PREDICTIVE ANALYICS ON IPL DATASET

Predictive Analytics

Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future.

STATISTICAL ANALYSIS :

A subdomain of Predictive Analysis is the study of collection, organization, analysis, interpretation and presentation of data. It begins with the identification of the process or population in consideration. The population is collection of process at various times known as the time series and data from each of the observation serves as a member of overall group.

Descriptive Analysis: It summarizes the population data in consideration by describing what was observed in the sample graphically or numerically. Numerical descriptiors are mean and standard deviation for continuous data types. Frequency and percentage are more useful for describing categorical data.

To draw inference about the population represented Inferential Statistics uses patterns in the simple data

PREDICTIVE ANALYSIS PROCESS:

1.DEFINE PROJECT

2.COLLECTION OF DATA

3.DATA ANALYSIS

4.STATISTICS

5.MODELLING

6.DEPLOYMENT

1.DEFINITION OF THE PROJECT:

A Predictive/Statistical Analysis project that determines the outcome of an IPL Match. There are various variables that have taken into consideration Toss, Batting First, Fielding First, Venue, Batsmen Strike rates, Bowling economy. etc. The project includes two programming language:

Python Programming Language has been used for Collection of Data, Data Manipulation and Stastical Modelling.

C Programming Language has been used for Deployment, a menu based system/interface which when provided the exact details of the match, including the two teams, venue, toss result. etc. Predicts the outcomes of the match in no time.

There were plenty of datasets available on the Internet, however we chose the most reliable Dataset available , they were from www.kaggle.com which included :

1) Matches.csv (A csv file of size 106.00 kB)

2) deliveries.csv (A csv file of size 14 MB)

The matches.csv includes data under the following variables:

Id, season, city, date, team1, team2, toss_winner, toss_decision, result, dl_applied, winner, win_by_runs, win_by_wickets, player_of_match, venue, umpire1, umpire2, umpire3.

The deliveries.csv includes data under the following variables :

match_id, inning, batting_team, bowling_team, over, ball, batsman, non_striker, bowler, is_super_over, wide_runs, bye_runs, legbye_runs, noball_runs, penalty_runs, batsman_runs, extra_runs, total_runs, player_dismissed, dismissal_kind, fielder.

2.DATA COLLECTION

Data Mining along with Data Analysis prepares the data from multiple sources for analysis. This provides a complete overview and a thorough understanding of the data.

Under the umbrella of Data Manipulation following things were done:

---- Addition of a Column in the matches.csv to figure out which type of match is it, a pre-qualifier, a qualifier, a semifinal, a quarterfinal or a final. For this purpose, every last match of the season was given the value of 'final' under the newly created "type" column and henceforth for all the matches it was added.

---- Addition of "team score" and "team extra" columns for each match and each inning in the match.csv file.

---- Batsmen Aggregates for each match, each inning. Using the deliveries.csv file we have made a table called as batsmen, which has match by match and further inning by inning gives the details about the batsmen in that match (batsman name, runs, balls faced,4s,6s, SR, dismissal_kind, fielder).

---- Bowler Aggregates for each match, each inning. Using the deliveries.csv file we have made a table called as bowlers, which match by match and further inning by inning gives the details about the bowlers in that match (bowler name, over, wide_runs, noball_runs, runs,extras, wickets, Econ).

3.DATA ANALYSIS ,STASTISTICS AND MODELLING

Data analysis, also known as analysis of data or data analytics, is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

The Libraries imported for the Data Analysis and visualization are:

```
import numpy as np # linear algebra

import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline
```

1) Number of matches each season:

Let us first look at the number of matches played per season.

----sns.countplot(x='season', data=match_df)

plt.show()---

There is not much to tell, seeing the graph of the number of matches played per season, one can just tell that there is a spike for three years 2011, 2012, 2013 where the number of matches were more than 70, otherwise for most of the seasons the number of matches have been around 60.

2) Number of matches in each venue:

----plt.figure(figsize=(12,6))

sns.countplot(x='venue', data=match_df)

plt.xticks(rotation='vertical')

plt.show()----

It can be clearly seen from the graph that M Chinnaswamy Stadium in Bangalore is the place where most number of IPL Matches have taken place, closely followed by Feroz Shah Kotla in Delhi and Eden Gardens in Kolkata. There are also small spikes for matches that took place when IPL was shifter abroad like Sharjah Cricket Stadium, Green Park .etc.

3) Number of matches played by each team:

----temp_df = pd.melt(match_df, id_vars=['id','season'], value_vars=['team1', 'team2'])

plt.figure(figsize=(12,6))

sns.countplot(x='value', data=temp_df)

plt.xticks(rotation='vertical')

plt.show() ----

The number of matches played by each team is in direct corresponds to their presence in each IPL Season, the 7 teams which have been present right from IPL season 1 have each played around 120-140 matches depending on their performances. Mumbai Indians have played the most number of matches very closely followed by Royal Challengers Bangalore. The new teams have played extremely few matches till now like Rising Pune Giants and Gujarat Lions.

Henceforth this visualization tells us the fact, that since all the teams have not played equal amount of games we have to make sure that each inference of Statistical Analysis on the matches data for each team, we should also Normalize the data every time to make sure that no team gets unfair advantage.

It also tells us that since few teams have played very few matches, the chances of errors for such team would be a bit high, since the Data for such teams is quite dumb so to speak.

4) Number of wins per team:

----plt.figure(figsize=(12,6))

sns.countplot(x='winner', data=match_df)

plt.xticks(rotation='vertical')

plt.show()----

It can clearly be seen that Mumbai Indians lead the pack closely followed by Chennai Super Kings, however again it is important to Normalize the data, since it is unfair to award Gujarat Lions with such a less amount of wins because they have hardly played any matches yet. Hence we should focus of Wins percentage per team.

5) Win Percentage for each team:

---- The code due to the longer length has been excluded from the report, please refer the other file for the code ----

The "Winning Percentage" bar graph clearly states that Chennai Super Kings has best Winning Percentage of about 60% followed by Mumbai Indians with 57%, hence our normalized data provides much better insights of the data. Pune Supergiants and Delhi Daredevils are among the teams who have the least winning percentage with 35% and 42% respectively.

6) Toss

Almost 55% of the toss decisions are made to field first. Now let us see how this decision varied over time.

7) Season by Season Toss Decisions:

It seems during the initial years, teams wanted to bat first. Voila.! Look at the 2016 season, most of the toss decisions are to field first.

Since there is a very strong trend towards batting second let us see the win percentage of teams batting second.

8) Toss influencing the Team Winning:

Here we have compared two scenarios, that :

a) if the team has won the toss how likely it is that they will go on further to win the match also

b) if the team has lost the toss how likely it is that they will go on further to win the match

We have plotted two separate bar graphs and a combined bar graph for the comparison.

The Visualization depicts that for some teams like Gujarat Lions Toss has a huge influence, e.g. 75% of the match they end up winning once they have won the toss and on the other hand only 37% of the match they win once they lose the toss. On the other hand, for Sunrisers Hyderebad if they lose the toss they have better chance of winning (62%).

Now let us see once the teams win the toss, which decision is most likely.

9) Toss decision percentage:

Here we have plotted a pie chart, that indicates that once the teams elect to field first they end up winning 54.6% of the times. Henceforth, Fielding First looks a better option according to Statistical Measures.

Now let us analyse how this pattern been for each season.

10) Toss decision percentage season by season:

A double bar graph indicating the toss decisions by teams over the seasons, which clearly shows that the fielding first trend has only been popular for the previous three seasons.

Let us try to find out why is this trend?

11) Win percentage batting second:

It can be clearly seen from the pie chart that teams winning chasing the total is about 53.2%.

Now let us try to dive in team by team analysis, as that will only help us predict outcome of the match between two teams.

12) Probability of Winning if the Team is Batting First:

The plotted bar graph clearly shows that the probability of winning the match if the team is batting first for each team varies quite a bit. Chennai Super Kings has the highest winning average of about 58% followed by Mumbai Indians. Gujarat Lions, Daredevils and Rising Pune Supergiants fail terribly once they Bat first.

However, this doesn't give us the entire picture it is necessary to analyse this with the Fielding First dataset also.

13) Probability of winning if the Team is Batting Second:

The plotted bar graph shows that teams have done better batting second for example Gujarat Lions have won 80% of the matches Fielding First, unlike Batting First in which they performed terribly, hence it is obvious for them to select Fielding first if they manage to win the toss. However, Chennai Super Kings seems to perform almost the same batting second also of about 62% which is marginally greater than batting first.

It is important point to note that all the teams have done pretty well once they chase at total, with winning percentages being around 50%.

A combined bar graph shows us the clear picture; every team has better winning percentage when they are chasing a total. Daredevils chances of winning increases by 20% and for RCB increases by 12%. However, the most affected are Gujarat Lions and the Supergiants.

Now let us see how have teams performed after winning the toss and electing a particular decision and how has it affected them winning/loosing.

14) Probability of winning if the Team has won the toss and elected to Field

The plotted bar graph clearly shows that if the team has won the toss and elected to field their chances of winning are around 60%. The only anomaly here is the daredevils team who has a winning percentage of slightly below 50% when they elect to field. Hyderabad, Lions and Supergiants have best winning chances if they elect to field.

15) Probability of winning if the team has won the toss and elected to Bat

Chennai Super Kings leads this table having winning percentage of 65% if they elect to bat first, followed by Mumbai Indians with 55%. Lions and Giants seem to have won none of the matches if they win the toss and elect to Bat first.

Kings XI Punjab have done terribly bad having winning percentage of 20% once they win the toss and elect to Bat First.

A combined graph gives the better picture, except for Chennai Super Kings all and Mumbai Indians all teams do well, once they win the toss and elect to field first. Chennai on the other hand wins 65% of the matches after winning the toss by putting scores on the board first whereas for Mumbai Indians it's pretty even -- 55% winning chances when they bowl first or even if they bat first.

16) Team win percentage in each Venue:

There are 9 bar graphs each for a team, which has the winning percentage of that team in that venue.

This will help us analyse at a venue how has that team performed, irrespective of all other factors.

17) Team wins in different cities in each Season from 2010:

The following bar graphs beautifully depicts how have teams performed in each venue throughout the season, each team seems to have won pretty decent amount of matches at their home venue.

18) Venue wise influence of electing to Bat First or Field First:

In order to predict a correct result, it is important that we take into consideration of how the team has performed in that venue and also how toss decision has affected that team. Hence in order to do that we have plotted 9 graphs one for each team that shows what is the winning percentage if the team elects to Bat First or elect to Field first at each venue.

Chennai Super Kings for example have won all its matches at Bangalore electing to Bat first and won only 33% of matches electing to field first.

Hence if a situation arises of CSK electing to field first in Bangalore their chances of winning will get reduced.

Kolkata Knight Riders on the other hand have only won matches if they elect to field at Bangalore. At their home ground if they win the toss and elect to field first they win 81% of the matches.

---PLAYERS :

19) Top players of the match awardees

The bar graph shows that Chris Gayle has won the most number of player of match awards, followed by Yousuf Pathan and AB De Villiers.

20) Top Run Scorers

The bar graph shows that Virat Kohli, Suresh Raina are closely contesting to be the top scorer, we have showcased the top 10 scorers in all the IPL matches till now.

21) Batsman who have scored most number of Boundaries

Gautam Gambhir leads when it comes to scoring runs of boundaries with number of boundaries being 422.

22) Top Bowlers – bowlers with most number of dot bowls bowled.

Praveen Kumar turns out to be getting the most number of dot bowls followed by Dale Steyn and Lasith Malinga.

23) A heat map showing how the teams approach to scoring runs has also been included for further analysis.

4.DEPLOYMENT :

We have built a user-interface to check the prediction using C Programming Language. A 600 line code encapsulating all the analysis made above is provided with menu driven interface to make it convenient for the user.