INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

RETAIL ANALYSIS — WALMART'S TREND ASSESSMENT

Dr. Neetu Anand[1], Dr. Tripti Sharma[2], Dr. Kumar Gaurav[3], Kanishka Kharbanda[4], Yash Raj Singh[5]

[1]Associate Professor, MSIT, India

[2]Professor, MSIT, India

[3]Associate Professor, MSIT, India.

[4]Student, IGDTU

[5]Student, NIT Jamshedpur, India.

ABSTRACT

The retail industry is a sizable and important sector of the economy made up of businesses that sell completed goods to consumers. The U.S. GDP (gross domestic output) is largely derived from retail sales. Brick & mortar shop sellers, who engage in the sale of goods from actual places so that customers can make purchases there, are where the business initially got its start. e-tailers, where commodities are advertised online and quickly delivered to clients' doorsteps, have emerged over the past seven years. Online purchases make customers' lives easier than ever because they might be made with the simple push of a mobile button. Although the internet era has made the retail sector more convenient, the day-to-day issues that retailers encounter across many sectors require more than just human foresight to be properly managed. and offer a viable solution that takes into account variables like cost, volume, and time. The use of statistics and machine learning models is pervasive and has successfully entered the retail industry as well. The current period's retail data is a priceless resource that is used as an input to these analytical models to produce practical knowledge for the retail businesses. The main issues facing the retail industry include client segmentation, inventory management, sales forecasting, and promotion tactics. Each of these issues is dissected to comprehend the potential outcomes, and then each is addressed with a solution that is statistically sound.

Keywords: e-tailers, clustering, Machine Learning, Linear Regression, Time Series Analysis

1. INTRODUCTION

The retail industry is essential to contemporary culture. Retail businesses, both offline and online, are heavily reliant on the needs of consumers for a variety of goods and services. Earlier, the process of trading was used to make goods and services accessible. But today, purchasing and selling items has taken the place of trading, making retail outlets a crucial link in the supply chain and a lucrative industry for many. Retail continues to be a substantial contributor to the world economy, with a footprint worth several trillion dollars.

With the most recent technology developments, analytics in retail has become very popular because it helps the company make important decisions. It gives the shop the ability to develop standardised procedures that analyse consumer segments and product categories and can increase sales. In order to provide analytical insights that can be crucial for making marketing and procurement decisions, this project employs a variety of retail analytics principles. To analyse the retail business for a single retailer, we intend to use a number of machine learning approaches.

After carefully analysing the data, we have chosen to respond to three key queries that might be useful for the owner of this data set's retail store:

1. Customer segmentation: By classifying customers into different groups based on their purchasing habits and other relevant characteristics, the client can target a specific group of consumers with pertinent advertisements and deals (targeted marketing).
2. Sales forecasting: Examining data trends to project sales of products for the retail company to be offered in the next weeks.
3. Affinity/purchasing/Basket Analysis: Examining daily customer purchasing behaviour and understanding product affinity (i.e., how Product X sells in conjunction with Product Y/Z)[2].

There are precedents of numerous businesses utilising these strategies to boost earnings.

Recent completion of its most recent success story on market segmentation study was reported by Infiniti Research, a renowned provider of market intelligence solutions based in London. A major company in the money transfer industry sought to poll agents worldwide in a number of European nations. The client was able to enhance its agent management programme by learning what the agents value with the aid of a customer intelligence technology. They were able to guarantee a higher degree of satisfaction, improving both sales and customer satisfaction. Many e-commerce behemoths

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

e-ISSN : 2583-1062

Impact Factor : 5.725

use sales forecasting, including Amazon, which provides an automated tool called Amazon forecast that uses time-series data to forecast future sales and product demand.

There will be an effort to go through every issue raised here in as much detail as possible and to offer retail business owners any useful information.

2. EXPLORATORY ANALYSIS AND CHALLENGES

The Kaggle website served as the repository for the dataset used in this project. No issue statement has been developed, nor has any analysis been done on this data to produce any insights. The ideas and issues that will be covered in this paper are all self-described and were not given any outside inspiration. For each consumer who visited the store, there are a total of three different data sources: customer data, product mapping data, and transaction data. For analysis and modelling, the transaction data contains a total of 4 years' worth of data. We first conduct exploratory data analysis to look for trends in the data. Since 2011, Fig. 1 displays the sales activity of various store kinds across various months. With twice as many sales as the next-best competitor, the e-shop dominates. This makes sense given that e-commerce sites like Amazon and eBay were becoming more and more popular at the time. The abrupt fall in the sales data from 2014 is an intriguing finding to take note of from the graph. Since it will last through the end of 2014, this phenomenon is not transitory.

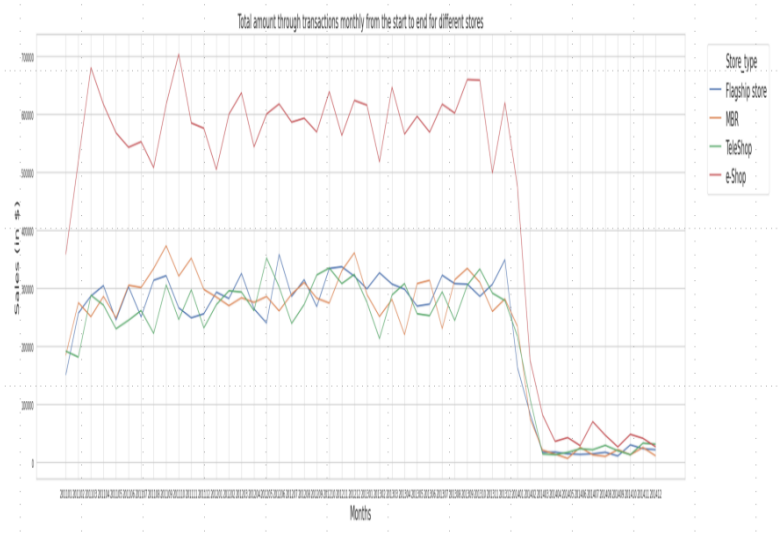


Fig. 1. Sales trend across months for different store types

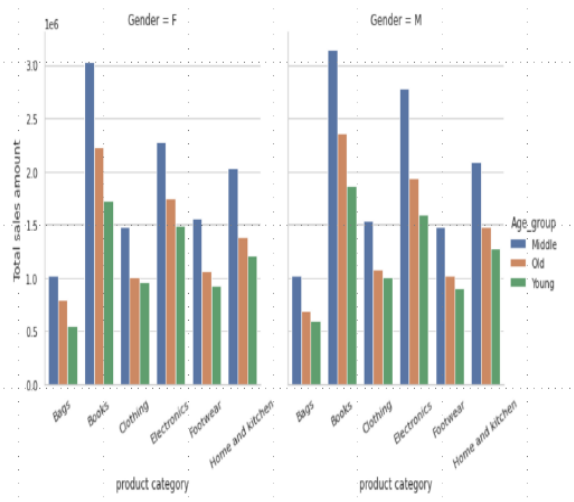


Fig. 2. Total sales amount for different product categories filtered on gender and age group

According to Fig. 2, middle-aged consumers account for the majority of sales across all product categories, which is to be expected given that this is the age at which most people start to become financially capable. The segmentation of books and bags as the highest and lowest sales categories for both genders (Male and Female) is another finding from

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

e-ISSN : 2583-1062

Impact Factor : 5.725

this visualisation. Figure 3 shows that teleshops and e-shops account for more than 60% of all transactions in the performance of retail stores.

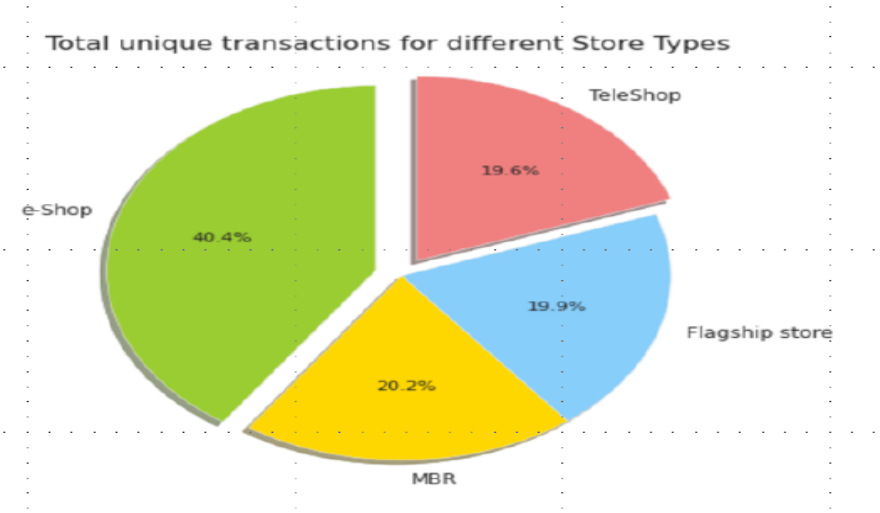


Fig. 3. Distribution of different store types

Approach: The data discrepancy is one of the main obstacles in any data science problem. The following are a few of them:

1. As we can see, there was a 144% decline in overall sales in 2014, which raises serious questions about the figures.
2. Transactions occasionally occur where the quantity is negative. Although the customer may view these as returns, there were still no transactions where the customer had previously purchased the item.
3. There is no information available regarding the several purchases the customer made in a single transaction. This would restrict giving a product-level suggestion at the SKU level.

Measures: We have done the following actions to modify the data and prepare it for further work in order to make the profiling and modelling process impartial:

1. Limit the data to those from 2011–2013.
2. Eliminate all transactions with negative amounts.
3. Sales, transactions, and quantity are some of the most important metrics that go through a process known as winsorization. To lessen the impact of outliers, a transformation is performed by reducing the extreme values of a data set.

4. PROBLEM-1 CUSTOMER SEGMENTATION

Retail establishments will be able to learn a lot about their clients thanks to customer segmentation, which will help them better meet their demands. This will also enable them to better design an acquisition strategy and personalise their communication to the customer's life cycle. K-means clustering is the approach employed in this situation. The selection of features was the initial step in this procedure. Since the data set only had a small number of numerical columns, an iterative method was used to choose the relevant variables for clustering.

Approach: Following is how the k-means algorithm operates:

1. Randomly select k data points (seeds) to serve as the initial centroids, or cluster centres.
2. Assign each data point to the closest centroid.
3. Recompute the centroids using the current cluster members.
4. If a convergence criterion is not met, the algorithm loops back to Step 2 and repeats itself until convergence is achieved.

Criterion for Stopping/Convergence

1. There was little to no re-clustering of data points.
2. No (or little) change in the centroids.
3. Little reduction in the sum of squared errors (SSE)

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} dist(x, m_j)^2$$

$$m_j = 1/n_j \sum_{x \in C_j} x$$

where,

C_j is the j^{th} cluster,

M_j is the centroid of cluster C_j ,

N_j is the number of points in cluster C_j ,

dist(x, m_j) is the distance between data point x and centroid m_j (generally Eucledian)

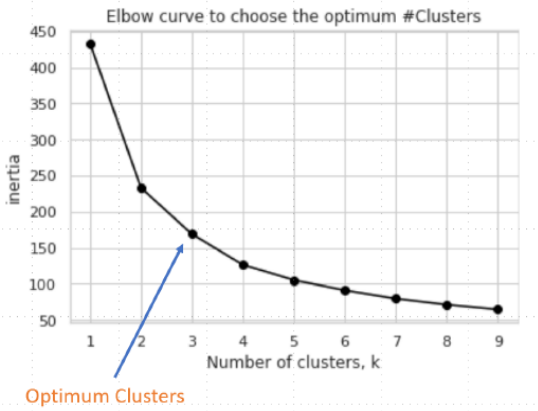


Fig. 4. Elbow curve to chose optimum number of clusters

The factors taken into account are Sales per Transaction (which provides information on the basket value), Age (the customer's age), Number of Transactions (which indicates the frequency of purchases), Quantity, and Quantity/Transaction (which indicates the size of the basket).The K means procedure was applied to several combinations of these variables, and the results showed that the following set of variables—Sales/transaction, Quantity, and the number of transactions—could be divided into 3 distinct clusters. The statistical test of the elbow curve, which displays the within the sum of square distances for various clusters in fig. 4, was used to determine the ideal number of clusters.

The elbow technique applies k-means clustering to the data set using a range of k values (from 1 to 10) and then calculates the average score for each value of k. The total of the squared distances between each point and its designated centre is used to calculate the distortion score. The ideal number of clusters was 3, as the decrease in inertia after the third cluster is not that substantial.

Results: Figures 5 and 6 depict the cluster in three dimensions. Although there is some overlap between the clusters when seen, a decision boundary is visible. The three clusters were deduced to be bargain hunters/seasonal buyers (BH) (lesser basket value and transactions), high spenders (HS) (with greater basket value), and regular shoppers (RS) (more frequent purchases) after carefully analysing the data points in these clusters.[7] From a business perspective, this makes sense, and it can be a valuable insight for the retail company when launching promotions or deals.

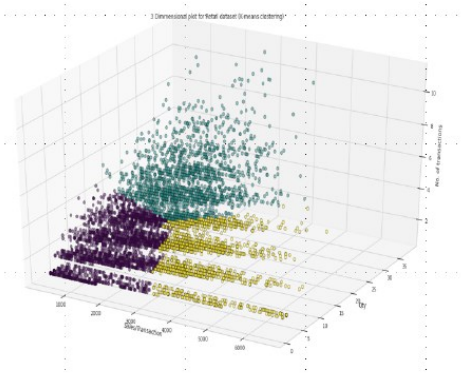


Fig. 5. 3-dimensional scatter plot for K-means

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

e-ISSN : 2583-1062

Impact Factor : 5.725

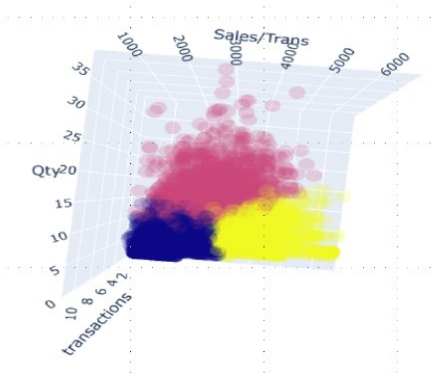


Fig. 6. Top view of the scatter plot for K-means

4. PROBLEM 2- SALES FORECASTING

What happens if the store wants to be well-equipped to deal with a sizable number of consumers at a special occasion? In order to build up logistics and inventory planning, it requires to have a very clearly defined approach[4]. Every business estimates its future sales before beginning this procedure, which can be quite helpful in advising the business on how to manage its resources and inventories effectively. Research shows that businesses with correct sales are 10% more likely to see annual revenue growth.

The retail store's weekly sales distribution is depicted in Fig. 7. The data shows several time series characteristics. The cycle pattern is indicated by the red boxes in the graph and is discussed in the following sections.

Fig. 7. Weekly sales trend

Approach:

Linear Regression

The forecasting issue can be approached in a variety of ways. Linear regression is a well-known statistical technique for forecasting an actual output with a value. In an effort to discover the line that minimises squared error and best fits the data, it attempts to simulate the relationship between a dependent variable and a number of independent variables. The following is the linear regression programme:

$$\min \sum_{i=1}^{n} (y^{(i)} - w^T x^{(i)} - b)^2$$

where,

$y^{(i)}$ is the dependent variable

$x^{(i)}$ is a matrix of independent variables

w and b are the coefficient and the intercept for the equation

The use of linear regression models for prediction is justified by the following four main hypotheses:

Linearity: The relationship between dependent and inde pendent variables should be linear

homoscedasticity: The variance of residuals should be same for any value of independent variable

No Auto correlation: The error terms shouldn't be correlated to one another

Normality: The errors should be normally distributed

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

e-ISSN : 2583-1062

Impact Factor : 5.725

Independence: The variables are independent of each other

Time Series Analysis

The phrase "time series analysis" describes a group of data points that are examined at regular intervals to ascertain the behaviour or pattern of the variables. We are able to make more progress in our pursuit of forecasting and prediction because to this analysis. The time dependency is the distinctive feature that sets linear regression apart from time series. Each observation depends on the others[5].

Stationarity is a crucial requirement that must be met in order to use the time series formulation. The supposition asserts that a time series' mean and variance remain constant over time. The modelling issue would be simpler to resolve once their consistency was guaranteed. Trend, seasonality, or cycles are the three types of patterns that may be found in the majority of time-series data.

Trend - A time series' overall behaviour is described by its trend. A time series has an upward trend if it has a positive long-term slope over time, and a downward trend if it has a negative slope.

Seasonality - A seasonal pattern is any variation in a time series brought on by events on the calendar. These occasions may be seasons, days of the week, or times of the day. Seasonality has set frequencies every year. The beginning and end of the seasonal trends fall within the same week. Think about the holidays Black Friday and Cyber Monday. The data will clearly show that the sales at this time are expected to increase.

Cycle - Cycles are characterised as increases and falls with variable magnitudes that can last more than a year. They don't repeat themselves. They typically result from outside forces, which makes them considerably more difficult to forecast. These patterns are used by time series forecasting to provide accurate forecasts. The breakdown of the weekly sales data into its many components is shown in Fig. 8 below.

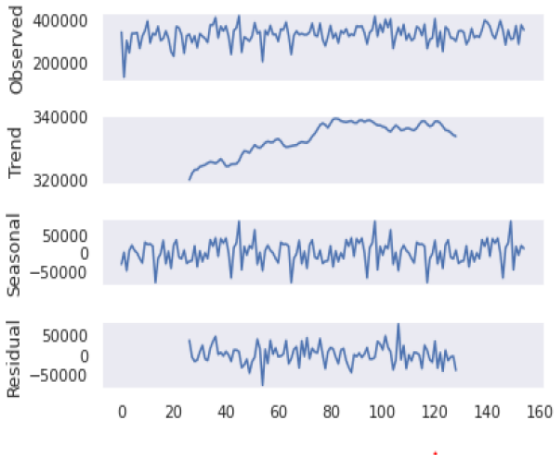


Fig. 8. Components of time series

5. METHODOLOGY

In this project, we used both of the aforementioned approaches to begin the modelling process. The ARIMA (Auto-Regressive Integrated Moving Average) is a particular class of time series that is implemented. The only predictors used in these models for forecasting are the forecast errors and the lag terms of the variables.

Predicting the retail store's sales value is our key goal. Since the statistics are more reliable at a lower level of granularity—weekly—we would like to estimate revenues at that level. The data spans 155 weeks between 2011 and 2013. The progression of this data leads to the ultimate predictions shown in Fig 9 as follows

1. The dataset is divided into a training set and testing set. The training set is the set on which the entire algorithm trains while simultaneously capturing all previous variants. Here, the test set serves as the validation set where we would learn how effective our model is. Training and testing sets are typically divided 80:20 as a general rule.
2. The time series and linear analytic presumptions are now being evaluated on this data. Due to the presence of auto-correlation in this situation, the conditions of linear regression would not be met; yet, we force-fit the model to grasp the baseline score using the oldest statistical method[3].
3. The ARIMA model incorporates the following parameters:

a. The number of lag observations to be taken into account in the model is indicated by the AR term (p).

b. The number of lag forecast errors to be considered is provided by MA (q).

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

e-ISSN : 2583-1062

Impact Factor : 5.725

c. The amount of differences between the observations that must be taken to keep the time series steady is known as the differencing term (d). All of these parameters were obtained from the plots of partial autocorrelation (PACF) and autocorrelation (ACF) presented in fig.9

4. By iterating through several combinations, we were able to determine that the values of p and q should be 3 or 4, according to the plots.

5. Various accuracy metrics are recorded to comprehend how both models function.

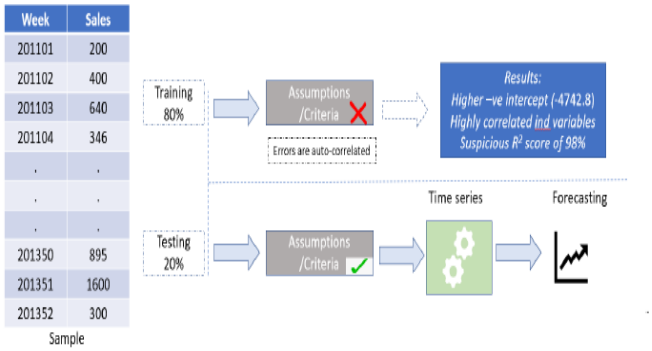


Fig. 9. Flowchart for Time series modelling

Forecasting Results: A linear equation with an extremely high -ve intercept of -4742.8 is produced by the linear regression. [1]This value is highly erratic (because sales cannot be negative), as it effectively indicates that the model would forecast this value as the sales for the future without the influence of any lag variables. Additionally, the R2 produces a score of 98%, which seems extremely dubious. Additionally, it should be noted that, in the case of linear regression, the data does not conform to the auto-correlation assumption[5].

Fig. 11's ARIMA model results show several intriguing findings that we can see. Most of the model's variables have p-values that are less than 0.05, making them highly significant. The AIC value (3047), which indicates how much information the model has lost, is the lowest of all the iterations that the model has chosen. The top model finally obtains a MAPE of 9.76%.

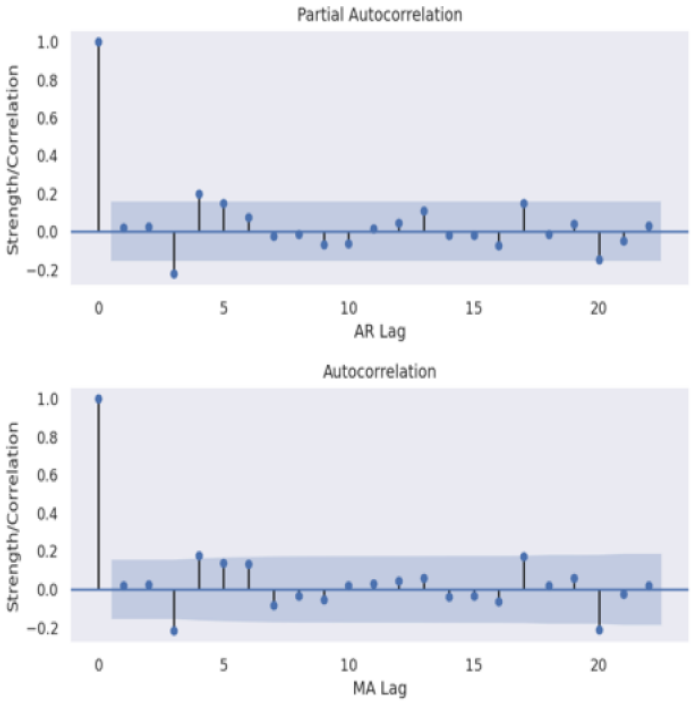


Fig. 10. PACF and ACF plots

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

ARMA Model Results

Dep. Variable:	total_amt	No. Observations:	126
Model:	ARMA(3, 4)	Log Likelihood	-1514.835
Method:	css-mle	S.D. of innovations	39394.206
Date:	Thu, 07 May 2020	AIC	3047.669
Time:	17:50:45	BIC	3073.196
Sample:	0	HQIC	3058.040

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.285e+05	3899.993	84.228	0.000	3.21e+05	3.36e+05
ar.L1.total_amt	0.0595	0.105	0.566	0.572	-0.146	0.265
ar.L2.total_amt	-0.2028	0.097	-2.092	0.039	-0.393	-0.013
ar.L3.total_amt	-0.8597	0.106	-8.141	0.000	-1.067	-0.653
ma.L1.total_amt	-0.0452	0.133	-0.339	0.735	-0.306	0.216
ma.L2.total_amt	0.3551	0.130	2.728	0.007	0.100	0.610
ma.L3.total_amt	0.6978	0.136	5.133	0.000	0.431	0.964
ma.L4.total_amt	0.2199	0.109	2.017	0.046	0.006	0.434

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.4621	-0.8884j	1.0014	-0.1737
AR.2	0.4621	+0.8884j	1.0014	0.1737
AR.3	-1.1600	-0.0000j	1.1600	-0.5000
MA.1	0.4765	-0.8793j	1.0001	-0.1710
MA.2	0.4765	+0.8793j	1.0001	0.1710
MA.3	-2.0631	-0.5385j	2.1322	-0.4594
MA.4	-2.0631	+0.5385j	2.1322	0.4594

Fig. 11. ARIMA - Summary

Results of Seasoned ARIMA (SARIMA): Even while the obtained parameters for the ARIMA model show good performance, there is always room to improve them by taking even more complex elements into account. The SARIMA model [3] incorporates seasonal variables to improve outcomes from the current model. The seasonal components are represented by P, D, and Q and are the counterparts of the ARIMA parameters p, d, and q. To determine the ideal seasonal settings, pdm Arima was used in a number of iterations (fig. 12). The outputs are displayed in the following figure.

According to SARIMA's findings, the model was able to account for greater variation than earlier models (fig. 13). As there are various offers available during various seasons and events, the seasonal component plays a significant impact in retail sales. However, the AIC (1817) and BIC (1828) of the model's MAPE offer 10.7%, somewhat higher than ARIMA model is drastically shrunk. Additionally, the p-values for the variables are below the 0.05 threshold for significance. As a result, the SARIMA model produces the best results for the data on retail sales, and the forecast is displayed in Fig. 14.

```
Performing stepwise search to minimize aic
Fit ARIMA(1,1,1)x(0,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(0,1,0)x(0,1,0,52) [intercept=True]; AIC=1874.011, BIC=1878.592, Time=0.366 seconds
Fit ARIMA(1,1,0)x(1,1,0,52) [intercept=True]; AIC=1824.251, BIC=1833.413, Time=3.562 seconds
Fit ARIMA(0,1,1)x(0,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(0,1,0)x(0,1,0,52) [intercept=False]; AIC=1885.167, BIC=1887.457, Time=0.296 seconds
Fit ARIMA(1,1,0)x(0,1,0,52) [intercept=True]; AIC=1836.528, BIC=1843.399, Time=0.450 seconds
Fit ARIMA(1,1,0)x(2,1,0,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(1,1,0)x(1,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(1,1,0)x(0,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(1,1,0)x(2,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(0,1,0)x(1,1,0,52) [intercept=True]; AIC=1854.608, BIC=1861.479, Time=3.085 seconds
Fit ARIMA(2,1,0)x(1,1,0,52) [intercept=True]; AIC=1824.047, BIC=1835.499, Time=4.926 seconds
Fit ARIMA(2,1,0)x(0,1,0,52) [intercept=True]; AIC=1839.002, BIC=1848.164, Time=1.133 seconds
Fit ARIMA(2,1,0)x(2,1,0,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(2,1,0)x(1,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(2,1,0)x(0,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(2,1,0)x(2,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(3,1,0)x(1,1,0,52) [intercept=True]; AIC=1818.201, BIC=1831.944, Time=8.691 seconds
Fit ARIMA(3,1,0)x(0,1,0,52) [intercept=True]; AIC=1834.289, BIC=1845.741, Time=1.787 seconds
Fit ARIMA(3,1,0)x(2,1,0,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(3,1,0)x(1,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(3,1,0)x(0,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(3,1,0)x(2,1,1,52) [intercept=True]; AIC=nan, BIC=nan, Time=nan seconds
Fit ARIMA(3,1,1)x(1,1,0,52) [intercept=True]; AIC=1819.950, BIC=1835.984, Time=13.449 seconds
Fit ARIMA(2,1,1)x(1,1,0,52) [intercept=True]; AIC=1824.203, BIC=1837.946, Time=12.377 seconds
Total fit time: 50.209 seconds
```

Fig. 12. Parameter Tuning - SARIMA

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

Statespace Model Results

Dep. Variable:	total_amt	No. Observations:	126
Model:	SARIMAX(3, 1, 0)x(1, 1, 0, 52)	Log Likelihood	-903.551
Date:	Fri, 08 May 2020	AIC	1817.101
Time:	00:06:35	BIC	1828.554
Sample:	0	HQIC	1821.665
	- 126		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6290	0.131	-4.811	0.000	-0.885	-0.373
ar.L2	-0.2707	0.079	-3.423	0.001	-0.426	-0.116
ar.L3	-0.2975	0.103	-2.891	0.004	-0.499	-0.096
ar.S.L52	-0.3747	0.033	-11.358	0.000	-0.439	-0.310
sigma2	2.663e+09	2.11e-11	1.26e+20	0.000	2.66e+09	2.66e+09

Ljung-Box (Q):	27.54	Jarque-Bera (JB):	96.07
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	1.13	Skew:	-1.60
Prob(H) (two-sided):	0.78	Kurtosis:	7.62

Fig. 13. SARIMA - Summary

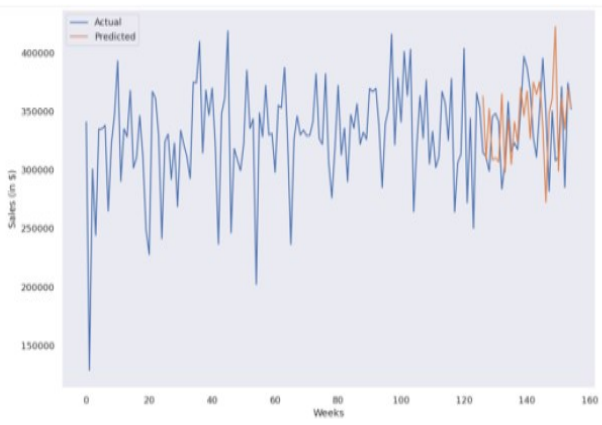


Fig. 14. SARIMA Forecast

6. CUSTOMER PRODUCT AFFINITY

Up to this point, two key facets of retail marketing were discussed. The former works with client segmentation, and the latter assists with sales forecasting to improve logistics and inventory control. The kind of products recommended to the consumer segment would be the missing link between these, increasing the likelihood of cross-selling and up-selling. This and MBA (Market Basket Analysis), a widely used method to determine the ideal combination of goods or services consumers frequently purchase, are quite similar. The best option given the limitations of the data was to use sophisticated data analysis techniques and mining to produce suggestions that were quite accurate.

Approach: The solution to this issue would be to comprehend how each retail store consumer purchases several subcategories. The top three subcategories that each consumer purchased, as well as the chain of subcategories they purchased, are the focus of a new line of study that we have proposed. We use customer segmentation to deliver recommendations to various groups in order to obtain a more accurate recommendation. The last stage would be to make offers to particular clients on the collection of desired subcategories. Figure 15 demonstrates this well.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

e-ISSN : 2583-1062

Impact Factor : 5.725

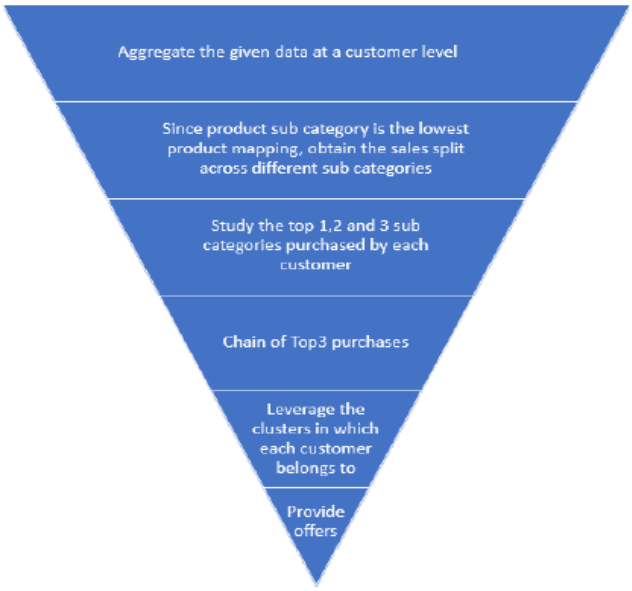
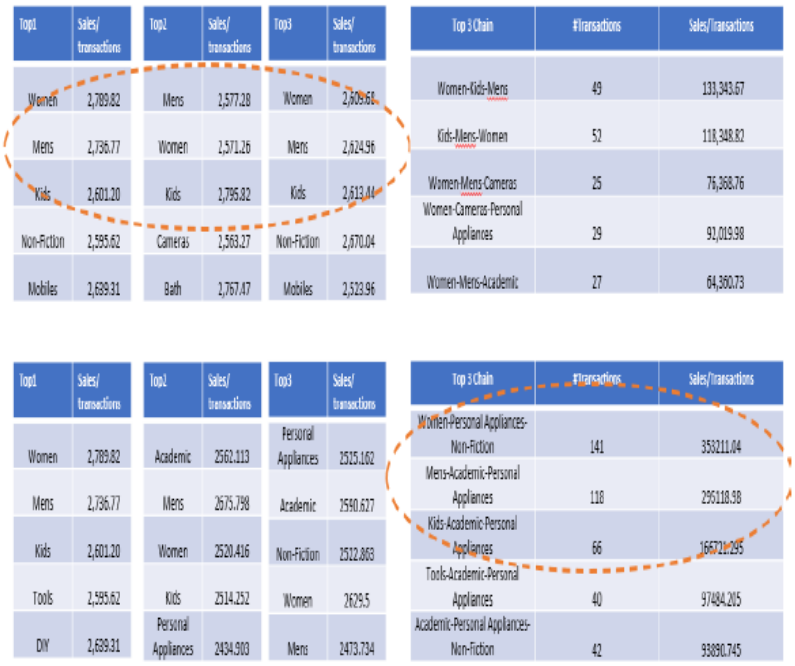


Fig. 15. Product Affinity Methodology

Results: The purchase analysis for the segment Regular Shoppers (RS) is displayed in the top two tables of Figure 16. It is evident that the top 3 and the chain belong to adults, adolescents, and children. This can assist us in determining, with some confidence, that regular shoppers are more inclined to shop in the clothing sector. Similar to the top 2 and 3 purchases, which reflect the academic and personal appliances categories, the chain is more consistent in the bottom 2 tables. Both Bargain Hunters (BH) and greater Spenders (HS) engage in this behaviour, and making such suggestions can result in greater sales conversions and increased profitability for the retailer. It is noticed that customer segmentation had a greater influence on customer grouping and on focusing the analysis among the clusters.



Top1	Sales/ transactions
Women	2,789.82
Mens	2,736.77
Kids	2,601.20
Non-Fiction	2,595.62
Mobiles	2,639.31

Top2	Sales/ transactions
Mens	2,577.28
Women	2,571.26
Kids	2,795.82
Cameras	2,563.27
Bath	2,767.47

Top3	Sales/ transactions
Women	2,605.61
Mens	2,624.96
Kids	2,613.44
Non-Fiction	2,670.04
Mobiles	2,523.96

Top 3 Chain	#Transactions	Sales/Transactions
Women-Kids-Mens	49	133,343.67
Kids-Mens-Women	52	118,348.82
Women-Mens-Cameras	25	76,368.76
Women-Cameras-Personal Appliances	29	92,019.98
Women-Mens-Academic	27	64,360.73

Top1	Sales/ transactions
Women	2,789.82
Mens	2,736.77
Kids	2,601.20
Tools	2,595.62
DIY	2,639.31

Top2	Sales/ transactions
Academic	2562.113
Mens	2675.738
Women	2520.416
Kids	2514.252
Personal Appliances	2434.903

Top3	Sales/ transactions
Personal Appliances	2515.162
Academic	2590.627
Non-Fiction	2512.863
Women	2629.5
Mens	2473.734

Top 3 Chain	#Transactions	Sales/Transactions
Women-Personal Appliances-Non-fiction	141	353211.04
Mens-Academic-Personal Appliances	118	295118.98
Kids-Academic-Personal Appliances	66	166721.295
Tools-Academic-Personal Appliances	40	97484.205
Academic-Personal Appliances-Non-fiction	42	93890.745

Fig. 16. Left - Cluster Regular Shoppers and Right - Cluster Bargain Hunters and High Spenders

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 363-373

e-ISSN : 2583-1062

Impact Factor : 5.725

7. CONCLUSION

In conclusion, this paper addresses a total of three issues. The statistical K-means method produces three groups of customers: Bargain Hunters (BH), High Spenders (HS), and Regular Shoppers (RS). This appears to be extremely pertinent as we can see that almost 41% of transactions have basket values that are higher than the average basket value for the specified period. In a similar vein, 33% of all customers visited more frequently than usual, on average, according to the data. In the flow of goods, inventory control and logistics management are extremely important for a retail store. Thus, forecasting tools aid retail establishments in being well-prepared for a massive influx of clients at particular times. The last issue suggests the kinds of goods that should be advertised to consumers.

This research is concluded with a brief examination of the sales of several subcategories. In a retail setting, the products recommended to frequent customers have the largest sales share while those recommended to spenders and bargain hunters have the lowest sales share. This demonstrates that the latter are highly seasonal and are rapidly seized by the customers to benefit from the promotions.

8. REFERENCES

[1] Nunnari and V. Nunnari, "Forecasting monthly sales retail time series: A case study", 2017 IEEE 19th Conference on Business Informatics (CBI), vol. 1, pp. 1-6, 2017.

[2] Niu Y. Walmart Sales Forecasting using XGBoost algorithm and Feature engineering. In 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), 2020: 458-461.

[3] P. Sobreiro, D. Martinho and A. Pratas, "Sales forecast in an it company using time series", 2018 13th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-5, 2018

[4] Wang, C.H.; Chen, T.Y. Combining biased regression with machine learning to conduct supply chain forecasting and analytics for printing circuit board. Int. J. Syst.Sci. Oper. Logist. 2022, 9, 143–154.

[5] Ulrich, M.; Jahnke, H.; Langrock, R.; Pesch, R.; Senge, R. Distributional regression for demand forecasting in e-grocery. Eur. J. Oper. Res. 2021, 294, 831–842

[6] Brviera-Puig, A.; Buitrago-Vera, J.; Escribá-Pérez, C. Internal benchmarking in retailing in retailing with DEA and GIS: The case of loyalty-oriented supermarket chain. J. Bus. Econ. Manag. 2020, 21, 1035–1057.

[7] K. C. Dewi, P. I. Ciptayani, N. W. D. Ayuni and I. B. P. S. Yudistira, "Modeling Salesperson Performance Based On Sales Data Clustering," 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022, pp. 390-396, doi: 10.1109/ISRITI56927.2022.10052816.