UNIVERSITY OF HOUSTON

EDS 6340 Introduction to Data Science

Group_10 Project Report

Analysis of Online Shopping Behavior using Clickstream Data for Online Shopping

Abstract

This research delves into the analysis of clickstream data obtained from an online store specializing in maternity clothing over a span of five months in 2008. The dataset encompasses various parameters including product category, photo placement on the webpage, IP address country of origin, and product pricing in US dollars. The primary objective of this study is to examine user behavior patterns within the e-commerce platform, specifically focusing on a comparative analysis of consumer purchasing habits between Poland and other European countries. Through the lens of web mining, particularly web usage mining, this research aims to dissect user behavior patterns online, with a specific focus on e-commerce platforms. Market basket analysis serves as a pivotal method for scrutinizing shopping patterns to enhance consumer satisfaction and bolster sales figures. The study employs methodologies such as association rules and sequence analysis to uncover underlying patterns in consumer behavior.

Utilizing transactional data and user preferences gleaned from a women's clothing e-shop, this research scrutinizes discrepancies between Polish clientele and those hailing from other European regions. Descriptive statistics unveil disparities in trendy clothing models and color preferences across distinct customer segments. Association rules analysis sheds light on product selection trends, elucidating favored clothing combinations among diverse customer groups. Sequence analysis unveils discrepancies in browsing patterns and decision-making processes between European and non-European customers. This research underscores the pivotal role of web usage mining in e-commerce studies, emphasizing its significance in comprehending user behavior and transactional trends. Furthermore, it discusses the practical applications of these findings in website enhancement, functionality refinement, and personalized marketing endeavors. Future research avenues are suggested, including expanding the analysis to encompass different website language versions and exploring additional analytical tools such as Markov chains for deeper insights.

Introduction

The advent of e-commerce has revolutionized the retail landscape, offering consumers unprecedented convenience and accessibility to a plethora of goods and services. With the exponential growth of online shopping platforms, understanding consumer behavior within these digital environments has become paramount for businesses striving to remain competitive and enhance profitability. This research embarks on a comprehensive exploration of user behavior patterns within an e-commerce ecosystem, utilizing clickstream data from a maternity clothing online store as the focal point of analysis.

The dataset under examination encapsulates a myriad of variables, ranging from product categorization to IP address geolocation, providing rich insights into user interactions and

preferences. By scrutinizing these data points, this study endeavors to unravel distinct patterns in consumer behavior, with a particular emphasis on discerning differences between Polish consumers and their counterparts from other European nations.

At the heart of this research lies the concept of web mining, a multifaceted approach aimed at extracting valuable insights from vast troves of web data. Within the realm of web mining, web usage mining emerges as a pivotal tool, enabling researchers to dissect user behavior patterns and glean actionable insights for businesses operating in the digital sphere. Central to this endeavor is the application of market basket analysis, a robust methodology for unraveling shopping patterns and identifying potential avenues for enhancing consumer satisfaction and driving sales growth.

Through the deployment of analytical techniques such as association rules and sequence analysis, this research endeavors to unearth hidden correlations and trends within the dataset. By elucidating nuanced differences in user behavior between Polish and other European consumers, this study aims to furnish e-commerce practitioners with actionable insights for refining website design, optimizing functionality, and devising personalized marketing strategies tailored to distinct customer segments.

As the digital landscape continues to evolve, the findings of this research are poised to inform future endeavors in e-commerce research, paving the way for enhanced understanding of user behavior and transactional dynamics in online retail environments. Moreover, by delineating potential avenues for further exploration, this study seeks to catalyze ongoing efforts to harness the power of web mining for driving business success in the digital age.

Data Description

Understanding user behavior in online environments necessitates the collection of diverse and detailed data points to unravel the intricacies of user interactions. The dataset utilized in this study spans a period of five months in 2008, encompassing information on clickstream activities within an e-commerce platform catering to maternity clothing. Each entry in the dataset encapsulates a wealth of specific details, providing insights into user behavior patterns, preferences, and decision-making processes.

- YEAR: The year 2008 serves as the temporal frame for the dataset, delineating the period under examination.
- MONTH: Ranging from April (4) to August (8), the MONTH variable categorizes data entries according to the chronological month of occurrence.
- DAY: This variable denotes the day number within the month, providing temporal granularity for user activities.

- ORDER: Reflecting the sequence of clicks during a single session, the ORDER variable offers insights into the chronological progression of user interactions within the e-commerce platform.
- COUNTRY: Indicating the country of origin of the IP address, the COUNTRY variable encompasses a diverse range of categories, spanning various geographic regions and areas.
- SESSION ID: Serving as a unique identifier for individual browsing sessions, the SESSION ID variable enables the aggregation of clickstream data at the session level.
- PAGE 1 (MAIN CATEGORY): This variable pertains to the main product category, including trousers, skirts, blouses, and sale items, facilitating the segmentation of user interactions based on product types.
- PAGE 2 (CLOTHING MODEL): Providing information about the specific product code for each item, the PAGE 2 variable enables granular tracking of user interactions at the product level.
- COLOUR: Reflecting the color of the purchased product, the COLOUR variable encompasses a range of color categories, capturing user preferences regarding product aesthetics.
- LOCATION: Denoting the position of product photos on the webpage, the LOCATION variable divides the screen into six distinct parts, offering insights into the visual presentation of products and its impact on user engagement.
- MODEL PHOTOGRAPHY: Categorized into "en face" and "profile," the MODEL PHOTOGRAPHY variable delineates the style of product photography employed on the e-commerce platform, potentially influencing user perceptions and purchase decisions.
- PRICE: Expressed in US dollars, the PRICE variable quantifies the cost of the purchased product, facilitating analyses of pricing dynamics and consumer spending patterns.
- PRICE 2: A binary variable indicating whether the price of a particular product exceeds the average price for the entire product category, offering insights into user preferences for premium-priced items.
- PAGE: Designating the page number within the e-store website, the PAGE variable aids in tracking user navigation patterns and assessing the efficacy of webpage layout and design.

This comprehensive array of variables provides a nuanced understanding of user behavior within the e-commerce platform, enabling the exploration of shopping patterns, preferences, and decision-making processes among diverse user segments. Through the meticulous analysis of these data points, this study seeks to elucidate key insights to inform strategic decision-making and enhance the overall user experience within the online retail environment.

Variable Table

Variable Name	Role	Type	Description	Units	Missing Values

year	Feature	Date	2008		no
month	Feature	Date	from April (4) to August (8)		no
day	Feature	Date	day number of the month		no
order	Feature	Integer	sequence of clicks during one session		no
country	Feature	Categorical	variable indicating the country of origin of the IP address		no
session ID	Feature	Integer	variable indicating session id (short record)		no
page 1 (main category)	Feature	Categorical	concerns the main product category		no
page 2 (clothing model)	Feature	Categorical	contains information about the code for each product (217 products)		no
color	Feature	Categorical	color of product		no
location	Feature	Categorical	photo location on the page, the screen has been divided into six parts		no
model photography	Feature	Categorical	variable with two categories		no
price	Feature	Integer	price	USD	no
price 2	Feature	Binary	variable informing whether the price of a particular product is higher than the average price for the entire product category		no
page	Feature	Integer	page number within the e-store website (from 1 to 5)		no

Overview of Table

Dataset Characteristics	**Multivariate, Sequential**
Subject Area	Business
Associated Tasks	Classification, Regression, Clustering
Feature Type	Integer, Real
Number of Instances	165474
Number of Features	14

Data Modeling

Preprocessing

One of the first steps in creating a model is making sure that the data set you are using is good. A good data set (clean data) is free of duplicates, does not have missing values, properly scaled (this depends on the type of model), and encoded (this also depends on the type of model). For this project's data cleaning, the following was implemented by group 10.

Search for missing values.

```
# Look for missing values

df.isna().sum()
```

```
year                        0
month                       0
day                         0
order                       0
country                     0
session ID                  0
page 1 (main category)      0
page 2 (clothing model)     0
colour                      0
location                    0
model photography           0
price                       0
price 2                     0
page                        0
dtype: int64
```

Search for duplicates.

```
# Look for duplicates

df.duplicated()
```

```
0         False
1         False
2         False
3         False
4         False
          ...
165469    False
165470    False
165471    False
165472    False
165473    False
Length: 165474, dtype: bool
```

Search for Linear Relationship between Variables

```
# Check the relationships between features

sns.pairplot(df)
```

Data Mapping

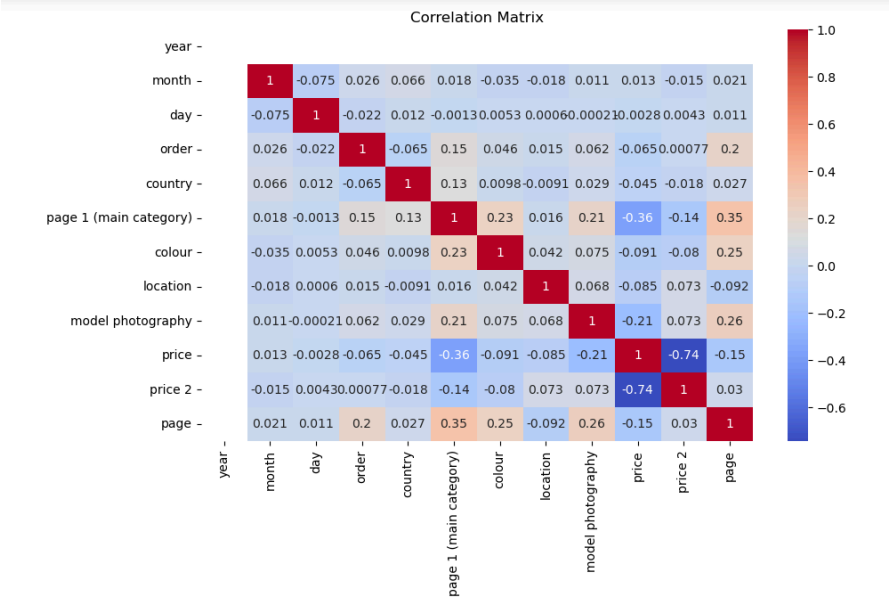
Some of the variables were given as integers. While this type of data is easy to model, there is a need to map them to their actual values. This helps us understand the data better. To map a column's data instances to another set of data instances, we use a python dictionary and the keyword 'map.' An example of the implementation is below.

```
model_map = {
    1: "En face",
    2: "Profile"}

df["model photography"] = df["model photography"].map(model_map)
```

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial phase in the data analysis process that involves exploring, summarizing, and visualizing data to gain insights, detect patterns, and identify relationships among variables. EDA serves as a foundational step before applying more complex statistical techniques or building predictive models. In the context of this project, EDA played a vital role in uncovering key trends, patterns, and correlations within the clickstream dataset from the maternity clothing e-shop. There were two main EDAs performed; a seaborn pair plot, and correlation matrix.



Models

KNN

The K-Nearest Neighbors (KNN) algorithm, a non-parametric method, was utilized to classify users based on their browsing and purchasing behaviors, leveraging similarities in feature space to identify clusters of similar users.

- Model training score: 97.43%.
- Model testing score: 96.07%.
- Model Accuracy: 96%.

Logistic Regression

This linear classification model will aid in predicting the likelihood of users from different countries purchasing specific clothing items, providing interpretable probabilities for user behaviors.

- Model training score: 60.37%.
- Model testing score: 60.48%.
- Model Accuracy: 60%.

Logistic regression is not a good model for this because the data is not linearly separable, and the target variable is a continuous variable.

Random Forest

Random Forest is an ensemble learning technique, will be employed to capture complex interactions among features and predict user preferences and purchasing decisions with improved accuracy and robustness.

- Model Accuracy: 66.60%.

ELM Regression

The Extreme Learning Machine (ELM) Regression model, known for its computational efficiency and fast learning speed, will enable rapid prediction of user preferences and product choices based on browsing history and session attributes.

- Model Accuracy: 38.15%.
- Model Accuracy after tuning hyperparameters: 35.38%.

Ensemble

Ensemble methods, such as bagging and boosting, was leveraged to aggregate predictions from multiple base models, enhancing predictive performance and robustness by combining the strengths of various algorithms. These machine learning models collectively empower the project to uncover intricate patterns and insights from the clickstream data, facilitating a deeper understanding of user behavior and informing strategic decision-making in the e-commerce domain.

- Model Accuracy: 79.95%.
- Model Accuracy after tuning hyperparameters: 83.60%.

XGBoost

XGBoost, an advanced implementation of gradient boosting, was integrated into the project's analytical framework. XGBoost stands out for its scalability, speed, and performance, making it well-suited for handling large-scale datasets and complex feature interactions. By employing an ensemble of decision trees trained sequentially, XGBoost iteratively minimizes a predefined objective function, enhancing model accuracy and generalizability. In the context of this project, XGBoost was utilized to predict user preferences, behavior patterns, and purchasing decisions with heightened precision and efficiency.

Feature Selection

Given the following excerpt of our data set, the goal is to predict the price people from different regions of the world are willing to pay for cloth items. After understanding the problem statement, looking at the EDA and correlation plot, we dropped the following variables from the model training.

```
# Let's drop the "session ID" becuase that would not be needed.
df.drop("session ID", axis = 1, inplace=True)

# Let's split our data into features (X) and target variable(y)

X= df.drop(["price", "year", "month", "page 2 (clothing model)"], axis = 1)
y = df["price"]
```

Result

Non-linearity:

KNN is a non-parametric algorithm, meaning it makes no assumption about the distribution of data. This makes it suitable for datasets where the decision boundary is highly non-linear, unlike logistic regression, which assumes a linear relationship between features and target.

No assumptions about data distribution:

Logistic regression assumes that the data is linearly separable, which might not always be the case. KNN, on the other hand, makes no such assumptions, making it more flexible and applicable to a wider range of datasets.

Local decision boundaries:

KNN makes predictions based on the majority class of its k-nearest neighbors. This allows it to capture local patterns in the data, which might be missed by logistic regression, especially if the decision boundary is complex and varies across the feature space.

Performance with imbalanced data:

KNN can perform well with imbalanced datasets because it does not rely on the class distribution of the data during training, unlike logistic regression which might struggle with imbalanced classes.

Interpretability:

While logistic regression provides coefficients that indicate the contribution of each feature to the prediction, KNN is less interpretable since it does not provide explicit coefficients. However,

in cases where interpretability is not the primary concern and predictive performance is key, KNN can be preferred.

Overall, KNN is chosen over logistic regression when dealing with non-linear, complex datasets where no assumptions about the data distribution can be made, and local patterns need to be captured effectively for this problem statement.

From the Exploratory Data Analysis (EDA) and Modeling, we can conclude that:

In our analysis of user behavior on our online store offering maternity clothing, we have uncovered key insights into customer preferences and purchasing patterns. We found significant differences between Polish customers and those from other European countries. Through descriptive statistics, we identified variations in trendy clothing models and colors among different customer segments.

Our machine learning models accurately predict customer behavior, enabling us to tailor personalized marketing strategies and optimize our website for improved user experience. By leveraging web mining techniques, we can enhance website design and functionality to increase consumer satisfaction and sales volume.

Moving forward, we recommend extending our analysis to different language versions of the website and exploring advanced analytical tools like Markov chains for deeper insights. These findings will inform future strategies for website optimization and personalized marketing efforts.

References

[Clickstream Data for Online Shopping - UCI Machine Learning Repository]