BMAN73701: COURSEWORK
Group 24

Abstract

The retail industry has undergone substantial growth through digitalisation lately. However, navigating the current market scenario requires retailers to figure out complex challenges related to technology, sustainability, competition, and changing consumer expectations. This report focuses on a South America-based firm known as ABC, operating within the retail sector. The company is facing issues in handling the inventory levels of products to meet customer demands due to varied factors. For this, a machine learning approach has been designed to predict optimal levels of inventory required, in order to keep a subtle balance between demand and wastage. Based on exploratory analysis, historical data from July 2015 has been utilised for employing various supervised machine learning algorithms with cross validation and hyperparameter tuning. Of the 33 product classes, diverse sales trends were observed among them with minimal similarity, leading to the decision of creating individual models for each product. XGBoost Regressor showed a remarkable performance with the least neg_root_mean_squared_error score of -46 and therefore was chosen as the final model. This report also discusses the challenges and suggestions for efficient management of inventory.

Introduction

Adapting to the modern trends is important for sustainability in a continuously developing retail environment which has evolved a lot from brick-and-mortar stores to present day online platforms. Globalisation has played a vital role in this major shift. Amidst growing competition and rising demand, retailers are bound to expand their business to new locations and introduce new products enabling them to enter new sectors and cater to global demand rather than serving locally. This in turn, might also help them in increased revenue generation. There is no specific solution to stay consistent amidst these changing dynamics however taking care of some factors in business might help the firms in the long run.

South America's prominent grocery retailer firm ABC, being familiar with the intricacies of changing consumer demands, wants to take this step and ensure the consistency of sales and product availability at all times. To achieve this, precise prediction of inventory is a major task as over prediction can lead to wastage while under prediction can lead to products being out of stock and eventually customer dissatisfaction.

They have a large dataset encompassing essential sales data for 33 distinct product types across 54 stores, spanning from 01/01/2013 to 15/08/2017 and wish to utilise it to identify customer purchase patterns and forecast sales using machine learning to make future decisions. Currently, they use subjective forecasting with limited automation techniques to achieve this without efficiently utilising the abundant historical data available.

The specific description of the dataset is as follows:

1) id: unique identifier of a record
2) date: the date of a record.
3) store_nbr: the store number.
4) product_type: the type of the product.
5) sales: the total sales for a specific product type at a specified store.
6) special_offer: the promotion effort, where a higher number signifies a more intense promotional effort.

id	date	store_nbr	product_type	sales	special_offer
0	2013-01-01	1	AUTOMOTIVE	0.0	0
1	2013-01-01	1	BABY CARE	0.0	0
2	2013-01-01	1	BEAUTY	0.0	0
3	2013-01-01	1	BEVERAGES	0.0	0
4	2013-01-01	1	BOOKS	0.0	0

Table 1. First few records of the sales dataset.

Table 1 shows the first five records of the dataset.

A few observations by looking at the dataset and applying basic summary statistics are:

1) There are not many inconsistencies in the data apart from the fact that the firm remains closed on 25th December, Christmas every year. There are entries for all products and store combinations even if the products are not sold or are inactive with sales value as 0.
2) Since ABC is a grocery retailer, it makes sense that most of the products would be low quantity selling products which is also reflected in the data.
3) More than 75% of the products are sold without any promotional offers.

Exploratory Data Analysis

Having observed the basic characteristics of the dataset, we now delve into the distribution of sales and explore key patterns in the variables. We did some basic feature engineering to extract columns such as day, month, year, day of week in order to analyse trends on different frequencies of time.

In this section, we have leveraged a variety of visualisations to unravel patterns of sales across different dimensions – analysing overall trajectories, examining individual store performances, and delving into sales trends for each product.

We start by plotting sales for individual stores over time to look for any underlying patterns.

1. **Sales Pattern of Stores over time** *(Figure 1)*
    - For an overview, the graphs for a few stores have been included in the report.
    - The foremost observation that comes out is a general trend of increase in sales year on year for most of the stores which suggests an overall growth.
    - Overall, similar sales patterns were observed in groups of stores indicating shared dynamics and potentially suggesting a reason for store clustering based on sales.
    - Notably, Store 52 is the latest launched store (2017) since there is continuous zero sales before that. This will be difficult to predict since the dataset for training would be very small for this outlet-product combination.
    - Store 20, 21, 22, 29, 42, 29 commenced their operations from 2014 as the sales in the year was 0 for the whole of 2013. This also suggests that a number of new stores were launched in 2014 and there were business decisions being taken causing major changes in sales patterns.
    - Considering customised forecasting techniques or adequate preprocessing of data for stores with different opening dates can contribute to better predictions. Especially, for the case of recently opened stores.

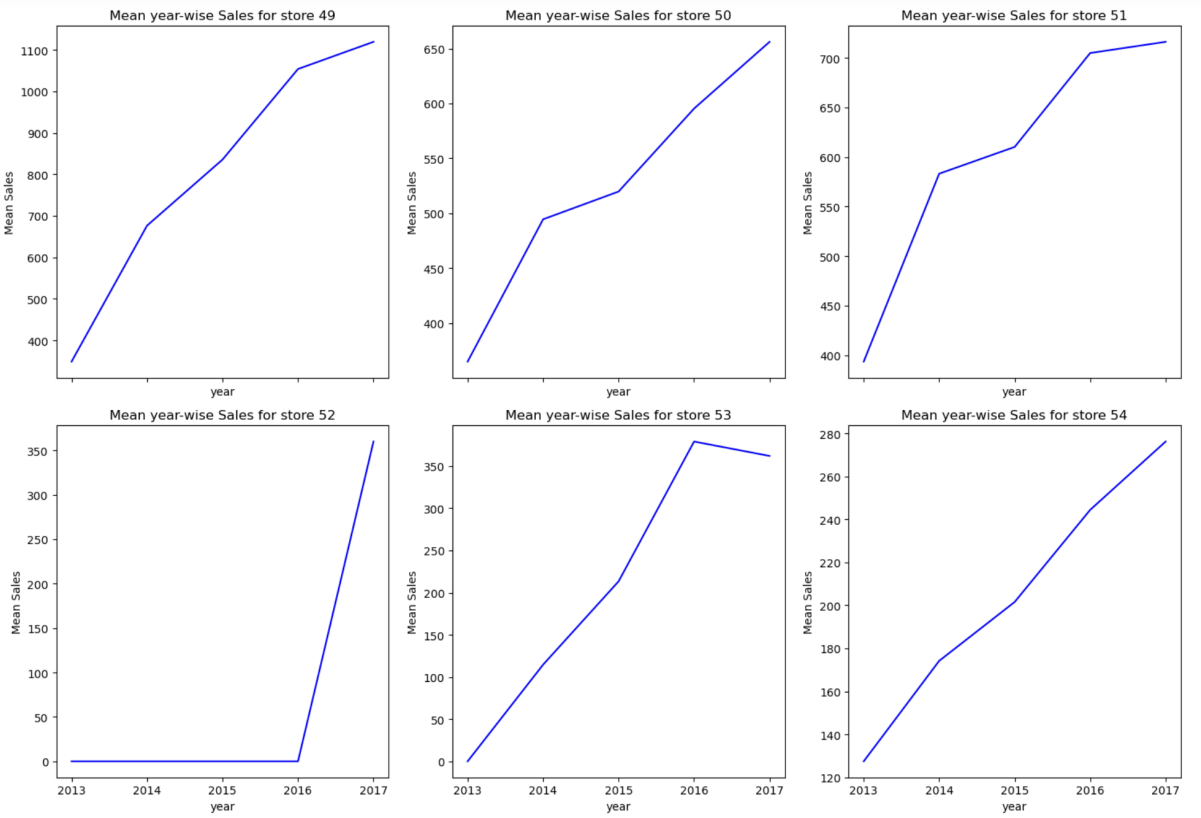


Figure 1. Yearly Mean sales over time for different stores

2. **Sales Pattern of Products over time** *(Figure 2)*
    - For an overview, the graphs for a few products have been included in the report.
    - Distinct sales patterns emerge for most products suggesting unique sales trends across the entire product catalogue. Tailored predictions for each product would be an efficient method. This aligns well with our aim of inventory management as with tailored predictions for each product, prediction could be even more precise.
    - A number of products sell in lower quantities such as Automotive, Beauty, Home Appliances, Home and Kitchen while higher selling products are daily usage product types such as grocery and dairy. The reason for low quantity sales in such products could be that these products are usually bought once by any customer and used for a long time. While the high selling products are mainly daily consumption or usage products which need restocking every now and then.
    - Another thing to note is that a few products began selling later than other products. Specifically, Books had zero sales up until 2014 i.e. new products launched.
    - A column start_date indicating the launch date of each product at each store was added to the dataset.

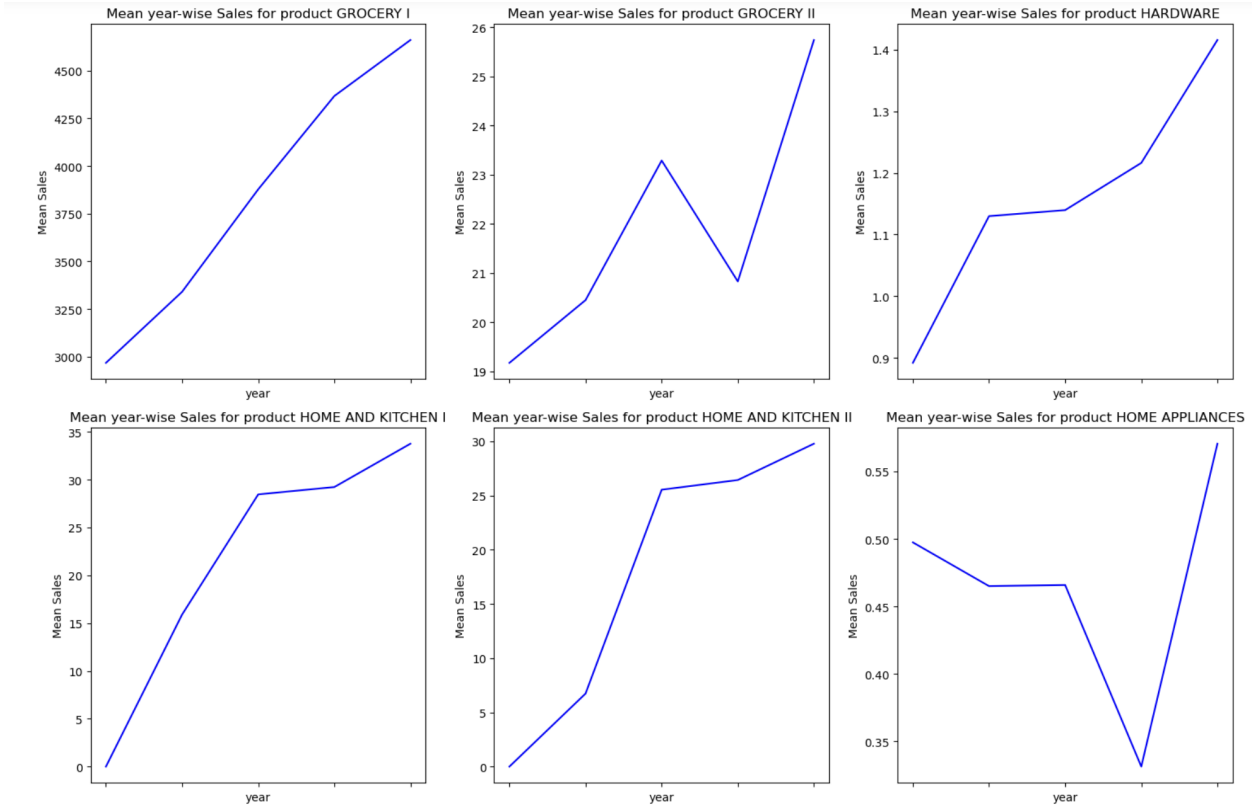


Figure 2. Yearly Mean sales over time for different products

3. **Overall Sales Patterns for each DOW over time + Anomaly Detection** *(Figure 3)*
    - An overall consistent year on year growth in sales is observed with comparatively lower sales in the year 2013 and high instability and considerable fluctuations in 2014 specifically. This could indicate a launching phase as we observed above that a number of stores and products were launched in the year 2014 potentially causing the deviations. The first half of 2015 observed a sudden dip (the only degrowth period) in sales which could be potentially attributed to some external influences affecting consumer purchasing patterns or market fluctuations. Post this, sales became quite stable suggesting a period of increased predictability of customer behaviour.
    - With the above information, we decided to trim our dataset and further explore the dataset post July 2015 as it would make more sense to further analyse trends on a stable dataset to conclude on patterns and eventual features for the predictive model.
    - Recognizing day-wise trends, weekends always marked the highest sales and Thursdays specifically emerged out as the lowest selling day of week.
    - High sales were observed during the start of the month, could be due to attainment of salaries.
    - Overall, highest sales were observed in the month of December followed by marginally low sales in January.

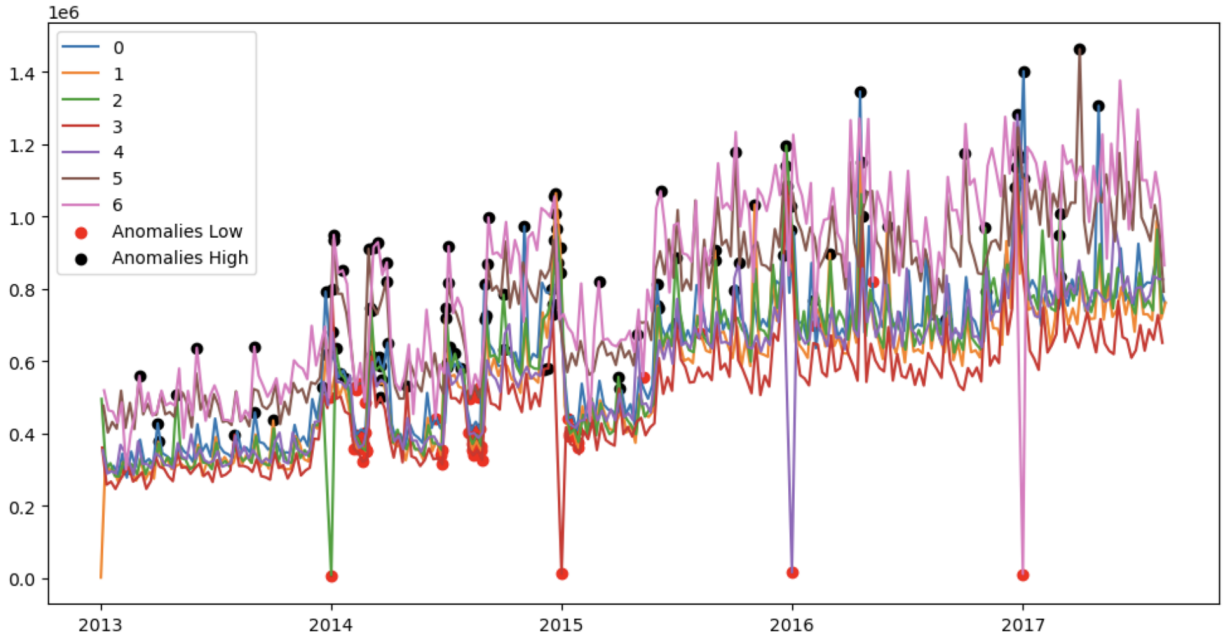


Figure 3. DOW-wise Overall Sales pattern over time

4. **Lag Features** *(Figure 4)*
    - Our dataset clearly follows a seasonal trend with repetition of trends at different frequency intervals which suggests that values at a certain point of time are related to the values at previous times. Lag features help the model understand and incorporate recent patterns and trends, making it better equipped to make accurate predictions.
    - As can be seen from *Figure 4*, lag feature lag_yoy_sales is somehow following the current actual values trend.
    - To include this feature, a feature called lag_yoy_sales representing the DOW's average from the previous year

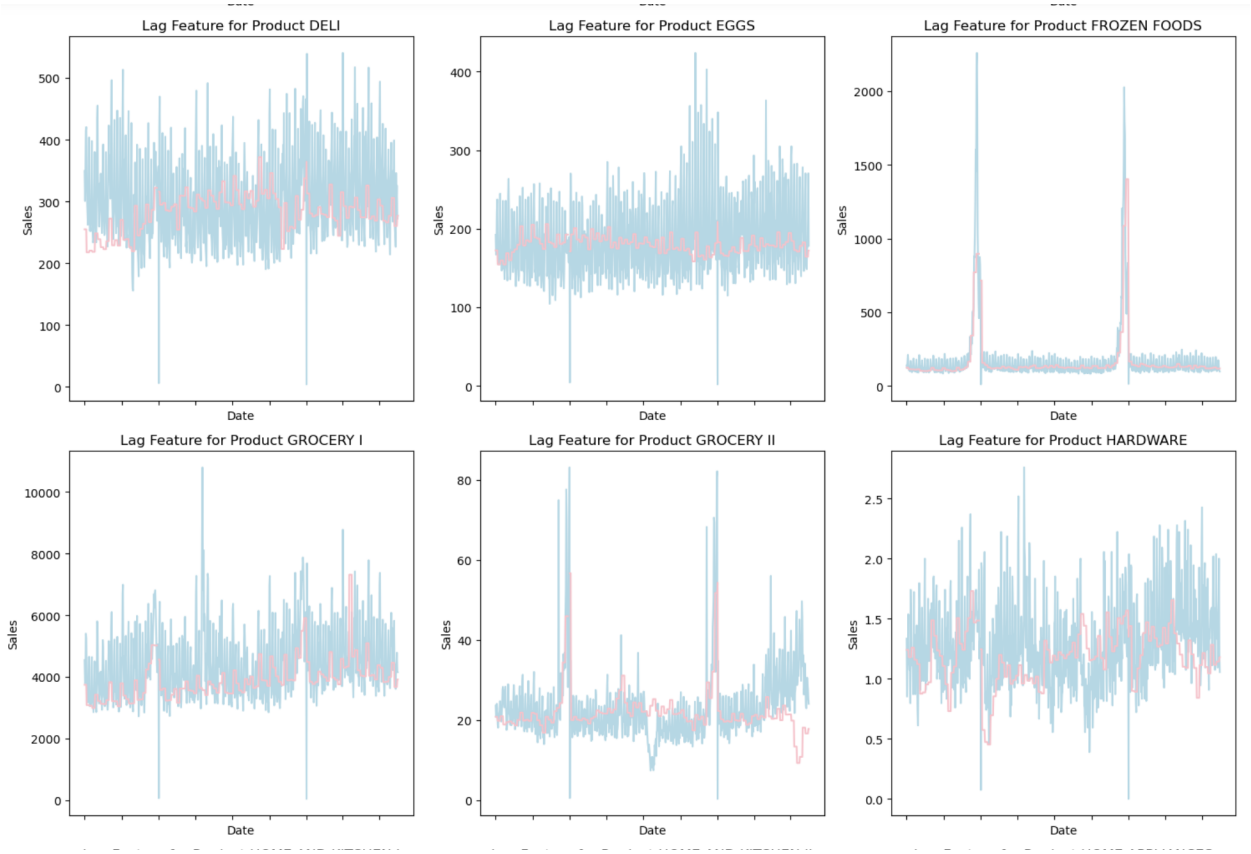


Figure 4. Understanding Lag Feature Trends

5. **Sales vs Special Offer**

   <u>Monthly Trend</u> *(Figure 5)*

   - A modest correlation is found between sales and special offers over months suggesting the impact of promotional offers on sales.
   - Peak offers are released around May and December suggesting festive offers for Easter and Christmas. Interestingly, even though the highest offers are released in May, December stands out with highest purchasing activity.

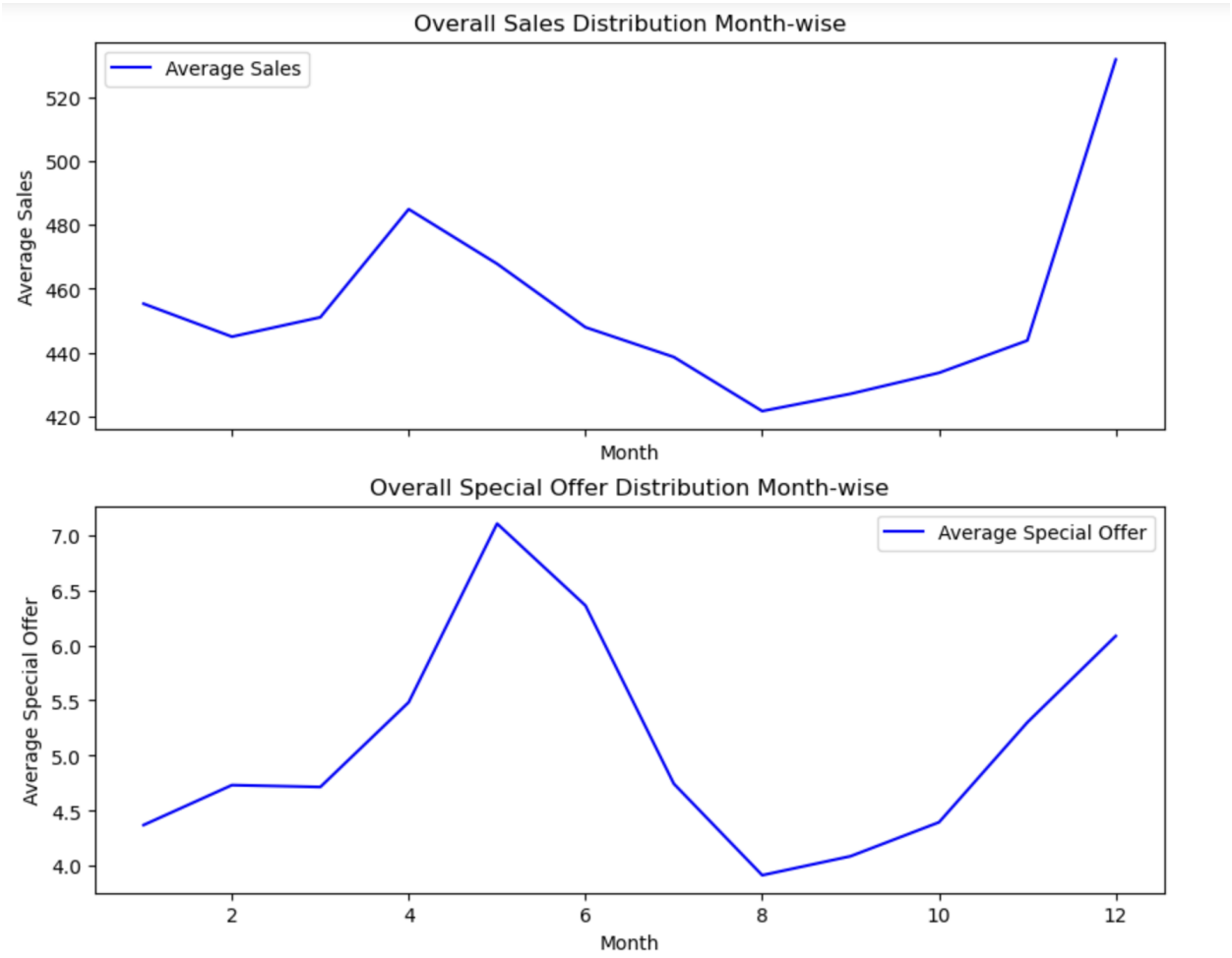


Figure 5. Monthly trends of average sales and special offers

DOW Trend *(Figure 6)*

- A highly consistent pattern for sales as well as promotions is observed each year with respect to DOW.
- Even though the offers on Wednesdays are the highest, sales are at its peak on weekdays suggesting the need for re-tailoring promotional strategies.
- Overall DOW exhibits a repetitive pattern for sales as well as special offers which can be an important feature for predictions.

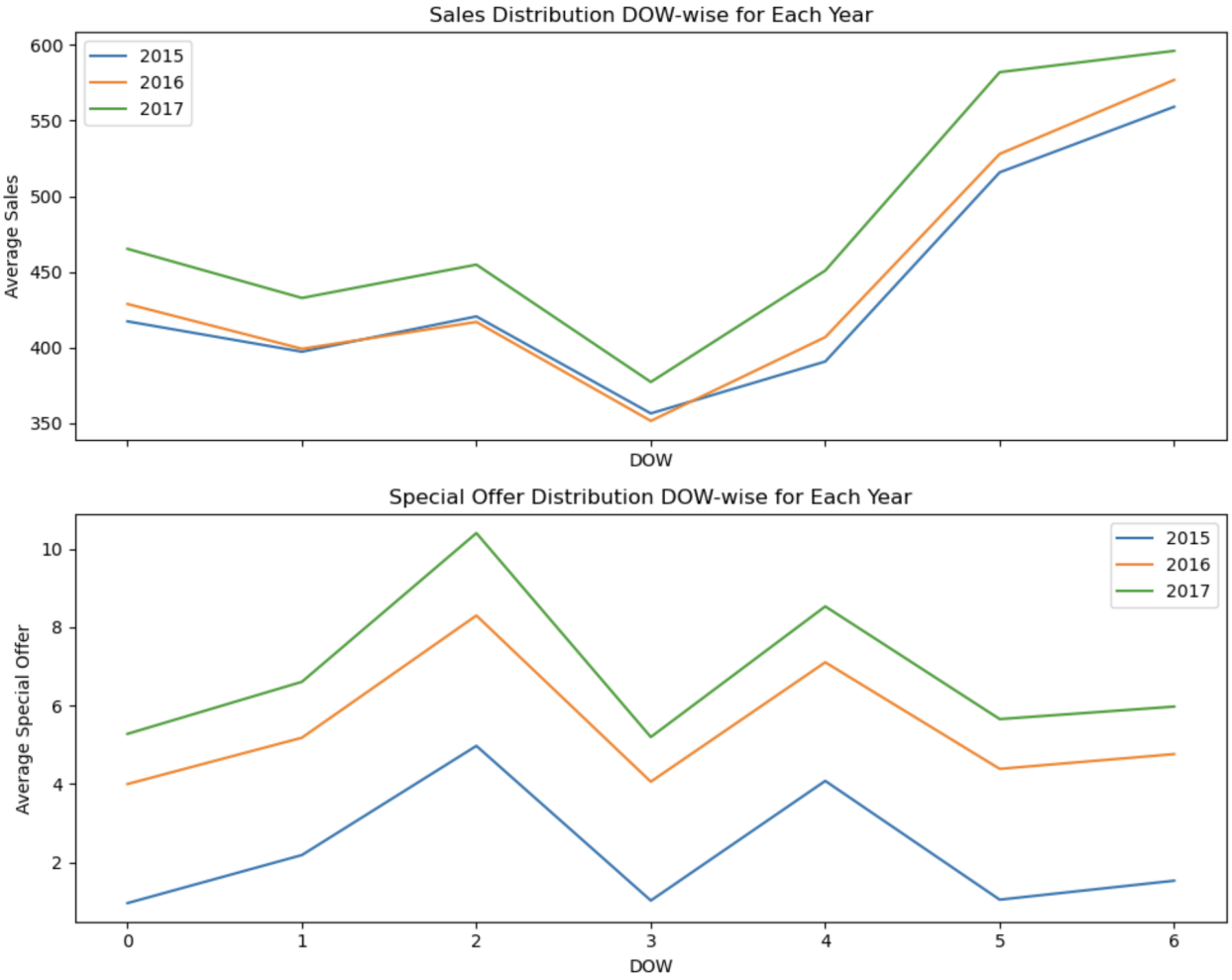


Figure 6. DOW wise trends of average sales and special offers for each year

With a broader understanding of sales patterns over time, we move to a more granular exploration on each product.

Product Specific Patterns + Clustering Possibility

1. **Sales Distribution Clusters**
    - For an overview, the graphs for a few products have been included in the report.
    - From initial exploration, it was observed that distinct products exhibited different sales ranges and patterns amongst them due to which we tried a clustering approach to categorise products into similar sales groups.
    - Three Clusters with values as 1 (Low), 2 (Medium) and 3 (High) were formed to represent their monthly average sales segmentation for each product as shown in *Figure 7*.
    - Similarly, a daily average sales cluster was also formed representing daily average sales groups for each product as shown in *Figure 8*.

-

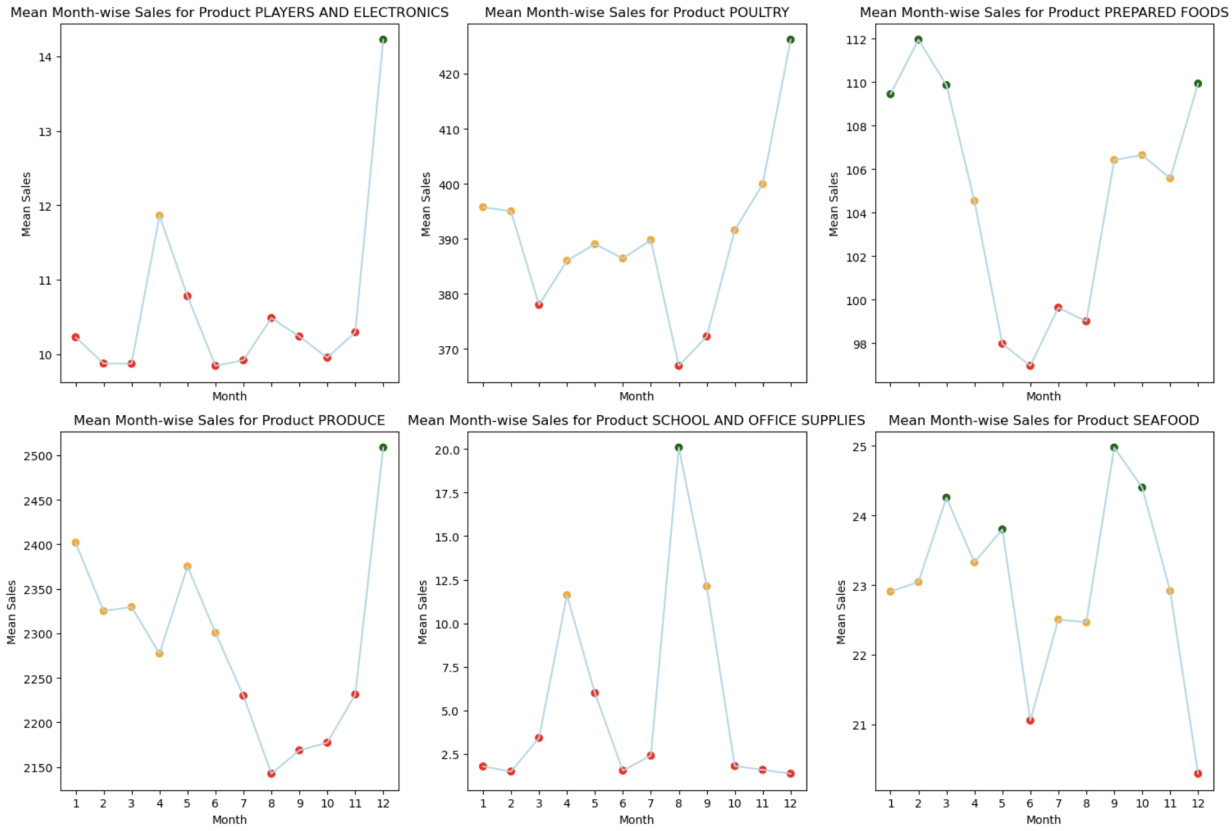


Figure 7. Monthly Average Sales clustering for each product

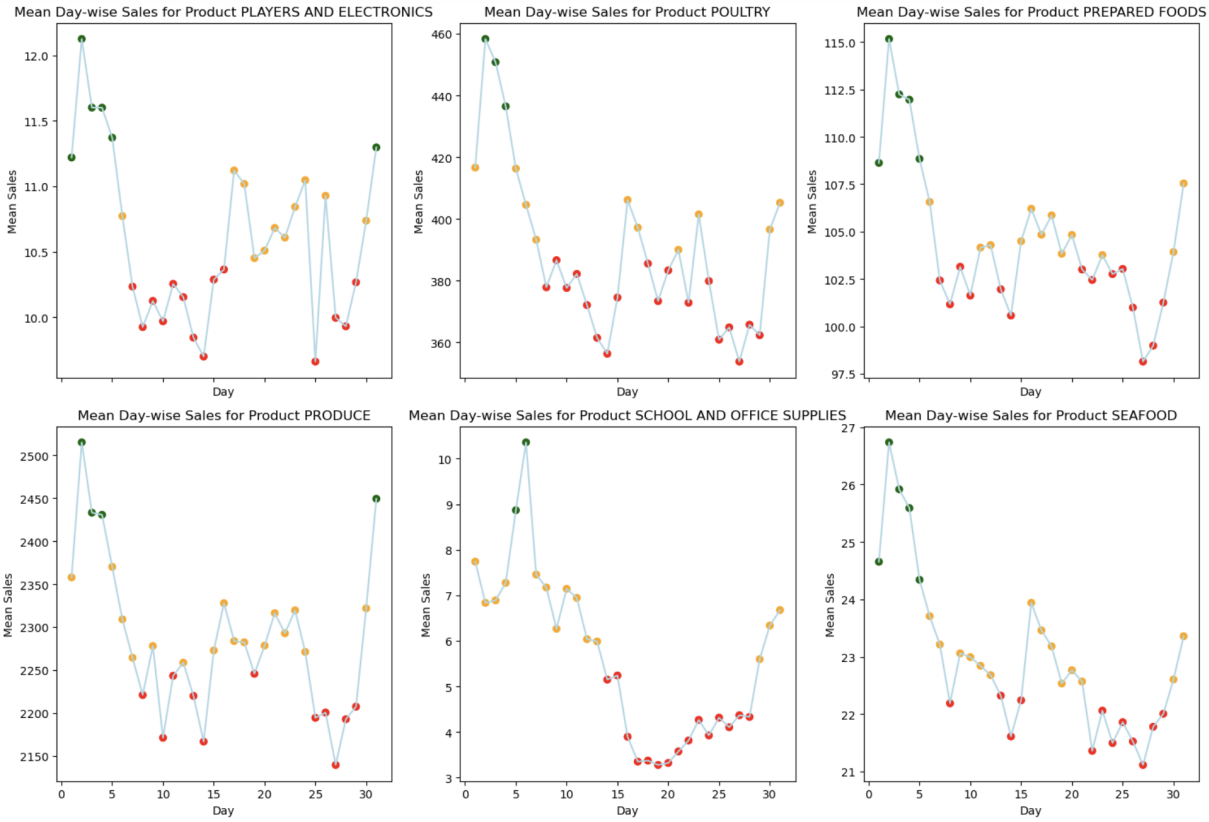


Figure 8. Daily Average Sales clustering for each product

2. **Most Selling Products** *(Figure 9)*
   - Upon plotting the yearly distribution of products as shown in *Figure 8*, approximately 80% of the sales came from just 5 products namely Grocery I, Beverages, Produce, Cleaning and Dairy. Among these, Grocery I & Beverages alone account for approximately 50% of the total sales each year.
   - Given their sales contribution, evaluating model performance for these products would be of paramount importance.

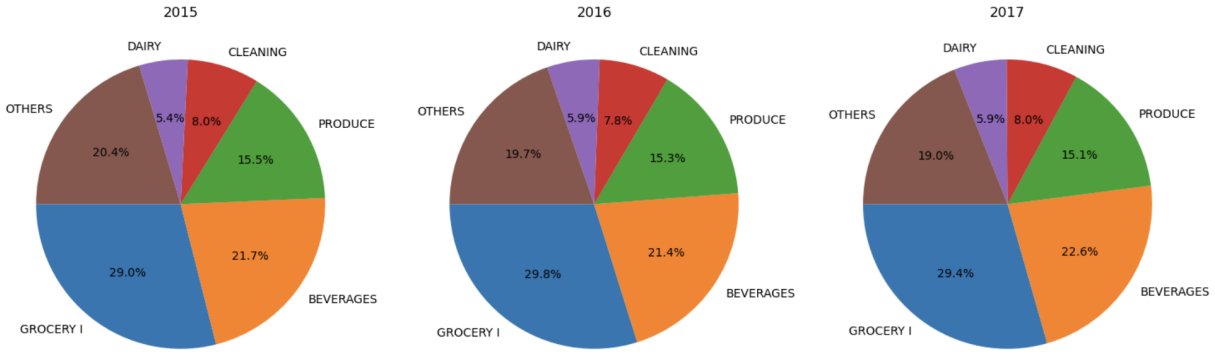


Figure 9. Sales Contribution percentage of different products for each year

3. **Correlation Matrix b/w special offer on products and sales** *(Figure 10)*

- Special Offers don't work equally effectively for all products and are generally less effective in increasing sales.
- School and Office supplies showed a standout correlation between sales & special offers each year underlining the feature importance of promotional offers only for selected products.
- Table 2 shows products with more than 20% correlation amongst sales & offers.

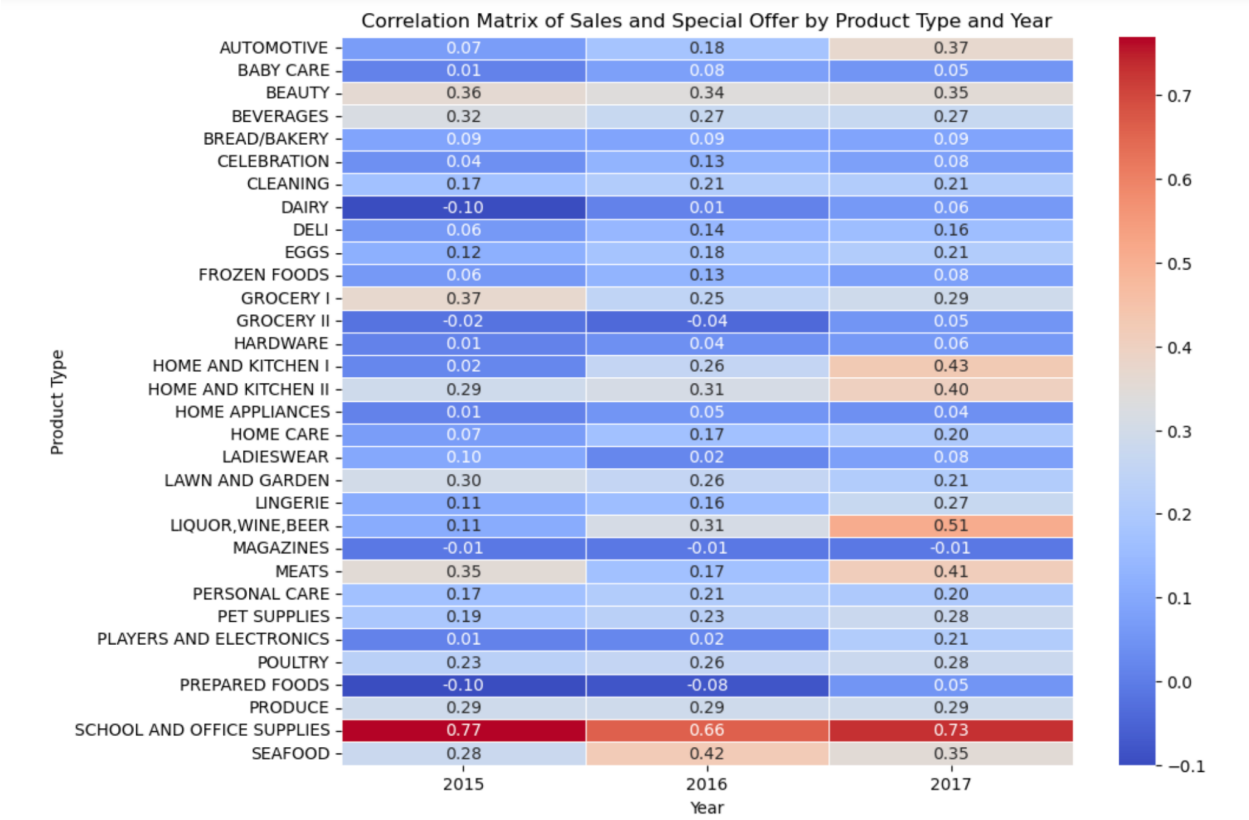


Figure 10. Correlation Matrix between sales and special offers by product for each year

year	2015	2016	2017
product_type			
BEAUTY	0.36	0.34	0.35
BEVERAGES	0.32	0.27	0.27
GROCERY I	0.37	0.25	0.29
HOME AND KITCHEN II	0.29	0.31	0.40
LAWN AND GARDEN	0.30	0.26	0.21
POULTRY	0.23	0.26	0.28
PRODUCE	0.29	0.29	0.29
SCHOOL AND OFFICE SUPPLIES	0.77	0.66	0.73
SEAFOOD	0.28	0.42	0.35

Table 2. Products with more than 20% correlation amongst sales & offers

4. **Special Offer Clusters** *(Figure 11)*
    - We formed monthly special offer clusters for products in Table n following the same trend of segments as for others with values as 1 (Low), 2 (Medium) and 3 (High) representing their monthly average offer segmentation for each month.

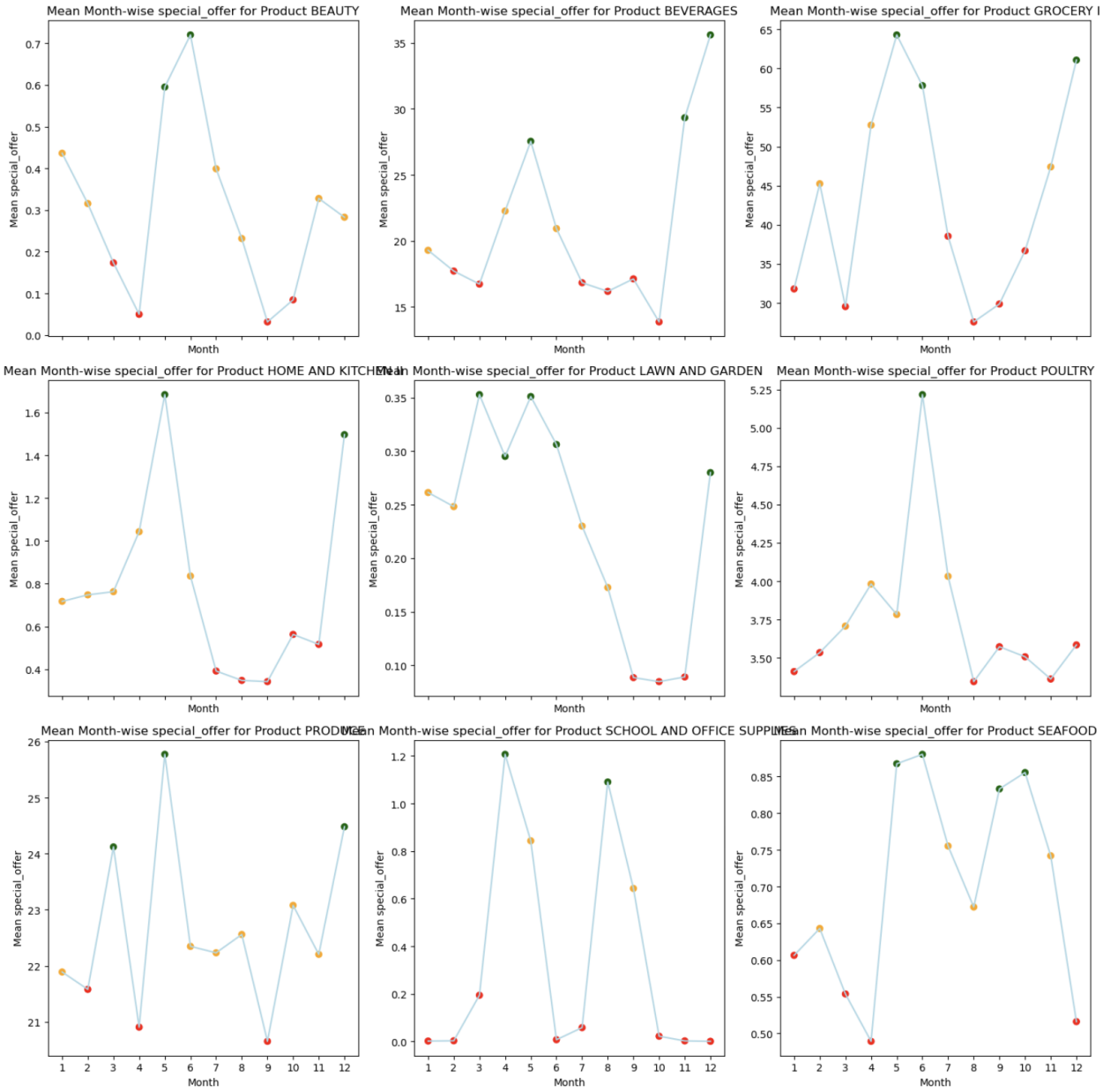


Figure 11. Special offer clusters for products with higher correlation between sales and offers

Store Specific Patterns + Clustering Possibility

1. **Correlation amongst stores** *(Figure 12)*
    - A notable correlation amongst all stores was found highlighting the importance of considering store clustering based on sales. Store 52 specifically had a low correlation with any of the stores since it opened in 2017 which needs to be considered while modelling.

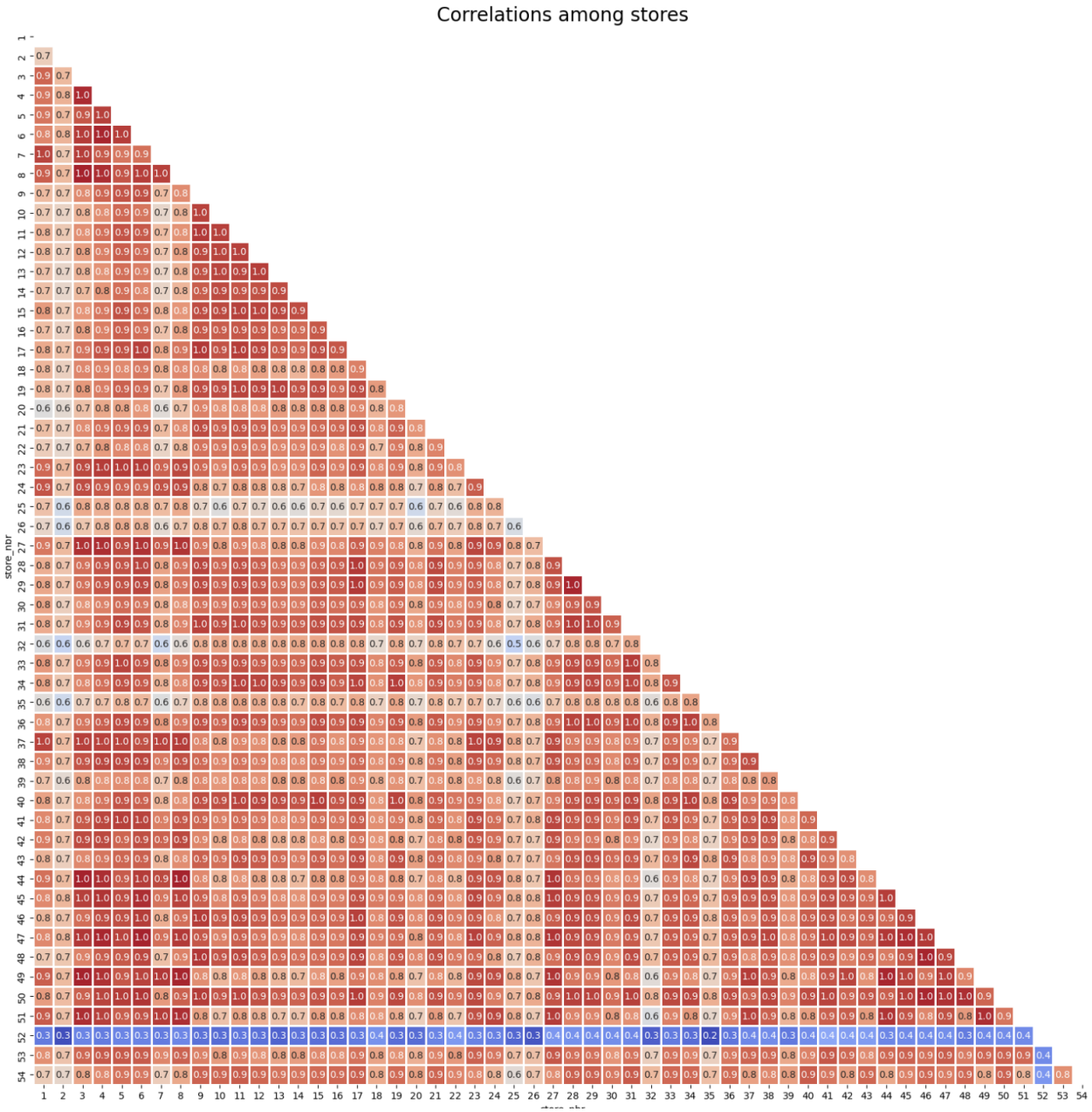


Figure 12. Correlation Matrix for sales of stores

- Recognizing the distinct sales patterns for different products and the aim being optimal inventory management, it seemed wiser to explore further store specific trends for each product individually.
- As shown in *Figure 13*, We created average monthly sales store clusters for each product with ordinal positioning i.e. 1 being the lowest selling segment. Around 7-12 clusters of stores were formed for each product. Plots for only 2 products have been included here for illustration.

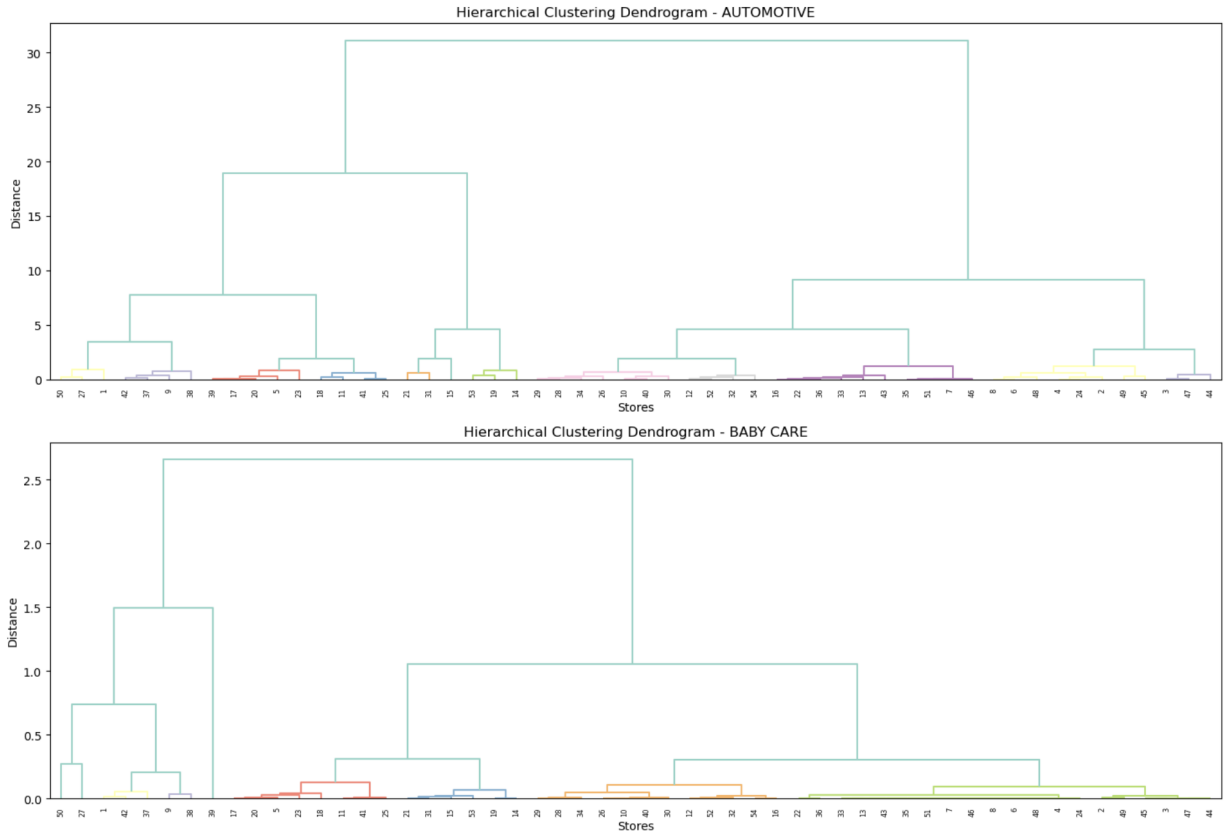


Figure 13. Clustering stores on the basis of monthly average for each product

2. **Products Unavailability in stores**
    - *Figure 14* shows a pictorial representation of which products are not sold in any particular stores.
    - Books are only sold in a limited number of stores while most of the stores sell all varieties of products.

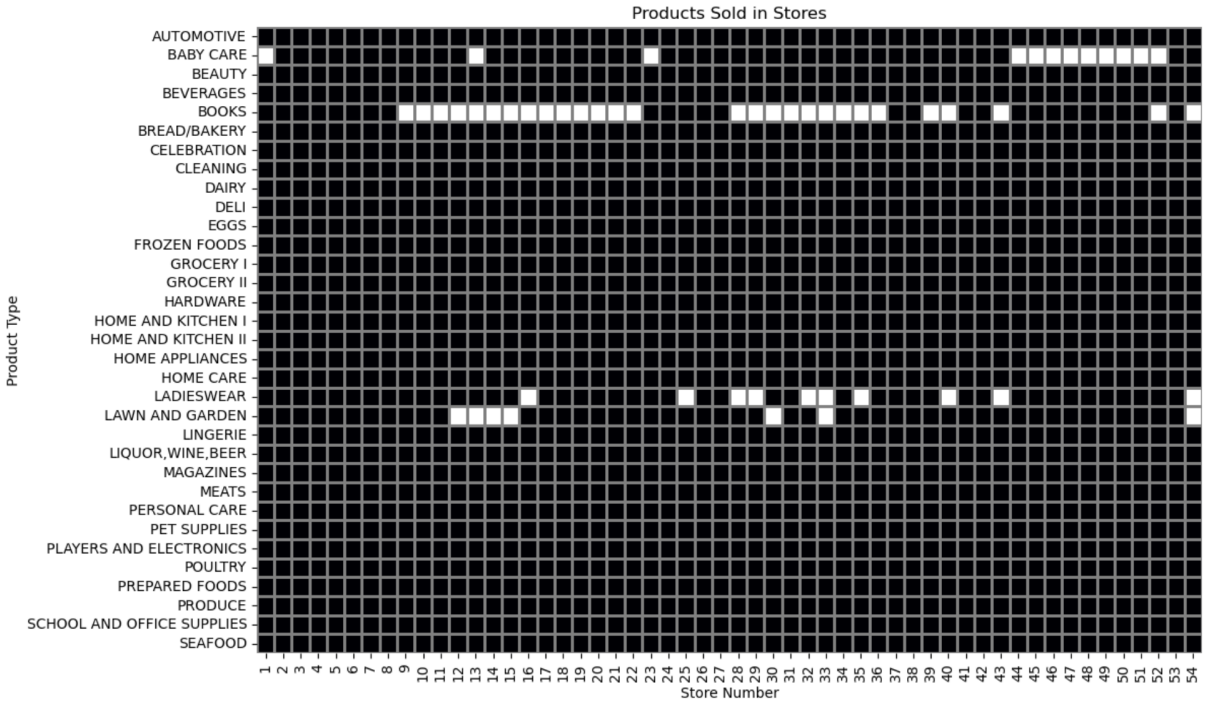


Figure 14. Heatmap illustrating product-store availability (white dots represent unavailability)

Outliers *(Figure 15)*

- Outliers were identified by finding unusual hike or dip in sales without any major offers on a particular day for each product, which can also be seen from the red dots in figure n. They were selected such that the sales on a particular day was 15 times the average sales of the particular product in that month and the sales jump or downfall was not attributed to promotions.
- Looking at School & Office Supplies, it had unusual highs and lows in its sales but minimal outliers due to the fact that they had promotional effect for its unusual sales.
- On the other hand, Home and Kitchen showed a lot of random behaviour which if not handled well, will make it difficult to predict its sales accurately.

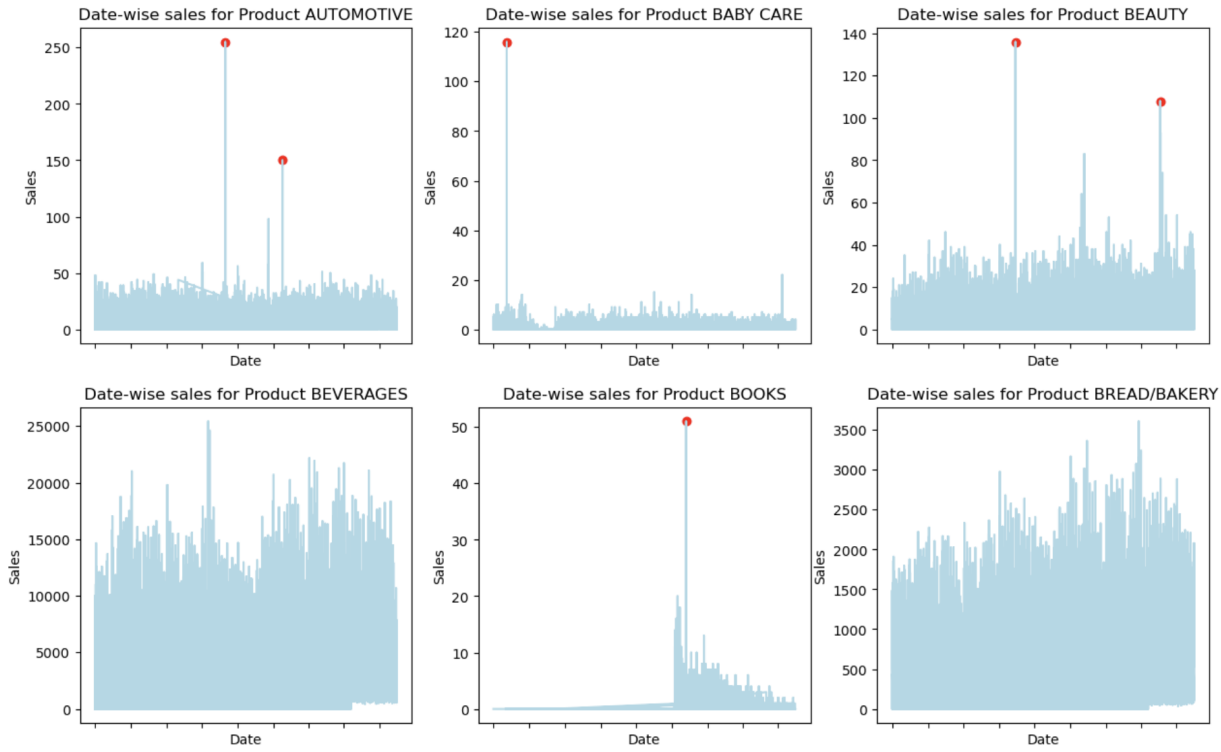


Figure 15. Highlighting sales outlier values (red points) for each product

To summarise, the comprehensive Exploratory Data Analysis conducted has provided key insights into the complexities of the intricate dynamics of this dataset. In light of these findings, we have laid the groundwork for subsequent modelling with the key takeaways being:

1. Dataset Trimming to contain data post July 2015 was strategically implemented to account for stable data with minimal random fluctuations for subsequent modelling.
2. Minimal similarity was observed between sales trajectories for different products suggesting to create different models for each product for better predictions.
3. Weekends showed a higher sales pattern in comparison to weekdays with a definite repetitive trend for each dow for sales as well as special offers.
4. High sales during the start of months which was probably due to salary releases and customers overall spending patterns.
5. Major hike in sales was observed in festive seasons around May and December in comparison to other months.
6. Many stores (store 52, opened in 2017) started functioning at a later date compared to other product and store combinations which needs to be accounted for, while modelling.
7. 5 products contributed to over 80% of the total sales and these products are daily usage products.
8. Special Offers are only effective for certain products and therefore any features related to special offer should only be used for selective products for which sales and offers are correlated.

9. Outliers tend to deviate from actual trends and therefore need to be removed from our dataset.

Basic Feature Engineering

To enhance the robustness and applicability of our future modelling efforts, we trimmed the dataset and refined the dataset to include the following possible additional features (as shown in *Table 3*) for the model feature set, some of which including dow, month and year were created during EDA. While the remaining were to be created during modelling as they could only be extracted from training dataset in order to avoid data leakage.

1. Dow: day of week
2. Month
3. Year
4. Start_date : date at which a product was launched at a store
5. Relative_monthly_sales_cluster: Cluster number in which the monthly average sales lies for that month (created ordinally, 1 being the lowest valued)
6. Relative_daily_sales_cluster: Cluster number in which the daily average sales lies for that day of month (created ordinally, 1 being the lowest valued)
7. Relative_monthly_offer_cluster: Cluster number in which the monthly average offer lies for that month (created ordinally, 1 being the lowest valued)
8. Relative_store_cluster: Cluster number in which the average sales of that store lies (created ordinally, 1 being the lowest valued)

	id	date	store_nbr	product_type	sales	special_offer	year	month	day	dow	is_weekend	start_date
1619838	1619838	2015-07-01	1	AUTOMOTIVE	5.0	0	2015	7	1	2	0	2013-01-02
1619839	1619839	2015-07-01	1	BABY CARE	0.0	0	2015	7	1	2	0	NaT
1619840	1619840	2015-07-01	1	BEAUTY	5.0	1	2015	7	1	2	0	2013-01-02
1619841	1619841	2015-07-01	1	BEVERAGES	2638.0	2	2015	7	1	2	0	2013-01-02
1619842	1619842	2015-07-01	1	BOOKS	0.0	0	2015	7	1	2	0	2016-10-13

Table 3. First 5 rows of the dataset after EDA & basic feature engineering

Modelling

In this section, we delve into the detailed explanation of the comprehensive modelling approach implemented to predict sales for the 16 days date range from 01/08/2017 to 16/08/2017. The framework involves execution of supervised machine learning algorithms including Random Forest, X Gradient Boost, and LightGBM to find the best working model for sales forecasting using cross validation and hyperparameter tuning. The structured pipeline includes data preprocessing to address the key issues identified during EDA, followed by feature engineering, hyperparameter tuning, and ultimately, the application of the selected model. Overall, 33 models were created for each of the unique product types to predict sales.

Note: Even though we deployed different models for each product, cross validation and hyperparameter tuning was only done on most critical products (highest selling products) to avoid the complexity of implementing different machine learning algorithms for test dataset for each product.

The key procedural steps undertaken during this phase include:

Data Preprocessing

The first step includes conditioning the dataset well enough for the model to identify trends and predict accurately. For this, a class 'ModelPreprocessing' was created, encompassing a suite of methods for dropping outliers, feature engineering to identify clusters and assessing inactive products. No encoding was required since we did not use any categorical columns in our feature set. Additionally, all numerical columns represented ordinal values. Regarding scaling all 3 employed models can work without scaling, for that reason scaling was also skipped. The functions in the class were namely:

1. drop_outliers: to drop outliers in the dataset
2. get_freq_group_cluster : to get relative daily sales cluster number & relative monthly sales cluster number
3. get_store_cluster: to get store cluster number
4. is_special_offer_effective: to assess whether the product is promotion sensitive. If yes, relative monthly offer cluster number
5. is_product_active: to assess whether the product has been active (available for sale) recently or not
6. set_lag_features: to include a lag feature representing the moving average of the DOW's average from the previous year

- Clustering includes techniques such as KMeans Clustering and Hierarchical Clustering to form clusters for sales and stores for each product.
- To avoid deviations in predicted sales, we have excluded the recent inactive products from the test dataset and we manually append their predicted sales as 0 with the model's predicted values.

Custom Grid Search for Hyperparameter Tuning

The heart of our modelling strategy lies in the custom class 'CustomGridSearch' created for applying grid search cross validation to orchestrate hyperparameter tuning. The need for a customised grid search method arose from the fact that our dataset was based on time and the sequence of time series was of utmost importance. Traditional validation sets are random subsets of train test splits which would not be possible to consider here. The dataset is firstly preprocessed and then the following methods are applied:

1. **Train-Test Split**

The number of folds and prediction days are passed to the method and accordingly sets of training and validation sets are created with prediction_days parameter as the number of days in validation sets and remaining data as the train set.

For illustration, to create 2 validation sets with 16 days prediction window:

The date ranges of validation set and train set would be:

Set 1:
train set: 01/07/2015 to 29/06/2017
validation set : 30/06/2017-15/07/2017

Set 2:
train set: 01/07/2015 to 14/07/2017
validation set : 15/07/2017-31/07/2017

2. GridSearchCV()

For selecting the best hyperparameters and cross validation, GridSearchCV from scikit-learn is deployed. The scoring method used for this was neg_root_mean_squared_error and finally the best hyperparameters and score obtained for each model was as shown in table .

3. Model Selection

A comparative graph for different models indicating the best scores for all 3 regressors can be seen from Figure 16. Detailed values can be referred from table 4. The lowest score indicating least deviation (nearest to zero) and best params were found from XGBoost Regressor with the same set of params for 4 out of 5 products_types passed for cross validation.

Best Params:

model	best_params	
XGBoost	{'XGB__max_depth': 10, 'XGB__n_estimators': 100}	4

product	model	best_score	best_params
BEVERAGES	RandomForest	-379.945847	{'RF__max_depth': 10, 'RF__n_estimators': 10}
BEVERAGES	XGBoost	-60.765024	{'XGB__max_depth': 10, 'XGB__n_estimators': 100}
BEVERAGES	LightGBM	-340.247965	{'LGB__max_depth': 5, 'LGB__n_estimators': 100}
CLEANING	RandomForest	-262.544810	{'RF__max_depth': 10, 'RF__n_estimators': 10}
CLEANING	XGBoost	-38.880899	{'XGB__max_depth': 10, 'XGB__n_estimators': 100}
CLEANING	LightGBM	-267.913110	{'LGB__max_depth': 5, 'LGB__n_estimators': 100}
DAIRY	RandomForest	-104.502614	{'RF__max_depth': 10, 'RF__n_estimators': 10}
DAIRY	XGBoost	-14.985900	{'XGB__max_depth': 10, 'XGB__n_estimators': 100}
DAIRY	LightGBM	-105.765573	{'LGB__max_depth': 5, 'LGB__n_estimators': 100}
GROCERY I	RandomForest	-820.853197	{'RF__max_depth': 10, 'RF__n_estimators': 10}
GROCERY I	XGBoost	-109.588523	{'XGB__max_depth': 10, 'XGB__n_estimators': 100}
GROCERY I	LightGBM	-795.383109	{'LGB__max_depth': 5, 'LGB__n_estimators': 100}
PRODUCE	RandomForest	-0.438669	{'RF__max_depth': 10, 'RF__n_estimators': 10}
PRODUCE	XGBoost	-6.321448	{'XGB__max_depth': 10, 'XGB__n_estimators': 100}
PRODUCE	LightGBM	-110.948429	{'LGB__max_depth': 5, 'LGB__n_estimators': 100}

Table 4. Best scores and params from Cross Validation for all 3 models for most selling products

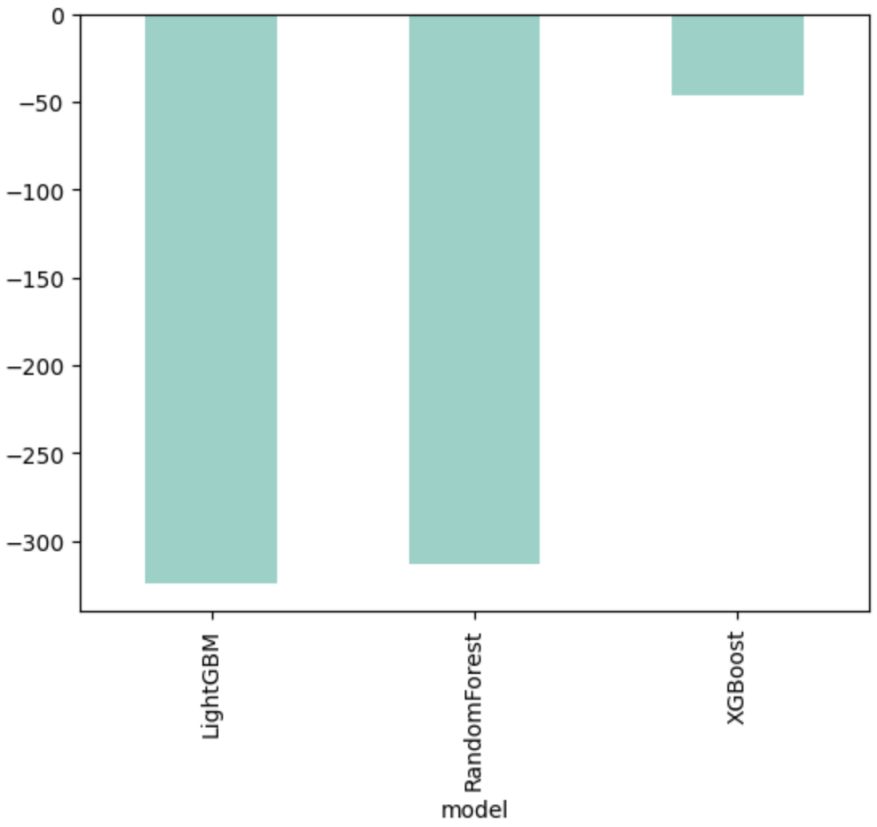


Figure 16. Model Score Comparisons

Model Application

Based on the score of cross validation, we selected the XGBoostRegressor since the neg_RMSE was the lowest (-46) out of all other models. A different model was employed for each of the products and individual predictions were made for them.

The Modelling for prediction on final test set is structured into several stages:

1. **Data Splitting**
   The dataset is divided into two subsets: a training set, comprising data until July 2017, and a test set, consisting of data for the next 16 days from 01/08/2017 to 16/07/2017. The pipeline has been crafted carefully to avoid data leakage and overfitting.

2. **Data Preprocessing**
   Similar to the cross validation pipeline, data is preprocessed to drop outliers, identify inactive products on store level and create clustering columns.

3. **Feature Selection**
   Apart from special offer clusters, all transformed and created features are passed to the model. The Features used for training the model include dow, year, month, lag_yoy_sales, relative_monthly_sales_cluster, relative_daily_sales_cluster, relative_store_cluster. Depending upon special offer effectiveness for a product, the column relative_monthly_offer_cluster is used as a feature.

4. **Model Training**
   An XGBoostRegressor model is trained on the preprocessed training set and fine-tuned with the pre-defined set of hyperparameters obtained from Grid Search cross validation to optimise predictive performance.

5. **Model Prediction**
   The trained XGBoostRegressor is applied to the preprocessed test set for sales prediction. The predictions are evaluated against the actual sales data, providing insights into the model's accuracy.

Performance Evaluation: Unveiling Insights into Sales Predictions

After putting our model to test, we now evaluate how well our model performed. In this comprehensive analysis, we explore various metrics and visualisations to dissect our model performance through multiple lenses. The Model takes between 1-2 minutes to run overall for all products with an average 1.2 seconds spent for each product’s prediction.

1. **Sales Comparison on product level**

   A bar chart is plotted for the products constituting majority of sales and remaining product types clubbed as others to illustrate a side by side comparison of cumulative actual sales vs predicted sales for the 16 day period. This visual (*Figure 17* and *Table 5*)allows us to identify patterns and discrepancies at a glance and we can see that the overall predictions are very close to the actual sales demonstrating a commendable performance.

product_type	avg_sales	avg_predicted_sales	mae	rmse
BEVERAGES	3469.02	3576.01	751.68	1084.93
CLEANING	1204.76	1301.95	314.43	527.65
DAIRY	832.77	884.40	118.71	182.63
GROCERY I	4580.49	4569.45	677.03	1027.26
PRODUCE	2290.86	2254.22	313.27	476.31
OTHERS	107.01	107.38	24.57	44.98

Table 5. Numerical Values for Sales v/s Predicted Sales

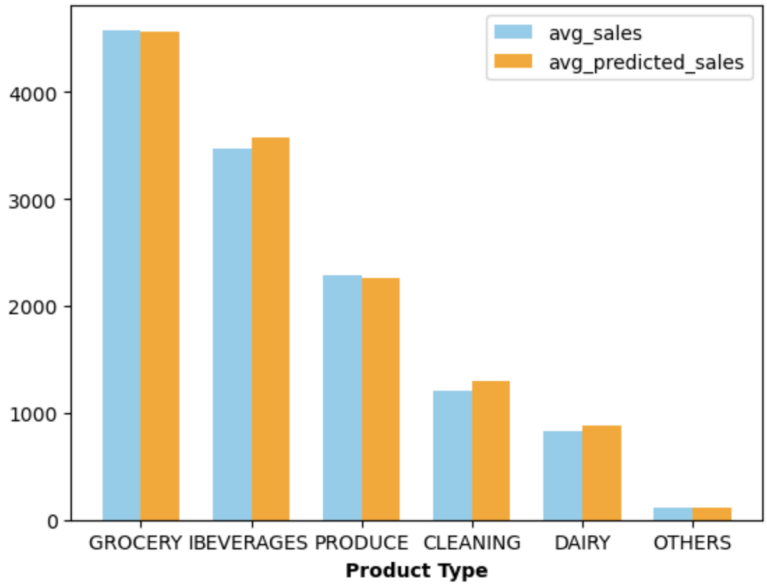
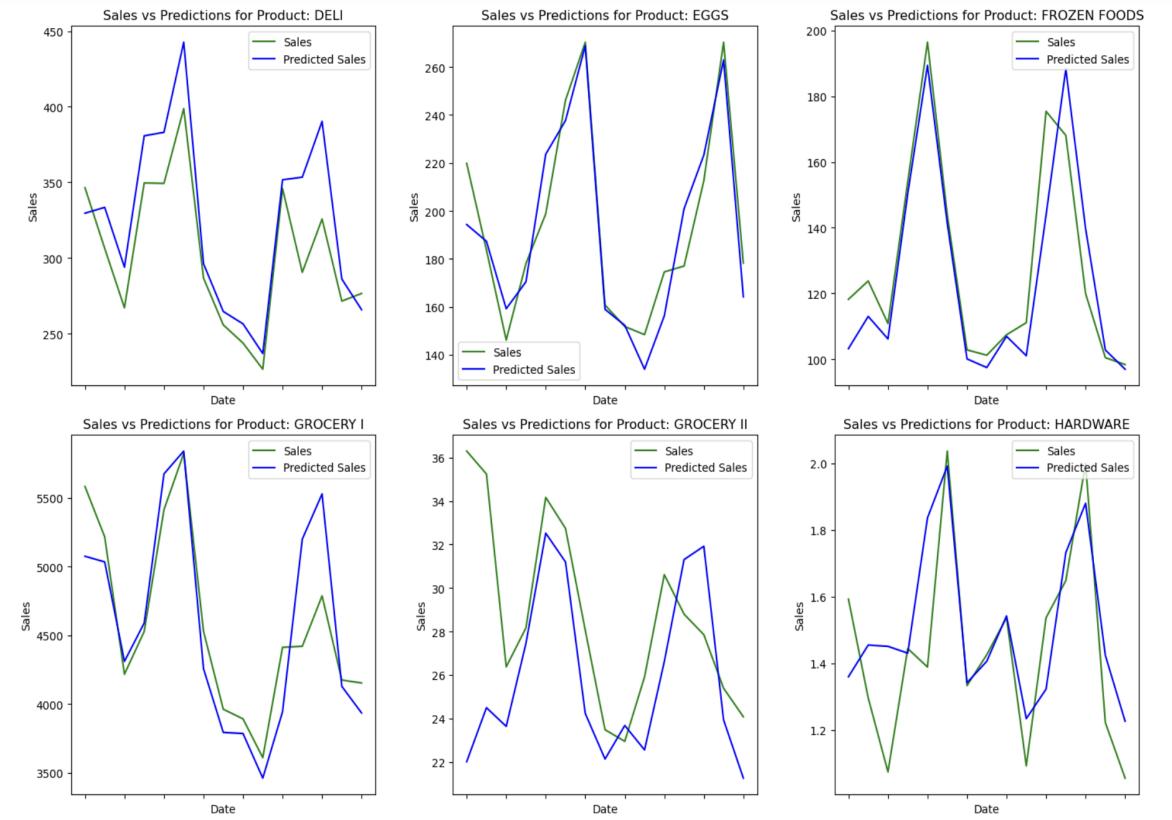


Figure 17. Sales v/s Predicted Sales comparison

2. **Line Plot analysis for actual vs predicted sales**

   To evaluate the model on a granular level, we plot a line chart to visualise trends and deviations between actual sales and predicted sales. Graphs for a few products are shown below for illustration in *Figure 18*. It can be observed from the plots that the model has identified sales patterns for most of the products and has worked really well.

However, the challenge specifically lies in sales forecasting for School and Office Supplies and Books with significant deviation from actual sales. This is attributed to sparse data and erratic sales patterns as observed during EDA.



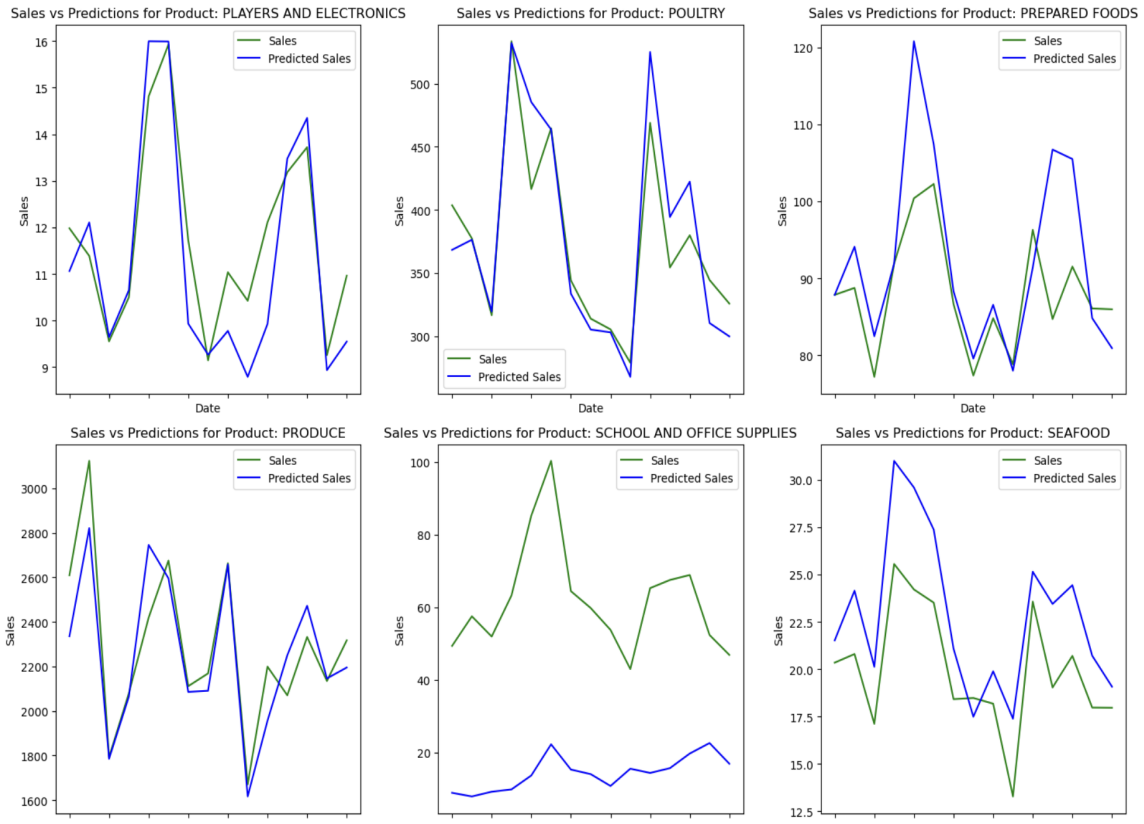


Figure 18. Date level actual vs predicted sales trends

3. **<u>Day-wise Percentage Deviation for all products</u>**

   Figure 19 shows day-wise percentage deviation of predicted sales from actual sales for the most selling products providing a quantitative measure of the model's accuracy. Of these, the model predicts well within the deviation of 20% which is an acceptable industry standard deviation for 4 products. Notably, Dairy products exhibit a higher deviation which is not a good sign and might need some more re tuning and in-depth analysis.

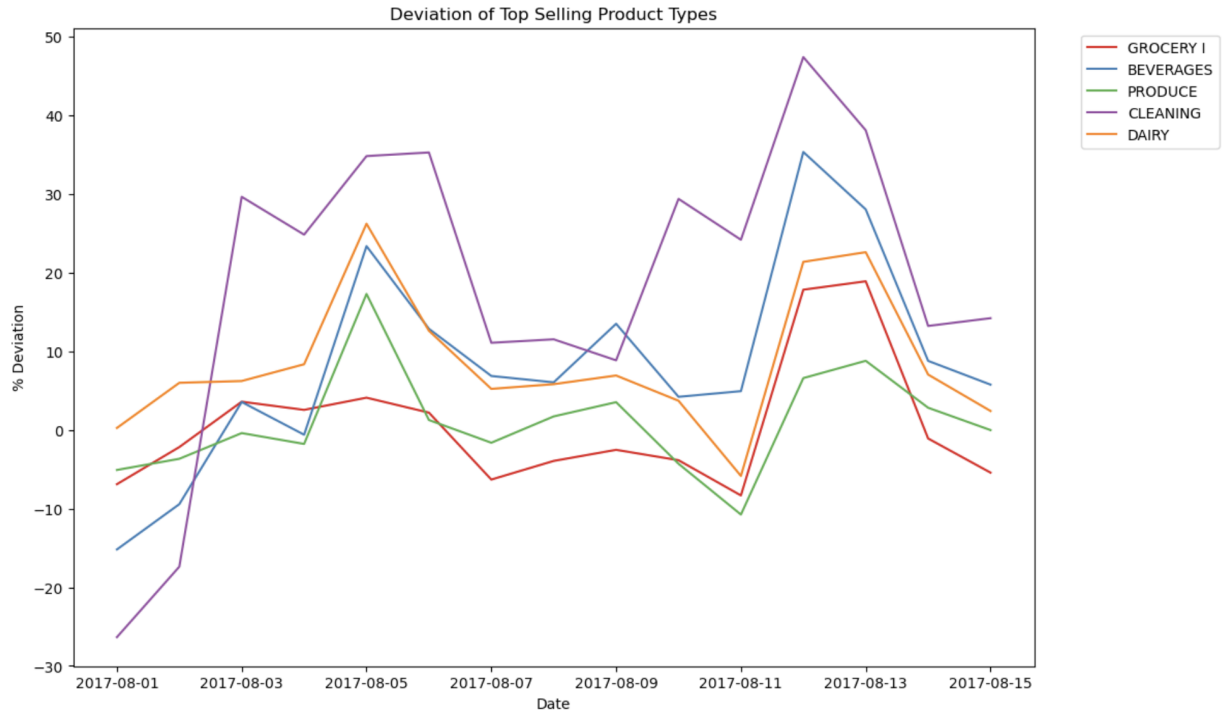


Figure 19. Date level actual vs predicted sales deviation percentages for most selling products

4. **<u>Shelf Life Sensitivity & Predictions</u>**

   A crucial aspect of maintaining stocks is understanding the shelf life of products. The model's performance is also scrutinised by assessing the predictions for low shelf life products. Based on domain knowledge, a list of products as shown in *Figure 20* was selected as sensitive products with low shelf life. The findings indicated a thumbs up for the model's working by aligning with the deviation under 20% for most of the products. However, Seafood emerged as an outlier, indicating a need for specialised attention to enhance predictions for this product category.

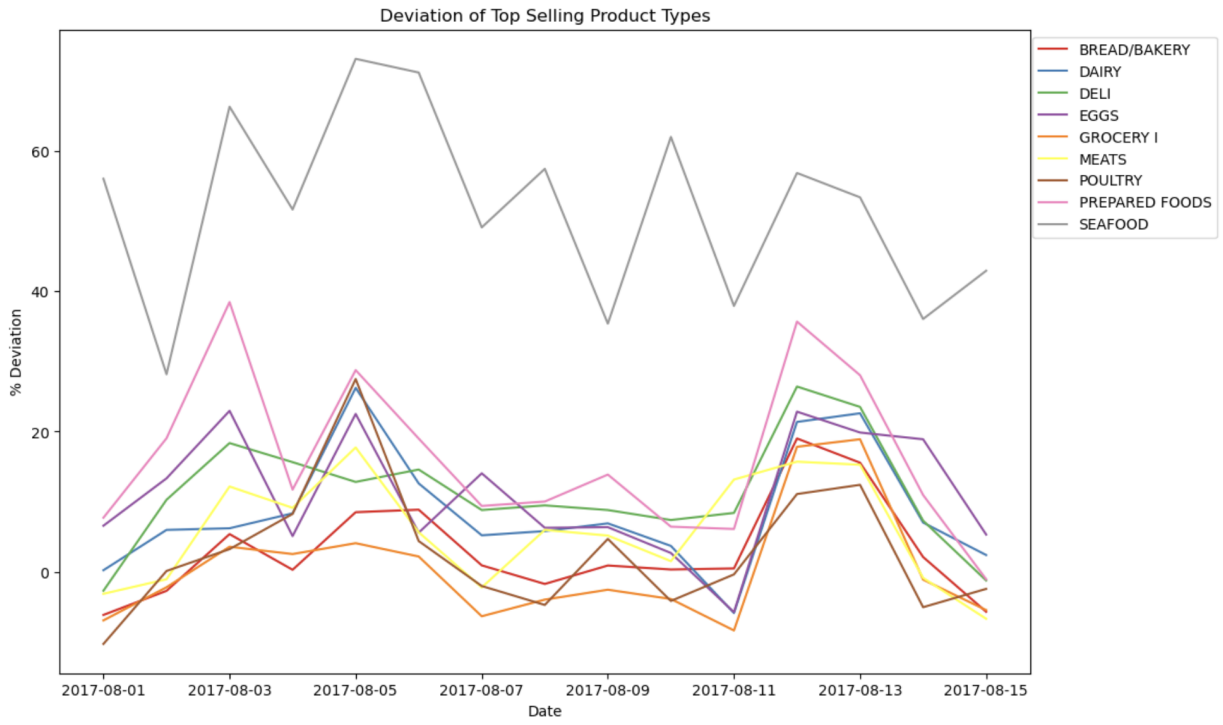


Figure 20. Date level actual vs predicted sales deviation percentages for low shelf life products

5. **<u>Day-wise Out of Stock Percentage Predictions for each product</u>**
   Preventing out of stock scenarios is paramount to prevent sales loss and customer dissatisfaction in the retail business. The *Table 6* shows Out of Stock (OOS) percentages for all products across all days. The data shows our model tends to predict sales lesser than the actual sales most of the times which needs to be taken care of. However, there is minimal wastage based on the predictions since predicted values are generally lower than the actual sales.

oos_flag

date product_type	2017-08-01	2017-08-02	2017-08-03	2017-08-04	2017-08-05	2017-08-06	2017-08-07	2017-08-08	2017-08-09	2017-08-10	2017-08-11	2017-08-12	2017-08-13	2017-08-14	2017-08-15
AUTOMOTIVE	50.00	48.15	40.74	59.26	33.33	61.11	59.26	53.70	46.30	40.74	64.81	31.48	46.30	37.04	62.96
BABY CARE	14.81	14.81	7.41	14.81	9.26	14.81	18.52	16.67	14.81	9.26	22.22	0.00	12.96	9.26	12.96
BEAUTY	48.15	38.89	44.44	48.15	62.96	53.70	61.11	64.81	59.26	57.41	61.11	51.85	48.15	51.85	61.11
BEVERAGES	79.63	64.81	46.30	44.44	24.07	48.15	37.04	37.04	37.04	44.44	42.59	14.81	27.78	31.48	31.48
BOOKS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.85	0.00	0.00	0.00	0.00	0.00	0.00
BREAD/BAKERY	57.41	61.11	37.04	51.85	35.19	46.30	50.00	66.67	50.00	46.30	59.26	22.22	38.89	40.74	51.85
CELEBRATION	57.41	33.33	35.19	42.59	40.74	40.74	35.19	37.04	33.33	62.96	42.59	24.07	44.44	57.41	37.04
CLEANING	94.44	83.33	40.74	24.07	14.81	16.67	29.63	35.19	35.19	12.96	20.37	11.11	24.07	33.33	40.74
DAIRY	59.26	35.19	40.74	33.33	11.11	37.04	37.04	38.89	29.63	35.19	68.52	20.37	29.63	29.63	46.30
DELI	57.41	33.33	25.93	31.48	31.48	31.48	35.19	25.93	35.19	46.30	42.59	7.41	25.93	37.04	51.85
EGGS	53.70	48.15	31.48	62.96	31.48	53.70	37.04	44.44	48.15	57.41	70.37	22.22	50.00	42.59	61.11
FROZEN FOODS	70.37	64.81	51.85	50.00	59.26	51.85	50.00	51.85	50.00	59.26	75.93	31.48	25.93	46.30	53.70
GROCERY I	64.81	57.41	42.59	40.74	40.74	51.85	70.37	50.00	66.67	53.70	75.93	29.63	25.93	44.44	61.11
GROCERY II	79.63	68.52	57.41	59.26	46.30	57.41	64.81	61.11	38.89	55.56	51.85	42.59	40.74	46.30	51.85
HARDWARE	48.15	33.33	25.93	40.74	33.33	42.59	42.59	38.89	38.89	38.89	42.59	40.74	37.04	35.19	33.33
HOME AND KITCHEN I	59.26	37.04	35.19	33.33	42.59	44.44	46.30	37.04	46.30	55.56	68.52	35.19	37.04	59.26	57.41
HOME AND KITCHEN II	59.26	51.85	38.89	46.30	40.74	50.00	42.59	51.85	42.59	38.89	55.56	27.78	37.04	57.41	53.70
HOME APPLIANCES	18.52	16.67	12.96	11.11	20.37	16.67	16.67	14.81	5.56	5.56	16.67	7.41	1.85	5.56	5.56
HOME CARE	77.78	48.15	33.33	11.11	9.26	25.93	29.63	25.93	20.37	29.63	38.89	9.26	14.81	20.37	31.48
LADIESWEAR	38.89	33.33	33.33	22.22	16.67	22.22	25.93	37.04	29.63	14.81	33.33	20.37	16.67	31.48	37.04
LAWN AND GARDEN	37.04	35.19	24.07	24.07	35.19	29.63	25.93	33.33	27.78	37.04	38.89	27.78	25.93	31.48	33.33
LINGERIE	38.89	35.19	42.59	37.04	42.59	37.04	31.48	40.74	37.04	51.85	57.41	38.89	35.19	44.44	44.44
LIQUOR,WINE,BEER	31.48	62.96	48.15	29.63	85.19	70.37	9.26	25.93	48.15	64.81	74.07	64.81	51.85	14.81	38.89
MAGAZINES	61.11	46.30	31.48	40.74	57.41	61.11	57.41	53.70	46.30	51.85	46.30	40.74	50.00	72.22	66.67
MEATS	62.96	62.96	44.44	50.00	24.07	55.56	66.67	44.44	50.00	48.15	46.30	37.04	35.19	53.70	74.07
PERSONAL CARE	70.37	57.41	29.63	50.00	35.19	40.74	61.11	50.00	51.85	38.89	75.93	22.22	18.52	40.74	57.41
PET SUPPLIES	44.44	38.89	42.59	35.19	44.44	44.44	44.44	44.44	44.44	33.33	61.11	27.78	31.48	37.04	40.74
PLAYERS AND ELECTRONICS	53.70	31.48	40.74	48.15	37.04	44.44	55.56	37.04	57.41	55.56	51.85	33.33	50.00	48.15	51.85
POULTRY	72.22	59.26	57.41	42.59	14.81	50.00	64.81	61.11	50.00	62.96	66.67	27.78	35.19	61.11	61.11
PREPARED FOODS	44.44	35.19	29.63	42.59	24.07	37.04	44.44	40.74	42.59	50.00	55.56	29.63	38.89	46.30	50.00
PRODUCE	68.52	62.96	46.30	48.15	16.67	64.81	59.26	59.26	38.89	62.96	70.37	35.19	38.89	44.44	64.81
SCHOOL AND OFFICE SUPPLIES	40.74	37.04	38.89	38.89	29.63	22.22	27.78	42.59	35.19	33.33	55.56	50.00	37.04	38.89	38.89
SEAFOOD	37.04	29.63	29.63	20.37	22.22	31.48	29.63	40.74	29.63	20.37	33.33	22.22	25.93	31.48	35.19

Table 6. Day level Out of Stock percentages for all products

6. Data Leakage & Overfit Examination

Clear distinction between training and test datasets ensuring that the training set precedes the test set as well as the validation sets chronologically has been maintained to avoid data leakage.

Conclusion: Towards Model Refinement and Optimization

In a nutshell, it can be concluded that the model has worked exceptionally well for some products while there still are areas of improvement for others. The model performance evaluation comprehensively underlines the areas of strengths and marks for improvement providing an aligned direction for specific products.

Roadmap for future model refinements

1. Dashboards could be created for continuous monitoring of the metrics evaluated during model performance evaluation to measure improvement over time.
2. In the critical facet of the retail industry, a subtle balance between OOS and wastage is of utmost importance, More so, it should be initially preferred to have a higher stock of inventory at all times to build a stronger customer base. To handle this, the inventory created could be a multiple of the predicted sales for a few months until the model starts to grasp the trends automatically.
3. Special attention to identify underlying trends to improve predictions for a few products such as School & Office Supplies, Seafood and books can be done.
4. Moving forward, a continuous feedback loop between model evaluation and refinement is recommended, fostering an iterative process that ensures the model evolves with changing data patterns and business dynamics.
5. Inter-store dispatching of products can also be done by creating a dashboard to manage the system with automation in order to efficiently utilise excess inventory.

Challenges Faced

One major challenge that still remains and incorporating which might improve the predictions are introduction of more lag features. We could not include recency based lag features such as values for a day before as the model was predicting values for n days in advance and for using the predicted values as lag values, we would have to predict one day at a time. Due to time constraints, we could not implement that completely and have removed that implementation from this report but we believe inclusion of recent lag features would significantly improve the sales prediction.

REFERENCES

1. Feature Engineering, https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10
2. K-Means clustering method, https://scikit-learn.org/stable/modules/clustering.html#k-means
3. Hierarchical Clustering, https://www.w3schools.com/python/python_ml_hierarchial_clustering.asp
4. X Gradient Boost Technique, https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/
5. Light GBM modelling technique, https://lightgbm.readthedocs.io/en/latest/Python-Intro.html