ROSSMAN SALES PREDICTION

1. Introduction

Rossman, a drug store chain, aims to predict sales from August 1, 2015, to September 17, 2015, for 1,115 of their stores. The variables for each of the available dataset are explained in Table 1 and Table 2, respectively.

Table 1: Train and Test Dataset Variables

Variable Name	Description
Store	Store number
DayOfWeek	Day of the week
Date	The given date
Sales	The turnover
Customers	The number of customers
Open	Store open status: 0 = closed, 1 = open
Promo	Indicates store-specific promo status
StateHoliday	Indicates a state holiday. a = public holiday, b = Easter holiday, c = Christmas, 0 = none
SchoolHoliday	Indicates public schools closure impact on (Store, Date)

Table 2: Store Dataset Variables

Variable Name	Description
Store	Store number
StoreType	Store models: a, b, c, d
Assortment	Assortment level: a = basic, b = extra, c = extended
CompetitionDistance	Distance (meters) to the nearest competitor store
CompetitionOpenSinceMonth	The approximate month the nearest competitor was opened
CompetitionOpenSinceYear	The approximate year the nearest competitor was opened
Promo2	A continuing promotion: 0 = not participating, 1 = participating
Promo2SinceWeek	The calendar week the store started participating in Promo2
Promo2SinceYear	The year the store started participating in Promo2
PromoInterval	The consecutive intervals in which Promo2 is restarted

The training set contains daily sales data from January 1, 2013, to July 31, 2015, while the store dataset details each store's type, assortment, competitors, and promotion activity by store number.

2. Methodology

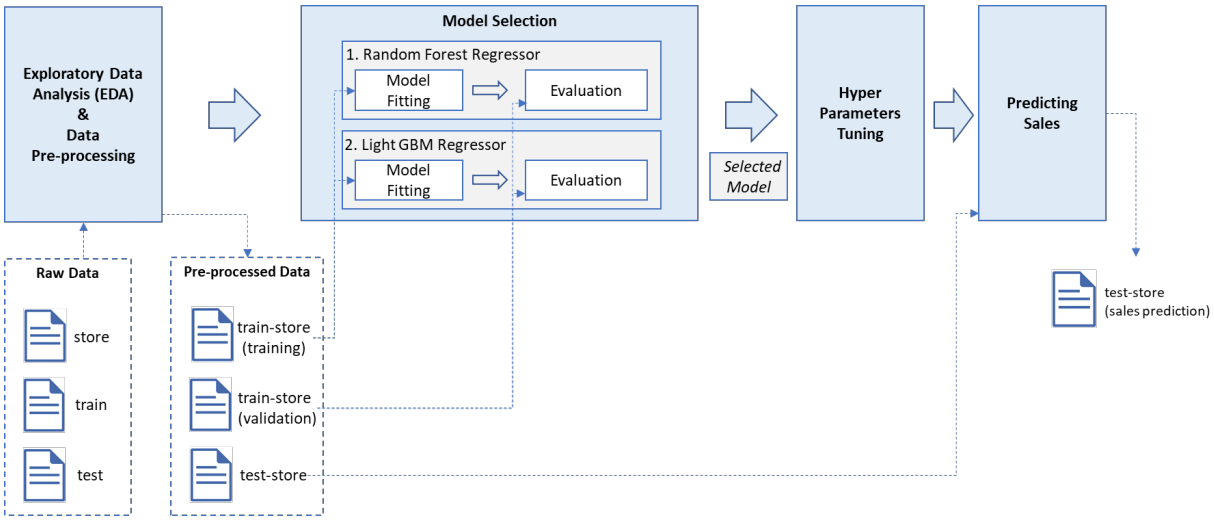


Figure 1: Prediction Workflow

The sales forecasting workflow depicted in Figure 1, started with examining the dataset through EDA to identify missing values and variables impact on sales, alongside data pre-processing. The training and test data were merged with store information for a comprehensive analysis, followed by imputation, encoding, and feature extraction. This pre-processed train dataset was then split and used for modelling.

Two predictive models, Random Forest and Light Gradient Boosting Machine, were developed and tested against the validation set to avoid overfitting. The hyperparameters of the chosen model was then fine-tuned and applied to forecast sales on the test-store dataset.

3. Exploratory Data Analysis (EDA) & Data Pre-processing

3.1. Data Cleaning & Preparation

3.1.1. Missing Data

Missing values were observed in both store and test datasets. For test dataset, only missing values in the Open variable were addressed, as Sales and Customers were intentionally left blank for predictions.

Table 3: Store Dataset Missing Values

Store	0
StoreType	0
Assortment	0
CompetitionDistance	3
CompetitionOpenSinceMonth	354
CompetitionOpenSinceYear	354
Promo2	0
Promo2SinceWeek	544
Promo2SinceYear	544
PromoInterval	544
dtype: int64	

Table 4: Test Dataset Missing Values

Store	0
DayOfWeek	0
Date	0
Sales	41088
Customers	41088
Open	11
Promo	0
StateHoliday	0
SchoolHoliday	0
dtype: int64	

Regarding promotion, Figure 2 shows that all missing values were caused by the stores not participating in the promo2, therefore these values were then imputed with zero.

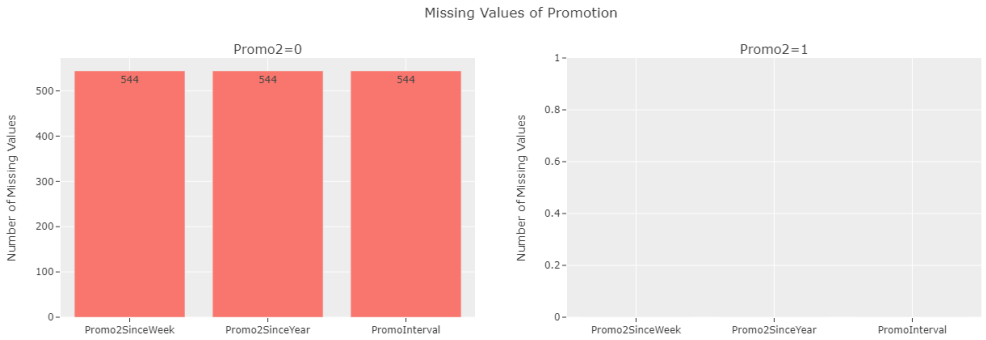


Figure 2: Missing Values of Promotion2

The matrix in Figure 3 highlights missing competition data with white stripes, notably three instances of missing CompetitionDistance with simultaneous absences in CompetitionOpenSinceMonth and CompetitionOpenSinceYear. Assuming missing data as no nearby competitors, missing CompetitionDistance were set to 100,000, above the highest record of 78,000, with the other two variables set to future dates, January 2016.

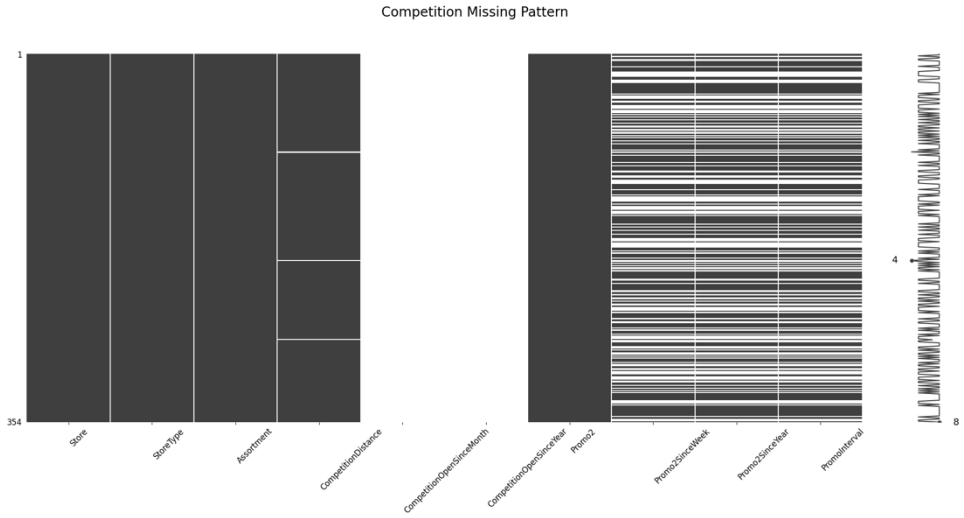


Figure 3: Missing Values Pattern of Competition

The remaining missing values of CompetitionOpenSinceMonth and CompetitionOpenSinceYear appear in the same records, yet their absence shows no relation with CompetitionDistance, StoreType, or Assortment. These might result from data processing errors or lack of information on competitors' opening dates. With over 30% of data missing, imputing these values with the mode might be preferable to prevent changes in the distribution.

Additionally, the test dataset shows a few missing Open variable entries. Comparing these dates with other stores' and store 622's historical openings suggests these stores were open, thus missing values could be set to one.

3.1.2. Merge Data

The train dataset was then merged with store to gain more insight into the relationship between store-related variables and sales.

3.2. Exploratory Data Analysis

3.2.1. Sales Performance

Figure 4 illustrates the monthly average trends of sales amounts and customers from January 2013 to July 2015.

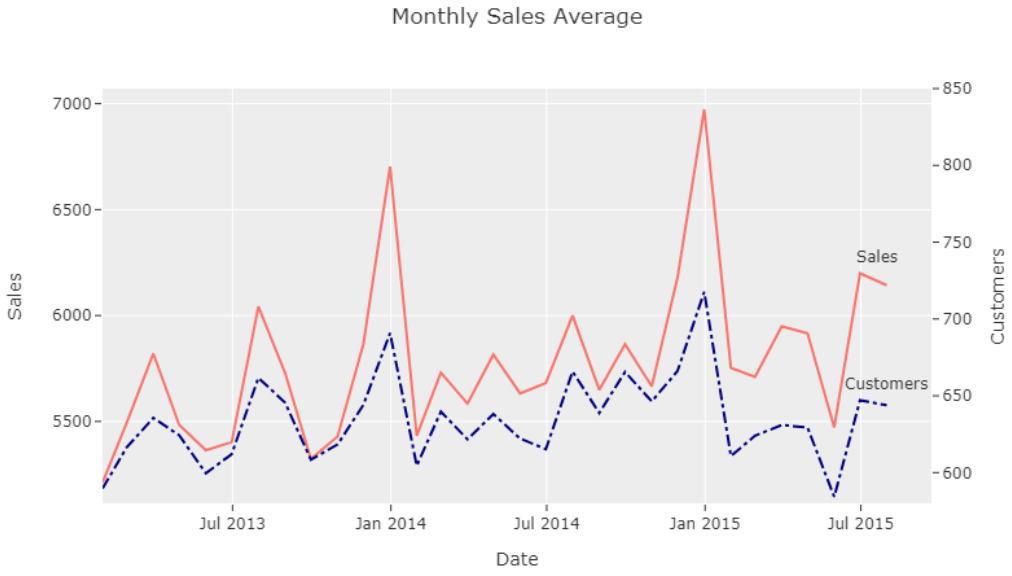


Figure 4: Monthly Sales and Customer Trends

The data indicates a gradual increase in sales and customers over this timeframe, with notable peaks observed. December consistently marked the highest peak, followed by peaks occurring between June and July, as well as March and April.

3.2.2. Store Type and Assortment

The dataset includes 1,115 stores with different types and assortments shown in Figure 5.

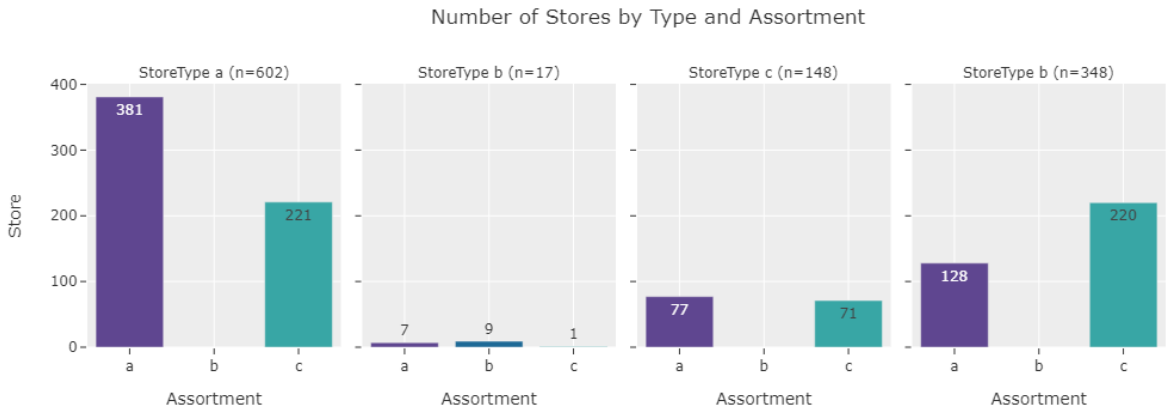


Figure 5: Number of Stores

Figure 6 shows the average sales for each store type and assortment in the training dataset. Store type b generally performed best, especially with an extended assortment level c, which had the highest sales.

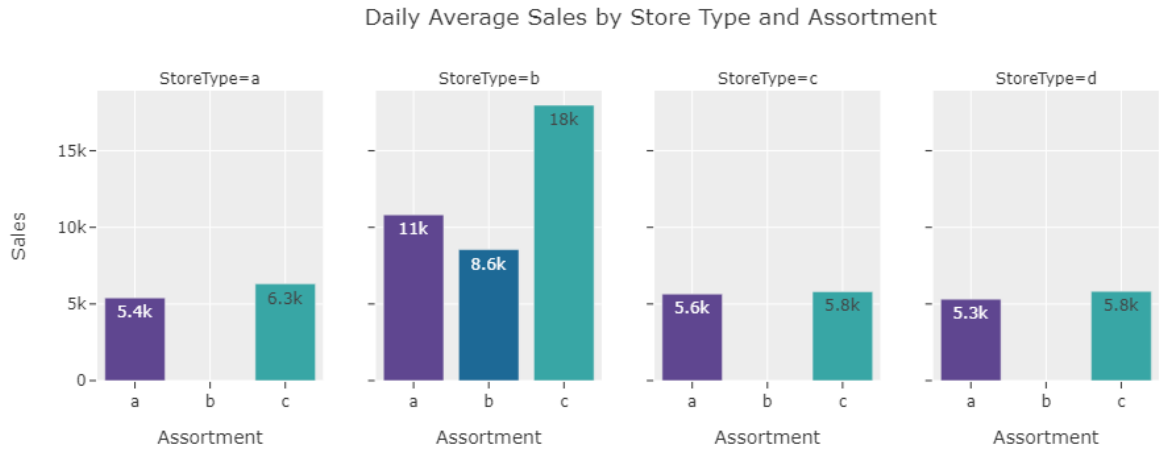


Figure 6: Daily Average Sales by StoreType and Assortment

Moreover, each store types displays similar trend for each day of week with the exception for store type b. Most of the sales dropped on Sunday due to closures, while store type b remained open and experienced a slight drop on Saturday.

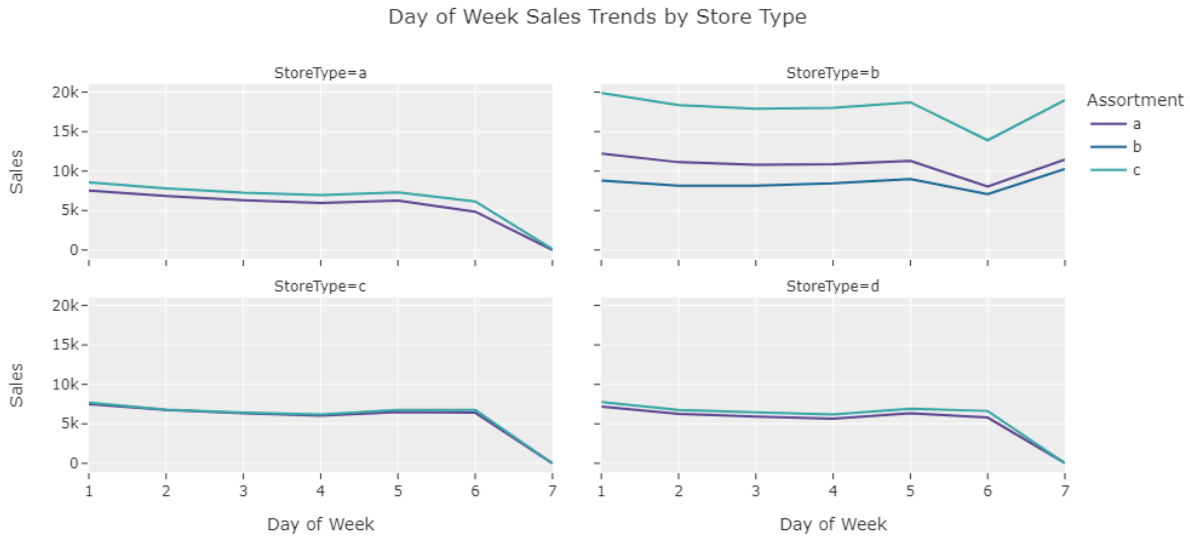


Figure 7: Day of Week Sales Trends

3.2.3. Promotion

There are two types of promotion, store-specific promotion and a continuing promotion program, promo2. The impacts of both promotion on sales are illustrated in Figure 8.

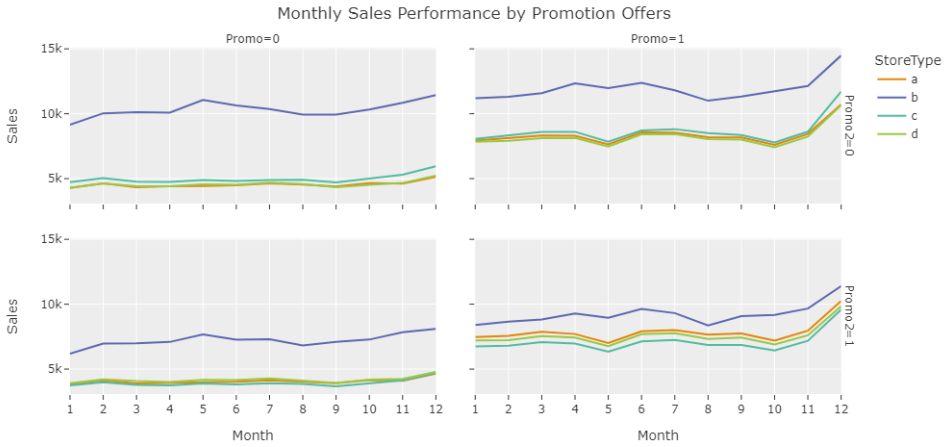


Figure 8: Promotion Impact on Sales

The sales during store-specific promotion were higher compared to the promo2. In fact, sales were slightly lower in promo2 participation. This might suggest promo2 will not be significant in predicting sales.

3.2.4. Competition

The competition related variables indicated the presence of competitors nearby the stores, and the time the competitors started to operate. Figure 9 displays the influence of CompetitionDistance on sales, grouped by distance intervals to discern potential variations across each interval.

Figure 9: Competition Distance Impact on Sales

Interestingly, the figure indicates the farther the competitors, the lower the sales. Despite seeming counterintuitive, nearby competitors might indicate higher demand in the area.

In addition, the CompetitionOpenSInceMonth and CompetitionOpenSinceYear could be transformed into CompetitionDuration by calculating the difference with the date of sales record. The relation to sales is shown in Figure 10.

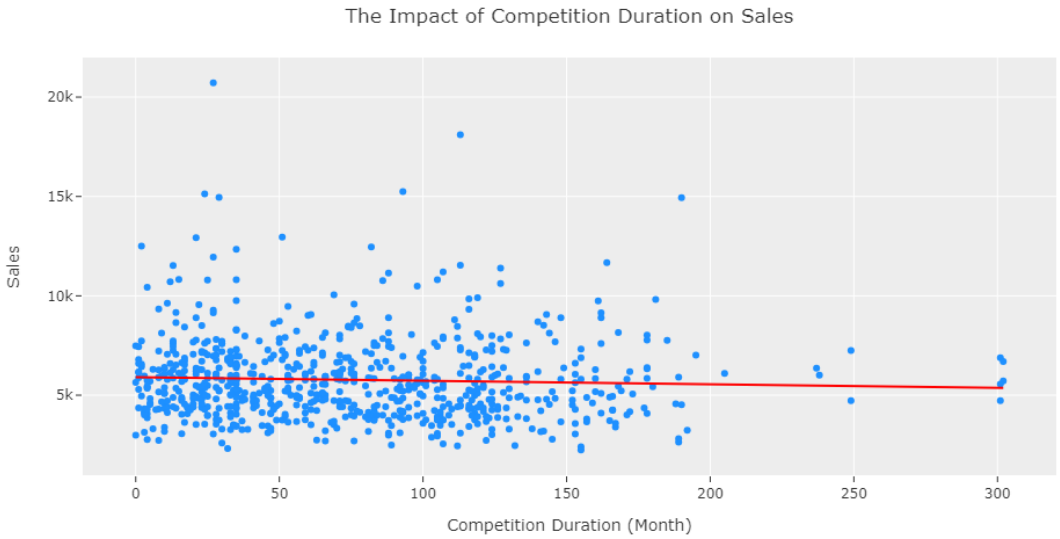


Figure 10: Competition Duration Impact on Sales

The figure indicates that longer competition durations are associated with lower sales. Due to around 30% of imputed duration data, this relationship might not be fully captured. Considering its weak association, excluding it from modelling might prevent biased predictions.

3.2.5. Holiday

There are two types of holidays in this dataset, StateHoliday which includes public holiday, Easter, and Christmas; and SchoolHoliday, to indicate the holiday of public schools. The sales performance comparison is as follows.

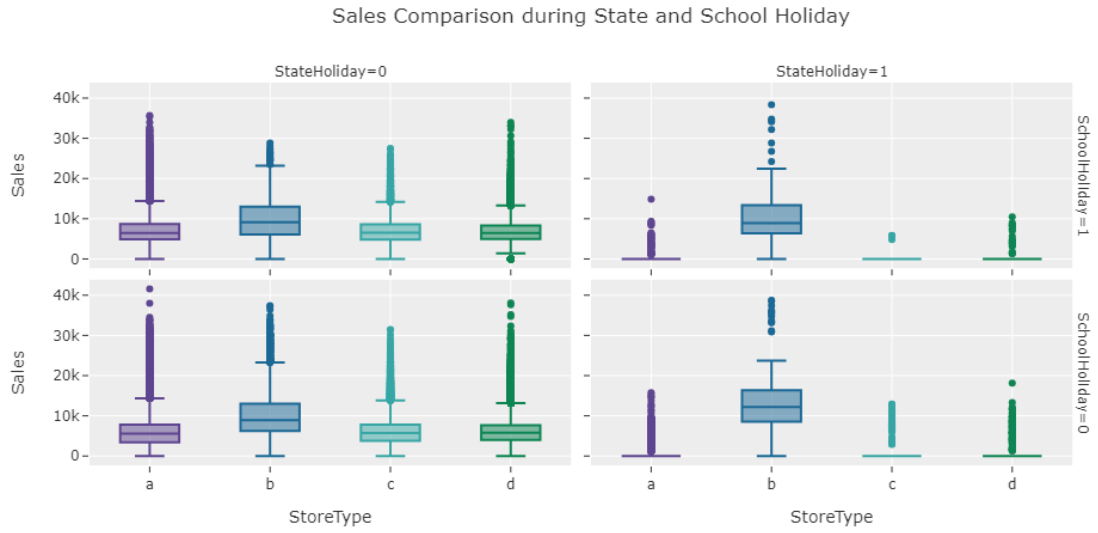


Figure 11: Sales Performance during State and School Holiday

During SchoolHoliday, there appears to be no significant impact on sales. However, during StateHoliday, most store types reported zero sales due to closures, except for store b, which

maintained relatively consistent sales. Therefore, the relation between holiday variables and sales could be explained by the Open variable.

3.3. Data Pre-processing

In this stage, some feature extractions, encoding, feature selection, and splitting the train dataset were carried out.

3.3.1. Feature Extraction

New variables were created from existing ones to enhance sales prediction accuracy. Apart from CompetitionDuration, other variables that could improve prediction accuracy include sales lag and rolling means. These two help in capturing historical sales trends. Sales lag represents sales from the previous n days. Given the objective of predicting sales for 48 days, the sales lag was set to 49 days, multiples of seven, to maintain consistency across day of week. This means the sales record considers the sales amount of the previous 49 days in the prediction model. Similarly, a sales lag of 365 days was also applied to capture annual trends. Additionally, rolling means might aid in smoothing the prediction and preventing overfitting, making it less sensitive to sudden changes, particularly those deviating from the usual pattern.

3.3.2. Encoding

To enable machine learning modelling, categorical variables like StoreType and Assortment must be encoded. For instance, store types a, b, c, and d were encoded as 1, 2, 3, and 4, respectively. This step is crucial as most machine learning libraries only handle numerical values.

3.3.3. Feature Selection

Only a subset of features was used for prediction to avoid overfitting. Learned from the EDA, these were the features utilized to build the model. Detailed explanation available in Appendix i.

Table 5: Selected Features

No.	Features	No.	Features
1	Store	**6**	Assortment
2	DayOfWeek	**7**	CompetitionDistance
3	Open	**8**	SalesLag49
4	Promo	**9**	SalesLag365
5	StoreType	**10**	RollingMeans7

3.3.4. Splitting Train Dataset

To enhance model generalization, the training dataset was divided into two based on the date. The validation dataset covered 48 days, same as test dataset, from June 14, 2015 to July 31, 2015, and used to evaluate the model's performance on unseen data.

4. Model Selection

Both Random Forest and Light GBM Regressor models were constructed and compared. Random Forest is a widely-used ensemble approach that is robust to overfitting and typically produces low bias predictions (Breiman, 2001). As a comparison, Light GBM, a gradient boosting algorithm known for its fast and efficient computation and ability to avoid overfitting (Ke et al., 2017), was also implemented. The prediction results are presented below.

Table 6: Evaluation Metrics Comparison

Metrics	Dataset	Model	
		Random Forest	Light GBM
RMSPE	Train	0.09	0.28
	Validation	0.16	0.19
MAE	Train	263.64	785.83
	Validation	661.66	752.75

Although the Light GBM model was able to avoid overfitting, its scores were still higher than those of Random Forest. Therefore, Random Forest was selected.

5. Hyperparameters Tuning

The hyperparameters of the random forest were then tuned incorporating time series cross-validation. In this stage, a randomized search was applied to find the best combination of hyperparameters. Although the results in Table 7 might appear worse, the model is more effective in avoiding overfitting, makes it more adept at predicting unobserved data. The prediction results are presented in Appendix ii.

Table 7: Evaluation Metrics After Hyperparameters Tuning

Metrics	Dataset	Random Forest
RMSPE	Train	0.29
	Validation	0.19
MAE	Train	836.37
	Validation	780.6

6. Prediction

The model was subsequently employed to forecast sales in the test-store dataset. The outcome of the prediction is presented below.

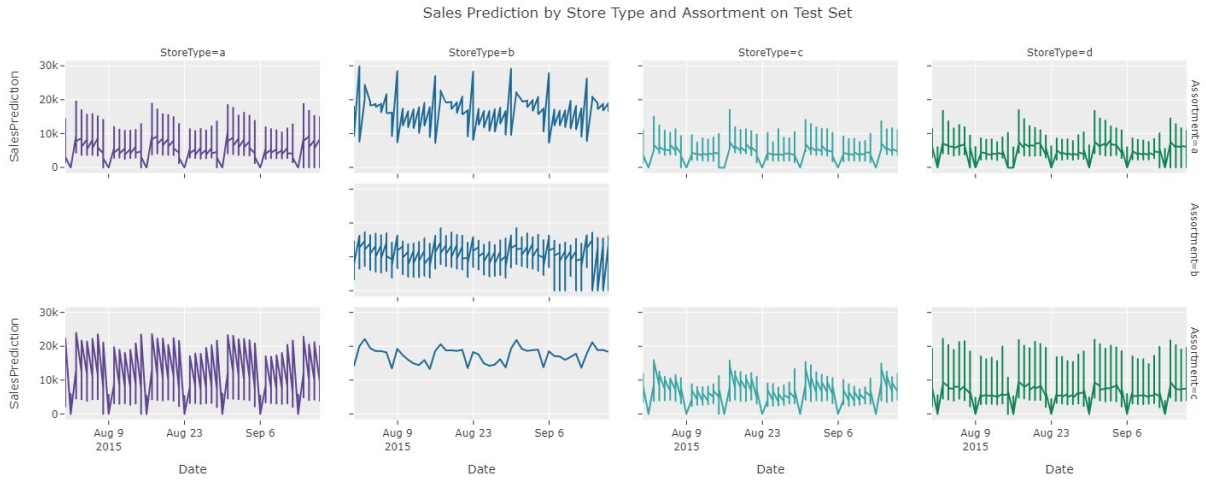


Figure 12: Sales Prediction Results

The forecast appears adept at identifying trends for each type of store and assortment. It could also detect weekly trends specific to each store type. Moreover, it takes into account the store's opening times, noticeable on Sundays.

7. Conclusion and Recommendation

7.1. Conclusion

1. The Random Forest model can produce sales predictions that capture the trends in sales using the provided variables and is able to avoid overfitting.
2. The model performance might be limited by a higher proportion of missing values, particularly in CompetitionOpenSinceMonth and CompetitionOpenSinceYear, which leads to uncertainty in data imputation and may result in a potential loss of information.

7.2. Recommendation

1. Updating predictions over shorter periods than 6-week might better capture sales changes.
2. Incorporating additional features, such as population numbers in store neighbourhoods, could offer valuable insights and enhance prediction accuracy.
3. Developing sales predictions for each product category could provide insightful information for inventory decision and resource allocation.

I. References

Breiman, L., 2001. Random Forests. Machine Learning, 45(1), pp.5-32. Available at: https://doi.org/10.1023/A:1010933404324 [Accessed 4 February 2024].

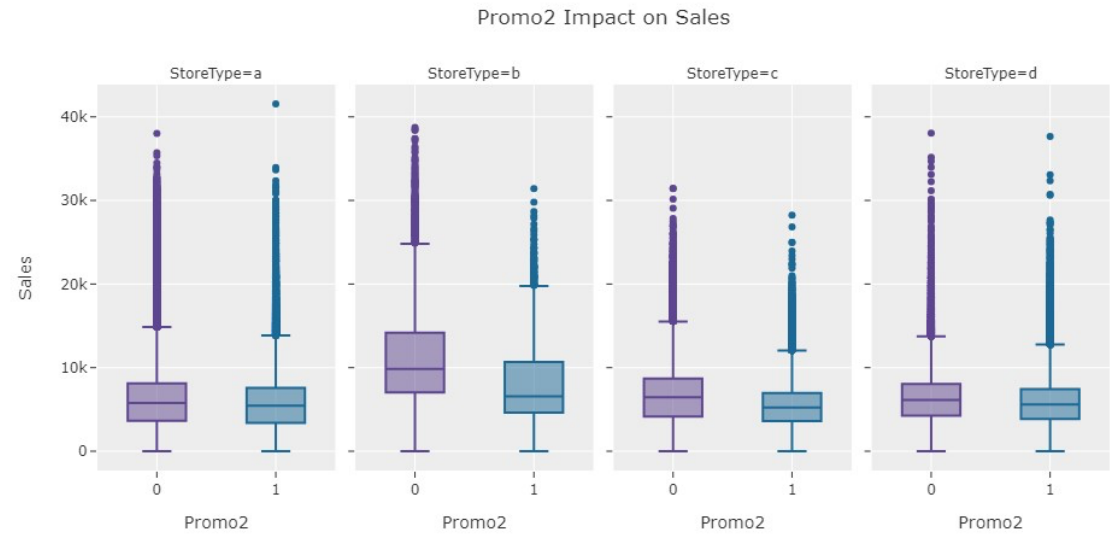
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). [pdf] Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf [Accessed 4 February 2024].

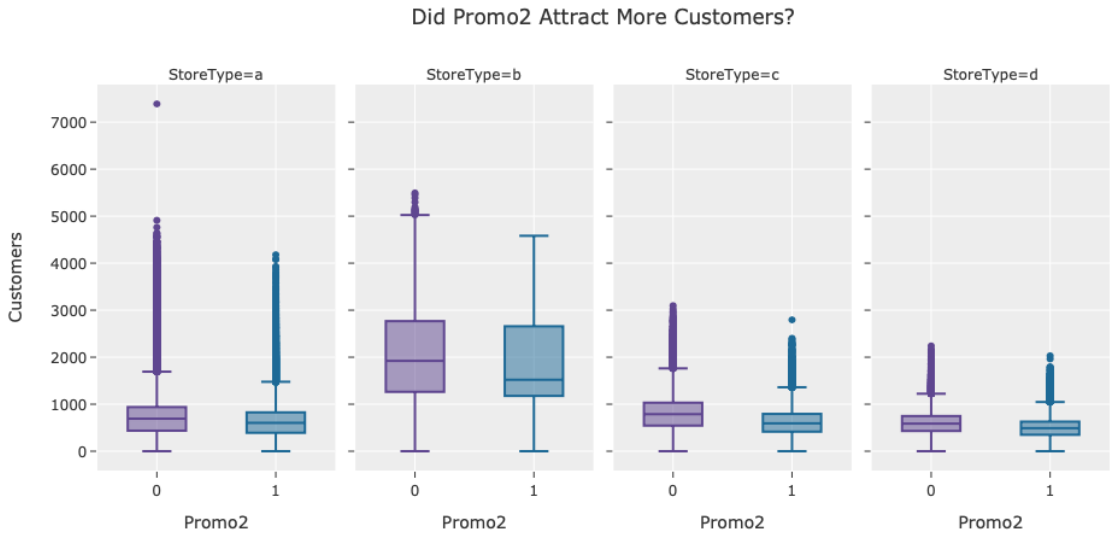
II. Appendix

i. Features Selection

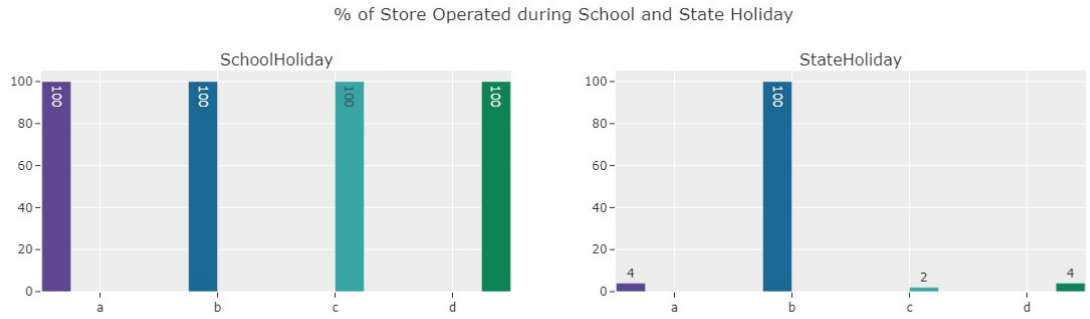
1. Dropped Feature: Promo2

   Promo2 shows no significant impact on sales and customers. Thus, all promo2 related variables were dropped.

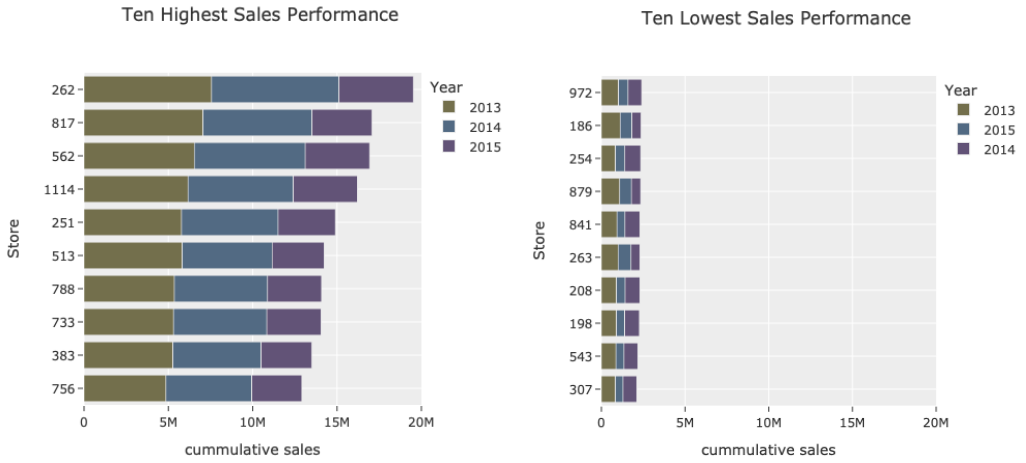




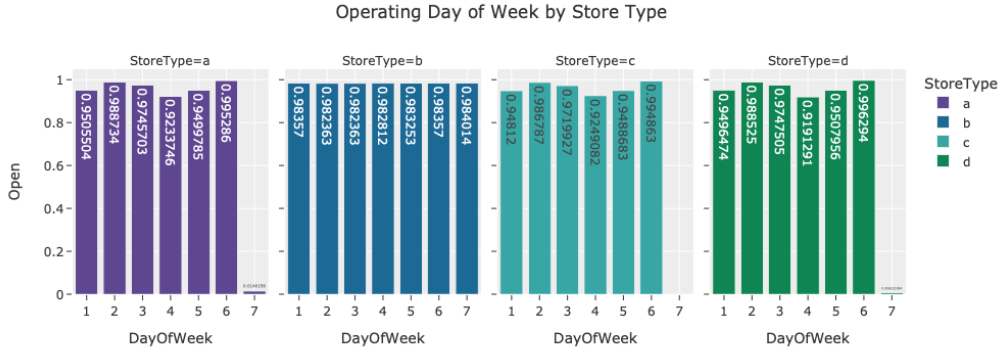
2. Dropped Feature: StateHoliday and SchoolHoliday
   The differences in sales during the StateHoliday explained by the closure of most stores. While during the school holiday, the differences were insignificant as explained in Figure 11.



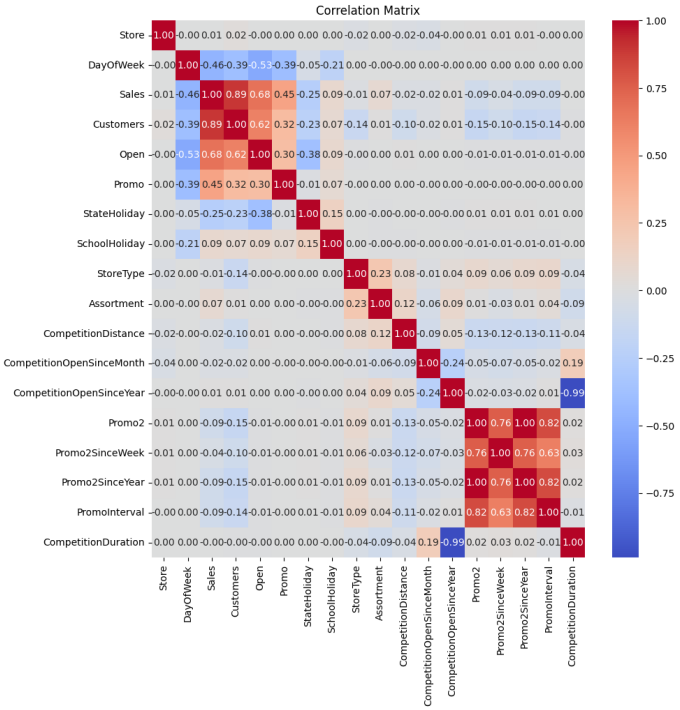
3. Kept Feature: Store
   Each store had different sales performance, which might be useful in predicting sales. These are the highest and lowest ten.



4. Kept Feature: Open
   The figure displays the average of opening time mapped in day of week. This supports the day of week trends where most sales in store a, c, and d dropped on Sunday due to closure operation.

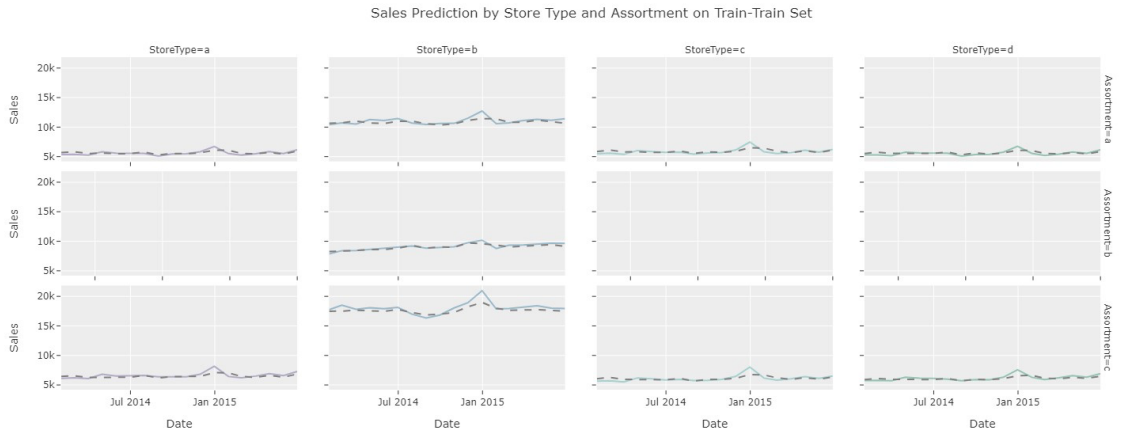


5. Correlation Matrix



ii. Prediction Results (After Hyperparameters Tuning)

1. Actual and Prediction of Sales (Training Set)



2. Actual and Prediction of Sales (Validation Set)