**A complement solution for retail shops**

Arti Sharma

Computer Science

*KIET Group of Institutions*

Ghaziabad, India

0000-0001-6540-1639

Priyambda Pandey

Computer Science

*KIET Group of Institutions*

Ghaziabad, India

priyambda.\*\*\*@kiet.edu

Aman Malik

Computer Science

*KIET Group of Institutions*

Ghaziabad, India

aman.1923co1031@kiet.edu

Gauri Dargar

Computer Science

*Kiet Group of Institutions*

Ghaziabad, India

[gauri.1923co1064@kiet.edu](mailto:gauri.1923co1064@kiet.edu)

***Abstract*—** **The issue of difficulties in controlling the inventory of small retailers with limited space and limited capital to buy just a few merchandises has just been led due to the continuous increase in the variety of products. In order to successfully manage the inventory of small stores, this article intends to investigate several machine learning algorithms for forecasting future product demand. [1] Demand forecasting is essential for the management of retail logistics procedures. The effectiveness of procedures not only in retail outlets but also throughout the supply chain they are a part of is highly influenced by the level of the forecasts which are provided. Now days, there is a large amount of information that can be cleansed and used for forecasting by utilizing the right machine learning algorithms.**

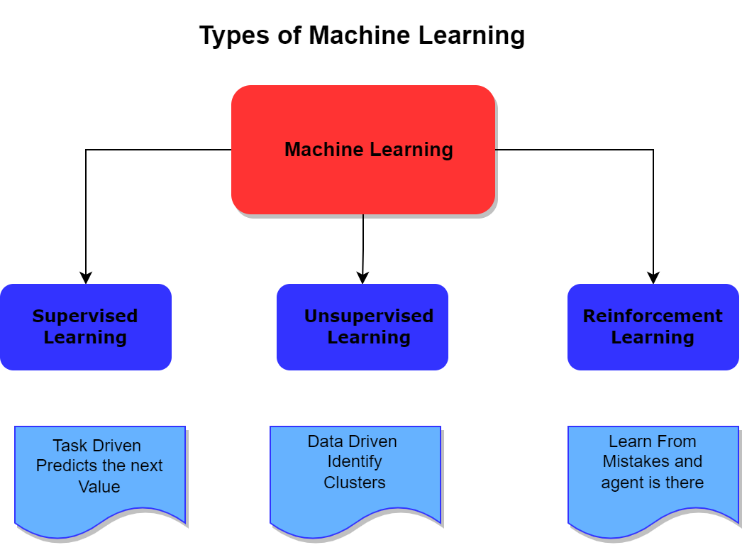
**This prediction system will help small retail shop owners in maximizing their profit margins. Along with forecasting sales, we aim for customer segmentation to take effective decisions to generate a good amount of revenue**.

Keywords—Machine Learning; Data Cleaning Prediction System; Demand Forecasting, customer segmentation, KNN algorithm.

# **Introduction**

Machine learning is a sub-domain of artificial intelligence that can be used to learn from past experiences and use that knowledge to take decisions based on the previous learning experiences[2]. Models that are generated using machine learning can become better with time as they learn more.

The basic idea behind machine learning is to predict outcomes without exactly being programmed to do so. Machine learning can be classified into three main types viz- supervised, unsupervised and reinforcement learning.



1. Types of Machine Learning Techniques

Because of its large amount of data and the various optimization problems it tries to present, which include optimal prices, discounts, recommendations, and stock levels, that can be settled using data analysis methods, retail is one of the most substantial business domains for machine learning and data mining applications. Making predictions for sales revenue over a predetermined period of time is known as forecasting models. Typically, records as well as other analytical information are taken account in sales forecasting to provide the most precise predictions. In order to fully understand and estimate customer demand, as well as other major financial factors, and to help businesses make critical supply decisions that will maximize profitability, sales prediction models specifically utilize predictive analytics on historical information.

As a by-product, several machine-learning methods have been created for this purpose. Using various machine-learning techniques, we attempt to calculate the revenues of a retail market in this paper. We also seek to identify the optimal algorithm for our specific problem statement. In our approach, we've included both conventional regression methods and boosting techniques, and we've found that the boosting algorithms produce superior results to the traditional regression algorithms.

Every retailing achievement depends on its ability to properly providing accurate at every significant retailer. Inventory management is assisted, product distribution among stores is better, over- and understocking at specific locations is reduced, reduced system, and most importantly, customer satisfaction and sales were significantly increased. Because of the increased competition in the retail industry these days, we should focus on other customer interests in in addition to sales forecasting. To split the huge data in the dataset into various groups as according their age, gender, purchase behaviour, income, and other characteristics, customer segmentation is necessary. These groupings are also alluded to as clusters. This enables it for us identify the age groups and products that sold in massive amounts.

Small retail shops can benefit from forecasting and customer analysis to efficiently manage their inventory the demand for the products will be predicted based on historical data.

# **RELATED WORKS**

[7] This paper discusses market and customer segmentation and techniques to achieve greater customer satisfaction through customer segmentation. Relationships between different variables are formed and tested. The importance of customer satisfaction is also described in the paper. A relationship is formed between customer satisfaction and different parameters like price, place, product, and market segmentation. The market needs to be carefully analyzed to segment customers into groups based on different parameters and these parameters need to be selected carefully. The paper puts forward how marketing mix can aid in market segmentation. The purchasing behavior of a customer can be predicted based on previous data and demographics of that customer. Thus, can be utilized by business firms like banks and manufacturing companies to plan their strategies.

[9] There is much more competition in the market than ever due to globalization. The major problem before companies is that they cannot visualize all the customers as a single entity. Here comes the role of dividing the market into segments, those segments of customers can be targeted individually rather than a one large segment. The paper discusses the steps involved in customer segmentation and different strategies to achieve better results. The customers can be divided into groups based on different characteristics like geographic situation of the customer, demographics, psychographics, and behavioral characteristics. It also discusses the benefits of segmenting customers which include avoiding losses, targeting specific markets, less competition, maximizing profits and many more. There is a need of extensive research in the field of market segmentation in different areas.

[10] The companies are facing huge losses due to wrong prediction trends of the market situation. This can be curbed by optimizing the operations of the supply chain. This paper discusses the main problems that arise due to overstock and stockouts. These include loss in revenue, lower customer contentment, etc. The paper proposes a solution in the form of improved demand forecasting based on historical data through various machine learning algorithms. Different machine learning algorithms such as time series forecasting, deep learning, etc. are employed and results are recorded, and an integration strategy of different models is used to predict future demands of products. After the integration of eleven different machine learning models the results were found to be promising and exciting.

[13] There is an increase in online purchases than ever recorded in history and we have huge amounts of customer data and sales data that is being generated every hour in very large quantities. The paper describes in detail the importance of predicting sales. There are many different algorithms for predicting sales using time series forecasting that are investigated. The main goal was to achieve increased accuracy in predicting future sales. Different machine learning models are examined, and their results are combined to achieve greater accuracy in prediction. It was found that regression algorithms performed better when compared to other algorithms.

[3] As there is an increase in number of customers and number of products rapidly as an effect of the competitive market and the world becoming a global village. It is quite burdensome to keep track of customers and visualize the whole market as one is not a piece of cake. The paper talks about how it is not possible to implement the same marketing strategy for all the customers, also all the customers don’t have same behavior, demographics, etc. Thus, there is a need of dividing the customers into groups or segments that can be targeted individually as a whole, thus maximizing revenue. K -means clustering algorithm is used for customer segmentation.

[4] There is more need for marketing strategies today than ever. Targeting customers has become more complex and laborious. The customers have many choices to choose from, the market is more contentious than ever. The paper examines customer segmentation as a great means of marketing. The more we know about the customer the more we can predict his buying patterns. Moreover, we don’t have to formulate a new strategy for every customer, but we can divide customers into different clusters using machine learning algorithms. K-means clustering algorithm is employed and performed well. Customer segmentation is a way to improve customer satisfaction and interaction with the customer.

[5] Production planning is a very important aspect in business endeavors. This paper discusses ARIMA model and LSTM model for predictive analysis of future sales of products. First ARIMA and LSTM are employed separately,

Then a hybrid model is tested which is a combination of ARIMA and LSTM both. It was found that the hybrid model performed better than the individual models in predicting future sales.

[6] The paper shows comparison of classical algorithms and machine learning algorithms in sales forecasting. The algorithms are judged based on root mean squared error (RSME), symmetric mean absolute percentage error (sMAPE) and mean absolute percentage error (MAPE). It was found that XGBoost outperformed all the algorithms. LSTM also performed quite well and is close to XGBoost. The ML algorithms were found to be more accurate in prediction than classical algorithms .it was also found that the efficiency of ML algorithms was more with larger datasets.

# **PROPOSED METHODOLOGY**

We have brought forth a strategy that utilizes user segmentation with sales forecasting to assist retail shop owners make more money.

We was using the dataset provided by the Kaggle website to perform our analysis. We will implement K-Means Clustering to our problem statement, which separates the data into different clusters based on their shared characteristics. The data will thereafter be displayed. Along with utilizing regression models and LSTM network for sales forecasting, we also employed the ARIMA model to analyze data from time series.

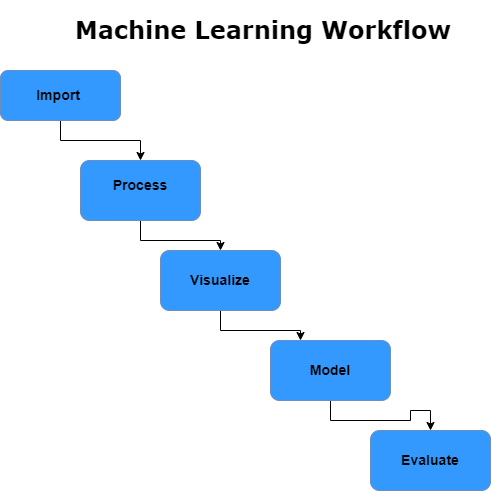


Fig 2. A basic machine learning workflow

Step 1: Data Collection

The first step is to collect data. For our work, we selected special set from Kaggle, which is split into two parts and contains a lot of info. One section focuses on customer segmentation and contains the product's information in addition to customer demo graphs. The other section is intended for forecasting and contains data on the types of products offered within every store and their specific sales. The customer information consist of approximately 25000 distinct customers plus their orders.

Step 2: Data Cleaning, Preprocessing, and EDA

After collecting the data, we need pre-process raw data values to discover its relevant characteristics. The main processing step is aggregating the data on a monthly for the time series forecasting process. The quantity of goods ordered, the average annual return rate, and the amount spent are some attributes that must be provided for the consumer segmentation part.

Step 3: Statistics, Machine Learning

Diverse statistical and machine-learning methodologies are used during the predictive analytics process. The two most important approaches that are usually utilized in analysis are regression analysis and probability theory. Similar to natural neural networks, decision trees are a machine learning methodology that is usually utilized in a variety of predictive analytics activities. The basis of every predictive analytics model is statistical and/or machine learning methods. In order to develop prediction models, analysts use the principles of machine learning and statistical methods. Machine learning approaches have an advantage over traditional statistical techniques, however any predictive algorithm must use statistical techniques. The forecasts in our situation are technically referred to as forecasting because we are working with time-series data.

Step 4: Models for forecasting sales.

This phase covers training the data on various regression models using machine learning and statistical methods using an example dataset. After development, the is tested on the test dataset, which is a sample of the main dataset collected, to determine its validity. If successful, the model is considered to be fit. Once fitted, the model is able to generate precise projections based on fresh data being fed into the system. The multi-model method has been chosen for a problem in many applications, and our problem is one of them.

* The best results were obtained from the XGBoost and LSTM models
* All models predicted within 2% of monthly mean sales for 12-month prediction.

Step 5: Segmenting the customers

Customer segmentation will be applied to the customer info using the K-means Clustering algorithm from python’s Scikit-learn.

**3.1 K-Meaans Clustering**

(Clustering via K-Means)

The K Means is a centroid-based clustering algorithm , which is iterative and strives to split the dataset into K distinct, non-overlapping clusters, where K is the total number of clusters present. In this Algorithm, the data values are clustered in such a manner that the sum of the squared distance which is mainly the Euclidean distance between the data points and centroid has to be minimized.

Every point is a part of a single cluster, but clusters cannot cross over.

**The algorithm's steps**

* Assign random starting items from the set D as the initial centroids.
* Repeat.
* Using the mean value of the objects in the cluster, put each object in the group to which it is most similar.
* Modify the cluster means by figuring out the overall average of the objects for each cluster.

**The following packages have been installed for the code execution-:**

* Pandas – Loading & preprocessing of Data
* NumPy – Numeric Data calculation
* Seaborn, Matplotlib and Plotly- Data Visualization
* Scikit- Machine Learning Library

**Customer Data Information:**

The various features in the dataset for customer segmentation includes-

Product Title

Product Type

Customer id

Order id

Net quantity

Gross sales

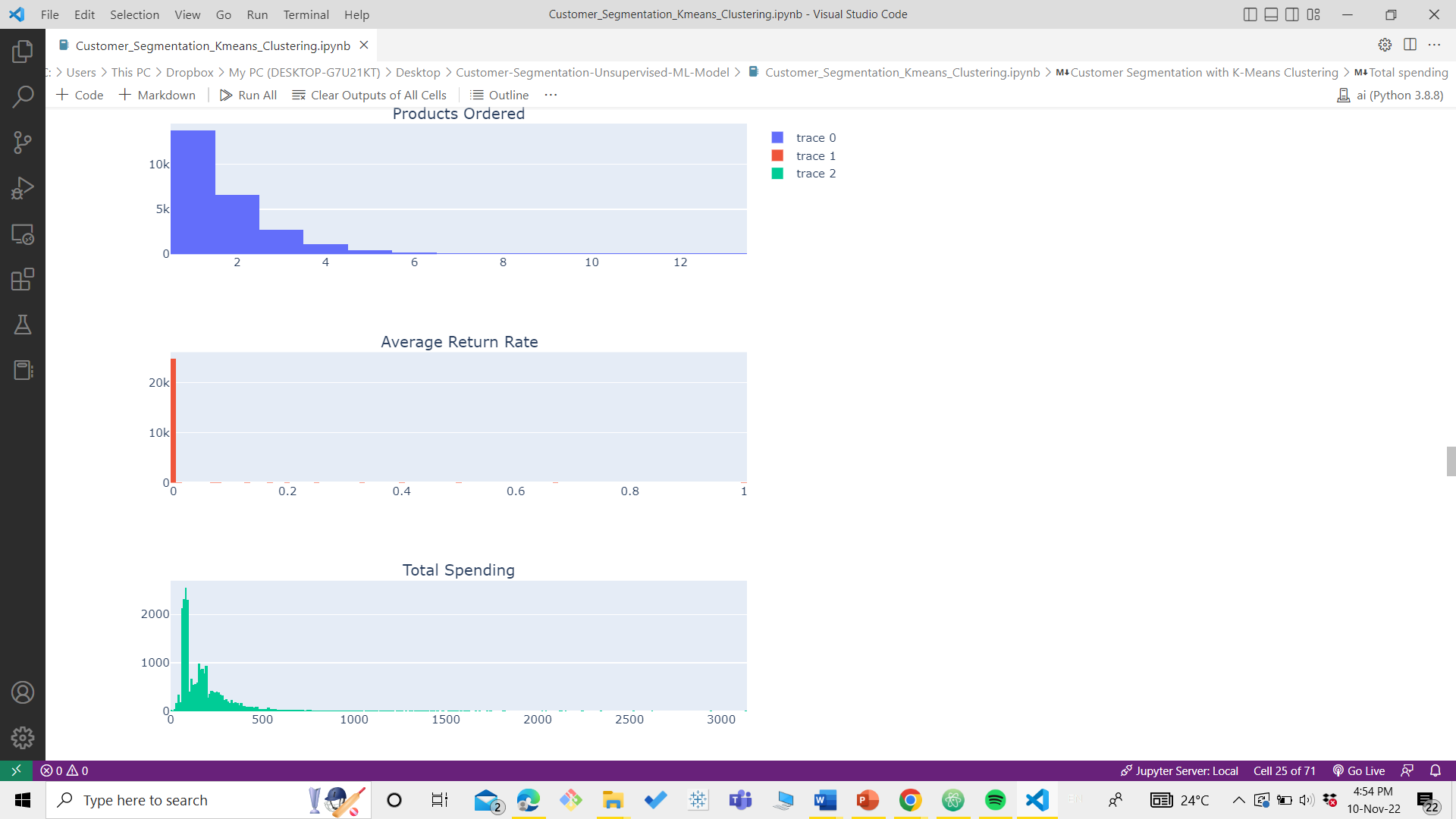
Day

Taxes and many more.

As mentioned above the features created that will be used for the segmentation are:

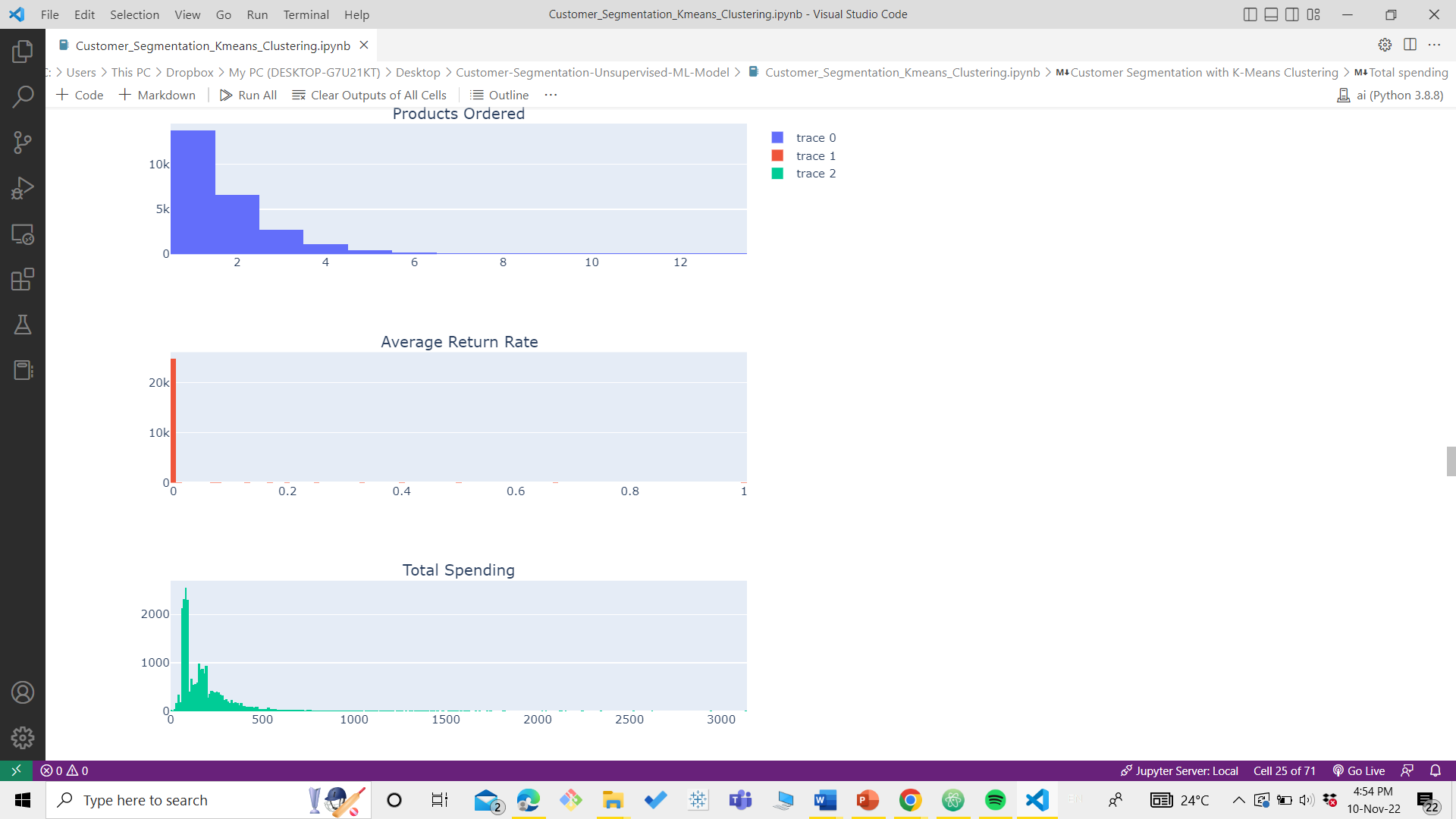
1. Products ordered or items purchased

It is the count of the products ordered in the product\_type column by a customer.



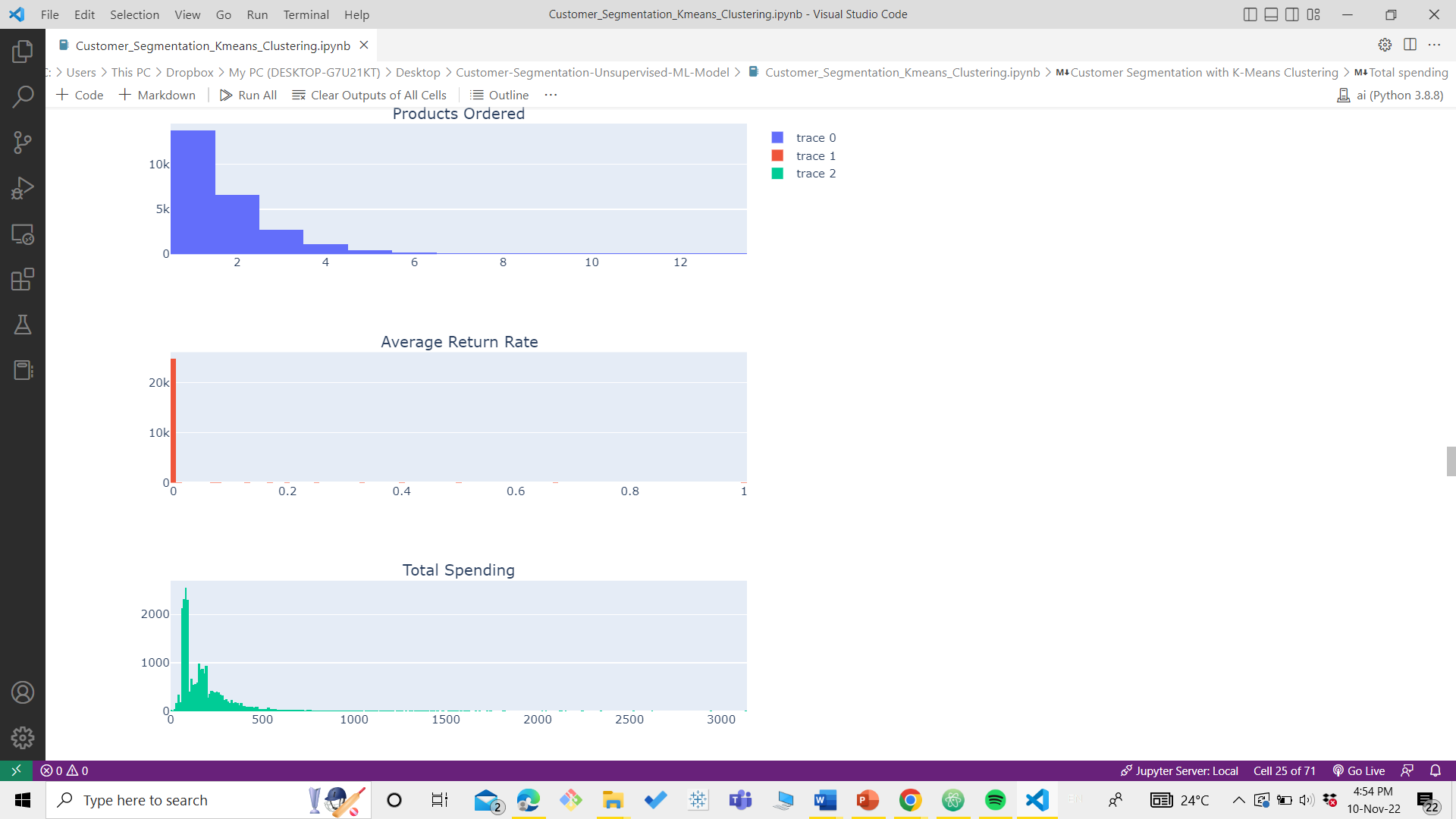
1. Average Return Rate:

It is the ratio of returned products to total orders. This ratio is initially calculated for every order, then it is averaged across each of a customer's orders.



1. Total Spending

Total spending is the aggregated sum of total sales value which is the amount after taxes and returns.



After calculating the main parameters to segment the customers we perform some feature Scaling.

Feature Scaling means to standardize the data in a fixed range, Scaling the features helps to smooth the gradient descent flow and aids algorithms in swiftly reaching the cost function minima. Without scaling features, the algorithm may be inclined toward the feature with higher magnitude values. Some of these techniques which are commonly used are

* Absolute Maximum Scaling
* Min-Max Scaling
* Normalization

In this project we have used **Log Transformation** in this each variable is replaced with its log value. Log transformation is used to make highly skewed distribution less skewed.

Using the K-means clustering algorithm we made a list of several K values and fitted our data on the values, from the list of 15 K values 4 was chosen to be as the optimal one.

**3.2 Sales Forecasting**

The other technique that maximize the profits we used is sales forecasting, with making predictions for future sales the shop owners can analyze the products that will give them better profits. A sales data is collected from the web which was preprocessed and fitted to various models from Linear regressions to LSTMs

**Data Used:**

The raw data that is used comprise of:

Date of purchase

Store number

Item id

Sales

During the preprocessing phase we modified the data as Monthly sales and Difference between the sales. This approach was adopted to get a better prediction from our data values.

We have generated csv file where each row represents a month and columns include sales, the dependent variable, and prior sales for each lag. Based on EDA (Exploratory Data Analysis), 12 lag features are generated. Data is used for regression modelling.

*Output*

*month1 sales  lag1  lag2  lag3 ... lag11 lag12*

**Training and testing data**

The train and test datasets are obtained with the help of train\_test\_split function provided in sklearn library.

**Data Scaling**

For the data scaling we have used Min Max Scaler which

Shrinks the data within a given range, It is present in the sklearn library of python.

The transformation is given by the formula:

**X scaled  = (X - X min)/ (X max – X min)**

With the scaled train and test data we moved forward for training the data on Linear Regression, Random Forest Classifier, XGBoost and LSTMs and we have calculated RMSE, MAE and R2 scores for each of them.

**Models for forecasting**

**1.Linear Regression**

Regression analysis is a statistical method that is used to determine the relationship between two variables that are reliant on one another and are independent in terms of both cause and effect. The basic goal of univariate regression is to build the linear connection equation between the dependent and independent variables and to assess the relationship between a dependent variable and one independent variable. Multi-linear regressions are regression models with one dependent variable and many independent variables[11].

When used to fit our dataset the results for root mean squared error ,mean absolute error and R2 found out to be:

RMSE: 16230.030810582312

MAE: 12429.0

R2 Score: 0.9908165968605961

Comparison between actual and predicted values are shown below in fig 5.

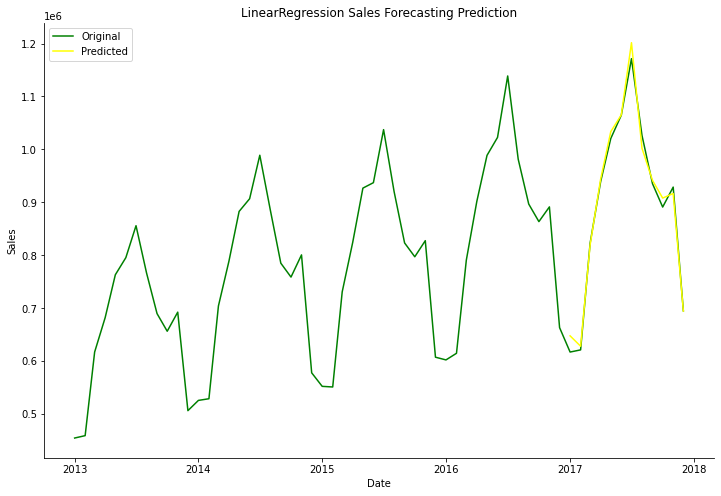


Fig 5. Comparison between actual and predicted values using a linear regression model.

**2.Random Forest Regression:**

Leo Breiman invented Random Forest. The features chosen throughout the induction procedure are at random. The predictions of the ensemble are combined (by majority vote for classification or by average for regression).[12]

The time series dataset must first be transformed into a supervised learning problem in order to employ Random Forest for time series forecasting. Additionally, k-fold cross-validation would result in results that are heavily biassed in favour of optimism, necessitating the employment of a specialised form of model evaluation known as walk-forward validation.

The root mean squared error ,mean absolute error and R2 score are found out to be

RMSE: 18035.41487139493

MAE: 15064.916666666666

R2 Score: 0.9885224485928914

The root mean squared error increased ,mean absolute error also increased while R2 score decreased .we have to move on to another model for better results with low error

Comparison between actual and predicted values are shown below in fig 6.

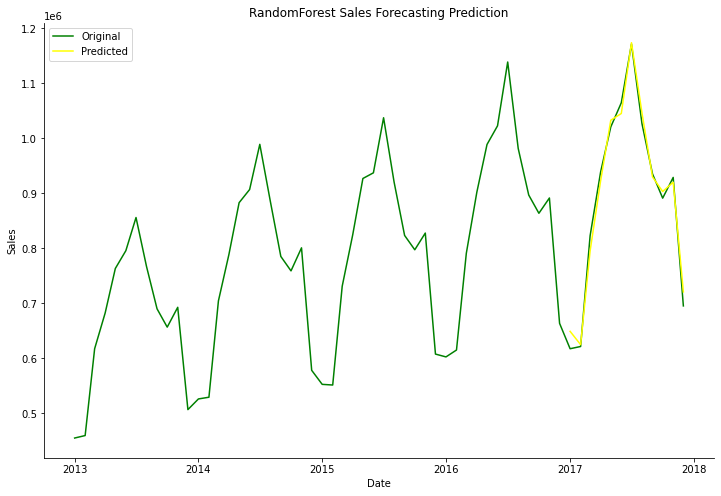


Fig 6. Comparison between actual and predicted values using a random forest regression model.

**3.XGBoost for Regression**

Extreme Gradient Boosting commonly known as XGBoost.It is written in C++ programming language to optimize the training of Gradient boosting the model is present in Python’s scikit learn library as well on platforms like Hadoop. It is frequently employed in a range of data mining situations and algorithmic contests. The algorithm's key benefits are its accuracy, adaptability, and automated handling of missing values. Gradient Boosting is implemented by XGBoost, which linearly combines several weak classifiers into a powerful classifier. XGBoostRegressor uses a number of trees to predict the vable of the dependent variable. The decision trees are combined to form a strong learner. XGBoost provides the concurrent selection of affect the output for faster operation, and model training takes substantially less time.[13].

XGBoost model gave the following results for root mean squared error ,mean absolute error and R2 score after fitting in our dataset.The root mean squared error decreased,mean absolute error also decreased while the R2 score incresed. We should explore other models which may give better results than this model.

RMSE: 15701.003359658262

MAE: 13342.666666666666

R2 Score: 0.9913013514225064

Comparison between actual and predicted values are shown below in fig 7.

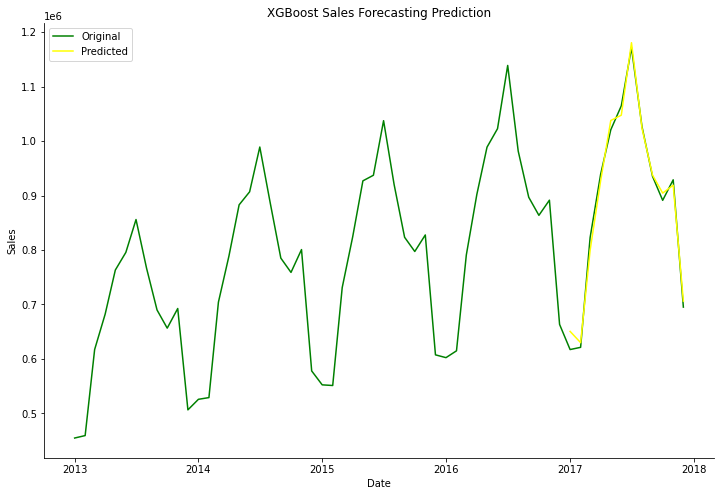


Fig 7. Comparison between actual and predicted values using decision xgboost regression model

1. **LSTM Network**

The Long Short-Term Memory Network (LSTM) is a particular kind of RNN that is good for processing time series and crucial events with a relatively long latency, but it is simple to slip into the local optimum. [14] The learning of long term dependencies in the data is achieved by the combination of four interacting layers.

The fundamental goal of LSTM cells is to remember the critical portions of the sequence that have already been seen while forgetting the less critical portions this is achived by so called Gates.

Here is a basic understanding of the architecture of LSTM is based on the concept of gates:

1. Forget Gate: dumps the information which is not useful
2. Learn Gate: the Event and the STM are grouped to apply the recently learned information to the input.
3. Remember Gate: Events are combined in this gate and works as updated LTM
4. Output Gate: to predict the output of the current event.

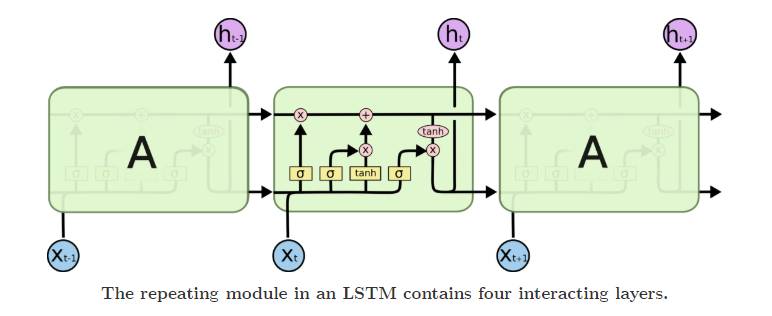


Fig 8. Basic Architecture of LSTM network [16]

Our LSTM is a sequential model the layers used are the basic LSTM layer with 4 units.

2 Dense layers for the network

Adam optimizer

The model is fit with the batch size of 1 and epochs for training are 200.

The scores obtained are as follows:

RMSE: 16483.005609313692

MAE: 12860.083333333334

R2 Score: 0.9904132856830274

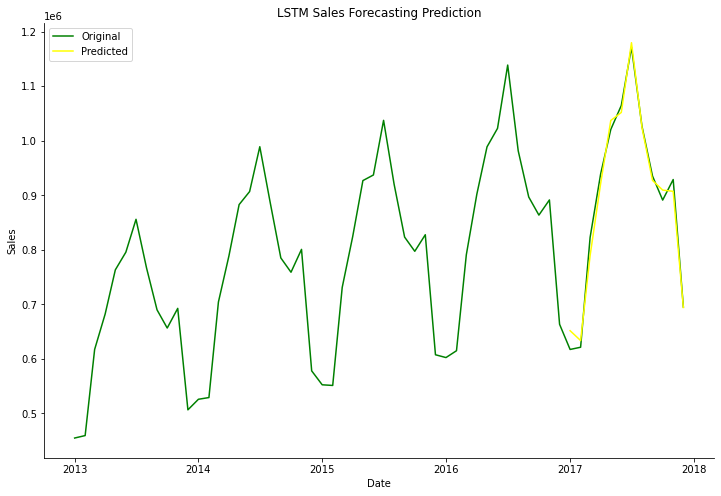


Fig 9. Comparison between actual and predicted values using decision LSTM model

**Comparison Table of the Models**

We have a comparitive study of the erros and score of the models on which we have tested our dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm Applied** | **Root mean squared error (RMSE)** | **Mean Absolute Error(MAE)** | **R2 Score** | **Remarks** |
| Linear Regression | 16230.0308 | 12429.0 | 0.9908 | High RMSE value |
| Random Forest Classifier | 18035.4148 | 15064.91 | 0.9988 | A very high error scores not a good fit |
| XGBoost | 15701.003 | 13342.66 | 0.991 | Performed much better than the above two |
| LSTMs | 16483.0 | 12860.083 | 0.9904 | Considered to be the best among all |

The R2 score (which is cofficient of Determination) it is a statistical measure that indicates the extend of variation in a dependent varaible due to an independent one in other words it represents the goodness of fit of a regression model. For all the models is approx to 0.99 and this falls under the accepted range.

**Time Series Forecasting**

**ARIMA**

Autoregressive Moving Average (ARIMA). Moving Average Using Autoregressive Regression (ARIMA). The ARIMA model is superior to the early AR, MA, and ARMA models in terms of application flexibility and the accuracy of the results it produces when simulating or forecasting.[15]

The ARIMA model predictions works on moving average where there is no constant mean of the data points, here the straight ARMA model cannot be used. In ARIMA model it predicts the difference from the time series from one timestamp to previous timestamp and we see to hover around some mean value. That’s how the problem of time series not being stationary is solved.

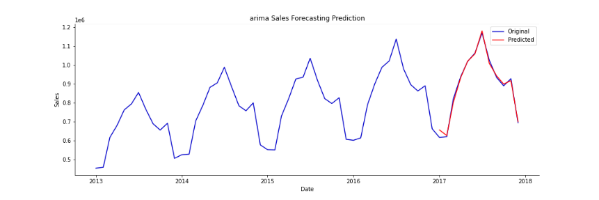


Fig 9.

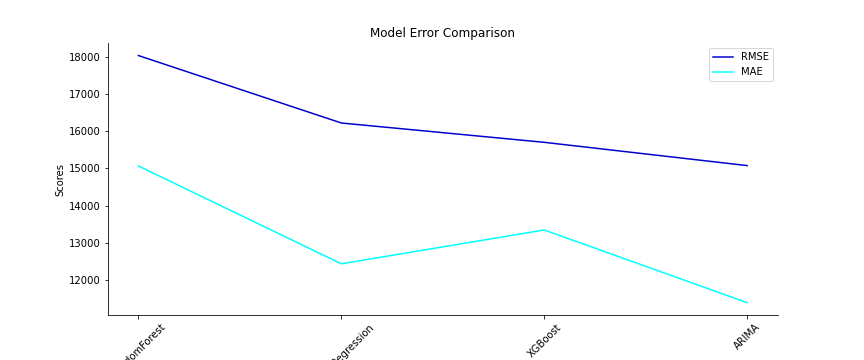
Now it has the same constant mean over time and has no seasonality. The ARIMA model takes 3 parameters i.e (p,d,q)

p- is the number of autoregressive terms

d- is the number of nonseasonal differences to remove seasonality

q- is the number of lagged forecast errors

**IV. RESULTS**



As it is known that a single algorithm never works best for all datasets. The accuracy of model varies as per the data values but we have tried our model on Linear regression, XGBoost, random forest, LSTM and the ARIMA (which is used for time series data. Among these XGBoost outperformed for multivariate and intermittent data.

Overall the LSTMs have performed better than the other models with minimum avg error scores.

**V. CONCLUSION**

We come up with a complement literature review based on various machine learning techniques used in the sales forecasting is provided in this paper. The main implementation involves forecasting the sales of products and segmenting the customers in order to mazimise the profits and sales. It will use the concepts of machine learning in predicting and clustering similar data points. Various machine learning models for predictive analysis have been tested and analyzed.XGBoost regression model gave the best performance in predicting future sales. we can predict future demand for products. This prediction system will be used to recommend products to small retail shops to keep in their inventory to help them maximize their profit.

# **References**

[1]. N.p., Web. 28 Nov. 2022.  
https://kkhsou.ac.in/eslm/E-SLM\_Main/1st Sem/Bachelor Degree/BBA/GBA S1 02 (Block 1).

[2] Çelik, Özer. (2018). A Research on Machine Learning Methods and Its Applications. 10.31681/jetol.457046.

[3] Patankar, Nikhil & Dixit, Soham & Bhamare, Akshay & Darpel, Ashutosh & Raina, Ritik. (2021). Customer Segmentation Using Machine Learning. 10.3233/APC210200.

[4] N, Hua & BA, Leu & Kumar, Rahul. (2022). Machine learning-based customer segmentation. International Journal of Information Security. 9.

[5] Yurtsever, Mustafa & çiçekli, Ural. (2022). A NOVEL HYBRID MODEL USING ARIMA AND LSTM FOR SALES FORECASTING.

[6] Haselbeck, Florian & Killinger, Jennifer & Menrad, Klaus & Hannus, Thomas & Grimm, Dominik. (2022). Machine Learning Outperforms Classical Forecasting on Horticultural Sales Predictions. 7. 100239. 10.1016/j.mlwa.2021.100239.

[7] Puwanenthiren Premkanth, - Market

Classification and Its Impact on Customer

Satisfaction and Special Reference to the

Commercial Bank of Ceylon PLC.‖ Global

Journal of Management and Business Publisher

Research: Global Magazenals Inc. (USA). 2012.

Print ISSN: 0975-5853. Volume 12 Issue 1

[8] Sulekha Goyat. "The basis of market

segmentation: a critical review of the literature.

European Journal of Business and Management

www.iiste.org. 2011. ISSN 2222-1905 (Paper)

ISSN 2222-2839 (Online). Vol 3, No.9, 2011

[9] Kilimci, Zeynep Hilal, et al. "An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain." *Complexity* 2019 (2019).

[10] vlyshenko, B. M. (2019). Machinelearning models for sales time series

forecasting. Data, 4(1), 15.

[11] Kaya Uyanık, Gülden & Güler, Neşe. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences. 106. 234–240. 10.1016/j.sbspro.2013.12.027.

[12] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.

[13] Zhang, Lingyu & Bian, Wenjie & Qu, Wenyi & Tuo, Liheng & Wang, Yunhai. (2021). Time series forecast of sales volume based on XGBoost. Journal of Physics: Conference Series. 1873. 012067. 10.1088/1742-6596/1873/1/012067.

[14] Yuzhen Wang et al 2019 J. Phys.: Conf. Ser. 1314 012215

DOI 10.1088/1742-6596/1314/1/012215

[15] AIP Conference Proceedings 1931, 030017 (2018); <https://doi.org/10.1063/1.5024076>

[16] http://colah.github.io/posts/2015-08-Understanding-LSTMs/