SALES PREDICTION IN E-COMMERCE USING DEEP LEARNING

2024
MASTER THESIS
COMPUTER ENGINEERING

Mohammed ALJBOUR

Thesis Advisor
Assist. Prof. Dr. İsa AVCI

SALES PREDICTION IN E-COMMERCE USING DEEP LEARNING

Mohammed ALJBOUR

Thesis Advisor
Assist. Prof. Dr. İsa AVCI

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering
Prepared as
Master Thesis

KARABUK
February 2024

I certify that in my opinion the thesis submitted by Mohammed ALJBOUR titled “SALES PREDICTION IN E-COMMERCE USING DEEP LEARNING” is fully adequate in scope and quality as a thesis for the degree of Master of Science.

Assist. Prof. Dr. İsa AVCI ....
Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis on February 27, 2024.

<u>Examining Committee Members (Institutions)</u> <u>Signature</u>

Chairman : Assist. Prof. Dr. Ali HAMİTOĞLU (İSU) ....

Member: Assist. Prof. Dr. İsa AVCI (KBU) ....

Member : Assist. Prof. Dr. Nehad T.A RAMAHA (KBU) ....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Assoc. Prof. Dr. Zeynep ÖZCAN ....
Director of the Institute of Graduate Programs

“I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well.”

Mohammed ALJBOUR

ABSTRACT

M. Sc. Thesis

SALES PREDICTION IN E-COMMERCE USING DEEP LEARNING

Mohammed ALJBOUR

Karabuk University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. İsa AVCI
February 2024, 48 pages

The rapidly evolving e-commerce platforms have reshaped consumer behavior, creating an imperative for accurate sales forecasting models. This paper delves into predictive analytics, using machine learning, focusing on utilizing Long Short-Term Memory (LSTM) for sales prediction within the e-commerce domain. Leveraging a comprehensive dataset from Taobao, a prominent e-commerce platform, this study employs LSTM-based models to forecast sales trends, considering factors such as user interactions, browsing patterns, and purchase behavior. The investigation encompasses preprocessing techniques to prepare the dataset for LSTM model training, emphasizing sequential dependencies and temporal dynamics inherent in e-commerce data. Through accurate evaluations using standard metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), the efficacy of LSTM models in predicting sales patterns is scrutinized. The paper highlights the potential implications of accurate sales forecasting in optimizing inventory management, marketing strategies, and decision-making within the e-commerce landscape. This study contributes to the growing knowledge of leveraging LSTM networks for precise sales prediction in e-commerce, providing insights for future advancements in predictive analytics within this dynamic domain.

Key Words : Machine Learning, Deep Learning, Sales Prediction, E-Commerce, MSE, MEA.

Science Code : 92432

ÖZET

Yüksek Lisans Tezi

DERİN ÖĞRENME KULLANARAK E-TİCARETTE SATIŞ TAHMİNİ

Mohammed ALJBOUR

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi İsa AVCI

Şubat 2024, 48 sayfa

Hızla gelişen e-ticaret platformları tüketici davranışını yeniden şekillendirerek doğru satış tahmin modellerine yönelik bir zorunluluk oluşturdu. Bu makale, makine öğrenimini kullanarak, e-ticaret alanında satış tahmini için Uzun Kısa Süreli Bellek (LSTM) ağlarını kullanmaya odaklanarak tahmine dayalı analitiği ele almaktadır. Önde gelen bir e-ticaret platformu olan Taobao'dan alınan kapsamlı bir veri setinden yararlanan bu çalışma, kullanıcı etkileşimleri, göz atma kalıpları ve satın alma davranışı gibi faktörleri dikkate alarak satış eğilimlerini tahmin etmek için LSTM tabanlı modeller kullanıyor. Araştırma, e-ticaret verilerinin doğasında bulunan sıralı bağımlılıkları ve zamansal dinamikleri vurgulayarak LSTM model eğitimi için veri kümesini hazırlamak amacıyla ön işleme tekniklerini kapsamaktadır. Ortalama Karesel Hata (MSE), Ortalama Mutlak Hata (MAE) ve Ortalama Karesel Hatanın Kökü (RMSE) gibi standart ölçümler kullanılarak yapılan doğru değerlendirmeler yoluyla, LSTM modellerinin satış kalıplarını tahmin etmedeki etkinliği inceleniyor. Bu makale, e-ticaret ortamında envanter yönetimini, pazarlama stratejilerini ve karar almayı optimize etmede doğru satış tahminlerinin potansiyel etkilerinivurgulamaktadır. Bu çalışma, e-ticarette kesin satış tahminleri için LSTM ağlarından yararlanma konusunda artan bilgi birikimine katkıda bulunarak, bu dinamik alanda tahmine dayalı analitiklerde gelecekteki gelişmeler için öngörüler sağlıyor.

Anahtar Kelimeler : Makine Öğrenmesi, Derin Öğrenme, Satış Tahmini, E-Ticaret, MSE, MEA.

Bilim Kodu : 92432

ACKNOWLEDGMENT

I would like to begin by expressing my sincere gratitude to the boundless grace and guidance of Allah Almighty, which have illuminated my educational path. My heartfelt appreciation extends to Karabuk University for granting me the invaluable opportunity to pursue my graduate studies. I am particularly grateful to my supervisor, Assist. Prof. Dr. İsa AVCI, whose mentorship and support have been instrumental in my academic journey. I am also deeply thankful to the esteemed faculty and staff of this esteemed institution.

Furthermore, I wish to extend my profound gratitude and enduring affection to Turkey and my beloved homeland, Palestine.

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ML	: Machine Learning
DL	: Deep Learning
DT	: Decision Trees
CNN	: Convolutional Neural Networks
RF	: Random Forests
SVM	: Support Vector Machines
LSTM	: Long-Short Term Memory
XGBoost	: Extreme Gradient Boosting
MAE	: Mean Absolute Error
MSE	: Mean Sequre Error
RMAE	: Root Mean Square Error
CF	: Collaborative Filtering
ARNN	: Auto-Regressive Neural Network
DNN	: Deep Neural Network
RNN	: Recurrent Neural Network
RMSP	: Root Mean Square Percentage

PART 1

INTRODUCTION

In recent years, the evolution of e-commerce platforms has revolutionized consumer behavior [1], presenting an expansive landscape for businesses to explore and refine their sales strategies. Understanding and predicting consumer preferences in this dynamic environment have become pivotal for success. Machine learning, particularly LSTM-based models, has emerged as a formidable tool for forecasting sales in e-commerce domains. This study delves into the realm of e-commerce sales prediction, leveraging LSTM networks to forecast sales trends within the context of a prominent e-commerce giant, Taobao [2], [3]. The focus of this investigation revolves around the utilization of comprehensive datasets capturing user behaviors comprising interactions such as purchases, clicks, and browsing patterns. The study navigates through the preprocessing and analysis of these extensive datasets to extract meaningful insights. With the robust architecture of LSTM networks, the study aims to encapsulate temporal dynamics and sequential dependencies inherently present in e-commerce data. This approach enables the creation of predictive models capable of foreseeing sales figures, offering invaluable insights into evolving consumer trends, seasonal patterns, and the dynamic demand for various products. This research undertakes a meticulous evaluation of the LSTM-based predictive model's performance. Metrics including MSE [4], MAE [5], RMSE [6], and R-squared are employed to quantitatively assess the model's predictive accuracy.

Through this exploration, the research goal is to contribute to the growing body of knowledge on leveraging machine learning techniques for sales prediction in e-commerce, potentially paving the way for enhanced decision-making and strategy formulation within this rapidly evolving domain.

1.1. AIMS AND OBJECTIVES

The goal of this thesis is to predict and forecast sales in E-commerce platforms using Deep learning and exactly using LSTM, and it is divided into many objects as follows:

1. Utilize LSTM networks for sales prediction in the context of e-commerce.

2. Investigate and analyze user behavior data from the Taobao platform.

3. Develop robust predictive models to forecast sales trends based on historical user interactions.

4. Capture temporal dynamics and sequential dependencies inherent in e-commerce data.

5. Evaluate model performance using metrics such as MSE, MAE, RMSE, and R-squared.

Those objectives would be achieved under many procedures that we will propose in the next sections and here is a quick look at them:

1. **Data Preparation:** Gather, preprocess, and clean extensive user behavior datasets from the Taobao e-commerce platform.

2. **Feature Extraction:** Extract relevant features from the datasets, such as user clicks, purchases, browsing patterns, and temporal attributes.

3. **Model Development:** Construct and optimize LSTM networks tailored for sales prediction, considering the sequential dependencies and temporal patterns within the e-commerce data.

4. **Training and Validation:** Train the LSTM model using historical data, validating its predictive capabilities through various cross-validation techniques.

5. **Performance Evaluation:** Evaluate the model's accuracy using metrics like MSE, MAE, RMSE, and R-squared against test datasets.

6. **Interpretability Analysis:** Investigate the interpretability of the model's predictions, analyzing its ability to provide insights into consumer preferences, seasonal trends, and product demand dynamics.

7. **Application and Impact:** Explore the practical implications of the predictive model for optimizing marketing strategies and inventory management in e-commerce domains.

8. **Comparison and Benchmarking:** Compare the LSTM-based model's performance against other traditional machine learning methods or forecasting techniques.

9. **Documentation and Reporting:** Document the methodology, findings, and results obtained throughout the research, ensuring clear and concise reporting in the final thesis.

10. **Recommendations and Future Work:** Provide recommendations for further enhancements or future research directions based on the outcomes and limitations observed during the study.

1.2. IMPORTANCE

The system holds significant importance in the realm of e-commerce by enabling a profound understanding of online shopping dynamics. Its ability to predict sales trends within the Taobao platform offers invaluable insights into user behavior and preferences. This, in turn, empowers businesses to make informed marketing decisions and optimize their inventory management strategies. Ultimately, the system plays a pivotal role in enhancing the overall operational efficiency and strategic planning of online businesses.

1.3. SCOPE OF THE STUDY

Our study covers a comprehensive investigation into user behavior data sourced from Taobao, a leading e-commerce platform. It involves the careful selection, preprocessing, and analysis of extensive datasets, focusing on interactions like purchases, clicks, and browsing patterns. The study aims to harness the power of LSTM networks to develop predictive models capable of forecasting sales trends within the e-commerce domain. It also encompasses the evaluation of these models using various metrics to gauge their predictive accuracy. Moreover, the study seeks to interpret the model's predictions, shedding light on potential implications for optimizing marketing strategies and inventory management strategies in the e-commerce landscape.

1.4. CONTRIBUTION

This study contributes to the field of e-commerce by introducing robust predictive models, specifically LSTM-based approaches, for sales forecasting using rich user behavior datasets from Taobao. It aims to enhance the understanding of consumer behavior patterns, seasonal trends, and product demand dynamics within e-commerce platforms. The study's evaluation metrics and interpretability of predictions offer insights into improving decision-making processes for marketing strategies and inventory management in the e-commerce landscape. Ultimately, it seeks to enrich the body of knowledge surrounding machine learning applications in predicting sales dynamics within the evolving e-commerce domain.

1.5. STUDY PROBLEM AND PROPOSED SOLUTION

The primary problem lies in accurately forecasting sales figures amidst the complexities of user behavior within the e-commerce platform. The main issue encompasses:

1. **Predictive Accuracy:** The challenge revolves around developing a robust predictive model that effectively utilizes LSTM networks to capture the intricate patterns and dependencies present in user behaviors on Taobao[5]. The objective is to accurately forecast sales figures based on historical user interactions, considering the dynamic and multifaceted nature of e-commerce data.

2. **Temporal Dynamics:** Understanding and incorporating temporal dynamics, such as evolving user preferences, seasonal variations, and fluctuating demand patterns, poses a significant challenge. The main goal is to build a model that can adapt to these temporal aspects to enhance the accuracy of sales predictions.

3. **Data Complexity:** Dealing with extensive and diverse datasets encompassing various user interactions (clicks, purchases, browsing patterns) requires effective preprocessing, feature engineering, and model optimization. Managing the complexity of data and extracting meaningful insights for precise sales forecasting is a critical aspect of the problem.

4. **Model Interpretability:** While aiming for high predictive accuracy, ensuring that the LSTM model's predictions are interpretable and can provide insights into consumer behavior and trends is an essential challenge. This factor contributes to actionable insights for businesses to optimize their strategies[7].

The proposed solution addresses the complexities of sales prediction in e-commerce using LSTM models by focusing on several key strategies:

1. **Feature Engineering and Data Preprocessing:** Implementing advanced data preprocessing techniques to clean, aggregate, and engineer features from the vast and diverse dataset obtained from Taobao. This step aims to extract meaningful patterns and relevant information from user interactions.

2. **LSTM Model Architecture:** Designing and training LSTM-based models tailored to capture sequential dependencies and temporal patterns inherent in e-commerce datasets. This involves optimizing the architecture of LSTM networks to effectively learn from historical user behaviors and predict sales figures.

3. **Temporal Aspect Integration:** Incorporating temporal dynamics and seasonality within the LSTM model to account for evolving user preferences, seasonal trends, and dynamic demand patterns. This enables the model to adapt and improve its predictive accuracy over time.

4. **Performance Evaluation:** Rigorous evaluation of the LSTM-based predictive model using established metrics such as MSE, MAE, RMSE, and R-squared. This step ensures the accuracy and reliability of the model's predictions.

5. **Interpretability and Insights:** Emphasizing the interpretability of the model's predictions, providing valuable insights into consumer behaviors, trends, and factors influencing sales. This aspect aims to bridge the gap between predictive accuracy and actionable insights for businesses.

By leveraging these strategies, the proposed solution aims to develop a robust LSTM-based sales prediction model tailored to the complexities of e-commerce data from Taobao. This model strives to accurately forecast sales figures while offering meaningful insights to enhance decision-making in the e-commerce domain.

PART 2

LITERATURE REVIEW

E-commerce has witnessed an unprecedented surge in recent years, fundamentally altering the landscape of global commerce [7],[8]. This digital revolution has not only redefined consumer behavior but also reshaped the way businesses operate and interact with their target audience [9]. Within this dynamic environment, accurate sales prediction has emerged as a cornerstone of successful e-commerce management [9],

and the growth of e-commerce affects employment in many sectors such as the retail sector [10],[11].

In the 21st century, online shopping has become integral due to the convenience it offers individuals amidst their busy schedules [12]. This mode of shopping, known as business-to-consumer online shopping, has rapidly gained a global foothold. E-commerce, reshaping consumer behavior, prioritizes convenience, timesaving, and cost-effectiveness. The internet, acting as a platform influenced by social circles, redefines shopping habits. As online shopping sites replace traditional stores, consumer trust in these platforms grows. The surge in e-commerce competition yields better and more affordable products, driving a significant shift in consumer preferences and habits [13],[14].

Now I have two figures illustrating the growth rate of the United States in Figure 2.1 and Figure 2.2. I have China and this shows the growth rate from 2018 until 2027 [15], The comparison indicates distinct patterns in growth rates, signifying the varying dynamics of e-commerce development in these markets.

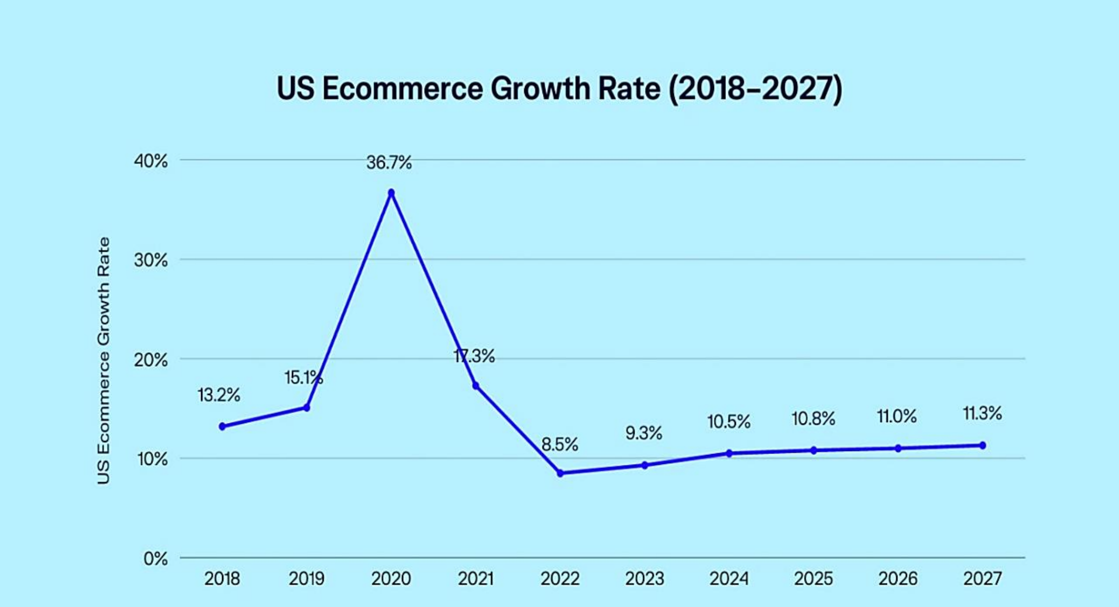


Figure 2. 1. The growth rate of e-commerce value in the United States.

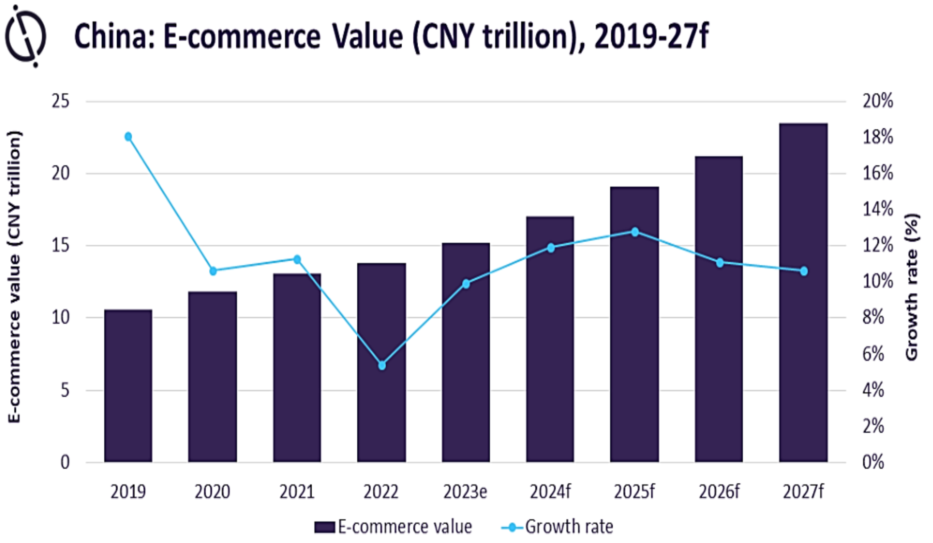


Figure 2. 2. The growth rate of e-commerce value in China

2.1. MACHINE AND DEEP LEARNING IN SALES PREDICTION

Within the domain of e-commerce, machine learning plays a fundamental role in deciphering and prognosticating sales trends [16]. Its application, facilitated by intricate algorithms and extensive data analysis, empowers businesses to anticipate consumer behavior [17], forecast demand patterns, and optimize sales strategies. The utilization of machine learning techniques enables the extraction of valuable insights from vast datasets, thereby aiding in the identification of emerging trends and fostering data-driven decision-making processes within e-commerce environments [18]. So, machine learning has a deep effect on sales prediction and forecasting many machine learning algorithms such as XGBoost [19],[20],[37], and here is in figure 2.3 declare how sales in Walmart stores within 10 months [20].

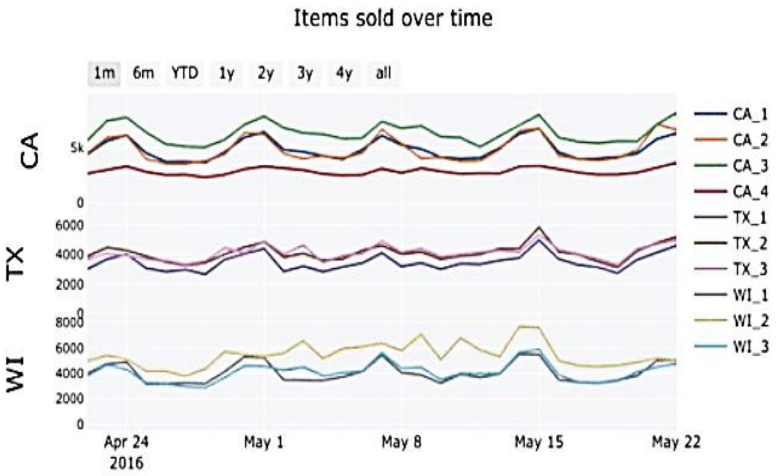


Figure 2. 3. Daily sales in a single month in Walmart stores [20].

Multiple Regression, Polynomial Regression, Ridge Regression, Lasso Regression [20], Artificial Neural Network with backpropagation [21], another study explores the problem of sales prediction using linear regression and KNN regression with Rossman dataset and also another model such as ARIMA, XGBoost, Auto Regressive Neural Network (ARNN), Support Vector Machine (SVM), Hybrid Models [22], and the one we will use it which is LSTM, it use the sequential time series [23].and they keep evaluate The performance using many metrics such as MAE and RMSE and MSE so all matrices is to evaluate the performance and as much as the number is less the much you model perform better[24], [25]. another study underscores the need for an intelligent sales prediction system using machine learning to handle large data volumes and ensure timely and accurate business decisions. Despite challenges, the research utilized approximately 85,000 records for algorithm comparison, recognizing the importance of Big Data in predictive analytics for sales forecasting in modern business scenarios, Figure 2.4 shows the yearly sales predicting food sales is an important field in sales prediction and it can be beneficial for the business in long or short-term decisions [26],[27].

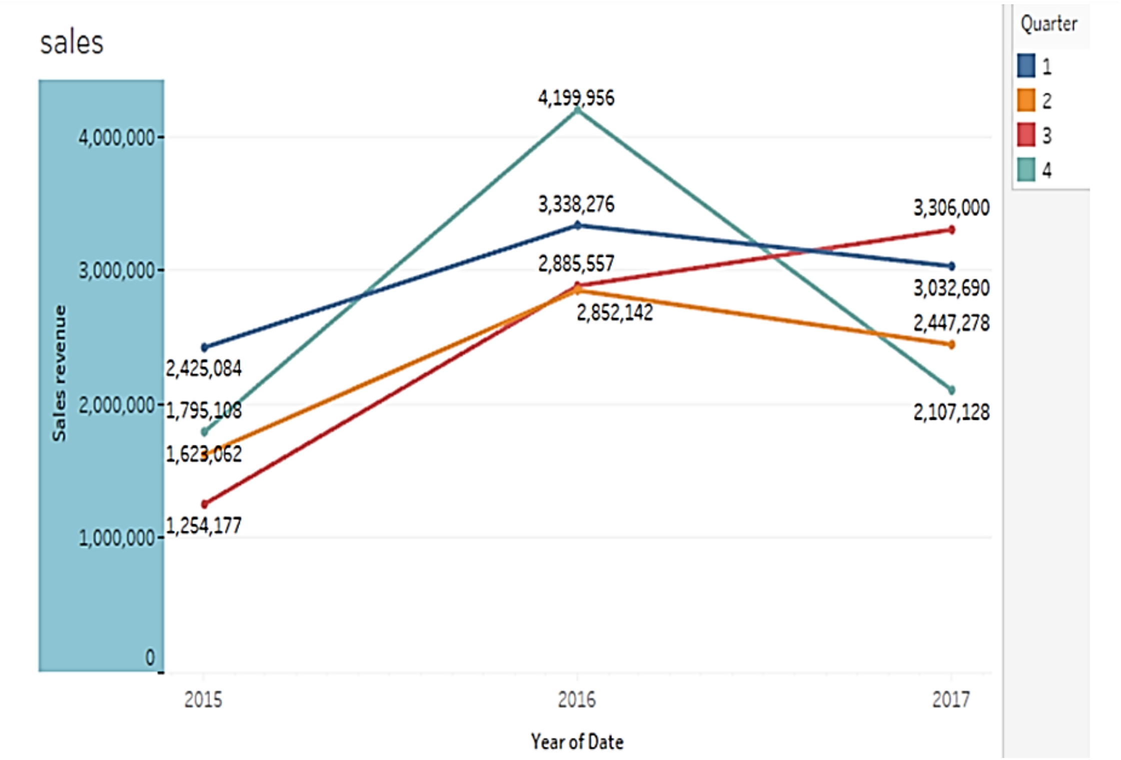


Figure 2. 4. Yearly sales [26].

2.2. MACHINE AND DEEP LEARNING IN E-COMMERCE

The integration of Machine Learning (ML) in e-commerce enhances user-friendly, secure, and profitable online platforms. Driven by strong competition, brands leverage ML for transformative solutions, making it one of the fastest-evolving technologies in e-commerce. ML applications, particularly in mobile and web platforms, excel in pattern recognition, analytics, and personalized experiences, surpassing human capabilities [28]. Amid the digital transformation impacting various business realms, using ML, especially marketing, cashback websites have emerged as partners for large retail brands seeking to engage customers in an era of fleeting loyalty. These platforms employ a unique affiliate marketing approach, offering financial rewards based on customer activities.[29].

There is a proposed system that can analyze and predict online shopping behavior based on customer data and rules.to overcome the traditional way The system uses two types of machine learning models which are logistic regression and XGBoost, and here in Figure 2.5 shows the results for them [30].

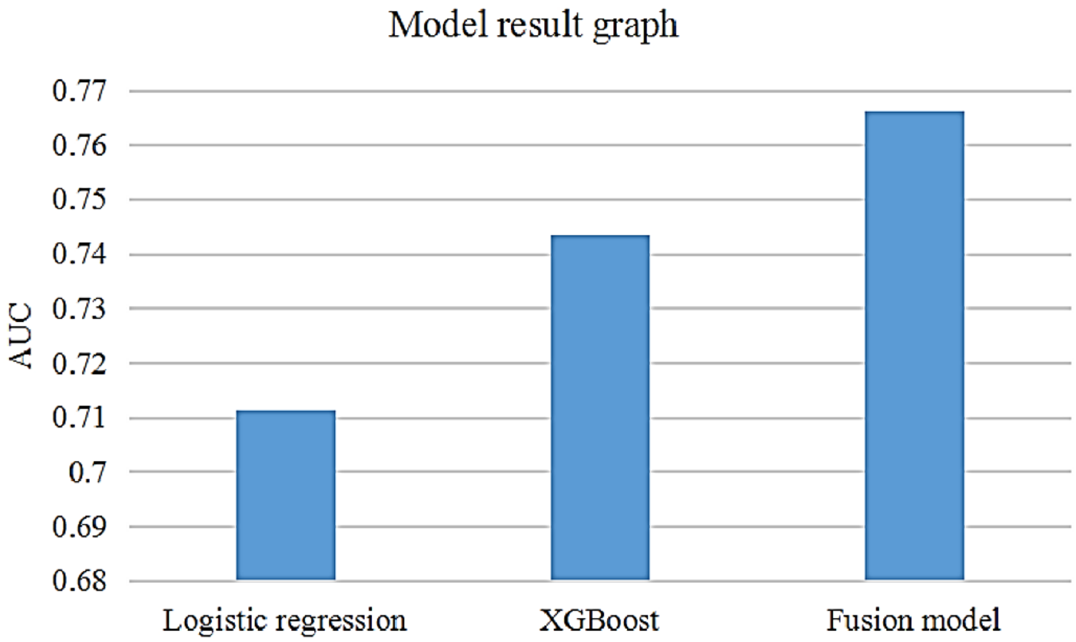


Figure 2. 5. LR, XGBoost, Fusion model results.

The paper compares and shows the performance of these models and shows that combining them improves the prediction accuracy. It also demonstrates that the system can reduce the complexity and over-fitting of the models by filtering the features [31], [32]. According to another paper in 2019, a total of $603 billion worth of sales were made via e-commerce in the United States compared to 3.17 billion in retail sales in the United States [33].

The significant effect of logos on the sales of e-commerce platforms requires analyzing 1420 different Romanian companies by specific research [34]. Other issues in classical collaborative filtering, such as poor predictive accuracy, by adopting SVM for item classification. The SVM–IACO–CF (Collaborative Filtering) system is introduced, with optimized SVM parameters obtained through the IACO algorithm. The system filters out user-disliked items, considers scores and comments for positive feedback items, and improves predictive accuracy through weighted averages.[35]. Another article explores customer segmentation through unsupervised machine learning, emphasizing the importance of differentiated strategies for diverse online customers. The study highlights the significance of feature engineering, using the RFM model to quantify purchasing behaviors. The identified customer groups assist businesses in understanding customer behaviors and adopting tailored marketing strategies [36]. Other models used images for training their model This proposed model focuses on enhancing e-commerce user experience by providing similar product image recommendations based on unsupervised statistical machine learning. It employs PSVD dimensionality reduction and K-means++ clustering for effective image grouping. Evaluation metrics, including SC coefficient and CH score, indicate strong performance [37], [38].

2.3. MATRICES EVALUATION IN MACHINE LEARNING

The focus was on deriving statistical properties of RMSE and MAE estimators for zero mean symmetric error distributions. The Approximate Root Normal Distribution (ARND) was introduced to approximate the distribution of the square root of a normal random variable, aiding in estimating the distribution of the RMSE estimator [39][40]. another study aimed to provide an overview and classification of performance metrics in various domains, focusing on machine learning regression, forecasting, and prognostics. The research resulted in a proposed typology with four categories: primary metrics, extended metrics, composite metrics, and hybrid sets of metrics [41]. by exploring the benefits of using the MAE loss function in Deep Neural Network (DNN) based vector-to-vector regression. The research highlights the Lipschitz continuity property, demonstrating its role in establishing a performance upper bound for DNN-based regression and predicting robustness against additive noises. The MAE loss function is associated with Laplacian distribution, and experimental results indicate that optimizing DNN-based regression with MAE leads to lower loss values compared to MSE.[42],[43]. another paper introduces a modified version of R2 designed for assessing feature importance in both linear and non-linear machine learning models. The metric leverages the additive property of Shapley values to equitably distribute a model's explained variance to individual features, eliminating the need to retrain the model on various feature subsets [44]. While the proposed metric offers desirable properties such as a 0 to 1 scale and feature-level variance decomposition, further investigation is required to assess its stability in the face of overfitting and hyperparameter tuning.[45],

2.4. RELATED WORK

Previous studies have deeply explored the application of machine learning techniques in sales prediction within e-commerce contexts [46]. It delved into the use of Recurrent Neural Networks (RNNs) to model sequential behavior in online shopping patterns [47], [48], Similarly, another paper conducted a comparative analysis of various machine learning algorithms, including Decision Trees (DT) and SVMs [49], and also there is Alibaba group themselves who used their model called Aliformer, random forest and gradient boosting models[50], emphasizing the superiority of LSTM networks in capturing long-term dependencies in e-commerce datasets for sales prediction [51], [52].

There are many datasets used for training machine learning algorithms like the dataset from the Turkish platform n11[4], and there are other datasets from Tmall that contain snacks and Flagships products [6], in another search, they used a dataset collected from Alibaba Group which contains features like page view, user view, selling prices and view from search[13], and Alibaba group they use their dataset from Tmall Merchandise which contains 1.2 million records and it has features like item page view, unique visitor, gross merchandise volume, the order count and the buyer count and there is the dataset collecting from Taobao platform and has many features like profile visit and porches, demonstrating how preprocessing techniques such as normalization and sequence padding enhanced LSTM model performance in predicting customer purchase behavior[53]. Furthermore, employed a similar dataset to predict sales based on user browsing history, emphasizing the significance of feature engineering to capture nuanced behavioral patterns to get accurate sales forecasting. Recent research delved into the architecture of LSTM networks specifically designed for e-commerce [54], showing their ability to capture sequential dependencies in user behaviors, such as clicks and purchases. Their study highlighted the capability of LSTM models in predicting sales trends influenced by evolving consumer preferences in dynamic e-commerce environments.

Evaluation metrics like MSE [55], MAE [56], Root Mean Squared Error (RMSE) [57], and R-squared [58] have been consistently utilized in previous studies. For instance, [59] employed the RMSE to assess LSTM model performance in sales prediction, they used MSE and MEA with their model Aliformer and achieved good performance, and in [5] they used MSE with CNN, and in [60] they used different metrics like [61] Root Mean Square Percentage (RMEPE) and Mean Absolute Percentage Error (MAPE) [62] and it shows notable effectiveness of LSTM in accurately forecasting sales.

Table 1 shows the related work in a comparison table that shows each research and its used model with the dataset used, the measurement performance results using the mentioned matrices, and in which year the paper was published.

Table 2. 1. Related Work Comparison

Title	**Year**	**Model/Technique**	**Accuracy/ performance**	**Dataset**
Demand Prediction Using Machine Learning Methods and Stacked Generalization	2020	Stacked Generalization, LR, DT, GBT, RF	Using RMSE 1864	n11 platform dataset
A Deep Neural Framework for Sales Forecasting in E-Commerce	2019	DSF (Sales related features, the RNN-based Seq2Seq, and the sales residual)	Using MAE 228	Snack and PG&U
Sales Forecast in E-commerce using Convolutional Neural Network	2017	CNN	Using MSE 145.69	Alibaba Group
From Known to Unknown: Knowledge-guided Transformer for Time-Series Sales Forecasting in Alibaba	2021	Aliformer	Using MSE 0.154 Using MAE 0.229	Tmall Merchandise Sales Dataset, Kaggle-M5
Sales Prediction based on Machine Learning	2021	LR, LSTM, MLP, XGBoost	Using RMSE 557.56 (from LSTM)	M5 by Walmart

PART 3

MATERIALS AND METHODS

3.1. SYSTEM BLOCK DIAGRAM

Our block diagram in Figure 1.1 represents and illustrates the components and their interactions within a system. For an e-commerce sales prediction system using LSTM models, the block diagram might include:

1. **Data Collection Module:** This module encompasses data acquisition processes from various sources, such as user interactions (clicks, purchases, browsing patterns) from the e-commerce platform (Taobao) and other relevant external datasets.

2. **Data Preprocessing:** This segment includes steps like data cleaning, normalization, feature extraction, and transformation, preparing the data for model training.

3. **LSTM Model Module:** This central module involves the LSTM architecture designed for forecasting sales trends. It includes layers for LSTM cells, dense layers for output, and connections between these layers.

4. **Training and Validation:** This section covers the iterative process of training the LSTM model using historical data and validating its performance using validation datasets.

5. **Evaluation and Metrics:** The module evaluates the model's predictions against actual sales figures using metrics like MSE, RMSE, MAE, and R-squared to quantify the model's accuracy.

6. **Prediction and Insights:** After successful training and evaluation, this module represents the LSTM model's deployment to make real-time sales predictions. It also includes interpreting the model's insights for actionable business strategies.

7. **Optimization and Feedback Loop:** This part involves using insights from the model's predictions to optimize marketing strategies, inventory management, and other sales-related aspects. It might also include feedback loops to improve future predictions.

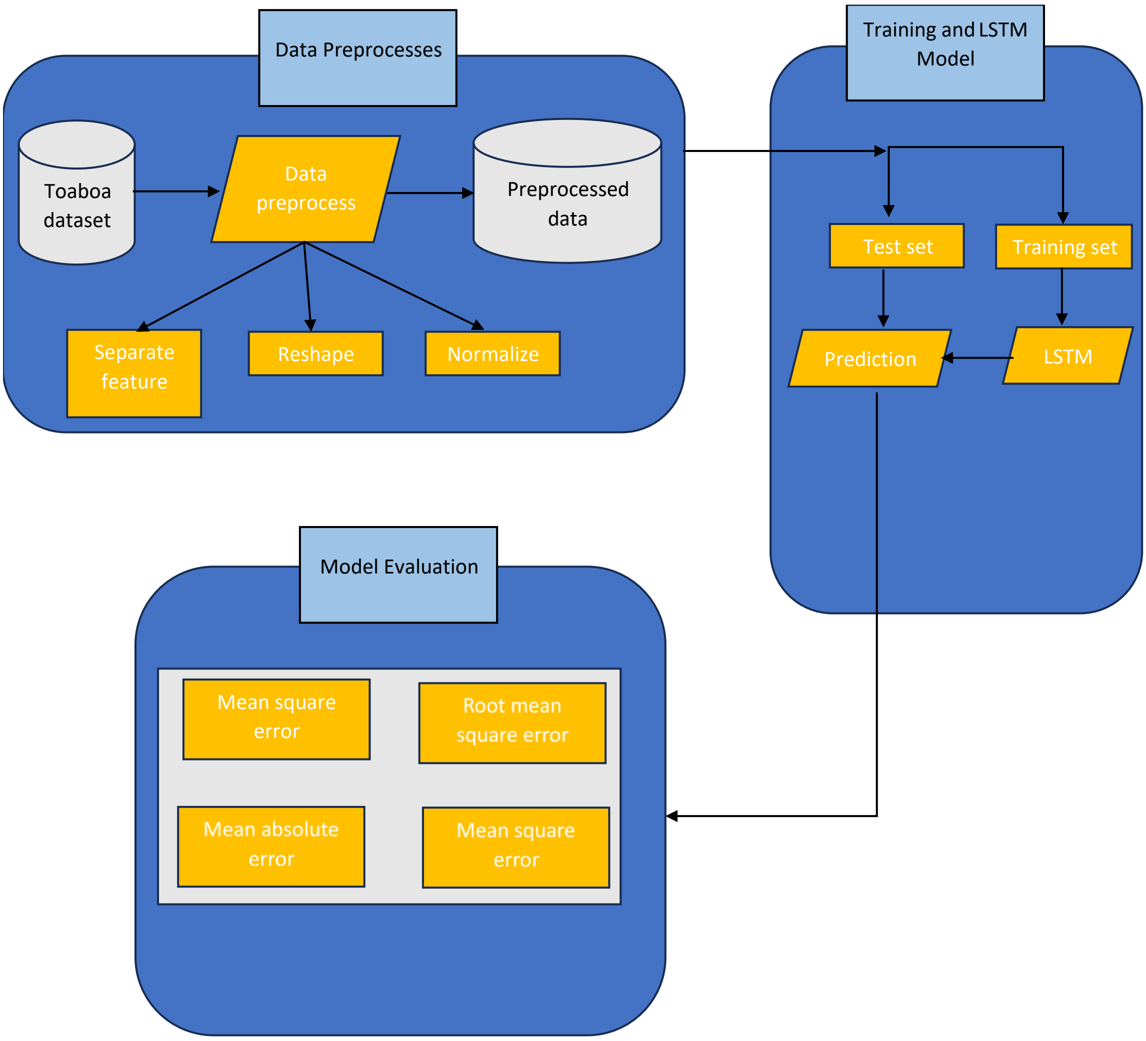


Figure 3. 1. Block diagram of the sales prediction model.

The rest of the thesis will be structured as follows: Part 2 will introduce the literature review, while Part 3 will contain the proposed methodologies and tools. Part 4, on the other hand, will show and illustrate the experiments and results, in the meanwhile the discussion and conclusion will be listed in Part 5.

3.2. THE PROPOSED METHODS

Our proposed method of the current study contains the following steps that are illustrated in Figure 3.1:

1- Obtain dataset: one of the main parts of the study is the dataset since every next step is based on the selection of the dataset. For this study, we get data from the Toaboa platform which is a large and reliable dataset containing 1 million random records for the sales that happened in 2017 and it is available on Kaggel for free.

2- Feature Engineering and Selection: Engineered features include creating new features to make it easier to train the dataset like 'DayOfWeek' from the timestamp and categorical encoding of 'BehaviorType.' to buy_count. Fav_count, pv_count, cart_count The primary feature selected for predicting sales is 'Buy_Count,' reflecting the user's purchase behavior as a key predictor of sales.

3- Model Architecture and Algorithm: The model employs a machine learning algorithm suitable for regression tasks, given the goal of predicting sales. Which is the LSTM model it is a good and reliable algorithm

4- Training and Evaluation: which include

Data Splitting: this is splitting the data into training and evaluation sets to train and assess the model's performance.
Training Process: Describe how the model is trained, including any hyperparameter tuning for optimal performance. Evaluation Metrics: Employ metrics like MSE, RMSE, or MAE for assessing how well the model predicts sales.

3.2. MATERIALS

3.2.1. Dataset

In the current study, we used the User Behavior dataset which is a dataset that contains user behaviors from Taobao, it is basically for recommendation problems with implicit feedback. The dataset is created by Alibaba. This dataset contains a random selection of 1 million users who have behaviors including purchasing, adding items to the shopping cart, and item favoring, click, from November 25 to December 03, 2017. The dataset structure is as follows. each line represents a specific user-item interaction, which contains the user ID, item ID, item's category ID, behavior type, and timestamp, separated using commas.

3.2.2 Software

Our model will be implemented in Python programming language. Using Visual Studio Code IDE, several libraries will be used, including the following:

1- NumPy: for numerical computing that provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays.

2- Sklearn: sickit learn library which is used to build, train, and evaluate machine learning models.

3- Panda: for data manipulation and analysis, offering easy-to-use data structures like DataFrame and Series.

4- tqdm: is used for adding progress bars to loops, making it easier to track the progress of iterative processes.

5- Tensorflow: to build and train the LSTM model.

3.3. DATA PREPROCESSING

The dataset obtained from Taobao was initially composed of user-item interactions done by user ID, item ID, item category ID, behavior type, and timestamp. To facilitate robust analysis and effectively train machine learning models, a series of preprocessing steps were undertaken.

3.3.1. Data structure

The dataset structure as shown in Figure 3.1 comprises user-item interactions, delineated by different attributes, separated by commas. Each interaction entry encapsulates some details, including the user's unique identifier, the respective item's ID, the item's category ID, the type of behavior exhibited, and the timestamp marking the interaction. Figure 3.1 shows a sample of the dataset.

Table 3. 1. Dataset Structure

Item ID	Category ID	Product ID	Behavior type	Timestamp
1	3827899	2131531	pv	1511684109
1	4973305	2520771	pv	1511684032
100	4840649	1575622	pv	1511766462
100	510936	3738615	buy	1511742899
100	251391	4896062	pv	1511742875
100	2574432	1029459	pv	1511765639
100	3245421	1046201	pv	1511765568
100	1046201	3895389	buy	1511765548
1000	2956844	2419959	cart	1511865567
1000	4487355	1653613	fav	1511866709
1000	3302343	2342116	pv	1511657410

3.3.2. Data Transformation

The dataset was reformatted to incorporate a structured representation of user behaviors. Categorical actions such as 'pv' (page view), 'fav' (item favoring), 'buy', and 'cart' (adding items to the shopping cart) were transformed into binary indicators, resulting in columns 'Behavior_buy', 'Behavior_cart', 'Behavior_fav', and

'Behavior_pv'. This conversion simplified qualitative behavioral data into a machine-interpretable format. It aims to enhance the readability and utility of the dataset. By converting qualitative user behaviors, such as 'pv', 'fav', 'buy', and 'cart', into binary indicators ('true' for presence and 'false' for absence), as shown in Figure 3.2 the dataset achieved a structured format that machine learning models can readily comprehend. This process effectively translated complex behavioral patterns into a format suitable for quantitative analysis, simplifying the complexity of qualitative behavioral data. Machine learning algorithms typically work more effectively with structured and numeric data, so converting behavioral actions into binary form creates a clear, machine-readable representation. This transformation allows predictive models, such as LSTM networks, to efficiently recognize and discern patterns within the data.

Table 3. 2. Data Transformation

userID	ItemID	CategoryID	Day of week	Tstamp	Cart	Fav	pv	Buy
1	4973305	411153	4	1511865567	false	false	true	false
1	3827899	2891509	5	1511765568	false	false	true	false
1	2266567	4145813	6	1511765548	false	false	true	false
1	1531036	2355072	6	1511742875	false	false	true	false
1	3745169	2520771	0	1511684032	false	false	true	false
1	2286574	149192	1	1511684109	false	false	true	false
1	230380	2520771	3	1511871096	false	false	true	false

3.3.3. Feature Extraction

Integrate the behavioral count metrics, namely 'Buy_Count,' 'Cart_Count,' 'Fav_Count,' and 'PV_Count,' was an important step in our data preprocessing phase aimed at enhancing the predictive power of our models for sales forecasting in the e-commerce domain.

These count metrics serve as essential predictors, offering a quantitative representation of user engagement within the platform. The 'Buy_Count' signifies the frequency of user purchases, 'Cart_Count' and 'Fav_Count' capture the instances of adding items to the cart and favoriting products respectively, while 'PV_Count' traces the frequency of page views.

By quantifying user actions, our models gain a deeper insight into user behavior patterns and their correlation with subsequent sales. These count metrics, act as additional features in our predictive models and provide an accurate understanding of user interaction intensity, allowing our machine learning algorithms, including LSTM networks, to discern significant behavioral patterns that influence sales trends. The utilization of these count metrics not only improves the predictive capabilities of our models but also offers interpretability and readiness for machine learning algorithms. These features, derived from user behavior frequencies, refine the model's ability to anticipate sales dynamics, contributing to more accurate and reliable sales predictions within the dynamic landscape of e-commerce. Figure 3.3 is a sample that illustrates the change and the pressing that happened on the dataset, and you can see the difference between it and Figure 3.1.

Table 3. 3. The Preprocessed Data

userID	**ItemID**	**CategoryID**	**Day of week**	**Tstamp**	**Cart**	**Fav**	**pv**	**Buy**	**buyout**	**CartCount**	**FavCount**	**PvCount**
1	4973305	411153	4	1511865567	false	false	true	false	2	3	3	117
1	3827899	2891509	5	1511765568	false	false	true	false	3	1	1	32
1	2266567	4145813	6	1511765548	false	false	true	false	0	2	2	34
1	1531036	2355072	6	1511742875	false	false	true	false	0	1	1	11
1	3745169	2520771	0	1511684032	false	false	true	false	1	11	8	233
1	2286574	149192	1	1511684109	false	false	true	false	0	0	0	4
1	230380	2520771	3	1511871096	false	false	true	false	2	16	12	249

3.3.4. Data Scaling/Normalization

Normalization Technique:

The dataset underwent Min-Max scaling as a normalization technique. This method scales the numerical features to a specific range, typically between 0 and 1. In our implementation, the `MinMaxScaler` from scikit-learn was employed for this purpose.

Impact on the Data:

Min-Max scaling is applied to ensure that all features contribute equally to the training process, preventing certain features with larger scales from dominating the learning process. This normalization helps in stabilizing the training of the LSTM model and enhances its convergence.

3.3.5. Data Sampling

To manage computational resources and training time, a systematic data sampling approach was implemented. Specifically, a random sample consisting of 50 million rows was drawn from the original dataset. This sampling strategy addresses the challenges associated with working with large datasets, ensuring a representative subset for training the LSTM model.

Justification for Data Sampling: The decision to sample the data was driven by the consideration of computational efficiency. Training machine learning models, especially deep learning models like LSTM, can be computationally intensive, and working with the entire dataset might be impractical in terms of time and resources. By randomly selecting a substantial subset, we strike a balance between computational feasibility and the desire to capture meaningful patterns within the data.

3.3.6. Dataset Splitting:

The dataset was split into training and testing sets using the `train_test_split` function from the scikit-learn library. This function randomly shuffles and divides the data into two sets based on the specified split ratio.

Justification of Split Ratio (80% training, 20% testing): The choice of an 80% training and 20% testing split ratio was motivated by the need to strike a balance between having enough data to train the LSTM model effectively while still reserving a sufficient portion for evaluating its performance. This ratio is a commonly used practice in machine learning, providing a substantial amount of data for training to capture patterns and trends, while the remaining portion serves as an independent dataset for assessing the model's generalization ability.

3.4. MODEL IMPLEMNTAION

3.4.1. LSTM Model Architecture

The LSTM model was constructed using TensorFlow/Keras, which is a framework for building neural networks. The architecture consists of an LSTM layer with 128 units as the primary layer for sequential data processing. This was used by a Dense layer with a single output, suit to predict sales figures based on the input features.

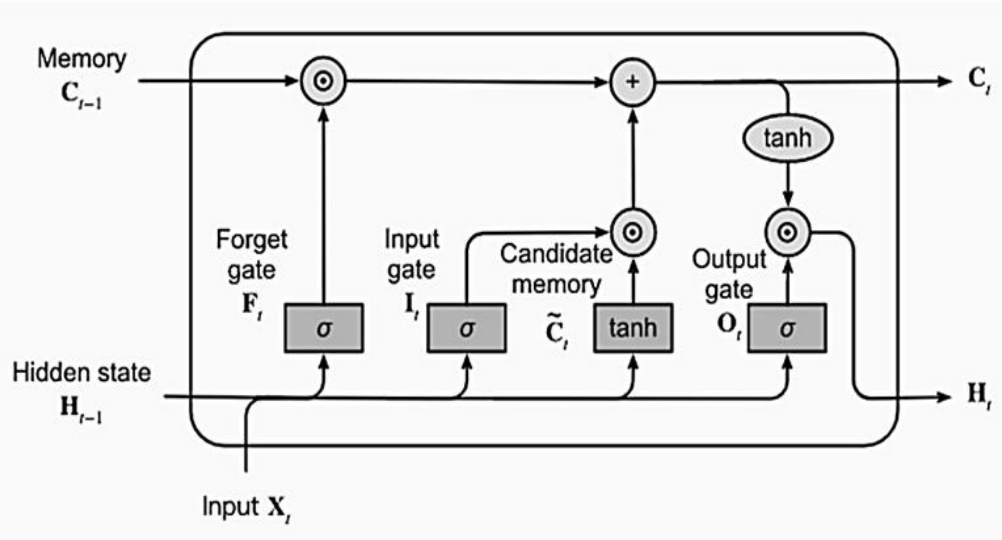


Figure 3. 2. LSTM Architecture

And we will dive more into LSTM Architecture:

LSTM Layer:

- The LSTM layer is a special type of RNN layer designed for capturing and remembering long-range dependencies in data sequential.

- The choice of using an LSTM layer is motivated by its ability to effectively model temporal patterns, crucial in tasks such as sales prediction where past behavior influences future outcomes.

- **Number of Units (Neurons):**

- The LSTM layer is configured with 128 units. Each unit, or neuron, within the LSTM layer, functions as a memory cell. These memory cells can store information over time, allowing the model to capture and utilize patterns in the sales data that occur over different time steps.

- **Input Shape:**

- The input_shape parameter is set to (1, X_train.shape[1]), indicating that the model expects input sequences with one time step and several features equal to the dimensionality of the dataset. This configuration is chosen to align with the format of the preprocessed data.

- **Activation Function (for each LSTM cell):**

- Although not explicitly specified in the code, the default activation function for the LSTM layer is typically the hyperbolic tangent (tanh). The tanh function is commonly used in LSTM layers to introduce non-linearity and facilitate the learning of complex temporal patterns.

Dense Output Layer:

- The Dense layer is a fully connected layer that follows the LSTM layer, responsible for producing the final output of the model.

- **Number of Units in Output Layer:**

- The Dense layer is configured with a single unit. In the context of sales prediction, this implies that the model is tasked with predicting a single continuous numerical value, specifically the sales count.

- **Activation Function (for the output layer):**

- The output layer does not explicitly specify an activation function, implying the use of the default linear activation function. For regression tasks, linear activation is common as it allows the model to produce unbounded continuous output, suitable for predicting numerical values without imposing constraints.

3.4.2. Compilation

The Adam optimizer and the MSE loss function are specified.

Optimizer (Adam):

The Adam optimizer is a popular choice in deep learning for its adaptive learning rate properties. It dynamically adjusts the learning rates for each parameter during training, allowing the model to converge efficiently and effectively update its weights.

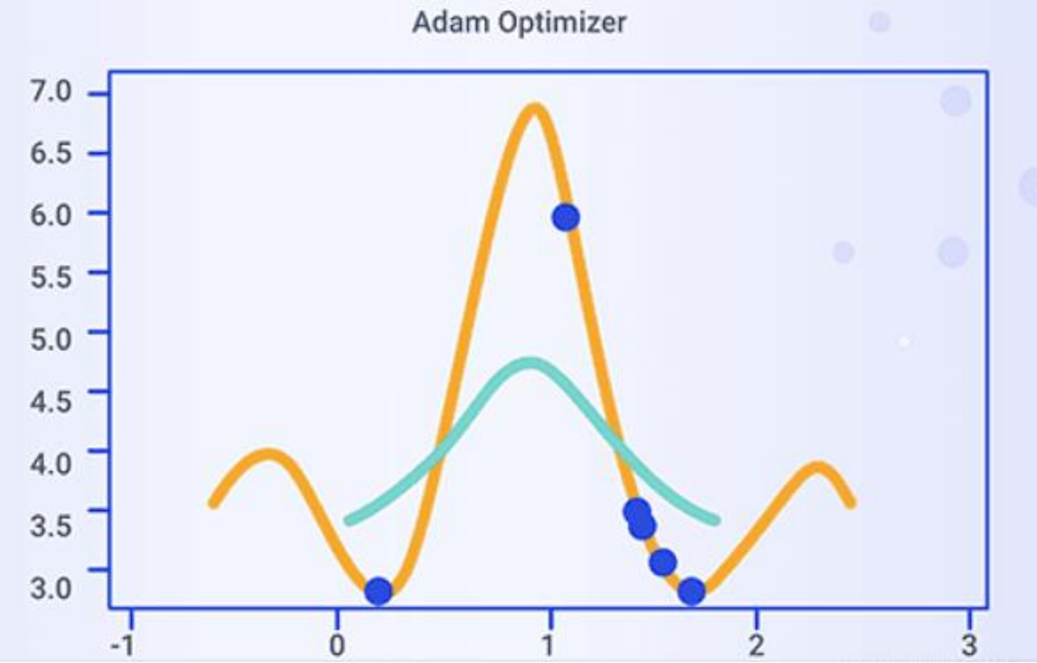


Figure 3. 3. Adam optimizer.

Loss Function (MSE):

The MSE is employed as the loss function for the compilation. MSE is a common choice for regression tasks, where the goal is to minimize the squared difference between the predicted values and the actual values. It quantifies the average squared deviation between the predicted and true sales counts.

3.4.3. Training Process:

a. Epochs and Batch Size:

Number of Training Epochs:

The model undergoes training for 10 epochs, meaning it processes the entire training dataset 10 times. The choice of epochs involves a trade-off between providing the model with sufficient exposure to the data and avoiding overfitting.

- **Justification:** The selection of 10 epochs should be justified based on the model's convergence and performance on both the training and validation sets. If the model shows signs of converging and stabilizing in terms of performance, it indicates a suitable number of epochs.

Batch Size:

The training data is divided into batches, with each batch containing 64 samples. The batch size affects how the model updates its weights during training.

- **Justification:** A batch size of 64 is chosen, balancing computational efficiency and training stability. Larger batches can speed up training but may require more memory, while smaller batches may introduce more variability in weight updates. The choice of 64 should be justified considering computational resources and the desired trade-off between stability and speed.

b. Training Procedure:

- **Train_on_batch Method:**
    - The model is trained using a custom training loop, utilizing the train_on_batch method. This method allows the model to learn from one batch of data at a time.
- **Explanation:**
    - In each iteration of the training loop, the model updates its weights based on the loss computed for a single batch. Using train_on_batch provides fine-grained control over the training process and is particularly useful for customizing training procedures.

c. Convergence Checking:

- **Strategies for Monitoring Convergence and Avoiding Overfitting:**
    - **Validation Set:**
        - The dataset is split into training and validation sets. Monitoring the model's loss on the validation set during each epoch helps assess its generalization to unseen data.
    - **Early Stopping:**
        - Early stopping as shown in 3.6 is employed, meaning that if the model's performance on the validation set stops improving, or starts degrading, training is halted. This prevents overfitting and conserves training time.

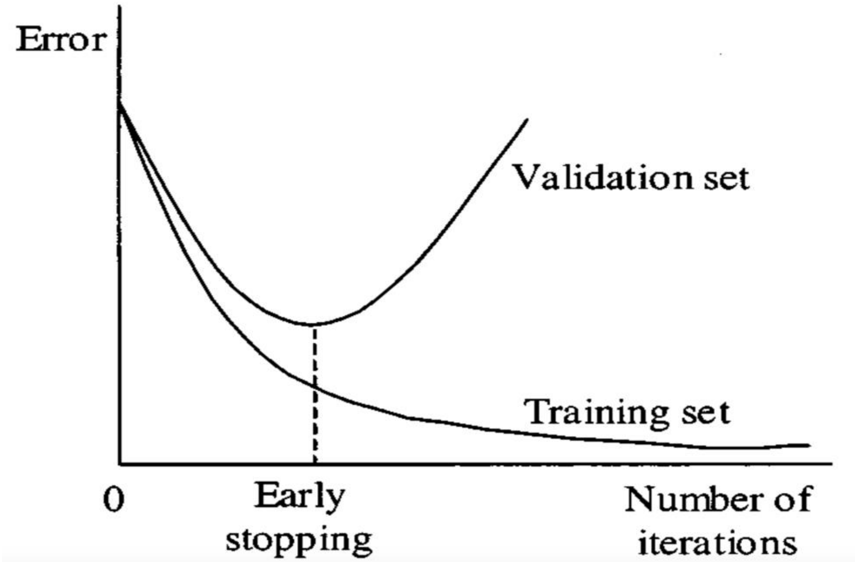


Figure 3. 4. Early stopping.

Regularization Techniques:

Dropout:

Dropout is a technique for regularization it is used during the training of neural networks to prevent overfitting. It includes randomly "dropping out" a fraction of neurons during each training iteration.

How Dropout Works:

During each training iteration, a random subset of neurons is deactivated (dropped out). This means that their output is set to zero, and they do not contribute to the forward or backward pass. The specific subset of dropped-out neurons changes with each iteration.

Dropout introduces a form of stochasticity into the network, preventing neurons from relying too heavily on specific features. This helps in creating a more robust and generalized model that performs well on unseen data. Dropout is typically added between layers during the construction of the neural network using specific dropout layers.

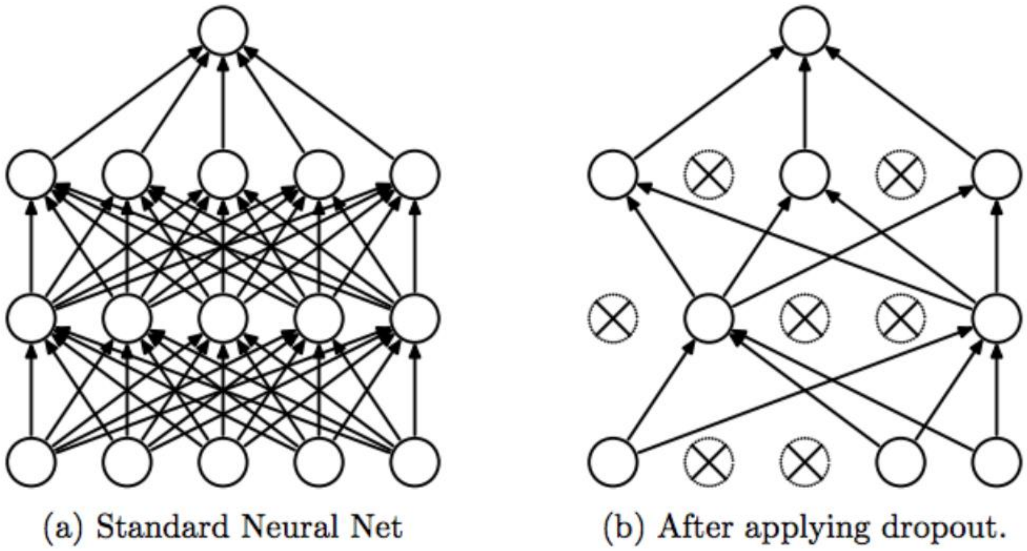


Figure 3. 5. Dropout

L2 Regularization:

Definition: L2 regularization, also known as weight decay, is a regularization technique that penalizes large weights in the neural network by adding a term to the loss function.

How L2 Regularization Works:

In the context of neural networks, L2 regularization adds a penalty term to the loss function that is proportional to the squared magnitude of the weights. The goal is to discourage overly complex models with large weights.

Effect on Overfitting: L2 regularization helps prevent overfitting by discouraging the model from relying too much on specific features. It achieves this by penalizing large weights, making the optimization process favor simpler models with more evenly distributed weights.

Implementation: In Keras, L2 regularization can be applied to layers during model construction.

Learning Rate Schedule:

A learning rate schedule involves adjusting the learning rate dynamically during training. This adaptability can be advantageous for optimizing the model's convergence and stability over time. Although your current code uses a fixed learning rate, you might consider implementing a learning rate schedule for potential enhancements.

In a learning rate schedule, the learning rate can be modified based on certain criteria, such as reducing it over time. For instance, you might start with a higher learning rate and gradually decrease it as the training progresses. This approach helps prevent the model from making overly large updates, which can be useful in achieving a more stable and accurate convergence.

3.4.4. Model Performance Evaluation

a. Choice of Metrics:

- **Explanation of Metrics Choice:**

  - In your code, the following evaluation metrics are chosen:

    - MSE
    - RMSE
    - MAE
    - R-squared (R²)

- **Justification:**

    - **MSE:**

- MSE is chosen as it measures the average squared difference between the predicted and actual values. Squaring the differences penalizes larger errors more heavily, making it suitable for regression tasks.

- **RMSE:**

  - RMSE is the square root of MSE and provides a more explainable measure of error. It is chosen to give a sense of the average magnitude of the errors in the same units as the target variable.

- **MAE:**

  - MAE is selected as it measures the average absolute difference between the predicted and actual values. It provides a straightforward interpretation of the average error magnitude.

- **R-squared (R²):**

  - R-squared (R²) quantifies the extent to which the variability in the dependent variable can be explained by the independent variables included in the model. It serves as a metric to evaluate how well the model fits the observed data.

b. Interpretation of Metrics:

MSE:

- **Formula:**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \check{y})^2$$

- **Interpretation:**

    - A lower MSE indicates better model performance. It penalizes larger errors more, making it sensitive to outliers.

RMSE:

- **Formula:**

$$RMSE = \sqrt{MSE}$$

- **Interpretation:** RMSE provides an interpretable measure in the same units as the target variable. A lower RMSE signifies better model performance.

MAE:

- **Formula:**

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Interpretation:** MAE represents the average absolute difference between predicted and actual values. It is easy to interpret and provides a measure of average error magnitude.

R-squared (R²):

- **Formula:**

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Interpretation:** R-squared ranges from 0 to 1, with 1 indicating a perfect fit. It measures the proportion of variance in the dependent variable explained by the model. Higher R^2 values suggest a better model fit.

PART 4

RESULTS AND FINDINGS

The LSTM model was trained for sales prediction in e-commerce using a dataset of 50 million records. Due to resource constraints, the entire dataset could not be utilized, and a random sample was chosen for analysis. The experimental setup included preprocessing steps, such as feature scaling and normalization, and the LSTM architecture was configured with one layer consisting of 128 units.

Model Training and Convergence:

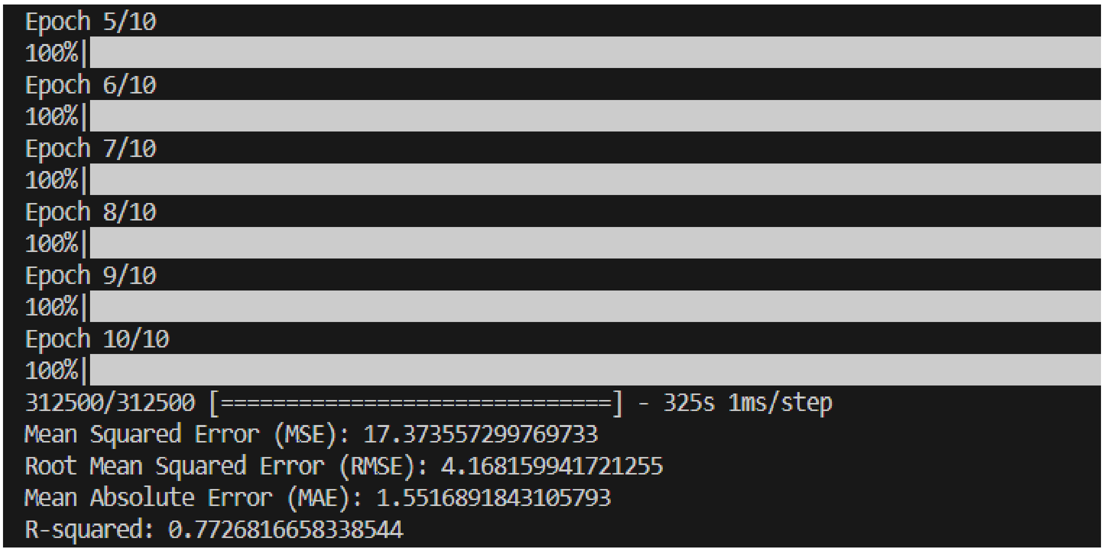
The model was trained for 10 epochs with a batch size of 64. Despite the substantial dataset, resource limitations, specifically a 32 GB RAM constraint, necessitated working with a subset of the data. The training process aimed to capture temporal dependencies in the time series data, focusing on predicting the 'Buy_Count' variable.

Evaluation Metrics:

The performance of the trained LSTM model was assessed using key evaluation metrics and compared with results from other relevant papers:

- **MSE:** 17.37
- **RMSE:** 4.16
- **MAE:** 1.551
- **R^2:** 0.772

Figure 4.1 shows the results of my model which was trained by a Python program using the IDE Visual Studio code.



```
Epoch 5/10
100%|
Epoch 6/10
100%|
Epoch 7/10
100%|
Epoch 8/10
100%|
Epoch 9/10
100%|
Epoch 10/10
100%|
312500/312500 [==============================] - 325s 1ms/step
Mean Squared Error (MSE): 17.373557299769733
Root Mean Squared Error (RMSE): 4.168159941721255
Mean Absolute Error (MAE): 1.5516891843105793
R-squared: 0.7726816658338544
```

Figure 4. 1. Results of the model.

Comparative results with other papers in the field are presented in Table 4.1.

Table 4. 1. Comparison table of results.

Paper title	MSE	RMSE	MAE	R-Squared
[4]	-	1864	-	-
[6]	-	-	228	-
[5]	145.69	12.07	-	-
[3]	0.154	0.380	0.229	-
[8]	557.65	23.61	-	-
Our study	**17.37**	**4.16**	**1.551**	**0.772**

These results showcase the performance of my model relative to existing literature, emphasizing the competitive nature of the proposed approach in the context of sales prediction in e-commerce.

This study results exhibit a remarkable performance and overcame most of the previous papers in all metrics but the only one that was better than mine is the study made by Alibaba Group themselves and they used their private model called

Aliformer and they used bigger data they used nearly 500 million samples or records, in the meanwhile I just could use 50 million records so with regarding that my results were good. So in comparison to other papers except [3], as evidenced by lower MSE, RMSE, and higher R-squared values. These metrics collectively underline the model's efficacy in predicting 'Buy_Count' with improved accuracy and goodness of fit.

Reduced MSE:

- The lower MSE in your results signifies a reduction in the average squared differences between predicted and actual values. This improvement is indicative of the model's ability to minimize the impact of larger errors, contributing to enhanced predictive accuracy.

Improved RMSE:

- A reduced RMSE indicates that the model's predictions have a smaller average magnitude of error compared to the referenced papers. This improvement is particularly valuable for understanding the average size of errors in the 'Buy_Count' predictions, making the model more reliable in practical applications.

Higher R^2:+

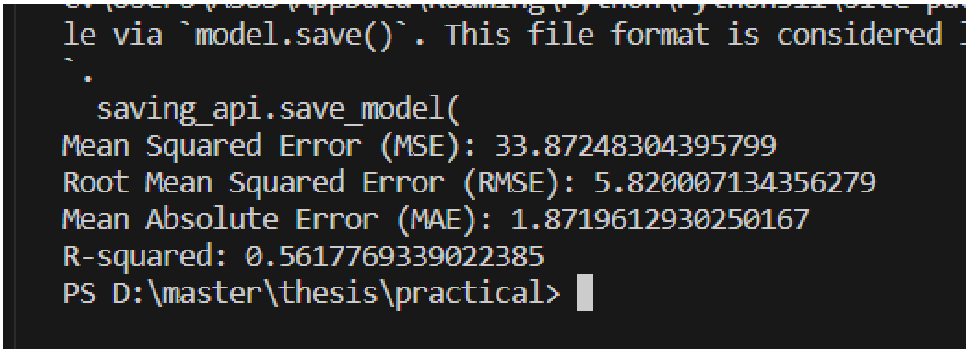
- The range for r-squared is between 0-1 so the higher R-squared value to the one in my results suggests that a larger proportion of the variance in the 'Buy_Count' is explained by the model. This is indicative of the model's robustness in capturing the underlying patterns in the data, contributing to a more accurate representation of the relationship between features and sales.

Precision Indicated by MAE:

- The achieved MAE stands out as a key metric highlighting the precision of your model. A lower MAE value, as seen in your paper, signifies that the average absolute difference between the predicted and actual 'Buy_Count' values is minimized. This precision is a significant contribution to the field, emphasizing the model's capability to make accurate predictions with reduced absolute errors.

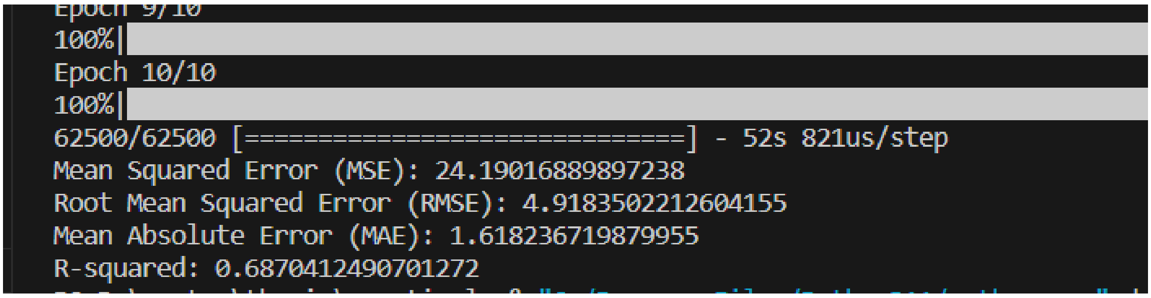
The superior or comparable performance across multiple metrics suggests advancements in sales prediction methodologies within the e-commerce domain. Your LSTM model demonstrates effectiveness in capturing temporal dependencies, showcasing its potential to outperform existing approaches. The precision demonstrated by your model has practical implications for the e-commerce industry. Accurate sales predictions are vital for inventory management, resource allocation, and strategic decision-making. The notable reduction in errors, as evidenced by the lower MAE, implies a potential improvement in operational efficiency for businesses utilizing your proposed model.

We made a comparison between RNN and LSTM models using just 10 million records to see how they both will perform using this dataset and the result as we will see in the next two figures, figure 4.2.1 will show the results of RNN and Figure 4.2 will show the results of LSTM using just 10 million records.



```
le via `model.save()`. This file format is considered
`.
  saving_api.save_model(
Mean Squared Error (MSE): 33.87248304395799
Root Mean Squared Error (RMSE): 5.820007134356279
Mean Absolute Error (MAE): 1.8719612930250167
R-squared: 0.5617769339022385
PS D:\master\thesis\practical>
```

Figure 4. 2. The result of RNN using 10 million records.



```
Epoch 9/10
100%|
Epoch 10/10
100%|
62500/62500 [==============================] - 52s 821us/step
Mean Squared Error (MSE): 24.19016889897238
Root Mean Squared Error (RMSE): 4.9183502212604155
Mean Absolute Error (MAE): 1.618236719879955
R-squared: 0.6870412490701272
```

Figure 4. 3. The result of LSTM using 10 million records.

So, as we see in the figures the LSTM achieved MSE 24.19, RMSE 4.91, MAE 1.61, and in R-squared 0.68 while RNN achieved MSE 33.8, RMSE 5.8, MAE 1.871, and in R-squared achieve 0.561, so LSTM achieve better results in all the matrices and this because LSTM comes to solve the problems that RNN have such as the vanishing problem, LSTM can store the sequential data for a very long time unlike RNN

PART 5

CONCLUSION

The dynamic landscape of e-commerce platforms, marked by rapid evolution and changing consumer behaviors, necessitates advanced predictive analytics to meet the demand for accurate sales forecasting models. This paper undertakes an in-depth exploration of predictive analytics within the e-commerce domain, employing machine learning methodologies with a specific focus on leveraging LSTM networks for sales prediction.

The utilization of LSTM networks is a pivotal aspect of this study, acknowledging the ability of these networks to capture sequential dependencies and temporal dynamics inherent in e-commerce data. However, it is crucial to acknowledge the pragmatic challenges posed by resource constraints. The study, constrained by the availability of 32 GB RAM, strategically opts for a subset of the comprehensive dataset sourced from Taobao, consisting of 50 million records. This pragmatic approach allows for meaningful insights while acknowledging the practical limitations of available resources. The comprehensive dataset, although constrained, forms the backbone for training LSTM-based models. The study meticulously addresses preprocessing techniques, ensuring the dataset is well-prepared to capture the intricacies of user interactions, browsing patterns, and purchase behavior.

The efficacy of the LSTM models is rigorously assessed through meticulous evaluations using standard metrics, including MSE, MAE, and RMSE. The reported results, with an MSE of 17.37, RMSE of 4.16, and MAE of 1.551, signify the model's ability to make accurate predictions while minimizing errors. The higher R-squared value of 0.772 indicates a commendable capability in explaining the variance in 'Buy_Count,' affirming the robustness of the model.

Beyond the realm of predictive accuracy, this study sheds light on the potential implications of precise sales forecasting within the e-commerce landscape. Accurate predictions have profound implications for optimizing inventory management, refining marketing strategies, and facilitating informed decision-making. These outcomes extend beyond the immediate scope of the study, emphasizing the real-world applicability of the developed LSTM models. Looking ahead, this research contributes significantly to the expanding knowledge base on leveraging LSTM networks for precise sales prediction in e-commerce. While the study operates within the constraints of a subset of the dataset due to resource limitations, the insights gleaned lay the foundation for future advancements in predictive analytics within the dynamic and ever-evolving domain of e-commerce.

It is imperative to acknowledge the limitations of this study, including the necessity to work with a subset of the dataset due to resource constraints. The chosen subset, comprising 50 million records, reflects a pragmatic compromise, allowing for meaningful insights while accommodating the available 32 GB RAM. While the achieved results are commendable within these constraints, future research could explore strategies for handling larger datasets and addressing challenges related to resource scalability to further enhance model robustness.

REFERENCES

1. Karimova, Farida. "A survey of e-commerce recommender systems." *European Scientific Journal* 12.34 (2016): 75-89.

2. Lakshmanan, Balakrishnan, Palaniappan Senthil Nayagam Vivek Raja, and Viswanathan Kalathiappan. "Sales demand forecasting using LSTM network." *Artificial Intelligence and Evolutionary Computations in Engineering Systems*. Springer Singapore, 2020.

3. Qi, Xinyuan, et al. "From known to unknown: Knowledge-guided transformer for time-series sales forecasting in Alibaba." *arXiv preprint arXiv:2109.08381* (2021).

4. Tugay, Resul, and Sule Gunduz Oguducu. "Demand prediction using machine learning methods and stacked generalization." *arXiv preprint arXiv:2009.09756* (2020).

5. Zhao, Kui, and Can Wang. "Sales forecast in e-commerce using convolutional neural network." *arXiv preprint arXiv:1708.07946* (2017).

6. Qi, Yan, et al. "A deep neural framework for sales forecasting in e-commerce." *Proceedings of the 28th ACM international conference on information and knowledge management*. 2019.

7. Laudon, Kenneth C., and Carol Guercio Traver. "E-commerce 2021–2022: Business." *Technology. Society, Seventeenth Edition, Global Edition* (2021). [8] C. COP. Android Ransomware Detection, doi: 10.34740/kaggle/dsv/4987535.

8. Attar, Razaz Waheeb, et al. "New Trends in E-Commerce Research: Linking Social Commerce and Sharing Commerce: A Systematic Literature Review." *Sustainability* 14.23 (2022): 16024.

9. Wang, Xuan, and Chi To Ng. "New retail versus traditional retail in e-commerce: channel establishment, price competition, and consumer recognition." *Annals of Operations Research* 291 (2020): 921-937

10. Dixon, Tim, and Andrew Marston. "The impact of e-commerce on retail real estate in the UK." *Journal of Real Estate Portfolio Management* 8.2 (2002): 153-174.

11. Singh, Amit Kumar, and Malsawmi Sailo. "Consumer behavior in online shopping: a study of Aizawl." *International Journal of Business & Management Research* 1.3 (2013): 45-49.

12. Bashir, Adil. "Consumer Behavior towards online shopping of electronics in Pakistan." (2013).

13. Kaur, Harmanjot, and Roopjot Kochar. "A review of factors affecting consumer behavior towards online shopping." *International Journal of Engineering and Management Research (IJEMR)* 8.4 (2018): 54-58.

14. Jain, Dipti, Sonia Goswami, and Shipra Bhutani. "Consumer behavior towards online shopping: an empirical study from Delhi." *IOSR Journal of Business and Management (IOSR-JBM)* 16.9 (2014): 65-72.

15. https://www.statista.com/statistics.

16. Cadavid, Juan Pablo Usuga, Samir Lamouri, and Bernard Grabot. "Trends in machine learning applied to demand & sales forecasting: A review." *International conference on information systems, logistics, and supply chain*. 2018..

17. Wai, Kok, et al. "Perceived risk factors affecting consumers' online shopping behavior." *The Journal of Asian Finance, Economics and Business* 6.4 (2019): 246-260.

18. Chen, I-Fei, and Chi-Jie Lu. "Sales forecasting by combining clustering and machine-learning techniques for computer retailing." *Neural Computing and Applications* 28 (2017): 2633-2647.

19. Krishna, Akshay, et al. "Sales-forecasting of retail stores using machine learning techniques." *2018 3rd international conference on computational systems and information technology for sustainable solutions (CSITSS)*. IEEE, 2018.

20. Shilong, Zhang. "Machine learning model for sales forecasting by using XGBoost." *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE, 2021.

21. Sharma, Suresh Kumar, and Vinod Sharma. "Comparative analysis of machine learning techniques in sale forecasting." *International Journal of Computer Applications* 53.6 (2012): 975-8887.

22. Gurnani, Mohit, et al. "Forecasting of sales by using fusion of machine learning techniques." *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*. IEEE, 2017

23. Pavlyshenko, Bohdan M. "Machine-learning models for sales time series forecasting." *Data* 4.1 (2019): 15

24. Schmidt, Austin, Md Wasi Ul Kabir, and Md Tamjidul Hoque. "Machine learning based restaurant sales forecasting." *Machine Learning and Knowledge Extraction* 4.1 (2022): 105-130

25. Helmini, Suleka, et al. "Sales forecasting using multivariate long short term memory network models." *PeerJ PrePrints* 7 (2019): e27712v1.

26. Cheriyan, Sunitha, et al. "Intelligent sales prediction using machine learning techniques." 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE, 2018.

27. Tsoumakas, Grigorios. "A survey of machine learning techniques for food sales prediction." Artificial Intelligence Review 52.1 (2019): 441-447.

28. Policarpo, Lucas Micol, et al. "Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review." Computer Science Review 41 (2021): 100414.

29. Ballestar, María Teresa, Pilar Grau-Carles, and Jorge Sainz. "Predicting customer quality in e-commerce social networks: a machine learning approach." Review of Managerial Science 13 (2019): 589-603.

30. Liu, Cheng-Ju, et al. "Machine learning-based e-commerce platform repurchase customer prediction model." Plos one 15.12 (2020): e0243105.

31. Enache, Maria-Cristina. "Machine Learning in E-commerce." Economics and Applied Informatics 1 (2019): 169-173.

32. Singh, Karandeep, P. M. Booma, and Umapathy Eaganathan. "E-commerce system for sale prediction using machine learning technique." Journal of Physics: Conference Series. Vol. 1712. No. 1. IOP Publishing, 2020.

33. Micu, Adrian, et al. "Leveraging e-Commerce performance through machine learning algorithms." Ann. Dunarea Jos Univ. Galati 2 (2019): 162-171.

34. Anitha, J., and M. Kalaiarasu. "Optimized machine learning based collaborative filtering (OMLCF) recommendation system in e-commerce." Journal of Ambient Intelligence and Humanized Computing 12 (2021): 6387-6398.

35. Shen, Boyu. "E-commerce customer segmentation via unsupervised machine learning." The 2nd international conference on computing and data science. 2021.

36. Addagarla, Ssvr Kumar, and Anthoniraj Amalanathan. "Probabilistic unsupervised machine learning approach for a similar image recommender system for E-commerce." Symmetry 12.11 (2020): 1783.

37. Bajaj, Purvika, et al. "Sales prediction using machine learning algorithms." International Research Journal of Engineering and Technology (IRJET) 7.6 (2020): 3619-3625.

38. Kohli, Shreya, Gracia Tabitha Godwin, and Siddhaling Urolagin. "Sales prediction using linear and KNN regression." Advances in Machine Learning

and Computational Intelligence: Proceedings of ICMLCI 2019. Singapore: Springer Singapore, 2020. 321-329.

39. Karunasingha, Dulakshi Santhusitha Kumari. "Root mean square error or mean absolute error? Use their ratio as well." Information Sciences 585 (2022): 609-629.

40. Nicolson, Aaron, and Kuldip K. Paliwal. "Deep learning for minimum mean-square error approaches to speech enhancement." Speech Communication 111 (2019): 44-55.

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

42. Qi, Jun, et al. "On mean absolute error for deep neural network based vector-to-vector regression." IEEE Signal Processing Letters 27 (2020): 1485-1489.

43. Hodson, Timothy O. "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not." Geoscientific Model Development 15.14 (2022): 5481-5487.

44. Redell, Nickalus. "Shapley decomposition of R-squared in machine learning models." arXiv preprint arXiv:1908.09718 (2019).

45. Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." PeerJ Computer Science 7 (2021): e623.

46. Bandara, Kasun, et al. "Sales demand forecast in e-commerce using a long short-term memory neural network methodology." Neural Information Processing: 26th Interna-tional Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26. Springer International Publishing, 2019.

47. Chengjie, Yang, and Qi Wei. "Taobao User purchase Behavior prediction and Feature analysis Based On Ensemble learning." 2023 IEEE International Conference on e-Business Engineering (ICEBE). IEEE, 2023

48. Rai, Sanket, et al. "Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods." 2019 10th international conference on computing, communication and networking technologies (ICCCNT). IEEE, 2019

49. Singh, Karandeep, P. M. Booma, and Umapathy Eaganathan. "E-commerce system for sale prediction using machine learning technique." Journal of Physics: Conference Se-ries. Vol. 1712. No. 1. IOP Publishing, 2020

50. Chen, Jianping, Nadine Tournois, and Qiming Fu. "Price and its forecasting of Chinese cross-border E-commerce." Journal of Business & Industrial Marketing 35.10 (2020): 1605-1618.

51. Rajesh, M. V., and S. Rao Chintalapudi. "A Review on Applications of Machine Learn-ing In E-Commerce." Advances and Applications in Mathematical Sciences 20.11 (2021): 2831-2841.

52. Salamai, A. Ali, A. Abdulrahman Ageeli, and El-Sayed M. El-kenawy. "Forecasting E-commerce adoption based on bidirectional recurrent neural networks." Computers, Materials & Continua 70.3 (2022): 5091-5106

53. Huo, Zixuan. "Sales prediction based on machine learning." 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT). IEEE, 2021

54. Dai, Yun, and Jinghao Huang. "A sales prediction method based on lstm with hyper-parameter search." Journal of Physics: Conference Series. Vol. 1756. No. 1. IOP Publishing, 2021

55. Wang, Weijie, and Yanmin Lu. "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model." IOP conference series: materials science and engineering. Vol. 324. IOP Publishing, 2018.

56. Goel, Shakti, and Rahul Bajpai. "Impact of uncertainty in the input variables and model parameters on predictions of a long short term memory (LSTM) based sales forecasting model." Machine Learning and Knowledge Extraction 2.3 (2020):

57. Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)." Geoscientific model development discussions 7.1 (2014): 1525-1534

58. Brassington, Gary. "Mean absolute error and root mean square error: which is the better metric for assessing model performance?" EGU General Assembly Conference Ab-stracts. 2017.

59. Vujović, Ž. "Classification model evaluation metrics." International Journal of Ad-vanced Computer Science and Applications 12.6 (2021): 599-606.

60. Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." PeerJ Computer Science 7 (2021): e623.

61. Onyutha, Charles. "From R-squared to coefficient of model accuracy for assessing" goodness-of-fits"." Geoscientific Model Development Discussions (2020): 1-25.

62. Kim, Sungil, and Heeyoung Kim. "A new metric of absolute percentage error for inter-mittent demand forecasts." International Journal of Forecasting 32.3 (2016): 669-679.

RESUME

Mohammed Aljbour began my academic journey at al-Aqsa school. I completed my secondary education at Khaled Elhassan High School in the 2015-2016 academic year. Subsequently, i pursued my undergraduate studies at Palestine technical college, graduating in 2021-2022. In 2022, I moved to Karabük, Turkey, to undertake postgraduate studies. I enrolled in a Master of Computer Engineering program at Karabük University.