ANALYSIS OF ONLINE MARKETPLACE SALES PREDICTING BASED ON MACHINE LEARNING ALGORITHMS: A CASE OF TURKISH E-COMMERCE SITE

A THESIS
SUBMITTED TO ABDULLAH GÜL UNIVERSITY
SOCIAL SCIENCES INSTITUTE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF (MASTER OF SCIENCE)

By
Ecem Kaya
06, 2023
Kayseri

Ecem Kaya
ANALYSIS OF ONLINE MARKETPLACE SALES PREDICTING BASED ON MACHINE LEARNING ALGORITHMS: A CASE OF TURKISH E-COMMERCE SITE
AGU
2023

ANALYSIS OF ONLINE MARKETPLACE SALES PREDICTING BASED ON MACHINE LEARNING ALGORITHMS: A CASE OF TURKISH E-COMMERCE SITE

A THESIS
SUBMITTED TO ABDULLAH GÜL UNIVERSITY
SOCIAL SCIENCES INSTITUTE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF (MASTER OF SCIENCE)

By
Ecem Kaya
06, 2023
Kayseri

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: Ecem Kaya
Signature :

REGULATORY COMPLIANCE

MSc. thesis titled Analysis of Online Marketplace Sales Prediction Based on Machine Learning Algorithms: A Case of Turkish E-commerce Site **has been prepared in accordance with the Graduate Thesis Preparation Guidelines of the Abdullah Gül University, Social Sciences Institute.**

Prepared By	Advisor
Ecem Kaya	Muhammed Sütçü

Head of the Data Science Program

Assoc. Prof. Umut Türk

ACCEPTANCE AND APPROVAL

MSc. thesis titled Analysis of Online Marketplace Sales Prediction Based on Machine Learning Algorithms: A Case of Turkish E-commerce Site and prepared by Ecem Kaya has been accepted by the jury in the Data Science Graduate Program at Abdullah Gül University, Social Sciences Institute.

09 /06 / 2023

JURY:

Advisor : Assoc. Prof. Muhammed Sütçü…….....

Member :Prof. Özgür Demirtaş…..…… .....

Member : Assoc. Prof. Umut Türk ……. .....

APPROVAL:

The acceptance of this MSc thesis has been approved by the decision of the Abdullah Gül University, Social Sciences Institute, Management Board dated ….. /….. / …….. and numbered ……..……. .

…….. /…….. /

……..

Director of Social Sciences Institute

Assoc. Prof. Umut Türk

Name Surname: Ecem Kaya
Program : Data Science (Master of Science)
Advisor : Assoc. Prof. Muhammed Sütçü
Thesis Title : Analysis of Online Marketplace Sales Prediction Based on Machine Learning Algorithms: A Case of Turkish E-commerce Site
Date and Place : June, 2023 – Kayseri, Turkey

ABSTRACT

Internet shopping has grown in popularity as more of our daily requirements have begun to be addressed online. Learning about the preferences and motivations of customers in the Turkish market and guiding e-commerce platforms to adapt their marketing strategies and increase customer satisfaction is important for both resource allocation and cost minimization. The purpose of this paper is to estimate future sales for popular e-commerce sites based on behavioral factors such as discounts, price or free shipping. Therefore, real-time and experiment-independent data are collected from the sales made by one of Turkey's most popular e-commerce sites. In order to produce predictions, we employ Artificial Neural Networks, Support Vector Regression, K-Nearest Neighbors Regressor, OLS regression, and Nu-Support Vector Regressor. The models developed using machine learning algorithms attempt to estimate the number of sales based on independent factors such as price, discount rate, and user ratings. As the result of this research, we calculate and compare the accuracy of the models with root mean squared errors and R^2.

Keywords: Big Data, Web Scraping, Online Marketplace, Sales Forecasting

Ad Soyad : Ecem Kaya
Anabilim Dalı, Program : Veri Bilimi (Yüksek Lisans)
Tez Danışmanı : Doç. Dr. Muhamed Sütçü
Tez Başlığı : Makine Öğrenimi Algoritmalarına Dayalı Çevrimiçi Pazar Yeri Satış Tahmininin Analizi: Türk E-ticaret Sitesi Örneği
Tarih ve Yer : Haziran, 2023 – Kayseri, Türkiye

ÖZET

Günlük ihtiyaçlarımızın çevrimiçi olarak karşılanmaya başlamasıyla birlikte internet alışverişinin popülaritesi de artmıştır. Türkiye pazarındaki müşterilerin tercihlerini ve motivasyonlarını öğrenmek ve e-ticaret platformlarına pazarlama stratejilerini uyarlamaları ve müşteri memnuniyetini artırmaları için rehberlik etmek hem kaynak temini hem de maliyet minimizasyonu açısından önemlidir. Bu makalenin amacı, popüler e-ticaret siteleri için indirim, fiyat veya ücretsiz kargo gibi davranışsal faktörlere dayalı olarak gelecekteki satışları tahmin etmektir. Bu nedenle, Türkiye'nin en popüler e-ticaret sitelerinden birinin yaptığı satışlardan gerçek zamanlı ve deneyden bağımsız veriler toplanmıştır. Tahmin üretmek için Yapay Sinir Ağları, Destek Vektör Regresyonu, K-En Yakın Komşular Regresyonu, OLS regresyonu ve Nu-Destek Vektör Regresörü kullanılmıştır. Makine öğrenimi algoritmaları kullanılarak geliştirilen modeller, fiyat, indirim oranı ve kullanıcı derecelendirmeleri gibi bağımsız faktörlere dayalı olarak satış sayısını tahmin etmeye çalışmaktadır. Bu araştırmanın sonucunda, modellerin doğruluğunu ortalama kare hatası ve R^2 ile hesaplıyor ve karşılaştırıyoruz.

Anahtar Kelimeler: Büyük Veri, Web Kazıma, Çevrimiçi Pazaryeri, Satış Tahmini

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

OLS	Ordinary Least Squares
ANN	Artificial Neural Network
KNN Regressor	K-Nearest Neighbors
SVR	Support Vector Regression
LSTM	Long-Short Term Memory
NU-SVR	Nu-Support Vector Regression
NLP	Natural Language Processing
MSE	Mean Squared Error
ReLU	Rectified Linear Unit
TUIK	Turkish Statistical Institute
TUBISAD	Turkish Informatics Industry Association

LIST OF TABLES

LIST OF FIGURES

LIST OF GRAPHS

1. INTRODUCTION

In recent years, the growth of e-commerce has been exponential, especially in emerging markets like Turkey. Online marketplaces have become increasingly popular, providing a platform for small and medium-sized businesses to reach a wider audience and sell their products locally or globally. The rate of individuals in the 16-74 age group ordering or purchasing goods or services over the Internet is announced as 44.3% in 2021 and in this sense, the volume of the e-retail market in Turkey is growing day by day (TUIK,2021).

In addition, according to the report prepared by the IT industry association TUBISAD and Deloitte, the volume of the e-retail sector in Turkey reached 161 billion TL in the first 6 months of 2021, growing by 49% compared to the same period of the previous year. (Deloitte Turkey, 2022). More vendors have started to sell on online marketplaces as well, thanks to the ease of entry and the large market potential. Therefore, to figure out consumers behavior and its determinants has become a critical challenge for e-commerce sector in sense of predicting future sales.

From the traditional marketing perspective, researchers have grouped consumer's buying behavior by consumer involvement, and differences between brands (Asseal,1984). On the other hand, new research shows that the spread of online markets leads to changes in the behavior of consumers. The ability of consumers to compare different products in the same class in terms of features and price, common distribution services, and return processes made easier by most service providers makes online shopping preferable (Huo,2021). There is a lack of literature that observes consumer behaviors toward the online marketplace in Turkey.

Machine learning algorithms shows great potential in overcoming the challenge of forecasting sales by providing accurate sales forecasts. In this thesis, we aim to analyze the performance of different machine learning algorithms in predicting online marketplace sales for a Turkish e-commerce site. The study will be based on data from a particular basket of products scraped from the site and will include preprocessing, feature selection and model training.

The main objective of this research is to investigate the accuracy of machine learning algorithms in predicting online marketplace sales and to identify the most effective algorithm for this purpose. In addition, we will explore the most significant features that contribute to sales prediction and their impact on the accuracy of the models.

This study has significant practical implications for online marketplace sellers and marketers as accurate sales predictions can help them make informed decisions regarding pricing, inventory management, and marketing strategies. Especially in small and medium-sized enterprises, since it will create positive outputs on stock management and cost, it has an impact on the profitability of the enterprises. Additionally, this work adds to the growing research on machine learning in the e-commerce industry, providing insights for future research in this area.

Identifying features correctly when shopping in the online market is the first stage of an accurate demand forecast. The previous rating both for products and retailers, product prices, promotions, or discounts that users have created affect the purchasing decisions of users (Chong, 2016). For example, on gittigidiyor.com that one of the most popular e-commerce sites in Turkey, users vote for products and vote for sellers, and there have been studies that have shown that you have influenced the customer's final decision on the online market**.**

To achieve the research objectives, we use a combination of quantitative and qualitative research methods. Firstly, we will conduct a thorough review of the literature on machine learning algorithms for sales prediction in e-commerce. This review helps to identify the most commonly used algorithms in this field and their strengths and weaknesses. For example, previous research predict product demand in a different concept. Kumar et al. offer a forecasting system for manufacturers to explore modern market trends, the season of the product, and the impact of the prediction on size with historical data.

Unlike previous articles, we predict demand based on consumers' spontaneous decisions with real-time data gathered from gittigidiyor.com. Hence, data which is

collected by this research will be real-time data, and data extraction techniques will be employed. The data extraction techniques are a set of technologies that provide to collect data from offline and online platforms and write them up to a database environment. For this part of the project, Beautifulsoap, a web scrapping library in Python, will be employed. This library will provide to take data from the gittigidiyor.com website and to be transferred it to the database thanks to the developed layer.

After that, we will use this data to train and test different machine learning algorithms, including Artificial neural networks, Support vector regression, Nu-Support Vector Regression, KNN regressor, and OLS. We will also evaluate the performance of these algorithms using metrics such as mean absolute error.

Ultimately, we'll evaluate the findings of our investigation by determining the most reliable method and crucial elements for sales forecasting. We will also emphasize the possible advantages and restrictions of utilizing machine learning algorithms for sales prediction in our discussion of the significance of our findings for online marketplace vendors and marketers.

Overall, with a focus on the Turkish market, this research will offer useful insights on the usage of machine learning algorithms for sales prediction in the e-commerce business. The study will increase our understanding in this area by demonstrating the potential of machine learning to enhance sales forecasting and support corporate decision-making.

1.1 . Motivation

The emergence of internet markets has completely changed how customers shop, creating both new opportunities and difficulties for firms. Investigating the factors that increase sales in this dynamic environment is urgently needed given the Turkish e-commerce sector's explosive rise. Understanding the motivations and interests of their target market is crucial for e-commerce businesses in order to

maximize their sales strategies. Therefore, this thesis aims to respond to the following question:

- How are online marketplace sales predicted reliably using machine learning algorithms?

This thesis aims to provide useful implications for Turkish e-commerce companies by answering these questions.

1.2. Contribution

- Machine Learning Algorithms for Sales Prediction:

Another significant contribution of this research is utilization of machine learning algorithms while predicting online market sales. By exploiting advance techniques such as artificial neural networks, k-nearest neighbor regressor, we aim to develop accurate and reliable prediction models. These models assist e-commerce business to predicts sales based on determinants which identify by consumer behaviors. On the other hand, models ensure that e-commerce business minimize their costs by data-driven decision making and optimizing their marketing promotions.

- Practical Implications for E-commerce business:

The results of this thesis have useful implications for Turkish businesses active in the e-commerce market. Identifying the determinants that affect customer behavior leads businesses to reach the right marketing strategy, launch campaigns for their targeted audience, and ultimately increase customer satisfaction by determining the optimum price. With these research findings, it has a positive impact on the decision-making processes and foresight capabilities of e-commerce platforms. These foresight capabilities enable Turkish e-commerce sites to reach the potential to increase their overall performance and competitiveness.

In conclusion, the primary aim of this study is to determine the determinants of customer behavior in the example of the e-commerce market in Turkey and to contribute to the existing literature by comparing sales forecasting algorithms using machine learning algorithms. The findings and practical implications of this research can guide e-commerce businesses to optimize their sales strategies and improve their competitiveness in the rapidly evolving online marketplace.

1.3. Applications

In this paper, we calculate the mutual information of each pair of variables to select the features we use in the models. Based on these calculations, we choose the feature we use in the prediction models. In this research, we compare five different models such as neural network, ols, etc. According to our results, nearest neighbor regression is the best forecasting algorithm that can make the best prediction on e-commerce data in Turkey.

The rest parts of the articles are organized as follows: In the second part of the article, we discuss the critical definitions of previous articles. In the third part, we present the implemented methodologies. In the fourth part, we demonstrate our empirical results in different models and discuss the distinct, and at the end of the article, we conclude our article.

2. LITERATURE REVIEW

In this section, the thoughts that the researchers add to the literature will be discussed in two parts. The first part will examine networks analysis, In second part the second part will examine consumer behaviors and, the relationship between e-commerce and consumer behaviors.

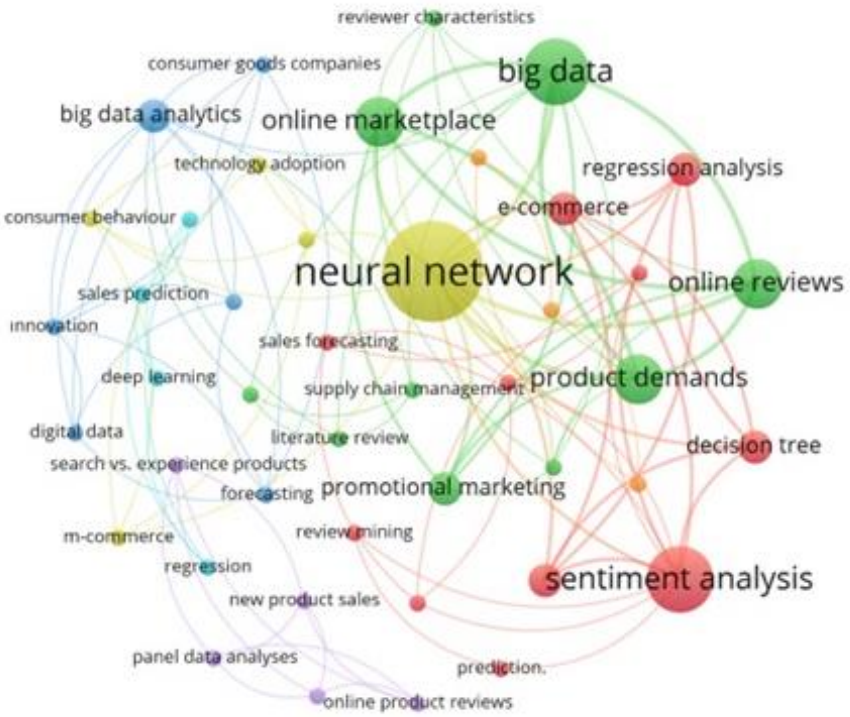
2.1. Network Analysis

In this part of this research, network analysis, which includes co-occurrence and co-author analysis, will be executed.

2.1.1 Co-occurrence Analysis

Co-occurrence is a type of analysis that examines the relationship between the keywords used by the articles. In this part, I investigate the main topic based on the keywords. According to the result, there are forty-five keywords in articles. The curves represent the linkage between the keywords. For example, “neural network” is linked with “big data”, it demonstrates that most of the article uses neural networks also use big data. Due to in fact that the most preferred technique is the neural network in this area the most used keyword is “neural network”. The other most used keywords are “big data”, “sentiment analysis”, “product demand” and “online marketplace”. These results correspond to the current literature. In Figure 1, the network between keywords is visualized with “VOSviewer” software.

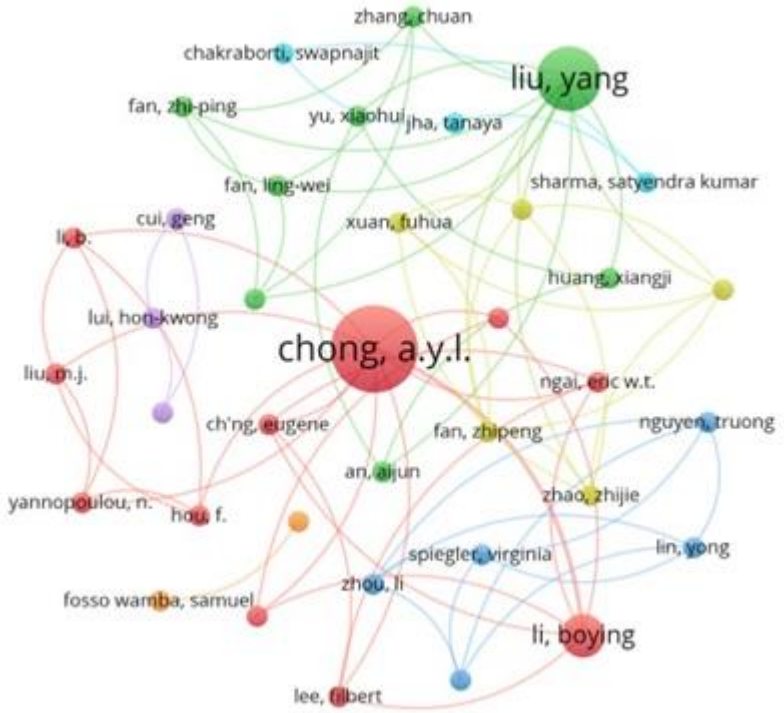
2.1.1.1. Co-occurrence Analysis



2.1.2 Co-author Analysis

Co-author is a type of bibliometric analysis that interprets the linkage between authors. In this technique, units of analysis are the authors. Even an author writes more than one article, in the analysis, they represent with one dot. However, the size of the represent how they are cited by other academicians. Like co-occurrence analysis, the curves imply the linkage between authors. There 33 authors and 5 clusters in the result set. According to the results of the analysis, the leading academician in this research field is Chong, A.Y.L. The other most-cited authors are Yang Liu and Boying Li. In Figure 2, co-author analysis is visualized with “Vosviewer”. In the following part of the research, the most cited author’s articles are reviewed.

2.1.2.1. Co-author Analysis



2.2. Literature Review

In this part of study, we discuss the previous literature. In first part, we examine the conceptual literature and in second part, we dispute the concept related with e-commerce.

Pricing is one of the crucial judgment factors for individuals' economic decision-making processes. For instance, Takemura (2019) explained the decision stages of purchasing in the sense of price which can affect people in the context of framing. People tend to evaluate the quality of goods from the perspective of price. Even brands that are evaluated to be of high quality are not bought if judged to be expensive, and luxury brands can be purchased if they are considered cheap. First, pricing knowledge serves as a clue to the quality level of goods. Second, it serves as a clue to the cheapness of the brand. Based on these clues, we obtain a "price response", which is a comprehensive assessment of price by consumers.

In addition to this, Lee and Chen-Yu (2018), conduct an online experiment that includes apparel products with four different levels of price promotion (%10, %30, %50, and %70) to show the impacts of promotions on consumers' behaviors. Consumers perceive the product that is lower price promotion as higher quality when the model conducted examines the price promotion effects on perceived quality directly. Even though the impacts are expected affirmative, high discount rates caused a negative frame and negatively affected consumer behavior. The limitation of this research is ignoring the economic gains of individuals and examining the effect of this gaining their consumption behavior.

Literature has a different perspective on consumer behaviors in circumstances of discounts. Such as Kristofferson et al. (2017) illustrate the mechanism of price promotion and consumer behavior in the context of scarcity promotion. Researchers claim that the individuals who have been exposed to that kind of promotion, tend to perform more aggressively these consumers have triggered the idea that other customers may be a competitive threat. Because the perception of scarcity forces them to avoid loss and encourages them to consume by creating negative framing. The

research obtains its data from the experiment to examine the level of aggressiveness of consumers on the occasion of interacting with these promotions. The researchers suggested alternative ways of shopping such as online.

Helmi et al. (2020), examine the impact of promotions but focused specifically on consumers' price differential behavior in online shopping. This research offers participants two different prices, regular and discounted, from online and offline shopping. An additively 6-question questionnaire is applied to see whether the participants notice the differences in these prices or not. They found that consumers are more efficient in understanding the price difference in online shopping than in face-to-face shopping. More importantly, this study proves that the price reduction in online shopping is a manipulative move to increase buying intentions. However, the main limitation of this research is to ignore the other criteria for price differentiation such as seasonality.

By drawing the concept of framing effect on prices, Chen et al. (2019) evaluate the impact of price reductions on customer behavior in the context of the singularity or multiplicity of price promotions. In this study, researchers explore the effects of price promotion framing messages on perceived value and online consumer purchase intention. For this purpose, they conduct an online experiment that assigned participants two different types of products which are low-price and high-price with the same amount of discount rate. Researchers state that participants respond differently to the same amount of discount ratio. Since the amount of monetary value that belongs to a high-price product is higher than the monetary amount of a low-price product. The perception of 'more profit' by the participants will create an illusion for them to prefer higher-priced products.

2.2.1. E-Commerce

The rising volume of e-commerce, especially in the last decade, has motivated researchers to study customers' online behavior. Shao and Yao (2018) aim to transform big data of customer behavior into meaningful wholes in real-time. The difference between this paper from the existing literature is that it provides continuity in real-time analysis through the Hadoop platform. In this respect, this study enables e-commerce initiatives to better understand their customers and deliver the services and products that their customers need.

Chong et al. (2017) investigates the contributions of product promotions and online reviews as determinants of consumer product demand. In the first step of the article, authors extract their data from Amazon.com and they try to estimate whether online review variables such as the value and volume of reviews, the number of positive and negative reviews, and online promotional marketing variables such as discounts and free deliveries affect electronic product demand in the Amazon.com. To make this assessment, they created a big data architecture that scrapes data from asynchronous I/O calls and writes them out into a database. This study uses big data architecture, sentiment analysis, and neural network modeling to investigate predictors of product sales in e-commerce. The result of this article demonstrates that both product promotions and online reviews are important predictors of product demand. This article provides better understanding of retailers or sellers to predict product demand according to online reviews and product promotions. The limitation of this study is that it includes only Amazon.com data and focuses only on electronic product sets.

Hou et al. (2017), examine the effect of online reviews and online reviewer characteristics on supply chain management. They state that figuring out determinants that influence online sales is crucial to managing supply chain management. They scrape their data from Amazon.com and perform sentiment analysis. Sentiment analysis is one of the text-mining methods and provides to identify emotions from texts. In this article, emotions are measured with 3 different polarities. If the text includes positive expressions in general, the result of the study would be 1, in negatives

expression would be -1 and the neutral expressions would be 0. After this application, the result of this analysis is included in the neural network as a control variable and neural network analysis tries to estimate the sales rank of sellers.

On the other hand, Sharma et al. (2019) investigate the performance of various modeling techniques with data from Amazon book sales. By analyzing reviews through sentiment analysis, the researchers assign a polarized sentiment score to reviews and thus investigate the impact of reviews on users' buying behavior. Suggesting that sentiment analysis is an important predictor in both linear and machine learning models, Sharma et al. prove that discount rate is a more important predictor than discount amount.

Chong (2013) conducted an online survey to understand the predictors of mobile commerce adoption, collecting data from 140 Chinese users. Neural network analysis is used to predict m-commerce adoption and the model is compared with the results of regression analysis. The neural network model outperformed the regression model in predicting adoption and captured non-linear relationships between predictors such as perceived value, trust, perceived enjoyment, personal innovativeness, users' demographic profiles effort expectancy. Multiple regression is used to compare results from both neural network and regression analysis to check whether there are differences in the predictors of M-commerce adoption. The neural network model outperformed the regression model and id able to show the importance of all predictors not identified by the regression model.

Contrary to behavioral variables, Huo (2021), which builds the regression model on time series, establishes two linear models, three machine learning models and two deep learning models in order to predict the sales volume of e-commerce sites. Contrary to the existing literature, Huo reveals that there is no obvious performance difference from the linear model in terms of performance of deep learning and machine learning models for the dataset he is working with.

In addition to all these studies, Zhao et al. (2019) analyze the purchasing decision of customers by collecting data on Toaboa.com, one of China's most popular online popular sites. Although this research also takes basic data such as price and discount

as determinants, its main focus is on guarantees such as refunds and product replacement provided by retailers and consumer reviews about products and retailers. In this study, artificial neural network models are used as in the previous literature. This research findings show that guarantees like refunds and product replacement provide consumers with a more comfortable and safer platform to shop online, because the products that consumers receive may sometimes differ from those displayed on the website. Since these guarantees reduce this risk, it increases the rate of consumption. Additionally, this research finds online reviews made by other consumers using products affect the purchasing decision of the customers.

Jothi et. al (2023) also investigate the sales prediction with Machine learning techniques in historical data. This study conducts the different methodologies: Support Vector Machine, Seasonal Autoregressive Integrated Moving Average with Exogenous Variables, and Multi-layer Perceptron. Results monitors that the most effective model is multi-layer perception for the historical sales datasets.

Zhang (2020) focuses on analyzing the short-term impact of e-commerce activities by determining the impact on sales to provide business insights. In this study uses data mining and analysis techniques to explore and quantify the impact of e-commerce promotion activities on sales volume. The main approach adopted in this study is to use data mining models such as Support Vector Regression to identify and predict causal relationships. Furthermore, it is compared with the classical multiple linear regression method. The findings show that, based on the current data set, the SVR model with the radial basis kernel exhibits the best prediction effect, with an error rate of 7%, and the generalization performance of the model is more robust than any other.

Zhang et al. (2021) examines the impact of customer reviews on demand for flash sales programs, using VIPS's flash sales program as an example. The literature on flash sales e-commerce model mainly focuses on research from the perspective of buyers, sellers and platforms. However, there is limited research on the impact of consumer reviews on consumer decision-making on flash marketing programs. The aim of this study is to use customer survey data from VIPS and apply sensitivity analysis techniques to improve demand forecasting in flash sales and e-commerce

models. Since this study contains historical data, it was carried out using an autoregressive model. At the end of the study, the authors showed that the number of comments affects the product demand at the right rate.

Mu (2019) aims to develop a big data-based prediction model for consumer purchase decisions on cross-border e-commerce platforms. With products sold on Tmall Global, it examines the influence of factors such as economy, culture, personal preferences, discounts, product categories, prices, product quality, shop ratings, online shopping convenience, and the number of comments on consumer purchase decisions. Regression analysis results reveal significant positive correlations between factors such as personal preferences, shop ratings, and the number of comments with pictures, and the sales volume of food products. A prediction model is created using the Multi-Layer Perceptron and association rules, successfully predicting purchase decisions for six different categories of food products. The study highlights the potential of big data analysis and understanding influencing factors in predicting consumer behavior and purchase decisions on cross-border e-commerce platforms.

Using deep learning techniques on a huge multidimensional data sample of web sessions, Chaudhuri et al. (2021) want to improve our understanding of online client purchase behavior for an e-commerce platform. As important predictors of online purchases by retail customers, this study used two independent sets of data, namely platform engagement and customer attributes. This article also compares the predictive capability of deep learning with other frequently used machine learning approaches for prediction, such as Decision Tree, Random Forest, Support Vector Machines, and Artificial Neural Networks. When applied to the same dataset, they discovered that deep learning outperformed machine learning techniques.

Bakir et al. (2018) aim to explore the applicability of LSTM neural networks in estimating time series data on phone prices in an e-commerce platform. The study focuses on a specific product, the Samsung Galaxy S7 (32GB), and uses daily price data from a specific marketplace amazon.fr. In this study, data obtained from Amazon.fr were analyzed using SVR and LSTM models. The results show that the multivariate LSTM model outperforms the multivariate SVR model in terms of

accuracy. The authors plan to further test this model on other market data, aiming to move towards real-time price prediction.

Gopalakrishnan et al. (2018) focus to analyze the sales of a large superstore and predict its future sales. The Linear Regression model is used by using historical sales data of a superstore. This study showed that linear regression using a historical data model can provide high accuracy. This methodology provides a comprehensive analysis of sales data and provides valuable information to improve business managers and increase product sales.

Raizada et al (2021) examine a study in which various machine learning algorithms were used to predict Walmart store sales. In the study, sales of various Walmart products were estimated using Linear Regression, Random Forest, KNN Regression, SVR and Extra Tree Regression algorithms. According to the results, it was seen that the Extra Tree Regression Model performed better than other supervised machine learning techniques and was able to predict the sales of the Walmart store with 98% accuracy. This study offers retail store owners a perspective on using methods like Extra Tree Regression or Random Forest Approach to forecast their sales, rather than analyzing different supervised machine learning algorithms.

3. METHODOLOGY

3.1. Data Collection

Data for this thesis is collected as suggested in previous literature. The data is collected based on specific criteria that emphasize products with a long shelf life and fast-consuming products that do not require fashion appropriateness, such as clothes. In this study, product selection is based mainly on small household appliances and personal care products.

For data collection purposes, a multi-step approach is adopted to ensure the inclusion of relevant products within the set criteria. The first step involved product selection, where products directly related to fashion or clothing are excluded. Instead, the focus is on small household appliances commonly used in households for a variety of purposes, such as kitchen appliances, cleaning devices and personal care tools.

To ensure affordability for the target consumers, the price range of the selected products is limited to 0-750 TL (Turkish Lira). This range is chosen in line with the budget constraints of a typical household and allows for a diverse and accessible set of products for analysis.

The collected data is subjected to a rigorous filtering process to include only small household appliances that fall within the defined price range and meet the criteria of non-fashionability, long shelf life and fast consumption. Descriptions and other available product information are utilized to verify the relevance of the products for the purpose of the thesis.

Data are collected during April and November of 2021 and final dataset consists of 11074 small home appliances that meet the specified criteria. Each sample in the dataset contains variables such as product name, price, discount amount and

percentage. The dataset represents a wide range of small household appliances in terms of functionality, purpose and price and provides a comprehensive basis for analysis and evaluation.

The data collection phase involved building a web crawler to extract data from gittigidiyor.com, a popular e-commerce website in Turkey. The crawler is developed using Python and the Beautiful Soup web scraping framework. The crawler is designed to scrape product information such as product names, prices, ratings, and reviews. The crawling process starts by sending HTTP requests to the gittigidiyor.com website using the requests library in Python. The HTML response receives from the website is then parsed using the Beautiful Soup library to extract product information from each product page.

The spider is configured to follow pagination links to scrape data from multiple pages of the website. We save the data to MongoDB to analyze data after the extraction process. The data collection period proceeds for about 7 months and is processable in extracting approximately 10,000 product records. The data is stored in a MongoDB database using a schema that is designed to capture all relevant product information.

Several data cleaning and validation methods carry out to guarantee the accuracy of the data. Duplicate records are eliminated, and incorrect or unnecessary material is discarded. The final dataset's quality, consistency, and completeness are examined in great detail.

In summary, the data collection phase involved developing a web crawler using Beautiful Soup to extract data from gittigidiyor.com and storing the data in a MongoDB database. The resulting dataset is carefully cleaned and validated to ensure its quality for further analysis.

3.1.1. Descriptive Statistics

	Product Price	Discounted Price	Discount Percentage	Free Shipping	Stocked Unit	Sold Unit	Number of Review	Product Ranking	Seller Raking	Discount Amount
count	11074	11074	11074	11074	11074	11074	11074	11074	11074	11074
mean	372,65	141,49	8,34	0,93	24,73	3,53	25,00	3,98	94,76	231,16
std	151,85	181,30	13,16	0,25	21,33	4,39	41,05	1,34	9,78	174,82
min	98,30	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	4,39
25%	250,00	0,00	0,00	1,00	5,00	0,00	3,00	4,00	94,00	71,00
50%	369,00	0,00	0,00	1,00	15,00	2,00	9,00	4,50	97,00	182,90
75%	469,75	282,17	17,00	1,00	50,00	5,00	29,00	4,70	99,00	368,90
max	709,08	658,80	46,00	1,00	50,00	20,00	260,00	5,00	100,00	705,60

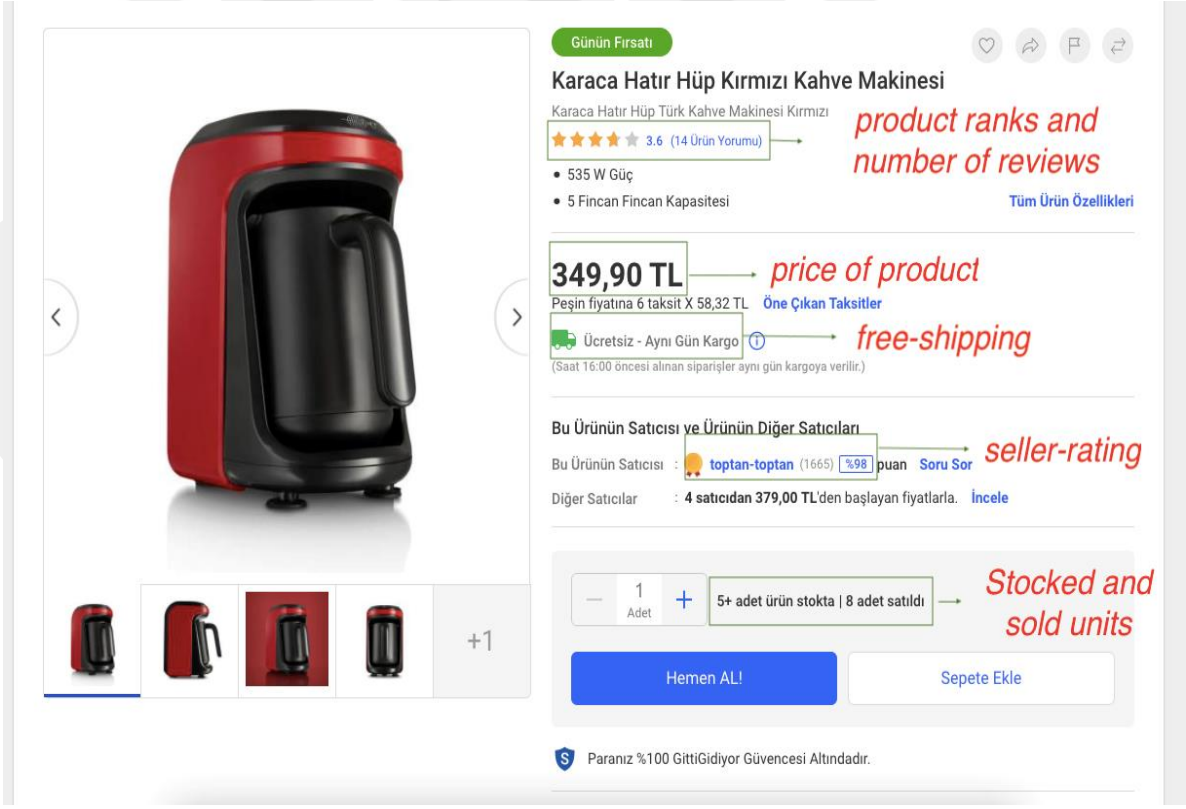
3.2. Methods

In this article, the relationship between consumer demand and other independent variables are identified and sales forecasting will be made in line with this relationship. To do this forecasting, OLS, KNN regressor, SVR, Nu-SVR and ANN regressions will be utilized. Performance evaluation of the models will be made according to MSE and R-squared results.

$Yj = \beta iXi+\varepsilon$ where I = 1,2,…,n and j = 1,2, …, J

where Yj refers consumer demand which is calculated by differentiation in sales quantity. X1 refers to product ranking, which is calculated from customer votes for the product, X2 refers to price of product and X3 refers to other control variables which are number of reviews, the rate of negative reviews, free shipping, seller rating and stocked unit. In figure1, variables can be seen in the forms the website uses.

3.2.1. Website View



3.2.2. Ordinary Least Squares

Definition: Ordinary Least Squares (OLS) is a statistical method used for estimating the parameters in linear regression models. It aims to find the line that minimizes the sum of squared differences between the observed target variable values and the predicted values.

Working Principle: The working principle of OLS involves fitting a linear regression model to the data by minimizing the sum of squared residuals. Here's a step-by-step explanation of the process:

Data Representation: OLS requires a dataset with input features and their corresponding target values.

Model Representation: Define the linear regression model as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta p Xp + \varepsilon$$

Y represents the target variable,

X_1, X_2, ..., Xp represent the input features,

β_0, β_1, β_2, ..., βp are the coefficients to be estimated,

ε represents the error term.

Residual Calculation: For each data point, calculate the residual as the difference between the observed target variable value and the predicted value from the linear regression model.

Objective Function: The objective is to minimize the sum of squared residuals. The objective function to be minimized is:

$$\text{Minimize: } \Sigma (Y - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta p Xp))^2$$

Optimization: To find the optimal values of the coefficients (β_0, β_1, β_2, ..., βp), differentiate the objective function with respect to each coefficient and set the derivatives to zero. This results in a system of linear equations known as the normal equations.

Solving the Normal Equations: Solve the normal equations to obtain the estimated coefficients. The solution can be obtained using matrix algebra techniques.

Prediction: Once the coefficients are estimated, use the linear regression model to make predictions on new, unseen data points by substituting the input feature values into the model equation.

Usage: Ordinary Least Squares (OLS) is commonly used in various domains, including:

Economics: OLS is widely used in econometrics to analyze and estimate economic relationships and models.

Finance: OLS is used to analyze the relationship between financial variables, such as stock prices, interest rates, and economic indicators.

Social Sciences: OLS is employed in social science research to analyze and understand the relationships between variables, such as education and income, crime rates, and demographic factors.

Marketing and Market Research: OLS is used to estimate the impact of marketing campaigns, pricing strategies, and customer behavior on sales and revenue.

How to Use OLS:

To use Ordinary Least Squares (OLS) effectively, follow these general steps:

Data Preparation: Prepare a dataset with input features and their corresponding target values for regression.

Model Specification: Determine the appropriate variables to include in the linear regression model based on domain knowledge and data exploration.

Model Training: Use the training dataset to estimate the coefficients of the linear regression model using the OLS method. This involves minimizing the sum of squared residuals.

Model Evaluation: Assess the performance of the OLS model using appropriate evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), or coefficient of determination (R-squared).

Assumptions Checking: Check the assumptions of OLS, such as linearity, independence of errors, constant variance (homoscedasticity), and absence of multicollinearity. Residual analysis and diagnostic tests can be used for this purpose.

Inference: Perform statistical inference on the estimated coefficients to determine their significance and assess the overall goodness of fit of the model.

Prediction: Once the model is trained and evaluated, use it to make predictions on new, unseen data points by substituting the input feature values into the model equation.

In conclusion, Ordinary Least Squares (OLS) is a statistical method used for estimating the parameters in linear regression models. It minimizes the sum of squared differences between the observed target variable values and the predicted values. OLS is widely used in economics, finance, social sciences, and market research for analyzing relationships between variables and making predictions.

3.2.3. K-Nearest Neighbor Regressor

Definition: K-Nearest Neighbors (KNN) Regressor is a non-parametric algorithm used for regression tasks. It predicts the value of a target variable by considering the average or weighted average of the values of its k nearest neighbors in the feature space.

Working Principle: The working principle of the KNN Regressor involves finding the k nearest neighbors of a given data point in the feature space and using their values to predict the target variable. Here's a step-by-step explanation of the process:

Data Representation: KNN Regressor requires a dataset with input features and their corresponding target values.

Distance Calculation: For each data point in the dataset, the distance between the input features of the target data point and the other data points is calculated. Common distance metrics include Euclidean distance, Manhattan distance, or Minkowski distance.

Neighbor Selection: The k nearest neighbors of the target data point are selected based on the calculated distances. These neighbors are the k data points with the shortest distances to the target data point.

Regression: The predicted value for the target data point is obtained by taking the average or weighted average of the target variable values of its k nearest neighbors. The weights can be assigned based on the inverse of the distances or other factors.

Prediction: Once the model is trained using the dataset, the KNN Regressor can make predictions for new, unseen data points by finding the k nearest neighbors and averaging their target variable values.

Usage: K-Nearest Neighbors (KNN) Regressor can be used in various domains, including: Housing Price Prediction: KNN Regressor can predict the price of a house based on its features, such as location, size, and number of rooms.

Stock Market Prediction: KNN Regressor can forecast stock prices based on historical market data and relevant indicators.

Demand Forecasting: KNN Regressor can be used to predict the demand for a product or service based on historical sales data and other factors.

Energy Consumption Prediction: KNN Regressor can estimate energy consumption based on historical usage patterns and environmental factors.

Anomaly Detection: KNN Regressor can identify anomalous data points by comparing their values with those of their nearest neighbors.

How to Use KNN Regressor: To use K-Nearest Neighbors (KNN) Regressor effectively, follow these general steps:

Data Preparation: Prepare a dataset with input features and their corresponding target values for regression.

Feature Scaling: Scale the input features to a common range, such as [0, 1] or [-1, 1], to ensure that they have a similar impact on the KNN Regressor model. This step is crucial as KNN is sensitive to the scales of features.

Choosing the Value of k: Determine an appropriate value of k, the number of nearest neighbors to consider. This value should be selected based on the characteristics of the data and the problem at hand. Typically, k is an odd number to avoid ties.

Model Training: Use the training dataset to train the KNN Regressor model. This involves storing the input feature vectors and their corresponding target variable values.

Prediction: Once the model is trained, use it to make predictions on new, unseen data points. For each prediction, find the k nearest neighbors in the feature space and use their target variable values to compute the predicted value, either by averaging or weighted averaging.

Model Evaluation: Assess the performance of the KNN Regressor model using appropriate evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), or coefficient of determination (R-squared).

Hyperparameter Tuning: Adjust the hyperparameters of the KNN Regressor model, including the value of k, distance metric, or weighting scheme, based on performance evaluation. Techniques like grid search or cross-validation can be used for this purpose.

In conclusion, K-Nearest Neighbors (KNN) Regressor is a non-parametric algorithm used for regression tasks. It predicts the value of a target variable by considering the average or weighted average of the values of its k nearest neighbors in the feature space. It can be used in various domains where regression-based predictions are required.

3.2.4. Support Vector Regression

Definition: Support Vector Regression (SVR) is a variant of Support Vector Machines (SVM) that is used for regression tasks. SVR aims to find a function that approximates the relationship between the input variables and the continuous target variable, while also minimizing the margin of error within a specified tolerance.

Working Principle: The working principle of SVR involves finding an optimal hyperplane in a higher-dimensional feature space that best fits the training data. Here's a step-by-step explanation of the process:

Data Transformation: SVR applies a mapping function to transform the original input space into a higher-dimensional feature space. This transformation is performed using a kernel function, such as the radial basis function (RBF), polynomial, or sigmoid.

Training Data Selection: SVR selects a subset of training instances, called support vectors, that are crucial for defining the regression model. These support vectors are the data points that lie closest to the margin or violate the margin boundary.

Margin Definition: SVR defines an ε-tube around the regression function, where ε is the specified tolerance. The regression function aims to fit the data points within this tube, allowing for a certain degree of error.

Optimization: SVR formulates the regression problem as a constrained optimization task. The objective is to minimize the margin violation and ensure that the error falls within the specified tolerance. The optimization problem involves solving a quadratic programming (QP) formulation.

Dual Formulation: By applying the Lagrange multipliers and the Karush-Kuhn-Tucker (KKT) conditions, SVR formulates the optimization problem in the dual space. This leads to the derivation of a set of equations that involve the inner products between the support vectors.

Prediction: Once the model is trained, SVR can make predictions for new input data points by computing the dot products between the support vectors and the input samples in the transformed feature space. The predictions are obtained by applying the learned regression function to the transformed data.

Usage: Support Vector Regression is commonly used in various domains, including:

Financial Forecasting: SVR can be applied to predict stock prices, exchange rates, or any other financial time series.

Energy Load Prediction: SVR can help forecast energy demand, enabling efficient energy management and resource planning.

Environmental Modeling: SVR is useful for predicting environmental factors such as pollution levels, temperature, or rainfall.

Engineering Applications: SVR can be employed for tasks such as predicting equipment failure, estimating product quality, or optimizing manufacturing processes.

Medical Data Analysis: SVR has been used for medical applications, including disease progression modeling, drug response prediction, and medical image analysis.

How to Use SVR:

To use Support Vector Regression effectively, follow these general steps:

Data Preparation: Prepare a dataset with input features and the corresponding target values for regression.

Feature Scaling: Scale the input features to a common range, such as [0, 1] or [-1, 1], to ensure that they have a similar impact on the SVR model.

Kernel Selection: Choose an appropriate kernel function based on the characteristics of the data and the problem at hand. The RBF kernel is a popular choice due to its flexibility.

Model Training: Use the training dataset to train the SVR model. This involves solving the optimization problem and determining the support vectors.

Model Evaluation: Assess the performance of the SVR model using appropriate evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), or coefficient of determination (R-squared).

Hyperparameter Tuning: Adjust the hyperparameters of the SVR model, including the kernel type, regularization parameter (C), and kernel-specific parameters like gamma or degree. This can be done using techniques like grid search or cross-validation.

3.2.5. Nu-Support Vector Regression

Definition: NuSupport Vector Regression (NuSVR) is a variant of Support Vector Regression (SVR) that introduces a parameter called "nu" to control the number of support vectors and the width of the ε-tube. NuSVR aims to find a function that approximates the relationship between input variables and the continuous target variable, while allowing for a flexible trade-off between the number of support vectors and the margin violation.

Working Principle: The working principle of NuSVR is similar to SVR, with the addition of the "nu" parameter. Here's a step-by-step explanation of the process:

Data Transformation: NuSVR applies a mapping function, typically using a kernel function like the radial basis function (RBF), polynomial, or sigmoid, to transform the original input space into a higher-dimensional feature space.

Training Data Selection: NuSVR selects a subset of training instances as support vectors, similar to SVR. These support vectors are the data points that lie closest to the margin or violate the margin boundary.

Margin Definition: NuSVR defines an ε-tube around the regression function, where ε is the specified tolerance. The regression function aims to fit the data points within this tube, allowing for a certain degree of error.

Optimization: NuSVR formulates the regression problem as a constrained optimization task, similar to SVR. The objective is to minimize the margin violation while controlling the number of support vectors within a range defined by the "nu"

parameter. The optimization problem involves solving a quadratic programming (QP) formulation.

Dual Formulation: By applying the Lagrange multipliers and the Karush-Kuhn-Tucker (KKT) conditions, NuSVR formulates the optimization problem in the dual space. This leads to the derivation of a set of equations involving the inner products between the support vectors.

Prediction: Once the model is trained, NuSVR can make predictions for new input data points by computing the dot products between the support vectors and the input samples in the transformed feature space. The predictions are obtained by applying the learned regression function to the transformed data.

Usage: NuSupport Vector Regression can be used in various domains, including:

Financial Forecasting: NuSVR can be applied to predict stock prices, exchange rates, or any other financial time series.

Energy Load Prediction: NuSVR can help forecast energy demand, enabling efficient energy management and resource planning.

Environmental Modeling: NuSVR is useful for predicting environmental factors such as pollution levels, temperature, or rainfall.

Engineering Applications: NuSVR can be employed for tasks such as predicting equipment failure, estimating product quality, or optimizing manufacturing processes.

Medical Data Analysis: NuSVR has been used for medical applications, including disease progression modeling, drug response prediction, and medical image analysis.

How to Use NuSVR:

To use NuSupport Vector Regression effectively, follow these general steps:

Data Preparation: Prepare a dataset with input features and the corresponding target values for regression.

Feature Scaling: Scale the input features to a common range, such as [0, 1] or [-1, 1], to ensure that they have a similar impact on the NuSVR model.

Kernel Selection: Choose an appropriate kernel function based on the characteristics of the data and the problem at hand. The RBF kernel is a popular choice due to its flexibility.

Model Training: Use the training dataset to train the NuSVR model. This involves solving the optimization problem while controlling the number of support vectors within the range defined by the "nu" parameter.

Model Evaluation: Assess the performance of the NuSVR model using appropriate evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), or coefficient of determination (R-squared).

Hyperparameter Tuning: Adjust the hyperparameters of the NuSVR model, including the kernel type, "nu" parameter, and kernel-specific parameters like gamma or degree. This can be done using techniques like grid search or cross-validation.

Prediction: Once the model is trained and evaluated, use it to make predictions on new, unseen data points.

In conclusion, NuSupport Vector Regression (NuSVR) is a variant of Support Vector Regression that introduces the "nu" parameter to control the trade-off between the number of support vectors and the margin violation. It can be used in various domains where accurate continuous value prediction is required while allowing flexibility in the number of support vectors.

3.2.6. Artificial Neural Network

Definition: Artificial neural networks (ANNs), commonly referred to simply as neural networks (NNs) or neural networks, are computing systems stimulated by the biological neural networks that make up animal brains. ANNs are based on a set of connected units or nodes called artificial neurons that loosely model neurons in the biological brain. ANNs consist of interconnected nodes, called artificial neurons or "nodes," organized into layers. The layers are typically categorized as input, hidden, and output layers. Each connection between neurons has an associated weight that determines the strength of the connection.

Working Principle: The working principle of ANNs involves a series of mathematical operations, often implemented through matrix multiplications and non-linear activation functions. Here's a step-by-step explanation of the process:

Input Layer: The input layer receives the initial data for processing. Each neuron in this layer represents a feature or attribute of the input data.

Hidden Layers: The hidden layers, placed between the input and output layers, perform intermediate computations. These layers help the network learn complex patterns and representations by transforming the input data through weighted connections and applying non-linear activation functions.

Output Layer: The output layer provides the final results or predictions based on the computations performed in the hidden layers. The number of neurons in this layer depends on the nature of the problem, such as regression or classification.

Weighted Connections: Each connection between neurons has an associated weight value. These weights represent the strength or importance of the connection. During training, the network adjusts these weights to optimize its performance by minimizing the error between predicted outputs and the expected outputs.

Activation Functions: Activation functions introduce non-linearities to the network, enabling it to model complex relationships. Common activation functions include the sigmoid, tanh, and ReLU (Rectified Linear Unit) functions.

Training and Learning: To use an artificial neural network effectively, a labeled dataset is needed to train. The following steps need to be taken for the training process:

Dataset Preparation: Prepare a dataset with input samples and their corresponding expected outputs or labels.

Forward Propagation: In the forward propagation step, the input data is fed through the network, and the weighted connections and activation functions are applied to generate predictions.

Error Calculation: Compare the predicted outputs with the expected outputs using an appropriate error or loss function. The choice of loss function depends on the specific problem, such as mean squared error (MSE) for regression or cross-entropy loss for classification.

Backpropagation: Backpropagation is a fundamental algorithm for training ANNs. It involves calculating the gradient of the loss function with respect to the network's weights. This gradient is then used to update the weights and reduce the prediction error. The process is repeated iteratively for multiple epochs until the network converges.

Applications: Artificial neural networks have found applications in various domains, including:

Pattern Recognition: ANNs are widely used for tasks such as image and speech recognition, object detection, and handwriting recognition.

Natural Language Processing (NLP): ANNs have been successful in applications like sentiment analysis, machine translation, and text generation.

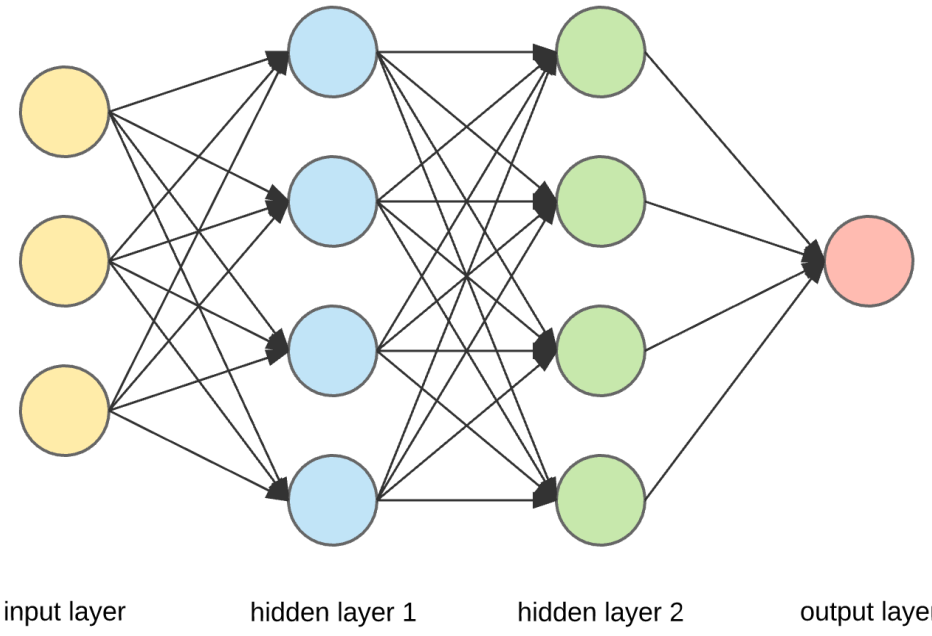
Time Series Analysis: ANNs can analyze and predict time-dependent data, making them useful for tasks like stock market forecasting, weather prediction, and anomaly detection.

Recommender Systems: ANNs can be utilized to build recommendation engines that suggest relevant products, movies, or content based on user preferences.

Medical Diagnosis: ANNs have been applied to medical data analysis, assisting in disease diagnosis, predicting patient outcomes, and drug discovery.

Autonomous Vehicles: ANNs play a crucial role in enabling autonomous vehicles to perceive and interpret their environment, helping with tasks like object detection and path planning.

3.2.6.1. Artificial Neural Network Diagram



4. APPLICATION

In this section, we train the proposed models of the study with the data from gittigidiyor.com, one of Turkey's most popular e-commerce websites. In this study, 5 different models are proposed in the light of the relevant literature: Artificial Neural Networks, Support Vector Regression, NuSupport Vector Regression, K-Nearest Neighbor Regressor, and Ordinary Least Squares. We compare the accuracy of these models according to mean squared error values.

4.1. Feature Selection

We conducted a feature selection process based on aggregate information criteria to ensure that the most relevant features are included in our analysis. Mutual information gives an idea about the lengths between the variables, the lengths and the strength of the relationship between them. Using mutual information, we aimed to identify the regulations to have the highest information impact in predicting online marketplace sales on the Turkish e-commerce site.

The cluster in our data started by calculating the common knowledge scores for each feature. Features considered include discount percentage, free shipping status, product rating, customer reviews, product category, seller rating, and product rankings. These features have been chosen based on their potential impact on customer behavior and sales results.

We aimed to overcome the dependencies of our variables that we will use in the regression with mutual knowledge scores. In order for our regression models to have higher MSE and R2 scores, our independent variables included in the regression should include the least number of users.

As a result of our analysis, although we determined that the excessive information score of the discount provisions and the product price is higher than the

other common information scores, we decided to include both variables in the model due to the high consumption of the score.

Using the general information criterion for data selection, we ensured that our analysis is focused on the most effective tools for providing online marketplace sales forecasts. This approach engages and depletes our predictive models by including key features that have an important bearing on customer behavior and purchasing decisions in the Turkish e-commerce space.

4.2. Data Cleaning

The collected data set undergo a comprehensive data cleaning process to ensure its quality, consistency and suitability for analysis. The first step in the data cleaning phase is to standardize the data formats. First, we extract data containing numeric expressions from string expressions and standardized formats and units to ensure consistency and facilitate meaningful analysis. For example, prices are converted into a consistent currency format, units of measurement are standardized, and categorical variables are coded appropriately.

We continued with data cleaning by identifying and organizing inconsistent or erroneous records. Then, we identify missing data. While the literature suggests several methods to fill in missing data, we choose to remove missing data from the data set. We use elimination for duplicate data that originated from the system during data collection and remove these records from the data set using product names and unique product IDs that we create before.

In the final stage of data cleaning, we address the detection of outliers, which are outliers that deviate significantly from the overall pattern of the data. As we considere that outliers would significantly bias our analyses, we fill the outliers with the median. Finally, we examine the categorical variables in the dataset and apply appropriate coding techniques. Typically, this involve converting categorical variables into numerical representations, such as one-shot coding or sequential coding, to ensure their inclusion in the analysis.

4.1.1. Mutual Information

	Product Price	Discounted Price	Discount Percentage	Free Shipping	Stocked Unit	Number of Review	Product Ranking	Seller Raking	Discount Amount
Product Price		2,07	1,31	0,22	1,31	2,69	2,1	1,57	4,67
Discounted Price	2,07		1,66	0,05	0,67	1,34	1,04	0,78	2,57
Discount Percentage	1,31	1,66		0,03	0,31	0,76	0,58	0,37	1,6
Free Shipping	0,22	0,05	0,03		0,03	0,09	0,07	0,06	0,2
Stocked Unit	1,31	0,67	0,31	0,03		0,55	0,37	0,31	1,47
Number of Review	2,69	1,34	0,76	0,09	0,55		1,93	0,81	2,84
Product Ranking	2,1	1,04	0,58	0,07	0,37	1,93		0,49	2,2
Seller Raking	1,57	0,78	0,37	0,06	0,31	0,81	0,49		1,73
Discount Amount	4,67	2,57	1,6	0,2	1,47	2,84	2,2	1,73	

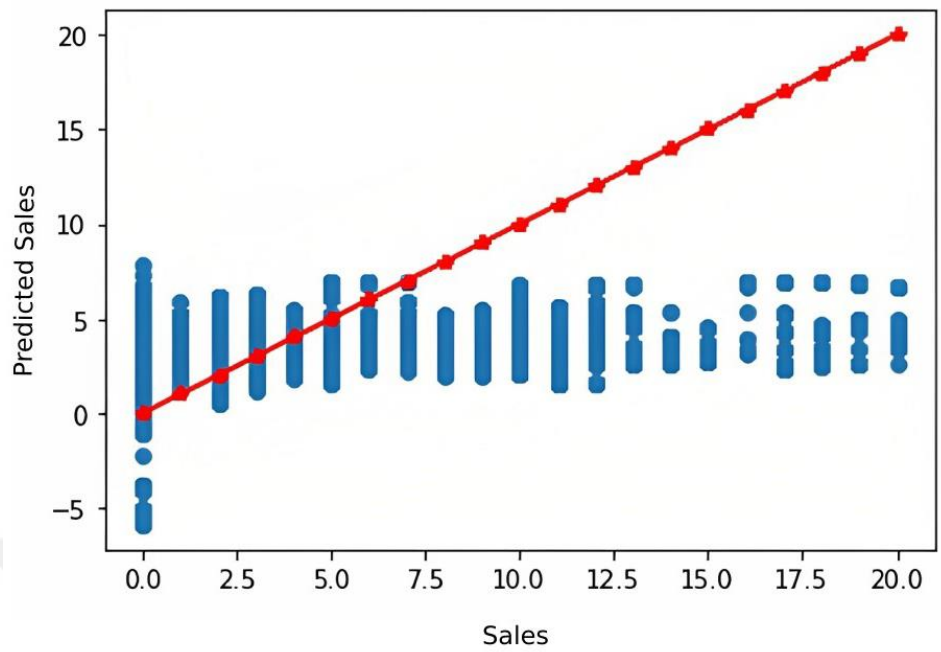
4.3. Ordinary Least Squares

This paper's goal is to forecast future sales using various statistical methods. OLS is one of these models. Sales are chosen as the dependent variable, while a collection of attributes, such as price, rating, and review count, are chosen as independent variables We can use OLS model in this paper, because our dependent variable is continuous.

An OLS regression model is fit to the training data, and then used to predict sales for the test data. As we mentioned before, the performances of model are calculated with mean squared error. For the OLS model, we also evaluate R-squared value.

The result of OLS model monitors that OLS model had an R-square of 0.110 and a MSE of 2.5. These findings imply that the model could account for 11% of the variation in sales and that, on average, its forecasts are within 2.5 units of the actual sales figures.The study provides evidence that an OLS regression model cannot be used to predict sales from the Gittigidiyor.com dataset with reasonable accuracy. The approach can be expanded to include more features, or other machine learning models can be used for better performance, as previous parts in this article.

4.3.1. OLS Predicted Values



The red line is a 45-degree line that is the cross-product of the y test and y test and blue dots represent the predicted values from model.

4.4. K-Nearest Neighbor Regressor

In this part of the study, we split the data collected from the website using a web crawler into test and training data sets in a 70/30 ratio. We build the KNN regressor model with the kd_tree algorithm for 5 neighbors. The model used price, product rank, number of review, free-shipping,seller rating and discounts to predict sales. The KNN regressor is chosen because it is a simple and effective method for regression problems.

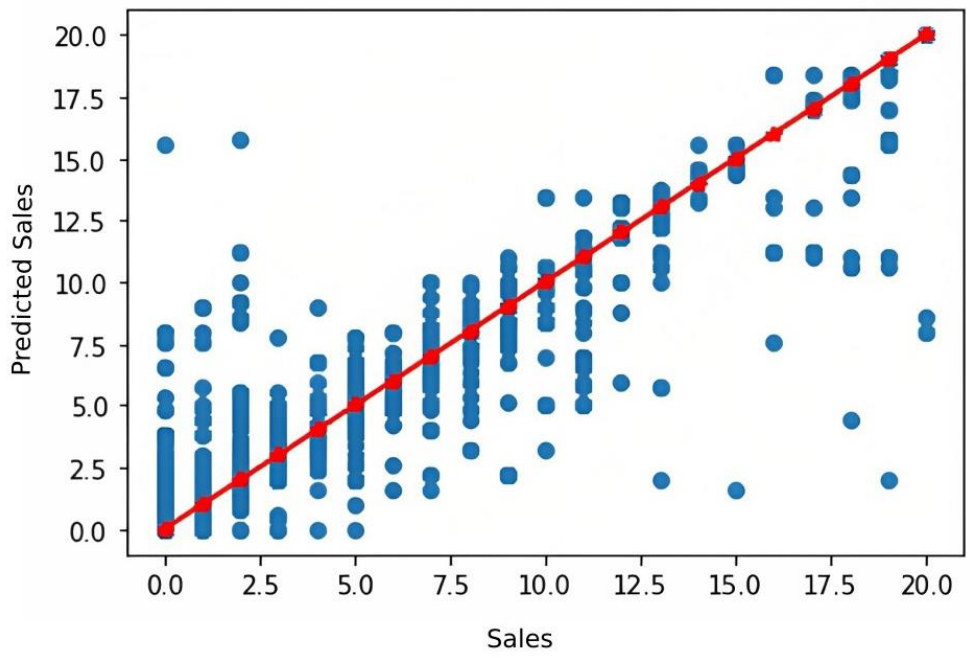
After deciding the attributes based on the common knowledge theorem mentioned in the methodology section, the KNN-regressor model is applied to the training data. Using the trained model, sales is predicted for the test dataset. The performance of the model is then evaluated using the mean squared error (MSE).

The results showed that the KNN regressor model has a weighted R-squared of 0.94 and a mean squared error of 1.27. This indicates that the model can explain 94%

of the variability in sales and accurately predicts within 1.27 units of the realized sales values on average.

In conclusion, the KNN-regressor model for sales forecasting is successfully tested on the Gittigidiyor.com dataset. The model demonstrates satisfactory performance and suggested that use cases can be a useful tool for optimizing sales strategies.

4.4.1. KNN Regressor Predicted Values



The red line is a 45-degree line that is the cross-product of the y test and y test and blue dots represent the predicted values from model.

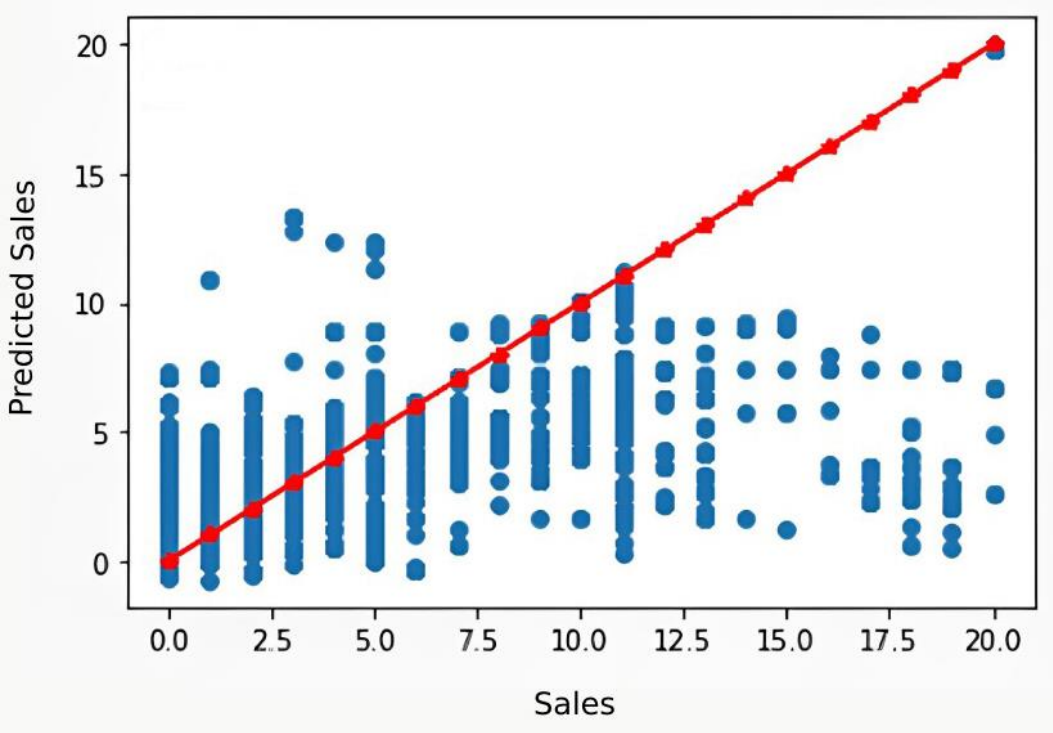
4.5. Support Vector Regression

As with the other models proposed in this study, we split the dataset 70% - 30% into a training set and a test set. We used cross-validation techniques such as k-fold cross-validation to ensure robustness and reduce overfitting.

After training the SVR model, we measured the performance of our predictions with test data. For forecasting online marketplace sales, the SVR model has an r2 score of 0.44 and an MSE value of 11,01. Thus, this indicates that the model can explain 44% of the variability in sales and accurately predicts within 11.01 units of the realized sales values on average.

To benchmark the performance of the SVR algorithm, we compare its results with other commonly used regression algorithms such as OLS and KNN regressor and Ann. While the SVR algorithm outperforms these traditional regression models in terms of predictive accuracy and robustness, it is not the most effective model for predicting online marketplace sales based on the identified determinants.

4.5.1. SVR Predicted vs Actual Values



The red line is a 45-degree line that is the cross-product of the y test and y test and blue dots represent the predicted values from model.

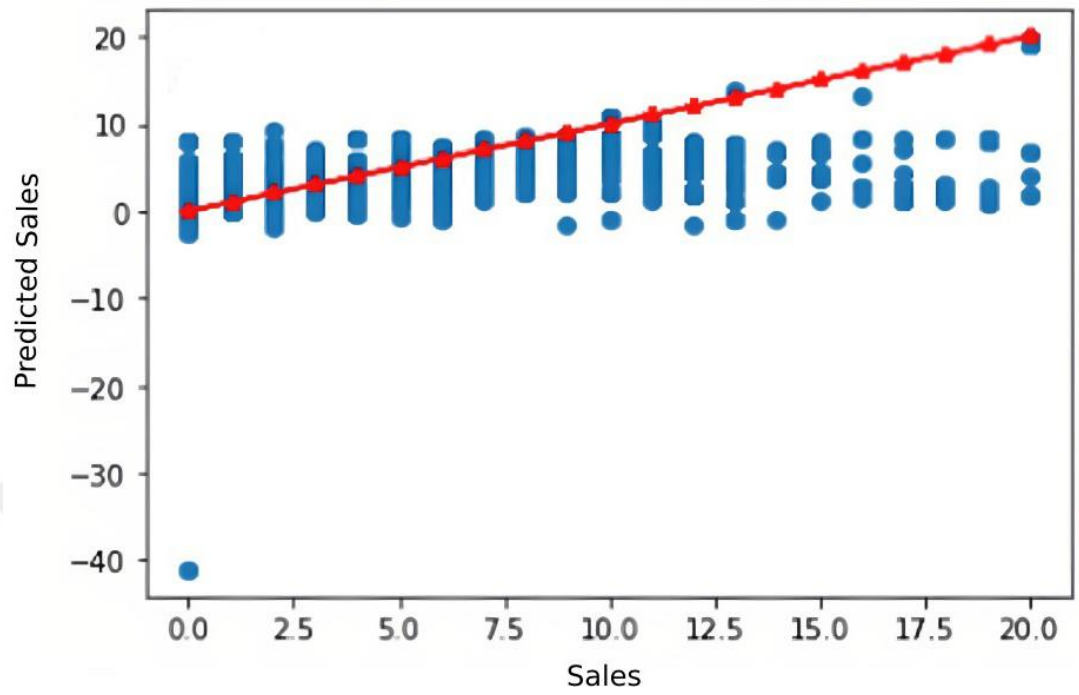
4.6. Nu-Support Vector Regression

As the other models we develop in this article, we apply same data processing steps. We separate data in two parts which are test and train in value of %30 and %70. We develop Nu- SVR model with radial basis function and gamma value of 0.22. We also use Mean Squared Error (MSE) and R-squared while measure the performance of Nu-SVR model.

The Nu-SVR model that developed the predict sales in Turkish E-bussines sector,

has value R-squared of 0,46 and MSE value of 10,86. It's obvious, Nu-SVR model is the more accurate model than SVR model for the this data set but it is not also most robust model for gittigidiyor.com dataset.

4.6.1. Nu-SVR Predicted vs Actual Values



The red line is a 45-degree line that is the cross-product of the y test and y test and blue dots represent the predicted values from model.

The results reveal that the Nu-SVR model outperforms the SVR model based on prediction accuracy and identified predictors. This performance difference can be attributed to several key differences between Nu-SVR and SVR:

4.6.2. Flexibility in Margin Width:

Nu-SVR allows for a flexible margin width determined by the parameter nu, which represents an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors. In contrast, SVR uses a fixed margin width determined by the epsilon-insensitive tube, making it less adaptable to different datasets.

4.6.3. Robustness to Outliers:

Nu-SVR exhibits improved robustness to outliers compared to SVR. By allowing a more flexible margin width, Nu-SVR can better accommodate outliers within the margin, leading to better performance in the presence of outlier data points.

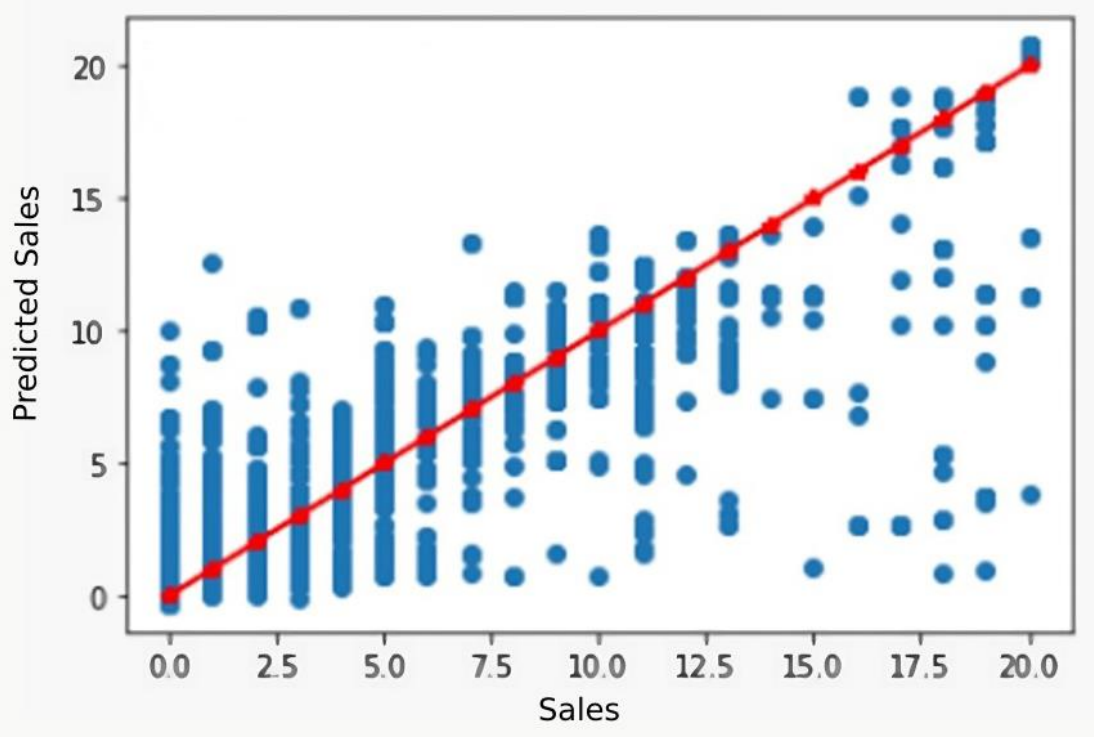
4.7. Artificial Neural Network

To evaluate the effectiveness of Artificial Neural Networks (ANNs) in predicting online marketplace sales based on identified determinants, we apply the most commonly used artificial neural network model in the literature using the Turkish e-commerce dataset (Leong et al. 2013; Tan et al. 2014, Hou et al. 2017). For the ANN model, we designed a feed-forward neural network architecture consisting of multiple hidden layers with varying numbers of neurons. The model is trained using backpropagation and gradient descent optimization techniques with rectified linear unit (ReLu) activation functions and learning rates.

We measured the prediction capabilities of the model and its ability to accurately capture the variability in sales data using the mean square error (MSE), coefficient of determination (R-squared). After training and testing the ANN model, we performed our measurements to predict online marketplace sales. We calculated the accuracy and reliability of the ANN model's predictions as 5.07 based on MSE and 0.74 based on R-square.

The obtained metrics show that the ANN model achieves a reasonable level of accuracy in predicting online marketplace sales based on the identified determinants. The low MSE values along with the high R-squared value indicate the ability of the model to capture the variability in the sales data and provide accurate predictions.

4.7.1. ANN Predicted vs Actual Values



The red line is a 45-degree line that is the cross-product of the y test and y test and blue dots represent the predicted values from model.

5. RESULTS

This section contains the findings of our empirical study that compared the accuracy of five different regression models at forecasting sales on online marketplaces using the identified determinants. Ordinary Least Squares (OLS), Support Vector Regression (SVR), Nu-Support Vector Regression (Nu- SVR), K-Nearest Neighbors (KNN) Regressor, and Artificial Neural Network are the models that are conducted. The Turkish e-commerce dataset is used for the evaluation, which includes sales information and pertinent factors including the availability of free shipping and discount percentages.

5.1 Model Results

METHODS \ MEASURES	R^2	MSE
Ordinary Least Squares	0.11	17.26
KNN Regressor	0.90**	1.81**
Support Vector Regression	0.44	11.01
Nu-Support Vector Regression	0.46	10.86
Artificial Neural Network	0.74*	5.07*

Based on the results, we observe that the KNN Regressor model achieved the lowest MSE values, indicating superior prediction accuracy compared to other models. In addition, the KNN Regressor model shows the highest R-squared value, indicating its ability to explain a significant portion of the variability in the sales data based on the identified determinants.

The performance of the remaining models varies. While the Artificial Neural Network provides reasonable predictive performance, the other models show slightly higher MSE values, and a lower R-squared value compared to the KNN Regressor model.

These findings suggest that the KNN Regressor model outperforms other regression models in predicting online marketplace sales based on the identified determinants. The KNN Regressor's ability to capture complex relationships and patterns in the data by exploiting the proximity of similar data points contributes to its superior performance in this context.

It is important to note that the performance of each model may vary depending on the specific dataset and experimental setup. However, based on our analysis of the Turkish e-commerce dataset, the KNN Regressor model emerges as the most promising model for predicting online marketplace sales.

Moreover, the KNN regressor model exhibits a low computational complexity compared to other models such as SVR and OLS. Therefore, we can say that the KNN regressor model outperforms other models in making sense of real-time, big data.

While the KNN regressor model monitors superior performance, the Artificial Neural Networks (ANN) model shows moderate performance in predicting online marketplace sales. The ANN model's ability to capture complex relationships and non-linear patterns lead it to perform better than other models. On the other hand, both Support Vector Regression (SVR) and Ordinary Least Squares (OLS) models underperform the ANN model. The fixed margin width in SVR and the linearity assumption in OLS lead to an increase in error terms.

In conclusion, our empirical study shows that the K-Nearest Neighbors (KNN) Regressor model outperforms other regression models in predicting online marketplace sales based on the identified determinants in the Turkish e-commerce context. The KNN Regressor model's ability to capture non-linear patterns, exploit the proximity of similar data points, and exhibit computational efficiency contribute to its superior performance.

These findings have practical implications for e-commerce businesses as they can use the KNN Regressor model to make accurate sales forecasts and inform decision-making processes related to marketing strategies, inventory management and resource allocation. However, it is important to note that the choice of the most appropriate regression model should take into account the specific characteristics of the dataset and the objectives of the forecasting task.

Further research could explore the performance of other advanced regression techniques or ensemble methods in predicting online marketplace sales. In addition,

investigating the impact of additional determinants or incorporating external factors such as customer demographics or product reviews could improve the prediction accuracy of the models.

5.2. Implications and Model Implementation

The developed model has important implications for various domains and offers valuable opportunities for its application in various fields to support decision-making and improve processes.

One of the areas where the model can be used is the model's ability to analyze the relationship between product features and prices, and their impact on pricing strategy and market positioning. Companies can leverage the insights provided by the model to determine optimal pricing strategies and market positioning. This includes identifying price points that maximize profitability while considering consumer affordability and the competitive environment.

Another area where the model can be applied is inventory management and supply chain optimization. By forecasting demand based on product characteristics and market conditions, the model can help companies make data-driven decisions on production levels, stock replenishment and distribution strategies. This helps minimize costs, reduce inventory outages and improve overall supply chain efficiency.

The model's analysis of consumer preferences and behavior also has implications for consumer insights and marketing campaigns. Companies can use the model's predictions to identify customer segments, personalize marketing messages and design campaigns that resonate with the target audience, ultimately increasing customer engagement and conversion rates.

Furthermore, the model's analysis of the small household appliances market has implications for market research and competitive analysis. It can help identify market trends, consumer preferences and competitors' product offerings, enabling companies

to make strategic decisions, differentiate their products and gain competitive advantage in the industry.

Potential applications of the model include e-commerce platforms and recommendation systems. By integrating the model into online platforms, personalized product recommendations can be created, user experience can be improved, customer satisfaction can be increased and sales can be increased.

It is important to note that the above mentioned areas are not exhaustive and the applicability of the developed model may extend to other areas depending on the specific context and requirements. It is recommended to explore additional applications and potential synergies with existing systems and processes through further research and collaboration with industry stakeholders.

6. CONCLUSION

In this study, we investigate the prediction of online marketplace sales using different prediction models using variables that enable the user to make decisions during shopping. Our findings provide valuable insights into the effectiveness of these models and their contribution to understanding customer behavior in the context of Turkish e-commerce.

In addition, while ANN and SVR are the most preferred models in the related literature, the K-Nearest Neighbors (KNN) Regressor model outperformed other regression models, including Linear Regression, Support Vector Regression (SVR) and Artificial Neural Networks (ANN). The KNN Regressor model showed superior prediction accuracy by capturing complex relationships and non-linear patterns by exploiting the proximity of similar data points. The computational efficiency of this model further enhances its suitability for real-time or large-scale prediction tasks.

In this study, he highlighted the importance of determinants that influence customers' purchasing behavior. In particular, factors such as discount percentages and free shipping have proved to be important in predicting purchasing behavior in the e-

commerce industry. By considering these determinants, e-commerce businesses can make informed decisions regarding marketing strategies, inventory management and resource allocation.

Although the KNN Regressor model works best for e-commerce data performance in Turkey, it is worth noting that it should consider the specific characteristics of the dataset and the objectives of the forecasting task. It is possible to build models with more different variables that yield higher results. For example, in this study, although we analyze customers behaviorally, we do not have access to their demographic data. There is no doubt that models built with demographic data of customers will yield higher results.

Our study contributes to the existing literature by providing empirical evidence on the effectiveness of different regression models in predicting online market sales. Furthermore, our findings shed light on the importance of the determinants of customer behavior in the Turkish e-commerce industry.

Future research can build on this study by exploring additional regression techniques, ensemble methods, or incorporating external factors such as customer demographics and product reviews to further improve the predictive accuracy of the models. Additionally, investigating the performance of these models in different e-commerce contexts or analyzing the impact of specific determinants on different product categories could provide further insights into customer behavior and sales forecasting.

In conclusion, our study highlights the importance of regression modeling in predicting online marketplace sales and emphasizes the value of considering the determinants that shape customer behavior. The findings have practical implications for e-commerce businesses seeking to optimize their sales strategies and improve customer satisfaction.

7. REFERENCES

[1] Assael, H. (1984). Consumer behavior and market action. Boston, MA: Kent.

[2] Bakir, H., Chniti, G., & Zaher, H. (2018). E-Commerce price forecasting using LSTM neural networks. International Journal of Machine Learning and Computing, 8(2), 169-174.

[3] Chaudhuri, N., Gupta, G., Vamsi, V., & Bose, I. (2021). On the platform but will they buy? Predicting customers' purchase behavior using deep learning. Decision Support Systems, 149, 113622.

[4] Chen, Y. F., & Cheng, R. C. (2019). Single Discount or Multiple Discounts?: Effects of Price Promotion Framing Messages on Online Consumer Purchase Intention. International Journal of Technology and Human Interaction (IJTHI), 15(1), 1-14.

[5] Chong, A. Y. L. (2013). Predicting m-commerce adoption determinants: A neural network approach. Expert systems with applications, 40(2), 523-530

[6] Chong, A. Y. L., Cheng, E., Liu, M. J., & Li, B. (2017). Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. International Journal of Production Research, 55(17), 5142-5156.

[7] Chong, A. Y. L., Li, B., Ngai, E. W., Ch'ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. International Journal of Operations & Production Management.

[8] Deloitte Turkey. (2022, February). E-ticaretin öne çıkan başarısı, tüketici davranışlarında değişim ve dijitalleşme. https://www.eticaretraporu.org/

[9] Gopalakrishnan, T., Choudhary, R., & Prasad, S. (2018, December). Prediction of sales value in online shopping using linear regression. In 2018 4th International Conference on Computing Communication and Automation (ICCCA) (pp. 1-6). IEEE.

[10] Helmi, M., Xiao, S., & Nicholson, M. (2020). The Effect of Price Promotion on Just Noticeable Difference in Multichannel Retailing.

[11] Hou, F., Li, B., Chong, A. Y. L., Yannopoulou, N., & Liu, M. J. (2017). Understanding and predicting what influence online product sales? A neural network approach. Production Planning & Control, 28(11-12), 964-975.

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

[13] Jothi, V. L., Aditya, B., Arthika, S., & Jayashree, S. Sales Prediction using Machine Learning Algorithm.

[14] Kim, M. J. (2017). How to promote E-commerce exports to China: an empirical analysis. KDI Journal of Economic Policy, 39(2), 53-74.

[15] Kristofferson, K., McFerran, B., Morales, A. C., & Dahl, D. W. (2017). The dark side of scarcity promotions: how exposure to limited-quantity promotions can induce aggression. Journal of Consumer Research, 43(5), 683-706.

[16] Lee, J. E., & Chen-Yu, J. H. (2018). Effects of price discount on consumers' perceptions of savings, quality, and value for apparel products: mediating effect of price discount affect. Fashion and Textiles, 5(1), 13.

[17] Leong, L. Y., Hew, T. S., Tan, G. W. H., & Ooi, K. B. (2013). Predicting the determinants of the NFC-enabled mobile credit card acceptance: A neural networks approach. Expert Systems with Applications, 40(14), 5604-5620.

[18] Mu, W. (2019). A big data-based prediction model for purchase decisions of consumers on cross-border e-commerce platforms. Journal Européen des Systèmes Automatisés, 52(4), 363-368.

[19] Raizada, S., & Saini, J. R. (2021). Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting. International Journal of Advanced Computer Science and Applications, 12(11).

[20] Shao, F., & Yao, J. (2018, January). The establishment of data analysis model about E-commerce's behavior based on Hadoop platform. In 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 436-439). IEEE.

[21] Sharma, S. K., Chakraborti, S., & Jha, T. (2019). Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach. Information Systems and e-Business Management, 17, 261-284.

[22] Takemura, K. (2019). Mental Accounting and Framing: Framework of Decisions in Consumer Price Judgment. In Foundations of Economic Psychology (pp. 121-156). Springer, Singapore

[23] Tan, G. W. H., Ooi, K. B., Leong, L. Y., & Lin, B. (2014). Predicting the drivers of behavioral intention to use mobile learning: A hybrid SEM-Neural Networks approach. Computers in Human Behavior, 36, 198-213.

[24] TUIK. Household Information Technology Usage Research. (2021) https://data.tuik.gov.tr/Bulten/Index?p=Hanehalki-Bilisim-Teknolojileri-(BT)-Kullanim-Arastirmasi-2021-37437

[25] Zhang, M., Wang, Y., & Wu, Z. (2021). Data mining algorithm for demand forecast analysis on flash sales platform. Complexity, 2021, 1-12.

[26] Zhang, Y. (2021). Sales forecasting of promotion activities based on the cross-industry standard process for data mining of E-commerce promotional information and support vector regression. Journal of Computers, 32(1), 212-225.

[27] Zhao, Z., Wang, J., Sun, H., Liu, Y., Fan, Z., & Xuan, F. (2019). What Factors Influence Online Product Sales? Online Reviews, Review System Curation, Online Promotional Marketing and Seller Guarantees Analysis. IEEE Access, 8, 3920-3931.