IPL Prediction Analysis

BIG DATA PROJECT

SNo	Name	USN	Class/Section
1	Sandeep PVN	01FB14ECS207	D
2	Shreyas Kathavathe	01FB14ECS232	D
3	Shuaib UR Rahman	01FB14ECS237	D
4	Siddharth Srinivasan	01FB14ECS241	D

Introduction

Our project is based on analyzing the various IPL and T20 matches from 2011-2015, and making use of the ball-by-ball probabilities as well as batsman-bowler probabilities obtained from programming using Map Reduce and Spark, so as to build a model that can predict various attributes associated with every ball accurately, such as number of runs scored, leg byes, wickets, dot balls etc. for the 2016 IPL Match.

We have made use of Big Data Applications such as Hadoop, HDFS, Hive, Spark and Hbase to store and manipulate this vast amount of data.

Related work

The collected data consists of player vs player statistics and individual player profiles: the batting strike rate, bowling economy rate etc. We extracted this data from the official Cricinfo website using a python script

We collected the datasets we would make use of(i.e IPL match data, batting & bowling profile, player-player profile, team profile) from various other websites (such as cricketarchive.com & cricsheet.org), and operated on them to obtain various cluster probabilities and pairwise probabilities.

We made use of K-means clustering for grouping similar players based on attributes such as average scoring rate, strike rate, economy rate, no. of wickets etc.

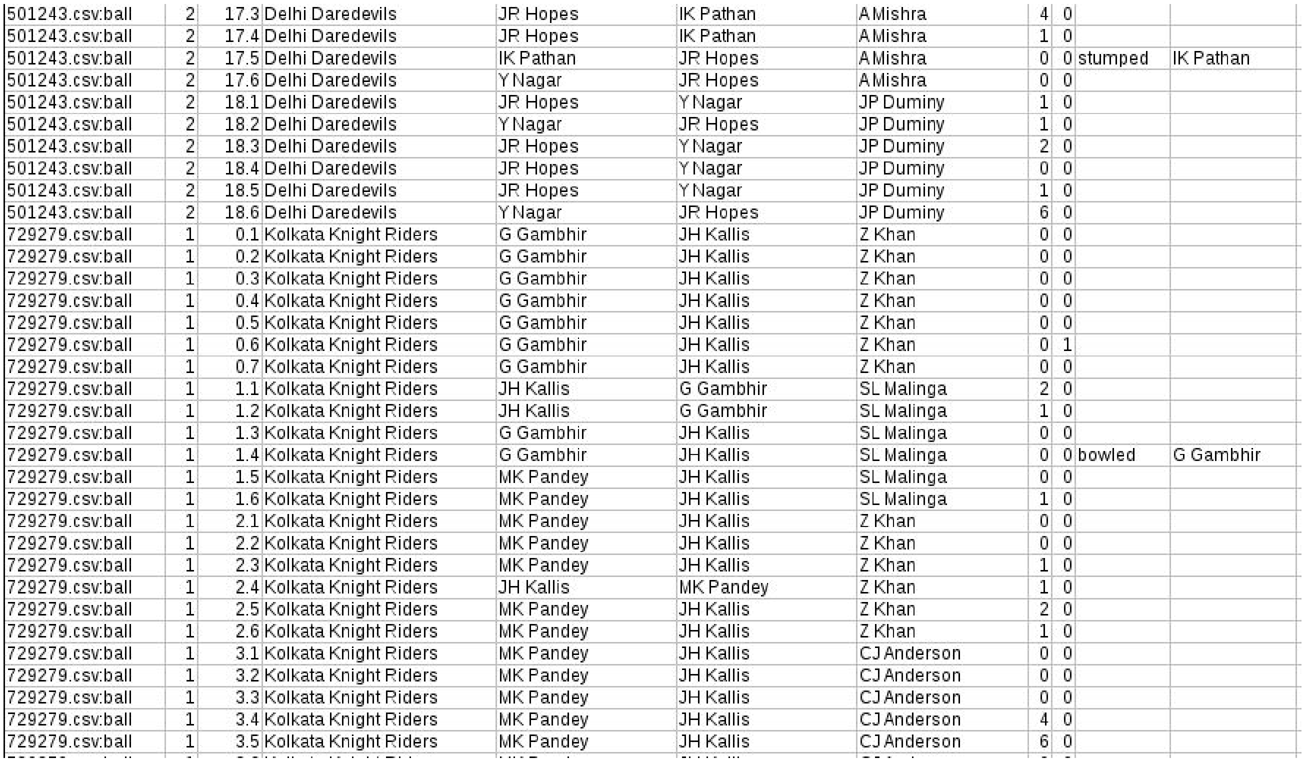
We also made use of formulae for weighted probabilities of players on the basis of giving more preference for recent matches

ALGORITHM/DESIGN

- We browsed through website sources and made scripts to extract all individual batsman & bowler profiles as well as player-player data (see BeautifulSoup in python). We then proceeded to load individual player profiles in HDFS and convert them to individual player probabilities of occurrence of various runs, dismissals, wickets etc. using Map Reduce.
- We created individual bowler clusters and batsman clusters by transferring individual player profiles to Hive and operating on them in Scala using Spark MLLib library based on number of runs & strike rate for clustering batsmen and runs given & economy rate for clustering bowlers.
- We then proceed to predict the ball by ball outcome based on these calculated probabilities given a batsman and a bowler, by selecting a random number between 0 and 1 & finding most probable number of runs for that number.

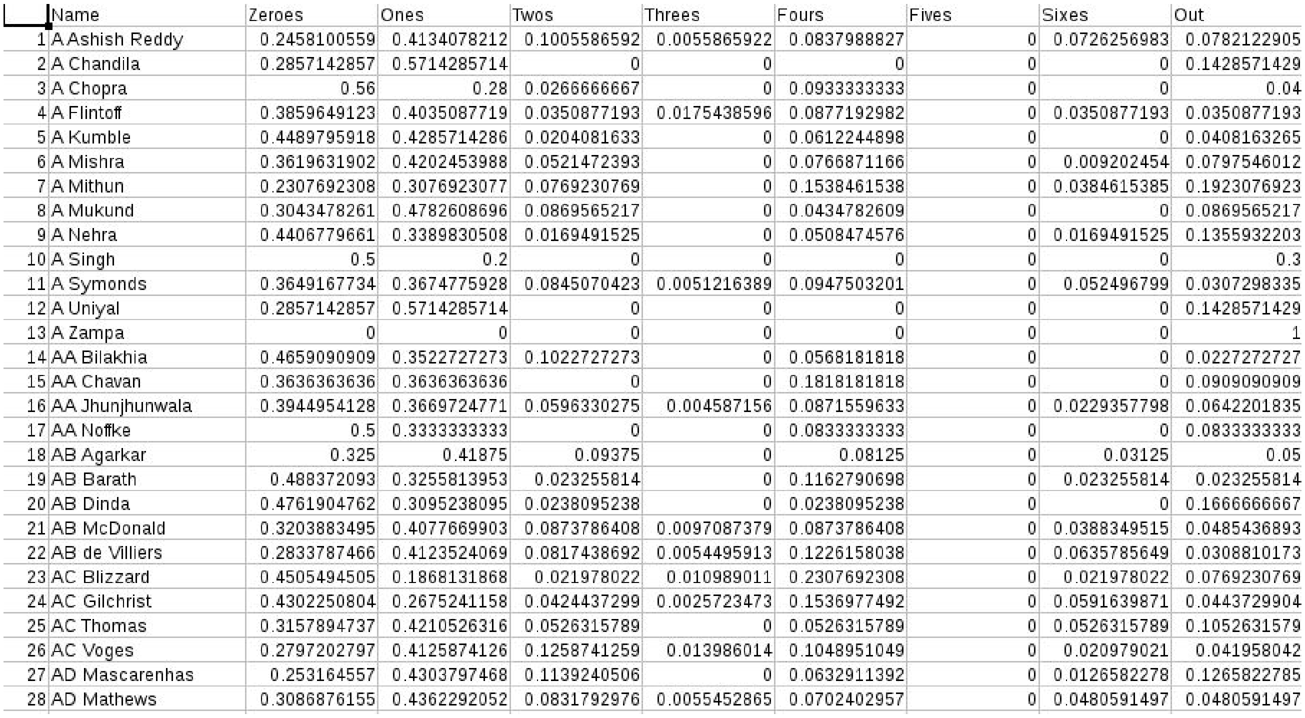
EXPERIMENTAL RESULTS

We obtained IPL data for various matches in several csv files (cricsheet.org), where each csv denotes an IPL match. We then proceeded to combine data in all csv files into a single csv file grouped by the csv file name. The fields for every ball of every match in the final csv are name of the csv file (to identify a given IPL match played at a given time), innings number, starting from 1, over and ball, batting team name, batsman, non-striker, bowler, runs-off-bat, extras,kind of wicket (if any), and dismissed played (if there was a wicket).



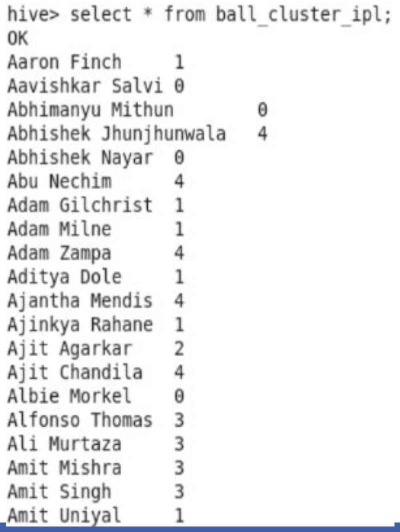
501243.csv:ball	2	17.3	Delhi Daredevils	JR Hopes	IK Pathan	A Mishra	4	0		
501243.csv:ball	2	17.4	Delhi Daredevils	JR Hopes	IK Pathan	A Mishra	1	0		
501243.csv:ball	2	17.5	Delhi Daredevils	IK Pathan	JR Hopes	A Mishra	0	0	stumped	IK Pathan
501243.csv:ball	2	17.6	Delhi Daredevils	Y Nagar	JR Hopes	A Mishra	0	0		
501243.csv:ball	2	18.1	Delhi Daredevils	JR Hopes	Y Nagar	JP Duminy	1	0		
501243.csv:ball	2	18.2	Delhi Daredevils	Y Nagar	JR Hopes	JP Duminy	1	0		
501243.csv:ball	2	18.3	Delhi Daredevils	JR Hopes	Y Nagar	JP Duminy	2	0		
501243.csv:ball	2	18.4	Delhi Daredevils	JR Hopes	Y Nagar	JP Duminy	0	0		
501243.csv:ball	2	18.5	Delhi Daredevils	JR Hopes	Y Nagar	JP Duminy	1	0		
501243.csv:ball	2	18.6	Delhi Daredevils	Y Nagar	JR Hopes	JP Duminy	6	0		
729279.csv:ball	1	0.1	Kolkata Knight Riders	G Gambhir	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	0.2	Kolkata Knight Riders	G Gambhir	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	0.3	Kolkata Knight Riders	G Gambhir	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	0.4	Kolkata Knight Riders	G Gambhir	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	0.5	Kolkata Knight Riders	G Gambhir	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	0.6	Kolkata Knight Riders	G Gambhir	JH Kallis	Z Khan	0	1		
729279.csv:ball	1	0.7	Kolkata Knight Riders	G Gambhir	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	1.1	Kolkata Knight Riders	JH Kallis	G Gambhir	SL Malinga	2	0		
729279.csv:ball	1	1.2	Kolkata Knight Riders	JH Kallis	G Gambhir	SL Malinga	1	0		
729279.csv:ball	1	1.3	Kolkata Knight Riders	G Gambhir	JH Kallis	SL Malinga	0	0		
729279.csv:ball	1	1.4	Kolkata Knight Riders	G Gambhir	JH Kallis	SL Malinga	0	0	bowled	G Gambhir
729279.csv:ball	1	1.5	Kolkata Knight Riders	MK Pandey	JH Kallis	SL Malinga	0	0		
729279.csv:ball	1	1.6	Kolkata Knight Riders	MK Pandey	JH Kallis	SL Malinga	1	0		
729279.csv:ball	1	2.1	Kolkata Knight Riders	MK Pandey	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	2.2	Kolkata Knight Riders	MK Pandey	JH Kallis	Z Khan	0	0		
729279.csv:ball	1	2.3	Kolkata Knight Riders	MK Pandey	JH Kallis	Z Khan	1	0		
729279.csv:ball	1	2.4	Kolkata Knight Riders	JH Kallis	MK Pandey	Z Khan	1	0		
729279.csv:ball	1	2.5	Kolkata Knight Riders	MK Pandey	JH Kallis	Z Khan	2	0		
729279.csv:ball	1	2.6	Kolkata Knight Riders	MK Pandey	JH Kallis	Z Khan	1	0		
729279.csv:ball	1	3.1	Kolkata Knight Riders	MK Pandey	JH Kallis	CJ Anderson	0	0		
729279.csv:ball	1	3.2	Kolkata Knight Riders	MK Pandey	JH Kallis	CJ Anderson	0	0		
729279.csv:ball	1	3.3	Kolkata Knight Riders	MK Pandey	JH Kallis	CJ Anderson	0	0		
729279.csv:ball	1	3.4	Kolkata Knight Riders	MK Pandey	JH Kallis	CJ Anderson	4	0		
729279.csv:ball	1	3.5	Kolkata Knight Riders	MK Pandey	JH Kallis	CJ Anderson	6	0		

From this we obtained various batsmen and bowler probabilities using Map Reduce, and it looks something like this.



	Name	Zeroes	Ones	Twos	Threes	Fours	Fives	Sixes	Out
1	A Ashish Reddy	0.2458100559	0.4134078212	0.1005586592	0.0055865922	0.0837988827	0	0.0726256983	0.0782122905
2	A Chandila	0.2857142857	0.5714285714	0	0	0	0	0	0.1428571429
3	A Chopra	0.56	0.28	0.0266666667	0	0.0933333333	0	0	0.04
4	A Flintoff	0.3859649123	0.4035087719	0.0350877193	0.0175438596	0.0877192982	0	0.0350877193	0.0350877193
5	A Kumble	0.4489795918	0.4285714286	0.0204081633	0	0.0612244898	0	0	0.0408163265
6	A Mishra	0.3619631902	0.4202453988	0.0521472393	0	0.0766871166	0	0.009202454	0.0797546012
7	A Mithun	0.2307692308	0.3076923077	0.0769230769	0	0.1538461538	0	0.0384615385	0.1923076923
8	A Mukund	0.3043478261	0.4782608696	0.0869565217	0	0.0434782609	0	0	0.0869565217
9	A Nehra	0.4406779661	0.3389830508	0.0169491525	0	0.0508474576	0	0.0169491525	0.1355932203
10	A Singh	0.5	0.2	0	0	0	0	0	0.3
11	A Symonds	0.3649167734	0.3674775928	0.0845070423	0.0051216389	0.0947503201	0	0.052496799	0.0307298335
12	A Uniyal	0.2857142857	0.5714285714	0	0	0	0	0	0.1428571429
13	A Zampa	0	0	0	0	0	0	0	1
14	AA Bilakhia	0.4659090909	0.3522727273	0.1022727273	0	0.0568181818	0	0	0.0227272727
15	AA Chavan	0.3636363636	0.3636363636	0	0	0.1818181818	0	0	0.0909090909
16	AA Jhunjhunwala	0.3944954128	0.3669724771	0.0596330275	0.004587156	0.0871559633	0	0.0229357798	0.0642201835
17	AA Noffke	0.5	0.3333333333	0	0	0.0833333333	0	0	0.0833333333
18	AB Agarkar	0.325	0.41875	0.09375	0	0.08125	0	0.03125	0.05
19	AB Barath	0.488372093	0.3255813953	0.023255814	0	0.1162790698	0	0.023255814	0.023255814
20	AB Dinda	0.4761904762	0.3095238095	0.0238095238	0	0.0238095238	0	0	0.1666666667
21	AB McDonald	0.3203883495	0.4077669903	0.0873786408	0.0097087379	0.0873786408	0	0.0388349515	0.0485436893
22	AB de Villiers	0.2833787466	0.4123524069	0.0817438692	0.0054495913	0.1226158038	0	0.0635785649	0.0308810173
23	AC Blizzard	0.4505494505	0.1868131868	0.021978022	0.010989011	0.2307692308	0	0.021978022	0.0769230769
24	AC Gilchrist	0.4302250804	0.2675241158	0.0424437299	0.0025723473	0.1536977492	0	0.0591639871	0.0443729904
25	AC Thomas	0.3157894737	0.4210526316	0.0526315789	0	0.0526315789	0	0.0526315789	0.1052631579
26	AC Voges	0.2797202797	0.4125874126	0.1258741259	0.013986014	0.1048951049	0	0.020979021	0.041958042
27	AD Mascarenhas	0.253164557	0.4303797468	0.1139240506	0	0.0632911392	0	0.0126582278	0.1265822785
28	AD Mathews	0.3086876155	0.4362292052	0.0831792976	0.0055452865	0.0702402957	0	0.0480591497	0.0480591497

We then performed clustering to group similar batsmen and bowlers based on their economy rate, no. of runs scored, strike rate etc. Bowler clusters are shown below.



```
hive> select * from ball_cluster_ipl;
OK
Aaron Finch     1
Aavishkar Salvi 0
Abhimanyu Mithun        0
Abhishek Jhunjhunwala   4
Abhishek Nayar  0
Abu Nechim      4
Adam Gilchrist  1
Adam Milne      1
Adam Zampa      4
Aditya Dole     1
Ajantha Mendis  4
Ajinkya Rahane  1
Ajit Agarkar    2
Ajit Chandila   4
Albie Morkel    0
Alfonso Thomas  3
Ali Murtaza     3
Amit Mishra     3
Amit Singh      3
Amit Uniyal     1
```

FUTURE ENHANCEMENTS

- ➢ The main focus is to develop a better model that is more accurate than the existing one we have, by operating on larger amount of datasets.
- ➢ Taking into account other miscellaneous factors such as weather, pitch, stadium size, player's form, etc. in real time.
- ➢ Making our code more usable out of the box so that it can be run with ease.
- ➢ Making it more friendly to use and easier to interpret

REFERENCES

- ➢ IPL player profiles - CricketArchive
- ➢ IPL match data – Crickinfo
- ➢ Web Scraping in Python – Scraper
- ➢ Spark MLLib – MLLib
- ➢ Weighted Probability - WikiHow