MACHINE LEARNING
IN

INDIAN PREMIER LEAGUE

SUNIL AYYAPPAN

THE KEY TAKEOUTS

01 PREDICTING HOME WIN

02 TWITTER SENTIMENT ANALYSIS

03 SELECTING PLAYING XI

RESTRICTIONS

PREDICTING HOME WIN

- 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 the 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 were 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.

DATA ACQUISITON

PREDICTING HOME WIN

The initial data set was taken from Kaggle. But I have added almost 50 columns to the original data. This was a long exercise as I had to open up each of the 450 matches and insert the new columns.

SELECTING PLAYING XI

The player auction data was collected from NDTV website. Again all the player statistics were collected manualy from Cricbuzz and Cricinfo website.

DATA CLEANING

- All team names were shortened for better visualization
- Most of the textual columns were converted to numbers. For example, home win and away win were converted to 1 or 0 and renamed the column as a Home win.
- Used stopwords to filter out some text for twitter sentiment analysis.
- For team selection, players who don't have records for a particular skill was filled with zero.
- Inconsistent column names.
- Teams changing names such as Rising Supergiant was changed to Rising Supergiants
- Ignoring matches which are not played on home ground. The 2013 season was ignored as it was played in South Africa. - Ignoring matches played between teams which are no longer part of IPL 2018.

DATA PRE-PROCESSING

PREDICTING HOME WIN

- From these datasets, features sets have constructed that form the input for our modeling.
- Some of the features were added using domain expertise.
- Statistical features were formed that represent the form or skill of a particular team.
- For each match, Home team and Away team were given points based on the number of batsmen, all-rounders, and bowlers in the team.
- For each batsman and bowler, 1 point is given as they have that particular skill. For an all-rounder, 2 points is given as they have two skills. So when a team has 4 batsmen, 3 allrounder and 4 bowlers the total points would be 4+(3*2)+4 = 14 points.

SELECTING PLAYING XI

- 5 features were added for calculating the batsmen's score.
- 4 features were added for calculating the bowler's skill

FEATURES PART ONE

PREDICTING HOME WIN

- Away_team
- HTBA - Home team # of batsman
- HTAL - Home team # of all-rounders
- HTBO - Home team # of bowlers
- HTSC - Home team score
- HTOV - Home team overs played
- HTWK - Home team wicket
- ATBA - Away team # of batsman
- ATAL - Away team # of all-rounders
- FTR - Full Time Result (H or A)

ATBO - Away team # of bowlers
ATSC - Away team score
ATWK - Away team wicket
N - 0 for Night games, 1 for Day games.
HTB - Home Team Batting (1 or 2)
ATB - Away Team Batting (1 or 2)
HTP - Home team points based on the team composition
ATP - Away team points based on the team composition

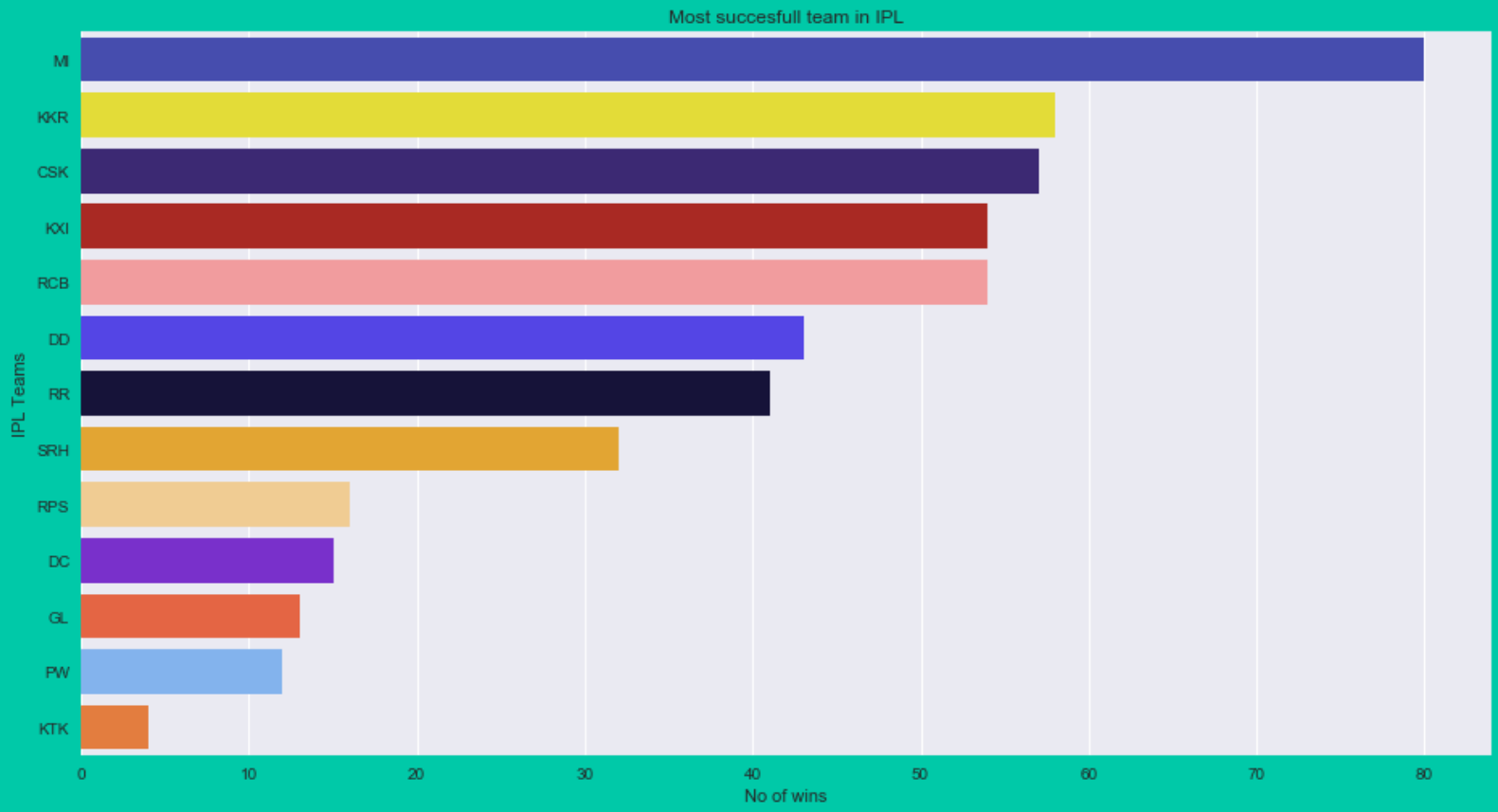
FEATURES PART TWO

SELECTING PLAYING XI

- **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
  - FastScorer = 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
  - WicketTaker= Total number of balls bowled / Total number of wickets taken
  - Consistent = Total number of runs conceded / Total number of wickets taken
- **An allrounder 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.**

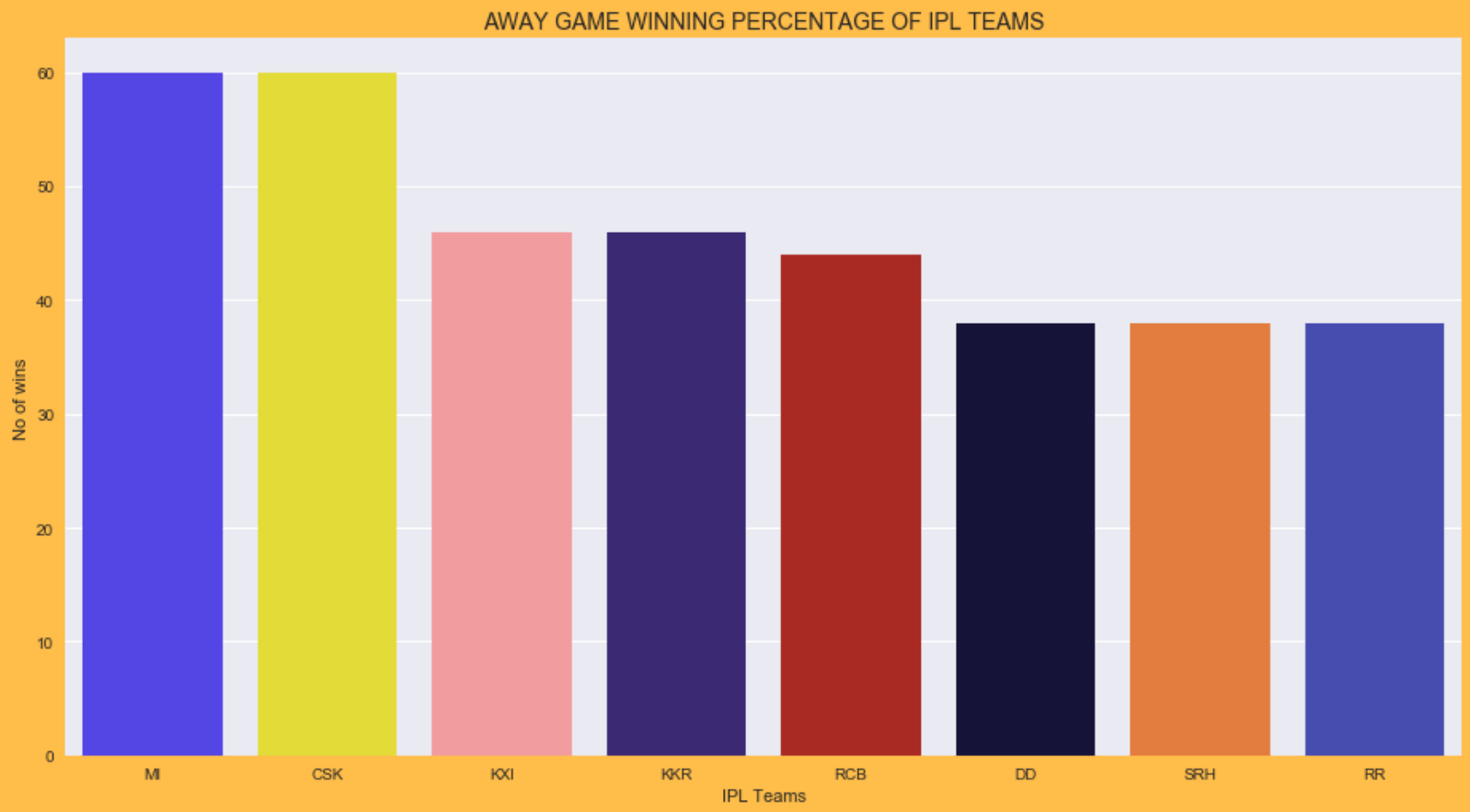
DATA EXPLORATION

MOST SUCCESSFUL TEAM IN IPL



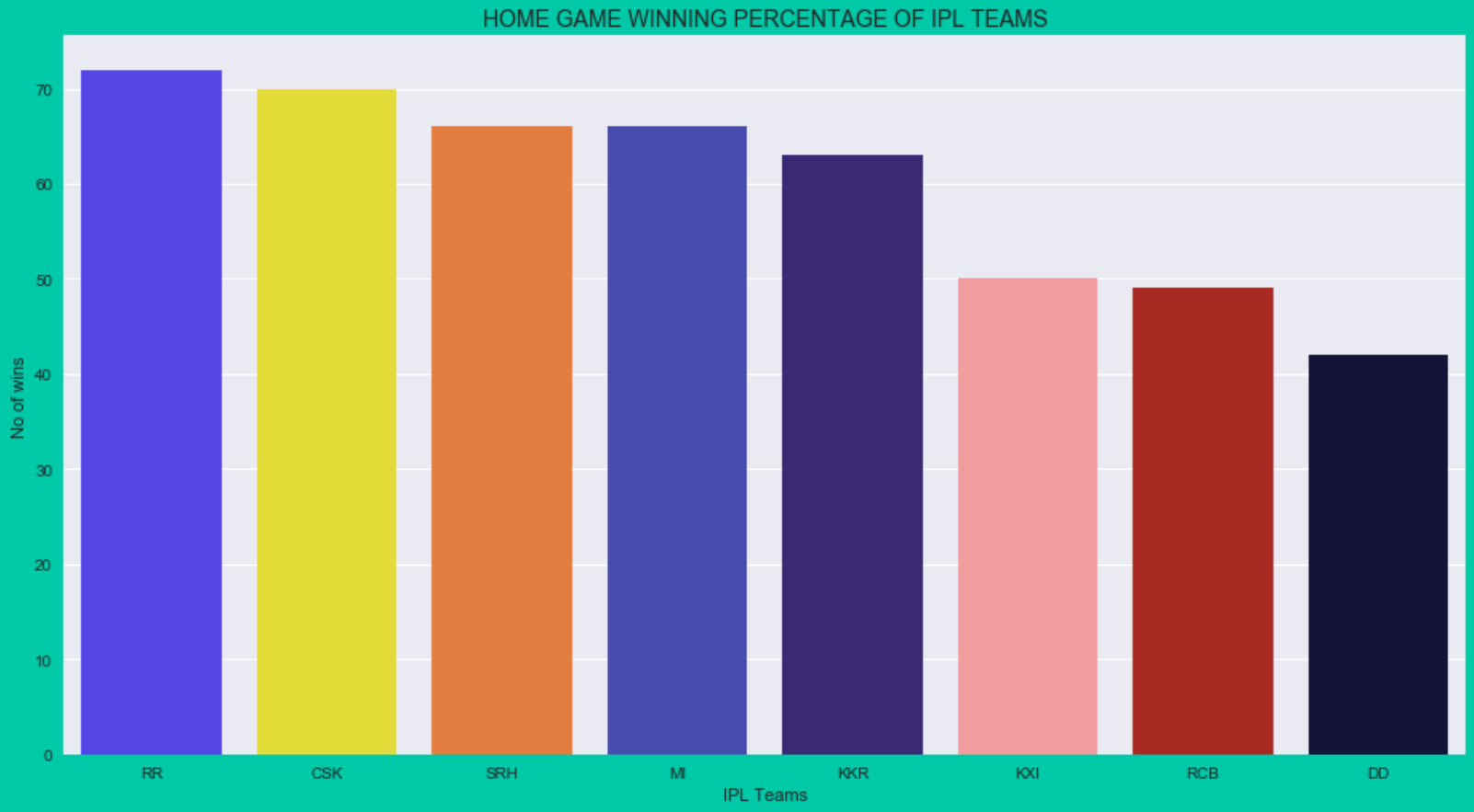
DATA EXPLORATION

PREDICTING AWAY GAME WINS



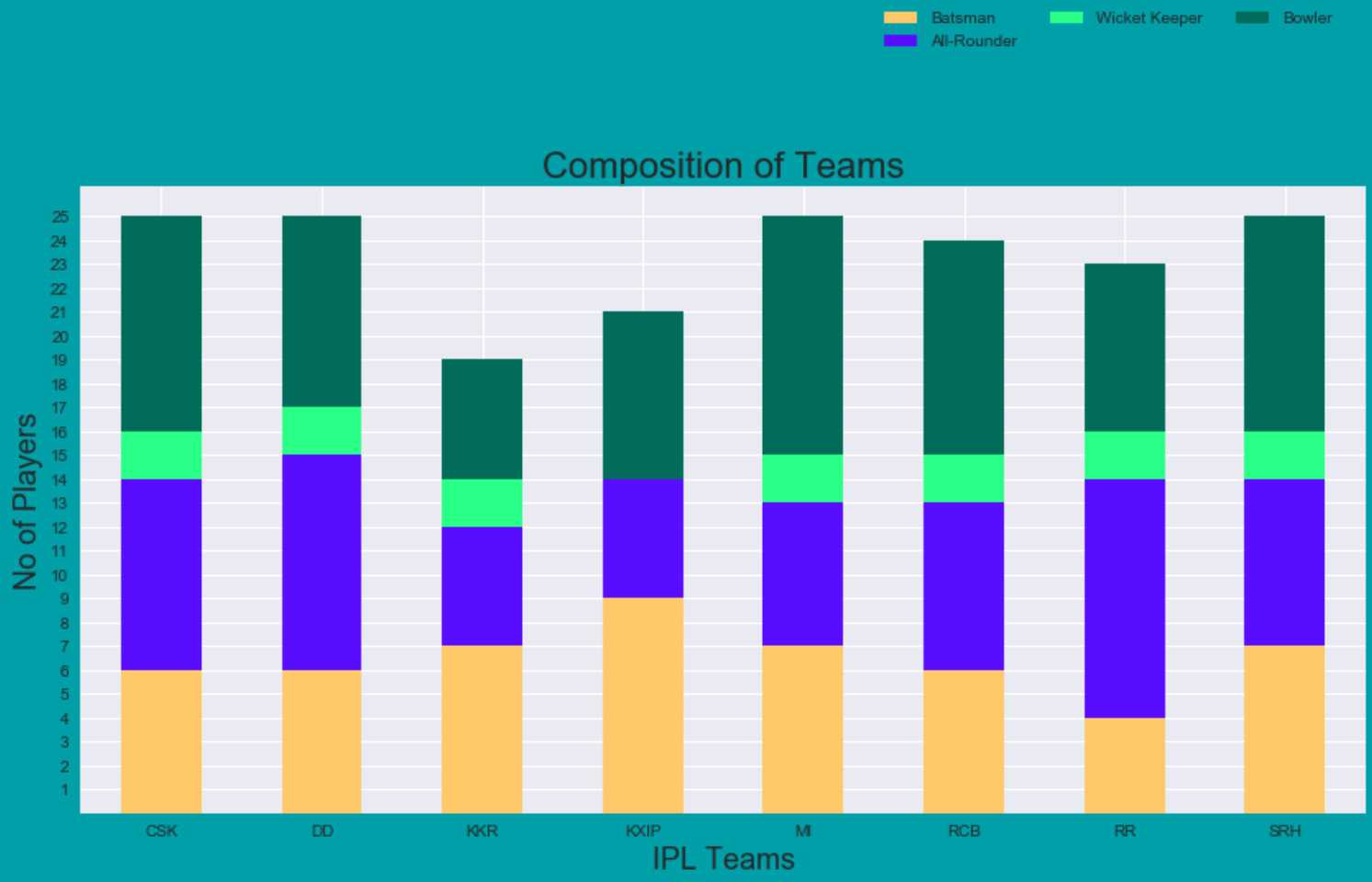
DATA EXPLORATION

PREDICTING HOME WINS



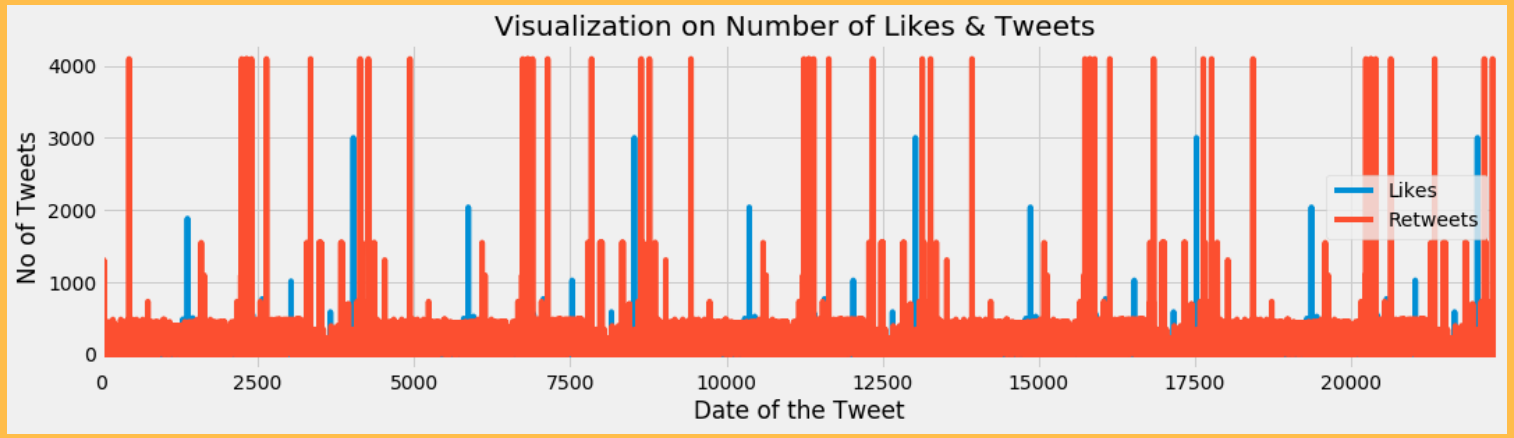
DATA EXPLORATION

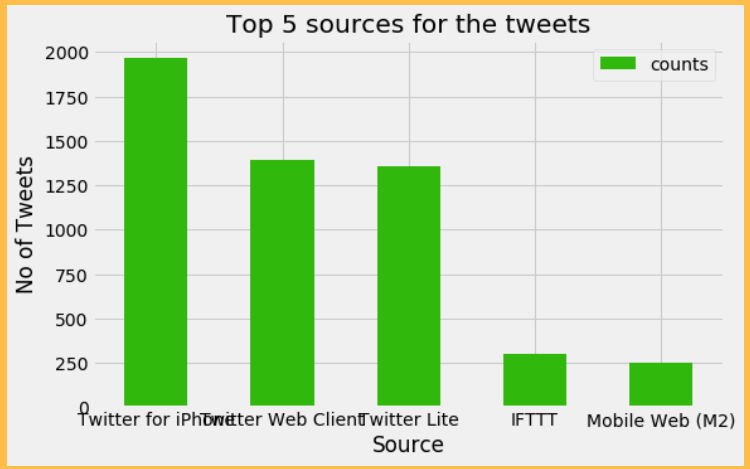
IPL TEAM COMPOSITION FOR 2018

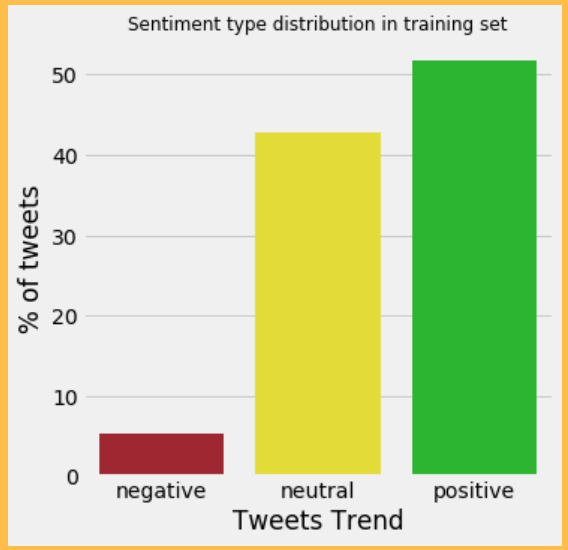


DATA EXPLORATION

TWEETS ANALYSIS







DATA EXPLORATION

TOP 30 PLAYERS WHO GOT PAID MORE IN IPL AUCTION 2018

MODELING

ALGORITHM USED FOR PREDICTING HOME WIN

- Gradient Boost Regressor
- Random Forest Classifier
- Random Forest Regressor

TWITTER SENTIMENT ANALYSIS

- Random Forest Classifier

SELECTING PLAYING XI

- Extra Tree Classifier for Feature selection.

PREDICTING HOME WIN

Based on the test results, both Random Forest classifier and Gradient boosting regressor yielded the best results.

	estimator	min_score	mean_score	max_score	std_score	alpha	learning_rate	max_depth	max_features	min_samp
111	RFC	0.890909	0.903331	0.918182	0.0112658	NaN	NaN	6	0.1	5
108	RFC	0.872727	0.903303	0.927273	0.0227528	NaN	NaN	6	0.1	3
109	RFC	0.881818	0.900246	0.918919	0.0151473	NaN	NaN	6	0.1	3
110	RFC	0.872727	0.897243	0.90991	0.0173382	NaN	NaN	6	0.1	3
81	RFC	0.881818	0.897243	0.90991	0.011633	NaN	NaN	4	0.1	3
112	RFC	0.863636	0.897215	0.918919	0.0240806	NaN	NaN	6	0.1	5
99	RFC	0.881818	0.894267	0.909091	0.01126	NaN	NaN	6	0.3	3
73	RFC	0.881818	0.89424	0.900901	0.00879103	NaN	NaN	4	0.3	3
104	RFC	0.881818	0.89424	0.900901	0.00879103	NaN	NaN	6	0.3	5
84	RFC	0.863636	0.894212	0.90991	0.0216231	NaN	NaN	4	0.1	5
74	RFC	0.881818	0.894212	0.90991	0.0117038	NaN	NaN	4	0.3	3
77	RFC	0.881818	0.894212	0.90991	0.0117038	NaN	NaN	4	0.3	5
106	RFC	0.881818	0.891182	0.90991	0.0132426	NaN	NaN	6	0.3	9
107	RFC	0.881818	0.891182	0.90991	0.0132426	NaN	NaN	6	0.3	9
76	RFC	0.872727	0.891182	0.90991	0.015181	NaN	NaN	4	0.3	5
113	RFC	0.854545	0.891155	0.918919	0.0270144	NaN	NaN	6	0.1	5
72	RFC	0.882883	0.888234	0.890909	0.00378359	NaN	NaN	4	0.3	3

RESULTS ON TWEETS

CONFUSION MATRIX ON TWEETS TEST RESULTS

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

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

	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

SELECTING PLAYING XI

PLAYING XI DERIVED FOR CSK USING MACHINE LEARNING

ORDER	PLAYER	ROLE	POINTS
1	Shane Watson	All-Rounder	102.22
2	Murali Vijay	Batsman	26.00
3	Dwayne Bravo	Batsman	101.38
4	Faf Du Plessis	Batsman	17.44
5	SK Raina	Batsman	38.54
6	MS Dhoni	Batsman/WK	37.17
7	Kedhar Jadhav	Batsman	17.11
8	R Jadeja	All-Rounder	106.44
9	Karn Sharma	Bowler	47.43
10	Harbhajan Singh	Bowler	112.06
11	Lungi Ngidi	Bowler	33.05

THINGS TO REMEMBER:

- Each player was ranked amongst their skill.
- All-rounder's score was derived using their bowling and batting skills.
- There is a 4 foreign player restrictions for each team.
- This point looks good for players who are new to IPL but has played a good amount of local tournaments or international matches.

IF YOU HAVE ANY QUESTIONS

HOLLER

SUNILAYYAPS@GMAIL.COM