IPL Match Win Predictor

Chirag Pandya

Intro

Got interested in Machine learning and came across the IPL dataset.

IPL - most sought after cricketing league in India

Predicting outcome of a match always crucial

Predictors can help team managers make informed decisions

Methodology

Data Exploration

Ensure the data is clean and can provide insights without errors

Feature Engineering

Explore suitable features available within the data set

Calculate more features based on existing data for the model to make better decisions

Machine Learning

Experimenting with various machine learning algorithms for accurate predictions

Data Source

IPL dataset on kaggle: https://www.kaggle.com/manasgarg/ipl

The dataset contains 2 files: deliveries.csv and matches.csv. 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.

Research scope: Predicting the winner of the next season of IPL based on past data, Visualizations, Perspectives, etc.

The data was read into R using the read.csv command. The files were also combined by “match_id” for calculating certain features

Exploring and cleaning

Matches.csv

```
'data.frame':   577 obs. of  18 variables:
 $ id              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ season          : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
 $ city            : Factor w/ 31 levels "","Abu Dhabi",..: 4 8 11 23 22 17 15 9 15 8 ...
 $ date            : Factor w/ 407 levels "2008-04-18","2008-04-19",..: 1 2 2 3 3 4 5 6 7 8 ...
 $ team1           : Factor w/ 13 levels "Chennai Super Kings",..: 7 1 10 8 2 5 2 1 2 5 ...
 $ team2           : Factor w/ 13 levels "Chennai Super Kings",..: 12 5 3 12 7 10 3 8 10 8 ...
 $ toss_winner     : Factor w/ 13 levels "Chennai Super Kings",..: 12 1 10 8 2 5 2 8 10 8 ...
 $ toss_decision   : Factor w/ 2 levels "bat","field": 2 1 1 1 1 1 1 2 2 2 ...
 $ result          : Factor w/ 3 levels "no result","normal",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ dl_applied      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ winner          : Factor w/ 14 levels "","Chennai Super Kings",..: 8 2 4 13 8 11 4 2 11 6 ...
 $ win_by_runs     : int  140 33 0 0 0 0 0 6 0 66 ...
 $ win_by_wickets  : int  0 0 9 5 5 6 9 0 3 0 ...
 $ player_of_match: Factor w/ 188 levels "","A Chandila",..: 26 100 101 117 42 171 181 107 185 76 ...
 $ venue           : Factor w/ 35 levels "Barabati Stadium",..: 15 23 9 35 8 27 24 16 24 23 ...
 $ umpire1         : Factor w/ 43 levels "A Nand Kishore",..: 6 26 4 38 9 4 19 16 6 4 ...
 $ umpire2         : Factor w/ 45 levels "A Nand Kishore",..: 29 40 14 13 19 28 5 14 23 5 ...
 $ umpire3         : logi  NA NA NA NA NA NA ...
```

Exploring and Cleaning

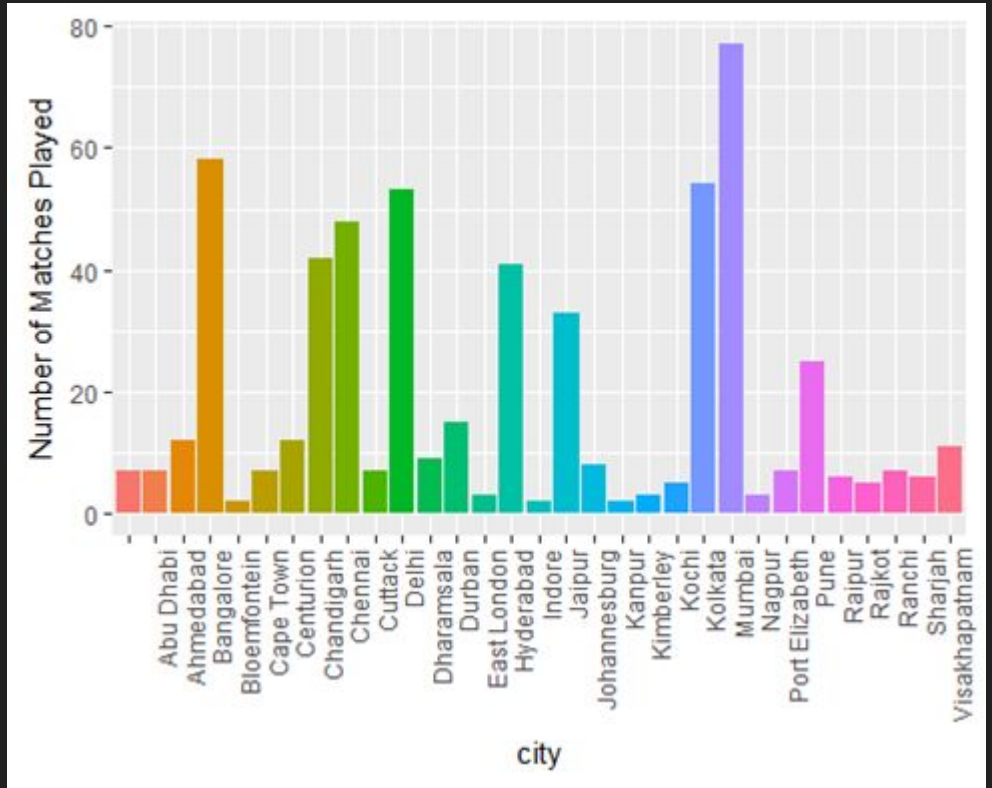
Deliveries.csv

```
'data.frame':   136598 obs. of  21 variables:
 $ match_id        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ inning          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ batting_team    : Factor w/ 13 levels "Chennai Super Kings",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ bowling_team    : Factor w/ 13 levels "Chennai Super Kings",..: 12 12 12 12 12 12 12 12 12 12 ...
 $ over            : int  1 1 1 1 1 1 1 2 2 2 ...
 $ ball            : int  1 2 3 4 5 6 7 1 2 3 ...
 $ batsman         : Factor w/ 436 levels "A Ashish Reddy",..: 354 61 61 61 61 61 61 61 61 61 ...
 $ non_striker     : Factor w/ 431 levels "A Ashish Reddy",..: 61 352 352 352 352 352 352 352 352 352 ...
 $ bowler          : Factor w/ 334 levels "A Ashish Reddy",..: 204 204 204 204 204 204 204 334 334 334 ...
 $ is_super_over   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ wide_runs       : int  0 0 1 0 0 0 0 0 0 0 ...
 $ bye_runs        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ legbye_runs     : int  1 0 0 0 0 0 1 0 0 0 ...
 $ noball_runs     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ penalty_runs    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ batsman_runs    : int  0 0 0 0 0 0 0 0 4 4 ...
 $ extra_runs      : int  1 0 1 0 0 0 1 0 0 0 ...
 $ total_runs      : int  1 0 1 0 0 0 1 0 4 4 ...
 $ player_dismissed: Factor w/ 413 levels "","A Ashish Reddy",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ dismissal_kind  : Factor w/ 10 levels "","bowled","caught",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ fielder         : Factor w/ 422 levels "","A Ashish Reddy",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Exploration and Cleaning - City Variable

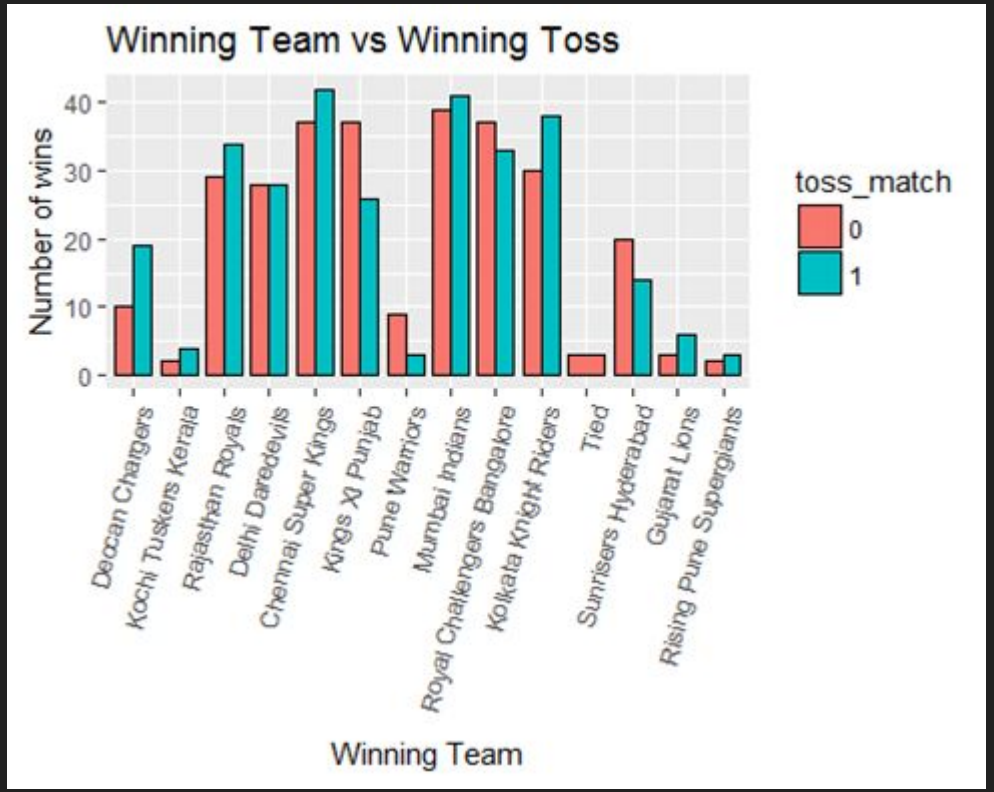
Cleaning the empty city variable

IPL 2009 was held in South Africa, hence the variance in plot



Exploration - To find important features

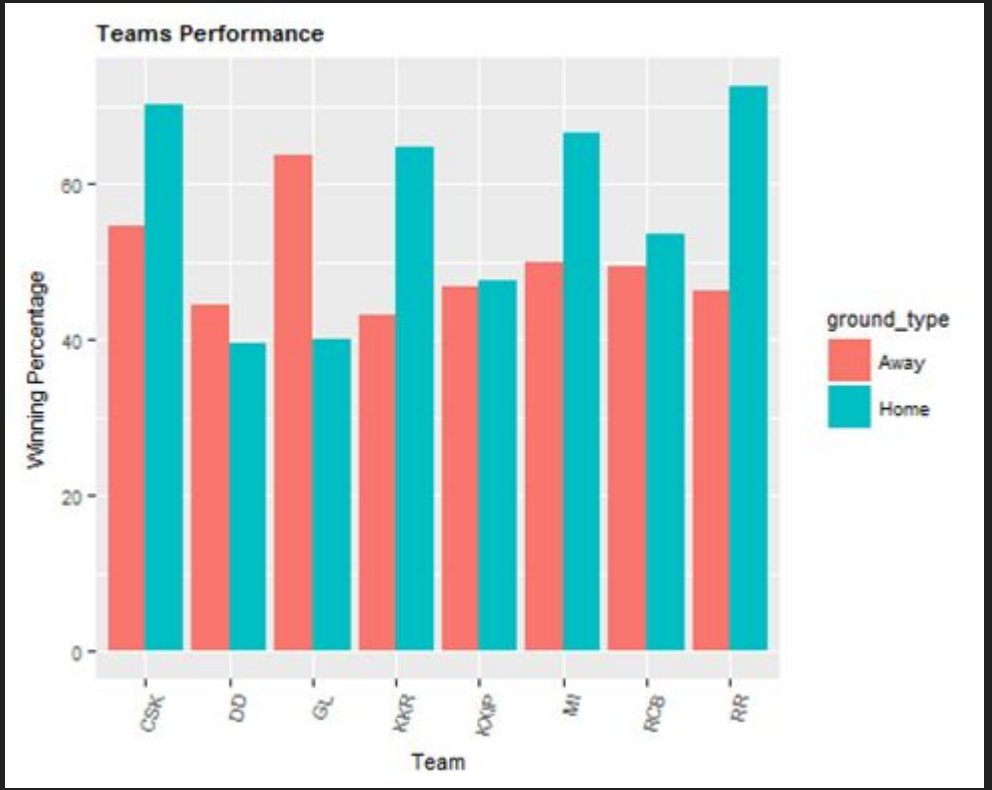
Effect of toss win on winning team



Exploration - To find important features

Home/Away ground advantage

Gujarat Lions - Less data hence the skew



Features list (Calcuated)

Name	Description
"t1ha"	Team 1 home ground = 1 away ground = 0
"t2ha"	Team 2 home ground = 1 away ground = 0
"t1tw"	Team1 toss win = 1 lose = 0
"t2tw"	Team2 toss win = 1 lose = 0
"t1twbf"	Team 1 toss win elected to bat = 1 field = -1
"t1twff"	Team 1 toss win elected to field = 1 bat = -1
"t2twbf"	Team 2 toss win elected to bat = 1 field = -1
"t2twff"	Team 2 toss win elected to field = 1 bat = -1

Features list (Calcuated)

Name	Description
"team1tm"	Team 1 total matches
"team1mw"	Team 1 matches won
"team1wp"	Team 1 win percentage
"team1wr"	Team 1 win record = matches won - matches lost
"team1rs"	Team 1 runs scored total
"team1rspi"	Team 1 runs scored per inning
"team1w"	Team1wickets total
"team1wpi"	Team1 wickets per inning

Features list (Calcuated)

Name	Description
"team2tm"	Team 1 total matches
"team2mw"	Team 1 matches won
"team2wp"	Team 1 win percentage
"team2wr"	Team 1 win record = matches won - matches lost
"team2rs"	Team 1 runs scored total
"team2rspi"	Team 1 runs scored per inning
"team2w"	Team1 wickets total
"team2wpi"	Team1 wickets per inning

Logistic Regression

Accuracy 62.5%

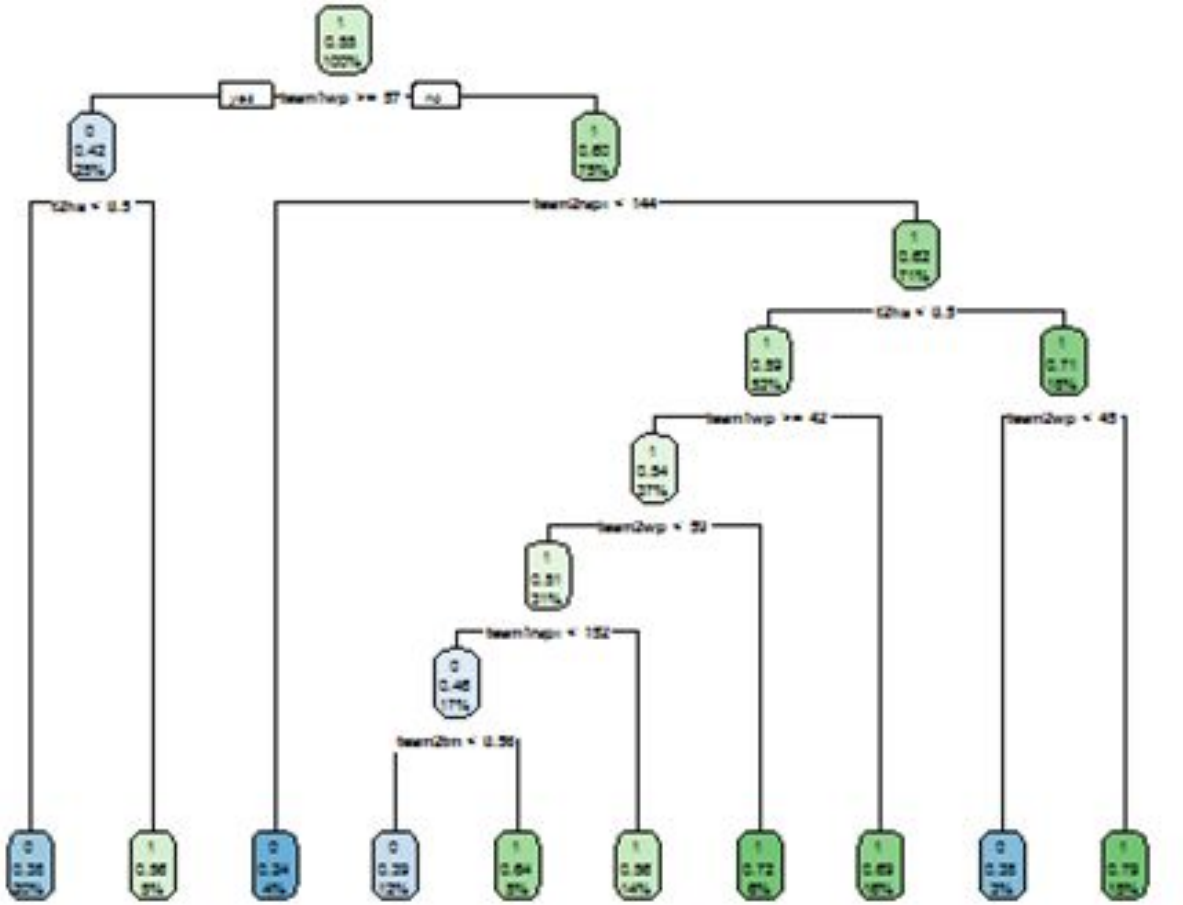
Confusion Matrix

	False	True
0	19	17
1	10	26

Decision Trees

Accuracy

58.33



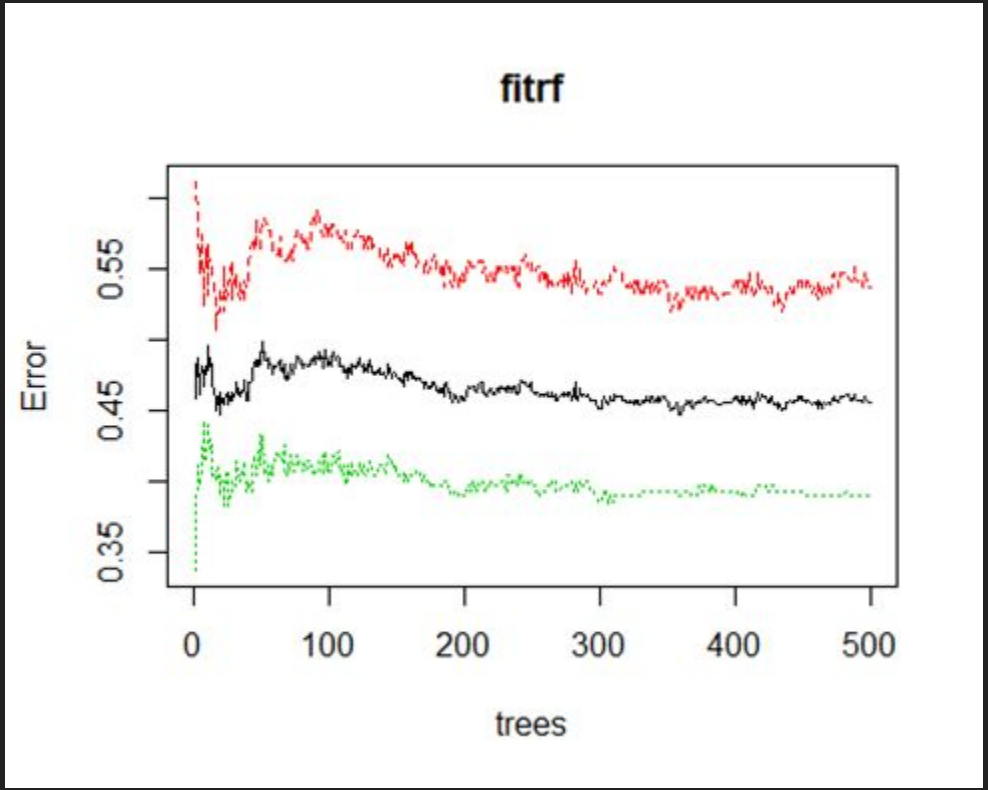
Random Forests

Accuracy: 51.33

Confusion Matrix

	False	True
0	16	17
1	20	19

Random Forests



Conclusion

Home ground advantage and runs scored per inning extremely significant

More features pertaining to strength of players in the team may improve predictions

Building a team strength index based on team player performance

Thanks!

Project Link:

https://github.com/chiragpandya88/IPL

Chirag Pandya
chiragpandya88@gmail.com