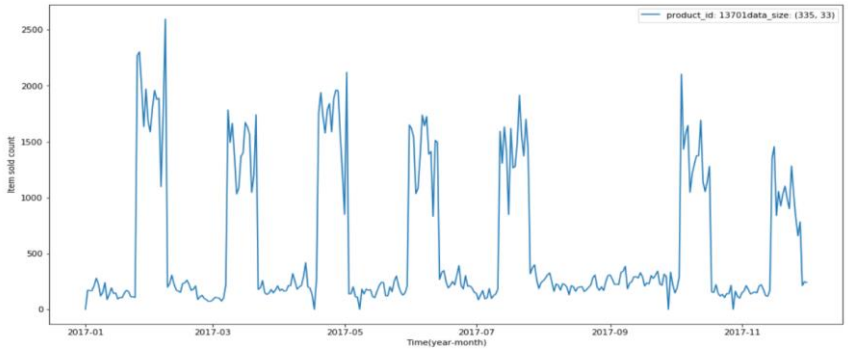
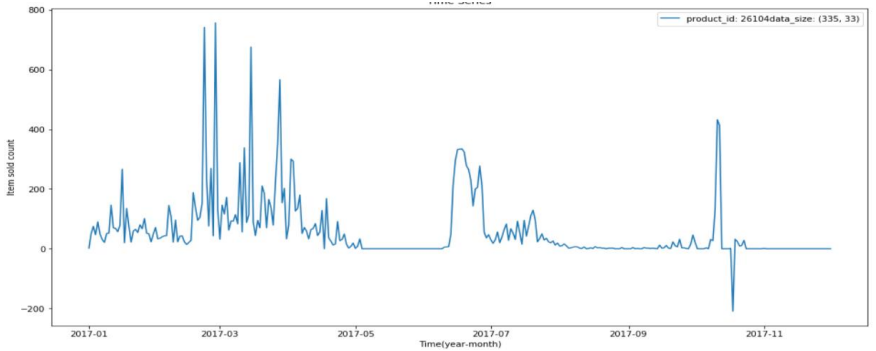
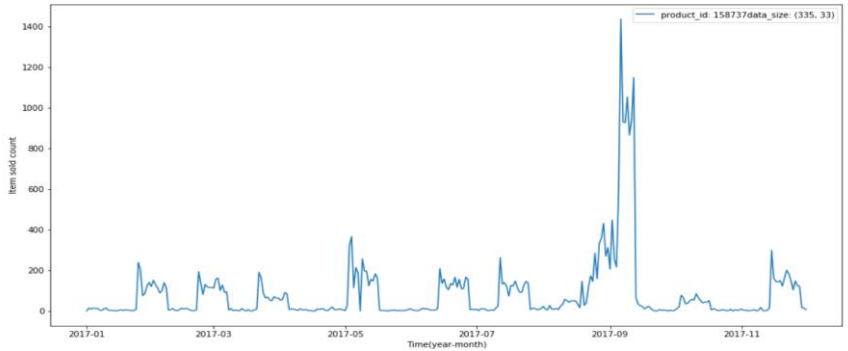
Sales Prediction Problem Challenge

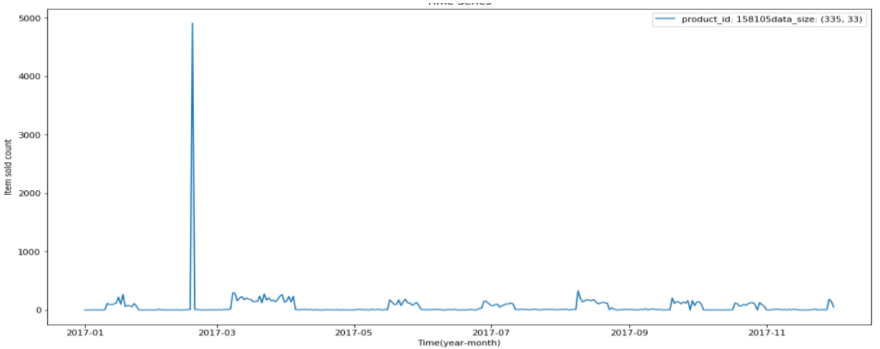
Bhavesh Bhansali

Sales Pattern Visualization



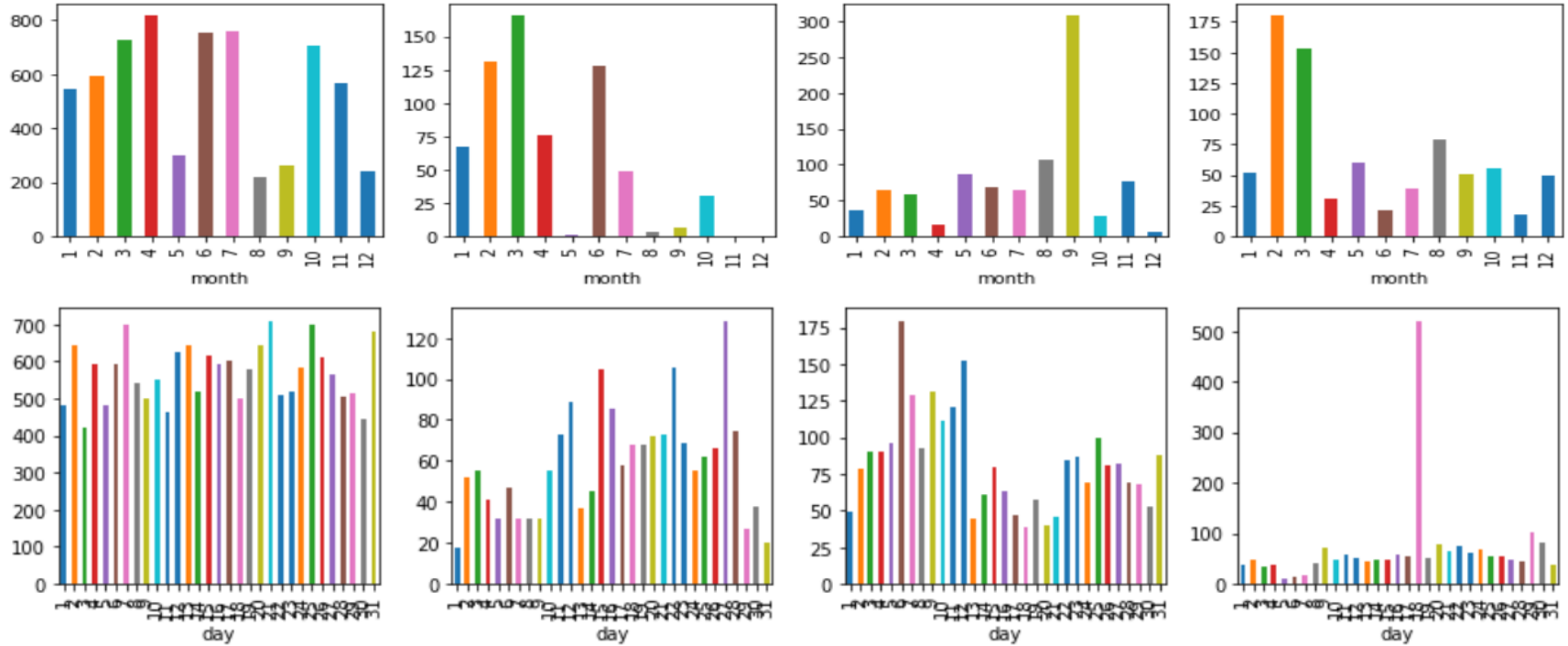






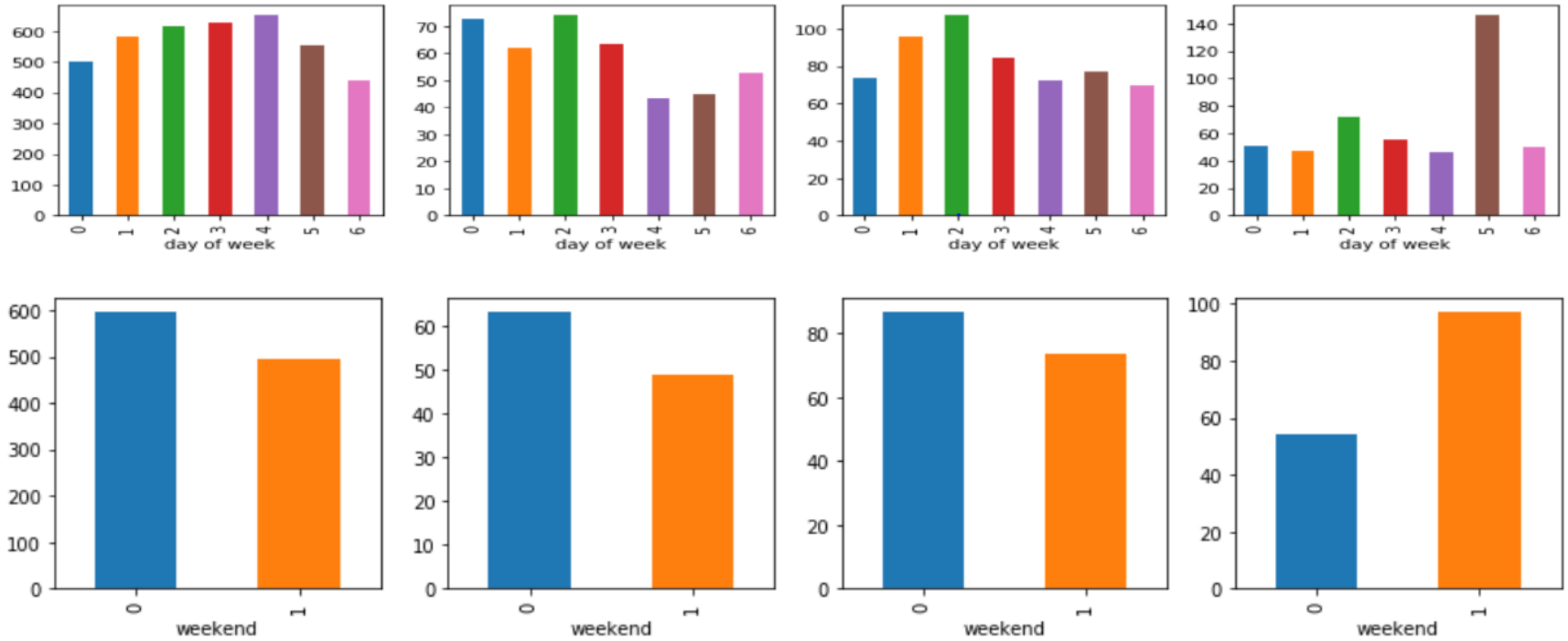
Sales Pattern: monthly and day of month

Item_ids: 13701, 26104, 158737, 158105

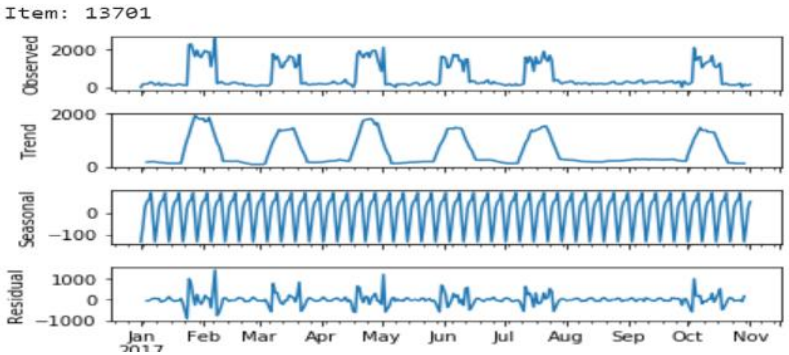


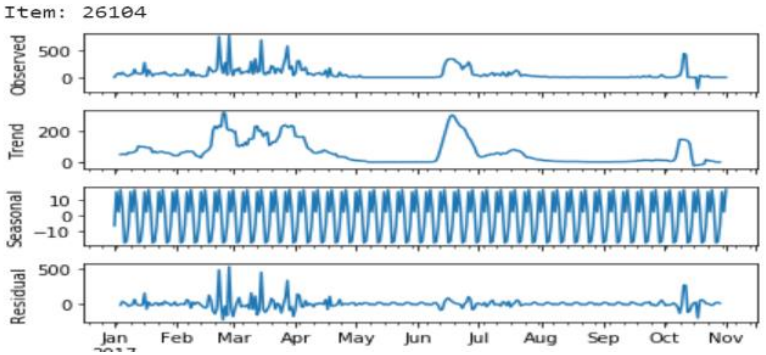
Sales Pattern: Day of week and Weekend/WeekDay

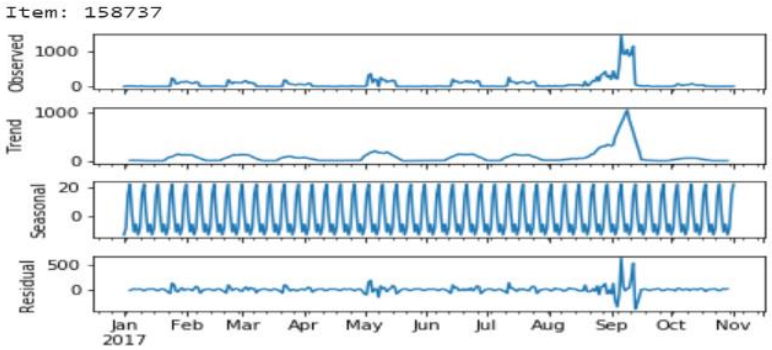
Item_ids: 13701, 26104, 158737, 158105

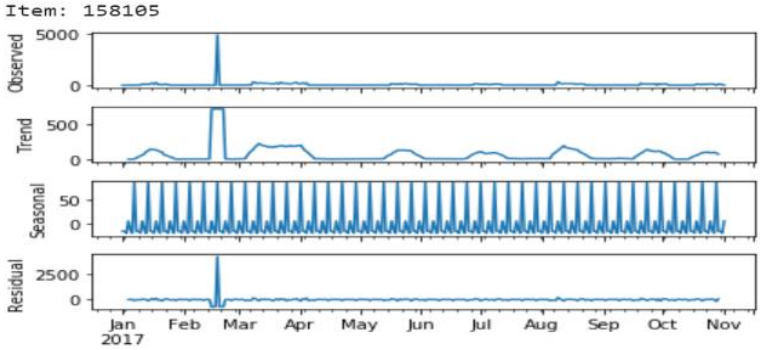


Time Series Components







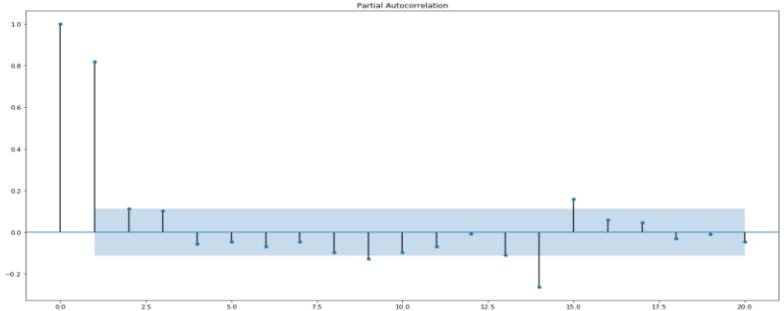


Time Series Model Analysis

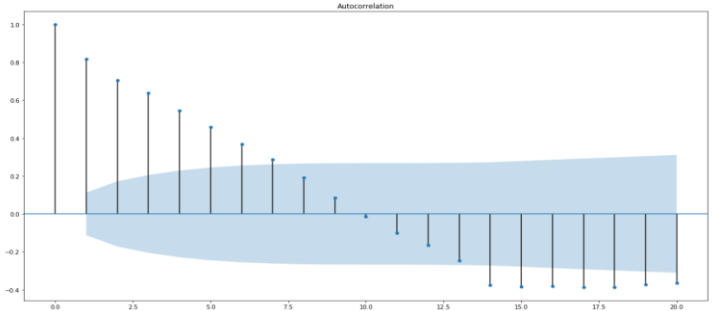
ITEM: 13701 (ARIMA: p=1, d=0, q=7) (For details: Notebook)

```
Results of Dickey-Fuller Test:
Test Statistic                  -5.785173e+00
p-value                          5.018897e-07
#Lags Used                       1.400000e+01
Number of Observations Used      3.200000e+02
Critical Value (1%)             -3.450952e+00
Critical Value (5%)             -2.870615e+00
Critical Value (10%)            -2.571605e+00
dtype: float64
```

Item: 13701
The Partial autocorrelation plot



Item: 13701
The autocorrelation plot




```
The RMSE of the ARIMA is 420.76471357226234
_____
___

The MSE of the ARIMA  is 177042.94418754795
_____
___

The MAE of the ARIMA  is 374.4606759981288
_____
___

_____
___

Sum of Original Sales for next 28 days:  16510.0
Sum of Predicted Sales for next 28 days:  [13759.73485805]
The difference between model original and predictions values of the ARIMA is [2750.26514195]
```

Deep Learning Models

DL Models

MLP, LSTM were tried without much fine tuning of hyper Parameters and number of layers.

Machine Learning Models

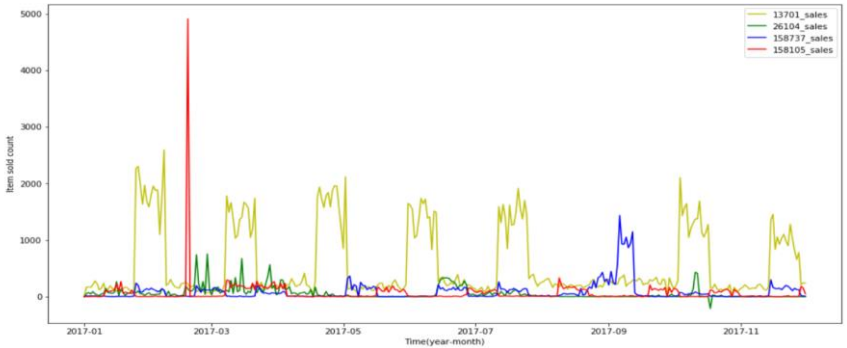
Data Assumptions

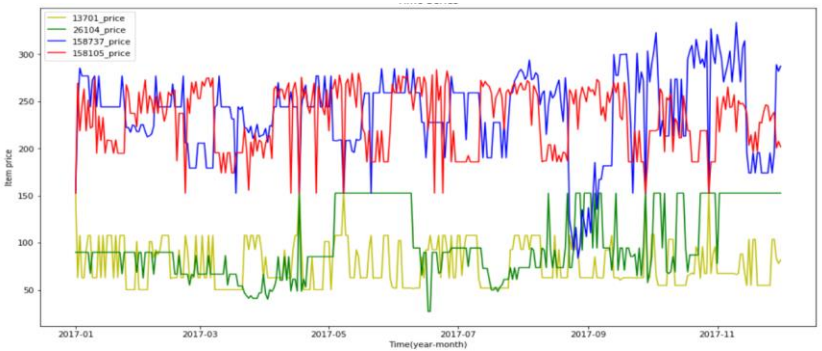
1. Assumed negative volume sold as an error and rectified it with positive sign. There can not be negative sales. However, it can be in case of return goods though.

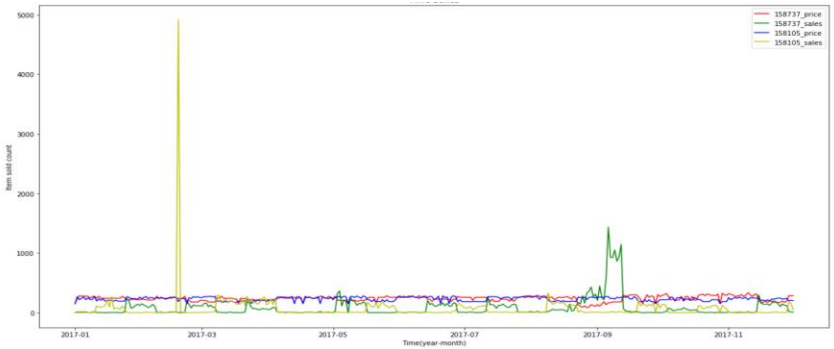
2. Assumed given data as observed data and did not consider any instance as an outlier (hence did not handle outliers).

3. Mean retail price is known in advance for next 30 days (this could be considered as recommended retail price or manufacturer recommended retail price)

Substitution/competitor items' analysis





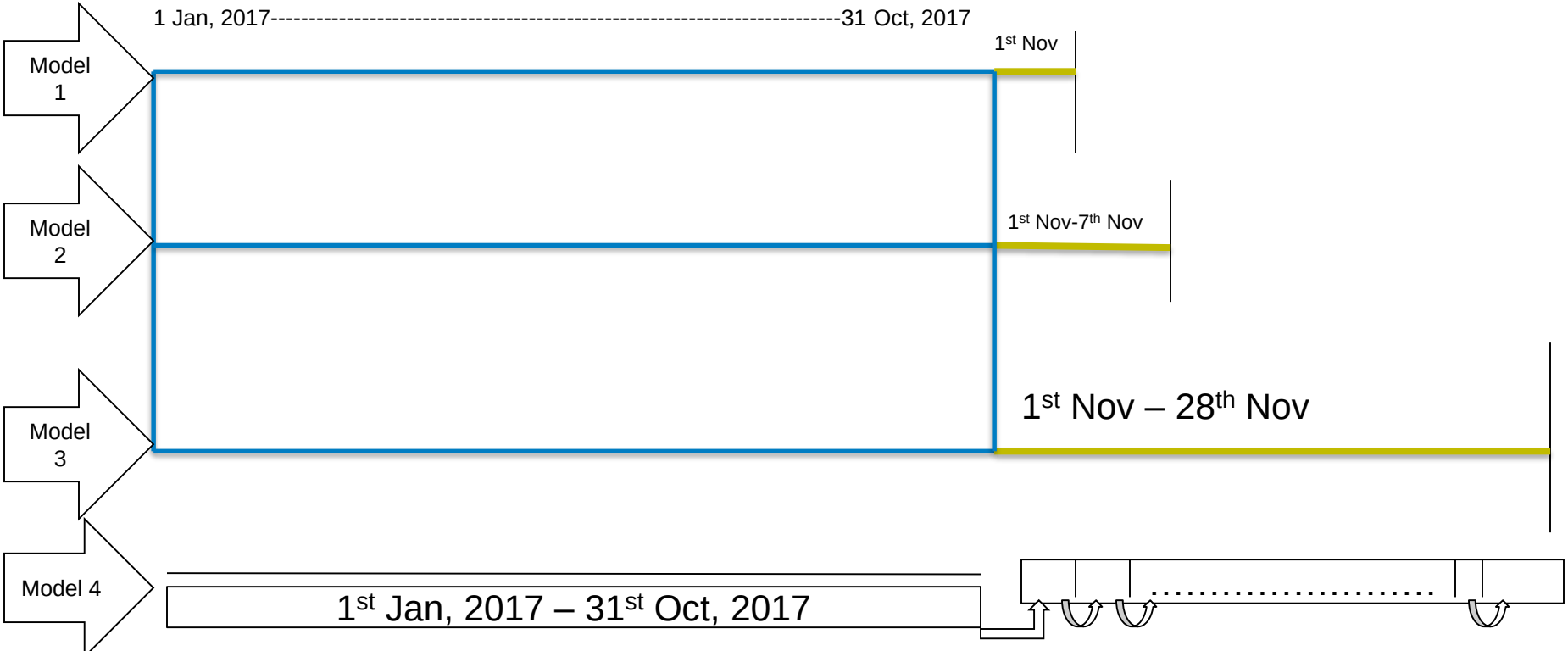


- Correlation Analysis

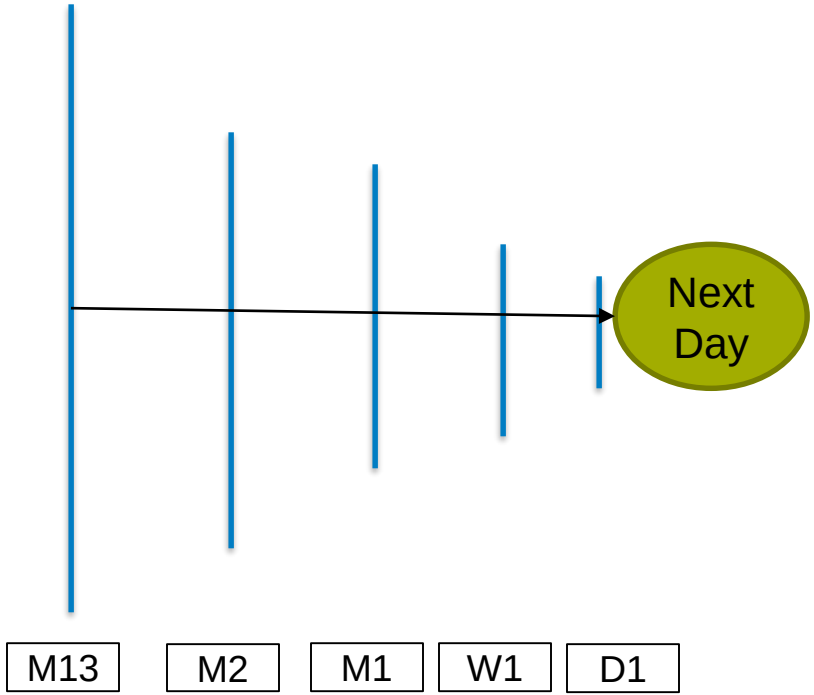
  Item 158737 and item 158105 sales were highly negatively correlated (one increases other decreases and vice versa)

- OLS fit (sales of one item with other items' sales-price and observed significant variables)

Data Split for Sales Forecast/Prediction



Model 1: One day Prediction/Forecast: Features



Given Features:

Last day's: Sales, Revenue, Cost, Maximum Price
Last week's Avg: Sales, Revenue, Cost, Maximum Price
Last 28 days: Sales, Revenue, Cost, Maximum Price
Between last 29 and 56 days: Sales, Revenue, Cost, Maximum Price
Last year Last 28 days: Sales, Revenue, Cost, Maximum Price
Mean retail price: assumption as it is available for next one day

Custom Features:

Time features: Month, day, weekend
Deviation: diff_day_before_and_last_weeks_avg
Deviation : diff_last_two_months_avg
Deviation : price_diff_max_price_in_previous_periods
Deviation : price_diff_retail_price_with_max_price_in_previous_periods
Deviation: price_diff_between_item_and_competitors
Stock_in_morning: Last day's remaining stock is next day's morning stock
Substitution: competitor items' price

Model 1: One day Prediction/Forecast: Linear Regression

```
13701
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'normalize': False, 'fit_intercept': True, 'copy_X': True}

The RMSE of the Linear Regression is 9.696306279465261
Original: [147.]
Prediction: [137.30369372]

The MAE of the Linear Regression is 9.696306279465261
```

```
26104
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'normalize': False, 'fit_intercept': True, 'copy_X': True}

The RMSE of the Linear Regression is 48.73741474132114
Original: [1.]
Prediction: [49.73741474]

The MAE of the Linear Regression is 48.73741474132114
```

```
158737
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'normalize': True, 'fit_intercept': True, 'copy_X': True}

The RMSE of the Linear Regression is 0.03412096073336812
Original: [10.]
Prediction: [9.96587904]

The MAE of the Linear Regression is 0.03412096073336812
```

```
158105
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'normalize': True, 'fit_intercept': True, 'copy_X': True}

The RMSE of the Linear Regression is 8.63738256302939
Original: [6.]
Prediction: [14.63738256]

The MAE of the Linear Regression is 8.63738256302939
```

Model 1: One day Prediction/Forecast: Support Vector Regression

```
_____________
Product: 13701
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'kernel': 'rbf', 'gamma': 0.001, 'degree': 4, 'C': 50}

Original: [147.]
Prediction: [134.65925937]
_____________________________________________________________________

The RMSE of the Support Vectors Regression is 12.340740631951519
_____________________________________________________________________

The MSE of the Support Vectors is 152.29387934509919
_____________________________________________________________________

The MAE of the Support Vectors is 12.340740631951519
_____________________________________________________________________
```

```
_____________
Product: 26104
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'kernel': 'rbf', 'gamma': 0.0001, 'degree': 3, 'C': 50}

Original: [1.]
Prediction: [13.36675048]
_____________________________________________________________________

The RMSE of the Support Vectors Regression is 12.366750482721267
_____________________________________________________________________

The MSE of the Support Vectors is 152.9365175018867
_____________________________________________________________________

The MAE of the Support Vectors is 12.366750482721267
_____________________________________________________________________
```

```
_____________
Product: 158737
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'kernel': 'rbf', 'gamma': 0.001, 'degree': 3, 'C': 50}

Original: [10.]
Prediction: [10.97290254]
_____________________________________________________________________

The RMSE of the Support Vectors Regression is 0.972902541140563
_____________________________________________________________________

The MSE of the Support Vectors is 0.9465393545577648
_____________________________________________________________________

The MAE of the Support Vectors is 0.972902541140563
_____________________________________________________________________
```

```
_____________
Product: 158105
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'kernel': 'rbf', 'gamma': 0.0001, 'degree': 4, 'C': 10}

Original: [6.]
Prediction: [22.41478016]
_____________________________________________________________________

The RMSE of the Support Vectors Regression is 16.414780160038468
_____________________________________________________________________

The MSE of the Support Vectors is 269.4450077023925
_____________________________________________________________________

The MAE of the Support Vectors is 16.414780160038468
_____________________________________________________________________
```

Model 1: One day Prediction/Forecast: Random Forest Regression

```
Product: 13701
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': 7, 'bootstrap': False}

Original: [147.]
Prediction: [108.09046591]
_____________________________________________________________________________________________

The RMSE of the Random Forest Regression is 38.90953409068182
_____________________________________________________________________________________________

The MSE of the Random Forest is 1513.9518431539304
_____________________________________________________________________________________________

The MAE of the Random Forest is 38.90953409068182
```

```
Product: 26104
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'n_estimators': 300, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_depth': 8, 'bootstrap': True}

Original: [1.]
Prediction: [8.21561866]
_____________________________________________________________________________________________

The RMSE of the Random Forest Regression is 7.2156186589913585
_____________________________________________________________________________________________

The MSE of the Random Forest is 52.06515263198425
_____________________________________________________________________________________________

The MAE of the Random Forest is 7.2156186589913585
```

```
Product: 158737
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'n_estimators': 400, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_depth': 8, 'bootstrap': True}

Original: [10.]
Prediction: [9.49957974]
_____________________________________________________________________________________________

The RMSE of the Random Forest Regression is 0.500420262248408
_____________________________________________________________________________________________

The MSE of the Random Forest is 0.2504204388687654
_____________________________________________________________________________________________

The MAE of the Random Forest is 0.500420262248408
_____________________________________________________________________________________________
```

```
Product: 158105
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_depth': 8, 'bootstrap': True}

Original: [6.]
Prediction: [11.54261844]
_____________________________________________________________________________________________

The RMSE of the Random Forest Regression is 5.542618436926704
_____________________________________________________________________________________________

The MSE of the Random Forest is 30.720619137359815
_____________________________________________________________________________________________

The MAE of the Random Forest is 5.542618436926704
```

Model 1: One day Prediction/Forecast: XGBoost Regression

```
Product: 13701
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'subsample': 0.8, 'n_estimators': 200, 'min_child_weight': 3, 'max_depth': 7, 'learning_rate': 0.1, 'gamma': 0,
'colsample_bytree': 0.7}

Original: [147.]
Prediction: [130.85118]
```

```
The RMSE of the XGBoost Regression is 16.148818969726562
```

```
The MSE of the XGBoost Regression is 260.7843541170005
```

```
The MAE of the XGBoost Regression is 16.148818969726562
```

```
Product: 26104
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'subsample': 0.7, 'n_estimators': 50, 'min_child_weight': 3, 'max_depth': 8, 'learning_rate': 0.1, 'gamma': 0,
'colsample_bytree': 0.8}

Original: [1.]
Prediction: [7.1524887]
```

```
The RMSE of the XGBoost Regression is 6.152488708496094
```

```
The MSE of the XGBoost Regression is 37.85311730817193
```

```
The MAE of the XGBoost Regression is 6.152488708496094
```

```
Product: 158737
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'subsample': 0.8, 'n_estimators': 300, 'min_child_weight': 4, 'max_depth': 8, 'learning_rate': 0.1, 'gamma': 0,
'colsample_bytree': 0.7}

Original: [10.]
Prediction: [10.8195305]
```

```
The RMSE of the XGBoost Regression is 0.8195304870605469
```

```
The MSE of the XGBoost Regression is 0.6716302192216972
```

```
The MAE of the XGBoost Regression is 0.8195304870605469
```

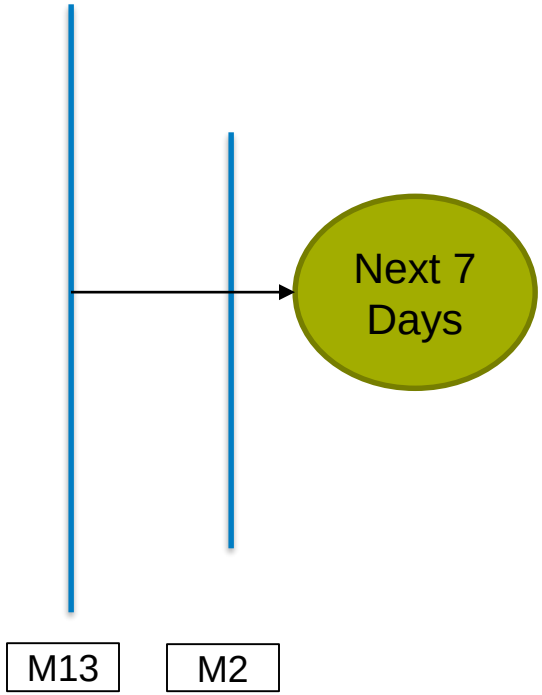
```
Product: 158105
# Tuning hyper-parameters for mean_squared_error

Best parameters set found on development set:

{'subsample': 0.7, 'n_estimators': 300, 'min_child_weight': 4, 'max_depth': 8, 'learning_rate': 0.05, 'gamma': 0,
'colsample_bytree': 0.8}

Original: [6.]
Prediction: [6.084872]
```

```
The RMSE of the XGBoost Regression is 0.08487176895141602
```

```
The MSE of the XGBoost Regression is 0.007203217164942544
```

```
The MAE of the XGBoost Regression is 0.08487176895141602
```

Model 2: Next 7 Days Prediction/Forecast: Features



Given Features:

Between last 29 and 56 days: Sales, Revenue, Cost, Maximum Price

Last year Last 28 days: Sales, Revenue, Cost, Maximum Price

Mean retail price: assumption as it is available for next one day

Custom Features:

Time features: Month, day, weekend

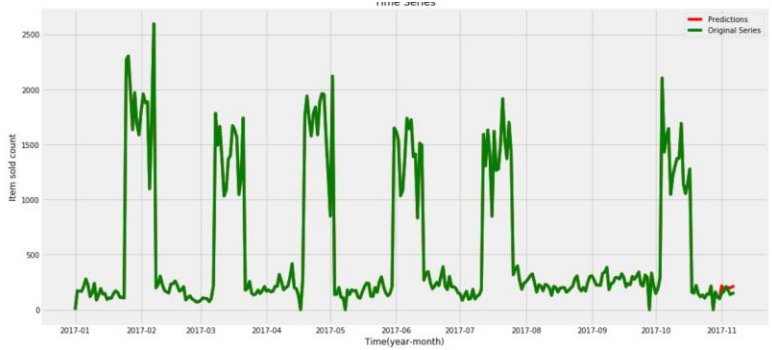
Deviation : price_diff_retail_price_with_max_price_in_previous_periods

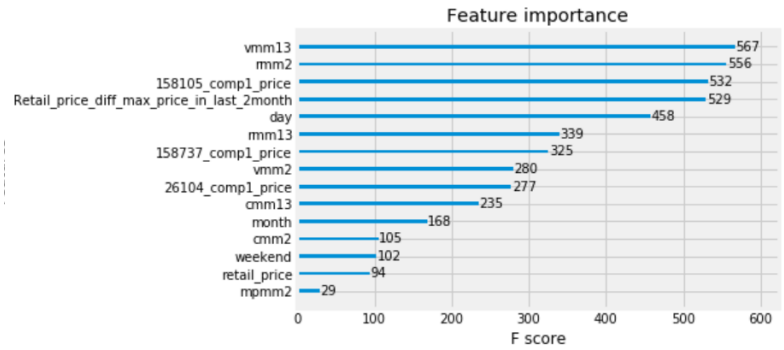
Deviation: price_diff_between_item_and_competitors

Substitution: competitor items' price

Model 2: Next 7 Days Prediction/Forecast: XGBoost Regression (Item1)

Item_id: 13701





The RMSE of the XGBoost Regression is 49.98765089394251

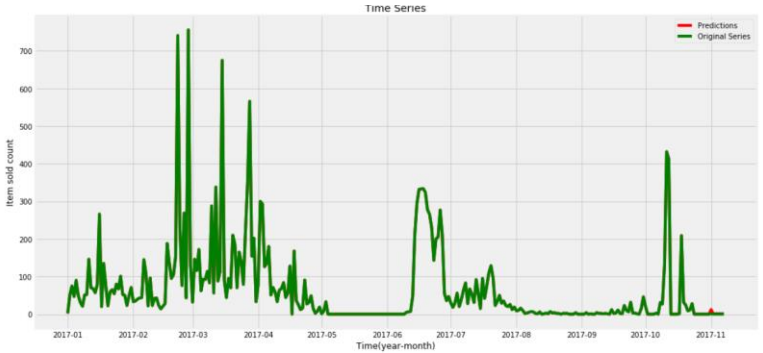
The MSE of the XGBoost Regression is 2498.7652418946714

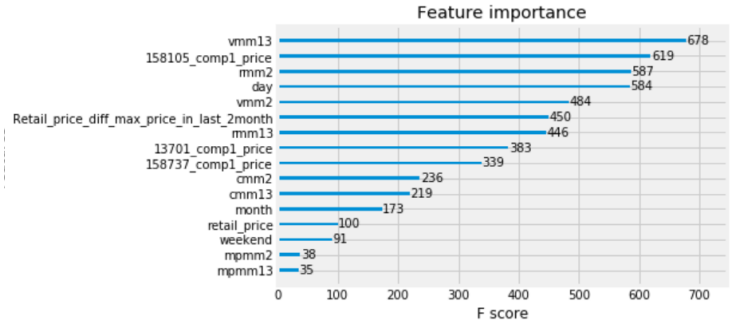
The MAE of the XGBoost Regression is 46.058606828962056

Sum of Original Sales for next 28 days: 1139
Sum of Predicted Sales for next 28 days: 1408.3726
The difference between model original and predictions values of the XGBoost Regression is -269.37255859375

Model 2: Next 7 Days Prediction/Forecast: XGBoost Regression (Item2)

Item_id: 26104

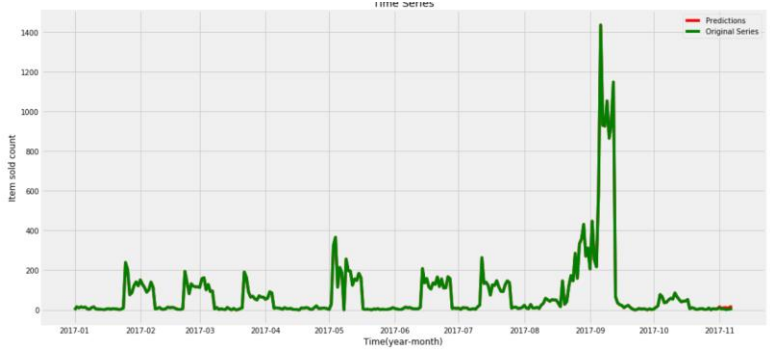


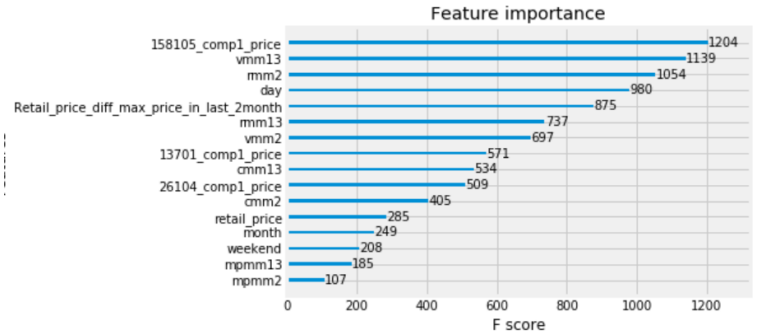



```
The RMSE of the XGBoost Regression is 4.403635421422245
_____________________________________________________________________________________

The MSE of the XGBoost Regression is 19.392004924804674
_____________________________________________________________________________________

The MAE of the XGBoost Regression is 2.6559190920421054
_____________________________________________________________________________________

_____________________________________________________________________________________

Sum of Original Sales for next 28 days:  1
Sum of Predicted Sales for next 28 days:  19.591434
The difference between model original and predictions values of the XGBoost Regression is -18.591434478759766
```

Model 2: Next 7 Days Prediction/Forecast: XGBoost Regression (Item3)

Item_id: 158737





The RMSE of the XGBoost Regression is 7.630082615568879

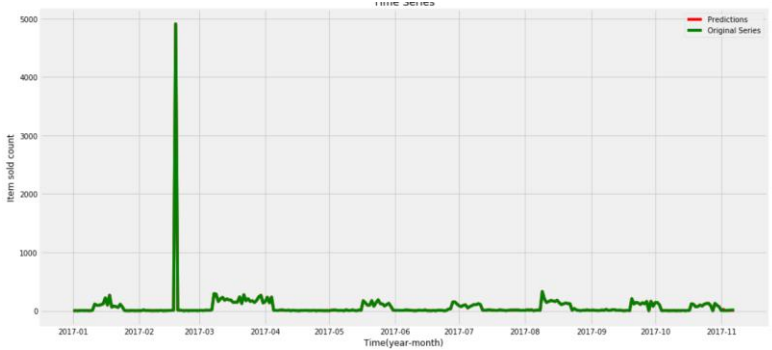
The MSE of the XGBoost Regression is 58.21816072040643

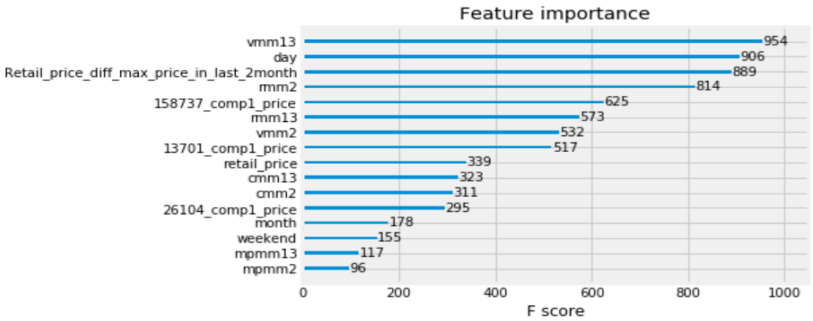
The MAE of the XGBoost Regression is 7.236275809151786

Sum of Original Sales for next 28 days: 33
Sum of Predicted Sales for next 28 days: 83.65393
The difference between model original and predictions values of the XGBoost Regression is -50.6539306640625

Model 2: Next 7 Days Prediction/Forecast: XGBoost Regression (Item4)

Item_id: 158105





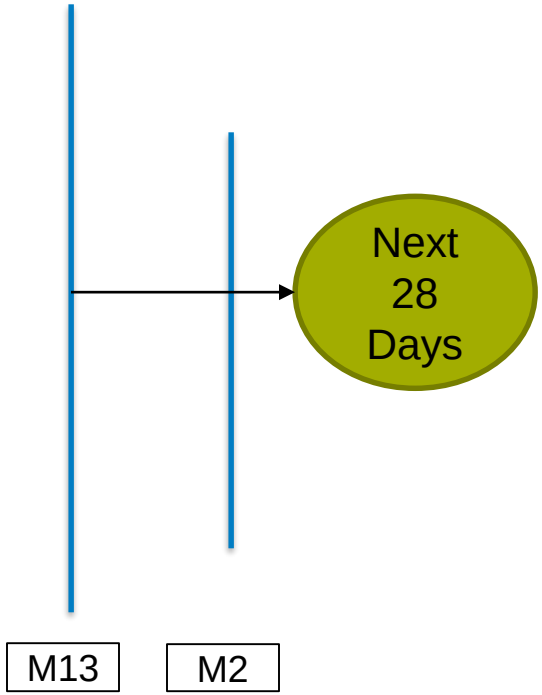
The RMSE of the XGBoost Regression is 10.287547893589707

The MSE of the XGBoost Regression is 105.83364166290201

The MAE of the XGBoost Regression is 8.074414696012225

Sum of Original Sales for next 28 days: 62
Sum of Predicted Sales for next 28 days: 49.536976
The difference between model original and predictions values of the XGBoost Regression is 12.463024139404297

Model 3: Next 28 Days Prediction/Forecast: Features



Given Features:

Between last 29 and 56 days: Sales, Revenue, Cost, Maximum Price

Last year Last 28 days: Sales, Revenue, Cost, Maximum Price

Mean retail price: assumption as it is available for next one day

Custom Features:

Time features: Month, day, weekend

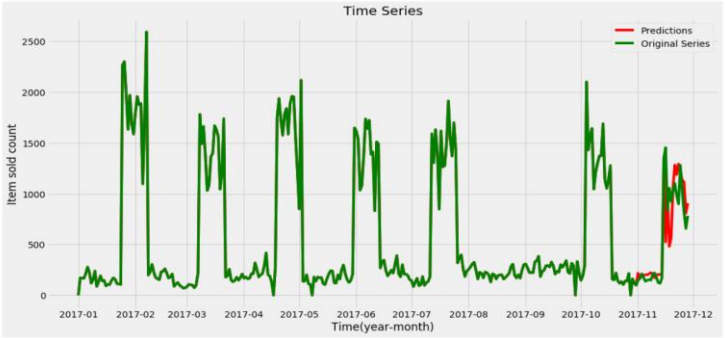
Deviation : price_diff_retail_price_with_max_price_in_previous_periods

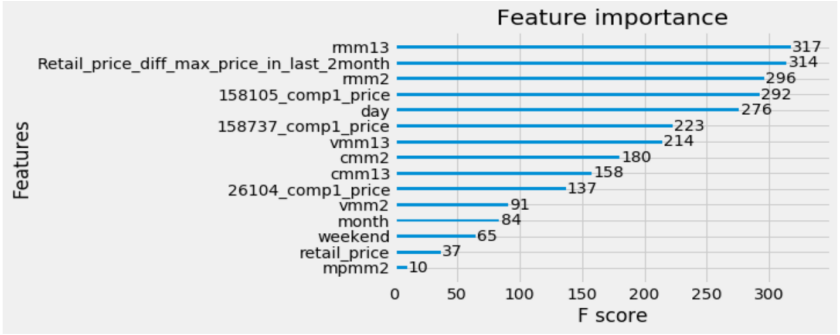
Deviation: price_diff_between_item_and_competitors

Substitution: competitor items' price

Model 3: Next 28 Days Prediction/Forecast: XGBoost Regression (Item1)

Item_id: 13701

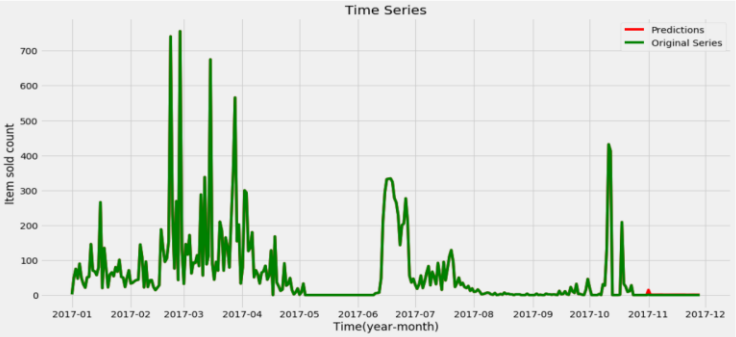


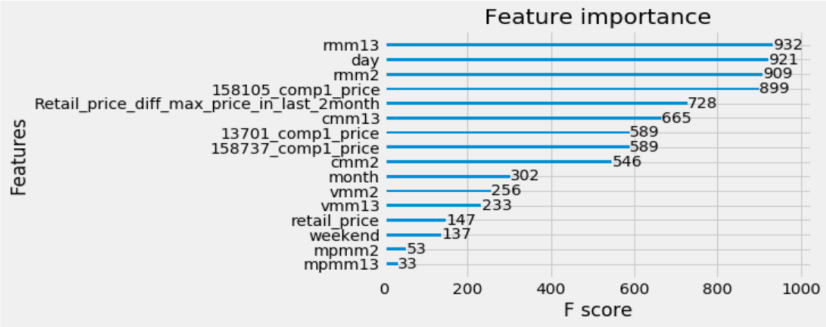



```
The RMSE of the XGBoost Regression is 251.62299997943518
_____________________________________________________________________________________________

The MSE of the XGBoost Regression is 63314.13411865083
_____________________________________________________________________________________________

The MAE of the XGBoost Regression is 154.82288306100028
_____________________________________________________________________________________________

_____________________________________________________________________________________________

Sum of Original Sales for next 28 days:  16510
Sum of Predicted Sales for next 28 days:  16633.7
The difference between model original and predictions values of the XGBoost Regression is -123.69921875
```

Model 3: Next 28 Days Prediction/Forecast: XGBoost Regression (Item2)

Item_id: 26104

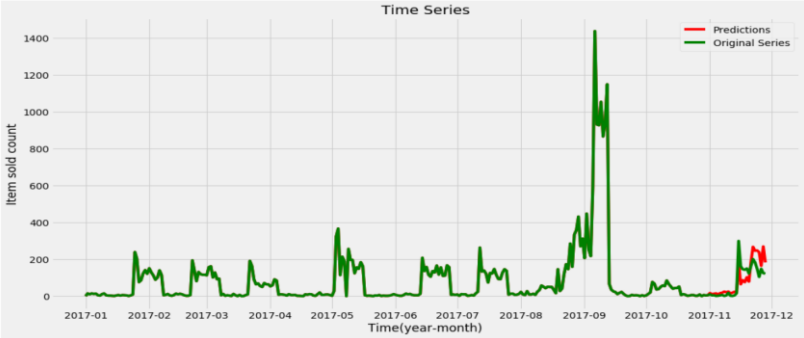


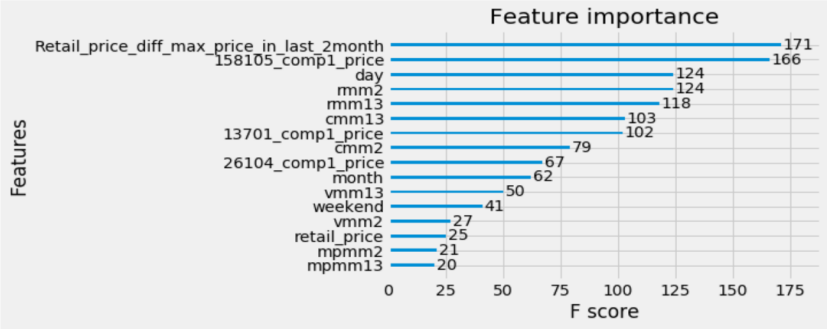



```
The RMSE of the XGBoost Regression is 2.7958344999777287
_____________________________________________________________________________________________

The MSE of the XGBoost Regression is 7.816690551265716
_____________________________________________________________________________________________

The MAE of the XGBoost Regression is 1.4935718753508158
_____________________________________________________________________________________________

_____________________________________________________________________________________________

Sum of Original Sales for next 28 days:  1
Sum of Predicted Sales for next 28 days:  42.820015
The difference between model original and predictions values of the XGBoost Regression is -41.82001495361328
```

Model 3: Next 28 Days Prediction/Forecast: XGBoost Regression (Item3)

Item_id: 158737

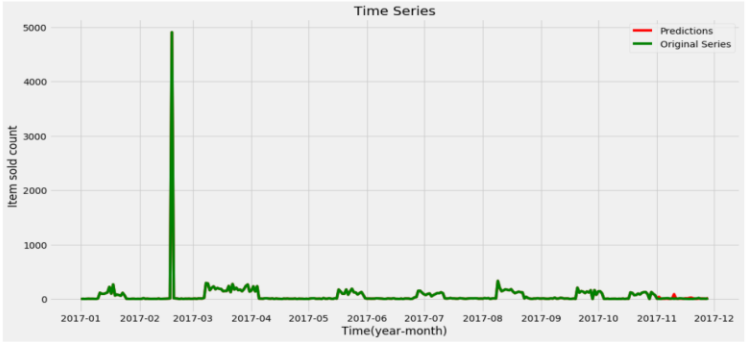


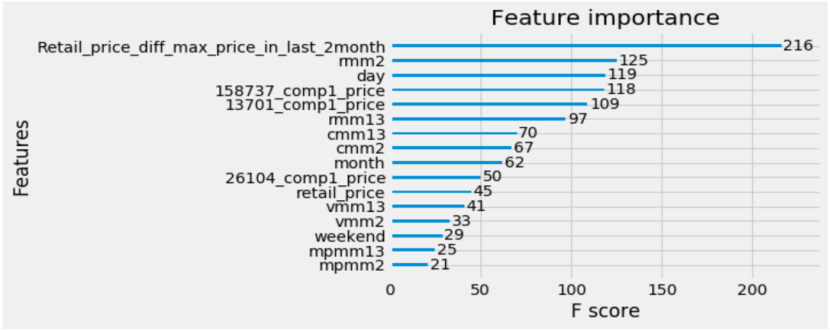



```
The RMSE of the XGBoost Regression is 59.102664135204
_____________________________________________________________________________________

The MSE of the XGBoost Regression is 3493.1249078787296
_____________________________________________________________________________________

The MAE of the XGBoost Regression is 42.93828340939113
_____________________________________________________________________________________

_____________________________________________________________________________________

Sum of Original Sales for next 28 days:  2296
Sum of Predicted Sales for next 28 days:  2644.8552
The difference between model original and predictions values of the XGBoost Regression is -348.855224609375
```

Model 3: Next 28 Days Prediction/Forecast: XGBoost Regression (Item4)

Item_id: 158105

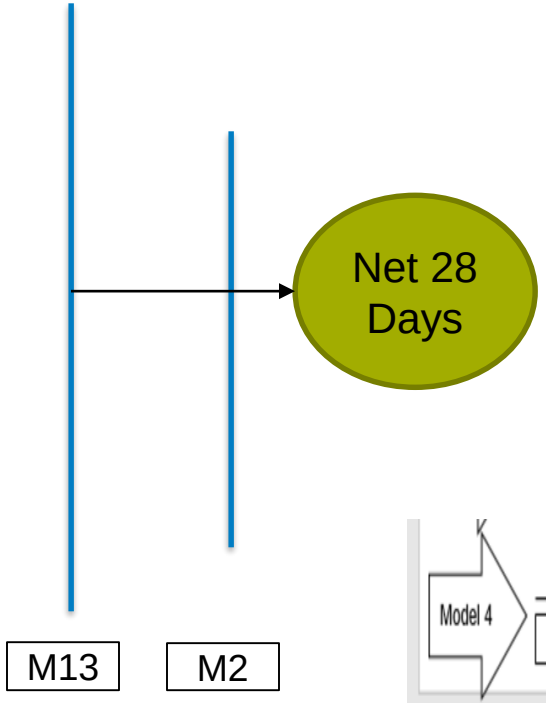





```
The RMSE of the XGBoost Regression is 19.30523689028837
_____________________________________________________________________________________________

The MSE of the XGBoost Regression is 372.6921713901509
_____________________________________________________________________________________________

The MAE of the XGBoost Regression is 10.585375377110072
_____________________________________________________________________________________________

_____________________________________________________________________________________________

Sum of Original Sales for next 28 days:  195
Sum of Predicted Sales for next 28 days:  333.50888
The difference between model original and predictions values of the XGBoost Regression is -138.50888061523438
```

Model 4: Next 28 Days Prediction/Forecast: Features

Predict for next one day and use that predictions to predict for next day and so on



Given Features:

Between last 29 and 56 days: Sales, Revenue, Cost, Maximum Price

Last year Last 28 days: Sales, Revenue, Cost, Maximum Price

Mean retail price: assumption as it is available for next one day

Custom Features:

Time features: Month, day, weekend

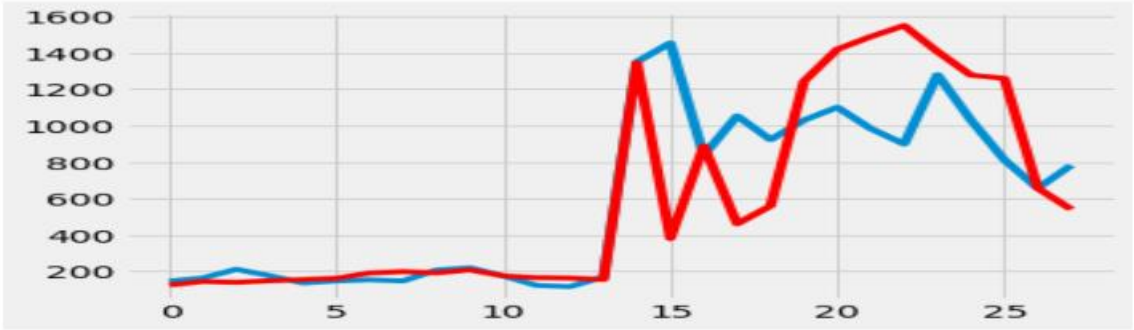
Deviation : price_diff_retail_price_with_max_price_in_previous_periods

Deviation: price_diff_between_item_and_competitors

Substitution: competitor items' price

Model 4: Next 28 Days Prediction/Forecast: XGBoost Regression (Item1)

Item_id: 13701




```
Sum of Original Sales for next 28 days:  16510.0
Sum of Predicted Sales for next 28 days:  [16822.215]
The difference between model original and predictions values of the is [-312.21484]
```

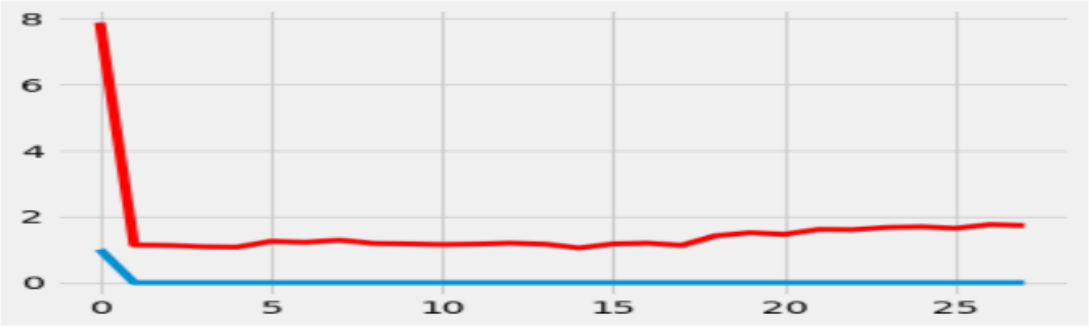
```
The RMSE of the XGBoost Regression is 317.1959296909404
```

```
The MSE of the XGBoost is 100613.26
```

```
The MAE of the XGBoost is 186.79834
```

Model 4: Next 28 Days Prediction/Forecast: XGBoost Regression (Item2)

Item_id: 26104




```
Sum of Original Sales for next 28 days:  1.0
Sum of Predicted Sales for next 28 days:  [43.728157]
The difference between model original and predictions values of the is [-42.728157]
```

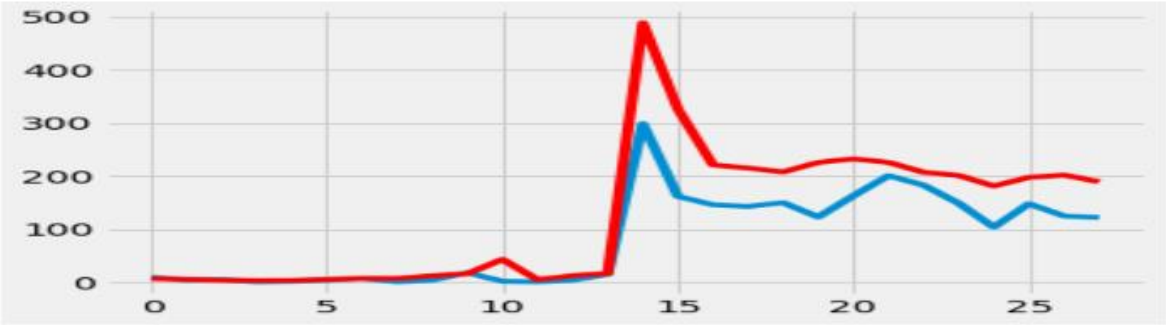
```
The RMSE of the XGBoost Regression is 1.8536375626710626
```

```
The MSE of the XGBoost is 3.4359722
```

```
The MAE of the XGBoost is 1.5260056
```

Model 4: Next 28 Days Prediction/Forecast: XGBoost Regression (Item3)

Item_id: 158737




```
Sum of Original Sales for next 28 days:  2296.0
Sum of Predicted Sales for next 28 days:  [3472.2683]
The difference between model original and predictions values of the is [-1176.2683]
```

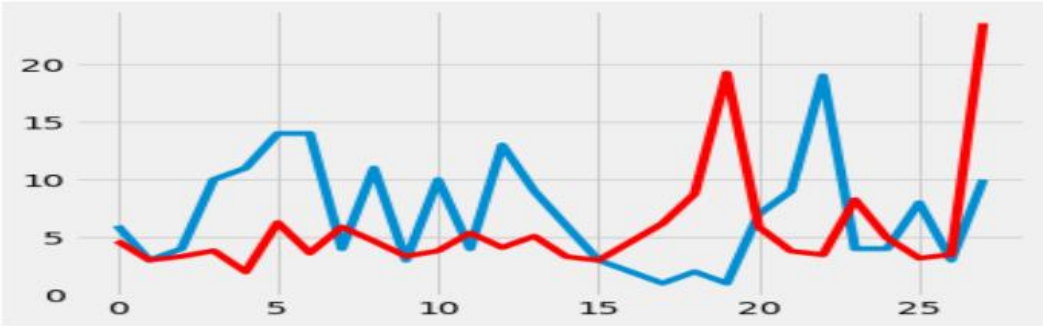
```
The RMSE of the XGBoost Regression is 64.80663094207644
```

```
The MSE of the XGBoost is 4199.8994
```

```
The MAE of the XGBoost is 42.434864
```

Model 4: Next 28 Days Prediction/Forecast: XGBoost Regression (Item4)

Item_id: 158105




```
Sum of Original Sales for next 28 days:  195.0
Sum of Predicted Sales for next 28 days:  [160.68086]
The difference between model original and predictions values of the is [34.319138]
```

```
The RMSE of the XGBoost Regression is 7.027136926326942
```

```
The MSE of the XGBoost is 49.380653
```

```
The MAE of the XGBoost is 5.2006383
```

Conclusion

In previous experiments, XGBoost worked best most of the time. However, ARIMA, DL or other ML models can be fine tuned to get more accurate predictions

Next Ideas

1. If items' data is sparse, it can be clustered with similar items and then can be forecasted.
2. Better strategy to select Competitor/Substitution/Similar items:
   - Clustering similar items based on item features
   - Clustering similar items based on price range
3. If item categories are known, we could add seasonal or sports events, i.e TV during world cups, swimming suite during summer, etc.
4. Deep learning models like to LSTM could be fine tuned given more data, which tend to capture long term relationship among data.
5. Ensemble of different (i.e, (RF_pred+XGBoost_pred+SVR_pred/3)) models or stacking can be applied to improve predictions.

Thank you ☺