Proposal

Prediction of Outcome of an IPL Match Winner

Clients

Our client is Cricket team Managers,Coaches,Analysts or sports score update and news websites like cricbuzz,cricinfo,sportskeeda etc. They will plan their strategies and team accordingly

Datasets used

We are using two datasets. One(**matches.csv**) which contains all the IPL matches results etc from the year 2008-2019. And Another one **(deliveries.csv**) which contains ball by ball details for every ball bowled from IPL 2008 to 2019. We got the data from kaggle platform

Approach

We will solve this problem using different features like Toss,Average First innings scores and Average first innings winning scores,Performance of various teams at Home and Away venues, No of boundaries hit by each team and does it affect their winning percentages.This is supervised problem as we will be using the past data to predict the match winners and this is a classification problem in which we will predict who will the match out of the two teams. We will be using **100 % of the set as training data** and then randomly choosing the 20% of the data out of this whole dataset and applying the different machine learning algorithms to predict who wins the match. Using Relative Analysis as % as the No of matches palyed by each team is different

We will be submitting the IPYTHON Notebook , Presentation and report on Github.

DATA WRANGLING

1.Cleaning Steps

- Filtered most important columns needed for analysis,removed umpire 1,umpire 2,umpire 3 ,eliminator .
- Also removed duplicate columns such as result and winner and combined them into one column winner
- Changed the ‘date’ column to datetime type in matches dataset
- Combined ‘Delhi Capitals’ in the same column as ‘Delhi Daredevils’
- Combined ‘Deccan Chargers’ in the same column as ‘Sunrisers Hyderabad’
- Changed ‘Rising Pune SuperGiants’ to Rising Pune SuperGiants’
- Changed Name of each IPL team and named them by their initials.

2.Dealing with missing values, if any?

- Filled ‘city’ names blank values with ‘Unknown’
- Filled ‘dl_applied’ Blank values to ‘No’
- Filled unknown player_of_matches with 'Not Available'
- Removed eliminator column after storing its appropriate values in winner column
-

3.Handling Outliers

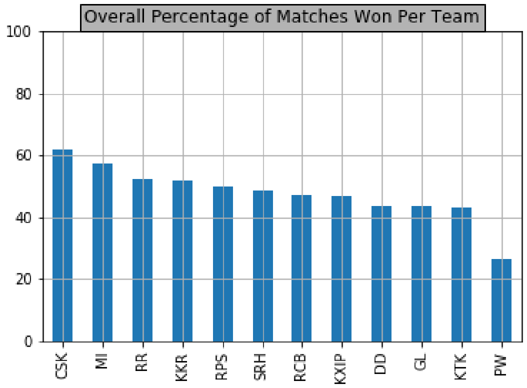
- Drop’ no result’ columns as it affected the statistics.
- There were six outliers
- Kept the ‘win_by_runs’ ,win_by_wickets’ ,’first_bat_score,’second_bat_score’ as it as filling them will affect the statistics.

DATA STORY

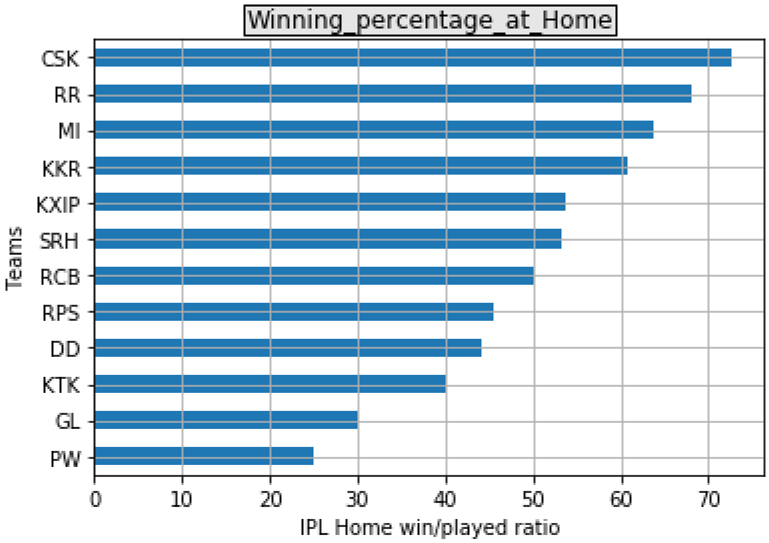
1. **Ask the following questions and look for the answers using code and plots:**
    1. **Can you count something interesting?**

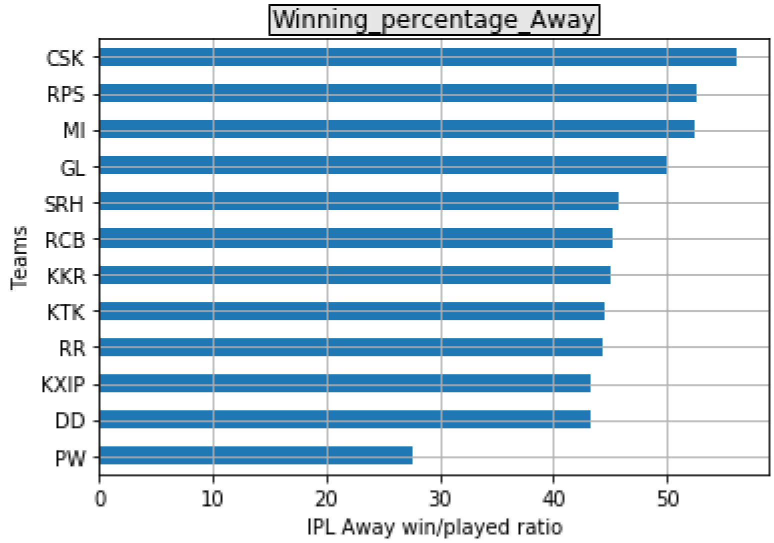
       Total matches played per IPL season over the 12 seasons of IPL,Total

Matches played and Won by each team, Overall Percentage of Matches Won by each team (Chennai Super kings Tops the List),Away and Home Venues Winning Percentage.



Chennai Super Kings is the most successful Home and Away side of the IPL with the highest percentage of Win in Away and Home Grounds .





While Delhi is the least successful side with the worst Home and Away records in the IPL played so far considering only the playing teams so far.(PW,RPS,GL,KTK not playing)

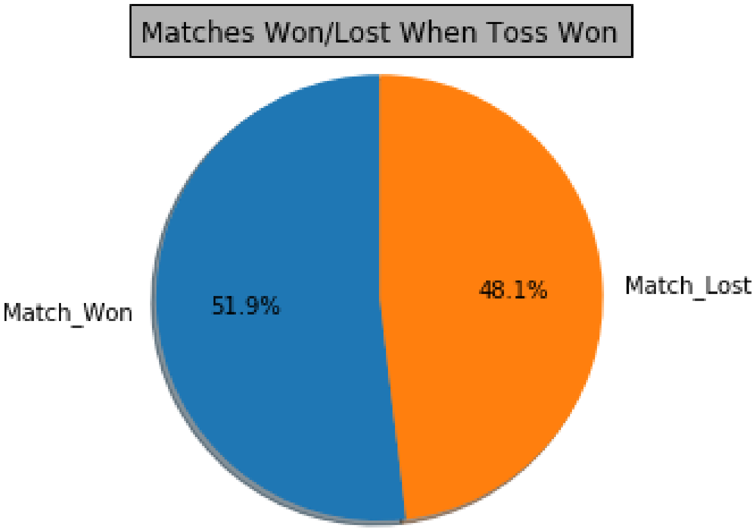
The Percentage of Overall win/played ratio is also the best for Chennai super king and worst for Delhi daredevils.

2. **Can you compare two related quantities?**

Compared and checked if there is any correlation between winning toss

And winning the match, Turns out there is not much correlation, the

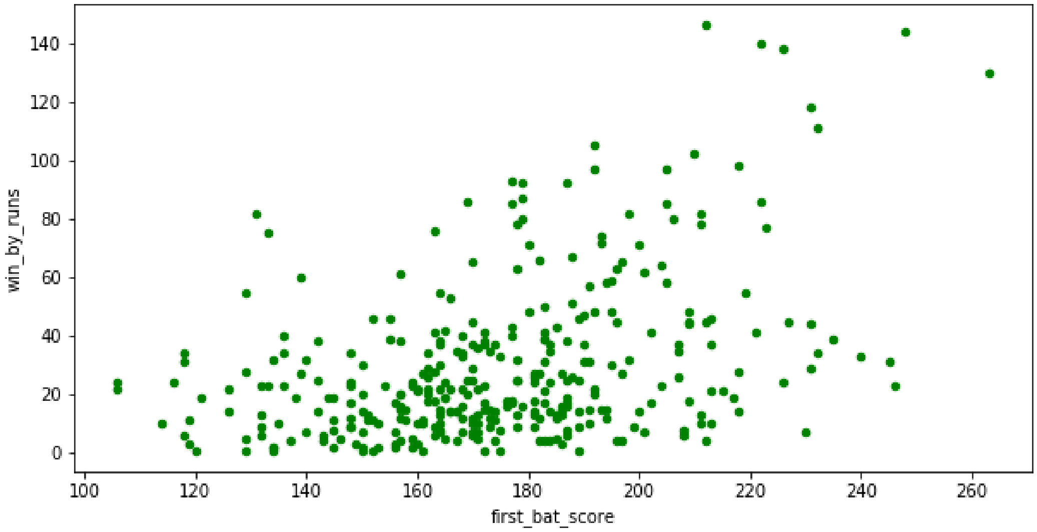
Chances of winning a match on winning or losing the toss are still close to 50%.



3. **Can you make a scatterplot?**

   To Find the Correlation between the Winning by_runs/Winning by Wickets and First bat_score

   It



   It looks like there is +ve correlation, but there is moderate correlation.

4. **Can you make a time-series plot?**

   Instead Plotted a Heatmap plot to check the average first innings score

   Of every team over different seasons of the IPL and the average first

   innings score of Each IPL season.

2. **Looking at the plots, what are some insights you can make? Do you see any correlations? Is there a hypothesis you'd like to investigate further? What other questions do the insights lead you to ask?**

- Chennai Super kings is the most consistent team of the IPL, while Delhi daredevils is Least consistent team of the IPL so far.
- Chennai Super kings is best side batting first while Mumbai Indians is the best Side chasing .
- Average first innings winnings score for Royal Challengers Bangalore is highest(184) , that means they need to score at least 184 to win if batting first.
- Average first innings they need to restrict the Opponent is highest for Gujarat Lions(164),: that means they need to restrict the Opposing team to at below or equal to 164 to successfully defend the total and win the match
- Most Successful team on winning the toss is Gujarat Lions, they win 65% of there matches if they win the Toss.

- The most successful team on losing the toss is Kings XI Punjab with a win ratio of 59% if they lose the Toss,Amazing right.

3. **Now that you've asked questions, hopefully you've found some interesting insights. Is there a narrative or a way of presenting the insights using text and plots that tells a compelling story? What are some other trends/relationships you think will make the story more complete?**

   Yes,Will be conveying the same through a Presentation.

   Yes, Need to check what is the winning percentage when teams are playing

   against each other and Is there any Trend that One team gets the better of the better

   Than most of the time.

DATA STORY (Continued)

1.Are there variables that are particularly significant in terms of explaining the answer to your project question?

- Toss_Winner
- First Innings_ Score
- Winning by Runs/Wickets
- Venue (Home/Away)
- No of Boundaries Hit
- Winning Percentage at Home and Away per Team

2.Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

- Moderate Positive Correlation(0.6) of Most Boundaries hitting Team hit winning the most no of Matches.
- Positive Correlation of 0.41(**between first_innings_score and win margin by runs**)shows as the Team score higher first innings scores the margin(win_by_runs) of winning also increases
- Negative Correlation of -0.29 (**between first_innings_score and win margin by wickets**) Shows as the First innings score increases, teams lose more wickets and win by_wickets decreases.

3.What are the most appropriate tests to use to analyse these relationships?

- Using Permutation tests to analyze the data by using different cases to support the above null hypothesis.
- Using frequency Interference and Bootstrap tests,
- The 95% Confidence Interval for first innings winning score are (166.43900536980695, 174.53069159989005) and [166.55284091, 174.58333333] respectively.

Building a Predictive Model

- ❖ **Find the right set of features and the most important features using Select KBest and ExtraTreesClassifier respectively**
  - ➢ 'Team1' - Who is the first team?
  - ➢ 'Team2' - Who is the second team?
  - ➢ 'Final_Home_winner' - Customized feature created using teams's Overall Winning %,Home and away winning percentage,Toss_win_Match_Win%,Toss_Loss_Match_Won%+matches_won_batting_first_%+matches_won_batting_second_%.
  - ➢ 'Toss_winner' - who wins the toss?
  - ➢ 'first_bowl_team - First Bowling Team
  - ➢ 'first_batting_team' - First_batting_team
- ❖ **Target Variable: 'Match_Winner'**
  - ➢ Predicting the match winner
- ❖ Labeling the Features and Target Variables
  - ➢ use LabelEncoder()
  - ➢ Using replace()
- ❖ Splitting the data into training_set=70% and testing_set=30% and equaling dividing the data using stratify=y as below
  - ➢ # Split into training and test set
  - ➢ X_train, X_test, y_train, y_test = train_test_split(X, y, test_size =0.3,random_state=42,stratify=y)

- We will be using supervised learning algorithms as our target labels are known - the match_winner. And we will be using different classification machine learning techniques by following the steps below
    - **Instantiating a Classifier**
    - **Fitting to the training set**
    - **Prediction the training set**
    - **Finding the accuracy score on the testing data**
    - **Cross-Validation Score**
    - **Finding the Classification report(if needed)**
    - **Tuning the classifier with different hypertuning parameters**

- **Different Supervised Machine learning techniques used**
    - Decision Tree
        - Default Parameters
        - Tuned with Randomized Search CV for Max_features,Minimum_Samples_leaf on the basis of Gini impurity and Entropy criterion
        - OneVsRestClassifier
    - Random Forest Classifier
        - Default parameters
        - Tuned with best n_estimators
        - OneVsRestClassifier
    - SVM(Support Vector machine)
        - Default parameters
        - Tuned with C and gamma Parameters

We also tried Logistic Regression but it did not converge due to Muti Class and Imbalanced Data.
Please find the result on the next page for Each Machine learning algorithms

RESULTS

Classifier	Tuned/Untuned /Method	Best_Parameters	Accuracy(%)
Decision Tree	**Untuned**		**53.69**
	Tuned	{'min_samples_leaf': 5, 'max_features': 6, 'max_depth': None, 'criterion': 'gini'}	**54.04**
	OnevsRestClassifier		**51.76**
Random Forest	**Untuned**		**57.04**
	Tuned	no of estimators: 22	**59.73**
	OnevsRestClassifier	no of estimators: 3	**61.74**
Support vector machine(SVM)	**Untuned**		**45.55**
	Tuned	Tuned Model Parameters: {'C': 100.0, 'gamma': 0.01}	**51.00**

As far as we can see from the following table,
Random Forest Classifier with Stratify Split performs the best with an accuracy of **approximately 62%** using OnevsRestClassifier and no of estimators as 26 using the Cross_Validation score as the Metric used.