MANCHESTER
1824

BMAN60422

DATA ANALYTICS FOR BUSINESS DECISION MAKING

Date: 10/05/2024

Word Count: 3,926
(excluding title page, executive summary page and table of contents)

TABLE OF CONTENT

1. Executive Summary

This report delves into the intricacies of daily sales forecasting for a prominent German drugstore chain encompassing 1,115 stores. Leveraging historical data from 2013 to 2015, an in-depth analysis is conducted to discern influential factors shaping sales dynamics and propose actionable strategies for enhanced performance.

Through the analysis of historical sales data, several pivotal factors influencing sales within the drugstore chain are identified. These encompass promotional activities, competitive landscape, holiday trends, seasonal variations, and the geographical positioning of individual stores.

Moreover, the report explores segmenting stores into distinct clusters based on mean sales. This stratification facilitates the development of tailored sales forecasting models. Various machine learning models are scrutinized for their efficacy in daily sales forecasting, with XGBoost emerging as the most promising candidate.

Based on the insights garnered, recommendations are proposed to support the drugstore chain's sales forecasting capabilities and profitability. Firstly, implementing the XGBoost model across the entire drugstore chain is advocated. This data-driven approach would optimize inventory management, staffing decisions, and marketing strategies. Additionally, leveraging insights from store cluster analysis is advised to craft targeted marketing campaigns and refine resource allocation strategies, ensuring tailored approaches for different store clusters to maximize campaign effectiveness and return on investment. Furthermore, exploring the integration of additional data sources, such as product information or customer demographics, is recommended to enrich sales forecasting models and enhance accuracy and granularity.

This data-centric approach to sales forecasting equips the drugstore chain with a powerful tool for gaining a competitive edge and thriving in the marketplace.

2. Introduction

This report focuses on the task of predicting daily sales for a period of six weeks across 1,115 drug stores in Germany, using data sourced from a leading European drugstore chain like Boots in the UK. Given the diverse product categories offered, precise sales forecasting is paramount for retailers, as it significantly impacts staffing, customer satisfaction, and overall profitability. The project addresses the intricate challenges inherent in retail sales forecasting, considering factors such as promotional activities, competitive dynamics, holiday seasons, seasonal variations, and geographical store positioning. Historical sales data spanning from January 1st, 2013, to July 31st, 2015, forms the basis for analysis, supplemented by comprehensive store-related information.

The primary objective of this endeavour is to extract actionable insights from data analytics methodologies to facilitate informed decision-making processes within the retail sector. A pivotal initial phase in this data-driven project is Exploratory Data Analysis (EDA), involving exhaustive examination of provided datasets to discern intrinsic characteristics, detect anomalies, and unveil underlying correlations among variables. Leveraging data visualization techniques and statistical analyses, EDA facilitates the extraction of valuable insights regarding sales determinants, thus informing subsequent data pre-processing and feature engineering endeavours.

By leveraging historical data and advanced analytics, prediction models can provide insights into future sales trends. This allows retailers to anticipate demand fluctuations and adjust inventory levels, optimizing operations and cost efficiency.

3. Data Pre-processing

3.1 Handling of Missing Values

In the context of exploratory data analysis (EDA) and machine learning, missing values in a dataset can introduce inaccuracies and biases, affecting the reliability of our analytical insights and predictive models. Therefore, it's crucial to carefully identify and address these missing values. They can distort key statistical metrics like the mean, median, and standard deviation, and impact the integrity of data visualizations. Additionally, traditional machine learning algorithms struggle to handle missing data, highlighting the need for proactive strategies to manage them.

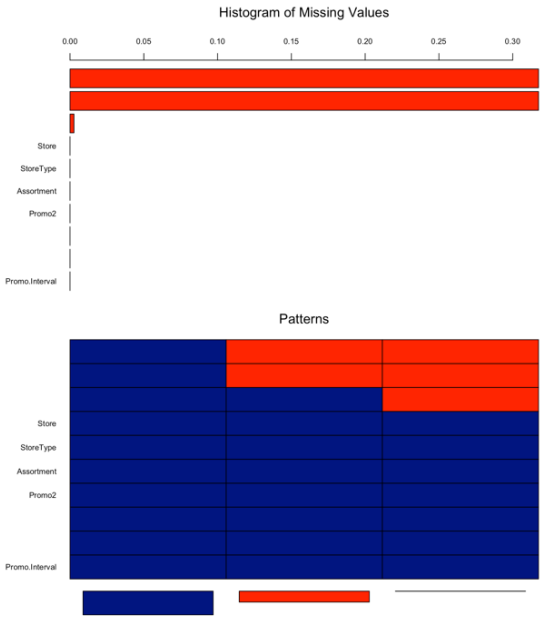


Figure 1: Histogram of Missing Values

In the provided dataset analysis, Missing values in several columns were identified: "Competition_Open", "Competition_Open_Sinceyear", and "Competition_Open_Sinceweek". To address this issue, we utilized the random forest imputation method to fill in the missing values. This approach ensures the completeness and accuracy of the dataset, facilitating robust analysis and modelling processes.

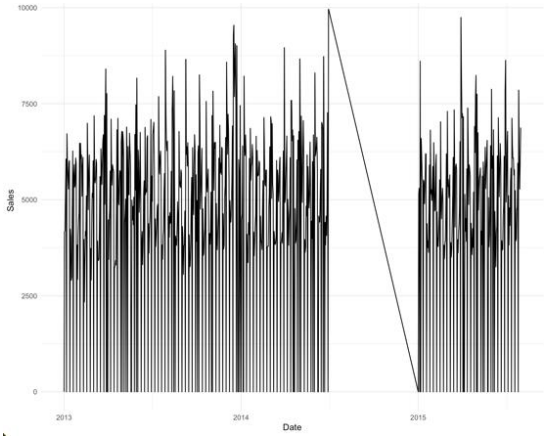


Figure 2: Missing Data (Store 181)

The dataset underwent analysis to verify the consistency of timestamps. This step was essential because the task involved calculating sales for specific dates. However, about 187 stores had missing timestamps, which could impede the overall process. To mitigate this issue, missing values were filled in using values from previous timestamps.

3.2 Changing the Date to Date -Time formatting

Converting the "Date" column to a date-time format is crucial for developing accurate machine learning models, especially for forecasting tasks. This step enables the models to understand relationships and extract patterns, improving their predictive capabilities.

3.3 Scaling and Encoding the Data

Scaling the data is crucial to prevent variables with larger values, such as competition distance, from dominating the predictive modelling process this was performed using the Standard Scalar package from Scikit learn in Python.

Additionally, encoding categorical variables is necessary to ensure the data remains usable. The columns "StateHoliday", "Assortment", ”StateHoliday”, “DayOfWeek” and “Promo.Interval” were encoded to facilitate predictive modelling. This function was implemented using the OneHotEncoding and the Label encoder functions available in Python programming. In subsequent sections of this report, references to these columns also denote all the encoded columns generated from them.

3.4 Outlier Detection

Identifying outliers in machine learning is paramount for ensuring the fairness of models and the precision of predictions. Outliers have a disruptive effect on statistical metrics, influence the behaviour of algorithms, and serve as indicators of potential data integrity concerns. Employing techniques like box plots, particularly in 'sales' data analysis, has led to the detection of 27,506 outliers. This approach aids in comprehending extreme data points, thereby enhancing decision-making processes, and bolstering the reliability of models.

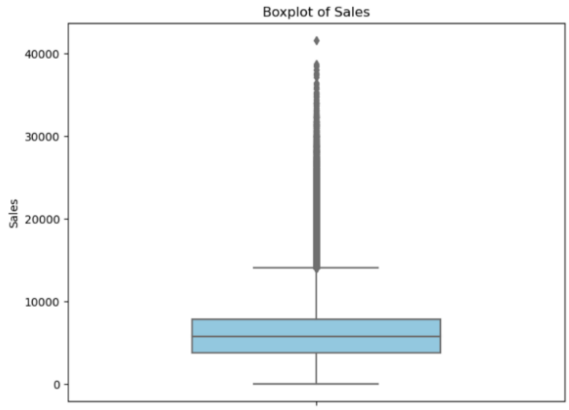


Figure 3: Boxplot of Sales

3.5 Splitting the Data into Train and Validation

Data splitting is a fundamental step in machine learning, involving the partitioning of available data into two subsets: training and testing. The training set is utilized to adjust the model's parameters, facilitating its learning process. Meanwhile, the testing set is employed to assess the model's performance on unseen data, thereby preventing overfitting and ensuring its ability to generalize to new instances. This approach is pivotal for evaluating a model's proficiency in handling real-world scenarios, as it enables unbiased assessment without the influence of training data. By allocating 70% of the data for testing and 30% for validation, the process ensures a balanced representation of the dataset across both sets. Metrics computed on the testing set, such as accuracy or error rates, provide valuable insights into the model's strengths and weaknesses, guiding further optimization efforts. Ultimately, data splitting fosters robust model evaluation, enhancing their generalization capabilities and reliability in real-world applications.

4. Descriptive Statistics

Based on the provided information, the dataset contains a substantial number of entries, with 1,048,575 observations. This extensive collection of sales data encompasses multiple stores and product types, offering a comprehensive perspective on retail operations. The dataset captures transactions beginning from January 1, 2013, and continues until August 1, 2015, providing a significant timeframe for analysis and insights into sales trends, patterns, and fluctuations over the specified period.

1. Sales Insights:
    - The 'Sales' column exhibits a wide range, with values ranging from 0.0 (potentially indicating days with no sales or store closures) to a maximum of 41,551.
    - The average sales amount to 5,767.83, accompanied by a high standard deviation, suggesting notable variability in sales figures across different stores and product categories.
    - The median sales amount to 5,740.0, which is lower than the mean, indicating a skewed distribution with several high-value sales transactions.

    - The store with the highest sales is Store 262, with a total sales amount of 19,516,842. Conversely, the store with the lowest sales is observed to be Store 208, with a total sales amount of 2,302,052.

2. Promotional Strategies and Sales:

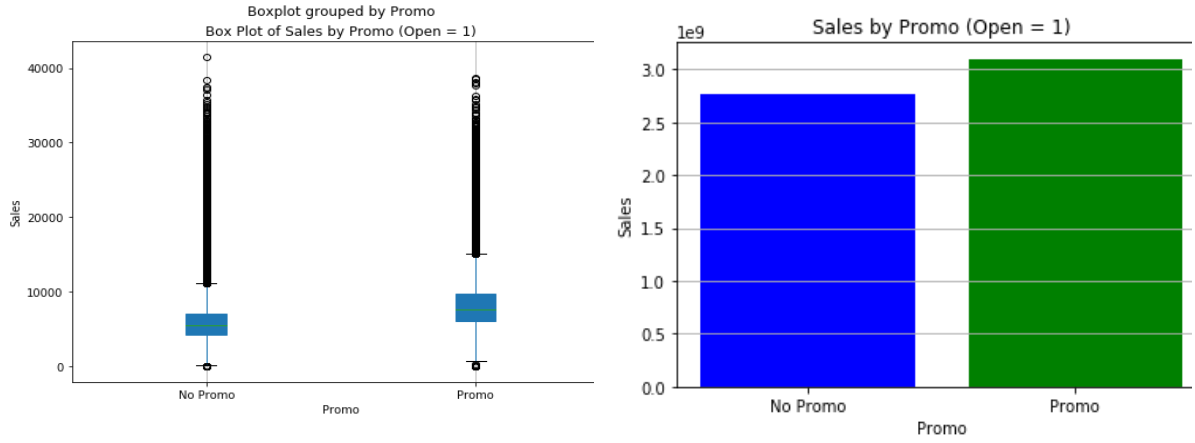


Figure 4: Sales by Promo (Open = 1) & Boxplot of Sales by Promo (Open = 1)

- The analysis confirms a direct link between promotions and sales, showing that stores with in-store promotions see higher average sales. This highlights the effectiveness of well-executed promotions in boosting customer engagement and driving purchasing decisions.

5. Data Analysis

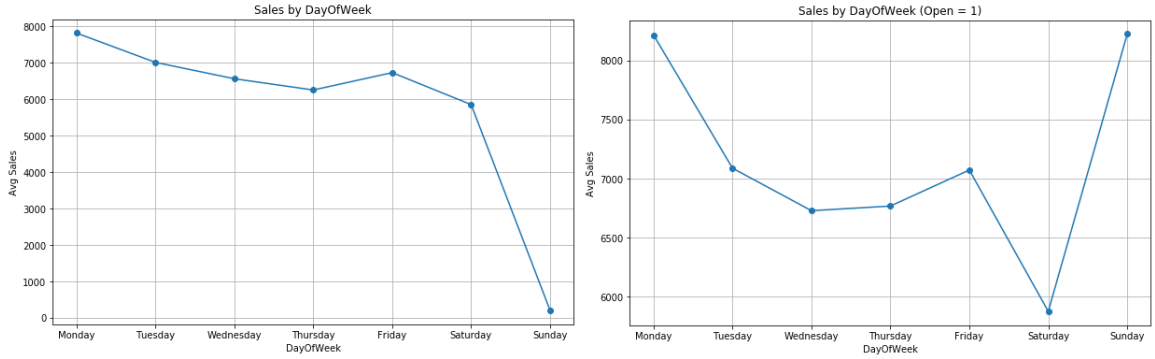


Figure 5: Sales by DayOfWeek (Open = 1) & Sales by DayOfWeek (All)

The analysis reveals a notable correlation between the day of the week and sales, with Sundays consistently exhibiting the lowest sales likely due to widespread store closures. Conversely, Mondays consistently outshine other weekdays, presenting a potential avenue for strategic adjustments in staffing or marketing initiatives to capitalize on this weekly peak. Interestingly, when analysing sales data for open stores, including Sundays, both Mondays and Sundays demonstrate a surge in sales, suggesting a shift in customer behaviour. This hints at some individuals opting for targeted shopping trips on traditionally non-peak days.

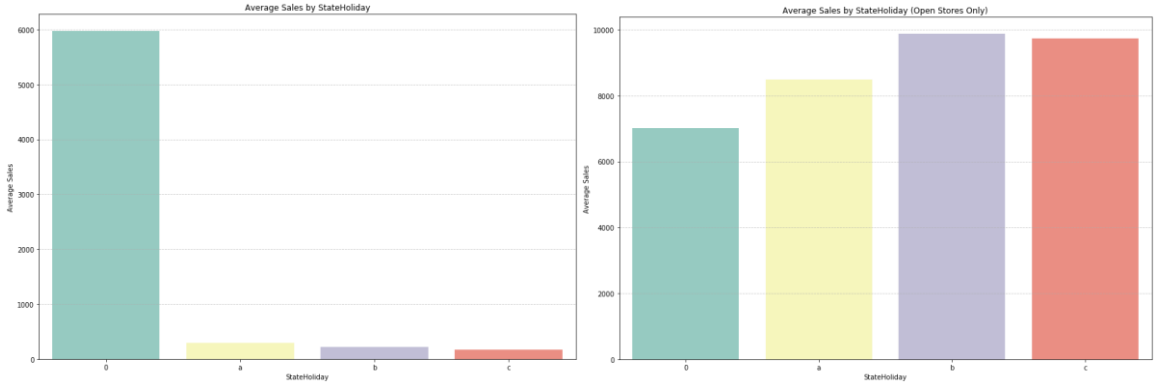


Figure 6: Average Sales by StateHoliday & Average Sales by StateHoliday (Open = 1)

Figure 7: Boxplot of Sales by ShcoolHoliday

Analysing sales about state holidays from Figure 6 provides valuable insights for optimizing store operations and promotional strategies during these periods. On average, stores without holidays tend to achieve higher sales, as they remain open more frequently. Interestingly, the

data indicates that sales peak on Easter and Christmas compared to public holidays for open stores. This understanding enables businesses to make informed decisions regarding staffing, inventory management and targeted marketing efforts during these critical sales seasons. Similarly, a similar trend is observed when analysing the impact of school holidays on sales, with sales remaining consistent during these periods.

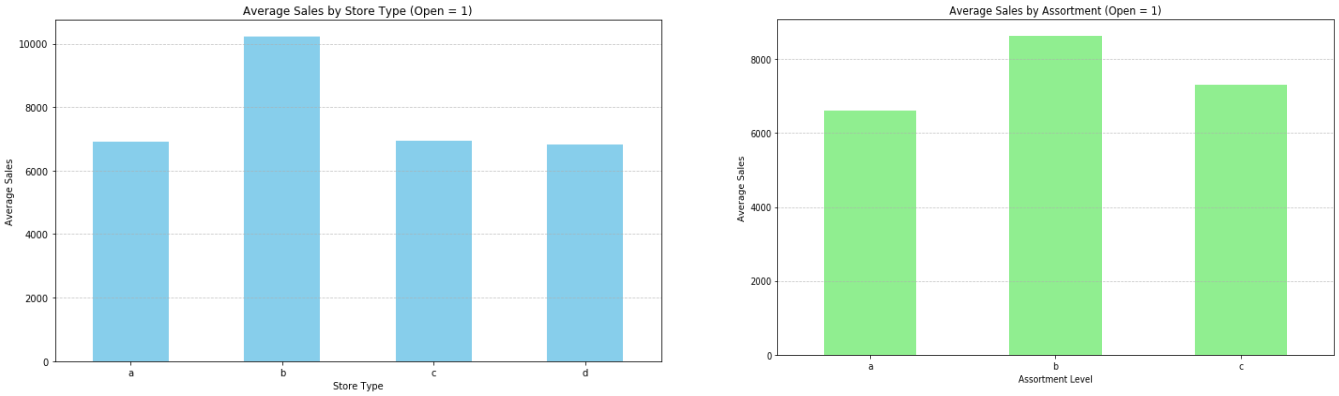


Figure 8: Average Sales by StoreType (Open = 1) & Average Sales by Assortment (Open = 1)

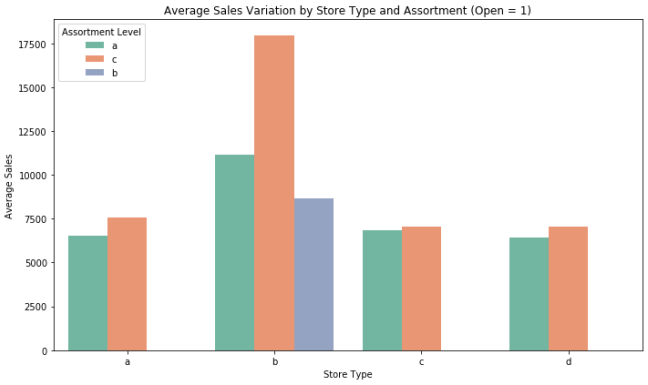


Figure 9: Average Sales by StoreType and Assortment (Open = 1)

In the concluding section of the report, a meticulous examination of the interplay between store type, assortment level, and sales unveils compelling insights. Figure 8 elucidates store type B as the preeminent performer during operational hours, and assortment level B as the frontrunner within its classification. However, the pinnacle finding materializes in Figure 9, where the amalgamation of StoreType B and Assortment C emerges as the unrivalled leader in overall sales. This comprehensive sales trend analysis underscores the criticality of dissecting these factors both individually and collectively to comprehend customer purchasing behaviour comprehensively. By discerning the impact of delivery fee dynamics, promotional endeavours, seasonal fluctuations, and intrinsic store attributes, organizations can orchestrate data-informed strategies to refine operations, amplify marketing endeavours, and attain enduring sales ascendancy.

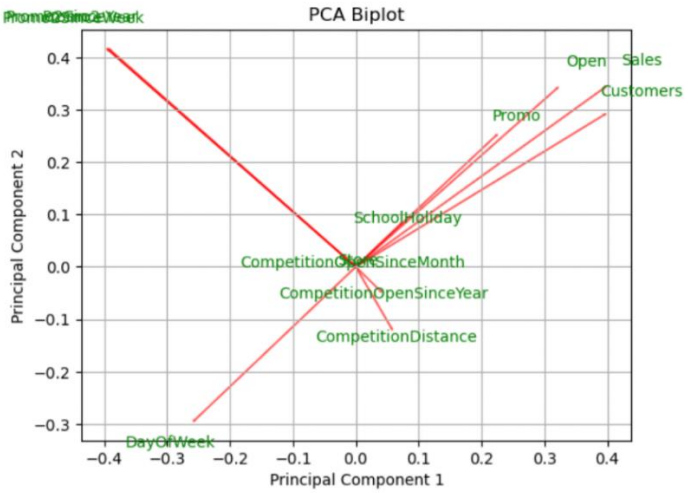


Figure 10: PCA

The PCA results unveil essential insights into the correlation structure among the variables under examination. Notably, DayOfWeek shows a negative correlation with Sales and Customers, contrasting with the high correlation observed between Sales and Customers. This suggests that the day of the week has a relatively weaker influence on the variance captured by the first two principal components compared to Sales and Customers. Furthermore, variables such as Store, State Holiday, Promo2, Promo2SinceWeek, and Promo2SinceYear exhibit a notable level of correlation. This implies a strong interrelation among these variables, signifying potential dependencies or shared influences within the dataset.

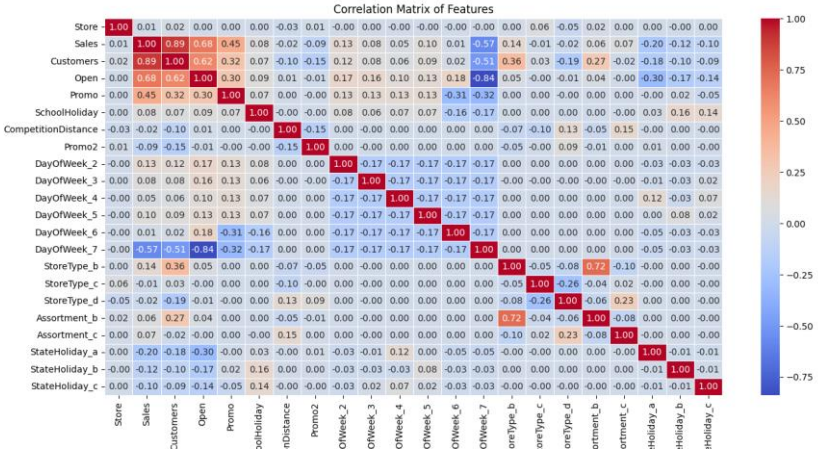


Figure 11: Correlation Matrix.

The correlation matrix provides more insights into the dataset from a pairwise perspective. It is observed that sales and customers have the strongest correlation, while open status has a considerable impact on them. Sunday has a strong negative correlation to open, which ultimately leads to lower sales and customers, which line up with the finding earlier that most stores close on Sunday. It is also found that public holiday has higher impact than Easter and Christmas which match the result from Figure 6. Besides the findings that support earlier analysis, two new insights are found based on the correlation matrix. There is high correlation between store type b and assortment b, and between promo2 and promo.interval starting January. From the finding, we conclude that the majority of store with assortment b is of type b, and the majority of store doing promo2 are renewing on a quarterly basis starting January every year.

5.1 Clustering of Stores

In the dataset provided, the number of individual stores (1115) posed a challenge for analysis and model-building. To address this issue, stores were categorized into distinct groups based on their mean sales, facilitating a more structured and efficient analysis process.

Clustering the stores based on mean sales allowed for targeted modelling and decision-making. Models tailored to store clusters could enhance prediction accuracy, while strategies developed for one cluster could be applied to similar-performing stores, optimizing resource allocation and marketing efforts.

The K-Means algorithm, an unsupervised learning technique, was utilized for clustering. Determining the optimal number of clusters was critical, achieved through the Within-Cluster Sum of Squares (WCSS) metric and the Elbow Method. Which indicated an 'elbow' point at four clusters, suggesting an appropriate representation based on mean sales patterns. The resulting scatter plot depicted a clear gradient in mean sales across clusters. Cluster 0 represented stores with the lowest mean sales, potentially smaller or struggling stores requiring targeted interventions. Cluster 1 exhibited moderate sales performance, while Cluster 2 showed a significant leap in sales, likely representing well-established stores in favourable locations. Cluster 3 comprised top-performing stores with the highest mean sales. This clustering analysis offers valuable insights into sales performance patterns across different store clusters, facilitating data-driven decision-making and targeted strategies to optimize store operations and enhance profitability.

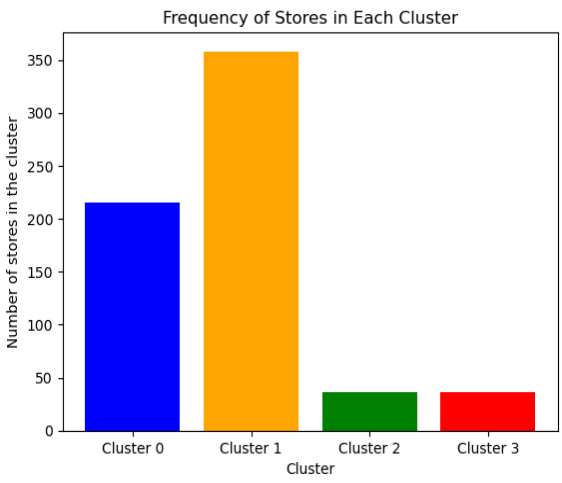


Figure 12: Stores in Each Cluster

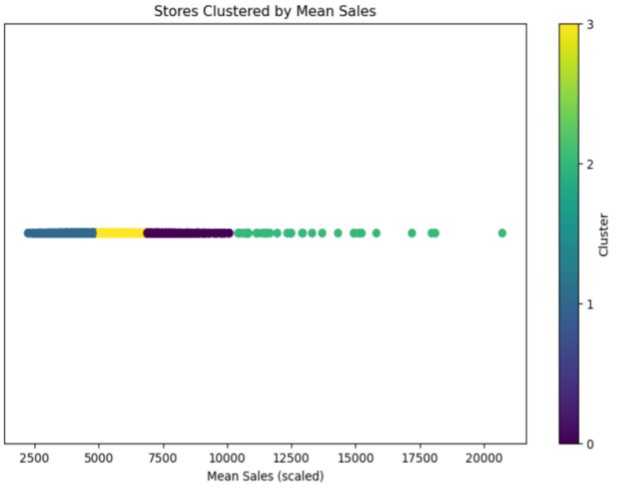


Figure 13: Clustering By K-Means

6. Predictive Modelling

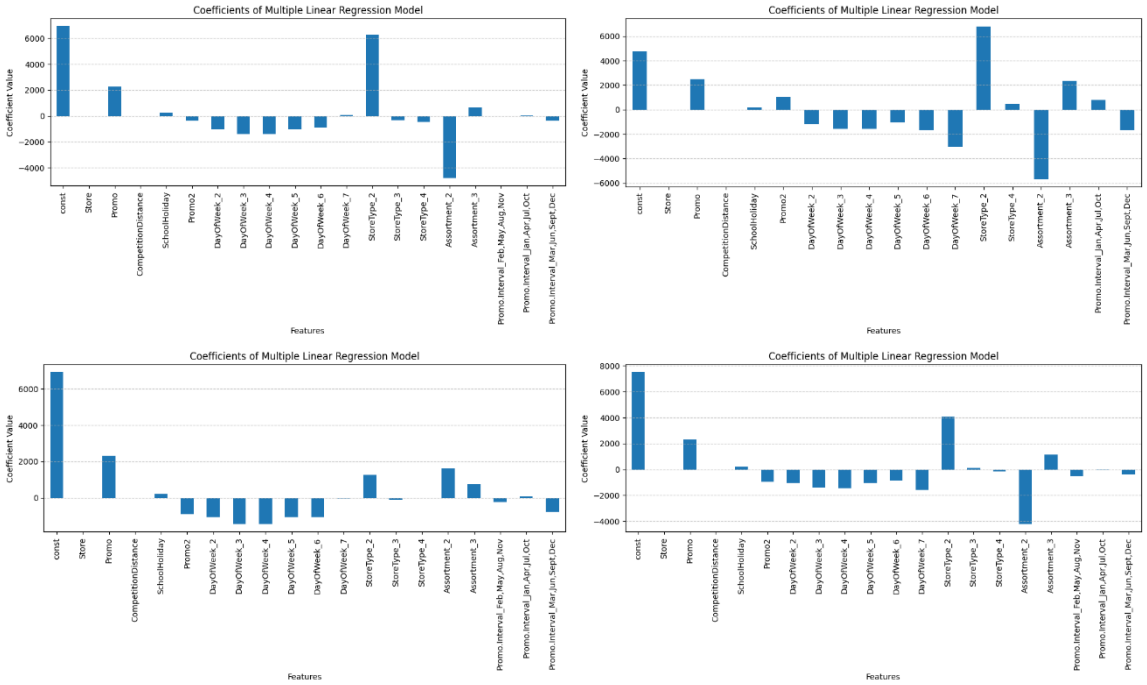
Each predictive model has its unique steps; however, a common step is maintained across all of them. To enhance the final accuracy of the prediction, as well as to accommodate the inability to deal with 0 in actual data from RMSPE, the records with open status as 0 are excluded from the training and prediction, and the sales on them are set to 0.

6.1 Multiple Linear Regression.

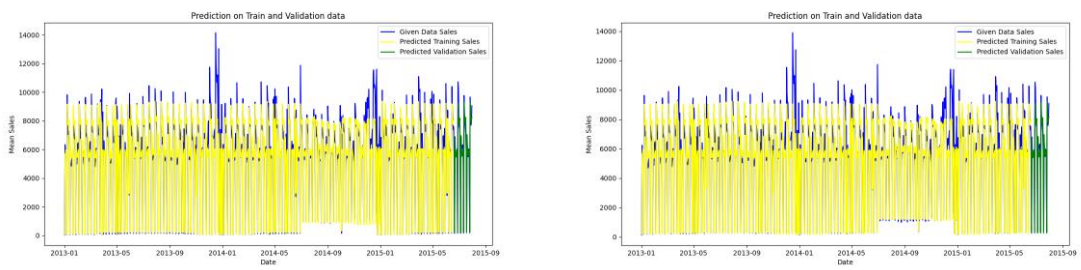
The multiple Linear Regression method was chosen to start with the simplest model. By

starting with multiple linear regression, we establish a baseline performance level that can be used to benchmark the performance of more advanced models. This helps in assessing whether the additional complexity of a more sophisticated model leads to significant improvements in predictive accuracy. The coefficients in multiple linear regression can provide information about the importance of each feature in predicting the target variable. This can guide feature selection or engineering efforts in more complex models.

The initial models for all clusters were built with all candidate features: 'Store', 'Promo', 'CompetitionDistance', 'SchoolHoliday', "Promo2", 'DayOfWeek', 'StoreType', 'Assortment', 'Promo.Interval'. These features were chosen after observing the correlation matrix, PCA, other exploratory data analysis graphs, and their applicability to MLR since it can only run on numerical predictors. Cluster 2 was a special case, with no store type b and promo.interval Feb, thus they are excluded from cluster 2 for all prediction. The MLR models are train with the training data of these features, and details coefficients between the predictors and sales are displayed in following figures.



For all clusters, 'CompetitionDistance' and 'SchoolHoliday' are found to be less correlated to the sales of a store. In addition, Promo.Interval is also proven to be not effective feature for clusters 0 and 3. By removing the less effectives features from the models of each cluster, the predictions on training, validation, and test data have been made.



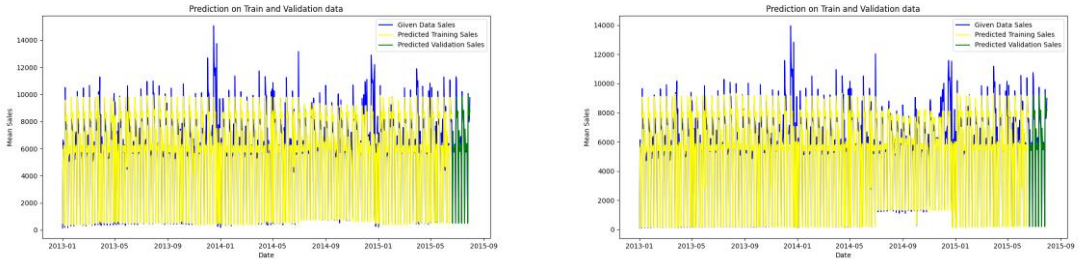


Figure 14: Actual sales (Blue) vs predicted sales on training data (Yellow) vs predicted sales on validation data (Green) of cluster 0 to 3 (from left to right, top to bottom)

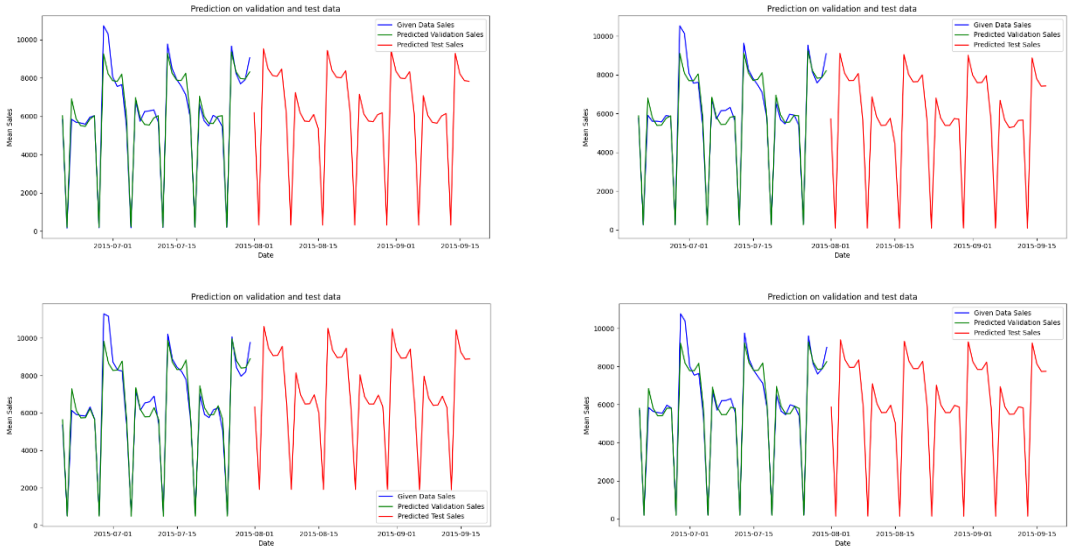


Figure 15: Actual sales (Blue) vs predicted sales on validation data (Green) vs predicted sales on test data (Red) of cluster 0 to 3 (from left to right, top to bottom).

The plotted means of predicted sales on training and validation data demonstrated a high level of accuracy when compared to the actual data. Despite noticeable errors during the early 2014 and 2015 when the sales peaked but the prediction could not catch, the prediction from MLR shows stable alignment with actual sales from all clusters.

Training	Cluster_0	Cluster_1	Cluster_2	Cluster_3
RMSE	2883	2599	2988	2697
RMSPE	52.4%	49.9%	57.06%	54.0%

Validation	Cluster_0	Cluster_1	Cluster_2	Cluster_3
RMSE	2717	2494	2820	2590
RMSPE	49.4%	41.1%	54.55%	46.4%

Figure 16: RMSE and RMSPE of prediction on training and validation data.

The RMSE and RMSPE for all clusters, both for training and validation data, revealed a significant disparity when compared with the actual data, despite the promising trends observed in the plots.

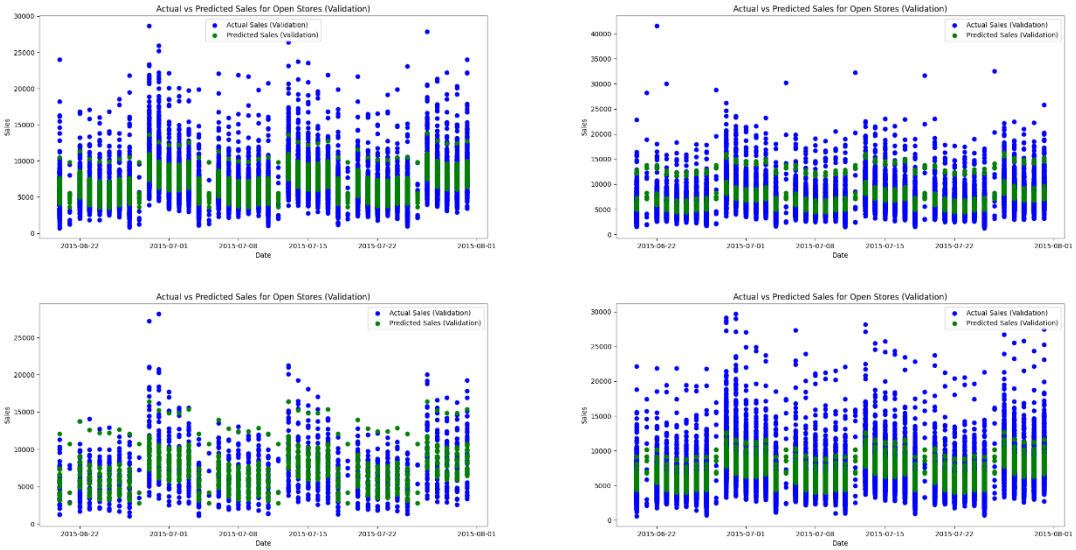


Figure 17: Actual sales (Blue) vs predicted sales on validation data (Green) of cluster 0 to 3 (from left to right, top to bottom)

MLR consistently predicts sales around the average of actual sales, however, fails to consider the high variation in sales between stores, leading to good averages but poor individual predictions. This suggests the necessity for more sophisticated prediction techniques capable of accommodating the dynamic nature of the dataset.

6.2 Random Forests.

Random Forest model was used, a predictive algorithm suitable for handling complex patterns in sales data due to its robustness and ability to effectively manage multiple types of input features. After setting up the model, the dataset was split into a training set for building the model and a validation set for testing its accuracy. Moreover, the GridSearchCV was used to get the best hyperparameters for the Random Forest Tree.

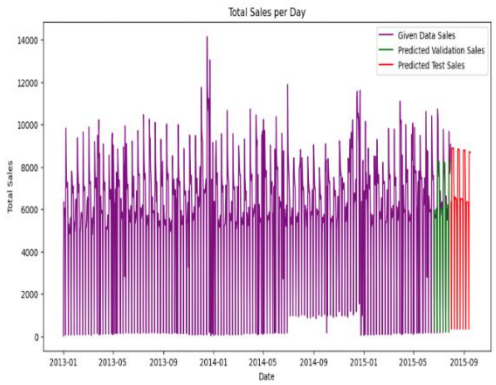


Figure 18: Prediction Model for Cluster 0

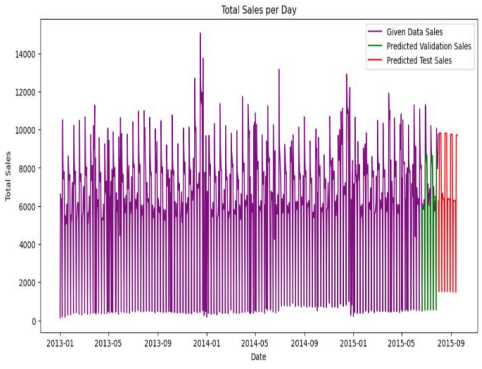


Figure 19: Prediction Model for Cluster 1

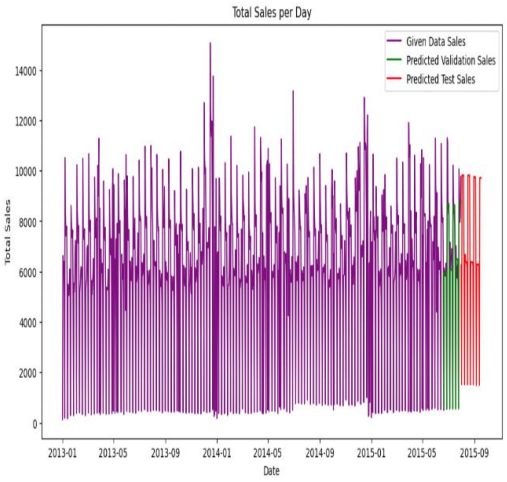


Figure 20: Prediction Model for Cluster 2

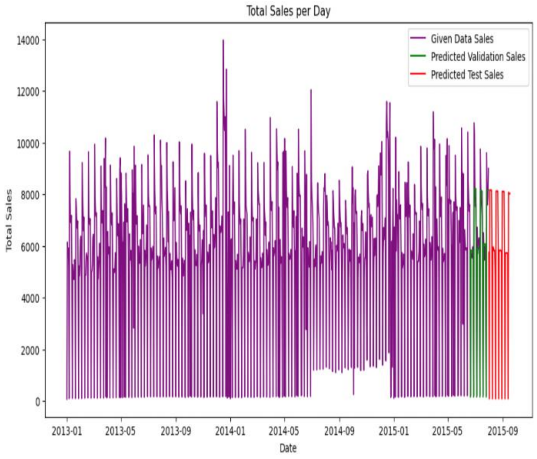


Figure 21: Prediction Model for Cluster 3

The key performance metric used to evaluate the model was RMSPE. For cluster 3, the results showed an RMSPE of approximately 30.83%, the highest among the clusters, indicating that on average, the model's predictions were within 30.83% of the actual sales figures. While this level of accuracy might be adequate depending on specific business needs, it also highlights potential areas for model improvement, such as adjusting model parameters or enhancing the data used for training.

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
RMSE_train	1398	1502	1367	1622
RMSPE_train	21.72%	22.05%	17.56%	28.95%
RMSE_validation	1437.9	1597.5	1441.1	1737.4
RMSPE_validation	23.45%	23.66%	19.8%	30.82%

Table 1: The result of RMSE and RMSPE for Random Forests model

6.3 XG Boost.

This section details the development of a predictive model to forecast store sales. The model was built using XGBoost, for its accuracy and ability to handle various data types.

The model used a variety of features extracted from the data to predict store sales. These features included numeric values like promotional activity (Promo, Promo2) and competition distance (CompetitionDistance), categorical variables like school holidays (SchoolHoliday) and day of the week (DayOfWeek), and additional features created by encoding categorical variables such as assortment type (Assortment), state holidays (StateHoliday), and store type (StoreType) into a one-hot format.

Before training the model, the data underwent preprocessing to handle the categorical variables. This involved converting them into a one-hot encoded format suitable for the model. Then, the data for open stores was divided into training and validation sets based on date ranges. The training set was used to train the model, and the validation set was used to assess its performance.

The selection of the XGBoost regressor for this task was not arbitrary. It was chosen for its robustness in handling various data types, its accuracy, and its ability to prevent overfitting through regularization techniques. The model was further fine-tuned by optimizing its hyperparameters using GridSearchCV to find the configuration that maximized its performance.

The model's performance was evaluated using two key metrics: RMSE and the RMSPE. Lower values of both metrics indicate superior model performance, implying that the predictions are closer to the actual sales values.

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
RMSE_train	1173	1278	1145	1289
RMSPE_train	15%	18%	14.5%	21.9%
RMSE_validation	1118.4	1310	1118.4	1342.2
RMSPE_validation	13.7%	19.5%	16%	23%

Table 2: The result of RMSE and RMSPE for XGBoost model

After training and validation, the model generated sales predictions for unseen test data. Finally, the predicted sales were compared with the actual sales data to assess how well the model generalizes to unseen data.

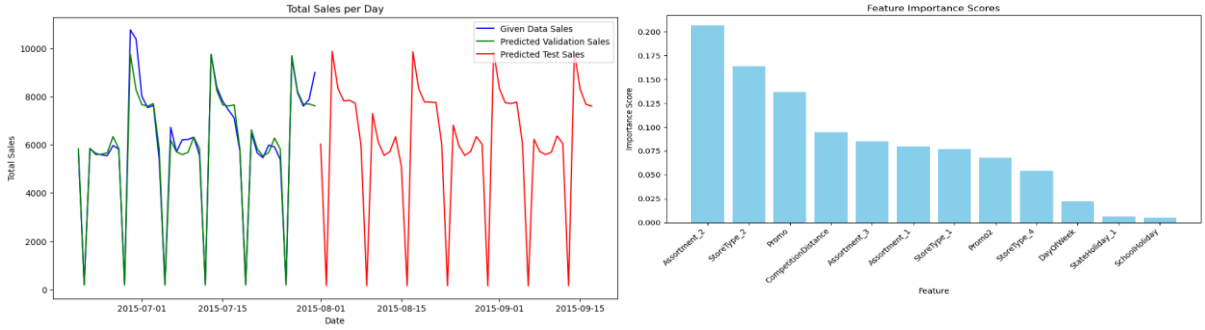


Figure 22: Prediction Sales & Importance of variables for Cluster

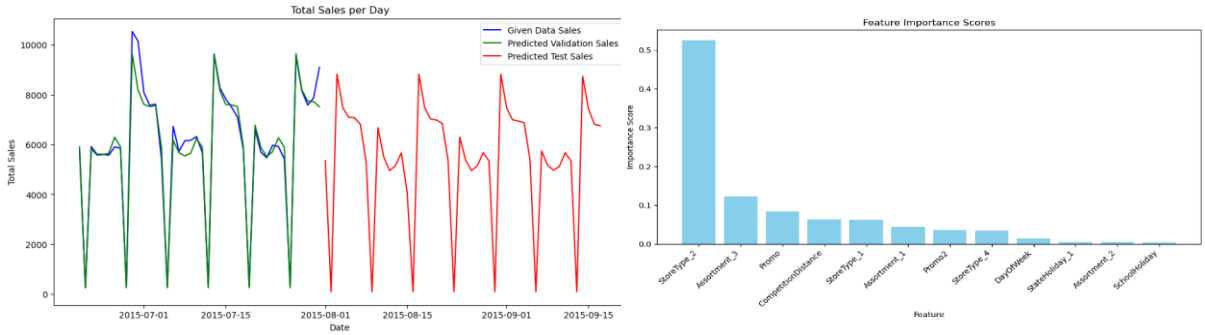


Figure 23: Prediction Sales & Importance of variables for Cluster

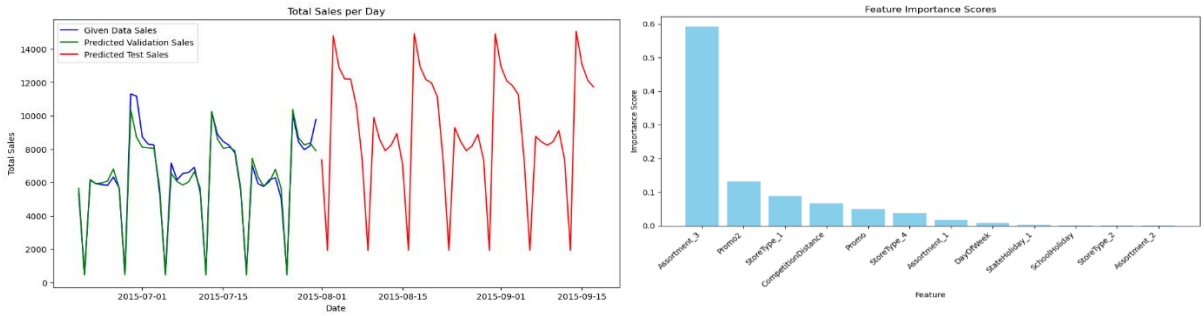


Figure 24: Prediction Sales & Importance of variables for Cluster 2

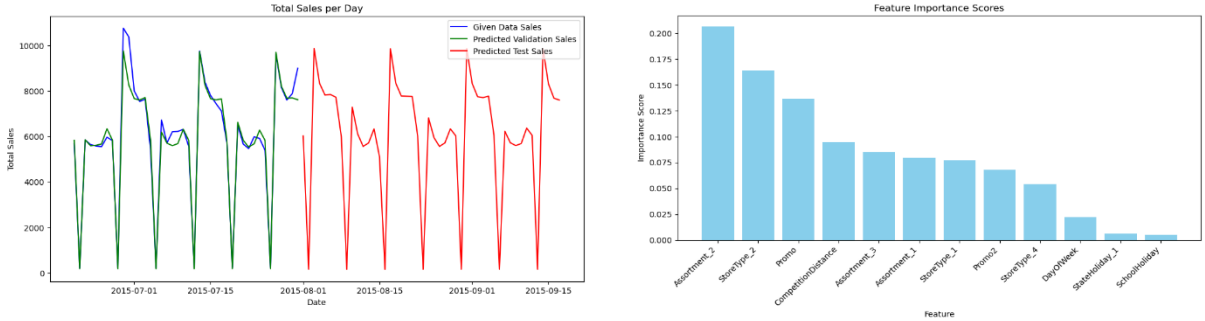


Figure 25: Prediction Sales & Importance of variables for Cluster 3

6.4 Neural Network.

6.4.1 Deep Neural Networks

In our first Neural Network model we created a Deep Neural Network with 1 hidden layer having 15 nodes and Rectified Linear Unit (ReLU) as activation function to be able capture the non-linear complex relationships between Sales and other features. The model was tested with epochs ranging from 40 to 70. 50 epochs and 20 batchsize were found to be the best performing parameters to the model. A key thing to note here is that since Neural Networks cannot process datetime values, Date values were subtracted from the first date in the train set to create a numerical 'Date_difference' variable.

Clusters	0	1	2	3
RMSE_train	1457.2	1606.9	1357.3	1725.7
RMSPE_train	25.57%	39.93%	19.45%	28.78%
RMSE_validation	1469.4	1596.2	1349.4	1727.6
RMSPE_validation	26.68%	41.27%	20.01%	29.86%

Table 3: The result of RMSE and RMSPE for Deep Neural Network model

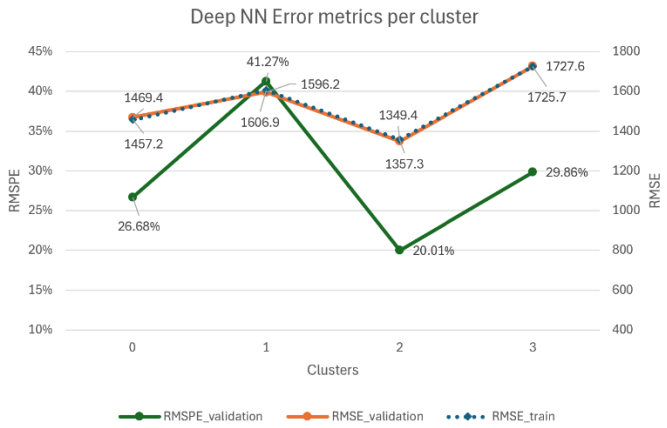


Figure 26: RMPSE & RMSE per cluster

Performance evaluation using RMSE and RMSPE showed that the model was neither under- nor over-fitted. Hence, we advanced to predict Sales for the test data. Similar to the train dataset, test was clustered, and predictions were made for the clustered data.

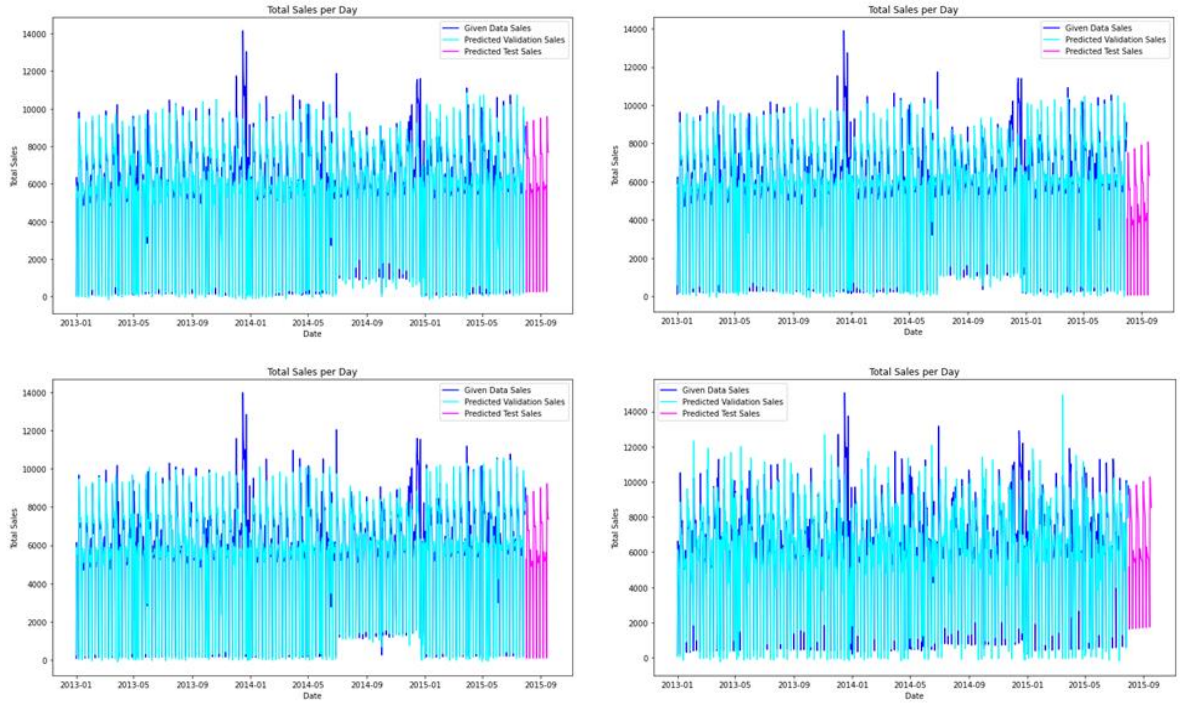


Figure 26: Predictions for clusters 0, 1, 2 and 3 (clockwise from top left)

To further improve upon the predictions, we proceeded to apply RNN to predict the sales.

6.4.2 Recurrent Neural Networks

Since the dataset is time-series data, the second Neural Network model option is long short-term memory (LSTM), which is a concept of recurrent neural network. Before training the model, all the numerical columns are normalized to ensure that each feature contributes approximately proportionally to the model's learning process. This prevents certain features from dominating solely based on their scale.

This model extracted several variables from the original data and created new variables:

- Original variables: Store, DayOfWeek, Open, Promo, StateHoliday, SchoolHoliday, StoreType, Assortment, Sales
- New variable: The dataset includes a new variable called "CompetitionStart*Distance". This variable is created by multiplying "CompetitionOpenSinceYear" and "CompetitionOpenSinceMonth" and then multiplying the result with "CompetetionDistance". The "CompetitionStart" variable is binary and only has values 0 and 1. It indicates the existence of competitors in the market. For example, if the competitors started their business in July 2014, the data before 1st July 2014 will be set as 0, and the data from 1st July 2014 until the last date of the training data will be set as 1. However, some numbers in "CompetitionStart" are set as 0, which means that even though the result of multiplying distance is 0, they do not exist in the market. To handle these cases, they will be assumed to be set as 10000, indicating that they are very far from the other stores. This would ensure that they do not affect the other stores.

Based on the relationship between DayOfWeek and Sales, it is clear that there is a weekly cycle. To make an accurate prediction, we will extract the previous sales data from the first day of the week to the seventh day, and use it to predict sales for the eighth day. During the training process, the data will be split into two parts: 80 percent for training and 20 percent for validation. Before calculating the RMSE (Root Mean Square Error) and RMSPE (Root Mean Square Percentage Error) values, the results will be transformed back into the original scale, as they were normalized at the beginning. During the process, if the store number of sales from the previous week does not match the predictive day, the data will be deleted because they are different stores.

To evaluate the accuracy of the model, we will use a table that shows the RMSE and RMSPE values, which are standard measures of accuracy.

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
RMSE_train	1248.5	1012.2	1293.4	1470.4
RMSPE_train	21%	16%	20%	25%
RMSE_validation	1351.5	1128.2	1373.4	1598.4
RMSPE_validation	23%	17%	22%	27%

Table 4: The result of RMSE and RMSPE for LSTM model

Below are the predictions from part of the validation data for 4 clusters.

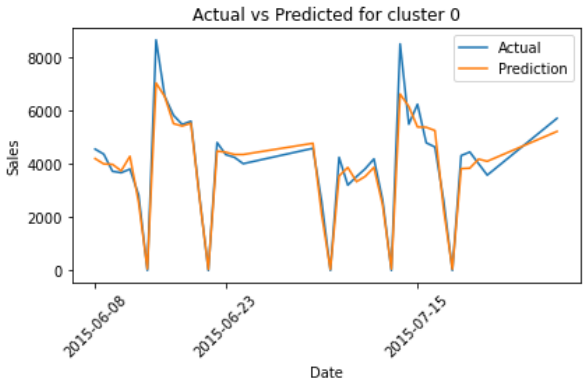


Figure 27: Prediction model for cluster 0

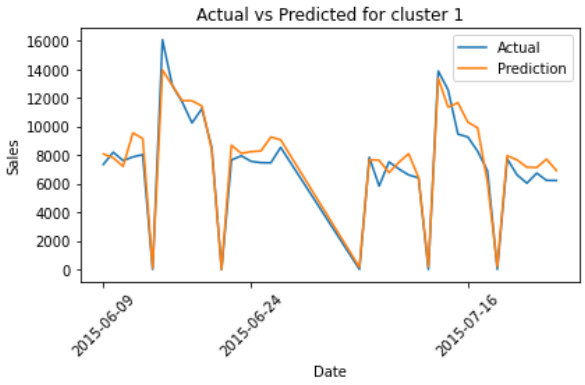


Figure 28: Prediction model for cluster 1

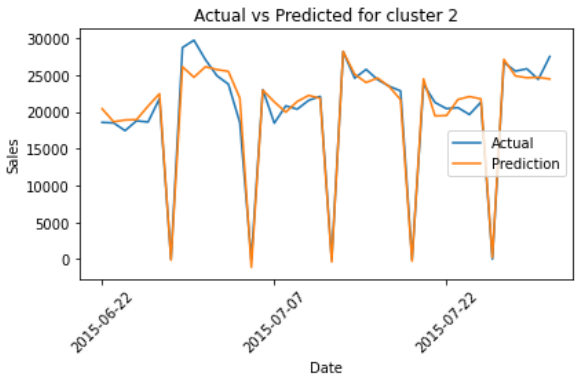


Figure 29: Prediction model for cluster 2

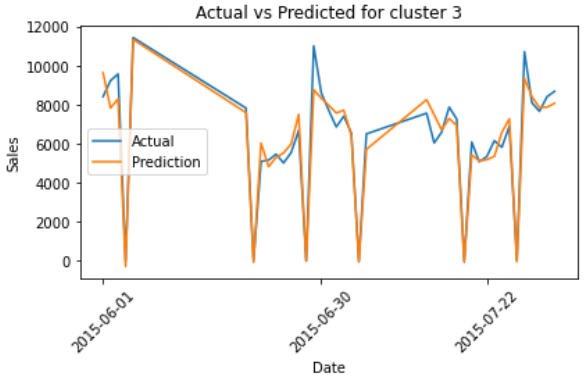


Figure 30: Prediction model for cluster 3

7 Conclusions and Recommendations

This report analysed sales data for 1,115 German drug stores to predict daily sales over six weeks. The report identified key factors influencing sales, including promotions, competition, holidays, seasonality, and store location, by leveraging historical data and a comprehensive data pre-processing approach. Clustering stores based on mean sales allowed for targeted model development and decision-making. This approach has the potential to optimize resource allocation and marketing efforts.

The report investigated the performance of multiple machine learning models, including multiple linear regression, Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks. While all models achieved a certain level of accuracy, XGBoost demonstrated the most promising results, with RMSE values ranging from 1118.4 to 1342.2 and RMSPE values between 13.7% and 23%.

Recommendations:

- Implement the XGBoost model for daily sales forecasting across the drugstore chain.
- Leverage insights from store cluster analysis to develop targeted marketing campaigns and optimize resource allocation.
- Explore the potential of incorporating additional data sources, such as products or customer demographics, to further enhance model accuracy.

By implementing these recommendations, the drugstore chain can leverage data-driven insights to optimize sales forecasting, improve decision-making, and achieve greater profitability.