JOB-A-THON - September 2021

## **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.**

## **Data Dictionary**

### **Train Data**

|  |  |
| --- | --- |
| Variable | Definition |
| ID | Unique Identifier for a row |
| Store\_id | Unique id for each Store |
| Store\_Type | Type of the Store |
| Location\_Type | Type of the location where Store is located |
| Region\_Code | Code of the Region where Store is located |
| Date | Information about the Date |
| Holiday | If there is holiday on the given Date, 1 : Yes, 0 : No |
| Discount | If discount is offered by store on the given Date, Yes/ No |
| #Orders | Number of Orders received by the Store on the given Day |
| Sales | Total Sale for the Store on the given Day |

### **Test Data**

|  |  |
| --- | --- |
| Variable | Definition |
| ID | Unique Identifier for a row |
| Store\_id | Unique id for each Store |
| Store\_Type | Type of the Store |
| Location\_Type | Type of the location where Store is located |
| Region\_Code | Code of the Region where Store is located |
| Date | Information about the Date |
| Holiday | If there is holiday on the given Date, 1 : Yes, 0 : No |
| Discount | If discount is offered by store on the given Date, Yes/ No |

### **Sample\_Submission**

|  |  |
| --- | --- |
| Variable | Definition |
| ID | Unique Identifier for a row |
| Sales | Total Sale for the Store on the given Day |

## **Evaluation**

The evaluation metric for this competition is [MSLE](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_log_error.html) \* 1000 across all entries in the test set.

### **Public and Private Split**

Test data is further divided into Public (First 20 Days) and Private (Last 41 Days). You will make the prediction for two months (61 days).

* Your initial responses will be checked and scored on the Public data.
* The final rankings would be based on your private score which will be published once the competition is over.

**Feature Engineering**

5 set of features are created.

1. Calander date and Holiday related features
   1. Year, month, day, day of week, week of month, quarter, etc.
   2. Number of holidays in current month, Number of days left to have a next holiday, number of days past from the previous holiday.
2. Store, Store type, location type, region related features
   1. Number of stores in the region, number of stores of same location type in a region, number of stores of same store type in a region, number of stores of same store & location type in a region
3. Discount related features
   1. **On same day**, number of stores with discounts for same store & location type in a region, number of stores with discounts for same location type in a region, number of stores with discounts in a region
   2. **For a store**, total number of discount days in current month, Number of days left to have a discount day, number of days past form previous discount day.
   3. **For current month,** number of stores with discounts for same store & location type in a region, number of stores with discounts for same location type in a region, number of stores with discounts in a region
4. #Orders Lags and exponential moving averages
   1. Lag variables of 61,62,63,64
   2. Exponential Moving Average
5. Sales Lags and exponential moving averages
   1. Lag variables of 61,62,63,64
   2. Exponential Moving Average

**Modeling**

Cat boost Model is used for Modeling.

* Sales and #Orders are Log transformed before modeling
* July 2018 and May 2019 are used as validation data for find out best Hyper parameters.
* Based on Feature Importance best features are selected