Philips - Time series forecasting

Sunday, April 19, 2020 11:02 AM

Data given:

- For 72 products over 36 months
- Few products are launched in that 36 months and do not have data for all the months
    - 25 product has data for all 36 months
    - 10 product has data for at least 24 months
    - Nearly 35 product has data only of less than one year

Although the ideal approach would be to treat all the 72 product as a univariate timeseries problem and fit a separate model for each of them, I tried a different approach here to try to fit a single model for all the products together.

Train set preparation:

To prepare the training set I took few features such as:

- Product sales in the last three months
- Product sales in last quarter
- Product sales in last half-year
- Product sales in the same period last year
- Moving average for three months
- Moving average for 6 months
- One hot encoded product variables
- One hot encoded month variables

The train set is taken for 36 months period from Jan-2016 to Dec-2018. Only the rows with more than 3 of the prepared features (except the OHE features) are taken for training set.

The features in the training set are then standardized (subtracted from mean and divided by standard deviation) to make them the same scale.

Correlation:

	Month-1	Month-2	Month-3	Quarter	Half_year	Yearly	MA3m	MA6m	Sales
Month-1	1.00	0.54	0.45	0.37	0.48	0.57	0.87	0.76	0.61
Month-2	0.54	1.00	0.56	0.43	0.37	0.35	0.78	0.76	0.44
Month-3	0.45	0.56	1.00	0.55	0.40	0.38	0.56	0.75	0.40
Quarter	0.37	0.43	0.55	1.00	0.43	0.31	0.45	0.71	0.33
Half_year	0.48	0.37	0.40	0.43	1.00	0.54	0.58	0.64	0.58
Yearly	0.57	0.35	0.38	0.31	0.54	1.00	0.66	0.62	0.69
MA3m	0.87	0.78	0.56	0.45	0.58	0.66	1.00	0.91	0.84
MA6m	0.76	0.76	0.75	0.71	0.64	0.62	0.91	1.00	0.75
Sales	0.61	0.44	0.40	0.33	0.58	0.69	0.84	0.75	1.00

The sales is highly co-related with the variables of the moving average of 6 months and 3 months.

Training:

The data is then fit into simple linear regression for the 36 months with the sales of the month as the y-variable and the above mentioned features as the x-variables.

Since the data is highly correlated, elasti-net regression will perform better than a normal linear regression, because of the L1 and the L2 regularization parameters.

The model has an RMSE of: 2.138

Making predictions on the 2019 data:
For making predictions on the 2019 data, it has to be done month on month, since only if we have Jan-2019's predictions with the rolling window. Hence once the prediction for Jan-2019 is available, it will then be used to forecast the value for Feb-2019 and so on.