Group: Action

Name: R S Waghmare

Competition: Skillenza, Unilever Data Science POC Use Case

Date: 31/3/2020

Hurdle 1

**Problem Definition:** One of the brands is going through some major changes in business execution plans and will like to know.

1. What are the major drivers for sales(EQ)?

Last month’s sales is the most important driver (Weight = 17)

Remaining drivers are as follows:

Sales 2 months ago, weight = 5

Sales 3 months ago, weight = 4

Sales 2 years ago, weight = 3

Sales 1 year ago, weight = 2

1. Knowing the drivers, how accurately we can predict future sales for next 6 periods?

The error in validation is -4 to +9 in average sales of 220 or about 5%

**Problem Solving:**

1. i. One Bayesian method: BSTS package in R
2. ii. One general ML/ forecasting approach: XGBoost package in Python

**Datasets used:**

1. Four-weeks period number and EQ for three years.
2. For the year 2018, 1 Target, 5 features and 13 records.

EQ (Target)

Last month’s sales

Sales two months ago

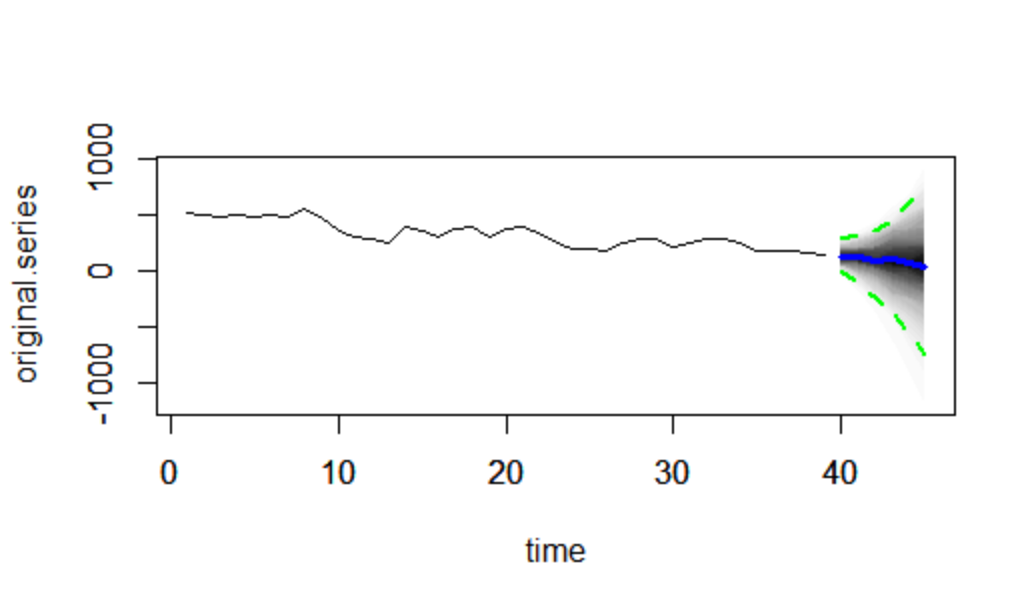
Sales three months ago

Sales 52 weeks ago

Sales 104 weeks ago

**Procedure:**

1. BSTS package is very user friendly. It just requires feeding 39 EQ values and extrapolate to a horizon of 6 periods. Only issue is very broad limits are posed on the mean values. This is shown in the graph on the next page:



1. The model is trained on 12 periods and prediction is made for the 13th period using xgboost.

For required forecasting, one period is taken at a time and the input data table is updated if necessary.

**Results:**

The results of the various methods are summarized below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Period | A | B | C | D | E |
| 2019 | 1 | 381.4756 | 374.8071 | 196.9334 | 130.6578 | 122.4047 |
| 2019 | 2 | 376.6873 | 375.5114 | 205.8174 | 137.9761 | 109.3634 |
| 2019 | 3 | 346.7937 | 351.1097 | 180.7403 | 137.9761 | 90.0666 |
| 2019 | 4 | 352.0007 | 345.0074 | 192.5326 | 130.6578 | 76.2474 |
| 2019 | 5 | 368.6203 | 362.7393 | 203.3145 | 130.6578 | 57.5309 |
| 2019 | 6 | 355.6983 | 359.1762 | 181.1540 | 130.6578 | 45.0585 |

Method Explanation

A Average of same period for the last three years

B Exponential smoothing of the same period in previous years with latest period

weight = 0.4

C Linear model for data same as that for XGBoost

D XGBoost predictions, error < 5%

E BSTS mean predictions

BSTS means are the least values, while averages of same past periods give highest predictions. XGBoost gives error estimation, which is just about 5% and these seem reasonable. There is one caution though, pandemic of Carona is not present in the past data and large scale currency would be distributed by the world Governments. The consequent lull and stimulus are quite new happenings for the period of study and they could substantially alter the actual sales values.

Hurdle 2

**Problem Solving:**

1. Random Forest Regressor is chosen as Naïve Bayes regression performs poorly.
2. XGBoost is chosen as alternative algorithm.

**Dataset Used:**

The full dataset is used as pandas dataframe in Python, for both the algorithms. Five category variables have unique values in both train and test sets. So, these are ignored.

**Procedure:**

As the datasets are small, the models are manually tuned and predictions are made.

**Results:**

RMSE error = 15 with xgboost and 24 with random forest algorithms.

Mean target in test set is 276. So, the % error with xgboost is 5.5 and random forest is 8.7.

**Discussion:**

The accuracy in hurdle 1 is confirmed. The top three important features are:

Avg\_no\_of\_items, weight = 43

Median\_Temperature, weight = 37

Social\_Search\_Working\_Cost, weight = 35

**Conclusion:**

XGBoost gives reliable estimates both from time series as well as full dataset with categories excluded.