Machine Learning in Indian Premier League

By Sunil Ayyappan

Table of Contents

1. About Indian Premier League

IPL competition is played amongst 8 elite clubs. The format of the game is very simple – each team will face the other 7 teams twice once in their home ground and then in other team's ground. At the end of all the 56 matches, the top four will qualify for the playoff. The first two ranked team will play against each other in Qualifier 1 and the third and fourth team will play against each other in Qualifier 2. The winner of Qualifier 1 one will be qualified for the final and the loser will play against the winner of Qualifier 2 in the Eliminator.

Group games do not necessarily end with one side winning. They can either have 'no result' if weather prevents the game from being finished or be 'Super-over if both teams end on the same score. In the knock out rounds, 'no result' games are replayed on a reserve day and 'tied' matches are decided by 'super-over, ensuring that there is always a winner.

In each IPL team, a total of 11 players will play the match and it must comprise of 7 Indians and 4 foreign players. The rule of 7 Indians is mandatory and was introduced to promote local players. So one of the challenges all the team faces in each match is to select the 4 foreign players from a total of 8 players.

2. Goal

Predicting a Home Team win

To predict whether a home team with given finite features will win a match

Sentiment analysis on IPL T20 hashtag

To analyze the sentiment expressed by the cricketing audience in twitter on the game day.

Selecting playing XI

To weight different sets of players skills and derive a point for each player and select the playing XI based on the points.

3. How data was collected?

There were no ready-made data available for this. Data was copied from cricketing website such as www.cricinfo.com and www.cricbuzz.com. This formed the core data from where the final set of data was derived using different techniques in python.

4. Data dictionary

IPL ALL MATCHES SUMMARY (Initial Data)

Field Name	Type	Description
match_id	int64	Unique id
Season	int64	Year
city	Object	City of the ground
date	Object	When the match was played
home_game	int64	Played in home ground?
home_team	object	Team name who are playing at home
away_team	object	Team who are visiting
toss_winner	object	Who won the toss
toss_decision	object	Field or Bat
winner	Object	Winner team name
win_by_runs	int64	By how much run they won
win_by_wickets	Int64	By how much wickets they won
player_of_match	object	Player of the match

5. Features added to the data

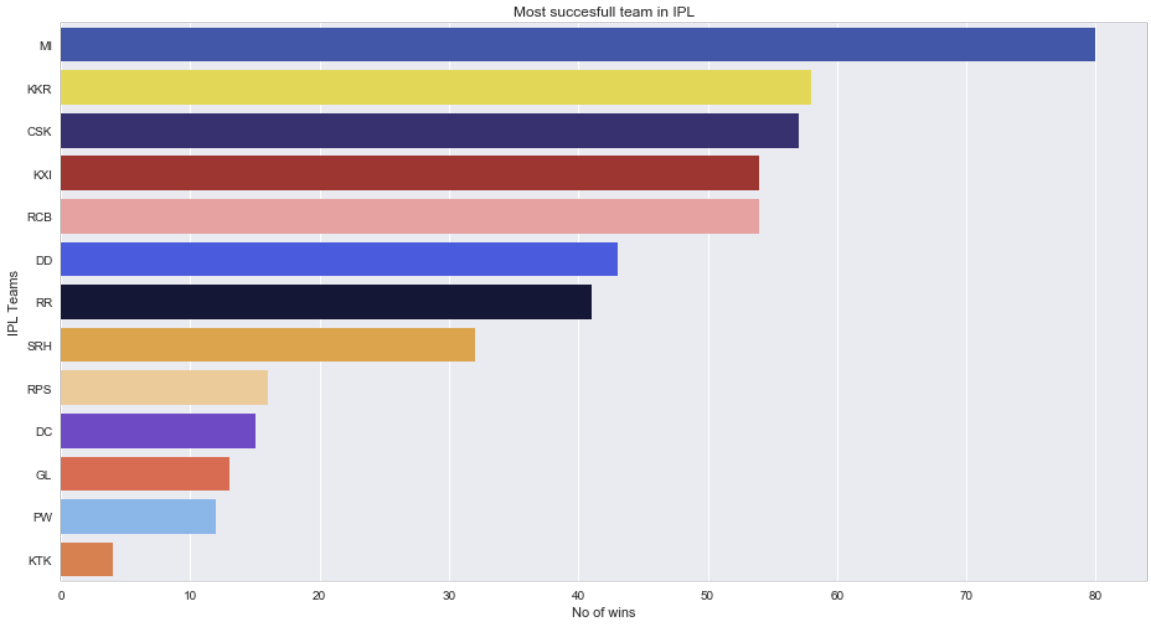
Field Name	Type	Description
AWP	int64	Away winning Percentage
HBRR	int64	Home batting run rate
HBER	Int64	Home Bowling Economy Rate
HBAR	int64	Home Batting Average Runs Scored
HBOA	int64	Home Bowling Average Runs Conceded
ABRR	Int64	Away Batting Run Rate
ABER	Int64	Away Bowling Economy Rate
ABAR	Int64	Away Batting Average Runs Scored
ABOA	Int64	Away Bowling Average Runs Conceded
HTB	Int64	Home team Batting First
ATB	int64	Away team Batting First
HTP	Int64	Home team total points based on their team
HTBA	int64	No of Batsmen in the home team
HTAL	Int64	No of All-rounders in the home team
HTBO	int64	No of Bowlers in the home team
HTWK	Int64	No of wickets lost by home team
HTOV	int64	No of overs played by the home team
ATP	Int64	Away team total points based on their team
ATBA	int64	No of Batsmen in the away team
ATAL	Int64	No of All-rounders in the away team
ATBO	int64	No of Bowlers in the away team
ATSC	Int64	Score of Away team
ATWK	int64	No of wickets lost by away team
ATOV	Int64	No of overs played by home team

6. POINTS TO NOTE

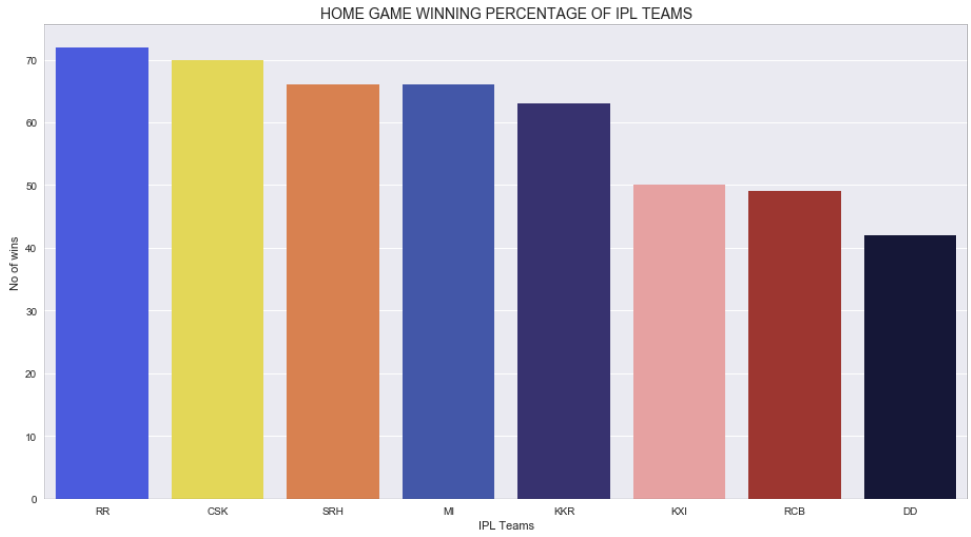
- In 2008 IPL started with 8 teams - Chennai Super Kings, Rajasthan Royals, Kolkata Knight Riders, Mumbai Indians, Kings XI, Deccan Chargers, Royal Challengers, Delhi DareDevils
- In 2011 - Two more teams were introduced - Pune Warriors and Kochi Tuskers. But after 2013, both these teams were terminated for breaching its terms of agreement.
- In 2013 - Deccan chargers were sold & renamed as Sunrisers Hyderabad.
- In 2016 - Two popular teams who were previous champions Chennai Super Kings and Rajasthan Royals was suspended for two years. Just to keep up with the number of matches two new teams Pune Supergiants and Gujarat Lions took their place and played 2016 and 2017 season
- In 2018 - Both the suspended teams are returning replacing the Pune Supergiants and Gujarat Lions
- In our datasets, you will be seeing this inconsistency in the team across the season. This is not a data issue but because of all the events happened.

7. DATA EXPLORATION

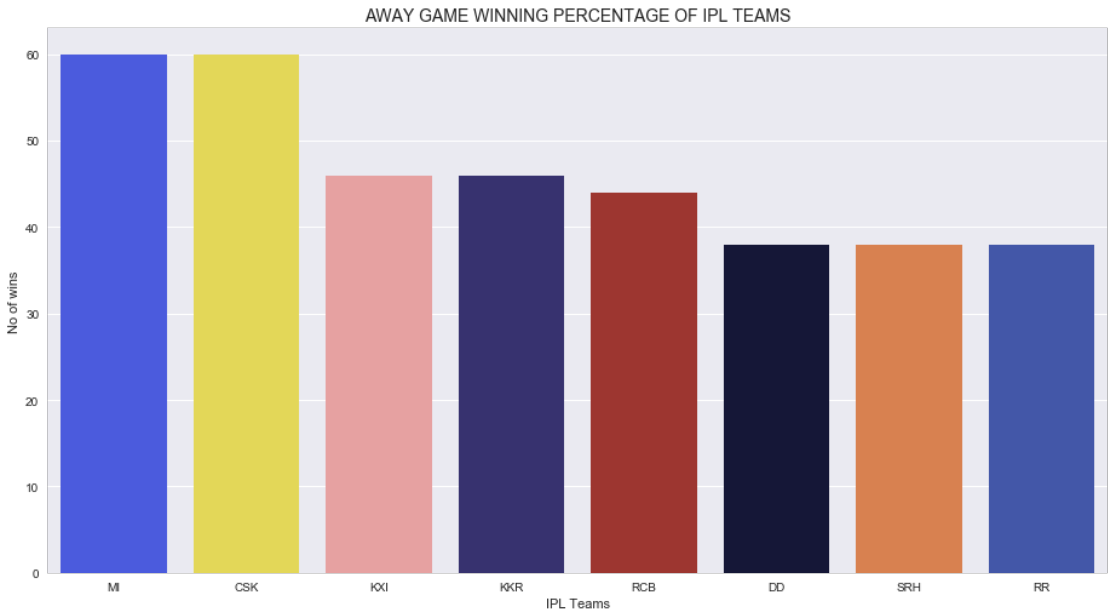
Mumbai Indians seems to be the most successful team in all of the IPL seasons with 80 wins in 10 seasons followed by Kolkata Knight Riders with 58 wins in 10 seasons and Chennai Super Kings with 57 in 8 seasons.



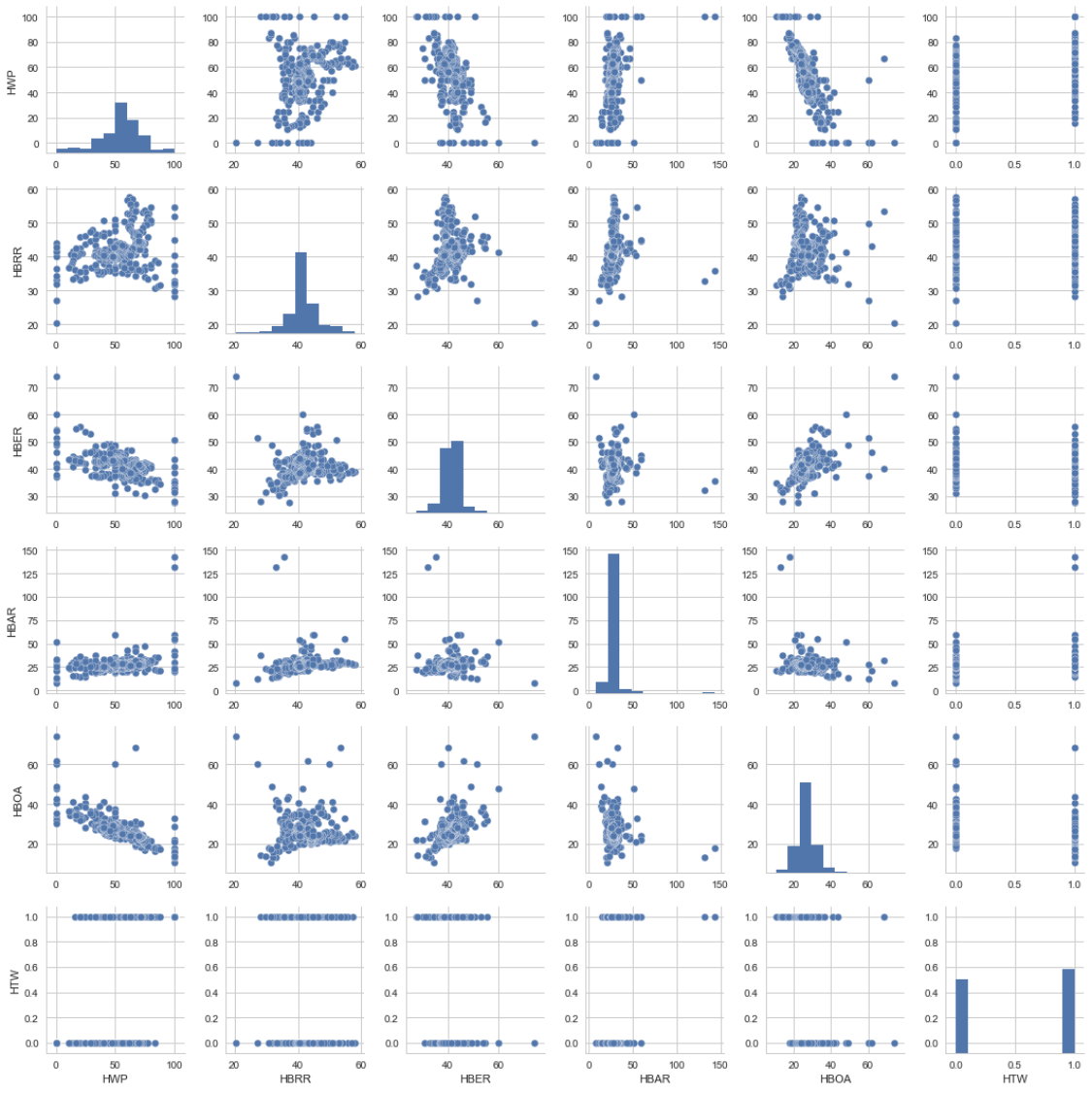
The winning percentage of RR & CSK has been terrific on their home ground. They have a highest winning percentage of 71% & 70.21% respectively. Even Sunrisers Hyderabad are closely behind with 65.71%. But one thing to notice is that all these three teams have played less games than the other 5 teams. That could be one of the reason why their win % is pretty good at home.



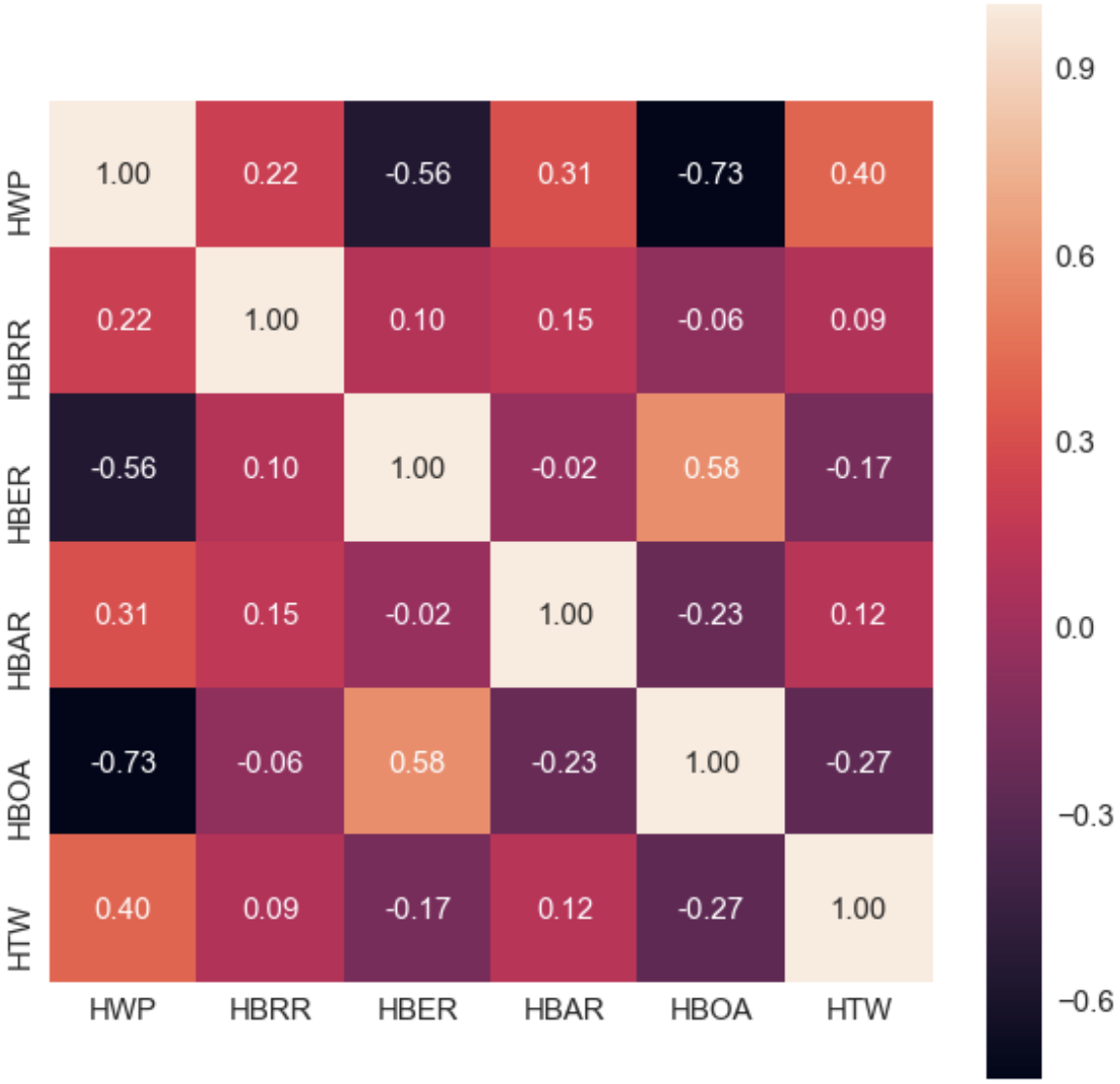
MI leads the way in winning away games with 60% closely followed by CSK with a 57%. CSK has been excellent in winning home games and second in winning away games. This is one of the reason why they were able to make it to all the playoffs in all the seasons they have played. Delhi has been poor both in away and home games and RR has been the worst traveler of all seasons.



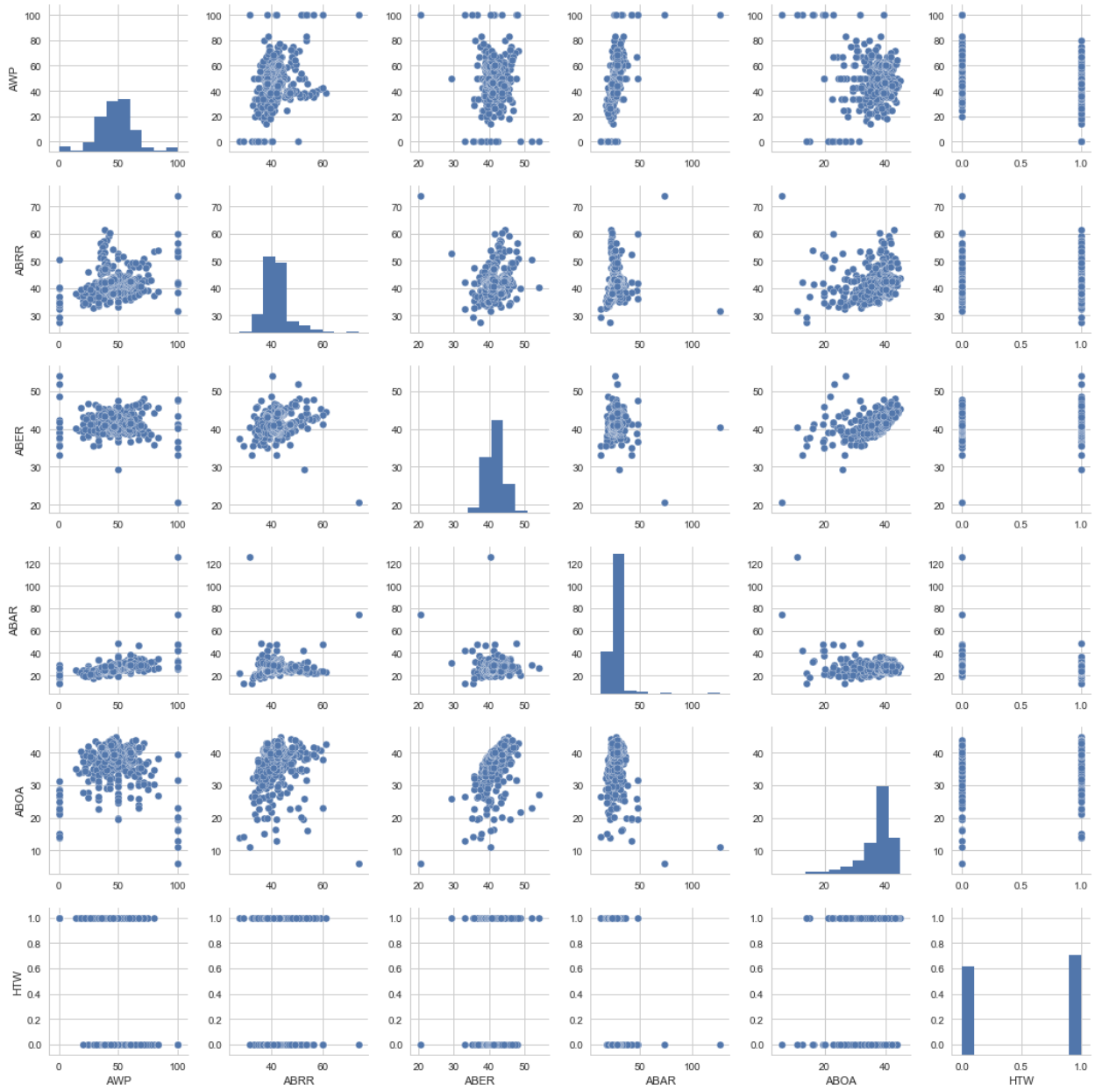
Below chart indicates how the newly added features for home team stack up against each other. You can positive relation Home bowling economy rate (HBER) and Home bowling average runs conceded (HBOA). We can also notice that HBER is normally distributed. Some things to note: ‘HTBO’ is the home team bowler.



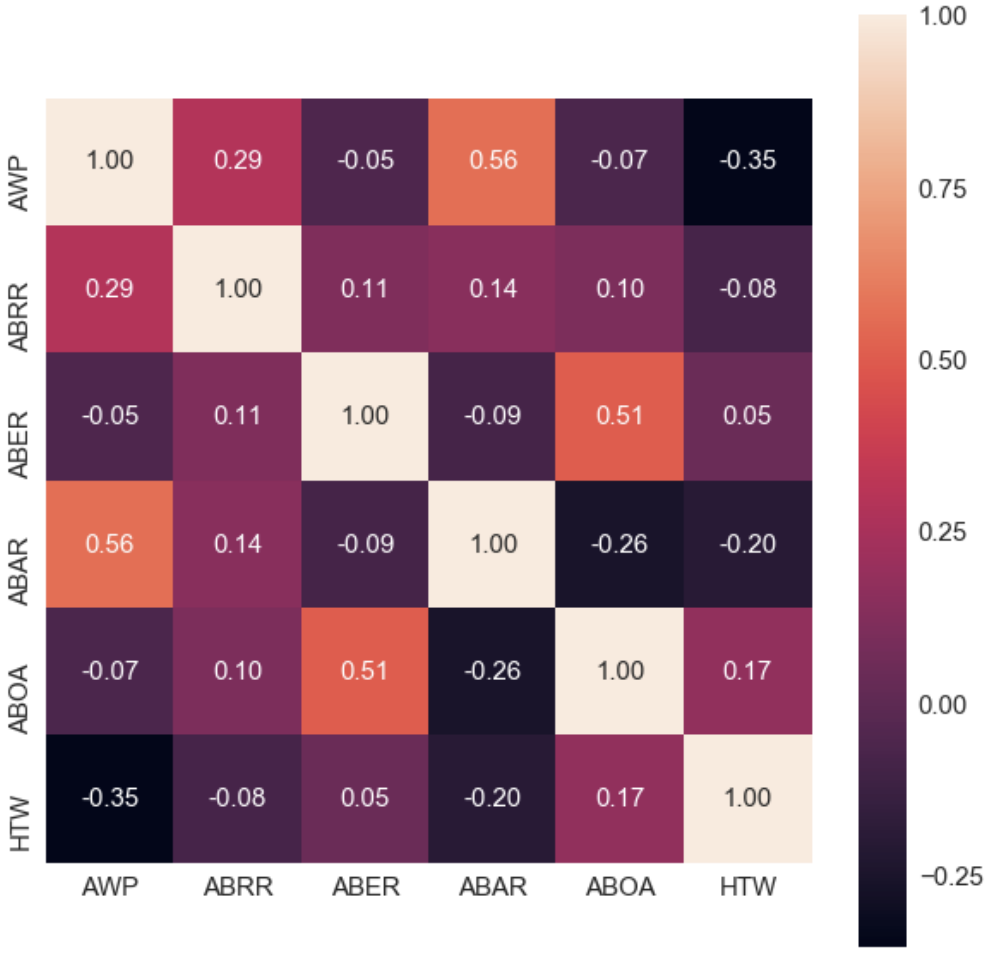
HTBA (Home Team Batsman) is the most negatively correlated with HTSC (Home Team Score) which is surprising because you normally expect a team to score more when they have more batsman in their team. But it looks like every time a team has played a match with more batsman, less they score



Similarly, we have done the pair plot and coefficient relation plot for away team also. Again, we see a positive relationship between ABER and ABOA, ABER and ABRR.



Again ATBA (Away Team Batsman) is the most negatively correlated with ATSC (Away Team Score). Also Away Team Score has a positive relation with Home Team Bowlers. So more specialized bowlers in opposition team, more runs you score.



8. Forecasting and evaluation

Helper function was used to pick the best model and adjust the hyperparameters. As you can see below, Random forest classifier performs the best. In all fairness to the linear models, I could have done more to make them better like scale and log transform some of the variables. I also tried setting n_estimators for RandomForest and Gradient Boost Regressor 1000. But for now I decided to move ahead with the highest scoring model from Gridsearch.

1. Gradient Boosting Regressor
2. Random Forest Classifier
3. Random Forest Regressor.

	estimator	min_score	mean_score	max_score	std_score	alpha	learning_rate	max_depth	max_features	min_samples_leaf	n_estimators
111	RFC	0.890909	0.903331	0.918182	0.0112658	NaN	NaN	6	0.1	5	100
108	RFC	0.872727	0.903303	0.927273	0.0227528	NaN	NaN	6	0.1	3	100
109	RFC	0.881818	0.900246	0.918919	0.0151473	NaN	NaN	6	0.1	3	500
110	RFC	0.872727	0.897243	0.90991	0.0173382	NaN	NaN	6	0.1	3	800
81	RFC	0.881818	0.897243	0.90991	0.011633	NaN	NaN	4	0.1	3	100
112	RFC	0.863636	0.897215	0.918919	0.0240806	NaN	NaN	6	0.1	5	500
99	RFC	0.881818	0.894267	0.909091	0.01126	NaN	NaN	6	0.3	3	100
73	RFC	0.881818	0.89424	0.900901	0.00879103	NaN	NaN	4	0.3	3	500
104	RFC	0.881818	0.89424	0.900901	0.00879103	NaN	NaN	6	0.3	5	800
84	RFC	0.863636	0.894212	0.90991	0.0216231	NaN	NaN	4	0.1	5	100
74	RFC	0.881818	0.894212	0.90991	0.0117038	NaN	NaN	4	0.3	3	800
77	RFC	0.881818	0.894212	0.90991	0.0117038	NaN	NaN	4	0.3	5	800
106	RFC	0.881818	0.891182	0.90991	0.0132426	NaN	NaN	6	0.3	9	500
107	RFC	0.881818	0.891182	0.90991	0.0132426	NaN	NaN	6	0.3	9	800
76	RFC	0.872727	0.891182	0.90991	0.015181	NaN	NaN	4	0.3	5	500
113	RFC	0.854545	0.891155	0.918919	0.0270144	NaN	NaN	6	0.1	5	800
72	RFC	0.882883	0.888234	0.890909	0.00378359	NaN	NaN	4	0.3	3	100
103	RFC	0.881818	0.888179	0.900901	0.00899568	NaN	NaN	6	0.3	5	500
75	RFC	0.872727	0.888179	0.900901	0.0116627	NaN	NaN	4	0.3	5	100
101	RFC	0.863636	0.885176	0.9	0.0155864	NaN	NaN	6	0.3	3	800
78	RFC	0.872727	0.885176	0.891892	0.00881178	NaN	NaN	4	0.3	9	100
82	RFC	0.863636	0.885176	0.9	0.0155864	NaN	NaN	4	0.1	3	500

9. MODEL & SCORE

The top 5 features were:

a. HWP – Home Win Percentage
b. HBOA – Home Bowling Average Runs conceded
c. ABOA – Away Bowling Average Runs conceded
d. HBER – Home Bowling Economy Rate
e. ABER – Away Team Bowling Economy Rate

f. Gradient Boost Regressor

```
MAE train: 0.010, test: 0.164
MSE train: 0.003, test: 0.063
R^2 train: 0.989, test: 0.748
```

g. Random Forest Classifier

```
MAE train: 0.003, test: 0.063
MSE train: 0.003, test: 0.063
R^2 train: 0.988, test: 0.748
```

h. Random Forest Regressor

```
MAE train: 0.115, test: 0.188
MSE train: 0.028, test: 0.071
R^2 train: 0.886, test: 0.716
```

10. Sentiment analysis from twitter

The objective of this exercise is to retrieve tweets on real time from couple of hashtags - #IPL and #IPL20 and analyze the sentiments of the cricket fans on a daily basis. This would give an idea on how the fans are reacting to daily results of their favorite teams.

I registered for a token with twitter API to get this on a real time. We are pulling the tweets on a real-time basis. We can also customize the pull based on our requirement.

The default columns we got from twitter are given below

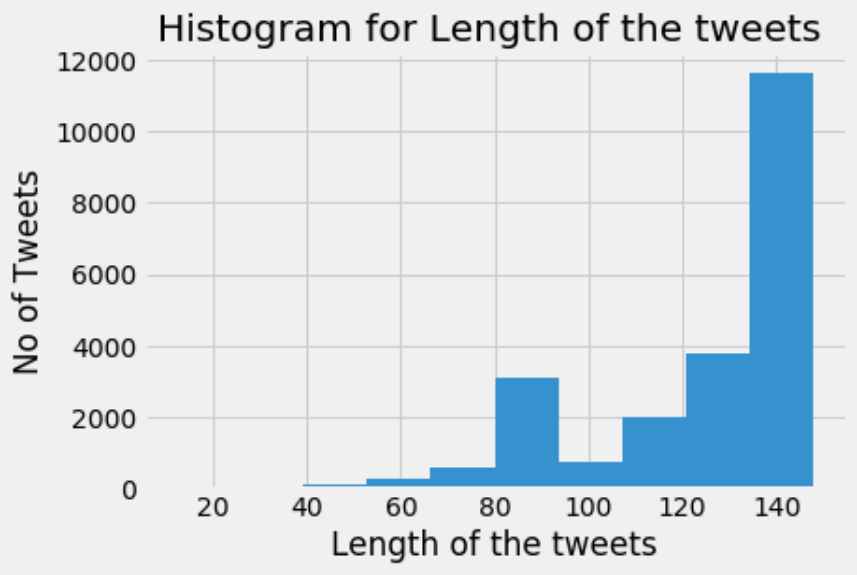
FieldName	Type	Description
Id	int64	Unique id
Created	Object	Date of the tweet
Tweet	Object	Content of the tweet
Source	int64	From where it was originated
fav_count	int64	How many likes
Retweet	Int64	How many retweet
ABER	Int64	Away Bowling Economy Rate

Using Utilities function we derived following values

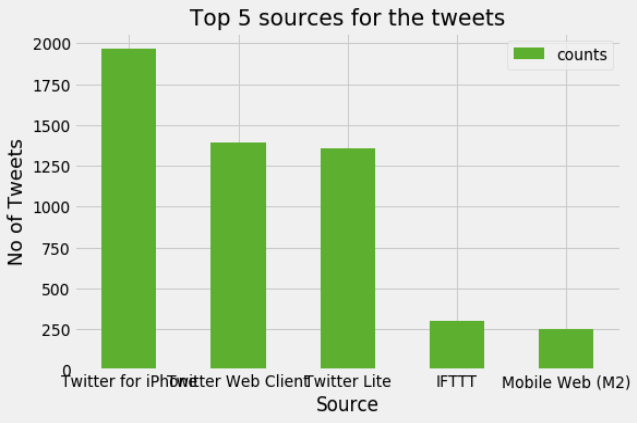
Field Name	Type	Description
human_date	Object	Friday 24, 2018
Month	Object	Month number
Day	Object	Day from the date
Day of Week	int64	Is it Monday, Tuesday etc?
Hour	int64	Time
len	Int64	Length of the tweet
Uppercase	Int64	Number of uppercase in tweet
hashtag_no	int64	Number of hashtags used in tweet
exclamation_no	int64	Number of hashtags used in tweet
question_no	int64	Number of question mark in tweet
mention_no	int64	Number of mentions in tweet
SA	int64	Sentiment analysis - 1, 0, -1

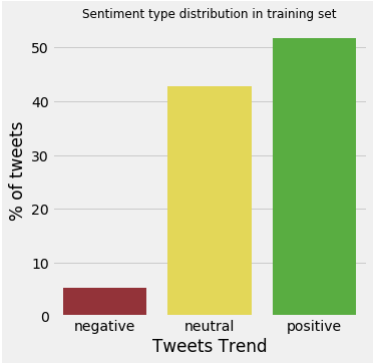
11. Exploratory analysis on tweets

Around 12k tweets were using the full length of 140 chars.



Top 5 sources are the usual suspects including IPhone and Android and based on the analysis more than 50% of the tweet are positive.





12. BAG OF WORDS

We derived the top 20 bag of words.

Word	Frequency	Word Encoded
#ipl	18663	b'#ipl'
:	16082	b':'
-	12543	b'-'
#ipl2018	7106	b'#ipl2018'
favorite	6883	b'favorite'
retweet	6766	b'retweet'
?	6754	b'?'
best	6163	b'best'
@iplcricket:	5264	b'@iplcricket:'
#vivoipl	4994	b'#vivoipl'
#psl	4794	b'#psl'
dhoni	4217	b'dhoni'
reply	4141	b'reply'
million	4114	b'million'
vs	3415	b'vs'
#csk	2802	b'#csk'
miller	2763	b'miller'
match	2658	b'match'
#rcb	2568	b'#rcb'

13. FORECASTING AND RESULTS

Random Forest classifier was used to train and test the tweets to predict the sentiment and below are the results.

Confusion Matrix

There are three possible predicted classes: whether the sentiment of the tweet is positive, negative or neutral.

	NEGATIVE PREDICTED	NEUTRAL PREDICTED	POSITIVE PREDICTED
NEGATIVE ACTUAL	2	8	2
NEUTRAL ACTUAL	0	47	2
POSITIVE ACTUAL	0	8	62

- The classifier made a total of 131 predictions
- Out of those 131 cases, the classifier predicted "positive" 64 times, neutral - 63 times and negative 2 times.
- In reality, there were 70 positive tweets, 64 neutral tweets and only were negative.
- So our model has predicted positive tweets correctly by 94%. Neutral tweets precision was predicted correctly 75% and negative tweets 100%.

	PRECISION	RECALL	F1-SCORE	SUPPORT
-1	1.00	0.17	0.29	12
0	0.75	0.96	0.84	49
1	0.94	0.89	0.91	70
AVG	0.87	0.85	0.83	131

14. Selection of playing XI

IPL franchises spend massive amounts of money is ensuring that they have the best batting and bowling options according to their game plan. An important problem is to select the playing XI from the available options.

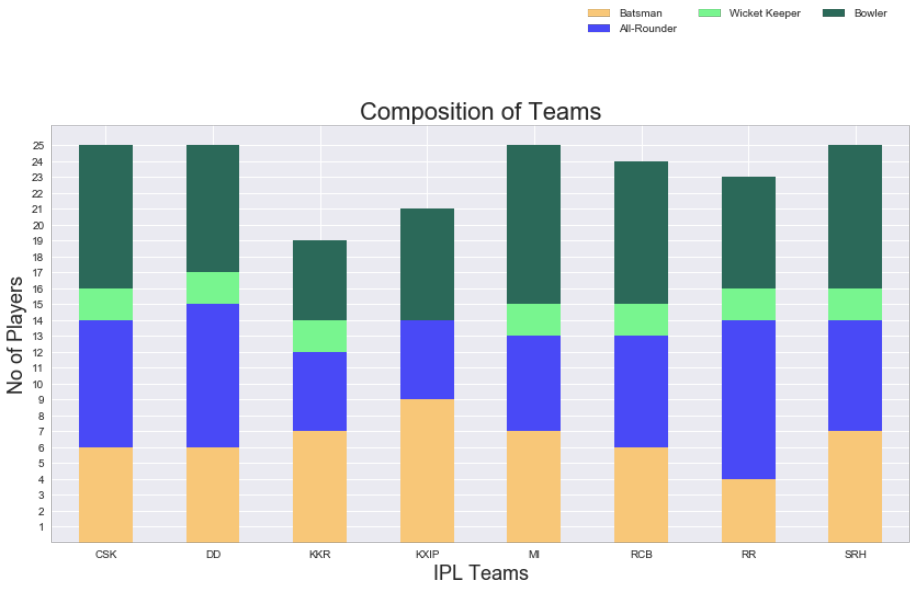
The present work focuses on Machine learning based data analytics to provide a good approach to solve this problem. A detailed performance ranking scheme is developed based on Random Forests Extra Tree Classifier to rank the players with respect to the other players in fray in IPL 11.

The ranking scheme provides percentile scores to the players for their batting and bowling performance and enables them to be compared against each other. The sum of the percentile scores for batting and bowling for the players in a given selection of Playing XI provide the fitness function for measuring the suitability of the team.

15. Interesting facts

G.Krishnappa who was a unknown name among Indian cricket fans jumped 3000% from his base price to earn millions in auction. Apart from that there were few surprises like Jaydav Unadkat getting 2 millions.

The visualization on composition of teams projects some interesting things like KXIP forgot to buy a wicket keeper in the auction. CSK, DD and RR are loaded with all-rounders.



16. New features

Eight new features were derived to evaluate the skills and form of batsman, bowler and all-rounders. There are few restrictions on this:

- Batsman and Bowlers were evaluated on their batting and bowling skills respectively.
- All-rounders were evaluated with their batting and bowling skills.
- The scored derived for each role must be compared with that role only. For example – A batsman with 70 score should not be compared with bowler having 100 points. Similarly you cannot compare all-rounder with bowler or batsman.

A batsman will be scored as per the below parameters

- Hard Hitter = (4*Fours* + *6*Sixes) / Balls faced by player
- Finisher = Number of times not out/Total number of innings played
- Fast Scorer = Total runs scored/ Total balls faced
  Consistent = Total runs scored/ Total number of innings in which he got out
- Running between wickets = (Total runs scored - (4*Fours* + *6*Sixes))/ Number of balls faced without boundary

A bowler will be scored as per the below parameters

- Economy= Total number of runs conceded / Total number of overs bowled
- Wicket Taker= Total number of balls bowled / Total number of wickets taken
- Consistent = Total number of runs conceded / Total number of wickets taken

An all-rounder will be calculated with both bowling and batting parameters

A wicket-keeper again will be calculated based on batting as each team does have only one or two keepers. So nothing more to select.

17. Modeling and results

RF Extra tree classifier was used to derive weightage for each features listed above. A function was used to derive points for batsman, bowler, wicket keeper and all-rounder using this weightage.

Then ran the formulas on couple of teams RCB and CSK to check whether it is yielding the correct results. As per the results like the one below, it did bring out the best batsman of the respective teams

player	player_score
SK Raina	30.638895
MS Dhoni	29.598371
Murali Vijay	20.720036
Faf du Plessis	14.048420
Kedar Jadhav	13.738108
Sam Billings	7.179614

player	player_score
Virat Kohli	29.228313
AB De Villiers	26.130119
Brendon McCullum	21.542071
Mandeep Singh	13.459726
Manan Vohra	11.987577
Sarfaraz Khan	9.253275

Similarly applied formulas for bowler, batsman and all-rounder and came out with the playing XI.

18. IPL Team selection using machine learning

Chennai super kings

Order	Player	Role	Derived Score
1	Shane Watson	All-Rounder	102.22
2	Murali Vijay	Batsman	26
3	Dwayne Bravo	All-Rounder	101.38
4	Faf du Plessis	Batsman	17.44
5	SK Raina	Batsman	38.54
6	MS Dhoni	Batsman	37.17
7	Kedar Jadhav	Batsman	17.11
8	R Jadeja	All-Rounder	106.44
9	Karn Sharma	Bowler	47.43
10	Harbhajan Singh	Bowler	112.08
11	Lungi Ngidi	Bowler	33.05

Delhi daredevils

Order	Player	Role	Derived Score
1	Prithvi Shaw	Batsman	8.37
2	Gautam Gambhir	Batsman	56.06
3	Rishabh Pant	Wicket Keeper	12.19
4	Shreyas Iyer	Batsman	14.24
5	Glenn Maxwell	Batsman	17.78
6	Chris Morris	All-Rounder	61.84
7	Shahbaz Nadeem	Bowler	53.31
8	Amit Mishra	Bowler	101.7
9	Trent Boult	Bowler	24.93
10	Mohammed Shami	Bowler	41.37
11	Kagiso Rabada	Bowler	17.4

Kolkata knight riders

Order	Player	Role	Derived Score
1	Sunil Narine	All-Rounder	86.92
2	Chris Lynn	Batsman	13.85
3	Robin Uthappa	Wicket Keeper	34.64
4	Nitish Rana	All-Rounder	12.24
5	Andre Russel	All-Rounder	46.51
6	Dinesh Karthik	Wicket Keeper	31.82
7	Ishank Jaggi	Batsman	21.07
8	Mitchell Starc	Bowler	29.43
9	Kuldeep Yadav	Bowler	21.65
10	Piyush Chawla	Bowler	99.39
11	Vinay Kumar	Bowler	83.96

MUMBAI INDIANS

Order	Player	Role	Derived Score
1	Evin Lewis	Batsman	13.47
2	Rohit Sharma	Batsman	37.54
3	JP Duminy	Batsman	25.2
4	Kieron Pollard	All-Rounder	82.84
5	Hardik Pandya	All-Rounder	45.47
6	Krunal Pandya	All-Rounder	43.24
7	Tajinder Dhillon	All-Rounder	25.13
8	Ishan Kishan	Wicket Keeper	9.23
9	Rahul Chahar	Bowler	13.8
10	Jasprit Bumrah	Bowler	48.46
11	Pat Cummins	Bowler	24.83

Rajasthan royals

Order	Player	Role	Derived Score
1	Darcy Short	All-Rounder	37.64
2	Rahul Tripathi	Batsman	11.06
3	Ajinkya Rahane	Batsman	29.04
4	Steve Smith	Batsman	22.39
5	Ben Stokes	All-Rounder	32.39
6	Sanju Samson	Wicket Keeper	18.92
7	Shreyas Gopal	All-Rounder	225.75
8	Ben Laughlin	Bowler	89.22
9	G Krishnappa	Bowler	29.7
10	Dhawal Kulkarni	Bowler	59.53
11	Jaydev Unadkat	Bowler	44.21

Kings xi punjab

Order	Player	Role	Derived Score
1	Chris Gayle	Batsman	30.43
2	Aaron Finch	Batsman	19.92
3	KL Rahul	Batsman/Wk	14.16
4	Yuvraj Singh	Batsman	28.65
5	Karun Nair	Batsman	16.68
6	Marcus Stoinis	All-Rounder	28.25
7	Axar Patel	All-Rounder	68.72
8	R Ashwin	Bowler	89.9
9	Ben Dwarshuis	Bowler	29.08
10	Mohit Sharma	Bowler	63.58
11	Barinder Sran	Bowler	22.79

Sunrisers hyderabad

Order	Player	Role	Derived Score
1	David Warner	Batsman	48.24
2	Shikhar Dhawan	Batsman	32.49
3	Manish Pandey	Batsman	25.57
4	Deepak Hooda	All-Rounder	34.29
5	Yusuf Pathan	Batsman	85.34
6	Shakib al Hasan	All-Rounder	55.36
7	W Saha	Wicket Keeper	22.19
8	Mohammad Nabi	All-Rounder	16.18
9	Bhuvneshwar Kumar	Bowler	76.87
10	Rashid Khan	Bowler	20.98
11	Sandeep Sharma	Bowler	52.29

ROYALS CHALLENGERS BANGALORE

Order	Player	Role	Derived Score
1	Brendon McCullum	Batsman	27.04
2	Parthiv Patel	Wicket Keeper	27.64
3	Virat Kohli	Batsman	36.72
4	AB De Villiers	Batsman	32.71
5	Sarfaraz Khan	Batsman	11.29
6	Pavan Deshpande	All-Rounder	43.81
7	de Grandhomme	All-Rounder	25.14
8	Washington Sundar	All-Rounder	18.22
9	N Coulter-Nile	Bowler	28.6
10	Yuzvendra Chahal	Bowler	50.85
11	Umesh Yadav	Bowler	78.33

19. CONCLUSION

So, I was able to generate all IPL teams from Machine Learning and looking at overall team and the restriction of 4 foreign players per team, 70% of each team matches with everyone's expectation. The rest of the 30% is where ML has helped out in picking the right players as they would have been under the cloud.

For example, Pavan Deshpande for RCB, Shreyas Gopal for RR, Ben Dwarshuis for Kings XI and Lungi Ngidi for CSK wouldn't been the obvious choice for the team selection committee unless backed by numbers. Shreyas Gopal and Pawan Deshpande are unknown players to the IPL audience and I would be fascinated to see whether these players will really make it to the playing eleven when the season starts on April 7th.

As far as I am concerned, since I am fanatic of cricket I will continue to collect, explore and learn more about the game through data. Some of the players doesn't have not enough records. I feel that addition of more data from Big Bash League, International T20, PSL, Caribbean League, our model will get better.