Market sales prediction by analyzing customer buying patterns using machine learning.

BY
Tamanna Rahman Himi
ID: 181-15-1794
AND
Prianka Binte Zaman
ID: 181-15-1958
AND
MD. Al-Amin
ID:181-15-1986

This Report conferred in Partial Fulfillment of the necessities for the Degree of Bachelor of Science in engineering and Engineering

Supervised By
Dr. S. M. Aminul Haque
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By
Tania Khatun
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

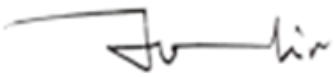
DECEMBER 2021

APPROVAL

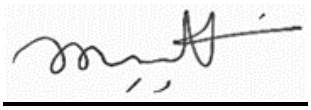
This Project titled **"Market sales prediction by analyzing customer buying patterns using machine learning.",** submitted by **Tamanna Rahman Himi, ID No: 181-15-1794, Priyanka Binte Zaman, ID No: 181-15-1958** and **MD. Al-Amin, ID No: 181-15-1986** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation was held on 18.01.2022.

BOARD OF EXAMINERS

Sheak Rashed Haider Noori **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Daffodil International University



Ohidujjaman **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Prof. Mohammad Shorif Uddin **External Examiner**
Professor,
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We herewith declare that this project has been done by us under the supervision of **Dr. S.M. Aminul Haque, Assistant Professor, and Department of CSE,** Daffodil International University. We even have an inclination to jointly declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or credentials.

Supervised by:

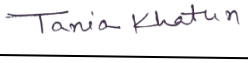
Dr. S. M. Aminul Haque

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised by:



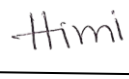
Tania Khatun

Sr. Lecturer

Department of CSE

Daffodil International University

Submitted by:

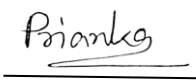


Tamanna Rahman Himi

ID: -181-15-1794

Department of CSE

Daffodil International University



Prianka Binte Zaman

ID: -181-15-1958

Department of CSE

Daffodil International University

MD. Al-Amin

ID: -181-15-1986

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First of all, we might praise the Almighty ALLAH for whom our final year analysis paper has been completed with success with no major hurdle.

We are extremely thankful and want our deep financial obligation to **Dr. S. M. Aminul Haque**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep knowing & keen interest of our supervisor within the field of "***Machine Learning***" to hold out this project. His vast endurance, studious steering, gradual animation, constant and diligent oversight, constructive sricture , valuable recommendation, reading several inferior drafts and correcting them in any respect stages have created it attainable to finish this project.

We would like to express our heartiest feeling to **Akteruzzaman Paramanik**, Lecturer, Department of CSE, and **Amir Sohel,** Lecturer, Department of CSE and **Tania Khatun**, Sr. Lecturer**,** Department of CSE, for their kind to facilitate the end of our project and to other faculty members and the staff of Daffodil International University. We might like to impart our all coursemates in Daffodil International University, who took part in discussion while completing our course work. Finally, we should acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

To improve a business, a company has to analyze the type of purchases they must keep track of the merchandise that are marketing the foremost in order that they will keep stock of these class merchandise and take away those forms of class that are marketing less. 'Sales' is the crucial success issue of a business. Increasing sales may be an excellent impact factor for a developing business. During this trendy time, it may be done by victimizing trendy technology like AI, machine learning, and deep learning. So, we are needing to do that job victimization machine learning by utilizing algorithms. In our research we have a tendency to act on however a mercantile establishment will get a lot of sales from its product victimization of its customer's previous product shopping for information. We've to preprocess victimization using totally different pre-processing techniques. Information exploration, data transformation and engineering play an important role in predicting correct results. This paper discusses a way to predict sales maximization by information analysis and the way to evaluate the effectiveness of machine learning techniques. Sales analysis of products is one of the major issues of identification buying frequency pattern. We proposed a model to predict seasonal products which is the "ARIMA" model. This model works to do time series analysis. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Our recommended models can be used to get an idea of which products need to be kept on a shop's shelves and which products are not for the advantage of the customer. Based on the customer's purchases for a few years this model will be able to recommend which products are more popular in which season.

Keywords: ARIMA,time series, forecasting, non-stationary .

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Nowadays business manages a large repository of knowledge. Data volume associated to extend any in an indicative manner. Measures are obligatory to extend the speed of the dealing process and also the expected increase in knowledge volume and client behavior. One among the objectives of this analysis work is to seek out reliable sales [18]. And this model will also be able to recommend which products are more popular in which season that use machine learning techniques to achieve the best possible revenue. Selling more products is the main objective for a company or a shopkeeper. If anyone wants to increase their sales, they should follow some tricks [19].

Research into the use of data mining in the retail business in other nations began far earlier than in the United States. A large portion of data mining in a company is devoted to analyzing information about consumers and products. Most of them are used for forecasting demand, inventory demand, retail location selection, and pricing research. In the world of commerce and industry, Wal-beer Mart's and diapers stand out as classic examples of basket analysis [20]. As a result, companies need to do thorough cost-benefit analyses to determine the most effective use of their resources for marketing and sales activities over time [21]. Decision support models that link expenses to customer behavior and anticipate the value of a client's portfolio will help decision makers. According to business knowledge and data, acquiring a new client costs five to 10 times as much as retaining an existing one [22]. While the specifics of maintaining consumers may vary greatly depending on the company setting, academics and practitioners [23] alike have paid significant attention to the issue [24]. The emergence of customer insights by marketing analysts in recent years has resurrected the subject of purchase prediction in consumer research. Customers' present state is not immediately visible at a time T and the accessible history information is suppressed at point T, making it difficult to accurately anticipate their future purchases in non-contractual contexts, as described in [25].

In order to grow a business, an organization can analyze the type of purchases of its customers in order to keep an eye on such products, store those products and remove

the products that sell the least. Between buying and selling, we offer machine learning based models by analyzing product purchases and comparing best results through purchase frequency patterns. Which product we proposed by our model needs to be kept more on their shelves and that merchandise that isn't for the convenience of the purchasers should be removed. We are hopeful that our model will be helpful to all stockholders.

1.2 Motivation

When we started research on the growing gravity of classifying the customers behavior and a large number of manageable forecast models and the data sources, we noticed that a small amount of work had been done. As has been done in the past on these issues, very few detailed essays have been published. In these matters, we thought that little research was done on these issues. The field of its research can be wide, bring a new perspective to it, see where it is. In a way of globalization international trade, and a myriad of international tv channels have exposed shoppers to new ideas. Looking from supermarkets, long thought-about a Western thought, is slowly being accepted by the people. We need to manipulate different parameters. Super Shops have so many products that sometimes they even couldn’t manage the product expiry date. For this people are sometimes buying date expiry products without their knowledge. That is why we have chosen Market Sales Prediction for research so that we can easily overcome this problem through our model.

1.3 Case of the Study

This analysis works to seek out the prediction mechanism for subsequent purchases that is enforced by victimization machine learning techniques to realize the most effective potential revenue. They shortly analyzed the idea of sales information and sales forecast and performance analysis, a best-suited prophetical model is recommended for the sales and seasonal sales trend forecast. The results are summarized in steps of the responsibility and accuracy of economical techniques taken for prediction and prognostication. they're making an attempt to search out the most effective algorithmic rule that shows the most accuracy in prognostication and future or next sales prediction.

We hope This paper suggests a machine learning approach to business improvement that predicts customer buying behavior so that customers can track what they buy from a superstore. The paper further shows that a customer will buy the next key by collecting customer purchase information using machine learning algorithms. Thus, the proposed thesis paper will predict the future potential of the customer buying pattern.

The main objective of this research is machine learning techniques to find reliable sales. Selling more products is the main objective for a company or a shopkeeper.

1.4 Research Questions

We know that problem recall is the first attempt to solve a problem. Inside a market economy plays a vital role in production by identifying the needs of the customers desired service and subsequent market penetration.

As our goal is to maximize the sales of the retail industry our main queries are:

Q1. “Can we solve the problem of wasting products by the expiration date?

Q2. “Can we propose a model to predict which products are more popular in which season?”

1.5 Research Objective

- To recognize what customer behavior is and the different types of consumers.
- To find out which product is more popular in which season.
- In certain seasons, it can be determined which products are more likely to be bought and sold.
- To predict which products are selling best and which products should be on their shelves.
- Reduce the space capacity problem of stores.
- This will benefit the Stakeholders.
- To understand the association between buyers, what the customer wants and the market related concept.
- New dimensions are being added for further exploration.

1.6 Expected Outcome

Our awaited outcome from this study is

- Increasing Sales.
- Decrease Product expiring rate.
- High Consumer satisfaction.
- Store capacity problem will be solved.
- Wasting space problem will also be solved.

1.7 Report Layout Chapter

The report has total 5 Chapters which will be followed given by instructions:

Chapter 1 of this research report has been summarized. The primary focus of this chapter is on introduction and discussion. This chapter does a good job of explaining why people are inspired. This sensitive study is also essential because it shows what the research questions will be and what the expected outcome of the investigation will be, as explained in the previous section.

Chapter 2 gives a brief introduction additionally as connected works that help us. to understand and implement the work. Also, we tend to mention the analysis outline, Scope of the issues and challenges that we tend to have to beat in our analysis work.

Chapter 3 gives the statistical methods of this work are discussed in the theoretical interpretation of the research. These methods are illustrated in this chapter, and the final section describes how the model is evaluated using the machine learning model.

Chapter 4 contains the results of this study and are detailed and discussed. Some research-related images have made it easier to understand the criteria for work

Chapter 5, this section contains the conclusions. This is important for achieving the whole division. The concept of a significant research study was presented. Also, what are the restrictions on conducting this research that will be useful to other scholars in the future.

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries/Terminologies

Retail store

A retail shop is a center where products are sold primarily to end users. It is generally owned and run by a retailer, although it can also be owned and operated by a manufacturer or someone other than a retailer. In other terms, a retailer, often known as a retail store, is a company whose major source of revenue is retailing. All of the actions involved in retailing are referred to as retailing.

Department store

A department store is a big retail establishment that sells a wide range of products. It offers a diverse range of products in each category and is structured into several divisions for purchasing, advertising, service, and management. Military canteens are examples of mass retailing departmental stores.

Super markets

A supermarket is meant to meet all of a person's food, washing, and housekeeping needs.It has a rather huge size. Its business model is low-cost, low-margin,high-volume, and self-service.

Super store

Consumers' entire demands for frequently purchased food and nonfood products are met by superstores.

Off price retailer

Leftover products, overruns, and anomalies purchased at discounted rates from manufacturers or other merchants are sold by an off-price store. There are three categories of off-price shops.

Catalogue showroom

Customers order items from a showroom catalogue. The products are then picked up at a store's merchandise pickup location

Machine Learning

Machine learning is a branch of artificial intelligence (AI) that allows computers to learn and develop on their own without having to be explicitly programmed. Machine learning is concerned with the creation of computer programs that can access data and learn on their own. Observations or data, such as examples, direct experience, or instruction, are used to seek for patterns in data and make better judgments in the future based on the examples we offer. The fundamental goal is for computers to learn on their own, without the need for human involvement, and to adapt their behavior accordingly. However, traditional machine learning algorithms treat text as a series of keywords, but a semantic analysis method replicates the human ability to comprehend the meaning of a document.

2.2 Related Works

E. Gummesson had mentioned the values of the long-term relationship in the middle of business and its consumers [25]. Though there have been several studies within the 90s to know client purchase patterns it was not enough. In the era of 2000, modern technology took over the businesses industries well of data storing. A company's sales study report shows a trend that occurs in sales knowledge over a time. The sales study report shows whether sales are increasing highly or decreasing. Sales analysis reports in large corporations may contain only helpful and category or data for the region. A small business is more interested in reducing sales. Specialized business with a single location. The general sales data was compact enough to use. The sales analysis report compares the actual sales with the estimated sales. Linear regression and logistic regression are two machine learning models that can easily fit this type of problem. A line of high-selling and low-selling products, quarters, and zones for a product.

A hybrid intelligent prediction method introduced by Liu Weixiao incorporates elements of both an artificial neural network (ANN) and an ANN. Through the use of correlation degree analysis, he was able to identify influential factors with high correlation degrees. After employing DGM (1,1) and ANN to make a forecast, the concept of polynomial residuals was presented [26]. Zheng Jun et al used clustering analysis technology [27] to optimize the categorization of items in logistics management, and data mining technology was used to tackle the problem of logistics network distribution. In order to overcome the Big Data challenge, Zhang et al. [28][29] developed a weighted combination approach and a Fuzzy RDF Model. In the past, most studies have predicted accuracy from the viewpoint of items, but few have examined client traits and purchase patterns. There is still a lot of potential for product-targeted sales strategy research to be done.

Marzia et al. had mentioned careful literature reviews associated with the appliance of prognostication analytics in client relationship management [30]. Xu and Walton projected an analytical CRM system for client information accusations [31]. Buckinx et al had projected a prediction model for the purchaser's future disbursement pattern. Guimei et al. had explored Alibaba sales information and projected prediction necessities and most significant options exploitation features of engineering and machine learning. As this can be the primary analysis on this subject of purchase behavior analysis exploitation machine learning, all our approaches are a unit distinctive.

2.3 Comparative Analysis and Summary

Comparative Analysis and Summary We wish to do our research from the perspective of our own country; however, we lack the necessary data. As a result, we'll be using Big Mart sales data for our research. As we investigate similar work, we can see that

they use the same data, but they leave out a few features for their work, and a few are worked on as is, but the accuracy in both terms is not sufficient. A few features must be eliminated, although they are not the same as in earlier efforts. We can forecast retail shop sales using these data

2.4 Scope & Challenges

When we explore other papers,

- There is no paper who directly deals with time analysis of super shops in Bangladeshi super shops
- Following a review of numerous research publications relevant to our study, it is evident that sales maximization utilizing a machine learning technique is a hot research issue. We also recognize that the job isn't over yet. There is a lot of work to be done, but we are certain that we can make significant improvements to the current one. As a result, we used machine learning to solve the problem in a novel way. Where can we achieve better results in selling items at a retail store?
- The main challenge was in data collecting. It's a major problem to develop any machine learning project.
- We face some problems to filter, clean, merge and find the best accuracy for huge data sets and applying the algorithm on a large data set was a real problem.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The subject of our research paper was to analyze the purchase data of an organization or Supershop and find out which products are selling best and which products should be on their shelves. This will increase the sales of the Supershop and reduce the space capacity problem of stores. It will be easy to make sure that the products do not expire. Which products are not for the convenience of the clients can be removed easily?

Our research topic is "**Market sales prediction by analyzing customer buying patterns using machine learning.**" and in this paper we use some tools like

- Microsoft Excel
- Jupyter notebook etc.

3.2 Working Process

A period format where measurements are recorded between normal time intervals. The motivation behind estimating data is to lay the groundwork for the improvement of the financial system, generation system, creation control and modern methods. The real goal is to achieve the best approximation, that is, to ensure that the average square of the deviation between the actual and determined attributes is as small as possible for each lead-time.

Much work has been done in recent decades to advance and improve the time format estimation models. The conventional model for determining the time format, for example, the Box-Jenkins or Auto regressed Coordinated Moving Normal (ARIMA) model, acknowledges that the considered time format has been created in a simple manner.

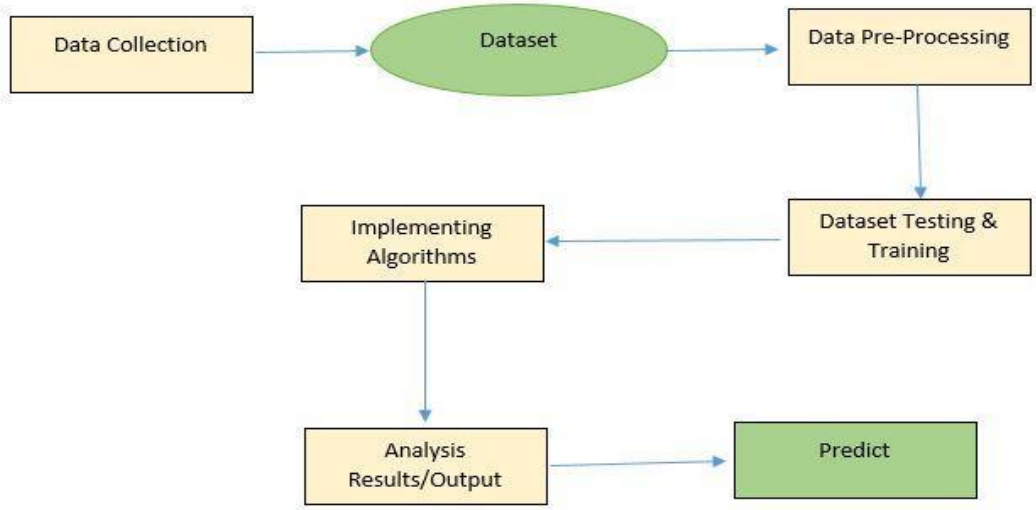


Figure-01: Data Processing Mapping 1

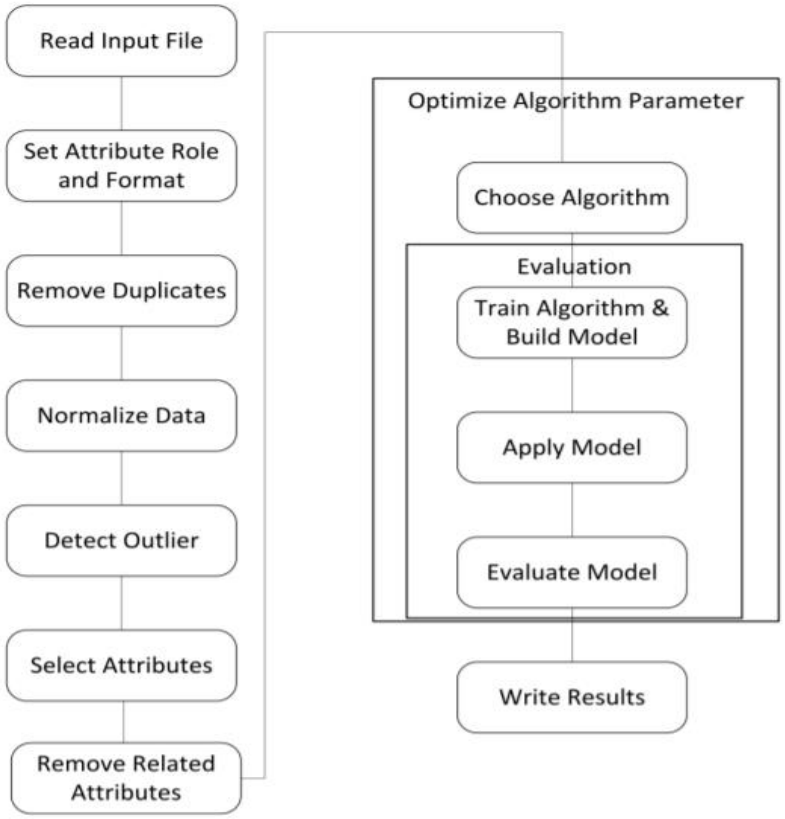


Figure-02: Data Processing Mapping 2

3.3 Data Collection

Data has been collected from **Swapno** super shop and **Family** super shop.

In our study, we used two local super shop dataset which had a variety of information about many products. The Dataset has about 200k purchase's information. Here, we were the ones who got the huge number of unique customers purchasing lots of unique products. In the dataset we have about 200000 unique products. We had pre-processed our data for the implementation. We deleted all unnecessary data and took the date, customer id, product id, date, category, product name, quantity, sell and price. Every product id and customer id are unique numbers. These two supershops every day contain huge amounts of data which helps predict which products are selling more in which season.

3.4 Data Collection Procedure

Data has been collected from **Swapno Super Shop- Mirpur 2**(November, 2021-December, 2021) and **Family Super Shop-Baluchar Bazar** (January, 2014- July, 2017). We are working on sales data from Swapno super shop and there are 192681 data variables. On the other hand, Family super shop was a furniture shop (2014 to 2015) but it's a totally super shop now. There is a huge amount of data in the dataset with the date, product code, product name, category, product id, quantity, price. Data is analyzed in a necessary format to find out the prediction of which product is selling more and which product should be more in the shelves of super shops. That's why it was needed for the integration of sales data. Then the data set is prepared for further work.

Swapno Super Shop Dataset:

Attribute	Details	Data Type
Invoice Date	it holds the date of product customer bought	Numeric
Outlet Name	Name of the outlate	String
Product Code	It holds products unique id	Integer
Product Name	It holds the name of the product	String
UOM	it holds the quantity of the products	String
Sales Qty	it holds the quantity of the sales	Numeric
Total	price of those particular products	Floating-point number.

Table 1: Swapno Supershop Dataset Details

InvoiceDate	OutletName	ProductCode	ProductName	UOM	SalesQty	Total
3/1/2021	F118 Mirpur 2 Outlet	3100174	Beef Premium Cube kg	KG	3.095	1857
3/1/2021	F118 Mirpur 2 Outlet	3100174	Beef Premium Cube kg	KG	'0.73	438
3/1/2021	F118 Mirpur 2 Outlet	2400080	ACI Pure Mustard Oil 200ml	EA	1	60
3/1/2021	F118 Mirpur 2 Outlet	2400217	Chinigura Rice Loose (P) kg	KG	1	90
3/1/2021	F118 Mirpur 2 Outlet	2501242	Aarong Dairy Matha 200ml	EA	1	25
3/1/2021	F118 Mirpur 2 Outlet	2901484	Tomato (Local)kg	KG	1.27	19.05
3/1/2021	F118 Mirpur 2 Outlet	2902799	Coriander Leaf(Dhoniapata Local)N	KG	'0.495	27.23
3/1/2021	F118 Mirpur 2 Outlet	2400372	Pusti Soyabean Oil 5ltr	EA	1	655
3/1/2021	F118 Mirpur 2 Outlet	2813634	Shwapno Orange Jelly 500g	EA	1	130
3/1/2021	F118 Mirpur 2 Outlet	2400029	ACI Pure Salt 1kg	EA	2	70
3/1/2021	F118 Mirpur 2 Outlet	2400602	Deshi Moshur Dal Loose (P) Kg	KG	2.01	194.97
3/1/2021	F118 Mirpur 2 Outlet	2401040	Sugar Loose Refined (Kg)	KG	2.005	140.35
3/1/2021	F118 Mirpur 2 Outlet	2603356	Wheel Clean&Fresh 2in1 D.Powder 500g	EA	1	45
3/1/2021	F118 Mirpur 2 Outlet	2603642	Vim DW.Liquid 500ml(Buy2 Get Gamla Free)	EA	2	440
3/1/2021	F118 Mirpur 2 Outlet	2603659	Vim Dishwash Liquid Refill Pack 250ml	EA	1	45
3/1/2021	F118 Mirpur 2 Outlet	2814105	Danish Dry Cake By350g Get Bis.210g Free	EA	1	120
3/1/2021	F118 Mirpur 2 Outlet	2815431	Ifad Kaju Del.250g(B2 G1 Orng.190g Free)	EA	1	150
3/1/2021	F118 Mirpur 2 Outlet	3002785	Savlon Aloevera AntiSpt HndW Refill200ml	EA	1	60
3/1/2021	F118 Mirpur 2 Outlet	2815235	Ifad Eggy Instant Noodles 480g(B2G1 Fre)	EA	1	260
3/1/2021	F118 Mirpur 2 Outlet	2400084	Fresh Refined Sugar 1 kg	EA	3	234
3/1/2021	F118 Mirpur 2 Outlet	2400643	Fresh Super Premium Salt 1kg	EA	2	64
3/1/2021	F118 Mirpur 2 Outlet	2401148	Bashundhara Atta 2 Kg	EA	1	72

Figure-03: Dataset of Swapno Supershop

Family Supershop:

Attribute	Details	Data Type
Customer Id	It holds customer unique id	Numeric
Date	it holds the date of product customer bought	Numeric
Product Id	It holds products unique id	Numeric (integer)
Category	it holds the category of the product	String
Product Name	It holds the name of the product	String
Sale	price of those particular products	Numeric (float)

Table 2: Family Supershop Dataset Details

Customer ID	Date	Product ID	Category	Product Name	Sale
152156	11/8/2016	10001798	Furniture	Bush Somerset Collection Bookcase	261.96
152156	11/8/2016	10000454	Furniture	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	731.94
138688	6/12/2016	10000240	Office Supplies	Self-Adhesive Address Labels for Typewriters by Universal	14.62
108966	10/11/2015	10000577	Furniture	Bretford CR4500 Series Slim Rectangular Table	957.58
108966	10/11/2015	10000760	Office Supplies	Eldon Fold 'N Roll Cart System	22.37
115812	6/9/2014	10001487	Furniture	Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood	48.86
115812	6/9/2014	10002833	Office Supplies	Newell 322	7.28
115812	6/9/2014	10002275	Technology	Mitel 5320 IP Phone VoIP phone	907.15
115812	6/9/2014	10003910	Office Supplies	DXL Angle-View Binders with Locking Rings by Samsill	18.5
115812	6/9/2014	10002892	Office Supplies	Belkin F5C206VTEL 6 Outlet Surge	114.9
115812	6/9/2014	10001539	Furniture	Chromcraft Rectangular Conference Tables	1706.18
115812	6/9/2014	10002033	Technology	Konftel 250 Conferenceÿphoneÿ- Charcoal black	911.42
114412	4/15/2017	10002365	Office Supplies	Xerox 1967	15.55
161389	12/5/2016	10003656	Office Supplies	Fellowes PB200 Plastic Comb Binding Machine	407.98
118983	11/22/2015	10002311	Office Supplies	Holmes Replacement Filter for HEPA Air Cleaner, Very Large Room, HEPA Filter	68.81
118983	11/22/2015	10000756	Office Supplies	Storex DuraTech Recycled Plastic Frosted Binders	2.54
105893	11/11/2014	10004186	Office Supplies	Stur-D-Stor Shelving, Vertical 5-Shelf: 72"H x 36"W x 18 1/2"D	665.88
167164	5/13/2014	10000107	Office Supplies	Fellowes Super Stor/Drawer	55.5
143336	8/27/2014	10003056	Office Supplies	Newell 341	8.56
143336	8/27/2014	10001949	Technology	Cisco SPA 501G IP Phone	213.48
143336	8/27/2014	10002215	Office Supplies	Wilson Jones Hanging View Binder, White, 1"	22.72
137330	12/9/2016	10000246	Office Supplies	Newell 318	19.46
137330	12/9/2016	10001492	Office Supplies	Acco Six-Outlet Power Strip, 4' Cord Length	60.34

Figure-04: Dataset of Family Supershop

3.5 Data Preprocessing

Collection of data was easier than the preparation for implementing algorithms. There was a huge number of data in the dataset.After analyzing the dataset, we found that some of our properties have zero values. And some unwanted items have content variables so there will be two things:

- Data Cleaning
- Feature Engineering

Where data cleaning will clean up unwanted data and solve feature engineering. The dataset identifies specialties and is prepared for use in appropriate development models.

cols = ['Order ID', 'Product ID', 'Category', 'Product Name']

furniture.drop(cols, axis=1, inplace=True)

furniture = furniture.sort_values('date')

furniture.isnull().sum()

This step includes removing columns we do not need, check missing values, aggregate sales by date and so on.

furniture = furniture.set_index('Order Date')

furniture.index

This step is for indexing with Time Series Data

3.6 Statistical Analysis

Swapno Supershop:

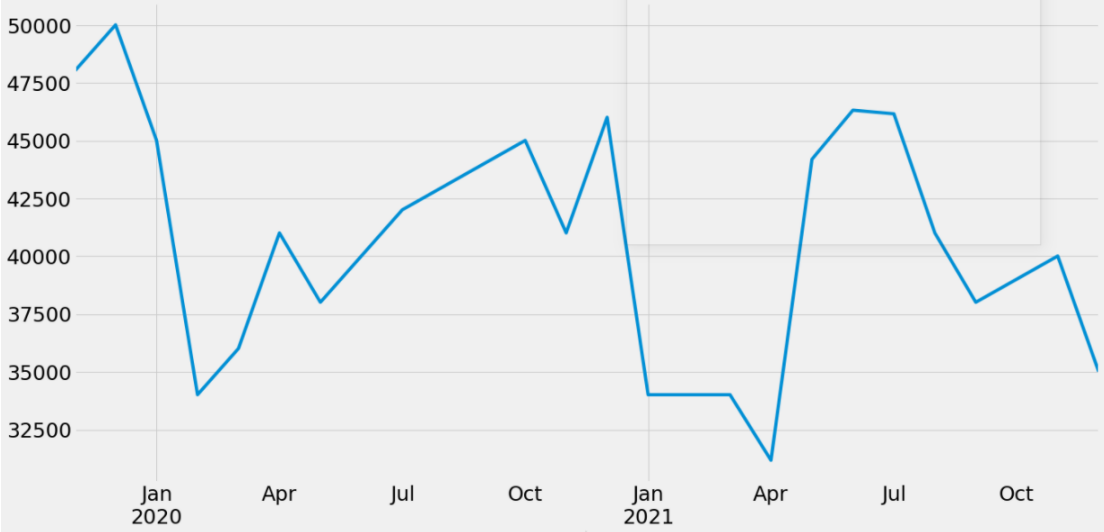


Figure-05: Visualizing Meat Sales Time Series Data1

Family Supershop:

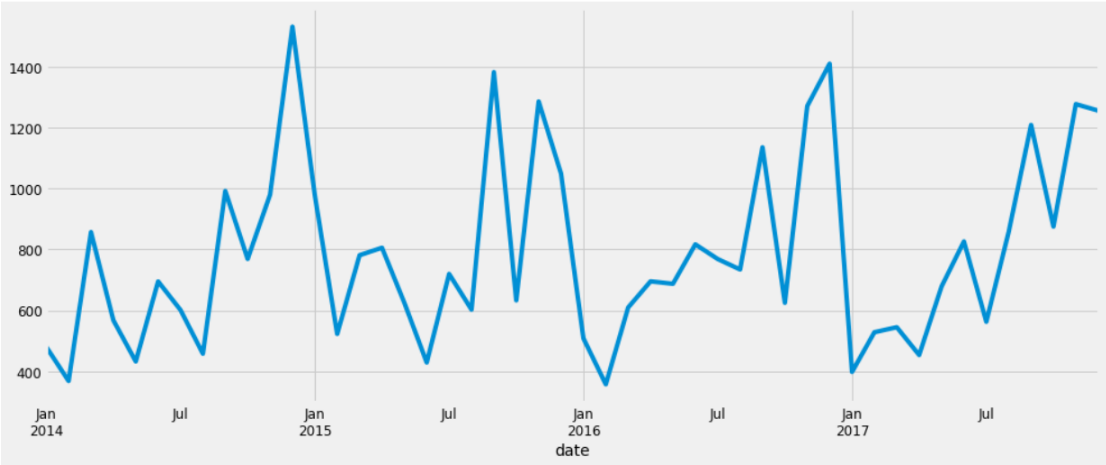


Figure-06: Visualizing Furniture Sales Time Series Data1

Some distinguishable patterns seem after we plot the info. The time-series has a seasonality pattern, like sales square measure continuously low at the start of the year and high at the tip of the year. There's a continuous Associate in Nursing upward trend at intervals any single year with some low months within the middle of the year. We

can conjointly visualize our information employing a technique referred to as time-series decomposition that enables decomposing our statistics into 3 distinct components: trend, seasonality, and noise.

Swapno Supershop:

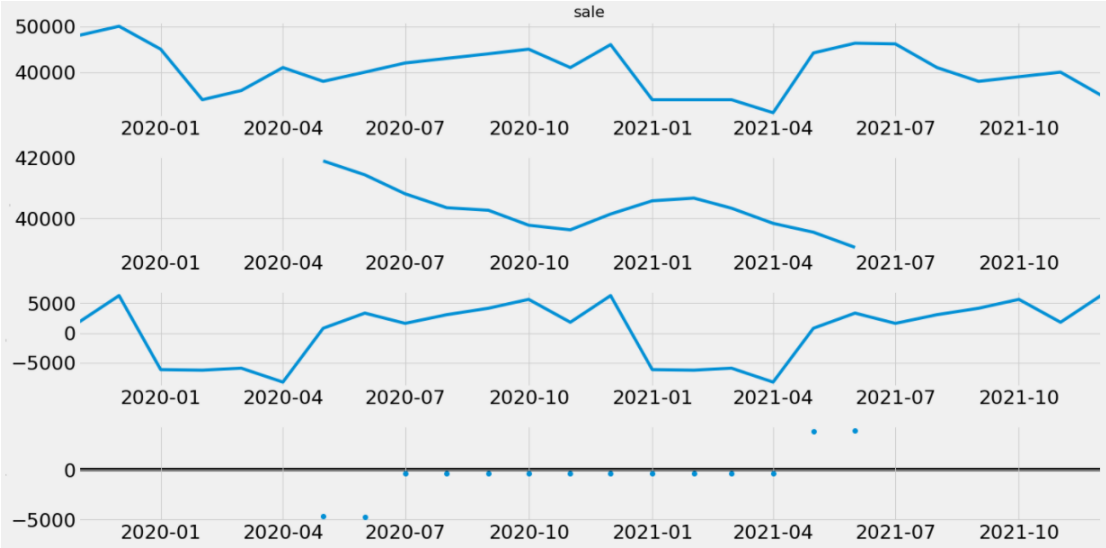


Figure-07: Visualizing Beef Sales Time Series Data2

Family Supershop:

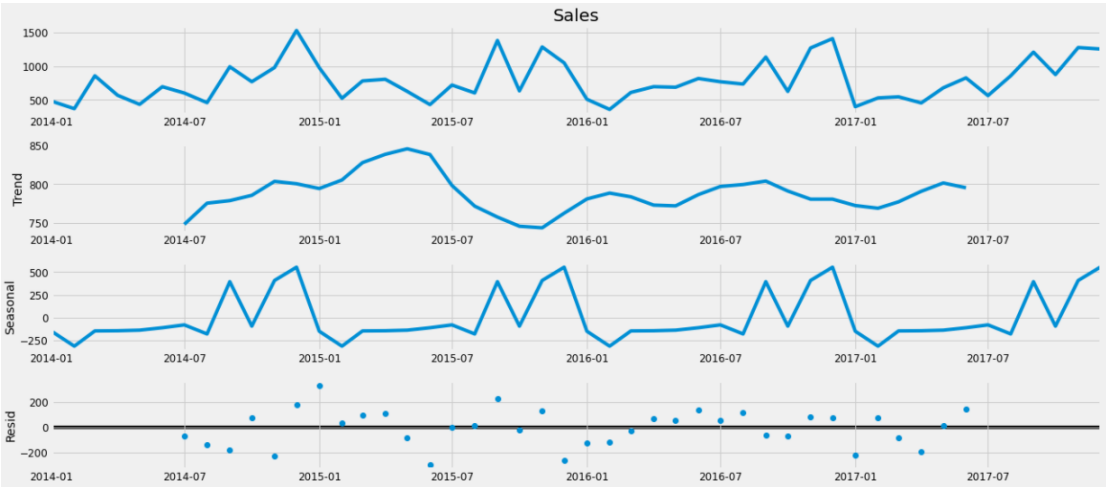


Figure-08: Visualizing Furniture Sales Time Series Data

The plot above clearly shows that the sales of furniture is unstable, along with its obvious seasonality.

3.7 Methodology

An autoregressive integrated moving average model could be a type of multivariate analysis that gauges the strength of 1 variable quantity relative to different ever-changing variables. The model's goal is to predict future securities or monetary market moves by examining the variations between values within the series rather than through actual values.

ARIMA is a best method for forecasting or predicting future outcomes based on a historical time series. It is based on the statistical concept of serial correlation, where past data points influence future data points.

An ARIMA model may be understood by outlining every of its parts as follows:

Autoregression (AR): refers to a model that shows an ever-changing variable that regresses on its own lagged, or prior, values.

Integrated (I): represents the differencing of raw observations to permit for the statistic to become stationary (i.e., knowledge values are replaced by the distinction between the info values and also the previous values).

Moving average (MA): incorporates the dependency between associate degree observation and a residual error from a moving average model applied to lagged observations.

ARIMA Parameters:

Each component in ARIMA functions as a parameter with a standard notation. For ARIMA models, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

- p: the number of lag observations in the model; also known as the lag order.
- d: the number of times that the raw observations are different; also known as the degree of differencing.
- q: the size of the moving average window; also known as the order of the moving average.

In a linear regression model, for example, the number and type of terms are included. A 0 value, which can be used as a parameter, would mean that particular component should not be used in the model. This way, the ARIMA model can be constructed to perform the function of an ARMA model, or even simple AR, I, or MA models.

Because ARIMA models are complicated and work best on very large data sets, computer algorithms and machine learning techniques are used to compute them.

Autoregressive Integrated Moving Average (ARIMA) and Stationarity

In an autoregressive integrated moving average model, the info square measures variations so as to form it stationary. A model that shows stationarity is one that shows there's constancy to the info over time. Most economic and market information show trends, therefore the purpose of differencing is to get rid of any trends or seasonal structures.

Seasonality, or once information shows regular and inevitable patterns that repeat over a year, may negatively have an effect on the regression model. If a trend seems and stationarity isn't evident, several of the computations throughout the method cannot be created with nice efficaciousness.

CHAPTER 4

EXPERIMENT RESULTS & DISCUSSION

4.1 Experimental Results and Analysis

	Dataset	Mean Squared Error	Year Of Data
Dataset 1	Swapno Dataset	42446227.08	2
Dataset 2	Family Dataset	39995.85	5

Table-03: Result Analysis

4.2 Experimental Results

Fitting the ARIMA model

We should always run model diagnostics to investigate any unusual behavior.

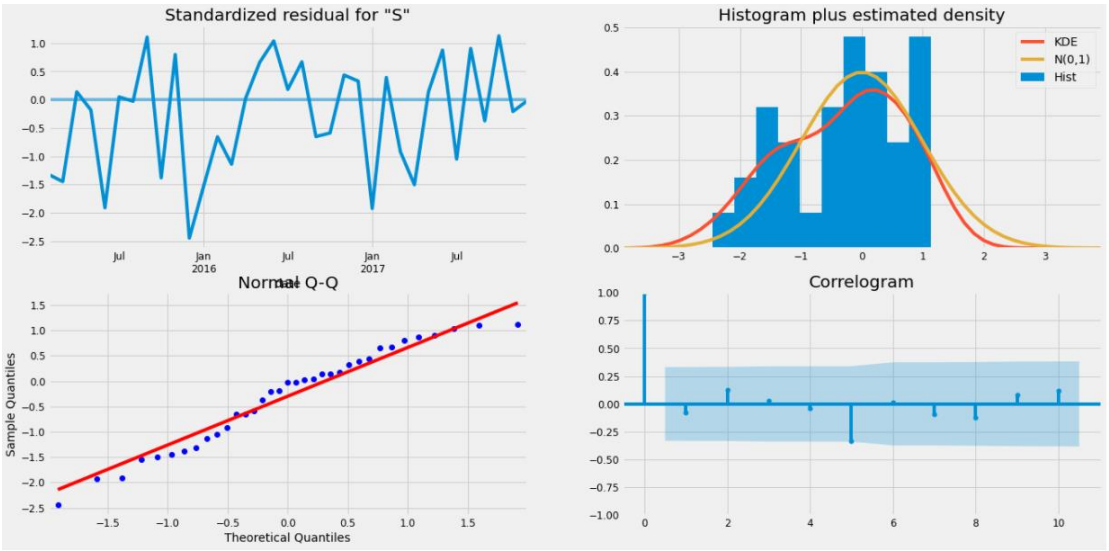


Figure-09

It is not perfect; however, our model diagnostics suggests that the model residuals are near normally distributed.

Validating forecasts

For Swapno Supershop:

Plotting Sales Forecast using ARIMA model:

The dataset we are using has large sales data from November, 2019 to December, 2021 and since we are using one product instead of a lot of products, the data is just sufficient to generate profit forecasts using the ARIMA model because it needs minimal 2-year data for forecasting. We tried to forecast sales from March, 2021 to December, 2021(9 month in total) to see the difference between observer and forecast.

Input: Beef Premium Cube kg

Output:

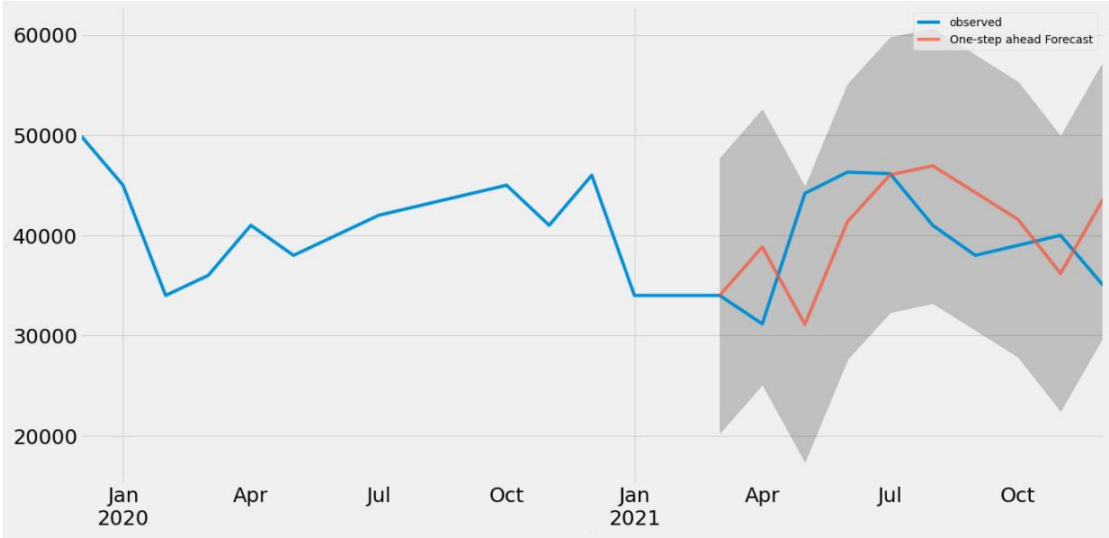


Figure-10: Time series forecasting with ARIMA

For Family Supershop:

Plotting Sales Forecast using ARIMA model:

The dataset we are using has large sales data from data from 2014 to 2017 and since we are using one category instead of individual products, the data is sufficient to

generate profit forecasts using the ARIMA model. We have 4 years data. We tried to forecast sales from January, 2017 to December, 2017(12 month in total) to see the difference between observer and forecast.

Input: Furniture

Output:

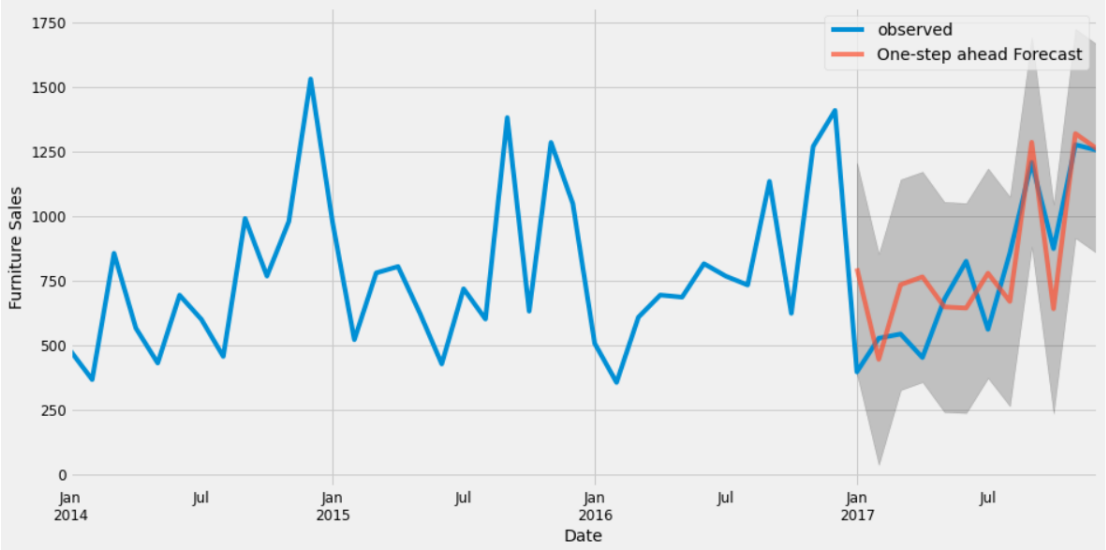


Figure-11: Time series forecasting with ARIMA

Plotting Future Sales Forecast using ARIMA model:

ARIMA model is used to analyze and forecast time-series data. The dataset we are using has large sales data from 2014 to 2017 and since we are using one category instead of individual products, the data is sufficient to generate profit forecasts using the ARIMA model. We tried to forecast profit from 2018 to 2026 (9 years in total).

Input: Furniture

Output:

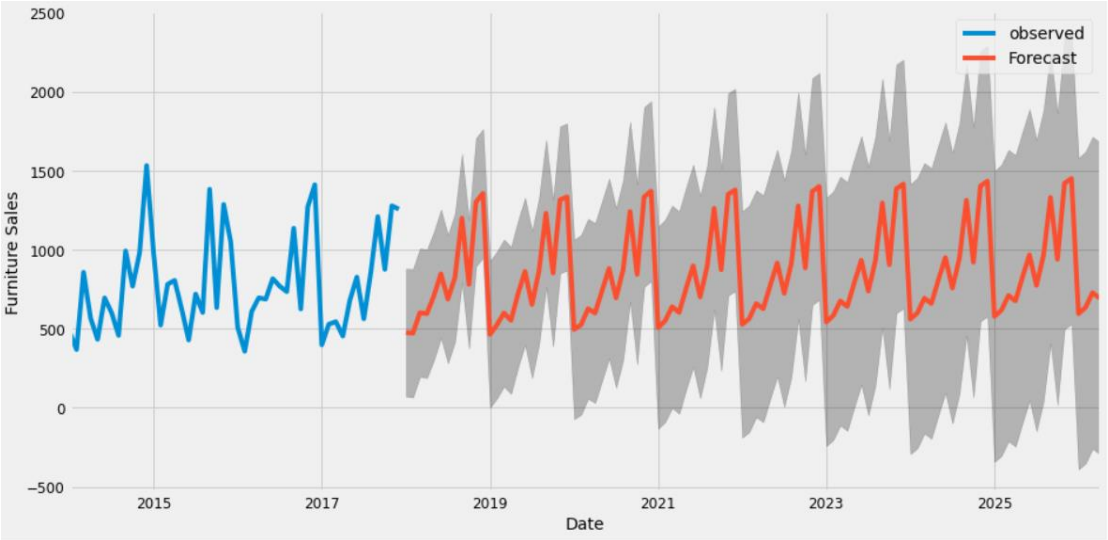


Figure-12: Producing and visualizing forecasts

Our model clearly captured the article of furniture sales seasonality. As we have a tendency to forecast any out into the longer term, it's natural for us to become less confident in our values. This can be mirrored by the arrogance intervals generated by our model, that grow larger as we have a tendency to move any out into the longer term. The on top of statistical analysis for articles of furniture makes ME interested in different classes, and the way they compare with one another over time. Therefore, we have a tendency to square measure aiming to compare statistics of articles of furniture and workplace providers.

Time Series of Furniture vs. Office Supplies

We are going to compare two categories' sales in the same time period. This means combine two data frames into one and plot these two categories' time series into one plot.

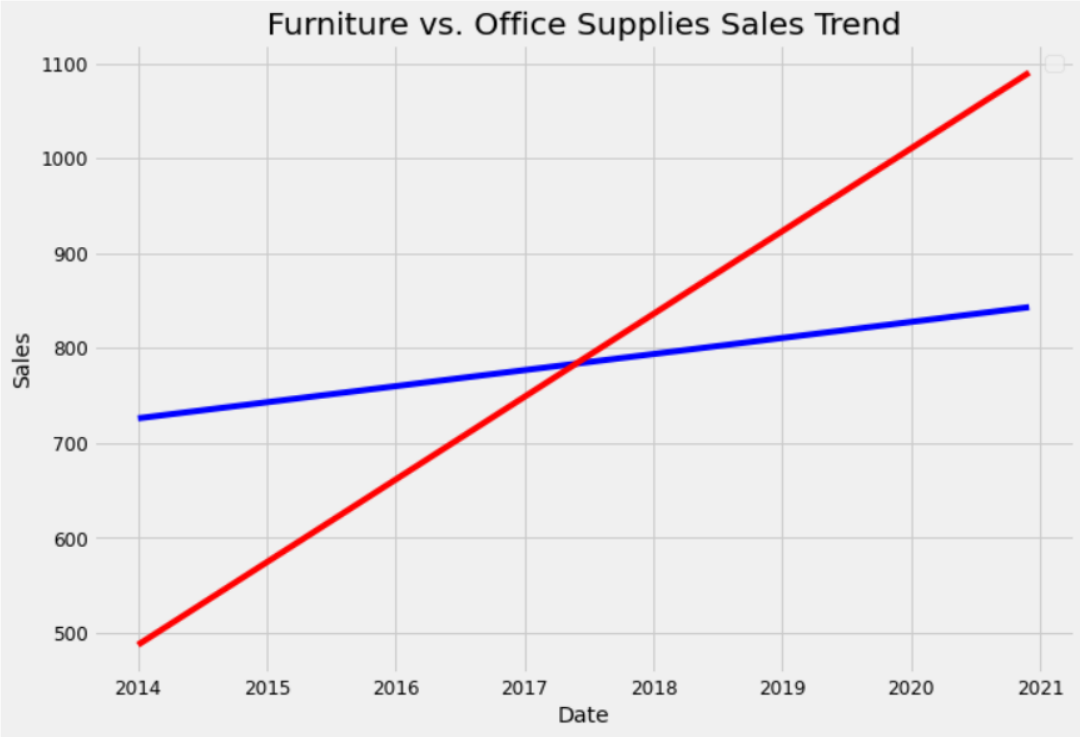


Figure-13: Trend of Furniture vs Office Supplies Visualization

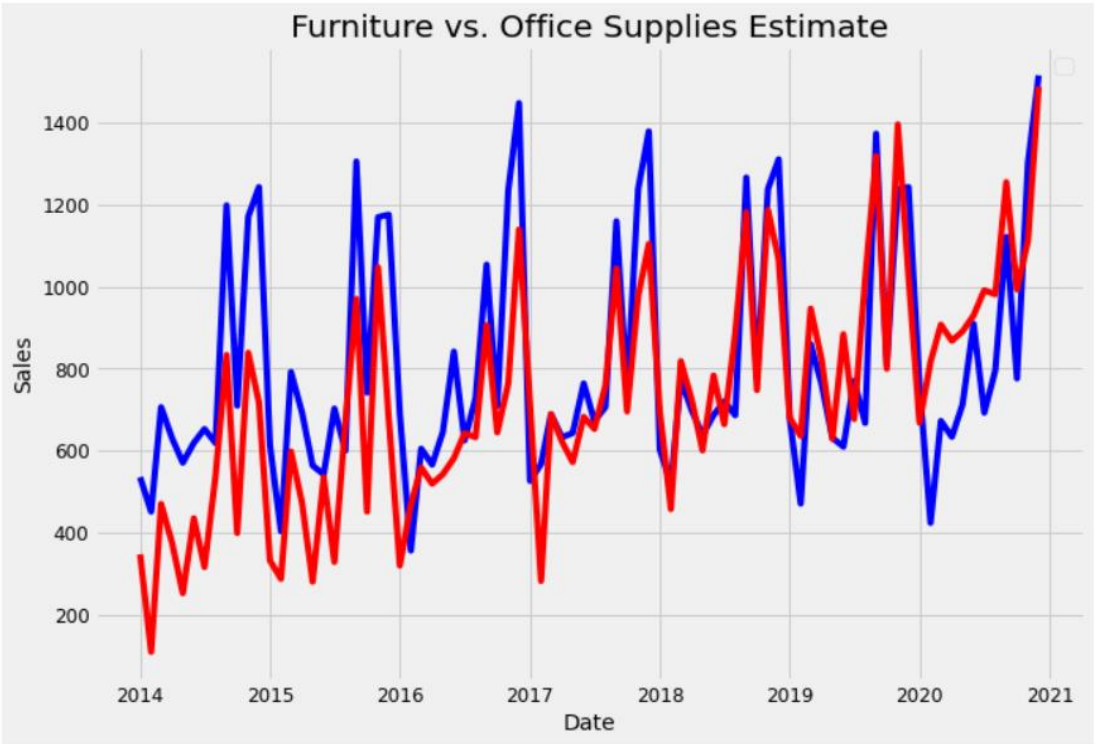


Figure-14: Furniture vs Office Supplies Estimate Visualization

We observe that sales of furniture and office supplies shared a similar seasonal pattern. Early in the year is the off season for each of the 2 classes. It appears summer time is quiet for office supplies too. In addition, average daily sales for furniture area

units over those of office supplies in most of the months. It's graspable, because the price of furniture ought to be a lot higher than what the office supplies. Often, office supplies passed furniture on average daily sales. Let’s determine once was the primary time workplace supplies’ sales surpassed those of furniture’s.

Trends and Patterns

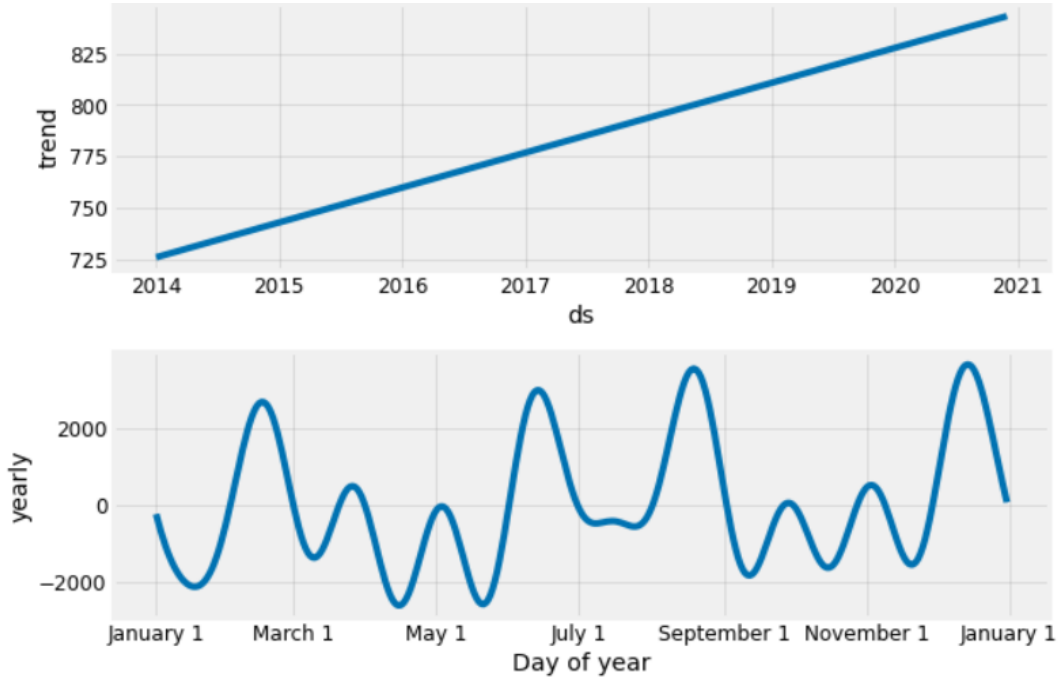


Figure-15: Trends and Patterns for Furniture

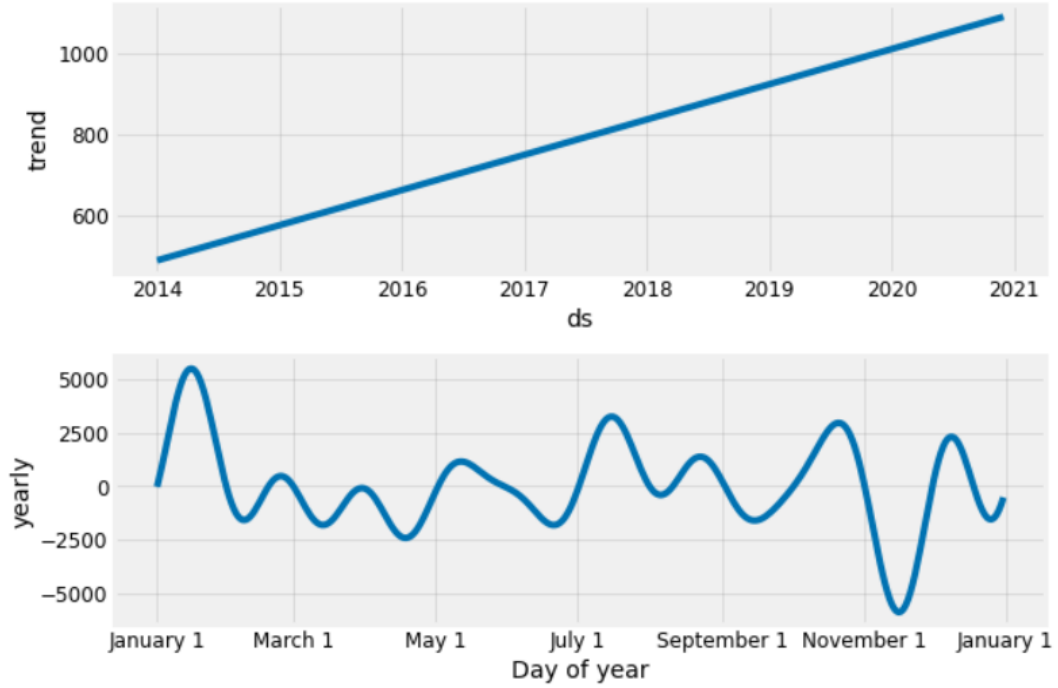


Figure-16: Trends and Patterns for Office Supplies

Good to examine that the sales for each furnishing and office supplies have been linearly increasing over time and can continue to grow, though office supplies' growth looks slightly stronger. The worst month for furnishings is April, the worst month for office supplies is Feb. The most effective month for furnishings is Dec, and also the best month for Office supplies is Oct. There are several time-series analyses we will explore from now on, like forecast with uncertainty bounds, amendment purpose and anomaly detection, forecast time-series with external data supply. We've just about started.

Time Series Modeling with Prophet

It was released by Facebook in 2017, forecasting tool Prophet is designed for analyzing time-series that display patterns on different time scales such as yearly, weekly and daily. It also has advanced capabilities for modeling the effects of holidays on a time-series and implementing custom changepoints. Therefore, we are using Prophet to get a model up and running.

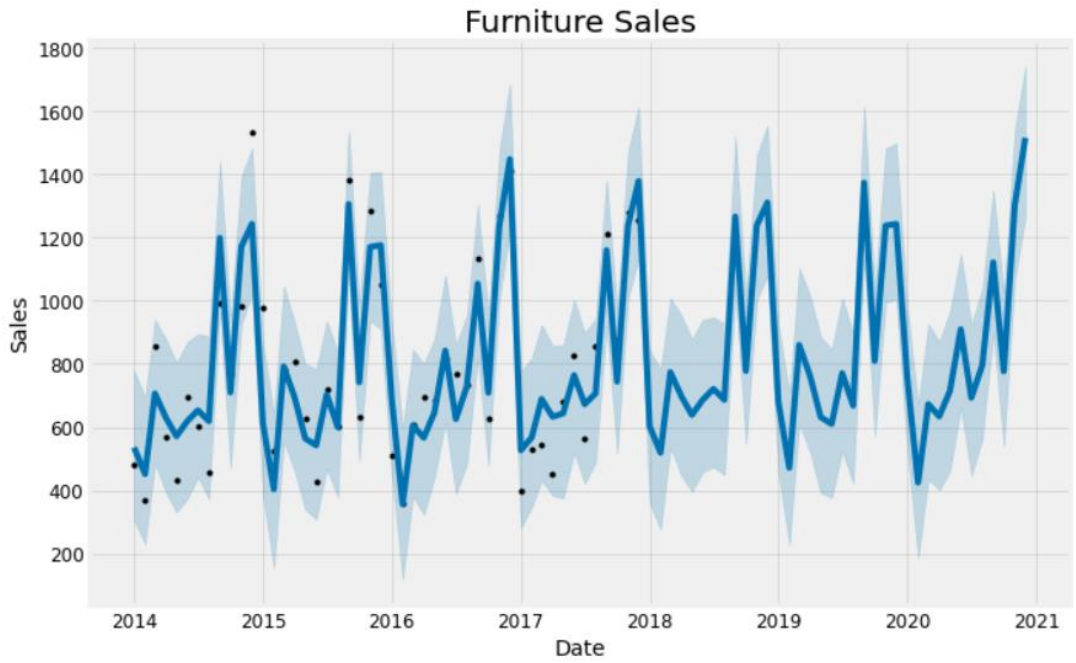


Figure-17: Time Series Modeling with Prophet (Furniture)

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

We think this research will highly impact retail shops business. The retailers will benefit from this research in a good positive way. The decision makers can take their decision from this research.

5.2 Impact on Environment

We show the way to maximize the sales, so we hope after taking the steps that we mentioned the stores will get their most impactful result thus the sales get higher in numbers. Customers will get on their product that day like most so they will be happy. This will help the retail industry of Bangladesh like the shop ‘Shopno’, ‘Agora’ ‘Meena Bazaar’ with their sales maximization.

5.3 Ethical Aspects

As we do this research to help the fintech industry specially the retail industry. We have no harmful intentions and do not apply and disservice resources.
We take the dataset from an private resource but we can safely use this so there is no chance of a private resource privacy leakage

5.4 Sustainability Plan

The dataset we have used here is from two Bangladeshi super shops data with a big number of samples and features from different locations stores. It gives us a convenient input. So, we can definitely say that the outcome is sustainable and it will be healthy for all of the retails.

CHAPTER 6

CONCLUSION & FUTURE WORK

6.1 Summary of the Study

The goal of our work is to seek out the factors that increase the sales additionally such as building a machine learning model that may predict the success rate of a brand new launching product of a retail search. Firstly, we have a tendency to preprocess our information set for gaining recent data to feed our model. Then attempt to perceive the insights that lie into the info. We have a tendency to do feature engineering for gaining additional melodious information which will impact our model in a positive manner. We've tried varied machine learning algorithms to seek out the foremost correct model.

6.2 Conclusions

The goal of this work is to use machine learning techniques to estimate future sales based on previous data. We looked at how different machine learning models are formed utilizing algorithms like Time analysis ARIMA, AR, Linear Regression, Decision Tree, Random Forest algorithm. These algorithms were used to forecast the final sales outcome. The mean squared error is used as evaluation metrics for comparing algorithms. Based on two dataset we can say in the ARIMA model if we have a lot of year data we can predict more accurately. We conclude that the ARIMA Model with enriched Dataset, are the best models with mean square value of 39995.85. So, we are hopeful that by this predicted result, sales maximization can be effective for retail business.

6.3 Implication for Further Study

There are several possibilities for further study of this research. In our work we focus on only sales maximization. But it is a common business fact that when the sales increase the profit increases proportionally. So, working with profit will be a big scope for further research. Another scope of this research will be profit forecasting. We have a plan to include them. But the inconvenient world situation did not favor us. Thus, we will do the rest in a near feature In sha Allah. Fintech researchers are also welcome to extend this study

REFERENCE

[1] Huang, Q., & Zhou, F. (2017). *Research on retailer data clustering algorithm based on Spark*. https://doi.org/10.1063/1.4977378

[2] Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(4), 580–585. https://doi.org/10.1109/tsmc.1985.6313426

[3] Maingi, M. N. (2015). A Survey on the Clustering Algorithms in Sales Data Mining. *International Journal of Computer Applications Technology and Research*, *4*(2), 135–137. https://doi.org/10.7753/ijcatr0402.1009

[4] Sastry, S. H., & Babu, M. S. P. (2013). Analysis & Prediction of Sales Data in SAP-ERP System using Clustering Algorithms. *Analysis & Prediction of Sales Data in SAP-ERP System Using Clustering Algorithms*. https://doi.org/10.5121/ijcsity.2013.1407

[5] Cohn, C. (2015, May 15). *A Beginner's Guide To Upselling And Cross-Selling*. Forbes. https://www.forbes.com/sites/chuckcohn/2015/05/15/a-beginners-guide-to-upselling-and-cross-selling/?sh=454a189e2912

[6] Korolev, M, & Reug, K. (2014). Gradient Boosted Trees to Predict Store Sales. https://cs229.stanford.edu/proj2015/193_report.pdf

[7] Värlander, S., & Yakhlef, A. (2008). Cross-selling: The power of embodied interactions. *Journal of Retailing and Consumer Services*, *15*(6), 480–490. https://doi.org/10.1016/j.jretconser.2008.01.003

[8] Johnson, J. S., & Friend, S. B. (2014). Contingent cross-selling and up-selling relationships with performance and job satisfaction: an MOA-theoretic examination. *Journal of Personal Selling & Sales Management*, *35*(1), 51–71. https://doi.org/10.1080/08853134.2014.940962

[9] Kamakura, W. A. (2008). Cross-Selling. *Journal of Relationship Marketing*, *6*(3–4), 41–58. https://doi.org/10.1300/j366v06n03_03

[10] Schmitz, C. (2012). Group influences of selling teams on industrial salespeople's cross-selling behavior. *Journal of the Academy of Marketing Science*, *41*(1), 55–72. https://doi.org/10.1007/s11747-012-0304-7

[11] Yu, T., Patterson, P., & de Ruyter, K. (2015). Converting service encounters into cross-selling opportunities. *European Journal of Marketing*, *49*(3/4), 491–511. https://doi.org/10.1108/ejm-10-2013-0549

[12] Wong, W., Leung, S., Guo, Z., Zeng, X., & Mok, P. (2012). Intelligent product cross-selling system with radio frequency identification technology for retailing. International Journal of Production Economics, 135(1), 308–319. https://doi.org/10.1016/j.ijpe.2011.08.005

[13] Bernazzani, S. (2021, December 23). *Cross-Selling and Upselling: The Ultimate Guide*. Hubspot. https://blog.hubspot.com/sales/cross-selling

[14] Dobreva, K. (2021, June 16). *A guide to cross-selling & upselling techniques*. Magento Blog: Technical Tips and e-Commerce Guides from Amasty. https://amasty.com/blog/a-guide-to-cross-selling-upselling-techniques/

[15] Shrivastava, V., & Arya, N. (2012). A study of various clustering algorithms on retail sales data. *Int. J. Comput. Commun. Netw*, *1*(2).

[16] C. (2020, January 7). *What Amazon Can Teach You About Cross-Selling*. Predictable Profits. https://predictableprofits.com/amazon-can-teach-cross-selling/

[17] Harrison, O. (2019, July 14). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Medium. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[18] Srivastava, T. (2020, October 18). *K Nearest Neighbor | KNN Algorithm | KNN in Python & R*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[19] Maillo, J., Luengo, J., García, S., Herrera, F., & Triguero, I. (2017, July). Exact fuzzy k-nearest neighbor classification for big datasets. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-6). IEEE.

[20] Sayli, A., Ozturk, I., & Ustunel, M. (2016). BRAND LOYALTY ANALYSIS SYSTEM USING K-MEANS ALGORITHM. *BRAND LOYALTY ANALYSIS SYSTEM USING K-MEANS ALGORITHM*. https://dergipark.org.tr/tr/download/article-file/333779

[21] A Machine Learning Based Method for Customer Behavior Prediction. (2019). *Tehnicki Vjesnik - Technical Gazette*, *26*(6). https://doi.org/10.17559/tv-20190603165825

[22] Berger, P. D., Bolton, R. N., Bowman, D., Briggs, E., Kumar, V., Parasuraman, A., & Terry, C. (2002). Marketing Actions and the Value of Customer Assets: A Framework for Customer Asset Management. Journal of Service Research, 5(1), 39–54. https://doi.org/10.1177/1094670502005001005

[23] Bhattacharya, C. B. (1998). When Customers Are Members: Customer Retention in Paid Membership Contexts. *Journal of the Academy of Marketing Science*, *26*(1), 31–44. https://doi.org/10.1177/0092070398261004

[24] Boles, J. S., Barksdale, H. C., & Johnson, J. T. (1997). Business relationships: an examination of the effects of buyer-salesperson relationships on customer retention and willingness to refer and recommend. *Journal of Business & Industrial Marketing*, *12*(3/4), 253–264. https://doi.org/10.1108/08858629710188072

[25] Burez, J., & van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, *32*(2), 277–288. https://doi.org/10.1016/j.eswa.2005.11.037

[26] Liu Weixiao. (2016). Hybrid intelligent model for fashion sales forecasting based on discrete grey forecasting model and artificial neural network. Journal of Computer Applications, 36(12), 3378-3384.

[27] Zheng Jun, Jin Yi, Yan JiDuo, et al. (2013). Application of Data Mining Technology in Logistics Management. Journal of Guiyang College Natural Sciences, 8(2), 32-34.

[28] Zhang, D., Sui, J., & Gong, Y. (2017). Large scale software test data generation based on collective constraint and weighted combination method. Tehnicki vjesnik, 24(4), 1041-1050. https://doi.org/10.17559/TV-20170319045945

[29] Zhang, D. (2017). High-speed Train Control System Big Data Analysis Based on Fuzzy RDF Model and Uncertain Reasoning. International Journal of Computers, Communications & Control, 12(4), 577-591. https://doi.org/10.15837/ijccc.2017.4.2914