Unilever Data Science POC Use Case

Team - 3_Idiots
Jang Bahadur
Satyam Satyajeet
Jai Bhagat

Problem Statement

One of our brands is going through some major changes in business execution plans and will like to know.

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

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

Assumption

- As training data was day wise we need to aggregate and make period (with 28 days) and then aggregate om this period to get mean of all the variables as test data was given from 2016 period wise
- generated periods and year based on previous assumption

What are the major drivers for sales(EQ)?

To find the best drivers two approaches were used.

1) Correlation Matrix
2) Feature importance of all the variables were done using **Light GBM** was done.

Steps Followed

- As data is numeric, no encoding techniques are used.
- As we are using tree based model, so no outlier treatment required.
- test,train split was 0.80,0.20.
- We used GBM, Random Forest, MLP, Light GBM modelling techniques
- Loss function is used as RMSE because data has fluctuation on both sides of the tail. RMSE is best metric to use in such scenerios.
- As NN and light GBM are good result.

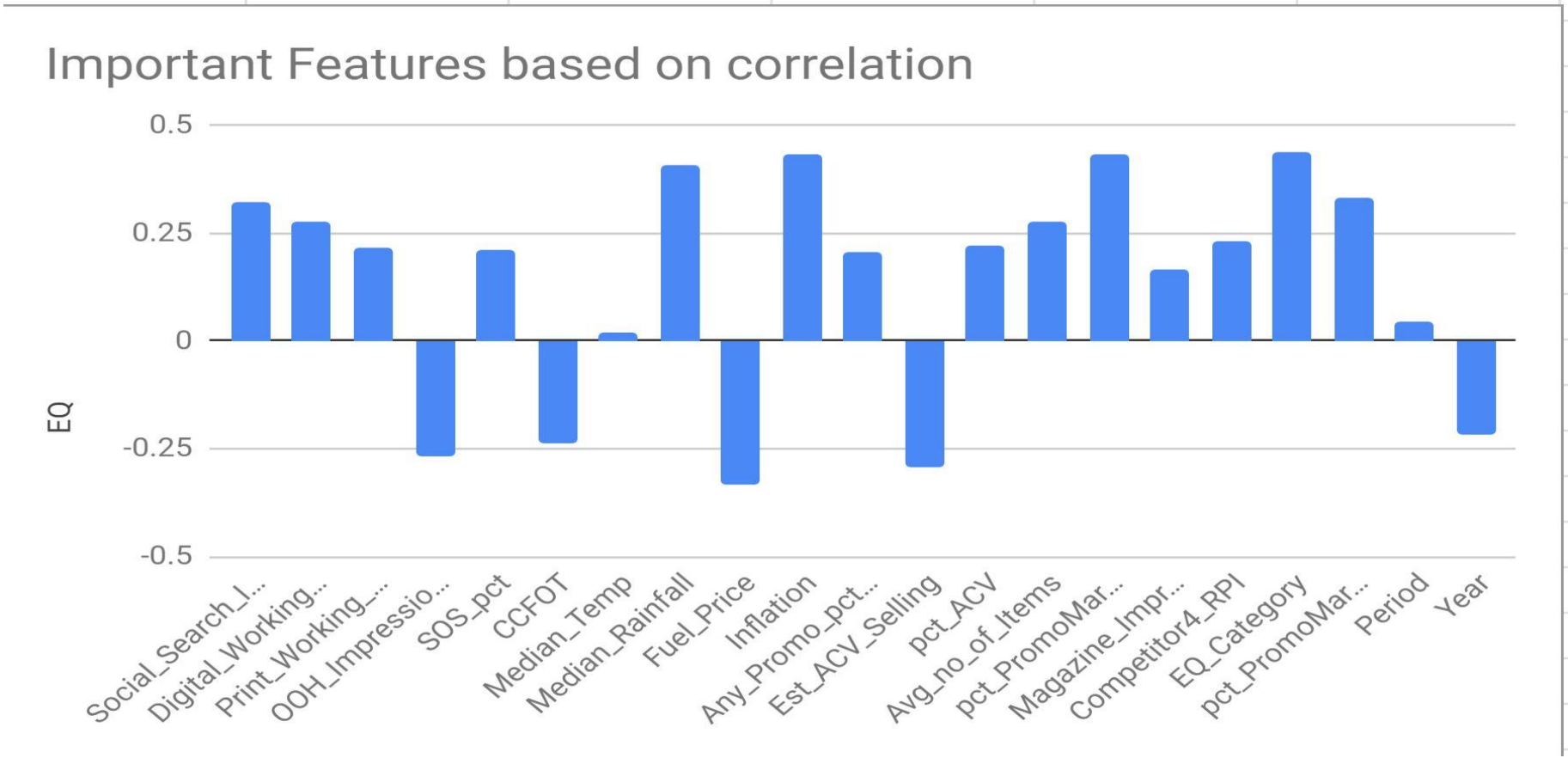
Cross validation Results

```
Results train  and validation MLP>>>>>>>
MAPE: 23.487187336113003 RMSE: 191.70474492371523 MAE: 141.7874795434258
MAPE: 22.061302350159078 RMSE: 186.7265188184021 MAE: 136.09448105926975
Results train  and validation RFR>>>>>
MAPE: 11.493171222589204 RMSE: 82.25009142403123 MAE: 65.13229458064178
MAPE: 26.384262713024153 RMSE: 190.69937976435656 MAE: 147.30680104999757
Results train  and validation GBM>>>>>
MAPE: 0.29572069720165445 RMSE: 11.17121865985714 MAE: 2.1514954497517444
MAPE: 26.09798822576949 RMSE: 196.3417355674708 MAE: 148.80896697961157
Results train  and validation LGB>>>>
MAPE: 13.968932565610343 RMSE: 116.33387567215631 MAE: 82.82071897488187
MAPE: 24.627128473731144 RMSE: 183.47789820592766 MAE: 141.60326428415013
```

Final Model

- NN is always predicting on the lower side
- Light GBM is predicting on the higher side
- So, we ensemble the models, NN(35%) and Light GBM(65%) on this proportion to keep the data in normal distribution.
- New data is just shared 1 day before so we are still working.

Correlation between Important Features



From Light GBM Importance

- Social search Impression
- Social Search Working Cost
- Median Rainfall
- Digital Impression
- Inflation

A combination of these 10 variables were used to build the models and results found were effective.

	feature	split	gain
12	Column_12	3576	20.033833
33	Column_33	2706	16.537004
0	Column_0	3207	10.684496
25	Column_25	3045	9.535060
14	Column_14	2630	9.359199
35	Column_35	1886	5.566309
34	Column_34	1343	4.334998
3	Column_3	663	2.676049
20	Column_20	1197	1.851965
1	Column_1	796	1.149503
8	Column_8	474	0.989520
9	Column_9	626	0.988552
30	Column_30	545	0.972841
2	Column_2	697	0.971315
27	Column_27	478	0.957077
24	Column_24	405	0.948287
36	Column_36	658	0.910098
18	Column_18	465	0.824994
31	Column_31	587	0.811600
7	Column_7	480	0.750276

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

Note :

- The final MAPE score achieved is 37 on test data

```
In [206]: pred_mlp=final_mlp.predict(X_te)
     ...: pred_lgb=final_lgb.predict(X_te)
     ...: pred=0.65*pred_lgb+0.35*pred_mlp
     ...: mape = np.mean(np.abs(y_te - pred) /y_te ) * 100
     ...: print('mape on test data after ensembling',mape)
mape on test data after ensembling 37.55760834532677
```

Model Framework