**Supplement Sales Prediction**

* Your Client WOMart is a leading nutrition and supplement retail chain that offers a comprehensive range of products for all your wellness and fitness needs.
* WOMart follows a multi-channel distribution strategy with 350+ retail stores spread across 100+ cities.
* Effective forecasting for store sales gives essential insight into upcoming cash flow, meaning WOMart can more accurately plan the cashflow at the store level.
* Sales data for 18 months from 365 stores of WOMart is available along with information on Store Type, Location Type for each store, Region Code for every store, Discount provided by the store on every day, Number of Orders everyday etc.
* Your task is to predict the store sales for each store in the test set for the next two months.

##### There are 2 datasets provided df\_train(train) and test(test) the working on both such dataset have done.

The total dataset contains 10 columns of which 6 are having a dtype of object and there are 3 int64 and 1 float64 column. the column "Sales" is the target column.

Of the 4 numerical columns-

* None of them are having missing values, as the count for all the columns are same.
* There might be skewness present in the columns.
* Holiday even though numerical is having 0 as lowest value and 1 as highest value and median 0 which means that it is indeed a categorical column.

#### Exploratory Data Analysis

* Univariate / bi variate / multivariate analysis done
* Feature engineering of the df\_train by splitting date column into 'year' ,'month', 'year'.

Location\_type L1 has the count of 85140 which is the highest in number.

Region\_code R1 has the count of 63984 which is the highest in number.

It can be seen that there is a disparity in the count of the number of holidays.

Store\_type S1 has the count of 88752 which is the highest in number.

It can be seen that the number of times discounts were given is less than the times were not given.

* The Location\_type L2 shows higest sales.
* There can be seen a rise in sales from L3 to L2 AND it decreases to L1 and L5 but it has increased to L4.

The 2 nd of the month has seen the highest sale.

It can be seen that the order and sales are having a positive relations it can be understood that with the increase in order the sales also increase.

For the Engineering and cleaning sections for both train and test datasets. The dataset had columns that were removed as they were not needed.

Dropped columns. In both datasets

'ID','Store\_id','Date','year','month','day',’#order

Checking corelation matrix for the understanding the multi collinearity.

It can be seen that the Multicollinearity are all in tolerable range but VIF has to be checked.

As random forest has the lowest MSLE value of 0.059782 it will be used for predicting with the test file.