SALES PREDICTION: AN INTEGRATED APPROACH USING MACHINE LEARNING

Aadika Bhatia(00213203120/IT-1/2020-24)

Japneet Singh Baluja(04713203120/IT-1/2020-24)

Lakshya Mishra(05413203120/IT-1/2020-24)

Manraj Singh(05913203120/IT-1/2020-24)

Under the guidance of:

Dr. Amandeep Kaur

Department of Information Technology & Engineering

Guru Tegh Bahadur Institute of Technology

Guru Gobind Singh Indraprastha University

Dwarka, New Delhi

Year 2020-2024

DECLARATION

We hereby declare that all the work presented in the dissertation entitled "Sales Prediction: An Integrated Approach Using Machine learning" for the partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in **Information Technology & Engineering,** Guru Tegh Bahadur Institute of Technology, Guru Govind Singh Indraprastha University, New Delhi is an authentic record of our own work carried out under guidance of **Dr. Amandeep Kaur.**

Date:

Aadika Bhatia(00213203120/IT-1/2020-24)

Japneet Singh Baluja(04713203120/IT-1/2020-24)

Lakshya Mishra(05413203120/IT-1/2020-24)

Manraj Singh(05913203120/IT-1/2020-24)

CERTIFICATE

This is to certify that dissertation entitled "Sales Prediction: An Integrated Approach Using Machine learning" which is submitted by **Aadika Bhatia(00213203120/IT-1/2020-24), Japneet Singh Baluja(04713203120/IT-1/2020-24), Lakshya Mishra(05413203120/IT-1/2020-24), and Manraj Singh(05913203120/IT-1/2020-24)** in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in **Information Technology**, Guru Tegh Bahadur Institute of Technology, New Delhi is an authentic record of the candidate's own work carried out by them under our guidance. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Dr. Amandeep Kaur   Dr. Savneet Kaur

(Project Guide)   Head of IT Department

ACKNOWLEDGEMENT

We would like to express our great gratitude towards our supervisor, **Dr. Amandeep Kaur** who has given us support and suggestions. Without their help we could not have presented this dissertation up to the present standard. We also take this opportunity to give thanks to all others who gave us support for the project or in other aspects of our study at Guru Tegh Bahadur Institute of Technology.

Date: 18/01/2023

Aadika Bhatia

(00213203120/IT-1/2024)

aadikabhatia30@gmail.com

Japneet Singh Baluja

(04713203120/IT-1/2024)

japneet8208@gmail.com

Lakshya Mishra

(05413203120/IT-1/2024)

lakshya764@gmail.com

Manraj Singh

(05913203120/IT-1/2024)

manrajsinghw08@gmail.com

ABSTRACT

Aim: This study provides an algorithm which deals in sales forecasting to enable sales organizations to make business decisions and predict sales forecast for short-term and long-term performance. The aim of the paper is to plan a model to foresee the offers of products through Machine Learning. In this we have old sales data of products of any store and from that sales data, we will predict the future sale of items.

Problem: Decision producers in people in general and private parts would profit by progressively precise expectation of interest for merchandise and enterprises. The target of this research work is to improve forecasting practices.

Method: This paper determining exploration to distinguish which strategies are helpful for interest anticipating, and which are not, and creates agendas of proof based estimating direction for demand forecasters and their customers.

Findings: This paper recognizes and portrays issues in foreseeing the precise outcomes and provides the hybrid approach to deal with actualizing the most appropriate strategies.

Usefulness: The goal of this paper aims the performance of the organization—more profit, more revenue, financial improvements. The evidence introduced in this paper recommends that by utilizing the hybrid approach, expectation will deliver request figures that are considerably more precise than those given by right now well-known strategies.
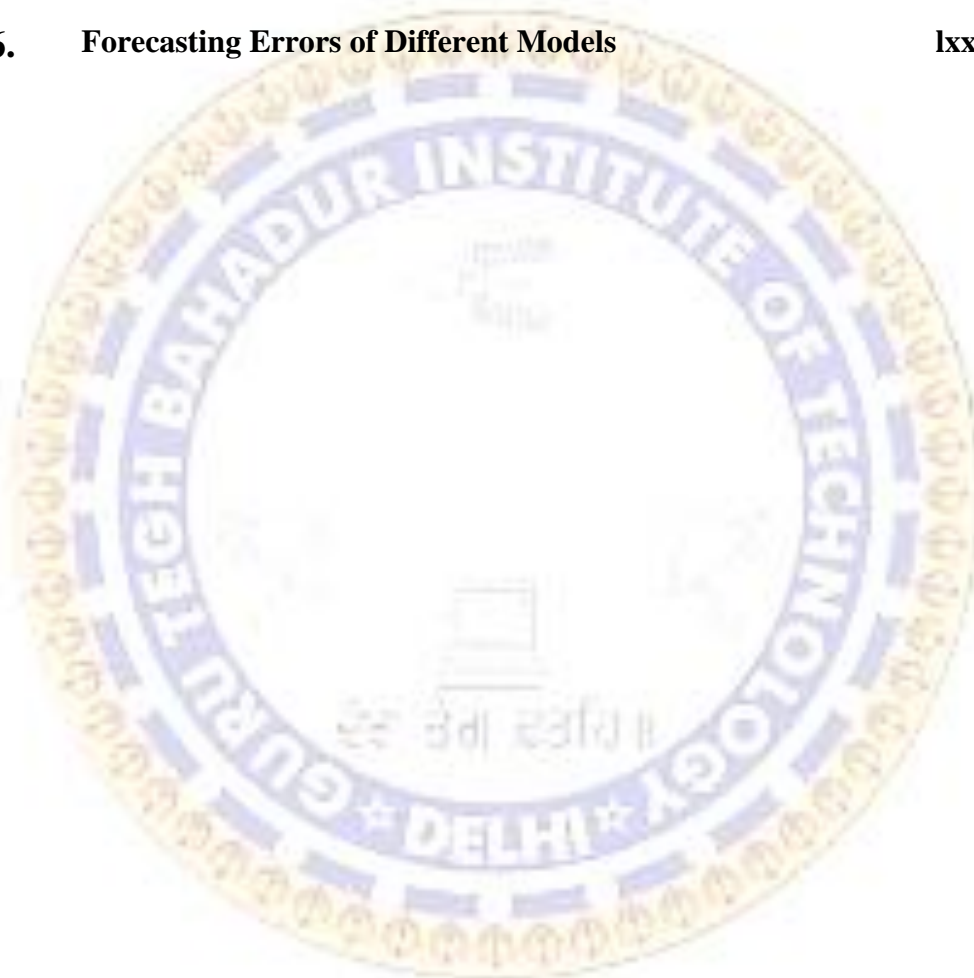
LIST OF FIGURES

LIST OF TABLES

CONTENTS

Introduction

1.1 INTRODUCTION

In this research work relapse investigation is utilized with feed forward neural system as a forecasting apparatus. Dataset containing prepared and test information is accessible on kaggle.com at https://www.kaggle.com/c/focused information science-foresee future-sales/information/. We are furnished with every day chronicled sales information. The errand is to estimate the aggregate sum of items sold each shop for the test set.

Forecasting is one of the significant parts of organization. The comer-stone of effective showcasing arranging is the estimation and forecasting to market request. As indicated by American Marketing Association, "Sales Forecasting is to find the of Sales, in money related or physical units, for a predefined future period under a proposed field-tested strategy or program and under an accepted arrangement of monetary and different powers outside the unit that the estimate is formed ."

A Sales forecasting is to measure the volume that a company or an organization can hope to achieve at a particular time frame. A business estimate isn't only a deals foreseeing. It is the demonstration that how much to be produced to fulfil the demand. It is the assurance of a company's offer in the market under a predefined future. In this manner deals anticipating demonstrates the plausible volume of offers.

Time series forecast for financial procedures is a point of expanding interest. One of most huge factors on which our organizations and authentic depends is Time. In any case, advancement has helped us manage the time with consistent improvements happening in all pieces of our lives. This examination is identified with the systems of anticipating and gauging future techniques. So as to decrease stock-keeping costs, an appropriate figure of the interest later on is essential. The technique we for the most part use, which handles time sensitive information that is only –TimeSeries Data and the models we develop for that is TimeSeries Modeling.

An ordered course of action of view of a variable or got object at likewise scattered time between time. Time arrangement is anything that's watched successively overtime at normal between time like hourly, well ordered, month to month, quarterly, and so on. Time arrangement information is tremendous when you are imagining something which is changing over the time utilizing past information. In timegame-plan examination the objective is to guess the future regard using the practices in the past information. There are many statistical techniques available for timeseries forecast however we have found few effectives ones which are recorded underneath:

Techniques Used:

- The Simple Moving Average

- Autoregressive Integration Moving Average

- Croston

- The Exponential Smoothing

- Neural Network

In this study, Autoregressive and Exponential Smoothing (SES) is utilized by utilizing the ideas of Machine Learning for a transient figure of the clearance of articles in stores. These methodologies are utilized beforehand in past research works. Presently the time has come to execute cross breed approach.

AI is an utilization of man-made brainpower (Artificial Intelligence) that enables systems to thus take in and improve for a reality without being unequivocally tweaked. AI bases on the progression of PC programs that can get to data and use it learn for themselves.

AI counts are normally delegated managed or unsupervised.

Directed AI calculations can apply what has been acknowledged in the past to new data using named advisers for envision future events. Starting from the examination of a known preparing dataset, the learning estimation makes an interpreted ability to make assumptions regarding the yield esteems inductions from datasets to depict hid structures from unlabeled data.

Unsupervised algorithms are utilized when the data used to get the information is neither arranged nor named. Unsupervised learning contemplates that the frameworks can infer a possibility to delineate a covered structure from unlabeled data. The framework doesn't understand the right results, yet it investigates the information and can draw.

In our investigation as research paper is concerned, we have to predict the ideas of different thing when all is said in done store using set away deals data. For figure distinctive count and procedure is used. In this examination Artificial Neural System is used.

Artificial Neural Networks are made of various hubs, which emulate natural neurons of human cerebrum. The neurons are associated by connections and they cooperate with one another. The hubs can take input information and perform straightforward tasks on the information. The consequence of these tasks is passed to different neurons. The yield at every hub is called its initiation or hub esteem.

The primary point is to allow the PCs learn automatically without human intercession or help and modify activities appropriately. The Modern Upset as an authentic idea has numerous inadequacies. Another idea—the "productive upheaval"— is proposed to put the Modern Transformation in a more extensive chronicled setting. The productive upset was a procedure of family unit-based asset reallocation that expanded both the supply of advertised items and work and the interest for market-provided merchandise. The innovative unrest was a family level change with significant interest side highlights that went before the Mechanical Upset, a supply-side marvel. It has suggestions for nineteenth-and twentieth-century financial history. Because of the purchase out arrangement of requesting, franchisee store chiefs perform request anticipating and stock control dependent on their business execution and experience. The point of this investigation is to give a calculation through which deals expectation could be made for the stores with the goal that stock could be kept up further.

1.2 Scope of Research

Sales forecasting utilizes past figures to foresee present moment or long-haul execution. It is a questionable movement, considering the way that such countless different factors can impact future deals: financial downturns, specialist turnover, changing examples and structures, extended test, creator surveys and various parts. Regardless, there are a couple of standard methods that can convey dependably exact deals gauges from year to year.

Without deals conjectures, it is uncommonly difficult for you to control the association the right way. You would not understand that the spring is reliably the slowest season, so you'd put a great deal in stock that would just sit on the racks. You wouldn't concentrate on industry agents who envision an incredible improvement in event deals, and you'd lose potential customers to the test, which duplicated it get-away deals control and displaying endeavors. This exploration focuses:

- How significant is sales determining to the financial related arranging and the executives of a business?
- Techniques and innovations that guarantee the most precise and reliable forecasts?
- To structure a model to foresee the offers of thing through Machine Learning".
- To just limit the estimate blunder
- In this we have old sales information of thing of any store and from that business information we need to foresee the future clearance of thing

1.3 OBJECTIVES OF THE STUDY

To predict the future sales to use it as the basis of planning time and resources.

To look after stock – A precise deals anticipating helps in evaluating the measure of stock required for future objectives. It helps in keeping the stock up for pinnacle periods.

Its aides in distinguishing Clients and Timelines: - Notwithstanding what is selling, an exact deals figure can recognize who is purchasing the most and when they are purchasing. On the off chance that our deals are regular, we know which months are moderate and can be utilized for prospecting.

1.4 SOURCE OF DATASET

In this investigation relapse examination is utilized with feed forward neural system as a determining instrument. Dataset containing prepared and test information is accessible on kaggle.com. We are given every day authentic deals information. The errand is to conjecture the aggregate sum of items sold in each shop for the test set.

1.5 STATISTICAL TECHNIQUES USED IN THE STUDY

1.5.1 Auto Regression Analysis

Regression analysis is a factual strategy utilized for the deciding the connection between a solitary ward (measure) variable and at least one autonomous (indicator) factors. In the proposed investigation the needy variable is predicted sale and independent variable is month and year.

For the proposed work Python will be used as programming language in perspective on its versatility. There are two basic ways to deal with perform straight backslide in Python—with Stats-models and scikit-learn. It is moreover possible to use the Scipy library. Stats models is a Python module that gives classes and abilities to the estimation of a wide scope of quantifiable models, similarly with respect to coordinating accurate tests, and quantifiable data examination. Scikit-adjust (sometime prior scikits.learn) is a free programming artificial intelligence library for the Python programming language.

1.6 PROBLEM FORMULATION

Increasing customer satisfaction: - In solicitation to keep your customers satisfied it is need to outfit them with the thing they need when they need it. This ideal position of deciding in business will help envision products on demand to fulfil customer orders with short interval of times.

Inventory Controls-The more definite the sales forecast, the better orchestrated your association will be to manage its stock, avoiding both over-burden and stock-out conditions. Stable stock in like manner means better organization of your age.

Marketing: - Sales Forecasting gives exhibiting a pushed look at future deals and offers the opportunity to design headways if it shows up deals will be fragile. In incredible cases, deals gauges may prompt stopping moderate moving items

2.1 MOTIVATION

Sales of business let it proceeded or shut. It pays workers, spread working costs. With a particular deals gauge accessible, you'll prepare for what's to return. Sales estimating is an important piece of the cash related arranging of a business. Sales estimates are likewise a big piece of beginning another business. On the off chance that your business estimate says that in the month of December you made 30 percent of your yearly deals, at that time, you've got to extend production in September to urge ready for the hurry. These thoughts should be communicated as numbers-misfortunes, benefits, and deals conjectures that the bank can without much of a stretch catch on.

It may likewise be savvy to put resources into increasingly occasional salesmen and begin a focused-on advertising effort directly in the wake of Thanks giving. One straightforward deals estimate can advise each other perspective regarding your business. It's a self-evaluation apparatus that utilizations past and current deals measurements to wisely anticipate future execution. Then you as business owner have to search for the business in same field and that are feasible. As such, a strategy is needed. You'll need to make concessions for the trouble of beginning starting without any outside help, this implies that initial months will be lean. At that point you'll have to persuade the bank that our business can beat the challenges in future.

Generally, all startup needs money or start-up cash-flow to buy all things needed to get off the lower level: office space, hardware, stock, worker pay rates and promoting. You can't simply stroll into a manage an account with a splendid thought and bunches of energy. A focal piece of that field-tested strategy will be the business conjecture. During business development, deals conjecture keep it up being a big estimation of your organization's wellbeing. Wall Street estimates the achievement of a corporation by how well it meets its quarterly deals gauges.

If an organization predicts hearty deals in the final quarter but only wins a large part of that estimate, it's an indication for the financial advisors of the company that the organization is not performing effectively, yet the board is confused for privately owned business, while pulling in new financial specialists, deals figures can be utilized to anticipate the potential degree of profitability. Since you won't have any past deals and experiences, it is needed to look into about similar organizations working in the equivalent topographical market with a comparative study.

2.1.1 Factors affecting forecast of sales:

Internal Factors: Internal factors that can effect are labor problems, price Changes new Product Lines, change in distribution method, inventory Shortage, working Capital Shortage production Capability shortage.

External Factors: External factors that can effect are styles and fashions Consumers earning, population changes, weather, direct and indirect competition, relative state of economy.

2.2 INFORMATIONS TO GENERATE A SALES FORECAST REPORT

Since the forecast relies on your relationship's past deals, it is fundamental to realize your dollar bargains volume for whatever timeframe that a huge drawn-out timeframe. To finish a thorough arrangements measure, you in addition need to think about most of the fragments, both inside and outside, that can impact bargains.

Mathematically, it is conceivable to measure manages some accuracy. Reasonably, notwithstanding, this exactness can be dulled due to outer market and money related parts that are outside your ability to control. Coming up next are a touch of the outside segments that can affect bargains:

- Seasonal business
- Economical relative condition
- Direct and contorted test
- Consumer pay

- Population changes

- Weather

- Political occasions

- Styles or structures

- Productivity changes

Deals anticipating requires enough separated examination of both the external and inside parts related to the business work. Inward factors that can impact deals are somewhat progressively controllable, for instance, labor issues, credit approach changes, sales motivation plans, production limit need, new item contributions, inventory insufficiencies, working capital need Price changes, change in movement strategy. It is basic to observe that if you sell more than one kind of thing or organization, you should set up an alternate deals gauge for every organization or thing gathering. The more drew in your business guess is, the more careful its outcome will be.

2.3 EVALUATION OF SALES FORECAST REPORT

A sales forecast should be performed, assessed, and differentiated and veritable execution results constantly. Think of it as a standard examination that keeps the riggings of your business running easily so your association can achieve a higher display record.

Though every business person's comfort level may be extraordinary, deals gauges should be driven month to month during the principal year, and quarterly after that. Although any forecast has a level of vulnerability, the more remote into the future you anticipate, the more noteworthy your vulnerability. When in doubt, there are three-time frames for sales forecasting:

1. Short-term conjectures are for less than a quarter of a year. They are utilized to settle on ceaseless choices about arranging, booking, stock and staffing under way, procurement, and logistics activities.

2. Medium-term forecasts have a range of a quarter of a year to two years. They are utilized for budgetary arranging, cost control, promoting new items, deals power pay plans, office arranging, scope quantification and procedure choice and circulation arranging.

3. Long-term forecasts spread over two years. They are utilized to choose whether to enter new markets, grow new items or benefits, extend or make new offices, or orchestrate long haul obtainment contracts.

Maybe the least troublesome procedure is to acknowledge that the rate augmentation (or decreasing) in arrangements will continue and that no market components will impact deals execution more later than previously.

Figure 1: Level of Sales Forecast

2.4 FORECASTING TECHNIQUES

Forecasting techniques should be qualitative and quantitative. The statistical techniques available for time arrangement estimate anyway we've discovered few effective ones which are recorded beneath:

- Simple Moving Average
- Auto regressive Integration Moving Average
- Exponential Smoothing
- Rolling Forecast

- Neural Network

An association must spend an idea to ensure that consumption doesn't surpass arranged income. Sales forecasting is that the beginning stage for business arranging exercises. Taking the medium-term Sales forecasting figure because the expressing point, spending plans are allotted to departments. Budgets are a way of control. The sales division spending plan is that the reason for the showcasing plan in accomplishing those forecasted sales.

2.4.1 SIMPLE MOVING AVERAGE

A simple moving average (SMA) is that the least troublesome quite arrangement of assessing. On a really basic level, a fundamental moving typical is dictated by including the prop 'n'

period's characteristics and after that isolating that number by 'n'. So, the moving ordinary regard is pondering as the guess for next period.

Moving midpoints can be used to quickly recognize on account of selling is moving in an uptrend or a downtrend depending upon the precedent gotten by the moving ordinary.

For example, a moving ordinary is used to smooth out irregularities (tops and valleys) to effectively perceive patterns.

Give us an opportunity to assume, we've a period plan data, to possess a prevalent appreciation on SMA,where, we've the graphical point of view on our data, therein we've twelve impression of Price with equal interval of your time . within the wake of plotting our data, it creates the impression that it's upward example with lot of peaks and valleys.
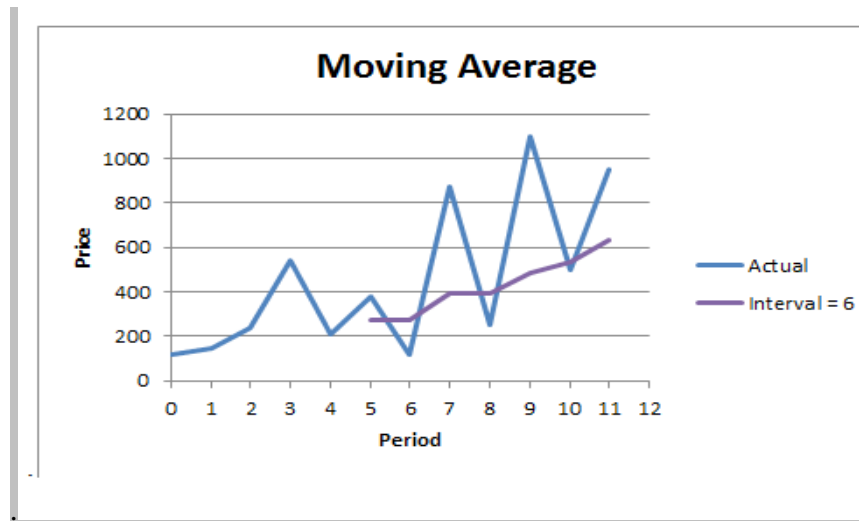
Figure 2: Moving Forecast

The tinier the between time, the closer the moving midpoints are to the veritable data centres. The SMA is on a very basic level oversee recorded data having progressively more zenith and valleys. Most likely it would be stock information, retail information and so on.

DIFFICULTIES WITH TRADITIONAL FORECASTING METHODS

Simple Moving Average

- Forecast based on mean computed for most recent observations.
- Equal weights is assigned to each past observation.

2.4.2 EXPONENTIAL SMOOTHING

This is the second outstanding system to convey a smoothed statistic. Exponential Smoothing doles out with lessening burdens to urge progressively settled observations.

•Exponential smoothing is regularly a technique for smoothing the data for removing an incredible piece of the confusion (unpredictable effect) from the data by providing a predominant sale.

2.4.2.1 Exponential Smoothing Methods

- **Simple Exponential Smoothing:**

If you have a period arrangement that can be portrayed utilizing an added substance model with steady dimension and non-uniform, you can utilize basic exponential smoothing for present moment estimate.

- **Holt's Exponential Smoothing:**

If, despite everything that you have a period plan that can be depicted using an additional substance model with growing or lessening design and no consistency, you can use Holt's exponential smoothing to make transient guesses.

- **Winter's Three Parameter Linear and Seasonal Exponential Smoothing:**

This is the second outstanding system to convey a smoothed statistic. Exponential Smoothing doles out with lessening burdens to urge progressively settled observations model with growing or reducing example and normality, Holt-Winters exponential smoothing can be used to get transient evaluations.

Graphical Views:
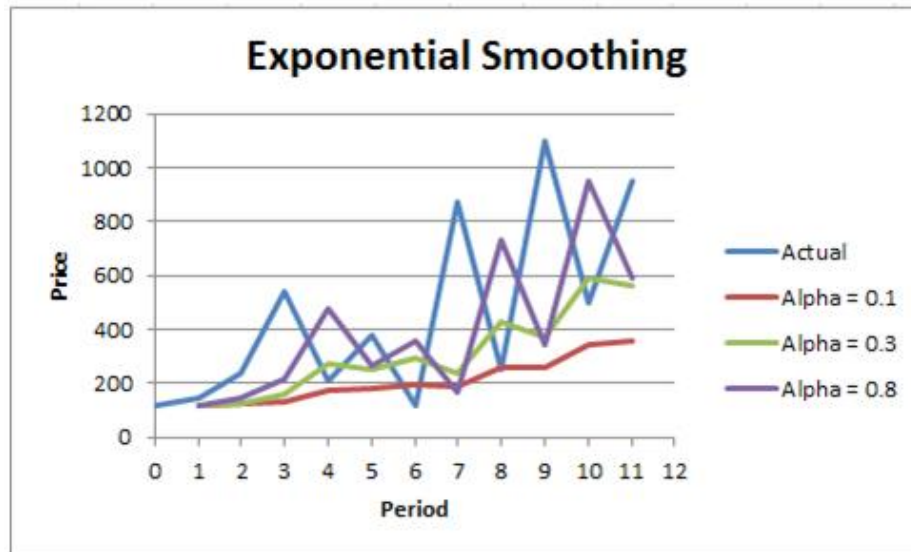
Exponential Smoothing:

Figure 3: Exponential Forecast

Here, we've alpha regard that's smoothing relentless and this technique is named an important exponential smoothing procedure which consider the opposite two elements reliable (for instance Regularity and Trend). Two folds (Holt's) Exp. Smoothing and Winter's Exp. Smoothing Methods overseeing two components as an example Pattern and Seasonality (for instance Beta and Gamma).

DIFFICULTIES WITH TRADITIONAL FORECASTING METHODS

Exponential Smoothing

- Recent observations have higher influence on the forecast.
- Exponentially decreasing weights are assigned to observations.

2.4.3ROLLINGFORECAST

A rolling forecast is an add/drop process for anticipating the future over a set period. Moving conjectures are frequently utilized in long haul climate forecasts, venture the board, store network the executives and budgetary arranging.

If, for instance, an association needs to foresee working costs a year ahead of time, the moving estimate's set period would be a year. After the main month had passed, that month would be dropped from the earliest starting point of the previous sales forecast and one more month would be added as far as possible of the figure. A moving average sales forecast is first in/first out (FIFO) process guarantees that the conjecture dependably covers a similar measure of time. Since a moving gauge window requires routine modifications, it is here and there alluded to as a persistent conjecture or an iterative forecast

Moving conjectures can be stood out from static forecasts and recursive figures. Static figures utilize a tally down procedure. A static forecasts for an association's yearly working costs, for instance, would in any case spread 12 explicit months - yet once those a year had passed, the primary figure would be changed and a totally new estimate would be made for the following a year.

2.4.4 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

Autoregressive Integrated Moving Average may be a determining system hooked in to exponential smoothing. ARIMA model is otherwise called the Box-Jenkins procedure. to urge the foremost ideal information about the parameter utilized in ARIMA relies upon recognizing verification strategy which was purposed by Box-Jenkins. ARIMA usage is identified with examples, normality, cycles, botches, and non-stationary pieces of an instructive accumulation when deciding. this is often the foremost well-known strategy and reasonable forever arrangement and commonly creates progressively precise outcomes than different strategies.

The parameters of the ARIMA model are characterized as pursues:

P: represent the slack request.
d: represent the extent normally.
Q: is that the size of moving normally

The ARIMA represents

- MA means moving Average.

- I for Integrated

- AR for Auto regression

DIFFICULTIES WITH TRADITIONAL FORECASTING METHODS

ARIMA

- Makes prediction based on previous p points (*auto-regressive*) and on the previous q residuals (*moving-average*).

- Inappropriate for Intermittent data as it allows values that aren't non-negative integers.

2.4.5 NEURAL NETWORK

An artificial neural network is AI that uses the info that ought to be within the sort of records. Its rules are often wont to recognize the precedent and make them truly adaptable and incredible.

Neural networks, has its own one-of-a-kind solidarity to get hugeness from befuddled or unsure data, and as a rule can be used to recognize the precedent and example in the data, which can't be unmistakable successfully any computer science methodology. Today Neural Network is equally important for Industrial Process Control, Risk Analysis, Data Visualization, Customer Search using Digital Marketing, Sales Forecasting etc.

Neural Network can be used in any type of research work get real results.

2.5 EVALUATION METRIC MODELS USED IN THE RESEACH WORK

Each machine learning model takes care of an issue with an alternate target, utilizing a perfect dataset and consequently, it is imperative to comprehend the setting before selecting a metric. Ordinarily, the results to the accompanying inquiry can help us pick the suitable metric:

- Type of task: Regression?

- Business objective?

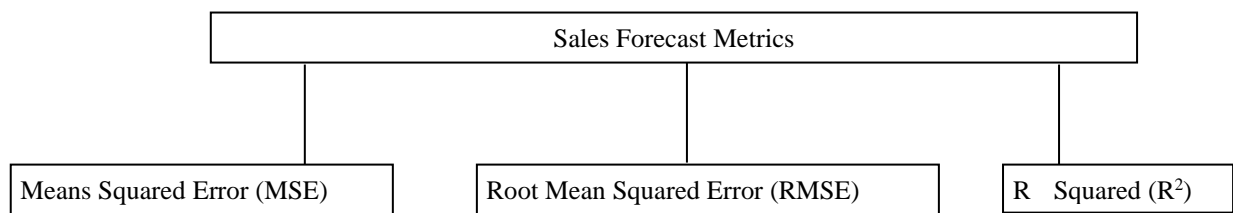- What is the dissemination of the objective variable?

Relapse Metrics

```
                        ┌─────────────────────────────────────────────┐
                        │            Sales Forecast Metrics            │
                        └─────────────────────────────────────────────┘
                             │                  │                │
         ┌───────────────────────┐  ┌───────────────────────────────┐  ┌───────────────────────┐
         │ Means Squared Error (MSE) │  │ Root Mean Squared Error (RMSE) │  │ R  Squared (R²) │
         └───────────────────────┘  └───────────────────────────────┘  └───────────────────────┘
```

Figure 4: Sales Metrics Used

2.5.1 MEAN SQUARED ERROR (MSE)

For regression evaluation, it is the most simple and popular metric. It is defined by the equation

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

where y_i is the actual expected output and $\hat{y}_i$ is the model's prediction.

MSE provides statistical results from gathered information and analyzed sales forecasts. The higher this value, the more terrible the model is. It is never negative, since we're squaring the individual expectation savvy blunders before summing them and zero for an ideal model.

This metrics is useful if we have startling qualities that we should think about. Very high or low esteem that we should focus.

Disadvantage: If we make a solitary exceptionally terrible forecast, the squaring will aggravate the mistake even and it might slant the measurement towards overestimating the model's disagreeableness. That is an especially hazardous conduct on the off chance that we have uproarious information (that is, information that for reasons unknown is not totally reliable)—even an "impeccable" model may have a high MSE in that circumstance, so it turns out to be difficult to pass judgment on how well the model is performing. Then again, if every one of the blunders are little, or fairly, littler than 1, than the contrary impact is felt: we may think little of the model's badness.

2.5.2 ROOT MEAN SQUARED ERROR TECHNIQUE

RMSE is square root of MSE.

$$\text{RMSE} = \sqrt{\frac{I}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} = \sqrt{MSE}$$

The root means square error (RMSE) a type of measure of error. It is a frequently used measure. To find the differences between the actual values observed and values predicted by an estimator or a model. Root mean square error is defined as the square root of differences between the values observed and values predicted. The RMSE estimates the magnitudes of the errors. It is a good measure of accuracy which is used to perform comparison forecasting errors from estimators of different topologies for a specific variable.

$$\text{MSE(a)} > \text{MSE(b)} \Leftrightarrow \text{RMSE(a)} > \text{RMSE(b)}$$

$$\frac{\partial RMSE}{\partial \hat{y}_i} = \frac{1}{2\sqrt{MSE}} \frac{\partial MSE}{\partial \hat{y}_i}$$

Gradient of RMSE with respect to i-th prediction

So, despite the fact that RMSE and MSE are extremely comparable as far as models scoring, they can be not promptly exchangeable for inclination-based strategies. We will most likely need to change a few parameters like the learning rate.

2.5.3 R SQUARED (R²)

The coefficient of assurance, or R² (occasionally read as R-two) is firmly identified with MSE, however have the benefit of being scale-free doesn't make a difference if the yield esteems extremely huge or little, the R² is the value between infinity and 1.

The model is worse than predicting the mean, if in case the r-squared is negative, .

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)}$$

The MSE of the model is computed as above, while the MSE of the baseline is defined as:

$$MSE(baseline) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

To make it clearer, this baseline MSE can be thought of as the MSE that the simplest **possible** model would get. In conclusion, R² represents the ratio, how good our model is by how good is the naive mean model.

2.6 THE FORECASTING ERRORS

Professor Paul Goodwin from the Management School at the University of Bath in England was the keynote moderator at the Analytics Conference in Copenhagen. He uncovered some fascinating exploration that he and his partners saw about the way during which organizations endeavor to foresee the longer term at different focuses in their supply chains. Shockingly, the examination found that despite the very fact that organizations put resources into and utilize investigative programming, the organizations balanced up to 93 percent of figures produced by the merchandise. Goodwin's introduction included extraordinary data from the examination that helped the group of spectators see why they ought to Let the PC make the conjecture - not the manager. He likewise sketched out nine normal forecasting mistakes:

2.6.1Anchoring and modifying –

When making an estimate people regularly begin with an underlying worth – the anchor and afterward change from this. Yet, estimates will generally be excessively near the grapple, no matter whether it is an unrealistic worth. Tying down can make individuals think a touch of upward patterns since they continue to be excessively near the newest worth, which may be especially serious for exponential patterns.

Solution: Use measurable techniques, as against judgment, to forecast patterns.

2.6.2 Seeing patterns in irregularity– Human creatures tend to ascertain deliberate examples even where there is nothing to urge. Individuals love narrating and are splendid at concocting clarification for irregular developments in diagrams.

Solution: Do not accept that you simply can complete superior employment you are your estimating framework.

2.6.3 Joining tons of weight to judgment with reference to statistical forecast – Even however proof shows judgment is a smaller amount exact than measurable figures individuals keep it up counting on their judgment.

Solution: Have more trust in factual techniques and alter them just once you are certain it is completely fundamental.

2.6.4 Recency Inclination – If the patterns are extraordinary, Companies would prefer not utilizing the data that goes more than a couple of years. In any case, factual techniques inserted in programming need a big dataset to give dependable estimates.

Solution: Give your product a possibility, at that point you should not change its gauges. Numerous factual techniques are applied for changes in patterns, and they are far more outlandish than manual judgment to find false new patterns in ongoing information.

2.6.5 Good faith inclination – People have an inborn predisposition toward hopefulness, analysts state.

Even though confidence predisposition is an indication of good emotional well-being, negative (descending) alterations are more fruitful than positive.

Solution: Break complex decisions into littler parts, for instance, modify for value decrease, advancement, and new clients independently as opposed to making a complete change. Furthermore, consider keeping a database of like deals and advancements that is evaluated the impacts of past unique occasions.

2.6.6 Political predisposition – Figures adjustment dependent on inner political inclination can make promoting directors look great when they give excessively hopeful gauges.

Solution: It is required to changes in accordance with that archived and audit their impact on precision.

2.6.7 Mistaking gauges for choices –

A conjecture is that the best gauge of what is going to occur later:" i feel we'll sell 200units." a choice may be a number intended to accomplish a minimum of one destination:" I figure we should always create 250 units on the off chance that request is out of the blue high, to regulate the likelihood of lost deals with the expenses of holding increasingly stock."
Solution: The estimate from the selection should not have a link. Get your gauge first, name it as a conjecture and afterward use as a reason for closing

2.6.8 Gathering inclinations – Having measurable gauges balanced by a gathering of people are often hazardous because the overwhelming majority do not feel great conflicting with cooperative choices.

Solution: Delphi Method can be used, wherein specialists provide estimates separately and secretly. The further effects of the surveying are then counted and insights nourished back. The surveys are continued and the procedure is rehashed until accord develops. Middle gauge is then utilized as the forecast.

HARDWARE
AND
SOFTWARE
REQUIREMENTS

SYSTEM ENVIRONMENT:

After analysis, some resources are required to convert the abstract system into the real one. The hardware and software selection begins with requirement analysis, followed by a request for proposal and vendor evaluation.

Software and real system are identified. According to the provided functional specification all the technologies and its capacities are identified.

Basic functions and procedures and methodologies are prepared to implement.

Some of the Basic requirements such as hardware and software are described as follows: -

HARDWARE AND SOFTWARE SPECIFICATION:

SOFTWARE REQUIREMENTS:

- Operating systems: Windows 10
- Coding Language: Python 3.8
- Web Framework: Google colab

HARDWARE REQUIREMENTS:

- System: Pentium i3 Processor
- Hard Disk: 500 GB
- Monitor: 15" LED
- Input Devices: Keyboard, Mouse
- Ram: 4 GB

SYSTEM ANALYSIS

PURPOSE:

The purpose of the system analysis is to comprehensively understand, evaluate, and improve the current state of the system under consideration. This involves identifying requirements, assessing existing processes, and proposing enhancements to meet specific objectives.

PROJECT SCOPE:

The project scope delineates the boundaries of the system analysis effort. It defines the functionalities, features, and constraints that are within the scope of the analysis. This ensures a clear understanding of what will be covered and sets realistic expectations for stakeholders regarding the deliverables of the analysis phase.

EXISTING SYSTEM:

The existing system refers to the current state of the system or processes in place before any modifications. This analysis involves a detailed examination of the strengths, weaknesses, opportunities, and threats associated with the current system. It includes an assessment of technologies, workflows, and any shortcomings that need to be addressed.

PROPOSED SYSTEM:

The proposed system outlines the improvements, changes, or enhancements recommended based on the findings from the analysis. This includes defining new features, functionalities, or technologies that can address the identified shortcomings and align with the goals and requirements of the stakeholders.

SYSTEM OVERVIEW:

The system overview provides a high-level description of the entire system, incorporating both the existing and proposed elements. It serves as a snapshot that encapsulates the key components, processes, and interactions within the system. This section offers a broad understanding for stakeholders who may not be intimately familiar with the technical intricacies but need a comprehensive view of the system's architecture and functionality.

TOOLS

AND

TECHNOLOGIES

1. **<u>NumPy</u>**

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

At the core of the NumPy package, is the ndarray object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance. There are several important differences between NumPy arrays and the standard Python sequences:

• NumPy arrays have a fixed size at creation, unlike Python lists (which can grow dynamically). Changing the size of an ndarray will create a new array and delete the original.

• The elements in a NumPy array are all required to be of the same data type, and thus will be the same size in memory. The exception: one can have arrays of (Python, including NumPy) objects, thereby allowing for arrays of different sized elements.

• NumPy arrays facilitate advanced mathematical and other types of operations on large numbers of data. Typically, such operations are executed more efficiently and with less code than is possible using Python's built-in sequences.

• NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently. This behaviour is called locality of reference in computer science. This is the main reason why NumPy is faster than lists. Also it is optimized to work with the latest CPU architectures.

• NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

• A growing plethora of scientific and mathematical Python-based packages are using NumPy arrays; though these typically support Python-sequence input, they convert such input to NumPy arrays prior to processing, and they often output NumPy arrays.

In other words, to efficiently use much (perhaps even most) of today's scientific/mathematical Python-based software, just knowing how to use Python's built-in sequence types is insufficient - one also needs to know how to use NumPy arrays.

The points about sequence size and speed are particularly important in scientific computing. As a simple example, consider the case of multiplying each element in a 1-D sequence with the corresponding element in another sequence of the same length. If the data are stored in two Python lists, a and b, we could iterate over each element:

```
c = []
for i in range(len(a)): c.append(a[i]*b[i])
```

This produces the correct answer, but if a and b each contain millions of numbers, we will pay the price for the inefficiencies of looping in Python. We could accomplish the same task much more quickly in C by writing (for clarity we neglect variable declarations and initializations, memory allocation, etc.)

```
for (i = 0; i < rows; i++) {
c[i] = a[i]*b[i];
}
```

This saves all the overhead involved in interpreting the Python code and manipulating Python objects, but at the expense of the benefits gained from coding in Python. Furthermore, the coding work required increases with the dimensionality of our data. In the case of a 2-D array, for example, the C code (abridged as before) expands to

```
for (i = 0; i < rows; i++) {
for (j = 0; j < columns; j++) {
c[i][j] = a[i][j]*b[i][j]; }
}
```

2. __PANDAS__

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open-source data analysis/manipulation tool available in any language. It is already well on its way toward this goal. Pandas is well suited for many different kinds of data:

● Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet.

● Ordered and unordered (not necessarily fixed-frequency) time series data.

● Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels.

● Any other form of observational / statistical data sets. The data need not be labelled.

at all to be placed into pandas data structure.

The two primary data structures of pandas, **Series** (1-dimensional) and **DataFrame** (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. For R users, Data Frame provides everything that R's data frame provides and much more. Pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Here are just a few of the things that pandas does well:

● Easy handling of missing data (represented as NaN) in floating point as well as non-floating-point data.

● Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects.

● Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for you in computations.

● Powerful, flexible group by functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data.

● Make it easy to convert ragged, differently-indexed data in other Python and NumPy data structures into DataFrame objects.

● Time series-specific functionality: date range generation and frequency conversion, moving window statistics, date shifting, and lagging.

3. MATPLOTLIB
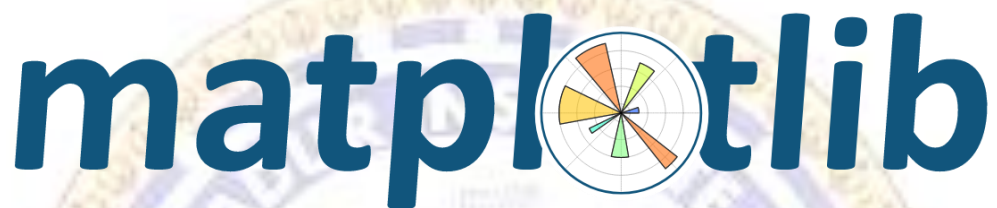
Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications. A Python matplotlib script is structured so that a few lines of code are all that is required in most instances to generate a visual data plot. The matplotlib scripting layer overlays two APIs:

● The **pyplot** API is a hierarchy of Python code objects topped by **matplotlib.pyplot**

● An OO (Object-Oriented) API collection of objects that can be assembled with greater flexibility than pyplot. This API provides direct access to Matplotlib's backend layers.

Matplotlib makes hard things possible and things easy .
 • Create publication quality plots.
 • Make interactive figures that can zoom, pan, update.
 • Customize visual style and layout.
 • Export to many file formats.
 • Embed in Jupyter Lab and Graphical User Interfaces.
 • Use a rich array of third-party packages built on Matplotlib.

Matplotlib and Pyplot in Python

The pyplot API has a convenient MATLAB-style stateful interface. In fact, matplotlib was originally written as an open source alternative for MATLAB. The OO API and its interface is more customizable and powerful than pyplot, but considered more difficult to use. As a result, the pyplot interface is more commonly used.
Now let's look at the various types of plots available in matplotlib :

Basic plots

Basic plot types, usually y versus x.

plot(x, y)   scatter(x, y)   bar(x, height)

stem(x, y)   step(x, y)   fill_between(x, y1, y2)   stackplot(x, y)
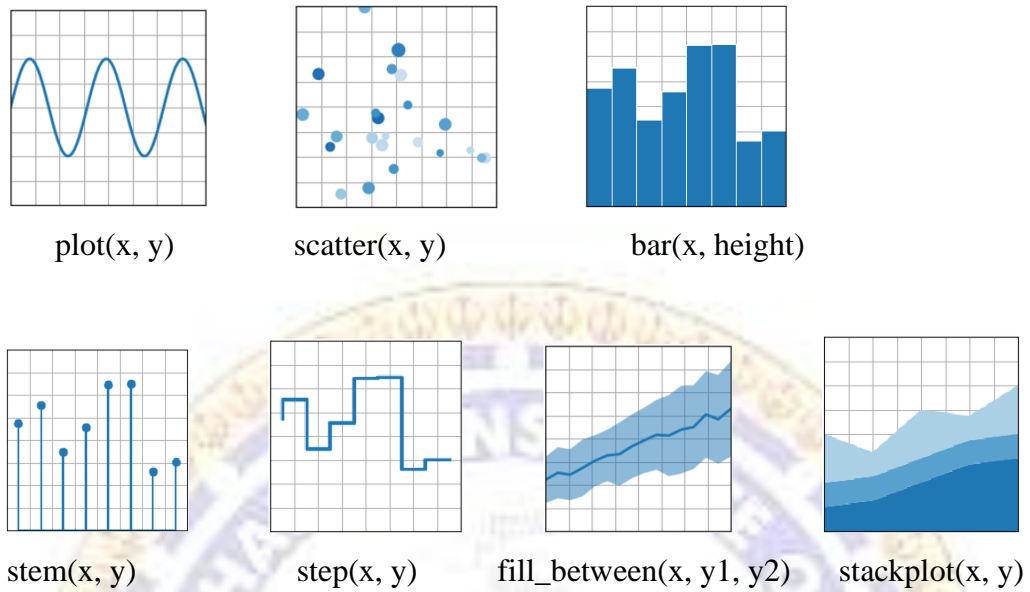
Figure 5

Plots of arrays and fields

Plotting for arrays of data Z(x, y) and fields U(x, y), V(x, y).

imshow(Z)   pcolormesh(X, Y, Z) contour(X, Y, Z)   contourf(X, Y, Z)

barbs(X, Y, U, V)   quiver(X, Y, U, V)   streamplot(X, Y, U, V)
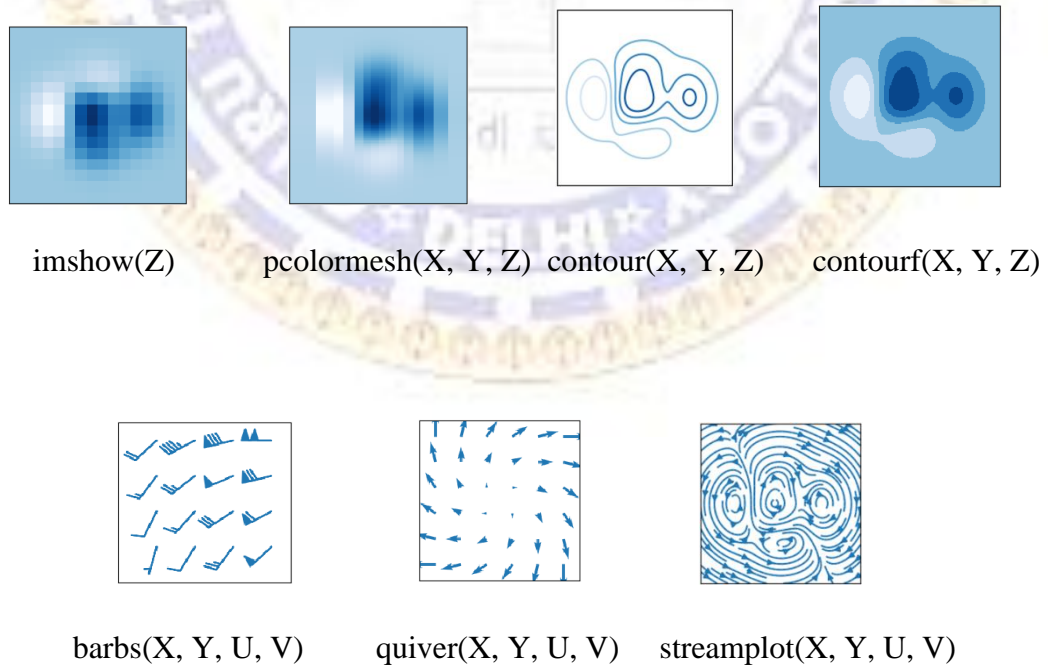
Figure 6

Statistics plots

Plots for statistical analysis.
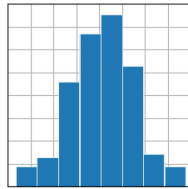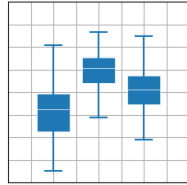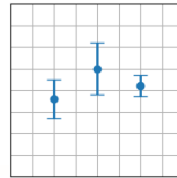
hist(x)   boxplot(X)   errorbar(x, y, yerr, xerr)   violinplot(D)
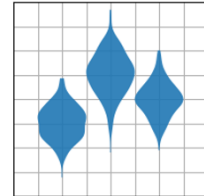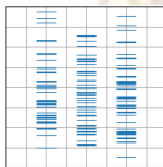
eventplot(D)   hist2d(x, y)   hexbin(x, y, C)   pie(x)
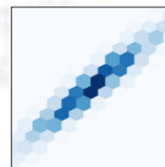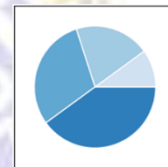
Figure 7

Unstructured coordinates

Sometimes we collect data z at coordinates (x,y) and want to visualize as a contour. Instead of gridding the data and then using contour, we can use a triangulation algorithm and fill the triangles.
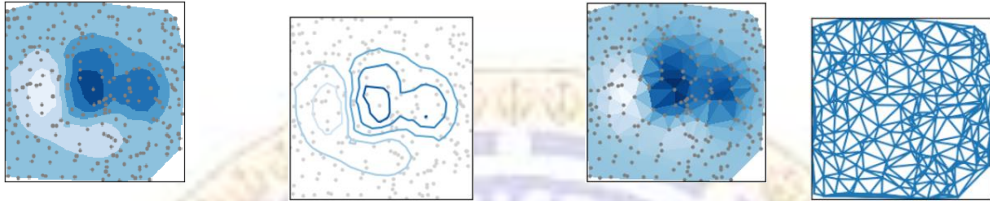
Tricontour(x, y, z)   tricontourf(x, y, z)   tripcolor(x, y, z)   triplot(x,y)

Figure 8

4. **SEABORN**

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.

Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for the same variables for better understanding of the dataset. Different categories of plot in Seaborn: Plots are basically used for visualizing the relationship between variables.

Those variables can either be completely numerical or a category like a group, class or division. Seaborn divides plot into the below categories –

● **Relational plots:** This plot is used to understand the relation between two variables.

● **Categorical plots:** This plot deals with categorical variables and how they can be visualized.

● **Distribution plots:** This plot is used for examining univariate and bivariate distributions

● **Regression plots:** The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analysis.

● **Matrix plots:** A matrix plot is an array of scatterplots.

● **Multi-plot grids**: It is a useful approach to draw multiple instances of the same plot on different subsets of the dataset.

● **Pair Plot :**

The Seaborn Pairplot function allows the users to create an axis grid via which each numerical variable stored in data is shared across the X- and Y-axis in the structure of columns and rows. We can create the Scatter plots in order to display the pairwise relationships in addition to the distribution plot displaying the data distribution in the column diagonally. The pairplot() function can also be used to showcase the subset of variables, or we can plot different types of variables on rows and columns.
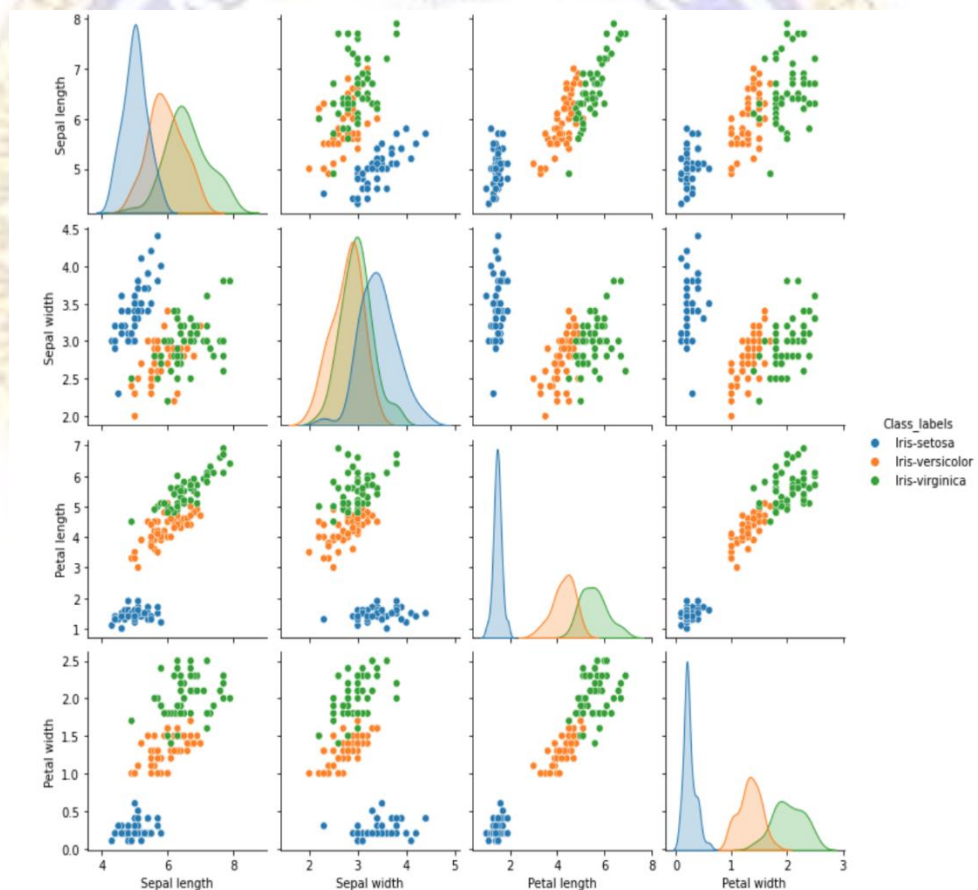
Figure 9

5. SKLEARN (SCIKIT LEARN)

It is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Here we imported the data set using pandas and explored the data using methods like head(),tail(), dtypes() and then try our hand at using plotting techniques from Seaborn to visualize our data.

Machine Learning:

SUPERVISED LEARNING:

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

How Supervised Learning Works?

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

The working of Supervised learning can be easily understood by the below example and diagram:
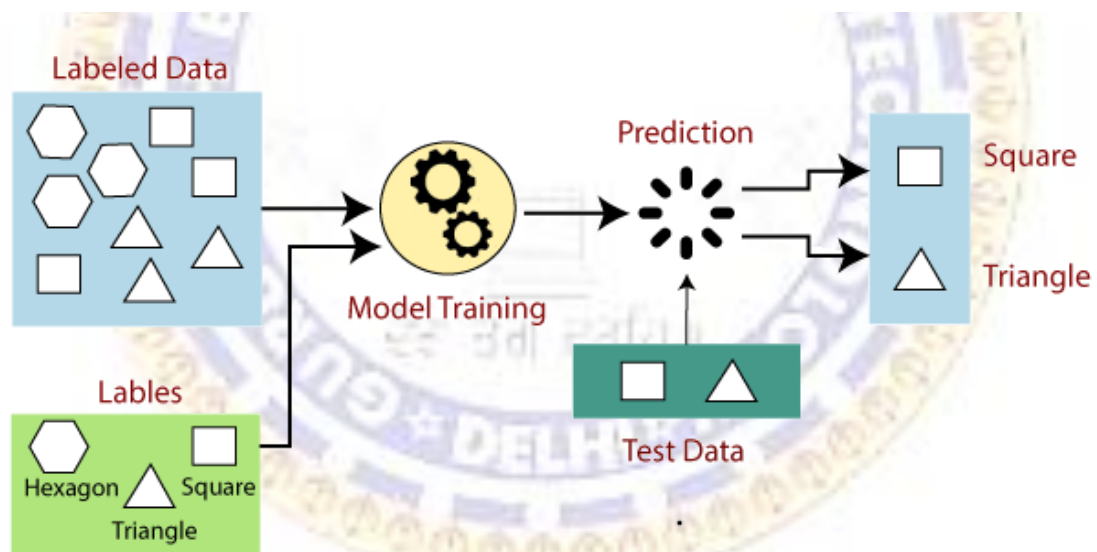

Figure 10

Types of supervised Machine learning Algorithms:

Supervised learning can be further divided into two types of problems:
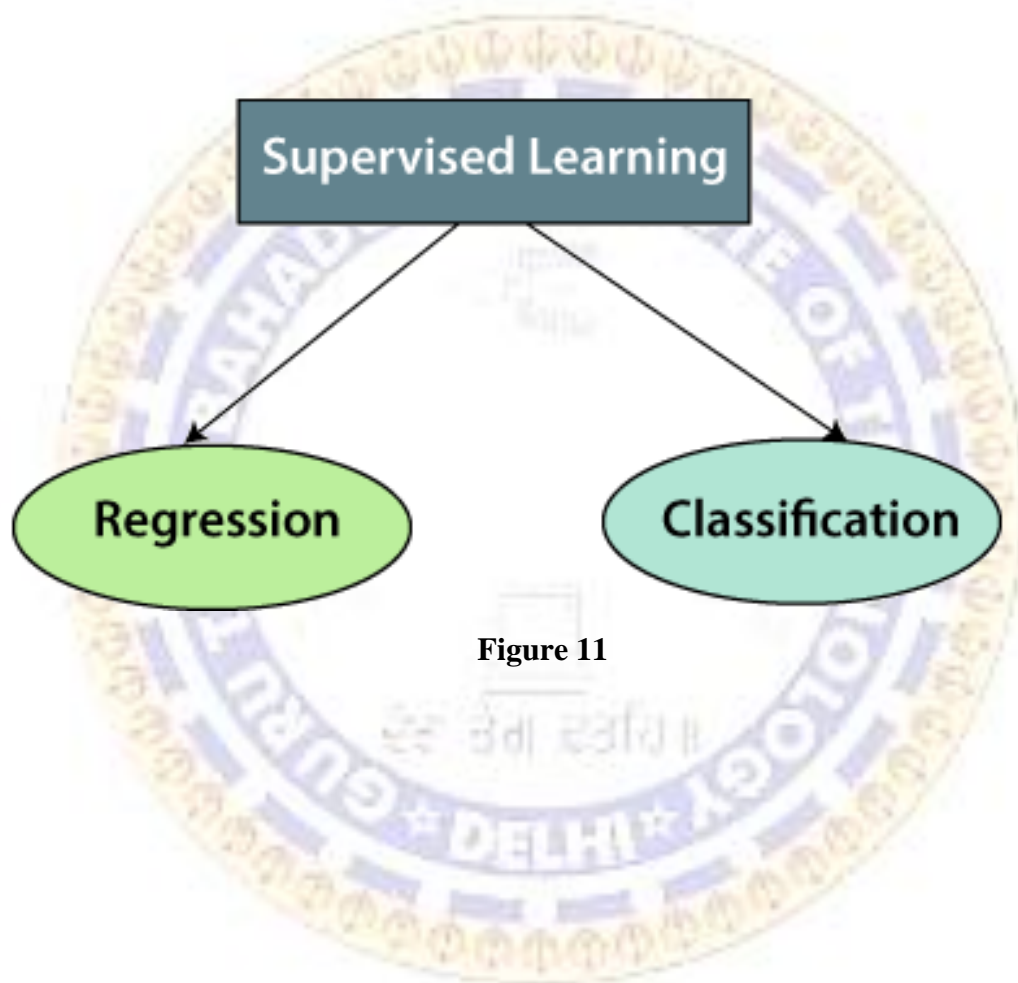
Figure 11

1. <u>Regression</u>

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

2. <u>Classification</u>

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

- Spam Filtering,
- Random Forest
- Decision Trees
- Logistic Regression
- Support vector Machines

UNSUPERVISED LEARNING:

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

Working of Unsupervised Learning

Working of unsupervised learning can be understood by the below diagram:
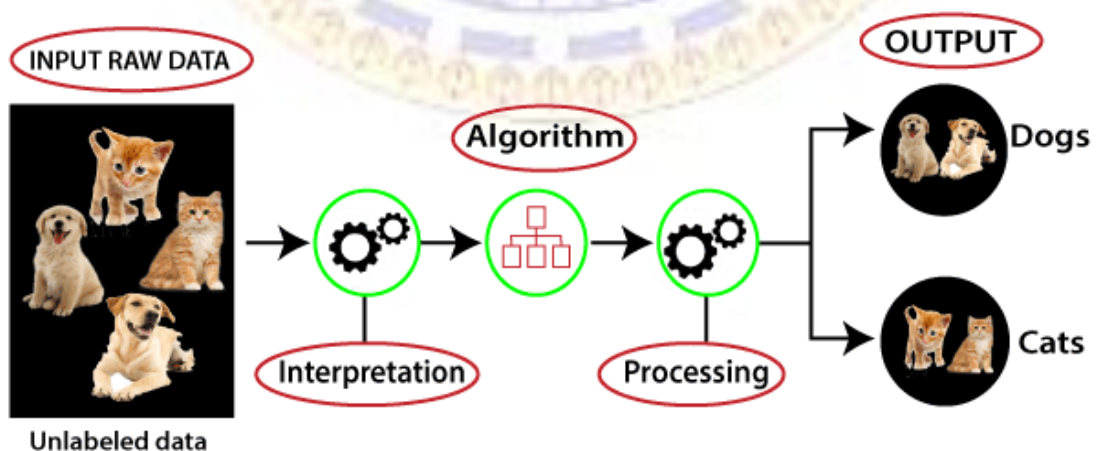
Figure 12

Here, we have taken an unlabelled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabelled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.

Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

Types of Unsupervised Learning Algorithm:

The unsupervised learning algorithm can be further categorized into two types of problems:
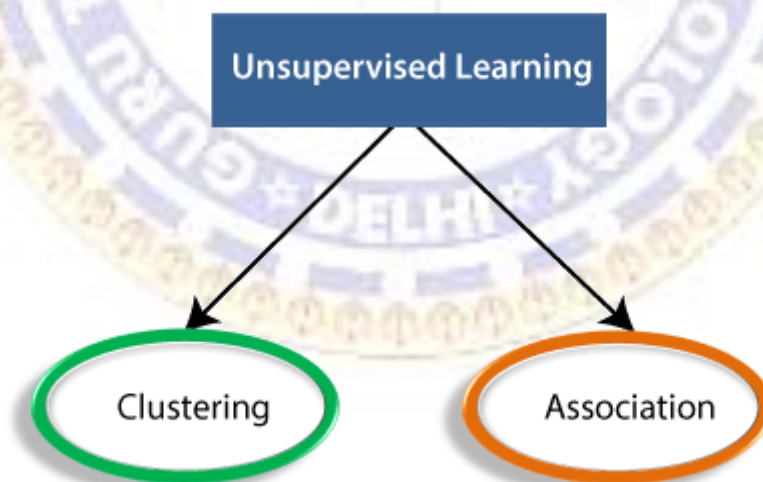
Figure 13

Clustering: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

Association: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Steps involved in Machine Learning
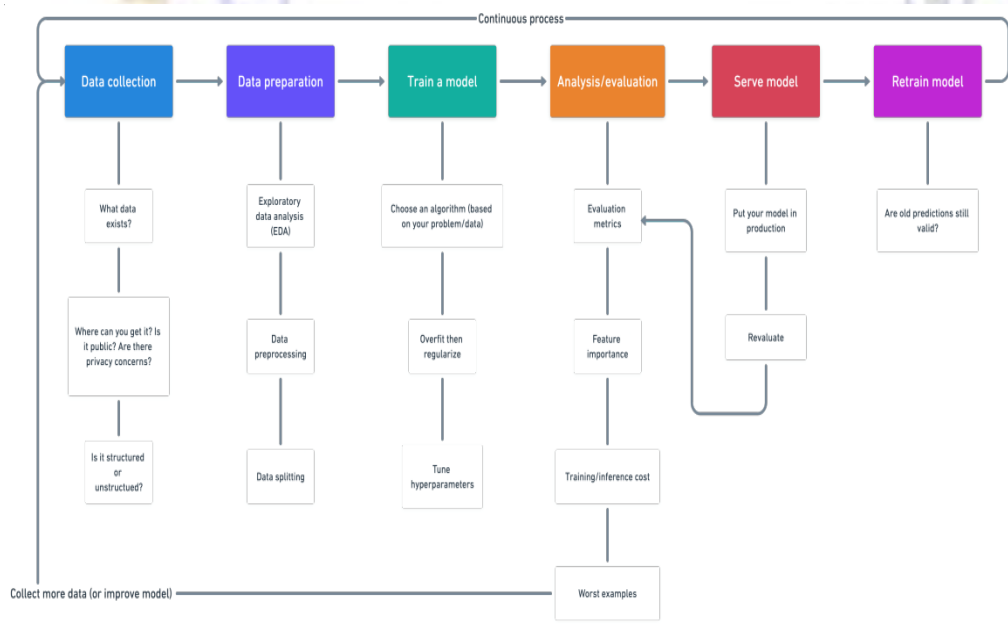
Figure 14

METHODOLOGY AND MATERIAL USED

For our investigation, we utilized store sales authentic information from the "Rossmann Store Sales" Kaggle rivalry. This information portrays sales in Rossmann stores. The figurings were directed within the Python code utilizing the principal libraries pandas, sklearn, NumPy, matplotlib, seaborn. to steer the examination, Python IDLE and Page as IDE was utilized.

4.1 ALGORITHMS USED IN THE RESEARCH

The following example shows the demand for the item for 6 months. To calculate the three months moving average and forecast for the 7th month.

Month   1   2   3   4   5   6
Demand (100's)   42   41   43   38   35   37

Define a constant where α, 00. exponentially smoothed moving average for the period it (Mt say) can be

$M_{t+1} = \alpha Y_t + (1-\alpha)M_t$, 00

So you'll see here that the exponentially smoothed moving normal considers the bulk of the past perceptions, check out the moving normal above where just a few of the past perceptions were considered. Besides, the exponentially smoothed moving normally for period t may be a direct blend of this esteem (Y_t) and therefore the past exponentially smoothed moving average (M_{t+1}).

The steady α is understood because the smoothing consistent and therefore the estimation of the load given to this perception (Y_t) in computing the exponentially smoothed moving normal Mt for period t (which is that the forecast for period t + 1). as an example, on the off chance that α = 0.2, at that time this demonstrates 20% of the load in creating figures is relegated to the newest perception and therefore the staying 80% to previous observations.

Exponential smoothing also can be viewed as a forecast continually updated by the forecast error just made.

For the demand data given above, the exponentially smoothed moving average with α = 0.2 and 0.8 as smoothing constant . For α = 0.2, we have.

$M_1 = Y_1 = 42$(we always start with $M_1 = Y_1$)

$M_2 = 0.2Y_2 + 0.8M_1 = 0.2(41) + 08(42) = 41.80$

$M_3 = 0.2Y_3 + 0.8M_2 = 0.2(43) + 0.8(41.80) = 42.04$

$M_4 = 0.2Y_4 + 0.8M_3 = 0.2(38) + 0.8(42.04) = 41.23$

$M_5 = 0.2Y_5 + 0.8M_4 = 0.2(35) + 0.8(41.23) = 39.98$

$M_6 = 0.2Y_6 + 0.8M_5 = 0.2(37) + 0.8(39.98) = 39.38$

Note here that it is usually sufficient to just work to two or three decimal placeswhen doing exponential smoothing. We use M6 as the forecast for month 7, i.e. the forecast for month 7 is 39.38 units.

We have the following for **α = 0.9**.

$M_1 = Y_1 = 42$

$M_2 = 0.9Y_2 + 0.1M_1 = 0.9(41) + 0.1(42) = 41.10$

$M_3 = 0.9Y_3 + 0.1M_2 = 0.9(43) + 0.1(41.10) = 42.81$

$M_4 = 0.9Y_4 + 0.1M_3 = 0.9(38) + 0.1(42.81) = 38.48$

$M_5 = 0.9Y_5 + 0.1M_4 = 0.9(35) + 0.1(38.48) = 35.35$

$M_6 = 0.9Y_6 + 0.1M_5 = 0.9(37) + 0.1(35.35) = 36.97$

i.e. 36.97 is the forecast for 7th Month.

4.2 DATASET FOR RESEARCH WORK

Dataset for simulation is downloaded from kaggle.com. Data is stored in the form of CSV files. These files contain comma separated values. A CSV file is used to store tabular data in the form of text file. Some services have given permissions to their users to export tabular data into a CSV file from the websites. A CSV file can be open in Excel, and nearly all databases have a tool to allow import from CSV file. In Excel, standard format of the data is followed in the form of rows and columns.

Moreover, each row is terminated by a newline to begin the next row. In each row, comma is separating the data.

Reading CSV File

To read CSV file, generate a reader object and we must have to use the reader function to.

Each row of the file can be read through reader function to create a list of columns. Now select the relevant colums.

```
#import required modules
import csv%to read csv files
with open('c:\data.csv','rt') as fil:
  rel_data = csv.reader(fil)
  for rows in rel_data:
     print(rows)
```

The data file used in the study is train.csv. This file contains data for 2013, 2014, 2015, 2016 and 2017. The following python statements are used to create data files having records of different item codes.

 # Reading data of a particular store from train.csv and writing in data.csv

```
import csv% to import libraries

import sys%to use system  files
```

#Enter item details for searching

```
Store_no = input('Enter store number')

item_no=input('Enter item number')
```

#read csv file

```
new_csv_file = csv.reader(open('train.csv',"r"))

f=open('data6.csv', 'w')
```

#reading csv file row wise

```
for rows in new_csv_file:

   if Store_no == rows[2] and item_no==rows[3]:
```

```
    #print (rows)
```

for l in rows:

f.write(l)

f.write(',')

f.write('\n')

f.close()

Train.csv file contains Ten lakh forty eight thousand records for twenty three stores. Each store is selling twenty items. The unit sale of each item in each store is give date wise.

ID	Date	Store_number	Item_number	Unit_Sales
0	1/1/2013	25	103665	7
1	1/1/2013	25	105574	1
2	1/1/2013	25	105574	2

Table 1

Form train.csv, ten files are extracted having data for ten items.
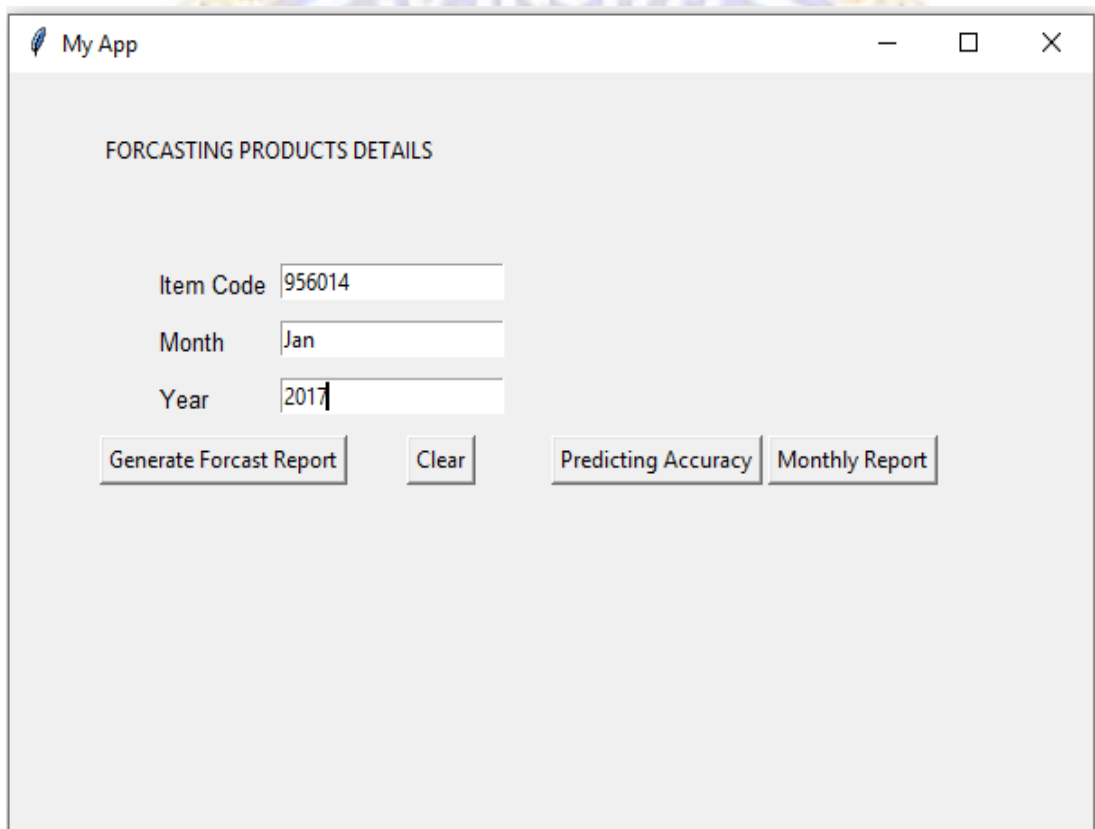
S.No	File name	Items.No
1	Data2.cvs	956014
2	Data3.cvs	956013
3	Data4.cvs	956012
4	Data5.cvs	956011
5	Data6.cvs	119629
6	Data7.cvs	123601

7	Data8.cvs	153267
8	Data9.cvs	122725
9	Data10.cvs	153229
10	Data11.cvs	119624

Table 2

5.1 IMPLEMENTATION

Simulator for Sales Forecasting

Figure 15: GUI for Simulator

Input to Simulator for Sales Forecasting

1. # Item Code = 956014
2. # Month = Jan
3. # Year 2017

   Output :
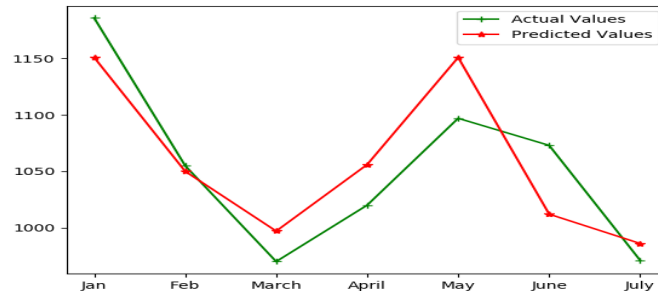
Figure 16: Predicted and actual values

1. s956014 Jan 2017

2. Actual sales of Jan [2013, 2014, 2015, 2016]

3. [1146.0, 1190.0, 1304.0, 1046.0]

4. Predicted value of Jan **in** 2017 using Simple Linear Regression =1124.9999999999927

5. Predicted values using Linear Regression

6. [1193.0, 1181.0, 1169.0, 1156.0, 1144.0]

7. combine list of monthly sale of Jan **in** [2013, 2014, 2015, 2016, 2017] using Rolling Forecast with Exponential Smoothing Method

8. [1146.0, 1190.0, 1304.0, 1046.0, 1157]

9. Predicted sale of Jan **in** 2017 using Linear Regression based on combine list of monthly sale using Rolling Forecast with Exponential Smoothing Method =1144.2000000000007

10. Final Predicted sale 1150.6000000000004

11. Coefficients: [[1.05617235]]

12. R2 score: 0.05

13. Mean squared error: 65.56

14. RMSE: 8.096817

15. actual= [1146.] observed= [1194.37060534]

16. actual= [1190.] observed= [1181.69653714]

17. actual= [1304.] observed= [1169.02246894]

18. actual= [1046.] observed= [1155.29222839]

19. actual= [1157.] observed= [1142.61816019]

20. average of observed values 1168.6

21. r2 calculated [0.04909595]

For Item Code: 956014

Sales for January month of the year 2013,2014,2015,2016 and 2017

DATE	2013	2014	2015	2016	2017
2	61	54	81	80	153
3	31	50	97	78	94
4	40	79	132	48	71
5	79	99	73	40	17
6	74	32	58	39	49
7	49	31	48	22	63
8	44	33	17	38	55
9	28	30	24	43	29
10	36	22	57	74	22
11	24	69	78	21	29
12	41	68	26	24	11
13	61	36	28	11	24
14	38	28	30	10	50
15	20	25	21	13	59
16	32	23	26	31	28
17	16	25	56	57	20
18	18	50	57	14	30
19	36	67	36	22	26
20	72	31	25	16	29
21	23	26	32	12	36
22	25	29	14	24	50
23	29	31	18	40	13
24	25	16	46	51	29
25	25	56	66	29	24
26	67	64	30	18	22
27	57	13	18	26	21
28	25	21	17	10	41
29	24	21	14	25	54
30	21	27	32	51	8
31	25	34	47	79	29
TOTAL	1146	1190	1304	1046	1186

Table 3: Per Day Sales for January Month

Actual values of four years

$Y_1 = 1146$

$Y_2 = 1190$

$Y_3 = 1304$

$Y_4 = 1046$

Regression line for actual data
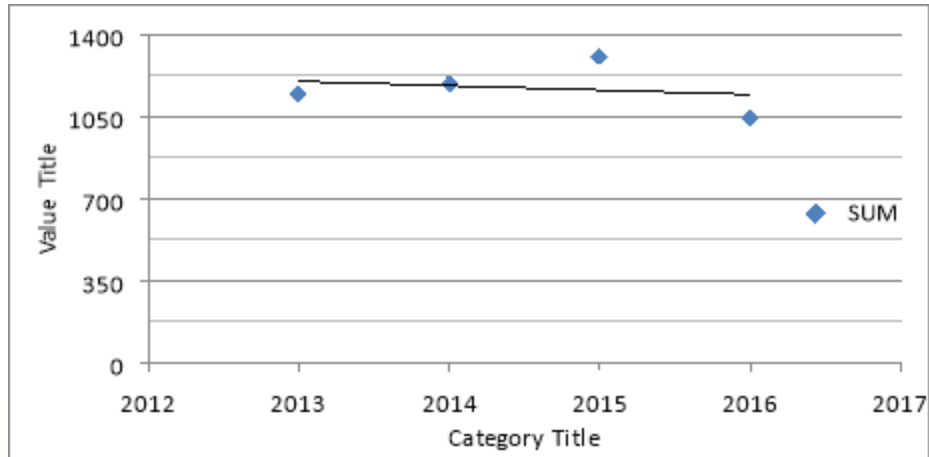
Figure 17: Regression Line

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1729.8	1729.8	0.106223	0.775427
Residual	2	32569.2	16284.6		
Total	3	34299			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	38641.2	114966.4	0.336109	0.768776	-45601 9	533301.6	-45601 9	533301.6
X Variable 1	-18.6	57.06943	-0.32592	0.775427	-264.15	226.9499	-264.15	226.9499

Table 4 : ANOVA

On substituting the independent variables in regression line equation: y=-18.6x+3841

Exponential Smoothing

Equation used for forecasting:

$S_{t+1}=\alpha y_t+(1-\alpha)S_t, 0<\alpha\leq1, t>0.$

$S_1=y_1=1146$

$S_2 = 0.2Y_2 + 0.8 S_1$

$= (0.2)(1190) + (1-0.2)(1146)$

$= 1154.8$

$= 1155$

$S_3 = 0.2Y_3 + 0.8S_2$

$= 0.2*(1304) + 0.8*(1155)$

$= 1184$

$S_4 = 0.2Y_4 + 0.8S_3$

$= 0.2*1046 + 0.8*1184$

$= 1156.8$

$= 1157$

Now S_4 is considered as the predicted sales for 2017.

Forecasting for 2018,

$Y_1 = 1190$

$Y_2 = 1304$

$Y_3 = 1046$

$Y_4 = 1157$

$S_1 = Y_1 = 1190$

$S_2 = 0.2*Y_2 + 0.8*S_1$

$= 0.2*13.4 + 0.8*1190$

$= 260.8 + 952$

$= 1212.8$

$S_3 = 0.2*Y_3 + 0.8*S_2$

$= 0.2*1046 + 0.8*1213$

$= 209.2 + 970.4$

$= 1179.6$

$S_4 = 0.2*y_4 + 0.8*S_3$

$= 0.2*1157 + 0.8*1179.6$

$= 231.4 + 943.6$

$= 1175$

Similarly forecasting for 2019,

We will consider Y_1 as 1304.

We will get S_5 as 1222.

Year	Actual Value	Value By Regression Line
2013	1146	1172.0
2014	1190	1174.0
2015	1304	1175.0
2016	1046	1177.0
2017	1157	1179.0
2018	1175	1181.0
2019	1222	1183.0

Table 5 : Sales Values

Average value with two approaches$=\dfrac{1222+1183}{2}$

$$=1202.5$$

$$=12030$$

The, Mean Square Error(MSE) = 52.53

The, Root Mean Square Error(RMSE) =7.24

SYSTEM DESIGN

USE CASE DIAGRAM:

Use case diagram consists of use cases and actors and shows the interaction between them.

The key points are:

1) Identify the primary actors involved in the sales forecasting system.
2) Define the main use cases for the system, considering actions and interactions between actors and the system.
3) Specify relationships between use cases and actors, highlighting the flow of information and actions.
4) Include a use case for "Relapse Investigation" using the feed-forward neural network as a forecasting tool.
5) Ensure that the use case diagram reflects the key elements of sales forecasting, such as data preparation, training, testing, and prediction.
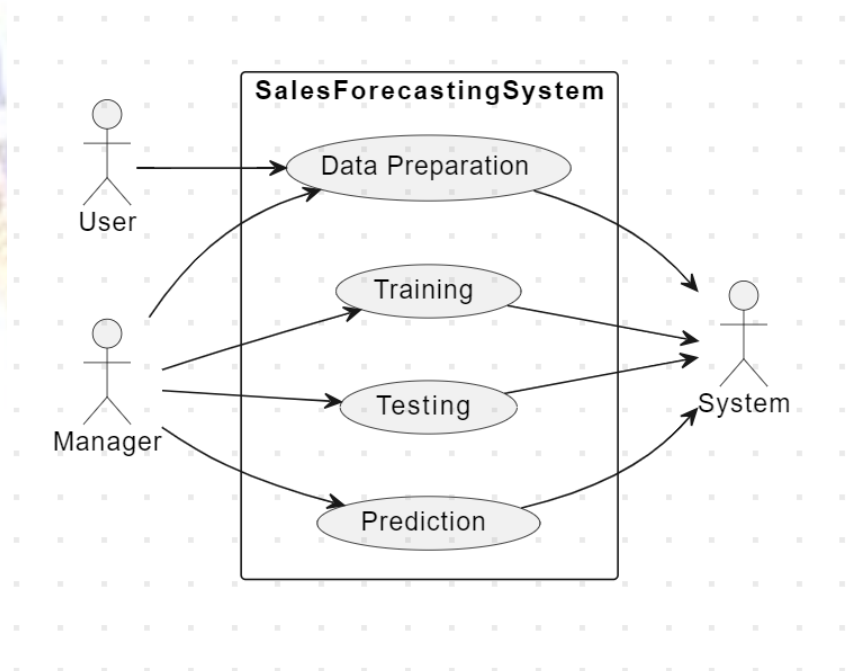
Figure 18

DATA FLOW DIAGRAM(DFD):
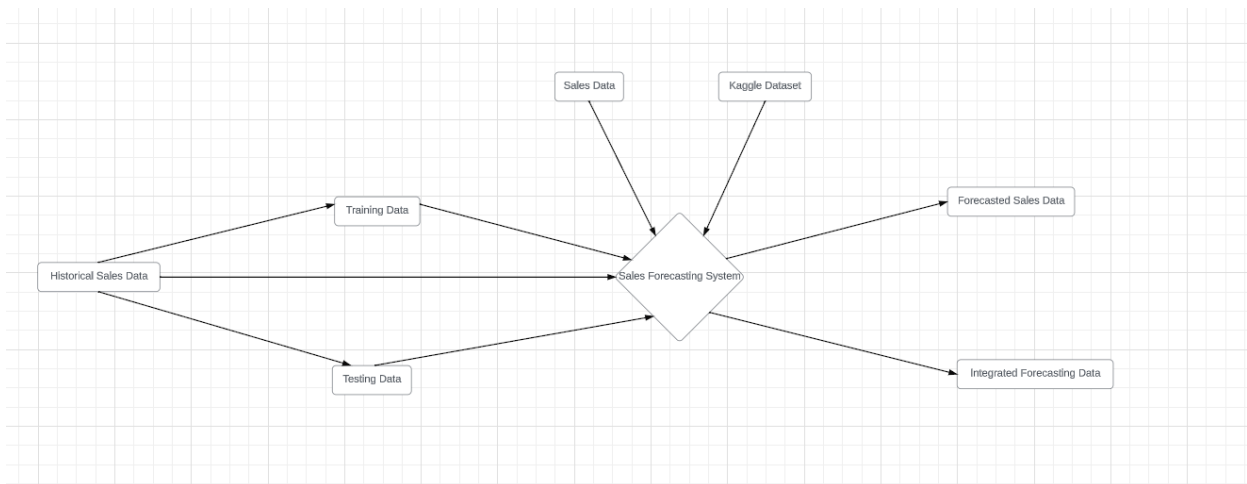
Figure 19

ER DIAGRAM:
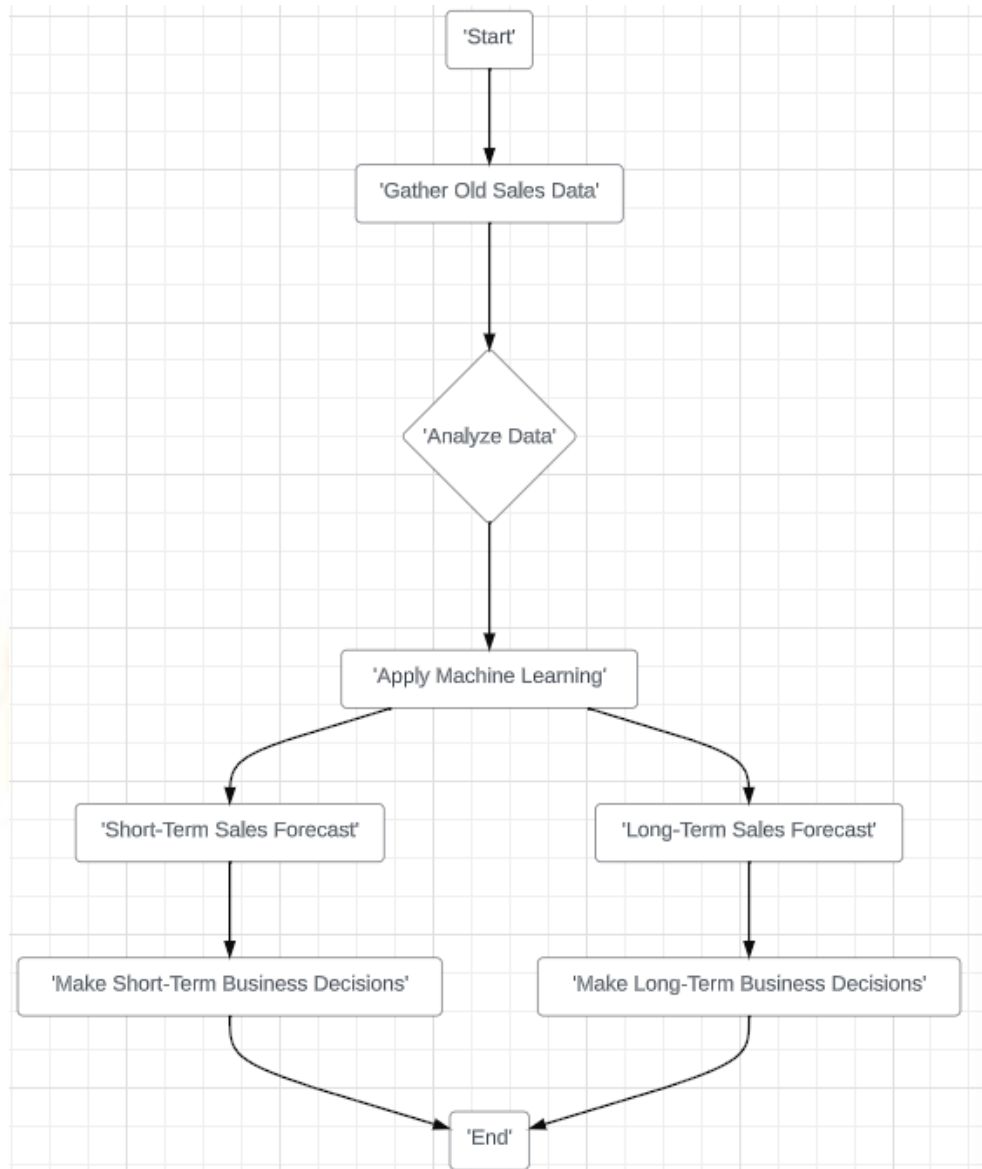
Figure 20

IMPLEMENTATION AND SYSTEM TESTING:

After all phase have been perfectly done, the system will be implemented to the server and the system can be used.

SYSTEM TESTING:

The goal of the system testing process was to determine all faults in our project. The program was subjected to a set of test inputs and many explanations were made and based on these explanations it will be decided whether the program behaves as expected or not. Our Project went through two levels of testing

1. Unit testing
2. Integration testing

UNIT TESTING:

Unit Testing is a software testing methodology where individual units or components of a software application are isolated and tested independently. The primary goal is to ensure that each unit functions as intended and meets the specified requirements. During unit testing, test cases are developed to validate the correctness of the smallest testable parts of the software, such as functions, methods, or modules. This process helps identify and rectify defects early in the development cycle, ensuring that each unit operates effectively before integration with other components

INTEGRATION TESTING:

Integration Testing, on the other hand, is a testing approach that focuses on evaluating the interactions and interfaces between different units or components of a

software system. The primary objective is to detect any issues that may arise when these units are combined and to ensure the smooth flow of data between them. Integration testing is critical for uncovering defects related to the integration points, data exchange, and communication between various parts of the system. It helps validate that the integrated components work together seamlessly, providing a more comprehensive assessment of the software's overall functionality.

CONCLUSION

CONCLUSION:

Most business specialists and data scientist concur that that sales forecasting should be a joint exertion. All around the best people to perform such activities are those most solidly included with the association's business works out.Contribution fuses direct relationship with customers, yet additionally an attention to economic situations. Counting key staff individuals from generation, stock administration and advertising advances a soul of cooperation and improves your capacity to make projections.

The expenses of data storage and storage security has made it feasible for little and medium sized organizations to deal with anticipating inside. Regardless, there are also various associations that you can contract with to support you. Discover them by securing referrals from your companions or checking trade disseminations. This study can help

- To minimize changeability
- Forecast results are improved
- Frameworks and methodologies are inter- related
- Improves client administration
- Decreases lead time
- Better information of customers
- Better control of inventory
- Enables firms to respond more rapidly to changing economic situations.

Python language is used to create simulator to compare the results and to check the accuracy of the prediction system. **As Regression Metrics is more efficient to use. T**he Mean Squared Error (MSE), R Squared (R²) and Root Mean Squared Error (RMSE) are calculated**. Auto Regression Moving Average and Exponential Smoothing algorithms in a Hybrid approach are giving the good results.**

REFERENCES

1) https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data

2) Jamal Fattah , Latifa Ezzine , Zineb Aman, "Forecasting of demand using ARIMA model", International Journal of Engineering Management, 7 June 2018.

3) Maobin Li, Shouwen Ji and Gang Liu," Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA", Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model",Mathematical Problems in Engineering, Volume 2018.

4) Bohdan M. Pavlyshenko,"Machine-Learning Models for Sales Time Series Forecasting",MDPI, January 2019

5) NavanshuKhare,SaadYunusSait,"Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models",International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018

6) Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).

7) Dermala N., Agrawal A.N., Credit card fraud detection using SVM and Reduction of false alarms, International Journal of Innovations in Engineering and Technology (IJIET) 7(2) (2016).

8) Bozarth CB and Handfield RB. Introduction to operationsand supply chain management, 4th ed. Raleigh: North CarolinaState University, 2016.

9) El Bahi Y, Ezzine L, Aman Z, et al.Modeling and forecastingof fuel selling price using time series approach: case study.In: Proceedings of the international conference on controldecision and information technologies, Thessaloniki, Greece,10–13 April 2018. IEEE.

10) Zineb Aman ,Haj El Moussami, and Abdeslam Lachhab,Forecasting of demand usingARIMA model, International Journal of EngineeringBusinessManagementVolume 10: 1–9, 2018,

11) Papacharalampous, G.; Tyralis, H.; Koutsoyiannis, D. Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. Water Resour. Manag. 2018, 32, 5207–5239

APPENDIX A

SCREENSHOTS

5.2 EXPERIMENTAL RESULTS

To find new approaches, it's important to seek out errors in previous approaches. The errors found within the previous approaches are given within the following Table 5.5, we've used a hybrid approach.

Model	Validation Error	Out–of-sample error
Extra Tree	14.6%	13.9%
ARIMA	13.8%	11.4%
Random Forest	13.6%	11.9%
Lasso	13.4%	11.5%
Neutral Network	13.6%	11.3%
Stacking	12.6%	10.2%

Table 6: Forecasting errors of different models

5.3 SALES GRAPHS

5.3.1 Sales Graphs for Item code:119624

Sales for January Month
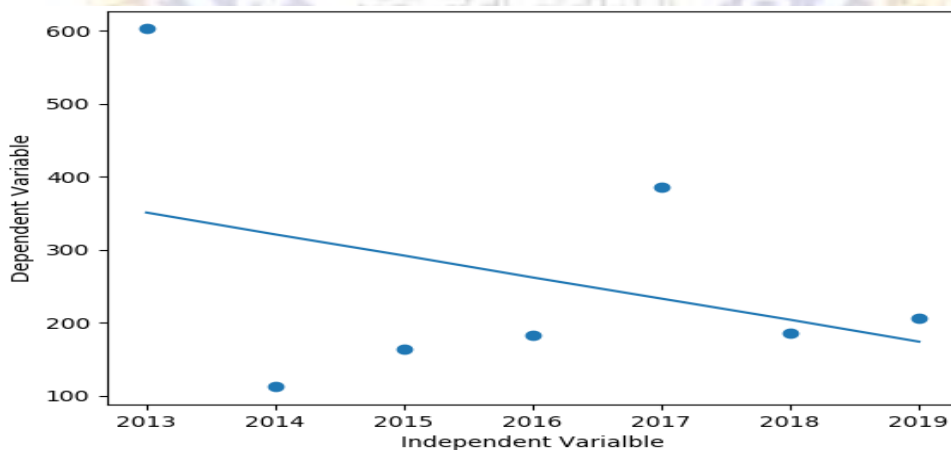
Figure 21 Sales for January Month for Item code 119624

Comparison of actual results with predicted values
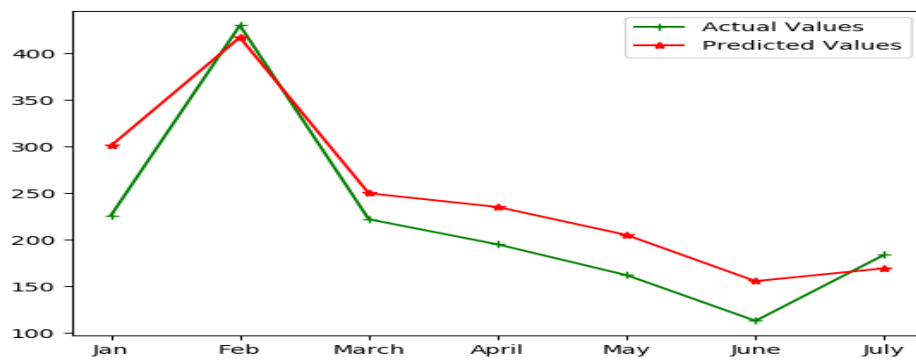
Figure 22 Comparison of results for Item code 119624

Monthly Sales
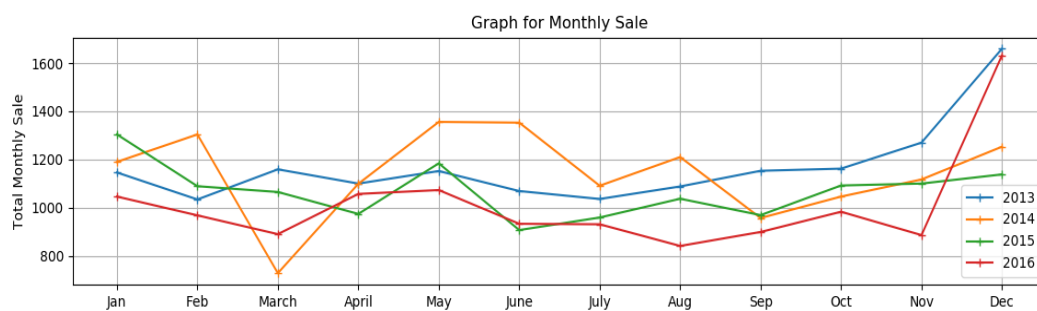
Figure 23. Monthly Sales for Item code 119624

Annual Sales
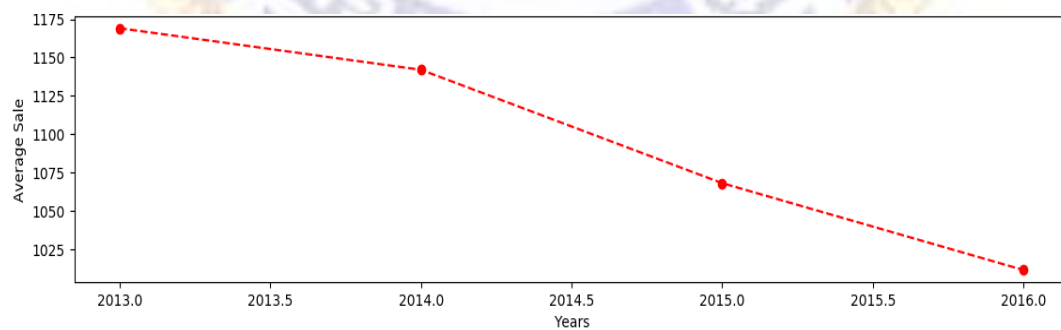
Figure 24 Annual Sales for Item code 119624

5.3.2 Sales Graph Item code:122425
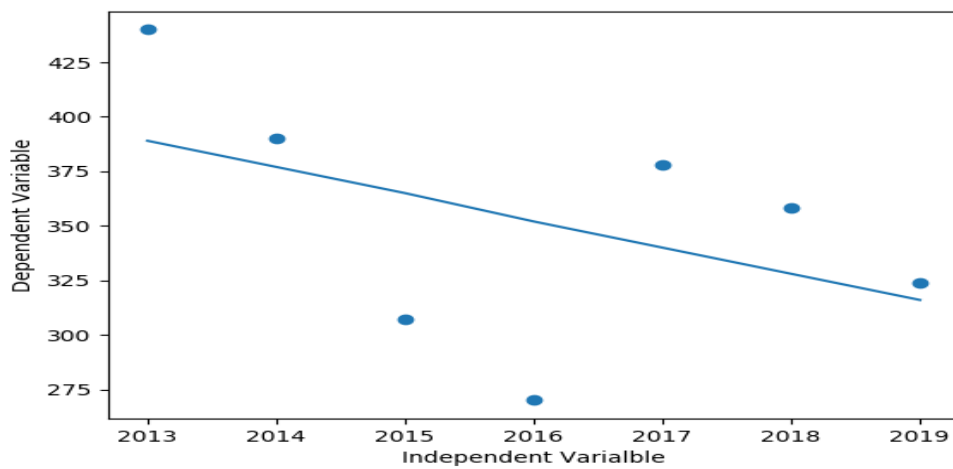
Figure 25 Sales for January Month for Item code 122425
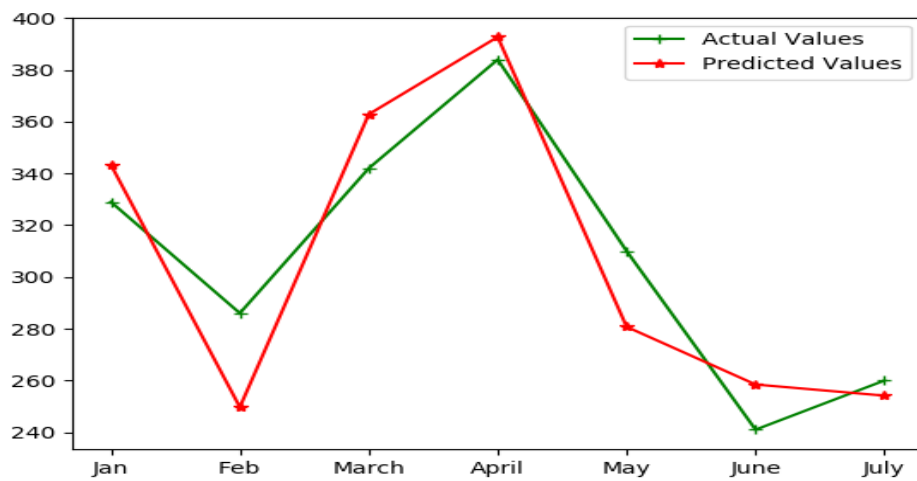
Figure 26 Sales for January Month for Item code 122425

Monthly Sales
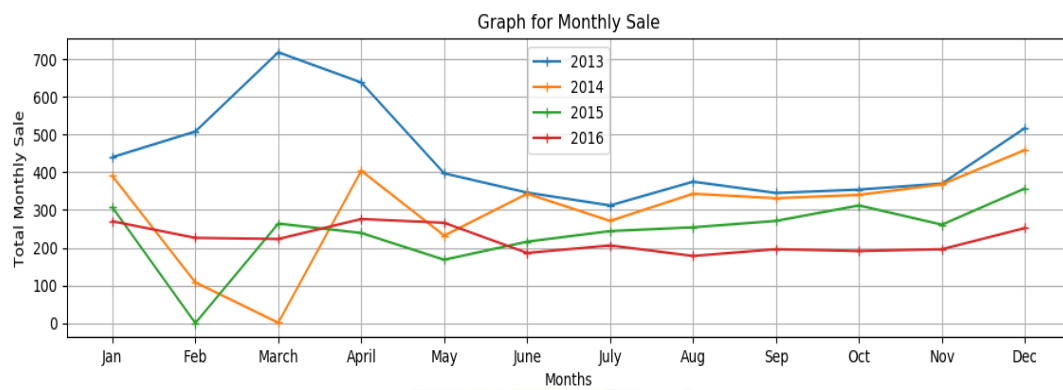
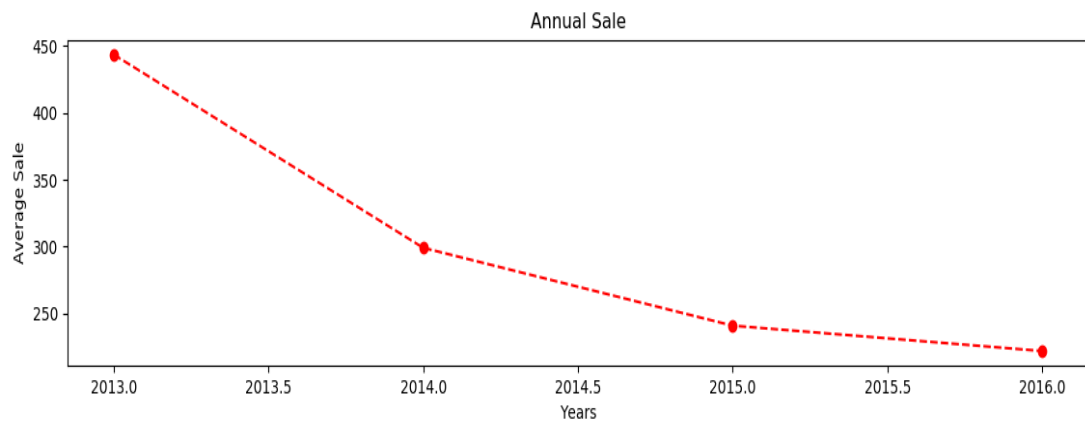**Figure 27 Comparison of results for Item code 122425**
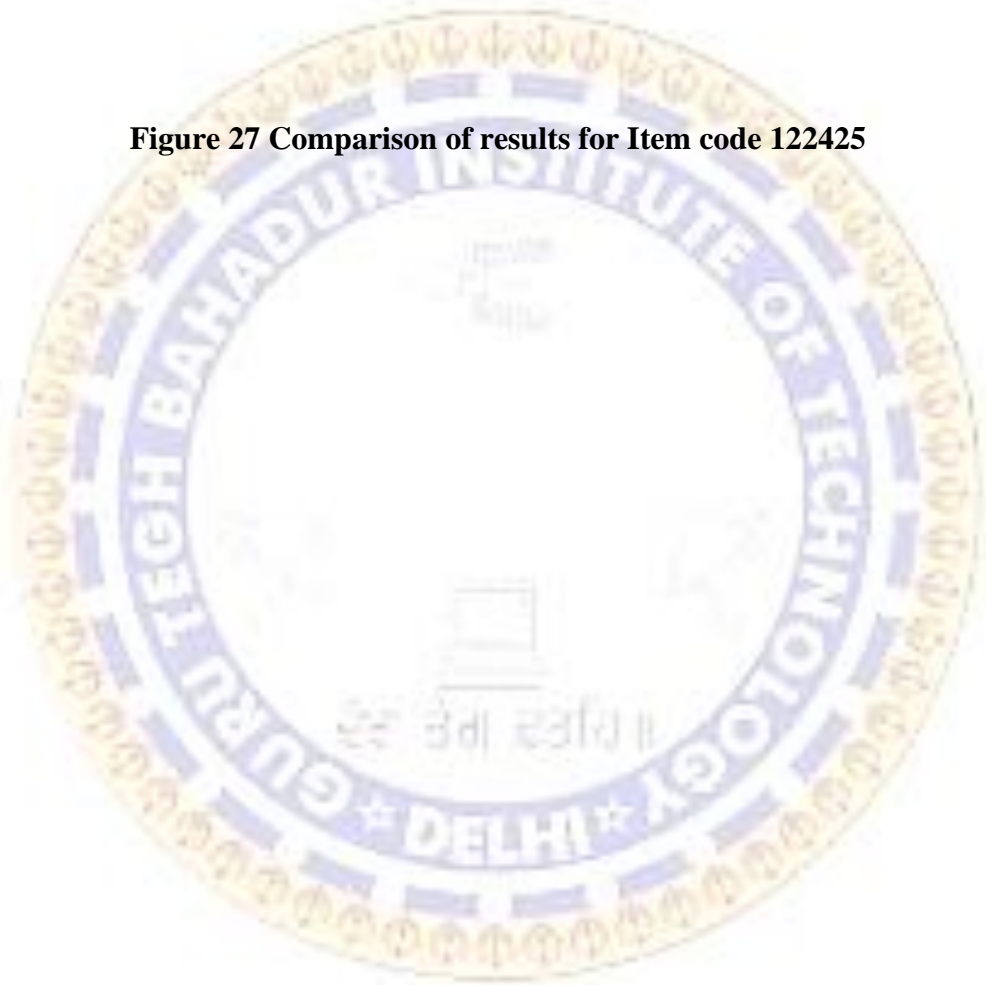
Annual Sales

Figure :28 Sales graphs for item code 122425

5.3.3 Sales Graphs for Item code:153267
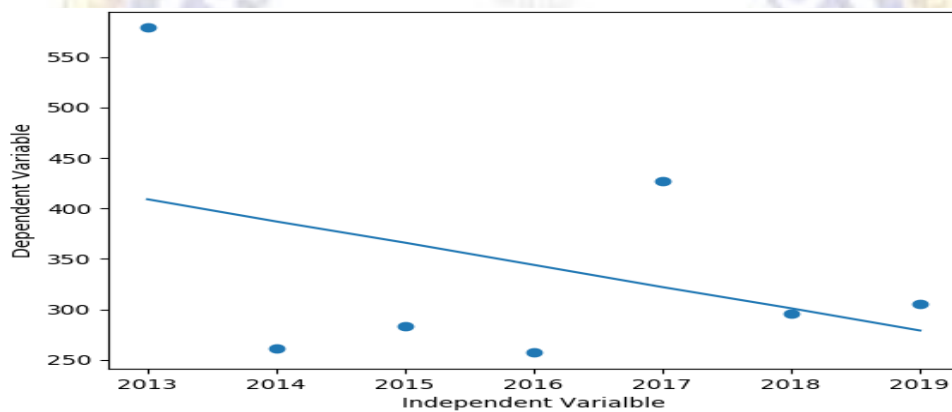
Figure: 29 Sales Graphs for Item code:153267
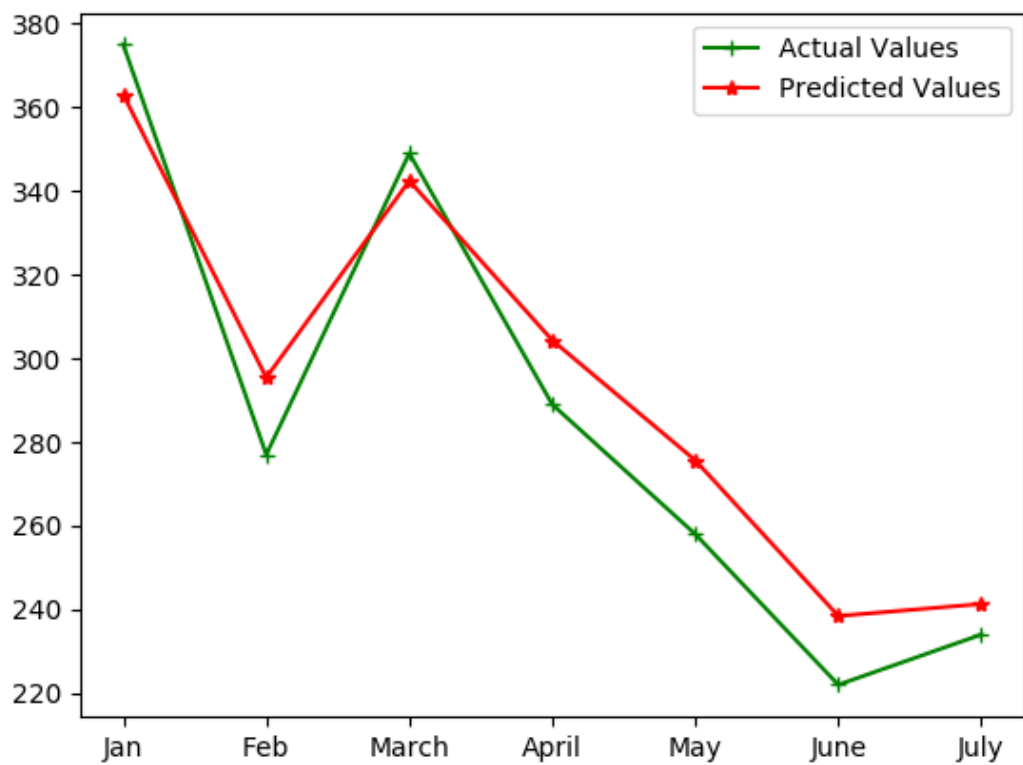
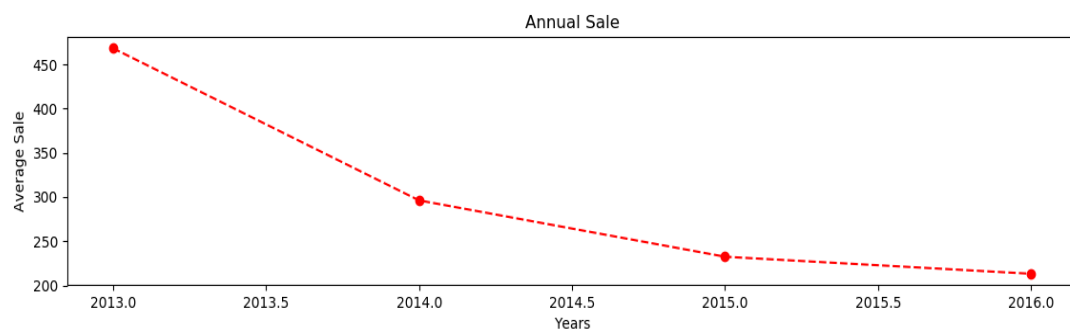**Figure : 30 Sales graphs for item code 153267**

Monthly Sales
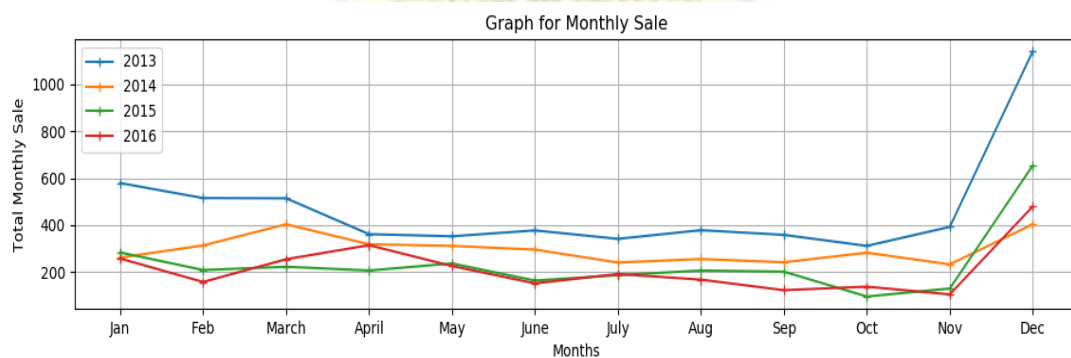
Figure : 31 Sales graphs for item code 153267

Annual Sales

Annual Sale

Figure : 32 Sales graphs for item code 153267

5.3.4 Sales Graphs for Item code:956011
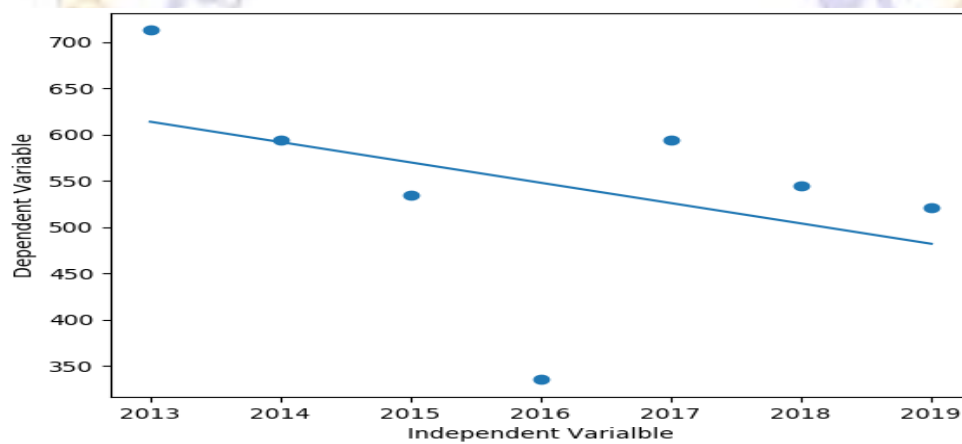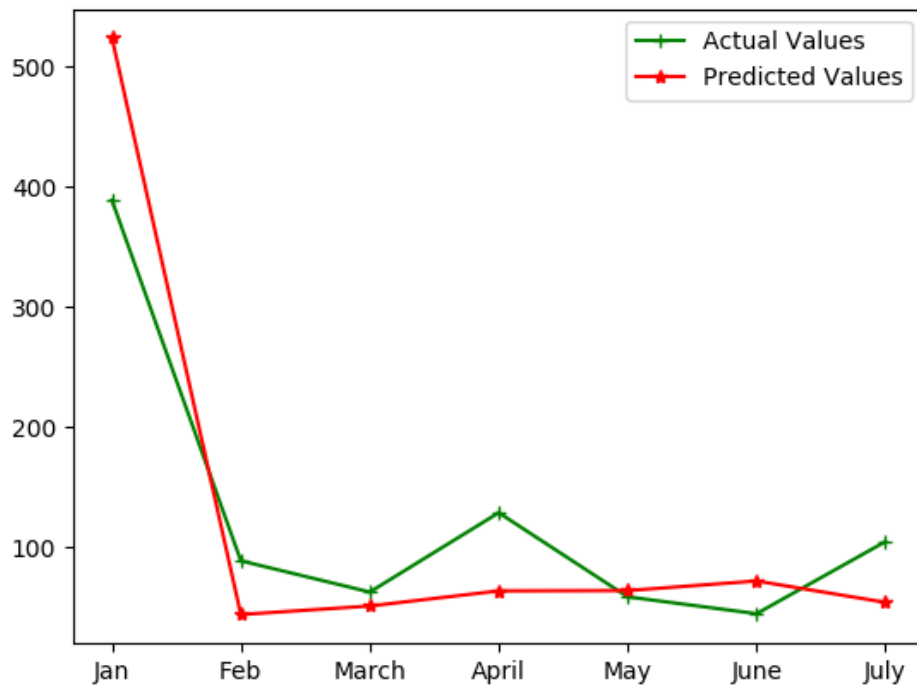

Figure : 33 Sales graphs for item code 956011

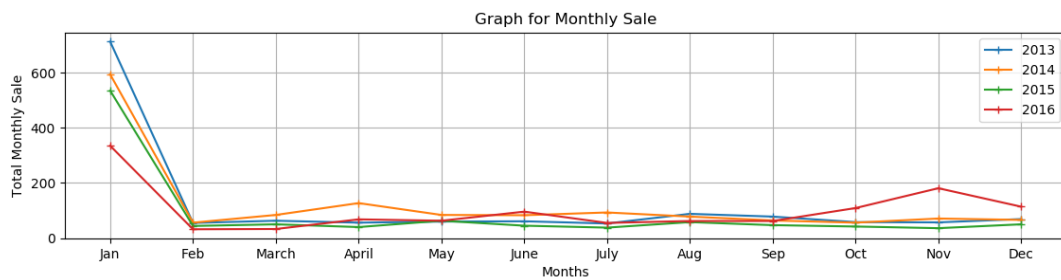**Figure : 34 Sales graphs for item code 956011**

Monthly Sales

Figure : 35 Sales graphs for item code 956011
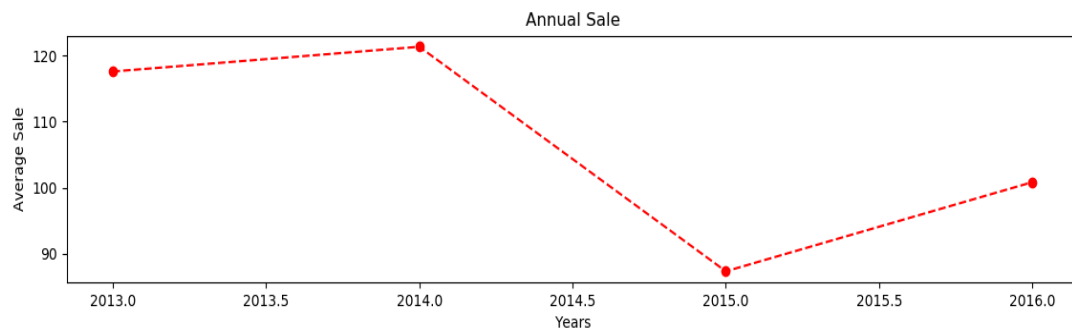
Annual Sales

Annual Sale

Figure : 36 Sales graphs for item code 956011
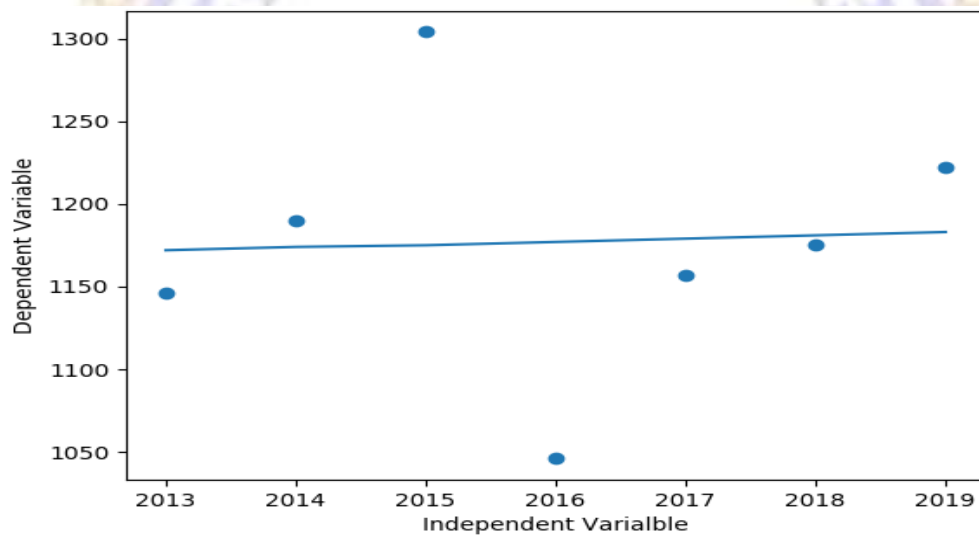
5.3.5 Sales Graphs for Item code:956014
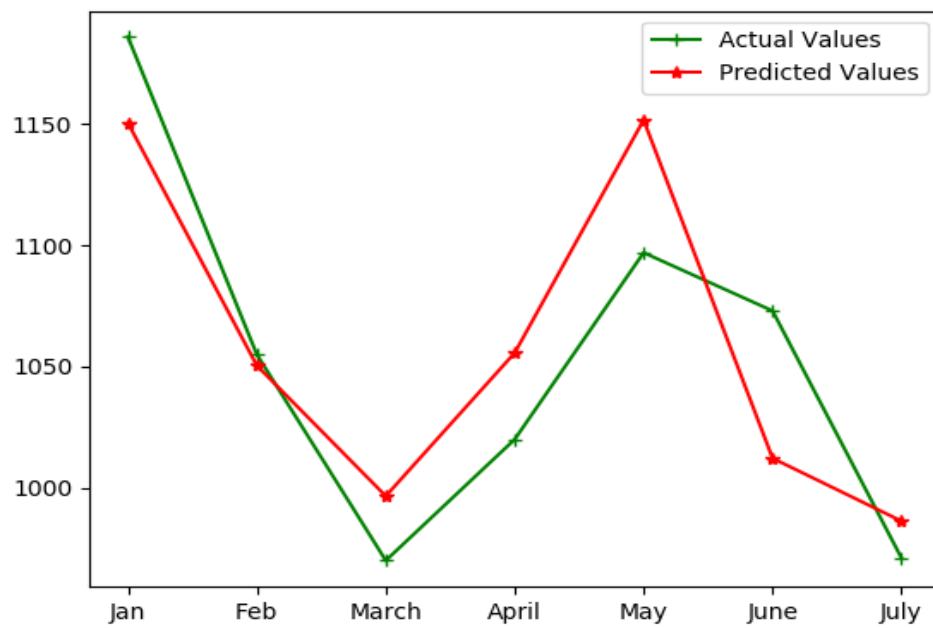

Figure : 37 Sales graphs for item code 956014

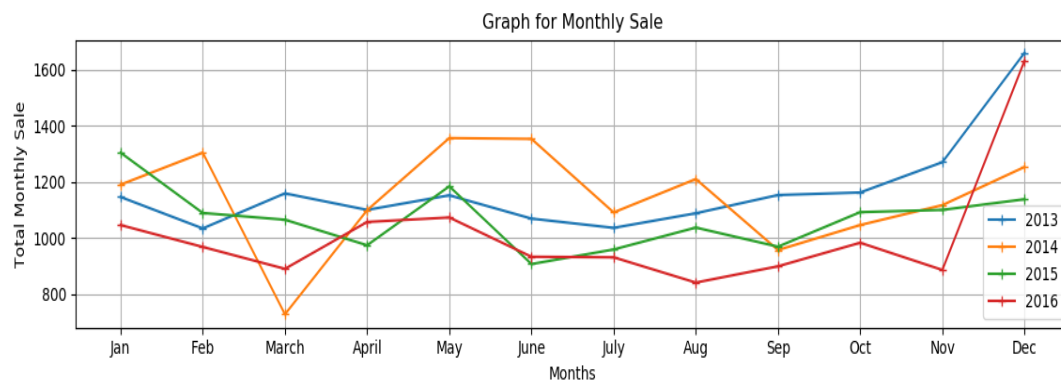**Figure : 38 Sales graphs for item code 956014**

Monthly Sales
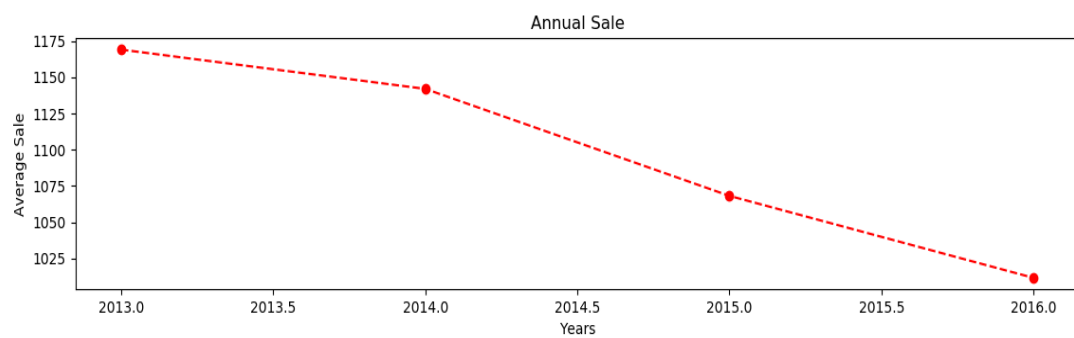

Figure : 39 Sales graphs for item code 956014

Annual Sales

Figure : 40 Sales graphs for item code 956014

APPENDIX B
SOURCE CODE

```python
from tkinter import *
from tkinter import messagebox

import myfilenew
import monthly_repo
import sevenmonthcom

def repo():
    try:

        print(item_id.get(),mon.get(),years.get(
))
        i=int(item_id.get())
        if i==956014:
            f="data2.csv"
        elif i==956013:
            f="data3.csv"
        elif i==956012:
            f="data4.csv"
        elif i==956011:
            f="data5.csv"
        elif i==119624:
            f="data6.csv"
        elif i==123601:
            f="data7.csv"
        elif i==153267:
            f="data8.csv"
        elif i==122425:
            f="data9.csv"
        elif i==153239:
            f="data10.csv"
        elif i==111397:
            f="data11.csv"
        if (int(years.get())<=2016) or
(int(years.get())>2023):

messagebox.showerror(title="Error",messa
ge="Please Enter correct values years
range should be(2017-
2023)",parent=window)
        else:

myfilenew.predict_data(f,mon.get(),int(y
ears.get()))
    except:

messagebox.showerror(title="Error",messa
ge="Please Enter correct values year's
range should be(2017-2021)\nPlease Enter
correct Month First Letter Should be
Capital",parent=window)
def clear():
    item_id.set('')
    mon.set('')
```

[xcv]

```python
        years.set('')

def monrepo():
    try:
        print(item_id.get())
        i=int(item_id.get())
        if i==956014:
            f="data2.csv"
            monthly_repo.monthly_data(f)
        elif i==956013:
            f="data3.csv"
            monthly_repo.monthly_data(f)
        elif i==956012:
            f="data4.csv"
            monthly_repo.monthly_data(f)
        elif i==956011:
            f="data5.csv"
            monthly_repo.monthly_data(f)
        elif i==119624:
            f="data6.csv"
            monthly_repo.monthly_data(f)
        elif i==123601:
            f="data7.csv"
        elif i==153267:
            f="data8.csv"
            monthly_repo.monthly_data(f)
        elif i==122425:
            f="data9.csv"
            monthly_repo.monthly_data(f)
        elif i==153239:
            f="data10.csv"
            monthly_repo.monthly_data(f)
        elif i==111397:
            f="data11.csv"
            monthly_repo.monthly_data(f)
        else:
messagebox.showerror(title="Error",message="Please Enter correct Item Code",parent=window)

        #print(f)

    except:

messagebox.showerror(title="Error",message="Please Enter correct Item Code",parent=window)


def accuracy():
    try:
```

```python
            print(item_id.get())
            i=int(item_id.get())
            if i==956014:
                f="data2.csv"
                sevenmonthcom.testing_data(f)
            elif i==956013:
                f="data3.csv"
                sevenmonthcom.testing_data(f)
            elif i==956012:
                f="data4.csv"
                sevenmonthcom.testing_data(f)
            elif i==956011:
                f="data5.csv"
                sevenmonthcom.testing_data(f)
            elif i==119624:
                f="data6.csv"
                sevenmonthcom.testing_data(f)
            elif i==123601:
                f="data7.csv"
                sevenmonthcom.testing_data(f)
            elif i==153267:
                f="data8.csv"
                sevenmonthcom.testing_data(f)
            elif i==122425:
                f="data9.csv"
                sevenmonthcom.testing_data(f)
            elif i==153239:
                f="data10.csv"
                sevenmonthcom.testing_data(f)
            elif i==111397:
                f="data11.csv"
                sevenmonthcom.testing_data(f)
            else:
messagebox.showerror(title="Error",message="Please Enter correct Item Code",parent=window)

        except:
messagebox.showerror(title="Error",message="Please Enter correct Item Code",parent=window)


window=Tk()

item_id=StringVar()
mon=StringVar()
years=StringVar()
```

[xcvii]

```python
window.title("My App")
window.geometry("800x400+700+200")

ptitle=Label(window,text='SALES
FORECASTINGT USING MACHINE
LEARNING',fg="red",bg="white")
ptitle.place(x=50,y=30)


l0=Label(window,text="Item
Code",width=20,font=("bold",10),anchor='
w')
l0.place(x=80,y=100)
e0=Entry(window,textvariable=item_id)
e0.place(x=150,y=100)
l1=Label(window,text="Month",width=20,fo
nt=("bold",10),anchor='w').place(x=80,y=
130)
e1=Entry(window,textvariable=mon).place(
x=150,y=130)
l2=Label(window,text="Year",width=20,fon
t=("bold",10),anchor='w').place(x=80,y=1
60)
e2=Entry(window,textvariable=years).plac
e(x=150,y=160)




b4=Button(window,text="Generate Forcast
Report",command=repo).place(x=50,y=190)
b5=Button(window,text="Clear",command=cl
ear).place(x=220,y=190)

b6=Button(window,text="Predicting
Accuracy",command=accuracy).place(x=300,
y=190)
b7=Button(window,text="Monthly
Report",command=monrepo).place(x=420,y=1
90)




window.mainloop()
```