



**UML501
MACHINE LEARNING**

PROJECT REPORT

ON

CRICKET ANALYSIS

SUBMITTED BY:

ISHAN SHARMA	101783059
PRABHJOT SINGH	101603232
PINAAK GOYAL	101603221

SUBMITTED TO:

DR. SINGARA SINGH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SEMESTER-V

BATCH 2016-2020

B.E. Computer Engineering

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY

ABSTRACT:


As one-day international (ODI) games rise in popularity, it is important to understand the possible predictors that affect the game outcome. The home-field advantage, coin-toss result, bat-first or second, and day vs day-night game format are such popular variables being considered in the cricket literature. This project focuses on a comprehensive study of quantifying the significance of those important predictors via the popular logistic regression and other model approaches. This study reveals the importance of the home-field advantage for major cricket playing nations in one-day international games but questions the uniformity of such factors under different playing conditions. Importantly, the home-field advantage is investigated further based on the opponent's geographical location. Conclusively, the CART approach provides interesting and novel interpretations for popular predictors in ODI games.

1. Introduction

Cricket has become one of the world's most popular outdoor sports. The International Cricket Council (ICC) identified 106 cricket playing nations which 10 of them are full members, 37 of them are associates, and the remaining 59 are affiliate members. One day international (ODI) is one of the three main types of cricket matches and is considered the most popular. One reason for its popularity is due to recent advances in technology which allows for a day-night format as opposed to the classical day-only form. In addition, umpire decisions and their review systems, more strict bowling rules, and options of power plays (Silva et al., 2015.) have turned it into a more offensive game. These changes have made 21st-century cricket more competitive and spectacular than ever before.

There are numerous factors that can affect a cricket game's outcome. Like in many other team sports, the home-field advantage is believed to be a critically important factor in cricket. It can be seen that the home-field effect is well researched in sport statistics literature. For instance, Clarke and Norman (1995) discuss the home-field effect in English soccer leagues, Pollard and Gómez (2014) discuss the home-field effect of men and women soccer leagues in Europe, and Karlis and Ntzoufras (1998) discuss a predictive modeling method to predict soccer outcomes using many factors including the home-field effect. Harville and Smith (1994) discuss the home court advantage in college basketball games. Levernier and Barilla (2007) discuss the home-field advantage in major league baseball using logit models, Vaz et al. (2012) discuss the effect of alternating home and away field advantage in rugby championship, and recently Ribeiro et al. (2016) discuss the microscopic, team-specific, and evolving features of home NBA games.

Cricket is increasingly popular among the statistical science community, but the unpredictable and inconsistent natures of this game make it challenging to apply in common probability models. However, numerous researchers successfully applied various statistical methods to cricket data. An early attempt of modeling cricket batsmen data can be found in Elderton and



Wood (1945). Interestingly, the home-field advantage in ODI cricket matches was discussed by De Silva and Swartz (1998) and they found that the effect is a significant advantage for the home team. They also addressed a continent based advantage when a game is played in a neutral venue. Fernando et al. (2013) applied a logistic regression method to address the home-field advantage in ODI games. The logistic regression modeling is a commonly used powerful and easily interpretable modeling technique. However, it may be questionable that the factors such as home-field advantage and the form of the game are uniformly influential regardless of the other factors such as coin-toss result and the factors associated with the opponent team. Fitting separate models for different forms may apprehend the use of remaining predictors. This question is further discussed via a machine learning based alternative modeling technique and the findings were compared to the results from logistic regression models.

1.1 Need of the System

Data science helps us to extract knowledge or insights from data- either structured or unstructured- by using scientific methods like mathematical or statistical models. In the last two decades, it has been one of the most popular fields with the rise of all big data technologies. A lot of companies have been using recommendation engines to promote their products/suggestions in accordance with users' interests such as Amazon, Netflix, Google Play. A lot of other applications like image recognition, gaming, or Airline route planning also involves the usage of big data and data science.

Sports is another field which is using data science extensively to improve strategies and predicting match outcomes. Cricket is a sport where machine learning has scope to dive into quite a large outfield. It can go a long way towards suggesting optimal strategies for a team to win a match or a franchise to bid a valuable player.

1.2 Applications of Proposed System

1. In proper analysis of winning record of any team and scope of improvement.
2. Lot of benefits to sports journalism.
3. Duckworth and Lewis techniques can be improved using it.

1.3 Challenges in development

We have created our project in python and while doing so there were some problems which we had to encounter. Mainly we had to learn about logistic regression model in detail, along with data preprocessing techniques, also collection of dataset was a very tedious process. We learnt to use them properly in our projects, about their working. Syntaxes etc. were also unknown regarding them and to create the project we had to learn about it.

2.0 Existing Work

In literature, Duckworth and Lewis proposed a solution, called the D/L method [1], to reset targets in rain interrupted matches which was adopted by the International Cricket Council (ICC) in 1998. Further, the use of Duckworth-Lewis resources to assess player's performances has been studied. The methods of graphical representation to compare players are presented considers the strength of opponent team, along with other factors, in modeling the performance of batsmen and bowlers. However, like in any sport, winning is the ultimate goal in cricket. It takes into account various factors affecting the game including home team advantage, day/night effect and toss, etc. They take into account both historical data as well as instantaneous state of a match while the game is still in progress. It studied the role of multiple factors including home field advantage, toss, match type (day or day and night), competing teams, venue familiarity, and season, etc., and applied Support Vector Machines(SVM) and Naive Bayes Classifiers for predicting the winner of a match. In literature, Duckworth and Lewis proposed a solution, called the D/L method, to reset targets in rain interrupted matches which was adopted by the International Cricket Council (ICC) in 1998. Further, the use of Duckworth-Lewis resources to assess player's performances has been studied. They take into account both historical data as well as instantaneous state of a match while the game is still in progress. They studied the role of multiple factors including home field advantage, toss, match type (day or day and night), competing teams, venue familiarity, and season, etc., and applied Support Vector Machines(SVM) and Naive Bayes Classifiers for predicting the winner of a match.

3.0 Working of the Proposed System

We have built cricket analysis software as our project. In this we accessed the match held on various aspects such as ground type, current ranking as many more.

Its working is basically like this:

First, we collected about 800 entries of data about matches where they were held and who won. We created a model based on eight features as mainly teams who played, place where match was held, type of pitch, team's strengths and their current ranking.

We used label encoder to find out to make numbers from strings. We used one hot encoder to introduce dummy variables. We created a logistic regression model to determine the winning team.

The user is to fill in details of match he wishes to check result for, and our system checks the validation of the input and prints the probability that the team will win match.

We also tested our model over other classification models but it didn't show promising results, the code is as follows:

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```

# Importing the dataset
dataset = pd.read_csv('Final1.csv')
dataset1 = pd.read_csv('originalDataset3.csv')

# Encoding categorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder
country_map = {country:i for i, country in
enumerate(dataset1.stack().unique())}

dataset['Team 1'] = dataset1['Team 1'].map(country_map)
dataset['Team 2'] = dataset1['Team 2'].map(country_map)
dataset['Winner']=dataset1['Winner'].map(country_map)

X = dataset.iloc[:, 0:8].values
y = dataset.iloc[:, 10].values

#Taking User Input
print("Enter Team 1:")
country1_name=input()
print("Enter Team 2:")
country2_name=input()
print("Enter the venue of ODI Match:")
Ground_name=input()

#Making dictionary
dict={}
dict['India']="Bat"
dict['Pakistan']="Bowl"
dict['Australia']="Bowl"
dict['England']="Bat"
dict['Sri Lanka']="Bowl"
dict['West Indies']="Bat"
dict['South Africa']="Bat"
dict['Bangladesh']="Bowl"
dict['New Zealand']="Bowl"

print("Enter Pitch type of ",Ground_name," Stadium:")
Pitch_Type=input()
print("Enter Current ODI Ranking of",country1_name,":")
Team1_Ranking=input()
print("Enter Current ODI Ranking of",country2_name,":")
Team2_Ranking=input()

if (Team1_Ranking>Team2_Ranking):
    benefit_ranking=0

```

```

else:
    benefit_ranking=1

country1_number=country_map.get(country1_name,"Country not found")
country2_number=country_map.get(country2_name,"Country not found")
country1_Strength=dict[country1_name]
country2_Strength=dict[country2_name]

if(Pitch_Type==country1_Strength):
    pitch_ranking=1
else:
    pitch_ranking=0
list1=[country1_number,country2_number,Ground_name,Pitch_Type,country1_Strength,country2_Strength,pitch_ranking,benefit_ranking]

list_new = np.array(list1)
X=np.vstack([X, list_new])
#print(X)

labelencoder_X_2 = LabelEncoder()
X[:, 2] = labelencoder_X_2.fit_transform(X[:, 2])
#print (X)

labelencoder_X_3 = LabelEncoder()
X[:, 3] = labelencoder_X_3.fit_transform(X[:, 3])
#print (X)

labelencoder_X_4 = LabelEncoder()
X[:, 4] = labelencoder_X_4.fit_transform(X[:, 4])
#print (X)

labelencoder_X_5 = LabelEncoder()
X[:, 5] = labelencoder_X_5.fit_transform(X[:, 5])
#print (X)

onehotencoder = OneHotEncoder(categorical_features = [0])
X = onehotencoder.fit_transform(X).toarray()
X = X[:, 1:]

onehotencoder = OneHotEncoder(categorical_features = [9])
X = onehotencoder.fit_transform(X).toarray()
X = X[:, 1:]

onehotencoder = OneHotEncoder(categorical_features = [19])
X = onehotencoder.fit_transform(X).toarray()
X = X[:, 1:]

X_pred_new = X[-1,:]

```

```

X = X[:-1,:]

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.10,
random_state = 0)

"""#Applying Decision Tree Classifier
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)"""

"""#xgBoost
import xgboost as xgb
classifier=xgb.XGBClassifier(seed=82)"""

"""#SVC
from sklearn.svm import SVC
classifier=SVC(random_state=912, kernel='rbf')
classifier.fit(X_train, y_train)"""

# Fitting Logistic Regression to the Training set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, y_train)

"""
# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)"""

"""
# Fitting Naive Bayes to the Training set
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, y_train)"""

# Predicting the Test set results
y_pred = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix

```

```

cm = confusion_matrix(y_test, y_pred)

from sklearn.metrics import accuracy_score
ac=accuracy_score(y_test, y_pred)
#print("Acuracy=",ac*100)

"""labelencoder_X_3 = LabelEncoder()
list[:, 3] = labelencoder_X_3.fit_transform(list[:, 3])
labelencoder_X_4 = LabelEncoder()
list[:, 4] = labelencoder_X_4.fit_transform(list[:, 4])
labelencoder_X_5 = LabelEncoder()
list[:, 5] = labelencoder_X_5.fit_transform(list[:, 5])"""

"""onehotencoder = OneHotEncoder(categorical_features = [0])
list = onehotencoder.fit_transform(list).toarray()
list = list[:, 1:]

onehotencoder = OneHotEncoder(categorical_features = [9])
list = onehotencoder.fit_transform(list).toarray()
list = list[:, 1:]

onehotencoder = OneHotEncoder(categorical_features = [19])
list = onehotencoder.fit_transform(list).toarray()
list = list[:, 1:] """

X_pred_final = X_pred_new.reshape(1, -1)
y1_pred = classifier.predict(X_pred_final)

for winner_country in country_map:
    if country_map[winner_country] is int(y1_pred):
        i=winner_country
        break

#y_name=country_map.get(y1_pred,"Country not found")
#print(type(list))
#print(y1_pred)

prob = classifier.predict_proba(X_pred_final)
prob_max = max(prob)
pin = max(prob_max)
print(winner_country, 'has high chances of winning with probability',pin*100)

```

3.1 Approach followed in proposed System with discussions on Machine Learning techniques used in System

Logistic Regression: In statistics, the logistic model (or logit model) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probity model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each dependent variable having its own parameter; for a binary independent variable this generalizes the odds ratio.

Logistic regression was developed by statistician David Cox in 1958. The binary logistic regression model has extensions to more than two levels of the dependent variable: categorical outputs with more than two values are modelled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier. The coefficients are generally not computed by a closed-form expression, unlike linear least squares.

4.0 Data Collection and Data Preparation, Cleaning of data, Preprocessing

The data collected available was from 75 but as the team ranking were not available from that time we reduced our data set and took entries after 2010. Later on there were more than 50 teams' data over the past 8 years. We eliminated the data of teams other than prominent 8 international teams.

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Clipboard Font Alignment Merge & Center Number Conditional Formatting Format as Table Styles Cells AutoSum Fill Clear Sort & Find & Filter

Calibri 11 A A Wrap Text General Normal Bad Good Neutral Calculation Check Cell

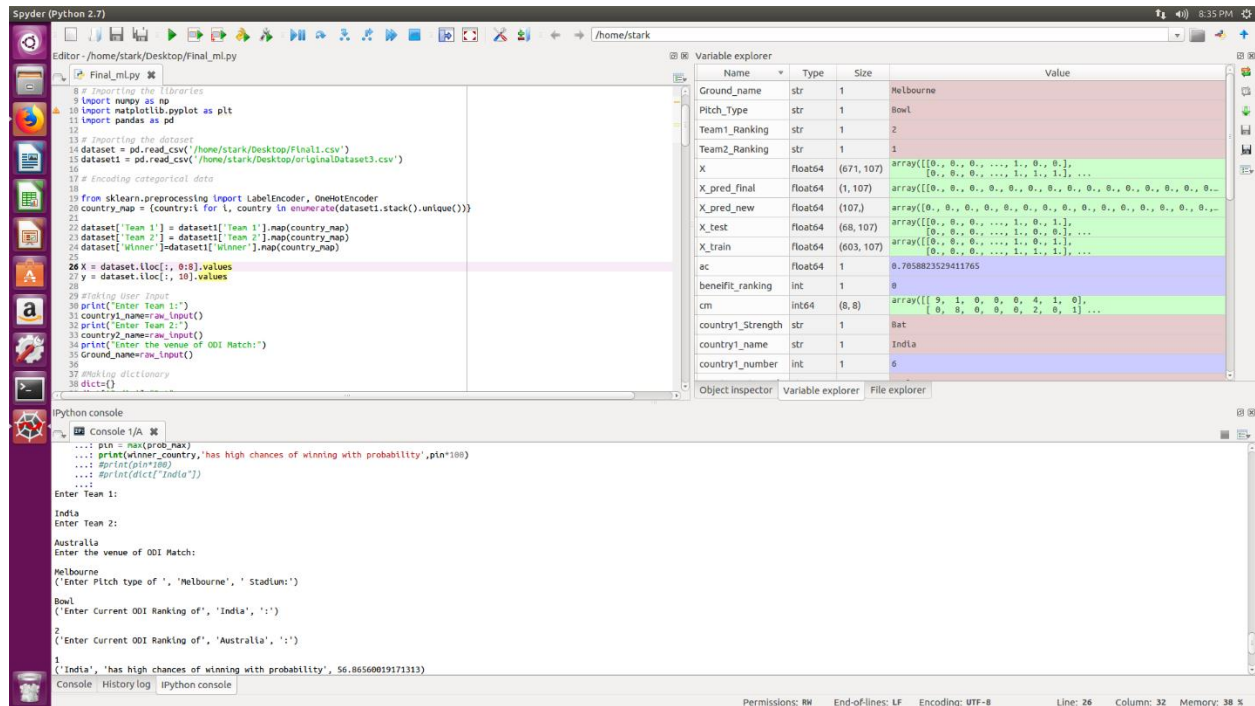
Clipboard Font Alignment Merge & Center Number Conditional Formatting Format as Table Styles Cells AutoSum Fill Clear Sort & Find & Filter

A1 Team 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	Team 1	Team 2	Ground	Pitch	Gooc	Team1	Go	Team2	Go	Pitch	Bene	Ranking	Bt	Team1	Rai	Team2	Rai	Winner						
2	Banglades	Sri Lanka	Dhaka	Bat	Bowl	Bowl		0	0	2	7	Sri Lanka		0										
3	India	Sri Lanka	Dhaka	Bat	Bowl			1	1	8	7	Sri Lanka		1										
4	Banglades	India	Dhaka	Bat	Bowl	Bat		0	0	2	8	India		0										
5	Banglades	Sri Lanka	Dhaka	Bat	Bowl	Bowl		0	0	2	7	Sri Lanka		0										
6	India	Sri Lanka	Dhaka	Bat	Bat	Bowl		1	1	8	7	India		1										
7	Banglades	India	Dhaka	Bat	Bowl	Bat		0	0	2	8	India		0										
8	India	Sri Lanka	Dhaka	Bat	Bat	Bowl		1	1	8	7	Sri Lanka		1										
9	Australia	Pakistan	Brisbane	Bowl	Bowl	Bowl		0	1	9	4	Australia		0										
10	Australia	Pakistan	Sydney	Bowl	Bowl	Bowl		0	1	9	4	Australia		0										
11	Australia	Pakistan	Adelaide	Bowl	Bowl	Bowl		0	1	9	4	Australia		0										
12	Australia	Pakistan	Perth	Bowl	Bowl	Bowl		0	1	9	4	Australia		0										
13	Australia	Pakistan	Perth	Bowl	Bowl	Bowl		0	1	9	4	Australia		0										
14	New Zeala	Banglades	Napier	Bowl	Bowl	Bowl		0	1	3	2	New Zealand		0										
15	Australia	West Indie	Melbourne	Bowl	Bowl	Bat		1	1	9	1	Australia		1										
16	New Zeala	Banglades	Dunedin	Bowl	Bowl	Bowl		0	1	3	2	New Zealand		0										
17	Australia	West Indie	Adelaide	Bowl	Bowl	Bat		1	1	9	1	Australia		1										
18	New Zeala	Banglades	Christchun	Bat	Bowl	Bowl		0	1	3	2	New Zealand		0										
19	Australia	West Indie	Sydney	Bowl	Bowl	Bat		1	1	9	1	no result		1										
20	Australia	West Indie	Brisbane	Bowl	Bowl	Bat		1	1	9	1	Australia		1										
21	Australia	West Indie	Melbourne	Bowl	Bowl	Bat		1	1	9	1	Australia		1										
22	India	South Afri	Jaipur	Bat	Bat	Bat		0	1	8	6	India		0										
23	India	South Afri	Gwalior	Bat	Bat	Bat		0	1	8	6	India		0										
24	India	South Afri	Ahmedaba	Bat	Bat																			

10 | Page

6.0 Testing of the Model



7.0 Results and Discussions

Finally, the built model is able to recognize winner team with winning probability. We can increase the scope of the project by expanding the database to more number teams and different format of cricket and by using a more user view-friendly U/I.

8.0 Conclusions and Future Scope

The project addresses the problem of predicting the outcome of an ODI cricket match using the statistics of 800 matches. The novelty of our approach lies in addressing the problem as a dynamic one, and using the participating players as the key feature in predicting the winner of the match. We observe that simple features can yield very promising result.

It future it can be used to predict the score,wickets etc of each team by taking into account the ball by ball record.



9.0 References

1. <http://www.howstat.com/cricket/Articles/RecentRecords.asp>
2. <https://www.kaggle.com/jaykay12/odi-cricket-matches-19712017>
3. <http://www.espncricinfo.com/rankings/content/page/211271.html>
4. https://en.wikipedia.org/wiki/ICC_ODI_Championship#Ranking_table