Product Demand Forecasting using Machine Learning Models

Purva Badve,Radha Krishna Boddu,Soumith Reddy Manyam,Varun Jaiswal

San Jose State University

DATA-270: Data Analytics Process

Dr. Eduardo Chan

October 11, 2022

1. Introduction

1.1 Project Background and Execute Summary

Based on what (Demand Forecasting for Inventory Control, 2014, p. 1), Eli Whitney began using assembly lines in 1797 in order to create muskets in large quantities for the United States government. All musket components were manufactured with the exact engineering tolerances, allowing them to be suitable for use in any musket. In this manner,the many sections and components all had the same appearance. (Vasudev, n.d.) When it comes to business planning and decision-making, every firm must face the constant challenge of change. For the company to meet its needs, some forecast is essential; the more reliable the forecasts, the greater the value of the outcomes for planning and decision-making.Once commercial manufacturing of the products became feasible and available for purchase by the general public, it became essential to establish sales projections in order to choose how many products to produce within the short-term planning perspective. Forecasting demand is the estimation of future demand and avoiding immediate threats to the sales and reputation of the company for a product based on the currently available data.

Management has battled with forecasting for years, but recent developments in computer technology has made it possible to implement complicated forecasting approaches relatively quickly. (Carbonneau et al,2007) indicates that in the constantly growing supply chain management sector, a precise forecasting system is crucial because it enables businesses to handle volatility in demand for their goods and resources effectively. Every company should have as its primary objective to keep inventory costs as low as possible while still keeping a level of stock that is high enough to meet demand from customers.

Managers in the manufacturing department of the organization must estimate consumer demand for the products to determine the number of raw materials, workforce, and money required to meet the demand. The company must plan and schedule how they will employ these resources before the customers place an order for the goods. Forecasting is necessary for the inventory control systems at places that hold inventory, such as stores, dealers, and distribution centers. In order to ensure that they have the correct number on hand to meet their customers' needs, the company must precisely forecast the demand for each product in advance.

The requirement for more storage space, the stock's depreciation, and the items' expiration will all result in losses for the company if there are too many things in the inventory. The company will experience a loss in sales if there is an insufficient inventory, which will then hurt brand perception. Accurate predictions are therefore necessary in order for the company to be able to respond to changes in demand and support organizational growth.

Understanding the data would be the first stage in our project because the accuracy of the demand estimate is highly dependent on the data's quality. Typically, this data must be cleaned, checked for gaps or other anomalies, confirmed to be accurate, and then restored. We create, clean, and relevance-check the data before organizing it into a comprehensive form. The business goal, the type, volume, quality of the data, and the forecasting period all affect the choice of machine learning models.

The following methods will be used to create the ML model and compare each model's outputs:

a. XGBoost

b. PROPHET

c. Linear Regression

d. ARIMA

e. SARIMA

f. Random Forest

PROPHET - Prophet is designed with business forecasting duties in mind. Prophet uses a model that accounts for trends that are non linear with cyclic periods that occur annually, monthly, daily, on weekends, and during holidays. It works well with historical data that spans numerous seasons and strongly seasonal time series.

XGBoost - (Chen & Guestrin, 2016a) The decision-tree-based ensemble deep learning algorithm XGBoost employs a gradient-boosting architecture. Gradient boosting which is a learning model based on integrating combination of estimates from a number of poorer and easier models in order to accurately predict a target variable. The primary benefit of XGBoost is that it can operate the same code in a large distributed environment and solve N number of issues.

Linear Regression - A statistical technique called linear regression uses historical data to forecast future values. It can help with handling issues with inflated costs and spotting underlying tendencies.

ARIMA - ARIMA forecasts future values based on historical data. The foundation of ARIMA is the idea that history repeats itself. To plan a short-term forecast, ARIMA can offer accurate projections. It clearly illustrates the results for demand, sales, planning, and manufacturing by providing expected numbers for user-specified timeframes.

SARIMA - A subgroup of linear regression models called ARIMA models makes an effort to predict the future values using the previous observations of the variable. A seasonal autoregressive integrated moving average is different from ARIMA in terms of the seasonal variations and trends. It makes it possible to record periodic properties.

Random Forest - Random forest is a more sophisticated technique for making numerous decision trees and combining them. The random forest model averages all individual decision tree estimations to provide more precise projections.

The project's goal is to develop a web application that lets customers enter a certain brand of alcohol and get an estimate of their sales projection and how many bottles they should buy and stock up in their inventory.

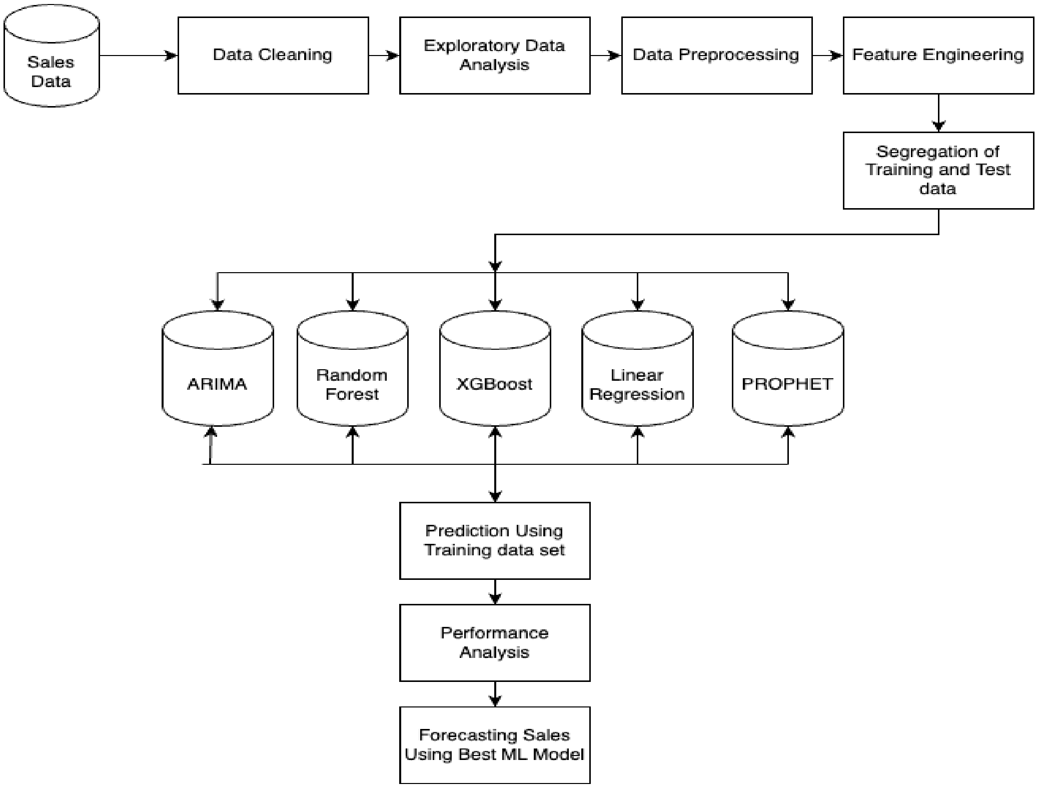
Additionally, we want to create a Visualization Dashboard that displays the dataset's major attributes as well as a trend in alcohol sales.

1.2 Project Approaches and Methods

We will undertake exploratory data analysis after data collection, preprocess the data, provide some straightforward visualizations, search for outliers, and impute missing values using the best method at our disposal. After careful examination, we will choose the features that are most relevant for our machine learning models. The next step is to train five machine learning models to a high model accuracy of more than 80%. Finding the ideal model and predicting sales are our goals.

Figure 1

Proposed Demand Forecasting Architecture


Sales Data
Data Cleaning
Exploratory Data Analysis
Data Preprocessing
Feature Engineering
Segregation of Training and Test data
ARIMA
Random Forest
XGBoost
Linear Regression
PROPHET
Prediction Using Training data set
Performance Analysis
Forecasting Sales Using Best ML Model

1.3 Project Requirements

1.3.1 Functional Requirements

In the section below, the overall functioning of the proposed system is discussed. The system must be capable of estimating demand and sales at the level of the retail establishment and ought to be able to recognize the name of a particular alcoholic beverage brand. If the user enters an inaccurate brand name, the system should be able to notify them. The AWS cloud's S3 buckets are where the data is initially kept. The data is then prepared for the data warehouse by connecting the AWS Glue to the data source. The advantage of AWS Glue is that it allows us to manage centralized data more effectively, and the ETL pipeline may be created, managed, and monitored visually. To store our 24.6 million records, we will use the AWS Redshift data warehouse. Due to its size, we cannot maintain and analyze our dataset locally on the System. For this reason, we shall use the AWS cloud's capabilities. We will create a VPC, or virtual private cloud, to authenticate only trusted users to keep our network secure. Using data from the Redshift data warehouse, Google Collab will be used to train our machine learning model. The micro-web framework Flask will link our trained model to a website we will create using React framework. The user can enter the data to get the sales forecast for that particular alcoholic beverage brand. To better understand the sales prediction, we also intend to create a visualization dashboard. In order to do this, we will link our Redshift data warehouse with Tableau, a visualization program..

1.3.2 AI-Powered Requirements

When using machine learning techniques, a sizable amount of data is required to train a trustworthy model. Our 26.4 million record dataset will be split into test and train data. We will dedicate a sizable portion of our data to training, and the remaining data will be used as test data to verify the accuracy of our model. Jira, a project management and issue tracking solution, will be used for project ma. This will help us in efficiently planning, managing, and tracking our work. The development and evaluation of models will be tracked using the version control software GIT. After each model is built, its accuracy will be evaluated using a variety of techniques, including the recall score, confusion matrix and ROC AUC. Five models—the ARIMA model, XGBoost, Random Forest Classifier, PROPHET, Linear Regression —will be constructed. We will try comparing every model and draw conclusions on which one is the best to employ.

1.3.3 Data Requirements

We got the dataset for the project from Iowa state’s website. We are also trying to use the publicly available data which lists the products and dates of each spirit purchase made by liquor license holders in IOWA from the month of January of 2012, to the present. The collection contains sales information with 24.6M entries and 24 columns . It includes information such as the number of units sold, an invoice, a vendor number, a sale, the name and location of the retailer etc.

1.4 Project Deliverables

The following deliverables would be implemented as part of the project. Each deliverable has a distinct estimation point.

1.4.1 Project Proposal

Description.: After having discussed a variety of cases and issue statements before settling on the one we felt we could best address. This would benefit a significant portion of society rather than just one particular sector. We used a tool called conceptboard to document our tasks break down.

Due Date.: 1 week

1.4.2 Project Breakdown

Description.: JIRA, a tool for project management, will be used. This will help us define our tasks clearly and create a timeframe that is appropriate for them. To break down our project into a number of stages, we will use a Work Breakdown Structure. Then, to keep track of the project and effort timeline, we will utilize Gantt and Pert charts. We will monitor the development and log any bugs we identify during testing or development..

Due Date.: 1 week

1.4.3 Data Collection

Description.: Our data would be gathered from the Iowa state's website. We will use AWS cloud to store and clean our data because it is too large—our dataset has 26.4 Million rows.

Due Date.: 1 week

1.4.4 Project Plan

Description.: The project is broken down into specific stages as follows. Understanding the problem would be the first step.The next step is to prepare the data. Building models is the third step.Training models is the fourth step.Model testing is the fifth step.Evaluation is the sixth step.

Due Date.: 4 weeks

1.4.5 Prototype

Description.: We will develop a single proposed system for which we want to finish forecasting using each of the five models. Specifically, the ARIMA, XGBoost, PROPHET, Linear Regression, and Random Forest Classifier models..

Due Date.: 7 week

1.4.6 Testing

Description.: As we test the model, we will take into account a range of inputs and analyze its overall performance. These evaluations might include curves, performance indicators, and possible instances of incorrect forecasts.

Due Date.: 2 Weeks

1.4.7 Final Developed Application

Description.: A handy website connecting Flask to our trained models will be the end outcome. The above mentioned website will project demand for the brand using data collected from consumers. Additionally, our project will feature an interactive dashboard that enables users to research the brand's patterns and examine its many aspects..

Due Date.: 10 week

1.4.8 Project Report

Description.: The project report would outline the project's goals, the steps taken to accomplish them, and a description of the challenges faced and the outcomes. It would present the topic's findings and give an overview.

Due Date.: 8 weeks

Table 1

Project Timeline

Deliverable	Description	Due Date
Project Proposal	Product Demand Forecasting using Machine Learning	09/19/2022
Project Breakdown Structure	Planning and work breakdown using Jira, Gantt and Pert chart	09/30/2022
Data Collection	Collected from IOWA state	09/30/2022

	website and processed in AWS	
Project Plan	Model training of 5 ML models	10/06/2022
Project Prototype	Model first demo	11/14/2022
Testing	Using various methods like confusion matrix, parameter tuning etc.	11/23/2022
Final Model	This includes model-website integration and Tableau dashboard	11/30/2022
Project Report	This includes all the documented records of all the findings.	12/06/2022

1.5 Technology and Solution Survey

1.5.1 Time Series Forecasting

Letham (2017) says that the analyst-in-the-loop technique involves calculated past forecasts to approximate out-of-sample effectiveness and find poor forecasts such that a professional analyst can discover the problem and correct the model. A common practice, to make forecasting on a global level while making the best utilization of both people and machines labor.To get started, we constructed the time series that used an adaptable configuration with parameter readings that are easy to understand by humans. After that, we present this model's predictions in addition to a set of feasible baselines based on various historical simulation timeframes.Whenever there is low performance or any aspects of the predictions that call for human participation, we communicate these potential issues in a prioritized manner to the appropriate person and signal them as potential concerns. After that, the analyst can examine the prediction and determine, based on the feedback received, whether or not to make any changes to the models.

So, based on the research, we can draw the conclusion that we use a straightforward, modular regression model, which frequently performs admirably with values that are default and allows analysts to easily pick the features which are relevant to the prediction problem and do any necessary adjustments. The last element is a platform for tracking and monitoring the accuracy of forecasts and trailing projections that need to be manually assessed to help analysts make small adjustments. The last element is very important since it aids analysts in determining when the model needs to be modified or when an alternative model could be more appropriate. Large-scale forecasting, as it is known, is the process of many analysts forecasting a huge diversity of time series using easy, adjustable models and flexible performance measurement.

Backtesting Approach: The term "backtesting" is used in the field of time-series forecasting to determine the accuracy of a forecasting method through the utilization of existing factual and historical knowledge for the purpose of comparison. Typically, the technique is repetitive and spread out over the course of a time that encompasses many of the periods in the historical data. Since the purpose of backtesting is to estimate the expected future correctness of a prediction approach, it is an excellent method for determining which prediction models should be considered the one that is most accurate.

The backtesting process starts with the choice of a range of threshold time frames from a period that the past data cover. After the records are cut off at each threshold. The prediction model is developed and tested based on the cut-off data, and finally, the predictions are contrasted to the original, not truncated data. Due to the fact that an unique forecasting model must be constructed for each threshold, backtesting typically consumes a substantial amount of computer resources. As a direct result, we usually see practitioners who train the prediction model only once, typically utilizing the complete scope of past datasets, and then move forward to the back-testing iterations. Back-testing commonly profits from this method because it is expected to considerably speed up the procedure.

(Kechyn, 2018) says “WaveNet is a generative model”. The Wavenet approach confirms that the prototype may produce sequences of real data in response to specific conditional parameters. The core concept behind wavenet design is dilated causal convolutions. Upsampled filters are used instead of convolution layers in dilated convolutions to address these issues. In the other words, dilated convolutions let you keep the dimension of the feature layouts from one layer to the next by only growing the field of sight of the kernel. Otherwise, you can get a broad perspective of the input with lower dimensionality.

(Dipti Srinivasan et al., 1995) Fuzzy neural networks (FNN) is a combination fuzzy neural technique that brings together parts of neural network designing with parts "of fuzzy logic and fuzzy set theory". Its goal is to predict potential electricity usage. The said combination of methods is based on a simple idea: the fuzzy knowledge-base models what is known about the system and its input variables, both quantitatively and qualitatively, and the neural network designs the connection between fuzzy inputs and outputs is hard to explain. This combination approach uses the strengths of both fuzzy systems and neural networks. For example, neural networks can generalize, degrade gracefully, find information from partial data, and learn from well-defined patterns. On the other hand, fuzzy systems can reason abstractly and act like humans when there is uncertainty or conflicting data. A clever way has been found to incorporate fuzzy logic with nns. Thus we can conclude from this paper that the FNN is a better approach in comparison to ANN whenever we want to predict anything that spans over the weekends and national holidays.

1.5.2 Comparison of Existing Techniques

(*Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting*, n.d.) Previously people have developed many techniques to determine what a supermarket will sell but all of them have some pros and cons. With the help of this paper they have tried to compare multiple ML Algorithms like "Multiple Linear Regression Algorithm, the Random Forest Regression Algorithm, the K-NN Algorithm, the Support Vector Machine (SVM) Algorithm, and the Extra Tree Regression", to develop forecasting algorithms that precisely predict sales. Predicting likely sales is determined by various factors, including past sales data, promotional programs, the week of a holiday, the weather, the gas price, the Consumer Price Index (CPI), and the labor force participation rate in the state.In an 80:20 split, the dataset was splitted into Training Samples and Test Samples. Following that, data pre-processing methodologies were used to prepare the data to feed into the blueprints. After choosing the features, the model was then trained. Test data were then used as input, and the features were measured.After feeding the test data into the trained models , the results of different ML models were compared. They used the Sklearn package to build the model. The authors chose characteristics and goals to plot on the X and Y axes so they could predict sales. Using an 80:20 ratio to split the data into sets for testing and training purposes. After that, they had to compare the Forecasting Models. They used Random Forest Regressor, Linear Regression, SVR,KNN Regressor, and Extra Tree Regressor along the Y-axis, to guess how much money the Walmart shop would make. The MAE is the sum of all error values. According to the findings of the study, the scenario of Support Vector Regression has the highest MAE, while the scenario of Extra Tree Regression has the lowest MAE. Extra Tree Regression has a cost, but it is substantially less than the cost of the other regression techniques. We may claim that it has been shown that the Extra Tree Regression Technique is the most accurate way for predicting future sales at Walmart Store, followed by the Random Forest Regression Technique, based on the data and the graphs utilized by the authors to form a conclusion. We can also draw that the SVMs and Linear Regressions models do not perform well for the demand forecasting.

If management is seeking to estimate sales for a short period of time and just has data from the last few years, regression techniques that just fulfill some requirements for constructing forecasting models is not one of the best choices.

1.6 Literature Survey of Existing Research

Applying random forest and GBM learning approaches (Islam, 2020) examined the prediction of likely back-order scenarios in the supply chain. Predictions based on machine learning algorithms are unsuitable for numerous commercial decision-making procedures due to their lack of precision and adaptability. Inaccurate predictions may result from using inaccurate data as inputs in the prediction system. By employing machine learning models to predict product back-orders, we seek to increase procedure transparency, give decision-makers more freedom and preserve a higher level of precision in the business domain of prediction. To cover the range of traits of real-time data that may emerge from a machine or user mistake, using a ranging method, many levels of predictive characteristics can be set.. The range is variable, providing decision-makers with flexibility. Tree-based deep learning is utilized to improve the model explanation. Distributed Random Forest (DRF) & Gradient Boosting Machines are utilized in this work to predict product back-orders (GBM). When the dataset is significantly skewed with random defects, they have seen a 20% increase in ML model accuracy using this range method. Four-level measurements reflect the levels of sales, inventory, forecast sales and lead time. Constructed model's decision tree is explored to appreciate the impacts of the ranging technique. As part of the research, they give the most likely back-order situations to assist with decisions related to business. They also demonstrate how this algorithm can estimate potential back-order items before their sale. The methodology developed in their study can be extended to diverse supply chain settings to forecast back-orders.

(Alves, n.d.) Inventory management is essential to hospital logistics since it guarantees that all drugs and emergency aid are delivered to clients on time. Currently, healthcare facilities manage this in an imprecise manner, not always employing industry standards and ignoring statistical or mathematical models that aid in forecasting seasonal variations in demand. With the use of machine learning algorithms, predictions may be made with greater precision, and statistical procedures that are more exhaustive can yield higher results. The Knowlogis project involves creating a dashboard that will enable the health board team's decision-making and serve as a valuable tool for running a hospital. This set-up will include a layer that predicts material requirements, anticipating issues that a person may not have anticipated. Accurate inventory forecasting enhances the efficiency of inventory management systems by greatly decreasing the risk of both surplus stock as well as the expenses connected to it, as well as a lack of stock. We look into how useful machine learning techniques are in prediction of hospital supply requirements. We evaluate them against standard procedures by analyzing how demand has changed over time at a large urban, multi-specialty facility in Portugal.

(van Steenbergen & Mes, 2020) It is considered that quantitative methodologies for the latest product forecasting mostly concentrates on comparable estimates to compensate for the lack of demand history. This indicates that these models generate predictions for forthcoming products based on data from similar products. This section focuses mostly on these quantitatively equivalent forecasting methodologies. The essential principle of analogous prediction is that resemblance between products results in the same demand pattern. There is no assurance that the demand for similar products in the past will transfer into a yearning for brand-new merchandise down the road; hence, it should be approached with caution.Diffusion models are frequently used to simulate the adoption of novel goods or emerging technologies, such as the adoption of electric vehicles in Europe. When used to anticipate demand for new products at the SKU level, these algorithms may result in significant errors. Presenting a model for early forecasting is Neelamegham and Chintagunta. This model was developed to predict the number of people who will watch movies during their opening week by using factors such as genre and the presence of movie stars. The model correctly predicted the nation of a given film. With a Bayesian approach, the level of uncertainty in the forecasts was measured. Thomassey and Fiordaliso developed a forecasting method for the garment industry's mid-term forecasting. A K-means algorithm is used first to classify the demand patterns of products already on the market into distinct profiles. These algorithms may produce considerable inaccuracies when forecasting demand for new items at the level of SKU. Neelamegham and Chintagunta present a model for early forecasting.

This model was built to predict the number of individuals who will watch films during their debut week by taking into consideration factors such as genre and the appearance of starlets. The notion of a specific film was accurately predicted by the model.

The degree of forecast uncertainty was assessed using a Bayesian methodology. Thomassey and Fiordaliso created a forecasting technique for the mid-term forecasting of the apparel sector. Using the K-means algorithm, the demand patterns for readily-available products are initially segmented into unique profiles. The authors used a probabilistic neural network to predict the demand profile. But in terms of performance, this model did not outperform the Naive Bayes classification. An ELM which is known as an extreme learning machine was used to predict clothing sales, and it outperformed earlier neural network-based methods. Investigations were done into the relationship between sales and other product characteristics like color, price, and size. The prediction technique proposed by the author compared the early deals of newish products to those of comparable items already on the markets.

As a result, this plan lacked a pre-launch forecast and instead relied on the early sales as data. Modifications were done to the time series to illustrate comparable trends at a different scale. This work also incorporates a combination model based on Random Forest and Two-step Clustering to boost prediction accuracy. Initially, data series are grouped into several unrelated groups using the two-step clustering process. Each cluster is then assigned as the input and output sets to construct the matching C-RF model. The testing set is then divided into the appropriate collection using the trained Two-step Clustering model, the prediction results are derived based on the C-RF model.

This paper proposes a hybrid forecast model based on Two-step Clustering and

Random Forest that considers demand-related characteristics. Data is separated into several groups using the Two-Step Clustering Algorithm. After that, the relevant C-RF models are built for various clusters using Random Forest. They have evaluated the performance of the C-RF model that was proposed against the single Random Forest. The experiment demonstrates that the C-RF outperforms the single model, and the demand forecast's precision has enhanced due to clustering. The e-commerce company has to train different forecasting models for various commodities. Clustering and other machine-learning approaches should be investigated and used in forecast models in the future.

(Mbonyinshuti et al., 2022)Utilized program data generated by e-LMIS, a digital electronic software used in Rwanda to manage medical supplies. It was collected and processed to ensure that the data would be utilized appropriately during the predictive modeling phase. This paper's time series analysis played a vital role in projecting the need for essential drugs to treat NCD. Our analysis gathered data from the health supply chain from 2015 to 2019, focusing on essential drugs used to treat NCDs. Our research centered on consumption statistics that might be used as a foundation for interest in forecasting based on various inventory control data contained in the dataset.

The Rwandan pharmaceutical supply chain was used to collect data for the eLMIS on NCD essential medication use.. The district level collection of all data is also permitted by the eLMIS system, allowing managerial level access for data of the pharmaceutical supply chain. In our study, district-level data were gathered and centrally synthesized. The dataset under review had roughly 500 products used in health institutions set by the government, most of which were essential drugs. The experimental results revealed that the RF model with precision of R-square is 0.78 on the training dataset and 0.71 on the testing dataset had the least error value.

The experiment focused on emphasizing the data that resulted from an estimate of needs in the future for critical NCD drugs based on past data on consumption. It has been shown that prototype quality enhances prediction. Given that the discrepancy between the train and test predictions is about 5%, the model is likely universal to other datasets. It is absolutely important to conduct additional research to justify the sustainable use of deep learning applications in various operations in the supply chain in light of this study's findings.

(Žunić et al., 2020) In order to categorize and get the model right, a prototype was made that marked a part of the model as "offline." The second part is an example of the subsidiary process that helps make precise sales forecasts. Data aggregation in the temporal domain at the production level, analysis of data to get rid of meaningless past data (for e.g., it was concluded not to utilize previous data which is greater than 6 years), and date filtering to get rid of irrelevant past data is a part of the analyzing of data and preprocessing process. The same unit can be used to convert between different amounts, such as pieces, packs, bundles, pallets, and so on (i.e., the conversion of daily sales data into monthly sales). The most important part of the suggested architecture for classifying product portfolios and predicting sales is a tool or method for making shipshape forecasts of time series from Prophet, a Facebook app.

The Prophet tool's usage in replicating how product sales in a portfolio change over time without adding any more regressors. Estimating monthly and quarterly revenues for the brand portfolio is the ultimate aim of modeling. Using PE, monthly or quarterly projections are made, and MAPE is used to figure out how accurate the forecasts are.

An expanding window backtesting approach is employed in calculating the desired level of accuracy in forecasting with a reliable estimate. It is possible to simulate historical settings by changing the present time to any moment in the past. As a result, it is possible to create a model using historical data, estimate forthcoming sales, and evaluate the model.

The framework that was suggested demonstrated the ability to generate near accurate quarterly and monthly sales forecasts and a lot of ability for classifying the portfolio of product into different groups in compliance with the level desired of forecasting: about half of the portfolio could be forecasted every month with a MAPE of 30% or lower.

(Jha & Pande, n.d.)they are trying to do time series analysis on data related to supermarket sales using the Prophet from FB. They gathered data for analysis, and they used the Additive model and the ARIMA model to do so. Both of these models have demonstrated better capability for future estimation. An ARIMA model was created in Python. The fit() method provides the model with the training data it requires to function properly.The predict() method is used to make predictions. when the dataset has been fully absorbed by the training model. The P>|z column, that highlights the significance of weight for different techniques including autoregressive, moving average, and the coef column, which shows how much each feature weights and how important it is, are the two most significant columns. MAPE ,MSE and RMSE are the three metrics which are used to analyze these models.The Facebook Prophet is a predicting model that is far more accurate. The Facebook Prophet forecasting model is better than other models because it is more accurate, fits better, and has a lower error rate.

(Smirnov & Sudakov, n.d.).They advise using ML algorithms on data indicating demand for a new product acquired from an online store when there is no sales history since the product is new in the market or never been utilized before. The attributes of the product, including category, price,text description and name, are sent to the algorithm as input. To address the regression issue, various gradient boosting methods like XGBoost, LightGBM, and CatBoost were used. To be able to draw solid conclusions about how accurate the suggested prototype was, it was required to choose the appropriate metrics. RMSE was applied during this inquiry. The RMSE is used by both forecasting time-series and conventional regression models to conclude the model's accuracy. It is possible to approach the difficulty of anticipating consumer demand for new products as a regression process. When performing this particular kind of job, the software is obligated to estimate numeric data based on the data that has been provided. The learning algorithm needs to produce the function f Rn->R to achieve this objective. Classification is a similar task to this one, but the output has a different structure. One of the best ways to deal with regression issues is through gradient boosting. Growing is a potent strategy that, by combining the output of several "base classifiers," creates a type of committee. This committee's performance may be significantly better than any of the base classifiers'.

Since it is computationally efficient, a gradient descent-based learning strategy was used. To start, it has been shown that, despite the fact that there are many variables involved, the margin of error reduces as learning progresses. This may indicate that the algorithm is receiving proper training. Furthermore, the inaccuracy on the test dataset hasn't changed much, which shows that the risk of overfitting has been reduced and that the model will perform as expected when applied to new data. Without any market research and regardless of the kind of products or historical data, they could estimate demand correctly.

(Krishna et al., n.d.) Researchers are attempting to predict a retail company's sales using a variety of ML techniques, and they are investigating various algorithms to see which one performs the best. Along with standard regression procedures, they have also included boosting techniques. Data exploration's primary objective is to locate product attributes and shop components that can affect sales as a result of projections constructed utilizing the products and businesses. When examining data, one must first study the dataset to learn more about the data that is present as well as the data that was assumed to be present. Three features are present in the dataset but are not theorized, nine parts were hypothesized but not discovered in the data, and six elements are both posited and found in the data. The data we have gathered contains both categorical and numerical variables, and both sorts of variables can be combined. The researchers typically manage the outliers and incomplete data are found in the dataset during the process known as "data preparation."We consider the mode as a feasible method for completing missing values since we are unable to compute an average for categorical variables. The phase of feature engineering, which follows, tackles the subtleties in the data and creates new variables for our data from existing variables that can be prepared for analysis.

(T. et al., 2018) Using the LabelEncoder as well as OneHotEncoder methods in sklearn preprocessing module, we could turn the category variables we were using into numeric values. Because the ML algorithms only accept numbers as input, this was necessary. The data that has been divided into two subsets with a 4:1 ratio between the two for training and testing has been segregated. In creating models for the goal of making predictions, different regression algorithms, such as “Multiple Regression, Polynomial Regression, Ridge Regression, and Lasso Regression,” have been utilized. In order to improve accuracy, XGBoost was also applied to the dataset also various boosting techniques, like AdaBoost were also applied.

The idea of multiple regression, is useful in forecasting the importantnce of the variable which is dependent using weight of a number of independent variables that act in the same way.

In polynomial regression, dependent and independent variables is classified as nth degree polynomial.This regression analysis is called "nth-degree polynomial regression." It is the illustration of a model which makes use of polynomial regression, which is a simple linear regression with an order of 1. Variable selection and regularization are carried out to improve the prediction precision and the readability of the findings. AdaBoost is an abbreviation for adaptive boosting. This algorithm's primary goal is to improve how the model being created operates. The weighted total of the results from the weak learners is used to create the algorithm's final output. The outcome is then considered to be this total.

It can be concluded that GradientBoost approach has least RMSE ioch equals 1088.64 and the method with the highest RMSE of 350.72 is ADABoost method. Algorithm with the most incredible R2 value of 0.59 is GradientBoost whereas the R2 value of the algorithm AdaBoost is 0.40. The GradientBoost approach is the best predictor for the dataset taken into consideration, as can be seen from the results produced. It has the highest R2 value (0.5), the lowest RMSE value (1088.64), and both.

(Nguyen et al., n.d.) Using historical sales data and seasonal economic considerations, They anticipate industry sales with a seasonal pattern using regression methods and artificial neural network models. Sales in the sector follow a way that changes throughout the year. The range of the short-term predictive model starts at one quarter, while the scope of the long-term predictive model goes up to twenty quarters. When creating the forecasting model, they did not depend solely on examining historical sales data. Instead, their model used historical sales data and economic indicators as predictor variables. As the indicators included in their sample data didn't show a repetitive pattern similar to the industry sales, they first deseasonalized the sales data using the LOESS smoothing approach. A regression method for data smoothing is referred to as "locally weighted scatterplot smoothing" (referred as "LOESS"). Using LOESS, we may assign a neighbor to each distinct point in the dataset. At its center is the neighborhood, consisting of k points on either side of the selected point. The following step is for LOESS to try and fit a quadratic regression curve to the nearest points. Physically closer points of time are considered more than those physically farther away. Regression quantifies a single equation to assess the effects of several independent factors on a single variable (referred to as a dependent variable). In their forecasting model, industry sales are the dependent variable, while economic indicators are the independent variables. The economic connections between the variables are evaluated using regression to determine their strength and direction. They consider time-lag values of economic indicators as a result. Lag values of economic indicators were used in the correlation calculations, indicator selection, and ANN prediction processes. Throughout the entire technique, this was done.

They constructed models for both long-term and short-term prediction Artificial Neural networks (ANNs) were used to create projections after the decision of specific economic variables as model inputs.The job of predicting significant sales based on input parameters that are not themselves seasonal was the most challenging part of this study within the context of the neural network approach. They established their target figures based on the previously provided deseasonalized sales data. In other words, rather than reflecting actual sales, the neural network modifies the economic indicators to reflect the seamless version of those sales. Deseasonalized sales have been the outcome of ANN as a result. They multiplied their projections by the appropriate coefficients to deseasonalize them and obtain forecasts for the sales volume that would take place. The reseasonalization coefficient for each quarter was estimated individually.

(Kohli et al., 2020) The study's main goal is to forecast future sales by analyzing the current sales of various departments within a large superstore, such as the sales, marketing, and financial departments, in order to help those departments increase their profits and build stronger brands in line with the current market trends. This study will also concentrate on increasing customer happiness. The method used to forecast future sales is the linear regression algorithm. The study uses a technique known as linear regression to carry out predictive analytics. The data gathered from many sources is stored in a single document. Following data collection, cleaning processes are carried out on the data, and other forms of noise in the data are also eliminated. The data is also appropriately classified, and any blank spaces or unnecessary data that might be present in the data are removed. Then, using the information that has been processed, a forecast is created using machine learning. Linear regression is used to achieve this objective. Thanks to this essay, we understand two fundamental concepts: Cost Function and Gradient Descent. These two concepts are related to the linear regression method. The search problem is rewritten as a minimization problem so that the difference between the expected answer and the actual answer is relatively less. Gradient descent is the next key concept you must understand to understand linear regression. The phrase "gradient descent" describes a method for lowering the cost function. The idea is to start with a set of values and incrementally change them to reduce costs. How to change the values is determined by the gradient's decline. Their calculations yielded findings with an average accuracy of approximately 84%.

(Zhang et al., 2003) To take advantage of each model's distinct strengths in nonlinear and linear modeling, the authors suggest a hybrid technique that blends ARIMA and ANN models. Studies using real data sets show that putting the two models together can significantly improve the precision of predictions than by models done alone. Conventional statistical models that are linear in nature include moving averages, exponential smoothing, and ARIMA. These models can use the linear function of past values to predict values in future. Over the last few decades, These models include bilinear, threshold auto-regressive (TAR), and auto-regressive heteroscedastic (ARCH) models. Although such nonlinear models have come a long way, they are still not very useful for solving typical forecasting problems. Due to the fact that nonlinearity in time series is not possible to be simulated but are created for nonlinear patterns. In an integrated auto regressive moving average model, it is presumed that the preceding observations and random errors when applied to a linear function gives the future value of the variable. Based on what was done by Yule and World prior to them, Box and Jenkins created a useful technique for building ARIMA models that substantially impacted time-series data analysis and forecasting. It ought to possess a few theoretical aspects of autocorrelation. It’s frequently possible to identify more than one potential model for particular time series by contrasting the empirical patterns of auto-correlation to theoretical ones. Box and Jenkins emphasized using testing data's autocorrelation and selective autocorrelation functions to make out the order of the model, ARIMA. Transformation of data frequently required to keep the time-series data constant during the identification step. It is simple to determine the parameters once a tentative model has been defined.

The parameters are chosen to have the lowest possible overall error rate. In order to achieve this, a nonlinear optimization technique might be applied. Making a diagnostic inspection to determine whether the model is adequate is the final stage in model construction. A model is typically built in three steps and iterated upon till a model is selected that is fit. The final model chosen is utilized to generate the predictions. There are numerous nonlinear structures that can be utilized to explain and predict a time series once the model format is no longer constrained to being linear. An one type of ANN also known as an artificial neural network is a type of model which can show how nonlinear effects within data can be repeated. The fact that ANN models are capable of approximating that can get close to a wide range of functions with high accuracy gives them a significant advantage over other kinds of nonlinear models.Their ability to process data from information simultaneously is what gives them strength. It is not necessary to know the final design of the model before construction. The network model is instead mostly dependent on how the data is set up. In their respective linear or nonlinear domains, the ARIMA model and ANN model have both performed admirably. But none of these is a universal strategy that applies to all circumstances. ANNs should not be used on any type of data at all. A amalgam approach which could handle linear as well as nonlinear modeling can be a good technique to tackle it because it is difficult to probably know all that is possible about the data in a real-world scenario. The method for the hybrid model that is being proposed has two stages: in the first phase, the linear element of the problem is analyzed by the ARIMA model. The second phase involves creating a neural network's model using the ARIMA's residuals. The linear model's residuals will demonstrate why data does not follow a single path, as the ARIMA model is unable to explain why this is the case. The training system based on GRG2 is used to create neural network models in this study, whereas the SAS=ETS system is used for all ARIMA modeling. Only one spot ahead of prediction is considered. The MSE and the MAD are used to gauge a forecast's accuracy . The Findings from three data sets from real-world demonstrate that the amalgam model is superior to every component element. If a hybrid model employs models that are radically unlike one another or that sharply disagree with one another, theoretical and empirical data in the literature suggests that there is less generalization error or variance in the hybrid model. The hybrid approach can also lessen model uncertainty, which is a common problem in time series forecasting and statistical inference. This is due to the possibility of data patterns becoming unstable or evolving over time. Additionally, the overfitting issue, which most often occurs with neural network models, could be resolved readily if the ARIMA model is applied to the data first.

(Akande et al., 2022)On the basis of sales information received from 45 Walmart stores predictions are made using the XGBoost algorithm. The open-source XGBoost package includes an element called XGBRegressor that may be used to predict sales in Python. The authors claim that the two most crucial factors to take into account while employing XGBoost are the execution speed and model consistency. To determine the model's correctness, As a result, after comparing various metrics and determining their valthe outcomes demonstrated that XGBoost did a respectable job of formulating predictions. As a result, the authors state in their conclusion that to predict sales, the XGBoost learning technique can be used, which may assist managers in determining the prices to be set for goods.

(Gumus & Kiran, 2017) XGBoost is a gradient boosting model. The authors try to find the factors which determine the elements affecting prices of the crude oil using XGBoost, and help in creating the estimates using those parameters. For the training data, the supervised learning issues to predict a target variation with a variety of attributes is primarily done by XGBoost. The cross-validation feature of XGBoost is also used In this instance, the original sample is split into n-folds of equal-sized subsamples. The cross-validation procedure is repeated n times, using accurate validation data from every nfold subsamples. The study report comes to the conclusion that component-wise gradient boosting, which incorporates autonomous selection of variables during the fitting process, increases the level of boosting.

(Zhang et al., 2021)To provide a precise forecast of the volumes of in-store sales in the upcoming years, the researchers first did a thorough analysis of time series for the retail industry volume of sales. They then fine tuned information with the help of feature engineering. The model is currently being improved to be more accurate by adding additional information, such as the recent weather and temperature. Two distinct datasets for different entity stores make up the source data. The time based data in the source needs to be processed. Still, the model is meant to forecast the number of sales over a specified period. The model's objective is to forecast sales volume for a specific time period. The period was divided into fifteen-minute segments, the sales volume was calculated by adding the order value for each segment, and the order volume was calculated by adding the total orders. The three features were used to differentiate the timestamp, gross merchandise volume (shortened GMV), and order volume. They compared the outcomes after testing the model's efficacy using order data. Three quarters which is equal to 75% is used as a training data set for the prototype and the last 25% is used as the testing dataset to measure how well the system works. These two dataset components are arranged in chronological order. The authors compare the performance of XGBoost with that of several different ML algorithms, such as the GBDT,LSTM,ARIMA algorithm and Prophet. The author's investigation demonstrates why, when it concerns forecasting the time series of transactional difficulties, the XGBoost technique is better compared to other methods. The GRBT, XGBoost, and other boosting methods require fewer data and features than deep learning approaches. They plan to anticipate the sales volume for the forthcoming period based on the data from the past. Despite having access to fewer features and data, the XGBoost models and GBDT perform better compared to others. XGBoost, In contrary to GBDT, includes a regularization element in objective function to manage complexity of model, avoid overfitting, and enhance generalizability.

(Bekal & Bari, 2021) The two fundamental problems with demand forecasting

are the cannibalization of current products and forecasting for a long time. Their study suggests a three-stage approach based on XGBoost to address the problems of product cannibalization and long-term error propagation.The proposed three-stage XGBoost-based framework's performance is compared to that of the standard XGBoost algorithm and is found to outperform. A tabular data format was used. The three-stage framework is used to train and predict each unique category of a product. The number of units sold is a constant real value. Three Datasets that are distinct were considered for the experiment, each from a different class of items. A tally shows how many units were sold during a specific week for each product category. In addition, the number of items in a certain category can vary from week to week. Depending on the categorical total, the model changed its forecasts for all products during stage 2 by raising or decreasing them. The final step is to guide each product's prediction within the category. The objective function is divided into two terms at stage 3. The sum limitation and the first word are mutually inclusive. This ensures that the total projections for each category and all devices remain consistent with the domain-specific category actuals. The second goal function term makes sure that, in accordance with the information obtained from stage 1, sales prediction for device sales grows or decreases as necessary. Our architecture's third and final stage, stage 3, verifies that the final predictions made for each device add up to the unconditional sum determined using domain knowledge. The input features also affect how the predictions generated at the level of each particular device are altered.

According to the authors, the studies show that the three-stage framework is consistently more effective at making long-term forecasts. They used a weighted accuracy metric to assess the performance differences between the two systems. The researchers discovered that, after putting the suggested framework into practice, the overall accuracy of their forecasts for outdated products dramatically increased. It improved to 67%, which was 38% in the baseline model

References

Akande, Y. F., Idowu, J., Misra, A., Misra, S., Akande, O. N., & Ahuja, R. (2022). Application of XGBoost Algorithm for Sales Forecasting Using Walmart Dataset. *Lecture Notes in Electrical Engineering*, 147–159. https://doi.org/10.1007/978-981-19-1111-8_13

Alves, C. M. F. (n.d.). Demand forecasting in a multi-specialty hospital setting: a comparative study of machine learning and classical statistical methods. *FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO*. https://repositorio-aberto.up.pt/bitstream/10216/114091/2/277665.pdf

Bekal, G., & Bari, M. (2021, November 25). An XGBoost-Based Forecasting Framework for Product Cannibalization. *Arxiv*. https://arxiv.org/pdf/2111.12680.pdf

Carbonneau, R., Vahidov, R., & Laframboise, K. (2007, October 1). Machine Learning-Based Demand Forecasting in Supply Chains. *International Journal of Intelligent Information Technologies*, *3*(4), 40–57. https://doi.org/10.4018/jiit.2007100103

Chen, T., & Guestrin, C. (2016a, August 13). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Comparative Analysis of Supervised Machine Learning Techniques *for Sales Forecasting*. (n.d.). https://thesai.org/Downloads/Volume12No11/Paper_12-Comparative_Analysis_of_Supervised_Machine_Learning_Techniques.pdf

Demand Forecasting for Inventory Control. (2014, December 4). Springer Publishing.

Dipti Srinivasan, Chang, C., & Liew, A. (1995, November). Demand forecasting using fuzzy neural computation, with special emphasis on weekend and public holiday forecasting. *IEEE Transactions on Power Systems*, *10*(4), 1897–1903. https://doi.org/10.1109/59.476055

Gumus, M., & Kiran, M. S. (2017, October). Crude oil price forecasting using XGBoost. 2017 International Conference on Computer Science and Engineering (UBMK). https://doi.org/10.1109/ubmk.2017.8093500

Islam, S. (2020, August 26). *Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques - Journal of Big Data*. SpringerOpen. Retrieved September 26, 2022, from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00345-2

Jha, B. K., & Pande, S. (n.d.). Time Series Forecasting Model for Supermarket Sales using FB-Prophet. *Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)*. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9418033

Kechyn, G. (2018, March 11). *Sales forecasting using WaveNet within the framework of the Kaggle. . .* arXiv.org. Retrieved September 26, 2022, from https://arxiv.org/abs/1803.04037

Kohli, S., Godwin, G. T., & Urolagin, S. (2020, July 26). Sales Prediction Using Linear and KNN Regression. *Algorithms for Intelligent Systems*, 321–329. https://doi.org/10.1007/978-981-15-5243-4_29

Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. (n.d.). Sales-forecasting of Retail Stores using Machine Learning Techniques. *Sales-Forecasting of Retail Stores Using Machine Learning Techniques*. https://sci-hub.se/10.1109/CSITSS.2018.8768765

Letham, B. (2017, September 27). Forecasting at scale. *Facebook, Menlo Park, California, United States*. https://doi.org/10.7287/peerj.preprints.3190v2

Mbonyinshuti, F., Nkurunziza, J., Niyobuhungiro, J., & Kayitare, E. (2022). Application of random forest model to predict the demand of essential medicines for noncommunicable diseases management in public health facilities. *Pan African Medical Journal*, *42*. https://doi.org/10.11604/pamj.2022.42.89.3383

Nguyen, G. H., Kedia, J., Snyder, R., Pasteur, R. D., & Wooster, R. (n.d.). Sales Forecasting Using Regression and Artificial Neural Networks. *Conference: Midstates Conference for Undergraduate Research in Computer Science and Mathematics*. https://www.researchgate.net/publication/280742365_Sales_Forecasting_Using_Regression_and_Artificial_Neural_Networks

Smirnov, P. S., & Sudakov, V. A. (n.d.). Forecasting newproduct demand using machine learning. *Journal of Physics: Conference Series*. https://iopscience.iop.org/article/10.1088/1742-6596/1925/1/012033

T., G., Choudhary, R., & Prasad, S. (2018, December). Prediction of Sales Value in Online shopping using Linear Regression. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. https://doi.org/10.1109/ccaa.2018.8777620

van Steenbergen, R., & Mes, M. (2020, December). Forecasting demand profiles of new products. *Decision Support Systems*, *139*, 113401. https://doi.org/10.1016/j.dss.2020.113401

Vasudev, S. R. (n.d.). Demand forecasting using statistical and machine learning algorithms. *Dublin Business School*, 9. https://esource.dbs.ie/bitstream/handle/10788/3714/msc_vasudev_s_2019.pdf?sequence=1&isAllowed=y

Zhang, G. (2003, January). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/s0925-2312(01)00702-0

Zhang, L., Bian, W., Qu, W., Tuo, L., & Wang, Y. (2021, April 1). Time series forecast of sales volume based on XGBoost. *Journal of Physics: Conference Series*, *1873*(1), 012067. https://doi.org/10.1088/1742-6596/1873/1/012067

Žunić, E., Korjenić, K., Hodžić, K., & Đonko, D. (2020, April 30). Application of Facebook's Prophet Algorithm for Successful Sales Forecasting Based on Real-world Data. *International Journal of Computer Science and Information Technology*, *12*(2), 23–36. https://doi.org/10.5121/ijcsit.2020.12203