ROLLING FORECAST OF SALES AND REVENUE DATA (ROSSMANN)

By Siddhant Singh

ACKNOWLEDMENT

Working on this project has been both difficult and fun to do it and it has been a very interesting experience for me. Through all of this, there have been several helpful people I would like to thank for their support.

My deep gratitude first goes to Prof. Adalbert Wilhelm, who expertly guided me through my graduate education and who shared the excitement of this work. His unwavering enthusiasm kept me constantly engaged with my research and his personal generosity helped make my time enjoyable.

My appreciation also extends to my supervisor, Andrej Pivcevic, whose mentoring and encouragement have been very significant. His insights have been very valuable and make up for a large part of this work. He steered me in the right direction at all times and gave me great support throughout my work process.

Furthermore, I would also like to thank my friends for giving me guidance, assistance, and sharing many laughs in the time that I spent with them who helped me a lot in finalizing this project within the limited time frame.

EXECUTIVE SUMMARY

- **Overview**

  Creating rolling financial forecasts of important financial KPI is currently hot topic in accounting and finance. However, it is difficult to create an accurate rolling forecast without automating the manual effort with machine learning techniques. Rossmann is one of the world's largest retailers, and precise sales estimates in several categories are critical to their success. In seven European nations, Rossmann runs more than 3,000 pharmacies. Promotions, competition, state and federal holidays, seasonality, and locality are just a few of the variables that affect store sales. The project forecasts store sales predictions using Rossmann's historical store sales data for different stores located in different regions leveraging various machine learning techniques.

- **Tasks/Goals**

  The aim/task of the project was to predict the store sales in advance up to a period of 6 weeks based on the dataset provided using the most suitable and efficient machine learning model.

- **Data Background**

  The dataset used for this project, 'Rossmann Store Sales', was provided by Rossmann on Kaggle.com which consists of 3 datasets namely 'train.csv', 'stores.csv', 'features.csv', & 'test.csv'.

  The 'train.csv' dataset provides information about Rossmann historical stores data including the sales data. The 'store.csv' dataset provides supplemental information about the stores. The 'test.csv' dataset provides historical data about the stores whose sales have to be predicted.

- **Approach/Methods Used**

  The approach/method used in this project to gather the results is by using Machine Learning Algorithms such as Random Forest Regression Model and LightGBM Regression Model to predict the number of sales of the Rossmann stores. The big dataset that was used in this project created various computing hurdles, forcing the approach to be modified in order to solve the problem. Identifying the correct variables on which to conduct the study was also a significant difficulty.

- **Results**

  In order to obtain the results, two machine learning models were implemented namely Random Forest Regression model and LightGBM Regression model. The result of the project was to predict the number of sales of the Rossmann stores using the dataset provided which was carried out successfully with the aforementioned Machine Learning Algorithms with one model having a slightly higher accuracy than the other one and also being the more efficient one.

REPORT

1. Introduction

Forecasts aren't just for meteorologists. Governments forecast economic growth. Scientists attempt to predict the future population. And businesses forecast product demand a common task of professional data scientists. Creating rolling financial forecasts of important financial KPI is currently hot topic in accounting and finance.[3] However, it is difficult to create an accurate rolling forecast without automating the manual effort with machine learning techniques. This project provides a detailed report on the forecasting of store sales predictions of a well-renowned multinational retail corporation called 'Rossmann'.

Rossmann is one of the world's largest retailers, and precise sales estimates in several categories are critical to their success. Because there are numerous elements that can influence sales in each department, it is critical to identify the important characteristics that influence sales and utilize them to construct a model that can help estimate sales with some accuracy.[2] This report discusses the key findings from the analysis that has been done by using Rossmann's historical store sales data for different Rossmann stores located in different regions leveraging various machine learning techniques.

The big dataset that was used in this project created various computing hurdles, forcing the approach to be modified in order to solve the problem. Identifying the correct variables on which to conduct the study was also a significant difficulty. To estimate future sales, two regression models were used in this project: Random Forest and LightGBM models. The regression models looked at a variety of aspects, starting with a comprehensive model that included all variables and working down to a reduced model that eliminated inconsequential variables.[9] Correlation plots, heatmaps, histograms, and other exploratory analysis were utilized to determine the essential variables for the regression equation.

2. Background of the Data

The dataset, 'Rossmann Store Sales', available on Kaggle was used for this project. We have weekly historical store sales data for different stores and store types for a two-year period in this dataset. We also have information particular to each store and area, such as store type, customers, Promo, school holiday, and so on. Based on these features and parameters the most efficient and accurate model is created to predict the store sales.

The dataset consists of 3 csv format files called 'train.csv', 'store.csv', & 'test.csv'.

- **'train.csv' –** This is the historical training data, which covers store sales data from 2013-01-01 to 2015-07-31.
- **'stores.csv' –** This file contains anonymized information about the different stores, indicating the type and size of the stores.
- **'features.csv' –** This file contains additional data related to store, department, and regional activity for the given dates.
- **'test.csv' –** This file is identical to 'train.csv', excluding the weekly sales. The sales for each triplet of store, department, and date in this file is being predicted.

3. Data Exploration

The subsections that follow are an attempt to analyze the dataset and identify useful attributes and parameters that can be utilized to forecast sales.

3.1 File 'train.csv'

The file 'train.csv' provides information about Rossmann historical stores data including the sales data. This csv file includes features such as, 'Store', 'DayOfWeek', 'Date', 'Sales', 'Customers', 'Open', 'Promo', 'StateHoliday', & 'SchoolHoliday', and a brief summary of how the dataset looks like is given below.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
0	1	5	2015-07-31	5263	555	1	1	0	1
1	2	5	2015-07-31	6064	625	1	1	0	1
2	3	5	2015-07-31	8314	821	1	1	0	1
3	4	5	2015-07-31	13995	1498	1	1	0	1
4	5	5	2015-07-31	4822	559	1	1	0	1
5	6	5	2015-07-31	5651	589	1	1	0	1
6	7	5	2015-07-31	15344	1414	1	1	0	1
7	8	5	2015-07-31	8492	833	1	1	0	1
8	9	5	2015-07-31	8565	687	1	1	0	1
9	10	5	2015-07-31	7185	681	1	1	0	1

Fig. 3.1: 'train.csv'

There are 1017209 rows with 9 features in the "train.csv" dataset, where 2 features are categorical (Date & StateHoliday) while the rest are numerical and contains the following fields:

- Store: a unique Id for each store
- Sales: the turnover for any given day (target variable).
- Customers: the number of customers on a given day.
- Open: an indicator for whether the store was open: 0 = closed, 1 = open.
- Promo: indicates whether a store is running a promo on that day.
- StateHoliday: indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. All schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- SchoolHoliday: indicates if the (Store, Date) was affected by the closure of public schools.

3.2 File 'store.csv'

The file 'store.csv' provides supplementary information about the stores. This csv file includes features such as, 'Store', 'StoreType', 'Assortment', 'CompetitionDistance', 'CompetitionOpenSinceYear', 'CompetitionOpenSinceMonth', 'Promo2', 'Promo2SinceYear', 'Promo2SinceWeek', & 'PromoInterval' and a brief summary of how the dataset looks like is given below.

	Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
0	1	c	a	1270.0	9.0	2008.0	0	NaN	NaN
1	2	a	a	570.0	11.0	2007.0	1	13.0	2010.0
2	3	a	a	14130.0	12.0	2006.0	1	14.0	2011.0
3	4	c	c	620.0	9.0	2009.0	0	NaN	NaN
4	5	a	a	29910.0	4.0	2015.0	0	NaN	NaN
5	6	a	a	310.0	12.0	2013.0	0	NaN	NaN
6	7	a	c	24000.0	4.0	2013.0	0	NaN	NaN
7	8	a	a	7520.0	10.0	2014.0	0	NaN	NaN
8	9	a	c	2030.0	8.0	2000.0	0	NaN	NaN
9	10	a	a	3160.0	9.0	2009.0	0	NaN	NaN

Fig. 3.2: ‘store.csv’

There are 1115 rows with 10 features in the "store.csv" dataset, where 3 features are categorical (StoreType, Assortment & PromoInterval) while the rest are numerical and contains the following fields:

- Store: a unique Id for each store
- StoreType: differentiates between 4 different store models: a, b, c, d
- Assortment: describes an assortment level: a = basic, b = extra, c = extended
- CompetitionDistance: distance in meters to the nearest competitor store
- CompetitionOpenSince[Month/Year]: gives the approximate year and month of the time the nearest competitor was opened
- Promo2: Promo2 is a continuing a promotion for some stores: 0 = store is not participating, 1 = store is participating
- Promo2Since[Year/Week]: describes the year and calendar week when the store started participating in Promo2
- PromoInterval: describes the consecutive intervals Promo2 is started, naming the months the promotion is started. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

3.3 File ‘test.csv’

The file ‘train.csv’ provides historical data about the stores whose sales have to be predicted. This csv file includes features such as, ‘Store’, ‘Id’, ‘Open’, ‘Date’, ‘Promo’, ‘StateHoliday’, & ‘SchoolHoliday’ and a brief summary of how the dataset looks like is given below.

	Id	Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday
0	1	1	4	2015-09-17	1.0	1	0	0
1	2	3	4	2015-09-17	1.0	1	0	0
2	3	7	4	2015-09-17	1.0	1	0	0
3	4	8	4	2015-09-17	1.0	1	0	0
4	5	9	4	2015-09-17	1.0	1	0	0
5	6	10	4	2015-09-17	1.0	1	0	0
6	7	11	4	2015-09-17	1.0	1	0	0
7	8	12	4	2015-09-17	1.0	1	0	0
8	9	13	4	2015-09-17	1.0	1	0	0
9	10	14	4	2015-09-17	1.0	1	0	0

Fig. 3.4: ‘test.csv’

There are 41088 rows with 8 features in the "test.csv" dataset, where 2 features are categorical (Date & StateHoliday) while the rest are numerical and contains the following fields:

- Id: an Id that represents a (Store, Date) duple within the test set
- Store: a unique Id for each store
- Open: an indicator for whether the store was open: 0 = closed, 1 = open.
- Promo: indicates whether a store is running a promo on that day.
- StateHoliday: indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. All schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- SchoolHoliday: indicates if the (Store, Date) was affected by the closure of public schools.

4. Data Preprocessing

The following subsections in the data preprocessing stage are a way to prepare and set the data for further exploration and analysis with the available datasets.

4.1 Data Quality Assessment

The following graph shows the correlation between every variable that has been used and to gain the most from the data provided.[1] These correlations were very useful in determining what were the driving forces that were affecting the number of sales for the Rossmann stores.

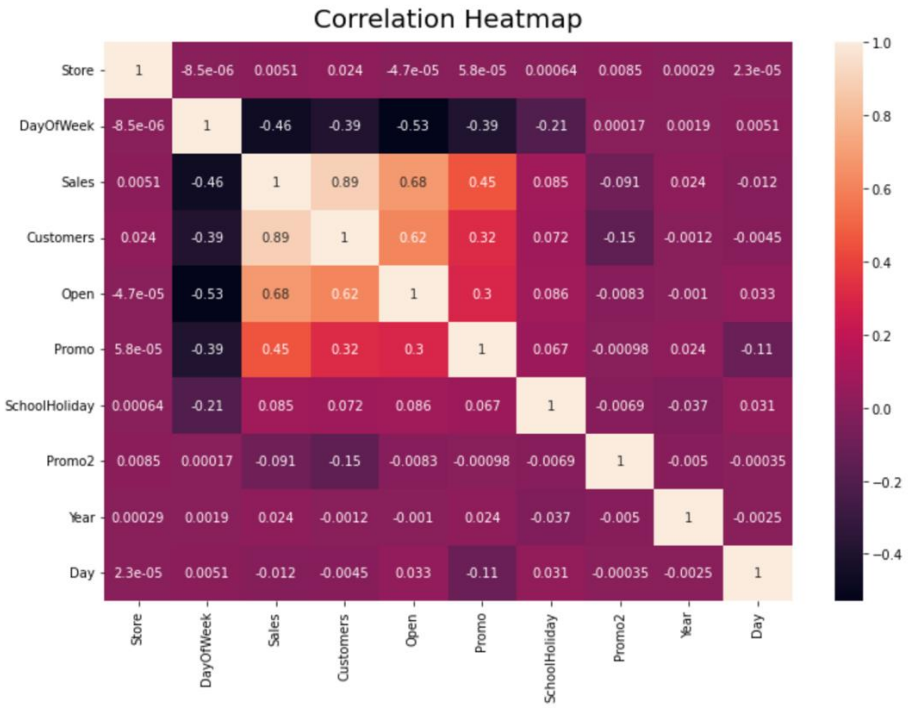

Fig. 4.1: Correlation Heatmap

4.2 Data Cleaning

One of the most important steps in the data preprocessing stage while having a large dataset that we have here in this project is to check for missing values and finding a way to deal with them.[8] As we can in the graphs below, there are no missing values in the file 'train.csv'.

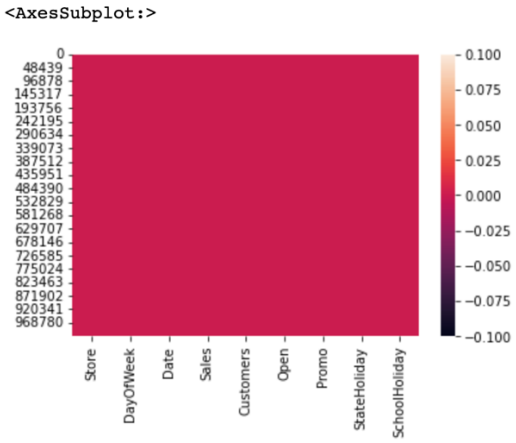

Fig. 4.2.1: NaN values in 'train.csv'

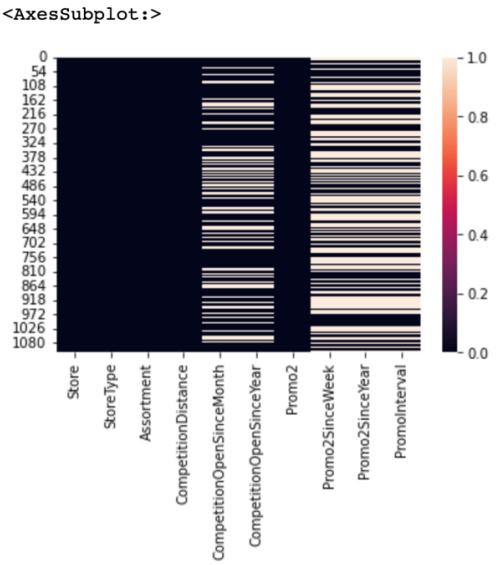

Fig. 4.2.2: NaN values in 'store.csv'

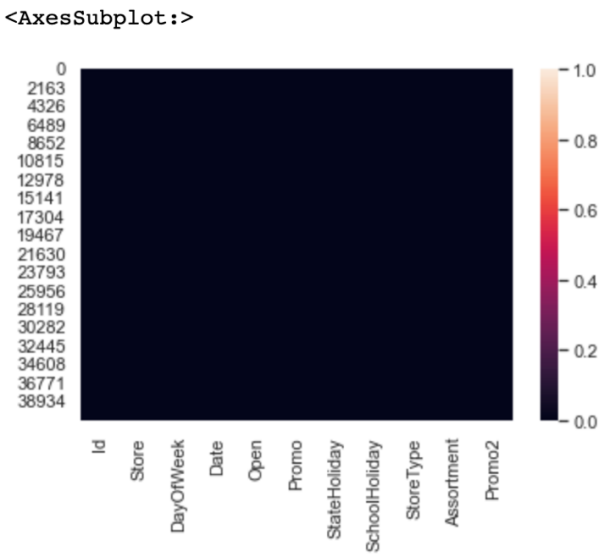

Fig. 4.2.3: NaN values in 'test.csv'

But as we can see above in Fig. 4.2.2, there are few features in the 'store.csv' file that have many missing values. The features 'CompetitionDistance', 'CompetitionOpenSinceYear', 'CompetitionOpenSinceMonth', 'Promo2', 'Promo2SinceYear', 'Promo2SinceWeek', & 'PromoInterval' have a lot a missing and even replacing the NaN values with either the mean or 0 could have an adverse effect on the accuracy and the efficiency of the model.[12] So, the best way of dealing with such problem is to drop these features altogether to achieve a higher accuracy.

The feature 'Open' in 'test.csv' file has a very few numbers of missing values. Most of the values in these two features are very similar. Therefore, all the NaN values in this feature is replaced with value '1'.[5]

4.3 Data Reduction

In order to get the most out of the data the following two steps had to be taken for the purpose of making it more useful to the model.

4.3.1 Merging the Datasets

A new dataframe 'df_merge' was created by merging three datasets i.e., 'train.csv' & 'store.csv'. As mentioned in subsection 4.2, the columns with many missing values were dropped.[15] A brief summary of the new merged dataframe is given below.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	Promo2
0	1	5	2015-07-31	5263	555	1	1	0	1	c	a	0
1	2	5	2015-07-31	6064	625	1	1	0	1	a	a	1
2	3	5	2015-07-31	8314	821	1	1	0	1	a	a	1
3	4	5	2015-07-31	13995	1498	1	1	0	1	c	c	0
4	5	5	2015-07-31	4822	559	1	1	0	1	a	a	0
5	6	5	2015-07-31	5651	589	1	1	0	1	a	a	0
6	7	5	2015-07-31	15344	1414	1	1	0	1	a	c	0
7	8	5	2015-07-31	8492	833	1	1	0	1	a	a	0
8	9	5	2015-07-31	8565	687	1	1	0	1	a	c	0
9	10	5	2015-07-31	7185	681	1	1	0	1	a	a	0

Fig. 4.3.1: Merged Dataset 'df_merge'

4.3.2 Splitting the 'Date' Column

There was a need of splitting the 'Date' column into three more columns i.e., 'Year', 'Month', & 'Day' for a better understanding & analysis of the data and then dropping the 'Date' column as it was no longer needed.[4] A brief summary of the dataframe after splitting and dropping the 'Date' column is given below.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	Promo2	Year	Month	Day	Months
0	1	5	2015-07-31	5263	555	1	1	0	1	c	a	0	2015	07	31	2015 - 07
1	2	5	2015-07-31	6064	625	1	1	0	1	a	a	1	2015	07	31	2015 - 07
2	3	5	2015-07-31	8314	821	1	1	0	1	a	a	1	2015	07	31	2015 - 07
3	4	5	2015-07-31	13995	1498	1	1	0	1	c	c	0	2015	07	31	2015 - 07
4	5	5	2015-07-31	4822	559	1	1	0	1	a	a	0	2015	07	31	2015 - 07
5	6	5	2015-07-31	5651	589	1	1	0	1	a	a	0	2015	07	31	2015 - 07
6	7	5	2015-07-31	15344	1414	1	1	0	1	a	c	0	2015	07	31	2015 - 07
7	8	5	2015-07-31	8492	833	1	1	0	1	a	a	0	2015	07	31	2015 - 07
8	9	5	2015-07-31	8565	687	1	1	0	1	a	c	0	2015	07	31	2015 - 07
9	10	5	2015-07-31	7185	681	1	1	0	1	a	a	0	2015	07	31	2015 - 07

Fig. 4.3.2: Dataset after splitting the 'Date' Column

4.4 Data Transformation

The encoding process was carried out in order to convert all the categorical data into integer format as machine learning models require numeric values to understand data. Therefore, the data from the converted categorical values was provided to the different models that are being used.[14]

In our dataset 'df_merge', we had two categorical variables ('StoreType' & 'Assortment') that needed to be encoded. Therefore, One-Hot-Encoder was used on those two variables i.e., 'StoreType' & 'Assortment' into integer format.[10] The result of the encoded data is shown below.

Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	...	Month	Day	Months	0	1	2	3	4	5
1	5	5263	555	1	1	0	1	c	a	...	07	31	2015 - 07	0.0	0.0	1.0	0.0	1.0	0.0
2	5	6064	625	1	1	0	1	a	a	...	07	31	2015 - 07	1.0	0.0	0.0	0.0	1.0	0.0
3	5	8314	821	1	1	0	1	a	a	...	07	31	2015 - 07	1.0	0.0	0.0	0.0	1.0	0.0
4	5	13995	1498	1	1	0	1	c	c	...	07	31	2015 - 07	0.0	0.0	1.0	0.0	0.0	0.0
5	5	4822	559	1	1	0	1	a	a	...	07	31	2015 - 07	1.0	0.0	0.0	0.0	1.0	0.0

Fig. 4.4: Dataset after One-Hot-Encoding

The encoded dataset 'df_merge' was further dealt with by splitting the dataset into training and test sets, with assigning 80 percent of data to the training set while the rest 20 percent to the test set.

5. Data Visualization

5.1 Sales

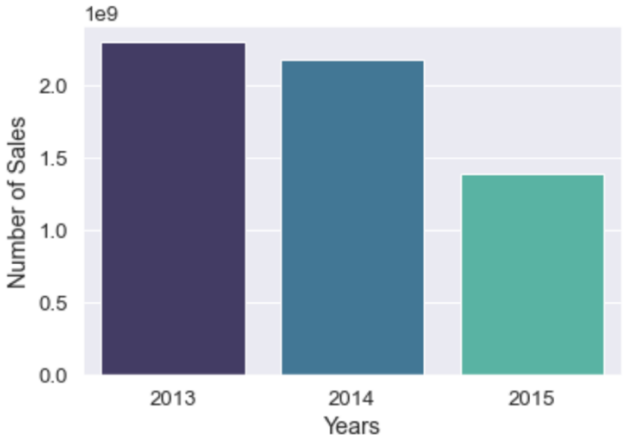

Fig. 5.1.1: No. of Sales Vs Years

Fig. 5.1.1 displays the number of sales that were made during three-year period from 2013-2015. The year 2013 had the maximum number of sales, while 2015 had the minimum.

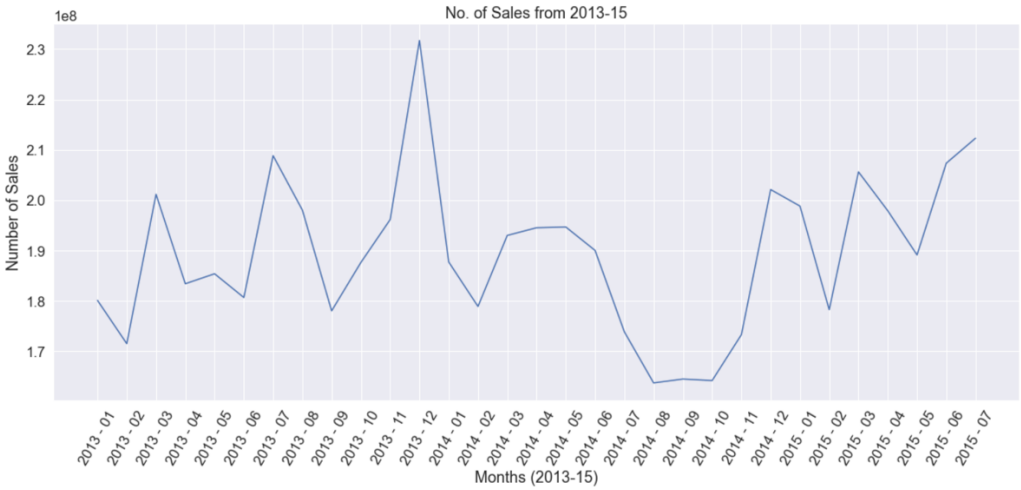

Fig. 5.1.2: No. of Sales in Months from 2013-15

Here is a graph about the number of sales that the Rossmann stores had over a period of three-years from 2013-2015. The graph depicts 3 phases. The 1st phase starts from Jan-2013 till May-2014 and shows some kind of linear trend and additive seasonality in the graph. But right after that 1st phase, from the start of the 2nd phase which is from June-2014 to Nov-2014, there is a sudden dip in the sales.

The 2st phase starts from June-2014 to Nov-2014 and also shows some kind of linear trend and additive seasonality.[7] The number of sales starts to increase from Dec-2014 to Jan-2015 right after the 2nd phase, which is quite similar what we see at the end of the 1st phase. The 3rd phase which starts from Dec-2014 and ends in July-2015 has no apparent trend but does show additive seasonality.

The graph shows a strong seasonality within each year and also shows some strong cyclic behavior with a period of about 4-5 months.

5.2 Store Types

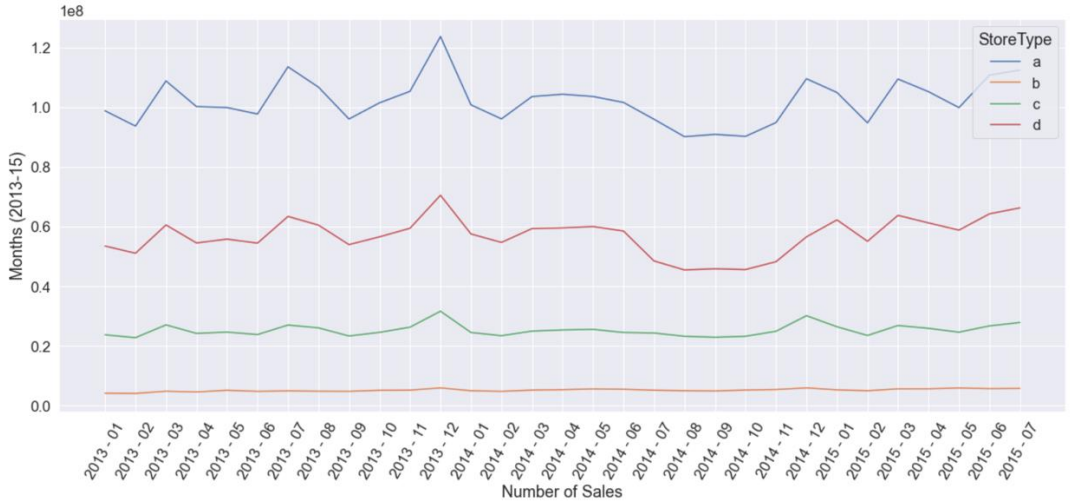

Fig. 5.2.1: No. of Sales by each Store Type

One of the major factors that was affecting the number of sales were the different store types. As we can see in the Fig. 5.2.1, it is quite evident from the graph that store type ‘a’ had the maximum number of sales while store type ‘b’ had the minimum.

This was mainly because of two reasons which are depicted on the figures below.

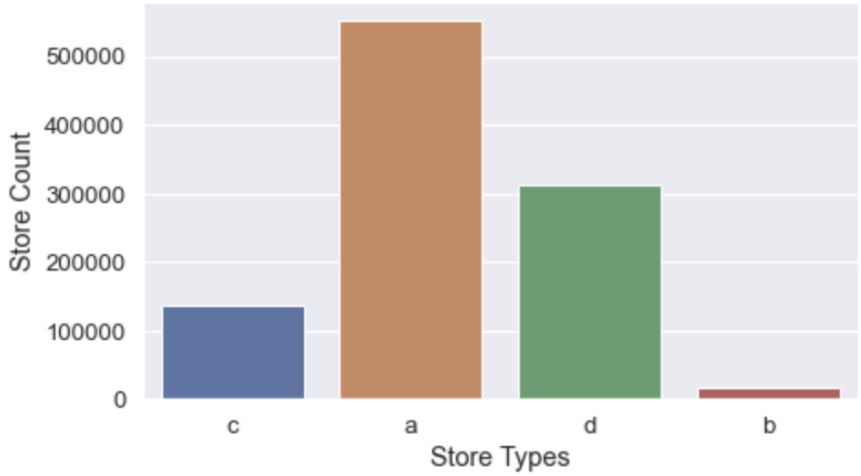

Fig. 5.2.2: No. of Store Types

The first reason behind high number of sales in store type ‘a’ can be seen in Fig. 5.2.2, the number of stores for the type ‘a’ were a lot more as compared to the other three types.[13]

Fig. 5.2.3: No. of Promos with each store types

The second reason for such high number of sales is shown in Fig. 5.2.3, where it talks about the promos of each store types. It is evident from the graph that compared to other three store types, the promos for the store type ‘a’ were a lot more which resulted in a greater number of sales. Even without the promos, store type ‘a’ had better sales numbers.

5.3 Customers

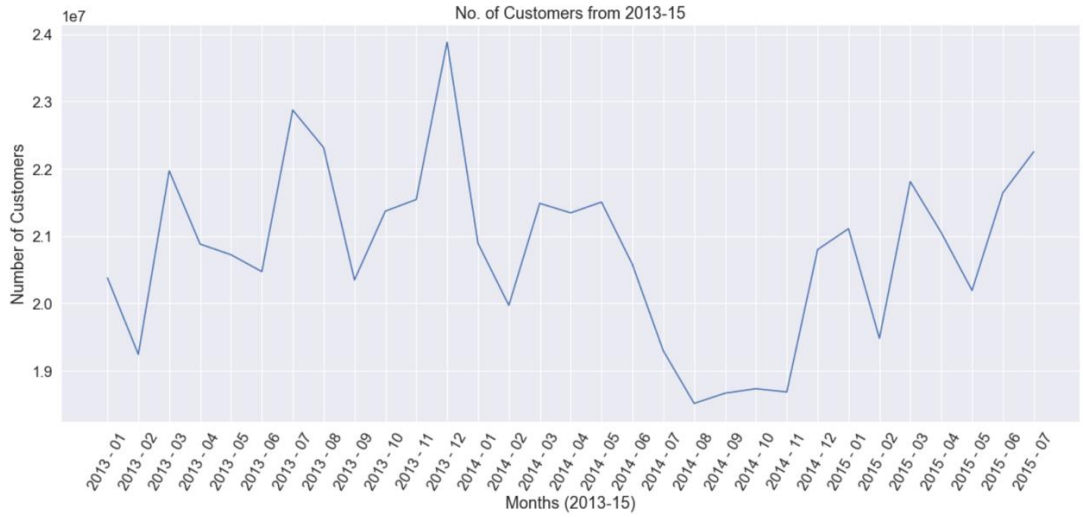

Fig. 5.3: No. of customers from 2013-15

Fig. 5.3 shows the number of customers over the course of 3 years from Jan-2013 to July-2015.[12] This graph is quite similar to the number of sales as it backs the data and shows why each phase had different sales numbers.

Fig. 5.3: No. of stores open from 2013-15

The above graph depicts the number of store that were open during 2013-15. As we can see the number of store that were open from May-2014 to Dec-2014 were less, which was the reason behind the low number of sales as well as less number of customers visiting the stores.

5.4 Store Promo

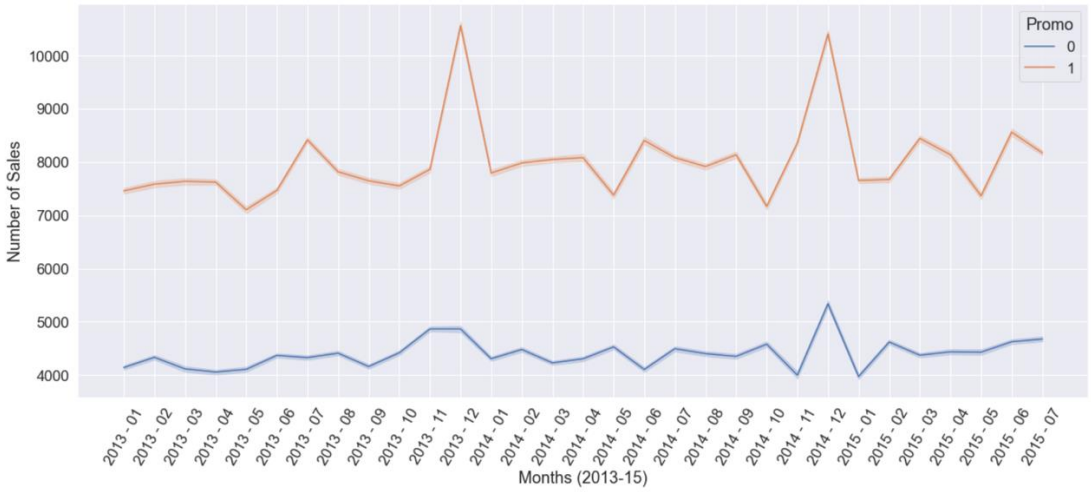

Fig. 5.4.1: No. of Promo from 2013-15

Even though the number of monthly sales were not that less but were still lower whenever there wasn't a promo available, while the sales numbers were definitely high when the stores had the promos as shown in Fig. 5.4.1.

Fig. 5.4.2: Holiday Sales for each Store Type

The above graph shows that sales were higher during school holidays as compared to the other normal days. It also shows that there was no difference in sales numbers both with and without promos during school holidays, but the numbers were higher when there wasn't a school holiday and no promo.

6. Results and Conclusions

In order to obtain the results, two machine learning models were implemented namely Random Forest Regression model and LightGBM Regression model. The following table shows how the two models performed based on the data that was provided.

MODEL	R-Squared	Mean Absolute Error	Mean Squared Error
Random Forest	0.9265484	577.8517184	1094347.20499
LightGBM	0.7554204	1323.586353	3643965.03248

Table 6.1: Model Evaluation

From the above model evaluation table, it is quite evident that Random Forest Regression Model performed better than LightGBM Regression model. To determine how the model performed, after splitting the dataset, a brief summary of the predicted sales values compared with the actual sales values of the test set is given below.

	Predicted	Actual
0	8401.7	9609.0
1	6590.9	6670.0
2	5236.7	4967.0
3	8219.2	8301.0
4	9253.7	8889.0
5	5557.8	6387.0
6	4336.8	4196.0
7	0.0	0.0
8	7461.8	7461.0
9	0.0	0.0
10	7469.6	6990.0
11	5337.7	5613.0
12	5550.8	5280.0
13	3422.2	3828.0
14	5914.0	5411.0

Fig. 6.1: Predicted Sales Values Vs Actual Sales Values

From the Fig. 6.1, it can be seen that even though there were few sales values that were not so close to the actual sales values, most of the sales values that the model predicted were quite close or similar to the actual values which is the reason that shows that the model performed decent enough.

Since the task was to predict the number of sales from the dataset ‘test.csv’ but there was no data about the actual sales values for the year 2015from this dataset, the only way to compare the predicted sales values from this dataset was to compare the values from previous years i.e., from 2013 & 2014.

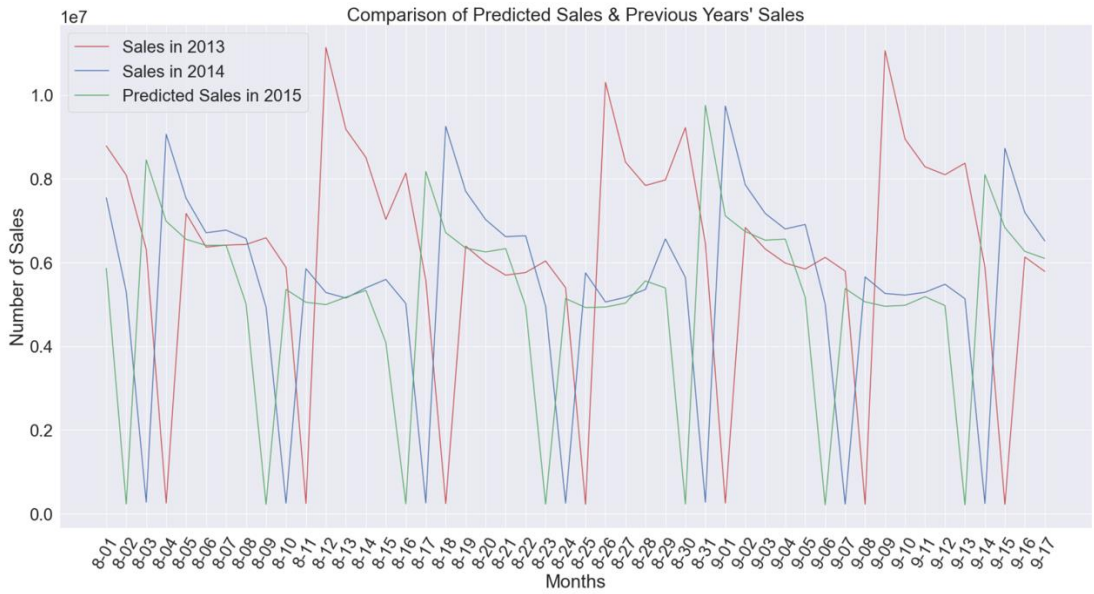

Fig. 6.2: Comparison of Predicted Sales & Previous Years' Sales

The above graph shows the comparison of the predicted sales values in the year 2015 and sales numbers in the years 2013 & 2014.[6] The predicted sales values were very similar to the previous years' sales values that means if there would have been data about the actual sales values from the 'test.csv' dataset, it would have been very close.

The conclusions that can be drawn from all the insights and inferences that the data provided are as follows:

- There were a lot of missing values in some attributes, which was the reason why these columns were dropped. If there was more data and information regarding these features then more insights could have been gathered regarding the store sales.
- The reason store type 'a' had the maximum number of sales was because there were a lot a greater number of stores for type 'a' in comparison to the other three store types i.e., 'b', 'c', & 'd'. This means if the number of stores were increased for types 'b', 'c', & 'd' then, it might also have an impact and can increase the sales values in these store types.
- One of the reasons behind the sudden dip in sales right after phase 1 & 2 was because there was a sudden decrease in the number of stores been opened during those periods which also resulted in a decrease in the number of customers. One way of countering this problem is to also open store sales for online marketing which could have a massive impact on the sales numbers.

REFERENCES

[1]. Predicting Store Sales — Random Forest Regression. Kelvin Prawtama.

[2]. Simple Guide on using Supervised Learning Model to forecast for Time-Series Data by Sue Lynn.

[3]. Simulation Based Sales Forecasting on Retail Small Stores. Hai Rong Lv Xin Xin Bai Wen Jun Yin Jin Dong.

[4]. Prediction of retail sales of footwear using feed forward and recurrent neural networks. Prasun Das Æ Subhasis Chaudhury.

[5]. Chen, C.-Y., Lee, W.-I., Kuo, H.-M., Chen, C.-W., & Chen, K.-H. (2010). The study of a forecasting sales model for fresh food. Expert Systems with Applications, 37(12), 7696–7702.

[6]. Chang, P.-C., Liu, C.-H., & Lai, R. K. (2008). A fuzzy case-based reasoning model for sales forecasting in print circuit board industries. Expert Systems with Applications, 34(3), 2049–2058.

[7]. Forecasting retailer product sales in the presence of structural change. Tao Huang, Robert Flides, Didier Sooparmanien.

[8]. A Comparative Analysis of Weekly Sales Forecasting Using Regression Techniques. Gopal Behera, Ashutosh Bhoi, Ashok Kumar Bhoi

[9]. A sales forecasting model for consumer products based on the influence of online word-of-Mouth.Ching-Chin Chern,Chih-Ping Wei,Fang-Yi Shen,Yu-Neng Fan.

[10]. Simulation Based Sales Forecasting on Retail Small Stores. Hai Rong Lv Xin Xin Bai Wen Jun Yin Jin Dong.

[11]. A Thorough Guide to Time Series Analysis. Fangyi Yu.

[12]. Building a sales prediction model for a retail store. Pablo Martin, Marina Castano and Roberto Lopez, Artelnics.

[13] Predicting sales using Machine Learning Techniques. Sai Nikhil Boyapati, Ramesh Mummidi.

[14] Forecasting Key Retail Performance Indicators Using Interpretable Regression. Belisario Panay, Nelson Baloian, Jose A. Pino, Sergio Penafiel, Jonathan Frez, Horacio Sanson, Gustavo Zurita.

[15] Sales-forecasting of Retail Stores using Machine Learning Techniques. Akshay Krishna, Akhilesh V, Animikh Aich, Chetana Hedge.