Capstone Project – Cricket Win Prediction

- NIMESH MARFATIA

Capstone Journey & Assessment:

Capstone Journey:

1st Mentor Connect – 4 Feb

PN1 Submission

2nd Mentor Connect – 18 Feb

PN2 Submission

3rd Mentor Connect – 4 Mar

Capstone Presentation

Final Report Submission

Assessment:

STAGES	GRADES
Project Notes – I	20
Project Notes – II	20
Capstone Presentation	20
Final report submission	40
Total Marks	100

Today’s Agenda

1) Discuss the business context.

2) Understanding Data Dictionary and Data Set

3) Data cleaning and pre - processing (like outlier treatment, missing value treatment etc.)

4) How to generate insights from EDA?

5) Discuss about any finer nuances that could be used to generate insights.

Project Notes -1

Criteria	Pts
1. Problem Understanding Defining problem statement, Need of the study/project, Understanding business/social opportunity	4
2. Data Report Understanding how data was collected in terms of time, frequency and methodology, Visual inspection of data (rows, columns, descriptive details), Understanding of attributes (variable info)	2
3. Exploratory Data Analysis Univariate, Bivariate Analysis, Missing Value, Outlier Treatment, Correlation, Multi collinearity, VIF, etc	10
4. Insights from EDA Data is Balanced, What to do?, Clustering or any other insights about data	4

Introduction

Board of Control for Cricket in India

The Board of Control for Cricket in India (BCCI) is the governing body for cricket in India and is under the jurisdiction of Ministry of Youth Affairs and Sports, Government of India. The board was formed in December 1928.

Total annual income

In FY 2019–2020, the total annual income of BCCI is estimated to be over INR 3,730 crore, including INR 2,500 crore from the IPL, INR 950 crore from bilateral cricket with other nations, and INR 380 crore from India's share of ICC revenue.

Revenue streams for BCCI:

ICC income share

Media rights

Sponsorship rights

Indian Cricket Team Performance

Test						
Matches	Won	Lost	Drawn	Tied	Win %	
555	164	171	219	1	29.54	
T20						
Matches	Won	Lost	Tied +W	Tied +L	NR	Win %
151	93	51	3	0	4	64.28
ODI						
Matches	Won	Lost	Tied	NR	W/L ratio	Win %
995	518	427	9	41	1.21	54.76

One Day Internationals Statistics

https://en.wikipedia.org/wiki/List_of_India_One_Day_International_cricket_records

T20 Internationals Statistics

https://en.wikipedia.org/wiki/List_of_India_Twenty20_International_cricket_records

Test Match Statistics

https://en.wikipedia.org/wiki/List_of_India_Test_cricket_records

Problem Understanding

Defining problem statement

Need of the study/project

Understanding business/social opportunity

Please Note:

- Cover each point, do not miss any point in the report.
- Write your observations / understandings on each point.

Problem Statement Given:

BCCI has hired an external analytics consulting firm for data analytics.

- The major objective of this tie up is to extract actionable insights from the historical match data and make strategic changes to make India win.
- Primary objective is to create Machine Learning models which correctly predicts a win for the Indian Cricket Team.
- Once a model is developed then you have to extract actionable insights and recommendation.

Business Objective

Also, below are the details of the next 5 matches, India is going to play.

You have to predict the result of the matches and if you are getting prediction as a Loss then suggest some changes and re-run your model again until you are getting Win as a prediction.

You cannot use the same strategy in the entire series, because opponent will get to know your strategy and they can come with counter strategy.

Hence for all the below 5 matches you have to suggest unique strategies to make India win.

The suggestions should be in-line with the variables that have been mentioned in the given data set.

Do consider the feasibility of the suggestions very carefully as well.

1. 1 Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.

2. 2 T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

3. 2 ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

Business Understanding, Objective & Scope

- India , as leading cricketing nation, plays all modes of cricket throughout the year. While the country had been successful in all formats of cricket, it is important for BCCI to better the standards to be the N0.1 team in the world.
- With this intention, using the historical data, we need to build a model which is accurate. Once done, we shall apply the model to the future matches.
- The critical need, then would be to investigate and remedy the matches that are predicted to be a loss.
- This needs to be done by tweaking the available parameters for the team.
- Thus the success metric for the project would be to make the team choose the right parameters, wherever opportunity is given, and win every match.

What type of problem is it?

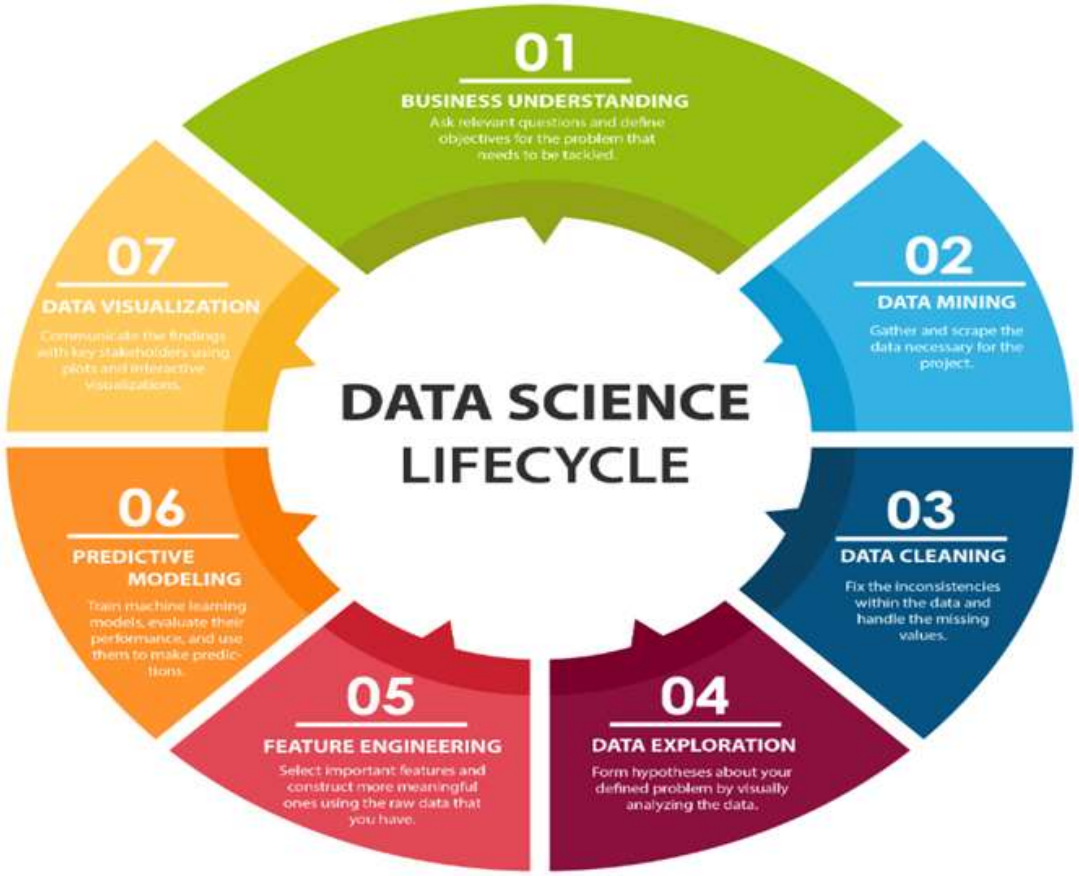
Supervised learning:

- You train the **model** using data which is well labelled.
- Accomplishing a task by providing **training**, input and output patterns to the systems.

Classification type:

- It specifies the class to which data elements belong to and is best used when the output has finite and discrete values.

Modelling Steps



Data Report

- Understanding how data was collected in terms of time, Frequency and methodology
- Visual inspection of data (rows, columns, descriptive details)
- Number of Rows & Columns (Variables)
- Understanding of attributes (variable info, renaming if required)

Data Report

- The dataset contains match details of 2930 matches India has played.
- In other words, 2930 rows of data.
- The dataset has 23 columns, or 23 distinct variables.
- The dataset has data pertaining to
  - 125 Test matches
  - 1865 Oneday internationals (ODI)
  - 870 20-20 (T20) matches
- There is no reference to the *date* on which the match was played. Thus we do not have a visibility on the players who could have influenced the match outcomes.
- The team composition used in each match is not very clear from the dataset.

Data Dictionary – Variable Description:

Variables	Description
Game_number	Unique ID for each match
Result	Final result of the match
Avg_team_Age	Average age of the playing 11 players for that match
Match_light_type	type of match: Day, night or day & night
Match_format	Format of the match: T20, ODI or test
Bowlers_in_team	how many full time bowlers has been player in the team
Wicket_keeper_in_team	how many full time wicket keeper has been player in the team
All_rounder_in_team	how many full time all rounder has been player in the team
First_selection	First inning of team: batting or bowling
Opponent	Opponent team in the match
Season	What is the season of the city, where match has been played
Audience_number	Total number of audience in the stadium
Offshore	Match played within country or outside of the country
Max_run_scored_1over	Maximum run scored in 1 over by team
Max_wicket_taken_1over	Maximum wicket taken in 1 over by team
Extra_bowls_bowled	Total number of extras bowled by team
Min_run_given_1over	Minimum run given by the bowler in one over
Min_run_scored_1over	Minimum run scored in 1 over by team
Max_run_given_1over	Maximum run given by the bowler in one over
extra_bowls_opponent	Total number of extras bowled by opponent
player_highest_run	Highest score in the match by one player
Players_scored_zero	Number of player out on zero run
player_highest_wicket	Highest wickets taken by single player in match

EDA – Exploratory Data Analysis

This is an important step in the complete Data Analysis and Model Development cycle. Let's look at some of the important activities to be performed in this phase

Univariate analysis

- Describe the data and find patterns that exist within it.

Bivariate analysis (relationship between different variables , correlations)

- Find out if there is a relationship between two different variables

Removal of unwanted variables

- Drop the features you do not find related to the Target variable.

EDA – Exploratory Data Analysis

Data Balancing

- Churn Variable: **Win 2457 (83.9%), Loss 473 (16.1%)**.
- Instances of one of the two classes is higher than the other, in another way, the number of observations is not the same for all the classes in a classification dataset.
- Oversampling using SMOTE (Synthetic Minority Over-sampling Technique).
- Under sampling using K-means algorithm

Correlation Test

- Check Multi Collinearity
- As there are multiple features in this data set which have a high correlation, it is important to handle multicollinearity.
- Dropping the features because of high correlation should be used as a last option if nothing really works out.

EDA – Exploratory Data Analysis

Missing Values:

- Numeric Variable – Mean / Median (if outliers are present)
- Categorical Variable – Most Common Class / Unknown Class
- If a feature has more than 15-20% missing value (15% is used as a norm, imputation of more than 15% is not recommended) and such features can be removed from the model building dataset.

Profiling of Continuous and Categorical Columns.

- Objective of this step is to
- Identify how variability in the continuous features is observed with respect to dependent feature
- For categorical feature what is the distribution of dependent variable class across each level.

EDA – Exploratory Data Analysis

Outlier treatment

There are multiple ways in which outliers can be treated. We can either use “Capping” technique or we can remove the outlier values.

Identify the outlier through Box Plot.

Dummy Variables:

Where a categorical variable has more than two categories, it can be represented by a set of dummy variables, with one variable for each category.

Business insights from EDA

Is the data unbalanced? If so, what can be done? Please explain in the context of the business

- Checking data balance is very critical for Classification problems.

Any business insights using clustering

- Clustering is generally done on Unsupervised data.
- If you wish to try you can try and build separate model for each cluster.

Any other business insights

- Add any other insights then you can add the same in this section.

Business Insights

- The team played most of the matches against South Africa and has a good probability of winning the game.
- Even though the no. of matches played against Zimbabwe are less. The chances of winning are slightly greater than 50%.
- Most of the matches are played in Rainy season and the chances of losing are less.
- Even though the no. of matches played in summer are not very less. The chances of winning are not as high as in summer.
- The number of matches played outside the country are less than the no. of matches played within the country.
- The chances of losing the match are high when played outside the country compared to when played in the country.
- As the Number of player out on zero run increases the chances of losing the match is also increasing.
- As the Highest wickets taken by single player in match increases the chance of losing the match is very less.
- When the Highest wickets taken by single player in match is very less there is high chance of losing the match.
- As the Average age of the playing 11 for that match increases there are high chances of winning.
- When the no. of full time bowlers in the team is less than 5 then there are 50-50 chances of winning the game.

Points to remember

- Understand and make clear the problem statement
- Explain the data dictionary in detail
- EDA should been done thoroughly.
- Do not paste only codes in your report.
- Focus should be on business report/notes.

Thank You