Data Analytics: Assignment 1

Duckworth-Lewis-Stern Method

Piyush Kumar Mishra (Sr No:- 19882)

Problem Statement: Using the first innings data alone in the above data set, find the best fit run production functions in terms of wickets-in-hand w and overs-to-go u. Assume the model is

$$Z\ (u, w) = Z_0(w)[1 - \ \exp(-Lu/Z0(w)$$

Use the sum of the squared errors loss function, summed across overs, wickets, and data points for those overs and wickets.

Data Cleaning:

Step 1:- First, I create one more data frame named ‘data’ in which the first Innings data is there only. In which I get 67794 rows out of 126768 rows.

Step 2:- Then I see that the ‘Runs.Remaining’ column is less than 0 for some of the rows and then through some of the matches in which the sum is less than 0. So I find there is a Match(65200) in which in 32 overs, the ‘Total. Runs’ column is 144 and in 33 over there is no run scored but then also they have added one run to it and score become 145. That’s why the ‘Runs. Remaining’ for this match is less than 0.

Step 3:- For the match id 65206, I have seen that there is a mistake in the calculation for ‘Total. Runs’. Total runs in that match was 196(by summing over all the overs) but in the ‘Innings.Total.Runs’ there, the entry is 197.

Step 4:- Here I assume that ‘Runs’ column is correct. To resolve the above two issues, I sum over all the Runs in each of the over and then create the ‘Actual_total_runs’ and ‘Actual_total_innings’, and from these two, I compute the Actual_Runs_Remaining. And insert all the 3 columns in the data frame.

Step 5:- I remove the ‘Innings.Total.Runs’ , ‘Total. Runs’ and ‘Runs.Remaining’ columns from the data frame.

Step 6:- I have also found the matches in which the first innings do not have 50 overs and the wickets remaining are greater than 0. Which means that the match has been stopped. So I remove all such matches data from my data frame. So after that I get 59236 data points from 65206 data points.

Step 7:- I have also checked a column named ‘Error.In.Data’, and there, I get 400 such entries out of 59236. And all of them are from the eight matches. So I remove these 8 matches also.So then I have 58836 data entries.

Step 8:- And I also check the matches in which the ‘Overs’ column does not start with 0. There is one match (410557) in which data for the first 15 overs is not present. So, I remove it. At last, I get 58801 data entries.

Approach Used

1. I initialize the Z_0 by taking the mean from each of the wickets remaining and the value of L is 1.The initial values of Z_0 is shown in the table.

W	$Z_0(w)$
1	9.944504896626768
2	19.387205387205388
3	34.872262773722625
4	52.90385617509119
5	72.30160692212608
6	97.6181504577418
7	126.19582875960484
8	160.95153506520177
9	192.9395895048667
10	231.2938824954573

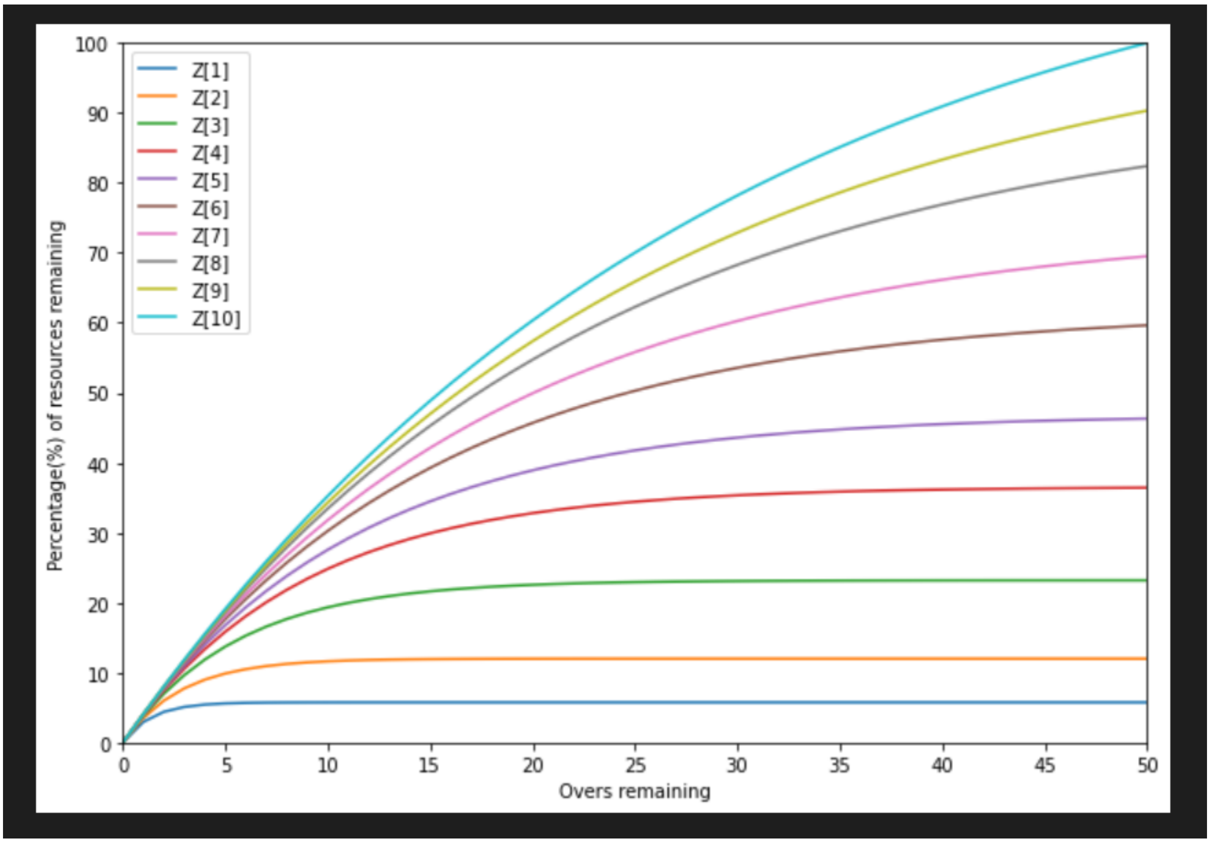
2. Then I used SciPy. Optimize library to minimize the loss function. In scipy.Optimize. minimize I have further tried different
   Methods like : TNC, L-BFGS-B, BFGS, POWELL, and COBYLA. A comparison of total normalized loss of all the methods is given as follows.

Method_used	Normalized_loss
TNC	1485.5468951956586
Powell	1865.746127514062
BFGS	1842.7490650292802
L-BFGS-B	1368.7382140119978
Cobyla	1403.82839666

3. After using the Scipy optimize with L-BFGS-B(because it gives the minimum normalized loss), I get the normalized loss of 1368.7382140119978. And the Z_0 after optimization is shown in the table.
   And the value of L is 10.372650438226092

W	Z_0(w)
1	14.34960831
2	29.92592743
3	57.89636048
4	91.23192065
5	116.98721067
6	154.20895673
7	184.64721876
8	229.72538047
9	261.43759163
10	305.59530588

4. Plots for 10 functions is shown below


Z[1]
Z[2]
Z[3]
Z[4]
Z[5]
Z[6]
Z[7]
Z[8]
Z[9]
Z[10]
Percentage(%) of resources remaining
Overs remaining