RETAIL DATA PREDICTIVE ANALYSIS USING MACHINE LEARNING MODELS

Capstone Project

Müjde Güner

İSTANBUL, 2020

MEF UNIVERSITY

RETAIL DATA PREDICTIVE ANALYSIS USING MACHINE LEARNING MODELS

Capstone Project

Müjde Güner

Advisor: Asst. Prof. Dr. Tuna Çakar

İSTANBUL, 2020

MEF UNIVERSITY

Name of the project: Retail Data Predictive Analysis Using Machine Learning Models
Name/Last Name of the Student: Müjde Güner
Date of Project Report Submission: 30/12/2020

I hereby state that the graduation project prepared by Müjde Güner has been completed under my supervision. I accept this work as a "Graduation Project".

30/12/2020
Asst. Prof. Dr. Tuna Çakar

I hereby state that I have examined this graduation project by Müjde Güner which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

30/12/2020

Director
of
Information Technologies Program

We hereby state that we have held the graduation examination of Müjde Güner and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature /Date
1. Asst. Prof. Dr. Tuna Çakar	……..
2. ……	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Müjde Güner	30/12/2020	

EXECUTIVE SUMMARY

RETAIL DATA PREDICTIVE ANALYSIS
USING MACHINE LEARNING MODELS

Müjde Güner

Advisor: Asst. Prof. Dr. Tuna Çakar

DECEMBER, 2020, 39 pages

Machine Learning (ML) is a popular field which deals with training the system with data (experience), performing some task (regression or classification) and evaluating the system with the desired performance metrics. ML automatically extracts useful and meaningful insights from the data. ML models for sales prediction applies computational intelligence in many real world applications such as stock market, production, economics, weather, retail, census analysis and so on. Sales prediction can be viewed as a regression problem and various algorithms can be applied.

In this project, real life data analysis has been done to predict the sales for four categories of products like Cold Cereal, Bag Snacks, Oral Hygiene Products, and Frozen Pizza. Exploratory Data Analysis (EDA) has been applied to the dataset to make exact predictions even during an unpredictable environment. The different phases of EDA used in this project are Data Preprocessing and Analysis, Feature Selection and Feature Extraction, Model Building and Regression Analysis, Clustering, Time Series Analysis and Model Evaluation using the Performance Metrics.

For outlier detection, InterQuartile Range (IQR) method is used. For Filter Based Feature Selection, Univariate Feature Analysis using SelectK-Best and SelectPercentile, Decision Tree Regressor method has been used. For Wrapper Based Feature Selection, Sequential Feature Selector method has been deployed.

For Regression Analysis, various algorithms such as Linear Regression, XGBoost Regression and Support Vector Regression (SVR) are analyzed. K-Means Clustering Algorithm has been used on the dataset to generate 4 different clusters. In Time Series Analysis, the week end date and average weekly basket attributes are analyzed, and the sequential data has been rendered for a given time period of occurrence.

In model evaluation phase, the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 and Adjusted R2 accuracy has been calculated and validated.

The project has been implemented in an open source software called Anaconda which includes Jupyter Notebook platform for scientific computations. Python programming language with different packages such as Numpy, Pandas, Scikit learn has been used.

Key Words: Machine Learning, Sales Prediction, Time Series, Predictive Analytics.

ÖZET

RETAIL DATA PREDICTIVE ANALYSIS
USING MACHINE LEARNING MODELS

Müjde Güner

Proje Danışmanı: Dr. Öğr. Üyesi Tuna Çakar

ARALIK, 2020, 39 sayfa

Makine Öğrenmesi (ML), sistemi verilerle eğitmek, bazı görevleri yerine getirmek (regresyon veya sınıflandırma) ve sistemi istenen performans ölçütleriyle değerlendirmekle ilgilenen popüler bir alandır. Makine Öğrenmesi, verilerden otomatik olarak yararlı ve anlamlı bilgiler çıkarır. Makine Öğrenmesi modelleri, satış tahmini, borsa, üretim, ekonomi, hava durumu, perakende gibi çeşitli alanlarda önemli bir rol oynar. Satış tahmini bir regresyon problemi olarak görülebilir ve çeşitli algoritmalar uygulanabilir.

Bu projede, kahvaltılık gevrek, poşet atıştırmalıklar, ağız hijyeni ürünleri ve dondurulmuş pizza gibi dört ürün kategorisinin satışlarını tahmin etmek için gerçek hayat veri analizi yapılmıştır. Öngörülemeyen bir ortamda bile kesin tahminler yapmak için veri setine Keşifsel Veri Analizi (EDA) uygulanmıştır. Bu projede kullanılan EDA'nın farklı aşamaları, veri ön işleme (Data Preprocessing) ve analizi, özellik seçimi (Feature Selection) ve özellik çıkarma (Feature Extraction), model oluşturma ve regresyon analizi, kümeleme (Clustering), zaman serisi analizi (Time Series Analysis) ve performans ölçütlerini kullanarak model değerlendirmedir.

Aykırı değer (Outlier) tespiti için InterQuartile Range (IQR) yöntemi kullanılır. Filtre tabanlı özellik seçimi (Filter Based Feature Selection) için, SelectK-Best ve SelectPercentile kullanılarak tek değişkenli (Univariate) özellik analizi, karar ağacı regresör (Decision Tree Regressor) yöntemi kullanılmıştır. Sarmalayıcı tabanlı (Wrapper Based) özellik seçimi için, sıralı özellik seçici (Sequential Feature Selection) yöntemi devreye alınmıştır.

Regresyon analizi için Lineer Regresyon, Destek Vektör Regresyonu (Support Vector Regression), XGBoost Regresyon modelleri analiz edilir. Kümeleme (Clustering) için veri setine K-Means kümeleme algoritması uygulanmış ve toplam 4 farklı küme elde edilmiştir. Zaman Serisi analizinde, hafta bitiş tarihi ve ortalama haftalık sepet (Average Weekly Baskets özellikleri analiz edilir ve sıralı veriler, belirli bir oluşum süresi için oluşturulur.

Model değerlendirme aşamasında, Ortalama Mutlak Hata (Mean Absolute Error), Ortalama Kare Hata (Mean Squared Error), Ortalama Kareli Hata (Root Mean Squared Error), R2 ve Ayarlanmış R2 (Adjusted R2) doğruluk hesaplanmış ve doğrulanmıştır.

Proje, bilimsel hesaplamalar için Jupyter Notebook platformunu içeren Anaconda adlı açık kaynaklı bir yazılımda uygulanmıştır. Numpy, Pandas, Scikit gibi farklı paketler ile Python programlama dili kullanılmıştır.

Anahtar Kelimeler: Makine Öğrenmesi, Satış Tahmini, Zaman Serileri, Tahminsel Analitik

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Overview of Project

Machine Learning (ML) is a widely popular field for training the system without explicitly programming. ML derives meaningful, useful, hidden insights and patterns from data. Time can be considered as an important factor of prediction and forecasting. Time series deals with the arrangement of observations in sequential manner based on time. Time Series Forecasting is a powerful forecasting tool which predicts future events or patterns based on past events or patterns. Patterns are entirely based on the historical data. Time series modeling works on time-based data to derive meaningful, useful insights to make informed decisions. Sales prediction or forecasting derives information about the future sales by taking expenses, profit and growth into account.

1.1.1. Sales Prediction Using Machine Learning

ML techniques renders most accurate predictions on sales data. For example, many business people do analysis with the time series data to analyze sales for next year or future events, website traffic or competition among their peers. The business people also make use of time series models for early prediction of sales and inventory in order to avoid overfilling and underfilling of materials.

In recent days, most of the global companies are taking the necessary preventive actions by forecasting the sales using ML models. Sales forecasting is also necessary to project the future budgets of the company. In today's highly competitive environment, companies need to predict their sales in order to secure their success for the future by taking preventive actions against possible sales loss.

Some of the ML models for sales prediction includes Linear Regression model, Support Vector Regression (SVR), Decision Tree Regression, ensemble models like XGBoost Regression, Random Forest Regression, Lasso Regression, K-Nearest Neighbour regression, Gradient Boost Algorithm and so on. Some of the Deep Learning models [6] for sales prediction includes Neural Network Regression, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN).

The dataset used for analysis in this proposed project is Breakfast at the Frat dataset. The dataset consists of sales and promotion information over a period of 156 weeks. It contains 25 columns and 538643 rows. The different product categories are Cold Cereal, Bag Snacks, Oral Hygiene Products and Frozen Pizza.

1.2. Problem Statement

Nowadays, most of the companies are facing difficulty in large scale data analysis and visualization. In an unpredictable environment, sales prediction plays a critical role for the companies to be successful in a highly competitive environment. So, this project aims in developing a model for sales prediction and time series analysis using ML techniques to forecast the sales and helps the companies in analyzing huge volumes of data.

1.3. Objectives

The ultimate aim of this proposed work is to design and develop a ML model with different phases of EDA and perform sales predictive analysis for time series data.

- To generate preprocessed dataset by performing data transformation, missing data analysis, correlation analysis and outlier detection.
- To select the highly relevant features from the dataset using filter based and wrapper-based feature selection methods.
- To build a ML model and perform regression analysis using the algorithms such as Linear Regression, Support Vector Regression and XGBoost Regression.
- To cluster the dataset using K-Means clustering technique for better visualization and analysis.
- To perform Time Series Analysis and generate Time Series graphs.
- To evaluate the developed ML model using the performance metrics.

1.4. Contributions

The major contributions of the proposed work can be summarized as follows:

- Large scale data analysis and visualization.
- Feature engineering by adding new feature to the dataset such as discount.
- Time series analysis
- Accuracy analysis of the ML models

1.5. Organization of Report

The rest of the report is structured as follows Section 2 discusses about the works related to sales prediction and time series analysis. Section 3 deals with system architecture. Section 4 describes about the dataset and analysis of dataset using preprocessing techniques. Section 5 explores about feature selection and extraction. Section 6 elucidates the process of model building, regression analysis and clustering. Section 7 discusses about time series analysis. Section 8 evaluates the proposed system using various performance metrics. Section 9 concludes the report.

2. LITERATURE SURVEY

This section surveys about the recent works related to sales prediction, time series analysis and the key challenges are portrayed below in Table 1.

Table 1: Related works

P. No.	Paper Title and Year	Author Name and Publication	TECHNIQUE / METHODOLOGY	ADVANTAGES	LIMITATIONS
1.	Sales Prediction Using Linear and KNN Regression, 2020 [1]	S. Kohli, G. T. Godwin and S. Urolagin In *Advances in Machine Learning and Computational Intelligence* (pp. 321-329). Springer, Singapore.	Data Preprocessing, Feature Selection, Predictive Analysis - Linear Regression, K-Nearest Neighbor Regression Model Evaluation using RMSE and MAPE.	Linear Regression – easy to fit, easy to interpret. Many firms allocate resources and also helps in cost optimization by generating maximum profit.	K-Nearest Neighbor Regression Overfitting model.
2.	Time Series Analysis for Sales Prediction, 2018 [2]	C. G. Chiru, and V. V. Posea In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications* (pp . 163-172). Springer, Cham.	ARIMA (Auto Regressive Integrated Moving Average) Method. solver and the optimization method.	Time information is provided to obtain the predictions.	Lack of ensemble models to eliminate prediction errors. Model convergence problem.
3.	Machine-Learning Models for Sales Time Series Forecasting, 2018 [3]	B. M. Pavlyshenko 2018 IEEE Second International Conference on Data Stream Mining and Processing.	Generalization of Machine Learning approaches. Stacking technique for regression and ensemble of single models.	Machine Learning generalization to reduce sales noise. stacking approach increases the validation accuracy.	For better model – Lasso regression with stacking approach.

4.	Sales Forecasting Newspaper with ARIMA: A Case Study, 2018 [4]	C. I. Permatasari, W. Sutopo and M. Hisjam In *AIP Conference Proceedings* (Vol. 1931, No. 1, p. 030017). AIP Publishing LLC.	Auto Regressive Integrated Moving Average (ARIMA) method to predict the newspaper count, minimize the number of returns, reduce the missed sales and restrict the oversupply.	ARIMA models are best to predict sales in short-term period.	No business competition analysis. Lack of user profile details for sales prediction.
5.	Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China, 2017 [5]	Y. Zhang, M. Zhong, N. Geng, and Y. Jiang In Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China. *PloS one*, *12*(5), p. e0176729.	This paper relies on sales prediction of automobiles Singular Spectrum Analysis (SSA) and Vector Auto Regressive model (VAR) for Electric Vehicles demand prediction in market.	This paper deals with the forecast model in three different aspects, such as model framework, model comparison and application. EV sales shows an increase in trend even with large fleets of vehicle. better forecasting performance.	No forecast of EV sales and model selection. Lack of unobserved heterogeneous variables.

2.1. Key Challenges

- Lack of ensemble learning models for predictions. So, the model convergence problem arises.
- Outlier detection and analysis.
- Complex model analysis.
- No proper sales forecasting.

3. SYSTEM ARCHITECTURE

This section describes about the major building blocks of retail data analysis system as shown in Figure 1.

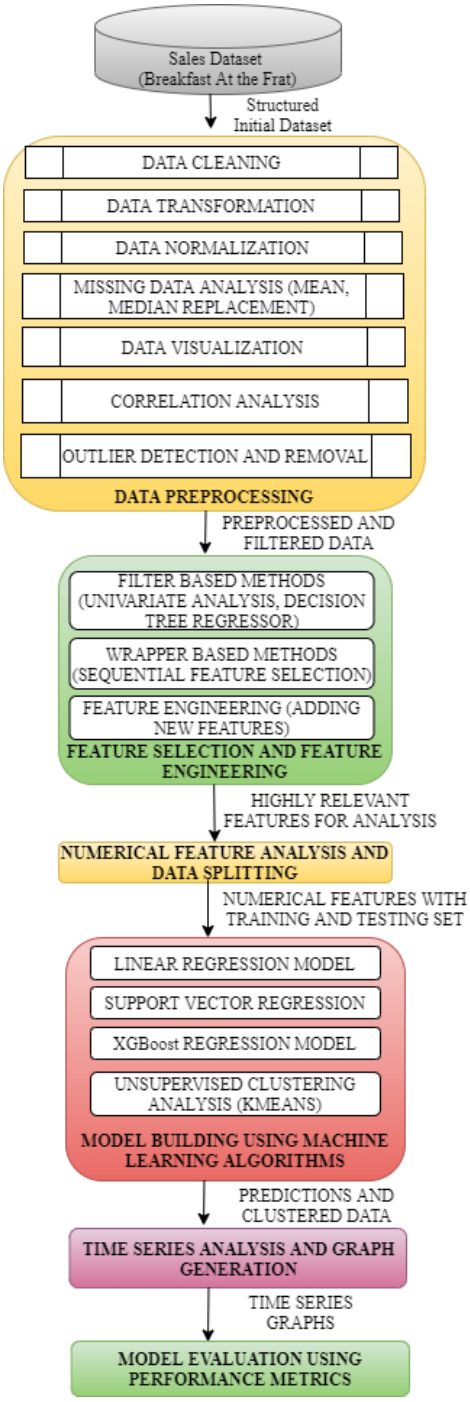


Figure 1: System Architecture for Predictive Retail Data Analysis

The project starts with the initial dataset (Breakfast at the Frat) analysis followed by data preprocessing. Data preprocessing is the first and foremost step for understanding the dataset and the relations between the features. Data preprocessing generates filtered data by using the techniques such as data cleaning, transformation, normalization, missing data analysis, correlation analysis, data visualization and outlier detection and removal.

Feature Selection deals with the process of choosing optimal and best features from the dataset for analysis. Filter Based Feature Selection methods namely Univariate Analysis and Decision Tree Regressor are used. In Univariate Analysis, ANOVA test has been applied and a series of P values has been calculated for each and every feature in the dataset. Decision Tree Regressor calculates the mean squared error value for each attribute. Wrapper Based method such as Sequential Feature Selector has been employed by calling Random Forest Regressor method. Feature Extraction deals with the addition of new features in the existing dataset. In this project, an attribute called discount has been added as a difference of base price and price.

The next step deals with the selection of numerical features from the dataset. There about 15 numerical features and 7 categorical features in the dataset. Then, the entire dataset is splitted into training sets such as (x_train, y_train) and testing sets such as (x_test, y_test). After data splitting, a ML model is developed using regression algorithms such as Linear Regression, XGBoost Regression and Support Vector Regression. Since the dataset consist of continuous values, regression algorithms are used to generate more accurate predictions.

The entire dataset is analyzed using the unsupervised K-Means Clustering Algorithm and four different clusters are generated. Time Series component analyzes the sequence of observations and generates time series graphs with respect to the time periods by taking into account the week end date and average weekly baskets attributes.

The proposed system is evaluated using certain performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2, Adjusted R2 and accuracy.

4. DATASET ANALYSIS

This section describes in detail about the dataset and its attributes taken for analysis.

4.1. Dataset Description

Breakfast at the Frat dataset is taken for analysis which contains sales and promotion information distributed over a period of 156 weeks. The dataset consists of 25 columns and 538643 rows. The features of the dataset are tabulated below as shown in Table 2.

Table 2: Description of Features

FEATURES	DESCRIPTION
WEEK_END_DATE	Week ending date
STORE_ID	Unique identifier for each store
UPC	Universal Product Code
UNITS	Units sold
VISITS	Number of specific purchases (baskets)
HHS	Number of household purchases
SPEND	Total spend (i.e., $ sales)
PRICE	Actual amount of the product at shelf
BASE_PRICE	Base price of product
FEATURE	Product published in in-store circular
DISPLAY	Product in in-store promotional display
TPR_ONLY	Reduction in Temporary price
DESCRIPTION	Product description
MANUFACTURER	Manufacturer
CATEGORY	Product Category

SUB_CATEGORY	Product Sub-category
PRODUCT_SIZE	Product quantity
STORE_NAME	Store name
ADDRESS_CITY_NAME	City
ADDRESS_STATE_PROV_CODE	State
MSA_CODE	Metropolitan Statistical Area
SEG_VALUE_NAME	Segment Value Name
PARKING_SPACE_QTY	Count of parking spaces in the Kroger parking lot
SALES_AREA_SIZE_NUM	Store Square footage
AVG_WEEKLY_BASKETS	Average weekly baskets sold in the store

4.1.1. Dataset Statistics and Visualization

Dataset Visualization is an important step for analysis since it describes graphically and statistically about the various features of dataset. In this project, Pandas Profiling and Sweetviz packages are deployed for visualization.

Pandas Profiling is a magical line of code for performing EDA process. Pandas Profiling package is imported and a profile report is generated which describes about the dataset and its attributes. Figure 2 gives the details about the dataset regarding the observations, variables, missing cells, duplicate rows and size occupied in the memory. It also displays the types of variables such as NUM(Numerical), CAT (Categorical) and BOOL(Boolean).

Dataset statistics

Number of variables	25
Number of observations	538643
Missing cells	366269
Missing cells (%)	2.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	102.7 MiB
Average record size in memory	200.0 B

Variable types

NUM	12
CAT	10
BOOL	3

Figure 2: Dataset Statistics

Sweetviz generates some meaningful visualizations by exploring the dataset. It is the dataframe which explains about the details of the dataset by displaying the number of categorical, numerical and text features.

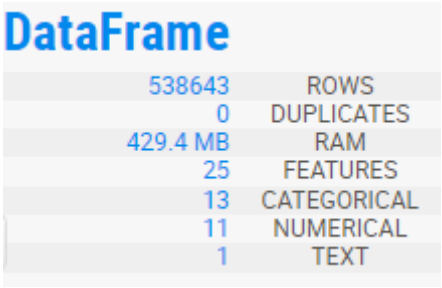


Figure 3: DataFrame Details

The concise summary of the dataset is displayed using dataset.info() method. The range index starts from 0 and it ends in 538642. The data columns are displayed along with the datatype of each column. There are about 10 columns in int64 datatype, 5 columns in float64 datatype, 10 columns in object datatype. PRICE, BASE_PRICE and PARKING_SPACE_QTY has missing values. Features of the dataset carry the information about a piece of data and the 13 categorical, 11 numerical and 1 text feature.

Figure 4 shows the value counts of the instances of the features CATEGORY and SUB_CATEGORY. In features CATEGORY, there are about four products such as COLD CEREAL, BAG SNACKS, ORAL HYGIENE PRODUCTS and FROZZEN PIZZA. COLD CEREAL has the highest value count of 174140. In features SUB_CATEGORY, there are about six sub categories such as PRETZELS, PIZZA/PREMIUM, ALL FAMILY CEREAL, MOUTHWASHES (ANTISEPTIC), KIDS CEREAL and Other values. PRETZELS has the highest value count of 130183.

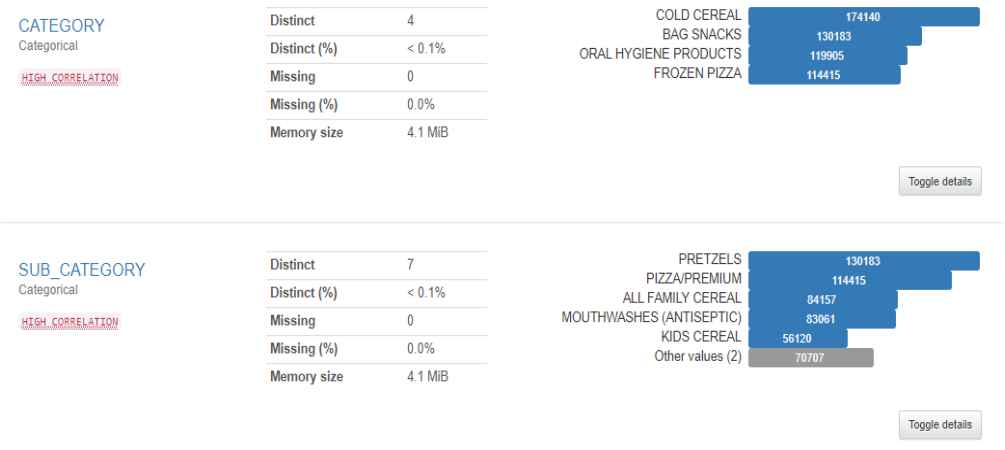


Figure 4: Sample of Two Categorical Features

Figure 5 visualizes the 4 different product categories and their sub categories with frequency of occurrences.

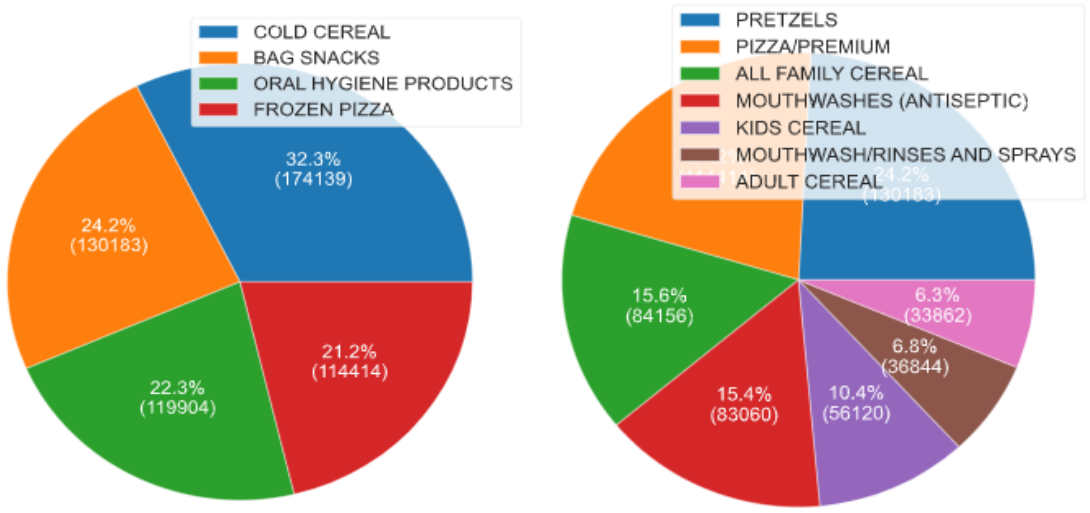


Figure 5: Pie Chart of the CATEGORY and SUB_CATEGORY of Products

Figure 6 displays the histogram of lengths graphically using the bars of different heights and gives the maximum length as 21, median length as 11, mean length as 13.197 and the minimum length as 10.

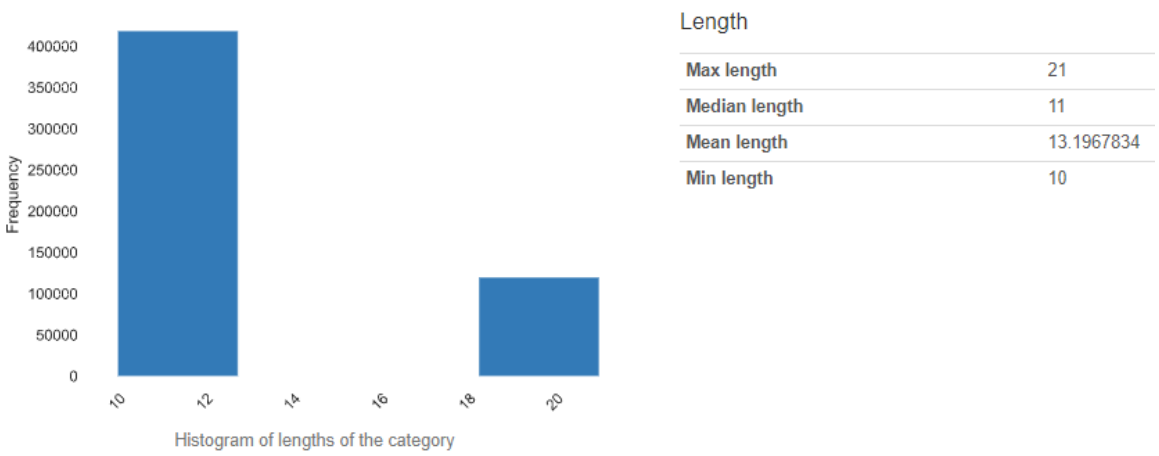


Figure 6: Length of the feature CATEGORY with Histogram

Figure 7 displays the quantile and descriptive statistics of the feature SPEND. Quantile Statistics are the cut points which divides the data into some continuous intervals with equal probabilities. The minimum value is 0. Nearly 95% of the dataset contains the

SPEND value of 4.59. Nearly 5% of the dataset contains the SPEND value of 169.86. Median is the middle value of dataset and so nearly 50% of the dataset contains the SPEND value of 32.13. The maximum value is 2952.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

From equation 1, range = 2952 – 0 = 2952.

Interquartile Range also called as midspread (50%). IQR can be obtained by subtracting 75th and 25th percentile. Q3 indicates upper quartile (75th percentile) and Q1 indicates lower quartile (25th percentile)

$$\text{Interquartile Range (IQR)} = Q3 - Q1$$

From equation 2, IQR = 67.83 – 13.47 = 54.36

Descriptive Statistics describes and summarizes about the features in the dataset. Standard deviation ($\sigma = 68.01$) describes the amount of deviation of one observation from a set of values. From equation 3, Coefficient of Variation is equal to standard deviation divided by mean value.

$$CV = \sigma / \mu = 1.28$$

Kurtosis is a Descriptive Statistical measure which describes the tallness and sharpness with respect to standard bell curve. The kurtosis value is 51.42. Skewness is also a descriptive statistical measure which tells the amount and direction of skew with respect to horizontal symmetry. Skewness value is 4.76. The average value (mean $\bar{X}$) spent by most of the customers is 53.28

$$\text{Median Absolute Deviation (MAD)} = |Xi - \bar{X}|$$

Where Xi – each value and $\bar{X}$ - mean

The first value of SPEND is X1 = 18.07, $\bar{X} = 53.28$

From equation 4, MAD of first value = |18.07-53.28| = 35.21

Sum is the total of SPEND values that is 28695356.36.

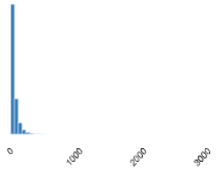
Variance is $\sigma 2 = 68.01 * 68.01 = 4625.36$

SPEND
Real number (ℝ≥0)

HIGH CORRELATION

Distinct	17523
Distinct (%)	3.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	53.27342295
Minimum	0
Maximum	2952
Zeros	24
Zeros (%)	< 0.1%
Memory size	4.1 MiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	4.59
Q1	13.47
median	32.13
Q3	67.83
95-th percentile	169.86
Maximum	2952
Range	2952
Interquartile range (IQR)	54.36

Descriptive statistics

Standard deviation	68.01168803
Coefficient of variation (CV)	1.27665324
Kurtosis	51.41641024
Mean	53.27342295
Median Absolute Deviation (MAD)	22.37
Skewness	4.758278311
Sum	28695356.36
Variance	4625.589709
Monotocity	Not monotonic

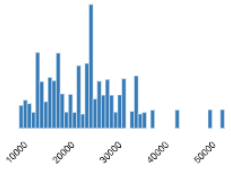
Figure 7: Quantile and Descriptive Statistic of SPEND feature

Figure 8 indicates the quantile and descriptive statistics of AVERAGE_WEEKLY_BASKETS feature.

AVG_WEEKLY_BASKETS
Real number (ℝ≥0)

Distinct	77
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	24643.62065
Minimum	10434.70513
Maximum	54052.51923
Zeros	0
Zeros (%)	0.0%
Memory size	4.1 MiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	10434.70513
5-th percentile	12706.53205
Q1	18196.82051
median	24766.80769
Q3	29435.62821
95-th percentile	38465.12821
Maximum	54052.51923
Range	43617.8141
Interquartile range (IQR)	11238.8077

Descriptive statistics

Standard deviation	8811.870791
Coefficient of variation (CV)	0.3575720839
Kurtosis	1.165892266
Mean	24643.62065
Median Absolute Deviation (MAD)	6410.52564
Skewness	0.9000699922
Sum	1.327411376e+10
Variance	77649066.84
Monotocity	Not monotonic

Figure 8: Quantile and Descriptive Statistic of AVERAGE_WEEKLY_BASKETS feature

Percentile values of numerical features are described in the codes. 25th, 50th, 75th, 85th, 95th and 99th percentile values are described along with count, minimum, maximum and standard deviation.

4.2. Data Preprocessing

Data preprocessing gives us the idea on how to produce meaningful insights about data to make informed decisions. Data preprocessing makes us understand the relationship between features in order to build ML models.

4.2.1 Data Conversion or Transformation

Data conversion or transformation is the process of data reconstruction into a format which is comfortable for analysis. Data conversion is a critical step of analysis which maps the data fields into some desired formats.

Data conversion is done by taking the difference of PRICE from BASE_PRICE features. By this way, discount feature is obtained.

$$\text{discount} = \text{BASE\_PRICE} - \text{PRICE}$$

For example, if the BASE_PRICE is 1.57 and PRICE is 1.39 then the discount value is 0.18. The discount feature is added to the dataset and the number of columns becomes 26. When Category of products with maximum BASE_PRICE is analyzed, it is seen that ORAL HYGIENE PRODUCTS has a maximum BASE_PRICE of 11.46 while the category of product named BAG SNACKS with the minimum BASE_PRICE of 0.55. It is also seen that the ORAL HYGIENE PRODUCTS is the category of products in which the units purchased by the customer is null. Whereas, COLD CEREAL is the category of products in which the units purchased by the customer is highest, that is 1800. The ORAL HYGIENE PRODUCTS has a maximum discount value of 8.77.

4.2.2. Data Normalization

Data normalization is the process of rescaling or normalizing the instances of the features according to the range value. Data normalization changes the values of the instances to a common scale value without distortions.

The price average is calculated by taking the maximum and minimum price values. The price average value is 5.73 for all the product categories. The range factor is calculated as shown in equation 6 by taking the difference of maximum and minimum price values.

$$\text{Range} = \text{PRICE.max()} - \text{PRICE.min()}$$

This range factor serves as the normalization value and the instances are normalized according to this range factor.

The instances of discount feature are rounded off to two decimals. Then the range factor is calculated as shown in equation 6 by taking the difference of maximum and minimum discount values. This range factor serves as the normalization value and the instances are normalized according to this range factor.

4.2.3 Data Cleaning and Missing Data Analysis

Data cleaning is the process of analyzing and correcting the inaccurate records from the dataset. Figure 9 indicates that the PARKING_SPACE_QTY feature has nearly 68% of missing data that is 366061 missing records. So, the PARKING_SPACE_QTY feature has been eliminated from the dataset using dataset.drop() method. After the removal of PARKING_SPACE_QTY feature, the number of columns becomes 25.

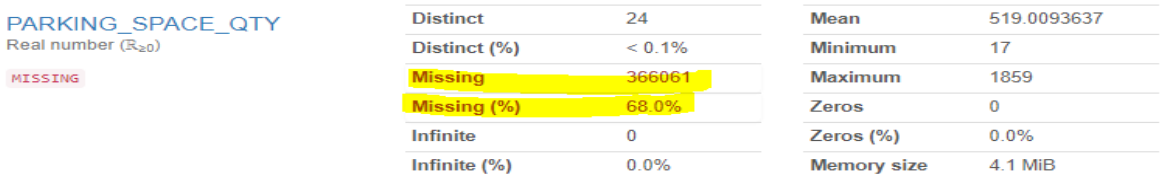


Figure 9: PARKING_SPACE_QTY feature

Three features such as MSA_CODE, UPC and STORE_ID which carry less importance for data analysis are also removed from the dataset. After the process of data cleaning, there are totally 22 columns in the dataset. Most of the dataset has inconsistent datapoints that is null values, so data reduction (missing data imputation) is an important step in preprocessing for handling missing value instances.

The total number of records in the dataset is 538643. The PRICE feature has 538620 records out of the total number of records which are indicated as missing values. Similarly, the BASE_PRICE feature has 538458 records out of the total number of records which are indicated as missing values. Then, the PARKING_SPACE_QTY has the highest percentage of missing values and it contains only 172582 out of the total number of records which are indicated as missing values.

The PRICE feature has 23 missing values. The BASE_PRICE feature has 185 missing values and the discount feature has 208 missing values. PARKING_SPACE_QTY feature has been removed from the dataset as a result of data cleaning step because of 68% of missing values.

The missing values in the dataset can be replaced by one of the three measures of central tendency such as mean, median and mode. Mean replacement is the most widely used technique compared to other two methods. In mean replacement, the missing observations are restored by the mean value of the entire feature column.

The arithmetic mean indicates the average as shown below:

$$\mu = \frac{\sum x}{N} \text{ for a population}$$

Mean replacement technique is used for filling the missing observations of discount and BASE_PRICE features in the dataset.

The median indicates the middle position of data when the data is in rank order.

Let n be the number of values. If n is odd, median indicates $\left(\frac{n+1}{2}\right)^{th}$ position.

If n is even, median indicates the average of the numbers in the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+2}{2}\right)^{th}$ positions.

Missing values of PRICE feature is filled with the median replacement technique. The total number of records in the PRICE feature is 538620. That is n = 538620, Since n is even, the average of the two numbers in the middle position is taken and it is replaced as the missing values.

When n value is divided by 2, that is 538620 / 2 = 269310th value (4.19).

The average of the middle two numbers is 4.19. So, the missing values of PRICE feature is replaced by 4.19.

4.2.4. Correlation Analysis

Correlation analysis is the process of determining the linear relationship or the association between the variables in the dataset. Correlation is a measure of how strongly one variable depends or influences the other variable. Statistical correlation measures the variation in X with respect to Y, and variation in Y with respect to X.

Figure 10 displays the highly cardinal and highly correlated variables. High cardinality refers to the columns with instances that are very uncommon or distinct or unique. From the dataset, the WEEK_END_DATE, DESCRIPTION, STORE_NAME and ADDRESS_CITY_NAME are the features with many distinct values and they are highly cardinal. High correlation indicates how one variable depends strongly on the other variables. For example, the SPEND feature is highly correlated with UNITS and VISITS features.

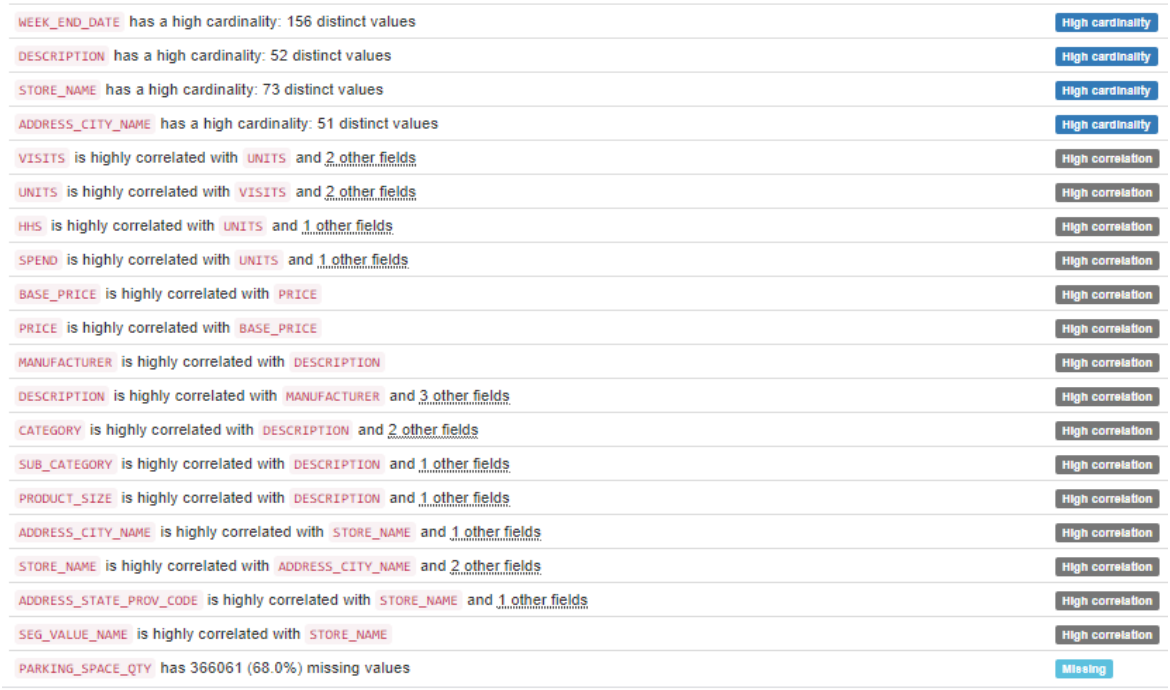


Figure 10: Features with High Cardinality and High Correlation

The Pearson's correlation coefficient (r) estimates the linear correlation between the variables X and Y. The range of r value is -1, 0 and +1, where -1 indicates total negative linear correlation, 0 indicates no linear correlation and 1 indicates total positive linear correlation.

To derive r for X and Y, divide the covariance of X and Y by multiplying their standard deviations as shown in equation 8.

$$r = \frac{1}{n-1}\sum\frac{(x_i-\overline{X})(y_i-\overline{Y})}{s_x s_y}$$

Figure 11 indicates the Pearson correlation matrix determination. The diagonal value is equal to 1 since the variables are connected to themselves. The UNITS feature has a total

positive correlation value of +1 with the VISITS and HHS feature. The AVG_WEEKLY_BASKETS with BASE_PRICE and PRICE has 0 correlation. The HHS feature and the PRICE feature has a negative correlation value of -0.2.

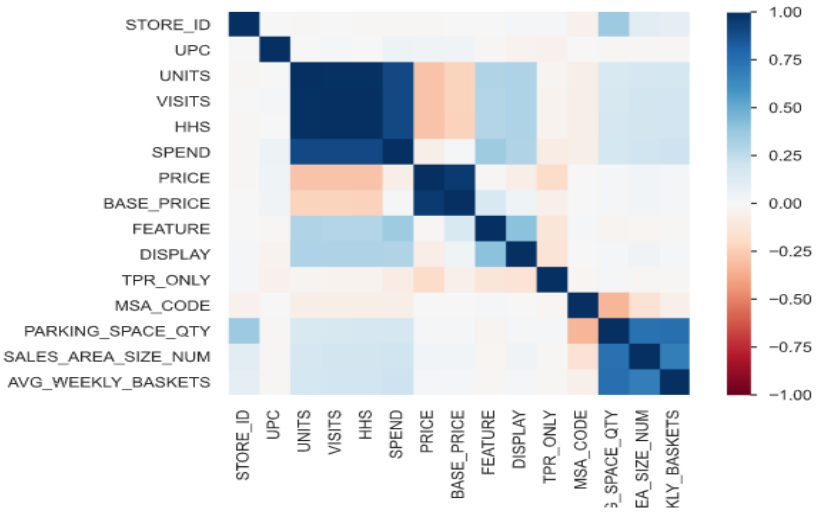


Figure 11: Pearson's correlation coefficient Matrix

Spearman's rank correlation coefficient (ρ) measures the nonlinear monotonic correlation between the variables. The range of r value is -1, 0 and +1, where -1 indicates total negative monotonic correlation, 0 indicates no monotonic correlation and 1 indicates total positive monotonic correlation.

Figure 12 displays the Spearman's rank correlation coefficient matrix. The BASE_PRICE and SPEND features has a correlation value of 0 which indicates no monotonic correlation. That is when the SPEND value increases the BASE_PRICE value decreases. The UNITS feature has a total positive monotonic correlation with the VISITS and HHS features indicating the value of +1. The PRICE feature has a total negative monotonic correlation with UNITS, VISITS and HHS.

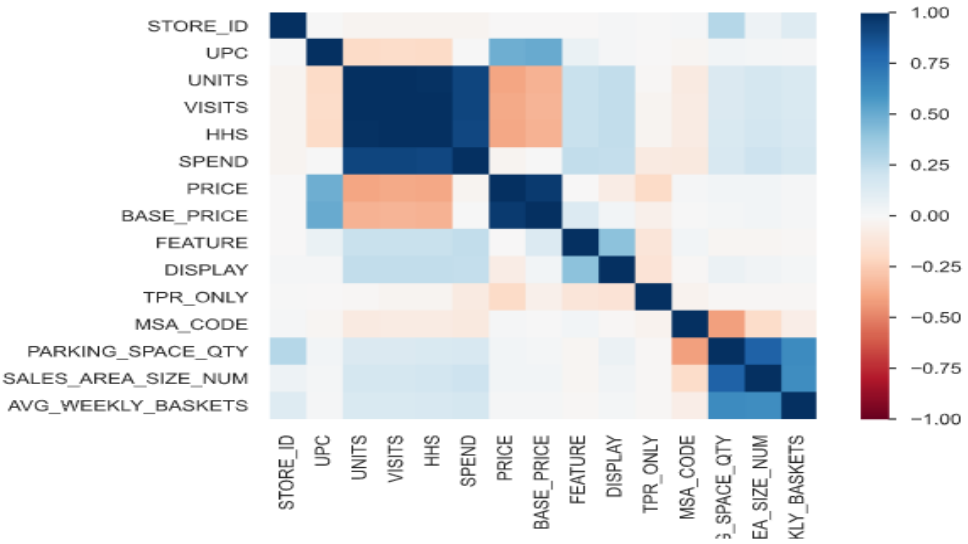


Figure 12: Spearman's rank correlation coefficient matrix

Phik (φk) correlation coefficient is a new correlation metric which can be applied to all the variables in the dataset such as categorical and numerical. In Phik (φk), all the variables are positively correlated. Fig 13 displays the Phik (φk) correlation coefficient matrix. The DESCRIPTION feature has a strong positive phik correlation with MANUFACTURER, CATEGORY, SUB_CATEGORY and PRODUCT_SIZE.

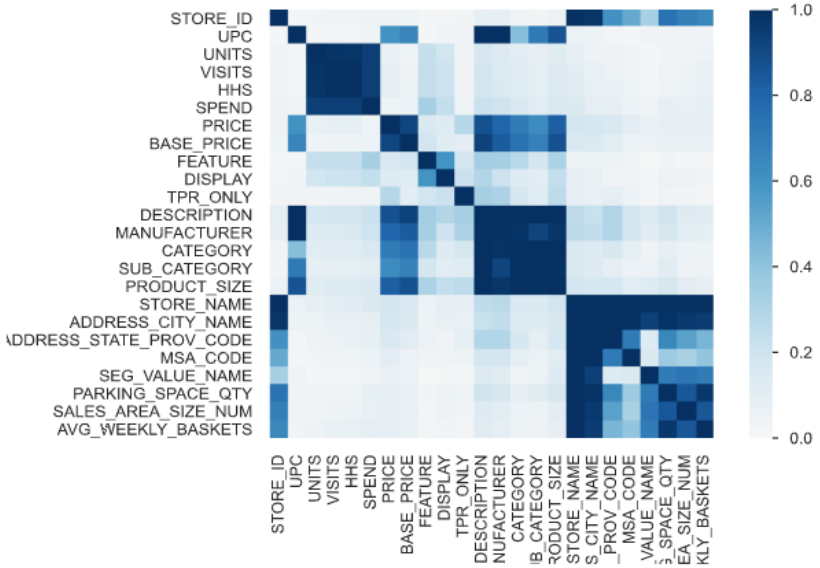


Figure 13: Phik (φk) correlation coefficient matrix

Association is the process of deriving relationship with the variables. Association rules can be generated by linking the variables. SQUARES are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is asymmetrical, indicating the skewness values (either left or right skewed). CIRCLES are numerical correlations (Pearson's) from -1 to 1. The trivial DIAGONAL is intentionally left blank for clarity. Fig 14 indicates the association matrix. The PRICE feature is linked or associated with the BASE_PRICE feature with a circle indicating strong numerical correlation. The PRODUCT_SIZE feature is categorically associated with the CATEGORY feature by SQUARE representation indicating a correlation ratio of 1.

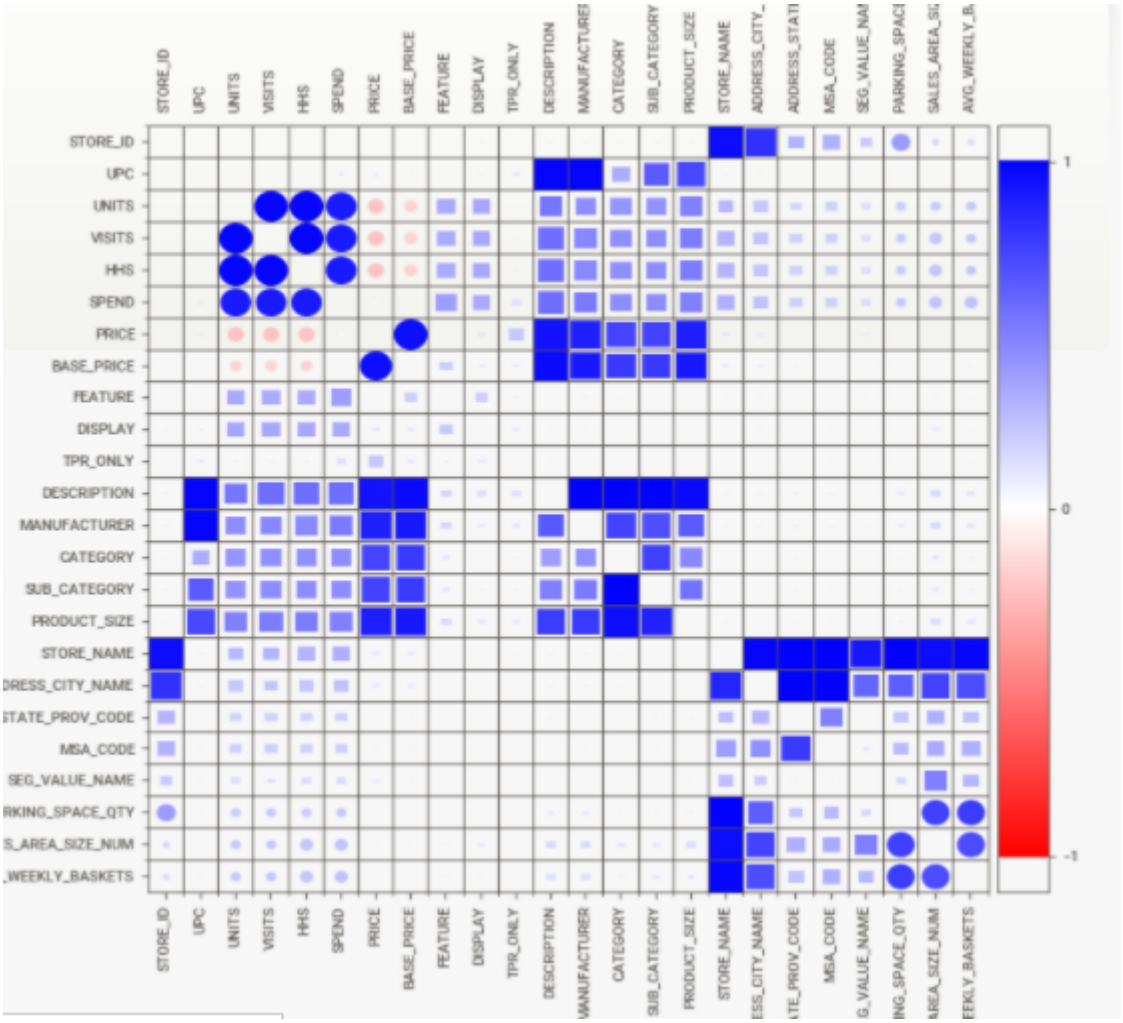


Figure 14: Association Matrix

4.2.5. Outlier Detection and Removal

An Outlier is a data object that deviates significantly from the other data objects in the dataset. Outlier detection and removal is the process of identifying the outliers from the dataset and handling or treating the outliers with some mathematical techniques such as Z-Score, InterQuartile Range (IQR) and so on. The visualization techniques for outlier detection and removal are box plot, scatter plot and so on.

A box plot as shown in Fig 15 consists of a rectangular box in center connected with lines (whiskers) from both ends. The box plot describes the median value (Q2 – 50th percentile), Q1(75th percentile) and Q3 (25th percentile), range, and IQR:

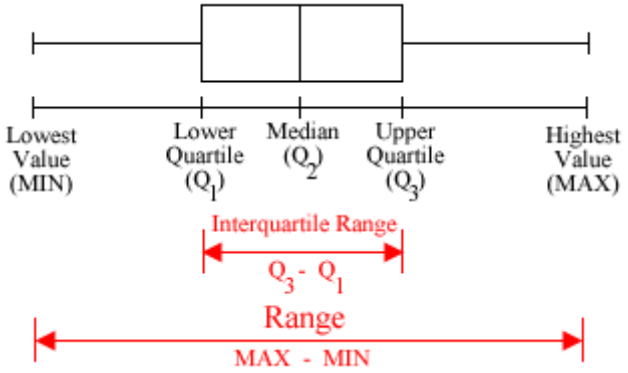


Figure 15: Box Plot

Figure 16 displays the outliers detected by box plots for the two features such as SPEND and AVG_WEEKLY_BASKETS.

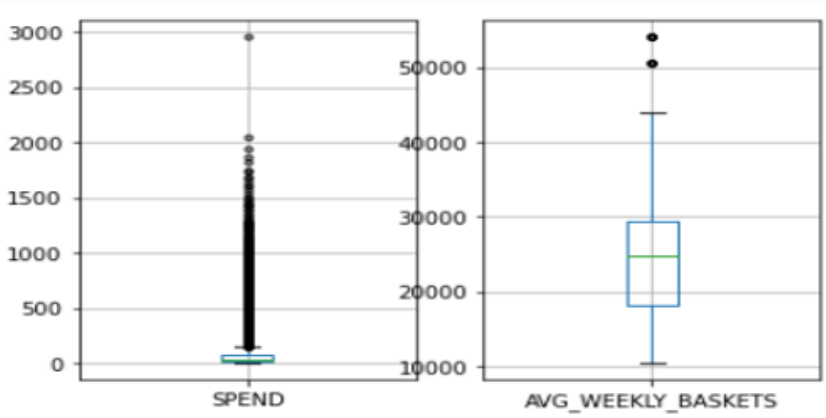


Figure 16: Box Plot for Detecting Outliers

Scatter Plot is a mathematical plot diagram which makes use of cartesian coordinate values between the two variables. It plots the data points in horizontal and vertical axis. Fig 17 describes about the scatter plot between two features such as SPEND and AVG_WEEKLY_BASKETS.

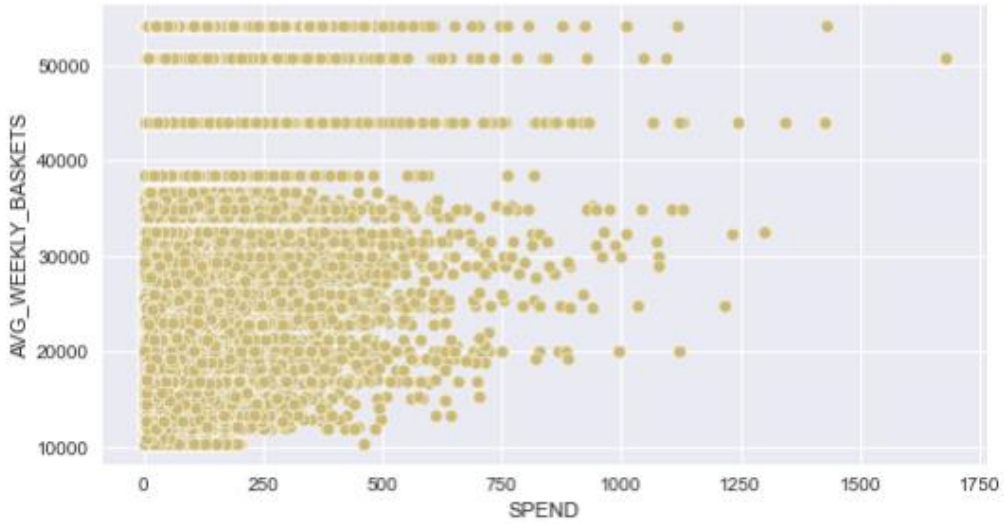


Figure 17: Scatter Plot for Detecting Outliers

An IQR also called as middle 50% is a statistical measure which is estimated by subtracting 75th percentile (Q3) and the 25th percentile (Q1) as shown in equation 9. The first quartile (Q1) separates the lower 25 percent from upper 75 percent and its value is equal to $\frac{n+1}{4}$ observation. The third quartile (Q3) separates the upper 25 percent from lower 75 percent and its value is equal to $\frac{3(n+1)}{4}$ observation.

$$IQR = Q3 - Q1$$

The shape of the dataset is reduced to 353589 rows and 18 columns after the detection of outliers using IQR method. The features in the dataset are sorted and then the Q1, Q3, and IQR score is calculated for each attribute in the dataset. The criteria are fixed, so that any data value that is within the range are treated as outliers and removed from the dataset. Nearly, 185054 that is 34.35% records are considered as outliers and are removed from the dataset.

4.2.6. Normalization by Min Max Scaling Method

Min Max scaling technique has been performed on the variable X. This technique transforms the values of each feature by scaling each feature values to a range of 0 to 1. The variable X has been allotted with the numerical features and y has been allotted with the target feature.

5. FEATURE ENGINEERING AND FEATURE SELECTION

5.1. Feature Engineering

Feature Engineering deals with the generation of new features from the existing features of the dataset. By performing feature engineering, the accuracy of the model can be improved and the high-quality features can be retained. Discount is a new feature which is added to the dataset as a result of feature engineering. Discount is the difference of BASE_PRICE and PRICE feature from the dataset. Discount is mainly calculated to increase the sales and also to attract more customers for buying the products.

Feature Selection chooses highly optimal, relevant and useful features from the dataset. Highly important features are selected from the dataset for further analysis. In other words, feature selection selects the optimal or best number of features from the entire dataset. Different methods for feature selection are filter, wrapper and embedded (filter and wrapper). Number of categorical features is 7 and the number of numerical features is 11 in the dataset after outlier detection.

The packages necessary for performing feature selection and selecting the numerical features are F_regression, SelectKBest, SelectPercentile, and train_test_split. F_regression calculates the regression score for each and every feature in the dataset. SelectKBest selects the highly best and significant features based on chi-square score. SelectPercentile selects the features based on the highest percentile score. Import train_test_split package splits the dataset into train and test datasets. Initially, the numerical features of the dataset are selected by taking the relevant datatypes. The selected number of numerical features is 14.

The entire dataset is splitted into training sets such as (x_train, y_train) and testing sets such as (x_test, y_test). The AVG_WEEKLY_BASKETS is considered as the target attribute since it has some continuous values and it is dropped and included as a target variable. The training and testing split is performed as 80% training and 20% testing. Random_state is any random integer value. At the end of splitting, the shape of X_train is (282871, 10), y_train is (282871,), X_test is (70718, 10) and y_test is (70718,).

5.2. Filter Based Method

Filter based method measures the relevance of features by calculating the correlation with the dependent variable. This method starts with the set of features and selects the best features by using a learning algorithm.

5.2.1. Select K-Best and Select Percentile Methods

Univariate feature selection methods work by determining the optimal features from the univariate test methods like Analysis of Variance (ANOVA). ANOVA deals with the collection of models with estimated methods. There are four different methods of univariate feature selection such as SelectK-Best, SelectPercentile, SelectFpr, SelectFdr, or family wise error SelectFwe, and Generic Univariate Selection. In this project, the two methods such as SelectK-Best and SelectPercentile is used for analysis. F_regression is the function used for performing feature selection on numerical data. SelectK-Best selects only the K highest scoring features. SelectPercentile selects the features according to a percentile of high scores.

As the output of SelectPercentile and SelectK-Best feature selection techniques, the percentile value is 50. The selected features after SelectPercentile feature selection technique is UNITS, VISITS, HHS, SPEND and PRICE. The SelectK-Best technique is performed with a score function called F_regression. The top 5 features after SelectK-Best are UNITS, VISITS, HHS, SPEND, BASE_PRICE.

5.3. Decision Tree Regressor Feature Importance

Feature Importance refers to the methodologies that give a score value to the initial features based on how meaningful they are at predicting a target feature. Classification and Regression Trees (CART) algorithm offers important scores using decision tree regressor class. Decision Tree regressor builds the regression model in the form of tree structure. It is a one-dimensional regression with decision tree.

80% of the dataset is allocated for training and 20% of the dataset is allocated for testing. Initially, the array is initialized, and the for loop is used. The for loop takes each feature of the dataset and builds a tree using DecisionTreeRegressor() function. For each feature in the training dataset, the decision tree algorithm generates a node which depicts the root node. The root node is considered to be the most important feature. The evaluation begins at the root node and repeats until the condition or "decision" is met. This step keeps

looping until a leaf node is identified. The leaf node contains the prediction or the output of the tree.

MSE values for each feature is calculated in the dataset. The features are sorted in the descending order. The feature with the higher MSE can be removed from the dataset. Since TPR_ONLY, DISPLAY, FEATURE are less relevant features, these features can be removed from the dataset. The lesser MSE values are most important features and they should not be removed from the dataset.

5.4. Wrapper Based Method

In Wrapper Based Feature Selection method, a subset of features will be selected from the entire dataset and a learning algorithm will be applied. This process is repeated until a best subset is selected. Based on the inferences, the features are removed or selected. Some of the methods of feature selection are sequential forward feature selection, recursive and backward feature elimination.

5.4.1. Sequential Feature Selection

Forward Sequential feature selection adds the features in a forward manner and backward sequential feature selection removes the features in a backward manner to generate a feature subset. It is also called as a greedy search algorithm. At each iteration, the sequential feature selection selects the optimal feature to add or remove based on the cross-validation score.

Sequential feature selection algorithm functions using the RandomForestRegressor() method which makes use of ensemble learning method that is bagging technique to make predictions. The number of features is initialized to 10 and the cross-validation score is equal to 3. This algorithm displays the most important 10 features of the dataset. The features are UNITS, VISITS, HHS, SPEND, PRICE, BASE_PRICE, FEATURE, DISPLAY, TPR_ONLY, discount.

Using sequential backend concurrent worker, the feature score for UNITS is 0.041, VISITS is 0.066, HHS is 0.088, SPEND is 0.091, PRICE is 0.093, BASE_PRICE is 0.094, FEATURE is 0.095, DISPLAY is 0.086, TPR_ONLY is 0.082 and discount is 0.080.

6. MODEL BUILDING AND ANALYSIS

Model building deals with the construction of statistical and probabilistic model that estimates the relationship between both dependent and independent features. 80% of the dataset is allocated for training and 20% of the dataset is allocated for testing.

6.1. Regression Analysis

Regression Analysis deals with statistical analysis and estimation of relationship between dependent feature Y and one or more independent features X. The dependent feature Y is the AVG_WEEKLY_BASKETS and the independent features X are UNITS, VISITS, HHS, SPEND, PRICE, BASE_PRICE, FEATURE, DISPLAY, TPR_ONLY, SALES_AREA_SIZE_NUM and discount.

6.1.1. Linear Regression Model

In this project, linear regression model is a predictive analysis technique that has been implemented between dependent feature y which is AVG_WEEKLY_BASKETS and independent features. The independent features are the numerical features selected from the dataset. The independent features are also called as inputs or predictors. The dependent feature is the average weekly sales of the products. The dependent feature is also called as output or response. Linear regression algorithm models the relationship between the features by fitting the linear equation.

The lr.fit() function takes the two arguments X_train and y_train fits the model for carrying out linear regression. The function lr.predict() predicts the quantities of the regression model.

The intercept value of the linear regression model using lr.intercept_ function and the regression coefficients of each numerical feature is calculated using lr.coef_ function. The function lr.intercept_ calculates the intercepts of the regression model for decision making. The function lr.coef_ contain the coefficients for the prediction of each of the targets.

The performance metrics of the linear regression model is displayed in Figure 18 below. The predicted value of AVG_WEEKLY_BASKETS for the test data provided with the independent features. The AVG_WEEKLY_BASKETS is the response feature

depending on the other features in the dataset. The function lr.predict (features) predicts the AVG_WEEKLY_BASKETS score for the given test data.

Algorithm	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R-Squared R2	Adjusted R-Squared	Accuracy
Linear Regression	5062.14	39782497.55	6307.33	0.70021	0.70020	12362.36

Figure 18: Evaluation Metrics for Linear Regression Model

6.1.2. XGBoost Regression Model

XGBoost Regression is the most commonly used algorithm for prediction and regression. The algorithm improves the accuracy of the model. This algorithm starts working by importing the XGBoost library and many other important libraries. The parameters of the algorithm are learning rate (shrinkage of step size) is 0.05, max_depth (depth of boosting round) is 6, subsample (subtrees which indicates the percentage of samples in the tree) is 1, n_estimators: number of trees to build.

An XGBoost regressor object is initially started by calling the XGBoost Regressor () method from the library called XGBoost with the hyper-parameter arguments. The cross validation score for the algorithm XGBoost is 0.9763. The evaluation metrics for the XGBoost regression model has been displayed in Figure 19.

Algorithm	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R-Squared R2	Adjusted R-Squared	Accuracy
XGBoost Regression	365.64	1698947.36	1303.44	0.988985	0.988984	2554.74

Figure 19: Evaluation Metrics of XGBoost Regression Model

6.1.3. Support Vector Regression Model

Support Vector Machine algorithm are very efficient algorithms for both classification as well as prediction problems. Support Vector Regression (SVR) predicts the linear relationship between two continuous variables. The training and testing sets are splitted, then fit method is used to fit the SVR model with the training set values. A non-linear function called SVR with Radial Basis Function (RBF) kernel is used for analysis. The RBF kernel results in optimized model. After training, the model is ready to do predictions. The predict method on the model passes X as a parameter to get the output as ypred. Evaluation metrics is described in Figure 20.

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R-Squared R2	Adjusted R-Squared	Accuracy
Support Vector Regression	4996.99	42382950.79	6510.22	0.67392	0.67391	12760.03

Figure 20: Evaluation Metrics for SVR Model

6.2. K-Means Clustering

Clustering is an unsupervised machine learning approach which deals with grouping of data points. The main aim of clustering is increasing intra class similarity and decreasing inter class similarity. K-Means is a partitioned clustering method in which the entire data is splitted into meaningful clusters so that each point belongs to one cluster only. Totally, four clusters are generated.

Pseudocode for K-Means Clustering Algorithm:

Step 1: Initialize or select the number of clusters.

Step 2: Selecting some points as the centroids.

Step 3: Perform iterations.

Step 4: Calculating the distance of all points from the centroids.

Step 5: Assigning the data points to the centroid whose distance value is less.

Step 6: Loop Step 3 to Step 5.

At the end of 4 iterations, the clusters are obtained.

Figure 21 displays the four different colors with different color combinations in which the cluster in red color has been fully filled.

```
0     125451
3     102247
2      83615
1      42276
```

0 -> Light Blue, 1-> Purple, 2-> Cream, 3-> Red

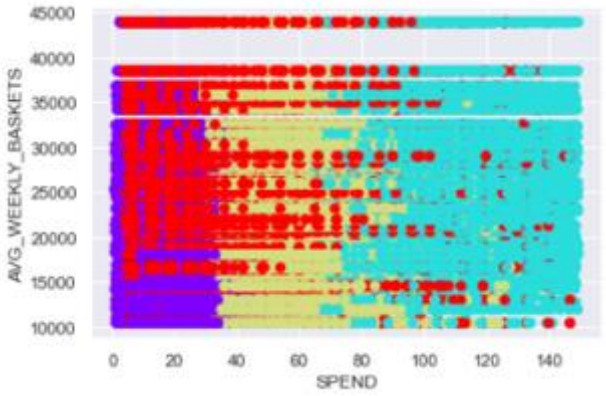


Figure 21: Clusters Generation

7. TIME SERIES ANALYSIS

Time Series consists of sequential observations of the similar variable(s) made over time. Time series Analysis performs analysis and generates sequential components based on time. The different frequency values are analyzed such as date, week, month and so on. The WEEK_END_DATE feature of the dataset is modeled as per the date frequency.

First value of the dataset is '2009-01-14 00:00:00' and last value of the dataset is '2012-01-04 00:00:00'.

Figure 22 displays that the average weekly baskets value is highest on 14/01/2009, the trend is downward and upward.

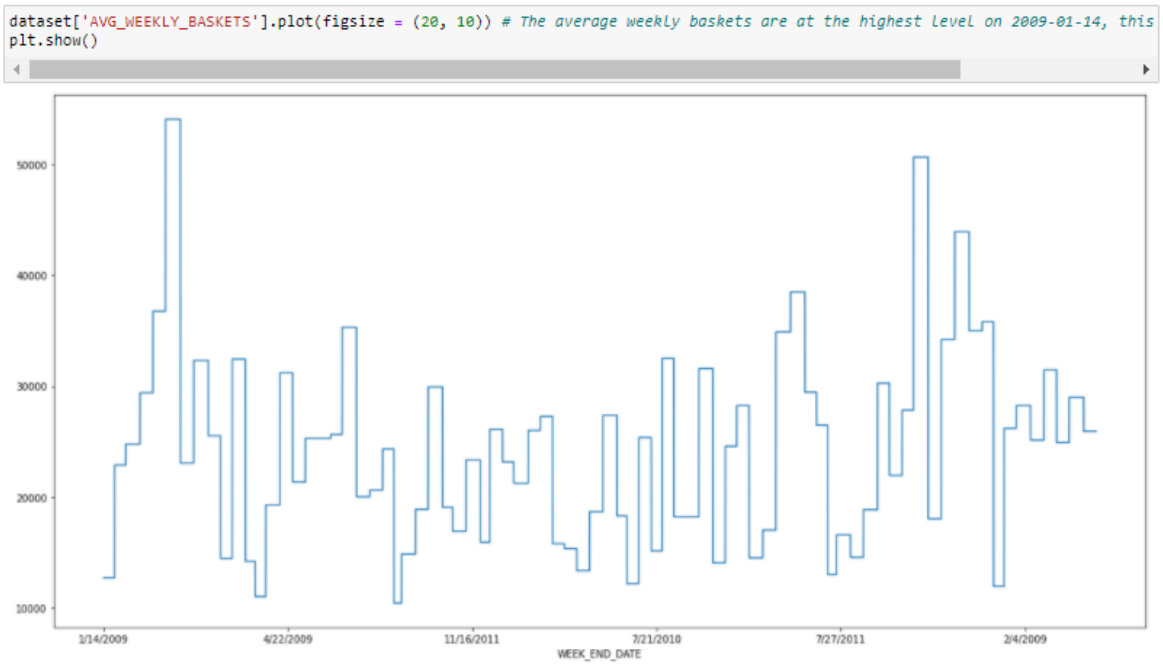


Figure 22: Trend Graph

Figure 23 clearly shows that the trend line of AVG_WEEKLY_BASKETS value decreases within the time period of 14/01/2009 to 22/01/2009. The trend predicted is downward trend.

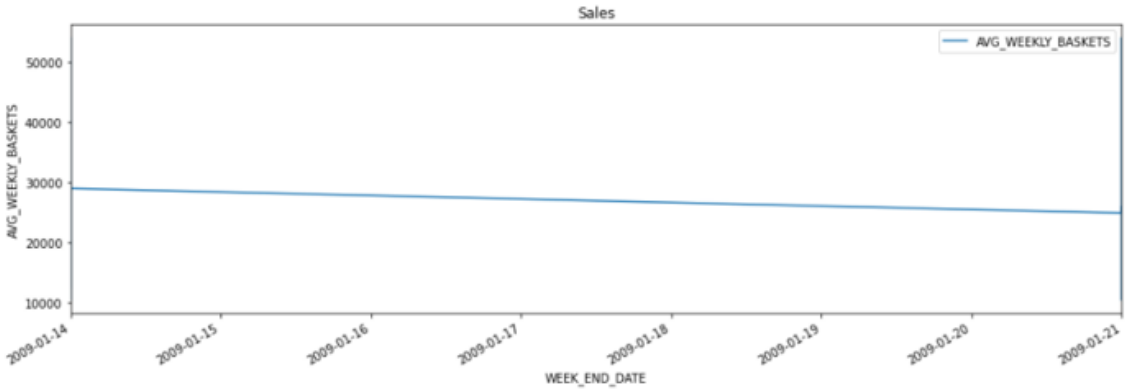


Figure 23: Downward Trend

Figure 24 clearly shows that the trend line of AVG_WEEKLY_BASKETS value increases within the time period of 14/07/2010 to 22/07/2010. The trend predicted is upward trend.

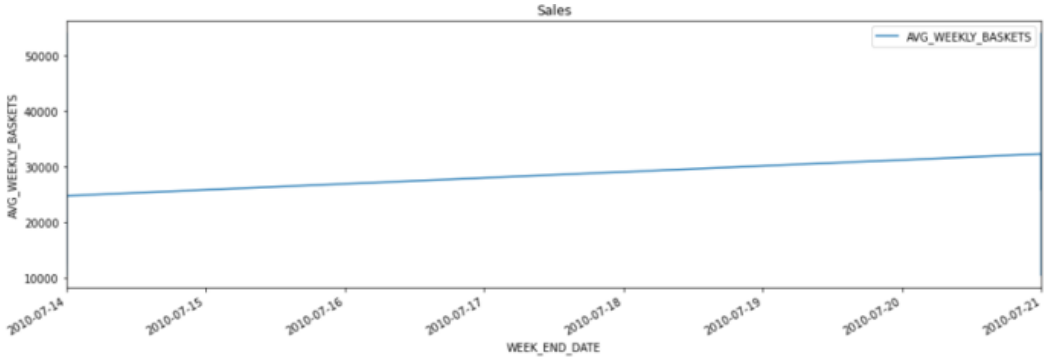


Figure 24: Upward Trend

Figure 25 displays the graph for seasonal first and fourth order difference data of AVG_WEEKLY_BASKETS.

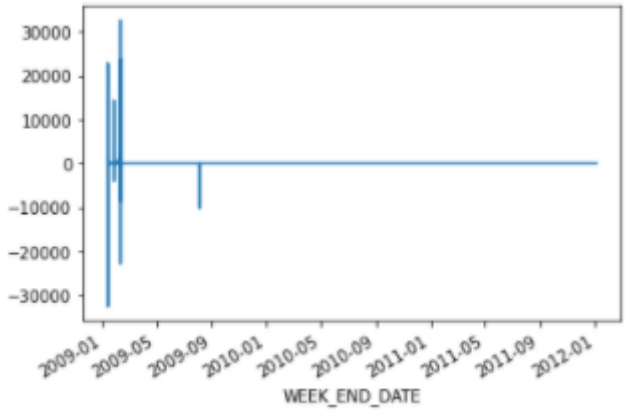


Figure 25: Seasonal First and Fourth Order Difference Data

Figure 26 explains the auto correlation function which is a conditional correlation. Autocorrelation describes the linear relationship between *delayed time series values*.

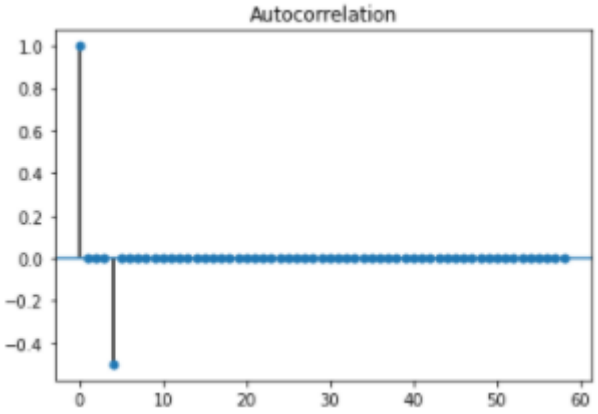


Figure 26: Auto Correlation Function

The entire dataset is splitted as 20% for testing and remaining 80% for training. y-axis depicts the AVG_WEEKLY_BASKETS and x-axis indicates the time (WEEK_END_DATE). The graph as shown in Fig 27 shows that the average is constant whereas the AVG_WEEKLY_BASKETS of the product increases and decreases randomly by a small margin. The average for each time period doesn't change and it is constant. In such a case the AVG_WEEKLY_BASKETS is taken for analysis which is similar to the average of all the previous days.

Simple Average technique predicts the expected value equal to the mean of all precedently examined points as shown in equation 10.

$$\hat{y}_{x+1} = \frac{1}{x}\Sigma_{i=1}^{x} y_i$$

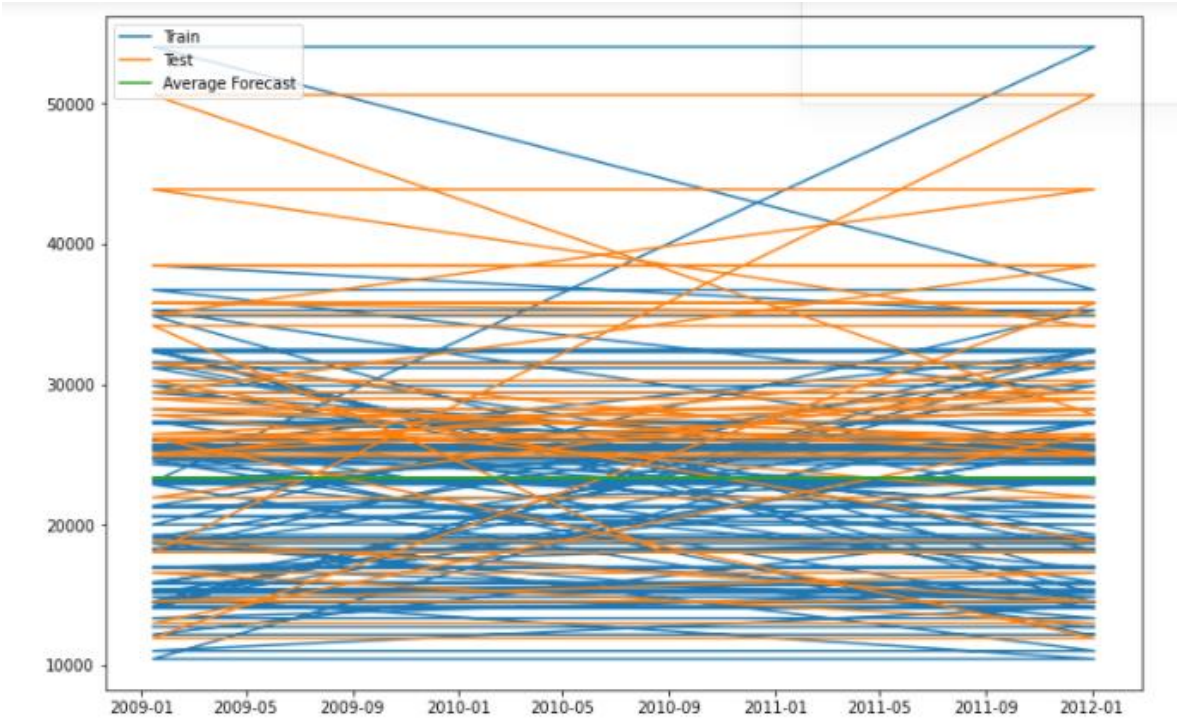


Figure 27: Simple Average Graph

The RMSE value is calculated as 10291.192232245114.

The prices of the starting period highly affect the prediction for the upcoming period. So, as an enhancement only the price mean average for last few time periods is taken. Such a forecasting technique which sets time period window for estimating the average is called Moving Average technique. Determination of the moving average deals with the creation of a "sliding window" of size n.

Using a simple moving average model, the upcoming value(s) in a time series are predicted on the basis of mean average of a constant finite count 'p' of the previous instances. So, for all i > p as shown in equation 11.

$$\hat{Y}_l = \frac{1}{p}\left(y_{i-1} + y_{i-2} + y_{i-3} \dots\dots + y_{i-p}\right)$$

The RMSE value is calculated as 14330.426511062706.

Fig 28 displays the Moving Average graph as below.

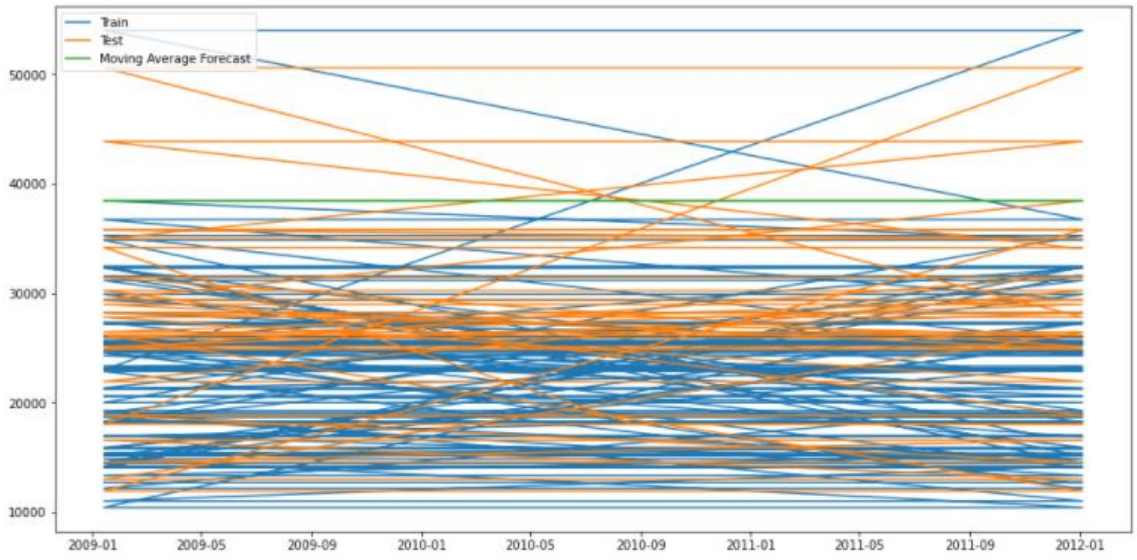


Figure 28: Moving Average Graph

8. MODEL EVALUATION

8.1. Evaluation Metrics

Mean Absolute Error – Arithmetic mean average of the absolute errors as shown in equation 12.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Absolute Error - Difference between the derived value of a quantity and its initial value.

Mean Squared Error – Mean of the square of the difference between initial or actual and derived or estimated values as shown in equation 13.

$$MSE = \frac{1}{n}\sum_{i=n}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Root Mean Squared Error - Square root of the Mean Squared Errors as shown in equation 14.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

R2 - Coefficient of determination is as shown in equation 15.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where RSS – sum of squares of residuals, TSS – total sum of squares.

Adjusted R-Squared is the updated or adjusted value of R-Squared where the number of parameters is adjusted. Using Root Mean Squared Error (RMSE) value, according to National Digital Elevation Guidelines (NDEP) and FEMA guidelines, a measure of accuracy is computed as:

Accuracy = 1.96 * RMSE

Table 3: Regression Model Comparison

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R-Squared R2	Adjusted R-Squared	Accuracy
Linear Regression	5062.14	39782497.55	6307.33	0.70021	0.70020	12362.36
XGBoost Regression	365.64	1698947.36	1303.44	0.988985	0.988984	2554.74
Support Vector Regression	4996.99	42382950.79	6510.22	0.67392	0.67391	12760.03

From Table 3, it is clear that the Support Vector Regression (SVR) model scores as the optimized and best model for the retail data analysis "Breakfast at the Frat" dataset.

Table 4: Time Series Analysis Comparison

Algorithms	Root Mean Squared Error	Accuracy
Simple Average	10291.19	20170.73
Moving Average Method	14330.43	28087.64

From Table 4, it is clear that the moving average method in time series analysis scores as the optimized and best model for the retail data analysis "Breakfast at the Frat" dataset.

CONCLUSION AND FUTURE SCOPE

Time Series Analysis for sales data mainly deals with sequential organization of data according to their time of occurrence. ML techniques are applied on the existing (training data) and new results are obtained for each test data. Regression based ML methods presents best results in comparison to time series analysis. The ultimate goal of this project is to predict the sales and perform time series analysis using different regression algorithms such as linear regression, Support Vector Regression and XGBoost Regression. Cluster analysis using K-Means Clustering Algorithm is also performed. Many ML algorithms are compared and evaluated using the standard performance metrics. Time Series Analysis using moving average method is also performed. After the Regression Analysis, the Support Vector Regression Algorithm is more accurate when compared to other algorithms. After the Time Series Analysis, the Moving Average Method is more accurate than the other method. In the future, ensemble models can be deployed for predicting the sales. Optimal Price Prediction for a particular product can also be focused.

REFERENCES

[1] S. Kohli, G. T. Godwin and S. Urolagin (2020) Sales Prediction Using Linear and KNN Regression. In Advances in Machine Learning and Computational Intelligence (pp. 321-329). Springer, Singapore.

[2] C. G. Chiru and V. V. Posea (2018, September). Time Series Analysis for Sales Prediction. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications (pp. 163-172). Springer, Cham.

[3] B. M. Pavlyshenko (2018). Machine-learning models for sales time series forecasting. Data, 4(1), 15.

[4] C. I. Permatasari, W. Sutopo and M. Hisjam (2018, February). Sales forecasting newspaper with ARIMA: A case study. In AIP Conference Proceedings (Vol. 1931, No. 1, p. 030017). AIP Publishing LLC.

[5] Y. Zhang, M. Zhong, N. Geng and Y. Jiang (2017). Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China. PloS one, 12(5), e0176729.

[6] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman (2017, September). Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) (pp. 1643-1647). IEEE.