Investigating Factors Affecting Consumer Purchase Intentions in the Online Retail Sector: An Analysis of Customer Data

Maiqi Huang[1*+], Qing Lian[2+], Siying Wu[3+]

[1]Department of economics, Henan University, Kaifeng, 475004, China
[2]Business School, The University of Sydney, Sydney, 2006, Australia
[3]Department of Accounting, Hubei University of Economics, Wuhan, 430205, China

*Corresponding author. Email: maidaren3@gmail.com

Abstract. This work delves into the various factors influencing online shopping behavior and the impact on customer purchase outcomes. With the ever-expanding e-commerce landscape, understanding the dynamics of consumer behavior is of paramount importance for online retailers seeking to enhance their revenue and customer loyalty. Five key factors are focused on: weekends, visitor types, special days, page values, and product-related activities. Through rigorous data analysis and regression models, this study identifies how these factors interact and affect online shopping behavior. The findings reveal that weekends play a crucial role in influencing consumer behavior, with increased social media usage during weekends offering opportunities for targeted promotions. Moreover, fostering customer loyalty through tiered benefits systems can bridge the gap between new and returning customers, ultimately contributing to sustainable growth. Strategically designed special day promotions and optimized webpage designs are shown to enhance user experiences and drive higher conversion rates. Additionally, informative and engaging content on product-related pages empowers potential buyers to make informed decisions. In conclusion, by implementing cohesive strategies that leverage these factors, online retailers can effectively improve conversion rates, enhance customer loyalty, and secure a competitive edge in the online retail landscape. This research contributes to a deeper understanding of e-commerce dynamics and offers practical recommendations for businesses in this domain.

Keywords: User Purchasing Behavior, Online Retail Websites, Conversion Rates, Optimization Strategies.

B. K. Kandel et al. (eds.), *Proceedings of the 2023 3rd International Conference on Business Administration and Data Science (BADS 2023)*, Atlantis Highlights in Computer Sciences 19,
https://doi.org/10.2991/978-94-6463-326-9_10

1 Introduction

1.1 Background and Context

In recent years, the online shopping landscape has experienced unprecedented growth, fundamentally reshaping the retail industry. The ease of access, convenience, and extensive product offerings have attracted consumers from diverse demographics. As technology continues to shape daily lives, businesses must adapt to this dynamic and competitive digital environment. Understanding the factors that drive consumer acceptance of online shopping is imperative for establishing a strong digital presence. By identifying both motivators and barriers in consumer decision-making, online retailers can tailor their strategies to effectively attract and retain customers.

1.2 Research Problem

The central challenge lies in deciphering the intricate relationship between various facets of online shopping and consumer purchasing behavior. By analyzing how elements such as Page Value, special days, visitor type, weekends, and product-related information impact user behavior, this research aims to uncover insights that empower online retailers to enhance the overall shopping experience. Understanding why consumers make certain choices and what aspects of online shopping may deter them will enable businesses to optimize their platforms and create a user-centric shopping environment.

1.3 Research Objectives

This research has two primary objectives:

1. To examine the factors influencing consumer purchasing behavior on online retail websites.
2. To propose optimization strategies based on data-driven insights to enhance conversion rates for online retailers.

1.4 Significance of the Study

This study's significance lies in its potential to inform and guide online retailers in a rapidly evolving digital marketplace. By gaining a deeper understanding of consumer decision-making processes, businesses can tailor their approaches to increase revenue and customer satisfaction. Additionally, this research contributes to the broader field of consumer behavior in the digital era, furthering understanding of the evolving relationship between consumers and online shopping platforms.

In the subsequent sections, this work will delve into the methodology, results, and discussion, offering a comprehensive analysis of our findings. By the end of this study, this work aims to provide insights that can contribute to the growth and success of online retailers in an increasingly digital-centric world, while also advancing our understanding of consumer behavior in the digital era.

2 Literature Review

When the literature is reviewed, it is seen that there are lots of studies looking into the factors influencing buyers behavior on online purchasing websites.

Berg et al. elucidate the concept of Page Value, its computation, and its significance for website proprietors and marketers in assessing the efficacy of individual webpages [1]. Dashthis.com expounds upon "Total Goal Value" and offers comprehensive instructions for its calculation [2].

Gao L delineates the evolution of e-commerce, emphasizing heightened competition and insights derived from a study on determinants of user shopping intentions for a traditional retailer in transition [3]. Moe et al. construct an individual-level model for evolving visitation patterns based on Internet clickstream data [4], while Park et al. formulate a stochastic timing model of cross-site visit behavior to elucidate the utilization of information from one site in explaining customer conduct on another [5].

Meer reports on a real-world application of customer development and retention using clickstream data from a financial institution in the Netherlands [6]. Hu et al. propose a unified model utilizing Web farming technology for comprehensive clickstream log collection during the entire user interaction process [7]. This meticulous recording of user clicks facilitates the modeling of user behavior, as demonstrated by Hanamanthrao et al. in their study, which extracts valuable customer insights from clickstream data [8].

Ghavamipoor et al. address the need for a QoS-sensitive model to formulate QoS-aware offers for customers, utilizing process mining to develop a QoS-CBMG [9]. The paper explores the trade-off between privacy and the business value of customer information, categorizing clickstream features as potential privacy threats, as articulated by Baumann et al. [10].

Somya et al. describe the architecture of an online shop application, the clickstream data recording module, and the data analysis methodology [11]. Gumber et al. introduce an innovative framework employing the Extreme Gradient Boosting ensemble method for predicting customer behavior [12]. Necula investigates the influence of time spent on reading product information on consumer behavior in e-commerce [13].

Li Guoxin, Li Yijun, and Li Bing employ the Technology Acceptance Model (TAM) to assess the impact of perceived usefulness, ease of use, and safety on online shopping attitudes and intentions among Chinese college students, offering marketing insights for online retailers [14]. Al Hamli and Sobaih find that product variety, payment methods, and psychological factors significantly influence online shopping during the COVID-19 pandemic, informing e-commerce businesses' crisis-driven marketing strategies [15].

Lastly, Daroch, Nagrath, and Gupta identify six factors hindering online purchases, including fear of bank transactions, convenience of traditional shopping, reputation, experience, insecurity, and insufficient product information, emphasizing the importance of trust in online shopping decisions [16].

In summary, these studies collectively contribute to understanding of various aspects related to online shopping, ranging from user behavior to factors affecting purchasing decisions. When the results are evaluated in overall terms, it is seen that some research

was conducted a long time ago. But online shopping is developing rapidly, there are still many factors that influence buyers behavior that remain unstudied. Therefore, this study is believed to contribute to the research to fill this gap and help online retailers maximize their profits.

3 Research Methodology

3.1 Dataset Overview

The data used for this study were collected from the online retailer, they used the tracking system to trace the users' behavior by using URL information of the pages visited by the user and updated in real-time when a user takes an action. The dataset comprises 12,330 sessions with 18 attributes, including 10 numerical and 8 categorical ones. These attributes provide information about user sessions on an e-commerce site. The dataset includes details on the number of different types of pages visited, the total time spent on these pages, bounce rates, exit rates, page values, special day indicators, as well as information about the user's operating system, browser, region, traffic type, visitor type, weekend visit indicator, and month of the year. The "Revenue" attribute can be used as the class label for classification and clustering tasks in a business context [17].

3.2 Data Collection

The data used in this study was obtained from the UCI Machine Learning Repository [17], which is a publicly accessible resource containing diverse datasets related to machine learning and data analysis. In the work, data collection process comprised several key steps: Firstly, the work began by identifying the most pertinent dataset for the research, which was the Online Shoppers Purchasing Intention Dataset available in the UCI repository. This dataset was deemed the most suitable for investigating the factors influencing consumer purchase intentions within the online retail sector. Next, the work proceeded with data selection. After downloading the dataset, a meticulous examination was carried out to pinpoint the specific variables and observations that directly aligned with the research objectives.

3.3 Data Cleaning and Processing

To improve the quality and reliability of data, the work undertook a comprehensive data preprocessing phase. This phase encompassed various tasks, one of which was the exclusion of the "other" category from the three types of visitors, as it was deemed irrelevant to the objectives of the research. Following data preprocessing, the work proceeded to transform categorical variables, including VisitorType, Weekend, and Revenue, into dummy variables.

3.4 Data Analysis Methodology

This study employed two statistical software packages, namely Excel and R, for data analysis.

Firstly, with regard to the independent variables under investigation, descriptive statistics were co ducted. Boxplots and histograms were utilized to gain comprehensive insights into the distribution of the independent variables within the two groups of the dependent varible, "Revnue" (True or False).

Secondly, to explore the relationships between all variables, a correlation matrix was co structed. This matrix facilitated the assessment of the strength and direction of associations among the variables.

Subsequently, the T-test was applied to determine whether significant differences existed in the independent variable concerning different values of the dependent variable. The T-test allowed for population inferences based on the collected samples, thus enabling an examination of the effect of the dependent variable on the independent variables.

Lastly, the Logit model was employed to analyze the significance of the independent variable on the dependent variable. Given that the dependent variable was binary (representing "purchase" or "non-purchase" events), the Logit model was well-suited for predicting the probability of purchase based on the influence of the independent variable.

Through this coherent research approach, the results could comprehensively analyze the relationships between the variables, delve into the impact of the independent variable on the dependent variable, and predict the probability of purchase behavior. These steps provided robust support for the validity of the study and the reliability of its conclusions.

4 Hypothesis

H0: Assume that PageValue, Visitor Type, Special Day, Product Related, Administration, Information, Product Related Duration, Administration Duration, Information Duration, Weekend have no effect on online users' willingness to shop (Revenue).

H1: Suppose PageValue, Visitor Type, Special Day, Product Related, Administrative, Information, Product Related Duration, Administrative Duration, Information Duration, Weekend will have a significant effect on online users' willingness to shop (Revenue).

5 Variables definition

The dataset comprises feature vectors from 12,330 distinct sessions, specifically structured to ensure that each session corresponds to different users over a one-year timeframe. This deliberate design aims to eliminate any bias towards particular campaigns, special occasions, user profiles, or time periods.

PageValue: The average value corresponds to the web page that a user accessed before finalizing an e-commerce transaction. This value is a metric tracked by "Google Analytics" for every page within the e-commerce site. The formula is:

$$PageValue = \frac{eCommerce\ Revenue + Goal\ Value}{Number\ of\ Unique\ Pageviews\ for\ Pages} \quad (1)$$

eCommerce Revenue: Total value of user-generated transactions

Goal Value: the dollar amount made from goal completions in Google Analytics. For example, if one of the goals is to drive more people to the product page of the eCommerce site, the total goal value would be indicative of the monetary value of this action.

Number of Unique Pageviews for Pages: Number of times a user visits a unique web page

Special Day: the proximity of site visits to specific occasions, such as Mother's Day or Christmas Day, where transactions are more likely to be completed. This attribute's value is calculated based on e-commerce dynamics, including the time gap between the order date and the delivery date. For instance, taking Valentine's Day as an example, its minimum value of zero corresponds to February 4, while its maximum value of 1 corresponds to February 14. As a special day draws nearer, the value of this attribute approaches 1.

Visitor Type: For returning or new visitor, 0 represents new visitor and 1 represents returning visitor.

Product Related: the number of webpages related to specific products visited by a user during a single session on an online platform. It quantifies the user's engagement with product-related content, indicating their level of interest and exploration of different products.

Administrative: the number of administrative pages that a visitor has accessed during their online session. These pages typically pertain to account management, settings, or other administrative functions on a website.

Information: the number of informational pages visited by a visitor during a particular session on a website. For instance, informational pages may include FAQs (Frequently Asked Questions), product guides, blog posts, customer reviews, educational resources, and other content aimed at guiding and assisting users in their decision-making process.

Product Related Duration, Administrative Duration, and Information Duration: The total time spent in each of these page categories.

Weekend: weekends are represented by 1, and non-weekends are represented by 0.

Revenue: The presence of buyers means revenue, the presence of revenue is represented by 1, and the absence of revenue is represented by 0.

6 Data Analysis and Findings

6.1 Descriptive Statistics

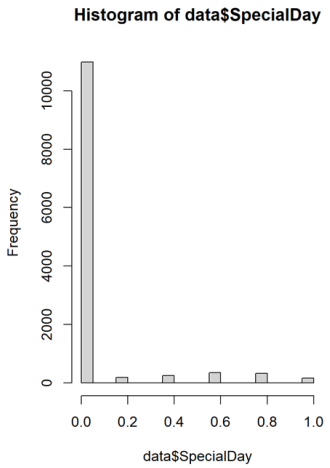

Fig. 1. Histogram of SpecialDay. [Owner-draw]

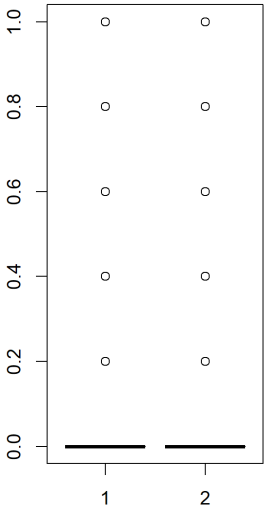

Fig. 2. Boxplot of SpecialDay. [Owner-draw]

As Figure 1 and Figure 2 show, the SpecialDay values are mainly distributed at 0, and it appears from the boxplot graph that whether or not to buy does not make a significant difference to the value of the SpecialDay.

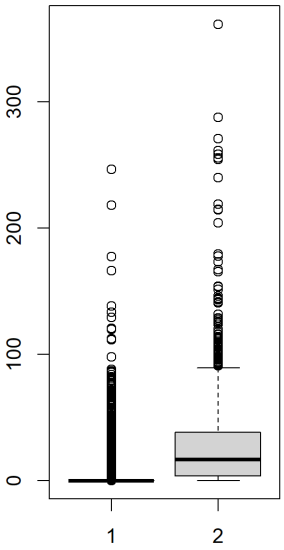

Fig. 3. Boxplot of PageValues. [Owner-draw]

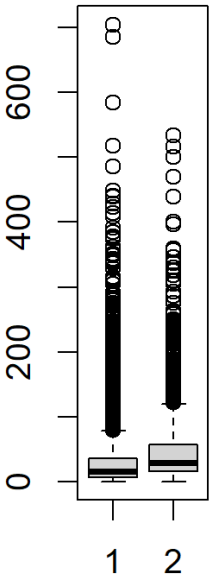

Fig. 4. Boxplot of ProductRelated. [Owner-draw]

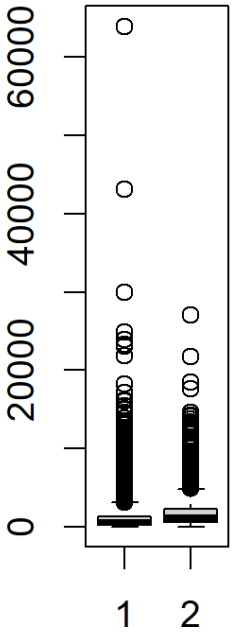

Fig. 5. Boxplot of ProductRelated_Duration. [Owner-draw]

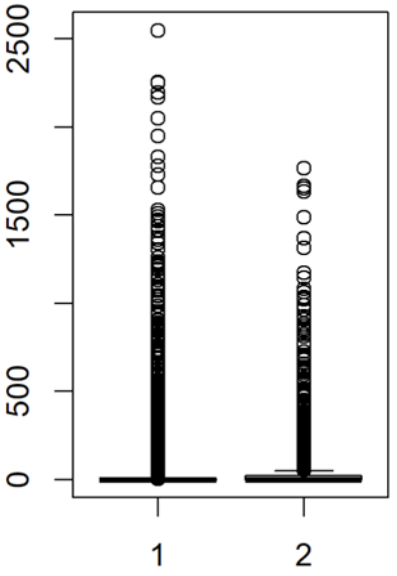

Fig. 6. Boxplot of Informational_Duration. [Owner-draw]

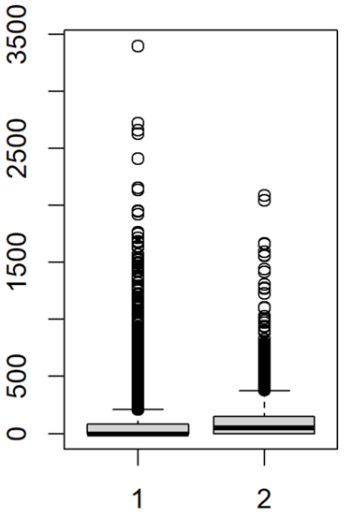

Fig. 7. Boxplot of Administrative_Duration. [Owner-draw]

As can be seen from Figure 3 to Figure 7, each independent variable has many outliers, even up to 10% of the sample size, so the work just leaves them. Apart from that, all variables are divided into two groups according to whether they purchased or not, some of the variables seem significantly different and some are less so, so a t-test is required to further investigate if there is a significant difference.

Table 1. Correlation matrix. [Owner-draw]

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	PageValues	SpecialDay	VisitorType	Weekend	Revenue
Administrative	1										
Administrative_Duration	0.602760249	1									
Informational	0.37753223	0.30398178	1								
Informational_Duration	0.256052957	0.238399142	0.618961349	1							
ProductRelated	0.431880719	0.290464457	0.373344571	0.279715348	1						
ProductRelated_Duration	0.374776005	0.357259934	0.386930023	0.347328208	0.860696838	1					
PageValues	0.103447812	0.071302836	0.052323078	0.033178645	0.057855426	0.054523028	1				
SpecialDay	-0.095561765	-0.073886194	-0.048848324	-0.030955723	-0.02494039	-0.037160584	-0.064057281	1			
VisitorType	-0.027820005	-0.024886948	0.054239707	0.043703082	0.124222643	0.117633974	-0.110140592	0.083669501	1		

Weekend	0.026027512	0.015181558	0.035295459	0.023439521	0.01523189	0.006684283	0.015109066	-0.01756871 2	-0.046509476	1	
Revenue	0.138901107	0.093536567	0.095665683	0.070412204	0.158780166	0.152192565	0.494959234	-0.08246849 1	-0.104876337	0.029870514	1

As shown in Table 1, ProductRelated and ProductRelated_Duration are 86% correlated, which means they are highly correlated, so care should be taken to avoid multicollinearity in subsequent modelling.

7 Regression Analysis

7.1 T-test

The work did t-tests on each of these independent variables, grouped according to whether they were purchased or not, and found that they were all significantly different, which means that all of these independent variables are affected by the dependent variable Revenue.

(For detailed test results please see the appendix.)

7.2 Logit Model

Model1 was put into all the independent variables and Logit regression analysis was performed on the dependent variables. The regression results show that the independent variables Weekend, VisitorType, SpecialDay, PageValues, ProductRelated, Administrative are significant. However, since ProductRelated and ProductRelated_Duration are highly correlated, multicollinearity needs to be dealt with.

Model2 chose ProductRelated among ProductRelated and ProductRelated_Duration, AIC=7552.4.

Model3 chose ProductRelated_Duration among ProductRelated and ProductRelated_Duration.

However, AIC=7572.3, is larger than in Model2 (AIC=7552.4), so the model is not selected and ProductRelated_Duration is eliminated from the next model and only ProductRelated is retained, which means that the model continues to be optimised on the basis of Model2.

Model2 shows that the least significant variable is Administrative_Duration, so Model4 is further optimised by removing Administrative_Duration from Model2 with AIC=7550.5.

Model4 shows that the least significant variable is Information_Duration, so Model5 is further optimised by removing Information_Duration from Model4 with AIC=7548.6.

```
Call:
glm(formula = data$Revenue.1 ~ data$Weekend + data$VisitorType +
    data$SpecialDay + data$PageValues + data$ProductRelated +
    data$Informational + data$Administrative, family = binomial)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.238092   0.080898 -27.666  < 2e-16 ***
data$Weekend         0.171118   0.069903   2.448   0.0144 *
data$VisitorType    -0.619444   0.082112  -7.544 4.56e-14 ***
data$SpecialDay     -1.066971   0.214411  -4.976 6.48e-07 ***
data$PageValues      0.086343   0.002364  36.532  < 2e-16 ***
data$ProductRelated  0.007344   0.000608  12.079  < 2e-16 ***
data$Informational   0.033972   0.022345   1.520   0.1284
data$Administrative  0.018987   0.009204   2.063   0.0391 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10541.9  on 12244  degrees of freedom
Residual deviance:  7532.6  on 12237  degrees of freedom

AIC: 7548.6

Number of Fisher Scoring iterations: 6
```

Fig. 8. Outcome of Model5. [Owner-draw]

As shown in Figure 8, Model5 shows that the least significant variable is Informational, so Model6 excludes Informational from Model5 and has an AIC = 7548.9, which is slightly higher as compared to Model5, so Model5 is finally chosen as the optimal model as it has the smallest value of AIC, indicating that it has the fits the best.

(For detailed test results please see the appendix.)

8 Discussion

8.1 Interpretation of Findings

```
                     Estimate Std. Error      z value Pr(>|z|)
(Intercept)         0.1066619   1.084260 9.659106e-13 1.000000
data$Weekend        1.1866310   1.072405 1.156433e+01 1.014472
data$VisitorType    0.5382437   1.085578 5.293410e-04 1.000000
data$SpecialDay     0.3440492   1.239132 6.899612e-03 1.000001
data$PageValues     1.0901797   1.002366 7.340673e+15 1.000000
data$ProductRelated 1.0073708   1.000608 1.761793e+05 1.000000
data$Informational  1.0345554   1.022597 4.573679e+00 1.137043
data$Administrative 1.0191688   1.009246 7.869893e+00 1.039883
```

Fig. 9. Interpretation of Model5. [Owner-draw]

Figure 9 presents the extent to which these factors affect the dependent variable, Revenue.

Weekend: The data analysis reveals a noteworthy 19% increase in potential buyers on weekends compared to weekdays. This can be attributed to the fact that people generally have more leisure time during weekends, which allows them to spend more time browsing through online catalogs. It is evident that weekdays are more hectic for individuals, and online shopping may not be a priority during these days. However, during weekends, people are more likely to explore online shopping as a means of leisure and stress release.

Returning Visitors: The research indicates a significant 46% decrease in potential buyers among returning visitors when compared to new visitors. This decline can be mainly attributed to the fact that returning visitors have already made their necessary purchases and may not need to buy again in the short term. However, it is important to consider that poor customer loyalty may also contribute to this decrease. Businesses should focus on building stronger customer relationships to encourage repeat purchases.

Special Days: On special days, potential buyers experience a substantial 66% decrease compared to non-special days. This phenomenon can be explained by the consumer behavior of planning ahead for special occasions to ensure timely delivery. As special days approach, concerns about delivery times may prompt customers to make their transactions earlier. Furthermore, people are less likely to focus on online shopping as they prioritize preparations for and celebration of the special day.

Page Value: The analysis demonstrates that as page value increases by 1, potential buyers show a corresponding increase of 9%. This finding indicates that higher page value, which could be attributed to better content quality or more relevant product offerings, positively influences potential buyers' purchase decisions. Users are more likely to make a purchase when they perceive the page content as valuable and relevant to their needs.

As shown in the formula (1), the higher Page Value gets, the lower Number of Unique Pageviews for Pages gets, therefore the outcome means that for a given revenue, the more times a user views a particular web page, the less likely they are to make a purchase. Too many navigation links on a webpage cause confusion for consumers and lead to the customers leaving to competitors.

Product Related: The data analysis reveals a positive correlation between the number of ProductRelated pages visited and the potential buyers. Specifically, for every additional ProductRelated page viewed, there is a 0.7% increase in potential buyers.This finding indicates that as users browse through more ProductRelated web pages, they gain access to additional details about the products they are interested in. Consequently, a higher level of product knowledge leads to an increased likelihood of customers making a purchase. Although the increase is 0.7%, which is relatively minor, it is not deemed substantial.

8.2 Limitations of the Study

The treatment of outliers, whether to retain or remove them, remains a debatable aspect in the analysis. Examining specific data entries is crucial to determine the nature of extreme values. Retaining outliers with valid reasons can provide valuable insights into

special circumstances related to the variable. However, illogical outliers can compromise the credibility of the sample data and should be removed, even if they constitute a significant proportion, such as 10% of the total.

The validity of the logit model could have been further strengthened through additional testing, such as conducting chi-square tests. Incorporating such assessments would have provided a more comprehensive evaluation of the model's performance and predictive capability.

The joint effect of variables on the dependent variable was not considered in this study. For instance, the relationship between Product Related and Page Value might influence each other. To explore these interactions, a new independent variable, such as Product Related * Page Value, could have been incorporated to fit a new model, yielding more nuanced insights.

The study acknowledges a limitation in terms of the theoretical foundation. While the data analysis and regression results provide valuable findings, a more robust theoretical framework could have enriched the interpretation of the results and facilitated deeper understanding of the observed relationships.

9 Conclusion

9.1 Summary of Key Findings

The study set out to find out the impact of various factors that affect online shopping on the customer's purchase results for the sake of helping online retailers maximize their profits. The thesis has provided a deeper insight into why variables such as weekend, visitor type, and special days affect consumer purchase behavior, and how online retailers can increase the number of buyers to achieve the goal of increasing revenue. The work use two methods of R and Excel to get the five factors that will affect the purchase behavior (Weekend, VisitorType, SpecialDay, PageValues, ProductRelated), but there are still limitations on issues such as treatment of Outliers and further testing of the validity. By implementing these strategic recommendations, online retailers can effectively attract and retain customers, drive sales during peak periods, and create a user-centric shopping environment, leading to long-term success and profitability in the competitive online retail landscape.

9.2 Recommendations for Online Retailers

In conclusion, several strategic approaches can be employed by online retailers to enhance their sales and customer loyalty.

Firstly, leveraging weekend promotion can capitalize on the increased social media usage during weekends. By offering targeted discounts and reward programs, such as points and coupons, retailers can effectively engage their target audience and boost sales.

Secondly, fostering customer loyalty through a tiered benefits system can bridge the gap between new and returning customers. This system, which includes VIP status and exclusive benefits, can lead to sustainable growth and a larger market share.

Additionally, special day promotion should be strategically designed with unique characteristics, offering customers ample time to plan their purchases and incorporating festival-themed incentives.

Furthermore, webpage optimization is crucial for enhancing user experience. Retailers should streamline webpage design, integrating essential product information while eliminating redundancy, ultimately reducing customer frustration and preventing potential loss due to perceived inconvenience or complexity.

Lastly, comprehensive product-related pages should offer informative and engaging content, empowering potential buyers to make informed decisions. Optimizing these pages will drive higher conversion rates and contribute to overall revenue growth.

By implementing these strategies cohesively, online retailers can effectively improve conversion rates and enhance their market position. These authors are listed in alphabetical order. They contributed equally to this work and should be considered co-first authors.

Appendix

Special Day----Buyers/Non-buyers

```
> ind.t.test<-t.test(data$SpecialDay ~ data$Revenue.1)
> ind.t.test

        Welch Two Sample t-test

data:  data$SpecialDay by data$Revenue.1
t = 12.941, df = 4174.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
95 percent confidence interval:
 0.03862953 0.05242392
sample estimates:
mean in group FALSE  mean in group TRUE
         0.06888824          0.02336152
```

Visitor Type----Buyers/Non-buyers

```
> ind.t.test<-t.test(data$VisitorType ~ data$Revenue.1)
> ind.t.test

        Welch Two Sample t-test

data:  data$VisitorType by data$Revenue.1
t = 9.9169, df = 2339.5, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
95 percent confidence interval:
 0.08037144 0.11999148
sample estimates:
mean in group FALSE  mean in group TRUE
          0.8771371           0.7769556
```

Page Values----Buyers/Non-buyers

```
> ind.t.test<-t.test(data$PageValues ~ data$Revenue.1)
> ind.t.test

        Welch Two Sample t-test

data:  data$PageValues by data$Revenue.1
t = -31.588, df = 1940.9, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
95 percent confidence interval:
 -26.28860 -23.21507
sample estimates:
mean in group FALSE  mean in group TRUE
           1.979402           26.731236
```

Product Related_Duration-----Buyers/Non-buyers

```
> ind.t.test<-t.test(data$ProductRelated_Duration ~ data$Revenue.1)
> ind.t.test

        Welch Two Sample t-test

data:  data$ProductRelated_Duration by data$Revenue.1
t = -14.366, df = 2328.6, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
95 percent confidence interval:
 -917.2191 -696.8882
sample estimates:
mean in group FALSE  mean in group TRUE
           1074.381            1881.434
```

Informational_Duration----Buyers/Non-buyers

```
> ind.t.test<-t.test(data$Informational_Duration ~ data$Revenue.1)
> ind.t.test

        Welch Two Sample t-test

data:  data$Informational_Duration by data$Revenue.1
t = -6.5926, df = 2330.1, p-value = 5.328e-11
alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
95 percent confidence interval:
 -35.67207 -19.31582
sample estimates:
mean in group FALSE  mean in group TRUE
           30.38243            57.87638
```

Administrative_Duration----Buyers/Non-buyers

```
> ind.t.test<-t.test(data$Administrative_Duration ~ data$Revenue.1)
> ind.t.test

        Welch Two Sample t-test

data:  data$Administrative_Duration by data$Revenue.1
t = -9.2838, df = 2413.6, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
95 percent confidence interval:
 -55.34678 -36.04322
sample estimates:
mean in group FALSE  mean in group TRUE
          73.88399           119.57899
```

Modle 1

```
Call:
glm(formula = data$Revenue.1 ~ data$Weekend + data$VisitorType +
    data$SpecialDay + data$PageValues + data$ProductRelated_Duration +
    data$ProductRelated + data$Informational_Duration + data$Informational +
    data$Administrative_Duration + data$Administrative, family = binomial)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.235e+00  8.100e-02 -27.598  < 2e-16 ***
data$Weekend                 1.735e-01  6.993e-02   2.481   0.0131 *
data$VisitorType            -6.245e-01  8.221e-02  -7.596 3.05e-14 ***
data$SpecialDay             -1.058e+00  2.143e-01  -4.936 7.97e-07 ***
data$PageValues              8.624e-02  2.363e-03  36.488  < 2e-16 ***
data$ProductRelated_Duration 4.870e-05  2.833e-05   1.719   0.0857 .
data$ProductRelated          5.633e-03  1.156e-03   4.875 1.09e-06 ***
data$Informational_Duration  2.480e-05  2.247e-04   0.110   0.9121
data$Informational           2.730e-02  2.698e-02   1.012   0.3115
data$Administrative_Duration -1.356e-04  1.978e-04  -0.686   0.4928
data$Administrative          2.335e-02  1.088e-02   2.146   0.0318 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10541.9  on 12244  degrees of freedom
Residual deviance:  7529.3  on 12234  degrees of freedom

AIC: 7551.3
```

Model 2

```
Call:
glm(formula = data$Revenue.1 ~ data$Weekend + data$VisitorType +
    data$SpecialDay + data$PageValues + data$ProductRelated +
    data$Informational + data$Informational_Duration + data$Administrative_Duration +
    data$Administrative, family = binomial)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.236e+00  8.098e-02 -27.617  < 2e-16 ***
data$Weekend                 1.707e-01  6.992e-02   2.442   0.0146 *
data$VisitorType            -6.202e-01  8.214e-02  -7.551 4.31e-14 ***
data$SpecialDay             -1.068e+00  2.145e-01  -4.981 6.31e-07 ***
data$PageValues              8.634e-02  2.363e-03  36.530  < 2e-16 ***
data$ProductRelated          7.330e-03  6.092e-04  12.034  < 2e-16 ***
data$Informational           2.840e-02  2.676e-02   1.061   0.2886
data$Informational_Duration  9.110e-05  2.178e-04   0.418   0.6757
data$Administrative_Duration -4.492e-05  1.851e-04  -0.243   0.8083
data$Administrative          2.023e-02  1.063e-02   1.902   0.0571 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10541.9  on 12244  degrees of freedom
Residual deviance:  7532.4  on 12235  degrees of freedom

AIC: 7552.4

Number of Fisher Scoring iterations: 6
```

Model 3

```
Call:
glm(formula = data$Revenue.1 ~ data$Weekend + data$VisitorType +
    data$SpecialDay + data$PageValues + data$ProductRelated_Duration +
    data$Informational + data$Informational_Duration + data$Administrative_Duration +
    data$Administrative, family = binomial)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.228e+00  8.091e-02 -27.532  < 2e-16 ***
data$Weekend                 1.790e-01  6.976e-02   2.566 0.010291 *
data$VisitorType            -6.047e-01  8.200e-02  -7.374 1.66e-13 ***
data$SpecialDay             -1.033e+00  2.139e-01  -4.831 1.36e-06 ***
data$PageValues              8.574e-02  2.353e-03  36.441  < 2e-16 ***
data$ProductRelated_Duration 1.671e-04  1.517e-05  11.013  < 2e-16 ***
data$Informational           3.304e-02  2.702e-02   1.223 0.221316
data$Informational_Duration -5.745e-05  2.255e-04  -0.255 0.798941
data$Administrative_Duration -2.905e-04  1.996e-04  -1.456 0.145426
data$Administrative          3.610e-02  1.059e-02   3.408 0.000654 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10541.9  on 12244  degrees of freedom
Residual deviance:  7552.3  on 12235  degrees of freedom

AIC: 7572.3

Number of Fisher Scoring iterations: 5
```

Model 4

```
Call:
glm(formula = data$Revenue.1 ~ data$Weekend + data$VisitorType +
    data$SpecialDay + data$PageValues + data$ProductRelated +
    data$Informational + data$Informational_Duration + data$Administrative,
    family = binomial)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.237e+00  8.093e-02 -27.644  < 2e-16 ***
data$Weekend                1.709e-01  6.991e-02   2.445   0.0145 *
data$VisitorType           -6.198e-01  8.212e-02  -7.548 4.43e-14 ***
data$SpecialDay            -1.067e+00  2.144e-01  -4.978 6.43e-07 ***
data$PageValues             8.633e-02  2.363e-03  36.531  < 2e-16 ***
data$ProductRelated         7.328e-03  6.094e-04  12.026  < 2e-16 ***
data$Informational          2.815e-02  2.676e-02   1.052   0.2928
data$Informational_Duration 8.644e-05  2.174e-04   0.398   0.6909
data$Administrative         1.893e-02  9.205e-03   2.056   0.0398 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10541.9  on 12244  degrees of freedom
Residual deviance:  7532.5  on 12236  degrees of freedom

AIC: 7550.5

Number of Fisher Scoring iterations: 6
```

Model 5

```
Call:
glm(formula = data$Revenue.1 ~ data$Weekend + data$VisitorType +
    data$SpecialDay + data$PageValues + data$ProductRelated +
    data$Informational + data$Administrative, family = binomial)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.238092   0.080898 -27.666  < 2e-16 ***
data$Weekend        0.171118   0.069903   2.448   0.0144 *
data$VisitorType   -0.619444   0.082112  -7.544 4.56e-14 ***
data$SpecialDay    -1.066971   0.214411  -4.976 6.48e-07 ***
data$PageValues     0.086343   0.002364  36.532  < 2e-16 ***
data$ProductRelated 0.007344   0.000608  12.079  < 2e-16 ***
data$Informational  0.033972   0.022345   1.520   0.1284
data$Administrative 0.018987   0.009204   2.063   0.0391 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10541.9  on 12244  degrees of freedom
Residual deviance:  7532.6  on 12237  degrees of freedom

AIC: 7548.6

Number of Fisher Scoring iterations: 6
```

Model 6

```
Call:
glm(formula = data$Revenue.1 ~ data$Weekend + data$VisitorType +
    data$SpecialDay + data$PageValues + data$ProductRelated +
    data$Administrative, family = binomial)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.2413915  0.0808931 -27.708  < 2e-16 ***
data$Weekend        0.1738115  0.0698708   2.488   0.0129 *
data$VisitorType   -0.6133984  0.0820082  -7.480 7.45e-14 ***
data$SpecialDay    -1.0720249  0.2143513  -5.001 5.70e-07 ***
data$PageValues     0.0864267  0.0023640  36.559  < 2e-16 ***
data$ProductRelated 0.0075682  0.0005903  12.820  < 2e-16 ***
data$Administrative 0.0226513  0.0088613   2.556   0.0106 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10541.9  on 12244  degrees of freedom
Residual deviance:  7534.9  on 12238  degrees of freedom

AIC: 7548.9

Number of Fisher Scoring iterations: 6
```

References

1. Berg, Annelieke. (2018). "What Is Page Value In Google Analytics? • Yoast". https: // yoast. com/what-is-page-value-in-google-analytics/.
2. Examples, KPI, and Total value. (2023). " What Is Total Goal Value And How To Calculate It | Dashthis ". https://dashthis.com/kpi-examples/total-goal-value/.
3. Gao L. (2021). Research on the strategy of enhancing users' purchase intention in Company A's online shopping mall. https: // kns. cnki. net/ KCMS/ detail/detail. aspx? Dbname =CMFD202301&filename=1021104937.nh
4. Moe, Wendy W., and Peter S. Fader. (2004), "Capturing Evolving Visit Behavior in Click-stream Data." Journal of Interactive Marketing 18, no. 1: 5–19.
5. Park, Young-Hoon, and Peter S Fader. (2004), "Modeling Browsing Behavior at Multiple Websites." Marketing Science (Providence, R.I.) 23, no. 3: 280–303.
6. Meer, Geoffrey Van. (2006), "Customer Development and Retention on a Web-Banking Site." Journal of Interactive Marketing 20, no. 1: 58–64.
7. Jia Hu and Ning Zhong, (2005), "Clickstream log acquisition with Web farming," The 2005 IEEE/WIC/ACM International Conference. In: Compiegne, France, pp. 257-263
8. R. Hanamanthrao and S. Thejaswini, (2017), "Real-time clickstream data analytics and visualization," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), In: Bangalore, India, pp. 2139-2144
9. Ghavamipoor, H., Hashemi Golpayegani, S.A. and Shahpasand, M. (2017), "A QoS-sensitive model for e-commerce customer behavior", Journal of Research in Interactive Marketing, Vol. 11 No. 4, pp. 380-397.
10. Baumann, Annika, Johannes Haupt, Fabian Gebert, and Stefan Lessmann. (2019). "The Price of Privacy: An Evaluation of the Economic Value of Collecting Clickstream Data." Business & Information Systems Engineering 61, no. 4: 413–431.
11. R. Somya, E. Winarko and S. Privanta, (2021), "A Novel Approach to Collect and Analyze Market Customer Behavior Data on Online Shop," 2021 2nd International Conference on Innovative and Creative Information Technology (ICITech), In: Salatiga, Indonesia, pp. 151-156
12. M. Gumber, A. Jain and A. L. Amutha, (2021), "Predicting Customer Behavior by Analyzing Clickstream Data," 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP), In: Chennai, India, pp. 1-6
13. Necula, S.-C. (2023), Exploring the Impact of Time Spent Reading Product Information on E-Commerce Websites: A Machine Learning Approach to Analyze Consumer Behavior. Behav. Sci.13, 439.
14. Li Guoxin, Li Yijun, Li Bing, et al. (2012) Influencing factors and evolution of online shopping behavior of e-commerce users in my country: An empirical study based on data from 2006 and 2009 [J]. Management Review, 24(07): 56- 62.
15. Al Hamli, S.S.; Sobaih, A.E.E. (2023) Factors Influencing Consumer Behavior towards Online Shopping in Saudi Arabia Amid COVID-19: Implications for E-Businesses Post Pandemic. J. Risk Financial Manag. 16(1), 36.
16. Daroch, B., Nagrath, G. and Gupta, A. (2021), "A study on factors limiting online shopping behaviour of consumers", Rajagiri Management Journal, Vol. 15 No. 1, pp. 39-52.
17. "UCIMachineLearningRepository". (2023).Archive. Ics. Uci. Edu. https:// archive. ics. uci. edu/dataset/468/on