**Machine Learning**

Consider a data set (Sales\_Dataset.csv) related to sales of items in shops. The fields in Sales\_Dataset.csv data set are:

Item\_Identifier: Code given to items

Item\_Weight: Weight of an item

Item\_Fat\_Content: Fat content of the item like low, high etc

Item\_Visibility: A continuous value indicating the visibility of the item for a customer

Item\_Type: Type of the item like dairy, vegetables etc

Item\_MRP: Maximum retail price of the item

Outlet\_Identifier: Code given to outlets

Outlet\_Establishment\_Year : Year of establishment of an outlet

Outlet\_Size: Size of an outlet like medium, high etc.

Outlet\_Location\_Type: Location of the outlet in a city like Tier1, 2 etc.

Outlet\_Type: Type of the outlet like supermarket, grocery shop etc.

Item\_Outlet\_Sales: Sale value of an item in outlet. It is a target variable.

Tasks:

* Use the Sales\_Dataset.csv file for fitting the models (Divide this file into training set and testing set)
* Treat the missing values appropriately
* Convert the categorical data into numerical data appropriately, if necessary.
* Normalize the data, if required.
* Fit linear regression model and compute RMSE and R-Square.
* Apply regularization techniques on the data to check whether they are better than the basic linear model. Loop through various regularization parameters to display and compare R-Square.
* Fit the residual plots in all cases.
* Give your analysis on which of the independent variables have significant impact on the Sales of an outlet. Call it as a best model.
* Using this best model, predict the values of sales for the data given in the file Test.csv and compute RMSE

**NOTE:**

* Give brief explanation about each task that you are going to do.
* Draw Graphs where and when required and give interpretation of such graphs.
* Graphs must have proper labels and legends.
* At the end of task, give your analysis on the entire dataset, about features, their importance etc.