DATA71011 Understanding Data and their Environment: ROSSMAN Sales Prediction Model

Word count: 2,103

1. Introduction

This report is to build a sales prediction model for ROSSMAN drug store chain using three datasets - Store, Test, and Train. In short, the Store dataset contains information about 1,115 drug stores, and Train dataset is a historical sales record from 01/01/2013 to 31/07/2015. The Test dataset has the same structure as Train but does not have any sales record, and sales need to be predicted using our model (Section 2 for details).
The main issue is pre-processing: handling missing values, duplicates, and data type issues to make the dataset appropriate for prediction model training and testing (Section 2). After merging Store and Train, three new variables will also be introduced based on the existing variables to see if they can improve sales prediction performance. In Section 3, We will discuss what prediction models can be used, and in Section 4 key factors on sales will be figured out based on the results of the chosen model (Section 4).

2. Exploratory data analysis

This section is to understand the characteristics of the dataset, identify issues, and do pre-processing. As mentioned in Section 1, three new variables – CompetitionDays, Promo2Days, Promo2Month – will also be introduced. After univariate analysis and pre-processing, a multivariate analysis will be made to investigate relationships between certain variables. The result of the multivariate analysis will provide insights into which variables can be useful predictors in terms of feature selection.

2.1. Univariate analysis

2.1.1. ‘Store’ dataset

1) Descriptions

The Store dataset contains information for 1,115 stores along with ten columns. Below are short descriptions of each column, along with datatypes and the number of missing values.

Figure 1. General description of Store dataset

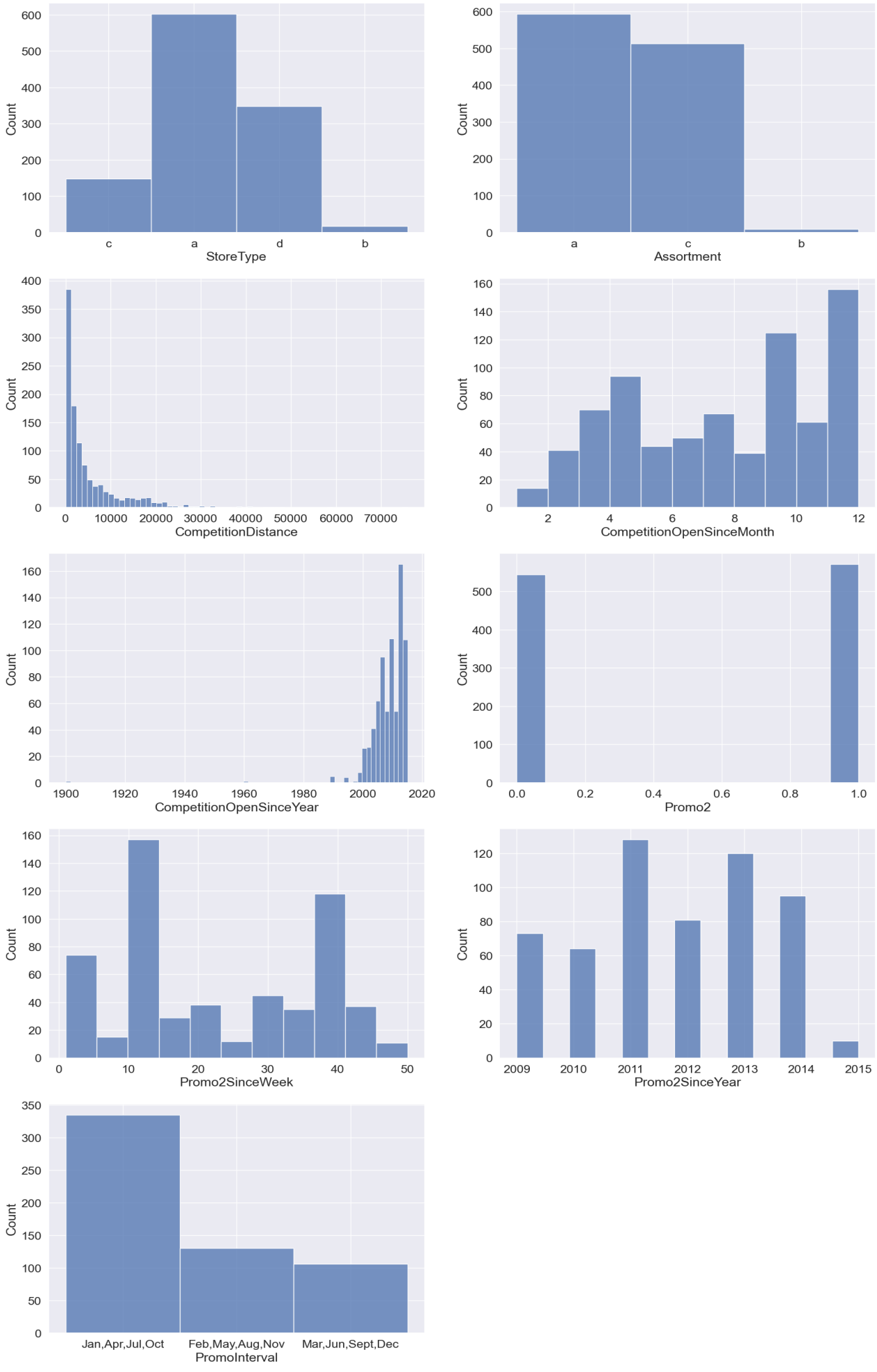
Column name	Description	Data Type	No. of missing values
Store	Store number	Int64	0
StoreType	Storetypes: a, b, c, d	Object	0
Assortment	Assortment Level: a=basic, b=extra, c=extended	Object	0
CompetitionDistance	Distance to the nearest competitor store (meters)	Float64	3
CompetitionOpenSinceMonth	The approximate month since the nearest competitor was opened	Float64	354
CompetitionOpenSinceYear	The approximate year since the nearest competitor was opened	Float64	354
Promo2	Status of participating coupon-based mailing campaign promotion: 0=not participating, 1=participating	Int64	0
Promo2SinceWeek	The calendar week when the store joined Promo2	Float64	544
Promo2SinceYear	The year when the store joined Promo2	Float64	544
PromoInterval	The intervals of Promo2; Term in which the coupon is issued: “Jan, Apr, Jul, Oct”, “Feb, May, Aug, Nov”, or “Mar, Jun, Sept, Dec”	Object	544

Figure 2 shows the statistical characteristics of the Store dataset except for Object type variables, and Figure 3 is histograms to show the distribution characteristics of all variables except the Store column.

Figure 2. Statistical summary of Store dataset

	Store	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
count	1115.00000	1112.000000	761.000000	761.000000	1115.000000	571.000000	571.000000
mean	558.00000	5404.901079	7.224704	2008.668857	0.512108	23.595447	2011.763573
std	322.01708	7663.174720	3.212348	6.195983	0.500078	14.141984	1.674935
min	1.00000	20.000000	1.000000	1900.000000	0.000000	1.000000	2009.000000
25%	279.50000	717.500000	4.000000	2006.000000	0.000000	13.000000	2011.000000
50%	558.00000	2325.000000	8.000000	2010.000000	1.000000	22.000000	2012.000000
75%	836.50000	6882.500000	10.000000	2013.000000	1.000000	37.000000	2013.000000
max	1115.00000	75860.000000	12.000000	2015.000000	1.000000	50.000000	2015.000000

Figure 3. histograms of Store dataset



2) Issues

2-a) Missing values

Figure 1 shows the number of missing values in some columns. First, the missing values in 'CompetitionDistance' appear to be incomplete records considering the definition of the variable; Each store must have the "nearest" competitor because there is no distance limitation in the definition. Therefore, we can impute the missing values with the median value corresponding to the StoreType, considering that the distribution of CompetitionDistance shows highly positive skewness in Figure 3 (skewness = 2.93). In the same context, the missing values in 'CompetitionOpenSinceMonth' and 'CompetitionOpenSinceYear' seem to be incomplete records, too. Considering each distribution, we can impute the missing values in 'CompetitionOpenSinceMonth' by its mean (skewness = -0.17), and the ones in 'CompetitionOpenSinceYear' by its median (skewness = -8.01). When a dataset shows skewed distribution, the median is better for imputation than the mean because mean values can be misleading (Steinberg, 2010, p. 73).

On the other hand, the missing values in 'Promo2SinceWeek, 'Promo2SinceYear', and 'PromoInterval' appears to be because the store was not participating in the coupon promotion, as the values are missing only when 'Promo2' = 0. We will not impute these missing values; instead, they will be dealt with by introducing three new variables – CompetitionDays, Promo2Days, and Promo2Month.

2-b) Categorical variables

As shown in Figure 3, there are four categorical variables in the Store dataset – StoreType, Assortment, Promo2, and PromoInterval. Except for Promo2 (binary), the other three variables need to be converted by one-hot encoding to make it easier to use them for prediction models.

2-c) Promo2SinceDate, CompetitionOpenSinceDate

Promo2SinceWeek and Promo2SinceYear contain one piece of information: they represent the time when Promo2 started. So, we can combine these two variables into one variable called 'Promo2SinceDate' (Datetime type) and tidy up the dataset. In the same way, we can combine CompetitionOpenSinceYear and CompetitionSinceMonth into 'CompetitionOpenSinceDate' (Datetime type). In Section 2.2.2, we will discuss how to use these variables when we merge Store and Train.

2.1.2. 'Train' dataset

1) Descriptions

The Train dataset contains 1,017,208 historical sales records with nine columns: Store, DayOfWeek, Date, Sales, Customers, Open, Promo, StateHoliday, and SchoolHoliday. Below are short descriptions of the dataset by each column, datatypes, and the number of missing values.

Figure 4. General description of Train dataset

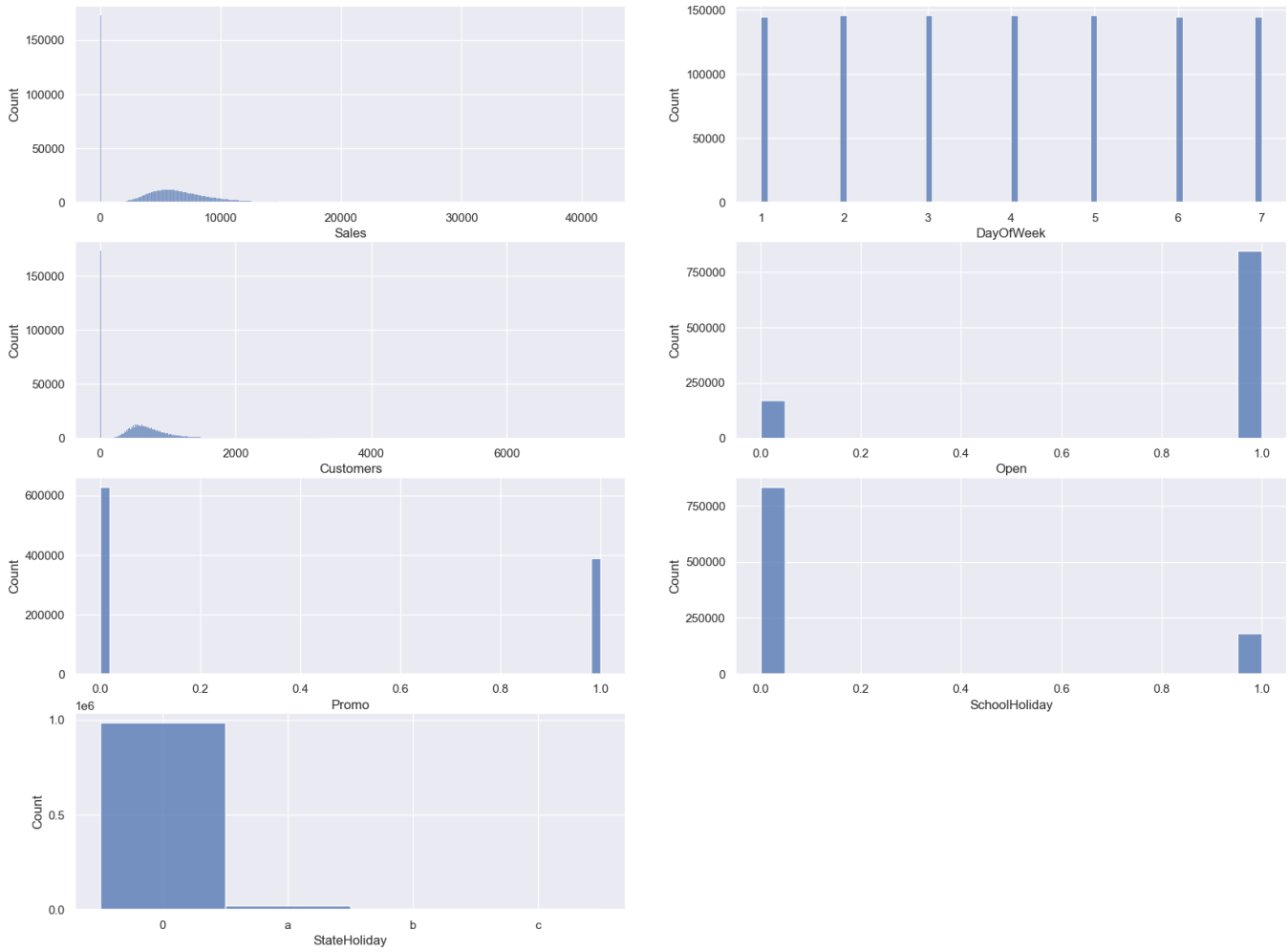
Column name	Description	Data Type	No. of missing values
Store	Store number	Int64	0
DayOfWeek	The day of the week: 1=Monday, 7=Sunday	Int64	0
Date	The date of record	Object	0
Sales	The sales record on of the day	Int64	0
Customers	The number of customers of the day	Int64	0
Open	Indicates if the store was open (1) or not (0)	Int64	0
Promo	Indicates if the store running a store-specific promo (1) or not (0)	Int64	0
StateHoliday	Indicates a state holiday: a=public, b=Easter, c=Christmas, 0=none	Object	0
SchoolHoliday	Indicates if the store was affected by the closure of public schools on that day	Int64	0

Figure 5 shows the statistical summary of the Train dataset except for Object type variables, and Figure 6 shows the distributions of all variables by histograms. It seems DaysOfWeek values are evenly distributed, while the 0 value takes up the majority of SchoolHoliday and StateHoliday. Also, there are approximately 170,000 rows with Open=0, which seems to be the same number as the rows with Customers=0 and with Sales=0. Therefore, we need to check the relationship between these rows.

Figure 5. Statistical summary of Train dataset

	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday
count	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06
mean	5.584297e+02	3.998341e+00	5.773819e+03	6.331459e+02	8.301067e-01	3.815145e-01	1.786467e-01
std	3.219087e+02	1.997391e+00	3.849926e+03	4.644117e+02	3.755392e-01	4.857586e-01	3.830564e-01
min	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.800000e+02	2.000000e+00	3.727000e+03	4.050000e+02	1.000000e+00	0.000000e+00	0.000000e+00
50%	5.580000e+02	4.000000e+00	5.744000e+03	6.090000e+02	1.000000e+00	0.000000e+00	0.000000e+00
75%	8.380000e+02	6.000000e+00	7.856000e+03	8.370000e+02	1.000000e+00	1.000000e+00	0.000000e+00
max	1.115000e+03	7.000000e+00	4.155100e+04	7.388000e+03	1.000000e+00	1.000000e+00	1.000000e+00

Figure 6. histograms of Train dataset



2) Issues

2-a) Records with ‘Open’ = 0

The Train dataset contains the sales record with 0 ‘Open’ values. When Open = 0, Sales and Customers values are always 0, too (See Appendix A). We can drop the rows with 0 Open values before using the Train dataset; The prediction model is to predict sales when the store is open.

2-b) Categorical variables

Excluding Open values, there are four categorical variables in the Train dataset – DayOfWeek, Promo, StateHoliday, and SchoolHoliday. Except for Promo and SchoolHoliday (binary), the other two variables will be converted by one-hot encoding.

2-c) ‘Customer’ variable

As discussed below, no sales or customers value is given in Test dataset. It means customer values are also the ones to be predicted along with Sales values. However, our prediction model only focuses on predicting Sales values; therefore, the Customer column will be dropped from the Train dataset.

2.1.3. ‘Test’ dataset

1) Descriptions

The Test dataset is a 41,088 x 9 dataset, which has the same structure as the Train dataset. Figure 3 shows the number of missing values in each column.

Figure 7. No. of missing values in Test dataset

Column name	No. of missing values
Store	0
DayOfWeek	0
Date	0
Sales	41088
Customers	41088
Open	0
Promo	0
StateHoliday	0
SchoolHoliday	0

2) Issues

2-a) Missing values

As shown in Figure 5, the Test dataset is missing 41,088 rows for columns ‘Sales’ and ‘Customers’. This seems to be because these two variables are the ones to be predicted. As aforementioned, the prediction model only focuses on prediction Sales values, so the Customer column will be dropped from the Test dataset.

2-b) Categorical variables

Like the Train dataset, there are four categorical variables in the Test dataset – DayOfWeek, Promo, StateHoliday, and SchoolHoliday. Therefore, except binary variables, DayOfWeek and StateHoliday will be converted by one-hot encoding.

2.2. Merging ‘Store’ and ‘Train’

2.2.1. Merging by ‘Store’ column

To train our prediction model, it seems necessary to merge Store and Train datasets by the Store column. It allows store information and past sales records to be used together to predict Sales.

2.2.2. Introducing new variables

In the merged dataset, there are three columns related to time: ‘Date’, ‘Promo2SinceDate’, and ‘CompetitionOpenSinceDate’. We can convert them into new variables that might be useful for prediction.

The first variable is ‘Promo2Days’, the number of days a store participated in Promo2 up to the record date. The second variable is ‘CompetitionDays’, which is the number of days a store was competing with another store until that date. Lastly, ‘Promo2Month’ is a new categorical variable whose value is the number of days since Promo2 coupons were issued. As the interval of Promo2 coupons is three months, Promo2Month takes values between 0 and 2.

2.2.3. Standardisation

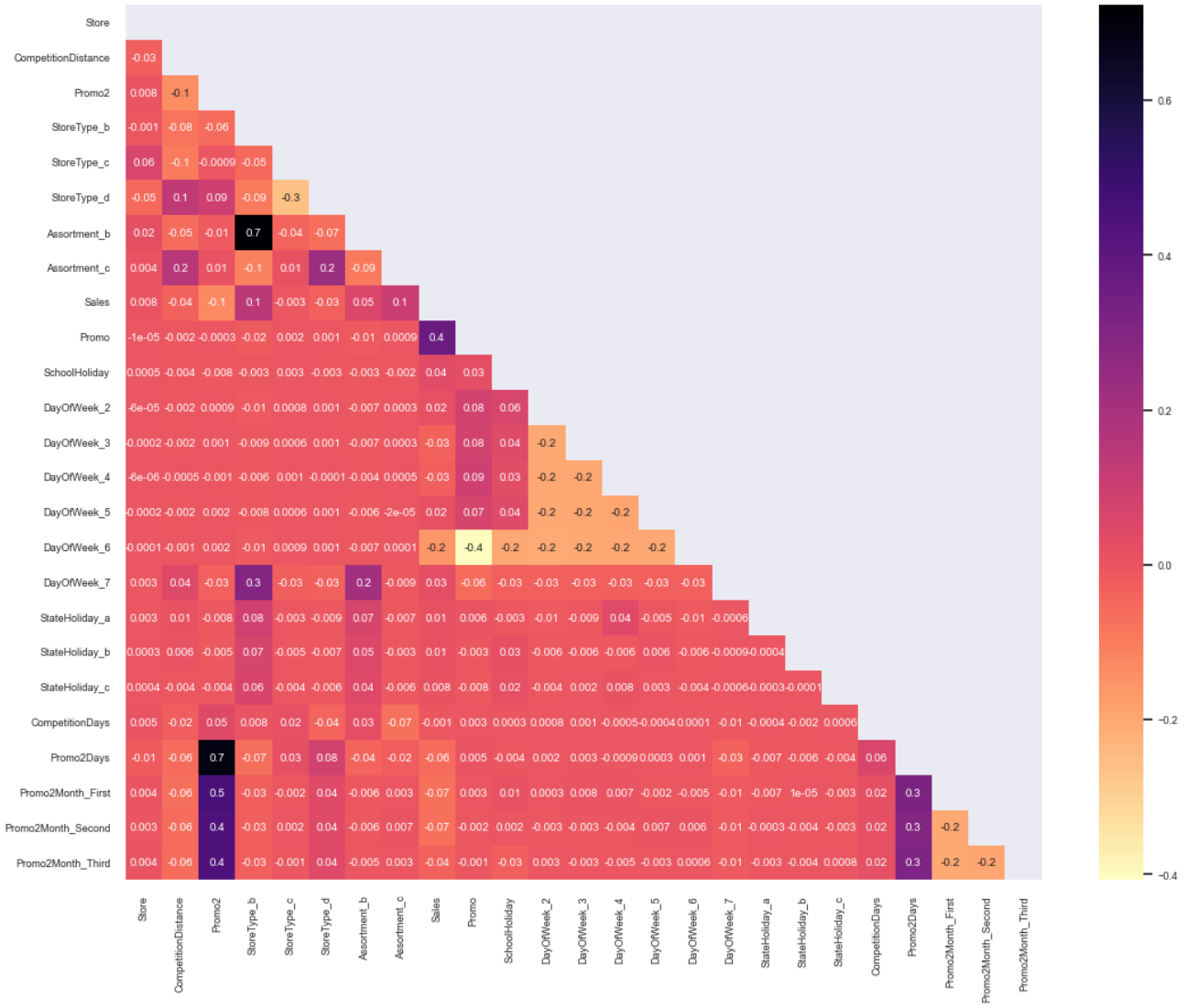
As numerical variables are in different units, we need to standardise the numerical variables in the dataset except for Store.

2.3. Multivariate Analysis

2.3.1. Pearson’s correlation coefficients between variables

We can investigate relationships between variables by calculating Pearson’s correlation coefficients. As shown in Figure 8, we can find that Promo2 and Promo2Days are highly correlated, and Promo2Days are less correlated with Sales than Promo2; therefore, we can drop Promo2Days from the dataset before model training.

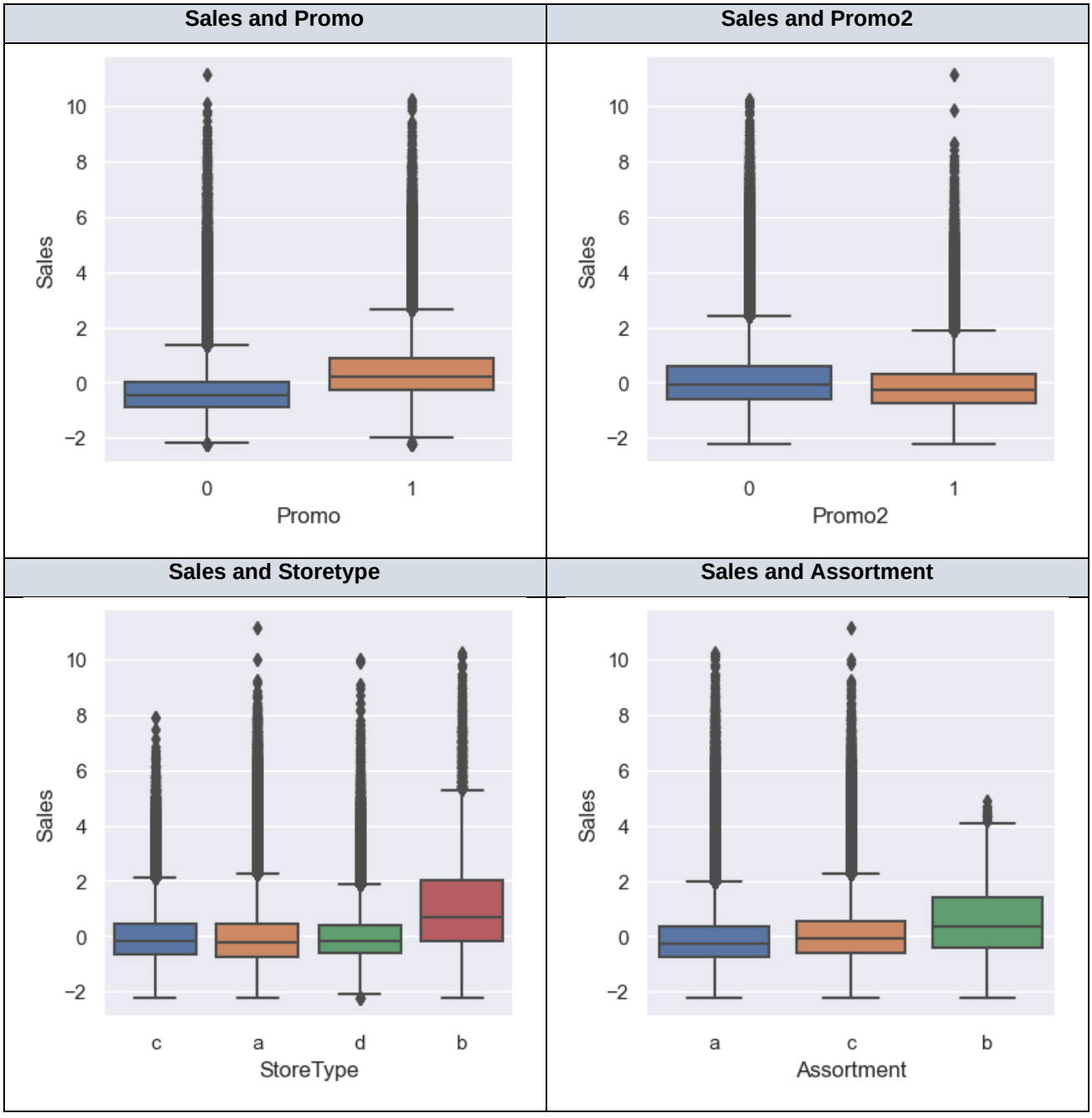
Figure 8. Pearson’s correlation coefficients between the variables

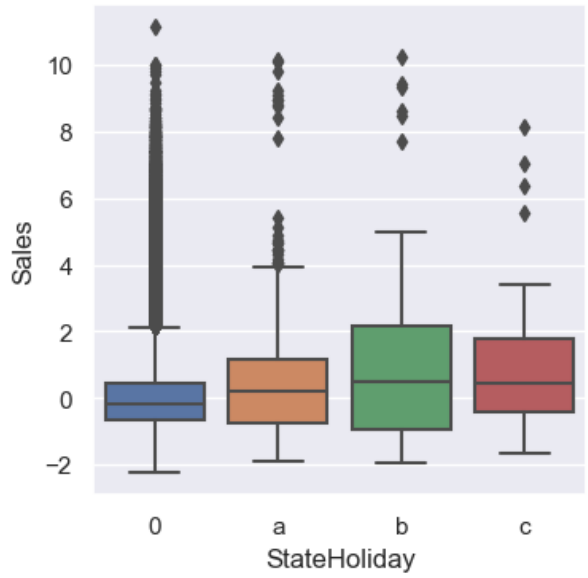
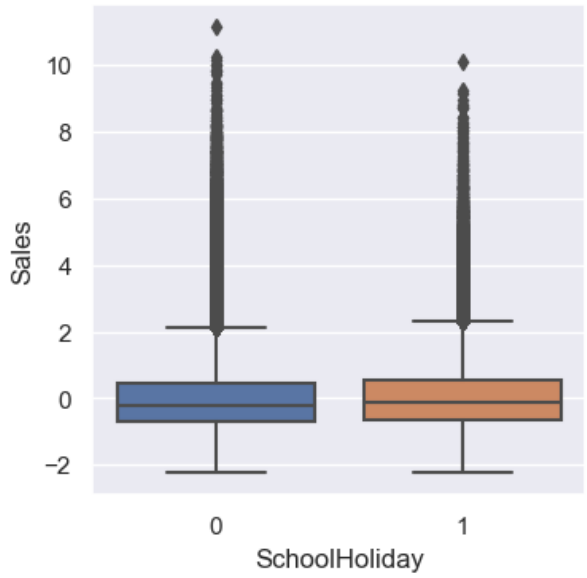
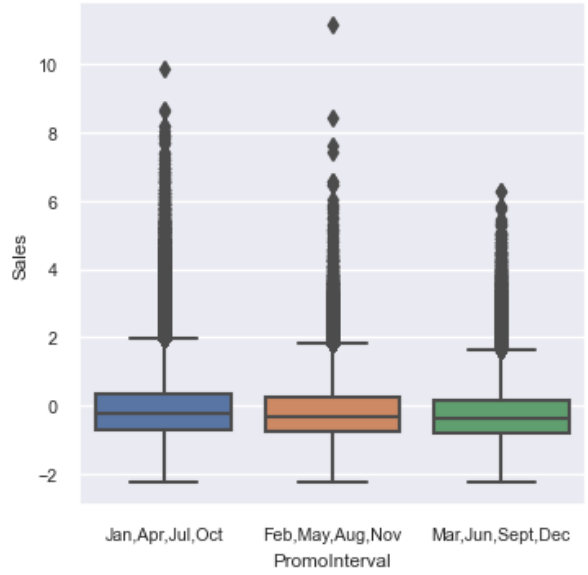
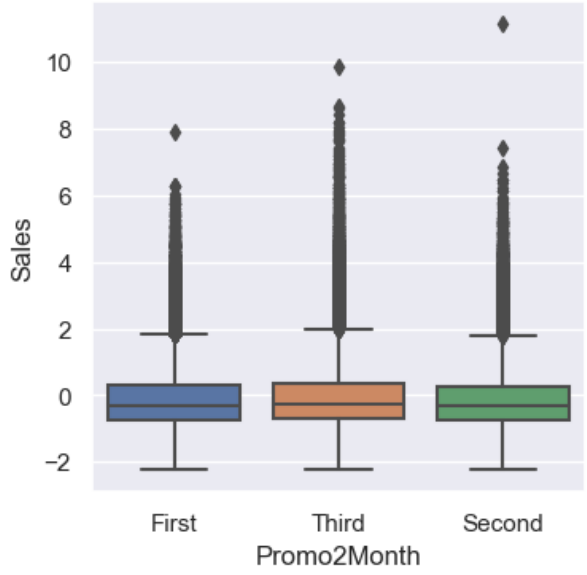
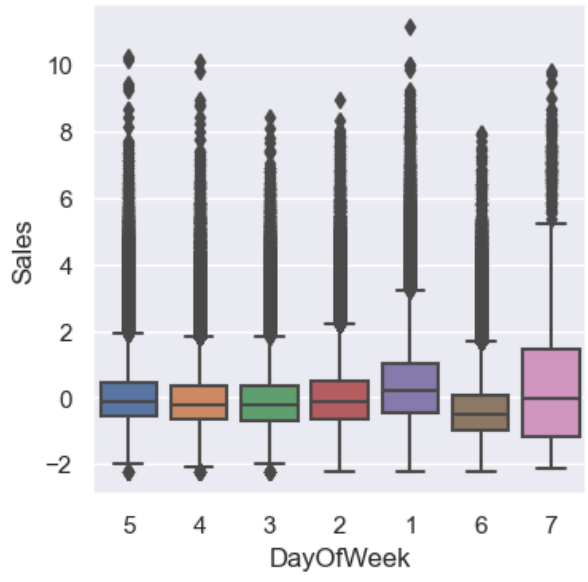


2.3.2. Relationship between Sales and categorical variables

We can investigate relationships between Sales and categorical variables by plotting box plots. As shown in Figure 9, it seems SchoolHoliday, PromoInterval, and Promo2Month do not make significant difference on Sales. Therefore, we can consider dropping these variables from the dataset for prediction.

Figure 9. Boxplots between categorical variables and Sales



Sales and StateHoliday	Sales and SchoolHoliday
	
Sales and PromoInterval	**Sales and Promo2Month**
	
Sales and DayOfWeek	
	

3. Prediction models

We can consider two types of models; Linear regression and Regression tree. Regression trees can be appropriate for handling datasets with a mixture of numerical and categorical variables. Meanwhile, linear regression models can be another option as we converted categorical variables by one-hot encoding. Linear regression can be more appropriate to clearly show the relationship between the target variable and predictors by parameters. Figure 10 shows the training dataset's statistical summary.

Figure 10. Statistical summary of dataset for training

	Store	CompetitionDistance	Promo2	StoreType_b	StoreType_c	StoreType_d	Assortment_b	Assortment_c	Sales
count	844392.000000	8.443920e+05	844392.000000	844392.000000	844392.000000	844392.000000	844392.000000	844392.000000	8.443920e+05
mean	558.422920	-8.933976e-15	0.498684	0.018431	0.133798	0.306462	0.009725	0.463376	1.649257e-17
std	321.731914	1.000001e+00	0.499999	0.134504	0.340435	0.461024	0.098136	0.498657	1.000001e+00
min	1.000000	-6.966392e-01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-2.240669e+00
25%	280.000000	-6.081792e-01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-6.753771e-01
50%	558.000000	-4.004905e-01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-1.889414e-01
75%	837.000000	1.828327e-01	1.000000	0.000000	0.000000	1.000000	0.000000	1.000000	4.524450e-01
max	1115.000000	9.026269e+00	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.114469e+01

Promo	DayOfWeek_2	DayOfWeek_3	DayOfWeek_4	DayOfWeek_5	DayOfWeek_6	DayOfWeek_7	StateHoliday_a	StateHoliday_b	StateHoliday_c
844392.000000	844392.000000	844392.000000	844392.000000	844392.000000	844392.000000	844392.000000	844392.000000	844392.000000	844392.000000
0.446352	0.170491	0.168093	0.159457	0.164189	0.170606	0.004255	0.000822	0.000172	0.000084
0.497114	0.376064	0.373949	0.366102	0.370447	0.376164	0.065092	0.028657	0.013103	0.009169
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

CompetitionDays
8.443920e+05
1.498642e-14
1.000001e+00
-9.552572e-01
-4.568179e-01
-1.682769e-01
2.792109e-01
2.242376e+01

3.1. Linear regression model

The model result is below when we train the prediction model using linear regression with 10-fold cross-validation.

$$Sales = -0.0046 - 0.059 * CompetitionDistance - 0.25 * Promo2 + 1.63 * StoreType_b - 0.033 * StoreType_C - 0.065 * StoreType_d - 0.98 * Assortment_b + 0.27 * Assortment_C + 0.74 * Promo - 0.34 * DayOfWeek_2 - 0.46 * DayOfWeek_3 - 0.46 * DayOfWeek_4 - 0.34 * DayOfWeek_5 - 0.34 * DayOfWeek_6 - 0.26 * DayOfWeek_7 + 0.068 * StateHoliday_a + 0.066 * StateHoliday_b + 0.29 * StateHoliday_c + 0.012 * CompetitionDays$$

$R^2 = 0.2194, RMSE = 0.8836$

The equation result clearly shows the relationships between predictors and Sales; However, the model performance is not desirable because R^2 is low (0.2194), and the RMSE value is high (0.8836).

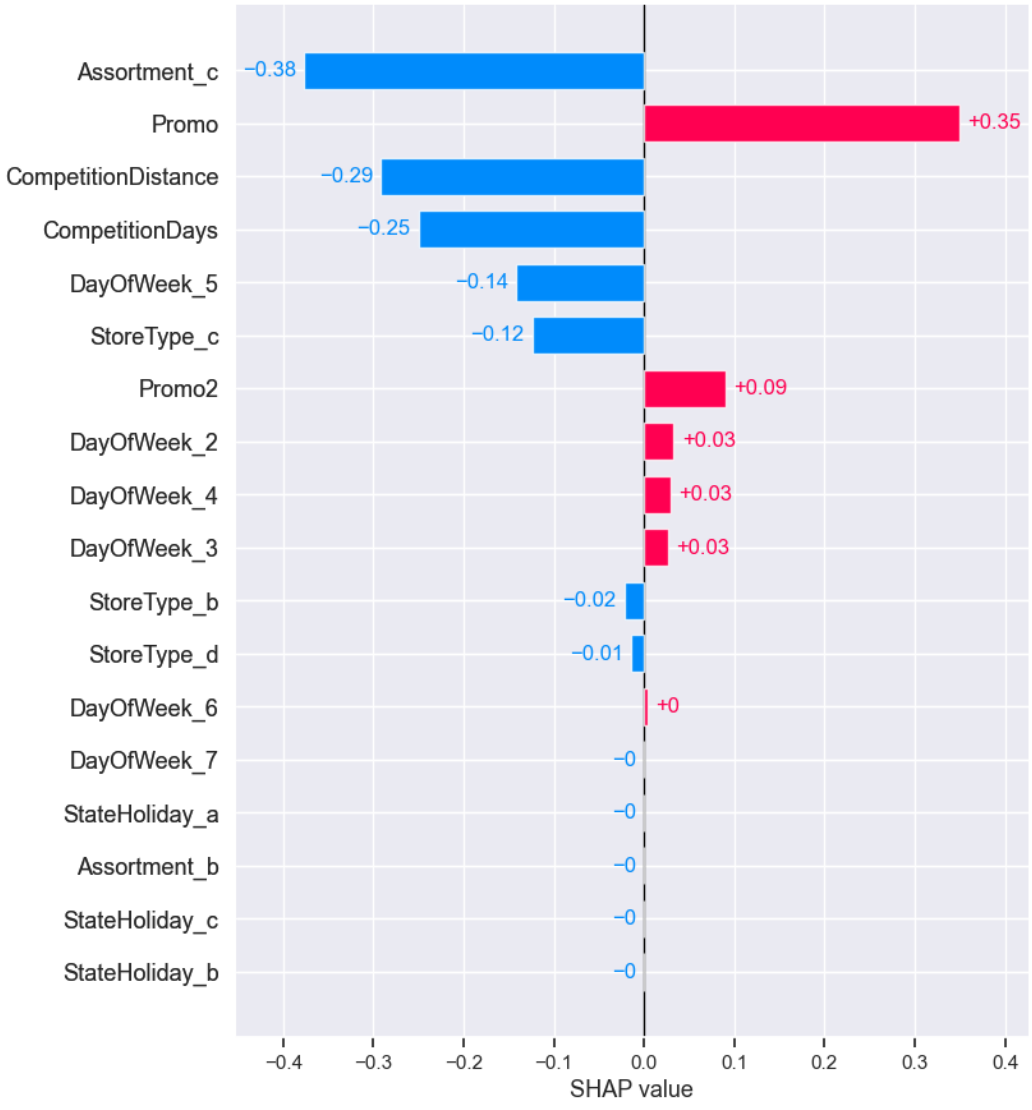
3.2. Regression tree model

Using the regression tree model, the prediction model showed better performance than the linear regression model (R^2 = 0.9056, RMSE = 0.3072). Further discussions will be made in the following Section 4.

4. Interpretation on the result of the regression tree model

Based on the result from Section 3, we can choose to use a regression tree for prediction. The shortcoming of using feature importance is that the values do not show the directions of each variable's impact on Sales. We can use SHAP (Shapley Additive exPlanations) values, which assigns feature importance values for each predictor, along with the direction of impacts (negative value for negative impact, and vice versa) (Lundberg and Lee, 2017). Figure 11 below shows the SHAP values of each variable. The values are calculated using the first 2000 rows of the training dataset because of the long computation time.

Figure 11. SHAP values of each predictor



In Figure 11, the most important factor is Assortment_c, which negatively impacts Sales. We need to investigate if there are any other significant characteristics of the stores of Assortment c type over the other types. Promotion seemed to affect Sales as intended positively, and it was the second important factor. Notably, CompetitionDistance is the third factor with a negative impact on Sales; This goes against the common thinking that sales will be less negatively affected the further away a competitor is, which requires further investigation. And the SHAP value of CompetitionDays shows that the longer the competition lasts, the greater the negative impact on sales; it also requires further research.
Meanwhile, the SHAP value of DayOfWeek_5 shows that drug sales tend to get lower on Fridays than Mondays, while the ones of DayOfWeek_2, DayOfWeek_4, and DayOfWeek_3 show that the sales tend to get higher on Tuesday, Thursday, and Wednesday than Monday. It seems necessary to investigate which characteristics related to store type c have a negative effect on sales because StoreType_c was the sixth important factor. Promo2 (issuing coupons) seems to be doing its work, but it does not have much impact like Promo. Therefore, it looks necessary to discuss how this promotion2 method can be more effective and whether it is cost-effective. StoreType_b and StoreType_d seem to have a very little negative impact on Sales. The other variables in Figure 11 have almost no impact on Sales; it is remarkable that sales are not affected by state holidays.

5. The prediction result for Test dataset

After pre-processing Test dataset, we can use our prediction model for sales prediction (See Appendix A for details).

6. Limitations of the research

This research has several limitations, some briefly mentioned in the previous sections. Such as:

- More prediction models and more different hyperparameters need to be trained for better prediction performance. Especially, there might be a more appropriate model for this study to reflect any relationship between the variables, which can be revealed by more multivariate analysis.

- This study only predicts sales using other factors, and more detailed relationships behind them require another study.

7. Conclusion

In this study, a regression tree model was made to predict sales using factors such as Assortment types, promotion status, competition status, and day of week. According to the result, it seems that each promotion method is doing its part to some extent. And assortment c type stores tend to have lower sales than other types. Also, competition statuses such as competition distance or days are another critical factor on sales. The model seems well-performing; however, it leaves the need for further discussion as to why these results came out.

REFERENCES

- Steinberg, W. (2010). *Statistics Alive!* 2nd ed. California: SAGE.
- Lundberg, S. and Lee, S. (2017). ‘A Unified Approach to Interpreting Model Predictions’, *arXiv: 1705.07875 [cs.AI],* p. 1. [Online]. Available at: https://arxiv.org/abs/1705.07874 (Accessed 5 Feb 2023].