored in consequent overs remains constant, the probability doesn't exhibit any change.

One of the important factor to notice is that, the probability of CSK winning rises everytime when a player gets dismissed. For example, in 12th over of the second innings, two of the RCB players are dismissed. This indicates that there are less resources for RCB to win the match and a higher chance for CSK to win the match.

From Fig 2 and 3 we can see that, the cumulative scores of the teams is responsible for only a small change in the probability of CSK winning. This can be seen in 17th over of second innings, even though the cumulative runs scored by RCB rises to a consdirable extent, the fall in probability is very less. This is in accordance with the findings from the Fig 1 that the scores did not have much importance in our current task of predicting the winning probability of a team.

VI. CONCLUSION

In this project,one of the popular statistical modelling techniques, Logistic regression was discussed along with the Maximum Likelihood Estimation(MLE) of the parameters of the model. Subsequently, the features that are influential in the prediction of probability of winning the game of cricket in Indian Premier League(IPL) were investigated and described. Using the features extracted, the parameters

of the conditional probability of the team winning,a linear model on logistic transformation, were estimated using MLE. Following that, the importance of each such feature was analysed using the help of the estimated weights. We found out that the features with utmost importance were the run rate at each over and the number of players dismissed in total at the end of each over. Intuitively, these two features determines the resources a team has in order to win the game, which is an important factor in predicting the outcome of a match. The model was then tested on a game between Chennai super kings and Royal challengers Bangalore, to predict the winning probability of the CSK.

Even though, predicting the outcome of a game is very challenging, we can consider similar features into account to approximately predict the probability of a team winning. Future analysis could potentially include the game history of each team. Clearly, the team names that are playing, the strike rate of the batsman, the performance of the bowler and even the venue of the game are few of the most influential factors to be considered when we build a model team specific. The model proposed in this project is a generic model which can be applied to predict the probability of winning, irrespective of the teams.

REFERENCES

[1] William Elderton and George H Wood. "Cricket scores and some skew correlation distributions:(an arithmetical study)". In: *Journal of the Royal Statistical Society* 108.1/2 (1945), pp. 1–11.

[2] Viraj Phanse and Sourabh Deorah. "Evaluation and Extension to the Duckworth Lewis Method: A Dual Application of Data Mining Techniques". In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE. 2011, pp. 763–770.

[3] Mahesh Fernando, Ananda Manage, and Stephen Scariano. "Is the home-field advantage in limited overs one-day international cricket only for day matches?" In: *South African Statistical Journal* 47.1 (2013), pp. 1–13.

[4] Neeraj Pathak and Hardik Wadhwa. "Applications of modern classification techniques to predict the outcome of ODI cricket". In: *Procedia Computer Science* 87 (2016), pp. 55–60.

[5] Kalanka P Jayalath. "A machine learning approach to analyze ODI cricket predictors". In: *Journal of Sports Analytics* Preprint (), pp. 1–12.