Data Analytics Assignment - 1

Arman Gupta (19220)

Implementation Detail

Considering the data for the first inning only, I have analyzed the data carefully and find the following problems with the data -

1. Innings total run (col: *Innings.Total.Runs*) must be consistent with the total runs (col: *Total.Runs*) at the end of the inning but that is not the case for some matches.
2. Some matches are incomplete i.e. overs played are less than 50 and total outs in an inning is < 10.
3. *Total.Runs* and *Innings.Total.Runs* columns are not consistent with the *Run* Column which ultimately affects the Runs.Remaining columns which we have to use in our model.

Based on above problem, I have denoised the data using three ways as described below -

1. **Preprocessing 1 :**
   Problem : Innings total run (col: Innings.Total.Runs) must be consistent with the total runs (col: Total.Runs) at the end of the inning. If that is not the case, we need to correct the Total.Runs and Innings.Total.Runs columns to make these columns consistent and then these columns will be used to correct Run.Remaining column values.
   Solution : When we have above inconsistency, we are using runs (col : Runs) to calculate Total.Runs and Innings.Total.Runs columns, hence we are getting consistent Total.Runs and Innings.Total.Runs at the end of this preprocessing step.
   Above solution will be applied if the sum of all the runs (col : runs) for a particular inning at least matches with either Total.Runs at the end of the innings or Innings.Total.Runs.
   The thought process was that this *Run* column is at least consistent with one of the two columns (*Total.Runs and Innings.Total.Runs*). As a consequence , we are deleting 10 matches as these matches' data are completely inconsistent.
   [Matches Deleted : 65200, 64793, 66351, 64933, 64940, 217978, 267386, 385023, 424849, 506207]

2. **Preprocessing 2:**
   Problem: Since some matches are incomplete i.e. overs played in the inning are less than 50 and total outs in the inning are < 10. This may occur due to some external interruption or missing data.
   Solution: We are eliminating the incomplete matches from the data obtained after preprocessing 1 and calling the resulting data frame as complete_clean_df in the code.

3. **Preprocessing 3:**
   Problem: The Total.Runs and Innings.Total.Runs are not consistent with the Run Column which ultimately affects the Runs.Remaining columns which we have to use in our model.

Solution: Considering Runs as the correct column and deriving other columns from it. The reason for considering this option as I have observed that around 340 matches' data are inconsistent i.e the Runs column is not consistent with either Innings.Total.Runs or Total.Runs.

General Step: This step is common to all the processing steps:

1. Computing Over_Used_in_Future which indicates the number of overs remaining.
2. Adding a row for each match which indicates the 0th over i.e. starting of the inning.
3. Since for each wicket and over combination, we may have more than one value so I take the average of all those values so that the model can be trained fast.

Optimization:

$Z_0(w)$ $\forall w$, where 1<= w <=10, are initialized with the mean of Runs.Remaining column having Wicket.In.Hands = w and L is initialized with a scalar value 10. These initial parameters are then optimized using scipy.optimize.minimize() function present in Scipy library with method = 'BFGS' as hyperparameter. The model takes Over_Used_In_Future as the input parameter for a particular value of wickets in hand and trained against Average Runs.Remaining as target with the help of squared error loss function.

Results

[Note : The values have been rounded off to 2 decimal places.**]**

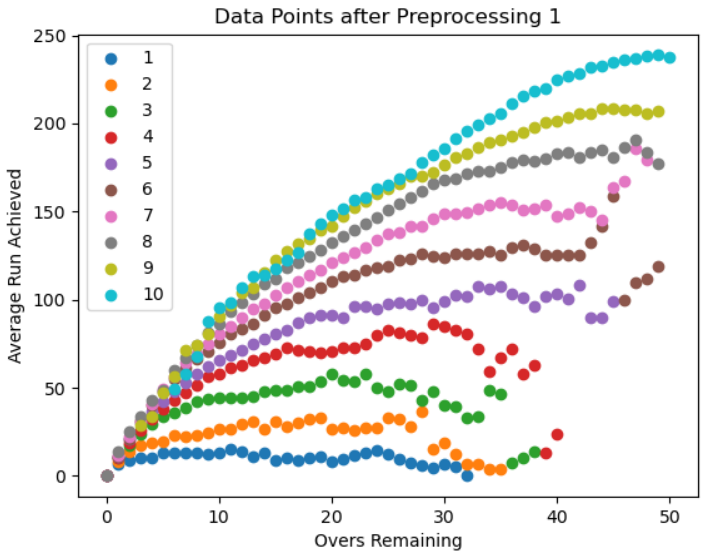
1. Results using Data after Preprocessing 1

$Z_0(1)$	10.28
$Z_0(2)$	24.00
$Z_0(3)$	44.31
$Z_0(4)$	71.40
$Z_0(5)$	103.17
$Z_0(6)$	133.20
$Z_0(7)$	170.30
$Z_0(8)$	205.13
$Z_0(9)$	238.27
$Z_0(10)$	276.63

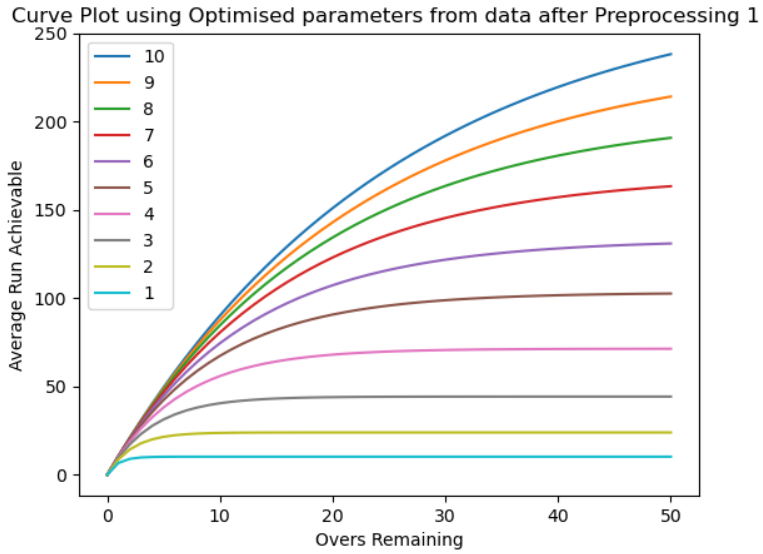
L (Slope) : 10.91
Normalized Squared Error : 51.67

Plotting data points-

The following figure has been created by plotting the data points from the data.



The below figure has been created using the optimized model parameters (Z(w) for all w ,where w indicates wickets in hand and slope(L)). We computed the Run Achievable using the model at each remaining over.

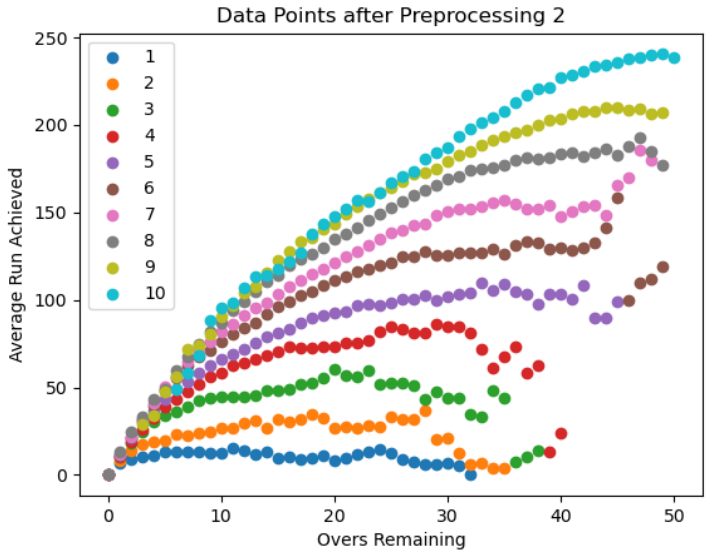


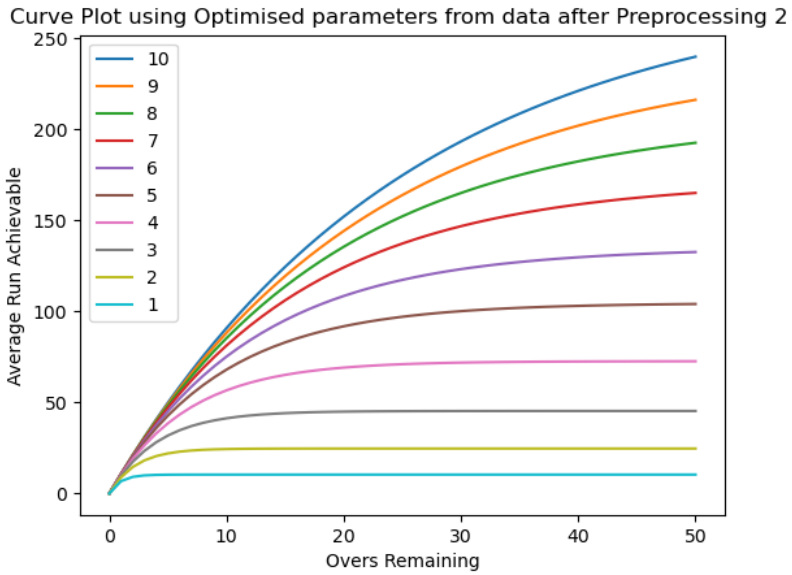
2. Result using Data after Preprocessing 2

$Z_0(1)$	10.39
$Z_0(2)$	24.70
$Z_0(3)$	45.31
$Z_0(4)$	72.66
$Z_0(5)$	104.60
$Z_0(6)$	134.89
$Z_0(7)$	172.10
$Z_0(8)$	207.15
$Z_0(9)$	240.68
$Z_0(10)$	278.52

L (slope) : 10.997
Normalized Squared Error : 54.16

Plotting data points-



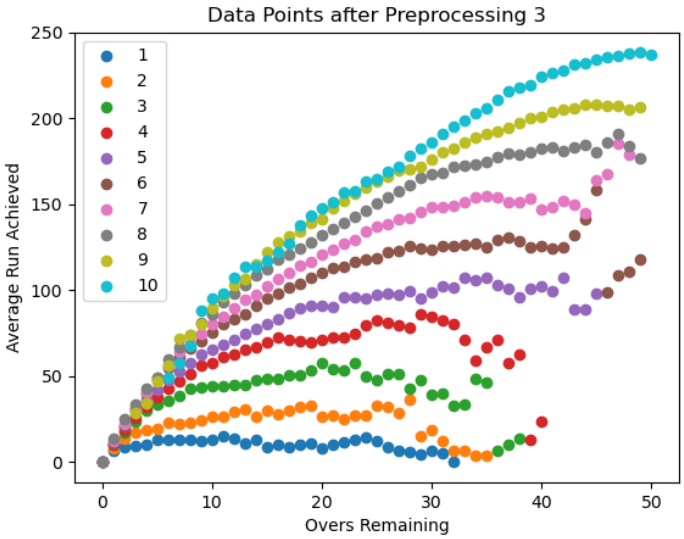


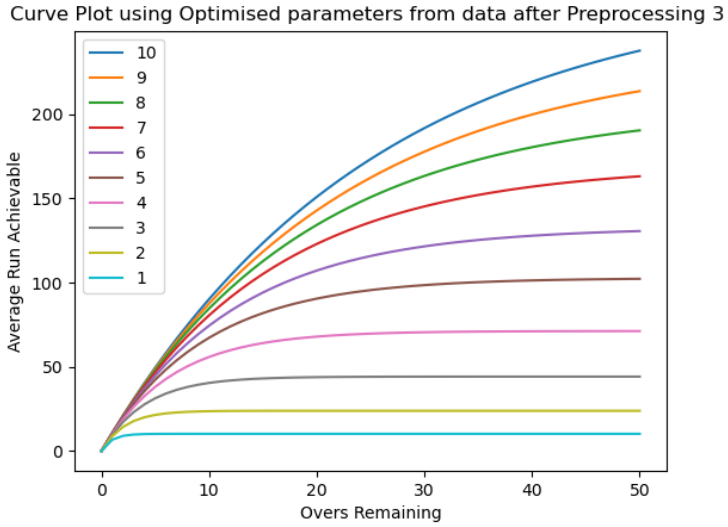
3. Result using Data after Preprocessing 3

$Z_0(1)$	10.24
$Z_0(2)$	23.92
$Z_0(3)$	44.21
$Z_0(4)$	71.29
$Z_0(5)$	102.81
$Z_0(6)$	132.84
$Z_0(7)$	170.05
$Z_0(8)$	204.67
$Z_0(9)$	237.66
$Z_0(10)$	275.94

L (slope) : 10.93
Normalized Squared Error : 51.90

Plotting data points-





Observations : Although the curve plots from all the three preprocessing steps look the same , the parameters are different for each of the preprocessed data. As we can see that we are getting the lowest normalized square error for data obtained using preprocessing 1.

Note : In above plots, the legend represents the wicket and the color corresponding to each wicket number indicates the color used for plotting the curve for that particular wicket.