FORECASTING FOR E-COMMERCE SALES USING SUPERVISED MACHINE LEARNING ALGORITHMS

AYÇELEN PAMUK

MEF UNIVERSITY

JANUARY 2024

MEF UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

MASTER'S IN INFORMATION TECHNOLOGIES

M.Sc. THESIS

FORECASTING FOR E-COMMERCE SALES USING SUPERVISED MACHINE LEARNING ALGORITHMS

Ayçelen PAMUK

ORCID No: 0009-0006-3273-5329

Thesis Advisor: Asst. Prof. Dr. Tuna ÇAKAR

JANUARY 2024

ACADEMIC HONESTY PLEDGE

This is to certify that I have read the graduation thesis and it has been judged to be successful, in scope and in quality and is acceptable as a graduation project Master's Degree in Information Technologies.

Name Surname: Ayçelen PAMUK

Signature:

ABSTRACT

FORECASTING FOR E-COMMERCE SALES

USING SUPERVISED MACHINE LEARNING ALGORITHMS

Ayçelen PAMUK

M.Sc. in Information Technologies

Thesis Advisor: Asst. Prof. Dr. Tuna ÇAKAR

January 2024, 69 Pages

The burgeoning landscape of e-commerce relies significantly on predictive analytics to drive operational efficiency and strategic decision-making. This thesis delves into the theoretical underpinnings of machine learning algorithms, showcasing their evolution and pivotal role in facilitating the growth of online commerce. At its core, this research centers on forecasting sales patterns through the analysis of an extensive e-commerce dataset.

Forecasting stands as a linchpin for various critical functions within e-commerce enterprises. Its multifaceted applications encompass inventory management, ensuring optimal stock levels and streamlined deliveries, financial planning through astute asset management, dynamic pricing strategies, and the enhancement of customer satisfaction via efficient delivery operations. Furthermore, forecasting plays a pivotal role in refining marketing endeavors, enabling tailored campaigns and judicious budget allocation. The integration of machine learning algorithms fortifies these functionalities.

Central to this research is the foundational task of sales prediction in the e-commerce realm, with a specific emphasis on integrating campaign variables. Leveraging six diverse machine learning algorithms, the study aims to discern the most accurate and explicable model. Remarkably, the investigation identifies LGBM as the

most suitable algorithm. Notably, the inclusion of campaign variables, an aspect seldom explored in prior studies concerning forecasting, yields intriguing insights. However, contrary to initial presumptions, the SHAP analysis reveals a lesser influence of campaign variables on the model's interpretability. Acknowledging this limitation, the study highlights the potential for augmenting model interpretability by employing clustering algorithms to effectively represent variables, as outlined in the limitations section.

Keywords: Sales Forecasting, Machine Learning, E-Commerce

Numeric Code of the Field: 92404

ÖZET

GÖZETİMLİ MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK E-TİCARET SATIŞLARININ TAHMİNLENMESİ

Ayçelen PAMUK

Bilişim Teknolojileri Tezli Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Tuna ÇAKAR

Ocak 2024, 69 Sayfa

E-ticaretin gelişimi için temel olan öngörüsel analitiklere büyük ölçüde bağımlı olan büyüyen e-ticaret manzarası, operasyonel verimliliği ve stratejik karar alma süreçlerini yönlendirmektedir. Bu tez, makine öğrenimi algoritmalarının teorik temellerine derinlemesine inerken, çevrimiçi ticaretin büyümesini kolaylaştırmadaki evrimini ve kilit rolünü sergilemektedir. Bu araştırma, geniş çaplı bir e-ticaret veri kümesinin analizi yoluyla satış desenlerini öngörme odaklıdır.

Öngörme, e-ticaret işletmelerinde çeşitli kritik işlevler için bir anahtar görev üstlenmektedir. Envanter yönetimini, optimal stok seviyelerini ve düzenli teslimatları sağlayarak, stratejik varlık yönetimi aracılığıyla finansal planlamayı, dinamik fiyatlandırma stratejilerini ve etkili teslimat operasyonlarıyla müşteri memnuniyetini artırmayı içeren çok yönlü uygulamaları kapsar. Ayrıca, öngörme, özelleştirilmiş kampanyaları ve akıllı bütçe tahsisini mümkün kılarak pazarlama çabalarını geliştirmede de temel bir rol oynamaktadır. Makine öğrenimi algoritmalarının entegrasyonu, bu işlevleri güçlendirmektedir.

Bu araştırmanın merkezinde e-ticaret alanındaki temel bir satış tahmini görevi bulunmaktadır ve özellikle kampanya değişkenlerini entegre etme odaklıdır. Altı farklı makine öğrenimi algoritmasını kullanarak, çalışma en doğru ve açıklayıcı modeli belirlemeyi amaçlamaktadır. Dikkat çekici bir şekilde, araştırma LGBM'yi en uygun algoritma olarak belirler. Önceki tahmin çalışmalarında nadiren keşfedilen kampanya değişkenlerinin dahil edilmesi, ilginç içgörüler sunar. Ancak, başlangıçtaki varsayımların

aksine, SHAP analizi, kampanya değişkenlerinin modelin açıklanabilirliği üzerinde daha az etkiye sahip olduğunu ortaya koymaktadır. Bu sınırlama farkındalığını kabul eden çalışma, değişkenleri etkili bir şekilde temsil etmek için kümeleme algoritmalarını kullanarak modelin açıklanabilirliğini artırma potansiyeline dikkat çekmektedir.

Anahtar Kelimeler: Satış Tahmini, Makine Öğrenmesi, E-Ticaret

Bilim Dalı Sayısal Kodu: 92404

Tüm süreç boyunca bana destek olan

aileme, Gati'ye ve arkadaşlarıma ithaf ediyorum.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ARIMA : AutoRegressive Integrated Moving Average

CAGR : Compound Annual Growth Rate

CBDT : Cause Based Decision Tree

CNN : Convolutional Neural Network

DNN : Deep Neural Network

EFB : Exclusive Feature Bundling

GBM : Gradient Boosting Machine

GOOS : Gradient-based One-Side Sampling

LS : Least Squares

LightGBM : Light Gradient Boosting Machine

LSTM : Long Short-Term Memory Networks

MAE : Mean Absolute Error

MAPE : The Mean Absolute Percentage Error

ML : Machine Learning

MSE : Mean Squared Error

NARNN : Nonlinear Autoregressive Neural Network

R^2 : R-squared

RMSE : Root Mean Squared Error

SARIMA : Seasonal Autoregressive Integrated Moving Average

SVM : Support Vector Machine

TL : Transfer Learning

WD : Weighted Decay

XGBoost : eXtreme Gradient Boosting Regression

INTRODUCTION

The uncertainty regarding future outcomes poses a persistent challenge for businesses. Accurate prognostication of forthcoming metrics is crucial for enterprises to effectively manage their operations, preempt potential issues, and optimize profitability. The evolving landscape of technological advancements affords companies the ability to anticipate future trends and outcomes, enabling them to make informed decisions and strategic adjustments to navigate uncertainties effectively.

Forecasting stands as a cornerstone solution for numerous enterprises, serving as a mechanism to project and anticipate forthcoming outcomes through the analysis of proprietary or outsourced data. This practice involves leveraging historical data, statistical models, and analytical tools to extrapolate future trends, enabling companies to make informed decisions, optimize resource allocation, and proactively strategize for future scenarios.

In the field of predictive analytics, data acquisition is a basic building block. Conventional approaches frequently face difficulties in collecting precise and accurate data in addition to the difficulties in guaranteeing effective storage. Nonetheless, the global environment and the paradigm shift toward digitization have driven businesses toward online platforms. This shift from physical storefronts to virtual ones is best illustrated by the e-commerce industry, which is becoming a significant domain. This represents a significant change in the way businesses operate. By the end of 2025, point estimates of $378.691 billion for US e-commerce retail sales are expected to represent 16.72 percent of all US retail sales, indicating a continuation of the upward trend [1]. Accurate prediction holds paramount importance within the realm of e-commerce due to its burgeoning market and immense growth potential.

The ideas of the digital economy and the Internet economy are connected to the concept of e-commerce. Though they have different focal points, all of these ideas are related to the use of new information and communication technologies for commercial endeavors [2]. Among the process of digitalization, enterprises are afforded the opportunity to meticulously monitor various facets of their operations, including customer

behaviors, inventory management, and the intricacies of the delivery process. This digital transformation allows companies to adeptly store extensive and detailed datasets encompassing multifaceted operational aspects.

Possessing comprehensive and precise datasets tends to significantly enhance the predictive capability and accuracy of forecasting models. Effective forecasting models play a pivotal role in aiding companies to optimize demand projections, comprehend prevailing market trends, facilitate financial planning, determine accurate pricing strategies, enhance customer satisfaction, and orchestrate strategic marketing initiatives.

RESEARCH FRAMEWORK & MOTIVATION

The primary objective of this research endeavor is to discern the most suitable machine learning algorithm for the prediction of next-day order counts based on historical data from the preceding day alongside the campaign variable. In pursuit of this objective, a comprehensive array of supervised learning techniques and data preprocessing methods have been employed. Moreover, an attempt has been made to optimize variable selection through the integration of industry expertise from professionals within the sector. Furthermore, the campaign variables were clustered into categories based on their sizes, delineating between small, medium, and large campaigns. The dataset used in this study, procured from an e-commerce company, comprises 1034 rows and 16 columns. The primary objective is to predict the next day's order count. Various methodologies such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), Ridge Regression, Polynomial Regression, XGBoost Regression, and LightGBM Regression were applied for forecasting purposes. Subsequently, a comparative analysis of these methods was conducted to ascertain the most effective approach for forecasting in this dataset.

1. LITERATURE REVIEW

1.1. Relevant studies

1.1.1. E-commerce

In response to the unprecedented global COVID-19 pandemic and the consequent imposition of lockdown measures to mitigate its spread, it has become evident through empirical data from Statista, a renowned German entity specializing in comprehensive market and consumer data analysis, that there has been a discernible and progressive shift in consumer behavior worldwide. Zwanka and Buff have observed that the implementation of lockdown measures has brought about notable alterations in many established routines and lifestyles, which encompass a transition to online shopping, remote working, the increased utilization of food delivery services from restaurants, virtual tourism experiences, and a notable shift toward online learning, among other noteworthy adaptations [3]. A noteworthy phenomenon emerging from this crisis has been the escalating proclivity of consumers to engage in online retail activities, encompassing a diverse array of product categories that encompass essential items such as food and personal care products, as well as more discretionary purchases, notably electronic goods and even luxury items. This shift towards e-commerce and digital transactions has not only been a pragmatic response to the exigencies of lockdowns and social distancing but also signifies a broader and lasting transformation in the contemporary consumer landscape, warranting further academic inquiry and analysis. Due to the COVID-19 pandemic and ensuing lockdowns, internet sales have experienced a significant increase in both volume and importance within the retail industry. This shift has not only led to higher sales figures but has also brought about notable changes in consumer preferences, supply chain management, marketing strategies, and technological advancements in e-commerce. Consequently, the surge in online sales is a crucial topic for academic analysis, as it represents a fundamental transformation in the contemporary retail sector.

In contrast to the $3.17 trillion in traditional retail sales recorded in the United States during 2019, it is noteworthy that e-commerce sales within the same period reached

a substantial figure of $603 billion [4]. Statista also presented an intriguing statistic: it was estimated that in 2021, the total global e-commerce retail sales amounted to approximately 4.9 trillion US dollars, constituting nearly 20% of the entire global retail e-commerce sales volume [5]. This marked discrepancy underscores the increasing significance of online commerce within the U.S. retail landscape, prompting an academic inquiry into the underlying factors and implications of this notable trend. Certainly, online transactions are anticipated to constitute 20.8 percent of total sales in 2023, reflecting a two-percentage-point expansion in the e-commerce market share compared to the 17.8 percent recorded two years earlier in 2021 [6]. Furthermore, this upward trajectory is forecasted to persist, with online sales projected to comprise 23 percent of total sales by the year 2025. According to the E-commerce Market Report and Forecast for both the global and United States markets, it is anticipated that the worldwide e-commerce market, which had a value of USD 7.75 trillion in 2022, will experience substantial growth, reaching an estimated USD 20.31 trillion by 2028, exhibiting a Compound Annual Growth Rate (CAGR) of 17 percent during this period [7].

As per a study published in January 2022 by Eurostat, the statistical agency of the European Union, it was observed that in the year 2021, a substantial majority of Internet users within the European Union engaged in online shopping, with 74% of these users participating in such activities. Among these online shoppers, a noteworthy 42% reported expenditure ranging from 100 to 500 Euros within the three months preceding the survey [8].

Moreover, across the various European Union member states, there was considerable variation in the percentages of Internet users who had undertaken online purchases during this same period. These percentages ranged from 42% in Bulgaria to a significantly higher 94% in the Netherlands, underscoring the divergent adoption rates of e-commerce practices across EU nations [9].

The global e-commerce landscape has exhibited sustained growth since the onset of the twenty-first century. This expansion is particularly noteworthy when considering

the statistics pertaining to e-commerce in Turkey, especially in the context of the unique challenges posed by the coronavirus (COVID-19) pandemic. The digital market for Turkish goods and services has emerged as one of the global markets characterized by rapid growth. In the year 2021, a substantial 64% of Turkish consumers actively engaged in online purchases. Consequently, there has been a substantial upsurge in the volume of e-commerce within the country, with a remarkable increase observed in the final quarter of 2021 when compared to the corresponding period in 2017 [10]. In Turkey, there was a remarkable 66% increase in e-commerce volume within the span of just one year, surging from 136 billion Turkish Liras (TL) in 2019 to 226.2 billion TL in 2020. Notably, domestic expenditure constituted the majority, comprising 91.4% of the total expenditures, whereas international purchases of Turkish e-commerce products accounted for a more modest 4.2%, equivalent to 9.3 billion TL [11].

In summary, e-commerce occupies a pivotal role as a cornerstone of global economies, signifying its status as an emerging structural pillar within the market landscape, and its enduring significance over the forthcoming years is underpinned by a multitude of compelling factors. To begin with, e-commerce enhances market accessibility and fosters economic inclusivity by providing a platform for businesses of various scales to reach a global audience, thereby contributing to economic growth. Furthermore, it facilitates data-driven decision-making, enabling businesses to refine their strategies and tailor offerings to meet consumer demands effectively, thus driving market efficiency and competitiveness. In essence, e-commerce's multifaceted contributions and adaptability position it as an indispensable driver of economic development in the coming years.

1.1.1.1. E-commerce advantages

E-commerce has experienced a recent global expansion due to its indisputable advantages and the growing demands of consumers. Numerous factors underpin the surge in e-commerce growth. One key reason is the lower overhead costs in e-commerce compared to physical stores. Physical stores require expenses for storage, physical space,

sales staff, and utilities like water and electricity [12]. In the realm of online retail, these costs are notably lower or even nonexistent. E-commerce businesses can leverage drop shipping or use their own homes or garage for storage, thus reducing overhead. Additionally, they typically require fewer employees than physical stores. The second reason is the continuous availability of e-commerce compared to the limited hours of operation for physical stores. Online stores are open 24/7, allowing customers to shop at their convenience. In contrast, physical stores have set hours and rely on in-person customers, which restricts their accessibility. The third factor is location flexibility. Physical stores are limited by their geographical reach, affected by travel and transportation constraints. E-commerce, on the other hand, can reach customers anywhere, as long as they have an internet connection and a device. Another factor and one of the most important ones, with digitalization and technology and e-commerce the business itself you have more advantage because of data-driven decisions. Last but not least, a significant determinant, and arguably one of the most pivotal, is the transformative potential of digitization, technology, and e-commerce, which bestows a distinct competitive edge to businesses. This advantage is particularly pronounced in the context of data-driven decision-making, where the accumulation and analysis of vast amounts of data empower businesses to make informed, strategic choices, enhancing their operational efficiency and market competitiveness.

1.1.1.2. Inventory Management

Compared to traditional stores, e-commerce businesses achieve significant cost savings in both inventory and operations. This is primarily due to the automated management of inventory through advanced online technology solutions [13]. E-commerce companies have the capacity to amass a wealth of information pertaining to consumer purchasing patterns, inventory dynamics, sales trends, and related metrics, owing to the opportunities presented by advanced technological tools. This reservoir of data empowers these companies to engage in precise sales forecasting, thereby facilitating effective inventory management. By virtue of this predictive capability, e-commerce businesses can proactively stock merchandise according to demand, mitigating the risk of

stockouts and ensuing client dissatisfaction. Moreover, astute inventory management not only fosters enhanced client retention and satisfaction rates but also yields cost savings, as it obviates the need for excess inventory expenditures. In contrast to offline shopping, which operates within finite parameters such as a countable number of cashiers and restricted product availability, e-commerce exhibits a distinct advantage in its capacity for order processing. E-commerce platforms have the ability to seamlessly manage and fulfill a significantly higher volume of orders. In situations where stock depletion may occur, e-commerce entities possess the agility to replenish their inventories swiftly or, alternatively, enable customers to place backorders for products they require. This heightened operational flexibility and adaptability in e-commerce exemplify its ability to efficiently meet customer demands, irrespective of inventory constraints.

1.1.1.3. Financial Planning

For e-commerce enterprises, the process of forecasting sales assumes paramount significance in the realm of financial planning. Precise projections of sales revenue play a multifaceted role, conferring numerous benefits upon sales managers. These include improved resource allocation, enhanced income estimation, judicious investment planning, adept problem identification and resolution, comprehensive business planning, and the formulation of informed decisions pertaining to business expansion and development [14]. Organizations can meticulously structure their expenditures and investment strategies predicated on their projected sales revenue. This diligent financial planning fosters superior cash flow management, heightened profitability, and overall enhanced financial stewardship.

1.1.1.4. Pricing

A commensurate investigation conducted by Su and Chen underscores the paramount importance of sales forecasting in the context of optimizing pricing strategies within the domain of e-commerce [15]. The authors posit that discerning the intricacies of price elasticities and demand patterns is of pivotal significance, as it empowers businesses

to make informed pricing determinations, ultimately contributing to the enhancement of profitability. Furthermore, the precision of sales predictions assumes a requisite role in the identification of price elasticity, a critical factor in pricing strategy refinement.

1.1.1.5. Customer Satisfaction

The precision of sales forecasting is instrumental in aiding organizations in the efficient management of inventory and resource allocation, thereby yielding enhanced customer satisfaction. In alignment with a study conducted by Hua, the accuracy of sales forecasts facilitates the assurance that products remain readily available to customers when needed, thereby diminishing the likelihood of stockouts and delivery delays [16]. In addition to the forementioned advantages, a comparative analysis of e-commerce in contrast to its offline counterpart underscores the distinctive feature of customer-provided product reviews in the e-commerce domain. These reviews assume a pivotal role in matters of brand credibility and customer contentment. They enable potential buyers to peruse the comments and experiences of prior customers, thereby enriching their understanding and appraisal of the product. Moreover, these reviews significantly influence the purchasing decisions of consumers, exerting a tangible impact on the choices they make. Simultaneously, product reviews serve as invaluable feedback for manufacturers and suppliers, furnishing them with insights to refine and enhance their offerings in response to consumer input. An additional strategy aimed at heightening customer satisfaction pertains to the implementation of a robust search engine within the e-commerce platform. The search engine functionality not only facilitates customers in locating specific products of their interest but also provides supplementary assistance by suggesting analogous or related items. Consequently, this feature augments the scope of available products, simultaneously enabling the platform to accumulate valuable user data. This data, in turn, contributes to the refinement and personalization of the e-commerce user experience, culminating in an elevated level of customer satisfaction. The ability for users to readily access and fulfill their specific needs is thereby enhanced, accentuating their overall contentment. Beyond the presentation of search results, an additional avenue to enhance customer satisfaction lies in leveraging their historical

purchasing patterns. By showcasing previous acquisitions and offering products aligned with their preferences or prior choices, e-commerce platforms can provide a personalized shopping experience. This tailored approach capitalizes on the familiarity and affinity customers have developed, thereby augmenting their satisfaction and cultivating a sense of individualized engagement.

1.1.1.6. Marketing

The capacity to strategically plan marketing and promotional initiatives represents another notable benefit derived from the proactive anticipation of e-commerce sales. By proactively forecasting demand, businesses can optimize the timing of their marketing campaigns to align with periods of heightened sales activity. This synchronization enhances the likelihood of customers engaging with and making purchases of items featured in the promotional materials, thereby contributing to heightened sales volumes. Moreover, businesses can judiciously orchestrate promotional activities and discounts, informed by sales forecasts, to stimulate sales during less active periods. Additionally, online retailers can leverage customer data to refine their marketing strategies, ensuring that promotional efforts are precisely tailored to the relevant demographic for product sales. Leveraging tools such as retargeting, which enable the collection of data regarding customers' prior visits, facilitates a strategic marketing approach. This technology allows businesses to retarget prospective customers as they navigate the internet across various web pages. These retargeted advertisements are instrumental in capturing the attention of potential buyers and subsequently redirecting them to the e-commerce platform where products are available for purchase. This practice harnesses the power of persistent engagement and influences customer behavior, ultimately guiding them towards the e-commerce website for product acquisition. An additional tool known as remarketing, which capitalizes on the knowledge of previous purchase behaviors, facilitates the creation of tailored campaigns aimed at retaining and engaging existing customers. This technology empowers businesses to effectively remind and re-engage customers through customized marketing initiatives, taking into account their prior purchasing history. By systematically rekindling customer interest and loyalty through these personalized campaigns, businesses can fortify their brand-consumer relationships and stimulate

recurrent sales, thus advancing the objectives of e-commerce enterprises. Timely and pertinent email communications are employed to keep customers informed about campaigns, discounts, product upgrades, offers, and the latest developments, fostering customer engagement and sustaining their interest in the e-commerce platform [17].

1.1.1.7. Forecasting

In accordance with the scholarly literature, forecasting can be aptly defined as the process of predicting, projecting, or estimating future events or conditions that lie beyond an organization's immediate sphere of control, thereby furnishing a foundational framework for managerial planning [18]. Forecasting, as an instrumental methodology, plays a pivotal role in enabling businesses and organizations to refine their data-driven decision-making processes. Notably, the foremost advantage inherent to e-commerce lies in its capacity to comprehensively gather data across diverse facets of its operations, thereby facilitating the enhancement of forecasting endeavors. With a robust repository of meticulously collected and detailed data, e-commerce entities are primed to refine and optimize their forecasting models, ultimately bolstering their decision-making capabilities.

Leveraging technology, organizations have the capacity to systematically amass data through the utilization of information systems. In the context of e-commerce, these systems meticulously catalog an array of critical information, including inventory status, product sales records, visitor traffic across web pages and applications, favored product categories, shopping cart contents, campaign utilization metrics, and a multitude of other variables. This wealth of stored data serves as a foundational resource for diverse forecasting endeavors, encompassing sales projections, demand estimations, risk assessments, growth prospects, revenue expectations, expense projections, and prevailing market trends, among others.

Of particular salience, the capacity to accurately forecast sales and demand emerges as a linchpin in operational efficiency. These forecasts not only facilitate the anticipation and calibration of inventory levels but also inform risk mitigation and growth strategies. Consequently, the act of forecasting sales and demand constitutes an

indispensable facet of organizational analytics, crucial for informed decision-making, and a vital component of the strategic insights delivered to executive teams and founders.

1.2.1. Machine Learning Based Forecasting

Arthur Samuel, a distinguished computer scientist and a trailblazer in the realm of artificial intelligence research, introduced the concept of 'Machine Learning' in 1959 [19]. He defined it as the capacity of a computer to acquire knowledge and make informed decisions without the need for explicit, rule-based programming. Machine learning methodologies can be classified into distinct categories predicated on their learning paradigms, notably supervised and unsupervised learning approaches.

1.2.1.1. Supervised Learning

Supervised learning, a fundamental category within machine learning, relies on the utilization of labeled instances and is designed to acquire knowledge through the assimilation of illustrative examples. In this learning paradigm, both the input and output variables are provided as part of the training data. The principal aim of supervised learning algorithms is to deduce the accurate label or output for unlabeled data points. Mathematically, a supervised learning algorithm can be succinctly represented as:

$$Y = f(x) \tag{1.1}$$

Herein:

$$x: input\ value$$
$$Y: predicted\ output$$

Supervised learning further subdivides into two primary domains: Classification and Regression. Classification techniques are employed for scenarios involving categorical output variables, whereas regression methods are geared toward tasks where continuous output variables are involved.

1.2.1.2. Classification

Classification methodologies serve to categorize or assign data to specific classes or categories based on prior knowledge. In this process, a model assimilates information from input data by assigning varying weights to the input features. Subsequently, this acquired knowledge is employed to classify new, previously unseen instances. Datasets used in classification tasks may exhibit a binary, two-class structure (e.g., determining whether a product has been sold or not) or may encompass multiple classes, thereby necessitating more complex multi-class classification approaches.

Kotsiantis conducted an empirical inquiry into classification problems characterized by discrete and unordered output values for individual instances. The research underscored the prominence of supervised classification as one of the most commonly employed tasks within this domain. Consequently, a diverse array of methodologies has been devised, including Logic and Perceptron-based techniques within the domain of Artificial Intelligence, while Bayesian Networks and Instance-based techniques have been developed in the realm of Statistics [20].

1.2.1.3. Regression

In the context of a regression task, the fundamental objective of the model is to effectuate an estimation and comprehension of the essential relationships that underlie the dependent and independent variables. This process, characterized by its predictive nature, harnesses the mechanisms of a continuous function to systematically evaluate how outputs evolve in response to specified inputs. It is noteworthy that in a majority of cases, discernible associations exist between the input and output variables. Within the domain of regression analysis, an array of methodologies is employed, with the most prevalent ones encompassing:

- Linear Regression
- Non-Linear Regression
- Logistic Regression
- Polynomial Regression

- Ridge Regression

Multiple regression studies have been conducted with the specific aim of unraveling the intricate web of relationships within the domain of sales features, thereby contributing to an advanced and comprehensive understanding of this economic domain.

1.2. Literature Review and Examples

Forecasting stands as an indispensable instrument for enterprises and institutions in the quest to anticipate forthcoming trends and outcomes. It encompasses the meticulous scrutiny of both historical and contemporary data, with the objective of predicting prospective events, encompassing aspects such as sales, demand patterns, and revenue projections. Within the ambit of this literature review, we will delve into the historical evolution and the paramount significance of the practice of forecasting. Emphasis will be placed on elucidating the manifold elements that contribute to its efficacy and utility in decision-making processes [21].

As previously elucidated, forecasting assumes a pivotal role in the strategic decision-making process of corporations and organizations. It serves as a linchpin for the judicious allocation of resources, the prescient identification of prospective risks and opportunities, and the meticulous anticipation of forthcoming demand patterns. With the aid of precise forecasting, businesses can proactively prepare for the future, enabling them to make informed determinations regarding investments, inventory management, and staffing. Furthermore, this strategic foresight equips organizations with the dexterity to adeptly respond to evolving market dynamics, thereby fortifying their competitive standing within their respective industries [22] [23].

A multitude of empirical studies have been undertaken in the realm of sales forecasting. These studies can be broadly categorized based on the utilization of sales forecasting techniques, which predominantly fall within two main domains: time series models and machine learning algorithms [24]. Time series models, which are extensively employed in the forecasting domain, facilitate the projection of future sales trends by

leveraging historical data observations. This category encompasses a diverse range of methodologies, including exponential smoothing and the ARIMA (AutoRegressive Integrated Moving Average) model family.

Despite their prominence, time series models exhibit certain limitations. These constraints are chiefly attributed to their underlying assumption of linear trends and their omission of external factors, such as pricing variations and promotional activities, which can exert a considerable influence on sales dynamics. Consequently, in numerous studies, univariate forecasting techniques are frequently adopted as benchmark models. These univariate techniques serve as a standard of reference for evaluating the performance of more intricate forecasting approaches, recognizing the necessity to account for external variables that extend beyond historical sales patterns [24].

In accordance with the research conducted by Zhao and Wang, the principal method employed in this study for sales forecasting was the Convolutional Neural Network (CNN) algorithm. Nevertheless, the research encompassed a comparative analysis involving alternative methodologies, namely the ARIMA, Deep Neural Network (DNN), Transfer Learning (TL), and Weighted Decay (WD) algorithms, with the aim of discerning the most accurate approach for sales prediction [25].

To enhance the predictive accuracy, the researcher incorporated two additional techniques: sample weight decay and transfer learning. Both of these techniques have demonstrated notable efficacy in experimental settings. Notably, the ARIMA model exhibited the highest average value, as indicated by the Mean Squared Error (MSE) boxplot. However, it is noteworthy that the CNN algorithm achieved success in its objective of autonomously extracting salient features and employing them for sales forecasting. This underscores the inherent capability of the CNN algorithm to discern and utilize pertinent characteristics, substantiating its effectiveness in the domain of sales prediction.

Drawing from their respective scholarly investigations, both Bandara, Li, Ji, and Liu, have elected to employ neural network techniques as an integral component of their

research endeavors [26][27]. The overarching objective of their individual research undertakings has been to develop comprehensive forecasting frameworks and systematic pre-processing methodologies tailored to address the intricate challenges posed by the e-commerce landscape, particularly with regard to the prediction of sales dynamics and business demand.

In these scholarly articles, the researchers have demonstrated a preference for the utilization of advanced algorithms, namely the Nonlinear Autoregressive Neural Network (NARNN) and the Long Short-Term Memory Networks (LSTM). In order to benchmark the efficacy of these neural network approaches, the ARIMA algorithm, a well-established time series analysis method, was also incorporated.

The outcomes of their comparative analyses have unveiled that both the NARNN and LSTM algorithms have exhibited superior performance when juxtaposed with the ARIMA model. Specifically, the NARNN algorithm manifested a noteworthy improvement of 21.12% in error rate, while the LSTM algorithm registered lower mean and median errors, further underscoring the potency of these neural network methodologies in enhancing sales prediction and demand forecasting within e-commerce contexts.

The primary challenge elucidated in the study by Pavlyshenko, revolves around the inherent complexities stemming from the extensive volume of transactional sales data, a substantial portion of which may exhibit missing values and outliers [28]. The unique characteristics of time series data, necessitating a considerable dataset to accurately encapsulate seasonality, contribute to this data quality issue. Moreover, the data necessitates a comprehensive adjustment to account for a diverse array of variables that can exert an influence on sales dynamics.

The fundamental objective of the time series analysis conducted in this study is to amalgamate a diverse array of time series algorithms, thereby augmenting the predictive capacity of the model. To this end, five distinct algorithms have been selected, encompassing ExtraTree, ARIMA, RandomForest, Lasso, and Neural Network, spanning

both time series and supervised learning paradigms.

The empirical results of the forecasting error analysis have illuminated that the ExtraTree algorithm exhibits the highest validation error in comparison to the other algorithms. Conversely, the Neural Network algorithm has registered the lowest validation error, thereby establishing itself as one of the preeminent algorithms for the purposes of sales prediction, as discerned from the findings of this comprehensive study.

In the comparative evaluation of three distinct algorithms, the researcher conducted an analysis employing two key metrics, namely the Mean Absolute Error (MAE) and the R-squared (R2) Score. The pivotal facet underpinning the achievement of this study lay in the employment of diverse categorization methods. The algorithms scrutinized in this research encompassed Extremely Randomized Tree (Extra Tree), Gradient Boosting, and Random Forest.

The empirical findings derived from this investigation have illuminated the Random Forest algorithm as the most promising among the alternatives. This determination is grounded in its attainment of the lowest MAE evaluation score, coupled with a notably high R2 score. These results collectively signify the superior accuracy of the Random Forest algorithm in relation to the other algorithms considered, thus establishing its preeminence in the context of the study [29].

A distinct scholarly investigation, conducted by Yu and colleagues, was dedicated to the examination of Amazon's sales estimation through the application of diverse statistical methodologies [30]. The principal objective of this research endeavor was to execute a comprehensive sensitivity analysis using three distinct approaches and subsequently discern the most reliable, accurate, and fitting strategy. The crux of this endeavor lay in the recognition that the precision of the chosen methodology is directly proportional to the efficacy of sales forecasting. In pursuit of this objective, the study harnessed three discrete methodologies, specifically: time-series decomposition, ARIMA, and Winters' exponential smoothing.

The outcomes of this research were adjudicated through the calculation of the Root Mean Square Error (RMSE), a widely recognized metric for assessing forecasting accuracy. Remarkably, the investigation revealed that each of the employed methodologies exhibited remarkably low forecasting error rates, implying their utility for effective sales forecasting within the context of the study.

In a research endeavor carried out by Bohanec, and his colleagues, the central focus is the exploration of machine learning algorithms for the purpose of sales prediction [31]. The primary thrust of this study revolves around an in-depth investigation of the predominant machine learning models that are conventionally employed in the realm of sales forecasting, with a specific emphasis on the identification of the foremost model currently in practical use.

The salient issue underscored in this research pertains to the judicious selection of an appropriate model, predicated on a nuanced understanding of the business domain, achieved through the strategic fusion of business intelligence and data-driven models. In the pursuit of this objective, a spectrum of machine learning methods or algorithms was harnessed, encompassing decision trees, neural networks, Naive Bayes, Random Forest, and Support Vector Machine (SVM). The evaluation criteria employed in this study were anchored in accuracy metrics, with the Random Forest method emerging as the preeminent choice due to its capacity to yield a high-accuracy model [32].

Drawing upon insights gleaned from prior scholarly literature, the present research endeavors to orchestrate a systematic comparison of machine learning models. The primary aim is to discern the optimal model by gauging its efficacy and accuracy, thereby advancing our understanding of how a multitude of variables exert influence on the predictive capacity of these models. This pursuit is grounded in the principles of empirical rigor and analytical precision, underpinning the imperative to identify the most fitting model within the framework of our investigation.

2. MATERIALS & METHODS

2.1. Model Development & ML Methods

2.1.1. Interaction Terms

Interaction takes place in statistical analysis when the impact of one independent variable on a dependent variable is influenced by different levels of a moderating variable [33]. In simpler terms, it means that the relationship between variables is not consistent across various levels of another variable, introducing complexity to the analysis.

The fundamental linear regression model is expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon \quad (2.1)$$

- y is the outcome of the dependent value.
- β_0 is the intercept term.

$\beta_1, \beta_2, \ldots, \beta_n$ are the parameters to be estimated corresponding to the independentvariables of $x_1, x_2, \ldots, x_n$ respectively.

- ε represents the random error term.

The effect of a one-unit change on the dependent variable is the marginal effect of the explanatory variable on the dependent variable [34]. Alterations in the marginal effect of one variable due to variations in the value of another variable are expressed through cross-partial derivatives or discrepancies, commonly known as interaction effects [35].

Upon the introduction of an interaction term into the model, its formulation would be articulated as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) + \ldots \beta_n x_n + \varepsilon \quad (2.2)$$

The incorporation of interaction terms in a model is crucial due to the underlying assumption in a model without interactions, where it is posited that the impact of each predictor on the outcome is independent of the presence or influence of other predictors within the model. The inclusion of interaction terms allows for a more nuanced and accurate representation of potential dependencies and conditional relationships among

predictors, providing a more comprehensive understanding of the complex dynamics at play in the model.

2.1.2. One-Hot Encoding

One-hot encoding is a crucial technique employed in Machine Learning (ML) Algorithms to transform categorical data into a numeric format. ML Algorithms are unable to process textual data directly. To enable their utilization with categorical data, the one-hot encoding method is utilized. This process involves converting categorical data into one-hot vectors, represented as sparse vectors where one element is designated as 1 while all other elements are set to 0. This technique serves as a prevalent method for representing strings with a finite set of values in ML applications [36].

A one-hot vector is a single-row (or one-dimensional) matrix, typically consisting of zeros except for a single cell containing the value 1, which is used to uniquely identify a particular categorical variable or word. This encoding technique offers an expressive representation for categorical data. The length of the vector is generally determined by the total count of unique tokens present in the dataset or context [37].

2.1.3. Autoregressive Integrated Moving Average

Autoregressive Integrated Moving Average (ARIMA) is predicting with time series data to present a significant challenge due to the combination of limited information availability and the presence of significant fluctuations in economic trends and conditions. ARIMA, being a commonly employed model, aims to tackle these complexities inherentin time series prediction [38]. Autoregressive models operate under the premise that the present value of a series, denoted as x_t , can be elucidated as a function of its preceding p values, x_{t-1}, x_{t-2}, …,
x_{t-p}. Here, the parameter p determines the requisite number of steps into the past necessary to predict the current value [39]. The model represents as follows:

$$x_t = x_{t-1} - 0.90x_{t-2} + w_t \quad (2.3)$$

Frequently, the process of attaining stationarity in a time series demands data manipulation. Stationarity is a fundamental prerequisite for the efficacy of an ARIMA model in forecasting. A key characteristic of a stationary time series is its ability to sustain consistent statistical properties over time, encompassing attributes like mean and autocorrelation structure. Prior to fitting an ARIMA model, common practices involve employing differencing and power transformations on the data. These methods aim to eradicate trends and stabilize variance, particularly when the observed time series exhibits heteroscedasticity and trend components [40].

2.1.4. Seasonal Autoregressive Integrated Moving Average

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model represents an advanced iteration derived from the ARIMA framework. While the ARIMA model solely relies on historical data to make predictions, the SARIMA model extends this approach by additionally incorporating and addressing any observed seasonality patterns within the time series data.To encapsulate and model the observed seasonality patterns, SARIMA incorporates an extended formulation that accounts for seasonal variations within the dataset [41]:

$$\underbrace{(1-\phi_1 B)}_{\substack{\text{Non-seasonal}\\ \text{AR(1)}}}\ \underbrace{(1-\Phi_1 B^4)}_{\substack{\text{Seasonal}\\ \text{AR(1)}}}\ \underbrace{(1-B)}_{\substack{\text{Non-seasonal}\\ \text{difference}}}\ \underbrace{(1-B^4)}_{\substack{\text{Seasonal}\\ \text{difference}}} y_t \;=\; \underbrace{(1+\theta_1 B)}_{\substack{\text{Non-seasonal}\\ \text{MA(1)}}}\ \underbrace{(1+\Theta_1 B^4)}_{\substack{\text{Seasonal}\\ \text{MA(1)}}} e_t. \tag{2.4}$$

2.1.5. Ridge Regression

In datasets with numerous variables, the likelihood of reduced model explainability due to multicollinearity tends to increase. It is widely acknowledged that collinearity has adverse implications for the least squares (LS) estimator, particularly in regression analyses. Numerous methodologies have been devised to mitigate this impact, with a significant emphasis on variable elimination strategies, involving the removal of one or more independent variables to improve the conditioning of the resultant correlation

matrix among the remaining factors. In contrast, ridge regression presents an alternative approach to address collinearity issues without necessitating the removal of any variables from the original set of independent variables [42].

McDonald's work in 2009 substantiated that a significant segment of the literature concerning ridge regression focuses on determining the optimal ridge parameter, denoted as 'k' for practical application. This pursuit aims to identify an appropriate k-value, or in instances where an exact optimal value is unattainable, to ascertain a k-value that ensures the ridge estimator exhibits a lower mean squared error compared to the LS estimator [43].

2.1.6. Polynomial Regression

A polynomial regression model, distinct from basic linear regression, possesses the capability to accommodate non-linear relationships by fitting polynomial functions. This model is utilized in situations where the relationship's intricacy exceeds the capacity of linear regression models to accurately capture and represent the underlying non-linear associations between variables [44]. The simplest non-linear polynomial model represents as follow:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots \beta_n x^n + \varepsilon \quad (2.5)$$

2.1.7. eXtreme Gradient Boosting (XGBoost) Regression

Recently, XGBoost, a machine learning technique, was developed and is now widely used across multiple fields. Its organized, portable, and adaptable design makes it appropriate for a wide range of uses [45]. XGBoost, an innovative algorithm merging Cause Based Decision Tree (CBDT) and Gradient Boosting Machine (GBM), has amplified the tree boosting technique's effectiveness in swiftly and accurately handling diverse data types. This algorithm boasts versatility in constructing regression and classification models, making it suitable for targeted datasets. Moreover, XGBoost excels in managing large datasets with numerous attributes and classifications, presenting effective solutions for novel optimization challenges, particularly when balancing efficiency with accuracy [46].

The first and second order gradients for the objective function "squared error" were determined for each training case in XGBoost at each boosting iteration. The scikit-learn compatibility of XGBoost was used to build the model. XGBoost's tree-based learners produced the best results when they were used with 1000 boosts, 0.08 "colsample by tree," and 0.04 learning rate. The XGBoost regression mechanism is shown in Fig. sayı [47].

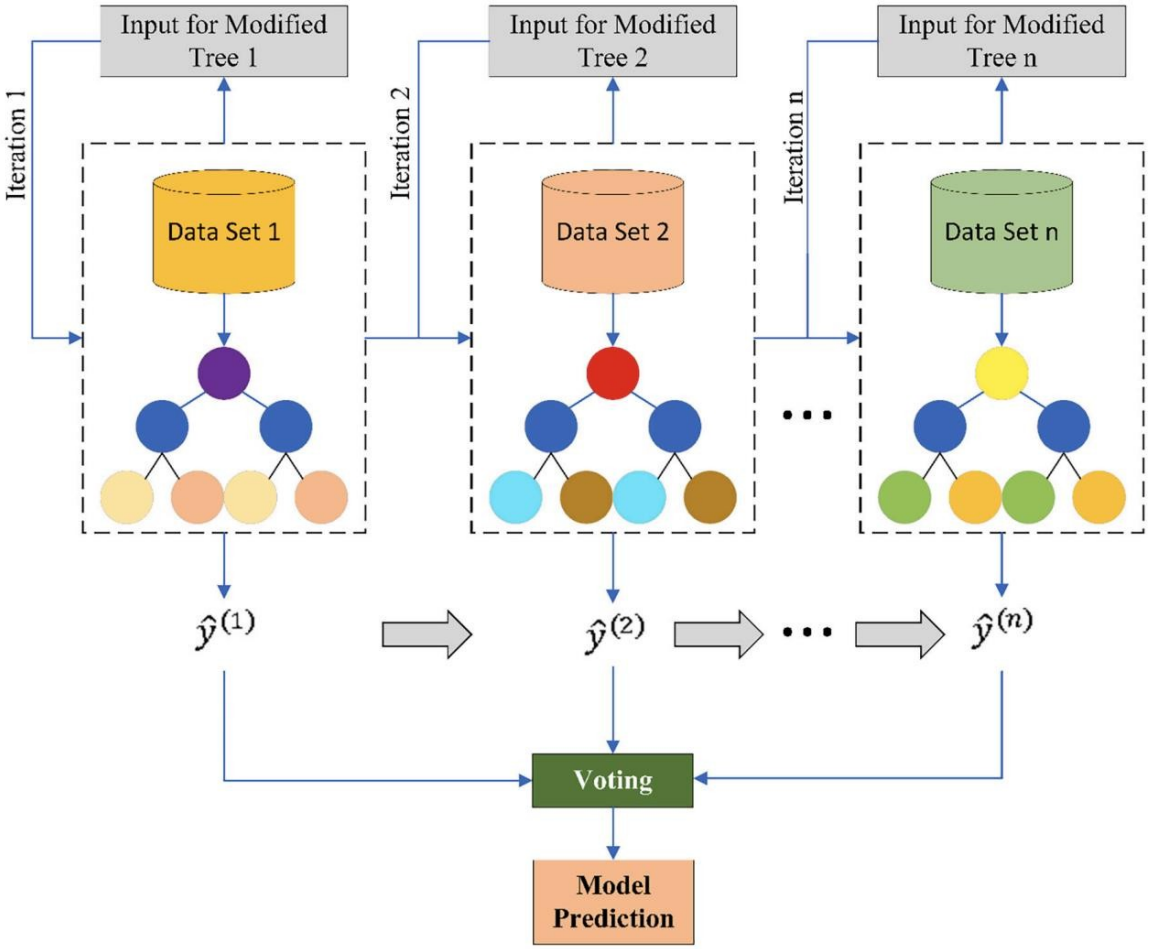

Figure 2.1: Modeling of XGBoost

Moreover, in two investigations conducted by Hamilton, Schlögl, and their colleagues XGBoost demonstrated superior performance compared to various other machine learning methodologies in forecasting. The evaluated models included Logistic Regression, Bayesian Regularized Neural Network, Pegasos SVM, Bagging Average Neural Networks, Deep Neural Network, and Gradient Boosting. Additionally, Shan, and his colleagues utilized artificial neural networks, integrating multiple XGBoost models, for predicting [48][49].

2.1.8. Light Gradient Boosting Machine (LightGBM) Regression

The LightGBM model, initially introduced by Microsoft, operates as a decision-tree-based algorithm [50]. It segments the input layer parameters into distinct sections to establish connections between the inputs and outputs. As depicted in Figure 1A, the primary characteristic of the LightGBM model is its utilization of leaf-wise-tree growth, a departure from the conventional level-wise-tree growth approach, which significantly accelerates the training process. In level-wise-tree growth, the tree structure progresses in a step-by-step level pattern [51].

Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS), two cutting-edge data sampling and classification techniques, have been combined to create this recently developed technique LightGBM [52]. Another cutting-edge machine learning-based data processing algorithm is used for more accurate residual value modeling and prediction. The LightGBM is used because of its exceptional skill, accuracy in data classification, and regression with a comparatively short processing time.

The algorithm's combination of functions enables efficient execution of tasks like data scanning, sampling, clustering, and classification, significantly outperforming traditional methods in terms of speed and accuracy. LightGBM is especially advantageous in scenarios prioritizing memory, processing time, and computational speed. It offers faster training, optimal memory utilization, ICU usage, satisfactory accuracy, parallel processing capabilities, and effective handling of large datasets. The ongoing investigation aims to evaluate and compare the performance of the newly introduced MDT-RV algorithm against XGBoost and LightGBM regression algorithms in predicting heavy equipment residual values [53].

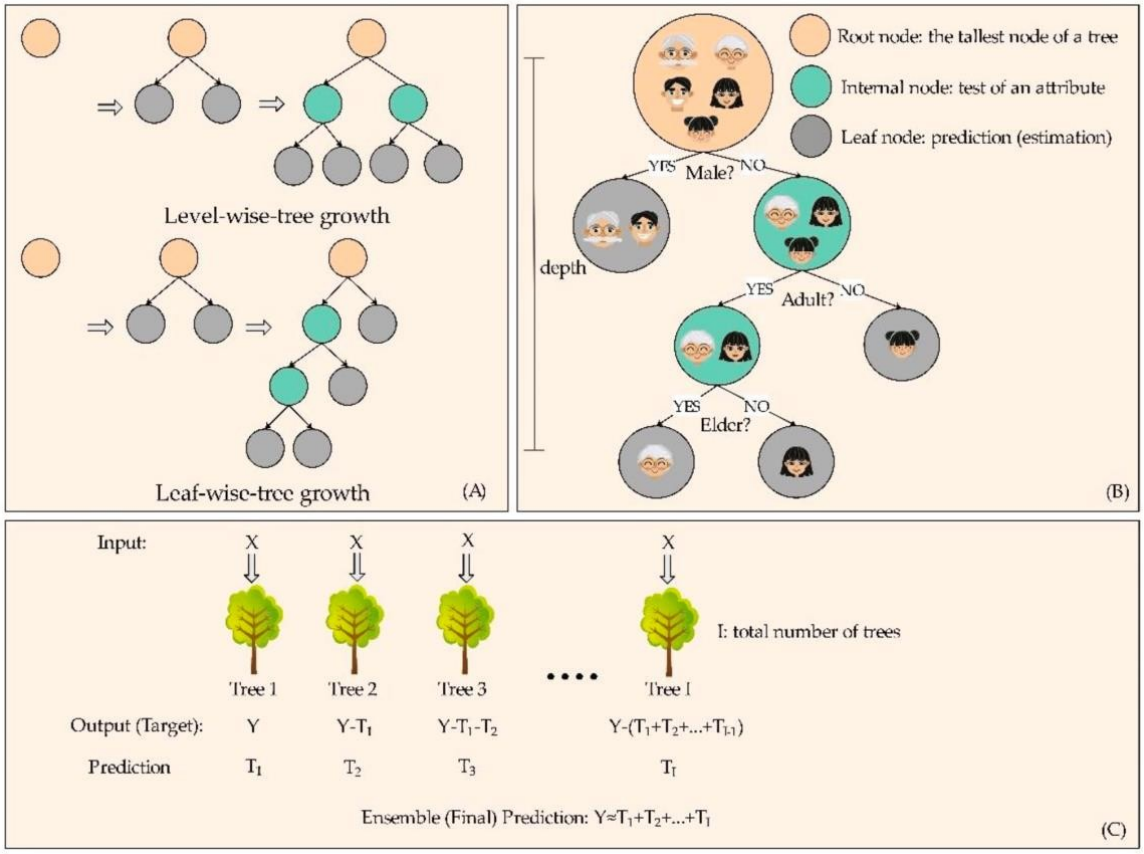

Figure 2.2: Schematic diagram of the LightGBM model: (A) growth tree structures; (B) an example of leaf-wise-tree growth conceptual algorithm and (C) Gradient Boosting Decision Tree algorithm

2.2. Evaluation of the Machine Learning Models

2.2.1. Root Mean Squared Error

The root mean-squared error (RMSE) is a standard metric used for model evaluation [54]. The RMSE is derived by dividing the sum of squared differences between predicted and observed values by the number of observations, followed by taking the square root of the resulting quotient. This mathematical operation serves as an evaluative metric, quantifying the dissonance between model-predicted values and the observed values in actuality. The RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Predicted_i - Actual_i)^2} \quad (2.6)$$

2.2.2. Mean Squared Error

The Mean Squared Error (MSE) serves as an evaluative metric for the effectiveness of an estimator, quantifying the level of error present in the models. It gauges the average squared disparity between actual and predicted values. It is primarily employed for identifying outliers within statistical models [55]. The MSE is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Predicted_i - Actual_i)^2 \quad (2.7)$$

2.2.3. Mean Absolute Error

The Mean Absolute Error (MAE) is a metric that quantifies the absolute difference between predicted and actual values, serving as a measure of the magnitude of errors [55]. The MAE is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Predicted_i - Actual_i| \quad (2.8)$$

2.2.4. Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) constitutes a metric designed to evaluate the precision of a forecasting model. This metric calculates the degree of accuracy in forecasting by determining the percentage difference between the forecasted quantities and the actual quantities [56]. The MAPE is calculated as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Predicted_i - Actual_i}{Actual_i}\right| \quad (2.9)$$

2.2.5. R-squared

The R-squared (R^2) is a statistical metric that denotes the fraction of the variance in a dependent variable explained by an independent variable within a regression model

[57]. While correlation delineates the intensity of the association between an independent and a dependent variable, R-squared elucidates the degree to which the variance of one variable clarifies the variance observed in the second variable. The R^2 is calculated as:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (2.10)$$

- R^2 : coefficient of determination
- RSS: sum of squares of residuals
- TSS: total sum of squares

2.3. Explainability of the Machine Learning Models

2.3.1. Multicollinearity

Multicollinearity is a method employed to elucidate the linear relationships among two or more variables, essentially assessing the statistical independence of those variables. While it is established that a high correlation signifies multicollinearity, the converse is not universally true [58]. It is noteworthy that multicollinearity may exist among variables even in the absence of a high correlation.

The foundational linear regression model is represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon \quad (2.11)$$

Where x's denote explanatory variables, β's represent regression coefficients, and ε signifies errors assumed to have a mean of 0 and a variance-covariance matrix of σ^2I. The least squares estimates β are given by:

$$b = (X'X)^{-1}X'Y \quad (2.12)$$

In the assessment of multicollinearity, regression coefficients may appear accurate despite the presence of multicollinearity or, conversely, may be inaccurate when

multicollinearity is absent. Elevated standard errors imply that the coefficients of certain independent variables may be deemed significantly different from zero. In simpler terms, multicollinearity, through the inflation of standard errors, can lead some variables to appear statistically insignificant even if they should be considered significant. A prominent indicator of multicollinearity is the manifestation of regression coefficients (b) with large standard errors or notable sampling variability.

Typically, we employ a diagnostic measure known as Variance Inflation Factors (VIF) to identify multicollinearity. The VIF serves as a tool for measuring and quantifying the extent to which variance is inflated. To understand the VIF value, the provided table below employs a specific rule for interpretation [59]:

Table 2.1: Brief information about VIF

VIF - value	conclusion
VIF = 1	Not correlated
1< VIF =< 5	Moderately correlated
VIF > 5	Highly correlated

In addition to indicating whether predictors are correlated, the square root of the VIF reveals the magnitude by which the standard error is increased. For instance, when the VIF is 9, it implies that the standard error of the predictor's coefficient is three times higher compared to the situation where the predictor is uncorrelated with other variables. The VIF is computed as follows:

$$VIF = \frac{1}{1-R_i^2} \tag{2.13}$$

Detecting multicollinearity is crucial prior to model development to mitigate the impact of elevated correlation, thereby enhancing the model's interpretability and explanatory power.

2.3.2. Shap

Shap, short for Shapley Additive Explanation, was conceptualized by Shapley within the framework of game theory [60].It provides a robust and illuminating metric for assessing the significance of a feature within a model. It is utilized for interpreting predictions, the method ranks feature importance, discerns the most crucial features, and facilitates the selection of optimal feature sets [61]. It works as considering any model employing a feature set to predict an output. In SHAP, the attribution of each feature to the model output is assigned based on their marginal contribution [62].

3. RESULTS

3.1. Dataset Description

3.1.1. Dataset

The dataset under consideration has been procured from a prominent textile company's e-commerce records spanning the years 2019 to 2021. The dataset is formed by combining two datasets: one containing transactional data based on dates and the other focused on campaign details. The campaign data set classifies campaigns into small, medium, and large categories based on criteria like product inclusivity, date duration, and discount rates. For example, a large campaign encompasses all products, spans a day, and offers a 20 percent discount. A medium campaign, on the other hand, may target only child and baby products with a 15 percent discount, while a small campaign could involve specific items like backpacks or a selected set of two hundred products with a 15 percent discount. At the outset, the dataset comprised 17 variables, each carefully selected to offer comprehensive insights into e-commerce dynamics. However, upon rigorous examination, the variable labeled "Holiday" was omitted from further analysis due to its adverse impact on predictive accuracy and increased error rates within the models. As a result of this strategic refinement, the dataset was streamlined to include 1,020 individual data points, incorporating 16 pertinent variables. Each variable underwent meticulous evaluation to ascertain its significance and analytical utility, ensuring the robustness and scholarly integrity of the ensuing investigation. These variables are elucidated as follows:

- Date: The specific date on which a transaction transpired.
- Day of the Date: An indicator signifying the day of the week associated with the transaction date.
- Year of the Date: An indicator designating the year corresponding to the transaction date.
- Month of the Date: An indicator delineating the specific month in which the transaction occurred.
- Web Session: The total number of web sessions recorded on the previous date.
- App Session: The total number of app sessions recorded on the previous date.

- Total Session: The overall volume of sessions (both web and app) on the previous date.
- Weather: The average temperature recorded on the date of the previous day transaction.Big Campaigns: The total number of significant marketing campaigns executed.
- Middle Campaigns: The aggregate count of intermediate-level marketing campaigns executed.
- Small Campaigns: The total number of minor marketing campaigns carried out.
- Total Campaigns: The cumulative count of all marketing campaigns executed.
- Average Price: The mean price of products sold on the previous date in question.
- Average Discount: The mean discount rate applied to products sold on the previous date.
- Dollar Currency: The average exchange rate of the U.S. dollar on the previous date.
- Transaction: The total number of orders placed on the given date.

This dataset serves as a valuable resource for conducting an academic exploration into various aspects of e-commerce, sales forecasting, and the multifaceted interplay between these variables.

3.1.2. Data Types

Before commencing the analysis, it is crucial to prepare the dataset adequately. The dataset comprises various data types, including five integers, one object, and ten floats. To enhance the dataset's suitability for analysis, it is necessary to change the data types ofthe "Year" and "Month" variables from their current format to object data types. This transformation is essential because these variables will be used as categories to assess theimpact of the year and month in the subsequent analysis.

Table 3.1: Name and Data Type of Variables

Data Type	Name of the Variable
int64	Date
object	Day of Week Name
int64	SessionsWEB
float64	SessionsAPP
int64	toplam_session
float64	WeatherAVG
float64	kamp_b
float64	kamp_k
float64	kamp_o
float64	kamp_toplam
float64	Ort_fiyat
float64	Ort_indirim
float64	Transaction
float64	dolar_kuru
int64	Year
int64	Month

3.2. Preliminary Data Analysis

3.2.1. Statistics of the dataset

This part presents a review of the fundamental statistics of the dataset, offering crucial insights into the properties and distribution of the data being analyzed. The dataset's summary statistics provide a basic comprehension of its numerical characteristics. For every numerical variable in the collection, the descriptive statistics comprise metrics like mean, standard deviation, minimum, maximum, and quartile values.

- Mean Values: The central tendency of the dataset can be inferred from the arithmetic mean of its numerical variables.
- Standard Deviation: The standard deviation illustrates the variability within each numerical feature by reflecting the dispersion of data around the mean.
- Minimum and Maximum Values: The dataset's range of values is shown by the values between the minimum and maximum values.
- Quartiles: Information about the central tendency and dispersion of the distribution can be gleaned from the quartile values, which include the median (50th percentile) and the lower and upper quartiles (25th and 75th percentiles).

Here is an overview table showcasing the main statistical details of the dataset:

	Date	SessionsWEB	SessionsAPP	toplam_session	WeatherAVG	kamp_b	kamp_k	kamp_o	kamp_toplam	Ort_fiyat	Ort_indirim	Transaction	dolar_kuru	index
count	1.021000e+03	1.021000e+03	1.021000e+03	1.021000e+03	1021.000000	1021.000000	1021.000000	1021.000000	1021.000000	1021.000000	1021.000000	1021.000000	1021.000000	1021.000000
mean	2.020131e+07	7.023096e+05	3.390517e+05	1.041361e+06	17.080205	0.226249	2.411361	1.226249	3.863859	38.171785	0.133579	82391.480901	7.302317	518.923604
std	8.036461e+03	2.509328e+05	2.254100e+05	4.402740e+05	6.849399	0.560756	1.976139	1.877299	3.063075	9.326231	0.060493	54855.783502	1.684509	300.259248
min	2.019030e+07	1.013220e+05	6.296700e+04	1.642890e+05	-0.555556	0.000000	0.000000	0.000000	0.000000	24.793028	0.031069	3.000000	5.332200	1.000000
25%	2.019111e+07	5.059420e+05	1.986770e+05	7.263740e+05	11.666667	0.000000	1.000000	0.000000	2.000000	30.211243	0.092881	40544.000000	5.879300	256.000000
50%	2.020081e+07	6.775590e+05	2.929530e+05	9.723290e+05	17.222222	0.000000	2.000000	1.000000	3.000000	36.882294	0.121688	77451.000000	7.022700	524.000000
75%	2.021042e+07	8.558230e+05	3.883810e+05	1.255654e+06	23.333333	0.000000	3.000000	2.000000	5.000000	42.336407	0.170682	105190.000000	8.278100	779.000000
max	2.021123e+07	2.240826e+06	1.885602e+06	3.146333e+06	30.000000	5.000000	12.000000	12.000000	18.000000	66.505517	0.362580	504141.000000	16.484800	1034.000000

Figure 3.1: Summary of Statistics

The variable "Transaction" exhibits the highest standard deviation due to its fluctuating nature influenced by various factors such as promotional campaigns, seasonal transitions, holiday periods, and technical issues, resulting in peak and downtrend periods. Conversely, the "Average Discount" variable portrays the lowest standard deviation, indicating its limited diversity and heightened data consistency. Notably, the dataset indicates a maximum campaign count of eighteen, with a prevalent occurrence of smaller to medium-sized campaigns. An inspection of the "Transaction" variable reveals outliers, with extreme values ranging between 3 and 504,141 and a substantial standard deviation of 54,861, illustrating a more dispersed distribution of data. Conversely, variables like "Average Weather" and "Average Price" demonstrate higher consistency in results, reflected by their standard deviation, mean, and quartile measures.

3.2.2. Correlation Matrix

The links between the variables in the dataset are shown visually in the correlation heatmap that is produced from the correlation matrix. In order to comprehend the relationships and interdependencies among the characteristics being studied, an analysis of the correlation heatmap is provided in this section.

A color gradient is used in the heatmap to depict the direction and intensity of the pairwise correlations between the variables. The colors' various intensities represent the correlation coefficients' magnitudes, with blue denoting a negative correlation and red denoting a positive correlation.

These indicators can be described as follows:

- Strength of Correlation: The heatmap's color intensity reflects how strongly the variables are related to one another. Lighter hues indicate lower connections, while darker hues indicate greater correlations.
- Positive and Negative Correlation: Positive correlations are shown in red tones, indicating that as one variable rises, the other also tends to rise. On the other hand, blue hues suggest negative correlations, suggesting an inverse relationship in which one variable tends to rise while the other falls.
- No Correlation: There is no discernible linear link between the variables in the areas that are colorless or neutral (almost white).

Upon examining the results, there emerges a weak correlation between sessions and date, whereas a robust correlation of 0.8 is observed between sessions and transactions. Notably, the average price exhibits a noteworthy correlation with both date and dollar currency, possibly indicating a connection between the rise in dollar currency and the concurrent inflation rates in Turkey. Moreover, a negative correlation between the average discount and average price is evident, aligning with expectations wherein an escalation in discounts leads to a reduction in average prices. Conversely, no substantial correlation is observed between transaction and campaign variables. Additionally, no significant correlation is detected between weather and transaction variables.

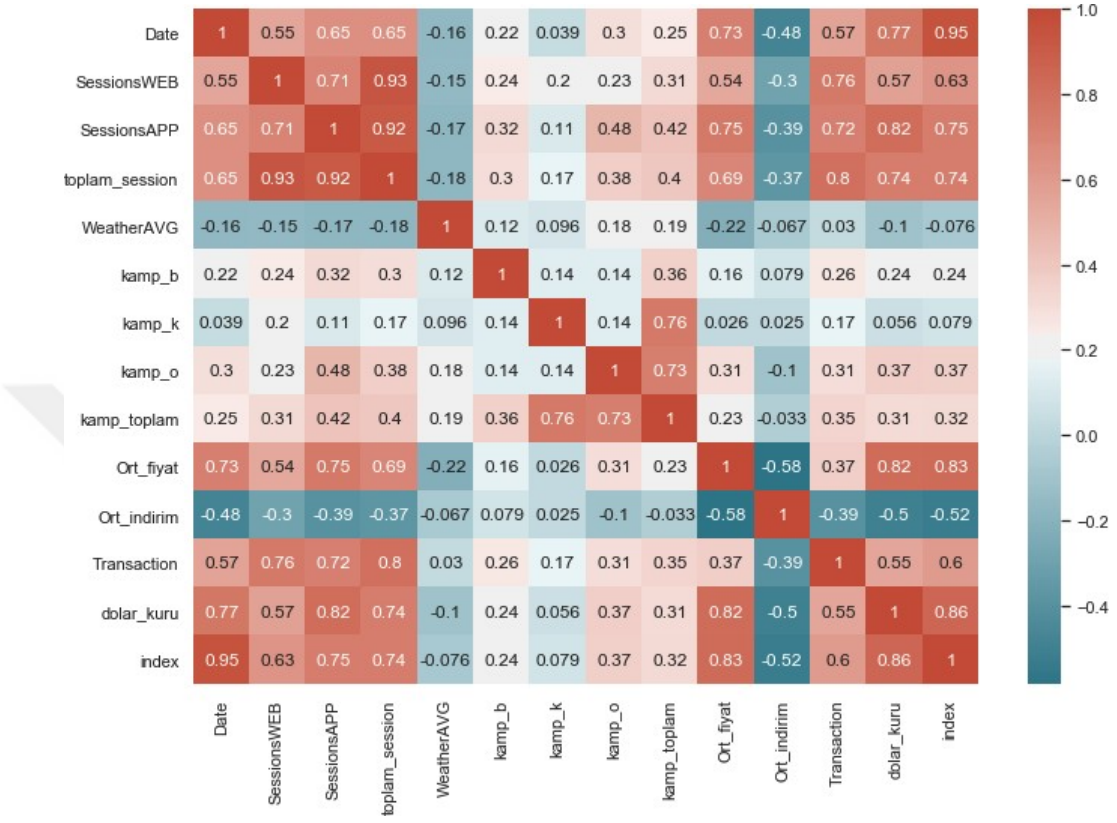

Figure 3.2: Correlation Matrix

3.2.3. Analysis of Multiple Columns

Pairplot visualization provides a thorough understanding of the connections between the dataset's many variables. It displays pairwise relationship scatterplots between the designated columns, illuminating possible patterns and associations betweenthem.

The relationship between two distinct columns from the dataset is represented by each cell in the grid of scatterplots that the pairplot displays. The off-diagonal cells display scatter plots that show the relationships between pairs of variables, while the diagonal cells display histograms or kernel density estimates for each individual variable.

I have selected four variables to scrutinize their interrelationships: transaction (the

target variable for prediction), total sessions (comprising both app and web sessions, offering a representation of both variables), total campaigns (encompassing the sum of small, medium, and large campaign sizes that impact sessions and transactions), and average discount (linked to average price, date, and dollar currency). These variables have been chosen based on their potential correlations and influence on the transactional outcome.

Upon examination, all variables within the dataset display a right-skewed distribution. The association between transaction and total session reveals a positive polynomial correlation: as the total session increases, so does the total transaction count. However, the relationship between total campaign and transaction appears to be widely dispersed, indicating a lack of discernible correlation. Similarly, transaction and average discount showcase a negative polynomial relationship; as the average discount increases, transaction levels demonstrate a decreasing trend. This occurrence might be attributed to the influence of product mix and seasonal effects. Typically, older seasons tend to exhibit higher discount rates. Therefore, a higher average discount could signify a product mix primarily comprising older seasons, potentially leading to a reduced variety of products for customers.

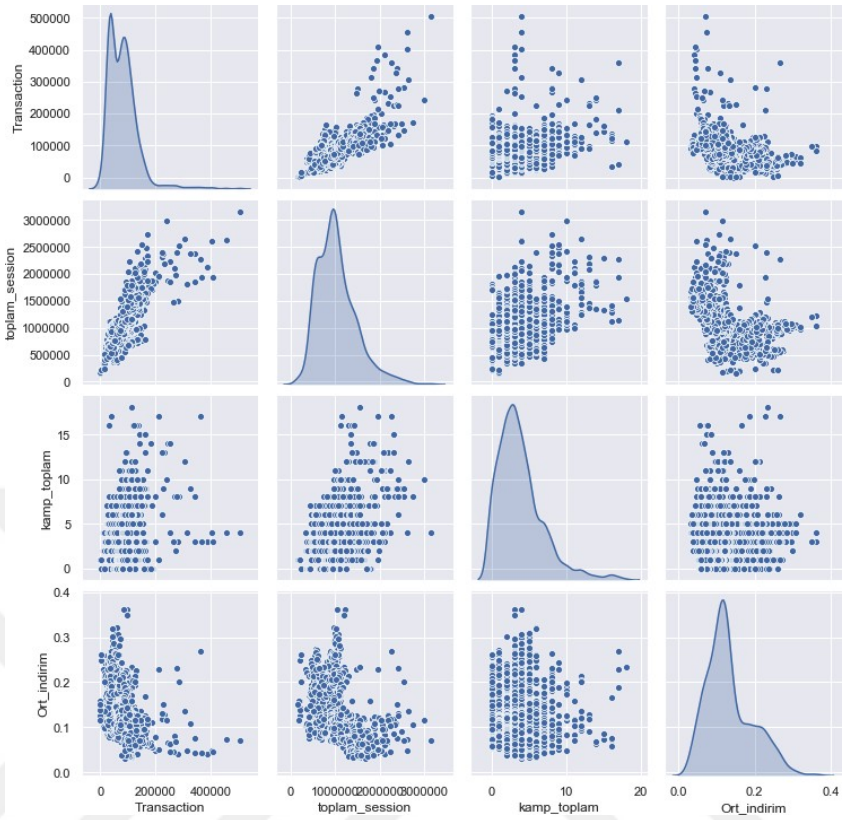

Figure 3.3: Multiple Columns Analysis

3.3. Data Preprocessing and Featuring Engineering

Before embarking on the analysis, it is imperative to conduct a comprehensive examination of the dataset to identify any missing values and determine the appropriate course of action for handling them. Upon close scrutiny, it has been ascertained that several variables contain null values, specifically:

- Five missing values in "App Sessions"
- Three missing values in "Weather"
- Ninety-six missing values in "Campaigns"
- Seven missing values in "Average Price"
- Eight missing values in "Transactions"

- Two hundred ninety-four missing values in "Dollar Currency"

The subsequent steps in the analytical process should involve a deliberate decision-making process regarding how to address these null values, whether through imputation or removal, contingent upon the specific analytical objectives and methodology.

Table 3.2: Missing Values

Name of the Variable	Total number of null values
Date	0
Day of Week Name	0
SessionsWEB	0
SessionsAPP	5
toplam_session	0
WeatherAVG	3
kamp_b	96
kamp_k	96
kamp_o	96
kamp_toplam	96
Ort_fiyat	7
Ort_indirim	0
Transaction	8
dolar_kuru	294
Year	0
Month	0

Upon conducting a comprehensive assessment of the distribution of the "Dollar Currency" variable, it was discerned that the null values exhibited a discernible pattern. This pattern appeared to be consistent and trend-driven. Consequently, it was determinedthat these null values are amenable to imputation using a method that aligns with the consistent trend observed. In this context, the "pad" method was judiciously employed, asit enables the filling of null values with the preceding value in the dataset. This methodological choice was made based on the rationale that it is well-suited to maintain the established trend within the "Dollar Currency" variable, contributing to the continuityand integrity of the dataset.

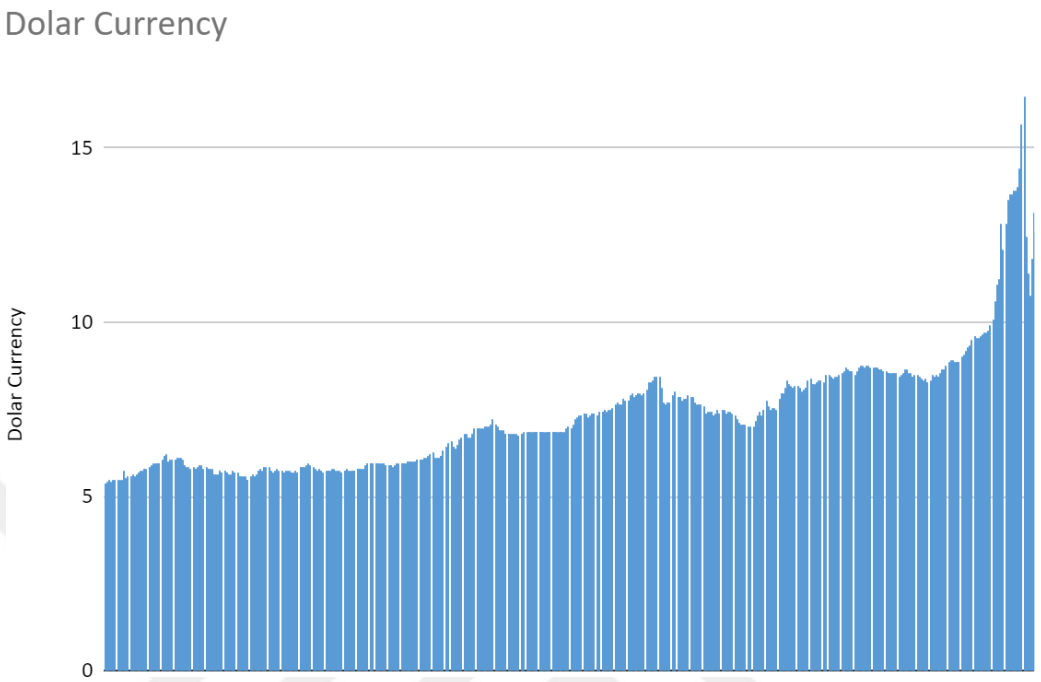

Figure 3.4: Dollar Currency Trend

Upon a thorough examination of the "Weather" variable, it was observed that the temperature values generally remained lower than those of the "Dollar Currency," with

the exception of a single anomalous instance. It is noteworthy that the "Weather" variable exhibited a distinct seasonal trend, characterized by limited day-to-day fluctuations.

In light of these observations and in keeping with the consistent trend exhibited by the "Weather" variable, a decision was made to employ a similar imputation method as was used for the "Dollar Currency." Specifically, the "pad" method was applied to fill the null values in the "Weather" variable with the preceding temperature value. This methodological approach was chosen to preserve the prevailing temperature trends and ensure the data's continuity and integrity.

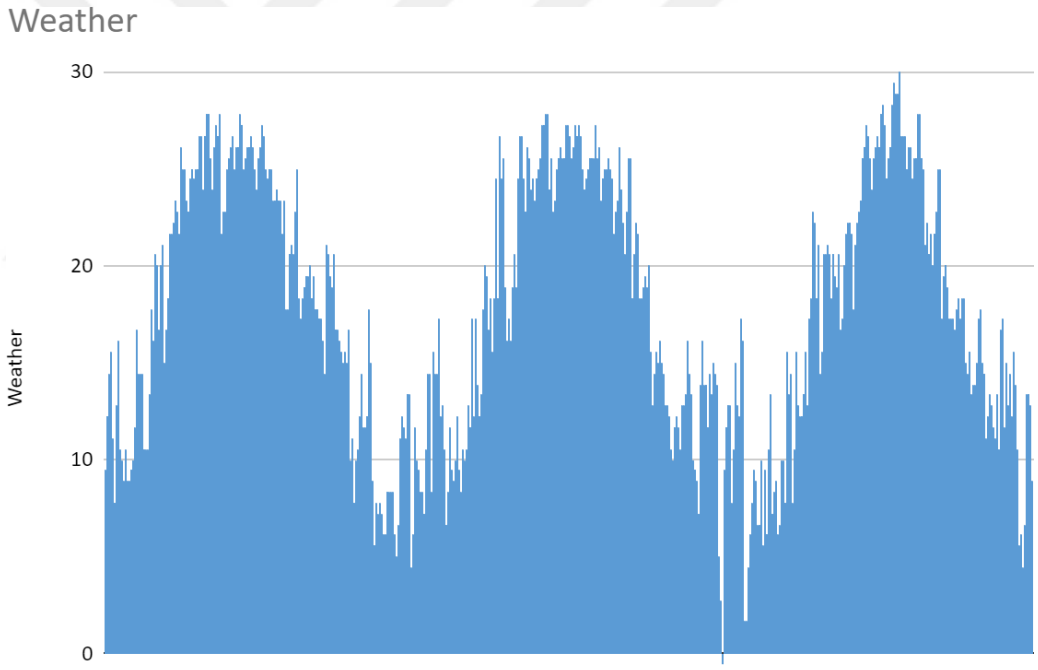

Figure 3.5: Weather Trend

In addressing the null values in the "Campaign" variable, a straightforward approach was taken. Null values in this context were interpreted to signify the absence of any campaign activity on the given date. Therefore, all null values in the "Campaign" variable were replaced with zeros.

However, when it comes to the "Transaction" variable, which serves as the target value for prediction and holds significant importance in the model's predictive quality, a

different strategy was employed. Null values within the "Transaction" variable were systematically dropped from the dataset. This measure was adopted to preserve the integrity and reliability of the transaction data, ensuring that the dataset used for predictive modeling remained complete and devoid of null values. Consequently, following the removal of null values from the "Transaction" variable, the dataset was rendered free of any remaining null values.

3.3.1. Explanation of Interaction Terms

Within my dataset, several independent variables not only exert an influence on the dependent variable but also exhibit interconnected relationships among themselves. To mitigate this effect, I introduce five modifiers:

- Web-App Ratio: Calculated as the ratio of total web sessions to total app sessions.
- WebSession Campaign: Found by multiplying the total number of campaigns by total web sessions.
- AppSession Campaign: Obtained by multiplying the total number of campaigns by total app sessions.
- PriceDiscount: Calculated by multiplying the average price by the average discount.
- DollarDiscount: Derived by multiplying the dollar currency by the average discount.

3.3.2. Explanation of One-Hot Encoding

One-hot encoding was applied to the date variables, wherein the dates were categorized into days of the week, months, and years. In response to the substantial surge in online sales due to the COVID-19 pandemic and the subsequent closure of physical stores, the year was incorporated as a categorical variable to account for this notable change. Recognizing the significance of transitional seasons in the textile retail industry and the distinct trends within different months and days of the week, I employed one-hot encoding for the Months, Years, and Days of the Week variables. To avoid multicollinearity in the subsequent analysis, I dropped 'Month_12', 'Day of Week

Name_Sunday', and 'Year_2019' from the dataset.

3.4. Model Performance Outputs

3.4.1. Implementation of ARIMA

In the ARIMA model, I utilized the date and transaction variables with a parameter setting of 7, 1, 1. The obtained results are as follows: The log likelihood was -11760, computed over 1025 observations. The Akaike Information Criterion (AIC) is 23540, the Bayesian Information Criterion (BIC) is 23589, and the Hannan-Quinn Information Criterion (HQIC) is 23558. Notably, there was one statistically significant outcome at Lag 5, with a p-value of 0.001.

ARIMA Model Results

Dep. Variable:	D.Transaction	No. Observations:	1025
Model:	ARIMA(7, 1, 1)	Log Likelihood	-11760.085
Method:	css-mle	S.D. of innovations	23253.092
Date:	Wed, 20 Apr 2022	AIC	23540.171
Time:	14:08:27	BIC	23589.495
Sample:	1	HQIC	23558.895

	coef	std err	z	P>|z|	[0.025	0.975]
const	38.5026	424.521	0.091	0.928	-793.544	870.549
ar.L1.D.Transaction	-0.8765	0.186	-4.702	0.000	-1.242	-0.511
ar.L2.D.Transaction	-0.3519	0.058	-6.041	0.000	-0.466	-0.238
ar.L3.D.Transaction	-0.2197	0.059	-3.697	0.000	-0.336	-0.103
ar.L4.D.Transaction	-0.2017	0.046	-4.392	0.000	-0.292	-0.112
ar.L5.D.Transaction	-0.1749	0.052	-3.341	0.001	-0.277	-0.072
ar.L6.D.Transaction	-0.0727	0.046	-1.590	0.112	-0.162	0.017
ar.L7.D.Transaction	0.0436	0.036	1.203	0.229	-0.027	0.115
ma.L1.D.Transaction	0.6663	0.184	3.612	0.000	0.305	1.028

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.2189	-0.4286j	1.2920	-0.4462
AR.2	-1.2189	+0.4286j	1.2920	0.4462
AR.3	-0.5486	-1.2839j	1.3962	-0.3143
AR.4	-0.5486	+1.2839j	1.3962	0.3143
AR.5	0.8752	-1.1298j	1.4292	-0.1451
AR.6	0.8752	+1.1298j	1.4292	0.1451
AR.7	3.4525	-0.0000j	3.4525	-0.0000
MA.1	-1.5009	+0.0000j	1.5009	0.5000

Figure 3.6: ARIMA Results

3.4.2. Implementation of SARIMA

For the SARIMA model, I employed the same dataset with a 12-month parameter

to capture seasonality. The obtained results are as follows: SARIMA yielded a 33% higher RMSE, a 77% higher MSE, a 75% higher MAPE, and the MAPE value is twice as much when compared to ARIMA. Hence, based on these metrics, ARIMA demonstrates superior performance compared to SARIMA.

Table 3.3: ARIMA and SARIMA Results

Models	RMSE	MSE	MAE	MAPE
SARIMA	63372	4016001802	55893	64.16%
ARIMA	47601	2265889800	31919	30.52%

3.4.3. Implementation of Polynomial Regression

During the initial analysis, it was evident that a polynomial relationship exists between the variables. Subsequently, I applied a polynomial regression model. Upon reviewing the results, the RMSE was calculated at 18899, the MSE at 357168492, the MAE at 12245, the MAPE at 16.84%, and the R-square value was determined as 79.4%. Upon examining the predicted and actual results, it is evident that the blue line inadequately captures the fluctuations in the left curve and on the right tail, which demonstrates a right-skewed pattern. However, it does reasonably well in approximating a flatter line.

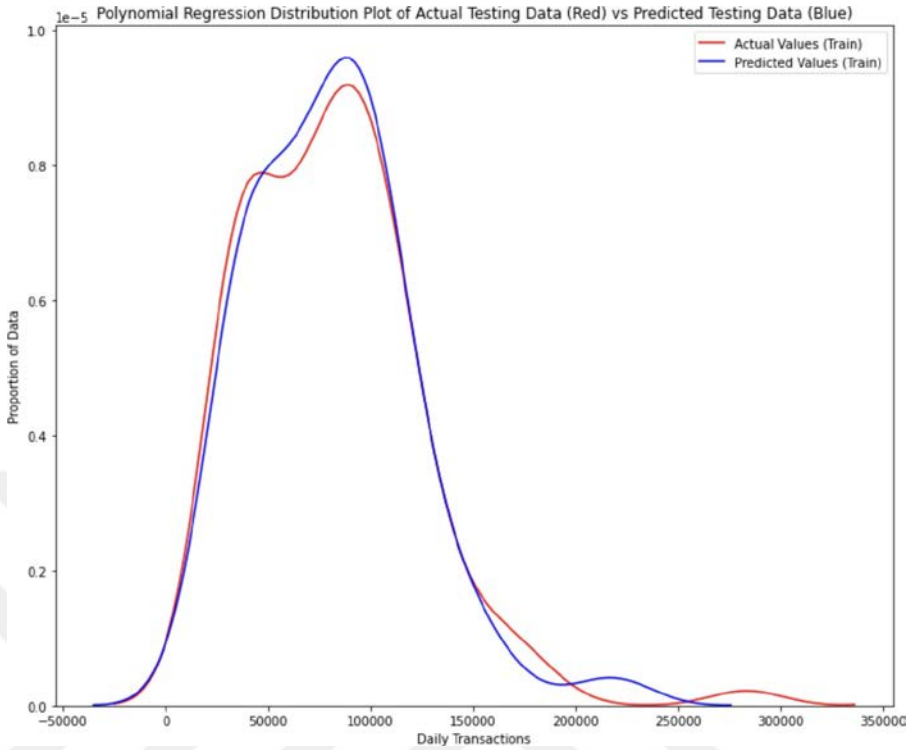

Figure 3.7: Polynomial Regression Actual and Predicted Line Graph

3.4.4. Implementation of Ridge Regression

Prior to employing ridge regression, I utilized grid search for linear algorithms to determine the most suitable parameter for the dataset. I conducted tests with parameters such as 0.001, 0.1, 1, 10, 100, 1000, 10000, 100000, fitting them to the training dataset. The results indicate that the parameter 0.001, along with the "normalize" set to false, emerges as the optimal choice for the dataset.

mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_alpha	param_normalize	params	split0_test_score	split1_test_score
0.004850	0.002457	0.001386	0.001381	10	False	{'alpha': 10, 'normalize': False}	-5.079142e+08	-2.553284e+08
0.003066	0.004432	0.001355	0.003039	1	False	{'alpha': 1, 'normalize': False}	-4.977715e+08	-2.511310e+08
0.015397	0.034295	0.000251	0.000637	0.001	True	{'alpha': 0.001, 'normalize': True}	-4.974556e+08	-2.508578e+08
0.002064	0.002147	0.000291	0.000874	0.1	False	{'alpha': 0.1, 'normalize': False}	-4.974643e+08	-2.476251e+08
0.002115	0.002538	0.002029	0.002400	0.001	False	{'alpha': 0.001, 'normalize': False}	-4.992915e+08	-2.467343e+08

split2_test_score	split3_test_score	split4_test_score	split5_test_score	split6_test_score	split7_test_score	split8_test_score
-3.003500e+08	-3.535729e+08	-4.132621e+08	-4.065661e+08	-4.368294e+08	-6.638811e+08	-5.284807e+08
-3.358776e+08	-3.584359e+08	-5.052859e+08	-3.970558e+08	-4.413984e+08	-7.050226e+08	-5.138669e+08
-3.367903e+08	-3.586260e+08	-5.101942e+08	-3.971459e+08	-4.414473e+08	-7.097046e+08	-5.130107e+08
-3.391610e+08	-3.600468e+08	-5.345795e+08	-3.999716e+08	-4.381969e+08	-7.187568e+08	-5.033747e+08
-3.388923e+08	-3.615763e+08	-5.409125e+08	-4.034921e+08	-4.362110e+08	-7.140629e+08	-4.986094e+08

split9_test_score	mean_test_score	std_test_score	rank_test_score
-7.988586e+08	-4.665044e+08	1.572376e+08	1
-7.539816e+08	-4.759827e+08	1.496469e+08	2
-7.530562e+08	-4.768289e+08	1.502090e+08	3
-7.481570e+08	-4.787334e+08	1.513611e+08	4
-7.480410e+08	-4.787823e+08	1.507199e+08	5

Figure 3.8: Ridge Regression Parameter Results

Identifying the parameters, I trained the ridge regression model and obtained the following results: RMSE of 18515, MSE of 342792969, MAE of 12428, MAPE of 17.18%, and an r-squared value of 80.3%.

Reviewing the graphical representation, similar to the polynomial regression, the model appears to struggle in capturing fluctuations but performs adequately in predicting smoother trends.

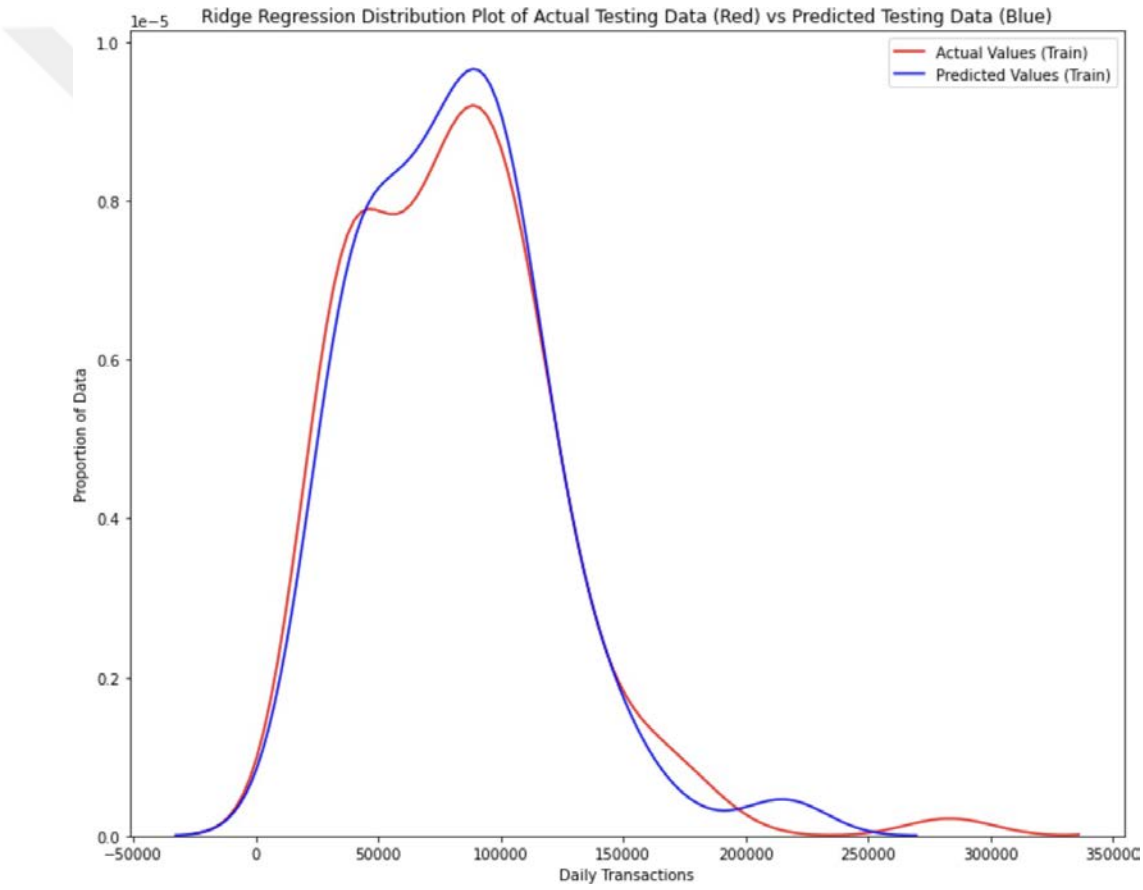

Figure 3.9: Ridge Regression Actual and Predicted Line Graph

3.4.5. Implementation of XGBoost

In preparation for training the XGBoost model, a hyperparameter tuning process was employed to identify the optimal parameters for the training dataset. The parameters

investigated include 'max_depth' in the range of 3 to 18, 'gamma' ranging between 1 to 9, 'reg_alpha' varying from 40 to 180, 'reg_lambda' spanning from 0 to 1, 'colsample_bytree' between 0.5 to 1, 'subsample' within the range of 0.5 to 1, 'min_child_weight' from 0 to 10, 'learning_rate' with a log-uniform distribution between log(0.005) to log(0.2), and 'n_estimators' between 100 to 1000 in increments of 200. This comprehensive exploration aims to ascertain the best parameters for the XGBoost model based on the training data.

Following the execution of the hyperparameter tuning model with the training dataset, the model identified the following optimized parameters:

- colsample_bytree: 0.7487886211452643
- gamma: 6.107950479155512
- learning_rate: 0.0627844179924045
- max_depth: 3.0
- min_child_weight: 8.0
- n_estimators: 800.0
- reg_alpha: 168.0
- reg_lambda: 0.7928207022353595
- subsample: 0.783904319732490

These parameters were determined to be the most suitable for enhancing the XGBoost model's performance based on the training dataset.

Following the implementation of the XGBoost regression model with the optimized tuning parameters, the model produced the following results:

- RMSE: 11918
- MSE: 142036826
- MAE: 8102
- MAPE: 9.88%R-squared : 91.8%

These metrics indicate the performance of the XGBoost regression model on the

dataset, demonstrating a strong predictive capability with high accuracy in explaining the variance of the dependent variable.

Comparing the predicted results represented by the blue line among polynomial, ridge regression, and XGBoost models, the XGBoost model demonstrates superior predictive performance. While it captures similar fluctuations as observed in the other models, it shows a different pattern on the right tail, displaying a more spiked behavior. Despite not perfectly replicating the observed wave, the XGBoost model exhibits improved predictive accuracy compared to polynomial and ridge regression models.

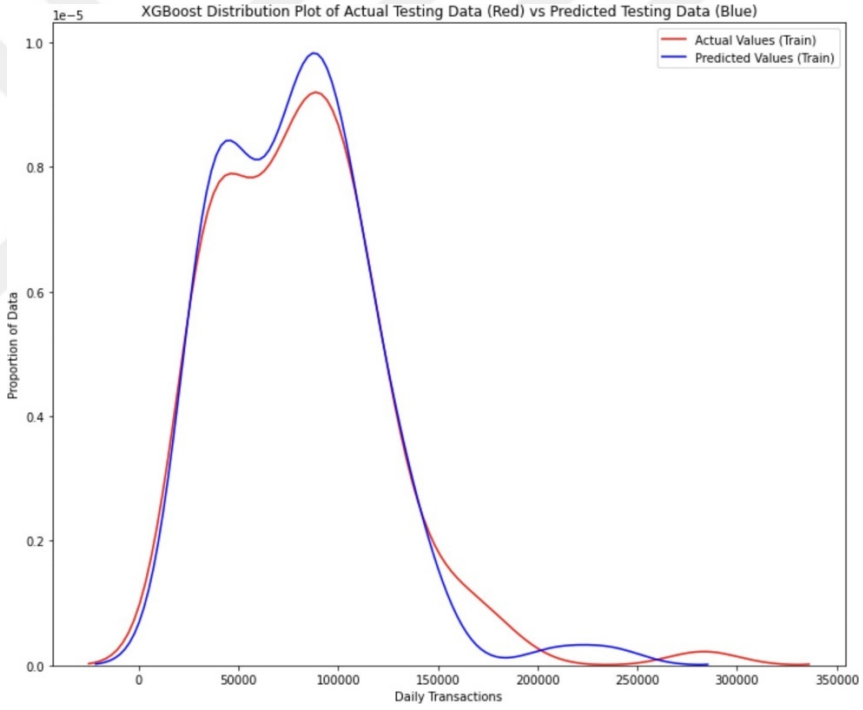

Figure 3.10: XGBoost Actual and Predicted Line Graph

3.4.6. Implementation of LightGBM

The LightGBM regression model produced the following results:

- RMSE: 9945
- MSE: 98899432
- MAE: 6869
- MAPE: 8.52%

- R-squared : 94.3%

These metrics indicate the performance of the LightGBM model in predicting the target variable. The lower the RMSE, MSE, MAE, and MAPE, the better, while a higher R-squared value signifies a better fit of the model to the data. Overall, these results suggest that the LightGBM model performed well in making predictions for the given dataset.

When examining the graphical representation, it becomes evident that the LightGBM model fits the fluctuations more closely and exhibits a slightly more undulating pattern. In particular, it appears to better fit the right-skewed aspect compared to the XGBoost regression model.

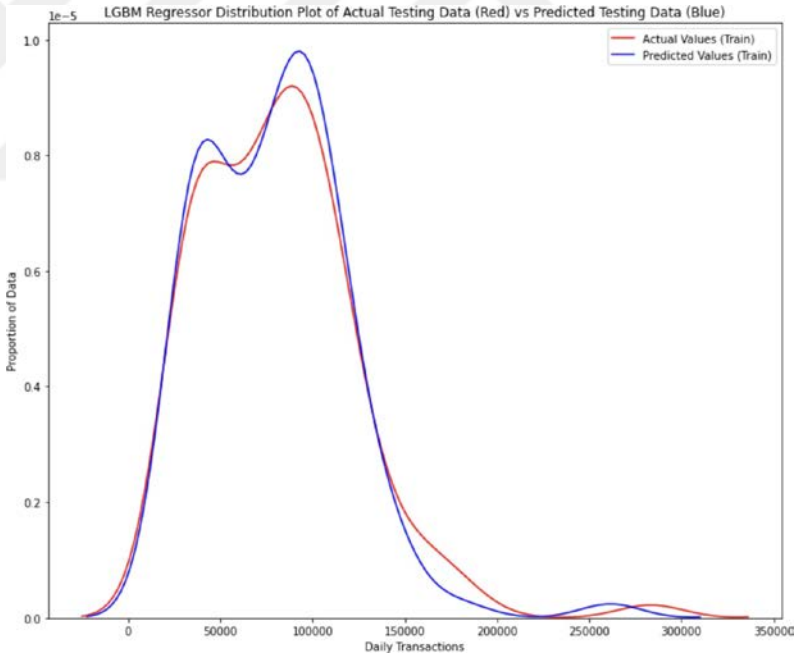

Figure 3.11: LightGBM Actual and Predicted Line Graph

3.5. Model Explainability

3.5.1. Evaluation of Multicollinearity

In conducting the multicollinearity analysis twice, I considered all variables initially, including those from one-hot encoding. In the first analysis, prior to dropping 'Month_12', 'Day of Week Name_Sunday', and 'Year_2019' variables, I observed

moderate multicollinearity between 'web-app ratio' and 'average weather.' Conversely, the majority of the other variables exhibited high multicollinearity.

Table 3.4: Multicollinearity Results Before One-Hot Coding

Index	Var	Vif
0	SessionsWEB	inf
29	Month_3	inf
22	Day of Week Name_Tuesday	inf
23	Day of Week Name_Wednesday	inf
24	Month_1	inf
25	Month_10	inf
26	Month_11	inf
27	Month_12	inf
28	Month_2	inf
30	Month_4	inf
20	Day of Week Name_Sunday	inf
31	Month_5	inf
32	Month_6	inf
33	Month_7	inf
34	Month_8	inf
35	Month_9	inf
36	Year_2019	inf
37	Year_2020	inf
21	Day of Week Name_Thursday	inf
19	Day of Week Name_Saturday	inf
1	SessionsAPP	inf
18	Day of Week Name_Monday	inf
2	toplam_session	inf
4	kamp_b	inf
5	kamp_k	inf
6	kamp_o	inf
7	kamp_toplam	inf
38	Year_2021	inf
17	Day of Week Name_Friday	inf
11	index	1261.61
16	Dolar-indirim	65.23
15	Fiyat-indirim	55.17
9	Ort_indirim	44.61
13	WebSession-Kampanya	39.37
14	AppSession-Kampanya	32.87
8	Ort_fiyat	31.49
10	dolar_kuru	29.22
12	Web-App Ratio	8.42

3	WeatherAVG	6.24

Dropping the specified variables ('Month_12', 'Day of Week Name_Sunday', and 'Year_2019'), the revised analysis revealed a shift in the results. There are now two variables exhibiting moderate multicollinearity, specifically 'Month_9' and 'Month_10'. Additionally, seven variables—'Month_11', 'Day of Week Name_Monday', 'Day of Week Name_Tuesday', 'Day of Week Name_Wednesday', 'Day of Week Name_Thursday', 'Day of Week Name_Friday', and 'Day of Week Name_Saturday'—demonstrate low multicollinearity.

Table 3.5: Multicollinearity Results After One-Hot Coding.

Index	Var	Vif
0	SessionsWEB	inf
2	toplam_session	inf
1	SessionsAPP	inf
4	kamp_b	inf
5	kamp_k	inf
6	kamp_o	inf
7	kamp_toplam	inf
12	index	1762.64
8	Ort_fiyat	612.50
11	dolar_kuru	561.47
17	Dolar-indirim	487.19
16	Fiyat-indirim	475.19
9	Ort_indirim	255.71
36	Year_2021.0	243.84
13	Web-App Ratio	114.78
14	WebSession-Kampanya	69.01
35	Year_2020.0	59.97
3	WeatherAVG	41.99
15	ppSession-Kampanya	36.81
24	Month_1.0	28.46
28	Month_3.0	22.48

10	Transaction	22.21
27	Month_2.0	21.25
29	Month_4.0	17.07
30	Month_5.0	16.11
31	Month_6.0	14.89
32	Month_7.0	13.88
33	Month_8.0	11.66
34	Month_9.0	07.02
25	Month_10.0	05.02
26	Month_11.0	2.70
20	Day of Week Name_Saturday	2.11
18	Day of Week Name_Friday	02.09
21	Day of WeekName_Thursday	02.07
23	Day of WeekName_Wednesday	02.04
22	Day of WeekName_Tuesday	02.03
19	Day of Week Name_Monday	02.02

3.5.2. Evaluation of SHAP

Certainly, when examining the SHAP results, it was revealed that the variable 'Total Session' had the most substantial impact on the predictive model, which aligns with expectations. Subsequently, 'App Sessions' emerged as the second most influential variable. In e-commerce, it is commonly known that app conversion rates tend to surpass web conversion rates, underscoring the importance of this variable. To moderate the model, 'App-Web Ratio' was introduced. The 'Index' variable ranked third in importance due to the expanding trend line in transactional activities within the growing e-commerce sphere.

Further impactful variables included 'Average Price', influenced by campaigns and concurrently associated with 'Average Discount'. This was followed by 'Dollar Currency', 'Average Weather', interaction variables such as 'Price-Discount' and 'App-Web Ratio', 'Fifth Month' (Ramadan holiday month), 'Average Discount', 'Dollar Currency-Discount', and 'Medium Campaigns'.

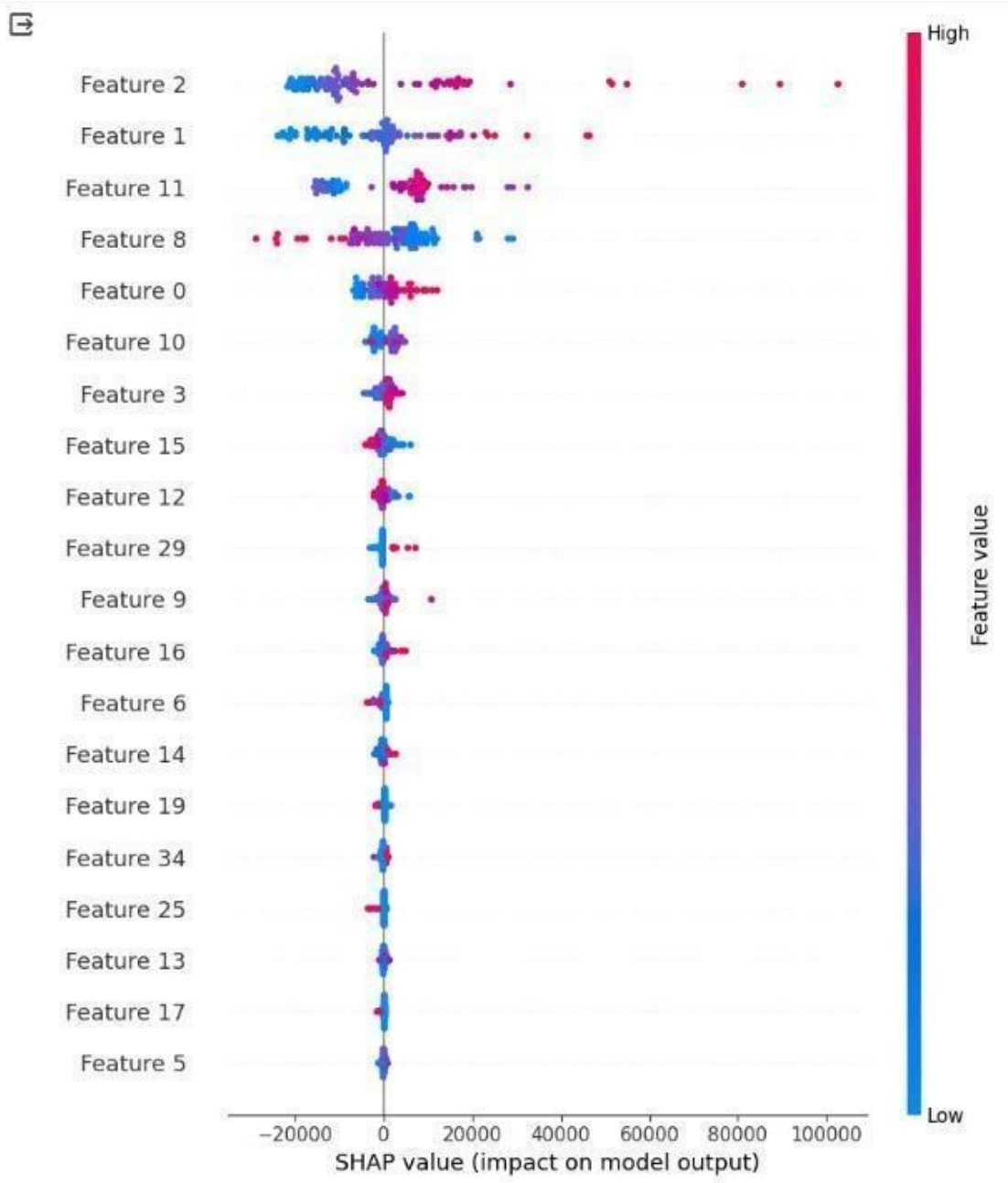

Figure 3.12: SHAP Results

4. DISCUSSION

The challenge of uncertain future outcomes remains a constant concern for businesses. Precisely forecasting upcoming metrics is pivotal for enterprises to efficiently oversee their operations, anticipate potential challenges, and enhance their profitability. The continuous advancements in technology empower companies to foresee future trends and results, facilitating informed decision-making and strategic adaptations to adeptly navigate uncertainties.

Technological advancements have revolutionized data collection and its diversity across various sectors, offering faster and more cost-effective methods compared to previous practices. This influx of rich and varied datasets enables accurate predictions. The realm of e-commerce, being greatly influenced by advancing technologies, has experienced significant advantages. Online retail allows companies to amass a vast array of diverse data, facilitating multifaceted problem-solving and optimization of operational processes.

One of the notable applications of this diverse dataset lies in inventory management, wherein companies achieve a balanced supply-demand equilibrium and efficiently manage their inventory levels. Moreover, they can track financial trends for effective financial planning, enabling astute management of investments and risk detection. Employing machine learning algorithms, companies can implement dynamic pricing strategies, tailoring prices based on customer behavior and seasonal fluctuations.

Utilizing diverse tools, companies can directly engage with customers via questionnaires or chatbots, discerning pain points in customer experiences and enhancing overall satisfaction. Additionally, these technologies bolster marketing efforts, enabling the segmentation of customers for targeted campaigns aimed at maximizing revenue.

Given the myriad of opportunities and components available, machine learning (ML) algorithms have played a pivotal role in leveraging these advancements. Within this landscape, forecasting stands out as a cornerstone, enabling predictions related to future

trends, sales trajectories, and demand projections. In the context of e-commerce, sales forecasting assumes paramount importance.

Sales forecasting represents a critical predictive measure for e-commerce entities, primarily due to its significant implications. It substantially aids in inventory management, determining the requisite volume of products to maintain adequate supply. Discrepancies between actual sales and target figures can trigger strategic marketing campaigns, allowing companies to realign their sales trajectory. Moreover, the ability to maintain optimal inventory levels contributes to heightened customer satisfaction, ensuring a consistent availability of desired products.

4.1. Discussion of models

In this research endeavor, sales forecasting within the e-commerce domain was executed through the implementation of six distinct models, namely ARIMA, SARIMA, Ridge Regression, Polynomial Regression, XGBoost, and LGBM. Within the ARIMA and SARIMA models, the focus lay on predicting forthcoming day-to-day sales patterns. These models were primarily reliant on discerning sales trends, inherently neglecting the incorporation of additional influencing factors that could potentially affect sales outcomes.

Upon analyzing the dataset using ARIMA and SARIMA models without the inclusion of additional variables, a comparative evaluation was conducted between the two. ARIMA demonstrated superior performance with 33.13% lower RMSE, 77.24% lower MSE, and 75.11% lower MAE in contrast to SARIMA. This discrepancy in results can be attributed to the nature of SARIMA, where the selection of parameters, in this case, 12 for monthly intervals, might not be optimally capturing the yearly trend influenced more significantly by season endings and holidays rather than monthly fluctuations.

However, upon comparative analysis with the remaining models, it became evident that both ARIMA and SARIMA models exhibited comparatively higher

prediction errors. This discrepancy underscores the necessity for a more comprehensive consideration of additional variables that exert influence on sales outcomes to attain enhanced accuracy and explanatory power in the forecasting models.

Incorporating a broader spectrum of influencing factors into the predictive models was crucial in enhancing the predictive capability of the alternative models. The selected variables were meticulously chosen, considering their potential impact on customer shopping behavior within the e-commerce domain. Recognizing that external determinants such as currency fluctuations, climatic conditions, and seasonal variations could significantly influence sales, these variables were thoughtfully incorporated.

Moreover, alongside these external variables, intrinsic factors including promotional campaigns, customer sessions, discount rates, and product prices were considered as essential contributors to sales performance. To further refine the models' predictive accuracy, the datasets underwent clustering of campaign variables, and interaction terms were introduced to capture complex relationships among the various predictors. This comprehensive approach aimed to create more nuanced and robust predictive frameworks capable of encapsulating multifaceted influences on sales outcomes.

Following the dataset development and preparation, a series of models were executed, and the outcomes indicated that LGBM exhibited the highest explainability along with the lowest error rate in contrast to the other models. Both LGBM and XGBoost are recognized as iterative algorithms that iteratively refine the model by minimizing error rates, thereby aiming to achieve optimal outcomes. Consequently, the anticipated superior performance of XGBoost and LGBM compared to Polynomial and Ridge Regression models aligns with their inherent iterative nature and focus on error reduction.

Table 4.1: Algorithms Comparison

MODELS	RMSE	MSE	MAE	MAPE	r2
RIDGE REGRESSION	18515	342792969	12428	17.18%	80.3%
POLYNOMIAL REGRESSION	18899	357168492	12245	16.84%	79.4%
XGBOOST	11918	142036826	8102	9.88%	91.8%
LGBM	9945	98899432	6869	8.52%	94.3%

Upon comparison between Ridge and Polynomial Regression models, Ridge Regression exhibited lower RMSE and MSE outcomes, albeit demonstrating higher errors in MAE and MAPE. Overall, Ridge Regression showcased enhanced explainability in comparison to Polynomial Regression, primarily due to its ability to account for multicollinearity among variables. The improved performance of Ridge Regression over Polynomial Regression is attributed to its capacity to address multicollinearity concerns effectively within the dataset.

In contrast to XGBoost, LGBM demonstrates superior performance for several reasons. LGBM benefits from the 'Gradient-based One-Side Sampling' (GOSS) or 'Exclusive Feature Bundling' (EFB) methods, which significantly enhance its performance with categorical data [63]. The dataset under consideration encompasses numerous categorical variables, thereby adversely impacting the efficacy of XGBoost but favoring LGBM due to its optimized handling of such data types. Furthermore, LGBM's implementation of vertically growing trees contributes to a reduction in loss and the potential development of deeper trees, further enhancing its modeling capacity.

4.2. Comparison with Earlier Studies

Numerous studies have explored sales prediction using diverse ML algorithms across various industries, such as e-commerce. Singh and her colleagues conducted a study on Amazon sales forecasts employing ARIMA models [64]. They specifically

applied both ARIMA and SARIMA methodologies, drawing comparisons between the two. Their findings suggested that SARIMA exhibited more favorable outcomes than ARIMA; however, they did not attain statistically significant results in terms of MAPE. Consequently, they concluded that additional variables should be considered to enhance sales forecasting accuracy.

In contrast, our study reveals ARIMA to yield superior results compared to SARIMA. Several factors might contribute to this discrepancy. Firstly, the dynamics of holidays and peak seasons could pose challenges, particularly as holiday schedules in Turkey are variable, making it difficult to capture seasonal effects accurately. Secondly, the post-Covid surge in online shopping trends may not be sufficiently reflected in the dataset, which spans only three years, potentially impeding the model's ability to capture this trend.

In an alternate study, a comparative analysis was conducted among three distinct models—LGBM, XGBoost, and Catboost—specifically for sales forecasting purposes [65]. Their findings revealed that Catboost exhibited superior accuracy and performance in comparison to the other models, while LGBM demonstrated the lowest R-squared value and accuracy. This variance in results might be attributed to variations in dataset size and composition of variables, thereby influencing the performance outcomes between these models. In a separate study focusing on forecasting demand within the fashion retail sector, various models including XGBoost, Gradient Boosted Regression Trees (GBRT), LGBM, Catboost, and Random Forest were employed. The study concluded that among these models, XGBoost exhibited the most favorable results and performance when applied to the specific dataset under analysis [66]. In a separate research endeavor focused on forecasting retail sales, a comparison was made among Ridge Regression, Polynomial Regression, Linear Regression, and XGBoost models. The analysis indicated that Ridge Regression exhibited superior performance compared to Polynomial Regression. Additionally, XGBoost outperformed all other algorithms by a substantial margin in predictive accuracy [67]. In a final study aimed at predicting consumer revenue, an evaluation was made among LGBM, XGBoost, Random Forest, and Gradient Boosting

models. The findings revealed that LGBM outperformed the other models in terms of predictive accuracy and effectiveness [68]. Research findings indicate that the performance of predictive models can be significantly influenced by various factors, such as diverse conditions, distinct variables, and the scale of the dataset.

4.3. Limitations of the Study

The study utilized a retail e-commerce dataset spanning three years, specifically covering the period from 2019 to 2021, a time frame marked by the advent of the COVID-19 pandemic and a subsequent surge in online shopping activity. A more extended dataset covering a duration beyond three years could potentially enhance the accuracy and predictability of the obtained results.

Moreover, the study exclusively focused on a restricted selection of six models categorized primarily by their use of independent variables, regression analysis, and iterative methodologies. While each model possesses its unique strengths and limitations, the direct comparison of all six models may not be entirely appropriate due to their divergent characteristics and functionalities. Nonetheless, it is imperative to delineate the strengths and weaknesses inherent in each of these models to provide a comprehensive understanding of their capabilities and limitations.

Moreover, certain variables may impact the independent variable in diverse ways. Nonetheless, including additional variables in the dataset could potentially enhance the predictability and explanatory power of the models.

4.4. Potential Contributions and Future Prospects

This research endeavors to transcend the confines of conventional scholarly inquiry within the realm of e-commerce dynamics. It does so by leveraging authentic transactional data spanning the tumultuous period from 2019 to 2021, coinciding with the onset of the COVID-19 pandemic. Rather than solely focusing on traditional sales forecasting methodologies, this study undertakes a pioneering exploration of campaign

variables that have been hitherto marginalized in academic discourse. By interrogating the efficacy and impact of these overlooked promotional initiatives, this research aims to deepen our understanding of their role in predicting future sales trends. Through this scholarly endeavor, we aspire to offer novel insights into the intricate interplay between promotional campaigns and predictive modeling within the ever-evolving landscape of e-commerce operations.

To enhance future studies, several strategies could be employed. Firstly, expanding the range of models utilized for comparison purposes would provide a more comprehensive understanding of which models perform optimally for sales prediction. Additionally, increasing the dataset size by incorporating more variables could improve the models' explanatory power and diminish multicollinearity issues. Augmenting the dataset's quality and quantity is imperative for yielding enhanced outcomes.

Moreover, in this study, campaign variables were manually clustered based on insights from industry professionals. For future investigations, employing clustering algorithms at the initial stages to categorize campaign variables using unsupervised learning techniques might yield more refined and data-driven clusters, potentially enhancing their impact as variables within the models. This approach could offer a more automated and data-centric approach to feature engineering, thereby potentially improving model performance.

CONCLUSION

This thesis encompasses the theoretical foundations of machine learning algorithms and emphasizes the evolution of technology in facilitating the growth and advantages of e-commerce. The principal objective of this study revolves around the prognostication of sales patterns through the utilization of an e-commerce dataset.

Forecasting serves as a pivotal tool for various facets within e-commerce enterprises. Its applications span across inventory management, facilitating optimized stock levels and efficient deliveries, financial planning through strategic asset management, dynamic pricing strategies, and enhancing customer satisfaction by streamlining delivery operations. Moreover, it plays a pivotal role in marketing activities, allowing customized campaigns and efficient budget management. The utilization of machine learning algorithms empowers these functionalities.

Sales prediction stands as a foundational step in this spectrum, significantly impacting various operational aspects of a company. This study focuses on predicting sales trends in the e-commerce sector, emphasizing the inclusion of campaign variables. Employing six distinct machine learning algorithms, it seeks to determine the most accurate and explainable model. The study found LGBM to be the most suitable algorithm. An intriguing aspect was the inclusion of campaign variables, considering the lack of previous studies exploring their impact on forecasting. Contrary to expectations, the SHAP analysis revealed a lesser impact of campaign variables on the model's explainability. The study acknowledges the potential to enhance model interpretability by utilizing clustering algorithms to represent variables effectively, as highlighted in the limitations section.

REFERENCES

[1] C. Tudor, “Integrated framework to assess the extent of the pandemic impact on the size and structure of the E-Commerce retail sales sector and forecast Retail Trade E-Commerce,” *Electronics*, vol. 11, no. 19, p. 3194, Oct. 2022, doi: 10.3390/electronics11193194.

[2] Y. Tian and C. Stewart, "History of E-Commerce," in Encyclopedia of E-Commerce, E-Government and M-Commerce, 2006.

[3] R. J. Zwanka and C. L. Buff, “COVID-19 Generation: A conceptual framework of the consumer behavioral shifts to be caused by the COVID-19 pandemic,” *Journal of International Consumer Marketing*, vol. 33, no. 1, pp. 58–67, May 2020, doi: 10.1080/08961530.2020.1771646.

[4] K. Singh, P. M. Booma, and U. Eaganathan, “E-Commerce system for sale prediction using machine learning technique,” *Journal of Physics*, vol. 1712, no. 1, p. 012042, Dec. 2020, doi: 10.1088/1742-6596/1712/1/012042.

[5] “Global retail e-commerce sales 2026 | Statista,” *Statista*, Sep. 21, 2022. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/#statisticContainer

[6] “Global e-commerce share of retail sales 2027 | Statista,” *Statista*, Aug. 29, 2023. https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/

[7] “E-Commerce Market Size, Growth, Revenue, Share, Report 2022-2028,” *ValuatesReports*. https://reports.valuates.com/market-reports/QYRE-Auto-6K6319/global-e-commerce

[8] “Home - Eurostat,” *Eurostat*. https://ec.europa.eu/eurostat

[[9] “Home - Eurostat,” *Eurostat*. https://ec.europa.eu/eurostat/statistics-explained

[10] “Topic: E-commerce in Turkey,” *Statista*, May 15, 2023. https://www.statista.com/topics/9411/e-commerce-in-turkey/#topicOverview

[11] M. Çevik, “Covıd-19 Sürecinde Türkiye’deki E-Ticaret Sitelerinin Ziyaretçi Sayılarının Yapay Sinir Ağları ile Tahmini,” *Turkish Studies*, vol. Volume 15 Issue 4, no. Volume 15 Issue 4, pp. 615–631, Jan. 2020, doi: 10.7827/turkishstudies.44299.

[12] D.-M. Petroşanu, A. Pîrjan, G. Căruţaşu, A. Tăbuşcă, D. Zirra, and A. Perju-Mitran, “E-Commerce sales revenues forecasting by means of dynamically designing, developing and validating a Directed Acyclic graph (DAG) network for deep learning,” *Electronics*, vol. 11, no. 18, p. 2940, Sep. 2022, doi: 10.3390/electronics11182940.

[13] H. Pan and H. Zhou, “Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce,” *Electronic Commerce Research*, vol. 20, no. 2, pp. 297–320, Apr. 2020, doi: 10.1007/s10660-020-09409-0.

[14] D.-M. Petroşanu, A. Pîrjan, G. Căruţaşu, A. Tăbuşcă, D. Zirra, and A. Perju-Mitran, “E-Commerce sales revenues forecasting by means of dynamically designing, developing and validating a Directed Acyclic graph (DAG) network for deep learning,” *Electronics*, vol. 11, no. 18, p. 2940, Sep. 2022, doi: 10.3390/electronics11182940.

[15] K.-W. Su, S. Chen, P. Lin, and C.-I. Hsieh, “Evaluating the user interface and experience of VR in the electronic commerce environment: a hybrid approach,” *Virtual Reality*, vol. 24, no. 2, pp. 241–254, Jul. 2019, doi: 10.1007/s10055-019-00394-w.

[16] N. Hua, S. Hight, W. Wei, A. B. Ozturk, X. R. Zhao, K. Nusair, and A. DeFranco, "The power of e-commerce: Does e-commerce enhance the impact of loyalty programs on hotel operating performance?," *International Journal of Contemporary Hospitality Management*, vol. 31, no. 4, pp. 1906-1923, 2019.

[17] M. Li, S. Ji, and G. Liu, “Forecasting of Chinese E-Commerce sales: An empirical comparison of ARIMA, nonlinear autoregressive neural network, and a combined ARIMA-NARNN model,” *Mathematical Problems in Engineering*, vol. 2018, pp. 1–12, Nov. 2018, doi: 10.1155/2018/6924960.

[18] C. Claycomb, K. N. S. Iyer, and R. Germain, “Predicting the level of B2B e-commerce in industrial organizations,” *Industrial Marketing Management*, vol. 34, no. 3, pp. 221–234, Apr. 2005, doi: 10.1016/j.indmarman.2004.01.009.

[19] L. SamuelA, “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, Jul. 1959, doi: 10.1147/rd.33.0210.

[20] S. Kotsiantis, “Supervised Machine Learning: A review of classification techniques,” *Informatica (Lithuanian Academy of Sciences)*, vol. 31, no. 3, pp. 249–268, Jan. 2007, [Online]. Available: https://dblp.uni-trier.de/db/journals/informaticaSI/informaticaSI31.html#Kotsiantis07

[21] S. Makridakis, R. M. Hogarth, and A. Gaba, “Why forecasts Fail. What to do instead,” *MIT Sloan Management Review*, vol. 51, no. 2, pp. 83–90, Dec. 2010, [Online]. Available: https://www.cimaglobal.com/Documents/Thought_leadership_docs/white-paper-hub/Why-Forecasts-Fail-What-to-do-Instead.pdf

[22] Y. Choi and H. Lee, “Data properties and the performance of sentiment classification for electronic commerce applications,” *Information Systems Frontiers*, vol. 19, no. 5, pp. 993–1012, Mar. 2017, doi: 10.1007/s10796-017-9741-7.

[23] Y. J. Lee, S.-J. Yang, and Z. Johnson, “Need for touch and two-way communication in e-commerce,” *Journal of Research in Interactive Marketing*, vol. 11, no. 4, pp. 341–360, Sep. 2017, doi: 10.1108/jrim-04-2016-0035.

[24] S. Ji, X. Wang, W. Zhao, and D. Guo, “An application of a Three-Stage XGBOOST-Based model to sales forecasting of a Cross-Border E-Commerce enterprise,” *Mathematical Problems in Engineering*, vol. 2019, pp. 1–15, Sep. 2019, doi: 10.1155/2019/8503252.

[25] K. Zhao and C. Wang, “Sales Forecast in E-commerce using Convolutional Neural Network,” *arXiv (Cornell University)*, Aug. 2017, doi: 10.48550/arxiv.1708.07946.

[26] K. Bandara, P. Shi, C. Bergmeir, H. Hewamalage, Q. Tran, and B. Seaman, “Sales demand forecast in e-commerce using a Long Short-Term Memory Neural Network methodology,” in *Lecture Notes in Computer Science*, 2019, pp. 462–474. doi: 10.1007/978-3-030-36718-3_39.

[27] M. Li, S. Ji, and G. Liu, “Forecasting of Chinese E-Commerce sales: An empirical comparison of ARIMA, nonlinear autoregressive neural network, and a combined ARIMA-NARNN model,” *Mathematical Problems in Engineering*, vol. 2018, pp. 1–12, Nov. 2018, doi: 10.1155/2018/6924960.

[28] B. M. Pavlyshenko, “Linear, machine learning and probabilistic approaches for time series analysis,” *IEEE First International Conference on Data Stream Mining & Processing*, Aug. 2016, doi: 10.1109/dsmp.2016.7583582.

[29] Y. Niu, “Walmart Sales Forecasting using XGBoost algorithm and Feature engineering,” *4th Asia-Pacific World Congress on Computer Science and Engineering*, Oct. 2020, doi: 10.1109/icbase51474.2020.00103.

[30] T. M. Le and S. Liaw, “Effects of pros and cons of applying big data analytics to consumers’ responses in an E-Commerce context,” *Sustainability*, vol. 9, no. 5, p. 798, May 2017, doi: 10.3390/su9050798.

[31] M. Bohanec, M. K. Borštnar, and M. Robnik–Šikonja, “Explaining machine learning models in sales predictions,” *Expert Systems With Applications*, vol. 71, pp. 416–428, Apr. 2017, doi: 10.1016/j.eswa.2016.11.010.

[32] M. Srinivas, G. Sucharitha, and A. Matta, *Machine Learning Algorithms and applications*. John Wiley & Sons, 2021.

[33] U. Andersson, Á. Cuervo-Cazurra, and B. B. Nielsen, “From the Editors: Explaining interaction effects within and across levels of analysis,” *Journal of International Business Studies*, vol. 45, no. 9, pp. 1063–1071, Nov. 2014, doi: 10.1057/jibs.2014.50.

[34] C. Ai and E. C. Norton, “Interaction terms in logit and probit models,” *Economics Letters*, vol. 80, no. 1, pp. 123–129, Jul. 2003, doi: 10.1016/s0165-1765(03)00032-6.

[35] P. Karaca-Mandic, E. C. Norton, and B. E. Dowd, “Interaction terms in nonlinear models,” *Health Services Research*, vol. 47, no. 1pt1, pp. 255–274, Aug. 2011, doi: 10.1111/j.1475-6773.2011.01314.x.

[36] S. Bagui, D. Nandi, S. Bagui, and R. J. White, “Machine Learning and Deep Learning for Phishing Email Classification using One-Hot Encoding,” *Journal of Computer Science*, vol. 17, no. 7, pp. 610–623, Jul. 2021, doi: 10.3844/jcssp.2021.610.623.

[37] M. Allahyari *et al.*, “A brief survey of text mining: Classification, clustering and extraction techniques,” *arXiv (Cornell University)*, Jul. 2017, [Online]. Available: http://export.arxiv.org/pdf/1707.02919

[38] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series,” *17th IEEE International Conference on Machine Learning and Applications*, Dec. 2018, doi: 10.1109/icmla.2018.00227.

[39] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and its applications*. 2011. doi: 10.1007/978-1-4419-7865-3.

[40] G. P. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model,” *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003, doi: 10.1016/s0925-2312(01)00702-0.

[41] E. Afrifa-Yamoah, “Application of ARIMA models in forecasting monthly average surface temperature of Brong Ahafo region of Ghana,” *International Journal of Statistics and Applications*, vol. 5, no. 5, pp. 237–246, Jan. 2015, [Online]. Available: https://www.researchgate.net/profile/Ebenezer_Afrifa-Yamoah/publication/282848214_Application_of_ARIMA_Models_in_Forecasting_Monthly_Average_Surface_temperature_of_Brong_Ahafo_Region_of_Ghana/links/561e2bd608aec7945a2541b6.pdf

[42] G. C. McDonald and R. C. Schwing, “Instabilities of regression estimates relating air pollution to mortality,” *Technometrics*, vol. 15, no. 3, pp. 463–481, Aug. 1973, doi: 10.1080/00401706.1973.10489073.

[43] G. C. McDonald, “Ridge regression,” *WIREs Computational Statistics*, vol. 1, no. 1, pp. 93–100, Jul. 2009, doi: 10.1002/wics.14.

[44] J. A. Stimson, E. G. Carmines, and R. A. Zeller, “Interpreting polynomial regression,” *Sociological Methods & Research*, vol. 6, no. 4, pp. 515–524, May 1978, doi: 10.1177/004912417800600405.

[45] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *Annals of Statistics*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[46] T. Chen and C. Guestrin, “XGBoost,” *In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, doi: 10.1145/2939672.2939785.

[47] A. Shehadeh, O. Alshboul, R. E. A. Mamlook, and O. Hamedat, “Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression,” *Automation in Construction*, vol. 129, p. 103827, Sep. 2021, doi: 10.1016/j.autcon.2021.103827.

[48] M. Schlögl, R. Stütz, G. Laaha, and M. Melcher, “A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset,” *Accident Analysis & Prevention*, vol. 127, pp. 134–149, Jun. 2019, doi: 10.1016/j.aap.2019.02.008.

[49] L. Shan, Z. Yang, H. Zhang, R. Shi, and L. Kuang, “Predicting duration of traffic accidents based on ensemble learning,” in *Lecture Notes in Computer Science*, 2019, pp. 252–266. doi: 10.1007/978-3-030-12981-1_18.

[50] G. Ke *et al.*, “LightGBM: a highly efficient gradient boosting decision tree,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 3149–3157, Dec. 2017.

[51] M. Gan, S. Pan, Y. Chen, C. Chen, H. Pan, and X. Zhu, “Application of the Machine Learning LightGBM model to the prediction of the water levels of the Lower Columbia River,” *Journal of Marine Science and Engineering*, vol. 9, no. 5, p. 496, May 2021, doi: 10.3390/jmse9050496.

[52] G. Ke *et al.*, “LightGBM: a highly efficient gradient boosting decision tree,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 3149–3157, Dec. 2017.

[53] S. Chi, S.-J. Suk, Y. Kang, and S. P. Mulva, “Development of a data mining-based analysis framework for multi-attribute construction project information,” *Advanced Engineering Informatics*, vol. 26, no. 3, pp. 574–581, Aug. 2012, doi: 10.1016/j.aei.2012.03.005.

[54] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.

[55] U. Batra, N. R. Roy, and B. Panda, *Data Science and Analytics: 5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, Gurugram, India, November 15–16, 2019, Revised Selected Papers, Part I*. Springer Nature, 2020.

[56] A. De Myttenaere, B. Golden, B. L. Grand, and F. Rossi, “Mean Absolute Percentage Error for regression models,” *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016, doi: 10.1016/j.neucom.2015.12.114.

[57] A. C. Cameron and F. Windmeijer, “An R-squared measure of goodness of fit for some common nonlinear regression models,” *Journal of Econometrics*, vol. 77, no. 2, pp. 329–342, Apr. 1997, doi: 10.1016/s0304-4076(96)01818-0.

[58] D. E. Farrar and R. R. Glauber, “Multicollinearity in Regression Analysis: The problem revisited,” *The Review of Economics and Statistics*, vol. 49, no. 1, p. 92, Feb. 1967, doi: 10.2307/1937887.

[59] J. I. Daoud, “Multicollinearity and regression analysis,” *Journal of Physics*, vol. 949, p. 012009, Dec. 2017, doi: 10.1088/1742-6596/949/1/012009.

[60] A. E. Roth, “Introduction to the Shapley value,” in *Cambridge University Press eBooks*, 1988, pp. 1–28. doi: 10.1017/cbo9780511528446.002.

[61] Y. Bi, D. Xiang, Z. Ge, F. Li, C. Jia, and J. Song, “An interpretable prediction model for identifying N7-Methylguanosine sites based on XGBOOST and SHAP,” *Molecular Therapy - Nucleic Acids*, vol. 22, pp. 362–372, Dec. 2020, doi: 10.1016/j.omtn.2020.08.022.

[62] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, “Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis,” *Accident Analysis & Prevention*, vol. 136, p. 105405, Mar. 2020, doi: 10.1016/j.aap.2019.105405.

[63] D. Zhang and Y. Gong, “The comparison of LightGBM and XGBOOST Coupling factor analysis and prediagnosis of acute liver failure,” *IEEE Access*, vol. 8, pp. 220990–221003, Jan. 2020, doi: 10.1109/access.2020.3042848.

[64] B. Singh, P. Kumar, N. Sharma, and K. P. Sharma, “Sales Forecast for Amazon Sales with Time Series Modeling,” *In 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, Jan. 2020, doi: 10.1109/icpc2t48082.2020.9071463.

[65] K. Alice, S. H. U. H. Andrabi, and S. Jha, “Sales Forecasting Based on Ensemble Learning,” *Authorea Preprint*, Aug. 2023, doi: 10.36227/techrxiv.24049452.

[66] P. K. Singh, Y. Gupta, N. Jha, and A. Rajan, “Fashion Retail: Forecasting demand for new items,” *arXiv (Cornell University)*, Jun. 2019, [Online]. Available:

https://arxiv.org/pdf/1907.01960

[67] P. Ranjitha and M. B. Spandana, “Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms,” *In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2021, doi: 10.1109/iciccs51141.2021.9432109.

[68] P. Ranjitha and M. B. Spandana, “Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms,” In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2021, doi: 10.1109/iciccs51141.2021.9432109.