Retail Demand Forecasting using Machine Learning Algorithms

Mehmet Nuri Bolat

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Technology
in
Information Technology

Eastern Mediterranean University
August 2023
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Technology in Information Technology.

Asst. Prof. Dr. Ece Çelik
Director, School of Computing and Technology

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Technology in Information Technology.

Asst. Prof. Dr. Hüsnü Bayramoğlu
Supervisor

Examining Committee

1. Assoc. Prof. Dr. Nazife Dimililer

2. Assoc. Prof. Dr. Kamil Yurtkan

3. Asst. Prof. Dr. Hüsnü Bayramoğlu

ABSTRACT

Understanding how to forecast a product's sales and demand is crucial for businesses that sell goods. Knowing how much demand will be in a given time gives them many benefits and gains. Many methods have been developed and used for demand forecasting from past to present. If we divide the methods used into two, traditional and machine learning methods are used for demand forecasting. We can say that traditional methods have left their place to machine learning due to less and slow data processing. Machine learning methods have the ability to process a lot of data faster and analyze the data it uses and provide a more accurate prediction by identifying hidden patterns in the data. The problem here is that there is no one "one-size-fits-all" prediction algorithm.

Typically, demand forecasting features consist of several machine learning approaches. Therefore, the choice of machine learning models depends on many factors such as business goal, data type, data quantity and quality, forecast time. Therefore, the main problem here is to determine which algorithm will be used with which parameters.

In this study, different machine learning methods and parameters was used and compared to select the most suitable machine learning algorithm and parameters according to the selected data set and provide more accurate predictions. Algorithms such as time series, linear regression, random forest was studied and external factors such as seasonal, regional and economic factors was used as parameters. The

algorithm with the best results will be chosen from models with or without external factors.

Keywords: Machine Learning, Demand Forecasting, Regression, Time Series

ÖZ

Ürün satan şirketlerin belirli bir ürünün satışını ve talebini öngörülebilmesi çok önemli bir yere sahip olmaktadır. Belirli bir sürede ne kadar talep olacağını bilmek onlara birçok fayda ve kazanç sağlamaktadır. Geçmişten günümüze talep tahmini için birçok yöntem geliştirilmiş ve kullanılmıştır. Kullanılan yöntemleri ikiye ayıracak olursak geleneksel ve makine öğrenmesi yöntemleri talep tahminleri için kullanılmaktadır. Geleneksel yöntemler daha az ve yavaş veri işlemesinden dolayı yerini makine öğrenimine bırakmıştır diyebiliriz. Makine öğrenimi yöntemleri çok fazla veriyi daha hızlı işleme özelliğine sahip olmaktadır ve kullanmış olduğu verileri analiz eder ve verilerdeki gizli kalıpları tanımlayarak daha doğru bir tahmin sağlar. Buradaki sorun, "herkese uyan" tek bir tahmin algoritmasının olmamasıdır.

Genellikle, talep tahmini özellikleri birkaç makine öğrenimi yaklaşımından oluşur. Bu nedenle makine öğrenimi modellerinin seçimi iş hedefi, veri türü, veri miktarı ve kalitesi, tahmin süresi gibi birçok faktöre bağlıdır. Dolayısıyla burada asıl sorun hangi algoritmanın hangi parametrelerle kullanılacağını belirlemektir.

Bu çalışmada seçilecek veri setine göre en uygun makine öğrenmesi algoritmasını ve parametrelerini seçmek ve daha doğru tahminler sunmak için farklı makine öğrenimi yöntemleri ve parametreleri kullanılıp karşılaştırılacaktır. Zaman serisi, doğrusal regresyon, rastgele orman gibi algoritmalar üzerinde çalışılacak ve mevsimsel, bölgesel, ekonomik faktörler gibi dış etkenler parametre olarak kullanılacaktır. En iyi sonuç veren algoritma ise dış faktörlerin dahil olmuş veya olmamış modellerden seçilecektir.

Anahtar Kelimeler: Makine Öğrenimi, Talep Tahmini, Regresyon, Zaman Serisi

DEDICATION

To my family.

ACKNOWLEDGEMENT

First of all, I would like to thank my family for their financial and moral support throughout my undergraduate and graduate education.

I would like to express my sincere and warm gratitude to Assist. Prof. Dr. Hüsnü Bayramoğlu for his continuous support in my graduate research, for his excellent and positive advice in the planning and development of this research, and for his generous time spent.

Likewise, I would like to thank my dear teacher Assoc. Prof. Dr. Nazife Dimililer, who gave her feedback and ideas during my research.

In addition, I would also like to thank the Managers of Trendbox, Real-Time Data Analysis firm in Istanbul/Turkey. They have supported me in exchanging ideas on machine learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARIMA	Autoregressive Integrated Moving Average
CPI	Consumer Price Index
KNN	K-Nearest Neighbors
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
R2	R- Squared
RMSE	Root Mean Squared Error
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous Factors
SVM	Support Vector Machines
WMAE	Weighted Mean Absolute Error
XGB	Extreme Gradient Boosting

Chapter 1

INTRODUCTION

The most precise and error-free method of estimating future demand for a service or product is called demand forecasting [1].

From past to present, companies selling products have tried to forecast the sales and demand of a particular product. Knowing how much demand will be for any product for a given period gives them many benefits and gains. In general, there are two key situations that companies want to avoid: understocking and overstocking. When a product's supply falls short of demand, this is known as understocking. Understocking occurs when a consumer comes to your store to purchase a product but realizes that it is out of stock. Overstocking, also known as stock surplus, occurs when a corporation orders or manufactures more products than are required. Because the products are expired or old, companies may be unable to use or sell them later. These problems lead to bigger problems as companies cannot control their inventories. The company may suffer a large amount of loss or even lead to bankruptcy with the frequent recurrence of the problem [2].

As mentioned above, overstocking and understocking are the most important problems of companies. These two problems have harmful consequences for the company. Customers may become frustrated and dissatisfied when a product is unavailable, and if it happens frequently, they may even lose interest in that

particular product. On the other hand, unnecessary costs are incurred for holding products that are not for sale. In a study conducted by Emir Zunic, other consequences about cost are mentioned. Because of lost sales, there are cost of stopping production, replanting, switching to other products, breaking deadlines, returning to production of the original product and related costs [3].

There are many benefits for companies if demand forecasts are done correctly. Generally, reliable sales forecasting benefit firms in a variety of ways, including improving corporate strategy, lowering expenses, and increasing profits [4]. Companies increase their sales by ensuring product availability and providing campaigns, and products are less spoiled by optimizing their stocks. Accurate forecasts enable the company's activities, such as manufacturing, finance, research and development, purchasing, and marketing, to be targeted appropriately and help them meet their goals [5]. One of the most important things in wholesale and distribution companies is stock optimization. With successful forecasts, companies can keep their stocks at a sufficient level, save money, and improve other processes such as warehousing, shipping and commissioning [3].

We can divide the methods developed for demand forecasting into two as Traditional methods and Machine Learning Methods. Traditional methods are divided into two as quantitative and qualitative. Expert opinion, market research, historical analogy, delphi method, end-use method are the examples of the qualitative forecasting methods. Moving average, exponential smoothing, adaptive smoothing, graphical methods are the example methods of the quantitative forecasting methods [1]. Many businesses today use artificial intelligence to complete a variety of activities, even ones that people cannot complete as quickly as a machine. Machine learning has

been utilized by businesses like Amazon and Alibaba to estimate product demand and manage inventory. There are many machine learning methods used for demand forecasting. We will review these methods in the next sections.

1.1 Problem Definition

With changing market dynamics and consumer demand, using some methods, retailers try to understand and predict the demand of a product or service to optimize supply decisions by corporate supply chain and business management. However, there are some problems which should be considered. The first problem here is the inadequacy of traditional methods such as Survey Method, Collective Opinion Method, Market Experiment Method and Traditional Statistical Methods. Generally, traditional methods process less data, have less data processing speed and sometimes, their forecasts are not accurate [6,7]. A more contemporary method is to use Machine Learning Algorithms for retail demand forecasting.

Machine learning is an algorithm or model that learns patterns using existing data and predicts similar patterns for new data. Compared to traditional demand forecasting methods, machine learning accelerates data processing speed, provides a more accurate forecast, automates forecast updates based on recent data, analyzes more data, identifies hidden patterns in data, creates a robust system and increases adaptability to changes [6, 7]. The issue in here is that there are no “one-size-fits-all” forecasting algorithm. Often, demand forecasting features consist of several machine learning approaches. Therefore, the choice of machine learning models depends on several factors, such as business goal, data type, data amount and quality, forecasting period, etc. So, the main problem here is to determine which algorithm to be used with which parameters.

1.2 Objectives

The main purpose of this study is to present the method that will provide the most accurate estimation according to the data set to be selected. To determine this, regression and time series machine learning algorithms were used in this study. The features to be used in algorithms significantly affect the accuracy of the model. Exploratory data analysis and feature selection techniques were applied to identify the correct features.

In this study, sales data of Walmart, one of the world's largest companies and a retail sales firm, was used. Weekly sales forecasts was made on the basis of stores and departments by using data from 45 different stores for the years 2010, 2011 and 2012. In the data set, store, department information, sales quantities, holiday weeks and factors that can affect sales; temperature, fuel price, promotional discounts, consumer price index and unemployment rate information. The steps to follow in the proposed approach are understanding data, exploratory data analysis, preprocessing of data, selection of machine learning models, model training, validation and observation of test results.

1.3 Structure of the Thesis

The conceptual framework diagram of this thesis project is given below. The remaining sections in this thesis was organized as follows; Literature Review, Methodology, Results and Findings and Conclusion.

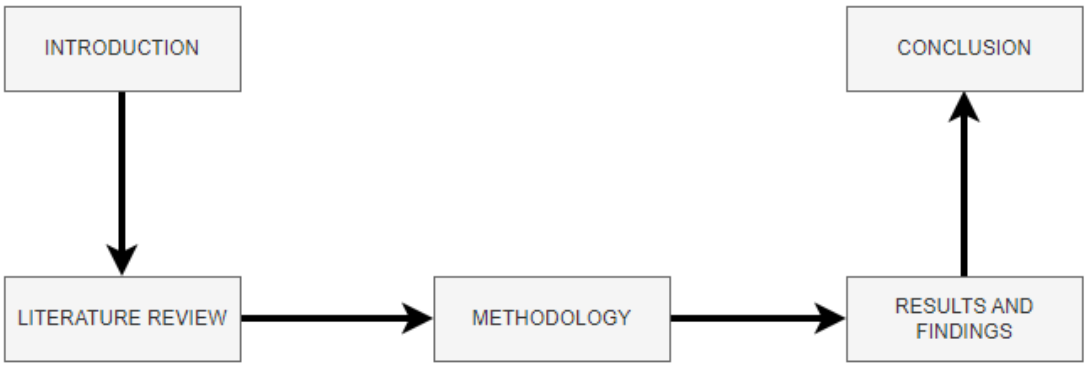


Figure 1: Conceptual Framework Diagram of Thesis Project

1.4 System Setup

For this research, python was used as the programming language and Jupyterlab was used as the development environment. In addition, libraries such as numpy, pandas, matplotlib, statsmodels, sklearn were used for coding.

Chapter 2

LITERATURE REVIEW

Reliable sales forecasts are vital for businesses. Accurate predictions benefit companies in many ways. Accurate forecasts allow companies to develop different business strategies, reduce costs, increase profits and increase customer satisfaction. Many methods have been developed and there have been lots of work to predict sales data from past to present. A lots of traditional sales forecasting methods have been tried and tested for decades. With the development of artificial intelligence, traditional methods of sales forecasting have been replaced by modern forecasting algorithms using machine learning.

The goal of the study area known as machine learning is to educate machines to do cognitive tasks that are similar to those of the human mind. Despite having cognitive abilities that are normally far more limited than those of the average person, they are able to digest enormous amounts of information fast and produce insightful business information [8]. When we compare Machine Learning Methods with Traditional Sales Forecasting Methods, we see that machine learning methods are more advantageous. Machine Learning algorithms provide acceleration of data processing speed, more accurate results and analyzing more data. There have been lots of work in the area of retail demand forecasting using machine learning. This part of the thesis provides a review of several related studies carried out on demand forecasting using different datasets, including Walmart data set.

The authors in [9] proposed an approach for demand forecasting on an e-commerce website. According to their proposed approach, there is a stack generalization-based model that uses multiple learning algorithms to improve forecasting performance in demand forecasting. In this stack generalization approach Random Forest, Gradient Boosted Trees, Decision Tree and Linear Regression algorithms have been used.In the first stage, these algorithms were used as first-level regressors. Then final prediction is made using best algorithm which gives the best prediction on first stage. Data from one of Turkey's most popular online e-commerce companies was used in the experiments. The data used was processed and arranged to present the weekly sales. In the sales data, the timestamp was used weekly and the popularity of each sales product was determined and used as an extra parameter. The disadvantage here is that external factors were ignored. A product sales may vary according to discounts and exchange rates. It has not been determined whether there is a discount for the product sold and other external factors have been completely ignored in their study. With external factors, there can be big differences in demands. According to the experimental results, some machine learning methods gave almost as good results as the stacked generalization method. With less data, proposed method showed more accuracy. But, performance of this approach can not be predicted with more data because the model was not been tested. Additionally, it is not known how the proposed model will provide results due to the variability of external factors [9].

In [10], a Demand Forecasting Model using 3 algorithms which are K-Nearest Neighbor, Decision Tree Classifier and Gaussian Naive Bayes was proposed. Data was collected, processed and raw data was converted to data set for model training and testing. After selection of features, selected algorithms were used and the best model was selected based on the results. The data used in the experiments were

collected from super and other stores. In the data set created, there are data for ten different products. In the experiments used, features such as seasonal weather, time, product category, customer behavior were taken into account. In their study, features and factors such as exchange rates, fuel prices, store location and region were not taken into account.

According to experimental results, their proposed method worked well in local market data and Gaussian Naive Bayes worked well with the used dataset. Mean Absolute Percentage Error and Mean Percentage Error were used as performance metrics. According to results, Mean absolute percentage error of Gaussian Naive Bayes has been less than Holt-Winters and Fuzzy Neural Network methods. Since the location of the stores, fuel prices, daily weather conditions were not taken into account, the accuracy of the recommendations was affected[10].

According to the study presented in [11], the time series approach model was used to forecast demand in a food company. In their study, historical demand data was used with many improved autoregressive integrated moving average (ARIMA) models. The most suitable model was chosen as ARIMA (1,0,1) and this model was selected according to Box–Jenkins time series procedure and four performance criteria. These performance criteria are Akaike criterion, Schwarz Bayesian criterion, maximum likelihood, and standard error. For this study, data has been collected from one of the Moroccan food company. Data set was arranged for ARIMA model containing timestamp and total number of the demand. For this model, coefficients of the algorithm were obtained using fast maximum likehood estimation algorithm by providing the main parameters of the algorithm which are number of autoregressive terms (p), number of differences (d) and number of moving averages (q). According

to the Autocorrelation Function and Partial Autocorrelation Function graphics, the appropriate model, namely the p, d, q values, were determined. In this study, univariate time series analysis was used but external factors were ignored. There are many factors which can affect the demand of food. Although the selected model showed high forecasting accuracy in the used data set, it is not known how the proposed model could yield results with different datasets. If the dataset used is not stationary, it is thought that the estimations to be made would be wrong [11].

Nowadays, time series algorithms are generally used for sales and demand forecasting. In a study conducted by Jason Brownlee in [12], a Random Forest Model for Time Series estimation was developed. In their study, the time series data was structured in accordance with supervised learning using a sliding-window representation. In structured data set, the number of demand and timestamp information was included. Although one-step and univariate estimation was used in this study, multivariate estimations can be easily made with the proposed method. A stationary dataset was used for the proposed method and it provided good results. However, it is not known how it could perform in non-stationary datasets. Additionally, external factors were not considered [12].

External factors are one of the most important problems in the accuracy of demand forecasting. Forecasts are affected by external factors such as price discounts, weather, seasonality, holidays, inflation, etc. In order to overcome this problems, the factors affecting the demand forecast should be taken into consideration. In [13], authors developed the Seasonal Autoregressive Integrated Moving Average with external variables (SARIMAX) model in order to estimate the daily sales of foods that can be broken in a retail store. SARIMAX model can handle stationary and non-

stationary time series with seasonal elements. In this study, the daily sales data of bananas were prepared for the forecasting model. The main parameters of the SARIMAX model were determined using the maximum likelihood method. Exogenous variables such as price discounts, holiday effects, and month effects were used in the existing dataset. Thus, the disadvantage of the SARIMA model, which only considers the trend and seasonality, was eliminated. Estimations were more accurate after taking the other external factors into account. At the end of their study, SARIMA and SARIMAX models have been compared. SARIMAX model has been performed much better then SARIMA which does not consider the external factors [13].

Another study with forecasting was done by [14]. In this study, Linear regression, Support Vector regression, Decision Tree, Ramdom Forest Regressor and Extra Tree Regressor were used as Machine Learning models. For each model, the best training parameters were selected. The results were compared to select an algorithm which provides the highest accuracy. However, the data set used for stock market prices does not include any external factors. In fact, factors such as inflation, unemployment, interest rate and money supply, directly affect the stock market prices [14].

Some of the recently machine learning algorithms for demand forecasting are Amazon's DeepAR+, Facebook's Prophet and CNN-QR. These algorithms are used to perform forecasting based on time series data. In [3], the authors examined these algorithms and compared their performances. In their experiments, monthly sales data were prepared, daily sales were collected, and monthly average prices and number of sales were determined for each product. In their proposed algorithm, the

trend and seasonality that may have an effect on the data behavior can be handled. According to the experiment results, Facebook's Prophet Algorithm showed better accuracy for forecasting the sale of frequent items with a longer history. This algorithm observed the effects of holidays, but other external factors are ignored. Amazon's DeepAR+ and CNN-QR algorithms provided good results for small amount of sale data, these algorithms has the principle of operation which is creating one model for all items [3].

The following studies discussed in the remainder of this section, uses Walmart dataset, where this thesis study was based on.

Anita et al. [15] used the Holt Winters algorithm to predict Walmart Sales. Data from 2010, 2011 and 2012 were used for training and they made sales forecasts for the next 39 weeks. In this study, Time Series algorithm was used for seasonality, trend and randomness observation. When the results of this study was examined, it has been seen that there are sales rates similar to the previous years, but the error rate could not be calculated because there are no real values of the estimates published in the article.

Gurudevi et al. [16] developed a sales forecasting system and suggested the use of regression algorithms. Random Forest was used in their study. In regression, Random Forest consists of different samples using multiple decision trees and majority votes are used for classification and averaging. In this study, Mean Absolute Error, Mean Square Error and Root Mean Square Error metrics were used to calculate accuracy rates. In this study, the performance of this study could not be

evaluated because the separation rates of the data for training and testing and the selected features were not clear.

Akande et al. [17] estimated sales using the Extreme Gradient Boosting algorithm. The dataset given for training was divided into two and 70% of the data was used for training whereas 30% for testing. The dataset used was cleaned and feature engineering was used. With the One-Hot encoding method, new columns were created for each value of the categorical data and 0 or 1 was assigned as the value. The Date Column is divided into different columns as year, month and day. MAE, R2, MSE and RMSE were used as performance metrics. According to the results they obtained, it has been seen that the XGBoost algorithm performed well. 1317 MAE, 3477 RMSE error values were obtained with hyperparameter adjustment. The disadvantage of this study is that there is no study on the relationship between features such as Correlation Analysis in the studies conducted for the proposed method, so it may have high error rates when different data are used.

In [18] Random Forest Algorithm in was used. The author used 20% of the dataset for cross validation and testing, while the remaining 80% was used for model training. In this study, data analysis, data cleaning, data preprocessing and feature selection was done. In model training, department, store, temperature, unemployment rate, fuel price, store size, markdowns, holiday flag, lagged sales, sales differences, lagged flag, pre-christmas flag and black friday flag features were used. MAE was used as the performance metric. In the study on the full dataset, 2573 MAE values were obtained.

Raizada [19] used Multiple Linear Regression, Random Forest Regression, K-NN, Support Vector Machine (SVM) and Extra Tree Regression algorithms to make sales forecasts. The results from all models were compared and the best model was determined. 80% of the data set was used for training and 20% for testing. After data preprocessing, the features were selected, the models were trained, and the tests were carried out. MAE, MSE and RMSE metrics were used for performance evaluation. According to the calculated error rates, the Extra Tree Regression algorithm provided the best results, followed by the Random Forest algorithm [19].

On the other hand, Catal et al. [20] used time series and regression algorithms in their studies. In addition to Bayesian Regression, Linear Regression, Neural Network Regression, Boosted Decision Tree Regression and Decision Forest Regression algorithms, Naive Method, Seasonal ETS, Non-Seasonal ARIMA, Seasonal ARIMA, Non-Seasonal ETS, Drift Method and Average Method time series algorithms were used. 2010 and 2011 data in the dataset were used for training and 2012 data were used for testing. In the evaluation of test results, RMSE, MAE and R2 were used as performance metrics. After data cleaning and feature selection, all selected models were trained and predictions were made. The results of all models were analyzed and the best performer were selected. According to the results, Boosted Decision Tree showed the best performance.

As seen in all studies reviewed, only one model was selected in each study. At the same time, studies in which external factors were examined in detail and the most suitable features were selected showed better performance.

Chapter 3

METHODOLOGY

This section contains detailed information about the proposed method to make the best estimation for the selected dataset. The remainder of the chapter provides detailed information about the proposed method, dataset, feature selection, machine learning and time series algorithms used.

3.1 The Proposed System

The proposed methodology in this study is a hybrid modeling. A hybrid model was created using machine learning regression algorithm and Prophet time series estimation algorithm to predict Walmart sales. After the dataset preprocessing and explatory data analysis stages offered by Walmart, separate training and test datasets were created for regression and time series modeling. First of all, the best regression algorithm was chosen with the K-Fold Cross Validation algorithm. After model training and prediction, the regression estimation and the time series results were combined.

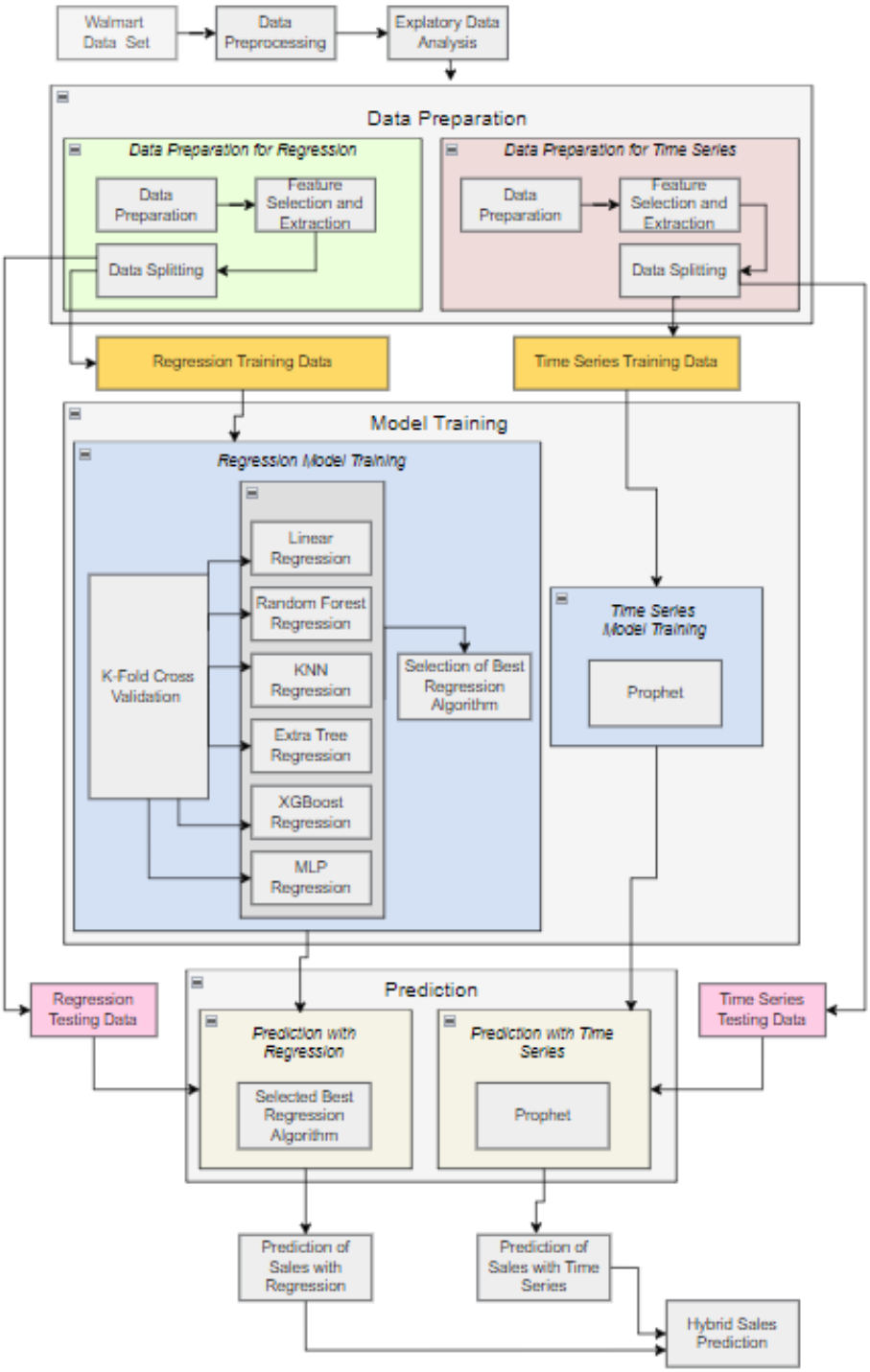


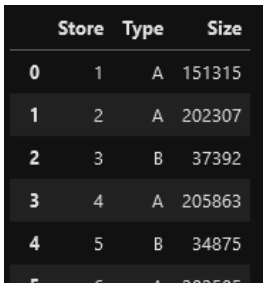
Figure 2: Block Diagram of Proposed Methodology

3.2 Understanding of Data

The dataset for this study was taken from a Kaggle competition created by Walmart. The dataset includes weekly sales data of different departments of 45 Walmart stores located in different locations. Since each store has different departments, it is our

main task to estimate the sales data of each store's department. Kaggle provides 4 different files for data sets. These are;

stores.csv: This file contains information about 45 stores and their type and capacity as shown in Figure 3.



	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

Figure 3: Store Data Set

train.csv: This file is the training dataset containing sales data between 2010-02-05 and 2012-11 as shown in Figure 4. Data fields in this file: Store (Store Number), Dept (Department number), Date (Week date), Weekly_Sales (Weekly Sales of the Related Store's Department), Is_Holiday (Information if the week is a special holiday week).

	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
...	...	...	...	...	...
421565	45	98	2012-09-28	508.37	False
421566	45	98	2012-10-05	628.10	False
421567	45	98	2012-10-12	1061.02	False
421568	45	98	2012-10-19	760.01	False
421569	45	98	2012-10-26	1076.80	False

Figure 4: Train Data Set

test.csv: This file has the same structure as the train.csv file as shown in Figure 5. It just doesn't include the Weekly_Sales field. Because this area is expected to be estimated. This dataset were not be used in the study since the sales values are not

known in the test dataset presented by Kaggle and it could not be evaluated. Instead of test data, train dataset was splitted and used as test and train.

	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
...	...	...	...	...
115059	45	98	2013-06-28	False
115060	45	98	2013-07-05	False
115061	45	98	2013-07-12	False
115062	45	98	2013-07-19	False
115063	45	98	2013-07-26	False

Figure 5: Test Data Set

features.csv: This file contains data about the external factors and weekly markdowns applied by the store according to the region where the store is located as shown in Figure 6. Data fields in this file: Store (Store number), Temperature (weekly average temperature of the region where the store is located), Fuel_Price (weekly average fuel price in the region where the store is located), CPI (Consumer Price Index for the relevant week), Unemployment (unemployment rate in the region where the store is located for the relevant week), MarkDown1:5 (discount type and number of discounts for the relevant week).

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
8185	45	2013-06-28	76.05	3.639	4842.29	975.03	3.00	2449.97	3169.69	NaN	NaN	False
8186	45	2013-07-05	77.50	3.614	9090.48	2268.58	582.74	5797.47	1514.93	NaN	NaN	False
8187	45	2013-07-12	79.37	3.614	3789.94	1827.31	85.72	744.84	2150.36	NaN	NaN	False
8188	45	2013-07-19	82.84	3.737	2961.49	1047.07	204.19	363.00	1059.46	NaN	NaN	False
8189	45	2013-07-26	76.06	3.804	212.02	851.73	2.06	10.88	1864.57	NaN	NaN	False

Figure 6: Features Data Set

One of the most important information in the presented data set is the holiday weeks. When these holiday weeks are examined, there are 4 important holiday weeks which are Super Bowl, Labor Day, Thanksgiving and Christmas. In 2010, 2011 and 2012, all holidays fall on the same week of the year as shown in Figure 7.

	Date	Holiday_Type	Year	Month	Day	WeekOfYear
0	2010-02-12	Super Bowl	2010	2	12	6.0
1	2011-02-11	Super Bowl	2011	2	11	6.0
2	2012-02-10	Super Bowl	2012	2	10	6.0
3	2013-02-08	Super Bowl	2013	2	8	6.0
4	2010-09-10	Labor Day	2010	9	10	36.0
5	2011-09-09	Labor Day	2011	9	9	36.0
6	2012-09-07	Labor Day	2012	9	7	36.0
7	2013-09-06	Labor Day	2013	9	6	36.0
8	2010-11-26	Thanksgiving	2010	11	26	47.0
9	2011-11-25	Thanksgiving	2011	11	25	47.0
10	2012-11-23	Thanksgiving	2012	11	23	47.0
11	2013-11-29	Thanksgiving	2013	11	29	48.0
12	2010-12-31	Christmas	2010	12	31	52.0
13	2011-12-30	Christmas	2011	12	30	52.0
14	2012-12-28	Christmas	2012	12	28	52.0
15	2013-12-27	Christmas	2013	12	27	52.0

Figure 7: List of Holidays with Date Informations

3.3 Data Preprocessing

3.3.1 Merging Data Sets

At this stage, 3 datasets which are store, features and train which contains sales were combined and a single dataset was obtained, as shown in Figure 8.

	Store	Dept	Date	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
0	1	1	2010-02-05	24924.50	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	1	1	2010-02-12	46039.49	True	A	151315	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106
2	1	1	2010-02-19	41595.55	False	A	151315	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106
3	1	1	2010-02-26	19403.54	False	A	151315	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106
4	1	1	2010-03-05	21827.90	False	A	151315	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
421565	45	98	2012-09-28	508.37	False	B	118221	64.88	3.997	4556.61	20.64	1.50	1601.01	3288.25	192.013558	8.684
421566	45	98	2012-10-05	628.10	False	B	118221	64.89	3.985	5046.74	NaN	18.82	2253.43	2340.01	192.170412	8.667
421567	45	98	2012-10-12	1061.02	False	B	118221	54.47	4.000	1956.28	NaN	7.89	599.32	3990.54	192.327265	8.667
421568	45	98	2012-10-19	760.01	False	B	118221	56.47	3.969	2004.02	NaN	3.18	437.73	1537.49	192.330854	8.667
421569	45	98	2012-10-26	1076.80	False	B	118221	58.85	3.882	4018.91	58.08	100.00	211.94	858.33	192.308899	8.667

Figure 8: Merged Data Sets

3.3.2 Cleaning of Data

When the data set was examined, it has been determined that there are null values in the MarkDown1-5 fields. Therefore, the null values in the data set were checked and replaced with 0, as shown in Figure 9.

	Store	Dept	Date	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
Data Type	int64	int64	object	float64	bool	object	int64	float64	float64	float64	float64	float64	float64	float64	float64	float64
Null Values (nb)	0	0	0	0	0	0	0	0	0	270889	310322	284479	286603	270138	0	0
Null Values (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	64.257181	73.611025	67.480845	67.984676	64.079038	0.0	0.0

Figure 9: Analysis of Null Values

3.3.3 Extraction of Informations from Date Column

It has been determined that the date column in the data set is in string format. First, the data type has been converted to DateTime format. For explotory data analysis, year, month, day and week number information was created using the date column, as shown in Figure 10.

	Store	Dept	Date	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Year	Month	Day	WeekOfYear
0	1	1	2010-02-05	24924.50	False	A	151315	42.31	2.572	0.00	0.00	0.00	0.00	0.00	211.096358	8.106	2010	2	5	5.0
1	1	1	2010-02-12	46039.49	True	A	151315	38.51	2.548	0.00	0.00	0.00	0.00	0.00	211.242170	8.106	2010	2	12	6.0
2	1	1	2010-02-19	41595.55	False	A	151315	39.93	2.514	0.00	0.00	0.00	0.00	0.00	211.289143	8.106	2010	2	19	7.0
3	1	1	2010-02-26	19403.54	False	A	151315	46.63	2.561	0.00	0.00	0.00	0.00	0.00	211.319643	8.106	2010	2	26	8.0
4	1	1	2010-03-05	21827.90	False	A	151315	46.50	2.625	0.00	0.00	0.00	0.00	0.00	211.350143	8.106	2010	3	5	9.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
421565	45	98	2012-09-28	508.37	False	B	118221	64.88	3.997	4556.61	20.64	1.50	1601.01	3288.25	192.013558	8.684	2012	9	28	39.0
421566	45	98	2012-10-05	628.10	False	B	118221	64.89	3.985	5046.74	0.00	18.82	2253.43	2340.01	192.170412	8.667	2012	10	5	40.0
421567	45	98	2012-10-12	1061.02	False	B	118221	54.47	4.000	1956.28	0.00	7.89	599.32	3990.54	192.327265	8.667	2012	10	12	41.0
421568	45	98	2012-10-19	760.01	False	B	118221	56.47	3.969	2004.02	0.00	3.18	437.73	1537.49	192.330854	8.667	2012	10	19	42.0
421569	45	98	2012-10-26	1076.80	False	B	118221	58.85	3.882	4018.91	58.08	100.00	211.94	858.33	192.308899	8.667	2012	10	26	43.0

421570 rows × 20 columns

Figure 10: Data Set after Extraction Features from Date Column

3.4 Explatory Data Analysis

3.4.1 Analysis of Negative Sales Values

Train data set contains 421570 sales record for 2010, 2011 and 2012 years. The values of the Weekly Sale field was checked at this stage. In Figure 11, the percentages of weekly sales value ranges are given. On the presented data set, 99.68% is greater than zero, 0.02% is zero, and 0.305% is less than zero, that is, negative values. Negative values here may indicate dirty data or sales returns.

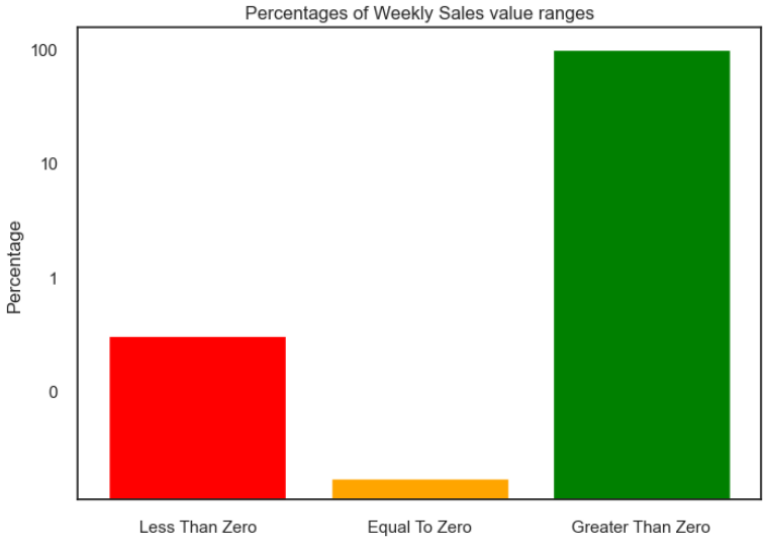


Figure 11: Percentages of Weekly Sales Value Ranges

3.4.2 Changes of Features by Date

Figure 12 shows the changes of main features which are Temperature, Fuel Price, CPI, MarkDown and Unemployement by date. The temperature seems seasonal. The air temperature decreases in winter and increases in summer. Fuel Prices, on the other hand, show a slightly positive trend. There has been an increase from 2010 to the last months of 2012. The Unemployment, on the other hand, shows a slightly negative trend in the given time period. There are some sudden increases in MarkDown values. It has been that all MarkDown types have started to be applied

since the last months of 2011. No inferences could be drawn from this figure regarding the consumer price index.

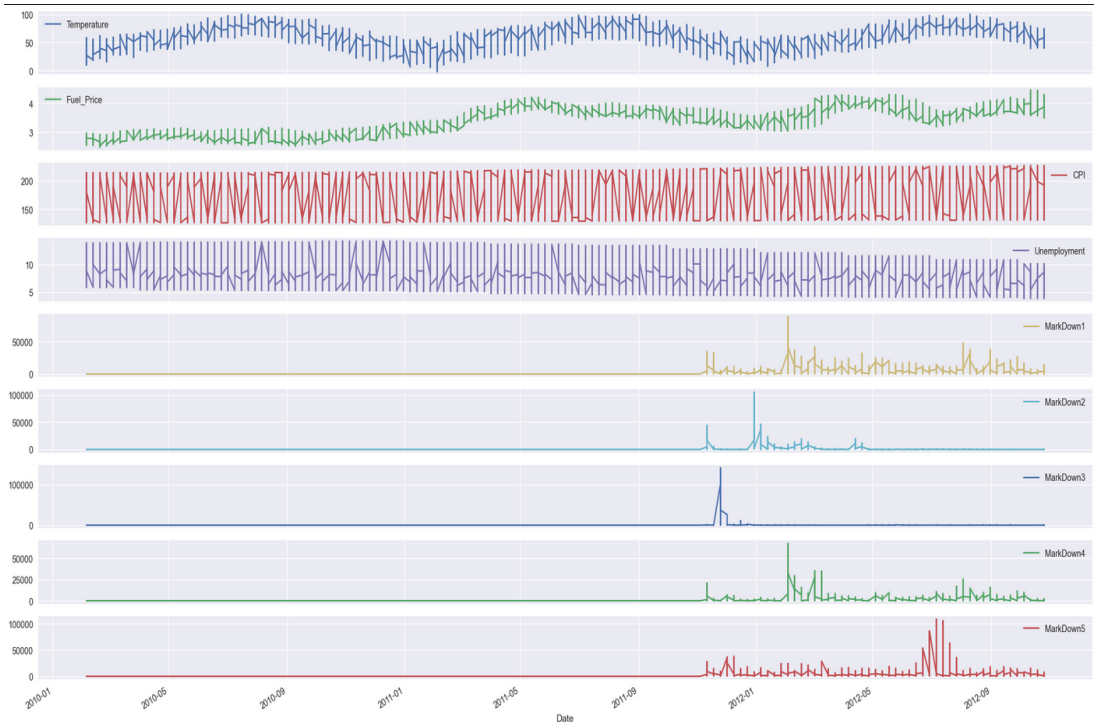


Figure 12: Changes of Main Features by Date

3.4.3 Mean and Median of Weekly Sales against Date

Figure 13 gives information about the mean and median of weekly sales by date. Here, the mean seems to be higher than the median. It seems that there are large sales values in the dataset that affect the mean. It is observed that some stores/departments make higher sales than others.

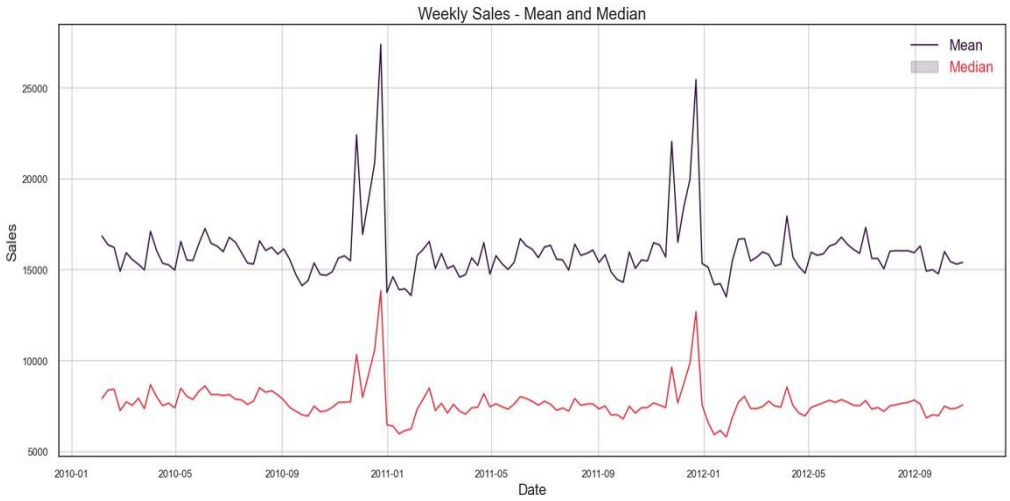


Figure 13: Mean and Median of Weekly Sales against Date

3.4.4 Average Weekly Sales per Year

Figure 14 gives information about the average weekly sales per year. When the sales are analyzed on a weekly basis, it is understood that the holidays have a significant effect on the increase in sales. Especially in 2010 and 2011, there is a great increase in sales during Thanksgiving (Week 47) and Christmas (Week 52) holiday weeks. At the same time, there is a big drop in sales after Christmas. When the graph is analyzed in detail, there is a slight increase in sales in the 13th week of 2010, the 16th week of 2010 and the 14th week of 2012. These weeks are Easter Days although these dates are not specified as holidays or special days in the dataset. Especially in 2012, when compared to the sales of other existing weeks, the highest sales were made in the 14th week of 2012.

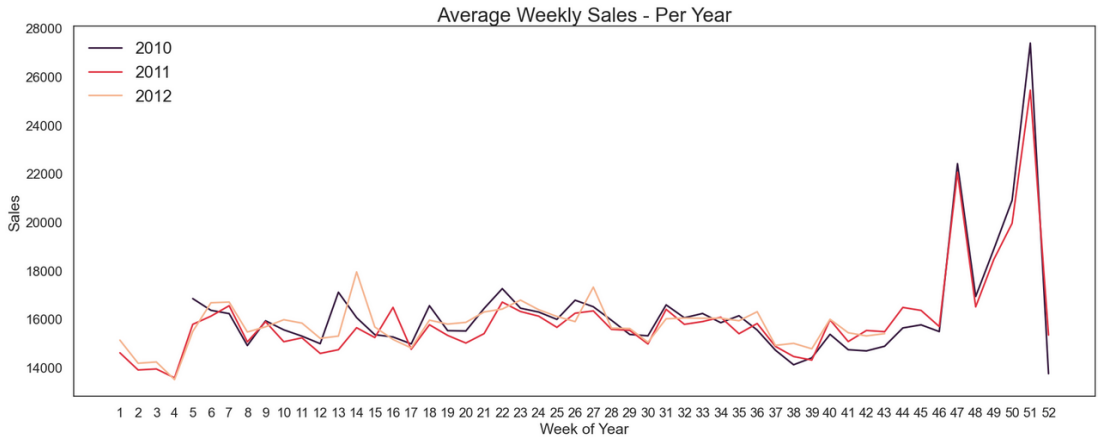


Figure 14: Average Weekly Sales per Year

3.4.5 Average Sales of Stores

Figure 15 presents the average sales information of each store for the years 2010, 2011 and 2012. There are 45 Walmart stores presented in the train dataset and each store has a different average sales level. Stores 2,4,13,14 and 20 showed the highest sales in all 3 years. Store size, store type, location, unemployment rate and temperature in the region, etc. could be listed as the factors for different stores having different sales levels.

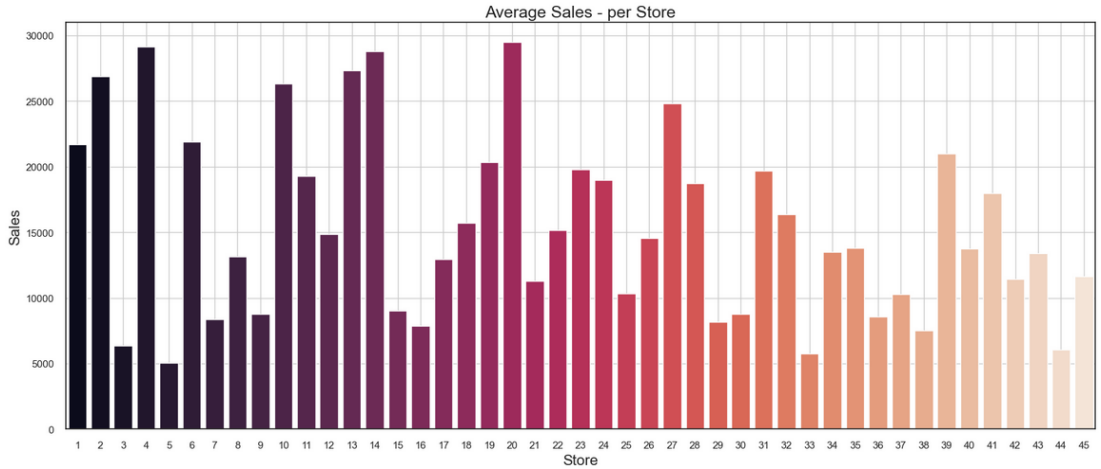


Figure 15: Average Sales of Stores

3.4.6 Average Sales of Store Types

Figure 16 presents the average sales information for each store type for the years 2010, 2011 and 2012. There are 3 different store types, A, B and C, that can be assigned to each store. According to the figure given below, the highest sales were made in A-type stores, and the lowest sales were made in C-type stores. In Table 1, the value ranges of the sizes of each store type are given. Using Table 1, we conclude that stores with larger sizes make more sales.

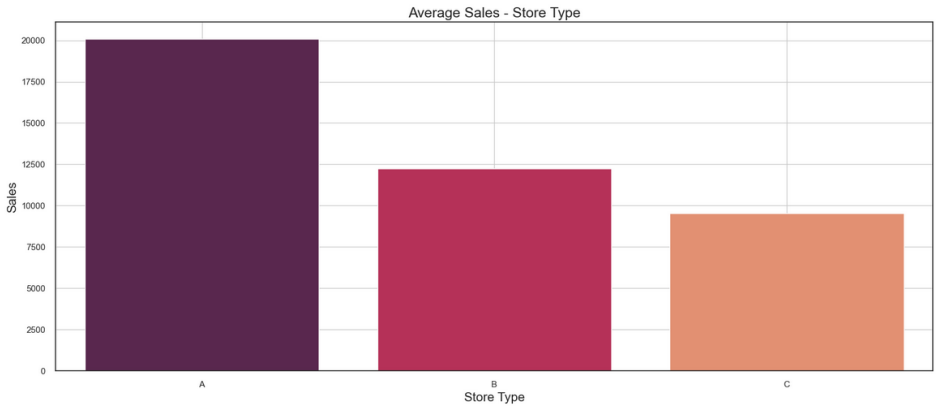


Figure 16: Average Sales of Store Types

Table 1: Minimum and Maximum Store Size Values

	Size	
Store Type	Min	Max
A	39690	219622
B	34875	140167
C	39690	42988

3.4.7 Average Sales per Department

Figure 17 presents the average sales information for each department for the years 2010, 2011 and 2012. The given departments show different levels in their average

sales. Departments 38,65,72,92 and 95 have the highest average sales. In the given dataset, the department numbers range from 0 to 100, however there are some numbers not in use. Department numbers are given in Table 2.

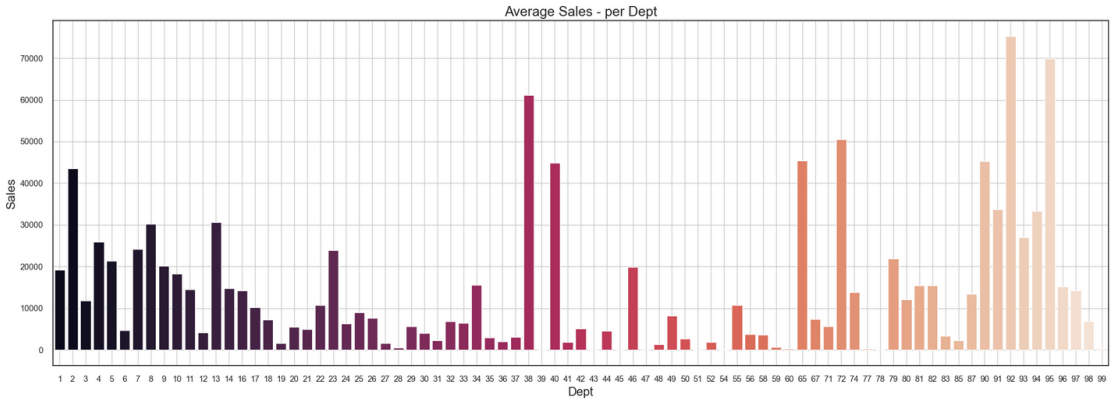


Figure 17: Average Sales per Department

Table 2: Department Numbers in Data Set

Dept	99 98 97 96 95 94 93 92 91 90 87 85 83 82 81 80 79 78 77 74 72 71 67 65 60 59 58 56 55 54 52 51 50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Dept – Not In List	0 15 53 57 61 62 63 64 66 68 69 70 73 75 76 84 86 88 89

3.4.8 Relationship between Sales and MarkDowns

Figure 18 shows the relationship between sales and markdown1-5. In general, sales are expected to increase as the markdown increases. However, when the graph is examined, there is no increase in sales as the markdown amount increases. Sales are decreasing on the Markdown 1, Markdown 2, Markdown 4 and Markdown 5 charts. There may be different factors in the low markdown values and high sales. It is not possible to establish a precise relationship when the Markdown 3 graph is examined. Therefore, it was concluded that there is no clear relationship between the MarkDown1-5 data and sales.

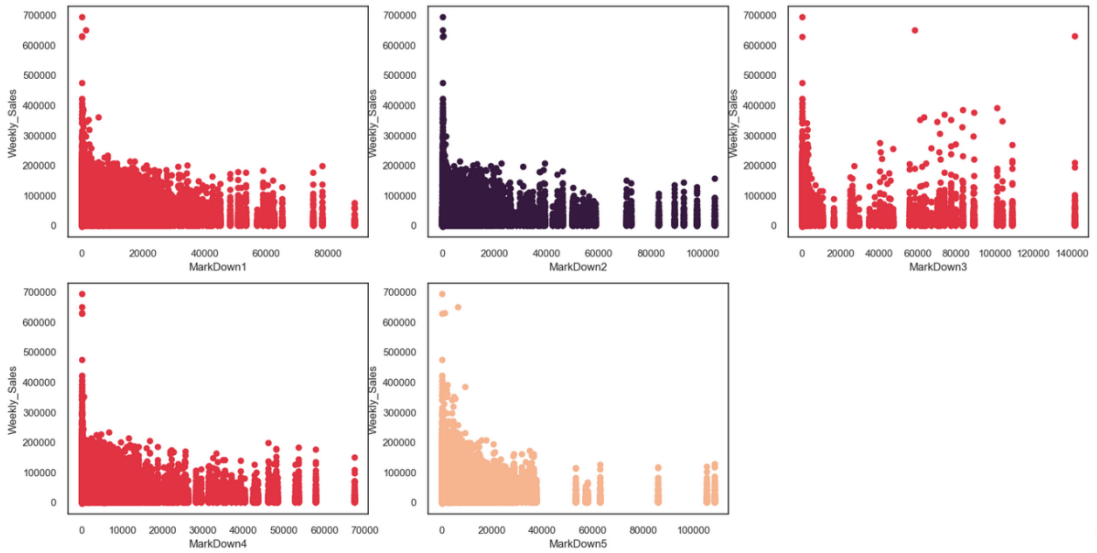


Figure 18: Relationship between Weekly Sales and MarkDown1-5

3.4.9 Relationship between Store Type, Size and Sales

Figure 16 and Table 1 provides information on the effects of store types and sizes on average sales. Figure 19 shows the relationship between Store Type, Store Size and Sales in detail. Here, we see that sales are linearly proportional to store size. Although some Type B stores have made more sales than Type A stores, it generally shows that sales increase as the size of the store increases.

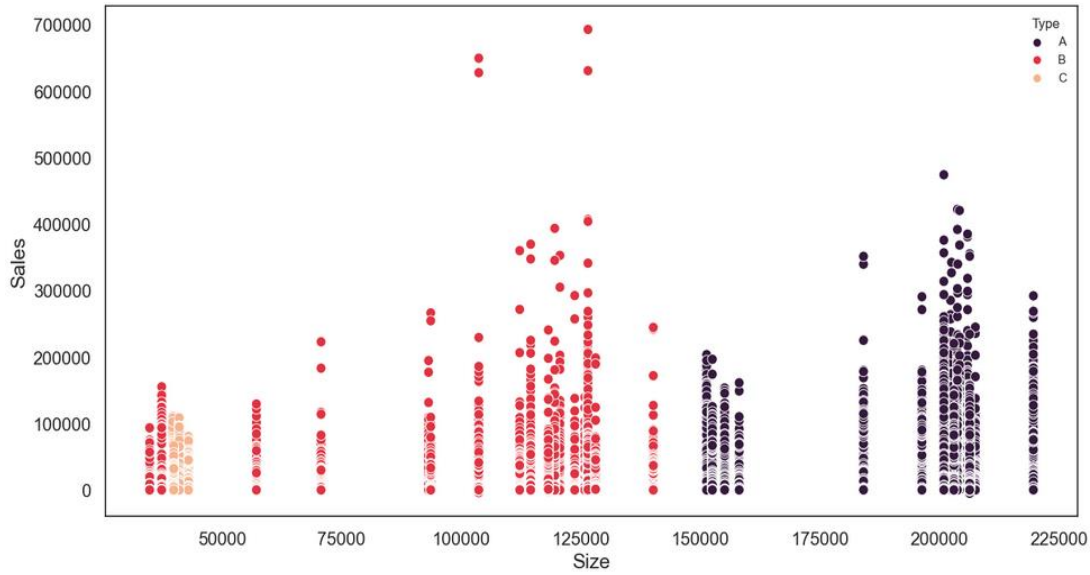


Figure 19: Relationship between Store Type, Size and Sales

3.4.10 Relationship between Store Type, Unemployment Rate and Sales

Generally, decreases in sales can be observed as the unemployment rate increases. Figure 20 shows the relationship between store type, unemployment rate and sales. According to the graph, when the unemployment rate is between 4 and 10, it is seen that the sales of type A and B stores are higher. As the unemployment rate increases, the sales of the B type stores decrease more than the A type store. On the other hand, the sales of the C type stores continued with the same average. Therefore, there is no clear relationship between unemployment rate and sales.

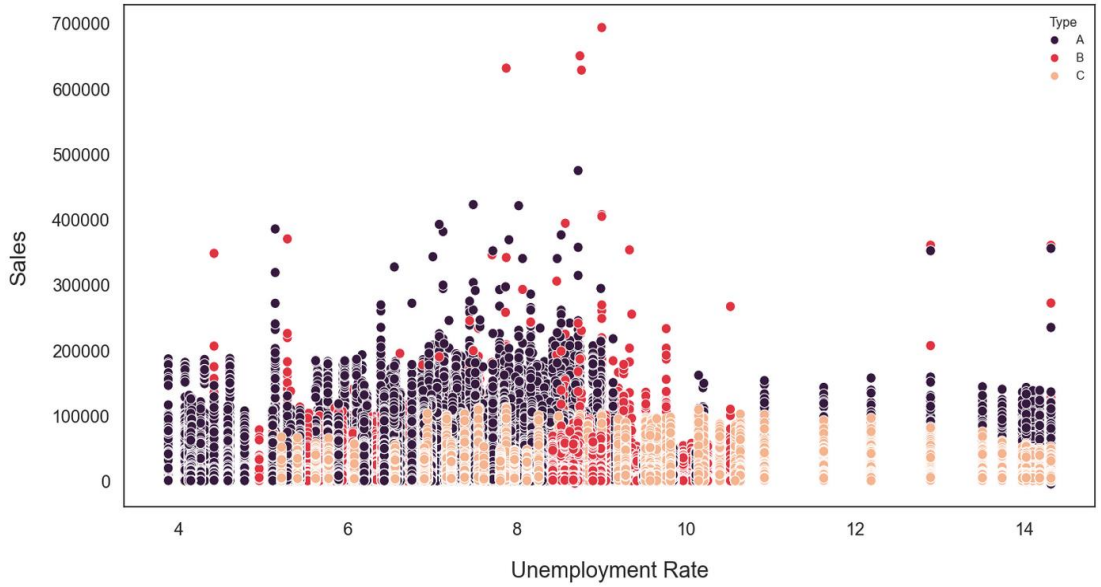


Figure 20: Relationship between Store Type, Unemployment Rate and Sales

3.4.11 Relationship between Store Type, Fuel Price and Sales

Figure 21 shows the relationship between store type, fuel price and sales. In this figure, there is no clear relationship between fuel price and sales. Although fuel prices are above 4.25 $, there is no decrease in sales of C type store. A slight decrease is observed in the A type store. In addition, A and B type stores had the highest sales fuel prices are between 2.75 $ and 3.75 $.

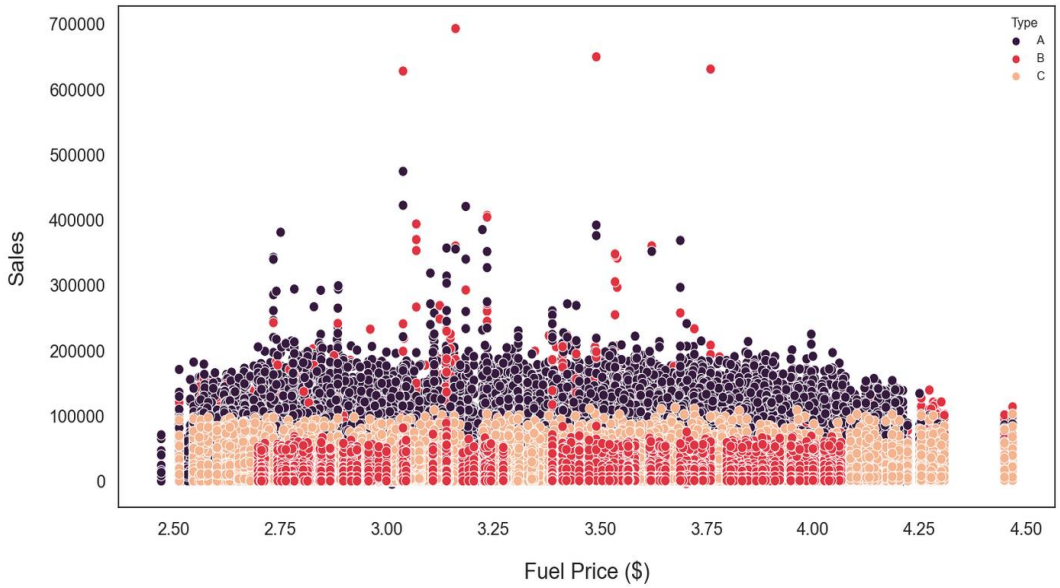


Figure 21: Relationship between Store Type, Fuel Price and Sales

3.4.12 Relationship between Store Type, CPI and Sales

The Consumer Price Index is a gauge of how prices for a market basket of consumer goods have changed on average over time for urban consumers. As the CPI increases, the prices of goods increase. This reduces a person's purchasing power. In other words, it means that person has to spend more money in order to maintain the same living standards. Figure 22 shows the relationship between store types, CPI and Sales. It seems that there are 3 different clusters here. However, there is no clear relationship between CPI and sales. Although the CPI is higher, there is no significant decrease in sales.

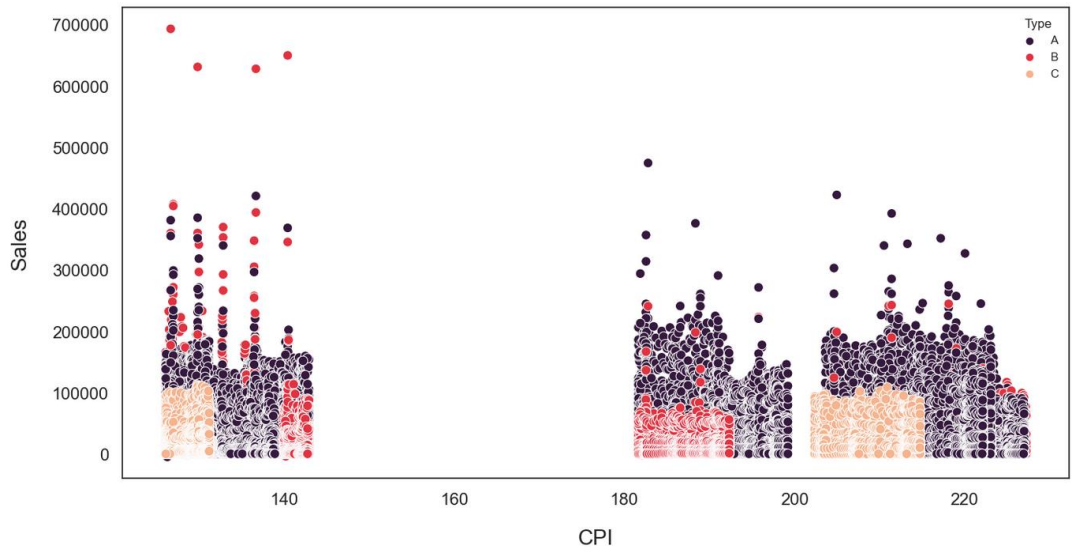


Figure 22: Relationship between Store Type, CPI and Sales

3.4.13 Relationship between Store Type, Temperature and Sales

Figure 23 shows the effect of temperature on average sales. In general, there seems to be a slight decrease in sales when the temperature is very low and high. However, there does not appear to be a direct relationship between the regional temperature and the weekly sales of the store. In general, most stores have sales between 30 and 90 degrees Fahrenheit.

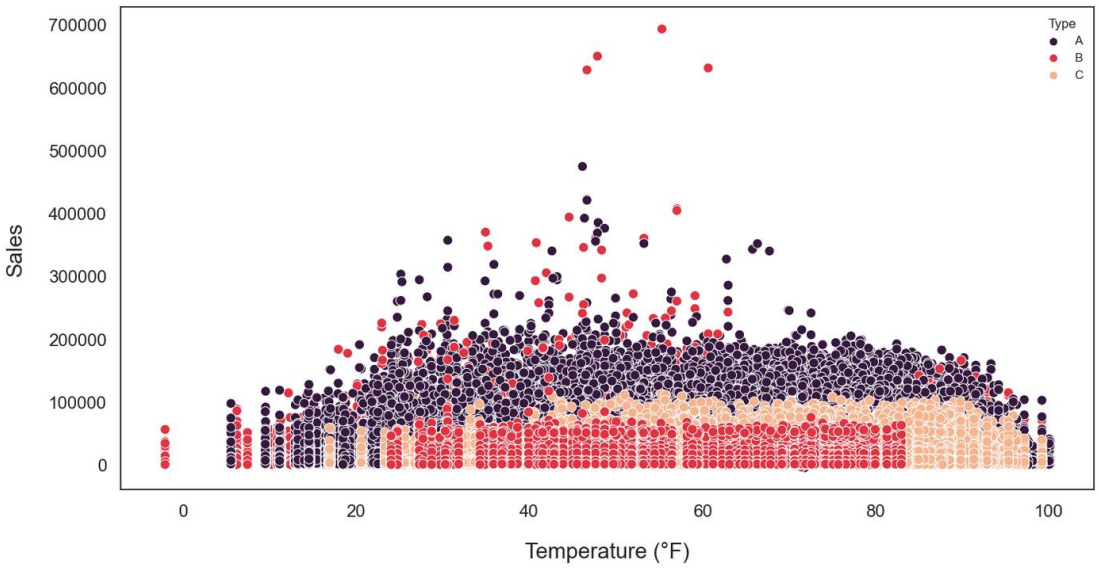


Figure 23: Relationship between Store Type, Temperature and Sales

3.4.14 Relationship between Store Type, Holiday and Sales

When the train dataset is examined, 92.96% of the data is for non-holiday days, and 7.04% for the holidays. In Figure 24, the percentages of the values of the IsHoliday field are given. In short, the sales data for non-holiday days is much higher since the percentage of holiday weeks in the data is lower. However, the average sales during holiday weeks are much higher. As seen in Figure 25, there is no increase or decrease according to the store types during the weeks with or without holidays. For example, there is no significant increase in the sales of the Type C store on holidays.

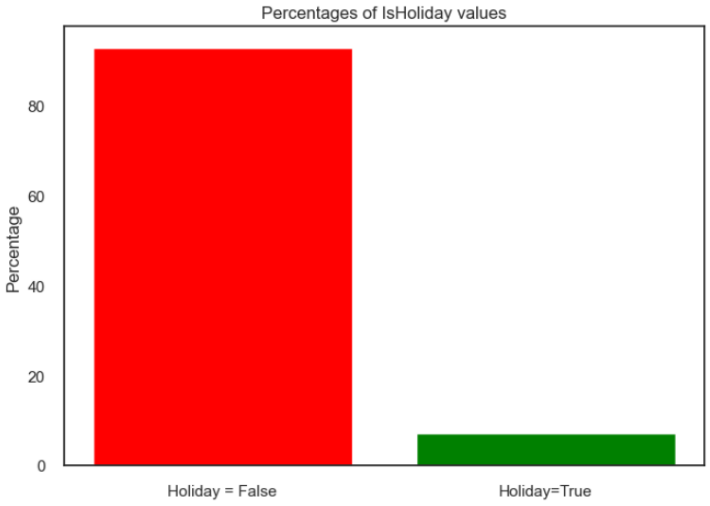


Figure 24: Percentages of IsHoliday Values

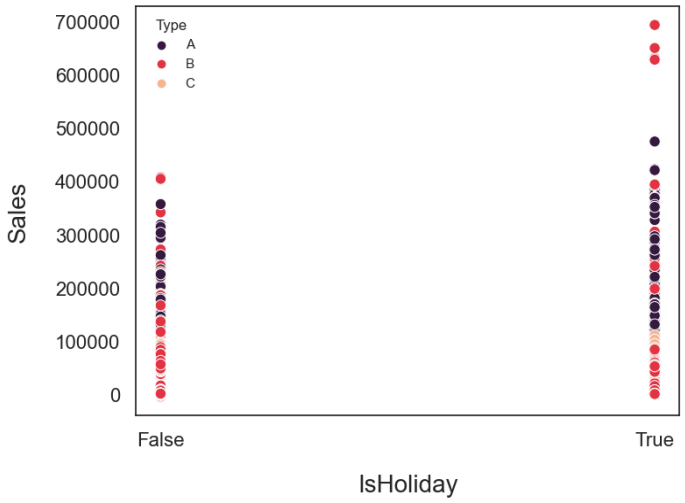


Figure 25: Relationship between Store Type, Holiday and Sale

3.4.15 Correlation Analysis

A linear correlation can be measured using Pearson's correlation coefficient (r). It gauges the relationship between two variables' strength and direction. Correlation number should be between -1 and 1.

Table 3: Pearson's Correlation Coefficient

Pearson Correlation Coefficient (*r*)	**Correlation Type**	**Interpretation**
Between 0 and 1	Positive correlation	If one variable changes, the other variable changes in the **same direction**.
0	No correlation	There is **no relationship** between the variables.
Between 0 and –1	Negative correlation	If one variable changes, the other variable changes in the **opposite direction**.

Persons correlation coefficient r is defined as

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}, \quad (1)$$

where,

n= number of the pairs of the stock

$\sum$xy = sum of products of the paired stocks

$\sum$x = sum of the x scores

$\sum$y = sum of the y scores

$\sum x^2$ = sum of the squared x scores

$\sum y^2$ = sum of the squared y scores.

Figure 26 gives information about the Correlation Matrix. Since the correlation values between the Weekly_Sales property and the MarkDown1-5 values are very close to 0, there is no clear relationship between them. There is a negative correlation between Weekly_Sales property and Fuel_Price, CPI and Unemloyment values. Although the values are close to 0, the correlation between them is negative. There is a high correlation between Fuel_Price and Year properties. On the other hand, there is a moderate correlation between Weekly_Sales and the Size and Type properties.

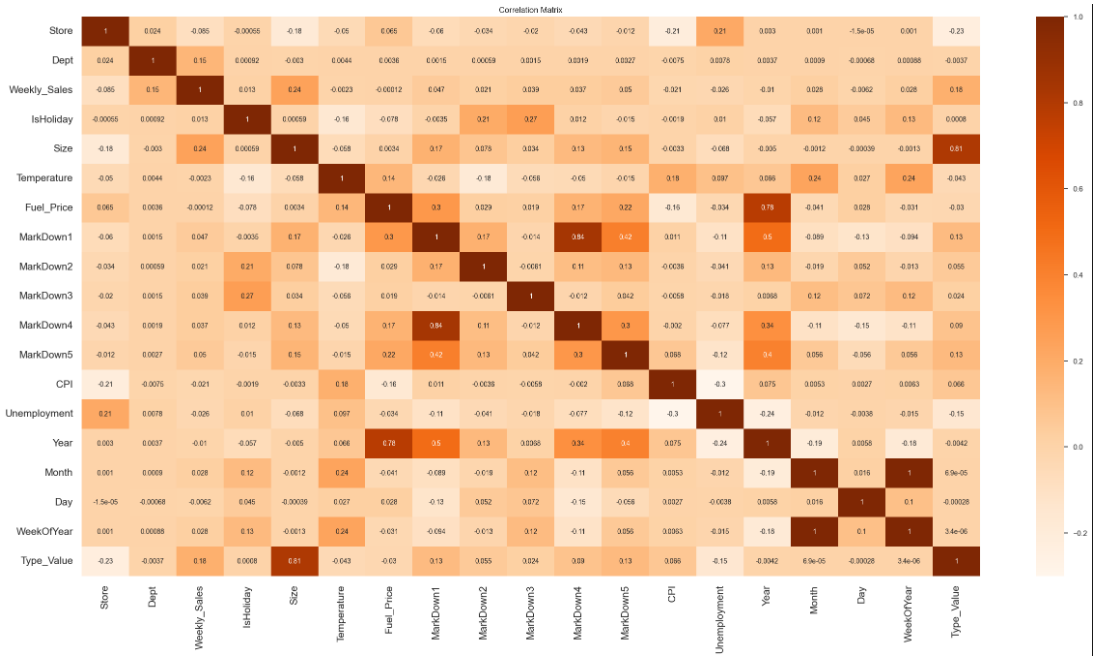


Figure 26: Correlation Matrix

3.5 Data Preparation

3.5.1 Preparation of Data Sets

3.5.1.1 Removing Negative Sales Date

When the train dataset was analyzed, sales data with negative values were seen. 1285 sales data with negative values were removed.

	Store	Dept	Date	Weekly_Sales	IsHoliday	Type	Size	Temperatu
846	1	6	2012-08-10	-139.65	False	A	151315	85.0
2384	1	18	2012-05-04	-1.27	False	A	151315	75.5
6048	1	47	2010-02-19	-863.00	False	A	151315	39.9
6049	1	47	2010-03-12	-698.00	False	A	151315	57.7
6051	1	47	2010-10-08	-58.00	False	A	151315	63.9
...	...	...	...	...	...	...	...	
419597	45	80	2010-02-12	-0.43	True	B	118221	27.7
419598	45	80	2010-02-19	-0.27	False	B	118221	31.2
419603	45	80	2010-04-16	-1.61	False	B	118221	54.2
419614	45	80	2010-07-02	-0.27	False	B	118221	76.6
419640	45	80	2011-02-11	-0.24	True	B	118221	30.3

Figure 27: Snapshoot of Train Data Set with Negative Sales

3.5.1.2 Completion of Missing Sales Data

Deficiencies in sales data were checked before model training and testing, and it was seen that there was no 143-week data for the entire store-department pair. In other words, there is no data for certain weeks for some store-department pairs. In this case, sales data with a sales value of 0 were created for each store-department pair that had no data on certain dates.

	Store	Dept	week_count
2628	6	36	142
2648	30	9	142
2650	23	54	142
2651	20	58	142
2652	28	32	142
...	...	...	...
3297	13	77	1
3298	16	78	1
3299	9	78	1
3300	27	39	1
3322	7	78	1

Figure 28: Number of Weeks of each Store-Department Pair having Sales Data

3.5.1.3 Adding Missing Holiday Data

During the exploratory data analysis, it was found that some important holidays were missing. There was an increase in sales data, especially on Easter Sunday. However, the weeks with the Easter Sunday were not specified as a holiday. Another missing date was the 25th of December. In the weeks of Christmas celebrations (51.Week), there is a great increase in sales. However, in the presented dataset, only the 52nd week was given as a holiday. The holiday dataset has been edited for the missing holiday weeks. Additional data has been added for Easter Sunday and Christmas and the IsHoliday field has been corrected in the training dataset.

	Date	Holiday_Type	Year	Month	Day	WeekOfYear
0	2010-04-04	Easter Sunday	2010	4	4	13.0
1	2011-04-24	Easter Sunday	2011	4	24	16.0
2	2012-04-08	Easter Sunday	2012	4	8	14.0
3	2013-03-31	Easter Sunday	2013	3	31	13.0
4	2010-12-25	Christmas	2010	12	25	51.0
5	2011-12-25	Christmas	2011	12	25	51.0
6	2012-12-25	Christmas	2012	12	25	52.0
7	2013-12-25	Christmas	2013	12	25	52.0

Figure 29: Missing Holiday Informations

3.5.1.4 Feature Selection and Extraction

At this stage, it has been determined which features of the dataset to be used in model training and testing. Since the proposed method in this study is hybrid, separate training and test data sets were created for the regression algorithms and the time series algorithm.

Information on the features used in the selected models after exploratory data analysis is given below:

- MarkDown1-5 properties will not be used because the correlation value is

unimportant.

- Fuel_Price feature has been removed since there is a high correlation between Fuel_Price and Year variables, and both will provide similar information.
- Holiday information will be used due to the change in sales during the holiday weeks. Since it is important which holiday is in the data, holiday information will be converted into indicator columns for regression algorithms and used. For the Time Series algorithm, holiday information was also injected.
- There is a negative correlation between the Temperature, CPI and Unemployment variables and the Weekly_Sales property. Because of this, they have been used.
- Type and Size variables have a moderate positive correlation value with Weekly_Sales.These two variables were used in regression algorithms.Since the Type variable is a categorical data, it will be converted to indicator columns and used. Since these two variables do not change depending on time and the algorithm has trained and predicted for each store and department, they did not used in the time series algorithm.
- In the regression algorithms, new and meaningful features were extracted from the Date column instead of the Date column with time information. By using only the Week_Number and Year information components, the Date information provided relevant patterns in the data.

- In time series algorithms, data is characterized by observations recorded at different points in time. That is, it is based on the nature of the data to be temporally ordered. Therefore, the Date column was used in the Time Series algorithm without dividing it into different components.

In summary, there are 12 main features in the dataset presented by Kaggle. According to exploratory data analysis and data selection, 10 of these features were selected. By transforming categorical data into indicator columns, 18 features were selected to be used in regression algorithms and 9 features in time series algorithms.

3.5.2 Data Splitting

Since test.csv dataset has no real sales value, the evaluations to be made with this dataset were uncertain. Since regression and time series were used in the study, the train.csv dataset was splitted differently. For regression, 80% of the data for the years 2010-2011 was used for training, 20% for validation, and all of the data for 2012 was used for testing. For the time series method, the part of the train.csv dataset for 2010 and 2011 was used for training, and the part for 2012 was used for testing. That is, 70% of all train.csv data for the time series was splitted for training and 30% for testing.

3.6 Model Selection and Implementation

We can divide the model selection into two. First of all, according to the literature study, it has been seen that many single regression algorithms work well. Therefore, Extra Tree Regression, XGB Regression, Random Forest Regression, KNN Regression, Linear Regression and MLP Regression algorithms were chosen. For Time Series Algorithms, only Prophet algorithm was used. All selected algorithms

were used with default hyperparameter values and no hyperparameter tuning was done.

3.6.1 Selected Regression Algorithms

3.6.1.1 Extra Tree Regression

Extremely Randomized Trees Regression, also referred to as ExtraTreeRegressor, is another name for Extra Trees Regression, which is an ensemble learning technique for regression applications. Given its many similarities to the Random Forest method, it is an extension of that model. To get precise regression predictions, Extra Trees Regression constructs a number of decision trees and aggregates their predictions.

Extra Trees Regression reduces overfitting compared to individual decision trees. By combining multiple trees and averaging their predictions, it reduces the variance of the model and provides more stable and reliable results. It is less sensitive to outliers in the training data compared to some other regression algorithms. It makes use of random splits and averages the predictions, which helps to dampen the impact of outliers. It is computationally efficient because it randomly selects a subset of features at each node for splitting, rather than considering all features like in regular decision trees. This reduces the training time and makes it suitable for datasets with a large number of features. It does not require feature scaling, such as normalization or standardization, as it is not affected by the scale of the input features. This saves preprocessing time and effort. It can handle missing data in the input features. It does this by randomly assigning values to the missing data during the training process. It can provide information about the importance of different features in the regression task. By evaluating the splits and the impact of each feature on the overall performance, it can help in feature selection and feature engineering.

3.6.1.2 XGB Regression

A potent machine learning technique used for regression tasks is called XGB Regression, also referred to as XGBoost Regression. Extreme Gradient Boosting, often known as XGBoost, is an improved version of the gradient boosting technique.

XGBoost is known for its high predictive performance. It is designed to handle large datasets and has a parallel and distributed computing capability, making it efficient for processing vast amounts of data. It includes regularization techniques to prevent overfitting, such as L1 and L2 regularization. This helps in reducing model complexity and improves generalization to unseen data. It provides a wide range of options for customization. You can specify various hyperparameters to control the model's behavior, such as learning rate, tree depth, and the number of boosting rounds. This flexibility allows you to fine-tune the model for your specific regression problem.It provides a feature importance analysis, which helps in understanding the relative importance of different features in the dataset. This analysis can assist in feature selection and feature engineering, leading to better model performance. It has built-in handling for missing values. During the tree construction process, it can automatically learn how to handle missing data, eliminating the need for manual imputation.

3.6.1.3 Random Forest Regression

Regression analysis and random forest concepts are used in the machine learning algorithm known as random forest regression. It is suitable for regression tasks since it can predict continuous numeric values. A group of decision trees is assembled in the Random Forest Regression algorithm, and each tree is trained using a random portion of the training data and a random subset of the features. The final forecast is

achieved by averaging, or taking the median, of the guesses from all the trees in the forest. During training, each tree in the forest separately generates predictions.

Random Forest Regression is less prone to overfitting compared to individual decision trees. By using multiple trees and aggregating their predictions, it reduces the impact of outliers and noise in the data. It can capture non-linear relationships between the features and the target variable. It can handle complex interactions and non-linear patterns in the data without requiring explicit feature engineering. It provides a measure of feature importance, which indicates the relative importance of each feature in making predictions. This information can be useful for understanding the underlying patterns in the data and selecting relevant features for future analyses. It can handle missing values in the dataset. During training, if a particular feature has missing values, the algorithm can still use the remaining features to make predictions. It can handle both numerical and categorical features. It can handle high-dimensional datasets and works well with large amounts of data. It is also relatively insensitive to the scale of the features, so there is no need for explicit feature normalization or standardization.

3.6.1.4 KNN Regression

A machine learning approach called KNN (k-nearest neighbors) regression is used for regression tasks where the objective is to predict a continuous numerical value. As an extension of the KNN classification method, it predicts a continuous output value based on the values of its k nearest neighbors rather than a class label.

KNN regression is relatively simple to understand and implement. It does not make any assumptions about the underlying data distribution and can handle both linear and nonlinear relationships. It is a non-parametric algorithm, meaning it does not

make any assumptions about the functional form of the relationship between the input variables and the output. This flexibility allows it to capture complex patterns in the data.

It provides transparency in its predictions. The output value is determined by the average or weighted average of the output values of the k nearest neighbors, making it easy to interpret and understand the reasoning behind the predictions. It is less sensitive to outliers compared to some other regression algorithms. Since the prediction is based on the average or weighted average of the k nearest neighbors, outliers have less impact on the final prediction. In this study, the default value of 5 was used for the k value.

3.6.1.5 Linear Regression

A statistical modeling method called linear regression is employed to determine the relationship between a dependent variable and one or more independent variables. The dependent variable can be predicted or explained by a linear combination of the independent variables, according to the assumption that there is a linear relationship between the variables. Simply said, the goal of linear regression is to identify the straight line that best captures the relationship between the input factors (also known as independent variables) and the output variable (also known as dependent variable).

Linear regression is a straightforward and easy-to-understand method. It provides a simple equation that represents the relationship between variables, making it accessible even to individuals with limited statistical background. The coefficients in the linear regression equation provide valuable insights into the magnitude and direction of the relationship between the independent variables and the dependent

variable. These coefficients can be interpreted to understand the impact of changes in the independent variables on the dependent variable.It can be used for prediction purposes. Once the relationship between variables has been established, the model can be used to predict the values of the dependent variable based on the values of the independent variables.

3.6.1.6 MLP Regression

A regression approach based on artificial neural networks is called MLP regression, commonly referred to as Multilayer Perceptron regression. It is a non-linear regression model that converts continuous output values from input data using numerous layers of interconnected nodes (neurons). MLP regression is a type of feedforward neural network in which data travels from the input layer via the hidden layers to the output layer in a single path.

MLP regression can capture complex non-linear relationships between input and output variables. This makes it suitable for solving regression problems that involve non-linear patterns. It can handle a wide range of input data types, including continuous, categorical, and binary variables. It can also handle multi-dimensional input data. It can learn and model intricate relationships in the data, making it capable of capturing both simple and complex patterns. The hidden layers in MLP allow for the extraction of high-level features from the input data. It has the ability to generalize from the training data to make predictions on unseen data. It can learn from examples and make predictions on new input data, even in the presence of noise or missing values.

3.6.2 Selection of the Most accurated Regression Model

The best Regression algorithm was selected with K-Fold Cross Validation. Here, after taking the average of the WMAE values obtained by each algorithm for each fold, the algorithm with the lowest average was selected.

3.6.2.1 K-Fold Cross Validation

K-Fold Cross Validation is a statistical and machine learning technique for assessing the efficacy of a predictive model. It entails dividing the dataset into k roughly equal-sized sections or folds. The model is trained and tested k times, with each evaluation utilizing the remaining folds as the training set and a different fold as the validation set. A final performance estimate of the model is then generated by averaging the outcomes from each iteration.

K-Fold Cross Validation provides a more reliable estimate of the model's performance compared to a single train-test split. It allows for a more comprehensive evaluation by using multiple validation sets, reducing the dependency on a specific split of the data. By utilizing all available data for both training and validation, K-Fold Cross Validation maximizes the use of the dataset. This is particularly useful when the dataset is limited or when the model's performance needs to be evaluated with a high degree of confidence. K-Fold Cross Validation helps to reduce bias in performance evaluation by averaging the results across multiple iterations. It mitigates the impact of any particular data split that may skew the performance evaluation.

K-Fold Cross Validation can aid in comparing and selecting between different models. By evaluating each model's performance on the same folds, it enables a fair comparison and facilitates informed decision-making regarding the choice of the best

model. More than one regression model was used in this study, and K-Fold Cross Validation is used for model selection. In the study, the k value was taken as 5. For each model, the average of the values taken in all folds was taken. The model with the lowest mean was selected and used as the most accurate regression model.

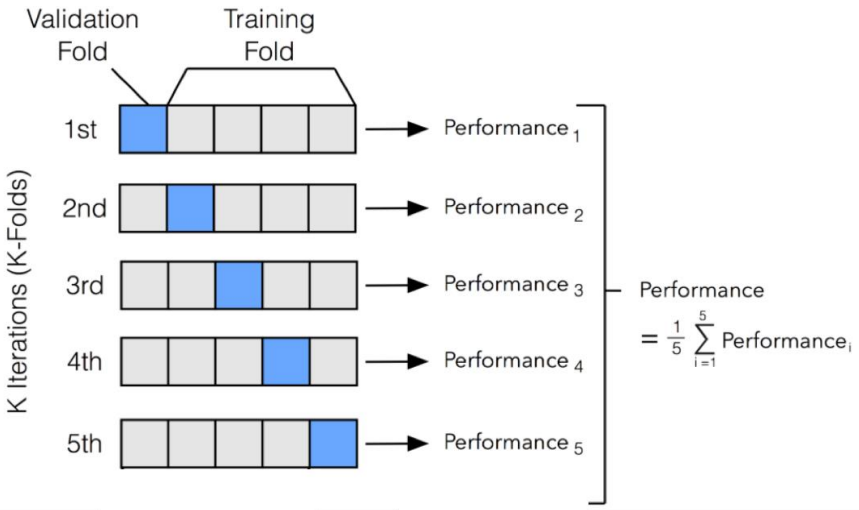


Figure 30: K-Fold Cross Validation (k=5)

3.6.3 Selected Time Series Algorithm

3.6.3.1 Prophet

The Prophet algorithm is a time series forecasting algorithm developed by Facebook's Core Data Science team. It is based on a generalized additive model (GAM) framework, which is a type of regression model that combines multiple predictor variables to make predictions. It is designed to provide accurate and efficient forecasts for time series data that exhibit various trends and seasonal patterns.

Prophet can handle a wide range of time series data, including those with missing values, outliers, and irregular sampling intervals. It accommodates both daily and

sub-daily observations and can incorporate multiple seasonalities. It automatically detects recurring patterns and seasonality in the data, making it convenient for forecasting time series with complex seasonal patterns. It can capture weekly, monthly, yearly, and even custom seasonalities. It uses a flexible trend model that can capture both linear and non-linear trends. It incorporates user-specified changepoints, which are time points where the trend undergoes significant shifts. This allows the algorithm to adapt to changes in the underlying patterns of the time series.

Prophet provides uncertainty estimation for its forecasts, allowing users to quantify the range of possible outcomes. This is achieved by modeling the intrinsic uncertainty in the data and accounting for the presence of outliers and irregularities. It decomposes the time series into several components, including trend, seasonality, and holiday effects. This decomposition enables users to understand the individual contributions of each component to the overall forecast, aiding in interpretation and analysis.

3.6.4 Hybrid Modelling

The Hybrid Modeling stage was actually found by averaging the estimates obtained by the best regression algorithm and the time series algorithm. The datasets of the predictions based on Store, Department and Date from the best regression and Prophet algorithm were combined, and then the regression predictions and time series predictions were averaged. The formula used for the average is given below.

$$(\text{Regression\_Prediction} + \text{TimeSeries\_Prediction}) / 2 \quad (2)$$

The rationale behind averaging is to take advantage of the strengths of both methods and potentially reduce the impact of the shortcomings of any individual estimates. Regression analysis captures the relationships between variables and considers the

influence of additional factors, while time series analysis captures temporal dependencies and patterns and focuses on modeling these patterns. Therefore, by averaging the estimates, we created a mixed estimate that includes both types of information. The results obtained using the methods in this chapter, provided in the next chapter of this thesis.

3.6.5 Evaluation of Models

The accuracy of the models proposed in the competition held by Walmart was requested to be made with the WMAE performance metric. Therefore, WMAE was used in this study. In addition, according to the literature review, MAE and RMSE performance metrics, which are the most used in other studies, were also used to evaluate the accuracy of the models in our proposed method and to compare them with other studies.

3.6.5.1 Mean Absolute Error

The average magnitude of errors between expected and actual values is measured using the metric known as mean absolute error, which is frequently used in statistics and machine learning. The average of the absolute differences between the predicted and the actual values is calculated by MAE. MAE is calculated as follows:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|, \quad (3)$$

where,

n = number of the data points,

y_i = actual value of the target variable for the ith data point,

$\hat{y}_i$ = predicted value of the target variable for the ith data point.

3.6.5.2 Weighted Mean Absolute Error

The Mean Absolute Error (MAE) and Weighted Mean Absolute Error are similar. The average magnitude of errors between predicted and actual values is gauged using

the metric known as WMAE. WMAE, on the other hand, introduces the idea of giving certain data points varying weights, allowing you to emphasize particular data points more than others in the error computation. WMAE is calculated as follows:

$$\text{WMAE}=\frac{1}{\sum_{i=1}^{n}\omega_i}\sum_{i=1}^{n}\omega_i\,.\left|y_i-\hat{y}_i\right|, \tag{4}$$

where,

n = number of the data points,

y_i = actual value of the target variable for the ith data point,

$\hat{y}_i$ = predicted value of the target variable for the ith data point,

ω_i = weight assigned to the ith data point. If the week is a holiday week, weight is 5 and 1 otherwise.

3.6.5.3 Root Mean Square Error

One of the most common metrics used to gauge how accurately our forecasting model's predicted values compare to the real or observed values occurs during the training of regression models or time series models is the root mean square error. RMSE stands for root mean square error, often known as residuals. It shows how dispersed the data is around the line of best fit. RMSE is calculated as follows:

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}, \tag{5}$$

where,

n = number of the data points,

y_i = actual value of the target variable for the ith data point,

$\hat{y}_i$ = predicted value of the target variable for the ith data point.

Chapter 4

RESULTS AND FINDINGS

4.1 Selected Most Accurate Regression Model

Many performance metrics were used to compare the results at different methods, however WMAE metric was mainly used to measure the performance of different method in this study. Smaller WMAE value means, higher accuracy. In the regression model selection, the model with the lowest WMAE was chosen as the most accurate model. The results are as shown in Figure 31. At the end of K-Fold Cross Validation, all WMAE values were averaged and overall performance was determined. When the average errors of each model were examined, the Extra Tree Regression model had the lowest error rate, as shown in Figure 32.

	Algorithm	modelName	Fold	MAE	MSE	RMSE	R2score	WMAE	WMSE	WRMSE	WR2score
0	0.0	extraTreesRegressor	0.0	1603.647051	1.230556e+07	3507.928555	3507.928555	1665.822825	1.303390e+07	3610.249726	3610.249726
1	0.0	extraTreesRegressor	1.0	1638.518341	1.262826e+07	3553.626325	3553.626325	1627.962159	1.209595e+07	3477.922922	3477.922922
2	0.0	extraTreesRegressor	2.0	2181.500350	3.256426e+07	5706.509879	5706.509879	2712.654125	5.149053e+07	7175.690106	7175.690106
3	0.0	extraTreesRegressor	3.0	1648.276701	1.495924e+07	3867.717425	3867.717425	1621.367763	1.400379e+07	3742.163567	3742.163567
4	0.0	extraTreesRegressor	4.0	2151.850230	2.653906e+07	5151.607219	5151.607219	2615.995084	4.280557e+07	6542.596320	6542.596320
5	1.0	XGBRegression	0.0	3024.517563	2.949914e+07	5431.310678	5431.310678	3132.186706	3.114494e+07	5580.764910	5580.764910
6	1.0	XGBRegression	1.0	2827.492222	2.281427e+07	4776.428975	4776.428975	2874.227227	2.390578e+07	4889.353814	4889.353814
7	1.0	XGBRegression	2.0	3634.409551	5.426276e+07	7366.325798	7366.325798	4460.243327	9.584686e+07	9790.141203	9790.141203
8	1.0	XGBRegression	3.0	2797.906759	2.353333e+07	4851.116107	4851.116107	2866.263910	2.475221e+07	4975.158808	4975.158808
9	1.0	XGBRegression	4.0	3713.714118	4.853850e+07	6966.957805	6966.957805	4322.039038	7.356621e+07	8577.074536	8577.074536
10	2.0	randomForestRegressor	0.0	6799.401128	1.493362e+08	12220.320332	12220.320332	6880.787998	1.516894e+08	12316.226505	12316.226505
11	2.0	randomForestRegressor	1.0	5947.633186	1.065086e+08	10320.302390	10320.302390	6000.881061	1.081668e+08	10400.328435	10400.328435
12	2.0	randomForestRegressor	2.0	7150.137066	2.085560e+08	14441.467386	14441.467386	8151.159290	3.222187e+08	17950.450910	17950.450910
13	2.0	randomForestRegressor	3.0	5865.273661	1.027188e+08	10135.026764	10135.026764	5910.637600	1.029504e+08	10146.448856	10146.448856
14	2.0	randomForestRegressor	4.0	7997.072589	2.340744e+08	15299.489019	15299.489019	9031.581912	3.463265e+08	18609.850882	18609.850882
15	3.0	knn	0.0	8372.198240	1.960839e+08	14002.996874	14002.996874	8498.774155	2.007115e+08	14167.268048	14167.268048
16	3.0	knn	1.0	7989.890676	1.790044e+08	13379.252262	13379.252262	7994.350389	1.797359e+08	13406.563147	13406.563147
17	3.0	knn	2.0	9710.740814	3.169571e+08	17803.289915	17803.289915	10905.548578	4.789684e+08	21885.347357	21885.347357
18	3.0	knn	3.0	8165.902994	1.814562e+08	13470.568662	13470.568662	8238.706028	1.841376e+08	13569.730241	13569.730241
19	3.0	knn	4.0	9386.020819	2.929256e+08	17115.069716	17115.069716	10478.308343	4.305083e+08	20748.693356	20748.693356
20	4.0	linearRegression	0.0	13157.192647	4.207198e+08	20511.454739	20511.454739	13370.587226	4.316312e+08	20775.736802	20775.736802
21	4.0	linearRegression	1.0	13494.627899	3.924781e+08	19811.059521	19811.059521	13517.899872	3.935456e+08	19837.982957	19837.982957
22	4.0	linearRegression	2.0	14161.143494	5.335439e+08	23098.569986	23098.569986	15539.842914	7.188317e+08	26811.036264	26811.036264
23	4.0	linearRegression	3.0	13422.490311	4.027009e+08	20067.408585	20067.408585	13578.280621	4.072854e+08	20181.311936	20181.311936
24	4.0	linearRegression	4.0	14320.721963	5.289468e+08	22998.843861	22998.843861	15470.322279	6.883143e+08	26235.744782	26235.744782
25	5.0	nn-Multi-layer Perceptron regressor	0.0	11090.519927	3.666821e+08	19148.944243	19148.944243	11245.084279	3.748403e+08	19360.792659	19360.792659
26	5.0	nn-Multi-layer Perceptron regressor	1.0	12133.765659	3.376758e+08	18375.956087	18375.956087	12164.392320	3.376410e+08	18375.011412	18375.011412
27	5.0	nn-Multi-layer Perceptron regressor	2.0	13970.674333	5.384504e+08	23204.533584	23204.533584	15025.498925	7.312088e+08	27040.872042	27040.872042
28	5.0	nn-Multi-layer Perceptron regressor	3.0	13393.368997	4.034431e+08	20085.892269	20085.892269	13465.795850	4.068276e+08	20169.967936	20169.967936
29	5.0	nn-Multi-layer Perceptron regressor	4.0	14377.058257	5.320664e+08	23066.564464	23066.564464	15266.623980	6.987043e+08	26433.015005	26433.015005

Figure 31: K-Fold Cross Validation (k=5) results of each fold for Regression Model Selection

Algorithm	modelName	MAE	MSE	RMSE	R2score	WMAE	WMSE	WRMSE	WR2score
0.0	extraTreesRegressor	1844.758535	1.979927e+07	4357.477881	4357.477881	2048.760391	2.668595e+07	4909.724528	4909.724528
1.0	XGBRegression	3199.608043	3.572960e+07	5878.427872	5878.427872	3530.992042	4.984320e+07	6762.498654	6762.498654
2.0	randomForestRegressor	6751.903526	1.602388e+08	12483.321178	12483.321178	7195.009572	2.062704e+08	13884.661118	13884.661118
3.0	knn	8724.950709	2.332855e+08	15154.235486	15154.235486	9223.137499	2.948123e+08	16755.520430	16755.520430
4.0	linearRegression	13711.235263	4.556779e+08	21297.467338	21297.467338	14295.386582	5.279216e+08	22768.362548	22768.362548
5.0	nn-Multi-layer Perceptron regressor	12993.077435	4.356635e+08	20776.378130	20776.378130	13433.479071	5.098444e+08	22275.931811	22275.931811

Figure 32: Average Performance Ratios for Regression Models

4.2 Prediction Results of Test Data

For the test results, firstly, the estimation results were taken from the best regression algorithm. Then, using the time series model, store-department based estimation results were obtained. Finally, all estimation results were combined and a hybrid result was obtained. Table 4 shows the regression, time series and hybrid forecasting error rates for 2012 test data.

Table 4: Regression, Time Series and Hybrid Forecasting WMAE and MAE Results

Type	WMAE	MAE
Regression	1688.7396635036057	1661.2677921432335
Time Series	2068.8280996671306	1986.2230297036044
Hybrid	1592.711355105404	1536.3057337044474

According to Table 4, the hybrid estimation results obtained using the regression and time series estimation results are more accurate and have a low error rate.

4.3 Visualization of Forecasting Results

4.3.1 Analysis of Average Sales per Store

When Figure 33 is examined, there is not much difference in store-based average sales. The patterns created by the average sales forecasts and actual sales of each store are very similar.

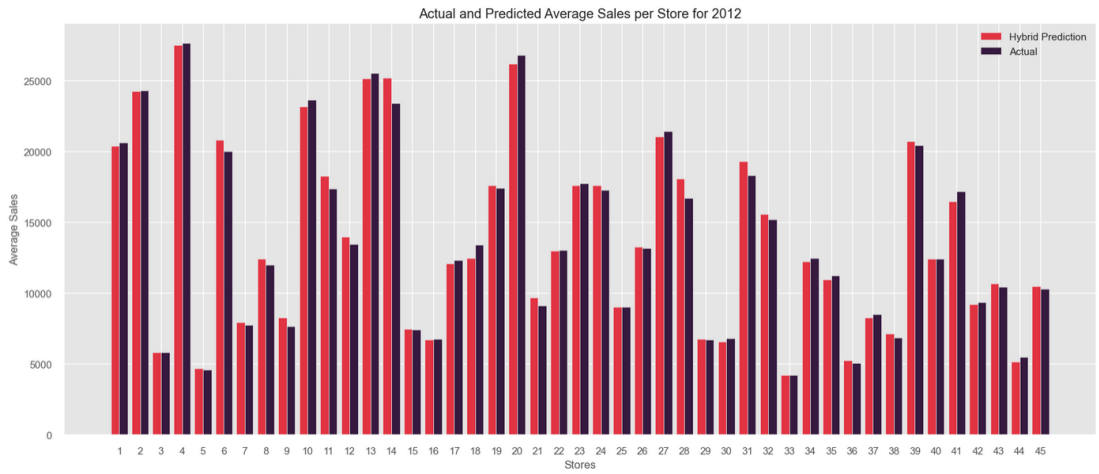


Figure 33: Actual and Predicted Average Sales per Store for 2012

4.3.2 Analysis of Monthly Sales based on Store-Department

Figures 34, 35, 36, 37 and 38 are the estimation results for some Store-Departments whose data came in healthy for 143 weeks. These graphs contain real, time series estimation, regression estimation and hybrid estimation results. The pink area seen in the graphs is the area of the lower and upper value ranges produced by the time series.

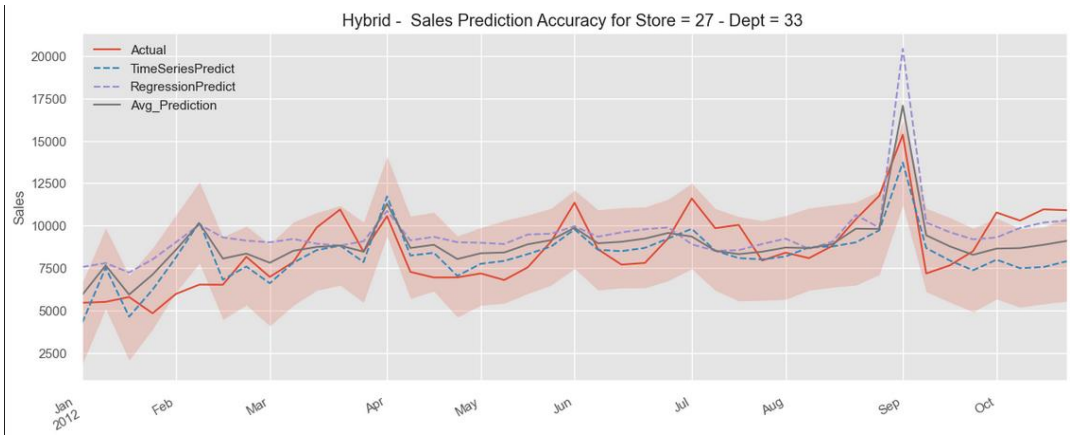


Figure 34: Monthly Actual and Predicted Sales Visualisation for Store 27 and Department 33

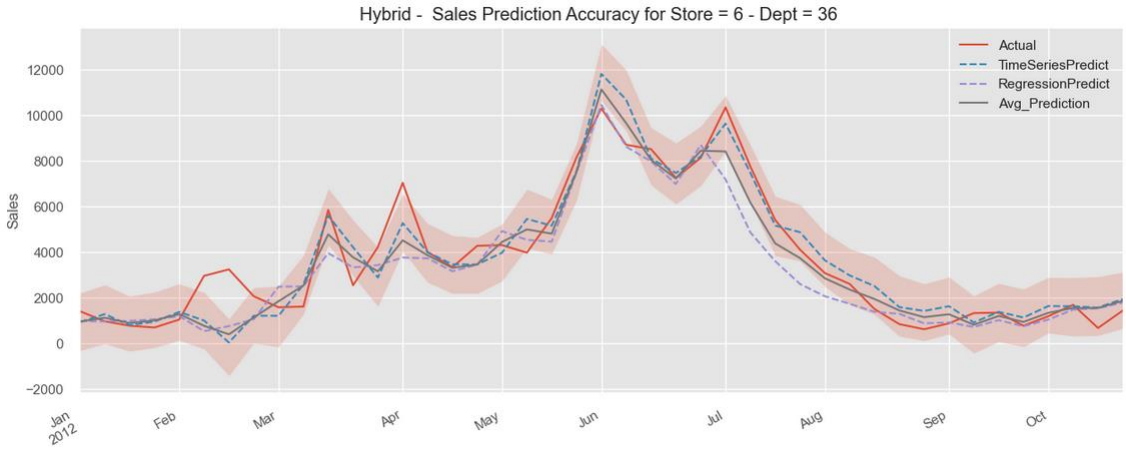


Figure 35: Monthly Actual and Predicted Sales Visualisation for Store 6 and Department 36

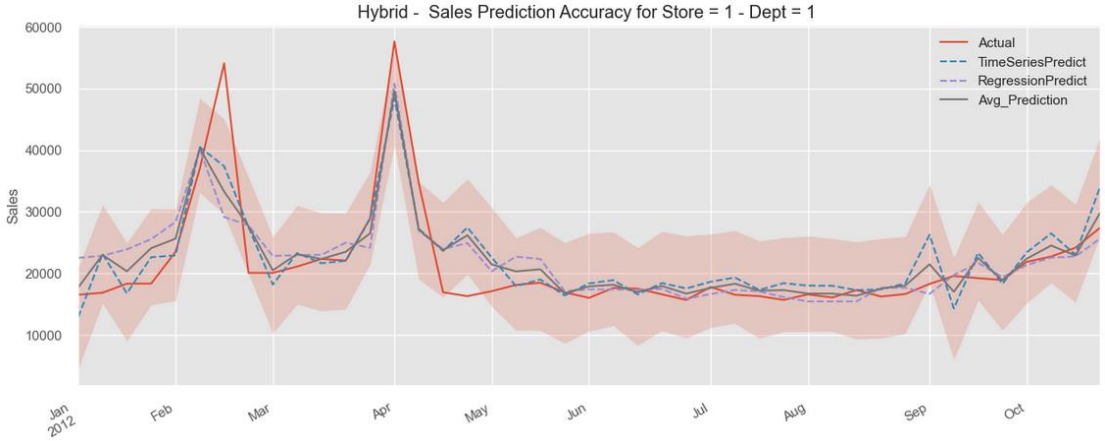


Figure 36: Monthly Actual and Predicted Sales Visualisation for Store 1 and Department 1

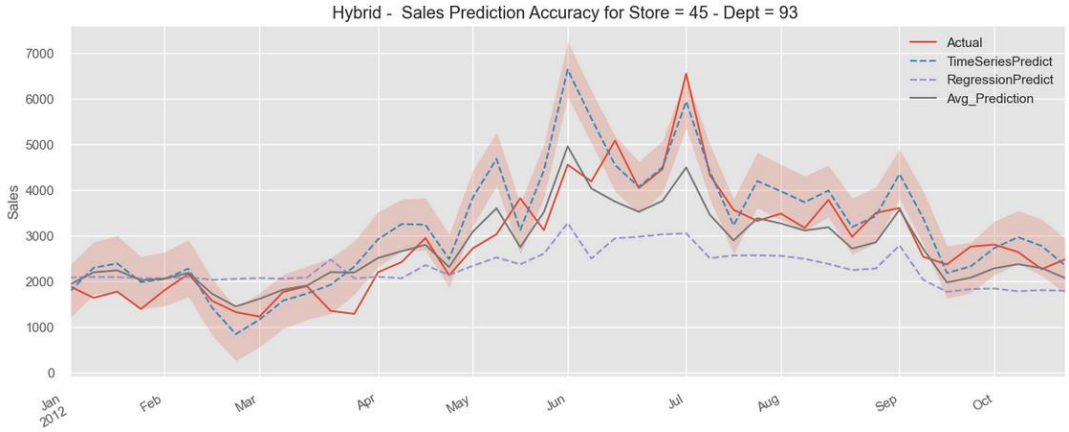


Figure 37: Monthly Actual and Predicted Sales Visualisation for Store 45 and Department 93

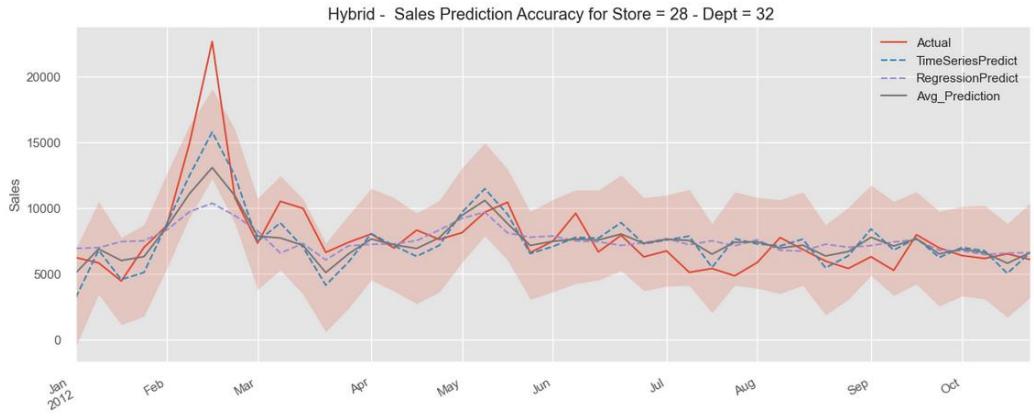


Figure 38: Monthly Actual and Predicted Sales Visualisation for Store 28 and Department 32

4.3.3 Analysis of Actual an Predicted Monthly Sales

In Figures 39, 40 and 41, monthly visualizations of regression, time series and hybrid estimation results with real values are given. When these graphs are examined, it is seen that with hybrid estimation, monthly sales estimation generates lower error rate and more accurate results.

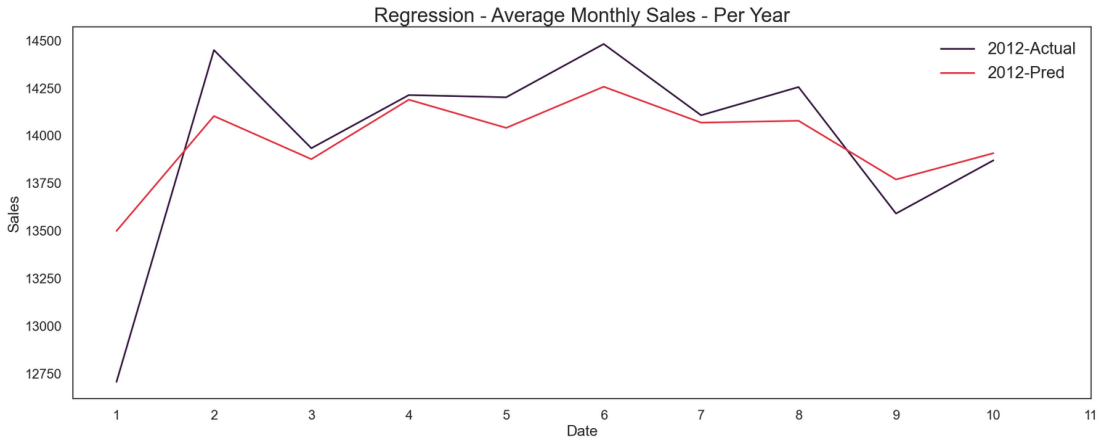


Figure 39: Visualisation of Actual Sales and Regression Forecast Values as Monthly

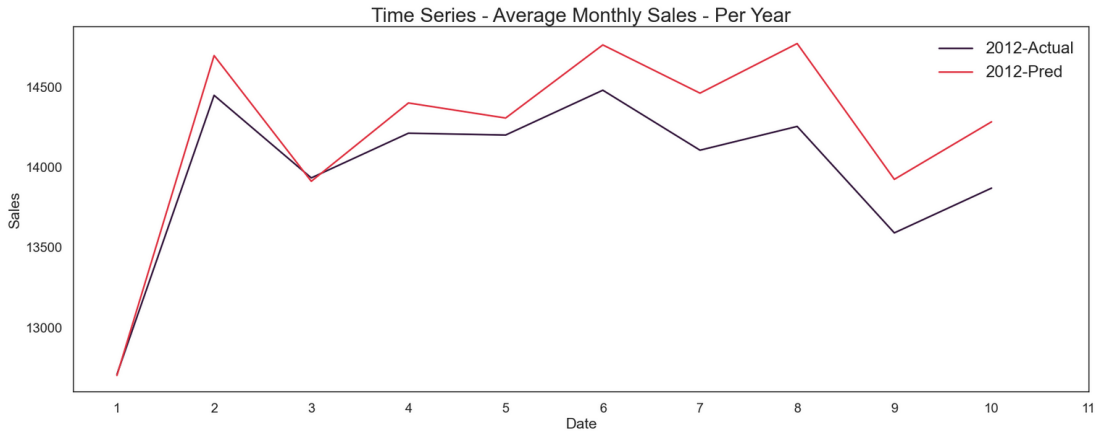


Figure 40: Visualisation of Actual Sales and Time Series forecast values as Monthly

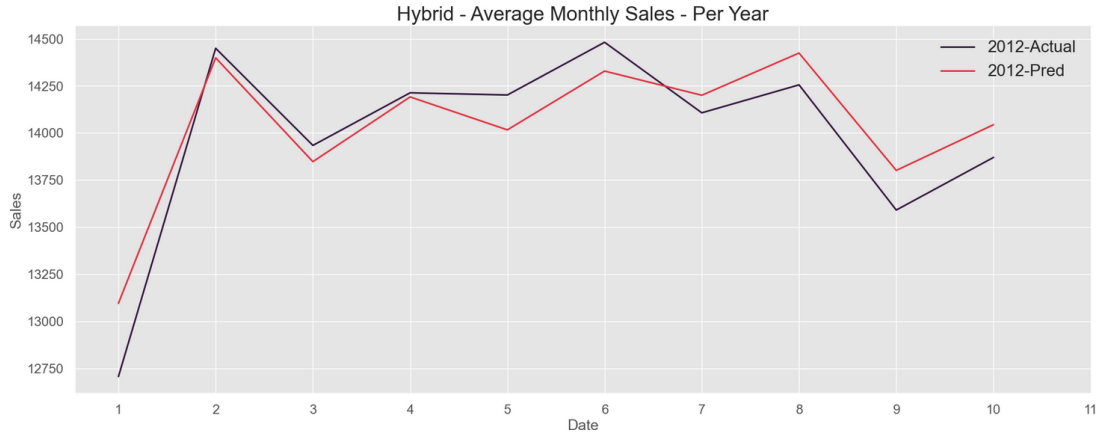


Figure 41: Visualisation of Actual Sales and Hybrid Forecast Values as Monthly

4.3.4 Analysis of Other Studies on Same Data Set

The performance results of the studies conducted on the Walmart data set, which we mentioned in the literature reviews in Chapter 2, were examined. The list of studies whose performance results are clearly stated is given in Table 5.

Table 5: Information About other Forecasting Studies on Walmart Data Set

Article	**Most Accurate Algorithm**	**About test and train sets**	**Performance**	**Hyperparameter Tuning**
[18]	Random Forest Regression	Train = %80, Test = %20	MAE = 2573	Applied
[17]	XGBoost Regression	Train = %70, Test = %30	RMSE = 3477.1 MAE = 1317.65	Applied
[20]	Boosted Decision Tree Regression	Data of unknown Single Store and its All departments	MAE = 1669.10 RMSE = 3696.59	Applied

Chapter 5

CONCLUSION

5.1 Overall Results

The main purpose of this study was to reveal the method that will provide more accurate predictions by selecting the most appropriate machine learning algorithms and parameters according to the data set to be selected. In this study, sales data of Walmart, one of the world's largest companies and retail sales company, were used.

With the Exploratory Data Analysis, the features to be used in the models were determined. The variables Store, Dept, Weekly_Sales, IsHoliday, Size, Temperature, CPI, Unemployment, Year, WeekOfYear, Holiday_Type_Christmas, Holiday_Type_Easter_Sunday, Holiday_Type_Labor_Day, Holiday_Type_Super_Bowl, Holiday_Type_Thanksgiving, Type_A, Type_B, Type_C were used in the regression algorithms. Year and WeekOfYear properties were extracted from Date property. Holiday and Type properties have been converted to indicator columns. In the Time Series algorithm, Date, Weekly_Sales, IsHoliday, Temperature, CPI, Unemployment and Holiday Days were used in the algorithm.

Hybrid Modeling method was used in this study. Regression methods capture the relationships between variables, while time series estimation methods take into account temporal dependencies and patterns in the data. By combining these two

approaches, it is aimed to increase the accuracy of the predictions by utilizing both information sets. In this study, Linear Regression, Random Forest Regression, KNN Regression, Extra Tree Regression, XGBoost Regression and MLP Regression methods were used and Extra Tree Regression was chosen as the most accurate method by using K-Fold Cross Validation. As the Time Series algorithm, only the Prophet method was used and the results from both algorithms were combined.

As a result, 1592.71 WMAE and 1536.306 MAE ratios were obtained for 30% test data with the proposed method. Compared to the studies in Table 5 given in Chapter 4, our proposed method performed better than the study in [18] and [20] without hyperparameter tuning. When the study in [17] is examined, there is a difference of approximately 200 MAE, and our proposed study presented slightly less accuracy.

5.2 Future Work

Before running a training task, hyperparameters are changes that can be applied to the behavior of an ML algorithm. They have a major impact on model training in terms of training time, infrastructure resource requirements (and associated cost), model convergence, and model correctness. The machine learning technique requires the hyper parameter tuning step. A model's ability to achieve the intended metric value can really be attributed to the hyperparameters that are chosen [21]. As a future work, hyperparameter tuning will be done on the algorithms used in this study.

REFERENCES

[1] Molodoria, A. (2022, October 26). How to apply machine learning to demand & sales forecasting in retail. Retrieved from https://mobidev.biz/blog/machine-learning-methods-demand-forecasting-retail

[2] Ali, M. A. (2022, March 30). Retail demand forecasting. Retrieved from https://scholarworks.rit.edu/theses/11093

[3] Žunić, E., Korjenić, K., Delalić, S., & Šubara, Z. (2021). Comparison analysis of Facebook's Prophet, Amazon's DeepAR+ and CNN-QR algorithms for successful real-world sales forecasting. *International Journal of Computer Science and Information Technology*, 13(2), 67–84. https://doi.org/10.5121/ijcsit.2021.13205

[4] 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. https://doi.org/10.1016/j.eswa.2010.04.072

[5] Mentzer, J. T., & Bienstock, C. (1998). *Sales Forecasting Management: Understanding the Techniques, Systems and Management of the Sales Forecasting Process*.

[6] Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine learning in predicting demand for fast-moving consumer

goods: an exploratory research. IFAC-PapersOnLine, 52(13), 737–742. https://doi.org/10.1016/j.ifacol.2019.11.203

[7] Taranenko, L. (2021,Jul 7). How to apply machine learning to demand forecasting. Retrieved from https://mobidev.biz/blog/machine-learning-methodsdemand-forecasting-retail

[8] Molodoria, A. (2022, August 18). 5 essential machine learning algorithms for business applications. Retrieved from https://mobidev.biz/blog/essential-machine-learning-algorithms-for-business-applications

[9] Tugay, R., & Oguducu, S. G. (2022). Demand prediction using machine learning methods and stacked generalization. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2009.09756

[10] Arif, M. A. I., Sany, S. I., Nahin, F. I., & Rabby, A. S. A. (2019). Comparison study: product demand forecasting with machine learning for shop. *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*. https://doi.org/10.1109/smart46866.2019.9117395

[11] Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*, 10, 184797901880867. https://doi.org/10.1177/1847979018808673

[12] Brownlee. (2020, November 2). Random forest for time series forecasting. Retrieved June 6, 2023, from https://machinelearningmastery.com/random-forest-for-time-series-forecasting/

[13] Arunraj, N. S., Ahrens, D., & Fernandes, M. (2016). Application of SARIMAX model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems*, 7(2), 1–21. https://doi.org/10.4018/ijoris.2016040101

[14] Polamuri*, S. R., Srinivasi, D. K., & Mohan, D. A. K. (2019). Stock market prices prediction using random forest and extra tree regression. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(3), 1224–1228. https://doi.org/10.35940/ijrte.c4314.098319

[15] Harsoor, A., & Patil, A. (2015). Forecast of sales of walmart store using big data applications. *International Journal of Research in Engineering and Technology*, 04(06), 51–59. https://doi.org/10.15623/ijret.2015.0406008

[16] Gurudevi, R. Patil (2022), Wal-Mart sales forecasting using machine learning. *International Journal of Creative Research Thoughts (IJCRT)*, ISSN:2320-2882, Volume.10, Issue 8, pp.b460-b463, Available at : http://www.ijcrt.org/papers/IJCRT2208184.pdf

[17] Akande, Y.F., Idowu, J., Misra, A., Misra, S., Akande, O.N., Ahuja, R. (2022). Application of XGBoost algorithm for sales forecasting using walmart dataset. In: Sengodan, T., Murugappan, M., Misra, S. (eds) Advances in

Electrical and Computer Technologies. Lecture Notes in Electrical Engineering, vol 881. Springer, Singapore. https://doi.org/10.1007/978-981-19-1111-8_13

[18] Colón, J.G. (2019). *Data Mining Techniques and Machine Learning Model for Walmart Weekly Sales Forecast*.

[19] Raizada, S., & Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting. *International Journal of Advanced Computer Science and Applications*, 12(11). https://doi.org/10.14569/ijacsa.2021.0121112

[20] Catal, C., Ece, K., Arslan, B., & Akbulut, A. (2019). Benchmarking of regression algorithms and time series analysis techniques for sales forecasting. *Balkan Journal of Electrical and Computer Engineering*, 7(1), 20–26. https://doi.org/10.17694/bajece.494920

[21] Kasture, N. (2020, November 16). Why hyper parameter tuning is important for your model ? Retrieved from https://medium.com/analytics-vidhya/why-hyper-parameter-tuning-is-important-for-your-model-1ff4c8f145d3