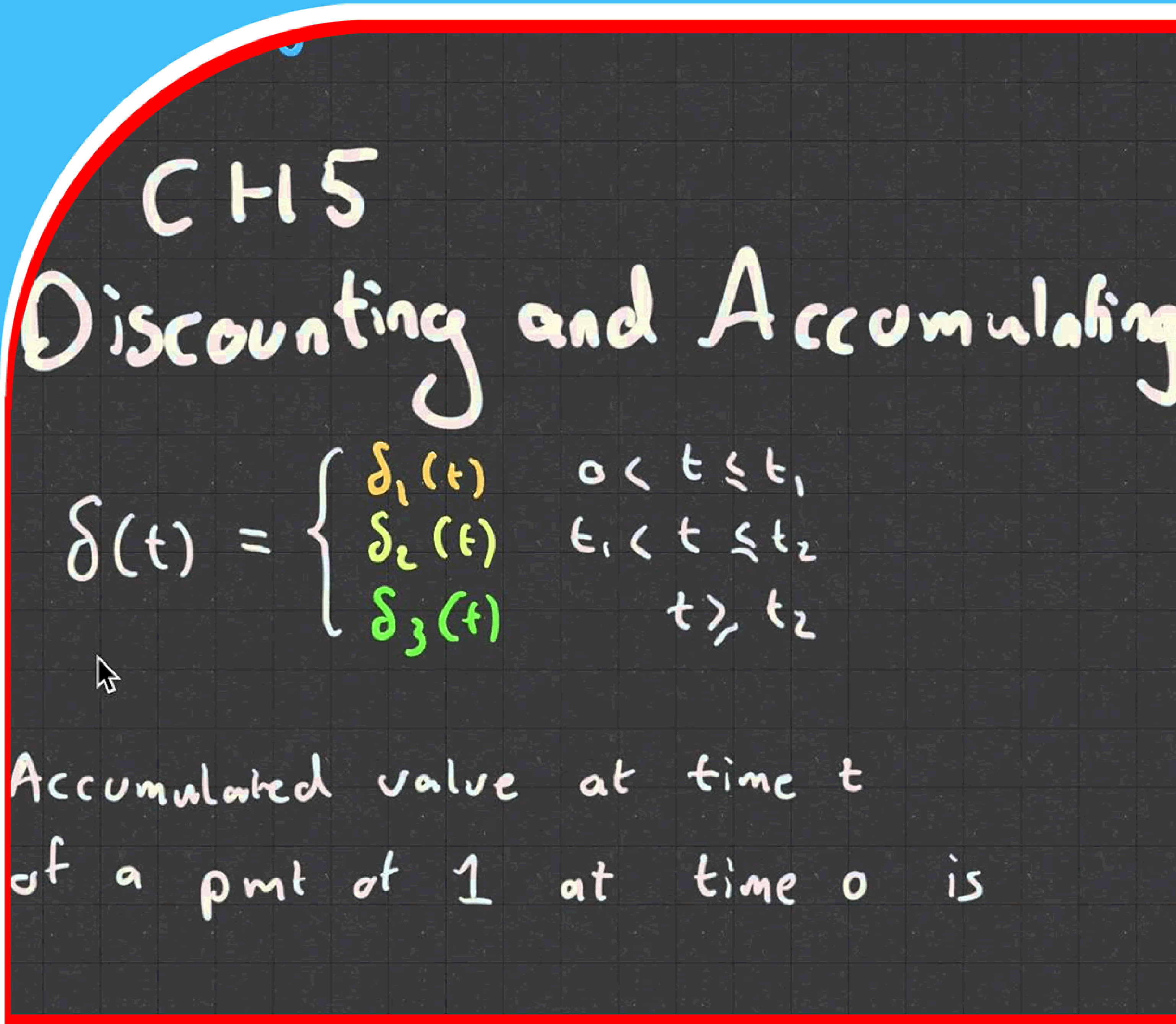
American Journal of
Statistics and Actuarial Science
(AJSAS)


CH5
Discounting and Accumulating
δ(t) = δ₁(t) 0 < t ≤ t₁
δ₂(t) t₁ < t ≤ t₂
δ₃(t) t ≥ t₂
Accumulated value at time t
of a pmt of 1 at time 0 is

Forecasting Retail Sales using Machine Learning Models

Oluwasola Oluwaseun Mustapha, Dr. Terry Sithole

AJPO

American Journal of Statistics and Actuarial Science

ISSN 2958-5244(Online)
Vol.6, Issue 1, pp 35–67, 2025

www.ajpojournals.org

Forecasting Retail Sales using Machine Learning Models

Oluwasola Oluwaseun Mustapha[1*], Dr. Terry Sithole[2]

Article history

Submitted 17.02.2025 Revised Version Received 16.03.2025 Accepted 18.04.2025

Abstract

Purpose: This paper's main objective is to examine common machine learning techniques and also time series analysis for sales forecasting in a bid to get the best fitted technique and give more logical hypotheses for raising future profit margins while obtaining historical in-depth understanding of prior demand utilising business intelligence software's like Tableau or Microsoft Power BI. The outcomes are laid forth with regards to dependability as well as precision of the various forecasting models that were employed.

Materials and Methods: In this project, a sales prediction is carried out on a 5 year store-item sales data for 50 different items in 10 different stores with a dataset obtained from Kaggle. This study focuses on using Machine Learning Methods including the Random Forest, Gradient Boosting Regression (XGBoost), Linear Regression and also the standard time series Autoregressive Integrated Moving Average (ARIMA) method were analysed and contrasted to measure the methods' effectiveness for prediction of Sales.

Findings: This study demonstrates the potential of machine learning algorithms in accurately forecasting sales, which can be extremely valuable for businesses in optimizing their operations, inventory management, and financial planning. By leveraging these predictive models, companies can make data-driven decisions to improve efficiency, reduce costs, and increase profitability. The findings also highlight the importance of selecting the most appropriate algorithm for a given dataset and problem, as well as the need for proper model tuning and validation to ensure reliable results. Furthermore, the study underscores the significance of understanding and interpreting error metrics like RMSE and MAE to effectively evaluate and compare model performance.

Unique Contribution to Theory, Practice and Policy: Factors such as Seasonality, Trend, Promotional offers and Randomity have been known to be important factors that affect the outcome of Sales Forecasting which is why the performances of the Mean Absolute Error (MAE), the Absolute error (R2) and the Root Mean Square Error (RMSE) are all compared in the different algorithms used, to help identify the best preferred algorithm to be adopted which turned out to be the XGBoost method.

Keywords: *Sales forecasting, Machine learning, Time series analysis, Random Forest, ARIMA, LSTM, XGBoost, Linear regression, RMSE, and MAE.*

https://doi.org/10.47672/ajsas.2679

INTRODUCTION

During the early years, a lot of businesses would usually embark on the production of goods without taking into account the sales numbers resulting from demand for such products. Competing in a market without taking into consideration the usage of Data about the demand for same or similar items in such a market in order to evaluate whether to expand or reduce quantity produced puts any manufacturer at the risk of losing out. Bajari, Nekipelov, Ryan, and Yang mentions that varying specifications are used by various businesses in gauging sales. Techniques that previously existed either fall short of capturing context-specific, erratic trends or failing to incorporate all available data in the face of a data shortage issue.

For every business to be successful, forecasting the right demand at each outlet is essential because it properly aids in management of stock, improves the distribution of products among stores, reduces excessive stocking and understocking at individual stores, which essentially, enhances profits and client happiness while bringing loss to a minimum [Berry and Linoff (2004)]. Given the limited shelf lives of many products in a lot of businesses, which results in income losses in both shortage and surplus scenarios, the importance of sales forecasting cannot be overemphasized [Jain, Menon, and Chandra (2015)].

Computer systems that can assist in making quality decisions by offering projections of future sales should be used in supporting the forecasting of sales as a result of the fact that human contingencies such as illnesses and deaths can usually cause scarcity of Manager who would usually make imprecise sales predictions on several occasions [Tsoumakas (2019a)]. Optionally, making use of machine learning algorithms in the creation of precise sales forecasting techniques by utilizing the abundance of available sales data is a better and easier process, because it is unbiased and adaptive to any data changes unlike what is attainable with human attributes of a Sales manager. Also a machine learning algorithm surpasses the imperfections of a human expert's accuracy.

How can we collect globally accepted data from sales (wholesale or retail) that a computer can analyse? Every time a sales transaction occurs (mostly between buyers and sellers) in a market, it is best practice to document the transaction (agreed-upon) price at the time. In a bid to estimate future sales and demand, several businesses have traditionally centred on analytical models like the linear regression, time series and the random forest models.

A set of samples taken over time in a chronological order is called a time series. Time series can be used to represent many different data-sets type, together with a regular sequence of the number of items dispatched from a facility, a weekly series of the number of traffic accidents, daily rainfall totals, regular measurements of the outcome of a chemical process, and so forth. Time series examples abound in a variety of disciplines, including finance, engineering, geology, astronomy, and sociology. One key inherent characteristic of time series is that of its dependent neighboring observations. It is highly practical to understand how observations in a time series interact with one another. Strategies for this dependence's study are the focus of time series analysis while also critically observing seasonality, irregularities and patterns. [Box, Jenkins, Reinsel, and Ljung (2015)]

Linear Regression is one basic important Statistical Modeling tool that represents regression functions as a collection of predictors. It is widely used because it offers a good approximation to underlying regression functions in the case of small sample size. It is safely referred to as the fundamental foundation of all contemporary modeling tools [Su, Yan, and Tsai (2012)]. In machine learning, feature engineering involves creating quantitative patterns of significant systems based on the subject matter expertise, allowing better analysis of data and an additional beneficial viewpoint [Li, Ma, and Xin (2017)]. Random Forest is a more advanced approach which allows for the merging of several trees to produce judgments. The random forest model eliminates the average of all individual tree decision predictions to produce projections that are more precise [Boyapati and Mummidi (2020)].

Regression rather than time series analysis is the better approach for predicting sales. Regression algorithms can frequently provide us with better outcomes than time series methods, as experience has shown. The time series can be examined for trends using machine learning algorithms [Pavlyshenko (2019)].

Research Questions

i. What limitations are evident in previous retail sales revenue projections research works?

ii. What are the most frequent and important characteristics or elements that affect the forecasting of retail sales?

iii. Which algorithms are most effective in solving issues with product sales forecasting?

iv. How are sales volume or market projections improved by machine learning algorithms over the classical statistical approach?

Research Objective(s)

Majorly, the ultimate objective of this study is to employ machine learning and conventional algorithms in determination of all categories of goods that sell more at a given business outlet, which will assist the shop owners to make better choices about stock procurement. In a bid to find out which machine learning algorithms perform the best on the provided data, a comparison analysis of four distinct classification machine learning algorithms would be conducted. Application of a conventional technique will also be covered in the further sections of this study for simple comparison purposes.

Why Forecast?

As forecasting involves the future, it greatly benefits a lot of sectors. For families, it equips the major provider of the home with proper planning based on past spending known as Budgeting. When projections are reliable and accurate, enterprises in the private sector can have significant confidence in their upcoming investments. Sales forecasts are vital components of many decision-making processes at the organisational level in a variety of functional areas, including administration, promotion, advertising, manufacturing, and financing. [Cheriyan, Ibrahim, Mohanan, and Treesa (2018)] of great importance is knowing that having a reliable sales forecast available to a firm offers the following advantages.

It directs how sales targets and anticipated profits should line up. Aids in making better future decisions. It assists in shortening the time required for setting up targets and planning market penetration. Setting criteria that can be utilised to identify trends in the future is beneficial. Also not forgetting how catastrophic erroneous forecasts can be for a business as it introduces disparities and gaps in the business strategic goals.

Forecasting Methods

The two most fundamental forecasting techniques with which businesses can conduct and execute sales forecasting are: the top-down (TD) sales forecast strategy and the bottom-up (BU) sales forecast strategy.

Upon a decision by a company to choose to use the bottom-up sales forecast strategy, it begins by concentrating on the anticipated number of units it believes it will probably sell then multiplying that number by the average cost per unit. The number of available sales agents, the number of potential retail sites, and digital presence are some other variables that can be integrated with this strategy. The bottom-up sales approach seeks to begin with the least elements of the projection and work its way up. The number of sales agents and retail sites, as well as the pricing of a product or quantity, can all be dynamically altered under the bottom-up sales forecast strategy. As a result, this method offers information that is relatively detailed [Ofoegbu (2021)]

Ofoegbu also mentions that, on the contrary, the top-down sales forecasting strategy focuses primarily on the global demand (total addressable market-TAM). By predicting the proportion of the overall market share they want to sell for the reviewed time, it calculates the potential market share that the company can obtain. The bottom-up strategy should be used first, followed by the top-down strategy, to guarantee that the prediction is practicable.

However, both of these strategies should be used to obtain solid and deployable sales projections. Over the past decade, numerous studies have explored various forecasting techniques, and their findings are well-documented in the literature. However, it is important to note that each forecasting method has its own limitations and drawbacks. For instance, the accuracy of standard statistical methods heavily depends on the characteristics of the time series data, which can significantly impact the precision of the forecast. On the other hand, artificial intelligence (AI) approaches have the potential to surpass traditional statistical forecasting models in terms of accuracy. Nevertheless, these AI methods often require more time and computational resources to generate results [Liu, Ren, Choi, Hui, and Ng (2013)]. Consequently, researchers have consistently recommended the simultaneous integration of different methodologies to develop a novel "hybrid technique" that can provide a reliable and effective forecasting outcome [Ofoegbu, 2021].

Achieving Precise Forecasts

Management of businesses ought to be able to integrate these 5 processes in their forecast criteria to guarantee an appropriate forecast is generated from a business perspective:

Past data availability: Historical data and patterns should be easily accessible for analysis. The foundation of a sound prediction begins with a breakdown of the data by revenue, transaction time, volume of sales, and other pertinent criteria that should be addressed. Next, this information is formed into a "sales operating margin," which is regarded as the anticipated number of sales for the evaluation period. In light of this, the CRISP-DM methodology’s deployment is essential at this point since it will assist the analyst in gathering data that is relevant to the business’s needs and objectives and ensure that the data gathered for the prediction is accurate [Shearer (2000)].

Address anticipated changes: This requires adjusting the sales operating margin to reflect anticipated future changes. This could involve adjustments to the pricing owing to rivalry or other causes, adjustments to the client base that are increasing or declining, incentives, adjustments to the outlets that are either incremental or reducing, and adjustments to the commodity itself [Liu et al. (2013)].

Recognize Market Dynamics: This is the identification of economic trends by estimating market occurrences which are expected to have an impact on item purchases, such as legislative changes and corporate culture. Moreover, when formulating a projection, it's imperative to take into account factors like periodicity, climate, and locality because they have an impact on the results. It will also guarantee that the forecast is executed using the best characteristics.

Observe the competition: Keen observation of the market to make sure that rivals' rewards are matched and that new competitors' strategies are studied.

Integrate the business plan of the company: To sustain the business objectives, make sure that all business strategies and predictions are executed with the forecast. At this point, the CRISP-DM methodology is once again relevant because no company or organisation can proceed forward with creating sales forecasts before even taking its objectives into account. By putting the CRISP-DM methodology into practice, you can be sure that the prediction is being driven by a carefully thought out strategy to boost business prowess and general consumer pleasure [Shearer (2000)].

Sales Forecasting Success

The viability and accuracy of sales forecasting depend on a variety of factors, including mechanisation, verifiable evidence, superior analysis tools, and effective and optimal cooperation. The potential for a forecast to be manufactured instantaneously gives the executives of an institution the capacity to amend and make better-informed decisions. In addition, a quality forecast ought to be data-driven, and predictive modelling manages to be more insightful than most subjective methodological tools. A strong forecast also serves as a centralised source with many perspectives, providing deeper understanding into the performance of a company and coordinating several business operations inside it. Likewise, a solid fore- cast acts as a foundation for additional information for future predictions that are refined and whose precision continually increases. As a result of having a better understanding of both business and market dynamics which may therefore influence the outcome of the sales analysed period, firms with high-quality forecasting methods and strategies eventually outperform their rivals.

Machine Learning Algorithms and the Performance Metrics Selection

It is not easy to choose an algorithm for each and every problem. Although there is no perfect method that solves all problems, a select handful are well known for sometimes outperforming other methods. The algorithms' accuracy won't be the same for all data kinds; it will vary depending on the kind of data. Machine learning techniques that were estimated in this work to be effective on the problems were simple linear regression, Long Short-Term Memory (LSTM), random forest regressor and the Gradient Boosting regression.

One must first examine the outcomes before determining the algorithm that performs best, and only then can we make a prediction. In our own instance, the mean square error (RMSE) would be quite significant in determining how well an algorithm has performed. Also in this study, mean absolute error (MAE) statistic will be utilised to calculate the average degree of errors. The MAE will be employed to measure the average absolute difference between the predicted and actual values. Unlike RMSE, MAE does not square the errors, treating all individual differences equally. As a result, MAE is less sensitive to outliers compared to RMSE. By using both metrics, we can gain a more comprehensive understanding of the algorithms' performance, considering both the overall magnitude of errors and their sensitivity to extreme values.

Prior to now, the majority of studies have concentrated on measures like the f-measure, mean absolute percentage error(mape) and accuracy but in this study, measurements such as mean squared error, mean absolute error, measures of variation (r2) are taken into account as they do not ignore the weightings, error significance and correlation between the actual and projected data.

Model training, testing, and validation: The models in this study will be trained, tested, and validated using a combination of techniques. First, the dataset will be split into training and testing sets using a train-test split approach. This involves randomly dividing the data into two subsets: a larger portion (e.g., 80%) for training the models and a smaller portion (e.g., 20%) for evaluating their performance on unseen data. Additionally, cross-validation techniques, such as k-fold cross-validation, will be employed to assess the models' generalization ability and to reduce the risk of overfitting. In k-fold cross-validation, the training data is further divided into k subsets, and the model is trained and validated k times, each time using a different subset for validation while the remaining subsets are used for training. This process helps to ensure that the models' performance is consistent across different subsets of the data.

Research Structure

The proposed model for this thesis study is created using the blueprint below.

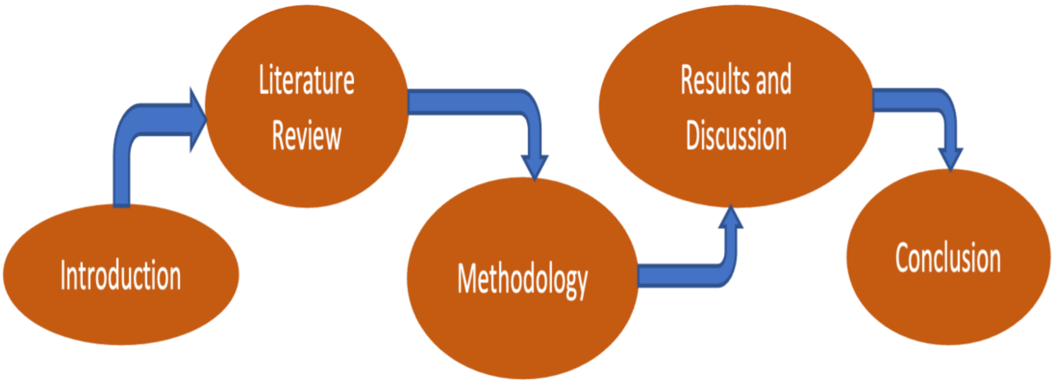


Figure 1.1: Research Structure Illustration

The Literature Review of various writers' research projects is found in Chapter 2. It explains the studies that has already been done in this field, as well as their drawbacks. Additionally, the flaws with the previous studies were examined. The suggested Methodology, which makes use of a variety of machine learning techniques, is presented in Chapter 3, along with a workaround for the drawbacks that are highlighted in the literature review.

The results and an understanding of it are included in Chapter 4. The Conclusion is then included in Chapter 5 so that scholars can realize that, in addition to this strategy, we can center on the expansion of revenue by utilizing machine learning approaches, which will be essential in enhancing corporate strategies.

LITERATURE REVIEW

”Fast Fashion Sales Forecasting with Limited Data and Time” (Choi, Hui, Liu, Ng, & Yu, 2014)

An industry technique frequently used in retailing of clothing is referred to as Fast clothing with its main goal being to provide the market steady supply of new products that showcase the newest fashions and assist manufacturers nail down the most popular designs. By creating a new algorithm known as the "Rapid Fashion Forecasting (3F) algorithm", that handles prediction using small dataset and in a constrained amount of time, the study explores the sales predictive model for fast fashion. The issue of prediction using small dataset and in a constrained amount of time is the subject of this research for the first time ever. A 4-period real sales dataset from a knitwear fashion company that utilises the principle of fast fashion is used for the study. The dataset is tested with the 3F Algorithm and the resulting outcomes are recorded. GM-EELMs with two neurons and two inputs are discovered to reach the time restriction of .5 seconds, whereas those with two neurons and seven inputs are discovered as shown by the research results that it is possible to complete the task in less than 20 seconds (and Not .5 s). The investigation goes to conclude that the 3F algorithm gives a more impressive prediction precision than the GM and can also perform excellently in the predictions of others like financial markets and short spanned seasonal products. [Choi et al. (2014)]

”Time Series Forecasting of Agricultural Products' Sales Volumes Based on Seasonal Long Short-Term Memory” (Yoo & Oh, 2020)

For a balance in agricultural supply and demand, this paper proposed developing a deep learning-based system (Seasonal Long short-time memory) in forecasting of agricultural sales. Sales of agricultural products are mostly time series, which makes it possible to project them using numerous traditional time series projection techniques. Majority of the time, machine learning or traditional statistical models are used to handle the supply and demand prediction problem for agricultural products. The time-series records comprising 5 products from a regional food shop's Point of sale system located at Wanju, South Korea (ChineseMallow, Jujube Mini Tomato, Onion, Lettuce, and Welsh Onion) from June 2014 to December 2019 that contained the fewest missing dates was selected for this study. In this study, the well-known models (Prophet, LSTM, SARIMA, as well as the suggested SLSTM) were implemented and each characteristics and execution results then compared. Three error measures which are the normalised mean absolute error (NMAE), root mean squared error (RMSE) and mean absolute error (MAE), have been adopted to appraise the forecasting performances. Lower accuracies were produced by the Prophet model's smaller predictions compared to actual values. Both LSTM and SLSTM predictions are more accurate than Prophet's predictions at predicting the direction of the actual data. These results explain why LSTM and SLSTM are more accurate than Prophet and auto arima. The paper came to the conclusion that in evaluating investigations with prophet, auto arima and regular LSTM models, the SLSTM with enhanced seasonal variation has a low predicting error rate and NMAE. [Yoo and Oh (2020)]. The Stacked LSTM (SLSTM) model is an extension of the standard LSTM architecture that incorporates multiple hidden LSTM layers. In a standard LSTM, there is a single hidden layer composed of LSTM units, which are responsible for capturing and learning the temporal dependencies in the data. By stacking multiple LSTM layers, the SLSTM model allows for a more complex and hierarchical representation of the temporal patterns. Each additional layer can learn increasingly abstract features from the previous layer's output, enabling the model to capture more intricate and long-term dependencies in the data.

”Machine-Learning Models for Sales Time Series Forecasting” (Pavlyshenko, 2019)

To forecast future purchases in this study, the "Rosemann Store Sales" dataset from Kaggle is used. The primary Python packages were used for the calculations. Analysis includes attributes like average sales price of past information, public and educational breaks, the range separating your store and those of your rivals, and the dispersion type of your store are taken into consideration in projecting knowing that Regression classifier captures patterns in the entire set of stores or consumer items. The data is analysed using different algorithms for machine learning, such as Random Forest, Neural Network, Arima model, Extra Tree model etc. The findings from the first level's models have non-zero parameters. The linear model and the Extra Tree model which are models from the Pythons' scikit-learn package, plus the Neural Network concept was implemented using the second stacking level. The 3rd level is where the result from level 2 was weighted and added. The analysis revealed that using regression models for sales prediction can commonly result in better outcomes than time series approaches. By taking into consideration the variances in outputs from numerous models with different permutations of inputs, stacking makes it possible to increase quality on the verification and on the out-of-sample sets of data. [Pavlyshenko (2019)]

”Time-Series Forecasting of Seasonal Items Sales using Machine Learning: A Comparative Analysis” (Ensafi, Amin, Zhang, & Shah, 2022)

With the use of various models, including SARIMA, Prophet Neural Networks and Exponential Smoothing, this papers objective is to perform time-series analysis of cyclical data. The best model in this inquiry is the one that produces the most accurate results. The dataset of the Superstore sales (Community.tableau.com, 2017) does not have any missing values and exhibits seasonality in its sales trend. This study's dataset, which comprises close to 10,000 data points and 21 attributes, describes the retail outlet volumes between 2014 and 2017. It includes sales data for office supplies, furniture, and technology items being the three categories. Variable of interest in this analysis is furniture sales, which exhibit seasonal trends. The most popular Indicators have been employed in this study's analysis to gauge forecast accuracy. The first Indicator to be used is MSE. Since the RMSE is on the same scale as the data, it can be preferred to the MSE as an indicator in this study. The third and final metric is MAPE, which has the benefit of being scale-independent. With a better MAPE and RMSE value, Stacked LSTM outperforms the other models according to all three metrics.

Out of 13 models, the CNN, Vanilla LSTM, Stacked LSTM, and Prophet models are among the five models that can predict seasonal time-series sales data is what this study concludes. [Ensafi et al. (2022)]. The reason why the Stacked LSTM outperforms other models in this study can be attributed to its deeper architecture, which allows it to better capture and model the complex seasonal patterns present in the sales data. The multiple layers of LSTM units enable the model to learn hierarchical representations of the time series at different levels of abstraction, from low-level patterns to high-level trends and seasonality. This deep learning approach is particularly effective in handling the non-linear and non-stationary characteristics of sales data, as it can adapt to changing patterns and extract meaningful features from the historical observations. In contrast, traditional models like ARIMA and standard LSTM may struggle to fully capture the complexity of the seasonal patterns, leading to less accurate predictions.

”Intelligent Sales Prediction Using Machine Learning Techniques” (Cheriyan et al., 2018)

Dataset gotten from a fashion store sales over three straight years is utilised. The sales forecast covers the three years from 2015 to 2017. The data mining approach goes through several stages of exploratory data analysis, including data interpretation, preparation, modelling, evaluation, and deployment. A smoothed average that has been vertically trendadjusted serves as the projection's basis and then seasonality-adjusted as well. Future revenue is predicted using machine learning methods like the Gradient Boost Tree (GBT), Decision Tree (DT), and Generalised Linear Model (GLM). Gradient Boost Algorithm offers 98% accuracy rate, Decision Tree follows closely as runner up with roughly 71% rate of accuracy, and Generalised Linear Model lastly comes in with 64% rate of accuracy, according to results obtained. Ultimately, if the three algorithms are empirically evaluated, Gradient Boosted Tree, which offers the highest prediction accuracy of all the algorithms, is shown to be the best fit for the theory. [Cheriyan et al. (2018)]

Further Literatures

It is possible to go back more than six decades in the history of sales forecasting [Boulden (1957), Winters (1960a)]. Ever since, numerous studies on sales forecasting have been released [Chen and Ou (2009), Lo (1994), Sakai, Nakajima, Higashihara, Yasuda, and Oosumi (1999), Wong and Guo (2010), Winters (1960b)], involving a robust range of real-world applications in the agriculture [Yoo and Oh (2020)], fashion [Liu et al. (2013)], furniture [Yucesan, Gul, and Erkan (2017)], and food [Tsoumakas (2019b)] industries, amongst others. Retailers generally concur that a retail product's initial sales are a reliable predictor of future sales [M. L. Fisher, Raman, and McClelland (2000)]. Nevertheless, there has been much focus on using early purchases to predict total retail product sales up to this point. Tanaka produced the sole piece of research that was discovered, and it forecasted long-term purchases based on early purchases and relationships between short-term and long-term cumulative purchases within related product groupings. Tanaka's work's predicting accuracy was heavily dependent on the choice of the sample population, which was made based on professional expertise, making it discretionary and perhaps untrustworthy. Further, Tanaka (2010) did not take into account how several influencing elements, such as manufacturing qualities and advertising strategies, may affect sales volume, making it unable to handle fluctuations in sales brought on by these factors.

The creation and enhancement of sales forecasting models for the retail industry have received a lot of attention over the past few decades as comprehensively discussed in M. Fisher & Raman, 2018. Like all companies, retailers must decide how to evolve strategically in the face of shifting market conditions and technology advancements. Forecasts are often a requirement for the conventional components of a sales approach, including market and competitive considerations in dynamic technological and legislative environments [Levy, Weitz, Grewal, and Madore (2012)]. The climate for small traditional vendors is also unstable, and there is ambiguity over the region and intended audience. Despite the fact that many of the issues that major merchants confront still exist (such as the digital service), there isn't much in the intensive information that even briefly summarizes the outcomes of the numerous small vendors' site choices. Accurate Economic Order quantity Unit quick demand estimates are essential in Just-in-time systems since they serve as the input elements for the supply chain scheduling process, which is often performed on a daily, weekly, or monthly basis. Since both underestimating quick demands and overestimating it are costlier due to the resulting supply constraints and poor customer service, overstocking, and goods depreciation, forecasting performance can significantly affect a firm's earnings outcomes [Sanders and Graman (2009)]. The fundamental approaches to forecasting item sales rely solely on historical sales data and linear modelling techniques. Conventional time series approaches are used in the retail sector [KALAOGLU et al. (2015)], including basic moving averages and the Auto Regressive Moving Average ARMA methodology. Another study area uses model-based forecasting to anticipate sales of products made, solely integrating promotional (plus other) data. [Fildes, Ma, and Kolassa (2022)], assert that these methods frequently depend on more advanced business theories or numerous linear models, whose components are related with periodicity, dated activities, climate, cost, and marketing features. The objective of new findings on projecting sales volumes has been to utilise a generic prediction model which can be applied to all the items under investigation. There is no guarantee that any method, no matter how sophisticated it may be, ranks higher for a diverse set of series than that of any other method, according to [Wolpert and Macready (1997)]”no-free-lunch theorem”. This indicates that it is nearly impossible that one particular method will be superior to others for all products and all coming years and emphasises how crucial it is to choose a method that fits the features of the problem. There hasn’t been a significant amount of study in the general forecasting community examining the advantages of various selection methods, differing the use of individual and collective choices and pairing [Fildes and Petropoulos (2015)].

You must first exclude the trend and seasonality from the data before using the autoregressive integrated moving average (ARIMA) models. For instance, you would have to eliminate this trend from the time series if you were analysing the number of regular visitors on your website and it was increasing by 15% monthly. To obtain the final forecasts, you would need to reintroduce the trend after the model has been developed and has begun to generate projections. Comparable to this, if one were attempting to forecast the monthly sales of water bottles, one would likely see significant seasonality: because it records high sales during the summer, and there is a high recurrence yearly. By calculating the change in the value at every time interval and the value from the previous year, for instance, one can be capable of eliminating this seasonality from the time series (technique referred to as differencing). Likewise, to obtain the final forecasts, one would need to re-introduce the seasonal pattern once the model has been developed and has made several projections. [Betel (2022)]

The Grey method (GM) has a reputation for being a strong contender for forecasting with limited past information. For instance, [C.-C. Hsu and Chen (2003a)] used the GM-based algorithm to project the demand for electricity despite having very little historical data, and they found some encouraging outcomes before an enhanced GM-based strategy was put out by [Yao, Chi, and Chen (2003)] also applying it to the forecasting of electricity demand. In order to create a projection technique for the integrated circuit industry Chen worked with the GM-based model and got favorable prediction results. Yao et al. (2003) discovered that their proposed enhanced GM approach performed extremely effectively with a largely small number of past data under a number of key situations that were particularly relevant to the electrical business. [Lin and Lee (2007), L.-C. Hsu (2009), Lei and Feng (2012), L.-C. Hsu and Wang (2007)] all carry out recent studies on predicting using GM models. Although the GM modelling techniques were thought to be an effective instrument for forecasting with little available data, it was still found to perform poorly, especially if the supporting data pattern is highly erratic and devoid of any recognizable trend [Deng (1989),C.-C. Hsu and Chen (2003b)].

[Tsao, Chen, Chiu, Lu, and Vu (2022a)] analyzed actual sales data as internal operational data from a business in the server industry. They created a revolutionary intelligent forecasting methodology that modifies demand forecasting findings by adding external information indices, stating that the suggested framework for intelligent forecasting offers a useful tool for businesses susceptible to inconsistent demand. Tsao et al. concluded that better forecasting outcomes should make it possible for a business to undertake manufacturing planning and control with more accuracy suggesting that when supply and demand are easily met, businesses become more competitive.

Using several data mining algorithms, [Jain et al. (2015)] did sales forecasts for stores in their project. Predicting sales for any retailer on any given day was the challenge at hand. Given that Data Mining applications have been popularly known for solving Demand predictions, Recommendation Systems, amongst a few others; the various Data Mining methods were applied to Retail Outlets and it was discovered that the XGBoost outperformed all others. When there are many time series available for commercial applications, data mining can be used for investigation and content discovery, as demonstrated in [Van Der Walt, Colbert, and Varoquaux (2011)] where time series data mining leverages univariate ARIMA models to a chain of a fast-food eatery.

One of the first mentions of the meta-learning terminology occured with [Prudêncio and Ludermir (2004)] in relation to forecasting many time series. With the traditional time series field, they offered a novel approach to obtaining testable theories which could be applied to enhance predictive ability as well as to offer a clearer knowledge of the time series analysis, irrespective of the fact that the principles of meta-learning may be used to other machine learning problems (regression problems as examples). Prudêncio and Ludermir examined two case studies with each employing a unique meta-learning strategy. While choosing necessary algorithms to predict passive time series in the first Case Study, utilising a single machine learning algorithm, and when choosing models for series in the second Case Study, the NOEMON technique was utilised. Taking into account the accuracy of the models chosen regarding the tests in the two case studies and their predicting abilities, the outcomes were adequate enough. A brand new meta-learner built on convolution neural network that could automatically retrieve properties from source purchase time series as well as its influencing variables through supervised learning was developed by [Ma and Fildes (2021)], employing a core forecaster mix that encompasses both personal and group forecasting techniques. Ma and Fildes were able to achieve a number of empirical conclusions from the forecasting experiments also to the ground-breaking meta-learning method described, as these are crucial in guiding the practise of commercial projecting.

A subfield of computer science called "machine learning" focuses on making it possible for computers to learn information without having to be specifically programmed [Samuel (1959)]. A more formalised concept explains Machine Learning as stating ”A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.” [(Mitchell, 1997), cited in Tsao, Chen, Chiu, Lu, and Vu (2022b)]. A few studies have been carried out utilizing machine learning methods including, estimation of previous sales loss and forecasting of upcoming product demand [Ferreira, Lee, and Simchi-Levi (2016)], employment of external social media data to project future demands for an online clothing company which reported that the random forest (RF) model outperformed all other techniques under consideration [Cui, Gallino, Moreno, and Zhang (2018)], and proposal of a predictive model that utilised clustering, classification, and the k-nearest neighbours algorithms for the demand for railway arrivals and departures during the pre-sale stage [Wei and Shan (2019)].

[Hyndman, Lee, Wang, & Wickramasuriya, 2018] defines a bundle of time series ordered hierarchically and capable of being averaged at various levels as a time series with levels.[Mancuso, Piccialli, & Sudoso, 2021] suggests a machine learning methodology in a study of multilevel time series forecasting while the method makes use of a deep neural network with convolutional layers that enable automatic extraction of properties of time series at any tier. Estimated coefficients accessible at any level of the hierarchy are combined with these features in the network. The findings of the study carried out by Mancuso et al., 2021 shows that the goal of combining all the relevant information through the hierarchy, both concealed and supplied by the parameter estimates, in a single machine, without the requirement for post-processing on the succession to reach optimal precision and homogeneity across sections, was achieved.

Customer retention describes the measures a firm takes to maintain customer loyalty and prevent customer churn, incorporating initiatives like personalised advertisement, membership initiatives, and customer relations. Most of the time, keeping current customers is simpler and cheaper than finding new ones [Derby (2018)]. A technique for predicting which clients will be kept and which will be lost one or more months in advance is proposed and evaluated by [Schaeffer and Sanchez (2020)]; the technique is tested using a variety of algorithms for machine learning (the Support Vector Mechanism, ADA Boost as well as Random Forests) that were identified to be effective in these situations. The results demonstrated that several techniques yield valuable classifications up to three months in advance. We can also conclude from the study carried out by [Saradhi and Palshikar (2011)] that identifying loyal customers is simpler than predicting customer attrition.

The pioneer study to look at the outcomes of over ten various predictive models, including both conventional and modern techniques like diverse Prophet and LSTM approaches was carried out by [Ensafi et al. (2022)]. The study examined the ability of Convolutional Neural Networks, primarily used for image recognition, to forecast sales of seasonal goods with results of the Ensafi et al. study showing that most of the neural network algorithms outperformed the traditional forecasting techniques. In a situation where a network system is in cycles, it is known as Recurrent Neural Network (RNN). They happen to be very important in forecasting time-series because of their capacity to retain memories of the past [Gamboa (2017)]. Two models were recommended by [Zhuge, Xu, and Zhang (2017)] namely the Emotional analysis model and LSTM; then reported that the LSTM model showed an outperformance in time series processing due to its ability to preserve both temporal properties of events and contextual information.

ARIMA, which stands for Auto Regressive Integrated Moving Average, is a popular model for time series forecasting. One of the key assumptions of ARIMA is that the time series data should be stationary, meaning that its statistical properties (such as mean and variance) do not change over time. However, many real-world time series data, including sales data, often exhibit non-stationary behavior, such as trends and seasonality. To address this issue, ARIMA employs a technique called differencing, which involves computing the differences between consecutive observations. The purpose of differencing is to remove the non-stationary components from the time series, making it more suitable for modeling. The number of times the differencing operation is applied is denoted by the "d" parameter in the ARIMA (p,d,q) notation.

MATERIALS AND METHODS

Data Overview

When choosing a dataset for further investigation, there are a number of elements to take into account, such as accessibility, anonymity, and data volume [Zhang, Zhang, Sun, and Liu (2018)]. We were able to retrieve 5 years of store-item sales data (Kaggle.com, 2019) from the Kaggle portal, which has no missing data, despite the fact that there are not many publicly released datasets with sufficient observations for execution of time-series forecasting. There are sales data from items from various stores that are included in this study. These data include information about the date of sales in each store, item number, store number and the amount of daily sales at each store. There are 4 attributes with 913000 instances each in the training dataset under analysis with every entry showing the daily sales volume at individual stores from 2013 to 2017, included. The training and testing portions of the dataset have been correctly separated.

Table 3.1: Description of Variables

Variable	Description
Date	Date of each sales occurrence. No holidays or Store closures are recorded
Store	An ID is assigned to each of the stores
Item	Each of the products are assigned an ID
Sales	The number of products that was sold on a specific store and date

Data Pre-Processing

One vital step for enhancing any project’s likelihood of success is to pre-process the data. In this study on sales forecasting, time-series grouping is employed, which greatly decreases computing capacity and improves prediction quality [Kotzur, Markewitz, Robinius, and Stolten (2018)]. The daily data used in this study is re-scaled into a month-by-month frequency because of the significant variation in the daily sales preceding projection. The sales volume and the date each piece of data was ordered are the most crucial components for performing sales forecasting. The JupyterLab and Google Colab environment in Python programming language was utilised for the execution of this task with an outlined approach shown below:

i. Importing and loading the required libraries and dataset.

ii. Performing an exploratory data analysis on the dataset to give a proper insight into the dataset.

iii. Using the various machine learning models on dataset.

iv. Viewing and comparing the performance output of the various models using appropriate indicators.

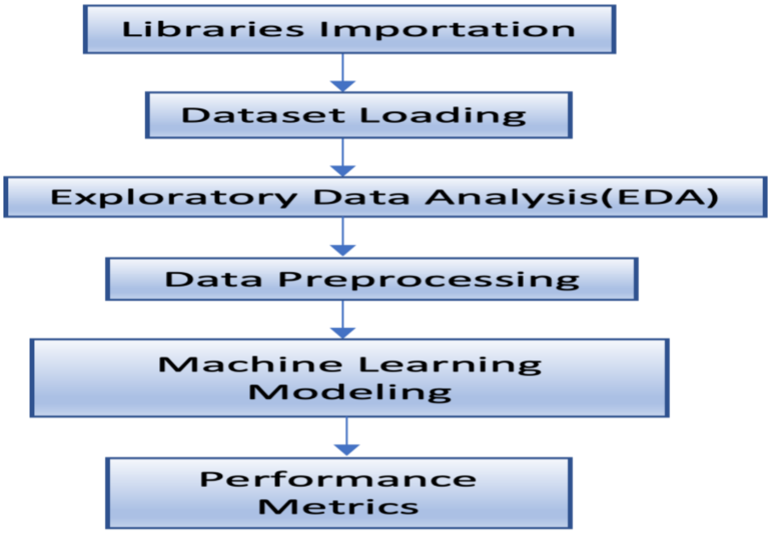


Figure 3.1: Process Flow

Importation of Necessary Libraries and Dataset

The LSTM-based recurrent neural network and all the python-based libraries required to read, analyse, and develop the dataset along with the machine learning models such as Pandas, NumPy, Keras, SKlearn, Matplotlib, Seaborn, etc. are imported. A brief overview of these libraries is given below:

i. Numpy: A multidimensional, regular collection of elements is called a NumPy array. Both the kind of components it contains and its structure define an array. Essentially, a NumPy array is basically a handy way to describe one or more blocks of computer memory so that the numbers they symbolise may be easily changed. NumPy offers a high-level framework for numerical processing without sacrificing efficiency, as explained earlier [(Van Der Walt et al., 2011)]

ii. Panda: Rich data formats as well as methods are provided by Pandas to make dealing with complex information quite quick, simple, and as descriptive as possible. As you'll see, it's one of the essential components that makes Python such a potent and effective environment for data analysis. Pandas combines NumPy's high-geared array- computing capabilities with the adaptability of spreadsheets and relational databases for data processing (such as SQL). In order to make it simple to mold, split, and divide data, do groupings, and choose data subsets, it offers robust indexing features. [McKinney (2012)]

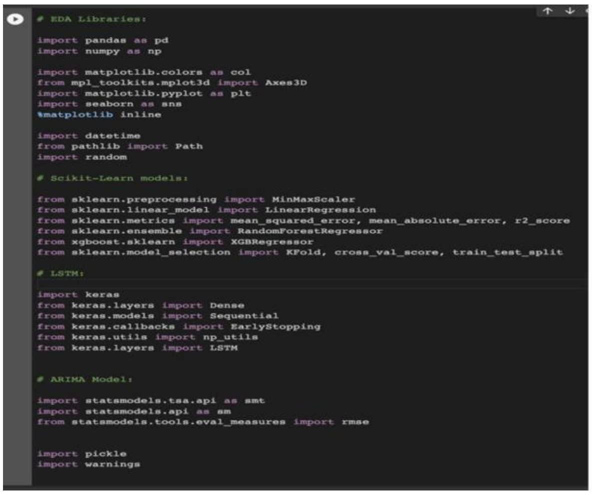
iii. Keras: Keras can be used with Tensorflow, a small deep learning framework in Python that's easy to learn. It enables programmers to take care of the minute intricacies of tensors, their structures, and their arithmetic elements while concentrating on the key ideas of deep learning, including building layers for neural networks. The backend for Keras needs to be TensorFlow, Theano, or CNTK. For deep learning applications, Keras can be used instead of TensorFlow, which is a very sophisticated framework. [Manaswi (2018)]

iv. Seaborn: An open-source python graphic presentation package called Seaborn is based on the core settings of Matplotlib. It makes some data visualisation processes accessible to the public, including those that are frequently used processes that map colour to a parameter or use acting needs around the universe. [Sial, Rashdi, and Khan (2021)]

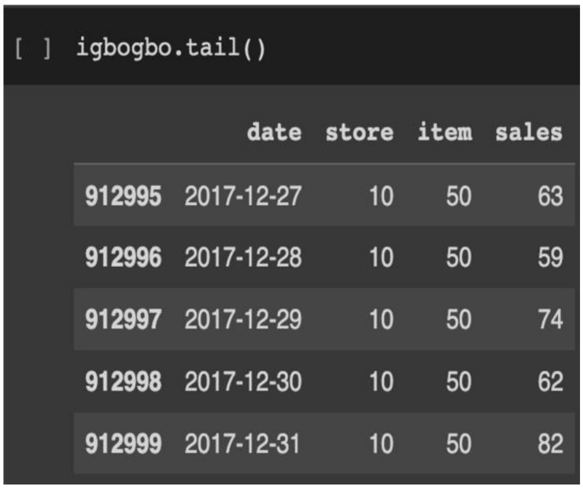
v. SciKit-learn: Using a standardised, mission architecture, Scikit-learn exposes a huge variety of machine learning algorithms, either supervised or unsupervised, making it simple to compare approaches for a particular application. It can be seamlessly implemented into systems outside the conventional scope of analyzing statistical data because it depends on the experimental Python framework. [Pedregosa et al. (2011)]

vi. Matplotlib: In data science, visualisation is an important process. Using visualisation, one may immediately comprehend data insights. Production-quality graphs are generated by the Python library for 2D plotting called Matplotlib. Giving room for both interactive and passive plots, also supporting a variety of output formats for image storage (PNG, PS, etc.). It offers a large selection of graphic layouts (bar charts, stacked pie charts, line charts, histograms, and many more). Because MATLAB was exception- ally good at graphing, Matplotlib was designed after MATLAB. Many people switched from MATLAB to Matplotlib due to their excellent level of similarity, extreme versatility, user-friendliness, and complete scalability. [Tosi (2009)]

After loading all the mentioned Libraries above, the retail sales dataset, in CSV format, was loaded into our python environment as illustrated below.



```
# EDA Libraries:

import pandas as pd
import numpy as np

import matplotlib.colors as col
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

import datetime
from pathlib import Path
import random

# Scikit-Learn models:

from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.ensemble import RandomForestRegressor
from xgboost.sklearn import XGBRegressor
from sklearn.model_selection import KFold, cross_val_score, train_test_split

# LSTM:

import keras
from keras.layers import Dense
from keras.models import Sequential
from keras.callbacks import EarlyStopping
from keras.utils import np_utils
from keras.layers import LSTM


# ARIMA Model:

import statsmodels.tsa.api as smt
import statsmodels.api as sm
from statsmodels.tools.eval_measures import rmse


import pickle
import warnings
```



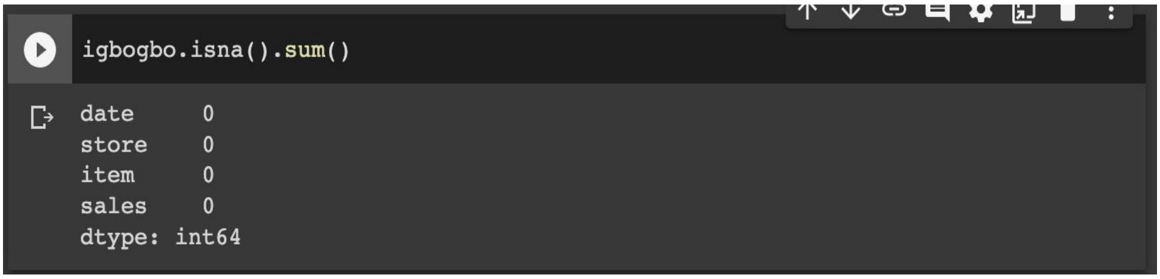
```
[ ]  igbogbo.tail()

              date  store  item  sales
912995  2017-12-27     10    50     63
912996  2017-12-28     10    50     59
912997  2017-12-29     10    50     74
912998  2017-12-30     10    50     62
912999  2017-12-31     10    50     82
```

Figure 3.2: Importation and Dataset Overview

Exploratory Data Analysis (EDA)

The next action taken to analyse the dataset was exploratory data analysis (EDA). The overview and illustration of the EDA's processes is as follows:

Finding missing/null values: The dataset contained no missing values.



```
igbogbo.isna().sum()
```

```
date     0
store    0
item     0
sales    0
dtype: int64
```

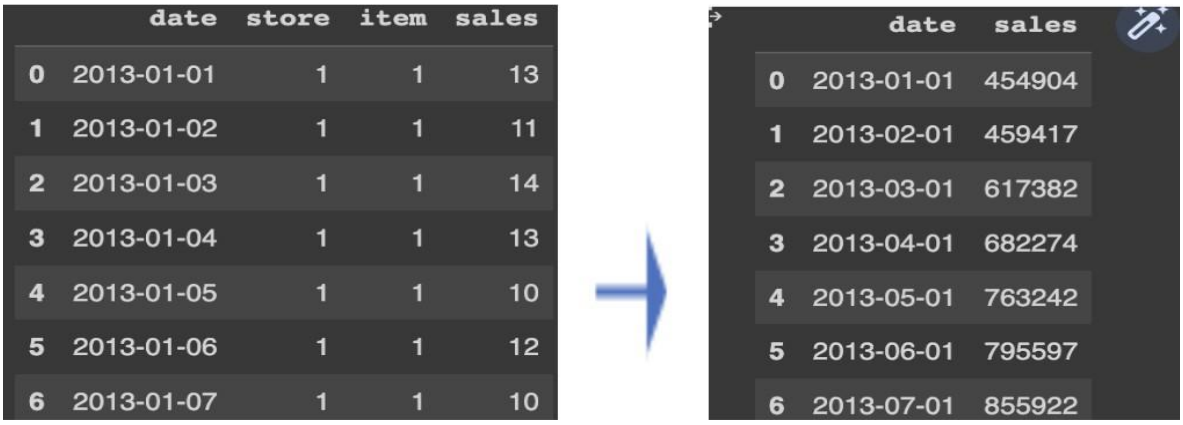
Figure 3.3: Null Values

Data Cleaning: After ascertaining that the dataset did not contain any null values, the following data cleaning process is implemented:

Dropping Irrelevant Columns: We decide to exclude the columns for the Store and Item ID’s since we are only interested in the total items sales across all of the stores.

Date Conversion: The dataset’s date column is an object datatype. So that it may be utilised for other analyses, it should be converted into a Date Time datatype.

Grouping: We will train our algorithms to forecast sales for the following month instead of the day after. Therefore, we must change our date to the month format and then add the total number of products sold throughout each month.



	date	store	item	sales
0	2013-01-01	1	1	13
1	2013-01-02	1	1	11
2	2013-01-03	1	1	14
3	2013-01-04	1	1	13
4	2013-01-05	1	1	10
5	2013-01-06	1	1	12
6	2013-01-07	1	1	10

	date	sales
0	2013-01-01	454904
1	2013-02-01	459417
2	2013-03-01	617382
3	2013-04-01	682274
4	2013-05-01	763242
5	2013-06-01	795597
6	2013-07-01	855922

Figure 3.4: Grouping Daily Sales into Monthly Sales

Visualisation: It is important to visualise our dataset and hence we proceed to visualise the total number of sales made across the various stores.

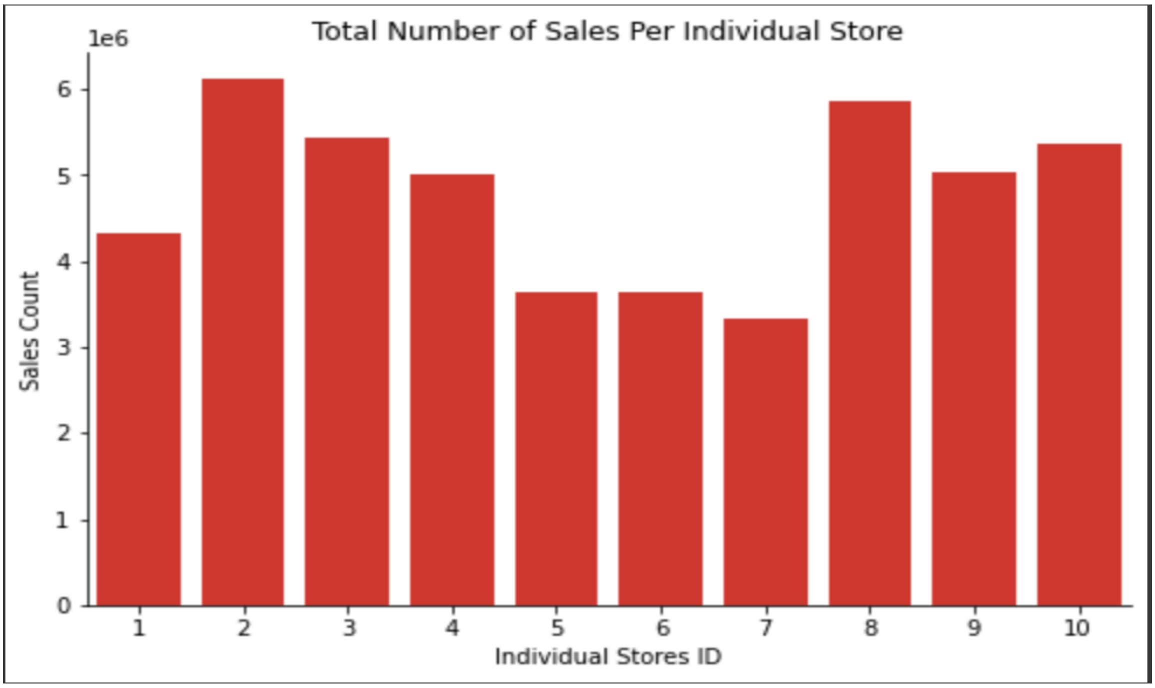


Figure 3.5: Total Individual Stores Sales

Stationarity: "Stationarity" is one of the most important factors to be aware of while working with time series data. The phrase "stationary series" usually refers to a set of data whose average, variance, and covariance do not vary with respect to time changes. A series must not demonstrate a trend in order to be considered stationary. Our data is not stationary because when we chart the monthly net sales volume, the mean monthly sales increase as time passes. To make our data frame stable, we will compute the disparity in purchases for every month and add it as a new column. The following image shows our data prior to and after the differencing procedure.

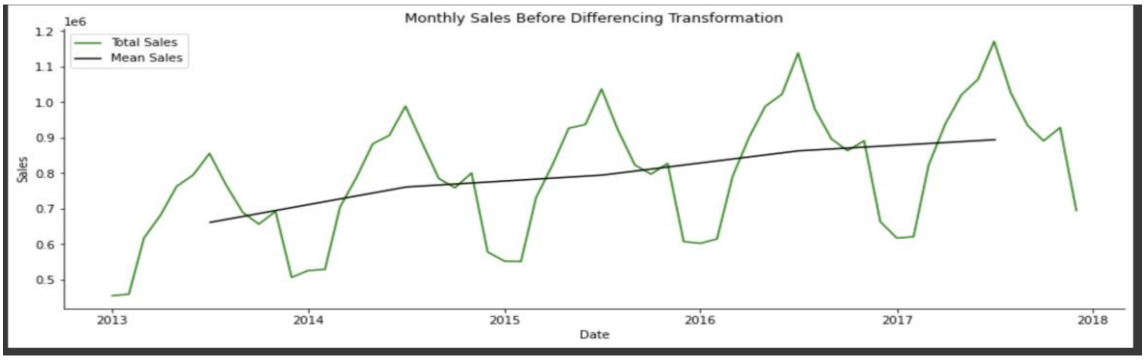


Figure 3.6: Before Transformation

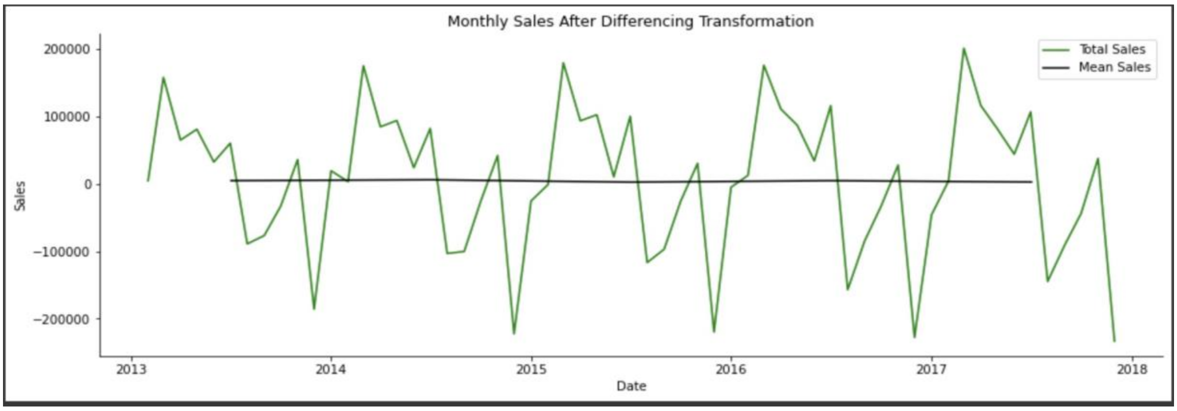


Figure 3.7: After Transformation

Models for Forecasting

The model selection procedure for this study was heavily influenced by the amazing Kaggle post titled: "Time Series - ARIMA, DNN, XGBoost Comparison."[Enolac5 (2018)] which is about comparing three models on a time series data set: ARIMA, DNN, and XGBoost. The nature of the set of data used to generate the analyses and the nature of the dataset used to generate our forecasts are substantially the same. Hence the optimism that if I completed other aspects of the project, such as data preparation and exploration, together with trying linear machine learning algorithms correctly, I could not make a mistake experimenting with the models mentioned above as they all produced outcomes that were quite compelling in the blog.

The sales of specific items in a set of stores must be predicted for 3 months. There isn't a perfect technique for data modelling that is applicable to all scenarios and datasets. Therefore, the best course of action would be to test out several machine learning algorithms, make some adjustments, and then tailor them to your particular circumstance. Some of the models we employed are discussed here:

Random Forest: The Random Forest (RF) classifier, happens to be an ensemble learning approach for classification which uses a subset of training cases and parameters that are randomly chosen to build several decision trees [Dey, Singh, and Singh (2016)]. The representation of a decision tree is a tree with a root and branches and nodes. According to Dey et al., the branches reflect relationships between the attributes that lead to the class labels, whereas the leaf nodes depict class labels. The Decision Tree's branches and nodes are connected in accordance with the data presented in the dataset [Dey et al. (2016)]. Due to the superior classification outcomes and processing speed, the adoption of the random forest (RF) classifier has drawn more attention during the past 20 years [Breiman (2001), Pal (2005)]. The predictions from an ensemble of decision trees used by the RF classifier produce accurate classifications [Breiman (2001)]. Furthermore, the variables with the best ability to distinguish between the target classes can be chosen and ranked using this classifier. [Lou et al., 2014] expressed this more formally as:

$$g\,(a) = f0(a) + f1(a) + f2(a) + f3(a) + \ldots..$$

Where the first specific model's f (1) is the number of the initial configuration's g. Any basic classifier in this case is a decision tree. Model assembly is a broad term describing the use of many models to enhance prediction accuracy.

Gradient Boosting (XGBoost): Gradient boosting is a regression method that involves boosting. Its basis is the idea that the best present state will reduce the highest estimation error when paired with earlier iterations. Creating objective outcomes to reduce inaccuracy is the main concept for this iteration. [Friedman (2001)]. In contrast to the Random Forest, which develops each base classifier individually using a small amount of data, GBRT employs a specific assembly method known as gradient boosting [Rokach (2016)].Gradient boosting is implemented with XGBoost, which also has a variety of additional features to improve speed and performance. By producing simultaneous decision trees, it aids in concurrency. This technique can evaluate extensive and complex models because it also has widespread computational power. It is an out-core processor since it analyses extensive and varied datasets. Resource utilisation is superbly managed by this method [Akanksha, Yadav, Jaiswal, Ashwani, and Mishra (2022)].

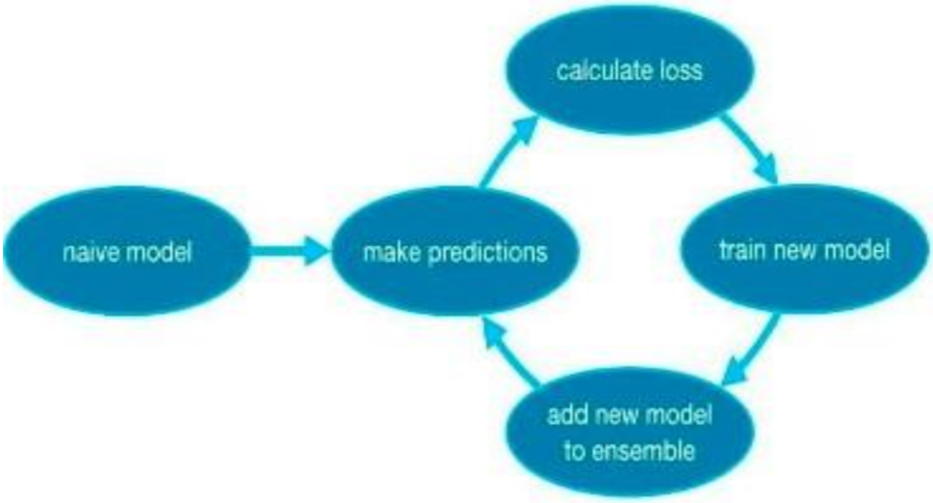


Figure 3.8: Cycle of Gradient Boosting [Alexis, 2022]

Long Short-Term Memory (LSTM**):** Establishment of the Long Short-Term Memory (LSTM) system was to address the fading gradients problem. To prevent it from disappearing, the target function’s gradient with regard to the signal from the state (the quantity linearly correlated to the input numbers determined during training with Gradient Descent) can be integrated into the RNN cell. This is the essential concept in the LSTM design[Sherstinsky (2020)].Where such action does not degrade it, LSTM will train to connect significantly less time lapses longer than a thousand sequential steps by mandating recurrent error movement via recurrent error conveyor belts within specific units.

”In comparison with real-time recurrent learning, backpropagation through time, recurrent cascade correlation, Elman nets, and neural sequence chunking, LSTM leads to many more successful runs, and learns much faster. LSTM also solves complex, artificial long-time-lag tasks that have never been solved by previous recurrent network algorithms” [(Hochreiter & Schmidhuber, 1997)].

Auto-Regressive Integrated Moving Average (ARIMA): Usually, whenever working with time series data, ARIMA modelling is the best option. However, ARIMA is most effective for data with a single variable, or for univariate time series. For instance, forecasting of stock prices. One method of applying ARIMA to a dataset involves grouping the time series data, projecting results after fitting ARIMA for each group in accordance with the identification columns. Due to this, the model must be fitted numerous times for each of the identifier’s distinctive values [Swami, Shah, and Ray (2020)]. In essence, the adaptability of ARIMA modelling allows for the creation of an appropriate model that is derived from the structure of the data. Derivatives of partial autocorrelation and autocorrelation enable for the finding of information like trend, random variations, cyclical component, repetitive patterns, and cointegration by roughly approximating the stochastic aspect of the time series. Deriving predictions of the series' anticipated values that are roughly accurate is therefore straightforward. [Ho and Xie (1998)].

Measurement of Forecast Error

The competence of the models is revealed through time-series forecasting evaluation criteria. Model evaluation can be done in a variety of ways [Hamzaçebi (2008)]. Three widely used metrics are employed in this study to estimate the model's competence; the root mean squared error (RMSE), the mean absolute error (MAE), and the R-squared score. They are briefly discussed below:

i. Root Mean Squared Error: A group of predictions' average errors are measured by its root mean squared error. As seen in Equation (1), each error's squares are summed before being averaged and the square root is taken. This makes sure that the direction of the error is inconsequential and that all errors have the same weight. Given that it is a quadratic function, global minima will always be reached. [Mitra, Jain, Kishore, and Kumar (2022)]

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - E_i)^2} \tag{3.3.1}$$

Where the obtained value is Oi, the prediction value is Ei, and we have n as the number of observations.

ii. Mean Absolute Error**:** A collection of predictions' average errors are calculated using this method. The error's positivity or negativity is neglected because it is absolute, and each individual error is given the same significance. It is simple to calculate MAE as Equation (3.3.2) illustrates. The aggregate of the observations is divided by the sum of the absolute values of the errors to obtain the "total error."[Wang and Bovik (2009)]

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - E_i| \tag{3.3.2}$$

Where the obtained value is Oi, the prediction value is Ei, and we have n as the number of observations.

iii. R-square Score: The coefficient of determination is a number that represents the amount of variability in the dependent variable that a framework can comprehend [Valbuena et al. (2019)]. The R-square score is used to evaluate the dispersed data relative to a line of best fit. Smaller disparities between the expected and true data are evidenced by higher R-square scores for related datasets. On a scale ranging from zero to 1, it evaluates the correlation between the predicted and actual data Mitra et al. (2022). An R-square value of 0.75, for instance, means that variation in the independent variable accounts for 75% of the variations in the dependent variable under study. Eq. (3) gives an illustration.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{total}}} = 1 - \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{\sum_{i=1}^{n}(O_i - \mu)^2} \tag{3.3.3}$$

The true value is Oi, the forecast value is Ei, and the mean is denoted as mu. The sum of squares for residuals is known as SS res, while the sum of squares for all data is known as SS total.

FINDINGS

The examination of the results is treated in this section. Each model is tested using 20% of the remaining data once it has been trained on a trainset, made from 80% of the total dataset.

ARIMA

The grid search approach is used to determine this ARIMA model's optimal order, which is (12, 0, 0). [Brownlee (2021)]. The forecasting for sales is shown in Fig 4.1. Below. Seasonal ARIMA (SARIMA) model is searched using a grid [Vincent (2021)]. The Auto-ARIMA grid search performance is assessed using a manually created grid search architecture. It discovered that (1, 1, 0) x is the ideal order (0, 1, 0, 12). The SARIMA model, which utilised Auto-ARIMA, has an RMSE value of 14959.89346608319.

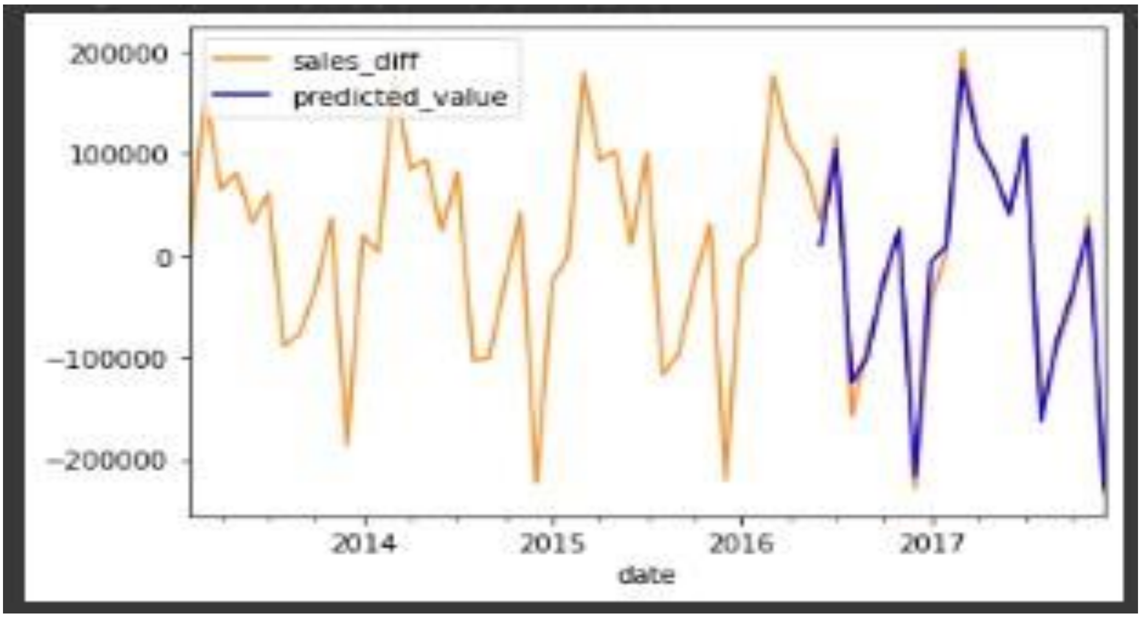


Figure 4.1: ARIMA

LSTM

Despite using the same system that has been developed on identical data, neural network techniques yield various outcomes because they seed arbitrary weights during the training phase. Utilized the Stacked LSTM and the average outcome of the training and testing procedures, which have been repeated 50 times, is used to create solid outcomes. The Stacked LSTM's MAE value is 12314.000000. The LSTM's R2 value is 0.992499. The RMSE is 14579.789785. The results of the LSTM model's sales predictions are shown in Fig 4.2. Below.

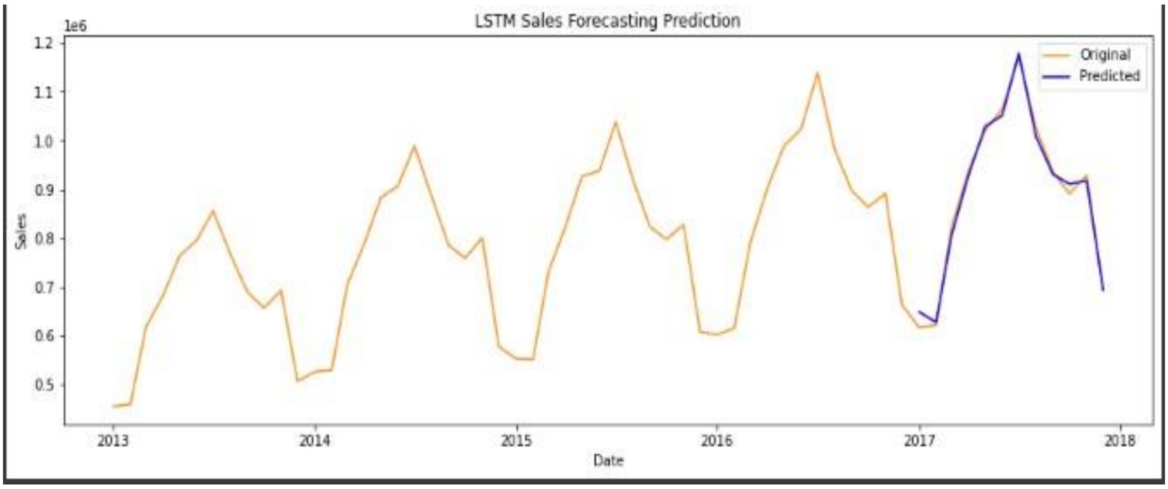


Figure 4.2: LSTM

Linear Regression

A linear correlation between a group of input and output parameters is sought after through linear regression. During the training phase, weighting associated with the input parameters are modified to ensure that the predicted and intended outcomes are closely aligned.

We have invoked Linear Regression using the Scikit-Learn library and passed the training set of data. Additionally, we use the test data and the linear regression models predict capacity to obtain the projected results. An RMSE of 16221.040790693221, MAE of 12433.0 and an R2 value of 0.990716 is obtained for this model. The result is shown in the Fig 4.3. Below.

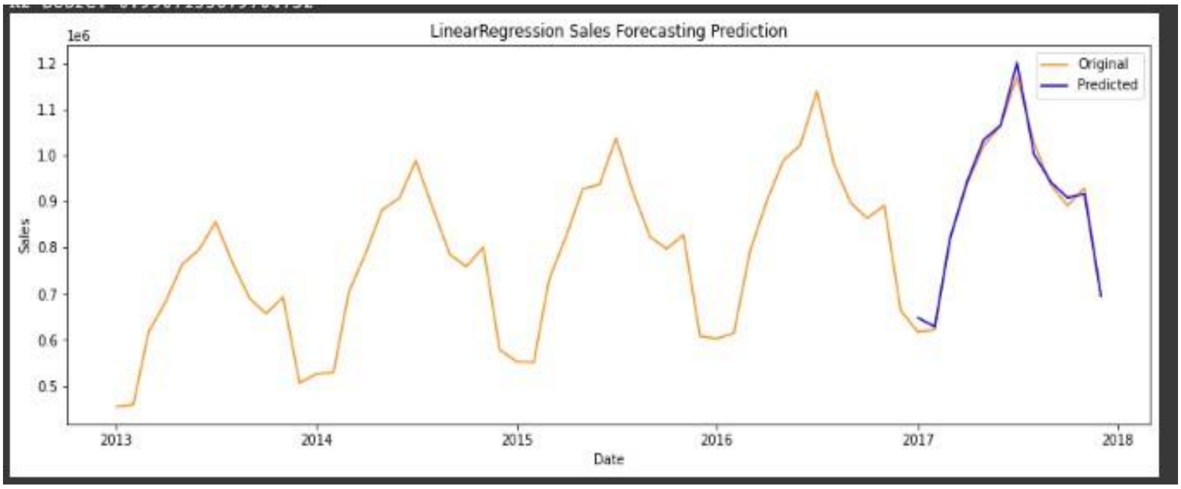


Figure 4.3: Linear Regression

Random Forest

The average of the outputs from each decision tree included in the Random Forest model serves as the ultimate output value in this collection of decision trees. With the help of the variables n estimators and max depth, we can define the number of decision trees and their total depth. After defining the Random Forest's features, we fit the model to the training data and forecast the resulting sale difference values and then compare the anticipated and actual sale values to gauge the model's performance. An RMSE of 17208.303725, MAE of 14559.250000 and an R2 value of 0.989551. We see the result in Fig 4.4. Below.

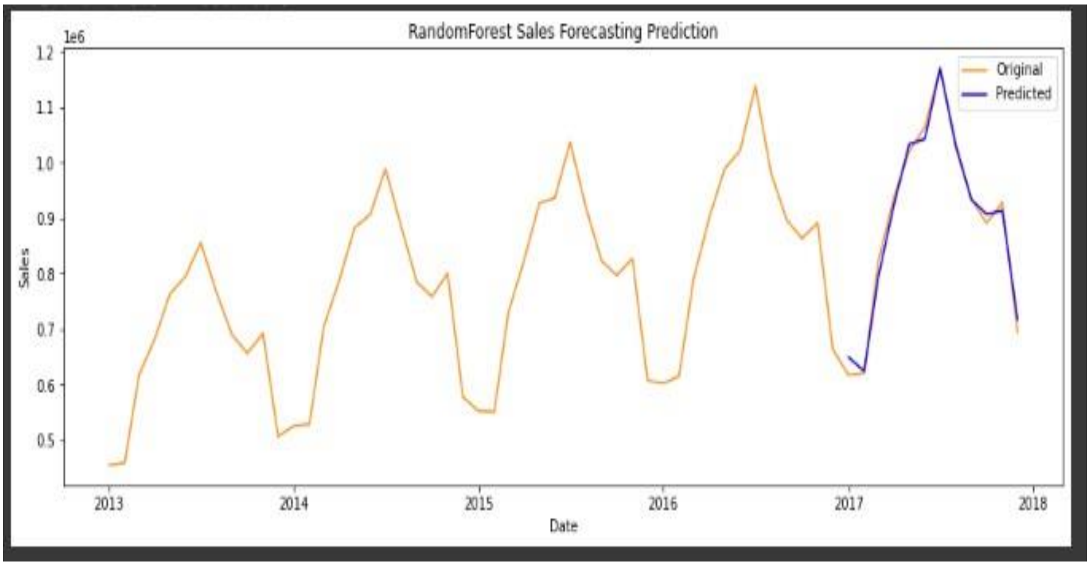


Figure 4.4: Random Forest

XGBoost

The gradient boosting approach known as the XGBoost uses the results of one decision tree to determine the learning variables for the subsequent decision tree, except the final decision tree produces a final result. The ensembled value of every decision trees that make up the XGBoost regressor is what makes up the ultimate result. By using the variables n estimators together with the learning rate, we can select the amount of decision trees and the velocity at which they learn. An RMSE of 13574.792632, an MAE value of 11649.666667 and an R2 value of 0.993498 is achieved with this algorithm. Let's visualize the result in Fig 4.5. below.

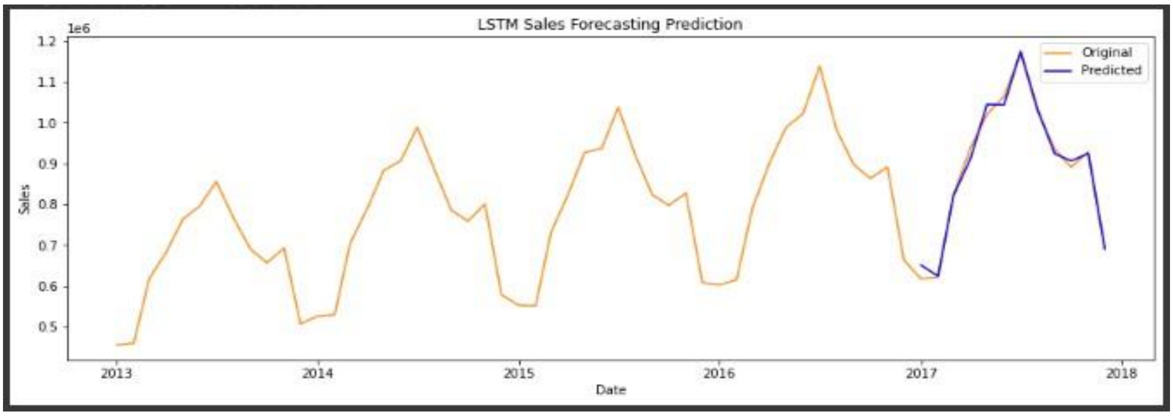


Figure 4.5: XGBoost

Result Comparison

Now that different models have been explored, we can compare the outcomes. By combining the outcomes from each model that was employed, a comparison can be established. Table 4.1 provides a convenient summary of the results.

Table 4.1: Summary of Error Scores

Model	RMSE	MAE	R2	Percentage Off
Random Forest	17208.303725	14559.250000	0.989551	1.63%
Linear Regression	16221.040791	12433.000000	0.990716	1.39%
ARIMA	14959.893466	11265.335748	0.983564	1.26%
LSTM	14579.789785	12314.000000	0.992499	1.38%
XGBoost	13574.792632	11649.666667	0.993498	1.3%

The Root mean square error (RMSE) and Mean Absolute Error (MAE) will be examined to compare model performance. Both of these metrics, albeit they have significantly distinct statistical and intuitive meanings, are frequently used to compare the performance of models. By taking the square root of the total sum of all squared errors, we may get the root mean square error (RMSE). When we square, greater errors do have greater influence on the total error, but smaller errors have less of an impact. Also, the mean absolute error, or MAE, indicates how far off the mark on average our forecasts are from reality. The weight of each error is the same in this situation. The formula for calculating the percentage of the prediction from the actual given in table 4.1 above is given as:

$$\text{Percentage Off} = \left(\frac{\text{Model MAE Score}}{\text{Average Monthly Sales}}\right) \times 100 \qquad (4.1.1)$$

The comparison of the models is shown in Figure 4.6 below.

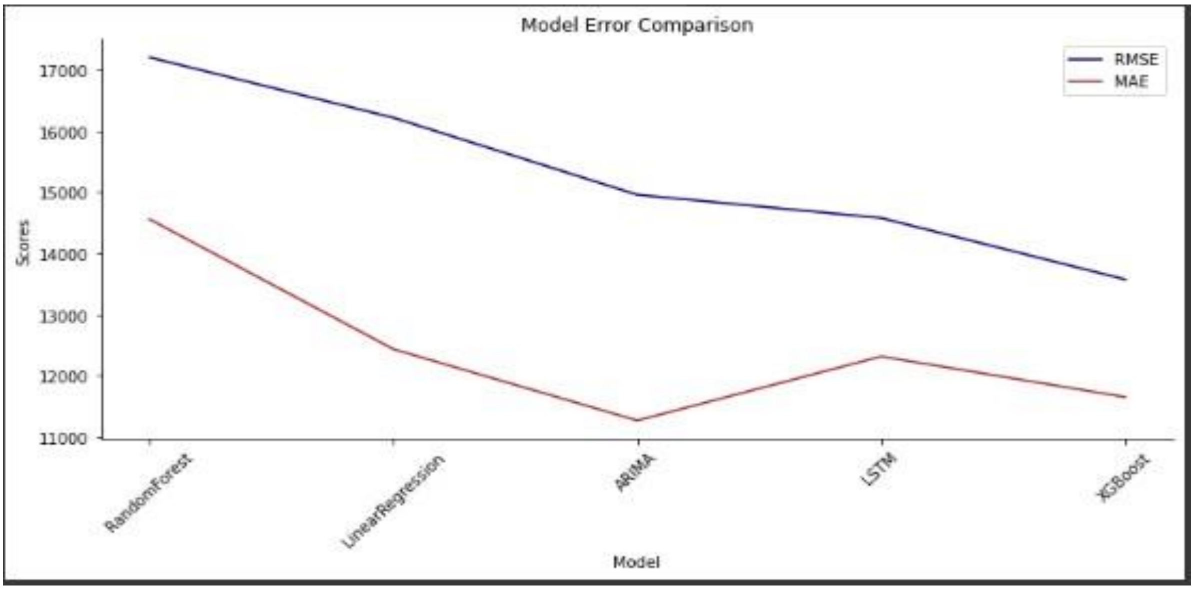


Figure 4.6: Error Comparison

Whilst the model’s outcomes appeared to be identical in their respective graphs seen earlier in the study, we observe from Fig. 4.6 that they actually differ in terms of accuracy.

CONCLUSION AND RECOMMENDATIONS

Conclusion

People have long been curious about forecasts. Though it can be challenging, numerous studies are being carried out to find strategies for making more precise forecasts. Because it affects so many firms today, sales forecasting needs to be addressed. To solve these problems, we can develop systems that can forecast sales based on the provided data. We discussed various machine learning (ML) techniques, their functionality in systems, and how accurate they are at forecasting sales in this study. Then, these approaches are evaluated against criterias from distinct perspectives. Five algorithms were used to analyse the precision of sales forecasting. For this study, the Kaggle platform’s store-item sales data was used. The prediction studies used the time-series and classification techniques; random forests, arima, lstm, xgboost and linear regression. The outputs of the models were accessed and contrasted using the R-squared score, Root Mean Square Error (RMSE), and Mean Absolute Error.

It was essential to comprehend errors, for which the RMSE and MAE measures were used. The best possible outcome would be an RMSE of 0, however in real financial datasets, this is almost impossible to obtain. We recognize that, generally, the XGBoost model outperformed the LSTM and ARIMA models by a small margin. It should be noted that all of the models discussed above were generated as simple as possible in order to show how they might be applied to sales forecasting, only minor tuning was done. For example, the LSTM and Random Forests may perform much better with hyperparameter optimization as well as many more nodes and layers.

This finding shows how machine learning algorithms can be used to estimate sales, which can be very useful for budgeting and tailoring. XGBoost likely outperformed the other models due to its ability to capture non-linear relationships and its robustness against overfitting. As a gradient boosting algorithm, XGBoost combines multiple weak learners (decision trees) to create a strong predictive model. This ensemble approach allows it to learn complex patterns in the data, including non-linear interactions between features. Additionally, XGBoost employs regularization techniques, such as L1 and L2 regularization, which help prevent overfitting by penalizing overly complex models. These properties make XGBoost particularly well-suited for handling the complexities and potential noise in retail sales data.

However, some key hyperparameters that could be considered for tuning include:

i. For Boost: max depth, learning rate, n_estimators, subsample, colsample bytree

ii. For Random Forest: n_estimators, max_depth, min_samples_split, min_samples_leaf

iii. For LSTM: hidden_layer_sizes, dropout_rate, learning_rate, batch_size, and epochs.

Recommendations for Future Work

This paper’s research can be expanded in numerous ways. First, it would be intriguing to be able to conduct a comparable analysis for sales forecasting by taking into account the data with its daily frequency rather than a monthly one. Indeed, retailers and distributors may find value in more frequent sales forecasts. Furthermore, a more thorough variation of the study could differentiate between customer segments because the sales tactics used by the businesses while interacting with consumers of different types and sizes do vary. Customer segmentation is a very intriguing and busy area of applied study, and it can definitely increase businesses’ profits. So, in the near future, the analysis can be continued by fusing time-series forecasting with customer classification techniques i.e. by knowing when and what each person buys, how individual clients recognise particular items separately (identifying supplementary and alternative commodities for each of them), and in what way the demand price elasticity for various products/customers correspond to one another are all factors that can also result in financial gains for businesses. Significant future study should incorporate discounts and holidays data into the model because they have a major impact on sales forecasting. There are also additional areas for this study’s future research. More comprehensive LSTM can be tested in order to enhance the outcomes. Adopting multivariate time-series forecasting is another option to investigate. Additionally, developing hybrid models, which mix traditional and modern forecasting methodologies, can be advantageous and represent a viable study area. Lastly, the machine learning algorithms can be attempted in Actuarial Employee Benefit Valuations as they are also mostly a month-by-month dataset of employee salaries trying to forecast the future liability needed to be reserved.

The forecasting techniques used in this study for retail sales could potentially be adapted to actuarial science because both domains involve analyzing time-series data to make predictions about future outcomes. In retail, the goal is to forecast future sales based on historical patterns, while in actuarial science, the aim is to predict future claims, premiums, or reserves based on past data. The underlying principles of time-series analysis, such as identifying trends, seasonality, and other patterns, are applicable in both fields. However, it's important to note that actuarial data may have its own unique characteristics and requirements, such as longer time horizons, different types of risks, and regulatory constraints. Therefore, careful consideration and domain expertise would be necessary to effectively apply these techniques in an actuarial context.

Limitations

Lack of a completely detailed dataset that captures metrics such as promotional offers and discounts does not allow for the study to have a complete picture of how accurate the models predictions are because they are very important metrics that should be considered in an exercise as such. The computers inability to run the python scripts was one restriction for this study. The majority of this study had to be executed on Google colab because the keras library kept shutting down the jupyter notebook kernel on the computer system that was used, which indicates that a very high processing system is needed for forecasting exercises.

There is a risk of data leakage and overfitting in this study, particularly given the small dataset and the use of powerful models like XGBoost. Data leakage can occur if future information is inadvertently included in the training data, leading to overly optimistic performance estimates. To mitigate this risk, it's crucial to ensure a strict separation between the training and testing data and to use techniques like time-based cross-validation to simulate real-world forecasting scenarios. Overfitting is another concern, especially with models that have high flexibility, such as XGBoost. Overfitting occurs when a model learns to fit the noise in the training data, resulting in poor generalization to new data. To address this, techniques like cross-validation, regularization, and early stopping can be employed. Additionally, collecting more data and using larger datasets can help reduce the risk of overfitting.

Acknowledgment

Firstly, all glory be to God Almighty for His favours in enabling me to satisfactorily conclude this project work. I am particularly appreciative of my project Supervisor who happens to double as my Personal tutor, Dr. Terry Sithole for her advice, persistence and feedback in ensuring my completion of this assignment. All my various module lecturers who have in one way or the other also impacted us as students thereby aiding in the production of this project work are also not left out. Lastly, I will always be appreciative of my brother, Babajide Mustapha and my entire family for their financial and psychological support. Their faith in me kept motivating me to continue.

REFERENCES

Akanksha, A., Yadav, D., Jaiswal, D., Ashwani, A., & Mishra, A. (2022). Store-sales forecasting model to determine inventory stock levels using machine learning. In *2022 international conference on inventive computation technologies (icict)* (pp. 339–344).

Alexis, C. (2022, January). *Time series - arima, dnn, xgboost comparison.* http://www.kaggle.com/code/alexisbcook/xgboost.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481–85.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Betul, G. (2022, September).*Time series-arima, dnn, xgboost comparison.* http://www.kaggle.com/code/badl071/forecasting-future-sales -using-machine-learning.

Boulden, J. B. (1957). Fitting the sales forecast to your firm. *Business Horizons*, *1*(1), 65–72.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Boyapati, S. N., & Mummidi, R. (2020). *Predicting sales using machine learning techniques.*

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brownlee, J. (2021, November). *How to grid search arima model hyperpa- rameters with python.* machinelearningmastery.com/grid-search-arima-hyperparameters-with-python.

Chen, F., & Ou, T. (2009). Gray relation analysis and multilayer functional link network sales forecasting model for perishable food in convenience store. *Expert Systems with Applications*, *36*(3), 7054–7063.

Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S. (2018). Intelligent sales prediction using machine learning techniques. In *2018 international conference on computing, electronics & communications engineering (iccece)* (pp. 53–58).

Choi, T.-M., Hui, C.-L., Liu, N., Ng, S.-F., & Yu, Y. (2014). Fast fashion sales forecasting with limited data and time. *Decision Support Systems*, *59*, 84–92.

Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, *27*(10), 1749–1769.

Deng, J. (1989). Introduction to grey theory system. *The Journal of Grey System*, *1*(1), 1–24.

Derby, N. (2018). Reducing customer attrition with machine learning for financial institutions. *Proceeding soft he SAS GlobalForum*, 1796.

Dey, A., Singh, J., & Singh, N. (2016). Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis. *International Journal of Computer Applications*, *140*(2), 27–31.

Enolac5. (2018, May). *Time series - arima, dnn, xgboost comparison.* http://www.kaggle

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning–a comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & service operations management*, *18*(1), 69–88.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701.

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1283–1318.

Fisher, M. L., Raman, A., & McClelland, A. S. (2000). Rocket science retailing is almost here-are you ready? *Harvard Business Review*, *78*(4), 115–123.

Fisher, M., & Raman, A. (2018). Using data and big data in retailing. *Production and Operations Management*, *27*(9), 1665–1669.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gamboa, J. C. B. (2017). Deep learning for time-series analysis. *arXiv preprint arXiv:1701.01887*.

Hamzaçebi, C. (2008). Improving artificial neural network sâ performance in seasonal time series forecasting. *Information Sciences*, *178*(23), 4550–4559.

Ho, S. L., & Xie, M. (1998). The use of arima models for reliability forecasting and analysis. *Computers & industrial engineering*, *35*(1-2), 213–216.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hsu, C.-C., & Chen, C.-Y. (2003a). Applications of improved grey prediction model for power demand forecasting. *Energy Conversion and management*, *44*(14), 2241–2249.

Hsu, C.-C., & Chen, C.-Y. (2003b). Applications of improved grey prediction model for power demand forecasting. *Energy Conversion and management*, *44*(14), 2241–2249.

Hsu, L.-C. (2009). Forecasting the output of integrated circuit industry using genetic algorithm based multivariable grey optimization models. *Expert systems with applications*, *36*(4), 7898–7903.

Hsu, L.-C., & Wang, C.-H. (2007). Forecasting the output of integrated circuit industry using a grey model improved by the Bayesian analysis. *Technological Forecasting and Social Change*, *74*(6), 843–853.

Hyndman, R., Lee, A., Wang, E., & Wickramasuriya, S. (2018). *hts: Hierarchical and grouped time series. R package version 5.1. 5.*

Jain, A., Menon, M. N., & Chandra, S. (2015). Sales forecasting for retail chains. *San Diego, California: UC San Diego Jacobs School of Engineering*.

Kaggle.com. (2019, September). *Store item demand forecasting challenge.* http://https:// www.kaggle.com/competitions/demand-forecasting-kernels-only/ data.

Kalaoglu, Ö. I˙., Akyuz, E. S., Ecemis¸, S., Eryuruk, S. H., Sümen, H., & Kalaoglu, F. (2015).

Kotzur, L., Markewitz, P., Robinius, M., & Stolten, D. (2018). Impact of different time series aggregation methods on optimal energy system design. *Renewable energy*, *117*, 474–487.

Lei, M., & Feng, Z. (2012). A proposed grey model for short-term electricity price forecasting in competitive power markets. *International Journal of Electrical Power & Energy Systems*, *43*(1), 531–538.

Levy, M., Weitz, B. A., Grewal, D., & Madore, M. (2012). *Retailing management* (Vol. 6).

Li, Z., Ma, X., & Xin, H. (2017). Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis today*, *280*, 232–238.

Lin, Y.-H., & Lee, P.-C. (2007). Novel high-precision grey forecasting model. *Automation in construction*, *16*(6), 771–777.

Liu, N., Ren, S., Choi, T.-M., Hui, C.-L., & Ng, S.-F. (2013). Sales forecasting for fashion retailing service industry: a review. *Mathematical Problems in Engineering*, *2013*.

Lo, T. (1994). An expert system for choosing demand forecasting techniques. *International Journal of Production Economics*, *33*(1-3), 5–15.

Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B., & Zhang, H. (2014). Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian naive bayes. *PloS one*, *9*(1), e86703.

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*(1), 111–128.

Manaswi, N. K. (2018). Understanding and working with keras. In *Deep learning with applications using python* (pp. 31–43). Springer.

Mancuso, P., Piccialli, V., & Sudoso, A. M. (2021). A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, *182*, 115102.

McKinney, W. (2012). *Python for data analysis: Data wrangling with pandas, numpy, and ipython*. "O'Reilly Media, Inc.".

Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, *18*(3), 11–11.

Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach. In *Operations research forum* (Vol. 3, pp. 1–22).

Ofoegbu, K. (2021). *A comparison of four machine learning algorithms to predict product sales in a retail store* (Unpublished doctoral dissertation). Dublin Business School.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, *26*(1), 217–222.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1), 15.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., others (2011). Scikit-learn: Machine learning in python. *The Journal of machine learning research*, *12*, 2825–2830.

Prudêncio, R. B., & Ludermir, T. B. (2004). Meta-learning approaches to selecting time series models. *Neurocomputing*, *61*, 121–137.

Retail demand forecasting in clothing industry. *Textile and Apparel*, *25*(2), 172–178.

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, *27*, 111–125.

Sakai, H., Nakajima, H., Higashihara, M., Yasuda, M., & Oosumi, M. (1999). Development of a fuzzy sales forecasting system for vending machines. *Computers & industrial engineering*, *36*(2), 427–449.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. doi: 10.1147/rd.33.0210

Sanders, N. R., & Graman, G. A. (2009). Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega*, *37*(1), 116–125.

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006.

Schaeffer, S. E., & Sanchez, S. V. R. (2020). Forecasting client retentionâa machine-learning approach. *Journal of Retailing and Consumer Services*, *52*, 101918.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13–22.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.

Sial, A. H., Rashdi, S. Y. S., & Khan, A. H. (2021). Comparative analysis of data visualization libraries matplotlib and seaborn in python. *International Journal*, *10*(1).

Su, X., Yan, X., & Tsai, C.-L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275–294.

Swami, D., Shah, A. D., & Ray, S. K. (2020). Predicting future sales of retail products using machine learning. *arXiv preprint arXiv:2008.07779*.

Tanaka, K. (2010). A sales forecasting model for new-released and nonlinear sales trend products. *Expert Systems with Applications*, *37*(11), 7387–7393. Tosi, S. (2009). *Matplotlib for python developers*. Packt Publishing Ltd.

Tsao, Y.-C., Chen, Y.-K., Chiu, S.-H., Lu, J.-C., & Vu, T.-L. (2022a). an innovative demand forecasting approach for the server industry. *Technovation*, *110*, 102371.

Tsao, Y.-C., Chen, Y.-K., Chiu, S.-H., Lu, J.-C., & Vu, T.-L. (2022b). an innovative demandforecasting approach for the server industry. *Technovation*, *110*, 102371.

Tsoumakas, G. (2019a). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, *52*(1), 441–447.

Tsoumakas, G. (2019b). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, *52*(1), 441–447.

Valbuena, R., Hernando, A., Manzanera, J. A., Görgens, E. B., Almeida, D. R., Silva, C. A., & García-Abril, A. (2019). Evaluating observed versus predicted forest biomass: R- squared, index of agreement or maximal information coefficient? *European Journal of Remote Sensing*, *52*(1), 345–358.

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in science & engineering*, *13*(2), 22–30.

Vincent, T. (2021, November). *Arima time series data forecasting and visualization in python*.www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecastingwith-arima-in-python-3.

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE signal processing magazine*, *26*(1), 98–117.

Wei, Z., & Shan, X. (2019). Research on forecast method of railway passenger flow demand in pre-sale period. In *Iop conference series: Materials science and engineering* (Vol. 563, p. 052080).

Winters, P. R. (1960a).Forecasting sales by exponentially weighted moving averages. *Management science*, *6*(3), 324–342.

Winters, P. R. (1960b).Forecasting sales by exponentially weighted moving averages. *Management science*, *6*(3), 324–342.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67–82.

Wong, W. K., & Guo, Z. (2010). A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. *International Journal of Production Economics*, *128*(2), 614–624.

Yao, A. W., Chi, S., & Chen, J. (2003). An improved grey-based approach for electricity demand forecasting. *Electric Power Systems Research*, *67*(3), 217–224.

Yoo, T.-W., & Oh, I.-S. (2020). Time series forecasting of agricultural products sales volumes based on seasonal long short-term memory. *Applied Sciences*, *10*(22), 8169.

Yucesan, M., Gul, M., & Erkan, E. (2017). Application of artificial neural networks using Bayesian training rule in sales forecasting for furniture industry. *Drvna industrija*, *68*(3), 219–228.

Zhang, C., Zhang, H., Sun, Q., & Liu, K. (2018). Mechanical properties of zr41. 2ti13. 8ni10cu12. 5be22. 5 bulk metallic glass with different geometric confinements. *Results in Physics*, *8*, 1–6.

Zhuge, Q., Xu, L., & Zhang, G. (2017). Lstm neural network with emotional analysis for prediction of stock price. *Engineering letters*, *25*(2).

License