Predicting sales of the E-commerce industry with Brazilian dataset using Deep learning algorithms

MSc Research Project
Data Analytics

Harsh Gupta
Student ID: x23173815

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Harsh Gupta
Student ID:	x23173815
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Prof. Jorge Basilio
Submission Due Date:	12/12/2024
Project Title:	Predicting sales of the E-commerce industry with Brazilian dataset using Deep learning algorithms
Word Count:	6397
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Harsh Gupta
Date:	25th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting sales of the E-commerce industry with Brazilian dataset using Deep learning algorithms

Harsh Gupta
x23173815

Abstract

Sales prediction is an important aspect for the e-commerce industry that involves inventory management, demand forecasting, and strategic planning. There are various challenges encountered by old traditional methods when dealing with dynamic, unstructured, and complex datasets, which affect the prediction accuracy. The deep learning models like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), help in identifying long-term patterns and sequential dependencies. The objective of this research is to propose a novel hybrid model, a combination of the LSTM model and the GRU model, respectively, to improve the prediction of sales in e-commerce, also capturing and ensuring the balance between long-term dependencies and lower computational efficiency. The study focused on the Brazilian e-Commerce data set, using various data preprocessing techniques, feature engineering, and a strong model evaluation to obtain accurate predictions. The results of Hybrid model was superior, when compared to LSTM and GRU architectures. The model achieved better R^2 value of 0.91 and even, the lower error metrics of MAE, MSE, RMSE values. The insights derived shows, Hybrid model's ability of capturing both short-term fluctuations and long-term patterns. This makes model more accurate for prediction of sales. This research contributes widely in the field of e-commerce industry and predictive modeling.

Keywords: deep learning, e-commerce, LSTM, GRU, Hybrid model

1 Introduction

1.1 Background and Motivation

There has been a significant growth in e-commerce industry, compared to other sectors. The role of e-commerce, focus on providing an online platform for seller and consumer. This helps to provide a proper customer convenience and satisfaction. This provides online platform for customers, which is more prominent than the offline stores. As per the (Matuszelański and Kopczewska; 2022), there was around $3.17 billion in 2019, e-commerce sales in the United States. This led to substantial growth in e-commerce industry. Some challenges like COVID - 19 pandemic also interrupted the growth for the companies for maintaining the operations, providing the services, and selling of products (Wei et al.; 2014).

The growth of e-commerce is rising, but there has to be accurate prediction for sales depending on historical data. The main goal of prediction, is to help the organizations for making better strategic decisions, enhancing customer loyalty, as well as demand forecasting. Some aspects based on customers, like habits, their loyalty and geolocation confirms the vital role of influencing sales trends. In the past years, various technologies are developed. One of them is 'Customer Relationship Management' (CRM), responsible for the relationship between customer retention and performance of sales, (Aljbour and Avcı; 2024). The CRM provides robust strategies, but is not able to deal with huge and unstructured datasets nature, which is also seen when developing traditional models.

The dataset for this research work is a public dataset collected form a public repository, with all proper ethical challenges and concerns. The data was produced by Olist store (Olist and Magioli; 2018). The data was collected in a systematic way. The nature of e-commerce industry is very dynamic leading to use of advance computational technique, that are capable for uncovering complex patterns. The old traditional methods fail to capture most of the multivariate dependencies when dealing with large datasets. This influence the optimal performance of models. Machine learning techniques also face various challenges, they are highly scalable but, face problems on handling temporal patterns and very large datasets. The output result is not as accurate as deep learning models. The main purpose of this research work, is to address all these challenges by implementing various deep learning models for enhancing sales prediction. The dataset is rich in various features like status of order, customer locations, customer behaviour, different product categories, trends of different customers. The models are mainly implemented in this research work, for achieving high prediction accuracy in handling large and unstructured datasets, like Brazilian e-commerce dataset. These models can be used in various industrial sectors like e-commerce and influence the prediction of sales, for better strategic decisions (Ginantra et al.; 2023).

Certain limitations have to be acknowledged, even though the research is capable of providing us with valuable insights for implementation of deep learning models. There is a year limitation for dataset from 2016-2018. The dataset is focused for Brazil region, may not be suitable for other regions or industries due to social economic factors. This also, suggests the addition of some external factors like strategic campaigns and marketing decisions.

1.2 Research Objectives

- To capture temporal and sequential dependencies, with help of designing and implementing deep learning models like: LSTM, GRU and a novel Hybrid model.
- To determine the most accurate and robust model for prediction of sales.
- To provide a strong framework which combines preprocessing of data, feature engineering, and evaluation for sales prediction and decision-making.

1.3 Research Question

- Which deep learning model performs best, among LSTM, GRU, and a novel Hybrid model for sales prediction based on evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-square, and Mean Absolute Percentage Error (MAPE)?

2 Literature Review

This section of research provides us with the, studies and information done by researchers and experts. The studies and evidence from past researchers and experts demonstrates, the e-commerce platform for sales prediction has drawn attention of business researchers and directing them for using various datasets and predictive modeling techniques to improve accuracy.

2.1 Purpose and Significance of Brazilian E-commerce Dataset in Sales Prediction

E-commerce is a digital platform, that is easily adopted worldwide across various industries. The evidence provided by (Singh et al.; 2020), shows that, the e- commerce platform is a mediator and provides the sales and purchase convenience to both - service provider and customers while influencing buyer's decisions. The findings also provides the statistical figures on how the online purchases have increased the value worth $603 billion in 2019 for total sales in the US retail industry. This finding gives an idea about e-commerce importance for improving sales across different sectors and achieving a strong business results. This main focus of the study is on Brazilian dataset, stated that global trends can be explored easily with e-commerce insights even though, regional aspects may vary. Forecasting is easy with the help of better sales prediction, providing the organization for strong and correct business decisions. However, this approach face challenges due to extensive data configuration, which affects the accuracy and specificity of the information. The choice of Brazilian e-commerce dataset is important as it provides with various features and aspects with multiple dimensions.

The rich-dataset was introduced and provide by the `Olist` retail store (Olist and Magioli; 2018). This store is one of the largest department stores in the Brazilian marketplace. So basically, `Olist` acts like a hub which links multiple small businesses together, through a single contract. This in return, helps the business for shipping products, and improving sales which is possible because of `Olist` logistics and store. The use of advanced machine learning algorithms is possible because of their well-organized structure as they are region specific. The main insights like, details about individual customer purchases are well-maintained across business segments, which helps for reliable order management and delivery processes. The (Pandey et al.; 2023) explained, customer loyalty towards brand is mostly dependent on continuous business operations and marketing influence. All this information feasibility depicts, that the Brazilian data can enhance customer loyalty with e-commerce contribution in which sales prediction plays a vital role for products and purchase-related attributes. Although the dataset is modern, it provides high sales prediction results with a correct choice of model, as it is relevant dataset which links Brazilian 'zip codes' to 'lat/long coordinates.

2.2 Prediction of E-commerce Sales Using Traditional Methods

The prediction of sales is a very important aspect, as it strongly affects how business make their decisions. Over, the past years, there has been a significant growth in e-commerce, drawing interest and attention for various areas of online business. These areas include-managing the workforce, financial resource management, supply chain optimization and improving sales performance. Considering, the evidence provided by (Wei et al.; 2014) presented that, sales forecasting has become more prominent but, is challenging because of factors like promotion rate, variation of price, and customer (user) preference. For many years, mostly the significance of methods like time-series forecasting based on historical data was taken into consideration. The research provided by (Wei et al.; 2014), tells us, online shopping plays an important role in reaching customers easily. In the past, businesses focused on both online and offline ways for enhancing customer convenience to buy products. But the global crisis like the dilemma of COVID-19 pandemic led to growth of e-commerce. (Wei et al.; 2014) proposed, as the growth of e-commerce has taken place, forecasting future sales have become important.

The use of time-series prediction is one of the approach that rely on web-search data for sales prediction purposes. In a study done by (Wei et al.; 2014), proposed the dataset was split into train and test data, on which the model was trained. Features like trend and seasonal factors were dropped providing more accuracy and specificity of the model. The insights suggests, the average of MAPE was approx. 4.84%. This was done for the forecasting of sales for 7 days. The basic approach of forecasting is provided by traditional models, but fails to compute with complex and unstructured dataset. This also results in the scalability of model in e-commerce industry. The advancement in this research proposed, information by high traffic websites can create problems for optimal accuracy. The study proposed by (Usmani et al.; 2017), there are various challenges for companies because of complex data, and influential behaviour of the consumer. The study introduced a recommender system for handling such issues. The systems are very smart and use advance techniques, and knowledge-based discovery methods for accurate prediction on customer preferences. This suggests, over the years sales prediction methods were evolved from historical-based data and analyzing customer preference or products for meaningful insights.

Coming to the other consideration, done by (Yuan et al.; 2018), sentiment analysis is also, an important thing for consideration. The study suggested, the mining of sentiments of customers from different social media platform for valuable insights. This provides the organization for improved decision- making. For the past years, volume of sales prediction provided solutions for increasing prediction accuracy. Now, there are various challenges for organization for dealing with large data, as traditional models are not that accurate. In a similar study, done by (Yuan et al.; 2014), forecasting the sales from e-commerce is an important step used in recent-times for managing inventory and product development. This study also stated, certain findings may not be suitable because of different complex factors in e-commerce. The focus of e-commerce have high priority towards analyzing customer behaviour, using clear evaluation criteria. In this method, it shows, a manual process of books for handling sales predictions is used. The results proposed were effective and efficient, proposed, forecasting accuracy has influenced using different books, and, also reliability on user behaviour is a strong way to predict sales trends.

2.3 Prediction of E-commerce Sales Using Machine Learning Methods

In the previous information, researchers focused on using different traditional methods for sales prediction in e-commerce. But, all these approaches are time-consuming and are not considered that effective as the data volume of e-commerce sales generation is increasing. Thus, this brings into picture the use of advanced methods like machine learning-based e-commerce prediction of sales. According to the information provided by (Singh et al.;

2020), prediction of sales and forecasting have evolved with newer technologies, in which machine learning plays an important role. The study proposed, the evidence on the use of various ML models in recent years, reducing the error rate and acquiring high prediction results. As per the study, by (Kulshrestha and Saini; 2020), the growth of online platforms has also improved customer satisfaction, which leads the customers to come back. This is possible, because e-commerce playing a vital role for managing inventories, and help predicting demand-supply chain for meeting customer product preferences.

The given approach is effective in promoting a healthy relationship between, the customers and e-commerce companies, but the use of large and unstructured data affects the rate at which the customer leaves, (Matuszelański and Kopczewska; 2022). After the situation is understood, (Kulshrestha and Saini; 2020), provides considerable evidence on the effectiveness of machine learning models, for, more accurate predictions. The study has performed an objective of using e-commerce data from different companies, which was divided into training (70%) and testing (30%) data. Focusing on the insights derived, a distinct result for companies to predict sales for the next quarters, and also provide analyses of total sold commodities. In the similar study, proposed by (Li et al.; 2019), growth of cross-border e-commerce between different enterprises, and large companies provides valuable insights, across different business areas. These insights are like managing operational costs, warehousing activities, development of products, and, also improving sales prediction. However, a similar problem arises due to an increase in sales, and large data volumes which leads the firms to face challenges in accurately predicting future sales. Hence, findings showed,with the use of ML algorithms,good prediction accuracy has been obtained, and directly leading to improved inventory management and sales performance of companies in the next quarters.

In today's world, people are mostly focused towards online platforms than physical stores. This has increased the opportunity for the e-commerce platform to become a priority in the business world. In this regard, (Huo; 2021), during the pandemic crisis in 2019, there were significant challenges for businesses in meeting customer demands, which affects the sales rate. However, with increase in e-commerce growth, has boosted the companies business potential to manage and handle activities online. This provides convenience to customers for more product choices and buying experience. Thus, it is very important for considering, the COVID-19 situation leading to importance of e-commerce. Amid this priority, the platform efficiency for effective sales performance requires, proper prediction of the sales volume, that is already managed through e-commerce platforms. The study proposed by (Huo; 2021), presented a comparative evaluation of machine learning (ML) models and deep learning algorithms in the prediction of sales. Experiments show that, the models show no improvement in predicting accuracy for sales forecasting. However, when both, data and price are included, leading to a better performance.

2.4 Prediction of E-commerce Sales Using Deep Learning Methods

The standard focus of the research mainly lies, on the knowledge of sales predictions. The studies, based on the evidence have clearly shown its importance. Since, the launch of `Amazon`, which is a first and one of the biggest e-commerce retail organization, the increasing growth of e-commerce influenced on how customers shop. According to the evidence, proposed by (Aljbour and Avcı; 2024), the consumer behaviour analysis has influenced e-commerce platform directly or indirectly, requiring more strong focus on the accurate sales forecasting, and prediction. Amid the focus, in this given study, a deep learning model, Long Short-Term Memory (LSTM) model was introduced and its ability of making predictions was examined. The studies suggests, LSTM can predict e-commerce sales, providing insights for predictions. The smaller businesses, like, start-ups may face challenges because of less resources.

As there has been exponential growth in the recent decade for the online business, the focus has shifted to virtual platforms rather than physical platforms. In the study by (Chen et al.; 2024), the user preference can be easily enhanced by accurate and efficient prediction of sales. In this regard, the decision necessary to consider online product development needs to be correlated with certain factors specific to industries, while leveraging improved prediction models. Considerably, the use of deep learning models has improved sales prediction compared to traditional machine learning methods. This can be seen, from the above study, which explains that models like the GRU architecture have proven to be more effective in capturing complex data patterns as compared to other models for prediction of sales (Chen et al.; 2024). The model is capable of dealing with large, unstructured datasets and also maintaining high efficiency. GRU's model architecture provides with fast computation when compared to LSTMs model, maintaining high accuracy which is very suitable for dynamic e-commerce environments. The study provides the facts about the model, having the ability to improve decision-making predictions in a dynamic and data-rich e-commerce environments. This is very beneficial for the e-commerce industry.

The information provided shows, both LSTM and GRU models performed well and can predict e-commerce sales accurately, although the data was unstructured, giving more accurate results compared to basic machine learning models (Aljbour and Avcı; 2024), (Chen et al.; 2024). However, its clear that no particular priority is given to the Brazilian e-commerce dataset using a hybrid model for sales prediction, although, the dataset is highly recognized due to its features and data specificity (Olist and Magioli; 2018). According to the information by (Petroșanu et al.; 2022), there has been increased significance of e-commerce because of digitalization. This led for business to make profit, increasing the use of e-commerce broadly. This studies shows, sales and their predictions plays a vital role for business increasing profits. The study above proposed, the use of "Directed Acyclic Graph Neural Network" (DAGNN) in prediction of e-commerce sales for increased accuracy and forecasting (Petroșanu et al.; 2022). The study also proposed that the model has the ability to scale up in dynamic and complex environments. The valuable insights provided that although the model is important, but limited information on dataset. Hence, this gives a rise to a new research on effective use of dataset, also, developing various advanced deep learning models for accurate prediction of sales.

3 Methodology

The research goal of our project is to predict e-commerce sales, using the Brazilian e-commerce dataset. The dataset is rich with various transactions, and operations data from the e-commerce industry. The methodology used for this research work is, (KDD) Knowledge Discovery from Database. The methodology process flow is shown in the Figure1. The process flow starts with preparation of data , analysis, selection of features, development of models and their evaluation through the use of various deep learning techniques. This is designed effectively, for ensuring the accurate sales predictions, which is possible because of rich features of the dataset. In this section, all steps are included:

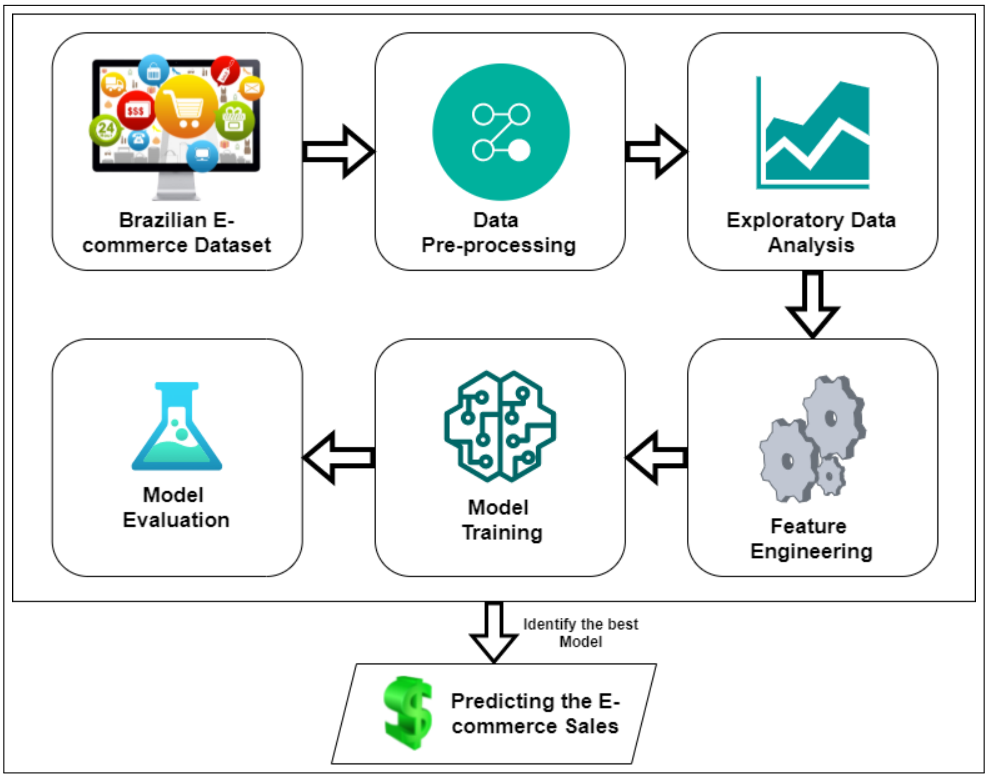


Figure 1: Methodology Flow Diagram for Predicting E-commerces Sales

3.1 Dataset Description

For the research, the dataset is collected from the Kaggle website, given by Olist store named the Brazilian e-commerce dataset (Olist and Magioli; 2018). This dataset provides, with an information for almost 100000 orders between the year 2016 and 2018 in Brazil for various marketplaces. The dataset is almost 126MB in size and has 9 inter-related different files.

Dataset Name	Description
olist_customers_dataset	customer IDs and locations.
olist_geolocation_dataset	geographic data (latitude and longitude).
olist_order_items_dataset	order items, including product IDs & prices.
olist_order_payments_dataset	payment methods & installments.
olist_order_reviews_dataset	customer reviews and ratings.
olist_orders_dataset	order status and timestamps.
olist_products_dataset	product details like weight and category.
olist_sellers_dataset	seller IDs and locations.
product_category_name	product category names in English.

Table 1: Data Set Names and Description

The files can be seen in Table 1, consisting of 117329 rows, and 52 columns. The analysis of orders can be done from different aspects, like pricing, details of payments, freight performance, locations of consumers, attributes of products, delay and reviews given by customers. The dataset has rich information for the customers, with the geolocation and mapped Brazilian ZIP codes to latitudes and longitudes,including ratings by customers, and their feedback after the purchase. The dataset seems to be a very perfect fit for this research, as it has rich information for e-commerce logistics, performance of products, customer trends.

3.2 Data Pre-processing

The use of data preprocessing is one the important step, that helps us to make the dataset clean, consistent, and effective for the further analysis required, for performance of predictive model. The first step of preprocessing involves checking for missing values, and the columns with missing data were identified. The missing values are handled using, a forward and backward filling imputation methods, including order_approved_at, order_delivered_carrier_date, and order_delivered_customer_date. 'No comment' for handling missing values of comments and 'Unknown' for product categories. The numerical columns like product weight and dimensions were handled using 52 median value imputation for more consistent data. The preprocessing also involves, checking for duplicate values for ensuring data quality. The format of some columns, like, Date columns were converted to datetime formats required for insights, as required for creating new features, like delivery_time and delay.

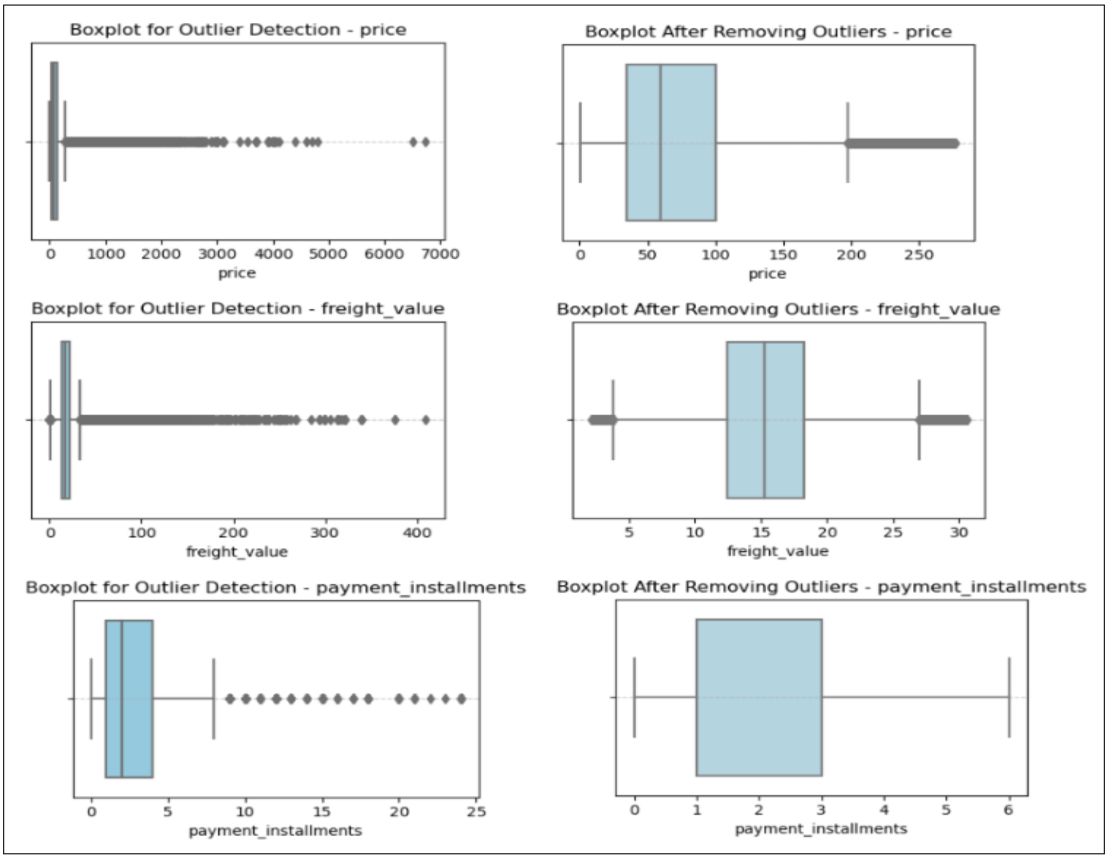


Figure 2: Boxplots Before and After Outlier Removal for Price, Freight Value, and Payment Installments

Outliers were identified, and removed using a method called Interquartile Range (IQR). This is done, for ensuring, that the extreme values should not affect the inference results or the performance of the model. The columns were detected, and outliers were removed from numerical columns like, price, freight_value, and payment_installments which can be seen in Figure 2 above. This leads to, reduction in noise and improving the reliability & stability of dataset. Boxplots were updated after removing outliers, displaying data has more accurate, ready and homogeneous distribution making it suitable for the modeling.

3.3 Exploratory Data Analysis

The purpose of Exploratory Data Analysis (EDA) is to gain a better understanding of the trends, patterns, structure and various relationships in the dataset. The use of EDA is done for identifying missing values, establish outliers and anomalies, behaviours, correlations, trends of the variables, which directly helps us for feature selection, development of models and data-driven decisions. This section will provide insights for, various variables using multiple charts and graphs.

The Figure 3 provides the insights, about, the relation between the product price and the amount for freight (logistics). The plot shows, as the value of product prices increase, the value of freight value gradually increases. But, the value of cheap products vary. Also, the expensive products almost have the consistent value. This draws that, there is no strong correlation between variables. This also, draws the fact that, freight value can also depend on other variables, like, shipping cost, shipping distance, weight of product and not only depend entirely on product price.

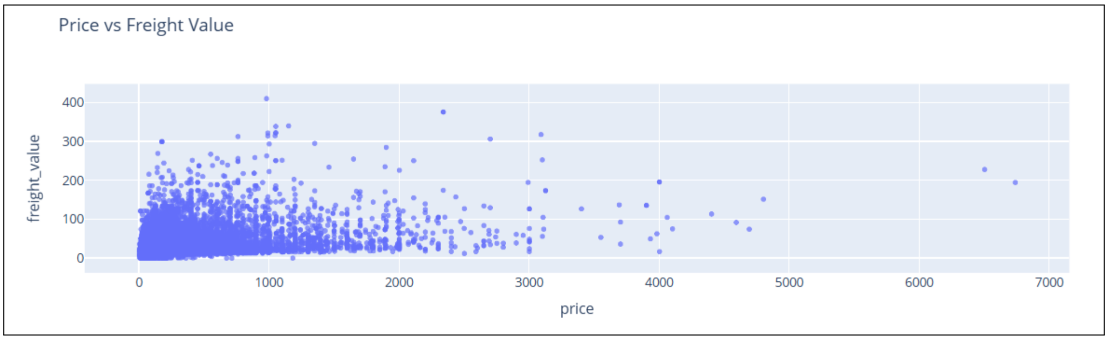


Figure 3: Relationship between Product Price and Freight Value

The bar chart and scatter plot, in the Figure 4 provides the insights of customer spending covering different product categories and the relation-link to the number of orders. The indication in the top plot suggests that most of the customers come under "Low' and "Medium " spending categories, whereas customers spending "High" or "Very High" are very less. The scatter plot in bottom provides the insights, the customer whose orders are fewer, mostly contribute to the higher spending, which indicates infrequent values, but the value of order is high. Thus, the insights suggests, most of the customers on online platform are cost-conscious where they like to spend more money in fewer transactions. This, information is very critical for organizations to create target marketing strategies.

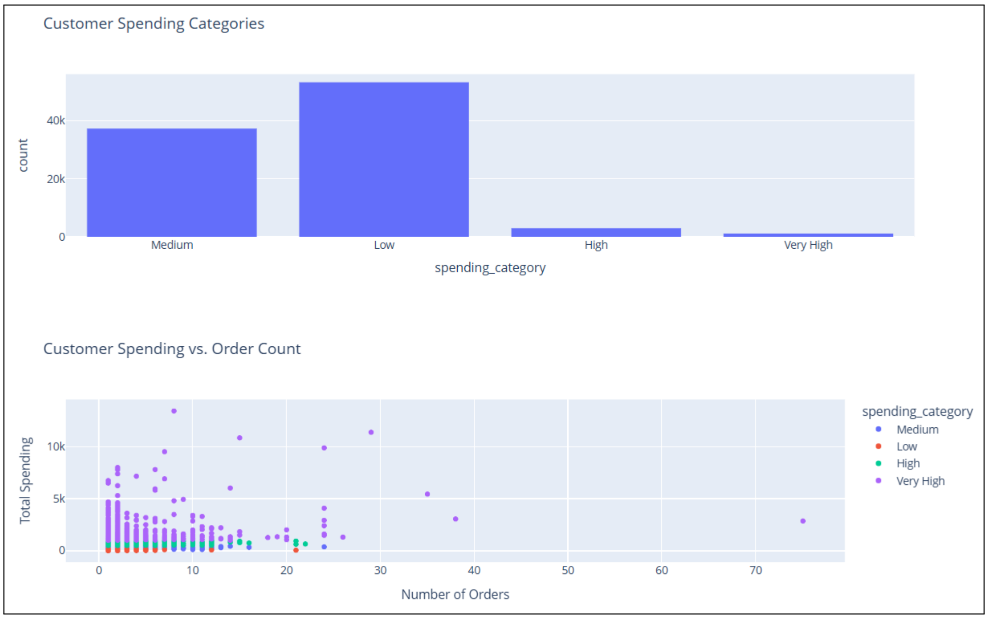


Figure 4: Analysis of Customer Spending Behavior

The Figure 5 below provides the information, of delay in average delivery across various states of customers, which is divided into early and late deliveries. The plot clearly provides the insights that Rorima (RR) and Amapá (AP) are the states, with highest average delivery delays. On the other hand, states like Mato Grosso do Sul (MS) and Rio Grande do Norte (RN) have very low average delivery delays. Mostly late deliveries only dominate most of the states, which can be seen from red bars. This suggests that logistic operations can be improved in this regions. The blue bars provide us with early deliveries, providing a benchmark, for a healthy supply chain practice.

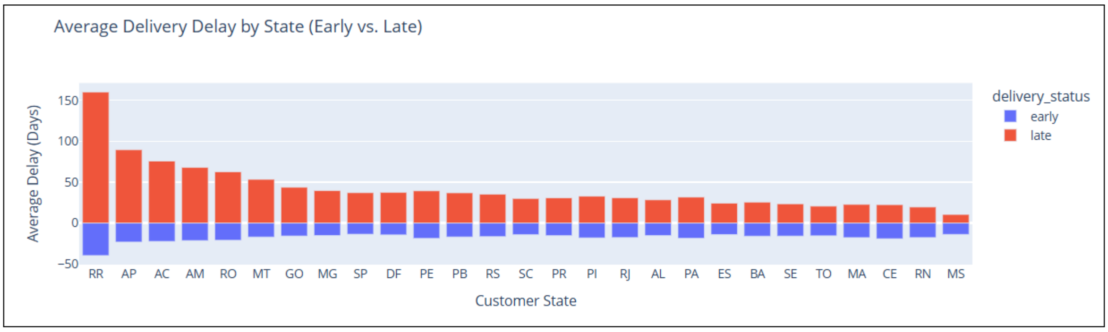


Figure 5: Average Delivery Delay by Customer State (Early vs. Late)

The Figure 6 below depicts, the trends in order volume on monthly basis from 2016 to 2018. The growth of graph month-on-month, indicates the direct growth of e-commerce rapidly over the years. The month of October and November 2017 showed the highest peak, may be because of seasonal promotions, holiday shopping and events. On the other hand, there is a sharp drop in peak in the month of August and September 2018, may be due to error & insufficient data or seasonal patterns. This insights, helps in planning for future inventories and target marketing strategies for the business organization.

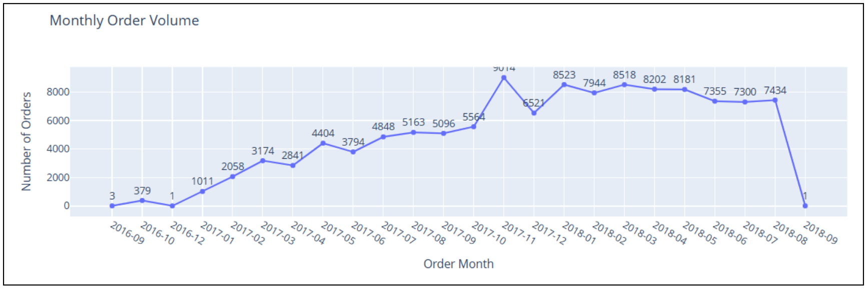


Figure 6: Monthly Order Volume Over Time

The Figure 7 is an area graph, about the month on month changes in order for the top 5 product categories respectively. It, can be seen from the graph, the categories like ”informatica_acessorios” and ”beleza_saude” (’computer and accessories’ & ’health and beauty’) has been consistently in demand by the customers. The peak is noted in late 2017, maybe due to seasonal trends or promotions. The other categories which include ”moveis_decoracao” and ”cama_mesa_banho” (’furniture & decor’ and ’bedding, bath items,tableware’), also shows a steady demand, but are very low on order volumes.

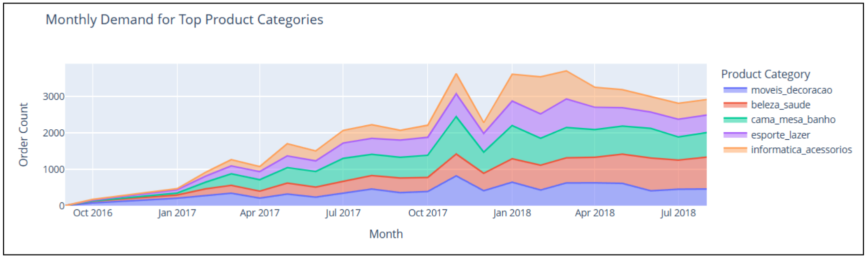


Figure 7: Monthly Demand for Top Product Categories

The chart in the Figure 8 is a sun-burst chart, useful for the breakdown of review score for new and repeat customers (1-5). The outer ring displays the review score from 1-5, whereas the inner ring is useful for the difference between, the new and repeat customers. The graphs provides insights of the consumer, having a high customer satisfaction for both, new and repeat, providing rating of at least 4 or 5. However, repeated customers moreover tend to give higher positive score. This insight is very useful for a high level of customer loyalty.

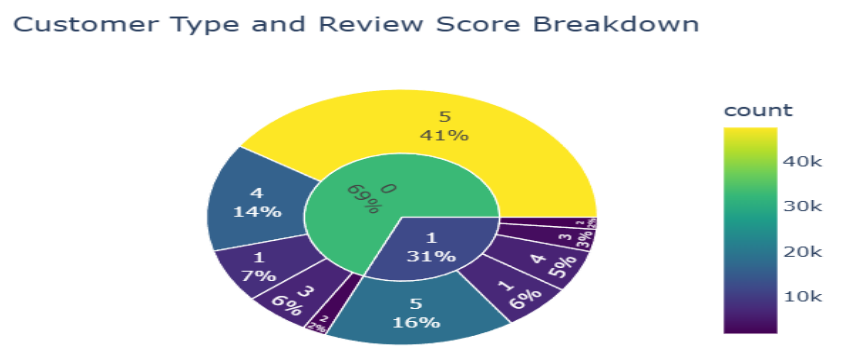


Figure 8: Customer Type and Review Score Distribution

The first plot in the Figure 9 is a histogram chart, depicts the distribution of the customer lifetime value (CLV). The insights suggests, most of the CLV is below 2,000, indicating the customers tend to purchase very less over the time. On the other hand, some extreme values of CLV suggests, some customer segments are of high value. The second plot in the Figure 9 highlights, average of the top 20 cities. The insights depicts, the cities like 'Agrestina' is leading with approx. 4,000 value, followed by Pirpirituba and Pianco. This insights helps to identify, the hotspot areas, target strategies and also manage the inventories level accordingly.

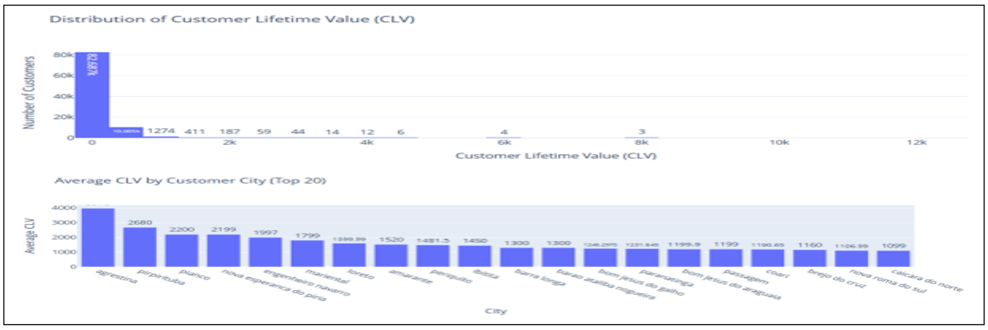


Figure 9: Customer Lifetime Value (CLV) Distribution and City-wise Analysis

3.4 Corelation Matrix and Feature Engineering

In this process, we started with identifying the categorical and numerical features. The co-relation matrix was plotted for all the values numerical in nature, and the heatmap was developed for checking the relations between variables. The main aim of this matrix is to avoid very highly related or multi-collinear features, by setting a threshold of 0.85. The correlation matrix is shown in Figure 10.

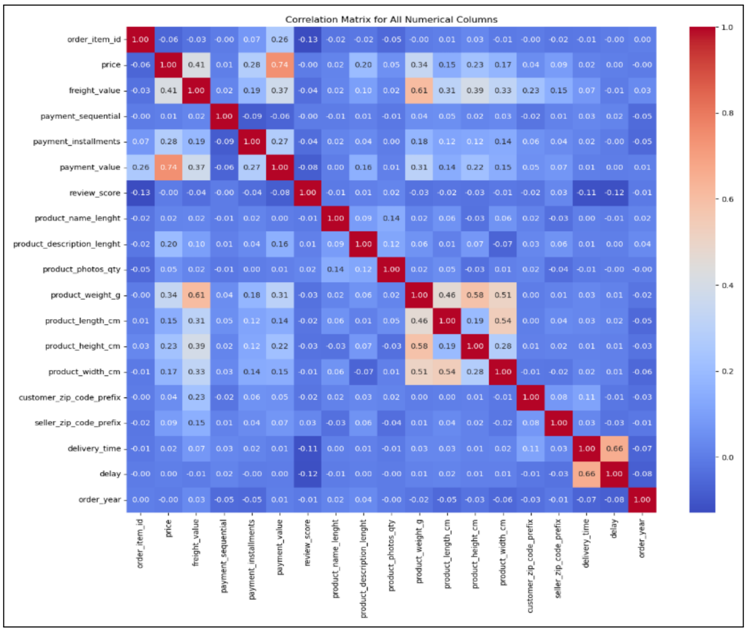


Figure 10: Correlation Matrix for Numerical Columns

Variable - order_purchase_timestamp is used for extraction of new features. The purpose of extracting these new features is for, the improvement of sales prediction for our model. The different features extracted are purchase_day, purchase_week, purchase_month, purchase_quarter, purchase_year, purchase_day_of_week, and purchase_is_weekend for sales prediction for the dataset. Variables like order_status, payment_type, product_category_name, and others are categorical, and they were transformed using label encoding, which makes them suitable for the model prediction. The method used for identification of top 15 features is Lasso Regression (Tibshirani; 1996), shown in the Figure 11. Some of the important features include, payment_value, customer_type, payment_type. The features selected were normalized using StandardScaler, and then added to the main dataset, for ensuring a better model training.

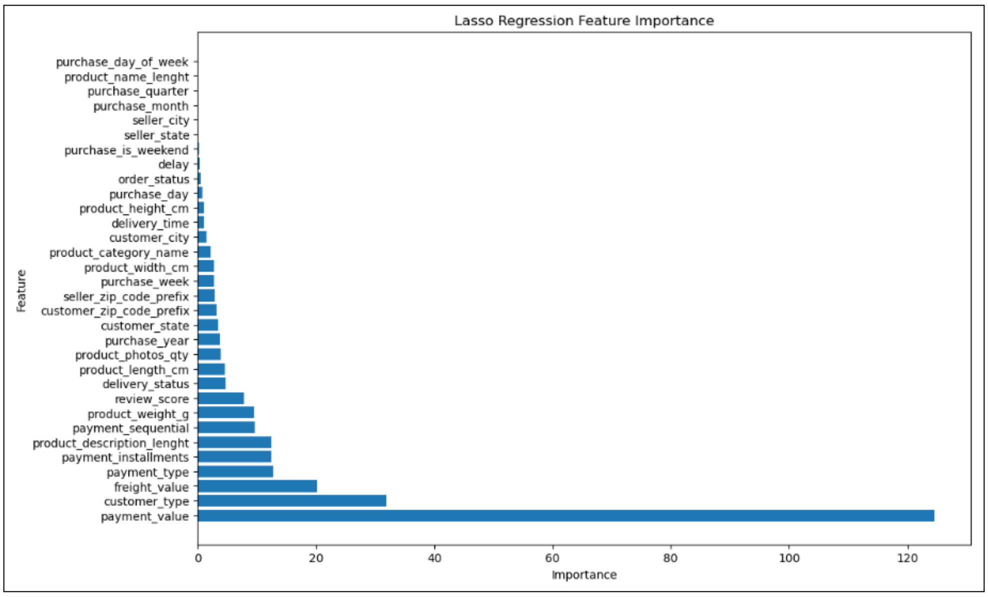


Figure 11: Lasso Regression Feature Importance

3.5 Model Training

This research work focuses on the model building, and training of advance deep learning models, that include, LSTM, GRU and a Hybrid model (combination of LSTM + GRU) for prediction of sales (Kulshrestha and Saini; 2020). The development of model was done using a sequential approach, and adding a dropout layer for prevention of overfitting of the model. Dense layer was also added for the regression output. The training phase of the model includes preprocessed, and the reshape data. Also, the Adam optimizer, and MSE (Mean Squared Error) for the loss function. All, the three models were trained with 30 epochs. The data split was validated for monitoring and ensuring the model accurate predictions for sales.

3.6 Model Evaluation

The evaluation of model is a very important and critical step in any of the machine learning predictive models. The main purpose of evaluation is, to check if the model is functional, valid and scalable. This research work focuses on performance analysis of all, the three models LSTM, GRU and Hybrid models using empirical metrics like MAE, MSE, RMSE, R-square and MAPE values (Chicco et al.; 2021). The evaluation metrics are useful for demonstrating, the scalable abilities of model for correct and incorrect predictions showcasing the strength and weakness of the models. The goal of research was achieved by comparing directly, the evaluation results of all the three models and making sure the model is scalable, robust and dynamic for real-world applications.

4 Design Specification

This section gives a detail on, the design specifications used for our research. The architecture and rationale behind the development and implementation of all the deep learning model are described, used for our prediction of sales in e-commerce industry. This research consist of, the three main models, named as LSTM, GRU and a Hybrid model specifically.

4.1 Long Short Term Memory (LSTM)

The first model for the research work used is, Long Short-Term Memory (LSTM) model. The model is a specialized and generalized form of recurrent neural network (RNN), designed for capturing and learning the long-term dependencies in a sequential form. The model like LSTM is very suitable for prediction in e-commerce industries because of its tendency of capturing temporal data, like, history of purchase and seasonal patterns. This all data is very essential, for correct accuracy, and identifying trends. The architecture of LSTM consist of, memory cells that are developed by multiple LSTM layers. This, allows the model to select a particular information, and forget it as needed with the time. In this research, we have implemented a two-layer LSTM model. The model consists of dropout layers to prevent overfitting. Dense layer is used to capture the output. This allows the model to be more accurate, and handle the temporal complex sequence of e-commerce data. The architecture of LSTM model is shown below in Figure 12.

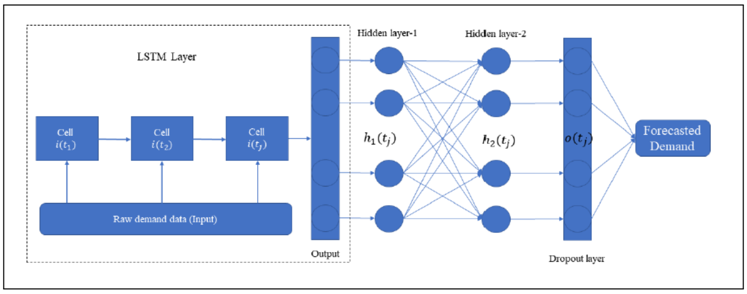


Figure 12: Long Short Term Memory Architecture Sardar et al. (2021)

4.2 Gated Recurrent Unit (GRU)

The second model for the research work used is, Gated Recurrent Unit (GRU) model. The model is a simplified and easy variant of the LSTM model. The computational intensity of GRU model is less. The GRU model is very suitable for prediction in e-commerce industries as they also offer the ability to capture temporal dependencies with fewer parameters. This allows, the fast training of the large datasets. This research work consist of, a two-stacked GRU layers, and a dropout layer for reducing overfitting. Dense layer is also added, for the prediction of the output. The GRU model consist of gated patterns, allowing the model's ability to learn complex patterns like, payment behaviour or delivery timings, widely used in e-commerce industries for analyzing sequential data. The architecture of the GRU model is shown below in Figure 13.

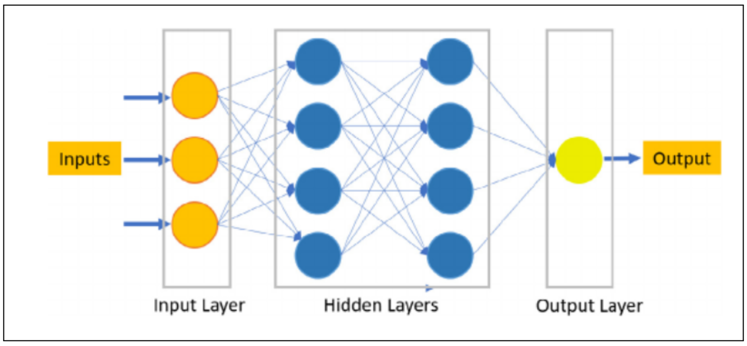


Figure 13: Gated Recurrent Unit (GRU) Architecture Kartal (2023)

4.3 Hybrid Model (LSTM + GRU)

The third model used for the research work is, a Hybrid model. The model is a combination of LSTM and GRU architecture for enhancing the sales prediction. The model uses both, traditional LSTM layers for capturing long-term dependencies, and GRU layers for handling, short-term patterns effectively. This makes the model more suitable for, handling complex e-commerce sales data for exhibiting both short-term fluctuations as well as long-term trends. The Hybrid model for this research work, consist of two different layers, that, includes LSTM layer and GRU layer. The use of LSTM layers for sequential learning, while GRU demonstrated enhanced computational efficiency. Dropout layer, and dense layer are included in the model to prevent overfitting and capture outputs. This hybrid approach allows the model for capturing diverse and complex patterns in the dataset, making the model more useful for e-commerce sales predictions. The architecture of the Hybrid model, is shown below in Figure 14.

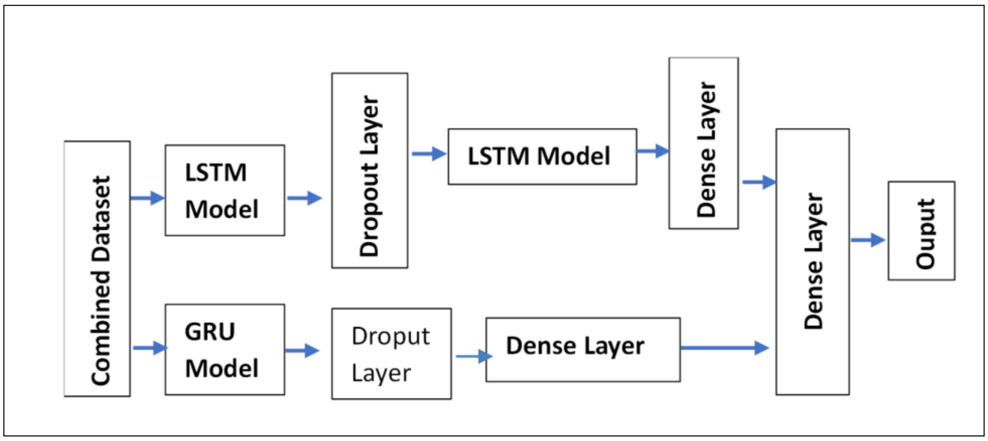


Figure 14: LSTM + GRU Architecture

5 Implementation

In this section of implementation, the goal of our research was, prediction of e-commerce sales with the use of deep learning techniques in a comprehensive, and a structured workflow. The libraries like pandas and numpy are used for easy manipulation of data for data preprocessing. It also includes, handling missing values and extraction of some new temporal features like purchase_day and purchase_year. The libraries like, matplotlib, seaborn and plotly are used for visualizations in the exploratory data analysis section, providing us with various insights on customer trends and behaviours, product categories and performance delivery. The use of, LabelEncoder was done to encode all the categorical variables into numerical variables that allows easy model training. The data was normalized, with the help of StandardScaler technique bringing the values between, 0 and 1. The selection of important features was performed by using the Lasso Regression technique, helping to identify important features for regression task. The deep learning models were implemented with the help of TensorFlow and its Keras Sequential API. The three deep learning models: LSTM, GRU and a Hybrid model were developed for capturing the long-term sequences and temporal patterns in e-commerce data. Overfitting was prevented by, dropout layer, the models compilation was done by, MSE loss function using Adam optimizer. The data integrity was maintained using a train_test_split pipeline, dividing into training and test phase. Reproducibility is achievable because of random seed. Model assessment was done using sklearn.metrics including MAE, MSE, RMSE, R-square, and, MAPE values. This values provide a perfect overview of the models capability. This implementation enhances, the use of various tools and libraries for a perfect framework for steps like data preprocessing, feature engineering, model training, and evaluation enhancing the prediction of sales in e-commerce. The system configuration used for research can be seen in Table 2.

Component	**Details**
Operating System	Windows
Platform	Jupyter Notebook
RAM	16 GB
CPU	8 Cores
Hard Disk	1TB SSD
Python Libraries	pandas, numpy, matplotlib, seaborn, plotly, sklearn, tensorflow, keras, random, os

Table 2: System Configuration Used in the Research

6 Evaluation

The evaluation of model is a very important part in any kind of research for the experts to gain knowledge. Model evaluation helps for determining the effectiveness, reliability, and accuracy of all the models implemented in any research. This section will include all the metrics required for the key evaluation: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (Coefficient of determination), and Mean Absolute Percentage Error (MAPE) (Botchkarev; 2018). The metrics mentioned, provides an accurate evaluations and the weakness of all the model implemented for prediction of sales.

6.1 Evaluation based on Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a vital metrics for the evaluation which does not consider, the directions of prediction error while measuring the average magnitude. In this research work, the value of MAE describes the estimate on how well our model performed in prediction of sales in e-commerce. The performance of the model is better when the value of MAE is less, compared to other models. The Figure 15 provides, the comparison of the Mean Absolute Error for all three models used for our research named as: LSTM, GRU, and Hybrid model. It is clear, the Hybrid model has outperformed, the other two models with the lowest Mean Absolute Error of 18.68, enhancing the sales prediction in e-commerce industry. This all shows, it has good ability for capturing the sales pattern when compared to other models. The second model which performed well was, LSTM model which had an MAE of 19.91. This tells us the fact, model is still not so bad in the performance for predictions and is capable. On the other hand, GRU model had the worst MAE value of 24.09, compared to other models.

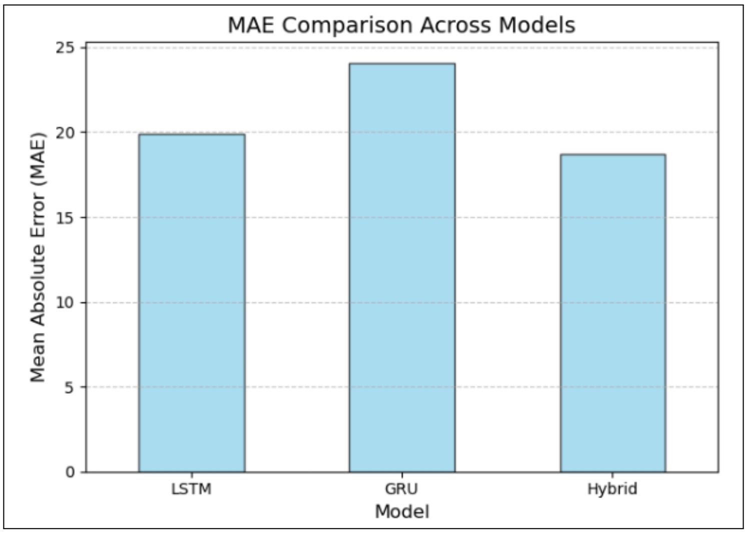


Figure 15: MAE Comparison Across LSTM, GRU, and Hybrid Models

6.2 Evaluation based on Mean Squared Error

Mean Squared Error (MSE) is also an evaluation metrics in regression analysis, helping in calculating the average square difference between, both, the actual and predicted values. The main focus of the metric is on squaring large errors. The evaluation is very crucial, indicating high accuracy with low MSE value. The Figure 16 is a bar chart, comparison of the MSE values over all, the three models. Out of all, the three models, the Hybrid model had the lowest MSE value of 3150, showcasing that the performance of model is efficient in predictions by minimizing the errors. The second model which followed closely, the Hybrid model was LSTM model. The model had a MSE value of 3278. On the other hand, GRU model was not able to be close enough and the MSE value was 4785. The results outlines, the use of a Hybrid model for prediction of sales.

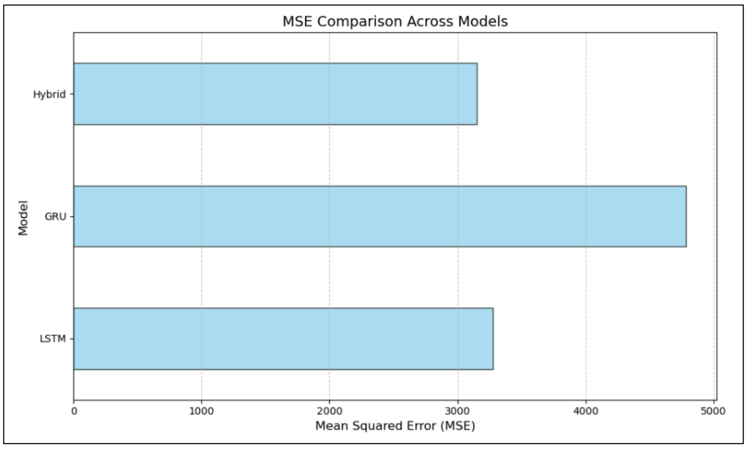


Figure 16: MSE Comparison Across LSTM, GRU, and Hybrid Models

6.3 Evaluation based on Root Mean Squared Error (RMSE)

Root Mean Square Error (RMSE) is one of the important metrics used in regression analysis. The error of our model is given by the RMSE values, by taking the square root of the mean square errors. The ability of metric is excellent when dealing with errors. This will predict accurate sales in e-commerce. The line graph in the Figure 17, depicts the comparison of the RMSE value for all, the three models. Again, the results suggests, the Hybrid model has the lowest RMSE value of 56.12. This shows, Hybrid model has the best prediction results, as compared to other models and dealing with minimal large errors. Again, LSTM model closely follows the Hybrid model with a RMSE value of 57.25. On the other hand, GRU model had the highest RMSE value of 69.17. The GRU model had a very huge deviations among the, predicted and actual values.

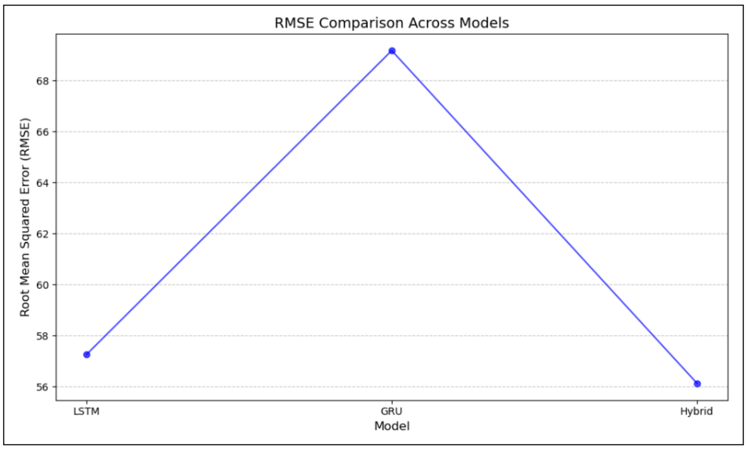


Figure 17: RMSE Comparison Across LSTM, GRU, and Hybrid Models

6.4 Evaluation based on R-Square

R-square or Coefficient of Determination is an important metric and plays a very vital role in regression analysis. The R-square value helps to evaluate the percentage of variance in the target variable explained by the model. When, value of R-square is high, it shows that the model fits very well and thus, making it very important metric for evaluation of the model in our research. The R-square metric is the most important metric among all the other metrics (Chicco et al.; 2021). The bar graph in the Figure 18, depicts the comparison of the R-square values for all, the three models namely: LSTM, GRU and Hybrid model respectively. The R-square value of the Hybrid model was the highest among all the other models, with value of 0.91. The results explains, the Hybrid model is the most accurate fit among all models, and also have the ability, for explaining most of the variance in the sales data. The trend for LSTM model continues, where, the model follows the Hybrid model, with the R-square value of 0.90. On the other hand, again, GRU model was not able to perform well, giving us the R-square value of 0.86. This indicates GRU was the weakest among all the three models.

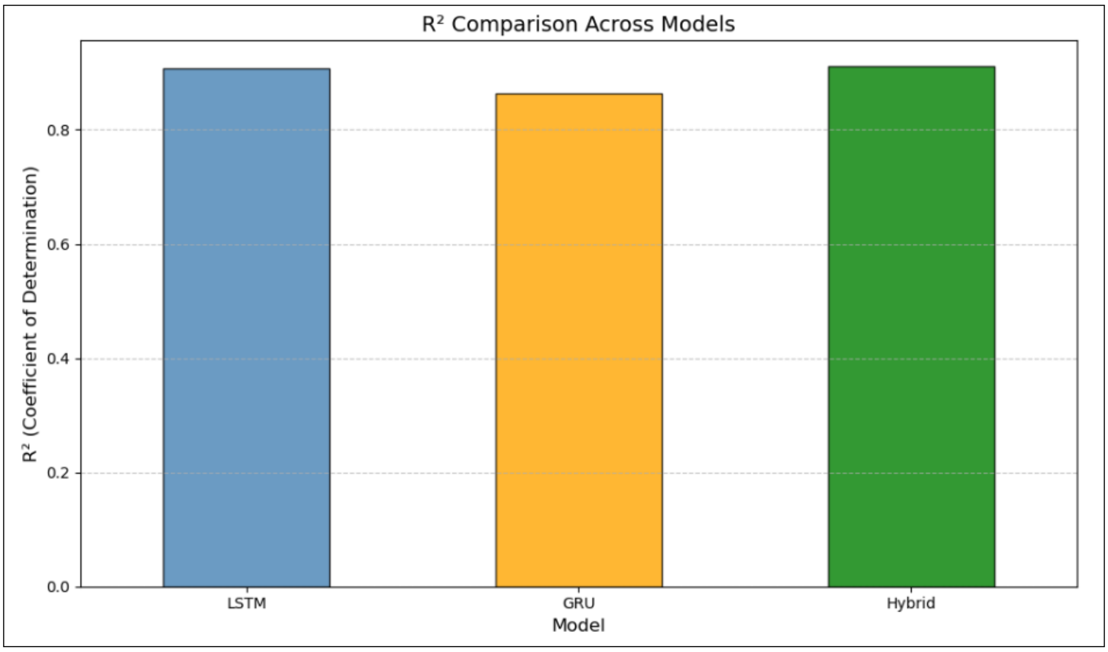


Figure 18: R-Squared Comparison Across LSTM, GRU, and Hybrid Models

6.5 Evaluation based on Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a key evaluation metrics for regression analysis. The value of MAPE measures, the average of percentage deviation between both predicted and actual values. It, is a normalized metric, most suitable for accuracy in prediction of sales within sales forecasting. The lower value of MAPE indicates, model has performed very accurately in prediction of sales. Figure 19 gives us the insights, on the comparison of the Mean Absolute Percentage Error (MAPE) for all the three models used for our research named as: LSTM, GRU, and Hybrid model. The insights are very interesting, as, the GRU model has the lowest MAPE value of 33.52. This suggests, the model is successful for reduction of relative errors of sales prediction. The second model followed by GRU model is, Hybrid model. The Hybrid model gives an almost similar MAPE value of 33.55 to GRU model. This indicates that the model has comparative performance. On the other hand, LSTM model had highest MAPE value of 37.96, indicating higher error predictions. The results proposed, that both GRU and Hybrid model are efficient for higher prediction error when compared to LSTM model.

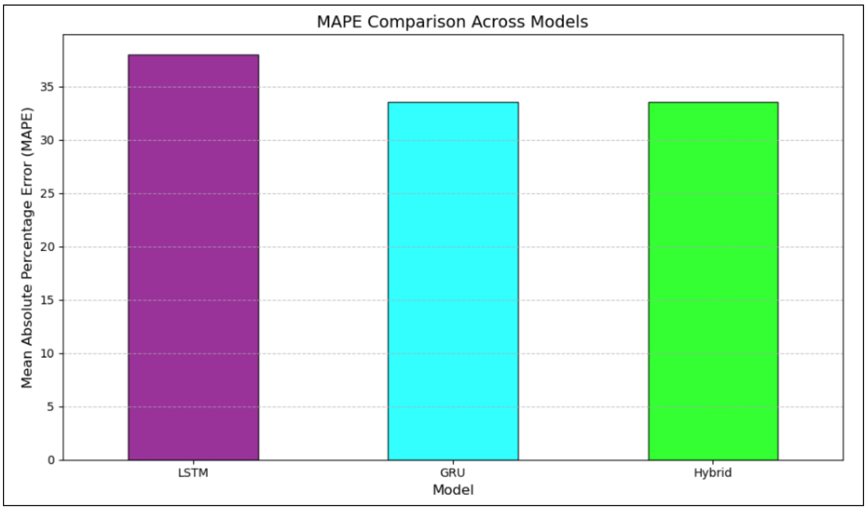
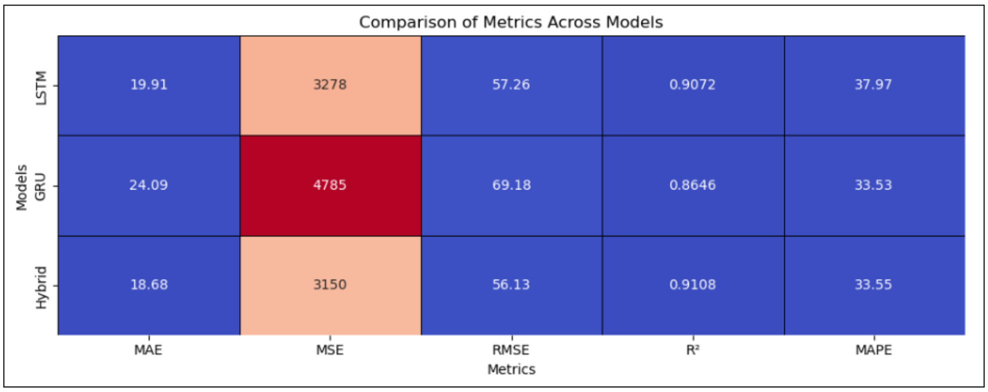


Figure 19: MAPE Comparison Across LSTM, GRU, and Hybrid Models

6.6 Discussion

The comparison of all the three models namely: LSTM, GRU and a Hybrid model, provides us to highlight the effective approach towards our research study in the field of e-commerce for sales prediction. The Hybrid model, is the best model as per the results. The model's MAE value was 18.68, MSE of 3150, RMSE was 56.12, and R-square was 0.91   91%. This shows the ability of a hybrid model for predicting sales performance accurately, also explaining the variance in sales data. The strength of Hybrid model lies in the combination of both LSTM and GRU architectures, as LSTM model architecture used for capturing long-term memory capabilities, along with the computational efficiency of GRU architecture. This also suggests, the model to deal with both short-term fluctuations and long-term patterns for e-commerce data, leading the model towards highly practical real-world applications.



Models	MAE	MSE	RMSE	R²	MAPE
LSTM	19.91	3278	57.26	0.9072	37.97
GRU	24.09	4785	69.18	0.8646	33.53
Hybrid	18.68	3150	56.13	0.9108	33.55

Figure 20: Comprehensive Comparison of Metrics Across LSTM, GRU, and Hybrid Models

Metric	LSTM	GRU	Hybrid Model (LSTM + GRU)
(MAE)	19.91	24.09	**18.68**
(MSE)	3,278	4,785	**3,150**
(RMSE)	57.25	69.17	**56.12**
(R^2)	0.90	0.86	**0.91**
(MAPE)	37.96	**33.52**	33.55

Table 3: Comparison of Model Performance

The performance of LSTM model was significant, with the MAE of 19.91, RMSE of 57.25. Metrics like MAPE of 37.96, proved models inability in handling percentage based errors. LSTM is able to capture temporal dependency, but, not good for computational efficiency. GRU model performed weak, with the highest MAE of 24.09, RMSE of 69.17, and R-square of 0.86. GRU model has simple architecture which provides faster training, but it fails to capture complex pattern in the data.The comparison metrics evaluation shown in Figure 20 with the help of a heatmap. Methodology and research in this study was achieved by data preprocessing, feature engineering and model evaluation. Our study provides the insights, Hybrid model is a better model as compared to individual architecture. Combination of LSTM and GRU architecture helps achieve Hybrid model to be, more scalable, efficient and accurate for prediction of sales and handle complexity of sequential data.

7 Conclusion and Future Work

The purpose of this research is implementing the advanced deep-learning models: LST, GRU, and a novel Hybrid model. The models used in this research are for prediction of sales in e-commerce. The performance of Hybrid model was the best, with highest R-square, and the lowest MAE, MSE and RMSE values. The results, showcased, ability of Hybrid model to capture both, short-term fluctuations and long-term dependencies when dealing with sales data. Although, the traditional approaches focus on single architecture, our research developed a Hybrid model that combines strengths of LSTM and GRU for scalability and prediction of model. The study on them makes the research well-defined,providing valuable contribution to the field of predictive modeling. The research showcase, the need of Hybrid model in accurate prediction of sales in e-commerce. The Hybrid model proposed in our study is efficient, but, also provides a flexibility and scalability for e-commerce companies, offering valuable tools for enhanced decision-making and improving operational efficiency satisfying all the objectives of research.

Our research achieved promising results. Although, results, are better for Hybrid model, but, future works brings external factors like marketing campaigns, seasonality and macro-economic factors, that can also be included for the accurate predictions. The use of Hybrid model on other datasets in the e-commerce industry would test its performance and scalability issues. The use of real-time prediction system, can be implemented for dynamic price prediction and inventory management. The attention towards combining the Hybrid model with ensemble techniques, for achieving more higher accuracy is also possible. The model can also be deployed on cloud, improving model computational performance in the areas where, relevant real-world applications are required. These additions will also add some values for building more deeper foundations and providing valuable insights in the established study for the e-commerce industry.

References

Aljbour, M. and Avcı, İ. (2024). Sales prediction in e-commerce platforms using machine learning, *International Conference on Forthcoming Networks and Sustainability in the AIoT Era*, Springer, pp. 207–216.

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology, *arXiv preprint arXiv:1809.03006* .

Chen, X., Zhang, L. and Zhang, Y. (2024). A time-series e-commerce sales prediction method for short-shelf-life products based on gru and lightgbm, *Electronics* **14**(2): 866.

Chicco, D., Warrens, M. J. and Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation, *Peerj computer science* **7**: e623.

Ginantra, N., Asana, I., Parwita, W. G. S. and Wiadnyana, M. L. D. (2023). Forecasting system analysis using gated recurrent unit neural network, *J. Syst. Manag. Sci* **13**: 470–482.

Huo, Z. (2021). Sales prediction based on machine learning, *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*, IEEE, pp. 410–415.

Kartal, S. (2023). Assessment of the spatiotemporal prediction capabilities of machine learning algorithms on sea surface temperature data: A comprehensive study, *Engineering Applications of Artificial Intelligence* **118**: 105675.

Kulshrestha, S. and Saini, M. (2020). Study for the prediction of e-commerce business market growth using machine learning algorithm, *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, IEEE, pp. 1–6.

Li, J., Wang, T., Chen, Z., Luo, G. et al. (2019). Machine learning algorithm generated sales prediction for inventory optimization in cross-border e-commerce, *International Journal of Frontiers in Engineering Technology* **1**(1): 62–74.

Matuszelański, K. and Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach, *Journal of Theoretical and Applied Electronic Commerce Research* **17**(1): 165–198.

Olist, D. and Magioli, F. (2018). Brazilian e-commerce public dataset by olist, *Kaggle* .

Pandey, T. N., Vasudev, A., Sagayanathan, D., Anjan, G., Arshad, D. and Patra, S. S. (2023). Predicting customer satisfaction in brazil e-commerce: A comparative study of machine learning techniques, *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, pp. 505–510.

Petroșanu, D.-M., Pîrjan, A., Căruţaşu, G., Tăbușcă, A., Zirra, D.-L. and Perju-Mitran, A. (2022). E-commerce sales revenues forecasting by means of dynamically designing, developing and validating a directed acyclic graph (dag) network for deep learning, *Electronics* **11**(18): 2940.

Sardar, S., Sarkar, B. and Kim, B. (2021). Integrating machine learning, radio frequency identification, and consignment policy for reducing unreliability in smart supply chain management, *Processes* **9**: 247.

Singh, K., Booma, P. M. and Eaganathan, U. (2020). E-commerce system for sale prediction using machine learning technique, *Journal of Physics: Conference Series* **1712**(1): 012042.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1): 267–288.

Usmani, Z., Manchekar, S., Malim, T. and Mir, A. (2017). A predictive approach for improving the sales of products in e-commerce, *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, IEEE, pp. 188–192.

Wei, D., Geng, P., Ying, L. and Shuaipeng, L. (2014). A prediction study on e-commerce sales based on structure time series model and web search data, *The 26th Chinese Control and Decision Conference (2014 CCDC)*, IEEE, pp. 5346–5351.

Yuan, H., Xu, W., Li, Q. and Lau, R. (2018). Topic sentiment mining for sales performance prediction in e-commerce, *Annals of Operations Research* **270**: 553–576.

Yuan, H., Xu, W. and Wang, M. (2014). Can online user behavior improve the performance of sales prediction in e-commerce?, *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, pp. 2347–2352.