**WALMART POINTS FORECASTING - APPROACH**

**Problem Statement:**

**2. Feature Selection:**

We have considered this problem as a multivariate time series analysis, as the sales forecast here depends not only the dates, but also depends on various other factors like geographic location, location specific stores, products and events set for special days.

The feature selection part therefore has been divided in two parts:

* Events to be considered from calendar.csv
* Features from sales\_train\_validation.csv

In order to infer the importance of events in impacting the sales for particular day, we have aggregated the sales of all the products at State level for a given day. The sales data was then mapped against the days having events and days not having events in calendar.csv. MinMaxScaler has been applied on the total sales to normalise the data, as there was noticeable deviations in the data.  
  
Here the events, no events days are taken as the predictor variables and the aggregated sales being the dependant variable. Then we have applied Ridge regression between the predictor variables and the dependant variable to calculate the coefficients so as to infer the importance of events and no events in forecasting the sales.

**2.1 Why Ridge Regression:**

Ridge regression was selected over Lasso regression, as in Lasso regression the coefficients become zero after some iterations. Based on the Ridge regression, the below events were identified as the important events that are positively and significantly correlated with the aggregated sales at state level.

| State | Features | Correlation\_Coefficient |
| --- | --- | --- |
| **CA** | month | 0.31 |
| **LaborDay** | **0.63** |
| No Event1 | 0.23 |
| **Purim End** | **0.422** |
| **SuperBowl** | **0.41** |
| No Event2 | -0.02 |
| Snap\_0 | -0.06 |
| Snap\_1 | 0.06 |
| **TX** | **LaborDay** | **0.41** |
| No Event1 | 0.09 |
| **Purim End** | **0.351** |
| **SuperBowl** | **0.42** |
| **Valentines Day** | **0.4** |
| No Event2 | 0.03 |
| Snap\_0 | -0.14 |
| Snap\_1 | 0.14 |
| **WI** | month | 0.39 |
| EidAlAdha | 0.38 |
| Halloween | 0.43 |
| **LaborDay** | **0.31** |
| No Event1 | 0.31 |
| **Purim End** | **0.3** |
| Valentines Day | 0.77 |
| Snap\_0 | -0.21 |
| Snap\_1 | 0.21 |

Based on the above approach, the below features were considered important at each state level.

\*From the above table it can be inferred that SNAP events did not have much effect on the total sales.

**There are 4 different types of events namely Sporting,Cultural,National and Religious.**

**From the table it can be seen that most of the times, sales is impacted only when Cultural and Religious events are set.**