Solution Sheet

1. Which model have you used for Total IPL 2020 Runs prediction for each player? Explain your model.

Ans. The model that I have used for prediction is

Multi Variate Linear Regression.

A multi variate linear regression model is an advanced version of a linear regression model. In a linear regression model, A dependent variable is guided by a single independent variable.

A simple equation for a linear regression model can be written as:

$$Y = a + bX$$

where Y is the dependent variable,
X is the independent variable,
b is the slope of the line,
and a is the y-intercept.

In real world scenarios, such as in the case given to me, the dependent variable Y is dependent on multiple input factors and not just one.

In multi variate linear regression model, multiple independent variables contribute to the dependent variable and hence multiple coefficients have to be determined and complex computation is introduced due to the added variables.

In this model, we don't throw in all the independent variables at a time and start minimizing the error function. First, we select the best possible independent variables that contribute well to the dependent variable. For this, we go on and construct a correlation matrix for all the independent variables and the dependent variable from the observed data. The correlation value gives us an idea about which variable is significant and by what factor. From this matrix we pick independent variables in decreasing order of correlation value and run the regression model to estimate the coefficients by minimizing the error function.

The equation of multivariate linear regression is as follows:

$$Y_i = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots. + \beta_n x_i^{(n)}$$

where Y_i is the estimate of i^{th} component of dependent variable y, where we have **n** independent variables and x_i^j denotes the component of the independent variable/feature.

Similarly cost function is as follows:

$$E(\alpha, \beta_1, \beta_2, \ldots, \beta_n) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - Y_i)$$

where we have **m** data points in training data and y is the observed data of dependent variable.

For example, the runs scored in 2020 would be the dependent variable Y and many independent variables like number of innings played, number of balls played, number of not outs etc. exist on which Y depends.