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. Walmart is one of the world's largest retailers, and precise sales estimates in several categories are critical to their success. The project forecasts store sales predictions using Walmart historical store sales data for 45 Walmart 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 10-12 months based on the dataset provided using the most suitable and efficient machine learning model.

- **Data Background**

  The dataset used for this project, 'Walmart Store Sales Forecasting', was provided by Walmart on Kaggle.com which consists of 4 datasets namely ‘train.csv’, ‘stores.csv’, ‘features.csv’, & ‘test.csv’.

  The ‘train.csv’ dataset provides information about Walmart historical stores data including the sales data. The ‘stores.csv’ & ‘features.csv’ datasets provide 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 Walmart 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 Walmart 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.

DETAILED 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 American multinational retail corporation called 'Walmart'.

Walmart 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 Walmart's historical store sales data for 45 Walmart 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, 'Walmart Store Sales Forecasting', available on Kaggle was used for this project. We have weekly historical store sales data for 45 stores and 99 departments for a three year period in this dataset. We also have information particular to each store and area, such as store size, unemployment rate, temperature, promotional markdowns, 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 4 csv format files called 'train.csv', 'stores.csv', 'features.csv', & 'test.csv'.

- **'train.csv' –** This is the historical training data, which covers store sales data from 2010-02-05 to 2012-11-01.
- **'stores.csv' –** This file contains anonymized information about the 45 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 Walmart historical stores data including the sales data. This csv file includes features such as, 'Store', 'Dept', 'Date', 'Weekly_Sales', & 'IsHoliday', and a brief summary of how the dataset looks like is given below.

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False
5	1	1	2010-03-12	21043.39	False
6	1	1	2010-03-19	22136.64	False
7	1	1	2010-03-26	26229.21	False
8	1	1	2010-04-02	57258.43	False
9	1	1	2010-04-09	42960.91	False

Fig. 3.1: 'train.csv'

There are 421570 rows with 5 features in this file and contains the following fields:

- Store: the store number
- Dept: the department number
- Date: the week
- Weekly_Sales: sales for the given department in the given store
- IsHoliday: whether the week is a special holiday week

3.2 File 'stores.csv'

The file 'stores.csv' provides supplementary information about the stores. This csv file includes features such as, 'Store', 'Type', & 'Size' and a brief summary of how the dataset looks like is given below.

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875
5	6	A	202505
6	7	B	70713
7	8	A	155078
8	9	B	125833
9	10	B	126512

Fig. 3.2: 'stores.csv'

There are 45 rows with 3 features in this file and contains the following fields:

- Store: the store number
- Type: the store type A or B
- Size: the size of the store

3.3 File 'features.csv'

The file 'train.csv' also provides supplementary information about the stores. This csv file includes features such as, 'Store', 'Date', 'Temperature', 'Fuel_Price', 'MarkDown1-5', 'CPI', 'Unemployment', & 'IsHoliday' and a brief summary of how the dataset looks like is given below.

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False
5	1	2010-03-12	57.79	2.667	NaN	NaN	NaN	NaN	NaN	211.380643	8.106	False
6	1	2010-03-19	54.58	2.720	NaN	NaN	NaN	NaN	NaN	211.215635	8.106	False
7	1	2010-03-26	51.45	2.732	NaN	NaN	NaN	NaN	NaN	211.018042	8.106	False
8	1	2010-04-02	62.27	2.719	NaN	NaN	NaN	NaN	NaN	210.820450	7.808	False
9	1	2010-04-09	65.86	2.770	NaN	NaN	NaN	NaN	NaN	210.622857	7.808	False

Fig. 3.3: 'features.csv'

There are 8190 rows with 12 features in this file and contains the following fields:

- Store: the store number
- Date: the week
- Temperature: average temperature in the region
- Fuel_Price: cost of fuel in the region
- MarkDown 1-5: anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
- CPI: the consumer price index
- Unemployment: the unemployment rate
- IsHoliday: whether the week is a special holiday week

3.4 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', 'Dept', 'Date', & 'IsHoliday' and a brief summary of how the dataset looks like is given below.

	Store	Dept	Date	IsHoliday
0	1	1	2012-11-02	False
1	1	1	2012-11-09	False
2	1	1	2012-11-16	False
3	1	1	2012-11-23	True
4	1	1	2012-11-30	False
5	1	1	2012-12-07	False
6	1	1	2012-12-14	False
7	1	1	2012-12-21	False
8	1	1	2012-12-28	True
9	1	1	2013-01-04	False

Fig. 3.4: 'test.csv'

There are 115064 rows with 4 features in this file and contains the following fields:

- Store: the store number
- Dept: the department number
- Date: the week
- IsHoliday: whether the week is a special holiday week

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 Walmart stores.

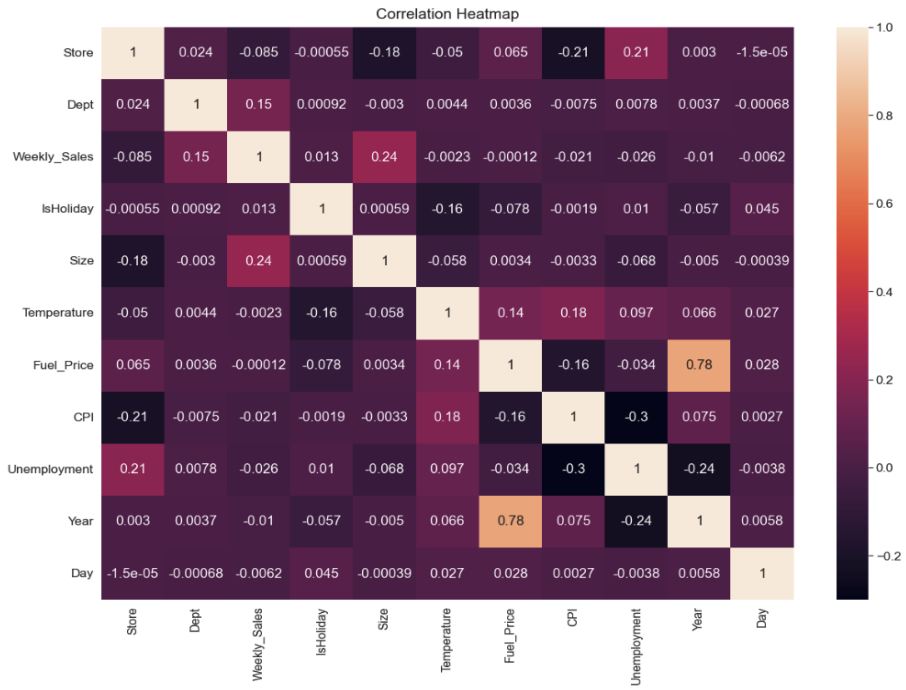

Fig. 4.1: Correlation Heatmap

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 files 'train.csv', 'stores.csv' & 'test.csv.

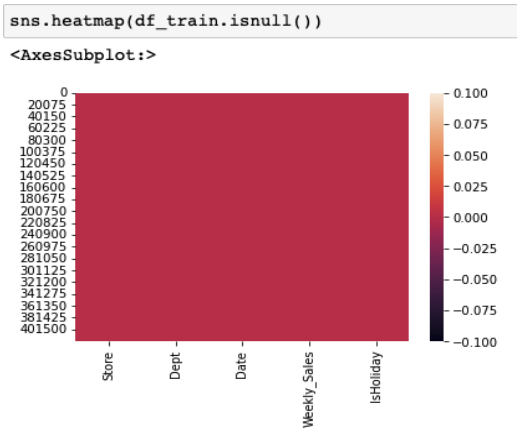

Fig. 4.2.1: NaN values in 'train.csv'

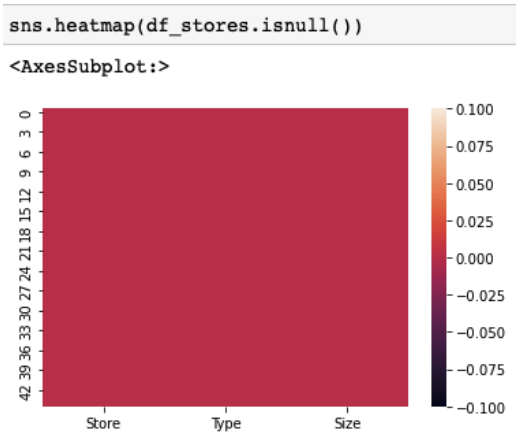

Fig. 4.2.2: NaN values in 'stores.csv'

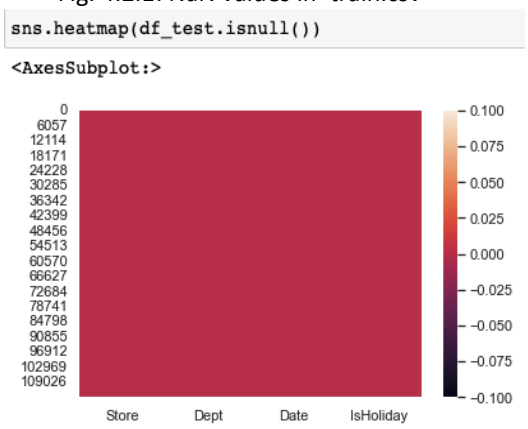

Fig. 4.2.3: NaN values in 'test.csv'

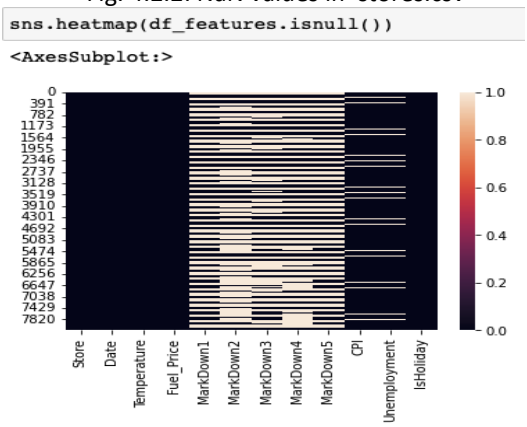

Fig. 4.2.4: NaN values in 'features.csv'

But as we can see above in Fig. 4.2.4, there are few features in the 'features.csv' file that have many missing values. The features MarkDown1, MarkDown2, MarkDown3, MarkDown4, & MarkDown5 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 features 'CPI' & 'Unemployment' have a very few numbers of missing values. Most of the values in these two features are very similar. Therefore, the mean of all the values is taken to replace the NaN values in these features.[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', 'stores.csv', & 'features.csv'. As mentioned in subsection 4.2, the columns with many missing values were dropped and the NaN values in columns 'CPI' & 'Unemployment' were replaced with their mean values.[15] A brief summary of the new merged dataframe is given below.

	Store	Dept	Date	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	CPI	Unemployment
0	1	1	2010-02-05	24924.50	False	A	151315	42.31	2.572	211.096358	8.106
1	1	1	2010-02-12	46039.49	True	A	151315	38.51	2.548	211.242170	8.106
2	1	1	2010-02-19	41595.55	False	A	151315	39.93	2.514	211.289143	8.106
3	1	1	2010-02-26	19403.54	False	A	151315	46.63	2.561	211.319643	8.106
4	1	1	2010-03-05	21827.90	False	A	151315	46.50	2.625	211.350143	8.106
5	1	1	2010-03-12	21043.39	False	A	151315	57.79	2.667	211.380643	8.106
6	1	1	2010-03-19	22136.64	False	A	151315	54.58	2.720	211.215635	8.106
7	1	1	2010-03-26	26229.21	False	A	151315	51.45	2.732	211.018042	8.106
8	1	1	2010-04-02	57258.43	False	A	151315	62.27	2.719	210.820450	7.808
9	1	1	2010-04-09	42960.91	False	A	151315	65.86	2.770	210.622857	7.808

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	Dept	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	CPI	Unemployment	Year	Month	Day
0	1	1	24924.50	False	A	151315	42.31	2.572	211.096358	8.106	2010	2	5
1	1	1	46039.49	True	A	151315	38.51	2.548	211.242170	8.106	2010	2	12
2	1	1	41595.55	False	A	151315	39.93	2.514	211.289143	8.106	2010	2	19
3	1	1	19403.54	False	A	151315	46.63	2.561	211.319643	8.106	2010	2	26
4	1	1	21827.90	False	A	151315	46.50	2.625	211.350143	8.106	2010	3	5
5	1	1	21043.39	False	A	151315	57.79	2.667	211.380643	8.106	2010	3	12
6	1	1	22136.64	False	A	151315	54.58	2.720	211.215635	8.106	2010	3	19
7	1	1	26229.21	False	A	151315	51.45	2.732	211.018042	8.106	2010	3	26
8	1	1	57258.43	False	A	151315	62.27	2.719	210.820450	7.808	2010	4	2
9	1	1	42960.91	False	A	151315	65.86	2.770	210.622857	7.808	2010	4	9

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 one categorical variable ('Type') & one Boolean variable ('IsHoliday') that needed to be encoded. Therefore, One-Hot-Encoder was used on those two variables i.e., 'Type' & 'IsHoliday' into integer format.[10] The result of the encoded data is shown below.

	Weekly_Sales	Store	Dept	Size	Temperature	Fuel_Price	CPI	Unemployment	Year	Month	Day	0	1	2	3	4
0	24924.50	1	1	151315	42.31	2.572	211.096358	8.106	2010	2	5	1.0	0.0	1.0	0.0	0.0
1	46039.49	1	1	151315	38.51	2.548	211.242170	8.106	2010	2	12	0.0	1.0	1.0	0.0	0.0
2	41595.55	1	1	151315	39.93	2.514	211.289143	8.106	2010	2	19	1.0	0.0	1.0	0.0	0.0
3	19403.54	1	1	151315	46.63	2.561	211.319643	8.106	2010	2	26	1.0	0.0	1.0	0.0	0.0
4	21827.90	1	1	151315	46.50	2.625	211.350143	8.106	2010	3	5	1.0	0.0	1.0	0.0	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

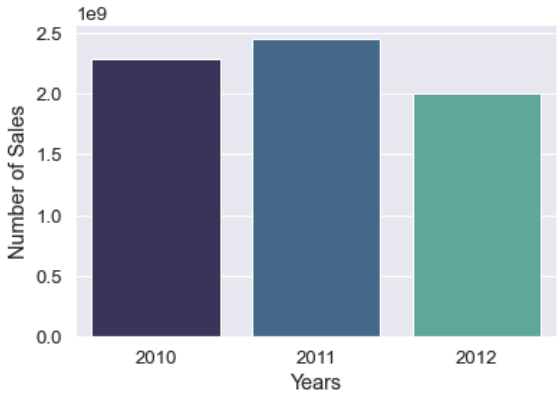

Fig. 5.1.1: No. of Sales Vs Years

Fig. 5.1.1 displays the number of sales that were made during three-year period from 2010-2012. The year 2011 had the maximum number of sales, while 2012 had the minimum.

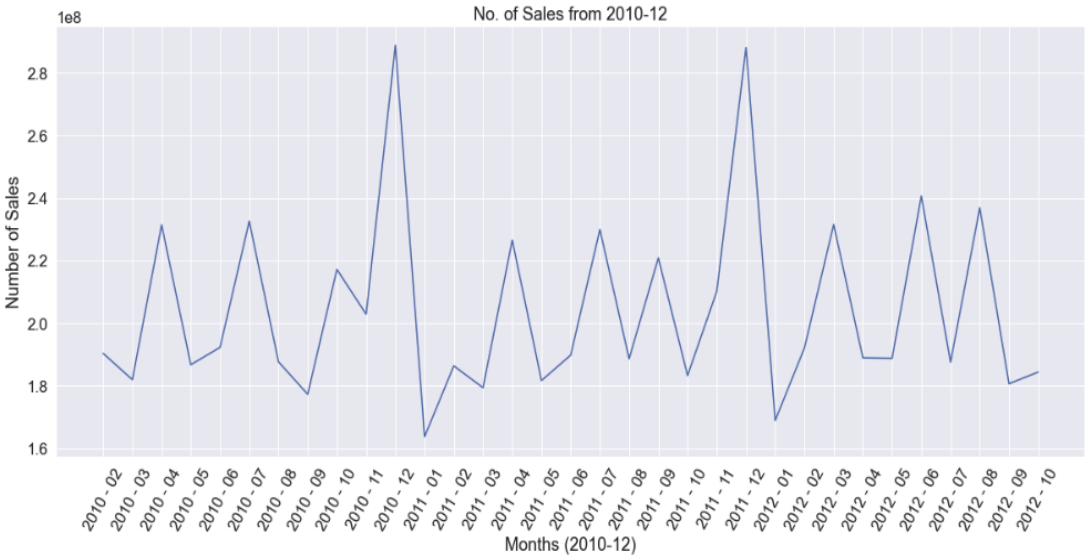

Fig. 5.1.2: No. of Sales in Months from 2010-12

Here is a graph about the number of sales that the Walmart stores had over a period of three-years from 2010-2012. The graph depicts 3 phases. The 1st phase starts from Feb-2010 till Dec-2010 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 Jan-2011 to Dec-2011, there is a sudden dip in the sales from Dec-2010 to Jan-2011.

The 2st phase starts from Jan-2011 to Dec-2011 and also shows some kind of linear trend and additive seasonality.[7] The number of sales decreases from Dec-2011 to Jan-2012 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 Jan-2012 and ends in Oct-2012 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 10-11 months.

5.2 Store Types

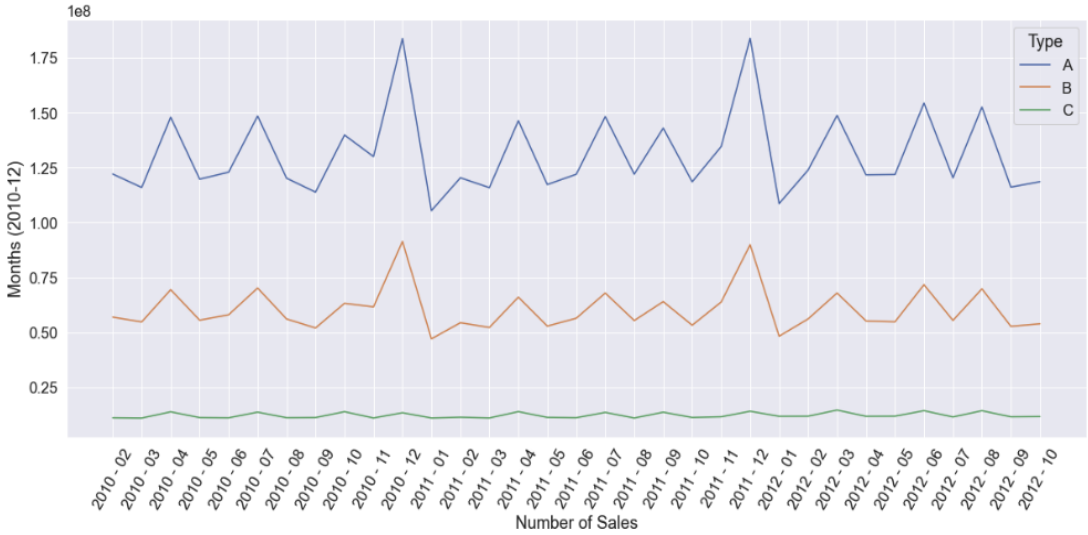

Fig. 5.2.1: No. of Sales by each Store Type

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 ‘C’ had the minimum.

This was mainly because of two reasons which is depicted on the bar graphs below.

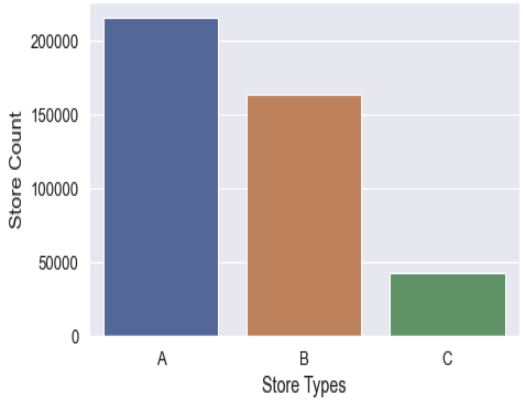

Fig. 5.2.2: No. of Store Types

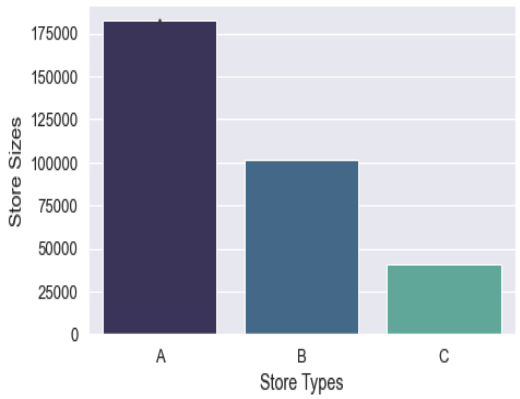

Fig. 5.2.3: Store Size for each Store Type

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 two types.[13] The second reason for such high number of sales is shown in Fig. 5.2.3, where it tells. About the sizes of each store types. It is evident from the graph that compared to other two store types, the store sizes for the type ‘A’ were larger.

5.3 Temperature

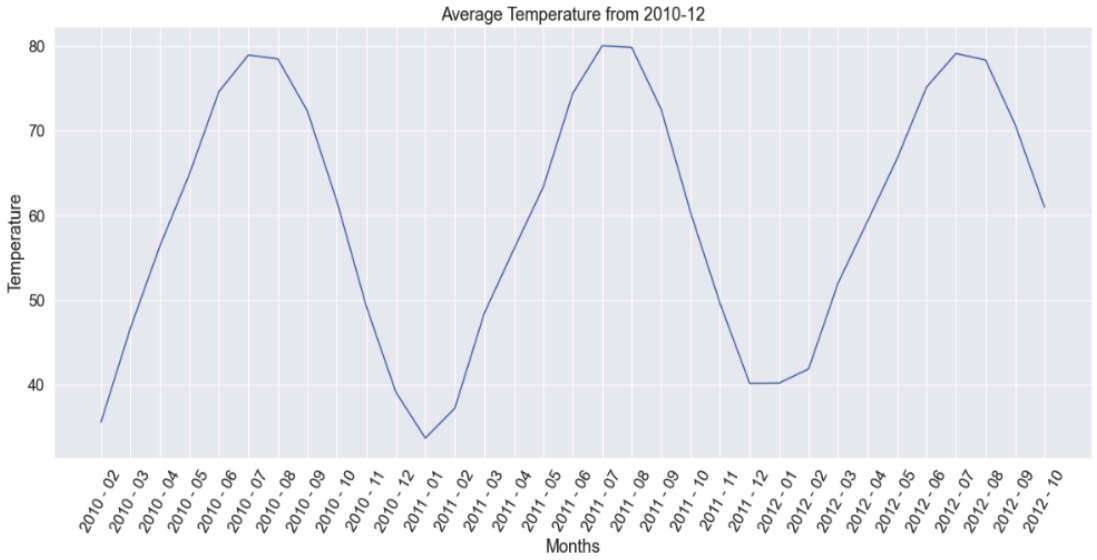

Fig. 5.3: Average Temperature from 2010-12

Fig. 5.3 shows the average temperatures in Fahrenheit over the course of 3 years from Feb-2010 to Oct-2012.[12] Average temperatures also played a key role in determining the number of sales each month as when the sales numbers dipped after phase 1 and phase 2, as depicted in the graph of Fig. 5.1, during both of these periods the average temperatures were close to 40-degree or below 40-degree Fahrenheit.

5.4 Holiday

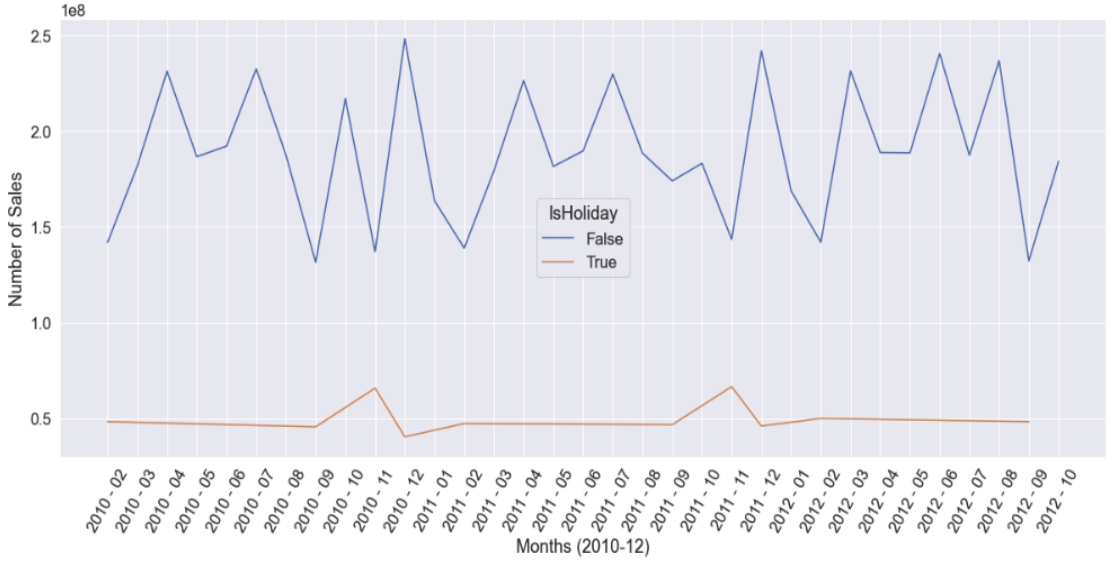

Fig. 5.4.1: No. of Sales with/without Holidays

Even though the number of monthly sales were not zero but were quite low whenever there was a holiday, while the sales numbers were definitely high there weren't any holidays as shown in Fig. 5.4.1.

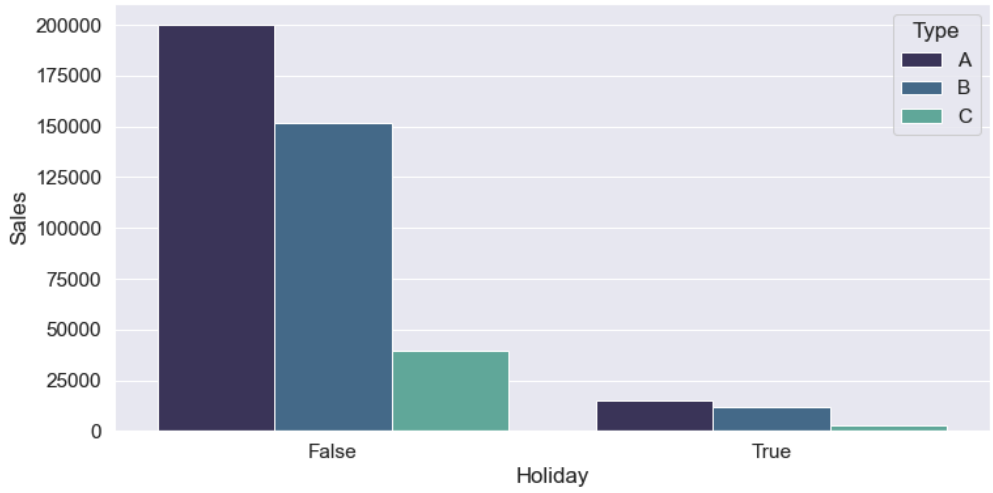

Fig. 5.4.2: Holiday Sales for each Store Type

The bar graph shows that sales for store type 'A' were higher as compared to the other two types no matter whether there was a holiday or not, while store type 'C' had the minimum number of sales and were close to zero for store type 'C' when there was a holiday.

5.5 Unemployment Rate & Fuel Prices

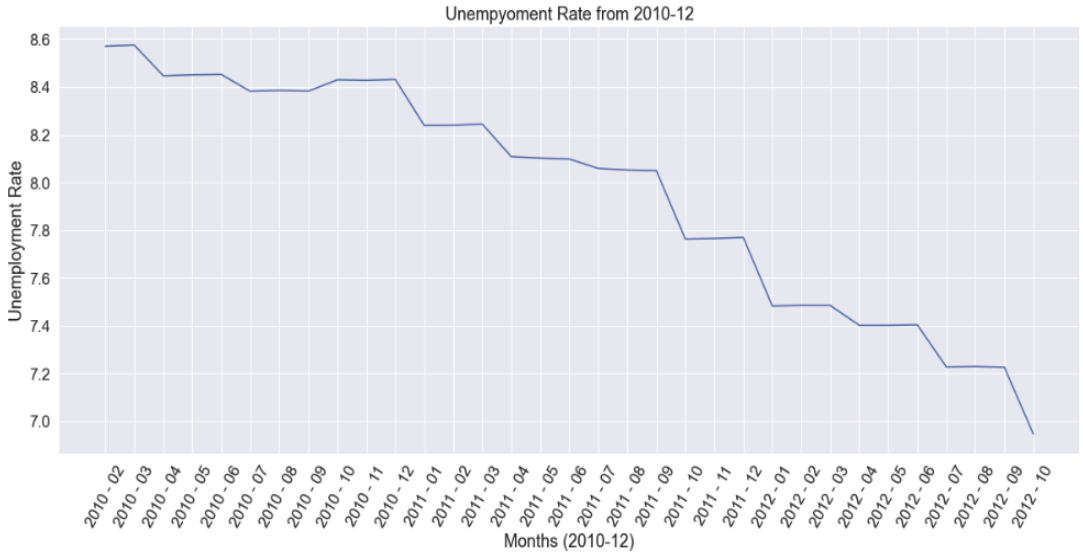

Fig. 5.5.1: Unemployment Rate from 2010-12

As it is quite evident from the Fig. 5.5.1, there appears to be a strong downward decreasing trend with strong seasonality in the unemployment rates during the three-year period from Feb-2010 to Oct-2012, but there is no presence of any kind of cyclic behavior.

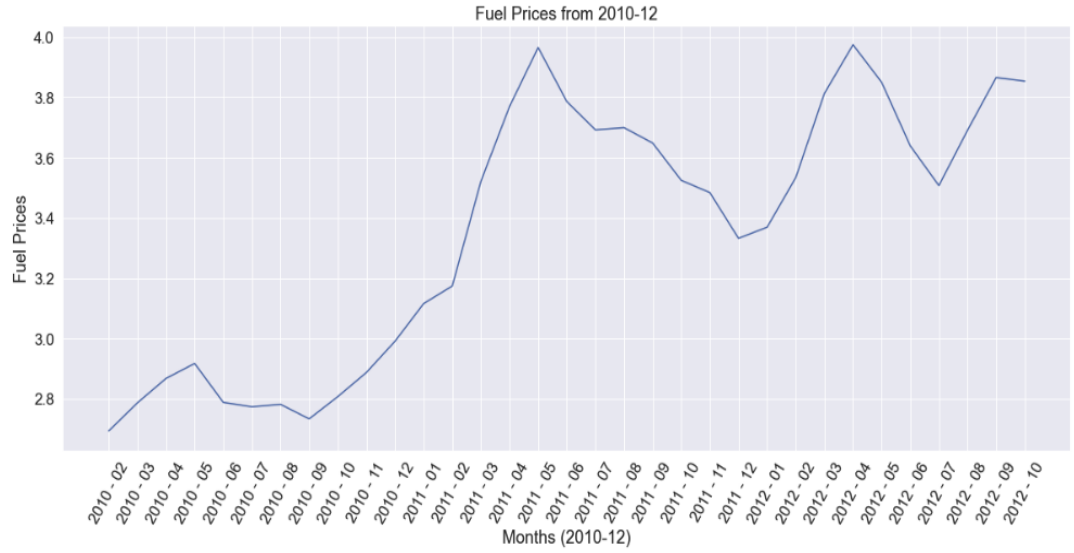

Fig. 5.5.2: Fuel Prices from 2010-12

As shown in Fig. 5.5.2, there was a sudden increase in the fuel prices from Sep-2010 where the prices were as low as $2.7 to May-2011 where they reached the highest with prices reaching close to $4. The prices seem to decrease a bit for a period of 8-10 months but then again when back up to $4 in Apr-2012.

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	0.9755639	1437.3701145	12648828.5941
LightGBM	0.9138105	4039.2787773	44614246.2539

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	321.089	202.76
1	13438.797	16482.00
2	47401.372	48167.29
3	27510.487	21581.64
4	1342.500	1315.00
5	8317.958	8683.26
6	52244.084	36639.21
7	5632.514	5631.81
8	7716.443	8159.89
9	6126.000	2420.00
10	4504.033	5194.21
11	54.404	28.92
12	25484.023	24604.93
13	524.823	452.22
14	7500.045	9405.37

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 from the year 2012-13 from 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 2010-12

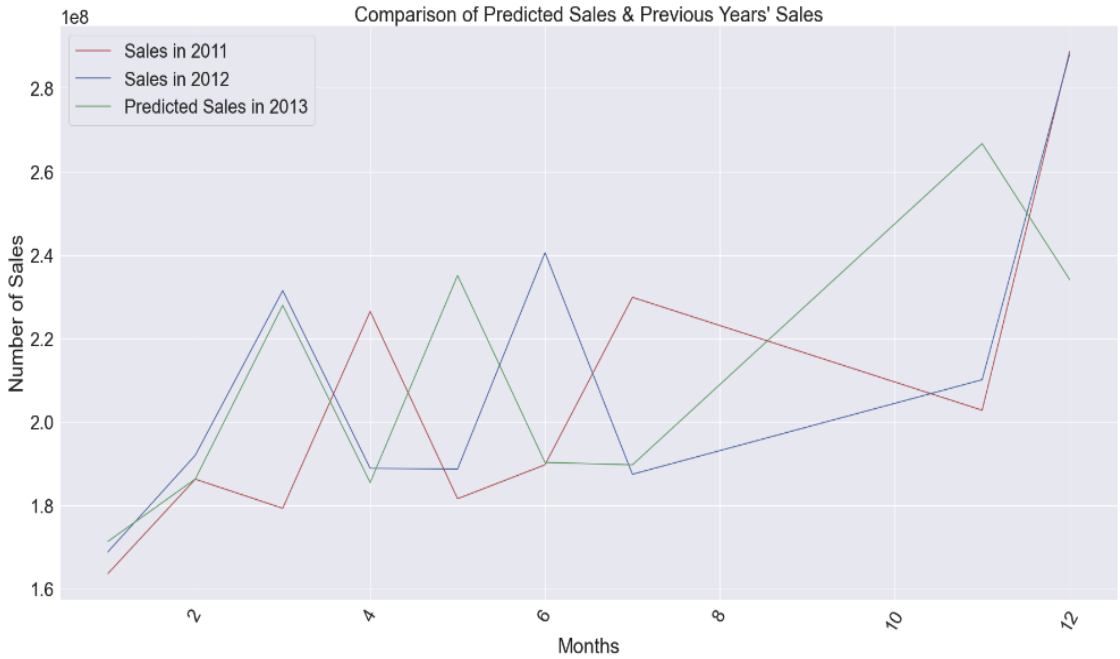

Fig. 6.2: Comparison of Predicted Sales & Previous Years' Sales

The above graph shows the comparison of the predicted sales values in the year 2013 and sales numbers in the years 2011 & 2012.[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 the attributes 'MarkDown1-5', which was the reason why these columns were dropped. If there was more data and information regarding the Markdowns 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' and the sizes of this type was also bigger in comparison to the other two store types i.e., 'B' & 'C'. This means if the sizes and the number of stores were increased for types 'B' & 'C' 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 average temperatures during those periods. 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.
- Two factors that were not affecting the sales that much were the unemployment rate and the fuel prices. As it was seen even though the unemployment rate went down and the fuel prices increased, they did not affect the sales numbers massively.

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.