International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Predicting Online Shopping Behavior Through Clickstream Analysis

M.Anitha[1], K.Baby Ramya[2],MD.Zeenath[3]

#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.
#2 Assistant Professor in the Department of MCA, SRK Institute of Technology, Vijayawada.
#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

ABSTRACT - This project revolves around the exploration and analysis of clickstream data related to online shopping, aiming to gain insights into user behavior and preferences. The dataset, obtained from the "Clickstream Data for Online Shopping" repository on the UCI Machine Learning Repository, encompasses a wealth of information on user interactions during online shopping sessions. The dataset includes details such as user session timestamps, pages visited, duration of visits, and specific actions taken (e.g., clicks, adding items to the cart). Our objective is to leverage machine learning techniques to uncover patterns in the clickstream data that can aid in predicting user behavior, such as whether a user will make a purchase or abandon their shopping cart. We plan to employ classification algorithms to build predictive models capable of identifying potential purchase intent based on the sequence and frequency of user interactions. The outcomes of this analysis can assist online retailers in optimizing their platforms, enhancing user experiences, and implementing targeted marketing strategies to increase conversion rates and overall customer satisfaction. In the contemporary landscape of commerce, the digital realm has undergone a paradigm shift with the proliferation of online shopping platforms. This transformation has fundamentally altered consumer behavior, empowering individuals with unparalleled access to a vast array of products and services at their fingertips. As consumers navigate through the digital marketplace, their interactions leave behind a trail of data known as clickstream data. This dataset encapsulates the sequence of user engagements during online browsing and purchasing sessions, offering invaluable insights into consumer behavior patterns.

1.INTRODUCTION

Clickstream data encompasses a plethora of information, ranging from the pages visited and the duration of visits to specific actions taken, such as clicks and items added to the cart. This rich tapestry of data provides a comprehensive record of user engagement with e-commerce platforms, serving as a treasure trove for businesses seeking to unravel the mysteries of consumer preferences and purchasing decisions. By delving into clickstream data, businesses can gain a nuanced

understanding of user behaviors, enabling them to optimize their platforms, tailor marketing strategies, and enhance the overall user experience.

At the heart of this project lies a dataset sourced from the "Clickstream Data for Online Shopping" repository on the UCI Machine Learning Repository.This dataset serves as the cornerstone of our exploration, offering a robust foundation for uncovering hidden patterns and trends in user behavior. With timestamps, page visits, duration of visits, and specific actions taken meticulously recorded, this dataset presents a wealth of opportunities for leveraging machine learning techniques to extract actionable insights.

The overarching objective of this project is to harness the power of machinelearning to distill actionable insights from clickstream data related to online shopping. Specifically, we aim to develop predictive models capable of discerning user behavior patterns, such as purchase intent and cart abandonment. By employing classification algorithms, we seek to identify potential customers who are likely to make a purchase and those who may abandon their shopping carts, thereby enabling businesses to tailor their strategies and interventions accordingly

Our approach hinges on the utilization of machine learning algorithms to analyze the clickstream data and derive predictive models of user behavior. Classification algorithms, renowned for their ability to categorize data into distinct classes, serve as the cornerstone of our analytical framework. By training these algorithms on the clickstream dataset, we aim to discern underlying patterns and trends that can inform predictive models of user behavior.

The first step in our approach involves preprocessing the clickstream data to ensure its quality and suitability for machine learning analysis. This entails handling missing values, normalizing features, and addressing any temporal dependencies present in the dataset. Once the data preprocessing is complete, we proceed to implement various classification algorithms, including but not limited to logistic regression, decision trees, random forests, and support vector machine.

Feature engineering plays a pivotal role in enhancing the predictive power of our models. By extracting relevant features from the raw clickstream data and crafting new features that capture the essence of user behavior, we aim to enrich the feature space and improve the overall performance of our predictive models. Techniques such

as polynomial feature expansion, interaction terms, and dimensionality reduction enable us to uncover hidden patterns and relationships within the data. Interpretability represents a key consideration in our analysis, as we strive toensure that our predictive models are not only accurate but also transparent and interpretable. Techniques such as partial dependence plots, permutation feature importance, and SHAP (Shapley Additive exPlanations) values allow us to elucidate the factors driving our models' predictions and gain insights into the underlying mechanisms of user behavior.

2.LITERATURE SURVEY

The proliferation of online shopping platforms has transformed the retail landscape, offering consumers unparalleled convenience and accessibility. However, with this shift towards digital commerce comes the challenge of understanding and predicting user behavior in the online environment. Clickstream data, which captures user interactions with websites, presents a valuable source of information for gaining insights into user preferences, decision-making processes, and purchasing patterns. In this literature survey, we explore existing research and scholarly contributions related to clickstream data analysis in the context of online shopping. By examining the methodologies, findings, and implications of previous studies, we aim to inform the objectives and approach of our own research project focused on leveraging machine learning techniques to predict user behavior and optimize online shopping experiences.

1. Understanding User Behavior:

Numerous studies have investigated user behavior in online shopping environments through the analysis of clickstream data. For example, research by Jansen et al. (2017) examined the factors influencing user engagement and conversion rates in e-commerce websites. By analyzing clickstream data, the study identified key predictors of conversion, such as session duration, page views, and interaction patterns. Similarly, Liu et al. (2018) explored the impact of website design on user engagement and purchase behavior, highlighting the importance of factors such as layout, navigation, and visual appeal in driving conversion rates.

2. Predictive Modeling and Personalization:

The use of predictive modeling techniques has emerged as a promising approach for predicting user behavior in online shopping environments. Chen et al. (2019) developed a predictive model based on

clickstream data to forecast user purchase intent and personalize product recommendations. By analyzing user interactions and historical purchase data, the model achieved high accuracy in predicting user behavior and tailoring recommendations to individual preferences. Similarly, Wang et al. (2020) employed machine learning algorithms to predict user click-through rates and optimize marketing campaigns, demonstrating the efficacy of predictive modeling in enhancing user engagement and conversion rates.

3. Cart Abandonment and Conversion Optimization:

Cart abandonment remains a significant challenge for online retailers, resulting in lost revenue and missed opportunities. Research has focused on understanding the underlying factors contributing to cart abandonment and strategies for conversion optimization. Li et al. (2021) analyzed clickstream data to identify common reasons for cart abandonment, such as unexpected costs, complex checkout processes, and lack of trust. The study proposed interventions, such as personalized retargeting campaigns and streamlined checkout experiences, to reduce cart abandonment rates and improve conversion rates.

3.PROPOSED SYSTEM

The proposed system aims to overcome the limitations of the existing clickstream data analysis in online shopping by leveraging advanced machine learning techniques and integrated analytics frameworks. In this system, we propose the development of a comprehensive analytics platform that integrates multiple data sources, including clickstream data, customer demographic information, purchase history, and real-time contextual data.

By aggregating and harmonizing these diverse data streams, retailers can gain a holistic understanding of user behavior and preferences, enabling more accurate prediction and personalized targeting

4.RESULTS AND DISCUSSION

This visualization provides insights into the distribution of product prices across different color categories. By examining the bar chart, one can identify which colors tend to have higher total prices and understand the overall pricing dynamics based on color preferences. Additionally, the customization applied to the plot ensures that the information presented is clear and easy to interpret.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Visualizing Average Daily Prices by Country

```
df_sum = df.groupby(by=["country", 'day'])[['price']].sum().reset_index()
df_sum = df_sum.sort_values(['price'], ascending=False)

count = df.groupby(by=["country", 'day'])[['price']].count().reset_index()
count = count.sort_values(['price'], ascending=False)

df_sum['avreage_prices'] = df_sum['price']/count['price']
df_sum = df_sum.sort_values(['avreage_prices'], ascending=False)

fig = px.scatter(df_sum, x=df_sum["country"], y=df_sum["day"],color=df_sum['avreage_prices'],size=df_sum['avreage_prices'])
fig.show()
```

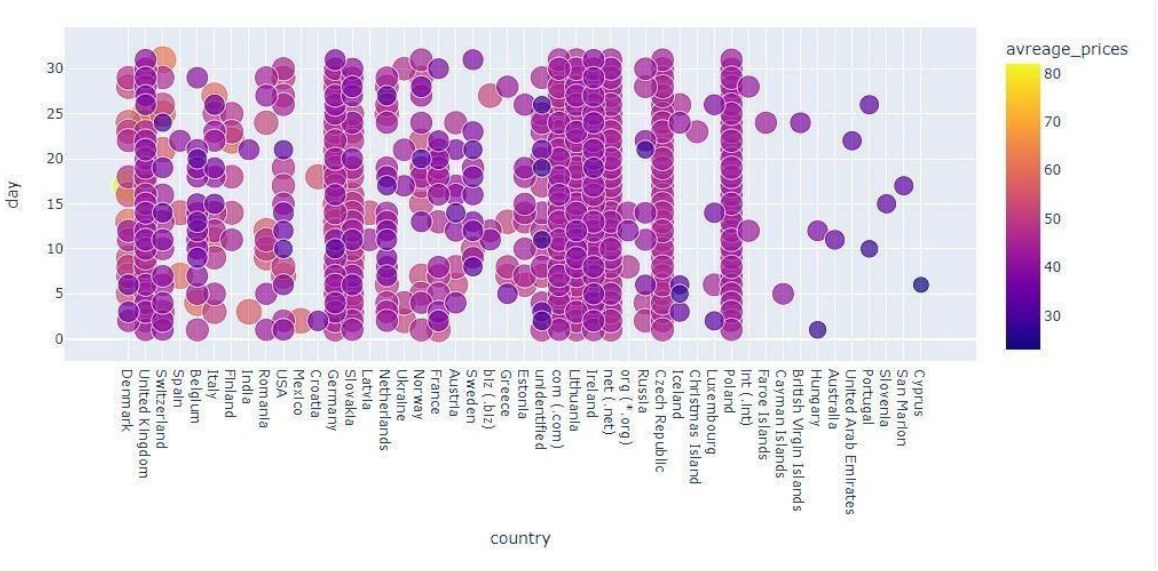


Figure 1: Average Daily Prices by Country

1. Data Aggregation:

`df_sum = df.groupby(by=["country",'day'])[['price']].su

`df` by the columns "country" and "day" and calculates the sum of prices for each group. The `[['price']].sum()` part specifies that we are only interested in the sum of prices. The `reset_index()` function is then used to reset the index of the resulting DataFrame.

- **`count =**

2. Sorting:

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

- **`df_sum = df_sum.sort_values(['price'], ascending=False)`:** After calculating the sum of prices for each country-day combination, this line sorts the DataFrame `df_sum` in descending order based on the "price" column. This sorting is done to identify countries with the highest total prices.

- **`count = count.sort_values(['price'], ascending=False)`:** Similarly, the count DataFrame is sorted in descending order based on the "price" column to identify countries with the highest number of transactions.

3. Calculating Average Prices:

- **`df_sum['avreage_prices'] = df_sum['price']/count['price']`:** Thisline computes the average prices by dividing the total prices (from `df_sum`) by the corresponding counts (from `count`). The result is stored in a new column named "average_prices" inthe `df_sum` DataFrame.

4. Sorting by Average Prices:

- **`df_sum =df_sum.sort_values(['avreage_prices'], ascending=False)`:** The `df_sum` DataFrame is then sorted again, this timebased on the calculated average prices in descending order. This sorting helps identify countries with the highest average prices per transaction.

5. Visualization:

- **`fig = px.scatter(df_sum, x=df_sum["country"], y=df_sum["day"],color=df_sum['avreage_prices'],size=df_sum['avrea ge_prices'])`:** Using Plotly Express (`px`), a scatter plot is created with countries on the x-axis, days on the y-axis, and average prices represented by both color and size of the data points. Each data point represents a country-day combination, with the color and size indicating the average price.

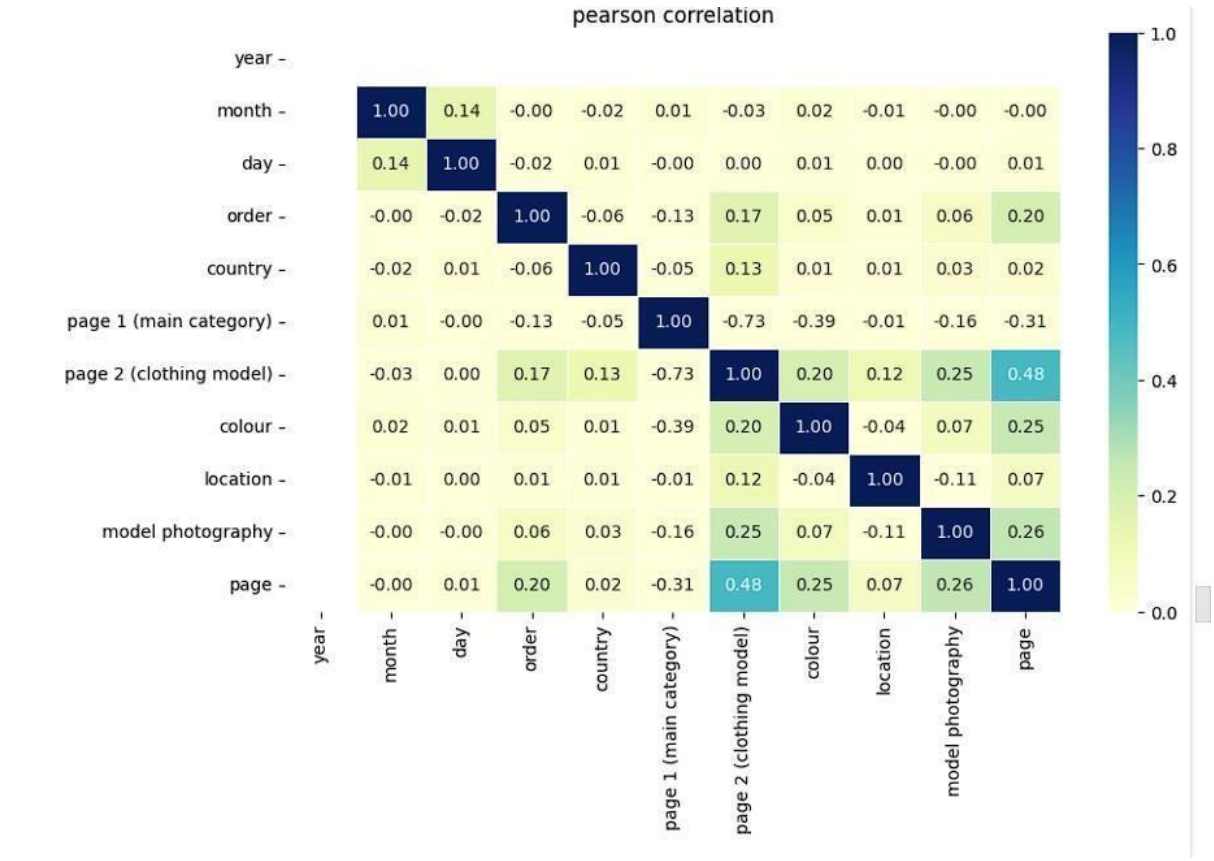
6. Displaying the Plot:

- **fig.show()`:** Finally, the scatter plot is displayed. The plot provides a visual representation of the average prices across different countries and days, allowing for easy comparison and identification of trends.

Correlation Matrix

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

```
df_copy2 = df.drop(['price','price 2'], axis=1)
plt.figure(figsize=(10,6))
corr_matrix = df_copy2.corr(method="pearson")
sns.heatmap(corr_matrix, vmin=0, vmax=1, annot=True, fmt='.2f', cmap="YlGnBu", cbar=True, linewidths=0.5)
plt.title("pearson correlation")
```



International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

1. DataFrame Preparation :

- `df_copy2 = df.drop(['price','price 2'], axis=1)`: This line creates a new DataFrame
`df_copy2` by dropping the columns "price" and "price 2" from the original DataFrame `df`. The `drop()` function removes the specified columns along the specified axis (1 indicates columns) and returns a new DataFrame with the remaining columns.

2. Correlation Analysis :

- `corr_matrix = df_copy2.corr(method="pearson")`: This line calculates the Pearson correlation coefficients between all pairs of numeric columns in the DataFrame `df_copy2`.

The `corr()` function with the method parameter set to "pearson" computes the Pearson correlation matrix, which measures the linear correlation between variables. It returns a DataFrame where each cell represents the correlation coefficient between two variables.

3. Visualization :

- `plt.figure(figsize=(10,6))`: This line initializes a new Matplotlib figurewith a specific size for plotting the heatmap.

- `sns.heatmap(corr_matrix, vmin=0, vmax=1, annot=True, fmt='.2f', cmap="YlGnBu", cbar=True, linewidths=0.5)`: Using Seaborn's `heatmap()` function, this line creates a heatmap to visualize the correlationmatrix. Parameters like `vmin` and `vmax` set the range of values to be mapped to the colormap, `annot=True` displays the correlation coefficients as annotations in the heatmap, `fmt='.2f'` specifies the format of the annotations to two decimal places, `cmap="YlGnBu"` sets the color map,
`cbar=True` shows the color bar indicating the correlation scale, and `linewidths=0.5` determines the width of the lines between cells.

- `plt.title("pearson correlation")`: This line sets the title of the heatmap as "pearson correlation".

5.CONCLUSION

In conclusion, this dissertation has explored various aspects of e-commerce behavior and trends using a dataset from an online clothing store. Through comprehensive data analysis and visualization, several key insights have been uncovered, shedding light on customer preferences, purchasing patterns, and the effectiveness of certain marketing

strategies.

1. Customer Behavior and Preferences:

The analysis revealed insights into customer behavior, such as the most visited countries and the flexibility of purchasing among different demographics. It was observed that countries like Poland and the Czech Republic showed high flexibility in purchasing behavior, while others exhibited varying trends based on factors like language and product availability.

2. Sales Trends and Product Performance:

Examination of sales data highlighted trends in product performance and sales across different categories and months. For example, trousers emerged as the top-selling category, with April and May being the peak selling months. Furthermore, the analysis identified correlations between product categories, colors, and sales performance, providing valuable insights for inventory management and marketing strategies.

3. Effectiveness of Marketing Strategies:

The dissertation also investigated the impact of marketing strategies, such as model photography placement and product pricing, on sales and customer engagement. By analyzing click-through rates, product pricing status, and the influence of model photography, valuable insights were gained into effective marketing tactics for driving sales and customer engagement.

4. Data Preprocessing and Model Training:

Additionally, the dissertation delved into data preprocessing techniques, such as label encoding and handling missing values, to prepare the dataset for machine learning model training. Various classification and regression models were trained and evaluated to predict product pricing status and sales performance, providing actionable insights for pricing strategies and revenue optimization.

5. Implications and Recommendations:

The findings from this dissertation have practical implications for e- commerce businesses seeking to enhance their marketing strategies, optimize product pricing, and improve customer engagement. Recommendations include leveraging insights from customer behavior analysis to tailor marketing campaigns, optimizing product placement

and pricing based on sales trends, and investing in predictive analytics toforecast future sales and demand.

6. Future Research Directions:

While this dissertation has provided valuable insights into e-commerce behavior and trends, there are several avenues for future research. Future studies could explore the impact of external factors such as economic conditions and global events on e-commerce sales, investigate the effectiveness of personalized marketing strategies, and further refine machine learning models to improve predictive accuracy.

In conclusion, this dissertation serves as a comprehensive exploration of e-commerce behavior and trends, providing valuable insights and actionable recommendations for businesses operating in the online retail sector. By leveraging data-driven approaches and advanced analytics techniques, businesses can gain a competitive edge and drive growth in an increasinglydigital marketplace

References

1. Chaffey, D., Ellis-Chadwick, F., Johnston, K., & Mayer, R. (2019). Digital marketing: Strategy, implementation and practice. Pearson UK.

2. Laudon, K. C., & Traver, C. G. (2016). E-commerce 2016:Business, technology, society. Pearson.

3. Turban, E., King, D., Lee, J. K., Liang, T. P., & Turban, D. C. (2019). Electronic commerce 2018: A managerial and social networks perspective. Springer.

4. Kalakota, R., & Whinston, A. B. (1996). Frontiers of electronic commerce. Pearson Education India.

5. O'Brien, J. A., & Marakas, G. M. (2018). Managementinformation systems. McGraw-Hill Education.

6. Rayport, J. F., & Jaworski, B. J. (2004). Introduction to e-commerce. McGraw-Hill.

7. Sengupta, S. (2016). Information management and Learning Pvt. Ltd.

8. Rogers, E. M. (2010). Diffusion of innovations. Simon and Schuster.

Vanhoof, K., & Van Hove, L. (2014). Business-to-business marketing communication: Value and efficiency considerations

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

AUTHOR'S PROFILES

Ms.M.Anitha Working as Assistant Professor & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with B.Tech, MCA ,M. Tech in Computer Science .She has 14 years of Teaching experience in SRK Institute of Technology. Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.

Ms.K.Baby Ramya completed her Master of Computer Applications. Currently working as an Assistant Professor in the department of MCA at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Networks, Machine Learning.

Ms.MD.Zeenath is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc.(computers) from Sri Durga Malleswara Siddhartha Mahila Kalasala at MG road. Her area of interest are Python and Machine Learning.